# Машинное обучение
# Семинар 3
# Матрично-векторное дифференцирование

Как правило, дифференцируемые модели обучаются с помощью градиентного спуска (или его модификаций), для чего важно уметь считать градиент функционала ошибки (loss-функции) по параметрам модели. Можно считать градиент покоординатно, а потом пристально смотреть на формулы и пытаться понять, как это может выглядеть в векторной форме. Однако на практике это очень неудобно и гораздо проще считать градиент напрямую — а для этого поможет знание градиентов для основных функций и основных правил матрично-векторного дифференцирования.

## 1 Дифференциал

### §1.1 Общее определение дифференциала

Прежде чем перейти к матрично/векторному дифференцированию, полезно вспомнить, что такое **дифференциал функции** в самом общем виде. Мы начнём с привычного случая одной переменной и постепенно обобщим понятие на произвольные нормированные линейные пространства (в которых можно мерить *расстояния*).

**Случай одной переменной.** Пусть $f : \mathbb{R} \to \mathbb{R}$. Говорят, что функция $f$ **дифференцируема** в точке $x$, если существует такое число $A$, что при $h \to 0$ выполняется разложение

$$f(x+h) = f(x) + A \cdot h + o(\|h\|). \tag{1.1}$$

В этом случае:

- число $A$ называется **производной** функции $f$ в точке $x$ и обозначается $f'(x)$;
- линейное отображение

  $$\mathrm{d}f_x(h) := f'(x) \cdot h$$

  называется **дифференциалом** функции $f$ в точке $x$.

*Интуиция.* Дифференциал — это не число, а *линейная функция*, которая описывает «основную, линейную часть» изменения функции. Переписав (1.1) в виде

$$f(x+h) - f(x) = \mathrm{d}f_x(h) + o(h),$$

мы видим, что $\mathrm{d}f_x(h)$ — это **наилучшая линейная аппроксимация** приращения функции $f(x+h) - f(x)$, а остаток $o(h)$ убывает асимптотически быстрее, чем $h$.

**Обобщение на линейные пространства.** Теперь перейдём от чисел к более общим объектам, а именно **нормированным** линейным пространствам (в которых можно мерить расстояния между точками).

**Определение нормы.** Пусть $V$ — линейное пространство. *Нормой* называется такая *неотрицательная функция* $\|\cdot\| : V \to [0, +\infty)$, удовлетворяющая следующим трем свойствам:

1. $\|x\| = 0 \Longleftrightarrow x = 0$
2. $\|\alpha x\| = |\alpha| \cdot \|x\|, \quad \forall\, \alpha \in \mathbb{R}, x \in V$
3. $\|x + y\| \leqslant \|x\| + \|y\| \quad \forall\, x, y \in V$ **(неравенство треугольника)**

Классическим примером нормы служит обычная евклидова норма (или $\ell_2$-норма) в $\mathbb{R}^d$ : $\|x\|_2 = \sqrt{x_1^2 + \ldots + x_d^2}$. Линейное пространство с заданной нормой называют **нормированным**. Легко видеть, что в нормированном пространстве можно определить расстояние между точками следующим образом: $\mathrm{dist}(x, y) = \|x - y\|$, таким образом нормированные пространства можно рассматривать как **метрические**.

**Дифференцируемость в общем виде.** Пусть $f : V \to U$ — функция из одного нормированного линейного пространства $V$ в другое $U$. Функция $f$ называется **дифференцируемой** в точке $x \in V$, если существует такое **линейное отображение** $L_x : V \to U$ такое, что при $h \to 0$:

$$f(x + h) = f(x) + L_x(h) + o(\|h\|).$$

Линейное отображение $L_x$ называется **дифференциалом** функции $f$ в точке $x$ и часто обозначается

$$\mathrm{d}f_x(h) := L_x(h).$$

**Обозначения.** В литературе встречаются разные формы записи дифференциала в точке $x$:

$$\mathrm{d}f_x(h), \quad \mathrm{d}f(x)[h], \quad \mathrm{d}f(h)$$

(в последнем случае $x$ подразумевается из контекста). Иногда пишут просто $\mathrm{d}f$, подразумевая, что определенное выражение выполняется не зависимо от того, какие значения принимают $x$ и $h$.

**Демистификация $\mathrm{d}x$** При обсуждении дифференциалов и дифференциирования очень часто возникает путаница с тем, что же означает $\mathrm{d}x$. Иногда говорят, что $\mathrm{d}x$ — бесконечно малое приращение аргумента (что бы это ни значило), а иногда, что это просто синоним $h$ из разложения $f(x + h) = f(x) + \mathrm{d}f_x(h) + o(\|h\|)$. Порой даже можно услышать что $\mathrm{d}f$ и $\mathrm{d}x$ – это просто мнемонические обозначения, не имеющие собственной сущности и необходимые просто в качестве удобной нотации.

