СХОДИМОСТЬ И ВЫЧИСЛИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ: МЕТОД ГРАДИЕНТНОГО СПУСКА В СРАВНЕНИИ С МЕТОДОМ НЬЮТОНА

Кахно А.А., Ступаков М.В., студенты

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Степанова Т.С. – канд. физ.-мат. наук, доцент

Аннотация. Работа посвящена сравнительному анализу методов численной оптимизации – метода градиентного спуска и метода Ньютона – с точки зрения их сходимости и вычислительной эффективности при решении задач минимизации функций.

Ключевые слова. Метод градиентного спуска, метод Ньютона, сходимость, оптимизация, итерация, градиент, матрица Гессе.

Оптимизация функций является важной задачей в математике и машинном обучении. Нейронные сети обучаются с помощью оптимизации, ведь благодаря ей мы можем найти наилучшее решение из всех возможных. В машинном обучении задачи оптимизации обычно решаются итерационными методами. Метод градиентного спуска и метод Ньютона – одни из наиболее известных среди них. Первый находит минимум за счёт движения в направлении наибольшего убывания функции, второй использует информацию о кривизне для более точного приближения. Рассмотрим подробнее, почему при обучении нейронных сетей используется именно метод градиентного спуска, а не метод Ньютона.

Метод градиентного спуска относится к итерационным методам первого порядка. Это значит, что для его использования необходима информация о первых производных функции. Метод основан на использовании градиента. Градиент – это вектор, который показывает направление и скорость наибольшего увеличения функции. В математике градиентом называют вектор, координаты которого являются частными производными функции, т.е., если задана функция $f(t_1, t_2, \dots, t_n)$, то градиент рассчитывается по следующей формуле:

$$\nabla f = \left(\frac{\partial f}{\partial t_1}, \frac{\partial f}{\partial t_2}, \dots, \frac{\partial f}{\partial t_n}\right). \quad (1)$$

Основная идея метода градиентного спуска заключается в том, что необходимо осуществлять переход из текущей точки в направлении, противоположном градиенту, т.е. в направлении наибольшего убывания функции. Итерационный шаг для метода градиентного спуска может быть представлен следующим образом:

$$\overline{x_{n+1}} = \overline{x_n} - \alpha \nabla f(\overline{x_n}), \quad (2)$$

где $\overline{x_n}$ – текущее приближение, α – размер шага, $\nabla f(\overline{x_n})$ – градиент функции в точке $\overline{x_n}$.

Метод Ньютона относится к методам второго порядка. Для его использования необходима информация как о первых, так и о вторых производных функции. В основе метода лежит квадратичная аппроксимация функции в окрестности текущей точки, т.е. приближение функции квадратным многочленом (параболой) вокруг заданной точки.

Итерационная формула для метода Ньютона:

$$\overline{x_{n+1}} = \overline{x_n} - [\nabla^2 f(\overline{x_n})]^{-1} \nabla f(\overline{x_n}), \quad (3)$$

где $\nabla f(\overline{x_n})$ – градиент функции (вектор первых производных) в точке $\overline{x_n}$, а $\nabla^2 f(\overline{x_n})$ – матрица Гессе (матрица вторых производных).

Проведем сравнительную характеристику методов.

При нахождении минимума функции с использованием метода градиентного спуска на каждом шаге требуется вычисление только градиента. Следовательно, стоимость одной итерации O(n) по времени и O(n) по памяти.

Метод Ньютона кроме градиента требует вычисления матрицы Гессе и матрицы, обратной ей. Следовательно, стоимость одной итерации составляет $O(n^2)$ на формирование гессиана и $O(n^3)$ на решение линейной системы, память – $O(n^2)$. Таким образом, при сопоставимом числе итераций метод Ньютона оказывается значительно дороже в вычислительном отношении.

Для того, чтобы можно было применить метод градиентного спуска, функция должна иметь непрерывный градиент, т.е. быть дифференцируемой. Для гарантии сходимости к глобальному минимуму функция должна быть ограничена снизу и являться выпуклой, что исключает наличие локальных минимумов.

