Методы оптимизации в машинном обучении, ШАД, весна 2017

Семинар 4: Метод Ньютона

6 марта 2017 г.

Внимание: Предварительная версия.

Задача 1. 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

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

Отсюда

$$u = x - (x^T u) \|x\|_2^{-2} x.$$

Осталось найти $x^T u$. Для этого умножим слева обе части полученного уравнения на x^T:

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

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 2. 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:

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

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

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

Отсюда

$$d = \frac{-g - \frac{x^T d}{\|x\|_2} x}{1 + \|x\|_2}.$$

Осталось найти $x^T d$. Для этого умножим слева обе части полученного уравнения на x^T:

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

Значит,

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

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 3. 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*.
