Descenso del Gradiente

Matemática IV - 2023

El propósito de este documento se centra en la explicación y desarrollo de uno de los métodos numéricos más utilizados para la optimización de funciones. En nuestro caso buscaremos minizar una función $f(x)$ utilizando un algoritmo que nos permita aproximar de forma iterativa lo mejor posible la solución analitica verdadera. Veremos solo uno de los métodos existentes para resolver este problema de optimización, el **descenso del gradiente**. Asi como su nombre lo indica, el gradiente de la función en cuestión estará involucrado en el proceso iterativo de minimización.

Los métodos de descenso, en general, utilizan la misma idea subyacente para conseguir una aproximación de la solución óptima: Partiendo de un punto inicial $x^{(0)}$, producir una secuencia de puntos $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ tales que se minimice cada vez mas la función, es decir,

$$f(x^{(0)}) \geq f(x^{(1)}) \geq f(x^{(2)}) \geq \cdots \geq f(x^{(k)})$$

Suponiendo siempre que $x^{(0)}$ no es el punto donde la funcion consigue su valor minimo.

Para construir dicha secuencia de puntos partiendo de uno inicial debemos decidir, en cierta forma, para dónde debemos trasladarnos. Podríamos movernos sobre una recta (en $\mathbb{R}^2$ cuando la función es de una variable) o sobre un plano (en $\mathbb{R}^3$ cuando la función es de dos variables). La *dirección* que elegiremos para buscar el punto mínimo estará dada por, el ya conocido, **gradiente** de la función.

Si bien hay múltiples formas de acercarse a la explicación del metodo de descenso del gradiente, en este caso buscaremos utilizar los conceptos de linealización y gradiente y lograr una aplicación directa de ellos para conseguir una solución algorítmica (paso a paso). Comencemos interpretando nuestro problema de forma simbólica: buscamos construir una secuencia de puntos $x^{(t)}$ que, tomando su anterior $x^{(t-1)}$, lo modifique de forma tal que en cada paso nos acerquemos un poco mas a la solución mínima. Dicha modificación la realizaremos utilizando un vector, que funcionará como una brújula y se ajustará a cada paso para indicar nuestro camino hacia el "norte" (la solución óptima). Llamando Δx a dicho vector indicador, veamos que podemos expresar a cada nuevo punto como una actualización del anterior de la siguiente forma:

$$x^{(t+1)} = x^{(t)} + \eta \Delta x^{(t)} \tag{1}$$

Como sambemos, si sumamos un vector u a otro v, obtenemos otro vector s que tendrá la punta de v "trasladada" $|u|$ unidades en la dirección de u. Teniendo esto en cuenta, podemos ver que la ecuación (1) indica una traslación del punto $x^{(t)}$ en la dirección de nuestro vector "brújula" $\Delta x^{(t)}$. El factor η (eta) indica un escalamiento del vector de direccionamiento y lo llamaremos *tamaño de paso* (también conocido como *tasa de aprendizaje* en el contexto del Deep Learning). Veremos que este factor será necesario para conseguir una aproximación correcta a medida que nos acercamos a la solución verdadera. No siempre querremos que nuestro punto se actualice trasladandose la mismas unidades en cada paso, menos aun cuando nos estemos acercando al punto mínimo, donde buscaremos en particualr realizar modificaciones pequeñas, para asegurarnos precisión y no "pasarnos de largo".

Tomando (1) , construiremos tantos puntos de forma tal que las actualizaciones cercanas a la solución verdadera sean cada vez menores. Pero... ¿cuantas debemos generar?

Notemos que cuando la función tome valores muy similares en dos aproximaciones consecutivas, como trataremos con funciones continuas, sería sensato pensar que la actualización fue muy pequeña y por ende nos encontramos cerca de la solución. Esta **condición de corte** tendrá la forma:

$$|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon$$

donde ε es un valor muy pequeño, en el orden de 10^{-6} o menor (cuanto menor mejor sera la aproximación, pero mas cantidad de pasos serán requeridos).

1 Dirección de actualización

Ya que tenemos una idea del procedimiento a seguir, intentemos identificar cómo debemos elegir el vector que indicará la dirección de actualización en cada paso.