На самом деле у $\mathrm{d}x$ есть и нормальное определение, а именно: $\mathrm{d}x$ — линейный оператор, который определяется в каждой точке $x \in V$ и для каждого аргумента $h \in V$ следующим образом: $\mathrm{d}x_x(h) = h$. Таким образом, писать, что $\mathrm{d}f_x(h) = f'(x)h$

это буквально то же самое, что и писать $\mathrm{d}f = f'\mathrm{d}x$. Разумеется, что т.к. этот оператор **не зависит** от точки $x$, то нет никакого смысла его с собой таскать, поэтому часто можно обойтись обозначением $\mathrm{d}x(h)$ или же просто $\mathrm{d}x$.

**Вопрос:** А что же в общем случае является аналогом производной?

## §1.2 Примеры дифференциалов и производных

Для ясности, посмотрим, как для функций различных размерностей и различного количества переменных устроены дифференциал, производная и разложение в точке.

1. **Одномерный случай:** $f : \mathbb{R} \to \mathbb{R}$

   В одномерном случае разложение должно иметь вид:

   $$f(x+h) = f(x) + Ah + o(h)$$

   Из курса анализа мы знаем, что:

   $$f(x+h) = f(x) + f'(x)h + o(h)$$

   Таким образом:

   - Разложение: $f(x+h) = f(x) + f'(x)h + o(h)$
   - Дифференциал: $\mathrm{d}f_x(h) = f'(x)h$ или $\mathrm{d}f(x) = f'(x)\mathrm{d}x$
   - Производная: $f'(x)$

2. **Многомерная функция** $f : \mathbb{R}^n \to \mathbb{R}$

   Так как линейные функции в $\mathbb{R}^n$ имеют вид $g(x) = \sum\limits_{j=1}^{n} a_j x_j$, то в многомерном случае разложение $f(x+h)$ должно иметь вид:

   $$\begin{aligned} f(x+h) &= f(x) + \sum_{j=1}^{n} a_j h_j + o(\|h\|_2) = \\ &= f(x) + \langle a, h \rangle + o(\|h\|_2) \end{aligned}$$

   Из курса анализа, мы знаем, что вектор $a$ — не что иное, как градиент функции $f$ в точке $x$, обозначаемый как $\nabla f(x) := (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^\top$

   - Разложение: $f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|)$
   - Дифференциал: $\mathrm{d}f_x(h) = \langle \nabla f(x), h \rangle$ или $\mathrm{d}f(x) = \langle \nabla f(x), \mathrm{d}x \rangle$
   - Градиент: $\nabla f(x) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^\top$

3. **Матричнозначная функция** $f : \mathbb{R}^{m \times n} \to \mathbb{R}$

   Так как матрица — это по сути вектор чисел, который запихнули в табличку, то все линейные функции в $\mathbb{R}^{m \times n}$ имеют вид:

   $$g(X) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} X_{i,j} = \langle A, X \rangle_F$$

для некоторой матрицы $A \in \mathbb{R}^{m\times n}$. Таким образом разложение $f(X+H)$ должно иметь вид:

$$f(X+H) = f(X) + \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}H_{i,j} + o(\|H\|_F) =$$
$$= f(X) + \langle A, H\rangle_F + o(\|H\|_F)$$

Здесь будет не лишним напомнить, что $\langle X,Y\rangle_F := \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} X_{i,j}Y_{i,j} = \mathrm{Tr}(X^\top Y)$ называется Фробениусовым скалярным произведением матриц $X$ и $Y$, а $\|X\|_F = \sqrt{\langle X,X\rangle_F}$ называется Фробениусовой нормой матрицы $X$. По аналогии с $\mathbb{R}^n$ несложно показать, что матрица $A$ — это не что иное, как *градиент* (да, в матричном случае $\nabla f(X)$ также называется градиентом):

$$A = \nabla f(X) = \left(\frac{\partial f}{\partial X_{i,j}}\right)_{i,j=1}^{m,n}$$

В данном случае:

- Разложение: $f(X+H) = f(X) + \langle \nabla f(X), H\rangle_F + o(\|H\|_F)$
- Дифференциал: $\mathrm{d}f_X(h) = \langle \nabla f(X), H\rangle_F$ или $\mathrm{d}f(X) = \langle \nabla f(X), \mathrm{d}X\rangle_F$
- Градиент: $\langle \nabla f(X), H\rangle_F$

4. **Еще одна многомерная функция** $f : \mathbb{R}^n \to \mathbb{R}^m$

   Линейные функции из $\mathbb{R}^n$ в $\mathbb{R}^m$ имеют вид $g(x) = Ax$, где $A \in \mathbb{R}^{m\times n}$. Таким образом разложение $f(x+h)$ должно иметь вид:

   $$f(x+h) = f(x) + Ah + o(\|h\|_2)$$

   для некоторой матрицы $A \in \mathbb{R}^{m\times n}$.

