ЛЕКЦИЯ 5
Глава 3. МЕТОДЫ ДЛЯ ЗАДАЧ БЕЗУСЛОВНОЙ ОПТИМИЗАЦИИ И НЕЛИНЕЙНЫХ УРАВНЕНИЙ
(продолжение)

Содержание лекции

1. Методы спуска (завершение)
   - Градиентные методы

Задача безусловной оптимизации

Задача

$$f(x) \to \min, \quad x \in \mathbb{R}^n, \tag{1}$$

где $f : \mathbb{R}^n \to \mathbb{R}$ — заданная функция.

Общая схема методов спуска

$$x^{k+1} = x^k + \alpha_k d^k, \quad d^k \in \mathcal{D}_f(x^k), \quad k = 0, 1, \ldots, \tag{2}$$

где параметры длины шага $\alpha_k > 0$ выбираются так, чтобы выполнялось по крайней мере

$$f(x^{k+1}) < f(x^k). \tag{3}$$

Градиентные методы

Градиентные методы
(предполагают дифференцируемость функции f на $\mathbb{R}^n$)

Это методы спуска, в которых $d^k = -f'(x^k)$.
Итерационная формула (2) принимает вид

$$x^{k+1} = x^k - \alpha_k f'(x^k), \quad k = 0,\, 1,\, \ldots \tag{13}$$

В частности, это методы первого порядка.

Градиентные методы — не ньютоновские, но:

- их изучение естественным образом приводит к мысли о необходимости лучших (ньютоновских) методов;
- идеи анализа глобальной сходимости распространяются на значительно более широкие классы методов спуска.

Градиентные методы

Алгоритм 1

Выбираем $x^0 \in \mathbb{R}^n$ и полагаем $k = 0$. Выбираем одно из трех основных правил одномерного поиска и необходимые для реализации этого правила параметры.

1. Вычисляем α_k в соответствии с выбранным правилом одномерного поиска по направлению $d^k = -f'(x^k)$.
2. Вычисляем x^{k+1} по формуле (13).
3. Увеличиваем номер шага k на 1 и переходим к п. 1.

Градиентные методы

Если $f'(x^k) = 0$, то текущая точка x^k является стационарной точкой задачи (1), и на практике метод останавливают. Чтобы не выделять случай $f'(x^k) = 0$ при теоретическом анализе, удобно считать, что при этом $x^k = x^{k+1} = \ldots$, что согласуется с алгоритмом 1.

При реализации алгоритм должен быть снабжен практическими правилами остановки.

Метод скорейшего спуска

Метод скорейшего спуска

Это градиентный метод с выбором параметров длины шага по правилу одномерной минимизации.

Для этого метода из (5) следует

$$\langle f'(x^{k+1}), f'(x^k)\rangle = 0,$$

т.е. направления $d^k = -f'(x^k)$ и $d^{k+1} = -f'(x^{k+1})$ ортогональны: «зигзагообразное» движение.
Должно настораживать!

Градиентные методы

Неравенство Армихо (6)

$$f(x^k - \alpha f'(x^k)) \leqslant f(x^k) - \varepsilon\alpha\|f'(x^k)\|^2.$$

Равенство (8) принимает вид

$$\bar{\alpha}_k = \frac{2(1-\varepsilon)}{\ell}.$$

Поэтому, согласно лемме 3, если градиент f липшицев, то

$$\alpha_k \geqslant \check{\alpha}, \tag{14}$$

где $\check{\alpha} > 0$ не зависит от k.

Глобальная сходимость

Теорема 1

Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ дифференцируема на $\mathbb{R}^n$, и ее производная липшицева на $\mathbb{R}^n$ с константой $\ell > 0$. Пусть в случае использования в алгоритме 1 правила постоянного параметра он удовлетворяет условию $\bar{\alpha} < 2/\ell$.
Тогда любая предельная точка любой траектории $\{x^k\}$ алгоритма 1 является стационарной точкой задачи (1).
Если предельная точка существует, или если функция f ограничена снизу на $\mathbb{R}^n$, то

$$\{f'(x^k)\} \to 0 \quad (k \to \infty). \tag{15}$$

Глобальная сходимость

Доказательство для правила Армихо

В силу неравенства Армихо $\{f(x^k)\}$ монотонно невозрастает. Тогда если $\{x^k\}$ имеет предельную точку, то $\{f(x^k)\}$ ограничена снизу, а значит сходится (если f ограничена снизу, то это верно и без предположения о наличии у $\{x^k\}$ предельной точки).
Из неравенства Армихо и (14) имеем

$$f(x^k) - f(x^{k+1}) \geqslant \varepsilon\alpha_k\|f'(x^k)\|^2 \geqslant \varepsilon\check{\alpha}\|f'(x^k)\|^2.$$

