Методы оптимизации, ФКН ВШЭ, зима 2017

Семинар 4: Методы градиентного спуска и Ньютона

31 января 2017 г.

Задача 1. Пусть $\beta \geq 0$ и $f : \mathbb{R}^n \to \mathbb{R}$ — функция

$$f(x) := \frac{1}{2+\beta}\|x\|_2^{2+\beta}.$$

Рассмотрим градиентный спуск с постоянной длиной шага $\alpha > 0$ для минимизации функции f, запущенный из точки $x_0 \in \mathbb{R}^n \setminus \{0\}$. Определите, для каких длин шага α метод будет сходиться к минимуму $x^* = 0$. Какова при этом будет скорость сходимости (линейная/сублинейная)?

Решение. Найдем градиент:

$$\nabla f(x) = \|x\|_2^\beta x.$$

Значит, итерация метода записывается следующим образом:

$$x_{k+1} = x_k - \alpha\|x_k\|_2^\beta x_k = (1 - \alpha\|x_k\|_2^\beta)x_k.$$

Обозначим $z_k := \|x_k\|_2$. Тогда

$$z_{k+1} = |1 - \alpha z_k^\beta| z_k. \tag{1}$$

Рассмотрим три возможные ситуации:

1. Длина шага α очень большая: $\alpha > 2z_0^{-\beta}$. В этом случае нетрудно понять, что z_{k+1} будет расходиться как минимум со скоростью геометрической прогрессии.

2. Длина шага $\alpha = 2z_0^{-\beta}$. В этом случае метод будет стоять на месте: $x_0 = x_1 = \dots$.

3. Длина шага $\alpha < 2z_0^{-\beta}$. Тогда будет монотонное убывание расстояний: $z_{k+1} \leq z_k$. Поскольку последовательность z_k ограничена снизу, то она имеет предел. Покажем, что этот предел в точности равен нулю. Обозначим предел через a. Пусть $a > 0$. Тогда, переходя в (1) к пределу, получим

   $$a = |1 - \alpha a^\beta| a.$$

   Поскольку $a \neq 0$, то отсюда

   $$|1 - \alpha a^\beta| = 1.$$

   Поскольку $\alpha > 0$ и $a > 0$, то

   $$\alpha a^\beta = 2 \qquad \Leftrightarrow \qquad \alpha = 2a^{-\beta}.$$

   Заметим, что это невозможно, поскольку $2z_0^{-\beta} < \alpha$ и z_k монотонно убывает. Итак, $a = 0$.

Заметим, что сходимость есть только в последнем случае, когда $\alpha < 2z_0^{-\beta}$. Скорость сходимости будет сублинейной, поскольку

$$\frac{z_{k+1}}{z_k} = |1 - \alpha z_k^\beta| \to 1.$$

(Аналогичная скорость сходимости будет и по функции.)

Задача 2. Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция

$$f(x) = \frac{2}{3}\|x\|^{3/2}.$$

Заметим, что эта функция является непрерывно дифференцируемой, однако не имеет липшицев градиент (в окрестности нуля). Рассмотрите поведение градиентного спуска на этой функции с постоянным шагом $\alpha > 0$, запущенного из точки $x_0 \in \mathbb{R}^n \setminus \{0\}$.

Решение. Посчитаем градиент:

$$\nabla f(x) = \|x\|_2^{-1/2} x.$$

Тогда итерация градиентного спуска:

$$x_{k+1} = x_k - \alpha\|x_k\|_2^{-1/2} x_k = (1 - \alpha\|x_k\|_2^{-1/2}) x_k.$$

Перейдем к $z_k := \|x_k\|_2$:

$$z_{k+1} = |1 - \alpha z_k^{-1/2}| z_k.$$

Отсюда видно, что если $\alpha < 2z_k^{1/2}$, то $z_1 < z_0$. Однако, это означает, что $z_2 > z_2$. А, значит, $z_3 < z_2$. Получается коллебательный процесс. Таким образом, «если повезет», то метод сойдется, а иначе будет бесконечно коллебаться. (Действительно, метод остановится, только если z_k в какой-то момент в точности станет равно нулю.)

Задача 3 (Градиентный спуск и седловые точки)**.** Пусть $A \in \mathbb{S}^n$ — симметричная невырожденная матрица, имеющая хотя бы одно строго положительное и хотя бы одно строго отрицательное значение ($n \geq 2$). Рассмотрите градиентный спуск с постоянной длиной шага $\alpha > 0$, запущенный из точки $x_0 \in \mathbb{R}^n \setminus \{0\}$. Определите, как будет вести себя метод в зависимости от α и x_0.

Решение. В этом случае единственная стационарная точка $x^* = 0$ является седловой.

