Gradiente descendente con momento

El método de optimización a pasos utiliza el valor de la derivada de la función $df(x_0)/dx$ para actualizar el valor de x_0 con el fin de acercarse a la solución aproximada x^*. Sin embargo, en muchos escenarios este esquema de optimización puede ser muy lento, particularmente cerca extremos locales donde la derivada $df(x_0)/dx \approx 0$. En estos casos, los pasos se vuelven cada vez más pequeños y se necesitan más iteraciones para alcanzar el mínimo. El gradiente descendente con momento es un método que acelera la optimización a pasos añadiendo un término de v_t que depende de la derivada del paso anterior $df(x_{t-1})/dx$, este término adicional se conoce como *término de momento*.

$$v_t = \gamma v_{t-1} + \alpha \nabla f(x_{t-1}) \tag{1}$$

Donde usualmente $\gamma = 0.9$ y el valor inicial del término de momento es $v_0 = 0$. El proceso para hallar la solución aproximada x^* se modifica de la siguiente manera:

1. Asignar un valor inicial para x_0 y evaluar $y = f(x_0)$. El valor inicial del término de momento es $v_0 = 0$.
2. Calcular el valor de v_1 utilizando la ecuación Ecuación 1.
3. Aumentar o reducir el valor de x_0 (según el resultado del paso 2).
4. Evaluar $y = f(x_1)$.
5. Repetir los pasos 2,3 y 4 incrementando en uno el subindice de x.

El motivo de la introducción del término de momento en el proceso de optimización es otorgarle una forma de *inercia* que acumula información sobre actualizaciones anteriores para orientar el siguiente paso. La regla de actualización para el método de gradiente descendente con momento se convierte en:

$$x_t = x_{t-1} - v_t \tag{2}$$

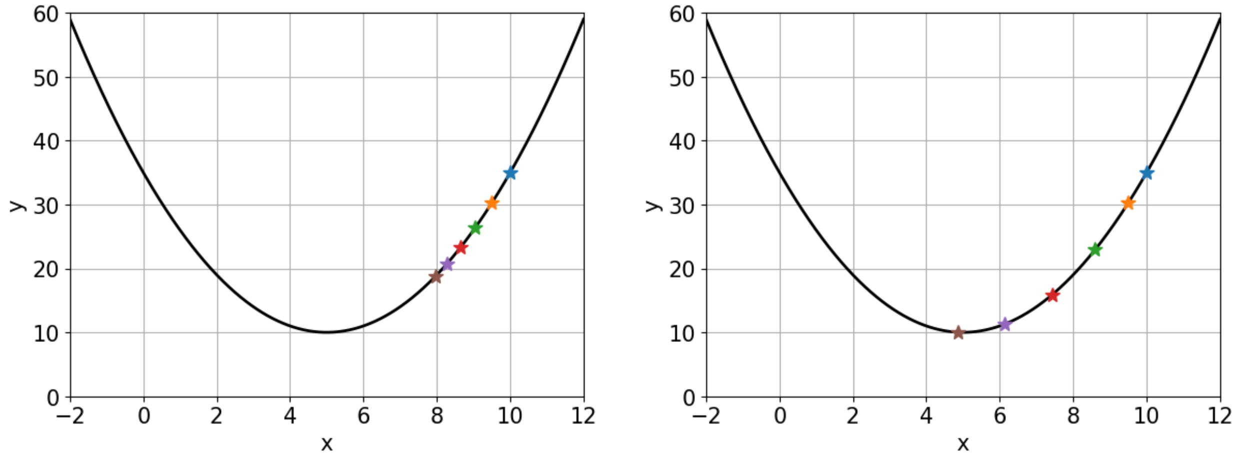
El esquema de gradiente descendente con momento permite reducir el número de iteraciones necesarias para encontrar un mínimo comparado con la optimización a pasos.

Ejemplo resuelto

Nuevamente, tómese como ejemplo la función $y = f(x) = (x-5)^2+10$ y considere un dominio $x \in [-2, 12]$. Utilizamos el esquema de gradiente descendente con momento para buscar el valor de x que minimiza la función $f(x)$. Definimos de manera arbitraria $x_0 = 10$ e inicializamos $v_0 = 0$ (Paso 1). Posteriormente, se evalúa la derivada en el punto x_0 y con este resultado se calcula el valor de v_1 (paso 2). Después se actualiza el valor de x ($x_0 \to x_1$) de acuerdo a la ecuación Ecuación 2 (paso 3). Utilizando el nuevo valor de x, se evalúa la función $y = f(x)$ (paso 4). El proceso desde el paso 2 hasta el paso 4 se repite t iteraciones hasta que se obtenga una solución aproximada x^* al problema de minimización. La Tabla 1 muestra los datos obtenidos al realizar los procedimientos de optimización a pasos y gradiente descendente con momento un total de 5 iteraciones.

método		$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
(GD)	x_0	10	9.5	9.05	8.645	8.280	7.952
(GD)	dy/dx	10	9	8.10	7.290	6.561	5.904
(GDM)	x_0	10	9.5	8.6	7.43	6.134	4.854
(GDM)	dy/dx	10	9	7.19	4.859	2.268	-0.291

Table 1: Tabla de valores de dy/dx para cada valor obtenido de x_t calculado con los métodos de gradiente descendente y gradiente descendente con momento.



(a)

(b)

Figure 1: a) Función $y = f(x) = (x-5)^2 + 10$ donde se ilustra el valor de x inicial (x_0) así como los valores de x obtenidos en cada iteración t utilizando el método de optimización gradiente descendente. b) Función $y = f(x) = (x-5)^2 + 10$ donde se ilustra el valor de x inicial (x_0) así como los valores de x obtenidos en cada iteración t utilizando el método de optimización gradiente descendente con momento.

En la Figura 1a se muestra el valor de x_t para 5 iteraciones utilizando el método de gradiente descendente, mientras que en la Figura 1b se muestra el valor de x_t para las mismas 5 iteraciones utilizando el esquema de gradiente descendente con momento. Además, en la Figura 2 se compara el valor de la función $y = f(x)$ obtenida para los valores x_t calculados con ambos métodos. Es posible observar cómo el proceso de optimización que incluye el término de momento converge significativamente más rápido al mínimo que la optimización a pasos. En este ejemplo se utilizaron $\alpha = 0.05$ y $\gamma = 0.9$. Nótese cómo aún con un valor de α pequeño, el gradiente descendente con momento alcanza el mínimo con pocas iteraciones.

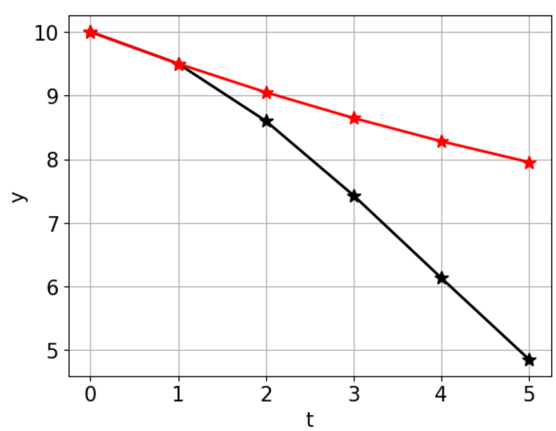


Figure 2: Comparación del valor de la función $y = f(x) = (x-5)^2 + 10$ evaluada en x_t utilizando el método de gradiente descendente (rojo) y el gradiente descendente con momento (negro) para 5 iteraciones.

Ejercicios