В силу сходимости $\{f(x^k)\}$ величина в левой части стремится к нулю, и значит, выполнено (15).

Для правил одномерной минимизации и постоянного параметра доказательство получается сведением к правилу Армихо.

Глобальная сходимость без липшицевости градиента

При отказе от требования липшицевости градиента трудность в том, что (14) может не выполняться ни при каком $\check{\alpha} > 0$: параметр длины шага может уменьшаться бесконечно.

Теорема 2

Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема на $\mathbb{R}^n$. Пусть в алгоритме 1 используется правило одномерной минимизации или правило Армихо.
Тогда любая предельная точка любой траектории $\{x^k\}$ алгоритма 1 является стационарной точкой задачи (1).
Если последовательность $\{x^k\}$ ограничена, то имеет место (15).

Глобальная сходимость

От градиентного метода нельзя ожидать большего, чем стационарность предельных точек траекторий: при попадании в неоптимальную стационарную точку метод не имеет средств, позволяющих из нее выбраться.

Предельная точка существует в случае ограниченности множества Лебега $L(x^0) = L_{f,\mathbb{R}^n}(f(x^0))$.
Если в теореме 2 множество $L(x^0)$ ограничено, то

$$\mathrm{dist}(x^k, S_0(x^0)) \to 0 \quad (k \to \infty),$$

где $S_0(x^0) = \{x \in L(x^0) \mid f'(x) = 0\}$.

Скорость сходимости

Типичный ингредиент анализа скорости сходимости: некоторая оценка расстояния до решения или множества решений.

Условие квадратичного роста в точке $\bar{x} \in \mathbb{R}^n$

$$f(x) - f(\bar{x}) \geqslant \gamma\|x - \bar{x}\|^2 \quad \forall x \in U \tag{16}$$

для некоторых окрестности U точки $\bar{x}$ и числа $\gamma > 0$.

Всегда имеет место при выполнении достаточного условия второго порядка оптимальности в стационарной точке $\bar{x}$, состоящем в том, что матрица $f''(\bar{x})$ в разложении

$$f(x) - f(\bar{x}) = \frac{1}{2}\langle f''(\bar{x})(x - \bar{x}),\, x - \bar{x}\rangle + o(\|x - \bar{x}\|^2)$$

положительно определена (подразумевается двукратная дифференцируемость функции f в точке $\bar{x}$).

Скорость сходимости

Условие квадратичного роста не может выполняться в неизолированном глобальном решении задачи (1).

Задача метода наименьших квадратов

Это задача (1), в которой

$$f(x) = \frac{1}{2}\|F(x)\|^2, \quad x \in \mathbb{R}^n, \tag{17}$$

где $F : \mathbb{R}^n \to \mathbb{R}^l$ — заданное отображение.

Множество глобальных решений этой задачи совпадает с множеством решений (при их наличии) уравнения

$$F(x) = 0.$$

Если $n > l$, то это множество обычно состоит из неизолированных точек.

Скорость сходимости

Имея в виду возможную неизолированность решений, условие квадратичного роста (16) естественно заменить оценкой расстояния до множества S глобальных решений:

$$f(x) - \bar{v} \geqslant \gamma(\mathrm{dist}(x,\, S))^2 \quad \forall\, x \in U. \tag{18}$$

Не подразумевает выпуклость f даже локально (вблизи $\bar{x}$).

