Metodos de Optimizacion NO lineal

Integrantes

Juan Pablo Oriana - 60621
Tomas Cerdeira - 60051
Santiago Garcia Montagner - 60352

Objetivos del trabajo

- Analizar 3 metodos de optimizacion NO lineal que minimicen el error
  - Gradiente descendiente
  - Gradiente conjugado
  - ADAM

- Dados:

$$W = \begin{pmatrix} W_0 \\ W_1 \\ W_2 \end{pmatrix} \qquad w = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix} \qquad w_0 = (w_{01}, w_{02})$$

Se busca **minimizar el error** e informar los a**rgumentos y algoritmos óptimos** que más se ajusten a los reactivos iniciales dados como “ciertos”.

Función F

- Dados:

$g(x) = \frac{e^x}{1+e^x}$ Función logística

ξ^k Valores de entrada, con $k = 1, 2, 3$

ζ Valores de salida

$$\Rightarrow F(W, w, w_0, \xi) = g(\sum_{j=1}^{2} W_j g(\sum_{k=1}^{3} w_{jk}\xi_k - w_{j0}) - W_0)$$

Donde el error resulta $E(W, w, w_0) = \sum_{\mu=1}^{3} (\zeta^\mu - F(W, w, \xi^\mu))^2$

Valores de entrada ξ

$$\xi^1 = \begin{pmatrix} 4{,}4793 \\ -4{,}0765 \\ -4{,}0765 \end{pmatrix} \qquad \xi^2 = \begin{pmatrix} -4{,}1793 \\ -4{,}9218 \\ 1{,}7664 \end{pmatrix} \qquad \xi^3 = \begin{pmatrix} -3{,}9429 \\ -0{,}7689 \\ 4{,}8830 \end{pmatrix}$$

$$\zeta^1 = 0 \qquad \zeta^2 = 1 \qquad \zeta^3 = 1$$

Objetivo de los métodos

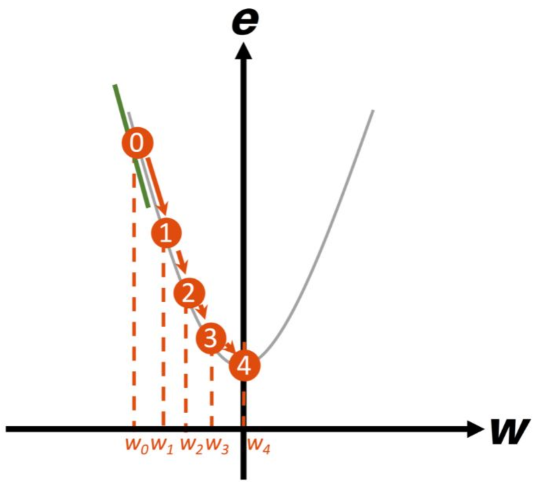
- Determinar el **valor de las variables independientes**, sujetas éstas en muchos casos a restricciones, **que maximizan o minimizan el valor de una función**

- Se realiza mediante la **penalización de la función objetivo**
  - **nueva función objetivo** que contiene a la original y a un término que cuando las restricciones estén próximas a no cumplirse incrementa el valor
    - es un **problema de minimización**

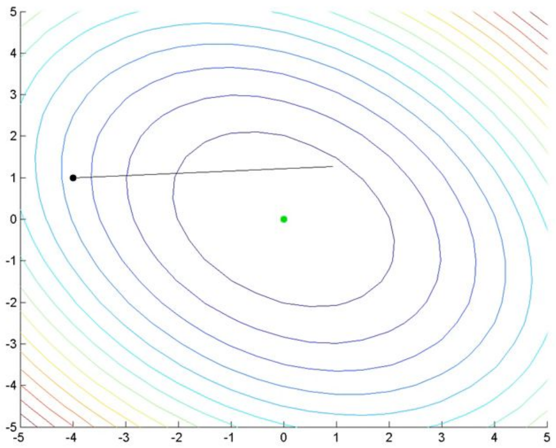
Metodo de Optimizacion NO lineal *gradiente descendente*

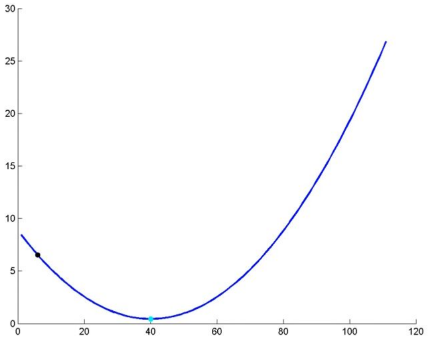
Gradiente Descendente

- Definiendo un **número de iteraciones** y una **tasa de aprendizaje**, se busca **llegar al mínimo de la función**.
    - Se escoge un valor de forma aleatoria de w
    - En cada iteración se actualiza el w = w - α . gradiente

