5. Семинар 2. Матрично-векторное дифференцирование. Теория

Никита Корнилов

5.1. Введение

Само название семинара говорит о том, что, скорее всего, мы будем учиться дифференцировать функции по переменным, которые являются векторами или же матрицами. Чтобы лучше погрузиться в эту тему, сначала обсудим интуицию и простейшие факты.

Начнем с самого простого примера функции – линейной функции:

$$f(x) = ax,$$

где $a \in \mathbb{R}$ - некоторое число, а $x \in \mathbb{R}$ - переменная. Все знают, чему равна производная такой функции: $f'(x) = a$.

Попробуем расширить этот пример. Представим функцию $f(x)$ как тоже линейную функцию, но от нескольких переменных $(x_1, x_2, \ldots)$. Тогда

$$f(x) = \sum_{i=1}^{n} a_i x_i.$$

Но эту функцию можно переписать и в другом виде, если обозначить $(a_1, \ldots, a_n)^\top$ за вектор a:

$$f(x) = \sum_{i=1}^{n} a_i x_i = \langle a, x \rangle = a^\top x.$$

На всякий случай отметим, что в общепринятой формализации вектор – это **столбец**.

Для такой функции мы можем посчитать набор частных производных:

$$\frac{\partial f}{\partial x_i} = a_i, \qquad i = \overline{1, n}.$$

Тогда можно составить вектор из частных производных, который будет называться **градиентом**:

$$\frac{df}{dx} = \nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)^\top.$$

Теперь попробуем еще больше расширить рассматриваемые функции, перейдя к переменным-матрицам. Если переменная x в функции $f(x)$ является матрицей, т.е. $x \in \mathbb{R}^{m \times n}$, то можно также составить матрицу из всех частных производных, которая тоже называется **градиентом**:

$$\frac{df}{dx} = \nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \dots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \dots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}$$

Из вышесказанного будет следовать замечательное свойство:

$$\frac{df}{dx^\top} = \left(\frac{df}{dx}\right)^\top = (\nabla f(x))^\top.$$

Теперь немного отойдем от определений, связанных с дифференцированием, и перейдем к вещи, связанной с линейной алгеброй, а именно, к **следу**. След матрицы A задается через сумму диагональных элементов матрицы, т.е. $\mathrm{Tr}(A) = \sum\limits_{i=1}^{n} A_{ii}$. Также обсудим его некоторые свойства:

- $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$
- $\mathrm{Tr}(cA) = c\mathrm{Tr}(A)$
- $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$
- $\mathrm{Tr}(A_1 \dots A_n) = \mathrm{Tr}(A_n A_1 \dots A_{n-1})$
- $\mathrm{Tr}(A^\top B) = \sum\limits_{i,j} A_{ij} B_{ij} = \langle A, B \rangle$, где $\langle A, B \rangle$ - скалярное произведение матриц.
- $\mathrm{Tr}(A) = \mathrm{Tr}(A^\top)$

Но зачем нужен след для дифференцирования? На самом деле, интуиция применения следа к дифференцированию довольно проста, и заключается в следующем. Рассмотрим некоторую функцию $f(x)$, которая может зависеть от вектора или матрицы, но которая выдает число, т.е. $f(x) \in \mathbb{R}$. Тогда $f(x) = \mathrm{Tr}(f(x))$, так как число это **матрица** размера 1×1. Накладывая след на значение функции, можно будет применять его различные свойства. Но для начала разберем пример.

Пусть $f(x) = \mathrm{Tr}(A^\top x)$, где x - матрица. Тогда

$$\nabla f(x) = A.$$

Чтобы это доказать, достаточно воспользоваться свойством следа: $\mathrm{Tr}(A^\top x) = \sum\limits_{i,j} A_{ij}x_{ij}$ и взять все частные производные.

Теперь, чтобы понять, как можно применять след, найдем $\frac{df}{dx^\top}$, если $f(x) = a^\top x$, где a - вектор.

$$\frac{df}{dx^\top} = \frac{d(a^\top x)}{dx^\top} = \frac{d(\mathrm{Tr}(a^\top x))}{dx^\top} = \frac{d(\mathrm{Tr}(xa^\top))}{dx^\top} = \frac{d(\mathrm{Tr}(ax^\top))}{dx^\top}$$
$$= \frac{d(\mathrm{Tr}((a^\top)^\top x^\top))}{dx^\top}.$$

Тогда, пользуясь предыдущим примером, получаем, что

$$\frac{df}{dx^\top} = a^\top.$$

Стоит отметить следующее:

а) Все функции, рассматриваемые в вышеуказанных примерах, являются линейными, но по аналогии можно действовать и с другими функциями (квадратичная, экспонента, логарифм и т.д.) с применениями обыкновенных правил дифференцирования.

б) Как и в математическом анализе, можно ввести понятие дифференциала: $df = \langle \nabla f(x), dx \rangle$ для случая, когда f возвращает скаляр.

в) Размерность градиента должна быть **такой же**, как и размерность переменной, по которой происходит дифференцирование.

К сожалению, такое введение в матрично-векторное дифференцирование не является формальным, хоть оно и правдиво. Поэтому перейдем к формальному знакомству с понятиями дифференциала и градиента.

5.2. Определения и полезные факты

5.2.1. Первая производная

Пусть U, V - конечномерные ЛНП. Основными примерами таких пространств являются действительные числа $\mathbb{R}$, векторы $\mathbb{R}^n$ и матрицы $\mathbb{R}^{n\times m}$, а также их декартовы произведения.

Рассмотрим функцию

$$f : X \to V, X \subset U.$$

Определение 5.1. Функция f дифференцируема в $x \in intX$, если

$$\exists L : U \to V : f(x+h) = f(x) + Lh + o(\|h\|), \|h\| \to 0.$$

Линейный оператор L называется производной f в точке x.

Если для любого линейного оператора $L : U \to V$ функция f в точке x не является дифференцируемой с производной L, то f недифференцируема в x. Если точка x не является внутренней, то понятие дифференцируемости не определено.

Замечание 5.2. Выбор нормы в определении 5.1 не имеет значения в силу топологической эквивалентности всех норм в конечномерных пространствах.

Предложение 5.3. Если функция f дифференцируема в точке x с производной L_1 и L_2, то $L_1 = L_2$.

Таким образом, производная, если существует, определена единственно. Будем обозначать её $f'(x)$.

Далее введём не менее важное определение дифференциала.

Определение 5.4. Дифференциалом $df(x)[h] \in V$ в точке $x \in X$ дифференцируемой функции f и с приращением h называется вектор $f'(x)[h]$.

Также для обозначения дифференциала $df(x)[h]$ используются

$$df(x)[h] \equiv Df(x)[h] \equiv f'(x)dx.$$

Часто на практике убирают приращения h, оставляя $df(x)$, и точку x, оставляя df, если понятно, о чём идёт речь.

Помимо дифференциала и производной у функции могут быть определены производные по направлению, отвечающие за изменения функции вдоль одного направления.

Определение 5.5. Производной по направлению h функции f в точке x называется

$$\frac{\partial f}{\partial h} := \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}.$$

В случае дифференцируемости f в точке x можно легко найти любую производную по направлению через обыкновенную производную.

Предложение 5.6. Пусть f дифференцируема в x. Тогда для произвольного направления h

$$\frac{\partial f(x)}{\partial h} = Df(x)[h] = f'(x)[h]. \tag{8}$$

В случае $U = \mathbb{R}^n$ если для направления вектора стандартного базиса

$$e_i = (0, \dots, 0, \underset{i}{1}, 0, \dots, 0) \in \mathbb{R}^n$$

существует двусторонний предел (5.5), то его называют частной производной по i-ой координате,

$$\frac{\partial f(x)}{\partial x_i} := \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}.$$

В случае $U = \mathbb{R}^{n \times m}$ индекс i заменяется на ij с сохранением определений. Заметим, что V не обязано равняться $\mathbb{R}$.

Как и в классическом матанализе, из дифференцируемости в точке x следует существование производных по всем направлением в точке x. Однако обратное неверно.

Пример 5.7. Пусть $f(x) = \|x\|_2$. Тогда её дифференциал равен

$$Df(0)[h] = \frac{\partial f}{\partial h}(0) = \lim_{t \to +0} \frac{t\|h\|_2}{t} = \|h\|_2,$$

но вторая норма нелинейная, что противоречит определению оператора производной. Таким образом, вторая норма не дифференцируема в нуле.

5.2.2. Градиент, матрица Якоби

В этом параграфе мы введём основные понятия для производных в стандартных пространствах U, V.

- В случае $f : \mathbb{R}^n \to \mathbb{R}$ линейную функцию $Df(x)[h]$ можно представить в виде

$$Df(x)[h] = \langle a_x, h \rangle, \quad \text{где } a_x \in \mathbb{R}^n \text{ зависит от } x.$$

Вектор a_x называется `градиентом` $f(x)$ в точке x и обозначается $\nabla f(x)$.

Подставив в качестве направлений $h = e_i$, получим явное значение градиента в стандартном базисе со стандартным скалярным произведением

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)^\top \in \mathbb{R}^n. \tag{9}$$

- В случае $f : \mathbb{R}^{n \times m} \to \mathbb{R}$ линейную функцию $Df(X)[H]$ можно представить в виде

  $$Df(X)[H] = \langle A_X, H\rangle, \quad \text{где } A_X \in \mathbb{R}^{n\times m} \text{ зависит от } X.$$

  Матрица A_X также называется `градиентом` $f(X)$ в точке X и обозначается $\nabla f(X)$.

