Формулы дифференцирования

Максим Каледин

16 ноября 2025 г.

В задачах ниже мы используем numerator layout, если пишем в знаменателе ∂X^T и denominator layout, если ∂X. В первом матрица Якоби (градиент) ∇f для функции $f : \mathbb{R}^n \to R$ – это вектор-строка (матрица $1n$), а матрица Якоби для $f : \mathbb{R}^n \to R^m$ – это

$$\nabla f = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j}.$$

В такой записи мы всегда должны помнить, какие значения проходят индексы i, j.

В первой нотации матрица Якоби вектор функции имеет количество столбцов, равное размерности X, во втором – размерности Y. Можно перейти из одной нотации в другую перестановкой размерностей. Для случая, когда мы имеем функцию f с матрицей на входе и скаляром на выходе – самый частый для нас случай – размерность производной в numerator (стараемся использовать, но не навязываем) – это размерность X^T, а в denominator – размерность X.

В общем случае порядок размерностей фиксируется заранее как считается удобнее. В суммировании Эйнштейна фиксируются сами размерности (они соответствуют свободным индексам), но не фиксируется их порядок. Оно позволяет легче получать формулы, но если в конце мы хотим записать формулу в виде матричных операций, мы должны определить для себя порядок размерностей, чтобы выражение записалось корректно.

1 Полезные хаки

След квадратной матрицы (или тензора с одинаковыми размерностями по каждому измерению)

$$\mathrm{Tr}A = \sum_i A_{i,i,i,\ldots,i}.$$

След можно много где увидеть, для нас центральная вещь – это

$$x^T y = \mathrm{Tr}(yx^T)$$

для векторов $x, y \in \mathbb{R}^d$. У него есть свойства, которыми можно пользоваться (это самые простые):

1. $\mathrm{Tr}(A) = \mathrm{Tr}(A^T)$.
2. $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, если матрицы перемножаются корректно.

В частности, евклидова 2-норма

$$\|x\|_2^2 = \sum_i x_i^2 = \mathrm{Tr}(xx^T).$$

Для определителя есть

1. Разложение по строке/столбцу;
2. Формула обратной матрицы;
3. ... что ещё базового можно вспомнить.

Ещё мы используем суммирование Эйнштейна: по повторяющимся индексам в произведении тензоров считается сумма по всем значениям, е сли не повторяется – свободная размерность.

Методология дифференцирования следующая.

1. Определяем размерности производной;
2. Записываем через Эйнштейна или используя другие тождества и считаем результат.

Погнали к задачам.

2 Производная следа, уровень 0

$f(X) = \mathrm{Tr}(AX)$, предполагается, что размерности корректно задают умножение.

Шаг 1 (определяем размерность). В колоночной нотации (numerator, я рекомендую как чуть более согласованную с общематематическим смыслом, но просто всегда сами пишите) размерность производной такой функции равна размерности X^T. Если $X = (X_{ts})_{ts}$, то производная – это тензор $D = (D_{st})_{st}$. Индекс t пробегает строки X, а s – столбцы.

Шаг 2 (вычисляем). Вообще след – идеальный случай для суммирования по Эйнштейну. Вот почему:

$$\mathrm{Tr}(AX) = A_{ij}X_{ji}.$$

Поэтому

$$\frac{\partial f}{\partial X^T} = (..)_{st} = \frac{\partial \left(A_{ij}X_{ji}\right)}{\partial X_{ts}}.$$

В числителе одна большая сумма, только для $j = t, i = s$ слагаемое даст ненулевую производную. Итого:

$$= A_{ts},$$

В матричной нотации

$$\frac{\partial f}{\partial X^T} = A.$$

3 Производная следа, уровень 0+

$f(X) = \mathrm{Tr}(AXB)$, предполагается, что размерности корректно задают умножение (в том числе, если переставлять).

Поскольку $\mathrm{Tr}(AXB) = \mathrm{Tr}(BAX)$, если с размерностями ок, то

$$\frac{\partial f}{\partial X^T} = (..)_{st} = (BA)_{ts},$$

в матричном виде

$$\frac{\partial f}{\partial X^T} = BA.$$

4 Производная следа, уровень 1

$f(X) = \mathrm{Tr}(XAX^T)$, предполагается, что размерности X, A корректно задают умножение.

Шаг 1 (определяем размерность). В колоночной нотации размерность производной такой функции равна размерности X^T. Если $X = (X_{ts})_{ts}$, то производная – это тензор $D = (D_{st})_{st}$.

Шаг 2 (вычисляем). Пользуемся суммированием Эйнштейна для записи матричного произведения:

$$XAX^T = (..)_{pz} = X_{pj}A_{jk}X_{zk}$$

Далее применяем след, суммируя по $p = z$. В итоге

$$\frac{\partial f}{\partial X^T} = (..)_{st} = \frac{\partial \left(X_{pj}A_{jk}X_{pk}\right)}{\partial X_{ts}}.$$

Тут посложнее. Но мы знаем, что слагаемое даст 0 после дифференцирования, если $p \neq t$, иначе будет не 0, всегда найдётся подходящий $j = s$.

