2.5.5. Методы возможных направлений

Указанные методы применяются для ***численного*** решения НП-задач.

Определение. Вектор с ненулевыми компонентами (ненулевой) S называется ***возможным направлением*** спуска в точке $X \in \mathfrak{R}$, если существует $\delta > 0$, такое, что $X + \lambda S \in \mathfrak{R}$ для всех $\lambda \in (0, \delta)$ и $f(X + \lambda S) < f(X)$.

Определение такого направления составляет краеугольный камень, лежащий в основании методов возможных направлений. Типичными представителями таковых являются алгоритмы, предложенные Зойтендейком (есть написание Заутендайка, в оригинале G. Zoutendijk) и Розеном.

2.5.5.1. Метод Зойтендейка [29, 30, 56]

Существуют три разновидности указанного метода:

- при линейных ограничениях;
- при нелинейных ограничениях и
- улучшенной сходимости при нелинейных ограничениях.

Случай линейных ограничений

Постановка задачи, в этом случае, такова

$$\min f(X)$$
$$\begin{cases} AX \le b, \\ HX = h, \end{cases}$$

где $f(X)$ – нелинейна, $A[m \times n]$, $H[l \times n]$ – матрицы, а $b[m]$ и $h[l]$ – вектора системы линейных ограничений.

Пусть в текущей точке все ограничения со знаком "≤" представимы в виде

$$\begin{cases} A_1 X = b_1, \\ A_2 X < b_2. \end{cases}$$

При этом считаем, что исходные матрицы блочные (вернее подлежат разделению или перестановке) по знакам отношений

$$\begin{cases} A^T = \begin{bmatrix} A_1^T & A_2^T \end{bmatrix}, \\ b^T = \begin{bmatrix} b_1^T & b_2^T \end{bmatrix}. \end{cases}$$

Вектор S будет являться направлением спуска в точке X при выполнении условий:

$$\begin{cases} A_1 S \le 0, \\ HS = 0, \end{cases}$$

как это показано на рисунке 2.24 ниже.

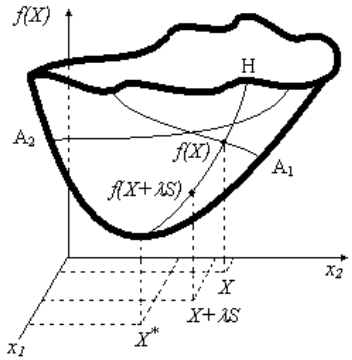

Рисунок 2.24 – К пояснению выбора направления

На изображении целевая функция показана жирной линией. Она напоминает кубок, линии на кубке есть проекции области ограничений на поверхность, описываемую целевой функцией. Из эскиза видно, что спуск должен выполняться «вдоль линии» (или не покидая гиперплоскости, в многомерном случае), по которой выполняются исходные ограничения вида «равенство», в сторону от активных ограничений (неравенства, выполняющиеся в точке в виде равенства, как нами было определено выше).

То есть, выбранное направление спуска S должно обеспечивать выполнение исходных ограничений модели $HX = h$ и устранять активность ограничений $A_1 X = b_1$.

На компоненты вектора S налагаются дополнительные ограничения (требования) нормировки:

- на величину элементов

$$-1 \le s_i \le 1,\ j = \overline{1,n}; \tag{2.103}$$

- на модуль вектора возможного направления S

$$S^T S \le 1, \|S\|^2 \le 1, \tag{2.104}$$

это ограничения объединяет условия $-1 \le s_j \le 1, j = \overline{1,n}$ и $\sum_{j=1}^{n} s_i \le 1, \sum_{j=1}^{n} s_i \ge -1$;

- на величину целевой функции ЗЛП

$$\nabla^T f(X) S \ge -1. \tag{2.105}$$

Поэтому возможна одна из трёх постановок задачи минимизации, которая может быть решена соответствующим методом линейного программирования:

$$\begin{array}{lll} \min \nabla^T f(X) S & \min \nabla^T f(X) S & \min \nabla^T f(X) S \\ \begin{cases} A_1 S \le 0, \\ HS = 0, \\ s_j \le 1, \\ s_j \ge -1,\ j = \overline{1,m}. \end{cases} & \begin{cases} A_1 S \le 0, \\ HS = 0, \\ S^T S \le 1. \end{cases} & \begin{cases} A_1 S \le 0, \\ HS = 0, \\ \nabla^T f(X) S \ge -1. \end{cases} \\ 1) & 2) & 3) \end{array} \tag{2.106}$$

Алгоритм Зойтендейка для случая линейных ограничений

Предварительный этап.

