Aspectos matemáticos y computacionales de los métodos de optimización de descenso coordinado

Juan Pablo Soto Quirós
jusoto@tec.ac.cr

Escuela de Matemática
Instituto Tecnológico de Costa Rica

12 de agosto del 2019

- La presente charla explicará algunos aspectos matemáticos y computacionales sobre uno conjunto de métodos iterativos para resolver problemas de minimización para funciones en varias variables, usando la técnica de **descenso coordinado**.
- En especial, daremos énfasis en tres métodos basados en el descenso coordenado:
  - Optimización Alternada
  - Mejora Máxima por Bloques
  - Proyección del Gradiente de Coordenadas por Bloques
- Al final, se mostrará una aplicación en un problema de optimización de matrices con rango restringido, utilizando la norma de Frobenius

- La presente charla explicará algunos aspectos matemáticos y computacionales sobre uno conjunto de métodos iterativos para resolver problemas de minimización para funciones en varias variables, usando la técnica de **descenso coordinado**.

- En especial, daremos énfasis en tres métodos basados en el descenso coordenado:
  - Optimización Alternada
  - Mejora Máxima por Bloques
  - Proyección del Gradiente de Coordenadas por Bloques

- Al final, se mostrará una aplicación en un problema de optimización de matrices con rango restringido, utilizando la norma de Frobenius

Función multivariable

Una función escalar de n variables $f : \mathbb{R}^n \to \mathbb{R}$, asigna a cada punto $(x_1, ..., x_n) \in \mathbb{R}^n$, un único número real denotado con $f(x_1, ..., x_n) \in \mathbb{R}$.

Problema

Sea $\mathbf{x} = (x_1, ..., x_n)$. El objetivo de esta presentación es explicar un conjunto de métodos iterativos que permitirán dar una solución al problema de minimización

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

- En resumen, los pasos para resolver el problema

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

en un curso de cáluclo en varias variables son los siguientes:

1. Calcular los puntos críticos (Gradiente).
2. Calcular el Hessiano Orlado (Hessiano Orlado).
3. Evaluar puntos críticos en el Hessiano Orlado.
4. Calcular determinantes de las submatrices principales.
5. Interpretar resultado.

- La solución del problema $\min\limits_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ se realizará a través de métodos iterativos.
- Es decir, dado un valor inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$, cada método iterativo generará una sucesión de puntos

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(k)}, ...\},$$

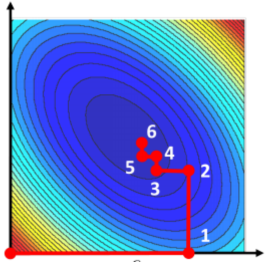
donde $\mathbf{x}^{(j)} \in \mathbb{R}^n$, para todo $j = 1, 2, ...$

- La sucesión de puntos puede tener tres criterios de convergencia:
  1. La sucesión converge a la solución del problema.
  2. La sucesión converge a a un punto que no es solución del problema.
  3. La sucesión no converge.

- Los métodos de descenso coordinado son un conjunto de algoritmo de optimización que minimiza sucesivamente a lo largo de las direcciones de coordenadas para encontrar el mínimo de una función.
- En cada iteración, el algoritmo determina una coordenada a través de una regla de selección de coordenadas
- Luego minimiza de manera exacta mientras fija todas las demás coordenadas o bloques de coordenadas.

- Sea $f(\mathbf{x})$ definida como $f:\mathbb{R}^n \to \mathbb{R}$ y sea $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$ un valor inicial.
- Sea $j = 1, 2, \dots, n$. La idea general de los métodos de descenso coordinado consiste en los siguientes pasos
  1. Definir una función de variable real $f_{x_j}:\mathbb{R} \to \mathbb{R}$, de la forma $f_{x_j}(x_j)$.
  2. Calcular el mínimo de la función f_{x_j}.
  3. Basado en algún criterio, se obtiene $x_j^{(1)}$.
- El proceso se repito para generar una sucesión $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(k)}, \dots\}$.

