

§7 МЕТОД СПУСКА

Пусть $f: X \rightarrow \mathbb{R}$ — дифференцируемая выпуклая функция, а значит, для всех $x, y \in X$ выполнено неравенство $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$. Для численного определения минимального значения функции f используют так называемый метод

спуска, который заключается в построении последовательностей векторов $(x^k)_{k \in \mathbb{N}_0}$, $(\Delta x^k)_{k \in \mathbb{N}_0}$ и действительных чисел $(t_k)_{k \in \mathbb{N}_0}$, таких что $x^{k+1} = x^k + t^k \Delta x^k \in X$ и $f(x^{k+1}) \leq f(x^k)$, где $x^0 \in X$ — заданный начальный вектор.

На практике, элементы последовательности $(x^k)_{k \in \mathbb{N}_0}$ вычисляются до тех пор пока не будет выполнен критерий остано-

ва. Например, для заданного $\varepsilon > 0$, таким условием является выполнение одно из неравенств: $\|x^{k+1} - x^k\| \leq \varepsilon$, $|f(x^{k+1}) - f(x^k)| < \varepsilon$, $\|\nabla f(x^k)\| \leq \varepsilon$.

Так как f — выпуклая функция, то из неравенства $\nabla f(x^k)^\top (y - x^k) > 0$ следует, что $f(y) > f(x^k)$. Таким образом, в методе спуска необходимо выполнение неравенства $\nabla f(x^k)^\top \Delta x^k \leq 0$. Будем говорить, что вектор Δx является направлением спуска в точке x , если $\nabla f(x)^\top \Delta x \leq 0$. Последовательность $(t^k)_{k \in \mathbb{N}_0}$ называется последовательностью размеров шагов. Метод спуска имеет множество вариаций, которые различаются способом вычисления направлений спуска $(\Delta x^k)_{k \in \mathbb{N}_0}$ и размеров шагов $(t_k)_{k \in \mathbb{N}_0}$.

Далее будем предполагать, что $X = \mathbb{R}^n$. Пусть Δx — направление спуска в точке x . Рассмотрим два способа определения размера шага t . Первый способ заключается в решении следующей задачи

$$t^* = \arg \min_{t \geq 0} f(x + t \Delta x)$$

и называется методом наискорейшего спуска (Exact Line Search, ELS). Второй

Алгоритм 7.2. Backtracking Line Search

Input: направление спуска Δx , $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$

Output: размер шага t

```

1  $t \leftarrow 1$ 
2 while  $f(x + t \Delta x) > f(x) + \alpha t \nabla f(x)^\top \Delta x$  do
3    $t \leftarrow \beta t$ 
4 return  $t$ 
```

способ называется Backtracking Line Search (BLS) и зависит от двух параметров $\alpha \in \left(0, \frac{1}{2}\right)$, $\beta \in (0, 1)$. Начиная с $t = 1$, будем уменьшать t , умножением на β , до тех пор пока не будет выполнено неравенство $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^\top \Delta x$. Нетрудно показать, что если выполнено неравенство $\nabla f(x)^\top \Delta x < 0$, то искомое t будет найдено за конечное число итераций. Действительно,

$$f(x + t\Delta x) = f(x) + t \nabla f(x)^\top \Delta x + o(t) = f(x) + \alpha t \nabla f(x)^\top \Delta x + th(t),$$

где $h(t) = \left(\frac{o(t)}{t} + (1 - \alpha) \nabla f(x)^\top \Delta x\right)$. При достаточно малом $t > 0$ функция $h(t)$ принимает отрицательные значения.

Зафиксируем точку x и определим направление ℓ^* , $\|\ell^*\| = 1$, наибольшего убывания, т.е. $\ell^* = \arg \min_{\ell \in \mathbb{S}^{n-1}} \frac{\partial f(x)}{\partial \ell}$, где $\mathbb{S}^{n-1} \stackrel{\text{def}}{=} \{x: \|x\| = 1\}$ — сфера единичного радиуса с центром в начале координат. Так как $\frac{\partial f(x)}{\partial \ell} = \nabla f(x)^\top \ell \geq -\|\nabla f(x)\|$, то направление наибольшего убывания ℓ^* сонаправлено с $-\nabla f(x)$. Метод спуска, в котором $\Delta x = -\nabla f(x)$, называется методом градиентного спуска. Перед тем, как провести анализ сходимости метода градиентного спуска, докажем ряд вспомогательных утверждений [1].