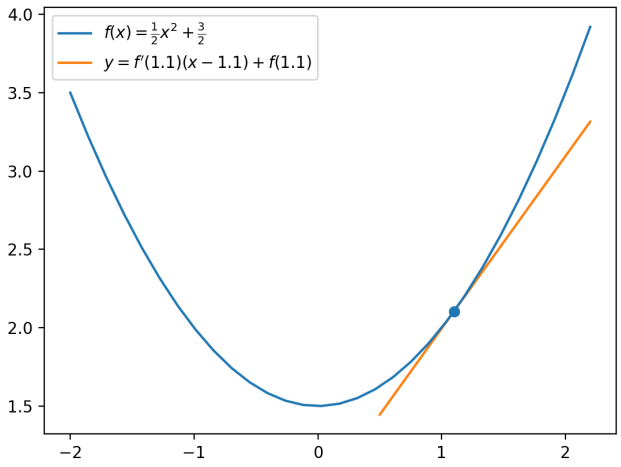
Como suele hacerse en ciencias que utilizan matemática aplicada como la ingeniería, cuando se cuenta con una función muy compleja, si ésta cumple ciertas propiedades, como la diferenciabilidad, es posible aproximarla mediante alguna función polinómica mucho más sencilla de estudiar. La más simple es la función de aproximación lineal, a la que llamamos *Linealización* y al denotabamos por L. Esta no es otra más que la aproximación de primer orden de la serie de Taylor a nuestra función en un punto dado.

Por cuestión de simplicidad, grafiquemos esta situación en $\mathbb{R}^2$: En funciones de una variable podemos calcular la recta tangente que pasa por un punto dado, que sirve como aproximación lineal de nuestra función en los alrededores del punto. Veamos un ejemplo de una función

cuadrática convexa,

$$f(x) = \frac{1}{2}x^2 + \frac{3}{2}$$

la cual ya sabemos que tiene un mínimo global (el vértice de la parábola). Si tomamos un punto inicial en el dominio $x_0 = 1.1$, podemos calular la recta tangente a la función en dicho punto como una aproximación. Gráficamente:



Si tomamos ahora a la tangente como una aproximación a la función, podríamos buscar su mínimo e interpretarlo como el mínimo de la función original $f(x)$. Esto no es del todo correcto... en primer lugar, una función lineal no tiene ningún mínimo global (si es que esta definida en un intervalo abierto) y, en segundo lugar, nada nos garantiza que el menor valor que elijamos en la recta tangente sea una buena aproximación del mínimo real. Bien sabemos que a medida que nos alejamos del punto donde se calucla la linealización, la aproximación tiende a empeorar y esto es claramente visible por medio del gráfico.

Podemos utilizar este método de aproximación lineal realizando algunas modificaciones:

1. Utilizar una función de aproximación que tenga un mínimo.

2. Que el mínimo de dicha función no este tan alejado del punto actual, caso contrario, podríamos estimar incorrectamente el valor mínimo de la función original al "pasarnos" mucho.

Luego, veamos que nosotros queremos que el paso $x^{(t+1)}$ sea igual al punto donde la función de aproximación alcanza su mínimo. Esta será calculada como una aproximación lineal de la función original $f(x)$ en el punto anterior $x^{(t)}$, de esta forma podríamos acercarnos cada vez más

al mínimo real. Luego:

$$x^{(t+1)} = \underset{\mathbf{x}}{argmin}\ L(\mathbf{x}) + \frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2 \tag{2}$$

donde $L(x)$ es la función de linealización (una recta, plano, o hiperplano) que esta definida como aproximación para el punto anterior de la iteración. Entonces, en general, y de forma vectorial, podemos definirla como:

$$L(\mathbf{x}) = (\nabla f(x^{(t)}) \cdot (\mathbf{x} - x^{(t)})) + f(x^{(t)})$$

Si $\mathbf{x}$ es un vector de n componentes (y f es una función de n variables) podemos expandir la forma anterior a:

$$L(\mathbf{x}) = L(x_1, x_2, \ldots, x_n) = \frac{\partial f(x^{(t)})}{\partial x_1}(x_1 - x_1^{(t)}) + \frac{\partial f(x^{(t)})}{\partial x_2}(x_2 - x_2^{(t)}) + \cdots + \frac{\partial f(x^{(t)})}{\partial x_n}(x_n - x_n^{(t)}) + f(x^{(t)})$$

Notemos que cuando:

