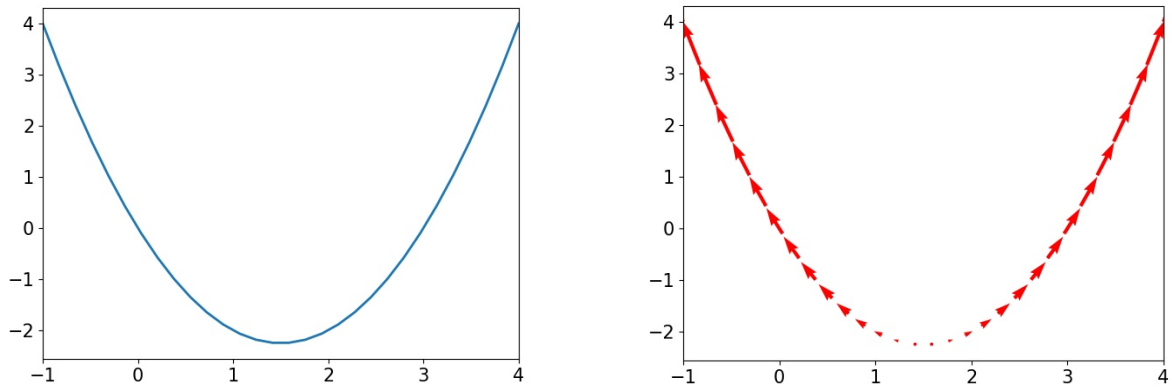
Metoda gradientului descendent

Gradientul descendent este un algoritm iterativ utilizat frecvent pentru soluționarea aproximativă a problemlor de optimizare atunci când acestea presupun minimizarea unei funcții-obiectiv. Noțiunea de *optimizare* se referă la identificarea celei mai bune soluții a unei probleme date (soluția optimă). Soluția se formulează în termenii valorilor unor parametri ai unei metode / ai unui sistem, scopul fiind identificarea valorilor parametrilor care conduc la optim. Adesea, problema de optimizare se formulează ca o problemă de minimizare a unei funcții ce descrie un comportament sau o stare nedorită a sistemului vizat. În acest caz, identificarea *soluției optime* înseamnă identificarea valorilor parametrilor sistemului care minimizează o funcție ce descrie eroarea, costul sau pierderile din cadrul acelui sistem.

Ideea principală a gradientului descendent este de a ajusta treptat valorile parametrilor în direcția opusă vectorului gradient, în ideea de a "coborî" în domeniul funcției către un punct de minim.

Considerăm funcția $f(x) = x^2 - 3x$, $f : [-1, 4] \to \mathbb{R}$ cu graficul din Fig. 1(a). Minimul acestei funcții se determină în punctul în care se anulează derivata $f'(x) = 2x - 3$, adică în $x_0 = 1.5$. Prin urmare, $f_{\min}(x) = f(x_0) = -2.25$. Prin contrast cu acest exemplu, funcțiile care descriu metode, reguli sau modele matematice utilizabile în practică fie sunt mult prea complexe, fie rădăcinile derivatelor lor sunt dificil de determinat, fie nu li se cunoaște expresia analitică. În aceste situații, pentru minimizarea lor se utilizează metode numerice, prin care se ajustează pas cu pas parametrii funcțiilor până când se ajunge la un minim aproximativ.