Градиентный спуск:

$$x_{k+1} = x_k - \alpha A x_k = (I_n - \alpha A) x_k.$$

Значит,

$$x_k = (I_n - \alpha A)^k x_0.$$

Рассмотрим спектральное разложение $A = Q\Lambda Q^T$, где $\Lambda := \mathrm{Diag}\{\lambda_1, \ldots, \lambda_n\}$ и $\lambda_1 \geq \cdots \geq \lambda_s > 0 > \lambda_{s+1} \geq \ldots \lambda_n$. Обозначим $\tilde{x}_k := Q^T x_k$. Тогда

$$\tilde{x}_k = (I_n - \alpha\Lambda)^k \tilde{x}_0.$$

В координатах:

$$\tilde{x}_{k,i} = (1 - \alpha\lambda_i)^k \tilde{x}_{0,i}.$$

Заметим, что для $\lambda_i < 0$ будет расходимость, кроме тех случаев, когда $\tilde{x}_{0,i} = 0$. То есть если x_0 *не* лежит в подпространстве, отвечающему отрицательным собственным значениям, то x_k расходится. Если же x_0 лежит в соответсвующем подространстве, то сходимость к седловой точке $x^* = 0$.

Задача 4. Примените классический метод Ньютона для минимизации функции

$$f(x) := \frac{1}{3}\|x\|_2^3, \qquad x \in \mathbb{R}^n.$$

Выпишите в явном виде как выражается k-я точка метода $x^{(k)}$ через начальную точку $x^{(0)} \neq 0$. Какова скорость сходимости последовательности $(x^{(k)})_{k=0}^{\infty}$? Соотнесите полученный результат с общей теоремой о скорости сходимости метода Ньютона.

Решение. Градиент и гессиан функции f мы уже считали на предыдущем семинаре:

$$\nabla f(x) = \|x\|_2 x, \qquad \nabla^2 f(x) = \|x\|_2 I_n + \|x\|_2^{-1} x x^T.$$

(Напомним, что приведенную формулу для $\nabla^2 f(x)$ в точке $x = 0$ надо понимать как 0.)

Выпишем как будет выглядеть итерация метода Ньютона для функции f:

$$x^+ = x - [\nabla^2 f(x)]^{-1} \nabla f(x) = x - (\|x\|_2 I_n + \|x\|_2^{-1} x x^T)^{-1} (\|x\|_2 x).$$

Вычислим отдельно $u := (\|x\|_2 I_n + \|x\|_2^{-1} x x^T)^{-1} (\|x\|_2 x)$. Для этого можно либо воспользоваться формулой Шермана-Моррисона, либо явно решить соответствующую линейную систему.

Решим явно следующую систему линейных уравнений относительно u:

$$(\|x\|_2 I_n + \|x\|_2^{-1} x x^T) u = \|x\|_2 x.$$

Раскрывая скобки, получаем

$$\|x\|_2 u + \langle x, u \rangle \|x\|_2^{-1} x = \|x\|_2 x.$$

Отсюда

$$u = x - \langle x, u \rangle \|x\|_2^{-2} x.$$

Осталось найти $\langle x, u \rangle$. Для этого умножим слева обе части полученного уравнения скалярно на x:

$$\langle x, u \rangle = \|x\|_2^2 - \langle x, u \rangle \qquad \Leftrightarrow \qquad \langle x, u \rangle = \frac{1}{2} \|x\|_2^2.$$

Значит,

$$u = x - \frac{1}{2} x = \frac{1}{2} x.$$

В итоге, $u = (1/2) x$, и итерация метода Ньютона принимает следующий вид:

$$x^+ = x - \frac{1}{2} x = \frac{1}{2} x.$$

Отсюда получаем, что

$$\boxed{x^{(k)} = \left(\frac{1}{2} \right)^k x^{(0)}}.$$

Это означает, что метод будет иметь линейную скорость сходимости с константой $1/2$ для любого $x^{(0)} \neq 0$. Почему скорость сходимости не квадратичная? Потому что теорема про метод Ньютона требует, чтобы гессиан в точке оптимума был невырожденным. В данном случае точка оптимума $x^* = 0$, и гессиан в этой точке — это нулевая матрица, которая, естественно, является вырожденной.

В предыдущей задаче проблема была с тем, что функция не сильно выпуклая. Исправим эту проблему с помощью добавления квадратичной добавки.