- f es una función de 1 variable:

  $$L(\mathbf{x}) = L(x) = \frac{df(x^{(t)})}{dx}(x - x_t) + f(x_t)$$

  y toma la forma de una recta tangente a la función $f(x)$ en el punto x_t

- f es una función de 2 variables:

  $$L(\mathbf{x}) = L(x_1, x_2) = \frac{\partial f(x^{(t)})}{\partial x_1}(x_1 - x_1^{(t)}) + \frac{\partial f(x^{(t)})}{\partial x_2}(x_2 - x_2^{(t)}) + f(x^{(t)})$$

  y toma la forma de un plano tangente a la función $f(x_1, x_2)$ en el punto $x^{(t)} = (x_1^{(t)}, x_2^{(t)})$

Finalmente, podemos escribir a nuestra fórmula (2) de la siguiente forma generalizada:

$$x^{(t+1)} = \underset{\mathbf{x}}{argmin}\ \underbrace{(\nabla f(x^{(t)}) \cdot (\mathbf{x} - x^{(t)})) + f(x^{(t)})}_{\text{linealización}} + \underbrace{\frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2}_{\text{termino cuadratico de corrección}} \tag{3}$$

Nos preguntamos ahora... ¿Qué significa el segundo término de la ecuación anterior? Este es una corrección de la ecuación de linealización que nos permitirá conseguir una función de aproximación con un mínimo y donde se garantiza una actualización del punto actual a uno cercano, evitando "alejarnos" mucho.

Analizando la forma de este término:

$$\frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2$$

Lo primero que debemos notar es el cálculo de la distancia al cuadrado entre los dos puntos (el anterior y el nuevo)

$$||\mathbf{x} - x^{(t)}||^2 = (x_1 - x_1^{(t)})^2 + (x_2 - x_2^{(t)})^2 + \cdots + (x_n - x_n^{(t)})^2$$

con lo que conseguimos una función con términos cuadráticos. Luego notamos el parámetro $\eta \in \mathbb{R}$ que funcionará como medio de penalización para actualizaciones lejanas. Veamos que si η es un numero positivo pequeño, es decir $0 < \eta < 1$, al estar en el denominador de la fracción, esto resultará en que el término en su totalidad tienda a ser un número grande. Luego, si recordamos, queremos *minimizar* la función que contiene este sumando y por lo tanto estamos buscando que el término sea lo más pequeño posible. De esta forma, la solución tenderá a ser un punto $\mathbf{x}$ muy cercano a $x^{(t)}$, con el fin de minimizar el numerador.

Notemos entonces que, en primer lugar, al añadir un término cuadrático estamos garantizando la existencia de un mínimo (ya que la función resultante es convexa) y, en segundo, que éste valor óptimo se encuentre cerca del punto de la iteración anterior. Finalmente nos resta determinar **dónde** se encuentra el mínimo de la función de aproximación del que tanto hablamos. ¿Como lo haremos...? Pues como ya sabemos, mediante la búsqueda de puntos críticos por medio del cálulo de las derivadas parciales.

1.1 Cálculo del mínimo de la función de aproximación

Primeramente debemos calcular las derivadas parciales de la función con respecto a cada una de las coordenadas del vector $\mathbf{x}$. Si tratamos con una función original de n variables, $\mathbf{x}$ tendrá n componentes:

Llamemos $H(\mathbf{x})$ a nuestra función de aproximación en el punto t-esimo de la iteración $x^{(t)}$:

$$H(\mathbf{x}) = (\nabla f(x^{(t)}) \cdot (\mathbf{x} - x^{(t)})) + f(x^{(t)}) + \frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2$$

Debemos hallar la derivada parcial de $H(\mathbf{x})$ con respecto a una componente cualquiera x_i del vector de entrada:

$$\begin{aligned}
\frac{\partial H}{\partial x_i} &= \frac{\partial}{\partial x_i}\left[(\nabla f(x^{(t)}) \cdot (\mathbf{x} - x^{(t)})) + f(x^{(t)}) + \frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2\right] = \\
&= \underbrace{\frac{\partial}{\partial x_i}\left[(\nabla f(x^{(t)}) \cdot (\mathbf{x} - x^{(t)}))\right]}_{(1)} + \underbrace{\frac{\partial}{\partial x_i}\left[f(x^{(t)})\right]}_{(2)} + \underbrace{\frac{\partial}{\partial x_i}\left[\frac{1}{2\eta}||\mathbf{x} - x^{(t)}||^2\right]}_{(3)}
\end{aligned}$$