Для целевой функции f из (17) метода наименьших квадратов оценка (18) имеет место при выполнении условия регулярности $\operatorname{rank} F'(\bar{x}) = l$, поскольку полная версия теоремы Люстерника при этом дает оценку

$$\mathrm{dist}(x,\, S) \leqslant c\|F(x)\| \quad \forall\, x \in U$$

для некоторых окрестности U точки $\bar{x}$ и числа $c > 0$.

Скорость сходимости

Для этих методов нет нужды в чисто локальных рассмотрениях.

Условие квадратичного роста относительно множества $S \subset \mathbb{R}^n$

Это (18), где U — окрестность множества S (а не точки $\bar{x}$).

Часто используют специальные окрестности множества S: при некотором $\delta > 0$

- $U = U(S, \delta) = \{x \in \mathbb{R}^n \mid \operatorname{dist}(x, S) \leqslant \delta\}$;
- $U = L_{f, \mathbb{R}^n}(\bar{v} + \delta)$;
- $U = L_{\|f'(\cdot)\|, \mathbb{R}^n}(\delta)$.

Ожидаемые результаты: (глобальная) сходимость к множеству S, либо (в идеале) к точке этого множества, а также получение оценок скорости сходимости.

Скорость сходимости

Естественно пытаться установить сходимость и скорость сходимости градиентных методов к множеству S_0 стационарных точек задачи (1).
Тогда условие квадратичного роста естественно заменить оценкой расстояния до этого множества:

$$\|f'(x)\| \geqslant \gamma \operatorname{dist}(x, S_0) \quad \forall x \in U, \tag{19}$$

где U — окрестность множества S_0.

Если $S = S_0$, а функция f достаточно гладкая, то из условия квадратичного роста (18) при $U = U(S, \delta)$ следует (19) (возможно, при других U и γ).

Скорость сходимости

Оценка (19) при $U = U(S_0,\, \delta)$ равносильна этой же оценке при $U = L_{\|f'(\cdot)\|,\,\mathbb{R}^n}(\delta)$ (возможно, с другим $\delta > 0$).
Всюду далее $U = L_{\|f'(\cdot)\|,\,\mathbb{R}^n}(\delta)$.
Тогда в предположениях теоремы 1, если $\bar{v} > -\infty$, то выполнено (15), а значит, для любой траектории алгоритма 1 все приближения, начиная с некоторого, лежат в U.

Скорость сходимости

Еще одно техническое требование:

Условие отделимости критических поверхностей уровня

Существует $\chi > 0$ такое, что для любых $\bar{x}^1, \bar{x}^2 \in S_0$, для которых $f(\bar{x}^1) \neq f(\bar{x}^2)$, имеет место $\|\bar{x}^1 - \bar{x}^2\| \geqslant \chi$.

Всегда выполняется в любом из случаев:

- S_0 гладко связно (при этом $f \equiv \text{const}$ на S_0);
- производная f липшицева на $\mathbb{R}^n$, и f принимает на S_0 лишь конечное число различных значений;
- f — квадратичная функция, имеющая критическую точку (при этом также выполняется оценка расстояния (19) при $U = \mathbb{R}^n$).

Скорость сходимости

Теорема 3

Пусть в дополнение к условиям теоремы 1 значение задачи (1) конечно, и выполнено (19) при $U = L_{\|f'(\cdot)\|, \mathbb{R}^n}(\delta)$ и некоторых $\delta > 0$ и $\gamma > 0$. Пусть, кроме того, выполнено условие отделимости критических поверхностей уровня. Пусть в алгоритме 1 используется правило Армихо или правило постоянного параметра.
Тогда любая траектория $\{x^k\}$ алгоритма 1 сходится к некоторой стационарной точке $\bar{x}$ задачи (1).
Скорость сходимости по функции линейная, а по аргументу геометрическая.

Градиентные методы

Главный недостаток методов этого класса: неприемлемо низкая скорость сходимости в случаях, когда поверхности уровня функции f «вытянуты» (т.е. имеет место «овраг», особенно с изогнутым «дном»).

