Методы оптимизации. Семинар 3. Векторное дифференцирование

Александр Катруца

Московский физико-технический институт

15 сентября 2020 г.

Напоминание

- Сопряжённые конусы
- Свойства сопряжённых множеств
- Опорная гиперплоскость
- Разделяющая гиперплоскость
- Лемма Фаркаша

Основные определения

Более подробно смотрите здесь. Пусть $f : D \to E$, производная $\frac{\partial f}{\partial x} \in G$:

D	E	G	Название
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	Производная, $f'(x)$
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	Градиент, $\frac{\partial f}{\partial x_i}$
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{m \times n}$	Матрица Якоби, $\frac{\partial f_i}{\partial x_j}$
$\mathbb{R}^{m \times n}$	$\mathbb{R}$	$\mathbb{R}^{m \times n}$	$\frac{\partial f}{\partial x_{ij}}$

Также квадратная $n \times n$ матрица вторых производных $\mathbf{H} = [h_{ij}]$ в случае $f : \mathbb{R}^n \to \mathbb{R}$ называется гессиан и равна $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Основная техника

Матричное представление функции: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}, \;\; \mathbf{A} \in \mathbf{S}^n$

Скалярное представление: $f = \sum\limits_{ij} a_{ij} x_i x_j$

$f'_k = \frac{\partial f}{\partial x_k}$

Скалярное представление градиента: $f'_k(\mathbf{x}) = 2 \sum\limits_j a_{kj} x_j$

Матричное представление градиента: $f'(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Примеры

1. Линейная функция: $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x}$
2. Квадратичная форма: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$
3. Квадрат ℓ_2 нормы разности: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$
4. Детерминант: $f(\mathbf{X}) = \det \mathbf{X}$
5. След: $f(\mathbf{X}) = \mathrm{trace}(\mathbf{A}\mathbf{X}\mathbf{B})$
6. $f(\mathbf{x}) = (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})$
7. $f(\mathbf{A}) = (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})$
8. $f(\mathbf{s}) = (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})$

Сложная функция: скалярный случай

- Пусть $f(\mathbf{x}) = g(u(\mathbf{x}))$, тогда $f'(\mathbf{x}) = \frac{\partial g}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$
- Важно смотреть на размерности и понимать как записывать $\frac{\partial g}{\partial u}$.

Сложная функция: скалярный случай

- Пусть $f(\mathbf{x}) = g(u(\mathbf{x}))$, тогда $f'(\mathbf{x}) = \frac{\partial g}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$
- Важно смотреть на размерности и понимать как записывать $\frac{\partial g}{\partial u}$.

Примеры

1. ℓ_2 норма вектора: $f(\mathbf{x}) = \|\mathbf{x}\|_2$
2. Экспонента: $f(\mathbf{x}) = -e^{-\mathbf{x}^\top \mathbf{x}}$

Сложная функция: векторный случай

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, где $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ — k-ый элемент градиента
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, где $\mathbf{J}$ — якобиан h

Сложная функция: векторный случай

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, где $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ — k-ый элемент градиента
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, где $\mathbf{J}$ — якобиан h

Примеры

- $h(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $g(\mathbf{u}) = \|\mathbf{u}\|_2^2$. Найти $f'(\mathbf{x})$
- $h(\mathbf{x}) = \cos(\mathbf{x})$ поэлементно, $g(\mathbf{u}) = \sum_i u_i$. Найти $\frac{\partial f}{\partial \mathbf{x}}$

Chain rule and autodiff[1]

Мотивирующий пример

- $f = h(g(\mathbf{x}))$, где $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ или $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^{k} \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j}$

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

Chain rule and autodiff[1]

Мотивирующий пример

- $f = h(g(\mathbf{x}))$, где $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ или $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^{k} \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Обобщение

- $f = f_L \circ \ldots \circ f_1$ — представление в виде графа
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

Chain rule and autodiff[1]

Мотивирующий пример

- $f = h(g(\mathbf{x}))$, где $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ или $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^k \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j}$

