

Многомерная оптимизация

Звонарев Никита

Слушатели: 4-й курс, специальность СМ-СМ

Санкт-Петербургский Государственный Университет
Кафедра Статистического моделирования



Санкт-Петербург – 2023 г.

$\mathcal{D} \subset \mathbb{R}^k$ — замкнутое связное множество, $f(x)$ — целевая функция,
 $f : \mathcal{D} \rightarrow \mathbb{R}$.

Задача:

$$x^* = \arg \min_{x \in \mathcal{D}} f(x).$$

Теперь $k > 1$.

Идея: свести задачу к одномерной оптимизации. (Как?)

Подход: итеративный, каждая итерация — в два этапа:

- 1 Поиск направления оптимизации (x_n — текущая точка, направление — p_n (как найти?)).
- 2 Шаг по направлению p_n либо одномерная оптимизация:
 $\alpha_n^* = \arg \min_{0 \leq \alpha \leq 1} f(x_n + \alpha p_n), x_{n+1} = x_n + \alpha_n^* p_n.$

$\mathcal{D} \subset \mathbb{R}^k$ — замкнутое связное множество, $f(x)$ — целевая функция, $f : \mathcal{D} \rightarrow \mathbb{R}$.

Задача:

$$x^* = \arg \min_{x \in \mathcal{D}} f(x).$$

Теперь $k > 1$.

Идея: свести задачу к одномерной оптимизации. (Как?)

Подход: итеративный, каждая итерация — в два этапа:

- 1 Поиск направления оптимизации (x_n — текущая точка, направление — p_n (как найти?)).
- 2 Шаг по направлению p_n либо одномерная оптимизация:
 $\alpha_n^* = \arg \min_{0 \leq \alpha \leq 1} f(x_n + \alpha p_n), x_{n+1} = x_n + \alpha_n^* p_n.$

Он же *coordinate descent*. Самое простое, что можно придумать.

В терминах p_n : $p_n = e_{(n \bmod k)+1}$, $e_i \in \mathbb{R}^k$ — i -й орт.

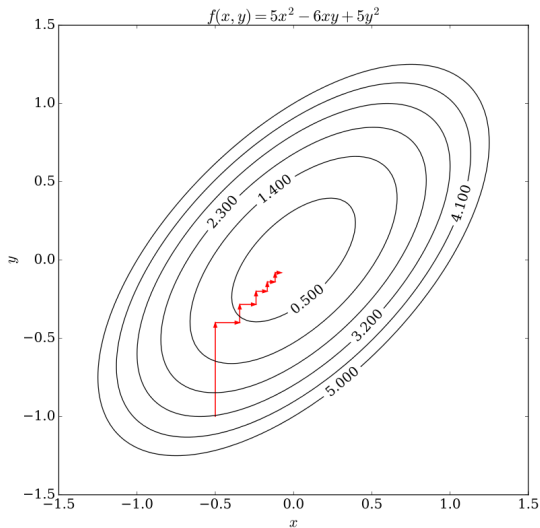
Схема итерации та же:

$$\alpha_n^* = \arg \min_{\alpha \in \mathbb{R}} f(x_n + \alpha p_n), \quad x_{n+1} = x_n + \alpha_n^* p_n.$$

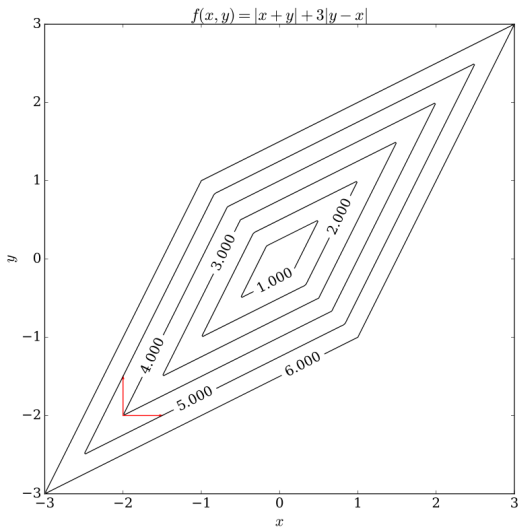
Что выполнено: $f(x_{n+1}) \leq f(x_n)$.