Для применения метода Ньютона функция должна быть дважды непрерывно дифференцируемой, то есть должны существовать как первые, так и вторые производные, матрица Гессе должна быть обратима, то есть быть невырожденной ($det(\nabla^2 f(\overline{x_n})) \neq 0$). Важно, чтобы функция обладала достаточной гладкостью.

Сходимость метода градиентного спуска является линейной. При выборе шага важно выбрать подходящий. При малом шаге алгоритм сходится к минимуму с высокой точностью, но этот процесс происходит медленно. Для достижения минимума при малом шаге требуется значительное количество вычислительных ресурсов и времени. Большой шаг – обучение идет быстро, но есть риск пропустить минимум.

Метод Ньютона имеет квадратичную сходимость при выполнении вышеперечисленных условий, что означает, что количество значащих цифр, которые можно считать верными, примерно удваивается на каждой итерации.

Рассмотрим минимизацию функции $f(x) = x^4 - x^2$ с начальным приближением $x_0 = 1{,}5$. Результаты вычислений представлены в таблице 1.

Таблица 1 – Сравнение итераций методов для функции $f(x) = x^4 - x^2$

№ итерации	Градиентный спуск ($\alpha = 0{,}05$), x_n	Метод Ньютона, x_n
0	1,5000	1,5000
1	0,9750	1,0800
2	0,8871	0,8400
3	0,8362	0,7331
4	0,8028	0,7084
5	0,7796	0,7071
10	0,7280	сходимость достигнута
20	0,7092	
30	0,7073	
34	0,7072	

Метод Ньютона достиг точности $\varepsilon = 10^{-4}$ за 5 итераций, а градиентному спуску с шагом $\alpha = 0{,}05$ для аналогичного результата потребовалось 34 итерации. Данный пример наглядно демонстрирует разницу в скорости сходимости двух методов, однако стоимость каждой итерации метода Ньютона в многомерном случае существенно выше.

Следует отметить, что сходимость обоих методов существенно зависит от выбора начального приближения. При этом у метода Ньютона она является локальной, тогда как метод градиентного спуска при выполнении условий выпуклости обладает глобальной сходимостью.

Сравнительная характеристика методов приведена в таблице 2.

Таблица 2 – Сравнение по критериям метода градиентного спуска и метода Ньютона

Критерий	Метод градиентного спуска	Метод Ньютона
Информация	Градиент функции в текущей точке	Матрица Гессе, градиент функции в текущей точке
Сходимость	Линейная	Квадратичная
Стоимость итерации	Низкая, необходимо лишь заново вычислять градиент функции	Высокая, необходимо вычислять заново градиент функции, матрицу Гессе, решить СЛАУ
Сложность	O(n)	$O(n^3)$
Предпочтительная размерность	Большая (1000+)	Наиболее эффективен в задачах малой размерности (2-10)
Число итераций	Большое	Малое
Лучшая область применения	Задачи с большим числом параметров	Задачи с малым числом параметров, но в которых требуется большая точность

Таким образом, при соблюдении рассмотренных условий метод Ньютона сходится к точке минимума функции с квадратичной скоростью, достаточной для большинства практических задач. Однако необходимость расчета и обращения матрицы вторых производных делает его неприменимым для задач высокой размерности. Современные нейронные сети содержат от миллионов до миллиардов параметров, что делает вычисление и хранение матрицы Гессе физически невозможным и окончательно исключает прямое применение метода Ньютона в таких задачах. В свою очередь, обладающий низкой стоимостью итераций метод градиентного спуска имеет и низкую скорость сходимости. Однако на практике при обучении нейронных сетей чаще используются усовершенствованные методы на основе градиентного спуска.

Список использованных источников:

1. Поляк, Б.Т. Введение в оптимизацию / Б.Т. Поляк. – М. : Наука, 1983. – С. 29–39.
2. Нестеров, Ю.Е. Введение в выпуклую оптимизацию / Ю.Е. Нестеров. – М. : МЦНМО, 2010. – С. 47–65.