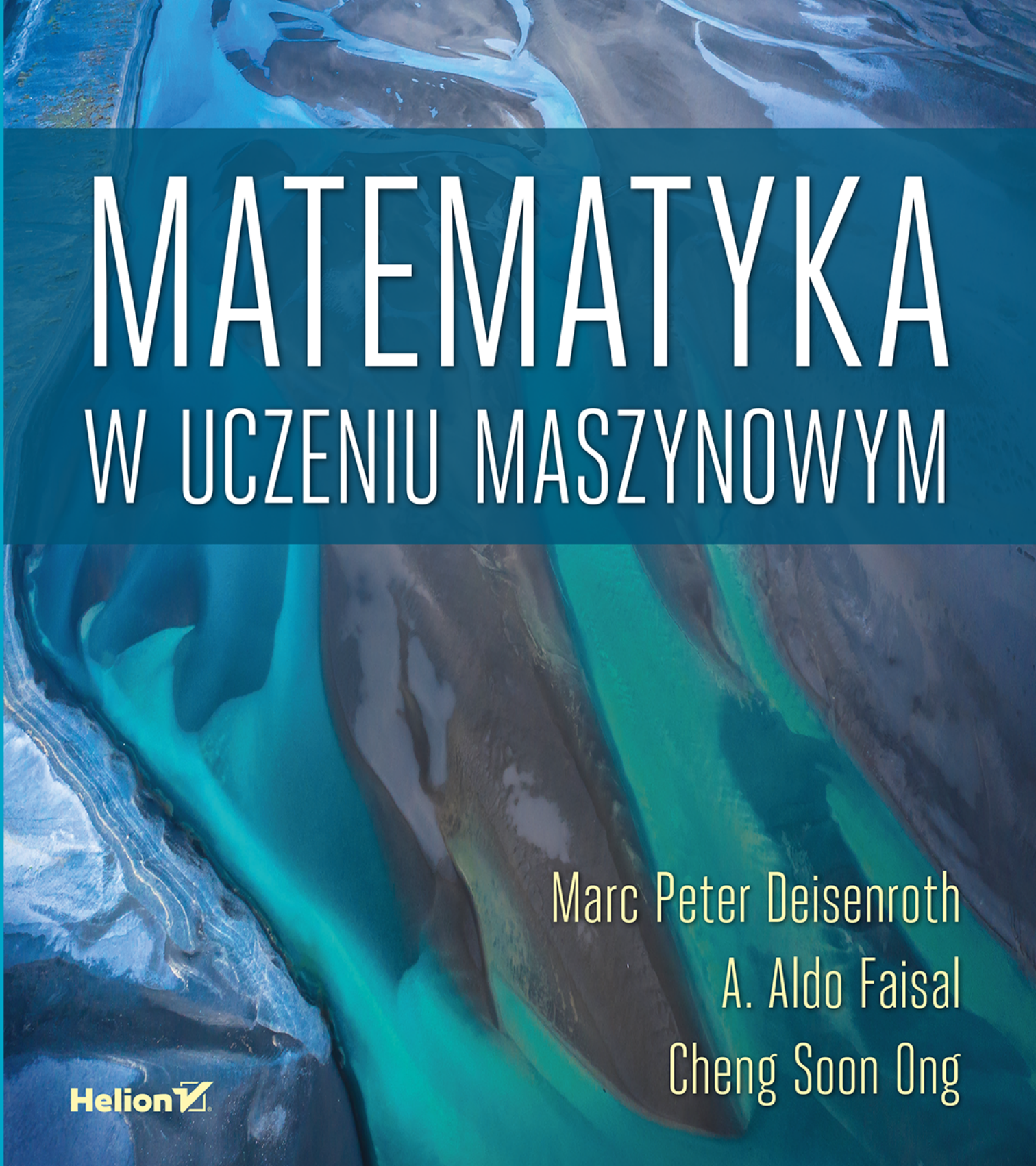

MATEMATYKA
W UCZENIU MASZYNOWYM
Marc Peter Deisenroth
A. Aldo Faisal
Cheng Soon Ong
Helion

Tytuł oryginału: Mathematics for Machine Learning

Tłumaczenie: Filip Kamiński

ISBN: 978-83-283-8459-0

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/mawuma
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

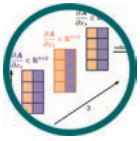
Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

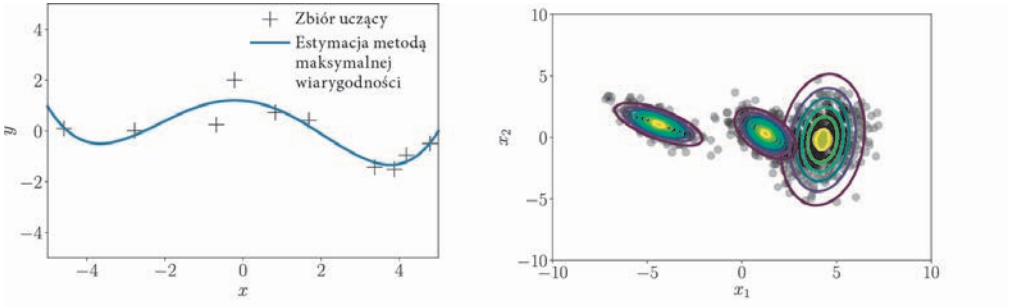
Spis treści

5

Rachunek wektorowy



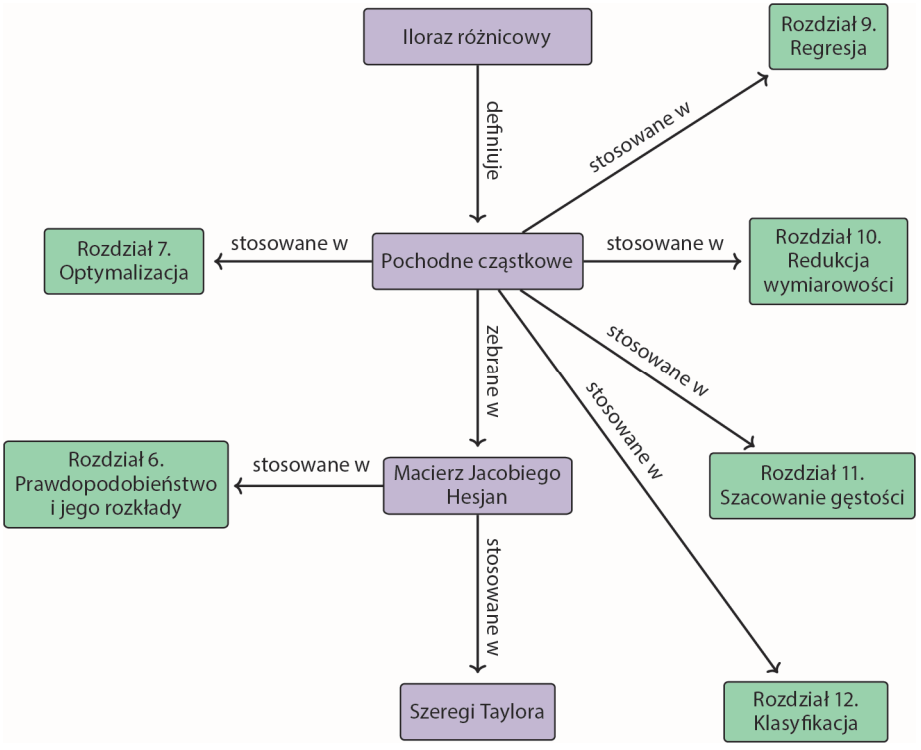
W wielu algorytmach uczenia maszynowego wykorzystuje się optymalizację funkcji celu względem parametrów modelu. Parametry te kontrolują to, jak dobrze model wyjaśnia dane. Problem doboru odpowiednich parametrów można sformułować w postaci problemu optymalizacyjnego (patrz podrozdziały 8.2 i 8.3). Przykładami takich problemów są: (i) regresja liniowa (rozdział 9.), w której przyglądamy się problemom dopasowania krzywych i optymalizujemy parametry liniowo zależnych wag w celu maksymalizacji wiarygodności; (ii) autoenkodery w sieciach neuronowych wykorzystywane do redukcji liczby wymiarów i kompresji danych; ich parametrami są wagi i biasy każdej warstwy; w przypadku autoenkoderów minimalizujemy błąd rekonstrukcji poprzez wielokrotne wykorzystywanie reguły łańcuchowej; oraz (iii) modele mieszanin rozkładów Gaussa (patrz rozdział 11.) wykorzystywane do modelowania rozkładów danych, w których, aby zmaksymalizować wiarygodność, optymalizujemy parametry opisujące położenie i kształt każdego składnika mieszaniny. Na rysunku 5.1 pokazano wybrane problemy, które rozwiązuje się zazwyczaj za pomocą algorytmów optymalizacji wykorzystujących informacje o gradientach (podrozdział 7.1). Na rysunku 5.2 pokazano, jak pojęcia z tego rozdziału są powiązane ze sobą i z innymi rozdziałami tej książki.



(a) Problem regresji: Dobierz parametry modelu tak, aby krzywa dobrze wyjaśniała obserwacje (zaznaczone krzyżykami)

(b) Szacowanie gęstości za pomocą modeli mieszanin rozkładów Gaussa: Znajdź średnie i kowariancje, które dobrze wyjaśniają dane (zaznaczone kropkami)

RYSUNEK 5.1. Rachunek wektorowy odgrywa kluczową rolę w (a) regresji (dopasowywanie krzywej) i (b) szacowaniu gęstości, czyli modelowaniu rozkładów danych



RYSUNEK 5.2. Mapa myśli prezentująca pojęcia przedstawione w tym rozdziale wraz z miejscami ich użycia w innych częściach książki

Centralnym punktem tego rozdziału jest pojęcie funkcji. Funkcja f to relacja, która wiąże ze sobą dwie wielkości. W tej książce wielkościami tymi są zazwyczaj dane wejściowe $\boldsymbol{x} \in \mathbb{R}^D$ i przewidywania (wartości funkcji) $f(\boldsymbol{x})$, o których zakładamy, że są wartościami rzeczywistymi (jeśli nie zaznaczono inaczej). Zbiór $\mathbb{R}^D$ nazywamy **dziedziną** funkcji f. Wartości $f(\boldsymbol{x})$ nazywamy **zbiorem wartości**, **obrazem** lub **przeciwdziedziną** funkcji f. Funkcje liniowe zostały omówione bardziej szczegółowo w punkcie 2.7.3. Do zapisu funkcji często wykorzystujemy następującą notację:

$$f: \mathbb{R}^D \to \mathbb{R}, \tag{5.1a}$$

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}). \tag{5.1b}$$

Równanie 5.1a mówi nam, że f jest odwzorowaniem prowadzącym z $\mathbb{R}^D$ do $\mathbb{R}$, a równanie 5.1b jawnie określa, że wejściu $\boldsymbol{x}$ przypisujemy dokładnie jedną wartość funkcji $f(\boldsymbol{x})$. Funkcja f przypisuje każdemu wejściu $\boldsymbol{x}$ dokładnie jedną wartość funkcji $f(\boldsymbol{x})$.

Przykład 5.1

Przypomnijmy, że iloczyn skalarny jest szczególnym przypadkiem iloczynu wewnętrznego (podrozdział 3.2). W powyższej notacji funkcję $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^2$ zapisalibyśmy jako

$$f: \mathbb{R}^2 \to \mathbb{R} \tag{5.2a}$$

$$\boldsymbol{x} \mapsto x_1^2 + x_2^2. \tag{5.2b}$$

W tym rozdziale omówimy sposoby obliczania gradientów funkcji. Często gradienty są narzędziem niezbędnym do uproszczenia procesu uczenia modeli uczenia maszynowego, ponieważ wskazują kierunek najszybszego wzrostu funkcji. Z tego powodu rachunek wektorowy jest jednym z podstawowych narzędzi matematycznych wykorzystywanych w uczeniu maszynowym. W tej książce zakładamy, że badane przez nas funkcje są różniczkowalne. Po zastosowaniu pewnych nieomówionych w tej książce definicji wiele zaprezentowanych tu podejść można rozszerzyć również na funkcje subróżniczkowalne (funkcje, które są ciągłe, ale w pewnych punktach nie są różniczkowe). W rozdziale 7. znajdziesz przykład takiego rozszerzenia na funkcje z ograniczeniami.

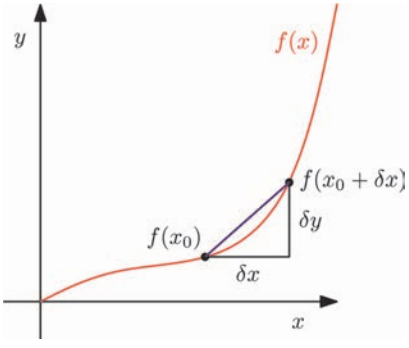
5.1. Różniczkowanie funkcji jednowymiarowych

Poniżej pokrótce omówimy różniczkowanie funkcji jednej zmiennej, które możesz znać z lekcji matematyki w szkole średniej. Zaczniemy od ilorazu różnicowego jednowymiarowej funkcji $y = f(x), x, y \in \mathbb{R}$. Następnie na jego podstawie zdefiniujemy pochodną.

Definicja 5.1 (iloraz różnicowy)**. Iloraz różnicowy**

$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x} \tag{5.3}$$

opisuje nachylenie siecznej przechodzącej przez dwa punkty na wykresie funkcji f. Na rysunku 5.3 są nimi punkty, których współrzędne na osi x są równe x_0 i $x_0 + \delta x$.



Rysunek 5.3. Średnie nachylenie funkcji f pomiędzy x_0 i $x_0 + \delta x$ jest równe nachyleniu siecznej (kolor niebieski) przechodzącej przez $f(x_0)$ i $f(x_0 + \delta x)$. To nachylenie to $\delta y / \delta x$

Jeżeli przyjmiemy, że f jest funkcją liniową, to za iloraz różnicowy można również uznać średnie nachylenie f pomiędzy x_0 i $x_0 + \delta x$. Jeśli f jest różniczkowalna, to w granicy $\delta x \to 0$ otrzymamy styczną do f w punkcie x. W takim przypadku styczna ta jest pochodną f w punkcie x.