Вопрос: это гарантирует сходимость к локальному минимуму?

Покоординатный спуск: пример, когда есть сходимость



Покоординатный спуск: пример, когда нет сходимости



Ой! Метод может “застрять” там, где нет локального минимума.
На гладких функциях может сходиться медленно.

Получается, в нём нет смысла?

Идею используют. Например, вместо спуска по одной координате разбивают координаты на два набора, $k_1 + k_2 = k$, и чередуют их. На итерации — многомерная оптимизация по подмножеству координат.

Зачем так делают? Когда напрямую $f(x)$ оптимизировать слишком сложно.

Применения: статистика (важный частный случай — *ЕМ-алгоритм*),
...

Он же *gradient descent*. f — дифференцируемая
 p_n — направление *антиградиента*: $p_n = -\nabla f(x_n)$.

Схема итерации та же:

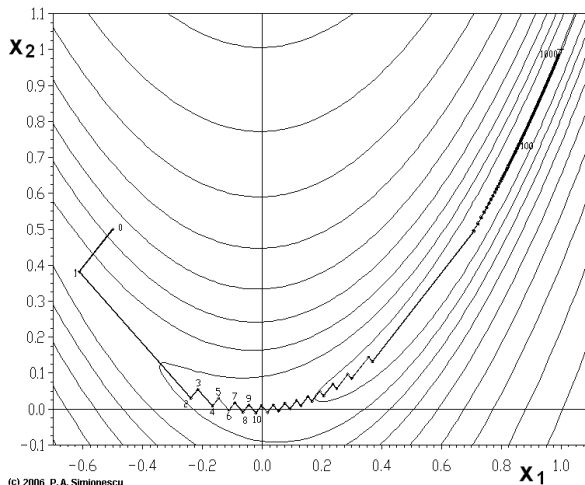
$$\alpha_n^* = \arg \min_{0 \leq \alpha < \infty} f(x_n + \alpha p_n), \quad x_{n+1} = x_n + \alpha_n^* p_n.$$

Можно выбрать такие малые α_n^* , что будет выполнено:
 $f(x_{n+1}) \leq f(x_n)$.

Сходимость выполняется в некоторых специфических случаях
(выпуклость f , липшицевость ∇f , ...).

Градиентный спуск: пример медленной сходимости

Rosenbrock function: $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$.



Градиентный спуск: применения

Применяют, когда более сложные методы слишком дороги по времени/памяти. В этом случае применяют не поиск по направлению, а шаг на определенное расстояние.

В статистике (точнее, машинном обучении) применяют вариант градиентного спуска к целевым функциям вида:

$$f(x) = \frac{1}{N} \sum_1^N f_i(x).$$

f_i — это

Вместо того, чтобы на каждой итерации оптимизировать f , случайным образом разыгрывается индекс i , и вычисляется шаг по антиградиенту $-f'_i$ с угасающим весом. Получили метод стохастического градиента (*Stochastic gradient descent, SGD*).

Модификации: AdaGrad, RMSProp, Adam, ... (2010-е)

Метод сопряжённых градиентов, квадратичная функция

Он же *Conjugated gradient*, CG.

Хорошее название: никаких градиентов в нём нет.

Рассмотрим частный случай: $\arg \min_{x \in \mathbb{R}^k} f(x) = \frac{1}{2}x^T A x - b^T x$, A — симметричная, положительно определенная матрица.

Эквивалентная задача: решение СЛАУ $Ax = b$.

Обозначим $r(x) = Ax - b$.

Рассмотрим систему из n сопряжённых (т.е. ортогональных с A) векторов $p_1, \dots, p_k$: $p_i^T A p_j = 0$, $i \neq j$, $p_i \neq 0$.

Возьмём произвольное x_0 , и найдём решение системы в виде разложения по векторам p_i , то есть, $x_{n+1} = x_n + \alpha_n p_n$. α_k можно найти, взяв минимум по направлению:

$$\alpha_n = -\frac{r_n^T p_n}{p_n^T A p_n}, \quad r_n = r(x_n).$$

Метод сопряжённых градиентов, квадратичная функция

$$x_{n+1} = x_n + \alpha_n p_n.$$

$$\alpha_n = -\frac{r_n^T p_n}{p_n^T A p_n}, \quad r_n = r(x_n).$$

Теорема

Для любого $x_0 \in \mathbb{R}^k$ последовательность x_n сходится к решению x^* в худшем случае за k шагов.