- Daremos énfasis en tres métodos basados en el descenso coordenado:
  - Optimización Alternada
  - Mejora Máxima por Bloques
  - Proyección del Gradiente de Coordenadas por Bloques



- El método de Optimización Alternativa (OA) actualiza solo una de las variables mientras las otras variables son "fijadas". El método OA también es conocido con el nombre de **descenso coordinado por bloques**.
- Existen varias reglas para seleccionar cual variable es actualizadas. Las tres reglas más utilizadas son:
  1. **Regla de Jacobian:** en cada iteración k, las n variables son actualizadas, utilizando exactamente los valores de la iteración anterior $k-1$.

$$x_j^{(k)} \in \arg\min_{x_j \in \mathbb{R}} f\left(x_1^{(k-1)}, \ldots, x_{j-1}^{(k-1)}, x_j, x_{j+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right),$$

para $j = 1, \ldots, p$. Esta actualización se puede hacer en paralelo.

2. **Reglas de Gauss-Seidel:** en cada iteración k, las n variables son actualizadas, utilizando exactamente los valores de la iteración anterior $k-1$, y los valores de la iteración k calculados anteriormente.

$$x_j^{(k)} \in \arg\min_{x_j \in \mathbb{R}} f\left(x_1^{(k)}, \ldots, x_{j-1}^{(k)}, x_j, x_{j+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right),$$

para $j = 1, \ldots, n$.

3. **Reglas Aleatorizada:** es una variación de la regla de Gauss-Seidel, pero la selección de la variable se hace de manera aleatoria, sin seguir un orden establecido.

Observación: De las 3 reglas mencionadas, la regla de Gauss-Seidel es la más utilizada.

Algorithm 1 Método de Optimización Alternada

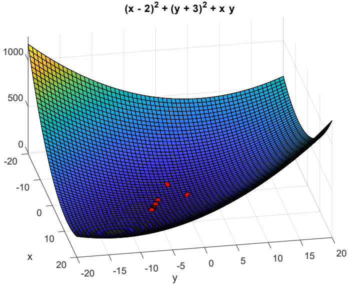
Requires: $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in \mathbb{R}^n$
Returns: $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)}) \in \mathbb{R}^n$

```
for k = 1, 2, ... do
    for j = 1, ..., n do
        x_j^(k) = Usando la regla de Jacobi o Gauss-Seidel
    end
    if el criterio de parada se cumple then
        return x^(k) = (x_1^(k), ..., x_n^(k))
    end
end
```

Ejemplo 1

Considere la función $f(x, y) = (x-2)^2 + (y+3)^2 + xy$. Aplicando el método OA con la regla de Gauss-Seidel y $\mathbf{x}^{(0)} = (1, 1)$, obtenemos los siguientes resultados.

k	$\mathbf{x}^{(k)}$	$f(\mathbf{x}^{(k)})$
0	$(1, 1)$	18
1	$(1.5, -3.75)$	4.8125
2	$(3.875, -4.9375)$	11.8633
3	$(4.4688, -5.2344)$	12.3040
$\vdots$	$\vdots$	$\vdots$
9	$(4.666, -5.333)$	12.3333



Ejemplo 2

Considere la función $f(x, y, z) = x^3 + y^3 + z^3 - 2xy - 2xz - 2yz$. Aplicando el método OA con la regla de Gauss-Seidel y $\mathbf{x}^{(0)} = (1, 1, 1)$, obtenemos los siguientes resultados.

k	$\mathbf{x}^{(k)}$	$f(\mathbf{x}^{(k)})$
0	$(1, 1, 1)$	3
1	$(1.1547, 1.1985, 1.2525)$	3.4366
2	$(1.2641, 1.2953, 1.3062)$	3.5392
$\vdots$	$\vdots$	$\vdots$
8	$(1.333, 1.333, 1.333)$	3.5556

Observaciones:

- El método converge a $(1.333, 1.333, 1.333)$, el cual no es un mínimo absoluto, sino un mínimo local.
- Si el valor inicial cambia a $\mathbf{x}^{(0)} = (-1, -1, -1)$, el método diverge.

