Tarea 7: Método de Región de Confianza

Optimización I

Maestría en Computación
Centro de Investigación en Matemáticas

Esteban Reyes Saldaña
esteban.reyes@citmat.mx

2 de abril de 2021

Resumen

En esta tarea se implementó el esquema general del método de Región de Confianza para la función de Rosembrock, la función de Wood y la función de Branin. El método de región de confianza busca un paso dentro de una bola de radio r que minimice al modelo cuadrático de una función dada. A diferencia de búsqueda en línea, este método obtiene la direccióny el tamaño de paso de manera simultánea. Se presenta a continuación una descripción general, así como el pseudocódigo de los métodos implementados. Finalmente se incluyen conclusiones observadas a partir de la experimentación.

1 Introducción

Los métodos de región de confianza, similar a los métodos de búsqueda en línea, para generar los pasos se basan en un modelo que aproxima a la función objetivo. En cada iteración, definen una región en la cual se confía que el modelo se ajusta bien a la función, a esta región se le llama **región de confianza**. El paso se calcula como un minimizador aproximado del modelo, $m_k(p)$ restringido a la región de confianza ($\mathcal{R}_c$)

$$p_k = \arg\min_p m_k(p) \text{ s.t } p \in \mathcal{R}_c$$

La **dirección de descenso** y el **tamaño de paso** se calculan al mismo tiempo, diferente a los métodos de búsqueda en línea que primero se buscan una dirección de descenso y luego el tamaño de paso. Así que la actualización queda dada por

$$x_{k+1} = x_k + p_k$$

Si el paso obtenido usando el método de región de confianza no produce un progreso en la minimización entonces el paso **no es aceptado**. Lo anterior es un indicativo de que el modelo no se ajustó bien a la función en la región de confianza y entonces se reduce el tamaño de la region de confianza y se recalcula el paso usando la nueva región de confianza.

1.1 ¿Qué se necesita?

- un modelo.
- Radio o tamaño de la región de confianza.
- Una medida para evaluar el ajuste del modelo en la región de confianza.

1.2 El Modelo

Se asumirá que el modelo m_k que será usado en la iteración x_k es cuadrático usando la expansión de Taylor de segundo orden, entonces

$$f(x_k+p) = f(x_k)+\nabla f(x_k)^T p+\frac{1}{2}p^T\nabla^2 f(x_k+tp)p$$

con $t \in (0,1)$. Si el modelo toma una aproximación del Hessiano B_k entonces

$$m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2}p^T B_k p$$

sujeto a $||p|| \leq \Delta_k$.

En general, la solución del problema puede ser muy difícil. En la práctica no se necesita resolver completamente el problema anterior y, por lo general, una solución aproximada al problema anterior es suficiente.
Un elemento importante en los métodos de región de confianza es la forma de calcular el radio Δ_k en cada iteración. Para ello se calcula la siguiente medida del ajuste

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$$

donde el numerador representa la **reducción en la función** y el denominador **la reducción en el modelo.**

Nota. La reducción del modelo siempre es positiva, dado que p_k minimiza al modelo. Entonces

1. Si $\rho_k < 0$ entonces $f(x_k) < f(x_k + p_k)$ y la función se incrementa, por lo que se debe **rechazar** el paso p_k.
2. Si $\rho_k \sim 1$ entonces el comportamiento de la función y el modelo concuerdan bastante bien en esta iteración, y es buena idea incrementar el radio de la región de confianza en la próxima iteración.
3. Si $0 \leq \rho_k < 1$ pero cercano a 1 entonces no se modifica el radio Δ_k en la próxima iteración.
4. Si $\rho_k < 0$ entonces se reduce el Δ_k en la próxima iteración, puesto que la función incrementó su valor o el ajuste del modelo no es bueno.

1.3 ¿Cómo calcular el paso?

Teorema 1.1. El vector p^* es una solución global del problema

$$\min_p m_k(p) = f(x_k) + g^T p + \frac{1}{2}p^T Bp$$

s.t. $||p|| \leq \Delta$ si y sólo si p^* es factible y existe $\lambda > 0$ y además se cumplen las condiciones

$$\begin{aligned}(B + \lambda I)p^* &= -g\\ \lambda(|p^*|| - \Delta) &= 0\\ B + \lambda I &\geq 0.\end{aligned}$$

Observaciones

- Una alternativa para aproximar la solución del problema anterior se basa en el punto de Cauchy.
- Una estrategia de aproximación es el método dogleg, el cual es una aproximación cuando B_k es definida positiva.