Definicja 5.2 (pochodna)**.** Bardziej formalnie: dla $h > 0$ **pochodna** f względem x jest zdefiniowana jako granica

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}. \tag{5.4}$$

Sieczna z rysunku 5.3 staje się w tej definicji styczną.
Pochodna wskazuje kierunek najszybszego wzrostu wartości f.

Przykład 5.2 (pochodna wielomianu)

Chcemy obliczyć pochodną funkcji $f(x) = x^n$, gdzie $n \in \mathbb{N}$. Być może wiesz już, że będzie ona równa nx^{n-1}, ale w tym przykładzie chcemy wyprowadzić tę wartość z definicji pochodnej jako granicy ilorazu różnicowego.

Z definicji pochodnej z równania 5.4 otrzymujemy

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{5.5a}$$

$$= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \tag{5.5b}$$

$$= \lim_{h\to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i}h^i - x^n}{h}. \tag{5.5c}$$

Zauważ, że $x^n = \binom{n}{0} x^{n-0}h^0$. Z tego powodu x^n upraszcza się, a w liczniku pozostaje jedynie suma indeksowana od jeden:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i}h^i}{h} \tag{5.6a}$$

$$= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i}h^{i-1} \tag{5.6b}$$

$$= \lim_{h\to 0} \binom{n}{1} x^{n-1} + \underbrace{\sum_{i=2}^{n} \binom{n}{i} x^{n-i}h^{i-1}}_{\text{dąży do 0 gdy } h\to 0} \tag{5.6c}$$

$$= \frac{n!}{1!\,(n-1)!} x^{n-1} = nx^{n-1}. \tag{5.6d}$$

5.1.1. Szereg Taylora

Szereg Taylora jest reprezentacją funkcji f w postaci nieskończonej sumy. Składniki tej sumy są określane przy pomocy pochodnych f w punkcie x_0.

$t^0 := 1$ dla wszystkich $t \in \mathbb{R}$.

Definicja 5.3 (wielomian Taylora)**. Wielomian Taylora** n-tego stopnia funkcji $f: \mathbb{R} \to \mathbb{R}$ w x_0 jest zdefiniowany jako

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \tag{5.7}$$

gdzie $f^{(k)}(x_0)$ jest k-tą pochodną f w x_0 (zakładamy, że pochodna istnieje), a $\frac{f^{(k)}(x_0)}{k!}$ to współczynniki wielomianu.

Definicja 5.4 (szereg Taylora)**.** Dla gładkiej funkcji $f \in C^\infty, f: \mathbb{R} \to \mathbb{R}$ **szeregiem Taylora** f w x_0 nazywamy

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{5.8}$$

$f \in C^\infty$ oznacza, że f jest ciągła i różniczkowalna nieskończenie wiele razy.

Dla $x_0 = 0$ otrzymujemy specjalny przypadek szeregu Taylora, nazywany **szeregiem Maclaurina**. Jeśli $f(x) = T_\infty(x)$, to f nazywamy **funkcją analityczną**.

Uwaga. Wielomian Taylora n-tego stopnia jest przybliżeniem funkcji, która niekoniecznie jest wielomianem. Wielomian Taylora jest podobny do funkcji f w otoczeniu punktu x_0. Wielomian Taylora n-tego stopnia jest dokładną reprezentacją funkcji wielomianowej f k-tego stopnia takiej, że $k \leqslant n$, ponieważ wszystkie pochodne $f^{(i)}$ dla $i > k$ są równe zero.

Przykład 5.3 (wielomian Taylora)

Rozważamy wielomian

$$f(x) = x^4. \tag{5.9}$$

Chcemy wyznaczyć wielomian Taylora T_6 w punkcie $x_0 = 1$. Zaczynamy od znalezienia współczynników $f^{(k)}(1)$ dla $k = 0, \ldots, 6$:

$$f(1) = 1 \tag{5.10}$$

$$f'(1) = 4 \tag{5.11}$$

$$f''(1) = 12 \tag{5.12}$$

$$f^{(3)}(1) = 24 \tag{5.13}$$

$$f^{(4)}(1) = 24 \tag{5.14}$$

$$f^{(5)}(1) = 0 \tag{5.15}$$

$$f^{(6)}(1) = 0. \tag{5.16}$$

Po podstawieniu otrzymujemy następujący wielomian Taylora

$$T_6(x) = \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.17a}$$

$$= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0. \tag{5.17b}$$

Po wymnożeniu i zmianie kolejności składników daje to

$$T_6(x) = (1 - 4 + 6 - 4 + 1) + x(4 - 12 + 12 - 4) + x^2(6 - 12 + 6) + x^3(4 - 4) + x^4 \tag{5.18a}$$

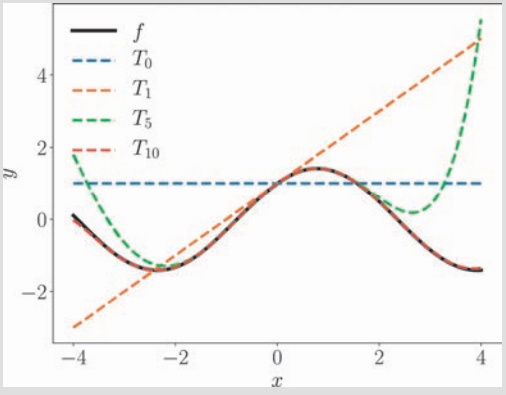
$$= x^4 = f(x). \tag{5.18b}$$

Otrzymaliśmy więc dokładną reprezentację naszej pierwotnej funkcji.

Przykład 5.4 (szereg Taylora)

Rozważmy funkcję z rysunku 5.4, która jest zdefiniowana w następujący sposób:

$$f(x) = \sin(x) + \cos(x) \in \mathcal{C}^\infty. \tag{5.19}$$



Rysunek 5.4. Wielomiany Taylora. Oryginalna funkcja $f(x) = \sin(x) + \cos(x)$ (czarna ciągła linia) jest aproksymowana wielomianami Taylora (kolorowe kreski) w otoczeniu punktu $x_0 = 0$. Wielomiany Taylora wyższego rzędu przybliżają funkcję f lepiej i bardziej globalnie. Wielomian T_{10} jest podobny do f w przedziale $[-4, 4]$

Chcemy znaleźć rozwinięcie f w szereg Taylora w punkcie $x_0 = 0$, czyli rozwinięcie f w szereg Maclaurina. Funkcja f ma następujące pochodne:

$$f(0) = \sin(0) + \cos(0) = 1 \tag{5.20}$$

$$f'(0) = \cos(0) - \sin(0) = 1 \tag{5.21}$$

$$f''(0) = -\sin(0) - \cos(0) = -1 \tag{5.22}$$

$$f^{(3)}(0) = -\cos(0) + \sin(0) = -1 \tag{5.23}$$

$$f^{(4)}(0) = \sin(0) + \cos(0) = f(0) = 1 \tag{5.24}$$

$$\vdots$$

Na powyższej liście pochodnych możemy zauważyć pewną prawidłowość. Współczynniki w szeregu Taylora będą równe jedynie ±1 (ponieważ $\sin(0) = 0$). Każda z tych wartości występuje w dwóch kolejnych wyrazach, po których następuje zmiana wartości na przeciwną. Ponadto $f^{(k+4)}(0) = f^{(k)}(0)$.

Dlatego pełnym rozwinięciem f w szereg Taylora w punkcie $x_0 = 0$ jest

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.25a}$$

$$= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \cdots \tag{5.25b}$$

$$= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \cdots + x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \cdots \tag{5.25c}$$

$$= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!}x^{2k+1} \tag{5.25d}$$

$$= \cos(x) + \sin(x). \tag{5.25e}$$

W powyższych przekształceniach wykorzystaliśmy reprezentację funkcji trygonometrycznych w postaci **szeregów potęgowych**:

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!}x^{2k} \tag{5.26}$$

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!}x^{2k+1}. \tag{5.27}$$

Na rysunku 5.4 pokazano kilka pierwszych wielomianów T_n dla $n = 0, 1, 5, 10$.

Uwaga. Szereg Taylora to szczególny przypadek szeregu potęgowego

$$f(x) = \sum_{k=0}^{\infty} a_k(x - c)^k, \tag{5.28}$$

w którym a_k są współczynnikami, a c jest stałą, która w szeregu Taylora przybiera formę określoną w definicji 5.4.

5.1.2. Zasady różniczkowania

Poniżej zamieściliśmy krótką listę podstawowych reguł różniczkowania, w których pochodną f oznaczyliśmy symbolem f'.

Pochodna iloczynu: $$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x) \tag{5.29}$$

Pochodna ilorazu: $$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \tag{5.30}$$

Pochodna sumy: $$(f(x) + g(x))' = f'(x) + g'(x) \tag{5.31}$$

Reguła łańcuchowa: $$(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x) \tag{5.32}$$

Zapis $g \circ f$ oznacza złożenie funkcji $x \mapsto f(x) \mapsto g(f(x))$.

Przykład 5.5 (reguła łańcuchowa)

Wykorzystajmy regułę łańcuchową do znalezienia pochodnej $h(x) = (2x + 1)^4$:

$$h(x) = (2x + 1)^4 = g(f(x)), \tag{5.33}$$

$$f(x) = 2x + 1, \tag{5.34}$$

$$g(f) = f^4. \tag{5.35}$$

Obliczamy pochodne f i g

$$f'(x) = 2, \tag{5.36}$$

$$g'(f) = 4f^3. \tag{5.37}$$

Pochodna funkcji h jest więc równa

$$h'(x) = g'(f)f'(x) = (4f^3) \cdot 2 \overset{(5.34)}{=} 4(2x + 1)^3 \cdot 2 = 8(2x + 1)^3. \tag{5.38}$$

Do wyznaczenia pochodnej wykorzystaliśmy regułę łańcuchową (równanie 5.32). W pochodnej $g'(f)$ wykorzystaliśmy definicję funkcji f.

5.2. Pochodne cząstkowe i gradienty

W podrozdziale 5.1 przedstawiliśmy różniczkowanie funkcji f zmiennej skalarnej $x \in \mathbb{R}$. Poniżej rozważymy ogólny przypadek, w którym funkcja f zależy od jednej lub więcej zmiennych $\boldsymbol{x} \in \mathbb{R}^n$, np. $f(\boldsymbol{x}) = f(x_1, x_2)$. Uogólnienie pochodnej na funkcje kilku zmiennych nazywamy **gradientem**.

Gradient funkcji f względem wektora $\boldsymbol{x}$ obliczamy, *analizując jedną zmienną naraz*. Pozostałe zmienne traktujemy jako stałe. Gradient jest zatem zbiorem **pochodnych cząstkowych**.

Definicja 5.5 (pochodna cząstkowa)**. Pochodne cząstkowe** funkcji $f: \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$ n-zmiennych $x_1, \ldots, x_n$ są zdefiniowane w następujący sposób:

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= \lim_{h\to 0}\frac{f(x_1+h,x_2,\ldots,x_n)-f(\boldsymbol{x})}{h}\\ &\vdots\\ \frac{\partial f}{\partial x_n} &= \lim_{h\to 0}\frac{f(x_1,\ldots,x_{n-1},x_n+h)-f(\boldsymbol{x})}{h}.\end{aligned} \tag{5.39}$$

Pochodne te zapisuje się w postaci wektora wierszowego:

$$\nabla_x f = \operatorname{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix}\frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n}\end{bmatrix} \in \mathbb{R}^{1\times n}, \tag{5.40}$$

gdzie n to liczba zmiennych, a 1 to wymiar obrazu/przeciwdziedziny/zbioru wartości funkcji f. W powyższych równaniach $\boldsymbol{x}$ jest wektorem kolumnowym: $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$. Wektor wierszowy z równania 5.40 nazywany jest **gradientem** funkcji f lub jej *macierzą Jacobiego*. Wektor ten jest uogólnieniem pochodnej z podrozdziału 5.1.

Uwaga. Powyższa definicja macierzy Jacobiego jest szczególnym przypadkiem ogólnej definicji macierzy Jacobiego (macierzy pochodnych cząstkowych) dla funkcji o wartościach wektorowych. Powrócimy do niej w podrozdziale 5.3.

Przykład 5.6 (obliczanie pochodnych cząstkowych za pomocą reguły łańcuchowej)

W obliczeniach możemy wykorzystać wyniki z różniczkowania funkcji jednej zmiennej. Każda pochodna cząstkowa jest pochodną względem jednej zmiennej.

Funkcja $f(x,y) = (x+2y^3)^2$ ma następujące pochodne cząstkowe

$$\frac{\partial f(x,y)}{\partial x} = 2(x+2y^3)\frac{\partial}{\partial x}(x+2y^3) = 2(x+2y^3), \tag{5.41}$$

$$\frac{\partial f(x,y)}{\partial y} = 2(x+2y^3)\frac{\partial}{\partial y}(x+2y^3) = 12(x+2y^3)y^2. \tag{5.42}$$

Do ich obliczenia wykorzystaliśmy regułę łańcuchową (równanie 5.32).

Uwaga (gradient jako wektor wierszowy). W literaturze nierzadko definiuje się gradient jako wektor kolumnowy. Wynika to z ogólnie przyjętej konwencji, która zakłada, że wektory są zazwyczaj wektorami kolumnowymi. Istnieją dwa powody, dla których zdefiniowaliśmy wektor gradientu jako wektor wierszowy. Po pierwsze, pozwoli nam to w konsekwentny sposób uogólnić gradient na funkcje o wartościach wektorowych $f\colon \mathbb{R}^n \to \mathbb{R}^m$ (gradient stanie się macierzą). Po drugie, taka reprezentacja pozwala nam zastosować regułę łańcuchową dla funkcji wielowymiarowych bez zwracania uwagi na wymiary gradientu. Oba zagadnienia omówimy w podrozdziale 5.3.