Пусть $A, B \in \mathbb{R}^{n \times n}$ — квадратные матрицы порядка n с вещественными элементами. Если матрица $C = A - B$ является неотрицательно определённой, т.е. для любого $x \in \mathbb{R}^n$ выполнено неравенство $x^\top C x \geq 0$, то будем писать $A \succeq B$. Для векторов будем использовать евклидову норму, а для матриц — спектральную.

Лемма 7.1. Пусть $f \in C^2(\mathbb{R}^n)$ и нашлось такое положительное число m , что $\nabla^2 f(x) \succeq mI$, $x \in \mathbb{R}^n$. Тогда для всех x, y справедливы неравенства

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \|y - x\|^2; \quad (7.1)$$

$$\min_{y \in \mathbb{R}^n} f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|^2; \quad (7.2)$$

$$\|x - x^*\| \leq \frac{2}{m} \|\nabla f(x)\|, \quad \text{где } x^* = \arg \min_{x \in \mathbb{R}^n} f(x). \quad (7.3)$$

▷ Согласно формуле Тейлора существует такая точка z , принадлежащая отрезку $[x, y]$, что $f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2} (y - x)^\top \nabla^2 f(z) (y - x)$.

Так как $\nabla^2 f(z) - mI$ — неотрицательно определённая матрица, то

$$(y - x)^\top (\nabla^2 f(z) - mI) (y - x) \geq 0,$$

т.е. $(y - x)^\top \nabla^2 f(z) (y - x) \geq m \|y - x\|^2$, а значит, имеет место неравенство (7.1).

При фиксированном x функция $g(y) = f(x) + \nabla f(x)^\top(y - x) + \frac{m}{2}\|y - x\|^2$ является строго выпуклой. Нетрудно видеть, что $\tilde{y} = x - \frac{1}{m}\nabla f(x)$ — стационарная точка функции g , а значит, в этой точке функция g принимает минимальное значение. Другими словами, $g(y) \geq g(\tilde{y}) = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$, $y \in X$. То, что $\min_{x \in \mathbb{R}^n} f(x)$ существует и достигается в единственной точке $x^* \in \mathbb{R}^n$ следует из неравенства (7.1) и строгой выпуклости функции f . Так как $f(y) \geq g(y)$, то выполнено неравенство (7.2).

В силу (7.2) имеем

$$\begin{aligned} 0 &\geq f(x^*) - f(x) \geq \nabla f(x)^\top(x^* - x) + \frac{m}{2}\|x^* - x\|^2 \geq \\ &\geq -\|\nabla f(x)\|\|x^* - x\| + \frac{m}{2}\|x^* - x\|^2. \end{aligned}$$

Из последнего, в частности, следует неравенство (7.3). \triangleleft

Лемма 7.2. Пусть $f \in C^2(\mathbb{R}^n)$ и нашлось такое положительное число M , что $\nabla^2 f(x) \preceq MI$. Тогда для всех $x, y \in \mathbb{R}^n$ справедливы неравенства:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^\top(y - x) + \frac{M}{2}\|y - x\|^2, \\ \inf_{y \in X} f(y) &\leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2. \end{aligned} \tag{7.4}$$

\triangleright Доказательство аналогично доказательству леммы 7.1. \triangleleft

Наконец, перейдём к анализу сходимости метода градиентного спуска.

Теорема 7.1. Пусть $f \in C^2(\mathbb{R}^n)$ и существуют такие положительные числа $m, M > 0$, что $mI \preceq \nabla^2 f(x) \preceq MI$, $x \in \mathbb{R}^n$. Обозначим $x^* = \arg \min_{x \in \mathbb{R}^n} f(x)$, а через $(x^k)_{k \in \mathbb{N}_0}$ — последовательность, построенную методом градиентного спуска. Тогда $f(x^k) - f(x^*) \leq c^k(f(x^0) - f(x^*))$, где $c = 1 - m/M$, если последовательность $(t^k)_{k \in \mathbb{N}_0}$ строится согласно процедуре ELS, и $c = 1 - 2\alpha m \min(1, \beta/M)$, если последовательность $(t^k)_{k \in \mathbb{N}_0}$ строится согласно процедуре BLS.