2 Método

2.1 Punto de Cauchy

El **punto de Cauchy** es el minimizador del modelo m_k a lo largo de la dirección del máximo descenso de la función, i.e., $-\nabla f(x_k)$, sujeto a la región de confianza.

ara hallar el paso, se resuelve el problema de opimización con restricciones

$$p_k^* = \arg\min_p m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2}p^T B_k p,$$

st. $||p|| \leq \Delta_k$ donde Δ_k es el radio de la región de confianza. El Punto de Cauchy, denotado como p_k^C, nos permite cuantificar el suficiente descenso del modelo.

1. Encontrar el punto p_k^S que resuelva la versión lineal
$$p_k^S = \arg\min_p f(x_k) + \nabla f(x_k)^T p$$
s.t. $||p|| \leq \Delta_k$.

2. Encontrar el parámetro τ_k que minimiza $m_k(\tau_k p_k^S)$ en la región de confianza, i.e.,
$$\tau_k = \arg\min_{\tau \geq 0} m_k(\tau p_k^S)$$
s.t. $||\tau p_k^S|| \leq \Delta_k$.

3. Calcular el punto de Cauchy haciendo
$$p_k^C = \tau_k p_k^S$$

2.2 Resolver el problema Lineal

Encontrar el punto p_k^S que resuelva la versión lineal
$$p_k^S = \arg\min_p f(x_k) + \nabla f(x_k)^T p$$
s.t. $||p|| \leq \Delta_k$.

Sabemos que la función decrece a lo largo de $-\nabla f(x_k)^T$ luego $p_k^S = -\lambda \nabla f(x_k)$ con $\lambda > 0$. Como $||p_k^S|| \leq \Delta_k$ entonces $\lambda \leq \dfrac{\Delta_k}{||\nabla f(x_k)||}$. Así que el **máximo descenso** se obtiene para $\lambda = \dfrac{\Delta_k}{||\nabla f(x_k)||}$, por lo que

$$p_k^S = -\frac{\Delta_k}{||\nabla f(x_k)||} \nabla f(x_k). \qquad (1)$$

2.3 Encontrar τ

Encontrar el parámetro τ_k que minimiza $m_k(\tau_k p_k^S)$ en la región de confianza, i.e.,
$$\tau_k = \arg\min_{\tau \geq 0} m_k(\tau p_k^S)$$
s.t. $||\tau p_k^S|| \leq \Delta_k$.

- Si $\nabla f(x_k)^T B_k \nabla f(x_k) \leq 0$ entonces $m_k(\tau_k p_k^S)$ decrece a lo largo de p_k^S, i.e., del $-\nabla f(x_k)$ y se toma a τ como el mayor valor posible, es decir, $\tau = 1$.
- Si $\nabla f(x_k)^T B_k \nabla f(x_k) > 0$ entonces $m_k(\tau_k p_k^S)$ es una función cuadrática convexa en τ. Si el mínimo se alcanza en el interior de la región de confianza entonces
$$\tau = \frac{||\nabla f(x_k)||^3}{\nabla f(x_k)^T B_k \nabla f(x_k)}.$$
En caso contrario la solución está en la frontera, $\tau = 1$ similar al caso anterior.

2.4 Dogleg

El Paso de Cauchy produce un suficiente descenso en el modelo $m_k(\cdot)$ lo que permite obtener convergencia global. Teniendo en cuenta que usar el Paso de Cauchy es equivalente a usar el Método del Máximo descenso con un tamaño de paso particular, el método de Dogleg utiliza la información dada por el Hessiano para mejorar la solución.
Una estrategia simple es calcular el paso completo
$$p_k^B = -B_k^{-1} \nabla f_k$$
siempre que B_k sea definida positiva.
El método de Dogleg minimiza el modelo cuadrático sin restricciones a lo largo del gradiente

$$p_k^U = \alpha \nabla f_k$$

minimiza el modelo cuadrático sin restricciones si B_k es positiva definida

$$p_k^B = -B_k^{-1} \nabla f_k$$

Luego calcula el tamaño de paso
$$p_k = F(p_k^U, p_k^B)$$
i.e., el tamaño de paso es una función que depende del paso completo y de la dirección de máximo descenso.