Przykład 5.7 (gradient)

Pochodne cząstkowe (pochodne f względem x_1 *i* x_2) funkcji $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$ to

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3 \tag{5.43}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2. \tag{5.44}$$

Gradientem f jest

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(x_1, x_2)}{\partial x_1} \frac{\partial f(x_1, x_2)}{\partial x_2}\right] = [2x_1 x_2 + x_2^3 x_1^2 + 3x_1 x_2^2] \in \mathbb{R}^{1\times 2}. \tag{5.45}$$

5.2.1. Podstawowe zasady obliczania pochodnych cząstkowych

W przypadku wielowymiarowym, w którym $\boldsymbol{x} \in \mathbb{R}^n$, nadal obowiązują podstawowe reguły różniczkowania, które znamy ze szkoły (np. reguła sumy, reguła iloczynu, reguła łańcuchowa; patrz też punkt 5.1.2). Podczas obliczania pochodnych względem wektorów $\boldsymbol{x} \in \mathbb{R}^n$ musimy zachować ostrożność, ponieważ w gradientach będą teraz występowały wektory i macierze, których mnożenie nie jest przemienne (tj. kolejność ma znaczenie, patrz punkt 2.2.1).

Pochodna iloczynu: $(fg)' = f'g + fg'$
Pochodna sumy: $(f + g)' = f' + g'$
Reguła łańcuchowa: $(g(f))' = g'(f)f'$

Oto ogólne reguły pozwalające obliczyć pochodną iloczynu i sumy oraz uogólniona reguła łańcuchowa:

Pochodna iloczynu:
$$\frac{\partial}{\partial \boldsymbol{x}}(f(\boldsymbol{x})g(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x}) \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.46}$$

Pochodna sumy:
$$\frac{\partial}{\partial \boldsymbol{x}}(f(\boldsymbol{x}) + g(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.47}$$

Reguła łańcuchowa:
$$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(g(f(\boldsymbol{x}))) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \boldsymbol{x}} \tag{5.48}$$

To tylko matematycznie niepoprawna intuicja. Pochodne cząstkowe nie są ułamkami.

Przyjrzyjmy się bliżej regule łańcuchowej. Przypomina ona do pewnego stopnia reguły mnożenia macierzy, które wymagają, aby „sąsiadujące" ze sobą wymiary macierzy były ze sobą zgodne (patrz punkt 2.2.1). Spojrzenie na równanie 5.48 od lewej do prawej pozwala zauważyć, że reguła łańcuchowa wykazuje podobne właściwości. ∂f pojawia się w „mianowniku" pierwszego czynnika i w „liczniku" drugiego. Jeśli pomnożymy te dwa „ułamki", to zauważymy, że mnożenie jest możliwe, ponieważ „wymiary" obu ∂f są „zgodne". W efekcie czynnik ∂f ulega „skróceniu" i w wyniku pozostanie jedynie $\partial g / \partial \boldsymbol{x}$.

5.2.2. Reguła łańcuchowa

Rozważmy funkcję $f: \mathbb{R}^2 \to \mathbb{R}$ dwóch zmiennych x_1 i x_2. Niech $x_1(t)$ i $x_2(t)$ będą funkcjami zmiennej t. Aby obliczyć gradient f względem t, musimy zastosować regułę łańcuchową (równanie 5.48) dla funkcji wielowymiarowych:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2}\end{bmatrix}\begin{bmatrix}\frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t}\end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}, \tag{5.49}$$

gdzie d oznacza gradient, a ∂ pochodne cząstkowe.

Przykład 5.8

Jeżeli $f(x_1, x_2) = x_1^2 + 2x_2$ oraz $x_1 = \sin \mathrm{t}$ i $x_2 = \cos t$, to

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{5.50a}$$

$$= 2\sin \mathrm{t}\frac{\partial \sin \mathrm{t}}{\partial t} + 2\frac{\partial \cos t}{\partial t} \tag{5.50b}$$

$$= 2\sin \mathrm{t}\cos t - 2\sin \mathrm{t} = 2\sin \mathrm{t}\,(\cos t - 1) \tag{5.50c}$$

jest pochodną f względem t.

Jeśli $f(x_1, x_2)$ jest funkcją x_1 i x_2, a $x_1(s,t)$ i $x_2(s,t)$ są funkcjami dwóch zmiennych s i t, to za pomocą reguły łańcuchowej możemy obliczyć ich pochodne cząstkowe w następujący sposób:

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}, \tag{5.51}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}. \tag{5.52}$$

Gradient otrzymamy z mnożenia macierzy:

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2}\end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}}\underbrace{\begin{bmatrix}\frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t}\end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}}. \tag{5.53}$$

Regułę łańcuchową możemy zapisać w postaci mnożenia macierzy.

Ten zwięzły sposób zapisu reguły łańcuchowej w postaci mnożenia macierzy ma sens tylko wtedy, gdy gradient jest zdefiniowany jako wektor wierszowy. W przeciwnym razie musielibyśmy transponować gradienty, tak aby zgadzały się wymiary mnożonych macierzy. Nie jest to skomplikowane, o ile gradient jest wektorem lub macierzą. Jednak gdy gradient stanie się tensorem (co omówimy w dalszej części rozdziału), znalezienie transpozycji nie będzie już trywialnym zadaniem.

Uwaga (weryfikacja poprawności implementacji gradientu). Do numerycznego sprawdzenia wyników obliczeń gradientów w programach komputerowych możemy wykorzystać definicję pochodnych cząstkowych w postaci granicy ilorazu różnicowego (patrz równanie 5.39). Do sprawdzenia poprawności implementacji możemy wykorzystać różnice skończone. Wybieramy małą wartość h (np. $h = 10^{-4}$) i porównujemy przybliżenie różnicy skończonej z równania 5.39 z naszą (analityczną) implementacją gradientu. Jeśli błąd jest mały, to prawdopodobnie nasza implementacja gradientu jest prawidłowa. „Mały" może oznaczać, że $\sqrt{\frac{\sum_i (dh_i - df_i)^2}{\sum_i (dh_i + df_i)^2}} < 10^{-6}$, gdzie dh_i jest przybliżeniem za pomocą różnic skończonych, a df_i jest różniczką funkcji względem i-tej zmiennej x_i.

5.3. Gradienty funkcji o wartościach wektorowych

Dotychczas omawialiśmy pochodne cząstkowe i gradienty funkcji $f: \mathbb{R}^n \to \mathbb{R}$ przekształcających wektory na liczby rzeczywiste. Poniżej uogólnimy pojęcie gradientu na funkcje o wartościach wektorowych (pola wektorowe) $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, gdzie $n \geqslant 1$ oraz $m > 1$.

Wektorem wartości funkcji $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, która przyjmuje wektor $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, jest

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m. \tag{5.54}$$

Powyższy zapis pozwala nam traktować funkcję o wartościach wektorowych $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ jako wektor funkcji $[f_1, \ldots, f_m]^\top$ postaci $f_i: \mathbb{R}^n \to \mathbb{R}$, których zbiorem wartości jest $\mathbb{R}$. Różniczkowanie funkcji f_i odbywa się za pomocą reguł różniczkowania omówionych w podrozdziale 5.2.

A zatem pochodną cząstkową funkcji o wartościach wektorowych $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ względem $x_i \in \mathbb{R}$ $(i = 1, \ldots n)$ jest wektor

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots x_n) - f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots x_n) - f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m. \tag{5.55}$$

Z równania 5.40 wiemy, że gradientem $\boldsymbol{f}$ względem wektora jest wektor wierszowy pochodnych cząstkowych. W równaniu 5.55 każda pochodna cząstkowa $\partial \boldsymbol{f} / \partial x_i$ jest wektorem kolumnowym. Gradient $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ względem $\boldsymbol{x} \in \mathbb{R}^n$ to zbiór tych wszystkich pochodnych cząstkowych:

$$\frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1}} & \cdots & \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}} \end{bmatrix} \tag{5.56a}$$

$$= \begin{bmatrix} \boxed{\frac{\partial f_1(\boldsymbol{x})}{\partial x_1}} & \cdots & \boxed{\frac{\partial f_1(\boldsymbol{x})}{\partial x_n}} \\ \vdots & & \vdots \\ \boxed{\frac{\partial f_m(\boldsymbol{x})}{\partial x_1}} & \cdots & \boxed{\frac{\partial f_m(\boldsymbol{x})}{\partial x_n}} \end{bmatrix} \in \mathbb{R}^{m \times n} \tag{5.56b}$$

Gradient funkcji $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ jest macierzą o wymiarach $m \times n$.

Definicja 5.6 (macierz Jacobiego)**.** Zbiór wszystkich pochodnych cząstkowych pierwszego rzędu funkcji o wartościach wektorowych $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ względem wektora $\boldsymbol{x} \in \mathbb{R}^n$ nazywamy **macierzą Jacobiego** (ang. *Jacobian*). Macierz Jacobiego $\boldsymbol{J}$ jest macierzą o wymiarach $m \times n$, którą definiujemy w następujący sposób:

$$\boldsymbol{J} = \nabla_x \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} \quad \cdots \quad \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}\right] \tag{5.57}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}, \tag{5.58}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad J(i,j) = \frac{\partial f_i}{\partial x_j}. \tag{5.59}$$

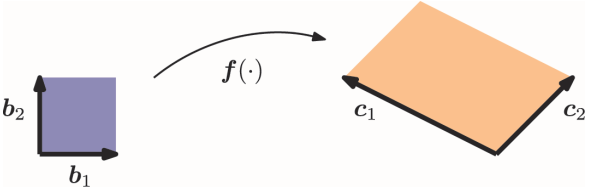
Szczególnym przypadkiem równania 5.58 jest funkcja $f: \mathbb{R}^n \to \mathbb{R}^1$, która odwzorowuje wektor $\boldsymbol{x} \in \mathbb{R}^n$ na skalar (np. $f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$). Macierz Jacobiego tej funkcji jest wektorem wierszowym (macierzą o wymiarach $1 \times n$, patrz równanie 5.40).

Uwaga. W tej książce używamy macierzy Jacobiego zapisanej w postaci licznikowej, tj. układzie, w którym pochodna $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$ funkcji $\boldsymbol{f} \in \mathbb{R}^m$ względem $\boldsymbol{x} \in \mathbb{R}^n$ jest macierzą o wymiarach $m \times n$, w której wiersze odpowiadają elementom $\boldsymbol{f}$, a kolumny elementom $\boldsymbol{x}$ (patrz równanie 5.58). Istnieje również postać mianownikowa, która jest transpozycją postaci licznikowej. W tej książce stosujemy postać licznikową.

W podrozdziale 6.7 zobaczysz przykład wykorzystania macierzy Jacobiego do zmiany zmiennych w rozkładach prawdopodobieństwa. Stopień skalowania spowodowany zmianą zmiennych jest określony przez wyznacznik macierzy Jacobiego (nazywany *jakobianem*).

W podrozdziale 4.1 wspomnieliśmy, że wyznacznik można wykorzystać do obliczenia pola równoległoboku. Jeśli mamy dwa wektory $\boldsymbol{b}_1 = [1,0]^\top$ i $\boldsymbol{b}_2 = [0,1]^\top$ reprezentujące boki kwadratu jednostkowego (zaznaczone kolorem niebieskim, patrz rysunek 5.5), to pole tego kwadratu jest równe

$$\left|\det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\right| = 1. \tag{5.60}$$



Rysunek 5.5. Jakobian funkcji $\boldsymbol{f}$ może być wykorzystany do ustalenia współczynnika skalowania niebieskiego obszaru w pomarańczowy

Wartość bezwzględna poniższego wyznacznika (patrz podrozdział 4.1) odpowiada polu powierzchni równoległoboku o bokach $\boldsymbol{c}_1 = [-2,1]^\mathsf{T}, \boldsymbol{c}_2 = [1,1]^\mathsf{T}$ (na rysunku 5.5 zaznaczony kolorem pomarańczowym)

$$\left|\det\left(\begin{bmatrix}-2 & 1\\ 1 & 1\end{bmatrix}\right)\right| = |-3| = 3. \tag{5.61}$$

Pole tego równoległoboku jest trzykrotnością pola kwadratu jednostkowego. Ten współczynnik skalowania możemy wyznaczyć, znajdując odwzorowanie, które przekształca jeden kształt w drugi. W języku algebry liniowej oznacza to przeprowadzenie zmiany zmiennych z $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ na $(\boldsymbol{c}_1, \boldsymbol{c}_2)$. W naszym przypadku odwzorowanie jest liniowe i wartość bezwzględna wyznacznika tego odwzorowania daje nam poszukiwany współczynnik skalowania.

Poniżej przedstawimy dwa podejścia pozwalające znaleźć to odwzorowanie. W pierwszym wykorzystamy fakt, że odwzorowanie jest liniowe, co pozwoli nam wykorzystać narzędzia z rozdziału 2. do zidentyfikowania odwzorowania. W drugim znajdziemy odwzorowanie za pomocą pochodnych cząstkowych i narzędzi, które omówiliśmy w tym rozdziale.

Podejście 1. W podejściu algebraicznym zauważamy, że $\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ i $\{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ są bazami przestrzeni $\mathbb{R}^2$ (patrz punkt 2.6.1). Interesującym nas odwzorowaniem jest zmiana bazy z $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ na $(\boldsymbol{c}_1, \boldsymbol{c}_2)$. W tym podejściu będziemy więc szukać macierzy transformacji, która realizuje tę zmianę bazy. W oparciu o informacje z punktu 2.7.2 otrzymujemy następującą macierz zmiany bazy

$$\boldsymbol{J} = \begin{bmatrix}-2 & 1\\ 1 & 1\end{bmatrix}. \tag{5.62}$$

$\boldsymbol{J}\boldsymbol{b}_1 = \boldsymbol{c}_1$ a $\boldsymbol{J}\boldsymbol{b}_2 = \boldsymbol{c}_2$. Wartość bezwzględna wyznacznika $\boldsymbol{J}$ jest poszukiwanym współczynnikiem skalowania. $|\det(\boldsymbol{J})| = 3$, co oznacza, że powierzchnia równoległoboku rozpinanego przez $(\boldsymbol{c}_1, \boldsymbol{c}_2)$ jest trzykrotnie większa niż powierzchnia kwadratu rozpinanego przez $(\boldsymbol{b}_1, \boldsymbol{b}_2)$.