   Можно показать, что эта матрица $A$ равняется:

   $$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

   и называется **матрицей Якоби**. Обычно матрицу Якоби обозначают $J(x)$ или $J_x$. Таким образом:

   - Разложение: $f(x+h) = f(x) + J_x h + o(\|h\|_2)$
   - Дифференциал: $\mathrm{d}f_x(h) = J_x h$ или $\mathrm{d}f(x) = J_x \mathrm{d}x$
   - Матрица Якоби:

     $$J_x = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

5. **Сверхобщий случай:** $f : \mathbb{R}^{d_1 \times \ldots \times d_k} \to \mathbb{R}^{D_1 \times \ldots \times D_m}$

   Здесь можно поступить аналогичным образом через разложение $f(X + H)$ в виде:

   $$f(X + H) = f(X) + L(H) + o(\|H\|)$$

   где $L(H)$ — линейная функция вида $L(H) = \sum_{i_1,\ldots,i_k} H_{i_1,\ldots,i_k} \cdot A_{i_1,\ldots,i_k}$.

   В качестве аналогов градиента будут возникать уже структуры высокой размерности (тензоры), с которыми работать уже не очень удобно. Поэтому мы будем ограничиваться предыдущими пунктами ($\mathbb{R}^{m \times n} \to \mathbb{R}$, $\mathbb{R}^n \to \mathbb{R}^m$), чего нам более чем хватит для задач машинного обучения.

# 2 Как дифференцировать?

В предыдущем разделе мы вспомнили, что такое дифференциал функции в общем виде. Хотя формальное определение важно, на практике оно не всегда помогает быстро вычислять дифференциалы и градиенты. В этом разделе мы познакомимся с техникой матрично-векторного дифференцирования, которая позволяет удобно находить дифференциалы и производные функций с векторными и матричными аргументами.

## §2.1 Основные правила дифференцирования

По аналогии с одномерным случаем можно вывести несколько базовых правил дифференцирования, которые удобно использовать для вычисления дифференциалов, градиентов и других производных. Ниже приведены некоторые из этих правил:

1. **Константа:** Если $f(x) = \mathrm{const}$, то её дифференциал равен нулю:

   $$\mathrm{d}f = 0$$

2. **Линейность оператора дифференцирования:** Дифференциал линейной комбинации функций равен линейной комбинации их дифференциалов:

   $$\mathrm{d}(\alpha f + \beta g) = \alpha \cdot \mathrm{d}f + \beta \cdot \mathrm{d}g$$

   где $\alpha$ и $\beta$ — константы.

3. **Дифференциал линейной функции:** Если функция $f$ является линейной, то дифференциал $f(x)$ совпадает с $f(\mathrm{d}x)$:

   $$\mathrm{d}f(x) = f(\mathrm{d}x)$$

   **Пример 2.1.** Пусть $f(X) = AXB$, где $A, B, X \in \mathbb{R}^{n \times n}$. Тогда:

   $$\mathrm{d}(AXB) = A\,\mathrm{d}X\,B$$

4. **Правило произведения:** Дифференциал произведения двух функций равен:

$$d(f \cdot g) = f\,dg + g\,df$$

Аналогичное правило работает и для скалярных произведений:

$$d\langle f, g\rangle = \langle df, g\rangle + \langle f, dg\rangle$$

5. **Цепное правило:** Пусть $z(x) = f(g(x))$ — сложная функция, то её дифференциал можно выразить как:

$$dz(x)[h] = df(g(x))[dg(x)[h]],$$

На первый взгляд это правило может выглядеть жутковато, однако применять его на практике довольно просто:

- Вычисляем $df(x)$ – какое-то выражение, содержащее $x$ и $dx$
- Заменяем $x \to g$, $dx \to dg$

Применение этого правила для матрично-векторного дифференцирования мы увидим ниже, а пока я приведу пример применения этого правила для одномерных функций.