  Аналогично подставив в качестве направлений $h = e_{ij}$, получим явное значение матрицы градиента в стандартном базисе со стандартным скалярным произведением

  $$\nabla f(X) = \left(\frac{\partial f}{\partial x_{ij}}(X)\right)_{i,j} \in \mathbb{R}^{n\times m}. \tag{10}$$

- В случае $f : \mathbb{R}^n \to \mathbb{R}^m$ линейный оператор $Df(x)[h]$ можно представить в виде

  $$Df(x)[h] = J_f(x)h, \quad \text{где } J_f(x) \in \mathbb{R}^{m\times n} \text{ зависит от } x.$$

  Матрица $J_x(x)$ называется `матрицей Якоби` $f(x)$ в точке x.

  Аналогично подставив в качестве направлений $h = e_i$, получим явное значение матрицы Якоби в стандартном базисе со стандартным скалярным произведением

  $$J_f(x) \equiv \frac{\partial f}{\partial x} := \left(\frac{\partial f_i}{\partial x_j}(x)\right)_{i,j} \in \mathbb{R}^{m\times n}. \tag{11}$$

- Во всех остальных случаях для построения производной достаточно найти все частные производные в виде тензора

  $$\frac{\partial f_{ij}}{\partial x_{kl}}(x).$$

Следует помнить, что из существования частных производных ещё не следует дифференцируемость. Однако на практике чаще всего все эти частные производные непрерывны, и, как следствие, функция дифференцируема.

Финальная таблица с каноническими видами

Вход \ Выход	Скаляр $\mathbb{R}$	Вектор $\mathbb{R}^n$
Скаляр $\mathbb{R}$	$df(x) = f'(x)dx$ $f'(x)$ скаляр, dx скаляр.	-
Вектор $\mathbb{R}^m$	$df(x) = \langle \nabla f(x), dx \rangle$ $f(x)$ вектор, dx вектор	$df(x) = J_x dx$ J_x матрица, dx вектор
Матрица $\mathbb{R}^{n' \times m'}$	$df(X) = \langle \nabla f(X), dX \rangle$ $\nabla f(X)$ матрица, dX матрица	-

Стоит отметить, что данная таблица верна и для произвольных скалярных произведений, а не только для стандартного.

5.2.3. Вторая производная

Пусть $f : U \to V$ дифференцируема в каждой точке $x \in U$. Рассмотрим дифференциал функции f при фиксированном приращении h_1 как функцию от x:

$$g(x) = Df(x)[h_1].$$

Определение 5.8 Вторая производная. Если в некоторой точке x функция g имеет производную, то она называется второй производной, а второй дифференциал имеет вид

$$D^2 f(x)[h_1, h_2] := D(Df[h_1])(x)[h_2]. \tag{12}$$

Можно показать, что $D^2 f(x)[h_1, h_2]$ билинейная функция по h_1, h_2. По аналогии определяется третий дифференциал $D^3 f(x)[h_1, h_2, h_3]$, четвёртый и так далее.

Определение 5.9 Непрерывная дифференцируемость. Если функция $Df(x)[h_1]$ является непрерывной по x, то f непрерывно дифференцируема.
Аналогично, если $D^n f(x)[h_1, \dots, h_n]$ непрерывна по x, то f n раз непрерывно дифференцируема.

В случае $f : \mathbb{R}^n \to \mathbb{R}$ второй дифференциал, как и любую билинейную функцию, можно представить с помощью матрицы

$$D^2 f(x)[h_1, h_2] = \langle H_x h_1, h_2 \rangle.$$

Матрица H_x называется `гессианом` функции f в точке x и обозначается $\nabla^2 f(x)$.

Предложение 5.10. Из формулы градиента верно, что

$$d(\nabla f(x)) = (\nabla^2 f)^\top dx \Leftrightarrow \nabla^2 f(x) = (J_{\nabla f})^\top.$$

В стандартном базисе гессиан имеет вид

$$\left(\nabla^2 f(x)\right)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

А также для дважды непрерывно дифференцируемой функции гессиан - симметричная матрица.

5.2.4. Теория дифференциалов

Напомним некоторые важные факты из курса математического анализа

Предложение 5.11. Пусть U, V — линейные пространства, $X \subset U, x \in X$ - внутренняя точка. Тогда справедливы следующие свойства:

- Линейность. Пусть $f : X \to V$ и $g : X \to V$. Если f, g дифференцируемы в точке x, при этом $c_1, c_2 \in \mathbb{R}$ числа, то $c_1 f + c_2 g$ дифференцируема в x и

$$d(c_1 f + c_2 g) = c_1 df + c_2 dg. \tag{13}$$

- Правило произведения. Пусть $\alpha : X \to \mathbb{R}$ и $f : X \to V$ функции. Если α, f дифференцируемы в точке x, то αf дифференцируема в точке x и

$$D(\alpha \cdot f)(x)[h] = (D\alpha(x)[h]) \cdot f(x) + \alpha(x) \cdot (Df(x)[h]) \tag{14}$$

для любых приращений h.

- Правило композиции. Пусть Y - подмножество V, $f : X \to Y$ - функция. Также пусть W линейное пространство, g :

$Y \to W$ - функция. Если f дифференцируема в точке x, g дифференцируема в точке $f(x)$, то их композиция $(g \circ f)(x) \equiv g(f(x))$ дифференцируема в точке x и

$$D(g \circ f)(x) = Dg(f(x))[df] \Longleftrightarrow Dg(f(x))[Df(x)[h]]. \quad (15)$$

- Правило частного. Пусть $\alpha : X \to \mathbb{R}$ и $f : X \to V$ - функции. Если α, f дифференцируемы в x и не обращается в 0 на X, то $(1/\alpha)f$ дифференцируема в x и

$$D\left(\frac{f}{\alpha}\right)(x)[h] = \frac{\alpha(x) \cdot (Df(x)[h]) - (D\alpha(x)[h]) \cdot f(x)}{\alpha(x)^2}. \quad (16)$$

- Правило произведения для матрично-значных функций. Пусть $f : X \to \mathbb{R}^{m \times n}$ и $g : X \to \mathbb{R}^{n \times k}$ - матрично-значные функции. Если f, g дифференцируемы в точке x, то $f \cdot g$ дифференцируема в x и

$$D(f \cdot g)(x)[h] = (Df(x)[h]) \cdot g(x) + f(x) \cdot (Dg(x)[h]). \quad (17)$$

Здесь подразумевается матричное умножение.

Следствие 5.12.

- Для дифференцируемых в точке x векторно-значных функций $f : X \to \mathbb{R}^n$ и $g : X \to \mathbb{R}^n$ функция $\langle f, g \rangle$ дифференцируема в x и

$$d(\langle f, g \rangle) = \langle df, g \rangle + \langle f, dg \rangle. \quad (18)$$

- Для дифференцируемой в точке x векторно-значной функции $f : X \to \mathbb{R}^n$ и линейного отображения $L : \mathbb{R}^n \to \mathbb{R}^m$ дифференциал и L перестановочны:

$$D(L \circ f)(x)[h] = L[Df(x)[h]]. \quad (19)$$

- Матрица Якоби сложной функции $g(f(x))$ равна произведению матриц Якоби композитов

$$J_{g(f(x))} = J_g J_f.$$

5.3. Примеры и задачи

5.3.1. Вычисление по определению

Пример 5.13. Табличные производные.

а) Для $f(x) = \langle c, x\rangle, x \in \mathbb{R}^n$ и приращения h считаем

$$\begin{aligned} f(x+h) - f(x) =& \langle c, x+h\rangle - \langle c, x\rangle \\ =& \langle c, h\rangle. \end{aligned}$$

Отображение $h \to \langle c, h\rangle$ является линейным, поэтому его можно принять за производную по определению

$$Df(x)[h] = \langle c, h\rangle.$$

б) Для $f(x) = \langle Ax, x\rangle, x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$ и приращения h считаем

$$\begin{aligned} f(x+h) - f(x) =& \langle Ax + Ah, x+h\rangle - \langle Ax, x\rangle \\ =& \langle (A + A^\top)x, h\rangle + \langle Ah, h\rangle. \end{aligned}$$

Заметим, что

$$\|Ah, h\| \le \|Ah\|\|h\| \le \|A\|\|h\|^2 = o(\|h\|),$$

где первое неравенство следует из Коши-Буняковского, а второе из согласованности матричной нормы. При этом $h \to \langle (A + A^\top)x, h\rangle$ линейный оператор.

Получается, что по определению

$$Df(x)[h] = \langle (A + A^\top)x, h\rangle.$$

в) Пусть $S := \{X \in \mathbb{R}^{n\times n} : \det(X) \ne 0\}$ и функция $f : S \to S$ обращает матрицу $f(X) = X^{-1}$. Для произвольного малого приращения H посчитаем

$$\begin{aligned} f(X+H) - f(X) =& (X+H)^{-1} - X^{-1} \\ =& (X(I_n + X^{-1}H))^{-1} - X^{-1} \\ =& ((I_n + X^{-1}H)^{-1} - I_n)X^{-1}. \end{aligned}$$

Отдельно оценим $(I_n + X^{-1}H)^{-1}$, для чего вспомним ряд Неймана. В нашем случае мы можем применить ряд Неймана в силу малости H

$$(I_n + X^{-1}H)^{-1} = I_n - X^{-1}H + \sum_{k=2}^{\infty}(-X^{-1}H)^k.$$

Оценим норму последнего слагаемого

$$\begin{aligned}\left\|\sum_{k=2}^{\infty}(-X^{-1}H)^k\right\| \leq& \sum_{k=2}^{\infty}\|(-X^{-1}H)^k\| \\ \leq& \sum_{k=2}^{\infty}\|-X^{-1}\|^k\|H\|^k \\ =& \frac{\|X^{-1}\|^2\|H\|^2}{1-\|X^{-1}\|\|H\|} \\ =& o(\|H\|),\end{aligned}$$

где первое неравенство получается из неравенства треугольника в пределе, второе – из свойств матричной нормы, а третье равенство – это сумма геометрической прогрессии.