Podejście 2. Podejście oparte na algebrze liniowej sprawdza się jedynie w przypadku przekształceń liniowych. W przypadku przekształceń nieliniowych (które odegrają ważną rolę w podrozdziale 6.7) stosujemy bardziej ogólne podejście, wykorzystujące pochodne cząstkowe.

W tym podejściu rozważamy funkcję $\boldsymbol{f}: \mathbb{R}^2 \to \mathbb{R}^2$, która dokonuje zmiany zmiennych. W naszym przykładzie $\boldsymbol{f}$ odwzorowuje współrzędne dowolnego wektora $\boldsymbol{x} \in \mathbb{R}^2$ w bazie $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ na współrzędne $\boldsymbol{y} \in \mathbb{R}^2$ w bazie $(\boldsymbol{c}_1, \boldsymbol{c}_2)$. Chcemy zidentyfikować odwzorowanie, aby ustalić, jak w wyniku przekształcenia $\boldsymbol{f}$ zmienia się powierzchnia (lub objętość). W tym celu musimy ustalić, jak zmienia się $\boldsymbol{f}(\boldsymbol{x})$, jeśli odrobinę zmodyfikujemy $\boldsymbol{x}$. Na to pytanie odpowiada dokładnie macierz Jacobiego $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{2\times 2}$. Możemy więc zapisać, że

$$y_1 = -2x_1 + x_2 \tag{5.63}$$

$$y_2 = x_1 + x_2. \tag{5.64}$$

Z geometrycznego punktu widzenia jakobian określa współczynnik powiększenia/skalowania przekształcanego obszaru lub objętości.

Otrzymaliśmy zależność funkcyjną pomiędzy $\boldsymbol{x}$ i $\boldsymbol{y}$, która pozwala nam obliczyć pochodne cząstkowe

$$\frac{\partial y_1}{\partial x_1} = -2, \qquad \frac{\partial y_1}{\partial x_2} = 1, \qquad \frac{\partial y_2}{\partial x_1} = 1, \qquad \frac{\partial y_2}{\partial x_2} = 1. \tag{5.65}$$

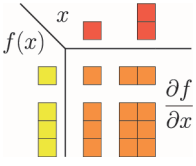
Po ich zapisaniu w macierzy otrzymujemy następującą macierz Jacobiego:

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}. \tag{5.66}$$

Macierz Jacobiego reprezentuje poszukiwaną przez nas transformację współrzędnych. Reprezentacja ta jest dokładna, jeżeli badane przekształcenie jest liniowe (tak jest w tym przypadku). Równanie 5.66 odpowiada dokładnie macierzy zmiany bazy z równania 5.62. Jeśli transformacja współrzędnych jest nieliniowa, to macierz Jacobiego lokalnie aproksymuje tę nieliniową transformację za pomocą odwzorowania liniowego. Wartość bezwzględna wyznacznika macierzy Jacobiego $|\det(\boldsymbol{J})|$ jest współczynnikiem skalowania powierzchni lub objętości podczas przekształcania współrzędnych. W tym przykładzie $|\det(\boldsymbol{J})| = 3$.

Jakobiany i przekształcenia zmiennych odegrają ważną rolę w podrozdziale 6.7, w którym będziemy przekształcać zmienne losowe i rozkłady prawdopodobieństwa. Przekształcenia te są niezwykle istotne w uczeniu głębokich sieci neuronowych z użyciem tzw. *reparametryzacji*, zwanej również **analizą nieskończonych perturbacji**.

W tym rozdziale omówiliśmy pochodne funkcji. Podsumowanie ich wymiarów pokazano na rysunku 5.6. Jeśli $f: \mathbb{R} \to \mathbb{R}$, to gradient jest po prostu skalarem (lewy górny róg rysunku). Dla $f: \mathbb{R}^D \to \mathbb{R}$ gradient jest wektorem wierszowym o wymiarach $1 \times D$ (prawy górny róg rysunku). Dla $\boldsymbol{f}: \mathbb{R} \to \mathbb{R}^E$ gradient jest wektorem kolumnowym o wymiarach $E \times 1$, a dla $\boldsymbol{f}: \mathbb{R}^D \to \mathbb{R}^E$ otrzymujemy macierz gradientu o rozmiarze $E \times D$.



RYSUNEK 5.6. Wymiary pochodnych (cząstkowych)

Przykład 5.9 (gradient funkcji o wartościach wektorowych)

Dane są:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \qquad \boldsymbol{A} \in \mathbb{R}^{M \times N}, \qquad \boldsymbol{x} \in \mathbb{R}^N$$

Aby obliczyć gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$, najpierw określamy jego wymiary. Ponieważ $\boldsymbol{f}: \mathbb{R}^N \to \mathbb{R}^M$, gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x} \in \mathbb{R}^{M \times N}$. Następnie wyznaczamy pochodne cząstkowe $\boldsymbol{f}$ względem każdego x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij} x_j \Longrightarrow \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{5.67}$$

i zapisujemy je w macierzy Jacobiego

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M \times N}. \tag{5.68}$$

Przykład 5.10 (reguła łańcucha)

Rozważmy funkcję $h: \mathbb{R} \to \mathbb{R},\ h(t) = (f \circ g)(t)$, gdzie

$$f: \mathbb{R}^2 \to \mathbb{R} \tag{5.69}$$

$$g: \mathbb{R} \to \mathbb{R}^2 \tag{5.70}$$

$$f(\boldsymbol{x}) = \exp(x_1 x_2^2) \tag{5.71}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix}. \tag{5.72}$$

Wyznaczmy gradient h względem t. Ponieważ $f: \mathbb{R}^2 \to \mathbb{R}$ i $g: \mathbb{R} \to \mathbb{R}^2$, możemy zauważyć, że

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1 \times 2}, \qquad \frac{\partial g}{\partial t} \in \mathbb{R}^{2 \times 1}. \tag{5.73}$$

Interesujący nas gradient obliczamy za pomocą reguły łańcuchowej:

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \tag{5.74a}$$

$$= \begin{bmatrix} \exp(x_1 x_2^2) x_2^2 & 2\exp(x_1 x_2^2) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t \sin t \\ \sin t + t \cos t \end{bmatrix} \tag{5.74b}$$

$$= \exp(x_1 x_2^2)(x_2^2(\cos t - t \sin t) + 2x_1 x_2(\sin t + t \cos t)), \tag{5.74c}$$

gdzie $x_1 = t \cos t$ i $x_2 = t \sin t$; patrz równanie 5.72.

Model ten omówimy znacznie bardziej szczegółowo w rozdziale 9., w którym przedstawimy go w kontekście regresji liniowej. W regresji liniowej wykorzystuje się pochodne minimalnokwadratowej funkcji straty L względem parametrów $\boldsymbol{\theta}$.

Przykład 5.11 (gradient minimalnokwadratowej straty w modelu liniowym)

Rozważmy model liniowy

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}, \tag{5.75}$$

gdzie $\boldsymbol{\theta} \in \mathbb{R}^D$ jest wektorem parametrów, $\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$ to cechy wejściowe, a $y \in \mathbb{R}^N$ to obserwacje. Zdefiniujmy funkcje

$$L(\boldsymbol{e}) := \| \boldsymbol{e} \|^2, \tag{5.76}$$

$$e(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}. \tag{5.77}$$

Interesuje nas gradient $\frac{\partial L}{\partial \boldsymbol{\theta}}$. Do jego wyznaczenia wykorzystamy regułę łańcuchową. L jest nazywana **minimalnokwadratową funkcją straty**.

Przed przystąpieniem do obliczeń określamy wymiary gradientu:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D}. \tag{5.78}$$

Reguła łańcuchowa pozwala nam zapisać gradient jako

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}}\frac{\partial e}{\partial \theta}. \tag{5.79}$$

d-ty element gradientu to

$$\frac{\partial L}{\partial \boldsymbol{\theta}}[1,d] = \sum_{n=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[n]\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[n,d]. \tag{5.80}$$

Korzystamy z faktu, że $\| \boldsymbol{e} \|^2 = \boldsymbol{e}^\top \boldsymbol{e}$ (patrz podrozdział 3.2), który pozwala zapisać, że

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2e^\top \in \mathbb{R}^{1\times N}. \tag{5.81}$$

Ponadto

$$\frac{\partial e}{\partial \theta} = -\Phi \in \mathbb{R}^{N\times D}. \tag{5.82}$$

Naszą pożądaną pochodną jest zatem

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2e^\top\Phi \overset{(5.77)}{=} \underbrace{-2(\boldsymbol{y}^\top - \boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top)}_{1\times N}\underbrace{\Phi}_{N\times D} \in \mathbb{R}^{1\times D}. \tag{5.83}$$

Uwaga. Ten sam wynik możemy otrzymać bez stosowania reguły łańcuchowej. Wystarczy spojrzeć na funkcję

$$L_2(\boldsymbol{\theta}) := \| \boldsymbol{y} - \boldsymbol{\Phi\theta} \|^2 = (\boldsymbol{y} - \boldsymbol{\Phi\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi\theta}). \tag{5.84}$$

To podejście sprawdza się w przypadku prostych funkcji, takich jak L_2, ale staje się niepraktyczne w bardziej złożonych przypadkach.

5.4. Gradienty macierzy

Tensory możemy traktować jako wielowymiarowe tablice.

W niektórych sytuacjach musimy obliczyć gradienty macierzy względem wektorów (lub innych macierzy). Obliczenia te dają w wyniku wielowymiarowe tensory. Tensory możemy traktować jako wielowymiarowe tablice, w których zapisano pochodne cząstkowe. Na przykład gradient macierzy $\boldsymbol{A}$ o wymiarach $m \times n$ względem macierzy $\boldsymbol{B}$ o wymiarach $p \times q$ to obiekt o wymiarach $(m \times n) \times (p \times q)$, tj. czterowymiarowy tensor $\boldsymbol{J}$, którego elementy są zdefiniowane jako $J_{ijkl} = \partial A_{ij} / \partial B_{kl}$.

Ponieważ macierze reprezentują odwzorowania liniowe, możemy wykorzystać fakt istnienia izomorfizmu (liniowego, odwracalnego odwzorowania) przestrzeni wektorowej $\mathbb{R}^{m\times n}$ macierzy $m \times n$ w przestrzeń $\mathbb{R}^{mn}$ mn-elementowych wektorów. Pozwala to przekształcić nasze macierze w wektory

Macierze można przekształcić w wektory, układając kolumny macierzy jedna za drugą (tzw. spłaszczanie).

o długościach mn i pq. Gradienty względem mn-elementowych wektorów to macierze Jacobiego o wymiarach $mn \times pq$. Oba podejścia pokazano na rysunku 5.7. W zastosowaniach praktycznych często chcemy przekształcić macierze w wektory, które pozwolą nam kontynuować pracę z macierzami Jacobiego. W przypadku macierzy Jacobiego reguła łańcuchowa (równanie 5.48) sprowadza się do prostego mnożenia macierzy, podczas gdy w przypadku tensorów Jacobiego musielibyśmy pilnować zgodności wymiarów, względem których sumujemy.

Przykład 5.12 (gradient wektorów względem macierzy)

Rozważmy następujący przykład:

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{f} \in \mathbb{R}^M, \qquad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \qquad \boldsymbol{x} \in \mathbb{R}^N. \tag{5.85}$$

Chcemy wyznaczyć gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{A}$. Ponownie rozpoczynamy od ustalenia wymiarów:

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} \in \mathbb{R}^{M\times(M\times N)}. \tag{5.86}$$

Z definicji gradient jest zbiorem pochodnych cząstkowych:

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}. \tag{5.87}$$

W obliczaniu pochodnych cząstkowych pomoże nam jawne zapisanie formuły iloczynu macierzy z wektorem

$$f_i = \sum_{j=1}^{N} A_{ij}x_j, i = 1, \ldots, M. \tag{5.88}$$

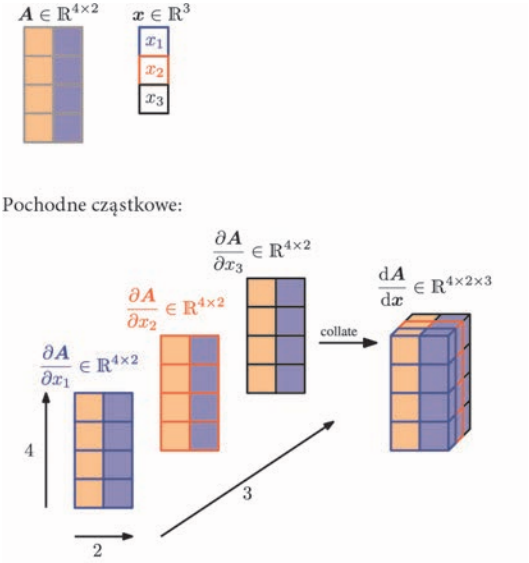
Pochodne cząstkowe mają więc następującą postać:

$$\frac{\partial f_i}{\partial A_{iq}} = x_q. \tag{5.89}$$

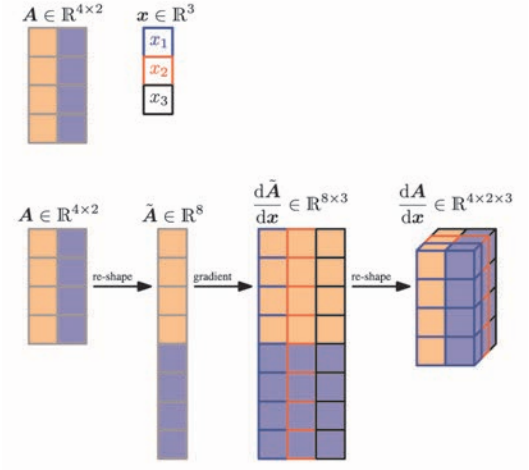
To pozwala nam obliczyć pochodne cząstkowe f_i względem wiersza macierzy $\boldsymbol{A}$:

