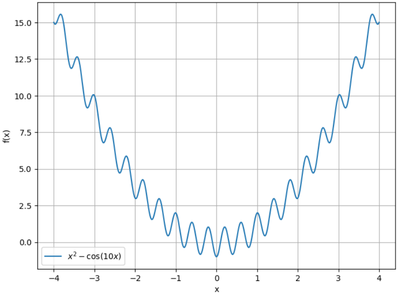
IIMAS
Universidad Nacional Autónoma de México

Agregando ruido a DG para llegar al mínimo global.

Aprendizaje máquina teórico.

Alejandro Antonio Estrada Franco

>Agregando ruido a DG para llegar al mínimo global



Para $f : \mathbb{R}^d \to \mathbb{R}$ el algoritmo que estudiaremos es el siguiente[1]:

Algoritmo 1 : *Decenso de Gradiente Ruidoso*; $x_1 \in \mathbb{R}^d; \alpha > 0; \sigma > 0$

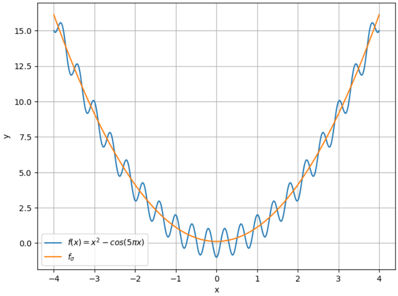
- para $t = 1, 2, \ldots$
- Generar una realización de la v.a. $U_t \sim Unif[B^2_{r=1}(0)]$
- Calcular $\nabla f(x_t)$
- $x_{t+1} = x_t - \alpha \nabla f(x_t) + \sigma U_t$

Definición 1: (componente σ-suave). Para todo $\sigma > 0$ y una función $f : \mathbb{R}^d \to \mathbb{R}$, la componente σ-suave de f se define como:

$$f_\sigma(x) = \mathbb{E}_U[f(x + \sigma U)] = \int_{B^2_{r=1}(0)} f(x + \sigma y) p_U(y) dy$$

definimos también:

$$r_\sigma(x) = f(x) - f_\sigma(x)$$



Algunas observaciónes y notación:

- Supondremos f continuamente diferenciable para obtener $\triangledown f_\sigma(x) = \triangledown \mathbb{E}_U[f(x+\sigma U)] = \mathbb{E}_U[\triangledown f(x+\sigma U)] = (\triangledown f)_\sigma(x)$; es decir:

$$\triangledown f_\sigma(x) = (\triangledown f)_\sigma(x).$$

- También denotaremos $\mathbb{E}_t[\cdot] = \mathbb{E}_{U_t}[\cdot]$, así mismo denominaremos $e_1 = x_1$, y $e_{t+1} = \mathbb{E}_t[x_{t+1}]$, además $\mathbb{E}[\cdot]$ denotará el valor esperado con respecto a la distribución conjunta de las uniformes (se toman independientes). Por ejemplo si x_t es el último punto del algoritmo, tenemos $\mathbb{E}[x_t] = \mathbb{E}_1 \ldots \mathbb{E}_{t-1}[x_t]$.

>Agregando ruido a DG para llegar al mínimo global

- Asumiremos también que la componente σ-suave f_σ es l-fuertemente convexa, y L-suave con $0 < l \leq L < \infty$. En consecuencia f_σ tiene un único minimizador x^*_σ, con $f^*_\sigma = f_\sigma(x^*_\sigma)$.
- Otro supuesto que se considera es que existen escalares M y μ tales que los gradientes $\triangledown f$ y $\triangledown f_\sigma$ satisfacen: $\mathbb{V}_U[\triangledown f(e + \sigma U)] \leq M_\sigma + \mu||\triangledown f_\sigma(e)||^2$
- Finalmente consideraremos que f es L_0-Lipschitz con respecto a la norma euclidiana en $\mathbb{R}^d$

>Agregando ruido a DG para llegar al mínimo global

Lema 1: Bajo las iteraciones de DGR se satisface para todo $n \in \mathbb{N}$: $e_{t+1} = e_t - \alpha\nabla f(e_t + \sigma U_{t-1}) + \sigma U_{t-1}$ mas aún se cumple también $\mathbb{E}_{t-1}[e_{t+1}] = e_t - \alpha\nabla f_\sigma(e_t)$.

Demostración: Para cada $t \in \mathbb{N}$ tenemos $x_{t+1} = x_t - \alpha\nabla f(x_t) + \sigma U_t$, tomando esperanzas con respecto a U_t; $e_{t+1} = x_{t+1} - \alpha\nabla f(x_t)$ de donde $e_{t+1} = x_{t+1} - \sigma U_t$ luego $e_{t+1} - e_t = -\alpha\nabla f(e_t + \sigma U_{t-1}) + \sigma U_{t-1}$ concluimos $\mathbb{E}_{t-1}[e_{t+1} - e_t] = -\alpha(\nabla f)_\sigma(e_t) \quad \square$.