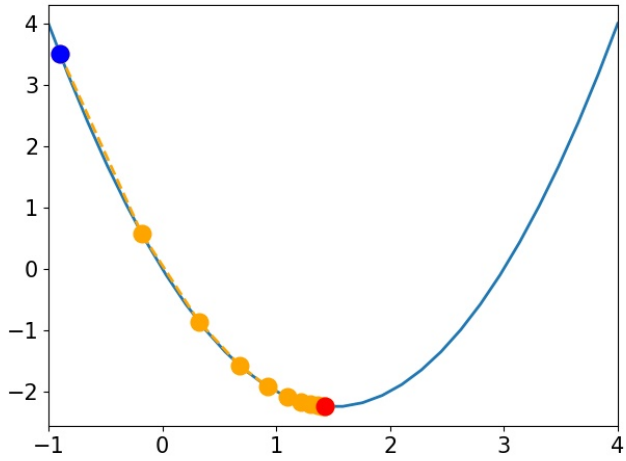


(a) **(b)**

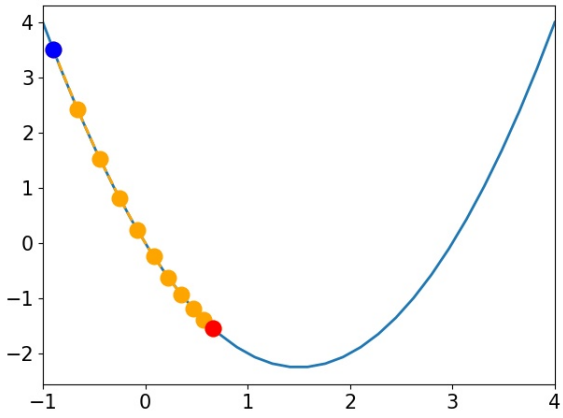
Fig.1. (a) Graficul funcției *f(x) = x² - 3x*; (b) Reprezentarea derivatei sub forma unor vectori tangenți la graficul funcției.

În Fig. 1(b) se prezintă o reprezentare a derivatei funcției anterioare, sub forma unor vectori tangenți la grafic. Sensul vectorilor este dat de semnul derivatei iar lungimea lor este proporțională cu modulului acesteia. Se observă faptul că, pentru orice x din domeniul funcției, vectorii indică direcția în care valoarea funcției crește, prin urmare $f(x + f'(x)) \geq f(x), \forall x \in [-1, 4]$. Prin contrast, $f(x - f'(x)) \leq f(x), \forall x \in [-1, 4]$. Așadar, prin ajustări repetate ale valorii lui x în sensul opus celui indicat de derivată, valoarea funcției va "coborî" către un minim. Metoda care exploatează acest rezultat se numește **gradient descendent** și, pentru funcții unidimensionale, se poate descrie astfel:

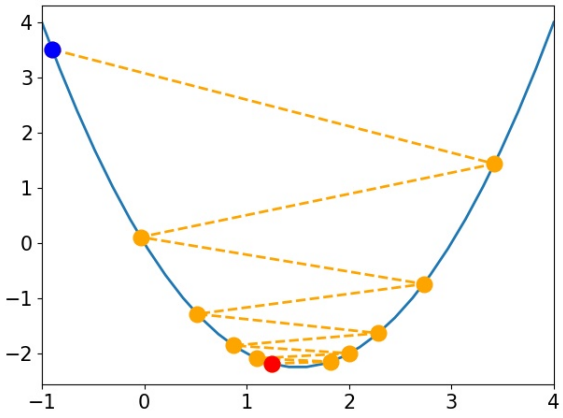
$x = x_{init}$
while not $\min(f(x))$:
  $x = x - \alpha f'(x)$



(a)



(b)



(c)

Fig. 2. Minimizarea funcției prin metoda gradientului descendent: (a) pornind de la valoarea inițială x = -0.9 (punctul albastru), se ajunge treptat la valoarea finală a lui x, în minimul funcției (punctul roșu); (b) Pentru valori prea mici ale ratei de învățare α, algoritmul nu converge, numărul iterațiilor fiind insuficient; (c) Pentru valori prea mari ale ratei de învățare α, x oscilează în vecinătatea minimului, existând șansa ca acesta să fie evitat.

Valoarea lui x se modifică din aproape în aproape, în decursul mai multor iterații, până când se consideră ca s-a atins min(f(x)). În această situație, spunem că algoritmul *converge* către soluția căutată. Parametrul α ("alfa") se numește *rată de învățare*, și se utilizează pentru a controla mărimea pasului cu care se ajustează x de la o iterație la alta (altfel spus, prin ajustarea lui α se stabilește viteza de convergență a algoritmului). Pașii se ilustrează în Fig. 2(a), unde, pornind din $x_{init} = -1$, se ajunge treptat la valoarea lui x pentru care f(x) este minimă. Algoritmul se încheie în momentul în care se respectă un *criteriu de convergență*, de exemplu:

- un număr maxim de iterații.
- în cazul în care se cunoaște minimul analitic al funcției (și interesează doar valoarea lui x pentru care se atinge acest minim), se consideră că algoritmul se încheie în momentul în care minimul determinat este aproximativ egal cu minimul cunoscut al funcției.
- în cazul în care nu se cunoaște minimul funcției, se consideră că algoritmul se încheie dacă în decursul ultimelor câtorva iterații minimele determinate sunt aproximativ aceleași.