Задача 5. Повторите аналогичные рассуждения для функции

$$f(x) := \frac{1}{3} \|x\|_2^3 + \frac{1}{2} \|x\|_2^2.$$

Решение. Вычислим градиент и гессиан функции f:

$$\begin{aligned} \nabla f(x) &= \|x\|_2 x + x = (1 + \|x\|_2) x, \\ \nabla^2 f(x) &= \frac{x x^T}{\|x\|_2} + \|x\|_2 I_n + I_n = \frac{x x^T}{\|x\|_2} + (1 + \|x\|_2) I_n. \end{aligned}$$

Направление d в методе Ньютона находится из системы $\nabla^2 f(x)d = -\nabla f(x)$. Обозначим для краткости $g := \nabla f(x)$, и найдем в явном виде d.

Нам нужно решить следующую систему линейных уравнений относительно d:

$$\left(\frac{xx^T}{\|x\|_2} + (1 + \|x\|_2)I_n\right) d = -g.$$

Раскрывая скобки, получаем

$$\frac{\langle x, d\rangle}{\|x\|_2} x + (1 + \|x\|_2)d = -g.$$

Отсюда

$$d = \frac{-g - \frac{\langle x,d\rangle}{\|x\|_2} x}{1 + \|x\|_2}.$$

Осталось найти $\langle x, d\rangle$. Для этого умножим слева обе части полученного уравнения скалярно на x:

$$\langle x, d\rangle = \frac{-\langle x, g\rangle - \frac{\langle x,d\rangle}{\|x\|_2}\langle x, x\rangle}{1 + \|x\|_2} = \frac{-\langle x, g\rangle - \langle x, d\rangle\|x\|_2}{1 + \|x\|_2} \qquad \Leftrightarrow \qquad \langle x, d\rangle = \frac{-\langle x, g\rangle}{1 + 2\|x\|_2}.$$

Значит,

$$d = \frac{-g + \frac{\langle x,g\rangle}{\|x\|_2(1+2\|x\|_2)} x}{1 + \|x\|_2}.$$

Заметим, что для произвольного вектора g формула выглядит довольно громоздкой. Тем не менее, надо помнить, что мы работаем не с произвольным вектором g, а именно с $g = \nabla f(x) = (1 + \|x\|_2)x$. Подставляя, получаем

$$d = \frac{-(1 + \|x\|_2)x + \frac{(1+\|x\|_2)\|x\|_2^2}{\|x\|_2(1+2\|x\|_2)}}{1 + \|x\|_2} = -x + \frac{\|x\|_2}{1 + 2\|x\|_2} x.$$

Таким образом, итерация метода Ньютона имеет вид:

$$x_+ = \frac{\|x\|_2}{1 + 2\|x\|_2} x.$$

Чтобы оценить скорость сходимости $(x_k)_{k=0}^\infty$, перейдем к нормам:

$$\|x_+\|_2 = \frac{\|x\|_2^2}{1 + 2\|x\|_2}.$$

Таким образом, получаем квадратичную скорость сходимости:

$$\|x_+\|_2 \leq \|x\|_2^2.$$

Тем не менее, сходимость будет для любой начальной точки x_0. Для больших $\|x\|_2$ имеем

$$\frac{\|x\|_2}{1 + 2\|x\|_2} \approx \frac{1}{2}.$$

Так что в первый момент времени уменьшение составляет $1/2$, а затем этот коэффициент монотонно уменьшается. Получаем глобальную сверхлинейную сходимость.

Задача 6. Выпишите в явном виде итерацию метода Ньютона для минимизации функции $f : \mathbb{S}^n_{++} \to \mathbb{R}$, заданной по формуле

$$f(X) := \operatorname{Tr}(CX) - \ln \operatorname{Det}(X).$$

Решение. Здесь нужно вспомнить, что на самом деле делает метод Ньютона — он минимизирует квадратичную модель функции. Вычислим производные функции f и запишем ее квадратичную модель:

$$Df(X)[H] = \operatorname{Tr}(CH) - \operatorname{Tr}(X^{-1}H),$$
$$D^2 f(X)[H, H] = \operatorname{Tr}(X^{-1}HX^{-1}H).$$

Тогда квадратичная модель функции имеет вид:

$$f(x+h) \approx f(x) + Df(X)[H] + \frac{1}{2}D^2 f(X)[H, H] = f(x) + \operatorname{Tr}([C - X^{-1}]H) + \frac{1}{2}\operatorname{Tr}(X^{-1}HX^{-1}H).$$

Наша цель — найти минимум этой модели по $H \in \mathbb{S}^n$. Для этого посчитаем градиент и приравняем нулю (функция выпуклая):

$$C - X^{-1} + X^{-1}HX^{-1} = 0.$$

Отсюда

$$H = X(X^{-1} - C)X = X - XCX.$$

Значит, итерация метода Ньютона имеет вид

$$X_{k+1} = X_k + H = 2X_k - X_k C X_k.$$

В линейной алгебре этот метод называется *методом Ньютона–Шульца.*