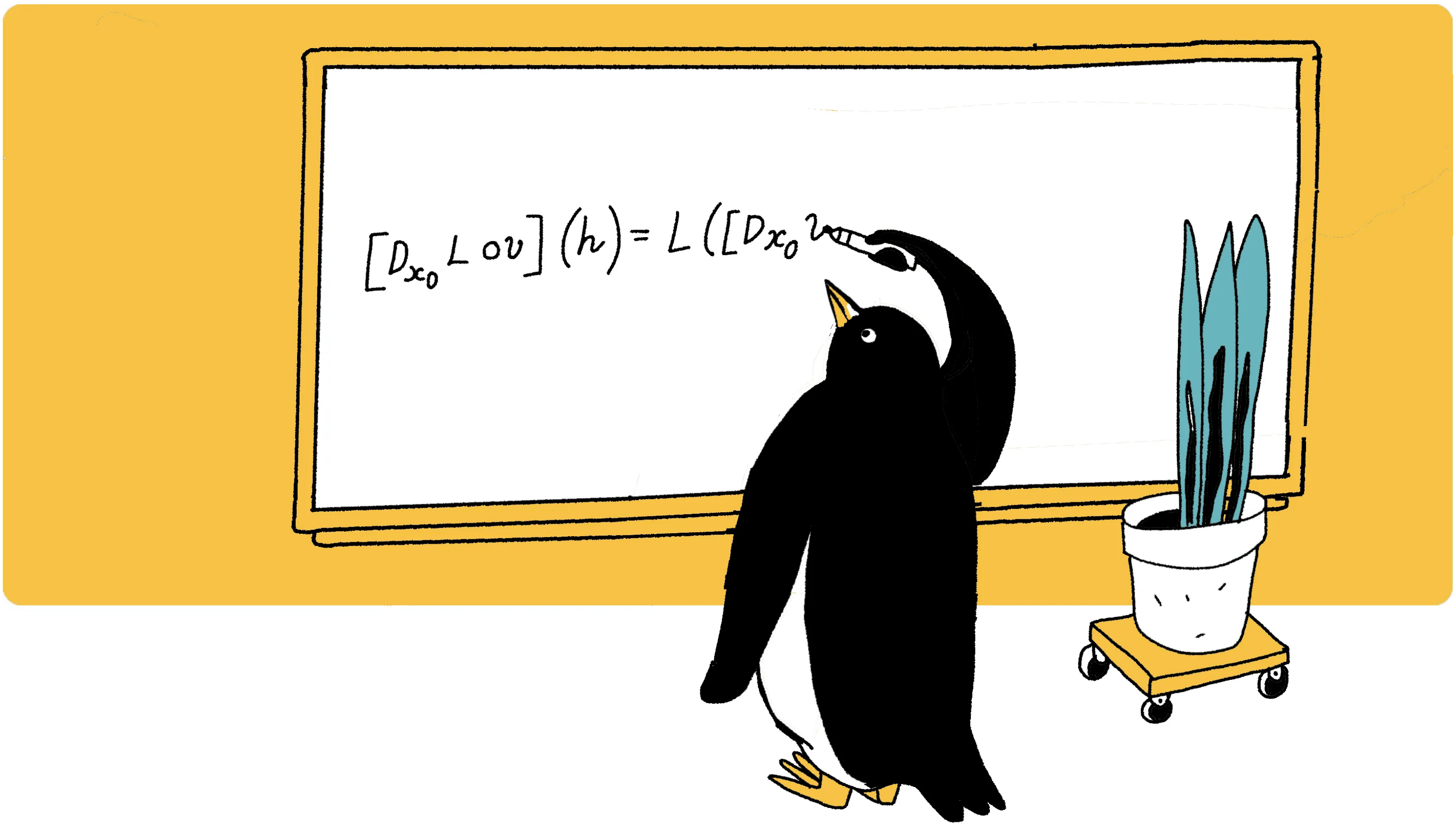

16.1. Матричное дифференцирование

Авторы

Федотов Станислав

Как дифференцировать матрицы и дифференцировать по матрицам: всё, что вам не рассказали про дифференцирование на матанализе

Любая задача машинного обучения — это задача оптимизации, а задачи оптимизации удобнее всего решать градиентными методами (если это возможно, конечно). Поэтому важно уметь находить производные всего, что попадается под руку. Казалось бы, в чём проблема: ведь дифференцирование — простая и понятная штука (чего не скажешь, например, об интегрировании). Зачем же как-то специально учиться дифференцировать матрицы?

