6. Interpretarea grafică a gradientului

Graficul unei funcții multivariabile
Gradienți locali
Regula de înlănțuire a gradienților

6.1 Gradientul unei funcții

Derivata, sau gradientul unei funcții, ne indică rata de modificare a funcției față de o variabilă, în jurul unei regiuni infinitezimal de mică, aproape de un anumit punct pentru care se evaluează derivata:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{6.1}$$

Fracția din partea stângă a Ecuației 6.1 nu reprezintă operația de împărțire, ci indică notația operatorului $\frac{d}{dx}$ aplicat funcției f, ce returnează derivata lui f. Atunci când h este foarte mic, funcția este aproximată printr-o linie dreaptă, unde derivata este panta acestei linii. Cu alte cuvinte, derivata unei funcții f față de o variabilă x indică senzitivitatea lui f față de valoarea lui x. Acest lucru este vizibil atunci cand rearanjăm expresia 6.1 astfel:

$$f(x+h) = f(x) + h\frac{df(x)}{dx} \tag{6.2}$$

Expresia 6.2 ne spune că valoarea funcției f crește sau descrește atunci cand valoarea lui x se modifică la $x+h$, valoarea funcției modificându-se cu cantitatea h înmulțită cu derivata funcției f.

Gradientul unei funcții poate fi obținut analitic, prin formulele de diferențiere, sau numeric. Gradientul numeric este încet, imprecis, însă ușor de programat, în timp ce gradientul analitic: este rapid, precis, însă susceptibil la erori. În antrenarea rețelelor neuronale se folosește gradientul analitic, însă implementarea gradientului analitic este verificată folosind gradientul numeric.

Gradientul unei funcții f se notează cu ∇f. Spre exemplu, gradientul unei funcții de 2 variabile $f(x, y)$ este:

$$\nabla f(x, y) = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right], \tag{6.3}$$

unde operatorul $\frac{\partial}{\partial x}$ indică derivata parțială a funcției $f(x, y)$ față de variabila x.

Pentru simplitate, vom folosi termenul "gradient față de x", în locul termenului complet de "derivată parțială față de x"

6.2 Interpretarea grafică a gradientului

În rețelele neuronale adânci, precum rețelele neuronale convoluționale, gradientul funcției de cost poate atinge dimensiuni foarte mari, aproape imposibil de scris sub formă de ecuații. Forma gradientului poate fi simplificată dacă îl reprezentăm ca și un graf computațional.

Fie funcția de trei variabile:

$$f(x, y, z) = (x + y)z \tag{6.4}$$

reprezentată grafic prin diagrama din Figura 6.1. Pentru valorile argumentelor:

$$\begin{aligned} x &= -2 \\ y &= 5 \\ z &= -4 \end{aligned} \tag{6.5}$$

funcția va returna valoarea $f(x, y, z) = -12$. Vom denumi ca și "propagare înainte" (feedforward propagation) calculul valori de ieșire a funcției, dându-se valorile variabilelor de intrare. Având valorile de intrare, calculăm valorile de ieșire a fiecărui nod din graf, de la stânga la dreapta.

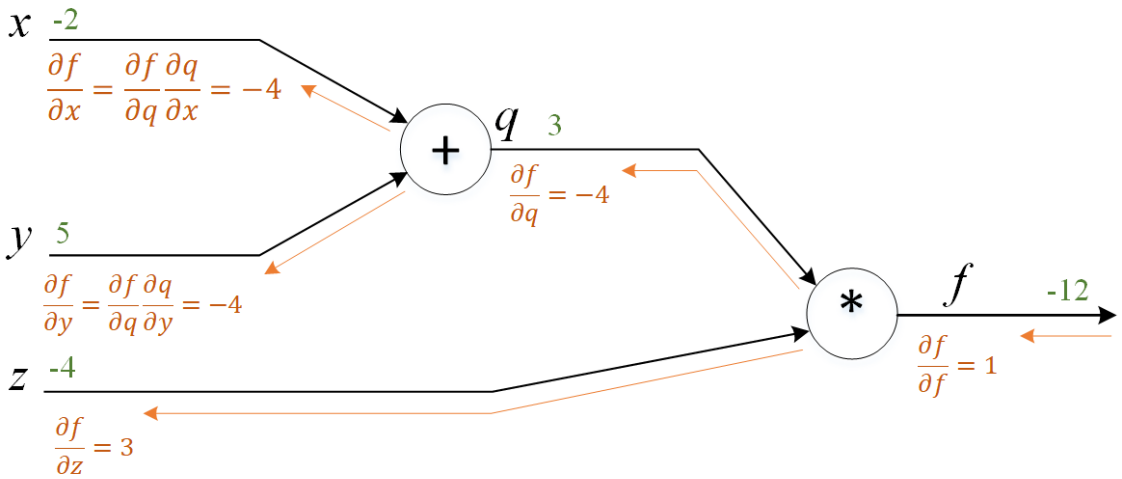

Fig. 6.1 Graful computațional pentru funcția $f(x, y, z) = (x + y)z$. Valorile de intare și ieșire din noduri sunt scrise în verde, deasupra liniei de propagare înainte (feedforward), iar gradienții în roșu, sub linie.