Нахождение точки X, удовлетворяющей системе ограничений задачи.

Итерация.

1. Найти множество активных ограничений в текущей точке и композицию матрицы системы ограничений $A^T = \left[A_1^T \quad A_2^T\right]$ и решить ЗЛП (2.106) вида 1, 2 или 3, по желанию.

2. *Проверка условия окончания.*

Если оптимальное значение целевой функции ЗЛП равно нулю, ***то*** текущие координаты X определяют точку Куна-Таккера.

3. ***Иначе*** решить задачу одномерной минимизации Коши: $\min_{0 \le \lambda \le \lambda_{\max}} f(X + \lambda S)$, в которой граничное значение параметра λ определяется как

$$\lambda_{\max} = \begin{cases} \min\limits_k \left\{ \dfrac{\widetilde{b}_k}{\widetilde{s}_k} \middle| \widetilde{s}_k > 0 \right\}, & \exists_k \widetilde{s}_k > 0, \\ \infty, & \forall_k \widetilde{s}_k \le 0. \end{cases} \quad \text{где } \widetilde{b} = b_2 - A_2 X, \qquad \widetilde{S} = A_2 S.$$

4. Перейти в новую текущую точку $X = X + \lambda \times S$.

Алгоритм не сложен, он представлен на рисунке 2.25.

Алгоритм Зойтендейка для случая нелинейных ограничений

Рассмотрим задачу НП-программирования вида

$$\begin{cases} \min f(X); \\ g_i(X) \le 0, \quad i = \overline{1,m}; \\ x_j \ge 0, \qquad\qquad j = \overline{1,n}. \end{cases}$$

Пусть текущая точка X является допустимой точкой, а $I = \{i : g_i(X) = 0\}$ – множество ограничений, активных в этой точке. Если

$$\begin{cases} \nabla^T f(X) S < 0, \\ \nabla^T g_i(X) S < 0,\ i \in I, \end{cases}$$

то S – вектор возможного направления спуска в этой точке (смотри рисунок 2.24).

Алгоритм имеет топологию, аналогичную представленной рисунком 2.25.

Предварительный этап.

Найти начальную точку X, для которой выполняются ограничения задачи: $g_i(X) \le 0, \quad i = \overline{1,m}$.

Основной этап (итерация).

1. Найти множество активных ограничений задачи в текущей точке $I = \{i : g_i(X) = 0\}$ и решить ЗЛП вида

$$\min z$$
$$\begin{cases}\nabla^T f(X)S - z \le 0,\\ \nabla^T g_i(X)S - z \le 0,\ i \in I = \{i : g_i(X) = 0\}.\end{cases}$$

при любом, оговоренном выше, условии нормировки компонентов вектора возможного направления S (2.103) – (2.105).

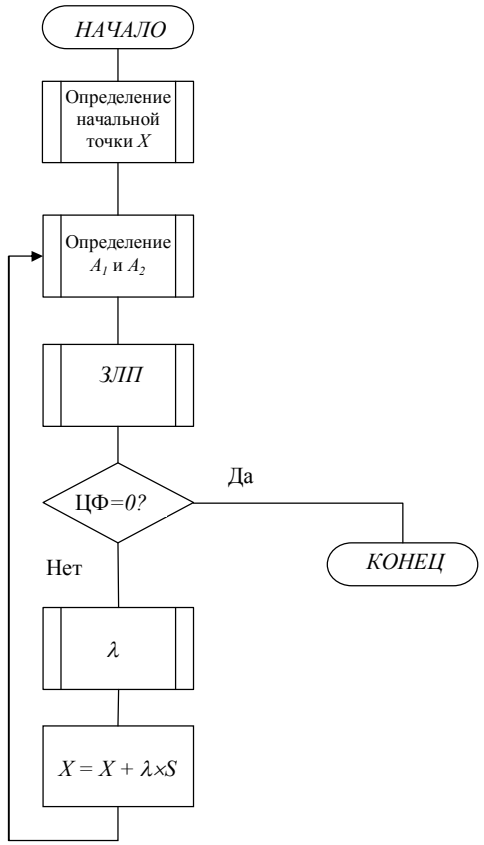

Рисунок 2.25 – Граф-схема алгоритма Зойтендейка

2. *Проверка условия окончания*.