**Пример 2.2.** Пусть $z(x) = \log \sin x$. Найти $dz(x)$

Мы имеем дело со сложной функцией: $z(x) = f(g(x))$, где $f(x) = \log x$, а $g(x) = \sin x$. Будем действовать по цепному правилу:

- Найдем дифференциалы $df$ и $dg$:
  - $df(x) = \frac{dx}{x}$
  - $dg(x) = \cos x\,dx$
- Возьмем $df$ и заменим $x \to g$, $dx \to dg$:

$$d(f \circ g)(x) = \frac{dg}{g} = \frac{\cos x\,dx}{\sin x} = \operatorname{ctg}(x)\,dx$$

**Задача 2.1.** *Пусть $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = x^\top Ax$. Найти градиент и дифференциал функции $f(x)$.*

Для начала, найдем дифференциал функции $f(x) = x^\top Ax$:

$$\begin{aligned} d(x^\top Ax) &= d(x^\top \cdot Ax) = \text{ используем правило произведения} \\ &= d(x^\top) \cdot Ax + x^\top \cdot d(Ax) = \text{ используем линейность } Ax \text{ и } x^\top \\ &= (dx)^\top Ax + x^\top A dx = \text{ перезаписываем через скалярное произведение} \\ &= \langle (A + A^\top)x, dx\rangle \end{aligned}$$

Таким образом, дифференциал функции $f(x) = x^\top Ax$ равен:

$$df = \langle (A + A^\top)x, dx\rangle$$

А градиент $\nabla f(x)$ — это просто левая часть скалярного произведения (так как дифференциал $\mathrm{d}f = \langle \nabla f(x), \mathrm{d}x \rangle$):

$$\nabla f(x) = (A + A^\top)x$$

Отметим, что если $A = A^\top$, то $\nabla f(x) = 2Ax$, что очень похоже на одномерный случай.

**Следствие 1.** Положив $A := I$, получаем $\nabla \|x\|^2 = 2x$, а $\mathrm{d}(\|x\|^2) = 2x\,\mathrm{d}x$

**Задача 2.2.** *Пусть $f(X) = X^{-1}$, $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$. Найти дифференциал и градиент функции $f$.*

Рассмотрим тождественную функцию $g(X) = X$. Тогда:

$$(f \cdot g)(X) = I.$$

Поскольку $f \cdot g$ — константа, то её дифференциал равен нулю:

$$\mathrm{d}(f \cdot g) = 0.$$

С другой стороны, по правилу дифференцирования произведения:

$$\mathrm{d}(f \cdot g) = \mathrm{d}f \cdot g + f \cdot \mathrm{d}g.$$

Из этого получаем:

$$\mathrm{d}f \cdot g + f \cdot \mathrm{d}g = 0,$$

или

$$\mathrm{d}(X^{-1}) \cdot X = -X^{-1} \cdot \mathrm{d}(X).$$

Так как $\mathrm{d}g(X) = \mathrm{d}X$, то:

$$\mathrm{d}(X^{-1}) = -X^{-1} \cdot \mathrm{d}X \cdot X^{-1}.$$

**Задача 2.3.** *Пусть $z(x) = f(g(x))$. Доказать цепное правило для $z(x)$, а именно, необходимо показать, что:*

$$\mathrm{d}z(x)[h] = \mathrm{d}f(g(x))[\mathrm{d}g(x)[h]]$$

Будем действовать по определению через разложение:

$$z(x+h) = z(x) + \mathrm{d}z(x)[h] + o(\|h\|)$$

Перепишем через $z(x) = f(g(x))$:

$$z(x+h) = f(g(x+h)) = f(g(x) + \mathrm{d}g(x)[h] + o(h))$$

Разложим $f(g(x) + \mathrm{d}g(x)[h] + o(h))$ в точке $g(x)$:

$$z(x+h) = f(g(x)) + \mathrm{d}f\Big(g(x)\Big)\Big[\mathrm{d}g(x)[h] + o(\|h\|)\Big] + o(\|h\|)$$

Так как $\mathrm{d}f(g(x))$ — линейная функция, то $o(\|h\|)$ можно вытащить из квадратных скобок:

$$z(x+h) = z + \underbrace{\mathrm{d}f(g(x))[\mathrm{d}g(x)[h]]}_{\mathrm{d}z(x)[h]} + o(\|h\|)$$

## §2.2 Таблица стандартных дифференциалов

Здесь приведена таблица со списком некоторых *стандартных* дифференциалов, достаточно часто используемых на практике. Некоторые из них мы уже вывели в предыдущем параграфе.