В итоге получаем разность

$$f(X+H) - f(X) = -X^{-1}HX^{-1} + o(\|H\|),$$

при этом отображение $H \to -X^{-1}HX^{-1}$ линейно. То есть по определению

$$Df(X)[H] = -X^{-1}HX^{-1}.$$

г) Пусть $S := \{X \in \mathbb{R}^{n\times n} : \det(X) \neq 0\}$ и функция $f : S \to \mathbb{R}$ считает определитель $f(X) = \det(X)$.

Для малого приращения H посчитаем приращение функции

$$\begin{aligned}f(X+H) - f(X) =& \det(X+H) - \det(X) \\ =& \det(X(I_n + X^{-1}H)) - \det(X) \\ =& \det(X)(\det(I_n + X^{-1}H) - 1).\end{aligned}$$

Отдельно оценим $\det(I_n + X^{-1}H)$. Пусть $\lambda_i(X^{-1}H)$ – собственные числа матрицы $X^{-1}H$ (в произвольном порядке и, возможно, комплексные). Тогда собственными числами матрицы $I_n + X^{-1}H$ будут

$1 + \lambda_i(X^{-1}H)$. Поэтому

$$\begin{aligned}\det(I_n + X^{-1}H) =& \prod_{i=1}^{n}[1 + \lambda_i(X^{-1}H)] \\ =& 1 + \sum_{i=1}^{n} \lambda_i(X^{-1}H) + \left(\sum_{1 \le i \le j \le n} \lambda_i(X^{-1}H)\lambda_j(X^{-1}H) + \dots \right),\end{aligned}$$

где $\dots$ обозначает всевозможные тройки, четверки и т.д. из $\lambda_i(X^{-1}H)$.

Для произвольной матрицы $A \in \mathbb{R}^{n\times n}$ все её собственные числа по модулю не превосходят её нормы. Действительно, для собственного числа $\lambda \in \mathbb{C}$ и единичного по норме собственного вектора $x \in \mathbb{R}^n$ верно

$$Ax = \lambda x \Rightarrow \|\lambda x\| = \|Ax\| \le \|A\|\|x\|.$$

Следовательно, всё выражение в скобках будет $o(\|H\|)$, поскольку в каждом слагаемом больше одного собственного числа. Таким образом,

$$\begin{aligned}\det(I_n + X^{-1}H) =& 1 + \sum_{i=1}^{n} \lambda_i(X^{-1}H) + o(\|H\|) \\ =& 1 + \mathrm{Tr}(X^{-1}H) + o(\|H\|).\end{aligned}$$

Подставив это выражение для приращения функции, получим

$$f(X + H) - f(X) = \det(X)\,\mathrm{Tr}(X^{-1}H) + o(\|H\|).$$

Далее мы будем работать со стандартным скалярным произведением. Мы доказали формулу $d(\det(X)) = \det(X)\langle X^{-\top}, dX\rangle$ только для обратимых X. Однако формулу для дифференциала $d(\det(X))$ можно получить и для произвольной матрицы из $\mathbb{R}^{n\times n}$. Эта формула называется `формулой Якоби` и записывается как

$$d(\det(X)) = \langle \mathrm{Adj}(X)^\top, dX\rangle, \tag{20}$$

где $\mathrm{Adj}(X)$ - присоединённая матрица к X.

Присоединённая матрица определяется как $\mathrm{Adj}(X)_{ji} = (-1)^{(i+j)}M_{ij}$, где M_{ij} - дополнительный минор, определитель матрицы, получившийся вычеркиванием i-ой строки и j-ого столбца из X. В случае невырожденной X выполнено $\mathrm{Adj}(X) = \det(X)X^{-1}$ и формулы переходят друг в друга.

Вспомним, формулу вычисления определителя через дополнительные миноры по i-ой строке

$$\det(X) = \sum_k x_{ik} \cdot (-1)^{(i+k)} M_{ik}.$$

Тогда градиент равен

$$\frac{\partial f}{\partial x_{ij}} = (-1)^{(i+j)} M_{ij} = \mathrm{Adj}(X)_{ji}.$$

Финальная табличка с правилами преобразования и табличными значениями выглядит так

Правила преобразования
$d(\alpha X) = \alpha dX$
$d(AXB) = AdXB$
$d(X+Y) = dX + dY$
$d(X^T) = (dX)^T$
$d(XY) = (dX)Y + X(dY)$
$d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
$d(g(f(x))) = g'(f)df(x)$
$J_{g(f)} = J_g J_f \iff \frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$

Таблица стандартных производных
$dA = 0$
$d\langle A, X\rangle = \langle A, dX\rangle$
$d\langle Ax, x\rangle = \langle (A + A^\top)x, dx\rangle$
$d\,\mathrm{Tr}(X) = \mathrm{Tr}(dX)$
$d(\det(X)) = \det(X)\,\mathrm{Tr}(X^{-1}dX)$
$d(X^{-1}) = -X^{-1}(dX)X^{-1}$

Стоит отметить, что данные таблицы верны и для произвольных скалярных произведений, а не только для стандартного.

Hint. Для запоминания формулы $d(X^{-1})$ через произведение X и X^{-1}

$$\begin{aligned}
&I = XX^{-1},\\
&dI = 0 = d(XX^{-1}) = (dX)X^{-1} + Xd(X^{-1}),\\
&d(X^{-1}) = -X^{-1}(dX)X^{-1}.
\end{aligned}$$

Однако это не является доказательством `существования` дифференциала.

5.3.2. Дифференцирование по вектору

Для начала попрактикуемся в подсчёте градиентов и вторых производных для функций вида $f(x) : \mathbb{R}^n \to \mathbb{R}$.

Начнём с самой простой и часто встречаемой квадратичной функции. Попробуем посчитать её производные прямым и дифференциальным метод и сравним их.

Пример 5.14. Квадратичная функция. Найдите первый и второй дифференциал $df(x), d^2f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,$$

где $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$.

Решение. □ Попробуем применить оба подхода для решения данной задачи.

- Сначала используем прямой метод и выпишем явную скалярную зависимость $f(x_1, \ldots, x_n)$

  $$\begin{aligned} f(x_1, \ldots, x_n) =& \frac{1}{2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}x_j + \sum_{i=1}^{n} x_i b_i + c \\ =& \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j + \sum_{i=1}^{n} x_i b_i + c. \end{aligned}$$

  Найдём частную производную по x_k

  $$\begin{aligned} f(x_1, \ldots, x_n) =& \frac{1}{2}A_{kk}x_k^2 + \frac{1}{2}\sum_{i\neq k} A_{ik}x_i x_k + \frac{1}{2}\sum_{j\neq k} A_{kj}x_k x_j \\ & + x_k b_k + \left(\frac{1}{2}\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_i x_j + \sum_{i\neq k} x_i b_i + c\right). \end{aligned}$$

  Взяв частную производную, получим

  $$\frac{\partial f}{\partial x_k} = \frac{1}{2}\cdot 2A_{kk}x_k + \frac{1}{2}\sum_{i\neq k} A_{ik}x_i + \frac{1}{2}\sum_{j\neq k} A_{kj}x_j + b_k$$

$$= \frac{1}{2}(Ax)_k + \frac{1}{2}(A^\top x)_k + b_k.$$

Подставив координатно, посчитаем градиент

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

Для подсчёта гессиана найдём двойную частную производную по x_k, x_l

$$\begin{aligned}\frac{\partial^2 f}{\partial x_l \partial x_k} =& \frac{\partial \frac{1}{2}\sum\limits_{i=1}^{n} A_{ik}x_i + \frac{1}{2}\sum\limits_{j=1}^{n} A_{kj}x_j + b_k}{\partial x_l} \\ =& \frac{1}{2}A_{lk} + \frac{1}{2}A_{kl} \\ =& \frac{1}{2}(A + A^\top)_{kl}.\end{aligned}$$

Следовательно, гессиан равен

$$\nabla^2 f(x) = \frac{1}{2}(A + A^\top).$$

- Теперь используем дифференциальное исчисление

$$\begin{aligned} df(x) = & \quad d\left(\tfrac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c\right) \\ = & \quad \tfrac{1}{2}\langle (A + A^\top)x, dx\rangle + \langle b, dx\rangle + 0 \\ = & \quad \left\langle \tfrac{1}{2}(A + A^\top)x + b, dx\right\rangle. \end{aligned}$$

Следовательно, приведя к стандартному виду $df = \langle \nabla f(x), dx\rangle$, получаем градиент

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

Далее для гессиана фиксируем первое приращение dx_1 у первого дифференциала и берём уже от него ещё один дифференциал

$d^2 f = d(df)$

$$
\begin{aligned}
&= d\left\langle \frac{1}{2}(A + A^\top)x + b, dx_1 \right\rangle \\
&= \left\langle d\left(\frac{1}{2}(A + A^\top)x + b\right), dx_1 \right\rangle + \left\langle \frac{1}{2}(A + A^\top)x + b, \underbrace{d(dx_1)}_{=0} \right\rangle \\
&= \left\langle \frac{1}{2}(A + A^\top)dx, dx_1 \right\rangle.
\end{aligned}
$$

Переносим и транспонируем матрицу в скалярном произведении, но поскольку $A + A^\top$ симметричная, она не меняется.

$$
d^2 f = \left\langle dx, \frac{1}{2}(A + A^\top)^\top dx_1 \right\rangle = \left\langle \frac{1}{2}(A + A^\top)dx_1, dx \right\rangle.
$$

Следовательно, приведя к стандартному виду $d^2 f = \langle \nabla^2 f(x) \cdot dx_1, dx \rangle$, получаем гессиан

$$
\nabla^2 f(x) = \frac{1}{2}(A + A^\top).
$$

Заметим, что в случае если A симметричная, то

$$
\nabla f(x) = Ax + b, \quad \nabla^2 f(x) = A.
$$

■

С помощью примера 5.14 можно найти также градиент функции невязки $f(x) = \frac{1}{2}\|Ax - b\|^2, \quad x \in \mathbb{R}^n$.