Оценка скорости к стационарной точке $\bar{x}$:

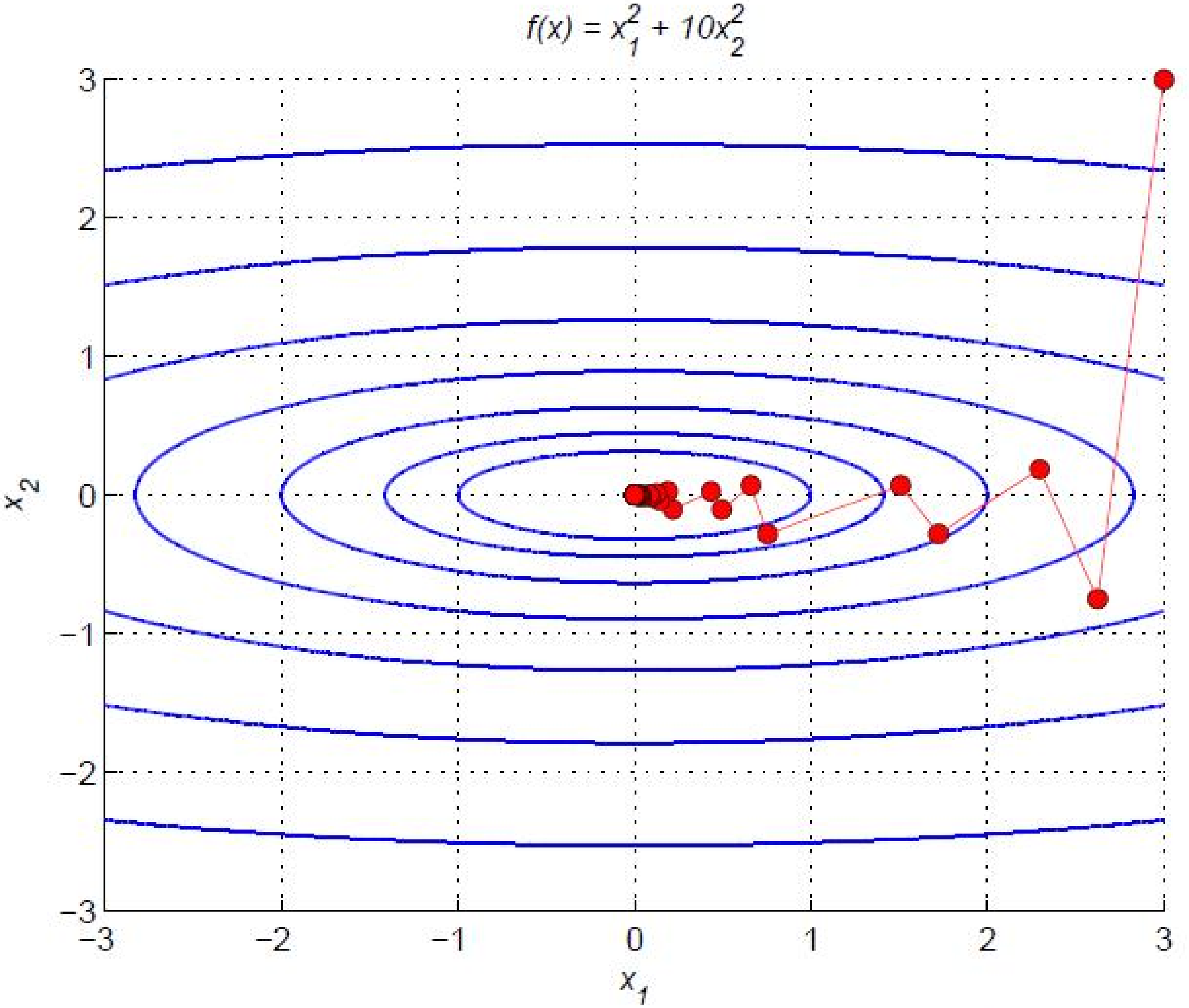
$$\|x^k - \bar{x}\| \leqslant c\frac{q^{k/2}}{1 - q^{1/2}}, \tag{20}$$