Obs: Două valori a, b sunt *aproximativ egale* dacă modulul diferenței lor este suficient de mic, adică dacă $|a - b| \leq \varepsilon$, unde ε este o constantă pozitivă de valoare mică (de exemplu 10^{-4}).

Valoarea ratei de învățare α se alege funcție de specificul problemei de minimizare:

- valorille prea mici cauzează o viteză de convergență prea mică, situație în care fie este nevoie de un număr foarte mare de iterații pentru a ajunge în minimul funcției, fie, pentru un număr rezonabil de iterații, acest minim nu se atinge (Fig. 2(b))
- valorile prea mari cauzează "salturi" mari în decursul iterațiilor. În acest caz, valorile funcției oscilează în jurul minimului și este posibil ca acesta să fie complet evitat (Fig. 2(c)).

Minime locale și globale

În Fig. 3 se ilustrează graficul funcției $f(x) = x\cos(3x)$, $f : [-1, 2] \to \mathbb{R}$. Funcția are două minime:

- un minim local f(x) = -0.187 pentru x = -0.287
- un minim global f(x) = -1.096 pentru x = 1.141
-

Obs: minim local = valoarea minimă a funcției într-o anumită vecinătate a lui x
minim global = valoarea minimă absolută a lui f(x) în întregul domeniu [-1, 2]

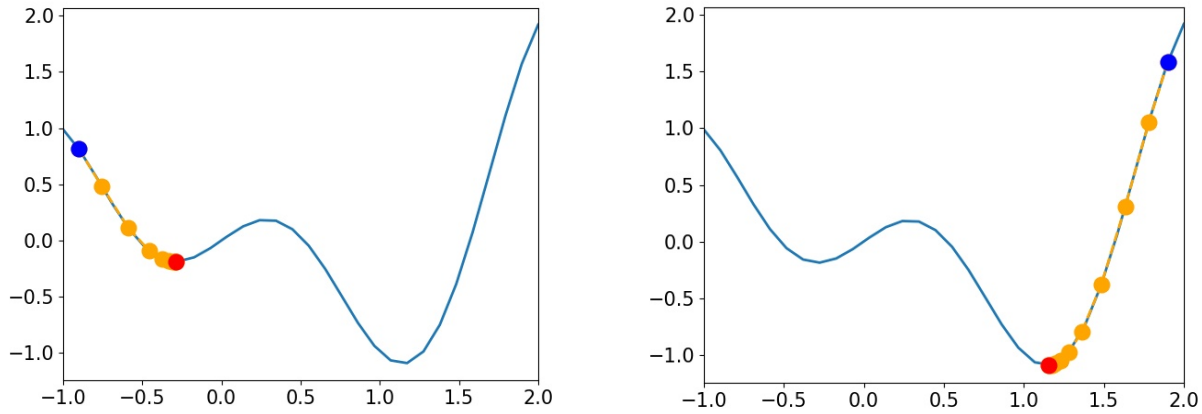


Fig. 3. Cazul unei funcții cu un minim local și unul global (a) pornind de la o valoare inițială *x = -0.9*, gradientul descendent identifică minimul local, neputând avansa către cel global; (b) pornind din valoarea inițială *x = 1.9* se ajunge în minimul global, întrucât din *x = 1.9* se poate "coborî" direct către acest minim.

Aplicând metoda gradientului descendent pentru identificarea minimului, în Fig. 3(a) se observă faptul că, pornind din punctul inițial x_{init} = *-0.9*, algoritmul va converge către minimul local *f(x) = -0.187*. Pentru a identifica minimul global, în Fig. 3(b) se pornește din x_{init} = *1.9*, situație în care minimul global *f(x) = -1.096* este corect identificat. Așadar, prin metoda gradientului dencendent se poate identifica minimul (local sau global) către care se poate "coborî" direct din punctul inițial. În Fig. 3(a) metoda se blochează în minimul local fără să se poată căuta mai departe un alt minim. Aceasta este o limitare a metodei gradientului descendent. O posibilă modalitate de a depăși această limitare este aplicarea algoritmului în repetate rânduri, pornind de la valori multiple ale lui x_{init}, în ideea că minimul global se va identifica în cel puțin un caz.