Минимум квадрата невязки позволяет найти решение системы линейных уравнений с минимальной ошибкой по норме.

В Машинном обучении этот метод известен под названием Метод Наименьших Квадратов(МНК). По матрице признаков A и вектору параметров x мы будем пытаться линейно приближать целевой вектор b.

Далее посмотрим на пример посложнее на применение правила дифференцирования сложной функции.

Пример 5.15. Найдите первый и второй дифференциал $df(x), d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$
f(x) = \ln\langle Ax, x \rangle
$$

где $x \in \mathbb{R}^n, A \in \mathbb{S}^n_{++}$.

Решение. $\square$ Найдём первый дифференциал

$$df = d\ln\langle Ax, x\rangle = \frac{1}{\langle Ax, x\rangle} d\langle Ax, x\rangle = \frac{2\langle Ax, dx\rangle}{\langle Ax, x\rangle} = \left\langle \frac{2Ax}{\langle Ax, x\rangle}, dx \right\rangle.$$

Теперь найдём дифференциал градиента

$$\begin{aligned} d\left(\frac{2Ax}{\langle Ax, x\rangle}\right) =& \frac{d(2Ax)\langle Ax, x\rangle - (2Ax)d\langle Ax, x\rangle}{\langle Ax, x\rangle^2} \\ =& \frac{2\langle Ax, x\rangle A dx - 4Ax\langle Ax, dx\rangle}{\langle Ax, x\rangle^2} \\ =& \left(\frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}\right) dx = J_{\nabla f} dx. \end{aligned}$$

Поскольку $\nabla^2 f = (J_{\nabla f})^\top$, а гессиан симметричен из-за непрерывности, то

$$\nabla^2 f = \frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}.$$

$\blacksquare$

Пример 5.16. Евклидова норма. Найдите первый и второй дифференциал $df(x), d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \|x\|_2, \quad x \in \mathbb{R}^n \setminus \{0\}.$$

Решение. $\square$ Найдём первый дифференциал

$$\begin{aligned} df(x) =& d(\langle x, x\rangle^{\frac{1}{2}}) \\ =& \left\{ dy^{\frac{1}{2}} = \frac{1}{2y^{\frac{1}{2}}} dy \right\} \\ =& \frac{d(\langle x, x\rangle)}{2\langle x, x\rangle^{\frac{1}{2}}} \\ =& \left\langle \frac{2x}{2\langle x, x\rangle^{\frac{1}{2}}}, dx \right\rangle \\ =& \left\langle \frac{x}{\|x\|}, dx \right\rangle. \end{aligned}$$

После этого приведём df к стандартному виду $df = \langle \nabla f, dx\rangle$ и получим градиент

$$\nabla f(x) = \frac{x}{\|x\|}.$$

Теперь посчитаем второй дифференциал, зафиксировав приращение dx_1 первого

$$\begin{aligned}
df^2(x) =& d\left(\left\langle \frac{x}{\|x\|}, dx_1 \right\rangle\right) \\
=& \left\langle d\left(\frac{x}{\|x\|}\right), dx_1 \right\rangle = \textit{Правило частного} \\
=& \left\langle \frac{dx\|x\| - xd(\|x\|)}{\|x\|^2}, dx_1 \right\rangle \\
=& \left\langle \frac{dx\|x\| - x\left\langle \frac{x}{\|x\|}, dx\right\rangle}{\|x\|^2}, dx_1 \right\rangle \\
=& \left\langle \frac{I_n\|x\| - \frac{xx^\top}{\|x\|}}{\|x\|^2} dx, dx_1 \right\rangle \\
=& \left\langle \left(\frac{I_n\|x\| - \frac{xx^\top}{\|x\|}}{\|x\|^2}\right)^\top dx_1, dx \right\rangle.
\end{aligned}$$

Представив $d^2 f$ в стандартной форме $\langle \nabla^2 f(x) \cdot dx_1, dx\rangle$, получим

$$\nabla^2 f(x) = \frac{I_n}{\|x\|} - \frac{xx^\top}{\|x\|^3}.$$

Заметим, что в точке $x = 0$ функция не является дифференцируемой. НО при этом мы можем посчитать производную по любому направлению h:

$$\frac{\partial f}{\partial h}(0) = \lim_{t\to 0} \frac{f(0+th) - f(0)}{t} = \lim_{t\to +0} \frac{||th||}{t} = ||h||.$$

Если бы функция была дифференцируем, то

$$df(x)[h] = ||h||,$$

а это НЕлинейная функция от h. ■

Пример 5.17. Куб Нормы. Найдите первый и второй дифференциал $df(x), d^2f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \frac{1}{3}\|x\|_2^3, \quad x \in \mathbb{R}^n.$$

Решение. □ Найдём первый дифференциал

$$\begin{aligned}
df(x) =& \frac{1}{3} d\langle x, x\rangle^{3/2} \\
=& \frac{1}{3}\cdot\frac{3}{2}\langle x, x\rangle^{1/2} d\langle x, x\rangle \\
=& \frac{1}{2}\langle x, x\rangle^{1/2}\cdot 2\langle x, dx\rangle \\
=& \langle x\|x\|, dx\rangle.
\end{aligned}$$

Приведя к стандартному виду $df = \langle \nabla f(x), dx\rangle$, получаем

$$\nabla f(x) = \|x\|x.$$

Найдём второй дифференциал

$$\begin{aligned}
d^2f(x) =& d(\|x\|\langle x, dx_1\rangle) \\
=& \underbrace{d(\|x\|)}_{=d\langle x,x\rangle^{1/2}}\langle x, dx_1\rangle + \|x\|d(\langle x, dx_1\rangle) \\
=& \left(\frac{1}{2}\langle x, x\rangle^{-1/2} 2\langle x, dx\rangle\right)\langle x, dx_1\rangle + \|x\|\langle dx, dx_1\rangle \\
=& \frac{1}{\|x\|}\langle x, dx\rangle\langle x, dx_1\rangle + \|x\|\langle dx, dx_1\rangle \\
=& \left\langle dx, \left(\frac{xx^\top}{\|x\|} + I_n\|x\|\right) dx_1\right\rangle.
\end{aligned}$$

Представив d^2f в стандартной форме $\langle \nabla^2 f(x)\cdot dx_1, dx\rangle$, получим

$$\nabla^2 f(x) = \frac{xx^\top}{\|x\|} + I_n\|x\|$$

Заметим, что в данная формула не определена в точке $x = 0$, поскольку ранее мы пользовались правилом дифференцирования функции $\sqrt{x}$, а её производная не определена в 0.

Однако можно найти вторую производную в $x = 0$ по определению. Зафиксируем приращение h_1 и рассмотрим

$$\begin{aligned}
&\lim_{h_2 \to 0} \frac{\|(Df[h_1])(0 + h_2) - (Df[h_1])(0)\|}{\|h_2\|} \\
&= \lim_{h_2 \to 0} \frac{|\langle (0 + h_2)\|0 + h_2\| - 0\|0\|, h_1 \rangle|}{\|h_2\|} \\
&= \lim_{h_2 \to 0} \frac{\|h_2\| |\langle h_2, h_1 \rangle|}{\|h_2\|} \\
&= \lim_{h_2 \to 0} |\langle h_2, h_1 \rangle| = 0.
\end{aligned}$$

Следовательно, по определению вторая производная в точке $x = 0$ равна 0. Можно даже сказать, что функция дважды непрерывно дифференцируема, потому что $\lim\limits_{x \to 0} \left(\frac{xx^\top}{\|x\|} + I_n \|x\| \right) = 0$. ■

Рассмотрим часто встречающую в Deep Learning и нейронных сетях функцию softmax, которая позволяет отобразить вектор из n координат в распределение вероятностей на n исходах. Например, многоклассовой классификации тем самым мы получаем вектор вероятностей принадлежности объекта каждому из n классов.