- Lo anterior nos pone a pensar en dos preguntas:

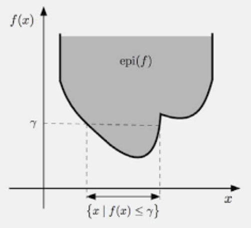
  1. **¿Cuándo el método AO converge?**
  2. **¿Si converge, a que valor converge?**

Algunas definiciones

Sea $f : \mathbb{R}^n \to \mathbb{R}$.

- El **epigrafo** de f se define como el conjunto

$$\text{epi}(f) = \{(\mathbf{x}, \gamma) : f(\mathbf{x}) \leqslant \gamma, \forall \mathbf{x} \in \mathbb{R}^n, \gamma \in \mathbb{R}\}.$$



- Una función f es **cerrada** si su epigrafo es cerrado.

Algunas definiciones (continuación)

- f es una función **apropiada** si no existe ningún punto $\mathbf{x}$ tal que $f(\mathbf{x}) = -\infty$ y cuyo dominio no es vacío.
- Un punto $\mathbf{x}^* \in \mathbb{R}^n$ se llama **mínimo coordenado** si

$$f(\mathbf{x}^*) \leqslant f(\mathbf{x}^* + \alpha \cdot \mathbf{e}_j),$$

para $j = 1, \ldots, n$ y $\alpha \in \mathbb{R}$. Aquí, $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ es la base canónica en $\mathbb{R}^n$.

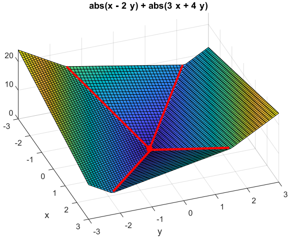


Figura: Gráfica de la función $f(x, y) = |x - 2y| + |3x + 4y|$.

Teorema de Convergencia del Método AO

Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función apropiada y cerrada, la cual es continua en su dominio. Si se cumple que

1. el problema $\min\limits_{x_j \in \mathbb{R}} f\left(x_1^{(k)}, \ldots, x_{j-1}^{(k)}, x_j, x_{j+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right)$ tiene una solución única, para todo $j = 1, \ldots, n$
2. para cualquier $\mathbf{x}^{(0)}$, el conjunto $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leqslant f(\mathbf{x}^{(0)})\}$ es un conjunto acotado

entonces la sucesión generada por el método AO converge a un mínimo coordenado.

Nota: Si $f(\mathbf{x}) = g(\mathbf{x}) + \sum\limits_{i=1}^{n} h_i(x_i)$, donde g es una función convexa y diferenciable, y cada h_i es convexa (no necesariamente diferenciable), entonces un mínimo coordenado de f es también un mínimo global de f.

- Otro método iterativo de optimización por bloques es el método conocido como **mejora máxima por bloques** (MMB).
- El método MMB actualiza cada variable usando un enfoque tipo *greddy*.
- El método MMB actualiza el bloque de variables correspondiente al bloque de mejora máxima usando la regla de Jacobi, pero compara el error de cada una de las actualizaciones.
- Al final, la actualización con el error mínimo es el nuevo elemento de la sucesión

Algorithm 2 Método de Mejora Máxima por Bloques

Requires: $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in \mathbb{R}^n$
Returns: $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)}) \in \mathbb{R}^n$

for $k = 1, 2, ...$ **do**
  for $j = 1, ..., n$ **do**
    $\widetilde{x}_j^{(k)} =$ Usando la regla de Jacobi
    $e_j^{(k)} = f(x_1^{(k-1)}, \ldots, x_{j-1}^{(k-1)}, \widetilde{x}_j^{(k)}, x_{j+1}^{(k-1)}, \ldots, x_n^{(k-1)})$
  end
  $i = \arg\min_{j=1,\ldots,p} \left\{ e_j^{(k)} \right\}$
  for $j = 1, ..., n$ **do**
    if $j = i$ **then**
      $x_j^{(k)} = \widetilde{x}_j^{(k)}$
    else
      $x_j^{(k)} = x_j^{(k-1)}$
    end
  end
  if el criterio de parada se cumple **then**
    return $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$
  end
end