Вопрос: как найти векторы p_n ?

$$p_n = -r_n + \beta_n p_{n-1},$$

$$\beta_n = \frac{r_n^T A p_{n-1}}{p_{n-1}^T A p_{n-1}}.$$

Nocedal, Jorge, Wright, S. Numerical Optimization.

Метод сопряжённых градиентов, алгоритм

$$r_0 = Ax_0 - b, p_0 = -r_0, n = 0.$$

Пока $r_n \neq 0$:

- $\alpha_n = \frac{r_n^T r_n}{p_n^T A p_n}$
- $x_{n+1} = x_n + \alpha_n p_n$
- $r_{n+1} = r_n + \alpha_n A p_n$
- $\beta_{n+1} = \frac{r_{n+1}^T r_{n+1}}{r_n^T r_n}$
- $p_{n+1} = -r_{n+1} + \beta_{n+1} p_n$
- Положить $n = n + 1$

Метод сопряжённых градиентов для произвольной функции

$f_i = f(x_i)$, $\nabla f_i = f'(x_i)$. f — дифференцируемая.

Метод Fletcher-Reeves:

$p_0 = -\nabla f_0$, $n = 0$.

Пока не выполнен критерий остановки:

- $\alpha_n = \arg \min_{0 \leq \alpha < \infty} f(x_n + \alpha p_n)$
- $x_{n+1} = x_n + \alpha_n p_n$
- Вычислить ∇f_{n+1}
- $\beta_{n+1}^{FR} = \frac{\nabla f_{n+1}^T \nabla f_{n+1}}{\nabla f_n^T \nabla f_n}$
- $p_{n+1} = -\nabla f_{n+1} + \beta_{n+1}^{FR} p_n$
- Положить $n = n + 1$

Метод Polak-Ribiere:

$$\beta_{n+1}^{PR} = \frac{\nabla f_{n+1}^T (\nabla f_{n+1} - \nabla f_n)}{\nabla f_n^T \nabla f_n}.$$

Многомерный метод Ньютона

$$p_n = -\nabla^2 f_n^{-1} \nabla f_n.$$

$\nabla^2 f_n$ — матрица Гессе (матрица вторых производных).

Теорема

Предположим, что f дважды дифференцируема, гессиан $\nabla^2 f(x)$ — липшицев в окрестности точки x^* , в которой выполнены достаточные условия локального минимума. Тогда

- 1 Если x_0 достаточно близко к x^* , то последовательность сойдётся к x^* .
- 2 Скорость сходимости x_k к x^* квадратичная.
- 3 $\|\nabla f_k\|$ сходится квадратично к 0.

Достаточное условие строгого локального минимума в точке x^* : $\nabla^2 f_n$ — непрерывна в открытой окрестности x^* , $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ положительно определена.

Липшицевость: $\|\nabla^2 f(x_1) - \nabla^2 f(x_2)\| \leq M\|x_1 - x_2\|$ для любых x_1, x_2 в некоторой окрестности x^* .

Многомерный метод Ньютона

$$p_n = -\nabla^2 f_n^{-1} \nabla f_n, \quad f_n = f(x_n).$$

Вопрос: сколько итераций требуется методу для сходимости на квадратичной функции?

Как вычислять градиент/матрицу Гессе: честно/численно (обсуждали).

Можно искать минимум по направлению p_n , либо задать длину шага: константа ($\alpha_n = 1$) — не самое лучшее решение.

Backtracking: $\alpha_n^{(0)} = 1, \alpha_n^{(j+1)} = \alpha_n^{(j)}/2$.

Если $f(x_n) \geq f(x_n + \alpha_n^{(j)} p_n)$, то $\alpha_n = \alpha_n^{(j)}$.

Когда можно обойтись без вычисления матрицы Гессе?