2.5 Minimizar el Modelo cuadrático

Si $p_k^U = \alpha \nabla f_k$, se busca llahar el tamaño de paso α del problema cuadrático sin restricciones. De donde

$$\alpha^* = -\frac{\nabla^T f_k \nabla f_k}{\nabla^T f_k B_k \nabla f_k}$$

Así que

$$\begin{aligned} p_k^U &= -\frac{\nabla^T f_k \nabla f_k}{\nabla^T f_k B_k \nabla f_k} \nabla f_k \qquad (2) \\ p_k^B &= -B_k^{-1} \nabla f_k. \qquad (3) \end{aligned}$$

Ahora el problema consiste en contrar el **paso óptimo** que minimiza el problema cuadrático con restricciones en la trayectoria Dogleg, es decir, en

$$\hat{p}(\tau) = \begin{cases} \tau p_k^U & \text{si } 0 \leq \tau \leq 1 \\ p_k^U + (\tau - 1)(p_k^B - p_k^U) & \text{si } 1 \leq \tau \leq 2. \end{cases} \qquad (4)$$

Se sabe que

- $Si ||p_k^B|| \leq \Delta_k$ entonces el tamaño de paso óptimo es
  $$p_k = p_k^B$$
  puesto que $m(\hat{p}(\tau))$ decrece a lo largo del camino Dogleg.
- En otro caso, hay que hallar el intercepto entre la trayectoria de Dogleg y la región de confianza.
  - Si $||p_k^U|| \geq \Delta_k$ entonces $p_k = p_k^C$.
  - De lo contrario se tiene que resolver la siguiente ecuación para τ
    $$||p_k^U + (\tau - 1)(p_k^B - p_k^U)||^2 = \Delta^2$$

Finalmente

$$p_k = \begin{cases} \tau^* p_k^U & \text{si } 0 \leq \tau^* \leq 1 \\ p_k^U + (\tau^* - 1)(p_k^B - p_k^U) & \text{si } 1 \leq \tau^* \leq 2. \end{cases} \qquad (5)$$

2.6 Pseudocódigo

2.6.1 Región de Confianza

Input: punto inicial x_0, $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$ y $\eta \in [0, \eta_1]$
Input: $\eta_1, \eta_2, \hat{\eta}_1, \hat{\eta}_2$
Obtener solución aproximada del modelo cuadrático con restricciones
Se calcula medida de confianza ρ_k
if $\rho_k < \eta_1$ entonces $\Delta_{k+1} = \hat{\eta}_1 \Delta_k$ **then**
else
  if $\rho_k > \eta_2$ y $||\rho_k|| = \Delta_k$ **then**
    $\Delta_{k+1} = \text{mín}\{\hat{\eta}_2 \Delta_k, \hat{\Delta}\}$
  else
    $\Delta_{k+1} = \Delta_k$
  end if
end if
if $\rho_k > \eta$ entonces $x_{k+1} = x_k + p_k$ **then**
else
  $x_{k+1} = x_k$
end if

2.6.2 Paso de Cauchy

Input: gradiente g_k, hessiano b_k y radio de confianza Δ_k
Output: paso p_k
$p_k^S = -\frac{\Delta_k}{||\nabla f(x_k)||} \nabla f(x_k)$
if $g_k^T b_k g_k \leq 0$ **then**
  $\tau_k = 1$
else
  $\tau_k = -\text{mín}\left(1, \frac{||g_k||^3}{\Delta_k g_k^T b_k g_k}\right)$
end if
$p_k = \tau_k p_k^S$

2.6.3 Paso Dogleg

Input: gradiente g_k, hessiano b_k y radio de confianza Δ_k
Output: paso p_k
$p_k^U = -\frac{g_k^T g_k}{g_k^T b_k g_k} g_k$
$p_k^B = -b_k^{-1} g_k$
if $||p_k^B|| \leq \Delta_k$ **then**

$p_k = p_k^B$
else if $p_k^U \geq \Delta_k$ **then**
$p_k = -\dfrac{\Delta_k}{||p_k^U||} p_k^U$
else
$d_f = p_k^B - p_k^U$
$a = d_f^T d_f$
$b = 2 * p_k^{BT} d_f$
$c = p_k^{UT} p_k^U - \Delta_k^2$
$\tau_k = 1 + \dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$
if $\tau_k \leq 1$ **then**
$p_k = \tau_k p_k^U$
else
$p_k = p_k^U + (\tau_k - 1) d_f$
end if
end if

2.7 Función de Rosembrock

$$f(x) = \sum_{i=1}^{N-1} [100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2]$$

donde $x = [x_1, \ldots, x_N]^T \in \mathbb{R}^N$.

