Машинное обучение, ФКН ВШЭ
Матричное дифференцирование

Задача 1. Пусть $f(X) = \ln \det X$, где $X \in \mathbb{R}^{n \times n}$. Найдите производную $\nabla_X f(X)$.

1. Применим цепное правило:

$$\nabla_X f(X) = \nabla_X \ln(\det X) = \frac{1}{\det X} \cdot \nabla_X (\det X)$$

2. Используем известную формулу для градиента определителя:

$$\nabla_X \det X = \det X \cdot X^{-T}$$

Это следует из дифференциала:

$$d(\det X) = \text{tr}((\det X \cdot X^{-1})^T dX) \quad \Rightarrow \quad \nabla_X \det X = (\det X \cdot X^{-1})^T = \det X \cdot X^{-T}$$

3. Подставим в цепное правило:

$$\nabla_X f(X) = \frac{1}{\det X} \cdot (\det X \cdot X^{-T}) = X^{-T}$$

Задача 2. Пусть $f(x) = x^T \exp(xx^T) x$, где $x \in \mathbb{R}^n$, а $\exp(B)$, $B \in \mathbb{R}^{n \times n}$. Матричной экспонентой обозначают ряд

$$I_n + \frac{B}{1!} + \frac{B^2}{2!} + \frac{B^3}{3!} + \frac{B^4}{4!} + \ldots = \sum_{k=0}^{\infty} \frac{B^k}{k!}.$$

Найдите производную $\nabla_x f(x)$.

Обозначим $A = xx^T$, $f(x) = x^T \exp(A) x$. Будем использовать следующий факт:

1. $\nabla_x (x^T B x) = Bx + B^T x$ для постоянной матрицы B

Вычислим этот градиент:
Применим правило произведения для нахождения $\nabla_x f(x)$.

$$f(x) = x^T \exp(A) x$$

Если бы $\exp(A)$ была постоянной, мы бы имели:

$$\nabla_x f(x) = \exp(A)x + \exp(A)^T x = 2\exp(A)x$$

Но $\exp(A)$ зависит от x, поэтому мы должны продифференцировать и ее тоже.

Вычислим дополнительный член по цепному правилу. Нам нужен $\nabla_x(x^T\exp(A)x)$ с учетом изменений в $\exp(A)$ из-за изменений в x.

Для этого воспользуемся формулой для направленных производных матричных функций: если $h(t) = x^T\exp(A+tB)x$, тогда $h'(0) = x^T\exp(A)B\exp(A)x$ В нашем случае нам нужно найти, как $\exp(A)$ меняется, когда x меняется в направлении v: если x меняется на $x+tv$, то $A = xx^T$ меняется на

$$(x+tv)(x+tv)^T = xx^T + t(xv^T + vx^T) + t^2vv^T$$

Изменение первого порядка для A равно $(xv^T + vx^T)$.

Для каждой компоненты v_i градиента вычисляем направленную производную в направлении e_i (единичный вектор). Это дает нам дополнительный член:

$$x^T\exp(A)(xe_i^T + e_ix^T)\exp(A)x$$

Объединяя все члены, полный градиент равен:

$$\nabla_x f(x) = 2\exp(A)x + 2\exp(A)xx^T\exp(A)x$$

Или:

$$\nabla_x f(x) = 2\exp(xx^T)x + 2\exp(xx^T)x\cdot x^T\exp(xx^T)x$$

Поскольку $x^T\exp(xx^T)x = f(x)$ - это скаляр, мы можем переписать это как:

$$\nabla_x f(x) = 2\exp(xx^T)x + 2f(x)\exp(xx^T)x = 2(1+f(x))\exp(xx^T)x$$

Итак, градиент равен:

$$\boxed{\nabla_x f(x) = 2(1+f(x))\exp(xx^T)x}$$

Задача 3. Пусть $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m, x\in\mathbb{R}^n$. Найдите производную $\nabla_x f(x)$ функции

$$f(x) = \sin\|Ax+b\|_2$$

Найдем градиент функции $f(x) = \sin\|Ax+b\|_2$, где $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m, x\in\mathbb{R}^n$. Для начала вспомним, что $\|y\|_2 = \sqrt{y^Ty}$. Обозначим $y = Ax+b$, тогда $f(x) = \sin\|y\|_2$. Применим цепное правило для нахождения градиента:

$$\nabla_x f(x) = \frac{df}{d\|y\|_2}\cdot\nabla_x\|y\|_2$$

Найдем каждую часть:

1. $\dfrac{df}{d\|y\|_2} = \cos\|y\|_2$

2. Для вычисления $\nabla_x\|y\|_2$ снова применим цепное правило:

$$\nabla_x\|y\|_2 = \nabla_y\|y\|_2\cdot\nabla_x y$$

   Известно, что $\nabla_y\|y\|_2 = \dfrac{y}{\|y\|_2}$, это градиент евклидовой нормы.