Lema 2: Se cumple $f_\sigma(\mathbb{E}[x_{t+1}]) \leq \mathbb{E}[f_\sigma(e_{t+1})]$

Demostración: De la convexidad de f_σ y la desigualdad de Jensen:

$$\begin{aligned} f_\sigma(\mathbb{E}[x_{t+1}]) &= f_\sigma(\mathbb{E}_1 \ldots \mathbb{E}_{t-1}\mathbb{E}_t[x_{t+1}]) \\ &= f_\sigma(\mathbb{E}_1 \ldots \mathbb{E}_{t-1}[e_{t+1}]) \\ &= f_\sigma(\mathbb{E}[e_{t+1}]) \leq \mathbb{E}[f_\sigma(e_{t+1})] \end{aligned}$$

$\square$.

>Agregando ruido a DG para llegar al mínimo global

Lema 3: Para todo $t \in \mathbb{N}$ se verifica lo siguiente:

$$\mathbb{E}_{t-1}||e_{t-1} - e_t|| \leq 2\alpha^2(\mu + 1)||\nabla f_\sigma(e_t)||^2 + \alpha^2 M$$

donde:

$$M = 2M_\sigma + 2\left(\frac{\sigma}{\alpha}\right)^2$$

Teorema 1: Para todo $t \in \mathbb{N}$ se cumple:

$$f_\sigma(\mathbb{E}[x_{t+1}]) - f_\sigma^* - \frac{\alpha LM}{2l} \leq \rho^t \left(f_\sigma(x_1) - f_\sigma^* - \frac{\alpha LM}{2l}\right)$$

Si

$$0 < \alpha \leq \frac{1}{2L(\mu + 1)}$$

donde $\rho = 1 - \alpha l \in \left(\frac{1}{2}, 1\right)$

Lema 4: Para todo $\sigma > 0$ se cumple $|f_\sigma(x) - f(x)| \leq L_0\sigma$
Demostración:

$$\begin{aligned} |f_\sigma(x) - f(x)| &= |\mathbb{E}_U[f(x+\sigma U) - f(x)]| \\ &\leq \mathbb{E}_U|f(x+\sigma U) - f(x)| \\ &\leq L_0\mathbb{E}_U[||\sigma U||] = L_0\sigma\mathbb{E}_U[||U||] \leq L_0\sigma \qquad \square. \end{aligned}$$

Lema 5: $f_\sigma^* \leq f^* + L_0\sigma$
Demostración: Tenemos que $f(x) = f_\sigma(x) + r_\sigma(x)$, entonces:

$$f^* = \min_{x\in\mathbb{R}^d} f(x) \geq \min_{x\in\mathbb{R}^d} f_\sigma(x) + \min_{x\in\mathbb{R}^d} r_\sigma(x) = f_\sigma^* + \min_{x\in\mathbb{R}^d} r_\sigma(x)$$

Deacuerdo al Lema 4:

$$-L_0\sigma \leq \min_{x\in\mathbb{R}^d} r_\sigma \leq L_0\sigma$$

Entonces tenemos $f^* \geq f_\sigma^* - L_0\sigma \qquad \square.$

>Agregando ruido a DG para llegar al mínimo global

Teorema 2: Para α y ρ como en el Teorema 1 se cumple para todo $t \in \mathbb{N}$:

$$f(\mathbb{E}[x_{t+1}]) - f^* \leq \rho^t M_1 + M_2$$

Donde:

$$M_1 = f_\sigma(x_1) - f_\sigma^* - \frac{\alpha L \left(M_\sigma + (\frac{\sigma}{\alpha})^2 \right)}{l},$$

$$M_2 = \alpha \left(\frac{LM_\sigma}{l} + \frac{L\left(\frac{\sigma}{\alpha}\right)^2}{l} + L_0 \left(\frac{\sigma}{\alpha}\right) \right) + L_0 \sigma$$

Demostración: Por el Lema 4 tenemos para todo $t \in \mathbb{N}$; $f(\mathbb{E}[x_{t+1}]) \leq f_\sigma(\mathbb{E}[x_{t+1}]) + L_0\sigma = f_\sigma(\mathbb{E}[x_{t+1}]) + \alpha L_0 \left(\frac{\sigma}{\alpha}\right)$, luego del Lema 5 $-f^* \leq -f_\sigma^* + L_0\sigma$ por lo que tenemos:

$$f(\mathbb{E}[x_{t+1}]) - f^* \leq f_\sigma(\mathbb{E}[x_{t+1}]) - f_\sigma^* + \alpha L_0 \left(\frac{\sigma}{\alpha}\right) + L_0 \sigma$$

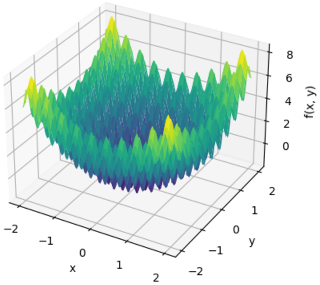
>Agregando ruido a DG para llegar al mínimo global

