Tema : Adagrad

A continuación se describe el funcionamiento del optimizador *Adagrad* (Gradiente adaptativo). Para esto se continua utilizando como ejemplo la búsqueda del mínimo global de la función $y = f(x) = (x-5)^2 + 10$ cuya derivada es $\frac{dy}{dx} = 2(x-5)$. El procedimiento a realizar para encontrar el mínimo de una función $f(x)$ es el ya descrito. La regla de actualización de para este optimizador es como sigue:

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{g_{t+1}} + \epsilon} \nabla f(x_t) \tag{1}$$

donde:

$$g_{t+1} = g_t + (\nabla f(x_t))^2 \tag{2}$$

Nótese que esta regla de actualización es similar a aquella del GD simple, con la diferencia que el factor de aprendizaje en pra esta estrategia es variable, teniendo un parámetro α adaptativo representado con α_a y que puede ser entendido como:

$$\alpha_a \equiv \frac{\eta}{\sqrt{g_{t+1}} + \epsilon} \tag{3}$$

pudiendo entonces expresar la Ecuación 1 de la siente manera:

$$x_{t+1} = x_t - \alpha_a \nabla f(x_t) \tag{4}$$

Como ejemplo se presentan en la Tabla 1 los valores resultantes del algoritmo de búsqueda para encontrar la ubicación del mínimo sobre la función mencionada. Para obtener estos resultados se aplico el algoritmo numérico de búsqueda sobre diez iteraciones empleando como parámetros: $\eta = 2.5$ para Adagrad y $\alpha = 0.1$ para GD. Asimismo, se propuso arbitrariamente $x_0 = 10$ como valor inicial de búsqueda. cabe destacar que previo a iniciar el algoritmo se define $g_t = 0$ (iteración 0, Ecuación 2). Entonces el primer valor de α_a dependerá unicamente del cuadrado del gradiente $f(x_t)$ considerando el valor inicial de x , es decir x_0. Conforme incrementa el número de iteraciones el parámetro g estará acumulando el cuadrado de los gradientes obtenidos en iteraciones anteriores, por tanto α_a irá disminuyendo su valor dado que α_a es inversamente proporcional a la raíz cuadrada de g_{t+1}. La variable ϵ tiene un valor pequeño (1×10^{-8} en este ejemplo) y se añade a fin de evitar una posible división entre cero.

La idea al emplear Adagrad es emplear un valor η con un valor lo suficientemente alto para aproximarse al mínimo de una forma mas acelerada en la primeras iteraciones y al estar acercandoce a la posición del mínimo el valor de α_a será cada vez mas pequeño permitiendo una aproximación mas suave al valor buscado.

Comparación entre métodos												
Método	valor	Iteración t										
		0	1	2	3	4	5	6	7	8	9	10
GD	x_t	10	9.0	8.2	7.56	7.04	6.63	6.31	6.04	5.83	5.67	**5.53**
$\alpha = 0.1$	$\frac{dy}{dx} = \nabla f(x_t)$	10	8.0	6.39	5.11	4.09	3.27	2.62	2.09	1.67	1.34	1.07
Adagrad	x_t	10	7.50	6.38	5.78	5.44	5.25	5.14	5.08	5.04	5.02	**5.01**
	$\frac{dy}{dx} = \nabla f(x_t)$	10	5	2.76	1.56	0.89	0.50	0.29	0.16	0.09	0.05	0.03
$\eta = 2.5$	g_t	0	100	125	132.63	135.08	135.87	136.13	136.22	136.25	136.25	136.26
	α_a	N/A	0.25	0.2236	0.2170	0.2150	0.2144	0.2142	0.2141	0.2141	0.2141	0.2141

Table 1: Valores obtenidos en la ejecución del algoritmo de búsqueda del mínimo global sobre la función $y = f(x) = (x-5)^2 + 10$ empleando las reglas de actualización de GD y Adagrad.