Обобщение

- $f = f_L \circ \ldots \circ f_1$ — представление в виде графа
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

Способы вычисления $\mathbf{J}_f$

- Справа налево — forward mode
- Слева направо — backward mode

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Реализация

- Выбираем элемент x_j

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Реализация

- Выбираем элемент x_j
- Задаём вектор $\mathbf{u} = \mathbf{e}_j$ — j-ый орт

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Реализация

- Выбираем элемент x_j
- Задаём вектор $\mathbf{u} = \mathbf{e}_j$ — j-ый орт
- Умножаем рекурсивно $\mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ справа налево

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Реализация

- Выбираем элемент x_j
- Задаём вектор $\mathbf{u} = \mathbf{e}_j$ — j-ый орт
- Умножаем рекурсивно $\mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ справа налево
- Умножение происходит одновременно с вычислением $f_L \circ \dots \circ f_1$

Forward mode

Основная идея

Вычислить $\frac{\partial f_i}{\partial x_k}$ для всех i и для заданного k, то есть вычислить j-ый столбец матрицы $\mathbf{J}_f$

Реализация

- Выбираем элемент x_j
- Задаём вектор $\mathbf{u} = \mathbf{e}_j$ — j-ый орт
- Умножаем рекурсивно $\mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ справа налево
- Умножение происходит одновременно с вычислением $f_L \circ \ldots \circ f_1$
- Для каждой f_i необходимо реализовать действие самой функции и умножение $\mathbf{J}_i$ на вектор

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k
- Задаём вектор $\mathbf{u} = \mathbf{e}_k$ — k-ый орт

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k
- Задаём вектор $\mathbf{u} = \mathbf{e}_k$ — k-ый орт
- Умножаем рекурсивно $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ слева направо

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k
- Задаём вектор $\mathbf{u} = \mathbf{e}_k$ — k-ый орт
- Умножаем рекурсивно $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ слева направо
- Сначала вычисляем f, потом произведение выше — два обхода графа

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k
- Задаём вектор $\mathbf{u} = \mathbf{e}_k$ — k-ый орт
- Умножаем рекурсивно $\mathbf{u}^\top \mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1$ слева направо
- Сначала вычисляем f, потом произведение выше — два обхода графа
- Для каждой f_i необходимо реализовать действие самой функции и умножение $\mathbf{J}_i^\top$ на вектор

Backward mode или backpropagation

Основная идея

Вычислить $\frac{\partial f_k}{\partial x_i}$ для всех i и для заданного k, то есть вычислить j-ую строку матрицы $\mathbf{J}_f$

Реализация

- Выбираем компоненту f_k
- Задаём вектор $\mathbf{u} = \mathbf{e}_k$ — k-ый орт
- Умножаем рекурсивно $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ слева направо
- Сначала вычисляем f, потом произведение выше — два обхода графа
- Для каждой f_i необходимо реализовать действие самой функции и умножение $\mathbf{J}_i^\top$ на вектор

Если $m = 1$, то $\mathbf{u} = 1$ и результат совпадает с градиентом!

Forward vs backward modes

Вычислительная сложность

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Forward vs backward modes

Вычислительная сложность

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Требуемая память

- Forward mode: не требует, все вычисления делаются в процессе вычисления f
- Backward mode: требует, промежуточные значения f_{i-1} надо сохранить для вычисления $\mathbf{J}_i^\top\mathbf{u}$

Forward vs backward modes

Вычислительная сложность

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5 C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top \mathbf{u}) \leq 4 C(f(\mathbf{x}))$

Требуемая память

- Forward mode: не требует, все вычисления делаются в процессе вычисления f
- Backward mode: требует, промежуточные значения f_{i-1} надо сохранить для вычисления $\mathbf{J}_i^\top \mathbf{u}$

Вывод

- Если $m \ll n$, используйте backward mode
- Если $m \geq n$, используйте forward mode

Различные реализации могут оптимизировать промежуточные вычисления!