Gradiente. Primero notemos que para $2 \leq j \leq N-1$ se puede reescribir $f(x)$ como

$$\begin{aligned} f(x) &= \sum_{i=1}^{j-2} [100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2] \\ &\quad + [100(x_j - x_{j-1}^2)^2 + (1 - x_{j-1})^2] \\ &\quad [100(x_{j+1} - x_j^2)^2 + (1 - x_j)^2] \\ &\quad + \sum_{i=j+1}^{N-1} [100(x_{i-1} - x_i^2)^2 - (1 - x_i)^2]. \end{aligned}$$

Notemos que el primer y útlimo término son constantes con respecto a x_j, entonces se anularán al calcular la correspondiente derivada parcial.
Para D_1 se tiene que

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x) &= 100(2)(x_2 - x_1^2)(-2x_1) - 2(1 - x_1) \\ &= -400(x_2 - x_1^2)x_1 - 2(1 - x_1). \end{aligned}$$

Para $D_j \in \{2, \ldots, N-1\}$,

$$\begin{aligned} \frac{\partial}{\partial x_j} f(x) &= [200(x_j - x_{j-1}^2)] + [200(x_{j+1} - x_j^2)(-2x_j)] \\ &\quad -[2(1 - x_j)] \\ &= 200(x_j - x_{j-1}^2) - 400(x_{j+1} - x_j^2)x_j \\ &\quad -2(1 - x_j) \end{aligned}$$

Para D_N,

$$\frac{\partial}{\partial x_N} f(x) = 200(x_N - x_{N-1}^2).$$

De lo anterior tenemos que

$$\nabla f(x) = \begin{bmatrix} -400(x_2 - x_1^2)x_1 - 2(1 - x_1) \\ \vdots \\ 200(x_j - x_{j-1}^2) - 400(x_{j+1} - x_j^2)x_j - 2(1 - x_j) \\ \vdots \\ 200(x_N - x_{N-1}^2). \end{bmatrix}$$

Hessiano. Para calcular el Hessiano, se debe considerar como casos particulares, aquellos que se apliquen a $\dfrac{\partial f(x)(x)}{\partial x_1}$ y $\dfrac{\partial f(x)(x)}{\partial x_N}$ (como se observó en el gradiente).
Para $D_{x_k x_1}$,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_1} = \begin{cases} 1200x_1^2 - 400x_2 + 2 & si \quad k = 1 \\ -400x_1 & si \quad k = 2 \\ 0 & e.o.c \end{cases}$$

Para $D_{x_k x_j}$ con $j \in \{2, \ldots, N-1\}$,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_j} = \begin{cases} 1200x_j^2 - 400x_2 + 2 & si \quad k = j \\ -400x_j & si \quad k = j+1 \\ -400x_{j-1} & si \quad k = j-1 \\ 0 & e.o.c \end{cases}$$

Para $D_{x_k x_N}$,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_1} = \begin{cases} -400x_{N-1} & si \quad k = N-1 \\ 200 & si \quad k = N \\ 0 & e.o.c \end{cases}$$

Así que el Hessiano tiene la forma

$$H(x) = \begin{pmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 & \ldots & 0 \\ -400x_1 & 1200x_2^2 - 400x_3 + 202 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & -400x_{N-1} & 200 \end{pmatrix}.$$

2.8 Función de Wood

Para $n = 4$ esta función está dada por

$$\begin{aligned} f(x) &= 100(x_1^2 - x_2)^2 + (x_1 - 1)^2 \\ &+ (x_3 - 1)^2 + 90(x_3^2 - x_4)^2 \\ &+ 10{,}1[(x_2 - 1)^2 + (x_4 - 1)^2] \\ &+ 19{,}8(x_2 - 1)(x_4 - 1). \end{aligned}$$

Gradiente. Derivando respecto a cada entrada obtenemos

$$\begin{aligned} \frac{\partial f(x)}{\partial x_1} &= 400(x_1^2 - x_2)x_1 + 2(x_1 - 1) \\ \frac{\partial f(x)}{\partial x_2} &= -200(x_1^2 - x_2) + 20{,}2(x_2 - 1) \\ &+ 19{,}8(x_4 - 1) \\ \frac{\partial f(x)}{\partial x_3} &= 2(x_3 - 1) + 360(x_3^2 - x_4)x_3 \\ \frac{\partial f(x)}{\partial x_4} &= -180(x_3^2 - x_4) + 20{,}2(x_4 - 1) \\ &+ 19{,}8(x_2 - 1). \end{aligned}$$