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^\top \in \mathbb{R}^{1\times 1\times N}, \tag{5.90}$$

$$\frac{\partial f_i}{\partial A_{k\neq i,:}} = \boldsymbol{0}^\top \in \mathbb{R}^{1\times 1\times N}. \tag{5.91}$$



(a) Podejście 1.: Obliczamy pochodne cząstkowe $\frac{dA}{dx_1}, \frac{dA}{dx_2}, \frac{dA}{dx_3}$. Każda z nich jest macierzą o wymiarach 4×2. Pochodne zapisujemy w tensorze o wymiarach $4 \times 2 \times 3$



(b) Podejście 2.: Przekształcamy (spłaszczamy) $A \in \mathbb{R}^{(4\times2)}$ w wektor $\tilde{A} \in \mathbb{R}^8$. Następnie obliczamy gradient $\frac{d\tilde{A}}{dx} \in \mathbb{R}^{8\times3}$. Tensor gradientu otrzymujemy poprzez zastosowanie przekształceń pokazanych na tym rysunku

RYSUNEK 5.7. Wizualizacja procesu obliczania gradientu macierzy względem wektora. Chcemy wyznaczyć gradient $\boldsymbol{A} \in \mathbb{R}^{4\times2}$ względem wektora $\boldsymbol{x} \in \mathbb{R}^3$. Wiemy, że $\frac{d\boldsymbol{A}}{dx} \in \mathbb{R}^{4\times2\times3}$. Do wyznaczenia gradientu stosujemy dwa podejścia: (a) zapisujemy pochodne cząstkowe w tensorze Jacobiego; (b) spłaszczamy macierz do wektora i obliczamy macierz Jacobiego, którą następnie przekształcamy w tensor Jacobiego

Musimy zwrócić uwagę na wymiary. Pochodną cząstkową f_i względem wiersza $\boldsymbol{A}$ jest tensor o rozmiarze $1\times1\times N$, ponieważ przeciwdziedziną f_i jest $\mathbb{R}$, a każdy wiersz $\boldsymbol{A}$ ma rozmiar $1\times N$.

Pochodne cząstkowe z równania 5.91 układamy w stos i otrzymujemy interesujący nas gradient (równanie 5.87):

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \\ \boldsymbol{x}^\top \\ \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \end{bmatrix} \in \mathbb{R}^{1\times(M\times N)}. \tag{5.92}$$

Przykład 5.13 (gradient macierzy względem macierzy)

Rozważmy macierz $\boldsymbol{R} \in \mathbb{R}^{M\times N}$ oraz funkcję $\boldsymbol{f}\colon \mathbb{R}^{M\times N} \to \mathbb{R}^{N\times N}$ zdefiniowaną jako

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}^\top\boldsymbol{R} =: \boldsymbol{K} \in \mathbb{R}^{N\times N}. \tag{5.93}$$

Chcemy znaleźć gradient $\mathrm{d}\boldsymbol{K}/\mathrm{d}\boldsymbol{R}$.

Aby rozwiązać ten trudny problem, najpierw zapiszemy to, co już wiemy:

Gradient ma wymiary

$$\frac{\mathrm{d}\boldsymbol{K}}{\mathrm{d}\boldsymbol{R}} \in \mathbb{R}^{(N\times N)\times(M\times N)}, \tag{5.94}$$

a więc jest tensorem. Ponadto dla $p, q = 1, \ldots, N$

$$\frac{\mathrm{d}K_{pq}}{\mathrm{d}\boldsymbol{R}} \in \mathbb{R}^{1\times M\times N}, \tag{5.95}$$

gdzie K_{pq} jest (p, q)-tym elementem $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$. Jeżeli i-tą kolumnę $\boldsymbol{R}$ oznaczymy przez $\boldsymbol{r}_i$, to każdy element $\boldsymbol{K}$ będzie określony przez iloczyn skalarny dwóch kolumn macierzy $\boldsymbol{R}$:

$$K_{pq} = \boldsymbol{r}_p^\top \boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}R_{mq}. \tag{5.96}$$

Obliczymy teraz pochodną cząstkową $\frac{\partial K_{pq}}{\partial R_{ij}}$:

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial R_{ij}} R_{mp}R_{mq} = \partial_{pqij}, \tag{5.97}$$

gdzie

$$\partial_{pqij} = \begin{cases} R_{iq} & \text{jeżeli } j = p, p \neq q \\ R_{ip} & \text{jeżeli } j = q, p \neq q \\ 2R_{iq} & \text{jeżeli } j = p, p = q \\ 0 & \text{w innych przypadkach} \end{cases}. \tag{5.98}$$

Z równania 5.94 wiemy, że interesujący nas gradient ma wymiar $(N \times N) \times (M \times N)$. Elementami tego tensora są ∂_{pqij} z równania 5.98 $(p, q, j = 1, \ldots, N$, a $i = 1, \ldots, M)$.

5.5. Tożsamości przydatne w obliczeniach gradientów

Poniżej prezentujemy kilka przydatnych tożsamości, które są często wykorzystywane w uczeniu maszynowym (Petersen i Pedersen, 2012). W poniższych równaniach $tr\,(\cdot)$ oznacza ślad macierzy (patrz definicja 4.4), $det\,(\cdot)$ wyznacznik (patrz podrozdział 4.1), a $\boldsymbol{f}(\boldsymbol{X})^{-1}$ odwrotność funkcji $\boldsymbol{f}(\boldsymbol{X})$ (przy założeniu, że taka odwrotność istnieje).

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{\top} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)^{\top} \tag{5.99}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \text{tr}\,(\boldsymbol{f}(\boldsymbol{X})) = \text{tr}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.100}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \det(\boldsymbol{f}(\boldsymbol{X})) = \det(\boldsymbol{f}(\boldsymbol{X})) \text{tr}\left(\boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.101}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} = -\boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} \tag{5.102}$$

$$\frac{\partial \boldsymbol{a}^{\top} \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^{\top} \boldsymbol{a} \boldsymbol{b}^{\top} (\boldsymbol{X}^{-1})^{\top} \tag{5.103}$$

$$\frac{\partial \boldsymbol{x}^{\top} \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\top} \tag{5.104}$$

$$\frac{\partial \boldsymbol{a}^{\top} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\top} \tag{5.105}$$

$$\frac{\partial \boldsymbol{a}^{\top} \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a} \boldsymbol{b}^{\top} \tag{5.106}$$

$$\frac{\partial \boldsymbol{x}^{\top} \boldsymbol{B} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\top} (\boldsymbol{B} + \boldsymbol{B}^{\top}) \tag{5.107}$$

Dla symetrycznych macierzy $\boldsymbol{W}$: $\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\top} \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\top} \boldsymbol{W} \boldsymbol{A}$ (5.108)

Uwaga. W tej książce zajmujemy się tylko śladami i transpozycjami macierzy. Wiesz jednak, że pochodne mogą być tensorami o większej liczbie wymiarów. W tym przypadku zwykły ślad i transpozycja nie są zdefiniowane. Ślad tensora o wymiarach $D \times D \times E \times F$ byłby macierzą o wymiarach $E \times F$. Jest to szczególny przypadek kontrakcji/zwężania tensorów. W podobny sposób, gdy „transponujemy" tensor, mamy na myśli zamianę dwóch jego pierwszych wymiarów. Jeżeli nie zamienimy macierzy na wektory, tak jak omówiono to w podrozdziale 5.4, to w przypadku obliczeń pochodnych funkcji wielu zmiennych $\boldsymbol{f}(\cdot)$ względem macierzy w równaniach od 5.99 do 5.102 pojawią się obliczenia na tensorach.

5.6. Propagacja wsteczna i różniczkowanie automatyczne

Dobre omówienie propagacji wstecznej i reguły łańcuchowej znajdziesz na blogu Tima Viery pod adresem *https://tinyurl.com/ycfm2yrw*.

W wielu rozwiązaniach z obszaru uczenia maszynowego do znalezienia odpowiednich parametrów modelu wykorzystujemy metodę gradientu prostego (podrozdział 7.1), która pozwala obliczyć gradient funkcji celu w odniesieniu do parametrów modelu. Gradient ten możemy wyznaczyć za pomocą reguł rachunku różniczkowego oraz reguły łańcuchowej (patrz punkt 5.2.2). Przykład takiego gradientu pokazaliśmy w podrozdziale 5.3, w którym przyjrzeliśmy się gradientowi minimalnokwadratowej straty w odniesieniu do parametrów modelu regresji liniowej.

Rozważmy funkcję

$$f(x) = \sqrt{x^2 + \exp\left(x^2\right)} + \cos\left(x^2 + \exp\left(x^2\right)\right). \tag{5.109}$$

Skorzystamy z faktu, że różniczkowanie jest operacją liniową. Po zastosowaniu reguły łańcuchowej otrzymujemy następujący gradient:

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} = {} & \frac{2x + 2x\exp\left(x^2\right)}{2\sqrt{x^2 + \exp\left(x^2\right)}} - \sin\left(x^2 + \exp\left(x^2\right)\right)\left(2x + 2x\exp\left(x^2\right)\right) \\ & -2x\left(\frac{1}{2\sqrt{x^2 + \exp\left(x^2\right)}} - \sin\left(x^2 + \exp\left(x^2\right)\right)\right)\left(1 + \exp\left(x^2\right)\right).\end{aligned} \tag{5.110}$$

Zapisywanie gradientów w powyższej jawnej postaci nie jest zbyt praktyczne i często prowadzi do bardzo długich wyrażeń opisujących pochodne. Z praktycznego punktu widzenia oznacza to, że przy nieuważnej implementacji koszt obliczenia gradientu może być znacznie wyższy niż koszt ewaluacji różniczkowanej funkcji. Stosowanie takiego podejścia może więc niepotrzebnie spowolnić nasze obliczenia. W przypadku uczenia głębokich sieci neuronowych wydajnym sposobem obliczania gradientu funkcji błędu w odniesieniu do parametrów modelu jest algorytm propagacji wstecznej (Kelley, 1960; Bryson, 1961; Dreyfusa, 1962; Rumelhart i in., 1986).

5.6.1. Gradienty w uczeniu głębokim

Obszarem, w którym silnie eksploatuje się regułę łańcuchową, jest uczenie głębokie, w którym funkcja $\boldsymbol{y}$ jest wielopoziomowym złożeniem funkcji

$$\boldsymbol{y} = (f_K \circ f_{K-1} \circ \cdots \circ f_1)(\boldsymbol{x}) = f_K\left(f_{K-1}(\cdots(f_1(\boldsymbol{x}))\cdots)\right), \tag{5.111}$$

Aby uprościć notację, omawiamy przypadek, w którym w każdej warstwie mamy identyczne funkcje aktywacji.

gdzie $\boldsymbol{x}$ to dane wejściowe (np. obrazy), a $\boldsymbol{y}$ obserwacje (np. etykiety klas). Każda funkcja $\boldsymbol{f}_i, i = 1, \ldots, K$ posiada własne parametry. W wielowarstwowej sieci neuronowej z i-tą warstwą sieci związana jest funkcja $\boldsymbol{f}_i(\boldsymbol{x}_{i-1}) = \sigma(\boldsymbol{A}_{i-1} + \boldsymbol{b}_{i-1})$, w której $\boldsymbol{x}_{i-1}$ jest wyjściem z $i-1$ warstwy, a σ jest funkcją aktywacji, na przykład funkcją sigmoidalną $\frac{1}{1+e^{-x}}$, tangensem hiperbolicznym lub funkcją ReLU. Proces uczenia takich sieci wymaga obliczenia gradientu funkcji straty L względem wszystkich parametrów $\boldsymbol{A}_j, \boldsymbol{b}_j$ $(j = 1, \ldots, K)$ modelu. Konieczne jest również znalezienie gradientu L względem danych wejściowych każdej warstwy. Rozważmy przykład, w którym $\boldsymbol{x}$ to dane wejściowe, a $\boldsymbol{y}$ to obserwacje. Struktura sieci jest określona przez

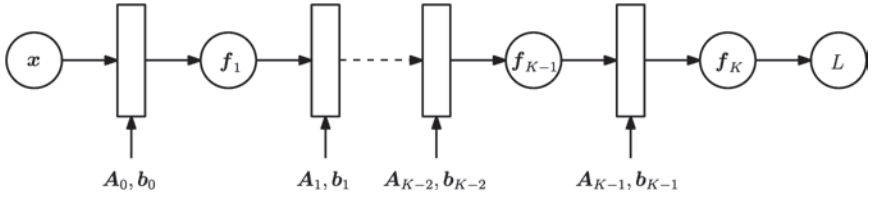
$$\boldsymbol{f}_0 := \boldsymbol{x} \tag{5.112}$$

$$\boldsymbol{f}_i := \sigma_i(\boldsymbol{A}_{i-1}\boldsymbol{f}_{i-1} + \boldsymbol{b}_{i-1}), i = 1, \ldots, K. \tag{5.113}$$

Sieć tę pokazano też na rysunku 5.8. Chcemy znaleźć parametry $\boldsymbol{A}_j, \boldsymbol{b}_j$ ($j = 0, \ldots, K-1$) minimalizujące kwadratową funkcję straty

$$L(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{f}_K(\boldsymbol{\theta}, \boldsymbol{x})\|^2, \tag{5.114}$$

gdzie $\boldsymbol{\theta} = \{\boldsymbol{A}_0, \boldsymbol{b}_0, \ldots, \boldsymbol{A}_{K-1}, \boldsymbol{b}_{K-1}\}$.