În următoarea fază, dorim să calculăm gradientul funcție $f(x, y, z)$ față de argumentele x, y, z. Pentru aceasta, introducem o funcție intermediară q ce calculează valoarea nodului

"+", și anume $q = x + y$. q este o valoare intermediară folosită în calculul valorii funcției f. Gradientul lui f reprezintă derivatele parțiale ale funcției f față de variabilele de intrare:

$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \tag{6.6}$$

Derivatele parțiale, sau gradienții, fiecărui nod din graful computațional sunt:

$$q = x + y : \ \frac{\partial q}{\partial x} = 1, \ \frac{\partial q}{\partial y} = 1 \tag{6.7}$$

$$f = qz : \ \frac{\partial f}{\partial q} = z, \ \frac{\partial f}{\partial z} = q \tag{6.8}$$

Vom denumi "gradienți locali" derivatele parțiale $\frac{\partial q}{\partial x}$ și $\frac{\partial q}{\partial y}$.

Gradientul funcției din Figura 6.1 se calculează prin "propagarea înapoi" (backpropagation), de la dreapta la stânga, pornind de la valoarea de ieșire a funcției și calculând derivatele fiecărui nod până la variabilele de intrare.

Pornind de la ieșire, gradientul funcției f față de ea însăși este identitatea $\frac{\partial f}{\partial f} = 1$. Gradientul lui f față de z este $\frac{\partial f}{\partial z} = q = 3$. Intuitiv, valoarea gradientului în acest caz, ne spune că influența variabilei z asupra valorii finale a lui f este pozitivă, având o pantă, sau "forță", egală cu 3. Daca incrementăm z cu o valoare h, atunci ieșirea grafului va fi crescută cu o valoare $3h$.

Gradientul lui f față de q ne spune că dacă q va crește, atunci ieșirea grafului va scădea. De exemplu, dacă q va crește cu valoarea h, atunci ieșirea grafului va scădea cu $4h$.

Gradientul funcției f față de variabila de intrare y se calculeaza folosind regula de înlănțuire, și anume:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y} = -4 \tag{6.9}$$

Influența lui y asupra lui q este 1 ($\frac{\partial q}{\partial y} = 1$). Influența variabilei de intrare y asupra ieșirii grafului este dată prin multiplicarea gradienților $\frac{\partial f}{\partial q} = -4$ și $\frac{\partial q}{\partial y} = 1$.

x și y au o influență pozitivă asupra lui q cu o pantă de valoarea 1. Crescând x cu o valoare h, implică creșterea lui q cu valoarea h.

Creșterea lui x va crește valoarea lui q, care la rândul lui va descrește valoarea lui f.

Figura 6.2 ilustrează un nod dintr-un graf computațional ce calculează valoarea de ieșire z folosind funcția de activare f. De îndată ce nodul primește valorile de intrare x și y, în timpul operației de propagare înainte, acesta poate calcula și valorile gradienților locali, și anume $\frac{\partial z}{\partial x}$ și $\frac{\partial z}{\partial y}$. Gradienții locali vor fi folosiți în calculul gradientului unei funcții finale L, funcție ce se află la ieșirea grafului. În cursul de față, L reprezintă funcția de cost.

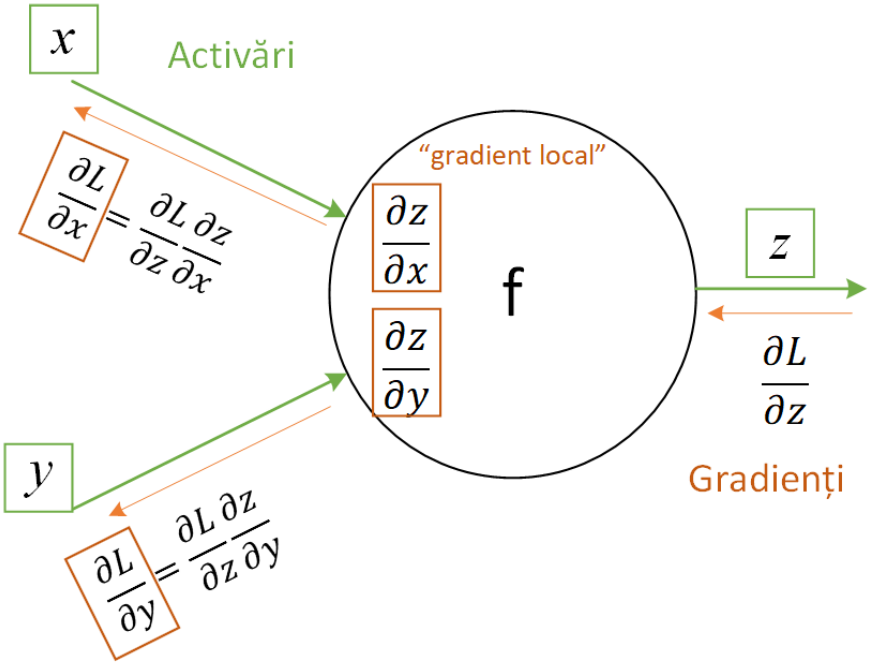

Fig. 6.2 Nod ce calculează funcția de activare f.