Ejemplo 2

Considere la función

$$f(x, y, z) = (x - 2)^2 + (y + 3)^2 + (x + y + z)^2.$$

Aplicando el método MMB con un valor inicial $\mathbf{x}^{(0)} = (1, 1, 1)$, obtenemos los siguientes resultados.

k	$\mathbf{x}^{(k)}$	$f(\mathbf{x}^{(k)})$
0	$(1, 1, 1)$	30
1	$(1, -2.5, 1)$	5.5
2	$(1, -2.5, 0.25)$	4.375
3	$(1.75, -2.5, 0.25)$	2.125
4	$1.75, -2.5, (-0.125)$	1.8438
$\vdots$	$\vdots$	$\vdots$
27	$(2.5, -2.5, -0.5)$	1

- Una desventaja del método MMB es el exceso de cálculo, lo cual hace que el método converja más lento, comparado con el método OA.
- Sin embargo, esta desventaja se puede reducir si se implementa en paralelo.
- Otra ventaja es que este método es menos restrictivo, en términos de convergencia.

Convergencia del Método MMB

Si el conjunto $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leqslant f(\mathbf{x}^{(0)})\}$ es un conjunto acotado, entonces la sucesión generada por el método MMB converge a un mínimo coordenado de f.

- El método de proyección del gradiente de coordenadas por bloques (PGCB) se basa en el método del descenso del gradiente.
- El método PGCB consiste en realizar un paso de proyección de gradiente con respecto a variable, tomado en un orden cíclico.
- Para implementar este método, se necesita conocer dos conceptos: Gradiente de una función y constante de Lipschitz.

Definiciones

- El **gradiente** de una función f se define como el vector

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{dx_1}, \ldots, \frac{\partial f(\mathbf{x})}{dx_n} \right).$$

- La **constante de Lipschitz** es el menor número real L que satisface

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leqslant L\|\mathbf{x} - \mathbf{y}\|,$$

para cada $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Algorithm 2 Método de proyección del gradiente de coordenadas por bloques

Requires: $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in \mathbb{R}^n$

Returns: $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)}) \in \mathbb{R}^n$

for $k = 1, 2, \ldots$ **do**

  $\mathbf{x}_{[0]}^{(k-1)} = \mathbf{x}^{(k-1)}$

  for $j = 1, \ldots, n$ **do**

    $\mathbf{x}_{[j]}^{(k-1)} = \mathbf{x}_{[j-1]}^{(k-1)} - \dfrac{1}{L} e_j e_j^T \nabla f(\mathbf{x}_{[j-1]}^{(k-1)})$

  end

  $\mathbf{x}^{(k)} = \mathbf{x}_{[p]}^{(k-1)}$

  if stopping criterion is satisfied **then**

    return $x^{(k)} = (x_1^{(k)}, \ldots, x_p^{(k)})$

  end

end

Convergencia del método PGCB

Sea f una función continua convexa con constante de Lipschitz, tal que exista un mínimo global $\mathbf{x}^*$. Entonces, la sucesión generada por el método PGCB converge a $\mathbf{x}^*$.

Observaciones:

- Los métodos explicados fueron diseñados para funciones $f : \mathbb{R}^n \to \mathbb{R}$.
- En el caso que se desea minimizar la función f bajo ciertas restricciones, el problema se interpreta de la siguiente manera: Sea $f : \mathbb{A} \to \mathbb{R}$, donde $\mathbb{A} \in \mathbb{R}^n$. Se desea resolver el siguiente problema de optimización
$$\min_{\mathbf{x}\in\mathbb{A}} f(\mathbf{x}).$$
- El método OA y PGCB incrementan las restricciones para garantizar la convergencia: necesitan que $\mathbb{A}$ sea convexo y cerrados. Además, utilizan proyecciones ortogonales al espacio $\mathbb{A}$, lo cual no siempre es fácil de obtener.
- Sin embargo, el método MMB no necesita agregar más restricciones en el caso de trabajar un problema con restricciones.