Рассмотрим случай $p = t$, тут имеем слагаемое $X_{tj}A_{jk}X_{tk}$, они дадут в случае $j = s, k \neq s$ сумму $\sum_{k\neq s} A_{sk}X_{tk}$. Если $j \neq s, k = s$, то после дифференцирования будет $\sum_{j\neq s} X_{tj}A_{js}$. Если же $j = s, k = s$, получим после дифференцирования $2X_{ts}A_{ss}$. Сложим все три результата, унифицируем индексы в разных суммах и получим (здесь суммирование по Эйнштейну в каждом из двух слагаемых)

$$A_{sj}X_{tj} + X_{tj}A_{js}.$$

Итого:

$$\frac{\partial f}{\partial X^T} = (..)_{st} = A_{sj}X_{tj} + X_{tj}A_{js} = (A_{sj} + A_{js})X_{tj},$$

Вспоминаем, какие значения пробегает индекс j, и определяем как правильно выписать в матричной нотации:

$$\frac{\partial f}{\partial X^T} = (A^T + A)X^T.$$

Интересный вопрос, что будет, если $f(X) = \mathrm{Tr}(X^TAX)$, Получится почти то же самое

$$\frac{\partial f}{\partial X^T} = X^T(A^T + A),$$

так что тут может не так очевидно транспонирования перекидывать, нужно быть внимательнее.

5 Производная следа, уровень 3

$f(X) = \mathrm{Tr}(XAX^TQ) = \mathrm{Tr}(QXAX^T)$, предполагается, что размерности X, A корректно задают умножение.

Шаг 1 (определяем размерность). В колоночной нотации размерность производной такой функции равна размерности X^T. Если $X = (X_{ts})_{ts}$, то производная – это тензор $D = (D_{st})_{st}$.

Шаг 2 (вычисляем). Пользуемся суммированием Эйнштейна для записи матричного произведения:

$$\mathrm{Tr}(XAX^TQ) = X_{pj}A_{jk}X_{bk}Q_{bp}.$$

В этой задаче тоже нужно определить, где 0 и ненулевые слагаемые после дифференцирования, в результате всё красиво свернётся в

$$A_{sj}X_{bj}Q_{bt} + X_{bj}A_{js}Q_{tb}.$$

Совмещаем размерности с шагом 1, определяем умножения, и в матричной нотации получим:

$$\frac{\partial f}{\partial X^T} = AX^TQ + A^TX^TQ^T.$$

6 Производная определителя, уровень 0

Для наших целей больше не пригодится, но возможно $+\varepsilon$ по сложности будет на КР, и $+2\varepsilon$ – в ДЗ.

Рассмотрим $f(X) = \det(X)$. На лекции мы использовали разложение по строке. Если возьмём произвольную строку s, то

$$\det(X) = \sum_i (-1)^{s+i} X_{si} M_{s,i},$$

где $M_{t,i}$ – это минор. В минор не входит X_{si}, для нас это очень удобно, потому что

$$\frac{\partial f}{\partial X^T} = (..)_{st} = (-1)^{s+t} M_{s,t}.$$

Справа стоит алгебраическое дополнение. Если перейти по формуле обратной матрицы

$$A^{-1} = \frac{1}{\det(A)} A^*,$$

где A^* – присоединённая (транспонированная матрица из алгебраических дополнений), то в результате увидим

$$\frac{\partial f}{\partial X^T} = (..)_{st} = \det(A)(A^{-1})^T.$$

7 А ещё не забываем классику

Формула Лейбница и формула дифференцирования композиции функций тоже работают, если правильно применить (когда видите матрицу, подумайте, как применяется).

8 Символ Кронекера

Это отдельный полезный трюк, если к нему привыкнуть. Все эти задачи решаются с ним ещё проще и про это есть отдельная детальная заметка (которая тоже почти появилась на гитхабе).

Символ Кронекера – это такая условная матрица, конкретно функция двух индексов i, j

$$\delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Она симметрична $\delta_{ij} = \delta_{ji}$. Её удобно использовать при суммировании Эйнштейна, полагая, что она – матрица подходящего размера, чтобы умножения были корректны. Давайте посмотрим как выглядит умножение на дельту:

$$\delta A = (..)_{ik} = \delta_{ij} A_{jk} = \sum_j \delta_{ij} A_{jk}$$

Только для одного индекса $j = i$ слагаемое ненулевое, получаем...

$$= A_{ik}.$$

Попробуем с другой стороны:

$$A\delta = (..)_{ik} = A_{ij} \delta_{jk} = \sum_j A_{ij} \delta_{jk} = A_{ik}.$$

По этой причине A_{ik} можно представить как суммму Эйнштейна по подходящему индексу с дельтой. Более того, мы увидели, что за счёт произвольно выбираемого размера дельты, её можно переставлять в произведении как угодно с чем угодно. А ещё она работает по такому принципу: она позволяет заменить суммируемый индекс на свободный в произведении двух матриц. К примеру, в выражениях выше в сумме Эйнштейна

$$\delta_{ij} A_{jk} = A_{ik},$$

мы заменили суммируемый индекс j на свободный i.

Попробуем на примере следа $f(X) = \mathrm{Tr}(AX)$:

$$\frac{\partial f(X)}{\partial X^T} = (..)_{st},$$

где s пробегает столбцы X, а t – строки.

Воспользуемся линейностью производной:

$$\frac{\partial f(X)}{\partial X^T} = (..)_{st} = \frac{\partial f(X)}{\partial X_{ts}} = A_{ij}\frac{\partial X_{ji}}{\partial X_{ts}}.$$

Чтобы был не 0 справа, должно быть $j = t, i = s$, это записывается как

$$\frac{\partial X_{ji}}{\partial X_{ts}} = \delta_{jt}\delta_{is}.$$

Итого:

$$\frac{\partial X_{ji}}{\partial X_{ts}} = A_{ij}\delta_{jt}\delta_{is} = A_{st}.$$

Согласуем размерности:

$$\frac{\partial f(X)}{\partial X^T} = (..)_{st} = A.$$

9 Conclusion

“I always thought something was fundamentally wrong with the universe”