| **Функция** | **Дифференциал** |
|---|---|
| $\langle A, X\rangle = \mathrm{Tr}(A^\top X)$ | $\mathrm{d}\langle A, X\rangle = \langle A, \mathrm{d}X\rangle$ |
| $\mathrm{Tr}(X)$ | $\mathrm{d}\,\mathrm{Tr}(X) = \mathrm{Tr}(\mathrm{d}X)$ |
| $X^\top$ | $\mathrm{d}(X^\top) = (\mathrm{d}X)^\top$ |
| $\det(X)$ | $\mathrm{d}\,\det(X) = \det(X)\,\langle X^{-1}, \mathrm{d}X\rangle$ |
| $x^\top A x$ | $\mathrm{d}(x^\top A x) = \langle (A + A^\top)x, \mathrm{d}x\rangle$ |
| $\\|X\\|_F^2 = \mathrm{Tr}(X^\top X)$ | $\mathrm{d}\,\\|X\\|_F^2 = 2\,\langle X, \mathrm{d}X\rangle$ |
| $X^{-1}$ | $\mathrm{d}(X^{-1}) = -X^{-1}(\mathrm{d}X)X^{-1}$ |

Таблица 1. Стандартные дифференциалы матрично-векторных выражений

**Задача 2.4.** *Найти дифференциал и градиент функции* $\log det(X)$ *для симметричной положительно определенной матрицы* $X$.

Здесь мы имеем дело с композицией двух функций: $f(x) = \log x$ и $g(x) = \det(X)$. Вспомним цепное правило: чтобы найти дифференциал $f \circ g$, нужно:

- Найти $\mathrm{d}f$
- Заменить $X \to g(X)$, $\mathrm{d}X \to \mathrm{d}g(X)$

Поехали:

$$\mathrm{d}(f \circ g)(X) = \frac{\mathrm{d}g}{g} = \frac{\det(X)\langle X^{-1}, \mathrm{d}X\rangle}{\det(X)} = \langle X^{-1}, \mathrm{d}X\rangle$$

Таким образом, получаем:

- **Дифференциал**: $\mathrm{d}\log\det(X) = \langle X^{-1}, \mathrm{d}X\rangle$
- **Градиент**: $\nabla \log\det(X) = X^{-1}$

**Задача 2.5.** *(**VIP пример**) Рассмотрим крайне важный пример матрично-векторной функции — функция потерь MSE для задачи линейной регрессии:*

$$L(w) = \frac{1}{N}\|Xw - y\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}(\langle x_i, w\rangle - y)^2$$

*где* $X \in \mathbb{R}^{N\times d}$ *— матрица «объект-признак»,* $y$ *— столбец целевой переменной, а* $w \in \mathbb{R}^d$ *— параметры модели (линейной регрессии).*

*Допустим мы хотим обучать линейную регрессию с MSE с помощью градиентного спуска. Необходимо найти градиент функции потерь по отношению к параметрам модели* $w$.

Найдем дифференциал функции $\mathrm{d}L(w)$:

$$\mathrm{d}L(w) = \mathrm{d}\left(\frac{1}{N}\|Xw - y\|^2\right) = \frac{1}{N}\mathrm{d}\langle Xw - y, Xw - y\rangle = \text{правило произведения}$$
$$= \frac{2}{N}\langle Xw - y,\, \mathrm{d}(Xw - y)\rangle$$

Так как $-y$ константа, а $Xw$ линейно, то $\mathrm{d}(Xw - y) = X\mathrm{d}w$:

$$\mathrm{d}L(w) = \frac{2}{N}\langle Xw - y, X\mathrm{d}w\rangle = \frac{2}{N}\langle X^\top(Xw - y), \mathrm{d}w\rangle$$

Таким образом, градиент $\nabla_w L(w)$ равен:

$$\nabla_w L(w) = \frac{2}{N}X^\top(Xw - y)$$

**Задача 2.6.** *$f(x) = \|x\|^{\frac{3}{2}}$. Найти градиент и дифференциал функции $f$.*

Найдем дифференциал:

$$\mathrm{d}\|x\|^{\frac{3}{2}} = \mathrm{d}(\|x\|^2)^{\frac{3}{4}} = \frac{3}{4}(\|x\|^2)^{-\frac{1}{4}}\,\mathrm{d}(\|x\|^2) = \frac{3}{4\|x\|^{\frac{1}{2}}}\langle 2x,\, \mathrm{d}x\rangle = \left\langle \frac{3x}{2\|x\|^{\frac{1}{2}}},\, \mathrm{d}x\right\rangle$$

Таким образом, дифференциал равен $\mathrm{d}f(x) = \left\langle \frac{3x}{2\|x\|^{\frac{1}{2}}},\, \mathrm{d}x\right\rangle$, а градиент $\nabla f(x) = \frac{3x}{2\|x\|^{\frac{1}{2}}}$