   Также $\nabla_x y = \nabla_x(Ax+b) = A^T$, поскольку b не зависит от x.

Таким образом,

$$\begin{aligned}\nabla_x \|y\|_2 &= \frac{y}{\|y\|_2} \cdot A^T \\ &= \frac{Ax+b}{\|Ax+b\|_2} \cdot A^T \\ &= A^T \frac{Ax+b}{\|Ax+b\|_2}\end{aligned}$$

3. Комбинируя полученные результаты:

$$\begin{aligned}\nabla_x f(x) &= \cos \|Ax+b\|_2 \cdot A^T \frac{Ax+b}{\|Ax+b\|_2} \\ &= A^T \frac{(Ax+b)\cos \|Ax+b\|_2}{\|Ax+b\|_2}\end{aligned}$$

Итак, градиент функции $f(x) = \sin \|Ax+b\|_2$ равен:

$$\boxed{\nabla_x f(x) = A^T \frac{(Ax+b)\cos \|Ax+b\|_2}{\|Ax+b\|_2}}$$

Задача 4. Рассмотрим симметричную матрицу $A \in \mathbb{R}^{n\times n}$ и её спектральное разложение $A = Q\Lambda Q^T$. Пусть $\lambda \in \mathbb{R}^n$ — это диагональ матрицы Λ (то есть вектор, составленный из собственных значений A). Найдите производные:

1. $\nabla_\lambda \operatorname{tr}(A)$

   Сначала вспомним, что след матрицы A определяется как сумма её диагональных элементов:

   $$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

   Также известно свойство следа: $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$.

   Используя спектральное разложение, имеем:

   $$\begin{aligned}\operatorname{tr}(A) &= \operatorname{tr}(Q\Lambda Q^T) \\ &= \operatorname{tr}(\Lambda Q^T Q) \\ &= \operatorname{tr}(\Lambda)\end{aligned}$$

   Поскольку Q — ортогональная матрица, то $Q^T Q = I$, и поэтому $\operatorname{tr}(\Lambda Q^T Q) = \operatorname{tr}(\Lambda)$.

   Теперь, $\operatorname{tr}(\Lambda) = \sum_{i=1}^n \lambda_i$, то есть сумма собственных значений.

   Поскольку $\operatorname{tr}(A)$ — это просто сумма элементов λ, то производная по λ будет:

   $$\boxed{\nabla_\lambda \operatorname{tr}(A) = \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \mathbf{1}_n}$$

   где $\mathbf{1}_n$ — вектор из единиц размера n.

2. $\nabla_Q \operatorname{tr}(A)$

   Снова используем спектральное разложение:

   $$\operatorname{tr}(A) = \operatorname{tr}(Q\Lambda Q^T)$$

   Чтобы найти $\nabla_Q \operatorname{tr}(A)$, нам нужно вычислить, как изменяется $\operatorname{tr}(A)$ при малом изменении Q.

   Поскольку Q — ортогональная матрица, то $Q^T Q = I$. При малых возмущениях Q мы должны сохранять это свойство.

   Вычислим:

   $$\begin{aligned}\operatorname{tr}(A) &= \operatorname{tr}(Q\Lambda Q^T) \\ &= \operatorname{tr}(\Lambda Q^T Q) \\ &= \operatorname{tr}(\Lambda)\end{aligned}$$

   Поскольку $\operatorname{tr}(\Lambda)$ не зависит от Q (только от собственных значений), то изменение Q не влияет на $\operatorname{tr}(A)$ при условии, что Q остаётся ортогональной.

   Таким образом:

   $$\boxed{\nabla_Q \operatorname{tr}(A) = O_{n\times n}}$$

   где $O_{n\times n}$ — нулевая матрица размера $n \times n$.

   Это объясняется тем, что след матрицы зависит только от её собственных значений, а не от базиса собственных векторов, представленного матрицей Q.

Задача 5. Рассмотрим задачу обучения линейной регрессии с функцией потерь Log-Cosh:

$$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} \ln(\cosh(w^T x_i - y_i))$$

Выпишите формулу для градиента $\nabla_w Q(w)$. Запишите ее в матричном виде, используя матрицу объекты-признаки X и вектор целевых переменных y. В матричном виде:

$$Q(w) = \frac{1}{\ell}\ln\left(\cosh(Xw - y)\right)$$

$$dQ = \frac{1}{\ell} X^T \tanh\left(Xw - y\right)$$

Задача 6. В случае многомерной Ridge-регрессии минимизируется функция со штрафом:

$$Q(w) = (y - Xw)^T(y - Xw) + \lambda w^T w,$$

где λ — положительный параметр, штрафующий функцию за слишком большие значения w.