- ***(1):***

$$\begin{aligned}
\frac{\partial}{\partial x_i}\left[(\nabla f(x^{(t)})\cdot(\mathbf{x}-x^{(t)}))\right] &= \frac{\partial}{\partial x_i}\left[\frac{\partial f(x^{(t)})}{\partial x_1}(x_1-x_1^{(t)})+\cdots+\frac{\partial f(x^{(t)})}{\partial x_n}(x_n-x_n^{(t)})\right]= \\
&= \frac{\partial}{\partial x_i}\underbrace{\left(\frac{\partial f(x^{(t)})}{\partial x_1}(x_1-x_1^{(t)})\right)}_{constante}+\cdots+\frac{\partial}{\partial x_i}\underbrace{\left(\frac{\partial f(x^{(t)})}{\partial x_n}(x_n-x_n^{(t)})\right)}_{constante}= \\
&= 0+\cdots+\frac{\partial}{\partial x_i}\left(\underbrace{\frac{\partial f(x^{(t)})}{\partial x_i}}_{constante}(x_i-x_i^{(t)})\right)+\cdots+0= \\
&= \frac{\partial f(x^{(t)})}{\partial x_i}
\end{aligned}$$

- ***(2):***

$$\frac{\partial}{\partial x_i}\left[\underbrace{f(x^{(t)})}_{constante}\right]=0$$

- ***(3):***

$$\begin{aligned}
\frac{\partial}{\partial x_i}\left[\frac{1}{2\eta}||\mathbf{x}-x^{(t)}||^2\right] &= \frac{1}{2\eta}\frac{\partial}{\partial x_i}\left[||\mathbf{x}-x^{(t)}||^2\right]= \\
&= \frac{1}{2\eta}\frac{\partial}{\partial x_i}\left[(x_1-x_1^{(t)})^2+(x_2-x_2^{(t)})^2+\cdots+(x_n-x_n^{(t)})^2\right]= \\
&= \frac{1}{2\eta}\left[\frac{\partial}{\partial x_i}\underbrace{(x_1-x_1^{(t)})^2}_{constante}+\frac{\partial}{\partial x_i}\underbrace{(x_2-x_2^{(t)})^2}_{constante}+\cdots+\frac{\partial}{\partial x_i}\underbrace{(x_n-x_n^{(t)})^2}_{constante}\right]= \\
&= \frac{1}{2\eta}\left[0+\cdots+\frac{\partial}{\partial x_i}(x_i-x_i^{(t)})^2+\cdots+0\right]= \\
&= \frac{1}{2\eta}\left[2(x_i-x_i^{(t)})\right]= \\
&= \frac{1}{\eta}(x_i-x_i^{(t)})
\end{aligned}$$

Finalmente, reemplazando (1), (2) y (3), conseguimos:

$$\frac{\partial H}{\partial x_i} = \frac{\partial f(x^{(t)})}{\partial x_i} + \frac{1}{\eta}(x_i - x_i^{(t)})$$

Como hallamos la expresión de la derivada parcial de H para cualquier variable x_i, si igualamos cada una de ellas a 0 podríamos conseguir el punto crítico donde hallaremos el valor mínimo de H. Entonces:

$$\begin{aligned}
\frac{\partial f(x^{(t)})}{\partial x_i} + \frac{1}{\eta}(x_i - x_i^{(t)}) &= 0 \\
\frac{1}{\eta}(x_i - x_i^{(t)}) &= -\frac{\partial f(x^{(t)})}{\partial x_i} \\
x_i - x_i^{(t)} &= -\eta \frac{\partial f(x^{(t)})}{\partial x_i} \\
x_i &= x_i^{(t)} - \eta \frac{\partial f(x^{(t)})}{\partial x_i}
\end{aligned}$$

De esta forma conseguimos una forma génerica de las coordenadas del vector $\mathbf{x}$ donde H alcanza su mínimo valor. Pero... mirando detenidamente, notemos lo siguiente:

$$\begin{aligned}
x_1 &= x_1^{(t)} - \eta \frac{\partial f(x^{(t)})}{\partial x_1} \\
x_2 &= x_2^{(t)} - \eta \frac{\partial f(x^{(t)})}{\partial x_2} \\
&\vdots \\
\underbrace{x_n}_{\mathbf{x}} &= \underbrace{x_n^{(t)}}_{x^{(t)}} - \eta \underbrace{\frac{\partial f(x^{(t)})}{\partial x_n}}_{\nabla f(x^{(t)})}
\end{aligned}$$