\triangleright Зафиксируем точку x и пусть $g(t) = f(x - t\nabla f(x))$. Из неравенства (7.4) следует, что $g(t) \leq f(x) + \left(\frac{Mt^2}{2} - t\right)\|\nabla f(x)\|^2$. Пусть t_E — размер шага, полученный согласно ELS. Тогда

$$g(t_E) \leq g\left(\frac{1}{M}\right) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2$$

(мы сравниваем со значением функции g в точке M^{-1} , так как функция $Mt^2/2 - t$ принимает там минимальное значение). Из (7.2) следует, что

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)).$$

Таким образом, $f(x - t_E \nabla f(x)) - f(x^*) \leq (1 - m/M)(f(x) - f(x^*))$.

Пусть t_B — размер шага, который полученный согласно процедуре BLS. Покажем, что $t_B = 1$ или $t_B \geq \beta/M$. Действительно, при $0 \leq t \leq M^{-1}$ справедлива цепочка неравенств

$$g(t) \leq f(x) - t \|\nabla f(x)\|^2 + \frac{Mt^2}{2} \|\nabla f(x)\|^2 \leq f(x) - \frac{t}{2} \|\nabla f(x)\|^2 \leq f(x) - \alpha t \|\nabla f(x)\|^2.$$

Если $t_B = 1$, то $f(x - t_B \nabla f(x)) \leq f(x) - \alpha \|\nabla f(x)\|^2$, а если $t_B \geq \beta/M$, то верно $f(x - t_B \nabla f(x)) \leq f(x) - \frac{\alpha\beta}{M} \|\nabla f(x)\|^2$. Таким образом,

$$f(x - t_B \nabla f(x)) - f(x^*) \leq f(x) - f(x^*) - \alpha \min\left(1, \frac{\beta}{M}\right) \|\nabla f(x)\|^2.$$

А значит, в силу неравенства (7.2) имеем

$$f(x - t_B \nabla f(x)) \leq (1 - 2\alpha m \min(1, \beta/M))(f(x) - f(x^*)). \quad \triangleleft$$

Предположение о том, что $\nabla^2 f(x) \succeq mI$, $x \in \mathbb{R}^n$, является достаточно сильным. Оказывается, что сходимость имеет место и без этого предположения. Мы не будем приводить формулировку и доказательство этого результата, так как он будет следовать из анализа, проведённого в §7 при обсуждении метода проекций градиента.

В зависимости от максимального порядка смешанных производных целевой функции f , которые участвуют в методе оптимизации, различают прямые методы поиска (или метод нулевого порядка), методы первого порядка, методы второго порядка и т.д. В прямых методах поиска используется информация только о самой функции и не используется информация о её производных. Как правило, такие методы применяются, когда аналитическое представление функции f не известно. Методы первого порядка при поиске решения используют информацию как о самой функции, так и о её производных первого порядка. Рассмотренный метод градиентного спуска является методом первого порядка. Далее будет рассмотрен метод Ньютона, который является методом второго порядка, так как в этом методе используются информация о самой функции и о её производных первого и второго порядков.

Пусть $f \in C^2(\mathbb{R}^n)$ — выпуклая функция, гессиан $\nabla^2 f(x)$ которой обратим в каждой точке $x \in \mathbb{R}^n$. Из последнего, в частности, следует, что $\nabla^2 f(x)$ — положительно определённая матрица для любого $x \in \mathbb{R}^n$. Выпишем формулу Тейлора второго порядка для функции f в окрестности некоторой фиксированной точки x :

$$f(x + v) = f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v + o(\|v\|^2).$$

Функция $h(v) = f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v$ принимает минимальное значение в точке $v^* = -\nabla^2 f(x)^{-1} \nabla f(x)$. Направление $\Delta x = v^*$ называется шагом Ньютона.

Минимальное значение функции h равно

$$h(\Delta x) = f(x) - \frac{1}{2} \nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x) = f(x) - \frac{1}{2} \lambda(x)^2 \approx f(x + \Delta x),$$

где $\lambda(x) \stackrel{\text{def}}{=} (\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$. Непосредственно проверяется, что

$$\lambda(x)^2 = -\nabla f(x)^\top \Delta x = \Delta x^\top \nabla^2 f(x) \Delta x.$$

Метод спуска, в котором направление спуска равно шагу Ньютона, называется

Алгоритм 7.3. Метод Ньютона

Input: начальное приближение $x \in \mathbb{R}^n$, $\varepsilon > 0$

Output: приближённое решение

```

1 repeat
2    $\Delta x \leftarrow -\nabla^2 f(x)^{-1} \nabla f(x)$ 
3    $\lambda(x)^2 \leftarrow \nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)$ 
4   Определить размер шага  $t$  согласно BLS
5    $x \leftarrow x + t\Delta x$ 
6 until  $\lambda(x)^2/2 > \varepsilon$ 
7 return x
```

методом Ньютона. Для определённости будем считать, что размер шага $t = t_B$ в этом методе выбирается согласно BLS.