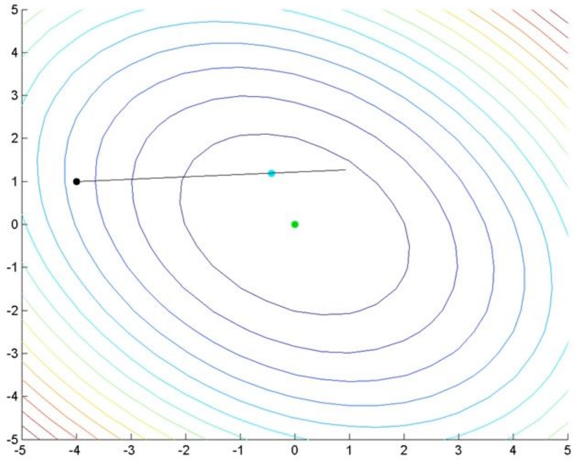


Ejemplo con curvas de nivel

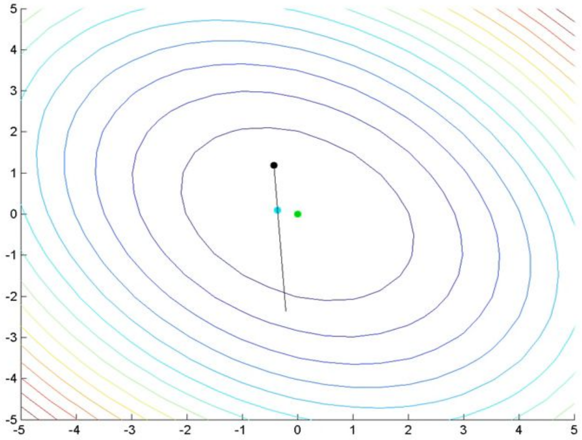




Sección de F sobre la gradiente



Ejemplo con curvas de nivel



Avanza en zig-zag

Gradiente Descendente Estocástico

- Método de **gradiente descendiente es costoso** ya que utiliza TODA la muestra de entrenamiento para el cálculo de dirección de decrecimiento.
- Surge, gradiente descendiente estocastico , que en cada iteracion toma una muestra de la población para el cálculo de la dirección

$$\sum_{\mu=1}^{k} \nabla_w E(\xi^\mu, \mathbf{w}^t), \; k << p.$$

- Los autores del método demuestran que este método converge, pero puede tardar más que el método determinístico

Metodo de Optimizacion NO lineal *gradiente conjugada*

Algunas definiciones

Definición 3.5.1. *Se dice que* $\{\vec{d}_k\}_{k=1}^{n}$ *son vectores mutuamente conjugados respecto a una matriz* G *simétrica y positiva definida si*

$$\vec{d}_k^{\,t}\, G\, \vec{d}_j = 0 \quad j \neq k. \tag{3.13}$$

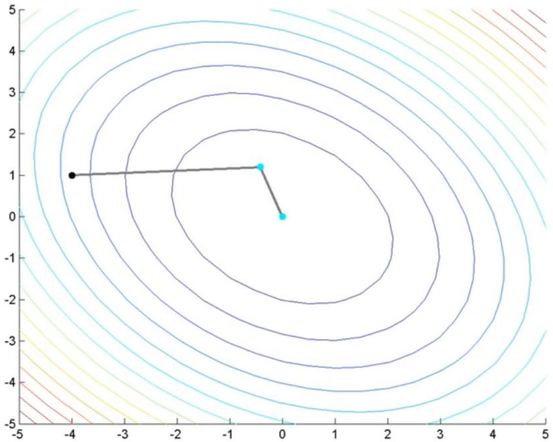
e.j.

$$(1,0)\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}(1/2,1) \quad = \quad \begin{pmatrix} 2 & -1 \end{pmatrix}(1/2,1) \quad = 0$$

OBS: En este caso, la matriz G vendría a ser H, la matriz Hessiana de f

Gradiente Conjugado

- Encontrando los dos **vectores conjugados**, encontramos una solución al problema en **menos pasos que el gradiente descendiente**.
    - Se **mueve por vectores ortogonales** que pasan por el x que minimiza a F



Referencias

Youtube - “Conjugated Gradient Method”

Patricia Saavedra Barrera - “Introducción a la optimización no lineal” (desde pág 68)

Metodo de Optimizacion NO lineal *ADAM (Adaptive Moment Estimation)*

ADAM (Adaptive Moment Estimation)

- Adapta la tasa de aprendizaje en cada paso
- La actualización del vector w se realiza coordenada a coordenada
- El algoritmo requiere:
  - $\propto$ Tasa de aprendizaje
  - β_1 y β_2 tasas de decaimiento. β_1 y β_2 ϵ [0,1)
  - f función objetivo estocástica
  - $m_0 = 0$. Donde m_0 es el primer vector de momentos
  - $v_0 = 0$. Donde v_0 es el segundo vector de momentos
  - w_0 parámetro inicial

Algoritmo

while w_t not converge do

$t := t + 1$

$g_t = \nabla f(w_{t-1})$

$m_t = \boxed{\beta_1} m_{t-1} + (1 - \boxed{\beta_1}) g_t$ → Actualiza la media móvil exponencial del gradiente

$v_t = \boxed{\beta_2} v_{t-1} + (1 - \boxed{\beta_2}) g_t^2$ → Actualiza la media móvil exponencial del gradiente al cuadrado

$w_t = w_{t-1} - \alpha \frac{m_t}{\sqrt{v_t + \epsilon}}$

Controlan las tasas de decaimiento exponencial de las medias móviles

Demo

Gracias por su atencion.