где

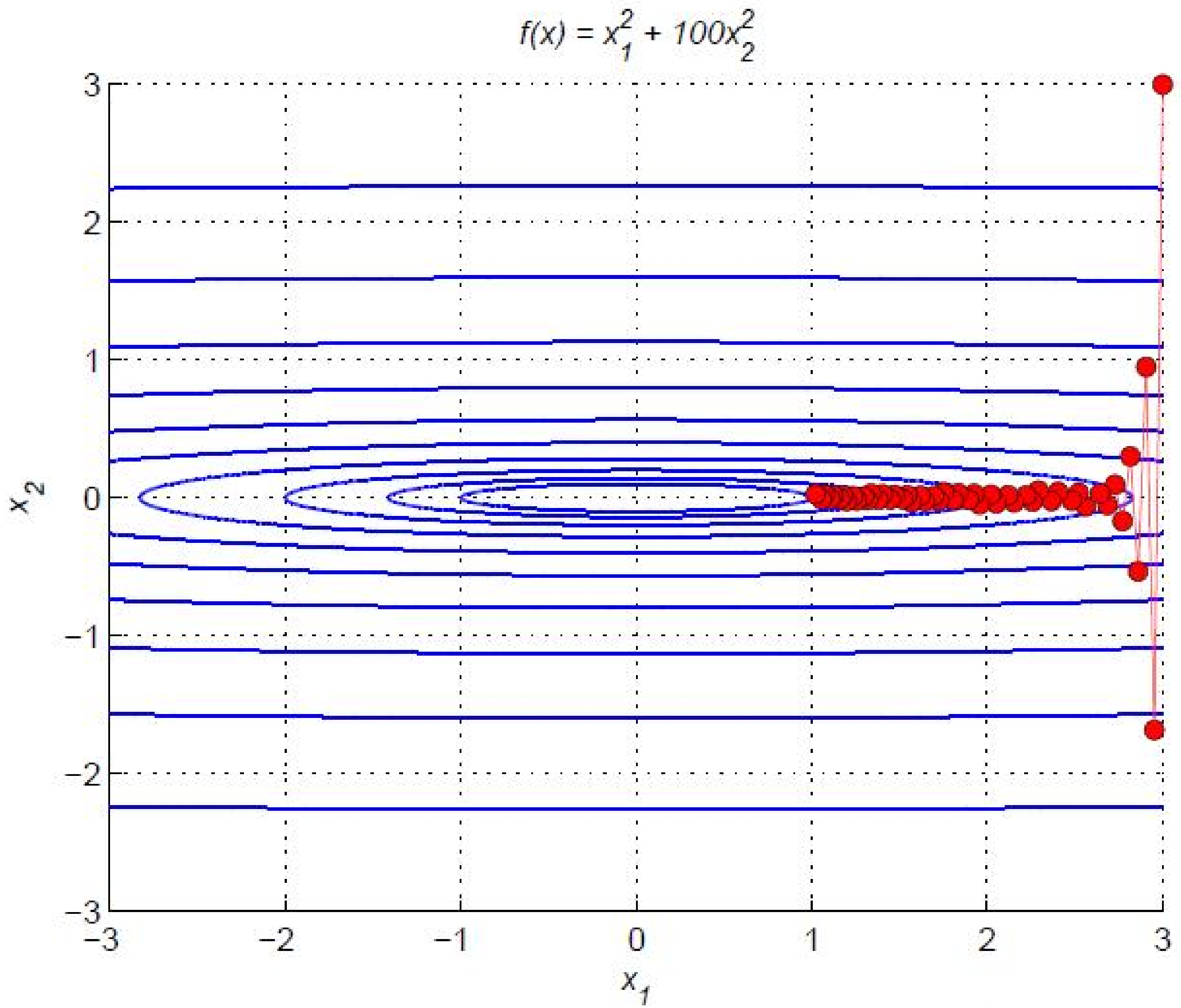
- $c > 0$ — константа;
- $$q = \frac{M}{1+M} \in (0, 1), \quad M = \frac{\ell\widehat{\alpha}}{2\varepsilon}\left(1 + \frac{1}{\check{\alpha}\gamma}\right)^2.$$

Если $\ell \gg \gamma$ (характеризует овраг), то $q \approx 1$.

Градиентные методы



Градиентные методы



Градиентные методы

На практике градиентные методы крайне неэффективны и должны использоваться лишь как «последнее средство», когда другие (более изощренные) методы не применимы.

Далее: гораздо более эффективные (ньютоновские) методы.