Пример: пусть $f = \|g(x) - b\|^2, g: \mathbb{R}^k \rightarrow \mathbb{R}^M, b \in \mathbb{R}^M$.

Метод Гаусса-Ньютона: $p_n = -(\nabla g(x_n))^+ (g(x) - b), \nabla g(x_n) \in \mathbb{R}^{M \times k}$ — якобиан, $F^+ = (F^T F)^{-1} F^T$ — псевдообращение Мура-Пенроуза.

На практике: методы *BFGS, L-BFGS-B* — используют приближение гессиана.

Многомерный метод Ньютона

$$p_n = -\nabla^2 f_n^{-1} \nabla f_n, \quad f_n = f(x_n).$$

Вопрос: сколько итераций требуется методу для сходимости на квадратичной функции?

Как вычислять градиент/матрицу Гессе: честно/численно (обсуждали).

Можно искать минимум по направлению p_n , либо задать длину шага: константа ($\alpha_n = 1$) — не самое лучшее решение.

Backtracking: $\alpha_n^{(0)} = 1, \alpha_n^{(j+1)} = \alpha_n^{(j)}/2$.

Если $f(x_n) \geq f(x_n + \alpha_n^{(j)} p_n)$, то $\alpha_n = \alpha_n^{(j)}$.

Когда можно обойтись без вычисления матрицы Гессе?

Пример: пусть $f = \|g(x) - b\|^2, g: \mathbb{R}^k \rightarrow \mathbb{R}^M, b \in \mathbb{R}^M$.

Метод Гаусса-Ньютона: $p_n = -(\nabla g(x_n))^+(g(x) - b), \nabla g(x_n) \in \mathbb{R}^{M \times k}$ — якобиан, $F^+ = (F^T F)^{-1} F^T$ — псевдообращение Мура-Пенроуза.

На практике: методы *BFGS, L-BFGS-B* — используют приближение гессиана.

Многомерный метод Ньютона

$$p_n = -\nabla^2 f_n^{-1} \nabla f_n, \quad f_n = f(x_n).$$

Вопрос: сколько итераций требуется методу для сходимости на квадратичной функции?

Как вычислять градиент/матрицу Гессе: честно/численно (обсуждали).

Можно искать минимум по направлению p_n , либо задать длину шага: константа ($\alpha_n = 1$) — не самое лучшее решение.

Backtracking: $\alpha_n^{(0)} = 1, \alpha_n^{(j+1)} = \alpha_n^{(j)}/2$.

Если $f(x_n) \geq f(x_n + \alpha_n^{(j)} p_n)$, то $\alpha_n = \alpha_n^{(j)}$.

Когда можно обойтись без вычисления матрицы Гессе?

Пример: пусть $f = \|g(x) - b\|^2, g: \mathbb{R}^k \rightarrow \mathbb{R}^M, b \in \mathbb{R}^M$.

Метод Гаусса-Ньютона: $p_n = -(\nabla g(x_n))^+(g(x) - b), \nabla g(x_n) \in \mathbb{R}^{M \times k}$ — якобиан, $F^+ = (F^T F)^{-1} F^T$ — псевдообращение Мура-Пенроуза.

На практике: методы *BFGS*, *L-BFGS-B* — используют приближение гессиана.

Возможные критерии остановки

- $\|\nabla f(\mathbf{x}_n)\| < \varepsilon$
- $\|\mathbf{x}_n - \mathbf{x}_{n-1}\| < \varepsilon$
- $\left\| \frac{f(\mathbf{x}_n) - f(\mathbf{x}_{n-1})}{f(\mathbf{x}_n)} \right\| < \varepsilon$

+ ограничение на число итераций! (чтобы метод не работал бесконечно долго, если он “застрял”).

Вопрос: а если мы используем численное вычисление градиента, то надо ли нам тогда обратить внимание на ε ?

Методы на выбор для семестрового задания

- 1 Покоординатный спуск
- 2 Градиентный спуск
- 3 Сопряженные градиенты Fletcher-Reeves
- 4 Сопряженные градиенты Polak-Ribiere
- 5 Ньютон (поиск по направлению)
- 6 Ньютон (backtracking)
- 7 Nelder-Mead (метод деформируемого многогранника)