Методы оптимизации в машинном обучении, ШАД, весна 2017

Семинар 1: Скорости сходимости и матричные вычисления

13 февраля 2017 г.

1 Скорости сходимости

В теории оптимизации мы хотим решить следующую задачу:

$$\min_{x \in E} f(x), \tag{1}$$

где $f : \mathbb{R}^n \to \mathbb{R}$ — функция, E — подмножество $\mathbb{R}^n$.

Решить задачу (1) означает найти точку $x^* \in E$, в которой достигается оптимальное значение $f^* := \inf_{x \in E} f(x)$ функции f на множестве E (в предположении, что такая точка существует):

$$x^* = \operatorname*{argmin}_{x \in E} f(x), \qquad f^* = f(x^*).$$

Но, к сожалению, довольно часто мы не можем решить задачу (1) точно, поэтому приходится довольствоваться *приближенным решением*, которое близко к оптимальному.

Наши методы обычно будут стартовать с некоторой начальной точки $x_0 \in E$ и итеративно строить последовательность $(x_k)_{k=1}^{\infty}$, где $x_k \in E$. От любого разумного метода мы хотим, чтобы последовательность соответствующих значений функции $(f(x_k))_{k=0}^{\infty}$ сходилась к оптимуму f^* при $k \to \infty$:

$$r_k := f(x_k) - f^* \geq 0, \qquad \lim_{k \to \infty} r_k = 0.$$

Но для того, чтобы сравнивать методы между собой, этого недостаточно. Введем понятие скорости сходимости.

Определение 1 (Линейная скорость сходимости)**.** Пусть $(r_k)_{k=0}^{\infty}$ — последовательность неотрицательных чисел, сходящаяся к нулю. Говорят, что $(r_k)_{k=0}^{\infty}$ имеет *линейную скорость сходимости* с параметром $q \in (0, 1)$, если существует $C > 0$, такое, что для всех $k \geq 0$ выполнено

$$r_k \leq C q^k.$$

Нижняя граница всех q, для которых $(r_k)_{k=0}^{\infty}$ имеет линейную скорость сходимости с параметром q, называется *константой линейной сходимости* последовательности $(r_k)_{k=0}^{\infty}$.

Замечание 1. Скорость сходимости — это ассимптотическое понятие. Константа C в определении может быть сколь угодно большой (и зависеть от параметра q). Сходимость последовательности $(r_k)_{k=k_0}^{\infty}$, начинающейся с некоторого номера k_0, влечет ту же скорость сходимости и для всей последовательности $(r_k)_{k=0}^{\infty}$, нужно лишь C увеличить должным образом.

Определение 2 (Сверхлинейная скорость сходимости)**.** Пусть $(r_k)_{k=0}^{\infty}$ — последовательность неотрицательныъ чисел, сходящаяся к нулю. Говорят, что $(r_k)_{k=0}^{\infty}$ имеет *сверхлинейную скорость сходимости*, если $(r_k)_{k=0}^{\infty}$ является линейно сходящейся с любым параметром $q \in (0, 1)$. Другими словами, $(r_k)_{k=0}^{\infty}$ сходится сверхлинейно, если константа ее линейной сходимости равна 0.

Определение 3 (Сублинейная скорость сходимости)**.** Пусть $(r_k)_{k=0}^{\infty}$ — последовательность неотрицательных чисел, сходящаяся к нулю. Говорят, что $(r_k)_{k=0}^{\infty}$ имеет *сублинейную скорость сходимости*, если $(r_k)_{k=0}^{\infty}$ не является линейно сходящейся ни для какого параметра $q \in (0, 1)$.

Скорость сходимости удобно устанавливать с помощью следующего теста.

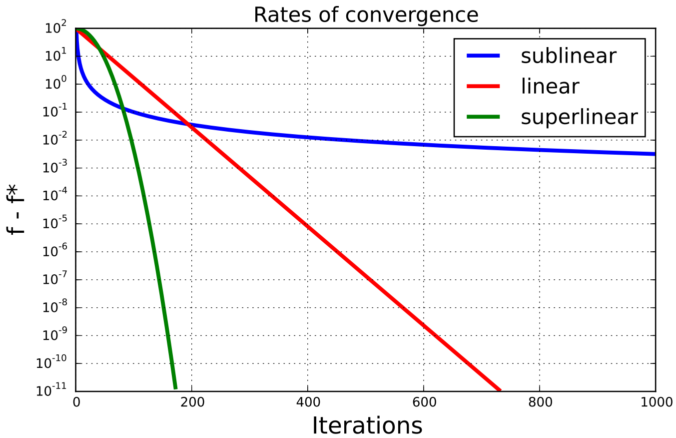

Рис. 1: Иллюстрация к определению скоростей сходимости.