Пример 5.18. Softmax. Найдите матрицу Якоби функции $s(x) = \operatorname{softmax}(x)$

$$\operatorname{softmax}(x) := \left(\frac{\exp(x_1)}{\sum_{i=1}^n \exp(x_i)}, \ldots, \frac{\exp(x_n)}{\sum_{i=1}^n \exp(x_i)} \right)^\top.$$

Решение. □ Считаем частные производные по определению

а) при $k \neq j$

$$\begin{aligned}
\frac{\partial s_k}{\partial x_j} &= \frac{\partial}{\partial x_j} \frac{\exp(x_k)}{\sum_{i=1}^n \exp(x_i)} \\
&= \exp(x_k) \frac{\partial}{\partial x_j} \frac{1}{\sum_{i=1}^n \exp(x_i)}
\end{aligned}$$

$$
\begin{aligned}
&= \exp(x_k) \frac{-1}{(\sum_{i=1}^n \exp(x_i))^2} \frac{\partial}{\partial x_j} \left(\sum_{i=1}^n \exp(x_i) \right) \\
&= - \frac{\exp(x_k) \exp(x_j)}{(\sum_{i=1}^n \exp(x_i))^2} \\
&= - s_k \cdot s_j,
\end{aligned}
$$

б) при $k = j$

$$
\begin{aligned}
\frac{\partial s_j}{\partial x_j} =& \frac{\partial}{\partial x_j} \frac{\exp(x_j)}{\sum_{i=1}^n \exp(x_i)} \\
=& \frac{\exp(x_j)(\sum_{i=1}^n \exp(x_i)) - \exp(x_j) \frac{\partial}{\partial x_j} (\sum_{i=1}^n \exp(x_i))}{(\sum_{i=1}^n \exp(x_i))^2} \\
=& \frac{\exp(x_j)}{\sum_{i=1}^n \exp(x_i)} - \frac{\exp(x_j) \exp(x_j)}{(\sum_{i=1}^n \exp(x_i))^2} \\
=& s_j(1 - s_j).
\end{aligned}
$$

Итого,

$$
J_{k,j} = \begin{cases} -s_k \cdot s_j, & k \neq j \\ s_j(1 - s_j), & k = j. \end{cases}
$$

■

Не менее часто в DL встречаются покоордиантные функции, которые применяются на выходе очередного слоя для каждого отдельного нейрона. Посмотрим, как через них считать градиент.

Пример 5.19. Покоординатные операции. Найдите градиент и гессиан функции $f(x) = h(g(x))$, где

$$
g(x) = \sin(x) \text{ поэлементно},
$$

$$
h(u) = \sum_{i=1}^n u_i.
$$

Решение. □ В этом примере нам в любом случае нужно будет применять именно первый подход для подсчёта матриц Якоби и градиентов. Действительно, входящие функции не являются стандартными, но являются достаточно легкими, чтобы считать частные производные напрямую.

Также полезно вспомнить правило матрицы Якоби сложной функции

$$J_f = J_{h(g)} J_g,$$

оно же с градиентами имеет вид

$$\nabla f = J_g^\top \nabla h.$$

Далее посчитаем матрицу Якоби покоординатной функции вида $g(x) = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}$

$$J_g = \mathrm{diag}(g'(x_1), \ldots, g'(x_n)) = \mathrm{diag}(g'(x)) = J_g^\top.$$

При умножении J_g на вектор удобно пользоваться поэлементным умножением матриц, обозначающимся $\odot$

$$(A \odot B)_{ij} = A_{ij} * B_{ij}.$$

Результат умножения J_g на вектор y равен

$$J_g y = \begin{pmatrix} g'(x_1) \\ \vdots \\ g'(x_n) \end{pmatrix} \odot y = g'(x) \odot y.$$

Заметим, что эта операция является довольно быстро вычислимой и легко поддаётся параллелизации.

Теперь приступим к непосредственному примеру

$$\begin{aligned}
&J_g = \mathrm{diag}(\cos(x_1), \ldots, \cos(x_n)) = \mathrm{diag}(\cos(x)), \\
&\{\nabla h(u)\}_j = \frac{\partial(\sum_{i=1}^n u_i)}{\partial u_j} = 1 \quad \to \quad \nabla h(u) = \mathbf{1}, \\
&\nabla f = J_g^\top \nabla h = \cos(x) \odot \mathbf{1} = \cos(x).
\end{aligned}$$

Теперь гессиан функции, который считается по формуле

$$\nabla^2 f(x) = J_{\nabla f}^{\top} = \mathrm{diag}(-\sin(x)).$$

$\blacksquare$

Логистическая регрессия – модель машинного обучения для двухклассовой классификации. Более подробно про саму модель и интуицию за ней можно почитать здесь. Её обучение может быть сведено к оптимизации функции, представленной в примере ниже.

Пример 5.20. Логистическая регрессия. Найдите первый и второй дифференциал $df(x), d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \ln(1 + \exp(\langle a, x\rangle)),$$

где $a \in \mathbb{R}^n$.

Решение. $\square$ Найдём первый дифференциал

$$\begin{aligned} d(\ln(1 + \exp(\langle a, x\rangle))) =& \{d \ln y = \frac{1}{y} dy\} \\ =& \frac{1}{1 + \exp(\langle a, x\rangle)} d(1 + \exp(\langle a, x\rangle)) \\ =& \{d \exp(y) = \exp(y) dy\} \\ =& \frac{1}{1 + \exp(\langle a, x\rangle)} \exp(\langle a, x\rangle) d(\langle a, x\rangle) \\ =& \left\langle \frac{\exp(\langle a, x\rangle)}{1 + \exp(\langle a, x\rangle)} a, dx \right\rangle. \end{aligned}$$

Для удобства введём функцию сигмоиды $\sigma(x) := \frac{1}{1+\exp(-x)}$. При этом заметим, что $\sigma(-x) = 1 - \sigma(x)$ и $\sigma'(x) = \sigma(x)(1 - \sigma(x))$. После этого приведём df к стандартному виду $df = \langle \nabla f, dx\rangle$ и получим градиент

$$\nabla f(x) = \sigma(\langle a, x\rangle) a.$$

Таким образом, градиент $\nabla f(x)$ - вектор коллинеарный вектору a с коэффициентом $\sigma(\langle a, x\rangle) \in (0, 1)$. В зависимости от точки x меняет лишь длина градиента, но не направление.

Теперь посчитаем второй дифференциал, зафиксировав приращение dx_1 первого

$$\begin{aligned}
d(df) =& d(\langle \sigma(\langle a, x\rangle)a, dx_1\rangle)\\
=& \langle d(\sigma(\langle a, x\rangle))a, dx_1\rangle\\
=& \langle \sigma'(\langle a, x\rangle)d(\langle a, x\rangle)a, dx_1\rangle\\
=& \langle \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle a, dx\rangle a, dx_1\rangle\\
=& \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle \langle dx, a\rangle a, dx_1\rangle\\
=& \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))(dx^\top aa^\top dx_1)\\
=& \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle aa^\top dx_1, dx\rangle.
\end{aligned}$$

Представив $d^2 f$ в стандартной форме $\langle \nabla^2 f(x)\cdot dx_1, dx\rangle$, получим

$$\nabla^2 f(x) = \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))aa^\top.$$

Заметим, что $\nabla^2 f$ - одноранговая матрица, пропорциональная $aa^\top$ с коэффициентом $\sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle)) \in (0, 0.25)$. Точка x влияет лишь на коэффициент. ■

Полезно понимать, как работать не только с векторным входом, но и со скалярным.

Пример 5.21. Дифференциал скаляра. Рассмотрим функцию скалярного аргумента α

$$\phi(\alpha) := f(x + \alpha p), \quad \alpha \in \mathbb{R},$$

$x, p \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ -дважды непрерывно дифференцируемая функция. Найдите первую и вторую производные $\phi'(\alpha), \phi''(\alpha)$ и выразите их через $\nabla f, \nabla^2 f$.

Решение. □ Важно помнить, что дифференцирование происходит не по стандартному вектору x, а по скаляру α со всеми вытекающими свойствами

$$\begin{aligned}
d\phi =& \{df = \langle \nabla f(y), dy\rangle\}\\
=& \langle \nabla f(x+\alpha p), d(x+\alpha p)\rangle\\
=& \langle \nabla f(x+\alpha p), d(\alpha)p\rangle\\
=& \langle \nabla f(x+\alpha p), p\rangle d(\alpha).
\end{aligned}$$

Заметим, что мы привили дифференциал к стандартному виду $d\phi = \phi'(\alpha) \cdot d\alpha$, то есть множитель перед $d\alpha$ – это производная

$$\phi'(\alpha) = \langle \nabla f(x + \alpha p), p \rangle.$$

Теперь вторая производная

$$\begin{aligned} d(\phi'(\alpha)) =& d\langle \nabla f(x + \alpha p), p \rangle \\ =& \{d(\nabla f(y)) = (\nabla^2 f(y))^\top dy\} \\ =& \langle (\nabla^2 f(x + \alpha p))^\top d(x + p\alpha), p \rangle \\ =& \langle (\nabla^2 f(x + \alpha p))^\top p d\alpha, p \rangle = \{\nabla^2 f(y) \\ =& (\nabla^2 f(y))^\top\}] \\ =& \langle \nabla^2 f(x + \alpha p) p, p \rangle d\alpha. \end{aligned}$$

Получается, что

$$\phi''(\alpha) = \langle \nabla^2 f(x + \alpha p) p, p \rangle.$$

■

5.3.3. Дифференцирование по матрице

Далее будем считать градиенты по матрицы для функций вида $f(X) : \mathbb{R}^{n \times m} \to \mathbb{R}$. В них активно применяются производные от таких матричных функций, как $\det, \mathrm{Tr}, X^{-1}$.

Пример 5.22. Фробениусова норма. Найти градиент $\nabla f(X)$ и дифференциал $df(X)$ функции $f(X)$

$$f(X) = \|AX - B\|_F, \quad X \in \mathbb{R}^{k \times n},$$

где $A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{m \times n}$.