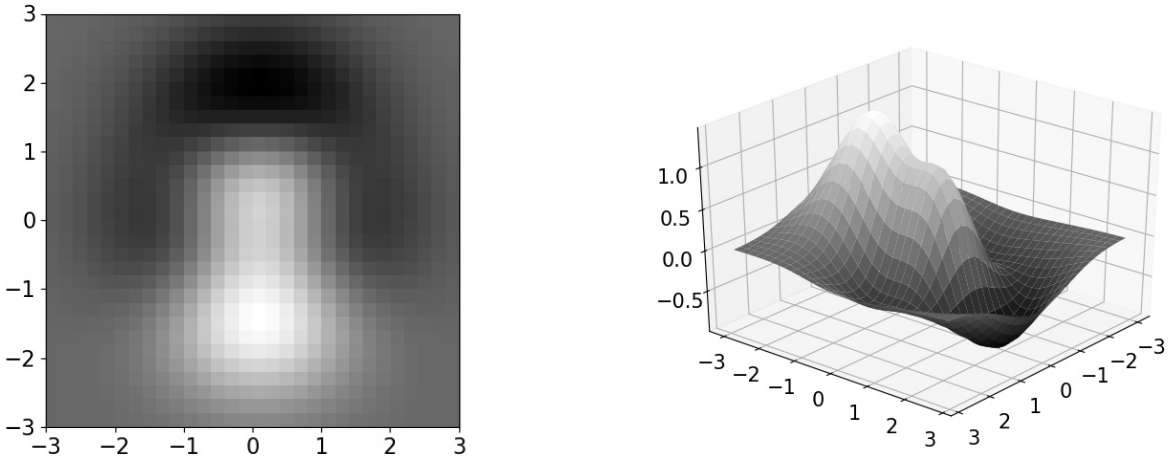


Fig. 4. Graficul funcției f(x, y): (a) reprezentare 2D sub forma unei imagini, unde valorile funcției sunt reprezentate prin valori grayscale; (b) reprezentare 3D sub forma unei suprafețe, unde valorile funcției sunt reprezentate pe axa verticală OZ.

Funcții multidimensionale

Considerăm funcția $f(x,y) = (1-x^2-y^3)e^{\frac{-x^2-y^2}{2}}$, $f:[-3,3]^2 \to \mathbb{R}$, cu graficul din Fig. 4.

Echivalentul derivatei pentru funcții multidimensionale (funcții cu cel puțin doi parametri) este **gradientul** – vectorul format din derivatele parțiale ale funcției în raport cu parametrii acesteia:

$$\nabla f(x,y) = \begin{bmatrix} \frac{\delta f(x,y)}{\delta x} \\ \frac{\delta f(x,y)}{\delta y} \end{bmatrix}$$

În Fig. 5 se ilustrează vectorii gradient ai funcției menționate. Se observă faptul că, în orice punct (x, y), direcția gradientului este cea în care funcția înregistrează creșterea maximă (i.e. direcția "cea mai abruptă" de pe graficul funcției). În mod similar, în direcția opusă gradientului, funcția înregistrează scăderea maximă (i.e. în direcția opusă gradientului "se coboară" cel mai rapid pe graficul funcției).