Для анализа метода Ньютона докажем вспомогательные леммы.

Лемма 7.3. Если для выпуклой функции $f \in C^2(\mathbb{R}^n)$ существуют такие числа m , $M > 0$, что $mI \preceq \nabla^2 f(x) \preceq MI$, то $f(x + t_B \Delta x) - f(x) \leq -\frac{\alpha \beta m}{M^2} \|\nabla f(x)\|^2$.

▷ Согласно условию леммы справедливы неравенства

$$\begin{aligned} \lambda(x)^2 &= \Delta x^\top \nabla^2 f(x) \Delta x \geq m \|\Delta x\|^2, \\ \lambda(x)^2 &= \nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x) \geq \frac{\|\nabla f(x)\|^2}{M}. \end{aligned}$$

Из неравенства (7.4) следует, что $f(x + t\Delta x) \leq f(x) - t\lambda(x)^2 + \frac{M}{2m} t^2 \lambda(x)^2$. Полагая в последнем неравенстве $t = m/M$, получаем, что

$$f(x + t\Delta x) \leq f(x) - \frac{m}{2M} \lambda(x)^2 \leq f(x) - \alpha t \lambda(x)^2.$$

Следовательно, $t_B \geq \beta m/M$. Таким образом,

$$f(x + t_B \Delta x) - f(x) \leq -\alpha t_B \lambda(x)^2 \leq -\alpha \beta \frac{m}{M} \lambda(x)^2 \leq -\frac{\alpha \beta m}{M^2} \|\nabla f(x)\|^2. \quad \triangleleft$$

Лемма 7.4. Пусть для функции $f \in C^2(\mathbb{R}^n)$ и чисел $m, M, L > 0$ верно, что

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \text{и} \quad \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad x, y \in \mathbb{R}^n.$$

Тогда $\|\nabla f(x + \Delta x)\| \leq \frac{L}{2m^2} \|\nabla f(x)\|^2$ и если $\|\nabla f(x)\| \leq 3(1 - 2\alpha)\frac{m^2}{L}$, то $t_B = 1$.

▷ Докажем первую часть утверждения леммы.

$$\begin{aligned} \|\nabla f(x + \Delta x)\| &= \|\nabla f(x + \Delta x) - \nabla f(x) - \nabla^2 f(x)\Delta x\| = \\ &= \left\| \int_0^1 (\nabla^2 f(x + t\Delta x) - \nabla^2 f(x))\Delta x dt \right\| \leq \frac{L}{2} \|\Delta x\|^2 = \\ &= \frac{L}{2} \|\nabla^2 f(x)^{-1} \nabla f(x)\|^2 \leq \frac{L}{2m^2} \|\nabla f(x)\|^2. \end{aligned}$$

Пусть $g(t) \stackrel{\text{def}}{=} f(x + t\Delta x)$, тогда получаем $g''(t) = \Delta x^\top \nabla^2 f(x + t\Delta x) \Delta x$. В силу условия леммы справедливы неравенства:

$$|g''(t) - g''(0)| = |\Delta x^\top (\nabla^2 f(x + t\Delta x) - \nabla^2 f(x)) \Delta x| \leq tL \|\Delta x\|^3$$

и $g''(0) = \lambda(x)^2 \geq m \|\Delta x\|^2$. Тогда $g''(t) \leq g''(0) + tL \|\Delta x\|^3 \leq \lambda(x)^2 + \frac{tL}{m^{3/2}} \lambda(x)^3$. Проинтегрировав дважды последнюю цепочку неравенств, учитывая равенство $g'(0) = -\lambda(x)^2$, и положив $t = 1$, получаем $f(x + \Delta x) \leq f(x) - \frac{1}{2} \lambda(x)^2 + \frac{L}{6m^{3/2}} \lambda(x)^3$.