Решение. □ Вычислим отдельно $d(\|X\|)$

$$\begin{aligned} d(\|X\|) =& d(\langle X, X \rangle^{\frac{1}{2}}) \\ =& \left\{ dy^{\frac{1}{2}} = \frac{1}{2y^{\frac{1}{2}}} dy \right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{d(\langle X, X\rangle)}{2\langle X, X\rangle^{\frac{1}{2}}} \\
&= \left\langle \frac{2X}{2\langle X, X\rangle^{\frac{1}{2}}}, dX \right\rangle \\
&= \left\langle \frac{X}{\|X\|}, dX \right\rangle.
\end{aligned}
$$

Тогда первый дифференциал равен

$$
\begin{aligned}
df(X) &= d(\|AX - B\|_F) \\
&= \left\langle \frac{AX - B}{\|AX - B\|}, d(AX - B) \right\rangle \\
&= \left\langle \frac{AX - B}{\|AX - B\|}, AdX \right\rangle \\
&= \mathrm{Tr}\left(\frac{(AX - B)^\top}{\|AX - B\|} AdX \right) \\
&= \mathrm{Tr}\left(\left(\frac{A^\top (AX - B)}{\|AX - B\|} \right)^\top dX \right) \\
&= \left\langle \frac{A^\top (AX - B)}{\|AX - B\|}, dX \right\rangle.
\end{aligned}
$$

Приведя к стандартному виду $df(X) = \langle \nabla f(X), dX \rangle$, получим

$$
\nabla f(X) = \frac{A^\top (AX - B)}{\|AX - B\|}.
$$

■

Пример 5.23. Найти градиент $\nabla f(X)$ и дифференциал $df(X)$ функции $f(X)$

$$
f(X) = \mathrm{Tr}(AXBX^{-1}), \quad X \in \mathbb{R}^{n \times n}, \det(X) \neq 0,
$$

где $A, B \in \mathbb{R}^{n \times n}$.

Решение. □ Перепишем след через скалярное произведение

$$
f(X) = \langle I_n, AXBX^{-1} \rangle.
$$

Найдём первый дифференциал

$$
\begin{aligned}
df(X) =& \langle I_n, d(AXBX^{-1})\rangle = \langle I_n, Ad(XBX^{-1})\rangle \\
=& \langle I_n, A(dX)BX^{-1} + AXd(BX^{-1})\rangle \\
=& \langle I_n, A(dX)BX^{-1} + AXB\cdot(-X^{-1}(dX)X^{-1})\rangle \\
=& \operatorname{Tr}(A(dX)BX^{-1}) - \operatorname{Tr}(AXBX^{-1}(dX)X^{-1}) \\
=& \operatorname{Tr}(BX^{-1}A(dX)) - \operatorname{Tr}(X^{-1}AXBX^{-1}(dX)) \\
=& \langle A^\top X^{-\top}B^\top - X^{-\top}B^\top X^\top A^\top X^{-\top}, dX\rangle.
\end{aligned}
$$

При работе со скалярными произведениями можно переходить к следу и обратно, для применения его полезных циклических и не только свойств. Главное, не забывать о транспонировании.

Приведя к стандартному виду $df(X) = \langle \nabla f(X), dX\rangle$, получим

$$\nabla f(X) = A^\top X^{-\top}B^\top - X^{-\top}B^\top X^\top A^\top X^{-\top}.$$

■

Пример 5.24. Найти градиент $\nabla f(X)$ и дифференциал $df(X)$ функции $f(X)$

$$f(X) = \operatorname{Tr}(AX^\top X).$$

Решение. □ Перепишем след через скалярное произведение для удобства

$$f(X) = \langle I, AX^\top X\rangle.$$

Найдём первый дифференциал

$$
\begin{aligned}
df(X) =& d\langle I, AX^\top X\rangle = \langle I, Ad(X^\top X)\rangle \\
=& \langle I, Ad(X^\top)X\rangle + \langle I, AX^\top dX\rangle \\
=& \langle I, A(dX)^\top X\rangle + \langle (AX^\top)^\top I, dX\rangle \\
=& \operatorname{Tr}(I^\top A(dX)^\top X) + \langle XA^\top, dX\rangle \\
=& \{\operatorname{Tr}(Y) = \operatorname{Tr}(Y^\top)\} = \operatorname{Tr}(X^\top dXA^\top) + \langle XA^\top, dX\rangle \\
=& \operatorname{Tr}(A^\top X^\top dX) + \langle XA^\top, dX\rangle \\
=& \langle XA, dX\rangle + \langle XA^\top, dX\rangle \\
=& \langle XA + XA^\top, dX\rangle.
\end{aligned}
$$

Приведя к стандартному виду $df(X) = \langle \nabla f(X), dX \rangle$, получим

$$\nabla f(X) = XA + XA^\top.$$

■

Пример 5.25. Логарифм определителя. Найдите первый и второй дифференциалы $df(X)$ и $d^2f(X)$, а также градиент $\nabla f(X)$ функции $f(X)$

$$f(X) = \ln(\det(X))$$

заданной на множестве $X \in \mathbb{S}^n_{++}$ в пространстве $\mathbb{S}^n$.

Решение. □ Заметим, что из положительной определённости следует $\det(X) > 0$, поэтому $f(X)$ определена корректно в каждой точке.

Найдём первый дифференциал

$$\begin{aligned} df(X) =& d(\ln\det(X)) \\ =& \frac{d(\det(X))}{\det(X)} \\ =& \frac{\det(X)\langle X^{-\top}, dX \rangle}{\det(X)} \overset{X \in \mathbb{S}^n_{++}}{=} \langle X^{-1}, dX \rangle. \end{aligned}$$

Приведя к стандартному виду $df(X) = \langle \nabla f(X), dX \rangle$, получим

$$\nabla f(X) = X^{-1}.$$

Теперь найдём второй дифференциал от первого с фиксированным приращением dX_1

$$d^2f(X) = \langle d(X^{-1}), dX_1 \rangle = -\langle X^{-1}(dX)X^{-1}, dX_1 \rangle.$$

Мы получили билинейную форму от dX, dX_1, выписать тензор производных в явном виде мы не будем.

Посмотрим, является ли эта форма отрицательно полуопределённой при фиксированной $X \in \mathbb{S}^n_{++}$. Для этого возьмём $H \in \mathbb{S}^n$ из исходного пространства.

Поскольку $X \in \mathbb{S}^n_{++}$, то $X^{-1} \in \mathbb{S}^n_{++}$. Матрицу X^{-1} можно разложить на произведение двух одинаковых матриц, обозначим их $X^{-1/2}$, т.е. $X^{-1} = X^{-1/2}X^{-1/2}$. Такое разложение можно получить, перейдя в базис из собственных векторов S, который всегда существует для симметричных матриц, при этом все собственные значения будут положительными:

$$X^{-1} = S\Lambda S^{-1} \quad \Rightarrow X^{-1/2} = S\sqrt{\Lambda}S^{-1}.$$

Тогда

$$\begin{aligned} d^2 f(X)[H,H] &= -\left\langle X^{-1}HX^{-1}, H\right\rangle \\ &= -\operatorname{Tr}(X^{-1}HX^{-1}H) \\ &= -\operatorname{Tr}(X^{-1/2}X^{-1/2}HX^{-1/2}X^{-1/2}H) \\ &= -\operatorname{Tr}(X^{-1/2}HX^{-1/2}\cdot X^{-1/2}HX^{-1/2}) \\ &= -\left\langle X^{-1/2}HX^{-1/2}, X^{-1/2}HX^{-1/2}\right\rangle \\ &= -\|X^{-1/2}HX^{-1/2}\|_F^2 \leq 0. \end{aligned}$$

По одному из критериев выпуклости, который мы узнаем на следующих семинарах, можно сказать по полученному неравенству, что $f(X) = \ln\det(X)$ является вогнутой на $\mathbb{S}^n_{++}$. ■

Пример 5.26. Найдите первый дифференциал $df(X)$ и градиент $\nabla f(X)$ функции $f(X)$

$$f(X) = \det(AX^{-1}B),$$

где A, X, B – такие матрицы с нужными размерностями, что $AX^{-1}B$ обратима.

Решение. □ Найдём первый дифференциал

$$\begin{aligned} df(X) =& d(\det(AX^{-1}B)) = \{d\det(Y) = \det(Y)\langle Y^{-\top}, dY\rangle\} \\ =& \det(AX^{-1}B)\langle (AX^{-1}B)^{-\top}, d(AX^{-1}B)\rangle \\ =& \det(AX^{-1}B)\langle (AX^{-1}B)^{-\top}, Ad(X^{-1})B\rangle \\ =& \{d(Y^{-1}) = -Y^{-1}(dY)Y^{-1}\} \\ =& -\det(AX^{-1}B)\langle (AX^{-1}B)^{-\top}, AX^{-1}(dX)X^{-1}B\rangle \\ =& -\det(AX^{-1}B)\operatorname{Tr}((AX^{-1}B)^{-1}AX^{-1}(dX)X^{-1}B) \\ =& -\det(AX^{-1}B)\operatorname{Tr}(X^{-1}B(AX^{-1}B)^{-1}AX^{-1}(dX)) \end{aligned}$$

$$= -\det(AX^{-1}B)\langle (X^{-1}B(AX^{-1}B)^{-1}AX^{-1})^\top, dX\rangle.$$

Приведя к стандартному виду $df(X) = \langle \nabla f(X), dX\rangle$, получим

$$\nabla f(X) = -\det(AX^{-1}B)X^{-\top}A^\top(AX^{-1}B)^{-\top}B^\top X^{-\top}.$$

■

Теперь посмотрим на случай, когда функция переводит матрица в матрицу, и записать производные в компактном виде через матрицу или вектор не получается.