En la Figura 1 se ilustra de forma visual los resultados mostrados en la Tabla 1, bajo los parámetros seleccionados, es evidente que el método de Adagrad logra una aproximación mas cercana al mínimo en menos iteraciones.

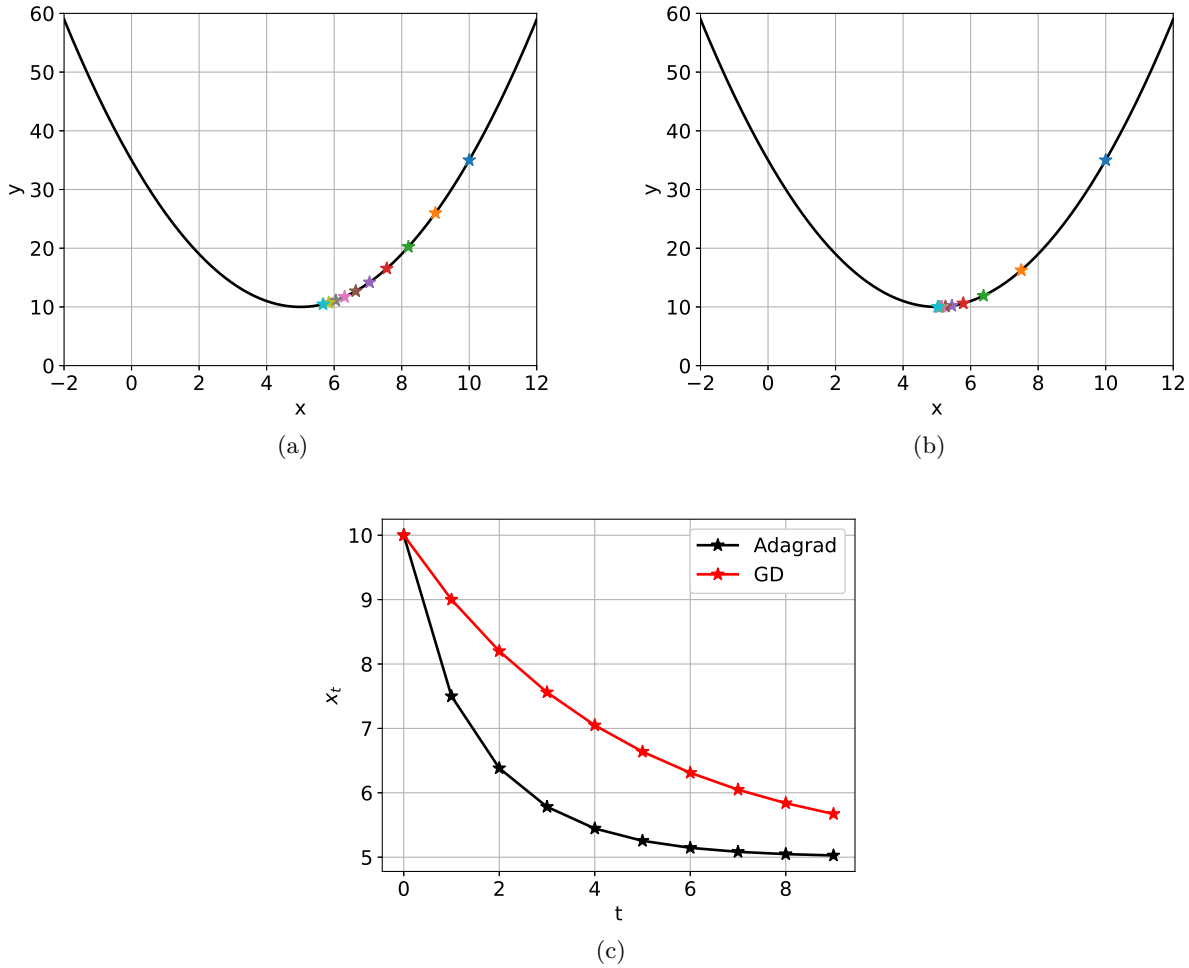
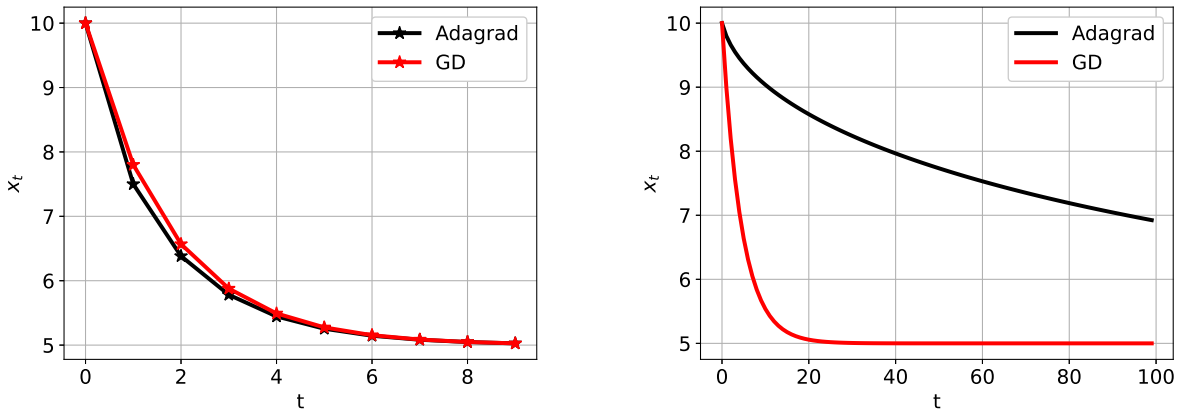