Да в принципе-то никаких проблем: в этом параграфе вы не узнаете никаких секретных приёмов или впечатляющих теорем. Но, согласитесь, если исходная функция от вектора x имела вид $f(x) = ||Ax -$

Содержание

не только эстетически приятно, но и благотворно сказывается на производительности наших вычислений: ведь матричные операции обычно очень эффективно оптимизированы в библиотеках, чего не скажешь о самописных циклах по i, j, k, s. И всё, что будет происходить дальше, преследует очень простую цель: научиться вычислять производные в удобном, векторно-матричном виде. А чтобы сделать это и не сойти с ума, мы должны ввести ясную систему обозначений, составляющую ядро техники матричного дифференцирования.

Основные обозначения

Вспомним определение производной для функции $f : \mathbb{R}^m \to \mathbb{R}^n$. Функция $f(x)$ дифференцируема в точке x_0, если

$$f(x_0 + h) = f(x_0) + [D_{x_0} f](h) + \bar{\bar{o}}\left(||h||\right),$$

где $\left[D_{x_0} f\right]$ — **дифференциал** функции f: линейное отображение из мира x-ов в мир значений f. Грубо говоря, он превращает «малое приращение $h = \Delta x$» в «малое приращение Δf» («малые» в том смысле, что на о-малое можно плюнуть):

$$f(x_0 + h) - f(x_0) \approx [D_{x_0} f](h)$$

Отметим, что дифференциал зависит от точки x_0, в которой он берётся: $[D_{x_0} f](h)$. Под $||h||$ подразумевается норма вектора h, например корень из суммы квадратов координат (обычная евклидова длина).

Давайте рассмотрим несколько примеров и заодно разберёмся, какой вид может принимать выражение $\left[D_{x_0} f\right](h)$ в зависимости от формы x. Начнём со случаев, когда f — скалярная функция.

- **Примеры конкретных форм $\left[D_{x_0} f\right](h)$, когда f — скалярная функция**

В примерах выше нам дважды пришлось столкнуться с давним знакомцем из матанализа: **градиентом** скалярной функции (у нескалярных функций градиента не бывает). Напомним, что градиент $\nabla_{x_0} f$ функции в точке x_0 состоит из частных производных этой функции по всем координатам аргумента. При этом его обычно упаковывают в ту же форму, что и сам аргумент: если x — вектор-строка, то и градиент записывается вектор-строкой, а если x — матрица, то и градиент тоже будет матрицей того же размера. Это важно, потому что для осуществления градиентного спуска мы должны уметь прибавлять градиент к точке, в которой он посчитан.

Как мы уже имели возможность убедиться, для градиента скалярной функции f выполнено равенство

$$\left[D_{x_0} f\right](x - x_0) = \langle \nabla_{x_0} f, x - x_0 \rangle,$$

где скалярное произведение — это сумма попарных произведений соответствующих координат (да-да, самое обыкновенное).

Посмотрим теперь, как выглядит дифференцирование для функций, которые на выходе выдают не скаляр, а что-то более сложное.

- **Примеры $\left[D_{x_0} f\right](h)$, где f — это вектор или матрица**

Простые примеры и свойства матричного дифференцирования

1. **Производная константы.** Пусть $f(x) = a$. Тогда

$$f(x_0 + h) - f(x_0) = 0,$$

то есть $[D_{x_0} f]$ — это нулевое отображение. А если f — скалярная функция, то и $\nabla_{x_0} f = 0.$

2. **Производная линейного отображения.** Пусть $f(x)$ — линейное отображение. Тогда

$$f(x_0 + h) - f(x_0) = f(x_0) + f(h) - f(x_0) = f(h)$$

Поскольку справа линейное отображение, то по определению оно и является дифференциалом $[D_{x_0} f]$. Мы уже видели примеры таких ситуаций выше, когда рассматривали отображения умножения на матрицу слева или справа. Если f — (скалярная) линейная функция, то она представляется в виде $\langle a, v\rangle$ для некоторого вектора a — он и будет градиентом f.

3. **Линейность производной.** Пусть $f(x) = \lambda u(x) + \mu v(x)$, где λ, μ — скаляры, а u, v — некоторые отображения, тогда

$$[D_{x_0} f] = \lambda[D_{x_0} u] + \mu[D_{x_0} v]$$

Попробуйте доказать сами, прежде чем смотреть доказательство.