Пример 5.27. Тензор производных. Найдите первый дифференциал и производную функции $f(A)$

$$f(A) = Ax,$$

где $A \in \mathbb{R}^{n\times m}$ – переменная, а $x \in \mathbb{R}^m$ - фиксированный вектор.

Решение. □ Дифференциал найти достаточно просто

$$df(A) = (dA)x.$$

Однако записать производную в виде вектора или матрицы уже не выйдет нужен трёхмерный тензор вида $\frac{\partial f_k}{\partial A_{ij}}(A)$ размерности $n \times n \times m$. В этом случае может быть полезен прямой подход и тензорное исчисление. Для этого выразим скалярную зависимость функции $f(A)$

$$f_k(A) = \sum_{l=1}^{m} A_{kl}x_l.$$

Теперь возьмём производную $\frac{\partial}{\partial A_{ij}}$

$$\begin{aligned}\frac{\partial f_k}{\partial A_{ij}}(A) &= \sum_{l=1}^{m} \frac{\partial (A_{kl}x_l)}{\partial A_{ij}} \\ &= \sum_{l=1}^{m} \delta(i,j=k,l)x_l \\ &= \sum_{l=1}^{m} \delta(i=k)\delta(j=l)x_l\end{aligned}$$

$$=\delta(i=k)x_j.$$

■

5.4. Приложения в задачах

Пример 5.28. (Поиск индуцированной нормы матрицы). Сперва введем определение индуцированной p-нормы матрицы.

$$\|A\|_p = \max_x \frac{\|Ax\|_p}{\|x\|_p}. \tag{21}$$

Именно из определения понятно, почему она называется индуцированной – она порождена p-нормой вектора. На всякий случай также введем p-норму вектора $x \in \mathbb{R}^n$, где $p \geq 1$:

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

Также можно заметить, что если в выражении $\frac{\|Ax\|_p}{\|x\|_p}$ заменить x на tx, где $t \neq 0$, то значение этого выражения не изменится. Действительно,

$$\frac{\|tAx\|_p}{\|tx\|_p} = \frac{|t|\|Ax\|_p}{|t|\|x\|_p} = \frac{\|Ax\|_p}{\|x\|_p}.$$

Из этого следует, что максимум из (21) можно искать не среди всех x, а только из x с определенным значением нормы, т.е. можно положить $\|x\|_p = 1$.
Здесь мы условимся, что далее мы рассматриваем только случай $p = 2$ (так как в этом случае наш пример будет нагляднее и проще).
Исходя из вышесказанного, задачу поиска индуцированной нормы матрицы A можно записать в следующем виде:

$$\begin{aligned} \text{Maximize} \quad & \|Ax\|_2^2 \\ s.t. \quad & \|x\|_2^2 = 1. \end{aligned} \tag{22}$$

На всякий случай уточним 2 детали. *s.t.* расшифровывается как *subject to*, т.е. *при следующих ограничениях*. Также для удобства мы возвели выражения в квадрат (это ничего не портит, так как нормы всегда неотрицательные).
Стандартный подход решения такой задачи - запись Лагранжиана:

$$L(x, \nu) = \|Ax\|_2^2 - \nu \left(\|x\|_2^2 - 1\right).$$

Тогда

$$\frac{\partial L}{\partial x} = 2A^\top Ax - 2\nu x.$$

Приравнивая $\frac{\partial L}{\partial x}$ к 0, получаем

$$\left(A^\top A\right) x = \nu x,$$

а значит искомый x это собственный вектор матрицы $A^\top A$, а ν – собственное значение. Но тогда, если расписать $\|Ax\|_2^2$ как $x^\top A^\top Ax$, и заменить $A^\top Ax$ на νx (как и было получено выше), то получим $x^\top \nu x$, а это в точности равно ν (в силу $\|x\|_2^2 = 1$). Как следствие, мы ищем максимально возможное ν. Вспоминая, что это собственное значение матрицы $A^\top A$, мы получаем, что этот максимум равен $\lambda_{max}[A^\top A]$. И поскольку мы возводили выражения в квадрат, то мы максимизировали **квадрат** индуцированной нормы матрицы. А значит,

$$\|A\|_2 = \sqrt{\lambda_{max}[A^\top A]}.$$

Замечание 5.29. Во-первых, внимательные читатели могли заметить, что не для всех пространств можно определить $\|A\|$ через максимум какого-то выражения. Это связано с рассматриваемыми пространствами – если они бесконечномерные, то требуется заменить max на sup. В методах оптимизации, как правило, этого не требуется. Во-вторых, трюк, связанный с переходом к ограничению $\|x\| = 1$ является классическим в такой дисциплине, как функциональный анализ, поэтому подробнее о данном переходе будет рассказано там.

Пример 5.30. (Вывод с наименьшей квадратичной ошибкой).
Рассмотрим следующую задачу. Пусть у нас есть какая-то модель,

которая для некоторого значения x возвращает y по следующему правилу: $y = Ax + n$, где $A \in \mathbb{R}^{m\times d}$ - некоторая матрица, а n - некоторый шум. Пусть эта модель выдала некоторое значение $\hat{y}$, и мы хотим понять, какой вход x для модели наиболее вероятен, чтобы получить этот $\hat{y}$. Формально это можно записать следующим образом:

$$\hat{x} = \underset{x:\hat{y}=Ax+n}{\arg\min}\ \|n\|_2^2.$$

Тогда, выразив n как $n = \hat{y} - Ax$, мы будем искать минимум $\|\hat{y} - Ax\|_2^2$. Для этого нам нужно посчитать градиент и приравнять его к 0.

$$\begin{aligned} d(\|\hat{y} - Ax\|_2^2) &= d(\langle \hat{y} - Ax, \hat{y} - Ax\rangle) = -\langle Adx, \hat{y} - Ax\rangle - \langle \hat{y} - Ax, Adx\rangle \\ &= -2\langle \hat{y} - Ax, Adx\rangle = -2\langle A^\top(\hat{y} - Ax), dx\rangle. \end{aligned}$$

Отсюда получаем, что градиент равен $\nabla f(x) = -2A^\top(\hat{y} - Ax)$. Приравнивая его к 0 мы получаем, что $\hat{x} = (A^\top A)^{-1}A^\top\hat{y}$, если матрица $A^\top A$ обратима. В противном случае ответ получить несколько сложнее.

Замечание 5.31. На самом деле, эта задача не что иное, как задача линейной регрессии, а решение выше – классический вывод решения задачи линейной регрессии. С данным объектом у Вас будет более детальное знакомство в курсе машинного обучения.

5.5. Автоматическое дифференцирование

5.5.1. Граф вычислений

Настало время поговорить о том, как происходит подсчёт градиентов в реальной жизни. Чаще всего функции, с которыми приходится иметь дело на практике представляют собой последовательность (дифференцируемых) параметрических преобразований. Таким образом, их можно представить в виде вычислительного графа (computational graph), где промежуточным вершинам соответствуют преобразования, входящим стрелкам – входные переменные, а выходным стрелкам – результат преобразования. Этот граф должен быть ациклическим, то есть DAG.

На рисунке 2 приведён вычислительный граф для логистической регрессии $f(x, \omega, y) = -\log(1 + \exp(-y\omega^\top x))$ – сплошные линии.

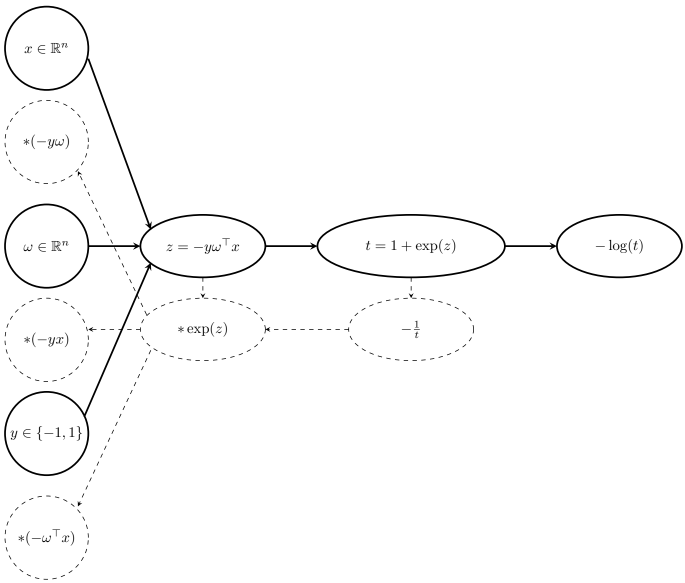

Рис. 2. Граф логистической регрессии

Вычисление значения функции по заданному входу часто называют прямым проходом, или же forward propagation (forward pass). На этом этапе происходит преобразование исходного вектора в целевой. Последовательно строятся промежуточные значения — результаты применения преобразований к предыдущим значениям слева направо. Именно поэтому проход называют прямым. В случае линейного графа можно записать его формулой

$$f(x) = u_m(u_{m-1}(\dots u_1(x)\dots)). \tag{23}$$

5.5.2. Backpropagation

Но как же считать производные в таких графах? Ответ может дать правило вычисления дифференциала сложной функции. Рассмотрим одну конкретную вершину в графе вида $u(x_1, \dots, x_d)$ и её дифференциал

$$du = \sum_{i=1}^{d} \frac{\partial u}{\partial x_i}(x_i)dx_i,$$

где $\frac{\partial u}{\partial x_i}(x_i)$ – частная производная по переменной x_i. Причём каждый следующий дифференциал dx_i можно расписать через предыдущую вершину (если, конечно, x_i не являются искомыми переменными), двигаясь по рекурсии в графе от детей к родителям. В случае линейного графа (23) получим итоговую формулу

$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial u_m}{\partial u_{m-1}}(u_{m-1})}_{\frac{\partial f}{\partial u_{m-1}}} \cdot \frac{\partial u_{m-1}}{\partial u_{m-2}}(u_{m-2}) \cdot \dots \cdot \frac{\partial u_1}{\partial x}(x). \tag{24}$$

Основная идея backward pass или backpropagation заключается в подсчёте формулы (24) **слева направо**.