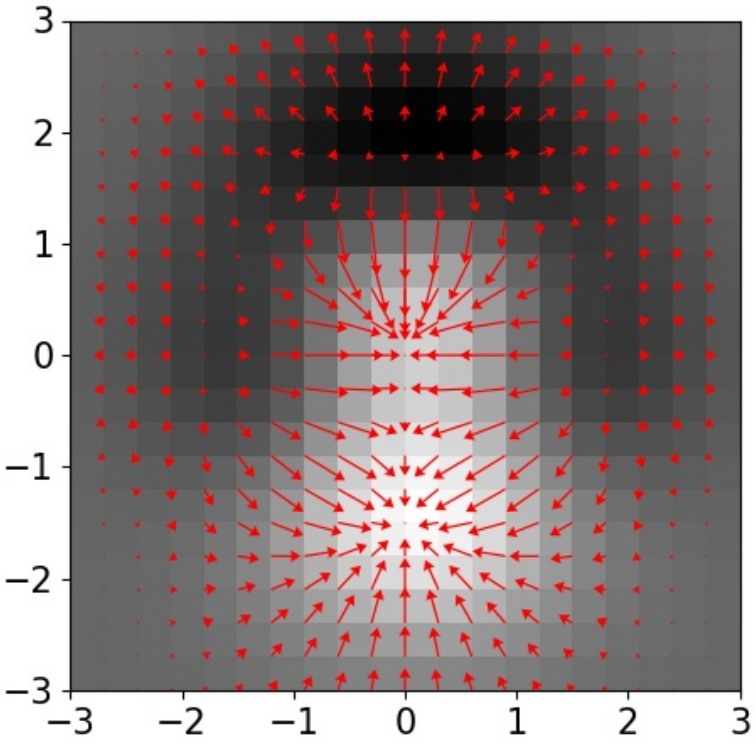


Fig. 5. Reprezentare a vectorilor gradient. Aceștia indică direcțiile în care funcția înregistrează creșterea maximă.

Pe același principiu ca și în cazul funcției unidimensionale *f(x)* prezentate anterior, aplicând **gradientul descendent** se poate identifica un minim al funcției *f(x, y)* ajustând x și y treptat, în sensul opus gradientului (sensul în care se găsește un minim):