Rysunek 5.8. Przejście do przodu przez wielowarstwową sieć neuronową w celu obliczenia straty L w funkcji wejść $\boldsymbol{x}$ i parametrów $\boldsymbol{A}_i, \boldsymbol{b}_i$

Bardziej szczegółowe omówienie gradientów w sieciach neuronowych znajdziesz w notatkach Justina Domkego pod adresem *https://tinyurl.com/yalcxgtv*.

Do wyznaczenia gradientów względem zbioru parametrów $\boldsymbol{\theta}$ potrzebne nam będą pochodne cząstkowe L względem parametrów $\boldsymbol{\theta}_j = \{\boldsymbol{A}_j, \boldsymbol{b}_j\}$ każdej warstwy $j = 0, \ldots, K-1$. Reguła łańcuchowa pozwala nam zapisać pochodne cząstkowe jako

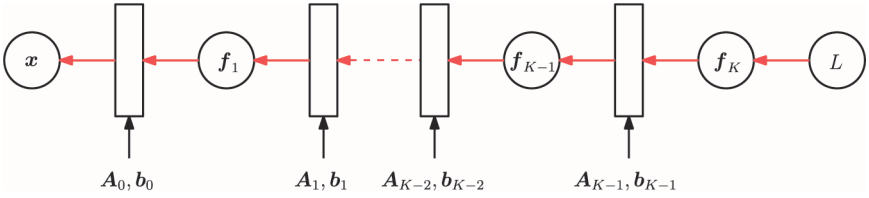
$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-1}} = \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{\theta}_{K-1}} \tag{5.115}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-2}} = \frac{\partial L}{\partial \boldsymbol{f}_K}\boxed{\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}}\frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{\theta}_{K-2}}} \tag{5.116}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-3}} = \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}}\boxed{\frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{f}_{K-2}}\frac{\partial \boldsymbol{f}_{K-2}}{\partial \boldsymbol{\theta}_{K-3}}} \tag{5.117}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_i} = \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}}\cdots\boxed{\frac{\partial \boldsymbol{f}_{i+2}}{\partial \boldsymbol{f}_{i+1}}\frac{\partial \boldsymbol{f}_{i+1}}{\partial \boldsymbol{\theta}_i}} \tag{5.118}$$

Składniki oznaczone kolorem pomarańczowym to pochodne cząstkowe wyjścia warstwy względem jej wejścia. Natomiast niebieskie składniki to pochodne cząstkowe wyjścia warstwy względem jej parametrów. Jeżeli obliczyliśmy już pochodne cząstkowe $\partial L/\partial \boldsymbol{\theta}_{i+1}$, to większość obliczonych wartości możemy wykorzystać ponownie do wyznaczenia $\partial L/\partial \boldsymbol{\theta}_i$. Dodatkowe składniki, które musimy obliczyć, zaznaczono ramkami. Na rysunku 5.9 pokazano, w jaki sposób gradienty są przekazywane wstecz sieci.



Rysunek 5.9. Przejście wstecz przez wielowarstwową sieć neuronową w celu znalezienia gradientu funkcji straty

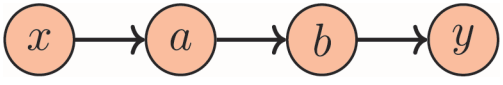
5.6.2. Różniczkowanie automatyczne

Różniczkowanie automatyczne różni się od różniczkowania symbolicznego i numerycznych metod aproksymacji gradientu np. przy użyciu różnic skończonych.

Okazuje się, że propagacja wsteczna jest szczególnym przypadkiem ogólnej techniki analizy numerycznej znanej jako **różniczkowanie automatyczne**. Różniczkowanie automatyczne możesz traktować jako zbiór metod służących do numerycznego (w odróżnieniu od symbolicznego) wyznaczania gradientów z dokładnością maszynową. Techniki różniczkowania automatycznego wykorzystują zmienne pośrednie i regułę łańcuchową. Różniczkowanie automatyczne wykorzystuje szereg elementarnych operacji arytmetycznych, takich jak na przykład dodawanie i mnożenie, oraz funkcje elementarne, takie jak sinus, cosinus, funkcja wykładnicza i logarytm. Zastosowanie reguły łańcuchowej pozwala wyznaczać gradient w sposób automatyczny. W ogólności różniczkowanie automatyczne dotyczy programów komputerowych i występuje przede wszystkim w dwóch wariantach: w przód i wstecz. Świetny przegląd różniczkowania automatycznego w kontekście uczenia maszynowego znajdziesz w pracy Baydina i in. (2018).

Na rysunku 5.10 pokazano prosty graf przedstawiający przepływ danych z wejść x do wyjść y przez zmienne pośrednie a i b. Gdybyśmy chcieli wyznaczyć pochodną $\mathrm{d}y/\mathrm{d}x$, to korzystając z reguły łańcuchowej, moglibyśmy zapisać, że:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}b}\frac{\mathrm{d}b}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}x}. \tag{5.119}$$



Rysunek 5.10. Prosty graf ilustrujący przepływ danych od x do y przez zmienne pośrednie a i b

Z intuicyjnego punktu widzenia warianty w przód i wstecz różnią się kolejnością wykonywania mnożenia. Ponieważ mnożenie macierzy jest łączne, możemy wybierać pomiędzy

W ogólnym przypadku pracujemy ze strukturami, którymi mogą być gradienty oraz macierze i tensory Jacobiego.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \left(\frac{\mathrm{d}y}{\mathrm{d}b}\frac{\mathrm{d}b}{\mathrm{d}a}\right)\frac{\mathrm{d}a}{\mathrm{d}x}, \tag{5.120}$$

i

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}b}\left(\frac{\mathrm{d}b}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}x}\right). \tag{5.121}$$

Równanie 5.120 to przykład metody różniczkowania automatycznego wstecz, ponieważ gradienty są w nim przekazywane od końca do początku grafu, tj. odwrotnie do przepływu danych. Równanie 5.121 jest przykładem metody w przód, w której informacje o gradiencie są przekazywane zgodnie z kierunkiem danych, tj. od lewej do prawej.

Poniżej skupimy się na różniczkowaniu wstecz, którego przykładem jest algorytm propagacji wstecznej. W przypadku sieci neuronowych, w których rozmiar wejścia jest często znacznie większy niż rozmiar wyjścia, różniczkowanie wstecz jest bardziej opłacalne obliczeniowo niż różniczkowanie metodą w przód. Zacznijmy od przykładu.

Przykład 5.14

Rozważmy funkcję

$$f(x) = \sqrt{x^2 + \exp\left(x^2\right)} + \cos\left(x^2 + \exp\left(x^2\right)\right) \tag{5.122}$$

z równania 5.109. Podczas implementacji tej funkcji na komputerze moglibyśmy zaoszczędzić trochę czasu, stosując tzw. **zmienne pośrednie**:

$$a = x^2 \tag{5.123}$$

$$b = \exp(a) \tag{5.124}$$

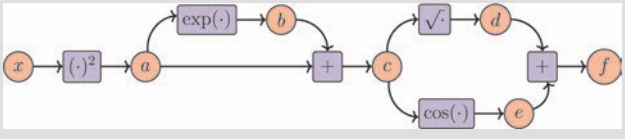
$$c = a + b \tag{5.125}$$

$$d = \sqrt{c} \tag{5.126}$$

$$e = \cos(c) \tag{5.127}$$

$$f = d + e. \tag{5.128}$$

W ten sam sposób rozumujemy również podczas stosowania reguły łańcuchowej. Zauważ, że powyższy zestaw równań wymaga wykonania mniejszej liczby operacji niż bezpośrednia implementacja funkcji $f(x)$ w postaci pokazanej w równaniu 5.109. Graf obliczeniowy tej funkcji, pokazujący przepływ danych i obliczenia wymagane do wyznaczenia wartości funkcji f, pokazano na rysunku 5.11.



Rysunek 5.11. Graf obliczeniowy z danymi wejściowymi x, wartościami funkcji f i zmiennymi pośrednimi a, b, c, d, e

Zbiór równań, które zawierają zmienne pośrednie, możemy traktować jako graf obliczeniowy. Ten sposób przedstawiania obliczeń jest powszechnie stosowany w implementacjach bibliotek przeznaczonych do tworzenia sieci neuronowych. Pochodne zmiennych pośrednich względem odpowiadających im danych wejściowych możemy obliczyć za pomocą definicji pochodnych funkcji elementarnych. Otrzymujemy następujące pochodne:

$$\frac{\partial a}{\partial x} = 2x \tag{5.129}$$

$$\frac{\partial b}{\partial a} = \exp(a) \tag{5.130}$$

$$\frac{\partial c}{\partial a} = 1 = \frac{\partial c}{\partial b} \tag{5.131}$$

$$\frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}} \tag{5.132}$$

$$\frac{\partial e}{\partial c} = -\sin(c) \tag{5.133}$$

$$\frac{\partial f}{\partial d} = 1 = \frac{\partial f}{\partial e}. \tag{5.134}$$

Analiza grafu z rysunku 5.11 pozwala zauważyć, że pochodną $\partial f / \partial x$ możemy obliczyć metodą wstecz, rozpoczynając obliczenia w wierzchołku wyjściowym:

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c} \tag{5.135}$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} \tag{5.136}$$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a} \tag{5.137}$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x}. \tag{5.138}$$

Zauważ, że do wyznaczenia $\partial f / \partial x$ zastosowaliśmy regułę łańcuchową. Po podstawieniu pochodnych funkcji elementarnych do powyższych równań otrzymujemy

$$\frac{\partial f}{\partial c} = 1 \cdot \frac{1}{2\sqrt{c}} + 1 \cdot (-\sin(c)) \tag{5.139}$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c} \cdot 1 \tag{5.140}$$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\exp(a) + \frac{\partial f}{\partial c} \cdot 1 \tag{5.141}$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a} \cdot 2x. \tag{5.142}$$

Jeżeli każdą z powyższych pochodnych potraktujemy jako zmienną, to proces wyznaczania pochodnej będzie miał podobną złożoność obliczeniową co proces obliczania wartości funkcji. Jest to sprzeczne z naszą intuicją, ponieważ wyrażenie matematyczne opisujące pochodną $\frac{\partial f}{\partial x}$ (równanie 5.110) jest znacznie bardziej skomplikowane niż wyrażenie opisujące funkcję $f(x)$ (równanie 5.109).

Różniczkowanie automatyczne jest formalizacją obliczeń z przykładu 5.14. Niech $x_1, \ldots, x_d$ będą zmiennymi wejściowymi, $x_{d+1}, \ldots, x_{D-1}$ zmiennymi pośrednimi, a x_D zmienną wyjściową. Graf obliczeniowy możemy wtedy zapisać jako

$$x_i = g_i\left(x_{\mathrm{Pa}(x_i)}\right) \text{ gdzie } i = d+1, \ldots, D. \tag{5.143}$$

W powyższym równaniu $g_i(\cdot)$ to funkcje elementarne, a $x_{\mathrm{Pa}(x_i)}$ to węzły rodzice zmiennej x_i w grafie. Pochodną tak zdefiniowanej funkcji możemy obliczyć za pomocą reguły łańcuchowej. Przypomnijmy, że z definicji $f = x_D$. Stąd

$$\frac{\partial f}{\partial x_D} = 1. \tag{5.144}$$

Dla pozostałych zmiennych x_i stosujemy regułę łańcuchową

$$\frac{\partial f}{\partial x_i} = \sum_{x_j : x_i \in \mathrm{Pa}\,(x_j)} \frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial x_i} = \sum_{x_j : x_i \in \mathrm{Pa}\,(x_j)} \frac{\partial f}{\partial x_j}\frac{\partial g_j}{\partial x_i}, \tag{5.145}$$

gdzie $\mathrm{Pa}\left(x_j\right)$ jest zbiorem węzłów rodziców wierzchołka x_j w grafie. Równanie 5.143 to przykład propagacji w przód, a równanie 5.145 reprezentuje propagację wsteczną gradientu w grafie obliczeniowym. W uczeniu sieci neuronowych propagacji wstecznej podlega błąd predykcji w odniesieniu do etykiety.

Różniczkowanie automatyczne metodą wstecz wymaga stworzenia drzewa składniowego.

Różniczkowanie automatyczne działa zawsze, gdy różniczkujemy funkcję dającą się przedstawić w postaci grafu obliczeniowego zawierającego różniczkowalne funkcje elementarne. Różniczkowana funkcja nie musi być nawet funkcją w sensie matematycznym, a jedynie programem komputerowym. Niestety, nie wszystkie programy komputerowe mogą być różniczkowane automatycznie. Nie jest to możliwe np. wtedy, gdy nie możemy wyznaczyć pochodnych funkcji elementarnych. Struktury występujące w kodzie, takie jak pętle `for` i instrukcje `if`, również wymagają większej uwagi.

5.7. Pochodne wyższych rzędów

Do tej pory omówiliśmy jedynie gradienty, czyli pochodne pierwszego rzędu. Czasami interesują nas też pochodne wyższych rzędów. Przydają się one np. w metodzie optymalizacji Newtona, która wymaga pochodnych drugiego rzędu (Nocedal i Wright, 2006). W punkcie 5.1.1 omówiliśmy szereg Taylora, pozwalający przybliżać funkcje za pomocą wielomianów. Przybliżenie takie możemy również wyznaczyć dla funkcji wielu zmiennych. W dalszej części tego rozdziału pokażemy, jak to zrobić. Zaczniemy jednak od wprowadzenia notacji.

Rozważmy funkcję $f: \mathbb{R}^2 \to \mathbb{R}$ dwóch zmiennych x, y. W przypadku pochodnych cząstkowych wyższego rzędu (oraz gradientów) stosujemy następujący zapis:

- $\frac{\partial^2 f}{\partial x^2}$ to druga pochodna cząstkowa f względem x.
- $\frac{\partial^n f}{\partial x^n}$ to pochodna cząstkowa n-tego rzędu f względem x.
- $\frac{\partial^2 f}{\partial y\, \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$ to pochodna cząstkowa otrzymana w wyniku różniczkowania względem x, a następnie względem y.
- $\frac{\partial^2 f}{\partial x\, \partial y}$ to pochodna cząstkowa otrzymana w wyniku różniczkowania względem y, a następnie względem x.

