Fundamentos Matemáticos del Aprendizaje Profundo

1er cuat. 2025
Clase 4

Repaso

TEOREMA 4.2.2. *Asumamos que* $f\colon D \subset \mathbb{R}^n \to \mathbb{R}$ *es de clase* C^2, *que* $\mathbf{x}^* \in D$ *es un mínimo local estricto de* f *y que existe un entorno* $\mathcal{V} \subset D$ *de* $\mathbf{x}^*$ *tal que* $\nabla f(\mathbf{x}) \neq 0$ *para* $\mathbf{x} \in \mathcal{V} \setminus \{\mathbf{x}^*\}$. *Entonces, dado* $\mathbf{x}_0 \in \mathcal{V}$, *existe una curva* $\gamma\colon [0,1] \to \mathbb{R}^n$ *tal que*

1. $\gamma(0) = \mathbf{x}_0$;
2. $\gamma(1) = \mathbf{x}^*$;
3. $\gamma'(t)$ *es perpendicular a* $\mathcal{S}_{f(\gamma(t))}$ *para* $t \in [0,1)$.

$$\mathcal{S}_c = f^{-1}(\{c\}) = \{\mathbf{x} \in D\colon f(\mathbf{x}) = c\}.$$

¿Cómo implementamos una versión discreta de este procedimiento?

Algoritmos de descenso (versión discreta)

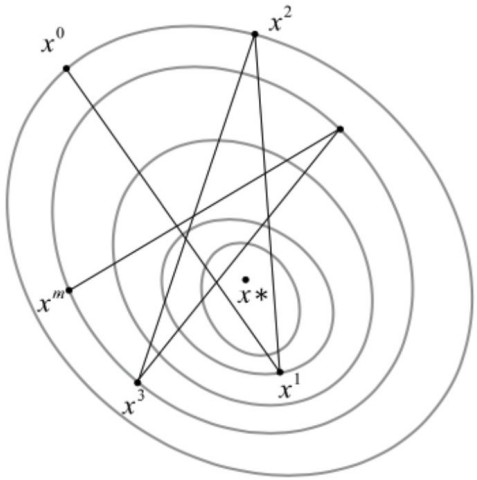
Buscamos $x_1,...,x_k,...$ tales que

1. $\mathbf{x}_k \in \mathcal{S}_{c_k}$;
2. $c_{k+1} < c_k$, para $k = 0, 1, \ldots, m-1$;
3. el segmento $[\mathbf{x}_k \mathbf{x}_{k+1}]$ es perpendicular a $\mathcal{S}_{c_k}$.

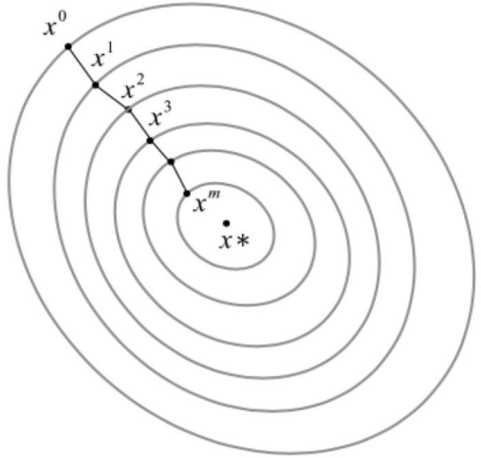
Empezamos con x_0 “cercano” al mínimo. Se elige un parámetro $\eta>0$ (chico)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}.$$

Elección de η



a



b

a. el paso η es grande

b. el paso η es chico

Criterio de parada

- Típicamente se para cuando $c_{k+1}<c_k$ para k=1,...,m-1 y $c_{m+1} > c_m$.

- Se espera que si η es más pequeño, m sea más grande.

- Cuando η→0, la poligonal $x_0x_1\ldots x_m$ aproxima la curva continua ɣ(t).

- Al momento de parar se tiene $\|\mathbf{x}_0 - \mathbf{x}^*\| - m\eta \leq \|\mathbf{x}_m - \mathbf{x}^*\| \leq \operatorname{diam}(\mathcal{S}_{c_m})$

Más sobre η

Al parámetro η se lo llama “tasa de aprendizaje”.

La tasa de aprendizaje no conviene en la práctica que sea constante

$$\eta_k \sim \delta \|\nabla f(x_k)\|$$

Luego:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \delta \nabla f(\mathbf{x}_k)$$

PROPOSICIÓN 4.2.4. *La sucesión* $\{\mathbf{x}_k\}_{k\geq 0}$ *definida en* (4.2.5) *resulta convergente si y sólo si la sucesión de gradientes tiende a 0,* $\nabla f(\mathbf{x}_k) \to 0$ *cuando* $k \to \infty$.