Утверждение 1. *Пусть* $(r_k)_{k=0}^{\infty}$ *— последовательность положительных чисел, сходящаяся к нулю.*

1. *Если* $\limsup_{k\to\infty} \frac{r_{k+1}}{r_k} =: q < 1$, *то* $(r_k)_{k=0}^{\infty}$ *имеет линейную скорость сходимости с константой* q.
2. *Если* $\lim_{k\to\infty} \frac{r_{k+1}}{r_k} = 0$, *то* $(r_k)_{k=0}^{\infty}$ *имеет сверхлинейную скорость сходимости.*
3. *Если* $\lim_{k\to\infty} \frac{r_{k+1}}{r_k} = 1$, *то* $(r_k)_{k=0}^{\infty}$ *имеет сублинейную скорость сходимости.*
4. *В остальных случаях данный тест ничего не говорит о скорости сходимости* $(r_k)_{k=0}^{\infty}$.

Особый случай сверхлинейной сходимости — сходимость порядка p.

Определение 4 (Сходимость порядка p)**.** Пусть $(r_k)_{k=0}^{\infty}$ — последовательность неотрицательных чисел, сходящаяся к нулю. Говорят, что последовательность $(r_k)_{k=0}^{\infty}$ имеет *сходимость порядка* $p > 1$, если существует $M > 0$, такое, что для всех достаточно больших k выполняется

$$r_{k+1} \le M r_k^p.$$

Сходимость порядка p для случая $p = 2$ имеет специальное название — *квадратичная сходимость*.

2 Матричные вычисления

Под вектором $x \in \mathbb{R}^n$ мы будем понимать вектор-столбец. Матрицу $A \in \mathbb{R}^{m\times n}$ можно умножить на такой вектор слева:

$$y = Ax, \qquad y \in \mathbb{R}^m.$$

Операция транспонирования матрицы:

$$A \in \mathbb{R}^{m\times n} \qquad \Rightarrow \qquad A^T \in \mathbb{R}^{n\times n}, \quad (A^T)_{ij} = A_{ji}.$$

Если мы хотим умножить матрицу на вектор справа, то вектор необходимо транспонировать:

$$u = v^T A, \qquad v \in \mathbb{R}^m, \ u \in \mathbb{R}^{1\times n}.$$

Также заметим, что для двух векторов одинакового размера $a, b \in \mathbb{R}^n$:

$$
\begin{aligned}
a^T b &= \sum_{i=1}^n a_i b_i \in \mathbb{R} && \text{— число,}\\
ab^T &= (a_i b_j)_{i,j=1}^{n,n} \in \mathbb{R}^{n\times n} && \text{— матрица.}
\end{aligned}
$$

Для квадратной матрицы $A \in \mathbb{R}^{n\times n}$ определена операция взятия следа

$$
\mathrm{Tr}(A) = \sum_{i=1}^n a_{ii}.
$$

Напомним основные три векторные нормы: ℓ_2-норма (евклидова норма), ℓ_1-норма и ℓ_∞-норма (норма Чебышева или равномерная норма):

$$
\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2}, \qquad \|x\|_1 = \sum_{i=1}^n |x_i|, \qquad \|x\|_\infty = \max_{1\le i\le n} |x_i|.
$$

Основные две матичные нормы: спектральная норма (или операторная норма) и норма Фробениуса:

$$
\|A\|_2 = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2}, \qquad \|A\|_F := \left(\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2\right)^{1/2} = [\mathrm{Tr}(A^T A)]^{1/2}.
$$

Некоторый свойства:

1. $(AB)^{-1} = B^{-1}A^{-1}$.
2. $(AB)^T = B^T A^T$.
3. $(A^{-1})^T = (A^T)^{-1} =: A^{-T}$.
4. $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.
5. $\mathrm{Tr}(A^T) = \mathrm{Tr}(A)$.
6. Если $x \in \mathbb{R} \simeq \mathbb{R}^{1\times 1}$, то $\mathrm{Tr}(x) = x$.

2.1 Собственные значения и спектральное разложение

Пусть $A \in \mathbb{R}^{n\times n}$ — квадратная матрица. Напомним определение собственного вектора и собственного значения.

Определение 5 (Собственные векторы и собственные значения)**.** Пусть $x \in \mathbb{C}^n \setminus \{0\}$ — ненулевой вектор с комплексными элементами и $\lambda \in \mathbb{C}$ — комплексное число. Вектор x называется *собственным вектором* матрицы A, отвечающим *собственному значению* λ, если

$$
Ax = \lambda x.
$$

Утверждение 2. *Любая матрица $A \in \mathbb{R}^{n\times n}$ имеет ровно n собственных значений с учетом кратности.*

Обозначим множество всех симметричных матрицы размера n через $\mathbb{S}^n$:

$$
\mathbb{S}^n := \{A \in \mathbb{R}^{n\times n} : A = A^T\}.
$$

Утверждение 3. *Если матрица симметричная, то все ее собственные значения вещественные.*

Согласно этому утверждению, для симметричной матрицы $A \in \mathbb{S}^n$ ее собственные значения можно отсортировать:

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A).$$

Будем обозначать через $\lambda(A) \in \mathbb{R}^n$ вектор из собственных значений матрицы A, осортированный по убыванию:

$$\lambda(A) := (\lambda_1(A), \ldots, \lambda_n(A)).$$

Утверждение 4. *Пусть $A \in \mathbb{S}^n$. Тогда существует ортонормированный базис $u_1, \ldots, u_n \in \mathbb{R}^n$, состоящий из собственных векторов матрицы A, отвечающих собственным значениям $\lambda_1(A), \ldots, \lambda_n(A)$ соответственно:*

$$u_i^T u_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, \qquad Au_i = \lambda_i u_i.$$

Пусть $U := [u_1, \ldots, u_n] \in \mathbb{R}^{n \times n}$ — матрица со столбцами $u_1, \ldots, u_n$. Ортонормированность базиса означает, что матрица U является *ортогональной*, т. е. $UU^T = U^T U = I_n$. Обозначим $\Lambda := \mathrm{Diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}$. Тогда получаем *спектральное разложение*:

$$AU = \Lambda U \qquad \Leftrightarrow \qquad A = U \Lambda U^T.$$

2.2 Положительно определенные матрицы

Одним из наиболее важных классов среди симметричных матриц является класс положительно определенных матриц.

Определение 6. Матрица $A \in \mathbb{S}^n$ называется *положительно определенной* (обозначение: $A \succ 0$), если для всех $x \in \mathbb{R}^n \setminus \{0\}$ выполнено

$$x^T Ax > 0.$$

Если соответствующее неравенство является нестрогим, т. е. $x^T Ax \geq 0$ для всех $x \in \mathbb{R}^n$, то матрица A называется *положительно полуопределенной* (обозначение: $A \succeq 0$).

Для удобства в дальнейшем будем использовать следующие обозначения

$$\mathbb{S}^n_+ := \{A \in \mathbb{S}^n : A \succeq 0\}, \qquad \mathbb{S}^n_{++} := \{A \in \mathbb{S}^n : A \succ 0\}.$$

Заметим, что $\mathbb{S}^n_{++} \subset \mathbb{S}^n_+ \subset \mathbb{S}^n$.

2.3 Ранг матрицы и линейные уравнения

Напомним определение *ранга* матрицы.

Определение 7 (Ранг матрицы)**.** Пусть $A \in \mathbb{R}^{m \times n}$. *Рангом матрицы* A (обозначение: $\mathrm{Rank}(A)$) называется размерность линейной оболочки ее столбцов, или, эквивалентно, размерность линейной оболочки ее строк.

Из определения сразу же следует, что ранг матрицы не может быть больше числа столбцов и числа строк этой матрицы: $\mathrm{Rank}(A) \leq \min\{m, n\}$.

Заметим, что матрица вида uv^T, где $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, имеет ранг 1. Поскольку ранг суммы матриц не превосходит суммы рангов соответствующих слагаемых, то матрица A вида

$$A = \sum_{i=1}^{k} u_i v_i^T, \tag{2}$$

где $u_i \in \mathbb{R}^m$, $v_i \in \mathbb{R}^n$ для $1 \leq i \leq k$, имеет ранг не больше k. Справедливо и обратное: если матрица $A \in \mathbb{R}^{m \times n}$ имеет ранг r, то для нее найдется разложение (2), в котором $k = r$ (это разложение называется *ранговым разложением*). Таким образом, справедливо следующее альтернативное определение ранга.

Утверждение 5. *Ранг матрицы $A \in \mathbb{R}^{m\times n}$ равен минимальному числу r одноранговых матриц, сумма которых образует матрицу A:*

$$A = \sum_{i=1}^{r} u_i v_i^T,$$

где $u_i \in \mathbb{R}^m$, $v_i \in \mathbb{R}^n$.

Ранг матрицы тесно связан с *системами линейных уравнений*

$$Ax = b,$$

где $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. Здесь возможны следующие три ситуации.

1. (Решений нет) Это возможно только в том случае, когда $b \notin \mathrm{Im}(A)$, где $\mathrm{Im}(A) := \{Ax : x \in \mathbb{R}^n\}$ — образ матрицы A. В частности, такое возможно лишь в случае *переопределенной системы* (когда $m \geq n$) и только если $b \neq 0$.

2. (Решение единственное) Это возможно только в том случае, когда $\mathrm{Rank}(A) = n$ и $b \in \mathrm{Im}(A)$. В частности, если $m = n$, то решение выражается через обратную матрицу: $x = A^{-1}b$.

3. (Решений бесконечно много) Это возможно только в том случае, когда $\mathrm{Rank}(A) < n$ и $b \in \mathrm{Im}(A)$. В этом случае множество всевозможных решений представляет из себя линейное многообразие $\hat{x} + \mathrm{Ker}(A)$, где $\hat{x} \in \mathbb{R}^n$ — произвольное (частное) решение, а $\mathrm{Ker}(A) := \{x \in \mathbb{R}^n : Ax = 0\}$ — ядро матрицы A.