4. **Производная произведения.** Пусть $f(x) = u(x)v(x)$, где u, v — некоторые отображения, тогда

$$[D_{x_0} f] = [D_{x_0} u] \cdot v(x_0) + u(x_0) \cdot [D_{x_0} v]$$

Попробуйте доказать сами, прежде чем смотреть доказательство.

Это же правило сработает и для скалярного произведения:

$$[D_{x_0} \langle u, v\rangle] = \langle [D_{x_0} u], v\rangle + \langle u, [D_{x_0} v]\rangle$$

В этом нетрудно убедиться, повторив доказательство или заметив, что в доказательстве мы пользовались лишь дистрибутивностью (= билинейностью) умножения.

5. **Производная сложной функции.** Пусть $f(x) = u(v(x))$. Тогда

$$f(x_0 + h) - f(x_0) = u(v(x_0 + h)) - u(v(x_0)) \approx$$

$$\approx [D_{v(x_0)} u] \left(v(x_0 + h) - v(x_0)\right) \approx [D_{v(x_0)} u] \left([D_{x_0} v] (h)\right)$$

Здесь $D_{v(x_0)}u$ — дифференциал u в точке $v(x_0)$, а $\left[D_{v(x_0)}u\right](\ldots)$ — это применение отображения $\left[D_{v(x_0)}u\right]$ к тому, что в скобках. Итого получаем:

$$[D_{x_0}u \circ v](h) = \left[D_{v(x_0)}u\right]([D_{x_0}v](h))$$

6. Важный частный случай: **дифференцирование перестановочно с линейным отображением.** Пусть $f(x) = L(v(x))$, где L — линейное отображение. Тогда $\left[D_{v(x_0)}L\right]$ совпадает с самим L и формула упрощается:

$$[D_{x_0}L \circ v](h) = L([D_{x_0}v](h))$$

Простые примеры вычисления производной

- Вычислим дифференциал и градиент функции $f(x) = \langle a, x\rangle$, где x — вектор-столбец, a — постоянный вектор.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

- Вычислим производную и градиент $f(x) = \langle Ax, x\rangle$, где x — вектор-столбец, A — постоянная матрица.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

- Вычислим производную обратной матрицы: $f(X) = X^{-1}$, где X — квадратная матрица.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

- Вычислим градиент определителя: $f(X) = \det(X)$, где X — квадратная матрица.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

- Вычислим градиент функции $f(x) = ||Ax - b||^2$. С этой функцией мы ещё встретимся, когда будем обсуждать задачу линейной регрессии.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

Примеры вычисления производных сложных функций

- Вычислим градиент функции $f(X) = \log(\det(X))$.

- Попробуйте вычислить сами, прежде чем смотреть решение.

- Вычислим градиент функции $f(X) = \mathrm{tr}(AX^T X)$.

- Попробуйте вычислить сами, прежде чем смотреть решение.

- Вычислим градиент функции $f(X) = \det\left(AX^{-1}B\right)$.

- Подумайте, почему мы не можем расписать определитель в виде произведения определителей

Вторая производная

Рассмотрим теперь не первые два, а первые три члена ряда Тейлора:

$$f(x_0 + h) = f(x_0) + [D_{x_0} f](h) + \frac{1}{2}\left[D^2_{x_0} f\right](h, h) + \bar{\bar{o}}\left(||h||^2\right),$$

где $\left[D^2_{x_0} f\right](h, h)$ — второй дифференциал, квадратичная форма, в которую мы объединили все члены второй степени.

Вопрос на подумать. Докажите, что второй дифференциал является дифференциалом первого, то есть

$$[D_{x_0}[D_{x_0} f](h_1)]\,(h_2) = \left[D^2_{x_0} f\right]\,(h_1, h_2)$$

Зависит ли выражение справа от порядка h_1 и h_2?

Этот факт позволяет вычислять второй дифференциал не с помощью приращений, а повторным дифференцированием производной.

Вторая производная может оказаться полезной при реализации методов второго порядка или же для проверки того, является ли критическая точка (то есть точка, в которой градиент обращается в ноль) точкой минимума или точкой максимума. Напомним, что квадратичная форма $q(h)$ называется положительно определённой (соответственно, отрицательно определённой), если $q(h) \geqslant 0$ (соответственно, $q(h) \leqslant 0$) для всех h, причём $q(h) = 0$ только при $h = 0$.