6.3 Reprezentarea grafică a funcției de regresie logistică

în graful computațional putem reprezenta orice fel de nod, atâta timp cât funcție ce o implementează este derivabilă.

Un alt exemplu de graf computațional este redat în Figura 6.3, pentru funcția de regresie logistică:

$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}} \tag{6.10}$$

Se poate observa că funcția de regresie logistică reprezintă modelul unui neuron cu n intrări x și n parametrii θ. **Gradienții din Figura 6.3 sunt exprimați față de ieșirea finală a funcției, folosindu-se regula de înlănțuire a gradienților.**

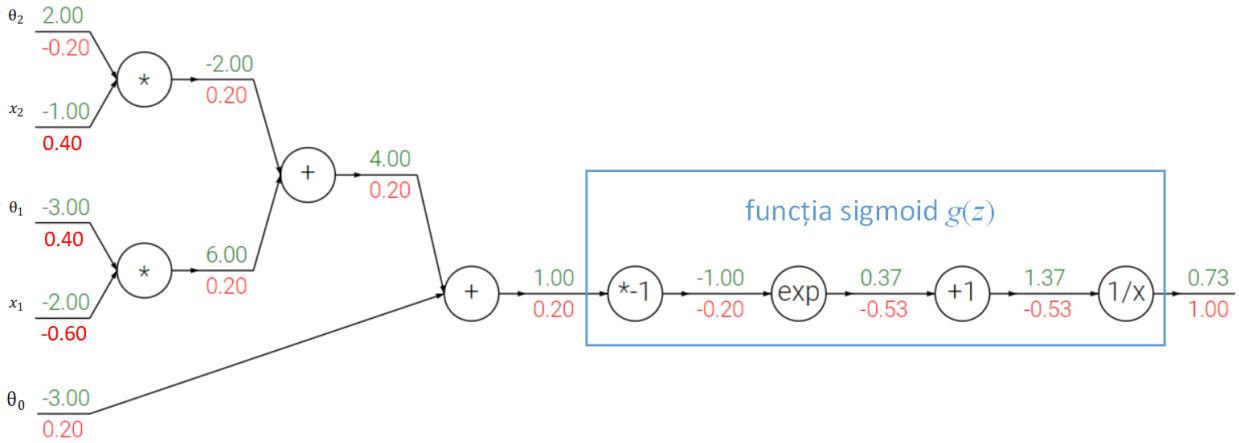

Fig. 6.3 Graf computațional pentru funcția de regresie logistică.

Derivatele nodurilor grafului din Figura 6.3 sunt:

$$\begin{aligned} f(x) = \frac{1}{x} \quad &\rightarrow \quad \frac{df}{dx} = -\frac{1}{x^2} \\ f_c(x) = c + x \quad &\rightarrow \quad \frac{df}{dx} = 1 \\ f(x) = e^x \quad &\rightarrow \quad \frac{df}{dx} = e^x \\ f_b(x) = bx \quad &\rightarrow \quad \frac{df}{dx} = b \end{aligned} \tag{6.11}$$

unde funcțiile f_c și f_b translatează (adună) intrarea cu o constantă c, respectiv scalează (înmulțește) intrarea cu o constantă b.

În cazul nodului $\frac{1}{x}$, gradientul este negativ deoarece atunci când valoarea de intrare în nod crește, valoarea de ieșire a nodului scade cu $\frac{1}{x}$. Următorul nod adaugă constanta +1, derivata acestuia fiind 0.

Gradientul local al nodului "+" este 1, ceea ce înseamnă că acest nod va copia către intrările sale gradientul primit la ieșire. Un nod plus distribuie pur și simplu gradientul primit către nodurile anterioare.

Pentru verificare, vom deriva funcția sigmoid, marcată cu albastru în graful din Figura 6.3:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{6.12}$$

$$\frac{dg(z)}{dz} = \frac{e^{-z}}{(1 + e^{-z})^2} = \left(\frac{1 + e^{-z} - 1}{1 + e^{-z}}\right)\left(\frac{1}{1 + e^{-z}}\right) = (1 - g(z))g(z) \tag{6.13}$$

Prin derivare, calculul gradientului devine mult mai simplu. Spre exemplu, în timpul propagării înainte, dacă funcția sigmoid primește la intrare valoarea 1.0, atunci ieșirea ei va fi 0.73. Utilizând Ecuația 6.13, gradientul local va fi $(1 - 0.73) \cdot 0.73 = 0.2$. În practică, vom grupa astel de operații pentru simplificarea calculului gradientului.

Bibliografie

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[2] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[3] M. Lutz, *Learning Python*, 2nd Ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2003.

[4] A. Ng, "Stanford cs229 - machine learning," 2008.

[5] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd Ed. Pearson Education, 2003.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.