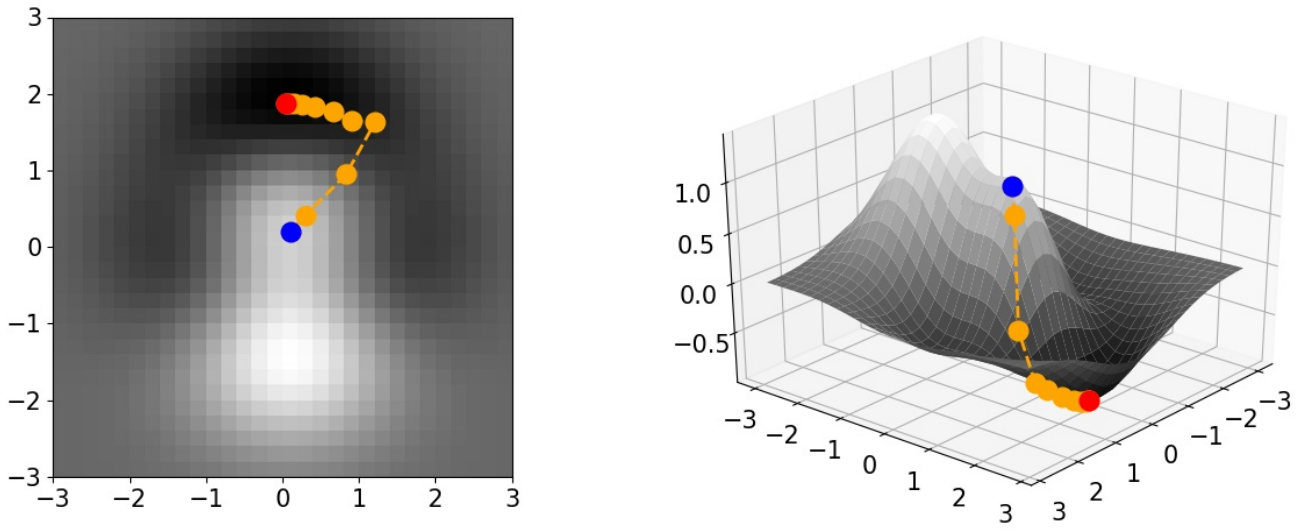
$x = x_{init}$

$y = y_{init}$

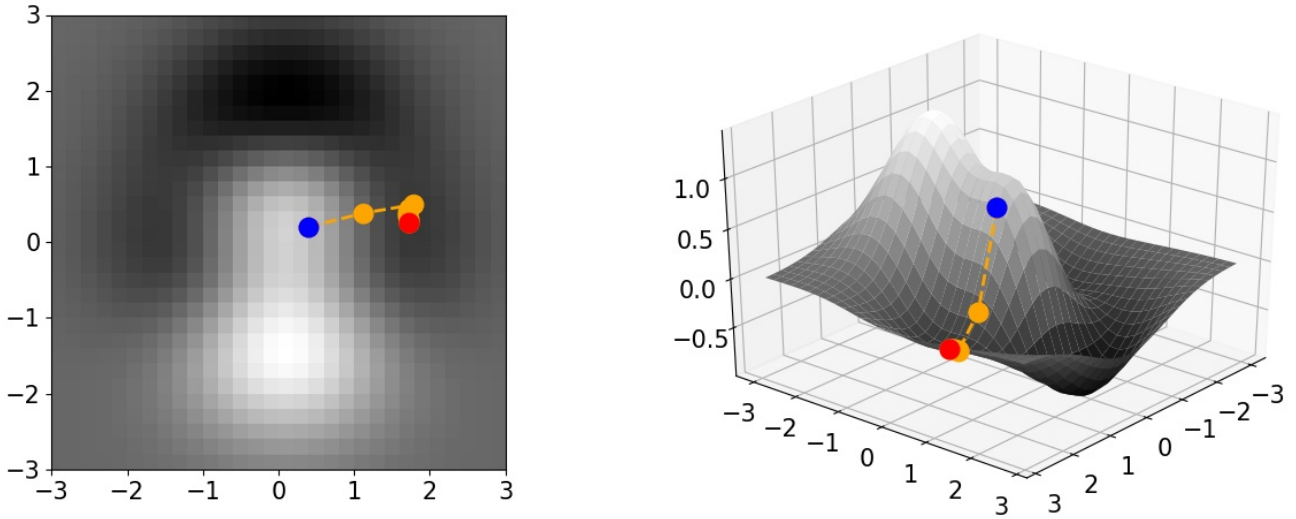
while not $\min(f(x, y))$:

$$x = x - \alpha \frac{\delta f(x, y)}{\delta x}$$

$$y = y - \alpha \frac{\delta f(x, y)}{\delta y}$$



(a)



(b)

Fig. 6. Minimizarea funcției prin metoda gradientului descendent. Pornind din punctul inițial (albastru), se ajunge în minim (roșu) după mai multe iterații ce presupun ajustări repetate ale parametrilor *(x, y)*. Funcție de valoarea inițială, se poate ajunge (a) în minimul global al funcției, sau (b) într-un minim local

Rolul ratei de învățare *α* și modalitatea de stabilire a criteriilor de convergență sunt aceleași ca în cazul funcției unidimensionale *f(x)* prezentate anterior. Pașii gradientului descendent sunt ilustrați în Fig 6. Pornind dintr-un punct inițial (x_{init}, y_{init}), în decursul mai multor iterații se avansează înspre minimul funcției pe traseul cel mai rapid (cel în care "se coboară" cel mai repede din punctul inițial în punctul minim). Funcție de poziția inițială, punctul minim în care se ajunge poate fi minimul global (Fig 6(a)) sau un minim local (Fig 6(b)), similar cu situația descrisă în cazul funcției unidimensionale.

Gradientul descendent se aplică în mod similar pentru funcții multidimensionale cu un număr oarecare de dimensiuni. Considerând funcția $f(x_1, x_2, ..., x_n): D_1 x D_2 x ... x D_n \to \mathbb{R}$, unde D_i sunt domeniile parametrilor $x_i, i = 1..n$, o iterație a gradientului descendent presupune, în mod similar cu situațiile descrise anterior, ajustarea parametrilor folosind gradientul funcției și rata de învățare *α*:

$$x_i = x_i - \alpha \frac{\delta f}{\delta x_i}, \ for \ i = 1..N$$

Aproximarea gradientului prin diferențe finite

Există numeroase situații în care gradientul analitic (formulele matematice ale derivatelor parțiale) fie este prea complicat pentru a se putea utiliza, fie nu se poate determina în fiecare punct (funcția nu este derivabilă – cazul frecvent al funcțiilor afectate de zgomot). În această situație, gradientul într-un punct se poate aproxima prin diferența dintre valorile funcției în două puncte situate de o parte și de alta în imediata vecinătate a punctului vizat. Un exemplu în acest sens este ilustrat în Fig. 7.