Теорема. Пусть функция $f : \mathbb{R}^m \to \mathbb{R}$ имеет непрерывные частные производные второго порядка $\frac{\partial^2 f}{\partial x_i \partial x_j}$ в окрестности точки x_0, причём $\nabla_{x_0} f = 0$. Тогда точка x_0 является точкой минимума функции, если квадратичная форма $D^2_{x_0} f$ положительно определена, и точкой максимума, если она отрицательно определена.

Если мы смогли записать матрицу квадратичной формы второго дифференциала, то мы можем проверить её на положительную или отрицательную определённость с помощью критерия Сильвестра.

Примеры вычисления и использования второй производной

- Рассмотрим задачу минимизации $f(x) = ||Ax - b||^2$ по переменной x, где A — матрица с линейно независимыми столбцами. Выше мы уже нашли градиент этой функции; он был равен $\nabla_{x_0} f = 2A^T(Ax - b)$. Мы можем заподозрить, что минимум достигается в точке, где градиент обращается в ноль: $x_* = (A^TA)^{-1}A^Tb$. Отметим, что обратная матрица существует, так как $\mathrm{rk}(A^TA) = \mathrm{rk}A$, а столбцы A по условию линейно независимы и, следовательно, $\mathrm{rk}(A^TA)$ равен размеру этой матрицы. Но действительно ли эта точка является точкой минимума? Давайте оставим в стороне другие соображения (например, геометрические, о которых мы упомянем в параграфе про линейные модели) и проверим аналитически. Для этого мы должны вычислить второй дифференциал функции $f(x) = ||Ax - b||^2$.

▸ **Попробуйте вычислить сами, прежде чем смотреть решение.**

Мы нашли квадратичную форму второго дифференциала; она, оказывается, не зависит от точки (впрочем, логично: исходная функция была второй степени по x, так что вторая производная должна быть константой). Чтобы показать, что x_* действительно является точкой минимума, достаточно проверить, что эта квадратичная форма положительно определена.

▸ **Попробуйте сделать это сами, прежде чем смотреть решение.**

- Докажем, что функция $f(X) = \log\det(X)$ является выпуклой вверх на множестве симметричных, положительно определённых матриц. Для этого мы должны проверить, что в любой точке квадратичная форма её дифференциала отрицательно определена. Для начала вычислим эту квадратичную форму.

▸ **Попробуйте сделать это сами, прежде чем смотреть решение.**

Чтобы доказать требуемое в условии, мы должны проверить следующее: что для любой симметричной матрицы X_0 и для любого симметричного (чтобы не выйти из пространства симметричных матриц) приращения $H \neq 0$ имеем

$$\left[D^2_{X_0} \log\det(X)\right](H, H) < 0$$

Покажем это явно.
Так как X_0 — симметричная, положительно определённая матрица, у неё есть симметричный и положительно определённый квадратный корень: $X_0 = X_0^{1/2} \cdot X_0^{1/2} = X_0^{1/2} \cdot \left(X_0^{1/2}\right)^T$. Тогда

$$\langle -X_0^{-1}HX_0^{-1}, H\rangle = -\mathrm{tr}\left(X_0^{1/2}\left(X_0^{1/2}\right)^T HX_0^{1/2}\left(X_0^{1/2}\right)^T H^T\right) =$$

$$-\mathrm{tr}\left(\left(X_0^{1/2}\right)^T HX_0^{1/2}\left(X_0^{1/2}\right)^T H^TX_0^{1/2}\right) =$$

$$= -\mathrm{tr}\left(\left(X_0^{1/2}\right)^T HX_0^{1/2}\left[\left(X_0^{1/2}\right)^T HX_0^{1/2}\right]^T\right) =$$

$$= -||\left(X_0^{1/2}\right)^T HX_0^{1/2}||^2,$$

что, конечно, меньше нуля для любой ненулевой H.

Параграф не прочитан

Отмечайте параграфы как прочитанные чтобы видеть свой прогресс обучения

Сообщить об ошибке

Предыдущий параграф

15.3. Методы оптимизации в Deep Learning

Следующий параграф

16.2. Матричная факторизация

Яндекс Практикум

Школа анализа данных

Программы в университетах

База знаний

Журнал

События

Партнерам

Сведения об образовательной организации

Пользовательское соглашение хендбуков

Рассылка

Бот