Hesjan (macierz Hessego) to zbiór wszystkich pochodnych cząstkowych drugiego rzędu. Jeśli $f(x, y)$ jest funkcją dwukrotnie różniczkowalną (a jej druga pochodna jest ciągła), to

$$\frac{\partial^2 f}{\partial x\, \partial y} = \frac{\partial^2 f}{\partial y\, \partial x}, \tag{5.146}$$

tj. kolejność różniczkowania nie ma znaczenia. Hesjanem tej funkcji jest symetryczna macierz

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\, \partial y} \\ \frac{\partial^2 f}{\partial x\, \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}. \tag{5.147}$$

Hesjan oznaczamy również symbolem $\nabla^2_{x,y} f(x, y)$. Dla $\boldsymbol{x} \in \mathbb{R}^n$ i $f: \mathbb{R}^n \to \mathbb{R}$ hesjan jest macierzą o wymiarach $n \times n$. Hesjan mierzy lokalną krzywiznę funkcji wokół punktu (x, y).

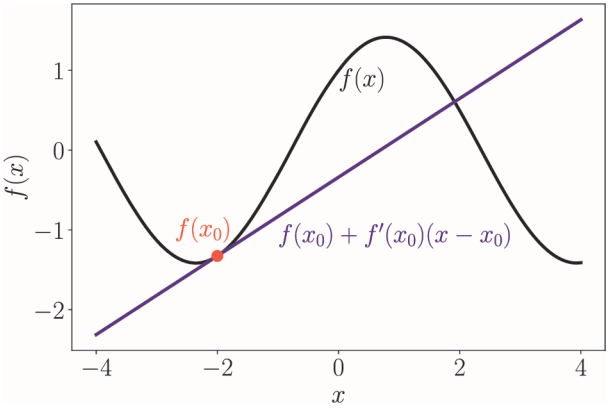
Uwaga (hesjan pola wektorowego). Jeśli $f: \mathbb{R}^n \to \mathbb{R}^m$ jest polem wektorowym, to hesjan jest tensorem o wymiarach $m \times n \times n$.

5.8. Linearyzacja i wielowymiarowe szeregi Taylora

Gradient ∇f funkcji f często wykorzystuje się do lokalnego liniowego przybliżania funkcji f w otoczeniu punktu $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0). \tag{5.148}$$

$(\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)$ jest gradientem f względem $\boldsymbol{x}$ obliczonym w $\boldsymbol{x}_0$. Liniową aproksymację funkcji f w otoczeniu $\boldsymbol{x}_0$ pokazano na rysunku 5.12. Oryginalna funkcja jest przybliżana za pomocą prostej. To przybliżenie jest lokalnie dokładne, ale im dalej od x_0, tym prosta gorzej przybliża funkcję f. Równanie 5.148 jest szczególnym przypadkiem rozwinięcia f w wielowymiarowym szeregu Taylora w otoczeniu punktu $\boldsymbol{x}_0$, w którym korzystamy jedynie z dwóch pierwszych wyrazów. Poniżej omówimy bardziej ogólny przypadek tego szeregu, który pozwoli uzyskać lepsze przybliżenia.



RYSUNEK 5.12. Liniowa aproksymacja funkcji. Oryginalna funkcja f jest linearyzowana w otoczeniu $x_0 = -2$ za pomocą rozwinięcia w szereg Taylora pierwszego rzędu

Definicja 5.7 (szereg Taylora funkcji wielu zmiennych)**.** Rozważamy funkcję

$$f: \mathbb{R}^D \to \mathbb{R} \tag{5.149}$$

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^D, \tag{5.150}$$

która jest gładka w punkcie $\boldsymbol{x}_0$. Jeżeli zdefiniujemy wektor różnic jako $\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0$, to **szereg Taylora funkcji wielu zmiennych** f w punkcie $\boldsymbol{x}_0$ będzie zdefiniowany jako

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \frac{D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)}{k!} \boldsymbol{\delta}^k, \tag{5.151}$$

gdzie $D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)$ jest k-tą (zupełną) pochodną f względem $\boldsymbol{x}$ obliczoną w $\boldsymbol{x}_0$.

Definicja 5.8 (wielomian Taylora)**. Wielomian Taylora** n-tego stopnia funkcji f w $\boldsymbol{x}_0$ zawiera pierwsze $n + 1$ elementów szeregu z równania 5.151 i jest zdefiniowany jako

$$T_n(\boldsymbol{x}) = \sum_{k=0}^{n} \frac{D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)}{k!} \boldsymbol{\delta}^k. \tag{5.152}$$

Wektor można zaimplementować jako tablicę jednowymiarową, a macierz w postaci tablicy dwuwymiarowej.

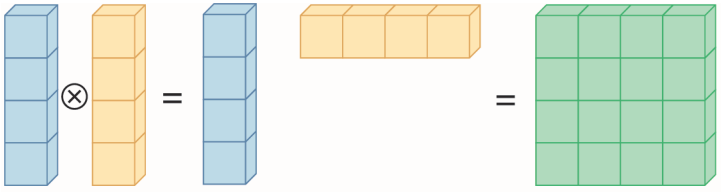
W równaniach 5.151 i 5.152 zastosowaliśmy nieco nieprecyzyjny zapis z użyciem $\boldsymbol{\delta}^k$, które jest niezdefiniowane dla wektorów $\boldsymbol{x} \in \mathbb{R}^D$ takich, że $D > 1$ *i* $k > 1$. Zauważ, że $D_{\boldsymbol{x}}^k f$ i $\boldsymbol{\delta}^k$ są tensorami k-tego rzędu, tj. k-wymiarowymi tablicami. Tensor k-tego rzędu $\boldsymbol{\delta}^k \in \mathbb{R}^{\overbrace{D \times D \times \ldots \times D}^{k \text{ razy}}}$ to wynik k-krotnego iloczynu zewnętrznego (oznaczanego symbolem $\otimes$) wektora $\boldsymbol{\delta} \in \mathbb{R}^D$. Na przykład

$$\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\top, \qquad \boldsymbol{\delta}^2[i,j] = \delta[i]\delta[j] \tag{5.153}$$

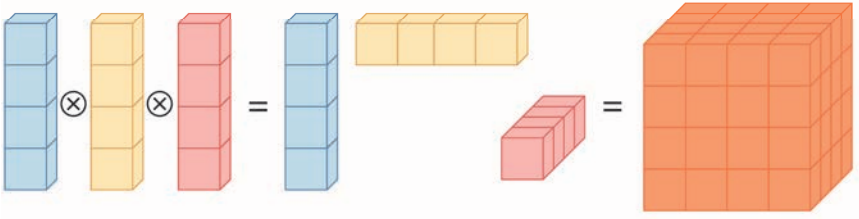
$$\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta}, \qquad \boldsymbol{\delta}^3[i,j,k] = \delta[i]\delta[j]\delta[k]. \tag{5.154}$$

Na rysunku 5.13 pokazano dwa takie iloczyny zewnętrzne. W ogólnym przypadku w szeregu Taylora pojawiają się następujące składniki

$$D_x^k f(\boldsymbol{x}_0)\boldsymbol{\delta}^k = \sum_{i_1=1}^{D} \cdots \sum_{i_k=1}^{D} D_x^k f(\boldsymbol{x}_0)[i_1, \ldots, i_k]\delta[i_1] \cdots \delta[i_k]. \tag{5.155}$$



(a) Iloczyn zewnętrzny wektora $\boldsymbol{\delta} \in \mathbb{R}^4$ daje $\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\mathrm{T} \in \mathbb{R}^{4\times4}$, czyli macierz



(b) Iloczyn zewnętrzny $\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta} \in \mathbb{R}^{4\times4\times4}$ daje tensor trzeciego rzędu (macierz trójwymiarową), tj. macierz z trzema indeksami (osiami)

Rysunek 5.13. Wizualizacja iloczynów zewnętrznych. Każdy iloczyn zewnętrzny wektorów zwiększa o jeden liczbę wymiarów wynikowej tablicy. (a) Iloczyn zewnętrzny dwóch wektorów daje macierz; (b) iloczyn zewnętrzny trzech wektorów daje tensor trzeciego rzędu

Element $D_x^k f(\boldsymbol{x}_0)\boldsymbol{\delta}^k$ zawiera wielomiany k-tego rzędu.

Teraz, gdy zdefiniowaliśmy już szereg Taylora dla pól wektorowych, wyznaczymy pierwsze wyrazy $D_x^k f(\boldsymbol{x}_0)\boldsymbol{\delta}^k$ rozwinięcia w szereg Taylora dla $k = 0, \ldots, 3$ i $\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0$:

$$k = 0\text{: } D_x^0 f(\boldsymbol{x}_0)\boldsymbol{\delta}^0 = f(\boldsymbol{x}_0) \in \mathbb{R} \tag{5.156}$$

$$k = 1\text{: } D_x^1 f(\boldsymbol{x}_0)\boldsymbol{\delta}^1 = \underbrace{\nabla_x f(\boldsymbol{x}_0)}_{1\times D}\underbrace{\boldsymbol{\delta}}_{D\times 1} = \sum_{i=1}^{D} \nabla_x f(\boldsymbol{x}_0)[i]\delta[i] \in \mathbb{R} \tag{5.157}$$

$$k = 2\text{: } D_x^2 f(\boldsymbol{x}_0)\boldsymbol{\delta}^2 = \mathrm{tr}\,(\underbrace{\boldsymbol{H}(\boldsymbol{x}_0)}_{D\times D}\underbrace{\boldsymbol{\delta}}_{D\times 1}\underbrace{\boldsymbol{\delta}^\mathrm{T}}_{1\times D}) = \boldsymbol{\delta}^\mathrm{T}\boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{\delta} \tag{5.158}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} H[i,j]\delta[i]\delta[j] \in \mathbb{R} \tag{5.159}$$

$$k = 3\text{: } D_x^3 f(\boldsymbol{x}_0)\boldsymbol{\delta}^3 = \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} D_x^3 f(\boldsymbol{x}_0)[i,j,k]\delta[i]\delta[j]\delta[k] \in \mathbb{R}. \tag{5.160}$$

W powyższych równaniach $\boldsymbol{H}(\boldsymbol{x}_0)$ jest hesjanem f w punkcie $\boldsymbol{x}_0$.

Przykład 5.15 (rozwinięcie funkcji dwóch zmiennych w szereg Taylora)

Rozważmy funkcję

$$f(x,y) = x^2 + 2xy + y^3. \tag{5.161}$$

Chcemy znaleźć jej rozwinięcie w szereg Taylora w punkcie $(x_0, y_0) = (1,2)$. Zanim zaczniemy, omówmy, czego możemy się spodziewać. Funkcja z równania 5.161 jest wielomianem trzeciego stopnia. Szukamy rozwinięcia w szereg Taylora, które samo w sobie jest kombinacją liniową wielomianów. Spodziewamy się więc, że w przypadku wielomianu trzeciego stopnia rozwinięcie nie będzie zawierało wyrazów czwartego lub wyższego stopnia. Oznacza to, że pierwsze cztery wyrazy z sumy z równania 5.151 powinny wystarczyć do uzyskania dokładnej reprezentacji funkcji z równania 5.161.

Proces wyznaczania rozwinięcia zaczynamy od wyrazu stałego i pochodnych pierwszego rzędu, które dane są wzorem

$$f(1,2) = 13 \tag{5.162}$$

$$\frac{\partial f}{\partial x} = 2x + 2y \Longrightarrow \frac{\partial f}{\partial x}(1,2) = 6 \tag{5.163}$$

$$\frac{\partial f}{\partial y} = 2x + 3y^2 \Longrightarrow \frac{\partial f}{\partial y}(1,2) = 14. \tag{5.164}$$

Stąd

$$D^1_{x,y} f(1,2) = \nabla_{x,y} f(1,2) = \left[\frac{\partial f}{\partial x}(1,2) \quad \frac{\partial f}{\partial y}(1,2)\right] = [6 \quad 14] \in \mathbb{R}^{1\times 2}, \tag{5.165}$$

co daje

$$\frac{D^1_{x,y} f(1,2)}{1!}\boldsymbol{\delta} = [6 \quad 14]\begin{bmatrix} x-1 \\ y-2 \end{bmatrix} = 6(x-1) + 14(y-2). \tag{5.166}$$

Zauważ, że $D^1_{x,y} f(1,2)\boldsymbol{\delta}$ zawiera tylko składniki liniowe, tj. wielomiany pierwszego rzędu.

Pochodne cząstkowe drugiego rzędu to

$$\frac{\partial^2 f}{\partial x^2} = 2 \Longrightarrow \frac{\partial^2 f}{\partial x^2}(1,2) = 2 \tag{5.167}$$

$$\frac{\partial^2 f}{\partial y^2} = 6y \Longrightarrow \frac{\partial^2 f}{\partial y^2}(1,2) = 12 \tag{5.168}$$

$$\frac{\partial^2 f}{\partial y\,\partial x} = 2 \Longrightarrow \frac{\partial^2 f}{\partial y\,\partial x}(1,2) = 2 \tag{5.169}$$

$$\frac{\partial^2 f}{\partial x\,\partial y} = 2 \Longrightarrow \frac{\partial^2 f}{\partial x\,\partial y}(1,2) = 2. \tag{5.170}$$

Po zapisaniu ich w macierzy otrzymujemy hesjan

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\,\partial y} \\ \frac{\partial^2 f}{\partial y\,\partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 6y \end{bmatrix}. \tag{5.171}$$

A zatem

$$\boldsymbol{H}(1,2) = \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{5.172}$$

A więc kolejny wyraz rozwinięcia w szereg Taylora to

$$\frac{D^2_{x,y} f(1,2)}{2!}\boldsymbol{\delta}^2 = \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{H}(1,2)\boldsymbol{\delta} \tag{5.173a}$$

$$= \frac{1}{2}\begin{bmatrix} x-1 & y-2 \end{bmatrix}\begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix}\begin{bmatrix} x-1 \\ y-2 \end{bmatrix} \tag{5.173b}$$

$$= (x-1)^2 + 2(x-1)(y-2) + 6(y-2)^2. \tag{5.173c}$$

W tym przypadku $D^2_{x,y} f(1,2)\boldsymbol{\delta}^2$ zawiera tylko wyrazy kwadratowe, tj. wielomiany drugiego rzędu.