В общем случае, пусть $u_1, \dots, u_m$ – вершины графа вычислений в топологическом порядке (т.е. родители идут перед детьми). Обозначим производную функции f по вершине u_i как

$$\overline{u_i} = \frac{\partial f}{\partial u_i}.$$

Общий алгоритм действий выглядит так

а) Произвести forward pass и сохранить все значения u_i как функции от их родителей.

б) Положить $\overline{u_m} = 1$ и для всех $i = m-1, \dots, 1$ посчитать

$$\overline{u_i} = \sum_{j \in \text{потомки}(u_i)} \overline{u_j} \frac{\partial u_j}{\partial u_i}.$$

Вычисление backpropagation на рисунке 2 показано пунктирными линиями.

Пример 5.32. Посчитаем шаг backpropagation в графе вычислений 3, где параметры x, ω - векторы, b - скаляр, а λ, t – заранее фиксированные константы.

Forward pass

$$\begin{aligned} z =& \omega^\top x + b, \\ y =& \sigma(z) = \frac{1}{1+\exp(-z)}, \\ L =& \frac{1}{2}(y-t)^2, \\ R =& \frac{1}{2}\omega^\top \omega, \end{aligned}$$

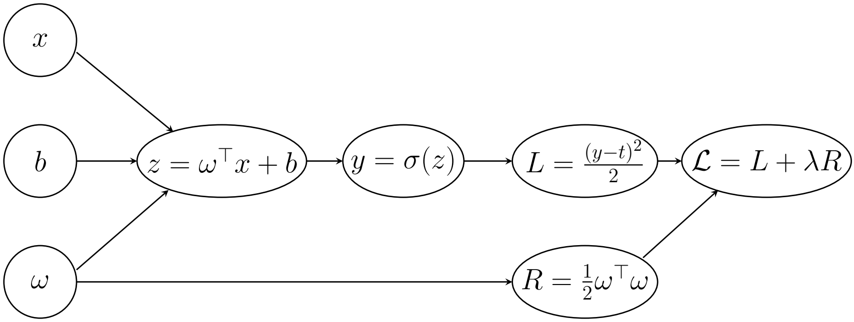

Рис. 3. Граф Вычислений

$$\mathcal{L} = L + \lambda R.$$

Backward propogation

$$
\begin{aligned}
\overline{\mathcal{L}} &= 1, \\
\overline{R} &= \overline{\mathcal{L}} \frac{d\mathcal{L}}{dR} = \overline{\mathcal{L}} \cdot \lambda = \lambda, \\
\overline{L} &= \overline{\mathcal{L}} \frac{d\mathcal{L}}{dL} = \overline{\mathcal{L}} \cdot 1 = 1, \\
\overline{y} &= \overline{L} \frac{dL}{dy} = y - t, \\
\overline{z} &= \overline{y} \frac{dy}{dz} = (y-t) \cdot \sigma'(z), \\
\overline{\omega} &= ((y-t) \cdot \sigma'(z)) \frac{dz}{d\omega} + \overline{R} \frac{dR}{d\omega} = ((y-t) \cdot \sigma'(z))x + \lambda\omega, \\
\overline{b} &= \overline{z} \frac{dz}{db} = (y-t) \cdot \sigma'(z), \\
\overline{x} &= \overline{z} \frac{dz}{dx} = ((y-t) \cdot \sigma'(z))\omega.
\end{aligned}
$$

5.5.3. Обсуждение backpropagation

- Для подсчёта backpropagation необходимо хранить ВСЕ промежуточные значения u_i на всех итерациях алгоритма. Это может быть существенным требованием к памяти, например, в больших нейронных сетях.

- Заметим, что вовсе необязательно для шага 2 полностью считать производную $\frac{\partial u_j}{\partial u_i}$, важно уметь быстро превращать градиент по выходу в градиент по входу (умножать якобиан на строку, см. главу 5.5.5). Например, можно вспомнить Пример 5.19 с покоординатными функциями, где результат считается просто покоординатным умножением.

- Причина, по которой на практике backpropagation работает так быстро, заключается в том, что вычисления с якобианами преобразований уже эффективно разработаны в рамках библиотек автоматического дифференцирования. Обычно мы даже не создаем и не сохраняем полный якобиан, выполняя matvec напрямую. См. главу 5.5.5.

- Отдельно стоит отметить нейронные сети, которые как раз и состоят из таких блоков. Отдельному блоку совершенно не надо знать, что происходит вокруг. То есть блок действительно может быть запрограммирован как отдельная сущность, умеющая внутри себя делать forward pass и backward pass, после чего блоки механически, как кубики в конструкторе, собираются в большую сеть, которая сможет работать как одно целое.

- Вычисление производной можно представить через ещё один граф вычислений, тем самым, сделав уже backward pass по графу производной, можно считать гессианы и производные высших порядков.

5.5.4. Forward propagation

Может возникнуть желание посчитать формулу (24) не слева направо, а справа налево, распространяя производную в графе в направлении forward pass от родителей к детям. Так производная по параметрам x будет считаться как

$$\frac{\partial u_i}{\partial x} = \sum_{u_j \in \text{родители } (u_i)c} \frac{\partial u_i}{\partial u_j}\frac{\partial u_j}{\partial x}.$$

При этом можно совершать проход вместе с подсчётом u_i и нужно будет хранить только один слой графа.

> **Пример 5.33.** Посчитаем шаг forward propagation в графе вычислений 3 из примера 5.32, где параметры x, ω - векторы, b - скаляр, а λ, t – заранее фиксированные константы.

Forward propogation

$$\frac{dz}{dx} = \omega,\ \frac{dy}{dx} = \sigma'(z)\omega,\ \frac{dL}{dx} = (y-t)\sigma'(z)\omega,$$

$$\overline{x} = (y-t)\sigma'(z)\omega.$$

$$\frac{dR}{d\omega} = \omega,\ \frac{dz}{d\omega} = x,\ \frac{dy}{d\omega} = \sigma'(z)x,\ \frac{dL}{d\omega} = (y-t)\sigma'(z)x,$$

$$\overline{\omega} = (y-t)\sigma'(z)x + \lambda\omega.$$

$$\frac{dz}{db} = 1,\ \frac{dy}{db} = \sigma'(z),\ \frac{dL}{db} = (y-t)\sigma'(z),$$

$$\overline{b} = (y-t)\sigma'(z).$$

Однако у этого алгоритма есть один нюанс: в forward pass нужно хранить производные $\frac{\partial u_i}{\partial x}$, а в backward pass $\frac{\partial f}{\partial u_i}$. Если размерность входа x намного больше, чем размерность выхода $f(x)$, то на каждом шаге forward pass нужно будет хранить и обсчитывать настолько же больше данных, чем в backward pass. Это характерно и для нейронных сетей, и для скалярнозначимых функций от тензорного входа в обычной оптимизации. При этом проблема с хранением всех значений u_i в backpropagation просто нивелируется. Если, наоборот, размерность выхода функции $f(x)$ намного больше, чем размерность входа x, то выгоднее использовать forward pass. В Примере 5.32 мы вплоть до вершины z хранили только скалярные производные, а градиенты появились только в конце.

Аналогично backward pass подсчёт производной в forward pass можно представить в виде графа вычислений и получать производные высших порядков.

5.5.5. Умножение гессиана на вектор и якобиана на строку

Пусть дана дважды непрерывно дифференцируемая функция $f : \mathbb{R}^n \to \mathbb{R}$ с симметричным гессианом $\nabla^2 f(x)$. Покажем, как можно эффективно считать

$$\nabla^2 f(x) \cdot u$$

для любого вектора $u \in \mathbb{R}^n$.

Считать полный гессиан и умножать его на вектор неэффективно особенно при большой размерности n. Но заметим, что

$$d(\langle \nabla f(x), u \rangle) = d(\nabla f)^\top u = dx^\top \nabla^2 f(x) \cdot u = dx^\top \nabla g(x),$$

где $g(x) = \langle \nabla f(x), u \rangle$. Таким образом вместо полного гессиана можно найти градиент функции вида g, с которой могут справиться библиотеки автоматического дифференцирования jax/autograd/pytorch/tensorflow (для градиента функции тоже можно построить граф вычислений). Особенно отметим jax, который эффективен для функций вида $g : \mathbb{R}^n \to \mathbb{R}$.

Аналогично для функции вида $f : \mathbb{R}^n \to \mathbb{R}^m$ и вектора $u \in \mathbb{R}^m$ считается значение $v^\top J_f(x)$ в точке x, а именно

$$d\langle u, f(x) \rangle = \langle u, J_f dx \rangle = \langle J_f^\top u, dx \rangle = \langle \nabla g(x), dx \rangle,$$

где $g(x) = \langle f(x), u \rangle$.

Именно поэтому вычисления из backpropagation можно эффективно реализовать на практике.