Так как $\|\nabla f(x)\| \leq 3(1 - 2\alpha)\frac{m^2}{L}$, то $\lambda(x) \leq 3(1 - 2\alpha)\frac{m^{3/2}}{L}$. Следовательно,

$$f(x + \Delta x) \leq f(x) - \lambda(x)^2 \left(\frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}} \right) \leq f(x) - \alpha \lambda(x)^2 = f(x) + \alpha \nabla f(x)^\top \Delta x.$$

Таким образом, $t_B = 1$. ◁

Теорема 7.2. Пусть для функции $f \in C^2(\mathbb{R}^n)$ и чисел $m, M, L > 0$ верно, что $mI \preceq \nabla^2 f(x) \preceq MI$ и $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$, $x, y \in \mathbb{R}^n$. Обозначим $x^* = \arg \min_{x \in \mathbb{R}^n} f(x)$, а через $(x^k)_{k \in \mathbb{N}_0}$ — последовательность, построенную методом Ньютона. Тогда для любого $\varepsilon > 0$ верно неравенство $f(x^k) - f(x^*) < \varepsilon$ при

$$k \geq N_\varepsilon \stackrel{\text{def}}{=} \frac{f(x^0) - f(x^*)}{\gamma} + \log_2 \log_2 \left(\frac{m^3}{L^2 \varepsilon} \right), \quad \text{где} \quad \gamma = \gamma(m, M, L).$$

▷ Пусть $\eta \stackrel{\text{def}}{=} \min(1, 3(1 - 2\alpha))\frac{m^2}{L}$. Если $\|\nabla f(x)\| \leq \eta$, то из леммы 7.4, в частности, следует, что $\|\nabla f(x + \Delta x)\| \leq \frac{\eta}{2}$. Действительно, так как $\frac{L}{m^2} \leq \eta^{-1}$, то

$\|\nabla f(x + \Delta x)\| \leq \frac{L}{2m^2} \|\nabla f(x)\|^2 \leq \frac{\eta}{2}$. Через s обозначим наименьшее натуральное число, такое что $\|\nabla f(x^s)\| \leq \eta$. Из леммы 7.3 следует, что $s \leq \frac{f(x^0) - f(x^*)}{\gamma}$, где $\gamma \stackrel{\text{def}}{=} \alpha\beta\eta^2 \frac{m}{M^2}$. Тогда при $k > s$ справедливо неравенство $\|\nabla f(x^k)\| < \eta$, а значит,

$$\frac{L}{2m^2} \|\nabla f(x^k)\| \leq \left(\frac{L}{2m^2} \|\nabla f(x^{k-1})\| \right)^2 \leq \left(\frac{L}{2m^2} \|\nabla f(x^s)\| \right)^{2^{k-s}} \leq \left(\frac{1}{2} \right)^{2^{k-s}}.$$

Из неравенства (7.2) следует, что $f(x^k) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x^k)\| \leq \frac{m^3}{L^2} 2^{-2^{k-s}} < \varepsilon$. \triangleleft

Упражнения

19. (Метод сопряжённых направлений) Пусть A — положительно определённая матрица и $f(x) = x^T A x - 2b^T x + c$. Векторы (направления) p и q назовём A -сопряжёнными, если $p^T A q = 0$. Для данного начального приближения x^0 и последовательности ненулевых сопряжённых направлений $(p^k)_{k=0}^{n-1}$ определим последовательность точек $(x^k)_{k=1}^n$ равенством $x^{k+1} = x^k + t^k p^k$, где $t^k = \arg \min_{t \geq 0} f(x^k + t p^k)$. Докажите, что

- (a) $t^k = -(p^k, \nabla f(x^k)) / (p^k, A p^k)$;
- (b) вектор $\nabla f(x^{k+1})$ ортогонален направлениям p^0, p^1, \dots, p^k ;
- (c) $\nabla f(x^n) = 0$, а значит, x^n — минимум функции $f(x)$;
- (d) последовательность направлений $(p^k)_{k=0}^{n-1}$ может быть определена как

$$\begin{cases} p^0 = -\nabla f(x^0), \\ p^k = -\nabla f(x^k) + \frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^{k-1})\|^2} p^{k-1}, \quad 1 \leq k \leq n. \end{cases}$$

Литература

1. Boyd S., Vandenberghe L. *Convex Optimization* – Cambridge University Press, 2004. – 730 с.