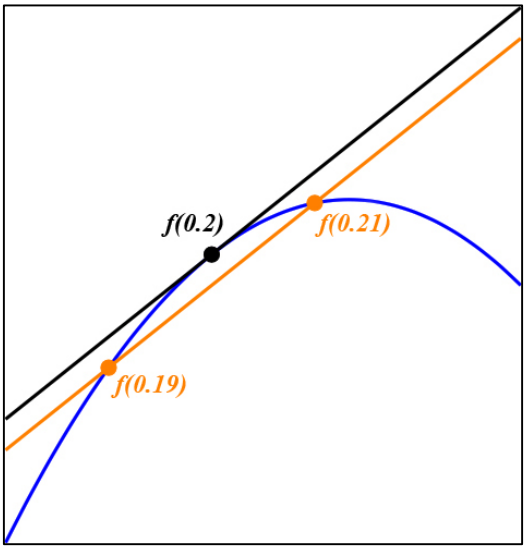


Fig. 7. Aproximarea prin diferențe finite a derivatei unei funcții f(x). Pentru x=0.2 se stabilesc doi vecini de o parte și de alta a sa, la distanța Δ=0.01. Derivata funcției f(0.2) în punctul x = 0.2 se aproximează prin diferența celor doi vecini: *f'(0.2) ≈ (f(0.2+0.01) – f(0.2-0.01))/(2*0.01)*

Pentru o funcție unidimensională *f(x)*, derivata se poate aproxima astfel:

$$f'(x) \approx \frac{f(x+\Delta) - f(x-\Delta)}{2\Delta}$$

Parametrul Δ (“delta”) are o valoare reală pozitivă și în general de mici dimensiuni. Valorile mai mici ale lui Δ determină o aproximare mai bună a gradientului analitic, iar valorile mai mari cauzează o aproximare mai grosieră a gradientului, dar mai robustă la zgomot, probleme cauzate de frecvențe de eșantionare prea mici (aliasing), etc.

Pentru o funcție bidimensională *f(x, y)*, gradientul se poate aproxima astfel:

$$\nabla f(x,y) \approx \begin{bmatrix} \frac{f(x+\Delta, y) - f(x-\Delta, y)}{2\Delta} \\ \frac{f(x, y+\Delta) - f(x, y-\Delta)}{2\Delta} \end{bmatrix}$$

În caz general, pentru o funcție n-dimensională $f(x_1,\ldots,x_i,\ldots,x_n)$, componenta *i* a gradientului se aproximează astfel:

$$\frac{\delta f(x_1,...,x_i,...,x_n)}{\delta x_i} \approx \frac{f(x_1,...,x_i+\Delta,...,x_n) - f(x_1,...,x_i-\Delta,...,x_n)}{2\Delta}$$

Cerințe:

Minimizați următoarele funcții folosind metoda gradientului descendent:

Obs:

- pentru fiecare funcție, determinați minimul global și reprezentați grafic pașii gradientului descendent, după modelul ilustrat în figurile din documentație.
- la fiecare utilizare a gradientului descendent, aplicați cele trei criterii de convergență descrise în documentație.
- pentru determinarea minimului global este posibil să fie necesară aplicarea gradientului descendent pornind din mai multe puncte inițiale, generate aleator sau repartizate egal în domeniul funcției.

1) $f(x) = x^4 - 7x^3 + 14x^2 - 8x, \quad f : [-0.2, 4.4] \to \mathbb{R}$

Folosiți derivata analitică.

Pentru verificare:

- minim local: f(x) = -1.383, x=0.392
- minim global: f(x) = -6.9141, x=3.326

2) $f(x) = \frac{\sin(\sqrt{x})}{x}, \quad f : [1, 40] \to \mathbb{R}$

Folosiți derivata aproximată prin diferențe finite.

Pentru verificare:

- minim global: f(x) = -0.0496, x=18.274

3) $f(x, y) = x^4 + 2x^2y - 21x^2 + 2xy^2 - 14x + y^4 - 13y^2 - 22y + 170, \quad f : [-4, 4]^2 \to \mathbb{R}$

Folosiți gradientul analitic.

Pentru verificare:

- minime globale:

f(x, y) = 0, (x, y) = (3, 2)

f(x, y) = 0, (x, y) = (-3.779, -3.283)

f(x, y) = 0, (x, y) = (-2.805, 3.131)

f(x, y) = 0, (x, y) = (3.584, -1.848)

4) $f(x, y) = (1 - x^2 - y^3)e^{\frac{-x^2-y^2}{2}}, \quad f : [-3, 3]^2 \to \mathbb{R}$

Folosiți gradientul aproximat prin diferențe finite.

Pentru verificare:

- minime locale:

f(x,y) = -0.44626, (x, y) = (1.732, 0)

f(x,y) = -0.44626, (x, y) = (-1.732, 0)

- minim global:

f(x, y) = -0.964, (x, y) = (0, 1.879)