Sea $\mathbb{R}_r^{m\times n}$ el conjunto de todas las matrices de tamaño $m \times n$, cuyo rango no es más grande que r.

Problema 1

Sean $A \in \mathbb{R}^{p\times q}$, $B_j \in \mathbb{R}^{p\times m_j}$ y $C_j \in \mathbb{R}^{n_j\times q}$, para $j = 1, 2, ..., k$. El problema consiste en calcular $\widehat{X}_j \in \mathbb{R}_{r_j}^{m_j\times n_j}$ tal que

$$\left\| A - \sum_{i=1}^{k} B_j \widehat{X}_j C_j \right\|_{fr} = \min_{X_1,\ldots,X_p} \left\| A - \sum_{i=1}^{k} B_j X_j C_j \right\|_{fr}$$

sujeto a $X_1 \in \mathbb{R}_{r_1}^{m_1\times n_1}, ..., X_p \in \mathbb{R}_{r_p}^{m_p\times n_p}$. Acá, $\|\cdot\|_{fr}$ representa la norma de Frobenius.

Se sabe además lo siguiente:

$X \in \mathbb{R}_r^{m\times n}$ si y solo si existen $M \in \mathbb{R}^{m\times r}$ and $N \in \mathbb{R}^{r\times n}$ tales que $T = MN$.

Por lo tanto, el Problema 1 se puede re-escribir de la siguiente manera:

Problema 2

Sean $A \in \mathbb{R}^{p\times q}$, $B_j \in \mathbb{R}^{p\times m_j}$ y $C_j \in \mathbb{R}^{n_j\times q}$, para $j = 1, 2, ..., k$. El problema consiste en calcular $\widehat{M}_j \in \mathbb{R}^{m_j\times r_j}$ y $\hat{N}_j \in \mathbb{R}^{r_j\times n_j}$ tal que

$$\left\| A - \sum_{i=1}^{k} B_j \widehat{M}_j \hat{N}_j C_j \right\|_{fr} = \min_{M_1, N_1, ..., M_k, N_k} \left\| A - \sum_{i=1}^{k} B_j M_j N_j C_j \right\|_{fr}$$

sujeto a $M_1 \in \mathbb{R}^{m_1\times r_1}$ y $R_j \in \mathbb{R}^{r_j\times n_j}$, para todo $j = 1, 2, ..., k$.

Consideremos las funciones $f_1 : \mathbb{R}_{r_1}^{m_1 \times n_1} \times ... \times \mathbb{R}_{r_k}^{m_k \times n_k} \to \mathbb{R}$ definida por

$$f_1(X_1, ..., X_p) = \left\| A - \sum_{i=1}^{k} B_j X_j C_j \right\|_{fr}$$

y $f_2 : \mathbb{R}^{m_1 \times r_1} \times \mathbb{R}^{r_1 \times n_1} \times ... \times \mathbb{R}^{m_k \times r_k} \times \mathbb{R}^{r_k \times n_k} \to \mathbb{R}$ definida por

$$f_2(M_1, N_1, ..., M_p, N_p) = \left\| A - \sum_{i=1}^{k} B_j M_j N_j C_j \right\|_{fr}.$$

En el siguiente ejemplo, aplicaremos los método OA, BBM y PGCB, para resolver el problema. Para eso, se debe saber que la solución al problema

$$\min_{X \in \mathbb{R}_r^{m \times n}} \|A - BXC\|_{fr}$$

está dado por $X = B^\dagger \lfloor BB^\dagger AC^\dagger C \rfloor_r C^\dagger$.