Пример

Для функции $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Найти $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$
$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Пример

Для функции $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Найти $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$

$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Forward mode

- Вычислим $\frac{\partial f}{\partial x_2}$
- $w_1 = x_1$, $w_2 = x_2$
- $\frac{\partial w_1}{\partial x_1} = 0$, $\frac{\partial w_2}{\partial x_2} = 1$
- $w_3 = 3w_2^2 \frac{\partial w_2}{\partial x_2}$
- $w_4 = \frac{\partial w_1}{\partial x_1} + w_3$
- $w_5 = -\sin(w_1 + w_2^3) w_4$
- $w_6 = 2\cos(w_1 + w_2^3) w_5$
- $w_6 = \frac{\partial f}{\partial x_2}$

Пример

Для функции $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Найти $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$
$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Forward mode

- Вычислим $\frac{\partial f}{\partial x_2}$
- $w_1 = x_1$, $w_2 = x_2$
- $\frac{\partial w_1}{\partial x_1} = 0$, $\frac{\partial w_2}{\partial x_2} = 1$
- $w_3 = 3w_2^2 \frac{\partial w_2}{\partial x_2}$
- $w_4 = \frac{\partial w_1}{\partial x_1} + w_3$
- $w_5 = -\sin(w_1 + w_2^3) w_4$
- $w_6 = 2\cos(w_1 + w_2^3) w_5$
- $w_6 = \frac{\partial f}{\partial x_2}$

Backward mode

- $w_0 = 1$
- $w_1 = \frac{\partial f_1}{\partial f_2} w_0 = 2 f_2 w_0$
- $w_2 = \frac{\partial f_2}{\partial f_3} w_1 = -\sin(f_3) w_1$
- $w_3 = \frac{\partial f}{\partial x_1} = \frac{\partial f_3}{\partial x_1} w_2 = w_2$
- $w_4 = \frac{\partial f_3}{\partial f_4} w_2$
- $w_5 = \frac{\partial f}{\partial x_2} = \frac{\partial f_4}{\partial x_2} w_4 = 3x_2^2 w_4$

Умножение гессиана на вектор

- Дан вектор $\mathbf{z}$, нужно вычислить $f''(\mathbf{x})\mathbf{z}$

Умножение гессиана на вектор

- Дан вектор $\mathbf{z}$, нужно вычислить $f''(\mathbf{x})\mathbf{z}$
- Вспомним, что $f''(\mathbf{x}) = (f'(\mathbf{x}))'$

Умножение гессиана на вектор

- Дан вектор $\mathbf{z}$, нужно вычислить $f''(\mathbf{x})\mathbf{z}$
- Вспомним, что $f''(\mathbf{x}) = (f'(\mathbf{x}))'$
- Вычислим градиент $f'(\mathbf{x})$ с помощью backward mode

Умножение гессиана на вектор

- Дан вектор $\mathbf{z}$, нужно вычислить $f''(\mathbf{x})\mathbf{z}$
- Вспомним, что $f''(\mathbf{x}) = (f'(\mathbf{x}))'$
- Вычислим градиент $f'(\mathbf{x})$ с помощью backward mode
- И вычислим $f''(\mathbf{x})\mathbf{z} = \mathbf{J}_{f'}\mathbf{z}$ с помощью forward mode

Умножение гессиана на вектор

- Дан вектор $\mathbf{z}$, нужно вычислить $f''(\mathbf{x})\mathbf{z}$
- Вспомним, что $f''(\mathbf{x}) = (f'(\mathbf{x}))'$
- Вычислим градиент $f'(\mathbf{x})$ с помощью backward mode
- И вычислим $f''(\mathbf{x})\mathbf{z} = \mathbf{J}_{f'}\mathbf{z}$ с помощью forward mode

Почему это хорошо?

- Полный гессиан не хранится — экономия памяти
- Выбор режимов вычисления градиентов и умножения гессиана на вектор обоснованы размерностями входа и выхода функций f и f'

Резюме

- Производная по скаляру
- Производная по вектору
- Производная по матрице
- Производная сложной функции
- Автоматическое дифференцирование