1. Найдите производную $\nabla_w Q(w)$ и выведите формулу для оптимального w.

   Дифференцирование:

   $$dQ = d(y - Xw)^T(y - Xw) + (y - Xw)^T d(y - Xw) + d(\lambda w^T w)$$

   Поскольку $d(y - Xw) = -X\,dw$, получаем:

   $$dQ = -dw^T X^T(y - Xw) - (y - Xw)^T X\,dw + 2\lambda w^T dw$$

Заметим, что первые два слагаемых — скаляры, поэтому транспонируем одно из них и сложим:

$$dQ = -2(y - Xw)^T X\, dw + 2\lambda w^T dw$$

Тогда градиент:

$$\nabla_w Q(w) = -2X^T(y - Xw) + 2\lambda w$$

Найдём оптимальное w из условия $\nabla_w Q(w) = 0$:

$$\begin{aligned} -X^T y + X^T X w + \lambda w = 0 \\ (X^T X + \lambda I) w = X^T y \\ \boxed{w = (X^T X + \lambda I)^{-1} X^T y} \end{aligned}$$

2. Найдите вторую производную $\nabla_w^2 Q(w)$. Убедитесь, что найдена точка минимума.

$$\nabla_w^2 Q(w) = \frac{d}{dw}\left[-2X^T(y - Xw) + 2\lambda w\right] = 2X^T X + 2\lambda I$$

Так как $\lambda > 0$ и $X^T X$ — положительно полуопределённая матрица, то $A = 2X^T X + 2\lambda I$ — положительно определённая.

Проверим по определению: для любого $z \in \mathbb{R}^n$:

$$z^T 2X^T X + 2\lambda I z = 2z^T X^T X z + 2\lambda z^T I z$$

$$\begin{aligned} z^T I z = z^T z = ||z||^2 > 0 \\ z^T X^T X z = ||Xq||^2 > 0 \end{aligned}$$

Значит, точка — действительно минимум.

3. Выпишите шаг градиентного спуска в матричном виде.
Обозначим шаг обучения через η.

$$w_{t+1} = w_t - \eta \nabla_w Q(w_t) = w_t - 2\eta\left[-X^T(y - Xw_t) + \lambda w_t\right]$$

$$\boxed{w_{t+1} = w_t + 2\eta\left(X^T(y - Xw_t) - \lambda w_t\right)}$$

Задача 7. Найдите симметричную матрицу X, наиболее близкую к матрице A по норме Фробениуса ($\sum_{i,j}(x_{ij} - a_{ij})^2$). Иными словами, решите задачу условной матричной минимизации

$$\begin{cases} ||X - A||_F^2 \to \min_X \\ X^T = X \end{cases}$$

Hint: Надо будет выписать лагранжиан. А ещё пригодится тот факт, что $\sum_{i,j}(x_{ij} - a_{ij})^2 = ||X - A||_F^2 = \mathrm{tr}((X - A)^T (X - A))$.
Выписываем Лагранжиан:

$$\begin{aligned} \mathcal{L} &= \sum_{i,j}(x_{ij} - a_{ij})^2 + \sum_{ij} \lambda_{ij}(x_{ij} - x_{ji}) \\ &= \mathrm{tr}((X - A)^T (X - A)) + \mathrm{tr}(\Lambda^T (X - X^T)) \\ &= \mathrm{tr}(X^T X) - 2\,\mathrm{tr}(X^T A) + \mathrm{tr}(A^T A) + \mathrm{tr}(\Lambda^T (X - X^T)) \end{aligned}$$

Найдём все необходимые нам дифференциалы:

$$\begin{aligned}
d[\mathrm{tr}(X^T X)] &= \mathrm{tr}(d(X^T X)) = \mathrm{tr}(X^T dX) + \mathrm{tr}(dX^T X) = \mathrm{tr}(2X^T dX) \\
d[\mathrm{tr}(X^T A)] &= \mathrm{tr}(A^T dX) \\
d[\mathrm{tr}(\Lambda^T X)] &= \mathrm{tr}(\Lambda^T dX) \\
d[\mathrm{tr}(\Lambda^T X^T)] &= \mathrm{tr}(\Lambda dX)
\end{aligned}$$

Выписываем в явном виде производную по X:

$$\frac{\partial \mathcal{L}}{\partial X} = 2X^T - 2A^T + \Lambda^T - \Lambda = 0$$

Нужно избавиться от Λ, давайте транспонируем уравнение:

$$\frac{\partial \mathcal{L}}{\partial X} = 2X - 2A + \Lambda - \Lambda^T = 0$$

А после прибавим его к исходному, тогда лишние части исчезнут:

$$4X - 2A^T - 2A = 0 \quad \Rightarrow \quad X = \frac{1}{2}(A + A^T)$$