Figure 1: Comparación visual de la aproximación al mínimo de la función $y = f(x) = (x-5)^2+10$ empleando el optimizador a) GD y b) Adagrad. En la figura c) se ilustra el valor de x en la iteración t (x_t), el cual debe alcanzar el valor de 5 que es el resultado de solución analítica para la búsqueda del mínimo.

La Figura 2 muestra la comparación visual de aproximamiento al mínimo global con distintos parámetros a los anteriormente mencionados. De acuerdo con la Tabla 1, la variable α_a toma un valor aproximado de 0.22, entonces, en la Figura 2a se visualiza la comparación de la actualización de x tomando el ejemplo anterior pero ajustando $\alpha = 0.22$ para GD. Nótese que el funcionamiento en esta ocasión es similar para ambos optimizadores dado que el factor de aprendizaje $\alpha_a \approx \alpha$. La diferencia en estos métodos radica a que GD tiene un parámetro α fijo, donde típicamente $\alpha \in (0, 1)$ que en otro caso ocasionaría la divergencia, en cambio, para Adragrad, la variable η no esta sujeta al un valor dentro de un rango, ésta puede ser de un valor mucho mas alto, lo cual, para funciones convexas no existiría divergencia dado la disminución de α_a en función del gradiente de x. Se alienta al estudiante a verificar lo anterior dicho. Adicionalmente, se hace mención de uno de los inconveniente de Adagrad, el cual radica la disminución indefinida de α_a; si a la variable η se le asigna un valor pequeño al inicia del algoritmo, la aproximación de x al mínimo sera

extremada mente lenta pudiendo incluso converger el algoritmo en una posición que no es el mínimo dado que α_a en algún momento pudiera llegar a tomar un valor de aproximadamente cero (Figura 2b)



(a) (b)

Figure 2: Comparación visual de la aproximación de x al mínimo global de la función $y = f(x) = (x-5)^2+10$ tomando como parámetros $\eta = 2.5$, $\alpha = 0.22$ (Figura a) y $\eta = 0.2$, $\alpha = 0.1$ (Figura b).