Es decir, en forma vectorial, conseguimos que el mínimo de H se alcanza en $\mathbf{x}$, que tiene la forma:

$$\mathbf{x} = x^{(t)} - \eta \nabla f(x^{(t)})$$

1.2 Reuniendo todo

Recordando lo explicado en la ecuación (3), vemos que $\mathbf{x}$ es en realidad el punto actualizado que estabamos buscando. Por lo tanto:

$$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)})$$

Maravillosamente, hemos encontrado nuestro vector "brújula" y es nada más y nada menos que el vector gradiente de la función evaluado en el punto anterior de la iteración: $\nabla f(x^{(t)})$

¡Conseguimos finalmente lo que buscabamos en la fórmula de actualización (1)!

2 Construcción del Algoritmo

Una vez vistos los fundamentos teóricos del método, es hora de construir un paso a paso del **descenso del gradiente**. Buscamos que, a partir de un punto inicial dado x_0, se construya una secuencia de puntos nuevos $x_1, \ldots, x_n$ de forma iterativa que minimice la función a cada paso. Pasemos entonces a construir el algoritmo:

1. Seleccionar un punto inicial x_0
2. Mientras no se haya llegado a una buena aproximación de la función:
   (a) Calcular el siguiente punto de la secuencia según la función de actualización (1.2):
   $$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)})$$

Pensando más en detalle, el punto 2 de nuestro problema implica definir una condición de corte. Esta será la misma que mencionamos previamente al inicio de la sección. Generaremos una secuencia de puntos hasta que dos consecutivos den un valor en la función muy similar, es decir, hasta que se cumpla:

$$|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon$$

El parámetro ε lo llamaremos **tolerancia** y será necesario seleccionarlo de forma previa. Por lo tanto, nuestro algorimo, quedría finalemente de la forma:

1. Seleccionar un punto inicial x_0
2. Seleccionar el valor de tolerancia ε
3. Seleccionar el tamaño del paso η
4. Mientras no se cumpla la condición de corte: $|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon$
   (a) Calcular el siguiente punto de la secuencia según la función de actualización (1.2):
   $$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)})$$

3 Ejemplo de Descenso de Gradiente

Veamos a continuación un ejemplo del algoritmo paso a paso. Por simplicidad y explicabilidad, tomaremos la función de una variable que hemos presentado previamente:

$$f(x) = \frac{1}{2}x^2 + \frac{3}{2}$$

De Matemática 2, sabemos que esta función tiene un único mínimo global y es fácilmente hallado mediante el cálculo de la derivada:

$$f'(x) = \frac{d}{dx}\left(\frac{1}{2}x^2 + \frac{3}{2}\right) = x$$

Igualando $f'(x)$ a 0 podemos hallar el punto crítico que resulta ser donde la función alcanza su mínimo. Luego, tenemos que en $x = 0$ se encuentra el mínimo de la función.

Comencemos el algoritmo:

- Seleccionamos el punto inicial $x_0 = 1.1$
- Seleccionamos $\varepsilon = 10^{-2} = 0.01$ y $\eta = 0.4$
- **Iteración 1:**
  - Como es la primera iteración, calculamos el siguiente punto x_1

$$x_1 = x_0 - \eta \nabla f(x_0) = 1.1 - 0.4 f'(1.1) = 1.1 - (0.4 \,.\, 1.1) = 0.66$$

- Verificamos la condición de corte:

$$|f(x_1) - f(x_0)| = |f(0.66) - f(1.1)| = |1.718 - 2.105| = 0.387$$

Como $0.387 > 0.01$ debemos continuar el algoritmo.

- **Iteración 2:**

  - Calculamos el siguiente punto x_2

$$x_2 = x_1 - \eta \nabla f(x_1) = 0.66 - 0.4 f'(0.66) = 0.66 - (0.4 \,.\, 0.66) = 0.396$$

  - Verificamos la condición de corte:

$$|f(x_2) - f(x_1)| = |f(0.396) - f(0.66)| = |1.578 - 1.718| = 0.14$$

Como $0.14 > 0.01$ debemos continuar el algoritmo.

- **Iteración 3:**

  - Calculamos el siguiente punto x_3

$$x_3 = x_2 - \eta \nabla f(x_2) = 0.396 - 0.4 f'(0.396) = 0.396 - (0.4 \,.\, 0.396) = 0.237$$

  - Verificamos la condición de corte:

$$|f(x_3) - f(x_2)| = |f(0.237) - f(0.396)| = |1.528 - 1.578| = 0.05$$

Como $0.14 > 0.01$ debemos continuar el algoritmo.