Algorithm 1 AO method for solving Problem (1)

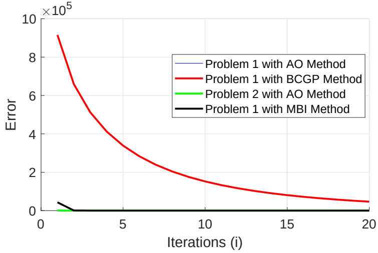
Requires: $A \in \mathbb{R}^{p\times q}$, $B_j \in \mathbb{R}^{p\times m_j}$, $C_j \in \mathbb{R}^{n_j\times q}$, $X_j^{(0)} \in \mathbb{R}_{r_j}^{m_j\times n_j}$, for $j = 1, ..., p$, and $tol \in \mathbb{R}^+$

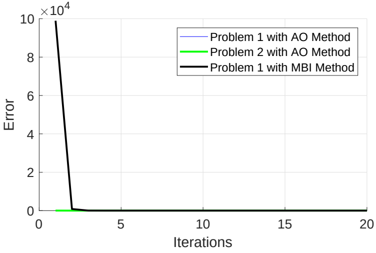
Returns: $X_j^{(i+1)} \in \mathbb{R}_{r_j}^{m_j\times n_j}$, for $j = 1, ..., p$

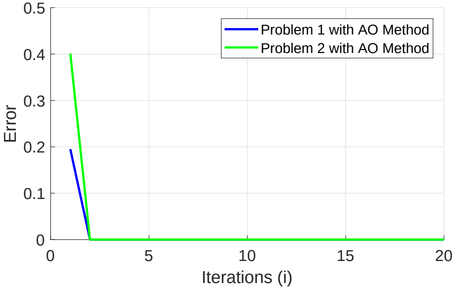
for $i = 0, 1, 2, \ldots$ **do**
    for $j = 1, ..., p$ **do**
        $A_j = A$
        for $k = 1, ..., p$ **do**
            if $k \neq j$ **then**
                $A_j = A_j - B_k X_k^{(i)} C_k$
            end
        end
        $P_B = B_j B_j^{\dagger}$ and $P_C = C_j^{\dagger} C_j$
        $X_j^{(i+1)} = B_j^{\dagger} \lfloor P_B A_j P_C \rfloor_{r_j} C_j^{\dagger}$
    end
    if $f_1(X_1^{(i+1)}, ..., X_p^{(i+1)}) < tol$ **then**
        return $X_1^{(i+1)}, ..., X_p^{(i+1)}$
    end
end

Ejemplo 1

Sean $A \in \mathbb{R}^{5\times 6}$, $B_j \in \mathbb{R}^{5\times 8}$, $C_j \in \mathbb{R}^{9\times 6}$, y $r_j = 4$, para $j = 1,2,3,4,5$. Aplicando los métodos anteriores a los Problemas 1 y 2 obtenemos la siguiente gráfica de errores





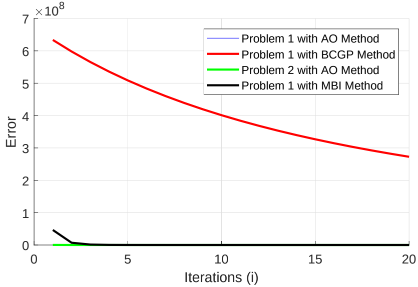


- Problema 1 - OA Método: $error = 1.9748 \times 10^{-9}$.
- Problema 1 - BCGP Método: $error = 6.1916 \times 10^{4}$.
- Problema 2 - AO Método: $error = 6.4526 \times 10^{-10}$.
- Problema 1 - MBI Método: $error = 7.8744 \times 10^{-10}$.

Observación: Después de 300 iteraciones, the método BCGP produce un error de 4.9256×10^{-4}.

Ejemplo 2

Sean $A \in \mathbb{R}^{50\times 60}$, $B_j \in \mathbb{R}^{50\times 80}$, $C_j \in \mathbb{R}^{90\times 60}$, y $r_j = 40$, para $j = 1, 2, 3, 4, 5$. Aplicando los métodos anteriores a los Problemas 1 y 2 obtenemos la siguiente gráfica de errores



- Problema 1 - OA Método: $error = 1.8235 \times 10^{-6}$.
- Problema 1 - BCGP Método: $error = 2.6819 \times 10^{8}$.
- Problema 2 - AO Método: $error = 7.6174 \times 10^{-7}$.
- Problema 1 - MBI Método: $error = 6.9444 \times 10^{-5}$.