Następnie wyznaczamy pochodne trzeciego rzędu

$$D^3_{x,y} f = \begin{bmatrix} \frac{\partial \boldsymbol{H}}{\partial x} & \frac{\partial \boldsymbol{H}}{\partial y} \end{bmatrix} \in \mathbb{R}^{2\times 2\times 2} \tag{5.174}$$

$$D^3_{x,y} f[:,:,1] = \frac{\partial \boldsymbol{H}}{\partial x} = \begin{bmatrix} \frac{\partial^3 f}{\partial x^3} & \frac{\partial^3 f}{\partial x^2\,\partial y} \\ \frac{\partial^3 f}{\partial x\,\partial y\,\partial x} & \frac{\partial^3 f}{\partial x\,\partial y^2} \end{bmatrix} \tag{5.175}$$

$$D^3_{x,y} f[:,:,2] = \frac{\partial \boldsymbol{H}}{\partial y} = \begin{bmatrix} \frac{\partial^3 f}{\partial y\,\partial x^2} & \frac{\partial^3 f}{\partial y\,\partial x\,\partial y} \\ \frac{\partial^3 f}{\partial y^2\,\partial x} & \frac{\partial^3 f}{\partial y^3} \end{bmatrix}. \tag{5.176}$$

Ponieważ większość pochodnych cząstkowych drugiego rzędu w hesjanie z równania 5.171 to stałe, jedyną niezerową pochodną cząstkową trzeciego rzędu jest

$$\frac{\partial^3 f}{\partial y^3} = 6 \Rightarrow \frac{\partial^3 f}{\partial y^3}(1,2) = 6. \tag{5.177}$$

Pochodne wyższego rzędu i pochodne mieszane trzeciego rzędu (np. $\frac{\partial f^3}{\partial x^2\,\partial y}$) zanikają:

$$D^3_{x,y} f[:,:,1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad D^3_{x,y} f[:,:,2] = \begin{bmatrix} 0 & 0 \\ 0 & 6 \end{bmatrix}. \tag{5.178}$$

A zatem

$$\frac{D^3_{x,y} f(1,2)}{3!}\boldsymbol{\delta}^3 = (y-2)^3. \tag{5.179}$$

Składnik ten zawiera w sobie wszystkie sześcienne wyrazy szeregu. (Dokładne) rozwinięcie f w szereg Taylora w punkcie $(x_0, y_0) = (1,2)$ jest zatem równe

$$f(x) = f(1,2) + D^1_{x,y} f(1,2)\boldsymbol{\delta} + \frac{D^2_{x,y} f(1,2)}{2!}\delta^2 + \frac{D^3_{x,y} f(1,2)}{3!}\delta^3 \tag{5.180a}$$

$$= f(1,2) + \frac{\partial f(1,2)}{\partial x}(x-1) + \frac{\partial f(1,2)}{\partial y}(y-2)$$
$$+ \frac{1}{2!}\left(\frac{\partial^2 f(1,2)}{\partial x^2}(x-1)^2 + \frac{\partial^2 f(1,2)}{\partial y^2}(y-2)^2\right. \tag{5.180b}$$
$$\left. + 2\frac{\partial^2 f(1,2)}{\partial x\,\partial y}(x-1)(y-2)\right) + \frac{1}{6}\frac{\partial^3 f(1,2)}{\partial y^3}(y-2)^3$$

$$= 13 + 6(x-1) + 14(y-2)$$
$$+ (x-1)^2 + 6(y-2)^2 + 2(x-1)(y-2) + (y-2)^3 \tag{5.180c}$$

W tym przypadku otrzymaliśmy dokładne rozwinięcie wielomianu z równania 5.161 w szereg Taylora, tzn. wielomian z równania 5.180c jest identyczny z pierwotnym wielomianem z równania 5.161. W tym konkretnym przykładzie wynik ten nie jest zaskakujący, ponieważ pierwotną funkcją był wielomian trzeciego rzędu, który wyraziliśmy poprzez kombinację liniową stałych oraz wielomianów pierwszego, drugiego i trzeciego stopnia (równanie 5.180c).

5.9. Materiały dodatkowe

Więcej informacji o różniczkowaniu macierzy oraz krótki przegląd związanej z nim algebry liniowej można znaleźć u Magnusa i Neudeckera (2007). Różniczkowanie automatyczne ma długą historię. Więcej szczegółów znajdziesz w pracach Griewanka i Walthera (2003), Griewanka i Walthera (2008) oraz Elliotta (2009), a także w pracach wymienionych w bibliografiach dołączonych do tych pozycji.

W uczeniu maszynowym (i w innych dziedzinach) często interesują nas wartości oczekiwane, czyli całki postaci

$$\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}. \tag{5.181}$$

Nawet jeśli $p(\boldsymbol{x})$ ma dogodną formę (np. jest funkcją gęstości rozkładu normalnego), to w ogólnym przypadku całka ta nie ma rozwiązania analitycznego. Rozwinięcie f w szereg Taylora jest jednym ze sposobów na znalezienie przybliżonego rozwiązania. Jeżeli przyjmiemy, że $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, czyli, że $p(\boldsymbol{x})$ jest funkcją gęstości prawdopodobieństwa rozkładu normalnego, to rozwinięcie w szereg Taylora pierwszego rzędu w otoczeniu $\boldsymbol{\mu}$ będzie lokalną linearyzacją nieliniowej funkcji f. Jeśli $p(\boldsymbol{x})$ jest funkcją gęstości rozkładu normalnego (patrz podrozdział 6.5), to w przypadku liniowych funkcji f możemy dokładnie wyznaczyć średnią (i kowariancję). Ta własność jest intensywnie

wykorzystywana przez **rozszerzony filtr Kalmana** (ang. *extended Kalman filter*; Maybeck, 1979) do bieżącej estymacji stanu w nieliniowych układach dynamicznych (zwanych również *modelami przestrzeni stanów*, ang. *state-space models*). Inne deterministyczne sposoby przybliżania całki z równania 5.181 to **transformacja bezśladowa** (ang. *unscented transform*, Julier i Uhlmann, 1997), która nie wymaga znajomości gradientów, oraz **aproksymacja Laplace'a** (MacKay, 2003; Bishop, 2006; Murphy, 2012), która wykorzystuje rozwinięcie lokalnego przybliżenia funkcji gęstości rozkładu normalnego w szereg Taylora drugiego rzędu (wymagana jest znajomość hesjanu) wokół jego mody.

Ćwiczenia

5.1. Oblicz pochodną $f'(x)$ funkcji

$$f(x) = \log(x^4)\sin(x^3).$$

5.2. Wyznacz pochodną $f'(x)$ sigmoidalnej funkcji logistycznej

$$f(x) = \frac{1}{1+\exp(-x)}.$$

5.3. Oblicz pochodną $f'(x)$ funkcji

$$f(x) = \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

gdzie $\mu, \sigma \in \mathbb{R}$ są stałymi.

5.4. Dla $n = 0, \ldots, 5$ znajdź wielomiany Taylora T_n funkcji $f(x) = \sin(x) + \cos(x)$ w punkcie $x_0 = 0$.

5.5. Rozważ następujące funkcje:

$$f_1(\boldsymbol{x}) = \sin(x_1)\cos(x_2), \quad \boldsymbol{x} \in \mathbb{R}^2$$
$$f_2(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{y}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$
$$f_3(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top, \quad \boldsymbol{x} \in \mathbb{R}^n$$

a) Jakie są wymiary $\frac{\partial f_i}{\partial \boldsymbol{x}}$?

b) Wyznacz gradienty/macierze Jacobiego powyższych funkcji.

5.6. Znajdź pochodną f względem $\boldsymbol{t}$ i pochodną g względem $\boldsymbol{X}$

$$f(\boldsymbol{t}) = \sin(\log(\boldsymbol{t}^\top \boldsymbol{t})), \quad t \in \mathbb{R}^D$$
$$g(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}), \quad \boldsymbol{A} \in \mathbb{R}^{D\times E}, \quad \boldsymbol{X} \in \mathbb{R}^{E\times F}, \quad \boldsymbol{B} \in \mathbb{R}^{F\times D},$$

gdzie $\mathrm{tr}(\cdot)$ oznacza ślad.

5.7. Za pomocą reguły łańcuchowej oblicz pochodne $\mathrm{d}f/\mathrm{d}\boldsymbol{x}$ następujących funkcji. Podaj wymiary każdej pochodnej cząstkowej. Szczegółowo opisz swoje działania.

a)

$$f(z) = \log(1+z), \quad z = \boldsymbol{x}^\top \boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^D$$

b)

$$f(\boldsymbol{z}) = \sin(\boldsymbol{z}), \quad \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{A} \in \mathbb{R}^{E\times D}, \quad \boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{b} \in \mathbb{R}^E,$$

gdzie $\sin(\cdot)$ jest obliczany dla każdego elementu wektora $\boldsymbol{z}$.

5.8. Oblicz pochodne $\mathrm{d}f/\mathrm{d}\boldsymbol{x}$ następujących funkcji. Opisz szczegółowo wykonywane kroki.

a) Wykorzystaj regułę łańcuchową. Podaj wymiary każdej pochodnej cząstkowej.

$$f(z) = \exp\left(-\frac{1}{2}z\right)$$
$$z = g(\boldsymbol{y}) = \boldsymbol{y}^\top \boldsymbol{S}^{-1}\boldsymbol{y}$$
$$\boldsymbol{y} = h(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\mu},$$

gdzie $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{S} \in \mathbb{R}^{D\times D}$.

b)

$$f(\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{x}\boldsymbol{x}^\top + \sigma^2\boldsymbol{I}), \boldsymbol{x} \in \mathbb{R}^D,$$

gdzie $\mathrm{tr}\,(\boldsymbol{A})$ jest śladem $\boldsymbol{A}$, czyli sumą elementów A_{ii} leżących na przekątnej. *Wskazówka: zapisz iloczyn zewnętrzny w jawnej postaci.*

c) Wykorzystaj regułę łańcuchową. Podaj wymiary każdej pochodnej cząstkowej. Nie musisz jawnie obliczać iloczynu pochodnych cząstkowych.

$$\boldsymbol{f} = \mathrm{tgh}(\boldsymbol{z}) \in \mathbb{R}^M$$
$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{b} \in \mathbb{R}^M,$$

gdzie $\mathrm{tgh}(\cdot)$ jest obliczany dla każdego elementu wektora $\boldsymbol{z}$.

5.9. Niech

$$g(\boldsymbol{z}, \boldsymbol{v}) \coloneqq \log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z}, \boldsymbol{v})$$
$$\boldsymbol{z} \coloneqq t(\boldsymbol{\epsilon}, \boldsymbol{v}),$$

gdzie p, q i t są funkcjami różniczkowalnymi oraz $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{z} \in \mathbb{R}^E$, $\boldsymbol{v} \in \mathbb{R}^F$, $\boldsymbol{\epsilon} \in \mathbb{R}^G$. Za pomocą reguły łańcuchowej wyznacz gradient

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{v}} g(\boldsymbol{z}, \boldsymbol{v}).$$

Uczenie maszynowe staje się wszechobecne. Dzięki coraz lepszym narzędziom służącym do tworzenia aplikacji szczegóły techniczne związane z obliczeniami i modelami matematycznymi są często pomijane przez projektantów. Owszem, to wygodne podejście, ale wiąże się z ryzykiem braku świadomości co do wszystkich konsekwencji wybranych rozwiązań projektowych, szczególnie ich mocnych i słabych stron. A zatem bez ugruntowanych podstaw matematyki nie można mówić o profesjonalnym podejściu do uczenia maszynowego.

Ten podręcznik jest przeznaczony dla osób, które chcą dobrze zrozumieć matematyczne podstawy uczenia maszynowego i nabrać praktycznego doświadczenia w używaniu pojęć matematycznych. Wyjaśniono tutaj stosowanie szeregu technik matematycznych, takich jak algebra liniowa, geometria analityczna, rozkłady macierzy, rachunek wektorowy, optymalizacja, probabilistyka i statystyka. Następnie zaprezentowano matematyczne aspekty czterech podstawowych metod uczenia maszynowego: regresji liniowej, analizy głównych składowych, modeli mieszanin rozkładów Gaussa i maszyn wektorów nośnych. W każdym rozdziale znalazły się przykłady i ćwiczenia ułatwiające przyswojenie materiału.

W książce między innymi:

- podstawy algebry: układy równań, macierze, przestrzenie afiniczne
- rachunek prawdopodobieństwa, sprzężenia, optymalizacja
- wnioskowanie z wykorzystaniem różnego rodzaju modeli
- regresja liniowa i redukcja wymiarowości
- maszyna wektorów nośnych i rozwiązania numeryczne

Marc Peter Deisenroth

kieruje zakładem sztucznej inteligencji w University College London. W swojej pracy badawczej zajmuje się efektywnym uczeniem, modelowaniem probabilistycznym i autonomicznym podejmowaniem decyzji.

A. Aldo Faisal

kieruje laboratorium Brain & Behavior w Imperial College London, gdzie jest również wykładowcą i członkiem Data Science Institute. W swoich badaniach zajmuje się zagadnieniami na styku neuronauki i uczenia maszynowego.

Cheng Soon Ong

jest głównym badaczem w Machine Learning Research Group i adiunktem na Australian National University. Koncentruje się na rozwijaniu statystycznych metod uczenia maszynowego.

Matematyka: koniecznie, jeśli chcesz zrozumieć istotę sztucznej inteligencji!

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-283-8459-0

9 788328 384590

Cena: 99,00 zł