Estimaciones para δ

$$f(\mathbf{x}_{k+1}) \simeq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k) \cdot (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}_k) Hf(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)^t$$
$$= f(\mathbf{x}_k) - \delta \|\nabla f(\mathbf{x}_k)\|^2 + \frac{\delta^2}{2} \nabla f(\mathbf{x}_k) Hf(\mathbf{x}_k) \nabla f(\mathbf{x}_k)^t.$$

Como $Hf(x_{min})>0$, podemos asumir que $Hf(x_k)>0$. Como $Hf(x_k)$ es simétrica se tiene

$$0 \leq \frac{\delta^2}{2} \lambda_{\min} \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{\delta^2}{2} \nabla f(\mathbf{x}_k) Hf(\mathbf{x}_k) \nabla f(\mathbf{x}_k)^t \leq \frac{\delta^2}{2} \lambda_{\max} \|\nabla f(\mathbf{x}_k)\|^2$$

Estimaciones para δ (2)

De donde

$$f(\mathbf{x}_{k+1}) \simeq f(\mathbf{x}_k) - \delta \|\nabla f(\mathbf{x}_k)\|^2 + \frac{\delta^2}{2} \nabla f(\mathbf{x}_k) H f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)^t$$

$$\leq f(\mathbf{x}_k) - \delta \|\nabla f(\mathbf{x}_k)\|^2 + \frac{\delta^2}{2} \lambda_{\max} \|\nabla f(\mathbf{x}_k)\|^2,$$

Luego $f(x_{k+1}) < f(x_k)$ si se cumple $-\delta + \frac{\delta^2}{2}\lambda_{\max} < 0 \iff \delta < \frac{2}{\lambda_{\max}}.$

Esto da una cota para la tasa de aprendizaje

Búsqueda estocástica

Caso determinístico: Se busca una curva $x'(t) = b(x(t))$ de manera tal que $f(x(t))$ alcance el mínimo de f lo más rápido posible.

$$b(x) = -\nabla f(x)$$

Variante estocástica: Se introduce un *ruido* o perturbación aleatoria al modelo determinista

$$dX(t) = b(X(t))\, dt + \sigma(X(t))\, dW(t)$$

Pero…. ¿Qué es dW(t)?

Movimiento Browniano y ruido blanco

DEFINICIÓN 4.3.1. Un movimiento Browniano, o proceso de Wiener, es un proceso estocástico W_t, $t \geq 0$, tal que

1. $W_0 = 0$ casi seguramente;
2. si $0 \leq u < s < t$, entonces $W_t - W_s$ y $W_s - W_u$ son independientes (el proceso tiene incrementos independientes);
3. $t \mapsto W_t$ es continuo casi seguramente;
4. los incrementos están normalmente distribuidos, con $W_t - W_s \sim N(0, |t - s|)$.

Un ruido blanco: $dW(t) \sim N(0, dt)$

Un movimiento Browniano m-dimensional es $W_t = (W_t^1, \ldots, W_t^m)$

Ecuación Diferencial Estocástica (SDE)

$$dX_t = \mathbf{b}(X_t)\,dt + \sigma(X_t)\,dW_t.$$

$\mathbf{b}\colon \mathbb{R}^n \to \mathbb{R}^n$ deriva (drift) $\sigma\colon \mathbb{R}^n \to \mathbb{R}^{n\times m}$ difusión (diffusion)

¿Cómo se resuelve una SDE? Método de Euler-Maruyama

$$X_{k+1} - X_k = \mathbf{b}(X_k)\delta + \sigma(X_k)(W_{t_{k+1}} - W_{t_k}).$$

Ecuación Diferencial Estocástica (SDE) (2)

$$X_{k+1} - X_k = \mathbf{b}(X_k)\delta + \sigma(X_k)(W_{t_{k+1}} - W_{t_k}).$$

Llamando $B_k := W_{t_{k+1}} - W_{t_k} \sim N(0, \delta)$, obtenemos

$$X_{k+1} = X_k + \mathbf{b}(X_k)\delta + \sigma(X_k)B_k$$

El método de descenso de gradiente determinístico $\mathbf{x}_{k+1} = \mathbf{x}_k - \delta\nabla f(\mathbf{x}_k)$

Luego, si tomamos $\mathbf{b}(x) = -\nabla f(x)$ se obtiene una perturbación estocástica