- **Iteración 4:**

  - Calculamos el siguiente punto x_4

$$x_4 = x_3 - \eta \nabla f(x_3) = 0.237 - 0.4 f'(0.237) = 0.237 - (0.4 \,.\, 0.237) = 0.142$$

  - Verificamos la condición de corte:

$$|f(x_4) - f(x_3)| = |f(0.142) - f(0.237)| = |1.510 - 1.528| = 0.018$$

Como $0.018 > 0.01$ debemos continuar el algoritmo.

- **Iteración 5:**
  - Calculamos el siguiente punto x_4

$$x_5 = x_4 - \eta \nabla f(x_4) = 0.142 - 0.4 f'(0.142) = 0.142 - (0.4 \,.\, 0.142) = 0.085$$

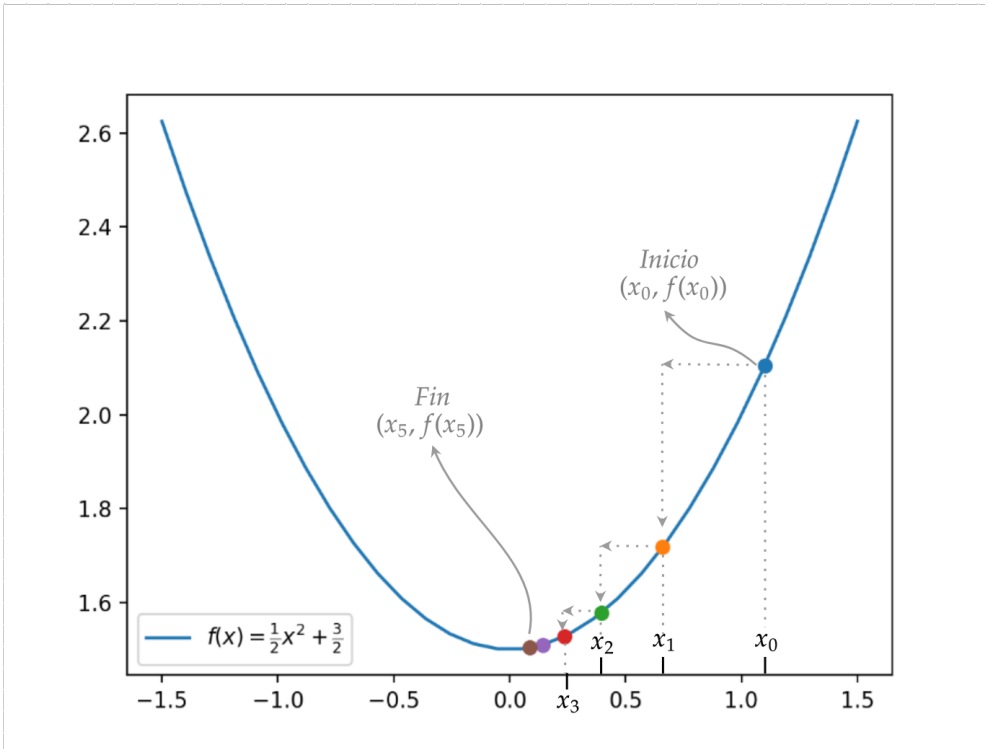
  - Verificamos la condición de corte:

$$|f(x_5) - f(x_4)| = |f(0.085) - f(0.142)| = |1.503 - 1.510| = 0.007$$

    Finalmente, como $0.007 < 0.01$ detenemos la ejecución del algoritmo.

  Por lo tanto, obtenemos que el mínimo aproximado de la función es $x^* = 0.085$ que resulta bastance cercano al verdadero $x = 0$. Cabe destacar que si eligieramos una tolerancia mucho menor, por ejemplo, en el orden de 10^{-5}, obtendríamos una aproximación mucho mas precisa.

Por último veamos de forma gráfica lo que estuvimos realizando:



iteration	prev	new	f(new)	error
1	1.10000	0.66000	1.717800	0.387200
2	0.66000	0.396000	1.578408	0.139392
3	0.396000	0.237600	1.528227	0.050181
4	0.237600	0.142560	1.510162	0.018065
5	0.142560	0.085536	1.503658	0.006503