Hessiano

$$H_f(x) = \begin{bmatrix} 1200x_1^2 - 400x_1 + 2 & -400x_1 & 0 & 0 \\ -400x_1 & 220{,}2 & 0 & 19{,}8 \\ 0 & 0 & 1080x_3^2 - 360x_4 + 2 & -360x_3 \\ 0 & 19{,}8 & -360x_3 & 2020{,}2. \end{bmatrix}$$

2.9 Función de Branin

Para $n = 2$ esta función está dada por

$$\begin{aligned} f(x) &= a(x_2 - bx_1^2 + cx_1 - r)^2 \\ &+ s(1 - t)\cos(x_1) + s. \end{aligned}$$

Gradiente. Derivando respecto a cada entrada obtenemos

$$\begin{aligned} \frac{\partial f(x)}{\partial x_1} &= 2a(x_2 - bx_1 + cx_1 - r)(-2bx_1 + c) \\ &- s(1 - t)sen(x_1) \\ \frac{\partial f(x)}{\partial x_2} &= 2a(x_2 - bx_1 + cx_1 - r) \end{aligned}$$

Hessiano

$$H_f(x) = \begin{bmatrix} \begin{matrix} 4ab(x_2 - bx_1^2 + cx_1 - r) \\ +2a(-2bx_1 + c)^2 - s(1 - t)cos(x_1) \end{matrix} & 2a(-2bx_1 + c) \\ 2a(-2bx_1 + c) & 2a \end{bmatrix}$$

3 Resultados

Se eligieron los parámetros

Parámetro	max_{iter}	τ_x	τ_f	τ_{grad}
Valor	10000	10^{-12}	10^{-12}	10^{-12}

Para la Región de Confianza se usaron los parámetros

Parámetro	Valor
Δ_k	0.1
Δ_{max}	0.2
η	0.1
η_1	0.25
η_2	0.75
$\hat{\eta}_1$	0.25
$\hat{\eta}_2$	2

3.1 Función de Branin

Para esta función se usaron los valores

Parámetro	Valor
a	1
b	$\frac{5{,}1}{4\pi^2}$
c	$\frac{5}{\pi}$
r	6
s	10
t	$\frac{1}{8\pi}$

El experimento se repitió 30 veces usando puntos iniciales aleatorios de la forma

$$x_0 = [x_0^1 + \eta_1, x_0^2 + \eta_2, \ldots, x_0^n + \eta_n]$$

donde $\eta_k \sim \mathcal{U}(-1, 1)$.

Los resultados para el promedio de tiempo fueron

Algoritmo	Dogleg	Newton-Cauchy Alterno	Newton Modificado
Rosembrock	0.2503 segundos	0.2154 segundos	2.3138 segundos
Wood	0.0294 segundos	0.0294 segundos	0.0548 segundos
Branin	0.0074 segundos	0.006 segundos	0.0630 segundos

Los resultados para el promedio de iteraciones fueron

Algoritmo	Dogleg	Newton-Cauchy Alterno	Newton Modificado
Rosembrock	89.50	83.53	447.00
Wood	117.96	126.36	163.16
Branin	28.60	28.03	358.80

4 Conclusiones

4.1 Dogleg

El paso de dogleg incorpora una dependencia al paso de Cauchy del Hessiano o paroximación del Hessiano B_k. Minimiza el modelo cuadrático sin restricciones a lo largo del gradiente y usando el modelo cuadrático sin restricciones si B_k es positiva definida. Luego calcula el paso p_k.

Para las funciones utilizadas se observó que la convergencia no siempre es global, esto se debe al radio de la región de confianza. Podría tomarse un Δ_k más grande que el que se usó para las pruebas y revisar si se llega al óptimo global. Sin embargo, mostró mayor eficiencia comparado con el método de Newton Modificado.

Se observó que la convergencia depende fuertemente del punto inicial y que en casi todas las iteraciones se llega al óptimo global. Excepto en la función de Rosembrock, justo como se observó en tareas pasadas.

4.2 Newton-Cauchy Alternado

Este método, a diferencia de las búsquedas en línea, calcula la dirección y el tamaño de paso simultáneamente. El radio de la región de confianza se incrementa solo si concuerda con el modelo de la función y se reduce sólo si la función incrementa su valor.

En la experimentación se observó que el costo computacional es bajo respecto al tiempo y las iteraciones comparado con el método de Newton Modificado. Además se observó que la convergencia fue global para las funciones probadas excepto para la función de Rosembrock.

En general, los métodos de región de confianza mostraron un menor tiempo de ejecución y un menor número de iteraciones comparados con Newton Modificado.