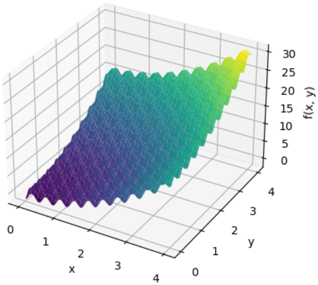
Del Teorema 1 tenemos:

$$\begin{aligned}
f(\mathbb{E}[x_{t+1}]) - f^* &\leq f_\sigma(\mathbb{E}[x_{t+1}]) - f_\sigma^* + \alpha L_0 \left(\frac{\sigma}{\alpha}\right) + L_0\sigma \\
&\leq \rho^t \left(f_\sigma(x_1) - f_\sigma^* - \frac{\alpha LM}{2l}\right) + \frac{\alpha LM}{2l} + \alpha L_0 \left(\frac{\sigma}{\alpha}\right) + L_0\sigma \\
&= \rho^t M_1 + \alpha \left(\frac{2M_\sigma + 2\left(\frac{\sigma}{\alpha}\right)^2}{2l} + L_0\left(\frac{\sigma}{\alpha}\right)\right) + L_0\sigma \\
&= \rho^t M_1 + M_2 \qquad \square.
\end{aligned}$$

>Agregando ruido a DG para llegar al mínimo global

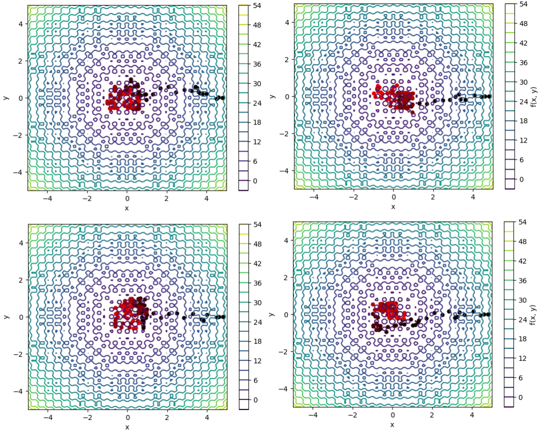
Ejemplo: Consideremos la función
$f(x, y) = x^2 + y^2 - cos(5\pi x) - cos(5\pi y)$





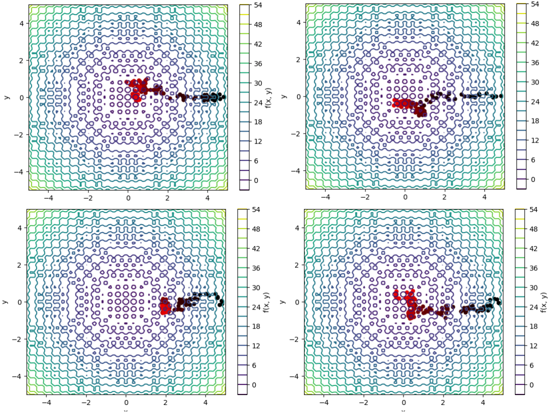
>Agregando ruido a DG para llegar al mínimo global

Se ilustran cuatro ejecuciones del algoritmo GDR con
$t = 100, \quad x_1 = (5, 0), \quad \sigma = 0{,}05, \quad \alpha = 0{,}02$



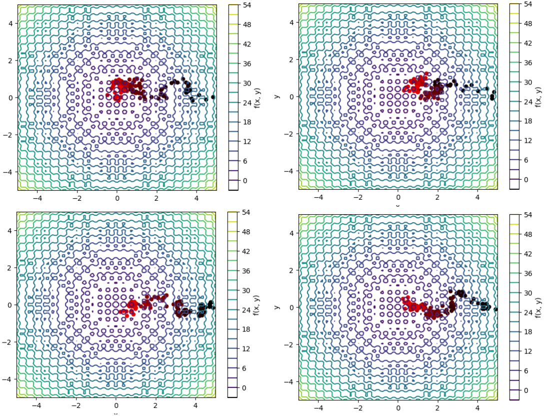
>Agregando ruido a DG para llegar al mínimo global

Se ilustran cuatro ejecuciones del algoritmo GDR con
$t = 100, \quad x_1 = (5, 0), \quad \sigma = 0{,}01, \quad \alpha = 0{,}015$



>Agregando ruido a DG para llegar al mínimo global

Se ilustran cuatro ejecuciones del algoritmo GDR con
$t = 100, \quad x_1 = (5, 0), \quad \sigma = 0{,}015, \quad \alpha = 0{,}015$



>Agregando ruido a DG para llegar al mínimo global

[1] Xuliang Quin, Xin Xiu y Xiaopeng Luo. «Global Convergence of Noisy Gradient Descent». En: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2022).