Если оптимальное значение целевой функции ЗЛП равно нулю, ***то*** X – оптимальное решение (точка Куна-Таккера) исходной НП-задачи.

3. *В противном случае* необходимо решить задачу одномерной минимизации Коши: $\min\limits_{0\le\lambda\le\lambda_{\max}} f(X+\lambda S)$, где параметр λ определяется из условия

$$\lambda_{\max} = \max\{\lambda : g_i(X+\lambda S) \le 0,\ i = \overline{1,m}\}.$$

4. Переместится в следующую точку $X = X + \lambda \times S$ и продолжить решение.

Алгоритм Зойтендейка для случая нелинейных ограничений улучшенной сходимости

Шаги, генерируемые вдоль направления поиска, стремятся, по мере итераций, к нулю. Это может, в отдельных случаях, вызвать остановку вычислительного процесса без достижения оптимума. Чтобы избежать несанкционированной остановки, ограничения эквивалентной ЗЛП усиливают.

Модифицированная ЗЛП имеет вид

$$\min z$$
$$\begin{cases}\nabla^T f(X)S - z \le 0,\\ \nabla^T g_i(X)S - z \le -g_i(X),\ i = \overline{1,m}.\end{cases}$$

При этом нововведения позволяют учитывать как активные, так и обычные ограничения в точке поиска. ЗЛП снабжается одним из условий нормировки (2.103) – (2.105).

2.5.5.2. Метод проекции градиента Розена

Метод предназначен для решения задач нелинейного программирования с линейной системой ограничений

$$\min f(X)$$
$$\begin{cases}AX \le b,\\ HX = h,\end{cases}$$

где $f(X)$ – нелинейная и ***дифференцируемая*** функция, A – матрица размером $[m \times n]$, H – $[l \times n]$-мерная матрица , а b – m-мерный вектор, а h – l-мерный вектор системы линейных ограничений.

Идея, лежащая в основе метода, состоит в следующем.

Пусть X – допустимая точка, удовлетворяющая системе ограничений, а направление спуска определяется градиентом $S = -\nabla f(X)$.

Движение вдоль направления спуска может, гипотетически, привести к ***нарушению допустимости***. Поэтому Розен предложил направление спуска строить по правилу

$$S = -P \times \nabla f(X),$$

где P – матрица проецирования или проецирующая (проектирующая) матрица, которая бы гарантировала сохранение допустимости текущей точки.

Пусть в допустимой точке часть неравенств активна, а другая часть – нет, то допускается представление системы ограничений по признаку активности в виде блочной матрицы:

$$\begin{aligned} &A_1 X = b_1, \quad A^T = \begin{bmatrix} A_1^T & A_2^T \end{bmatrix}, \\ &A_2 X < b_2, \quad b^T = \begin{bmatrix} b_1^T & b_2^T \end{bmatrix} \end{aligned} \qquad (2.107)$$

Матрица проецирования P должна удовлетворять условию $P \times \nabla f(X) \neq 0$ и блокировать возможные направления в сторону активных ограничений. Это будет соблюдаться, когда проектирующая матрица определена как

$$P = I - M^T \times (M \times M^T)^{-1} \times M, \qquad (2.108)$$

где I – единичная матрица, $M^T = [A_1{}^T H^T]$ – невырождена. Отметим, что выполняется тождество $M \times P = 0$, то есть $A_1 \times P = 0$ и $H \times P = 0$, таким образом, спуск в направлении ограничений, выполняемых как равенство, не производится.

Рассмотрим точку, в которой выполняется условие $P \times \nabla f(X) = 0$. Имеем

$$[I - M^T \times (M \times M^T)^{-1} \times M] \times \nabla f(X) = \nabla f(X) + M^T \times W = 0,$$

где $W^T = [U^T V^T]$, а

$$\nabla f(X) + A_1{}^T \times U + H^T \times V = 0.$$

Если компоненты вектора U неотрицательны, то X является точкой Куна-Таккера. В противном случае, можно построить новое направление спуска

$$S = -\tilde{P} \times \nabla f(X),$$

где $\tilde{P} = I - \tilde{M}^T \times (\tilde{M} \times \tilde{M}^T)^{-1} \times \tilde{M}$, в котором $\tilde{M}^T = [\tilde{A}_1^T H^T]$, а $\tilde{A}_1$ получено из A_1 путём вычёркивания строк, соответствующих компонентам $u_j < 0$.

Схема алгоритма показана на рисунке 2.26.

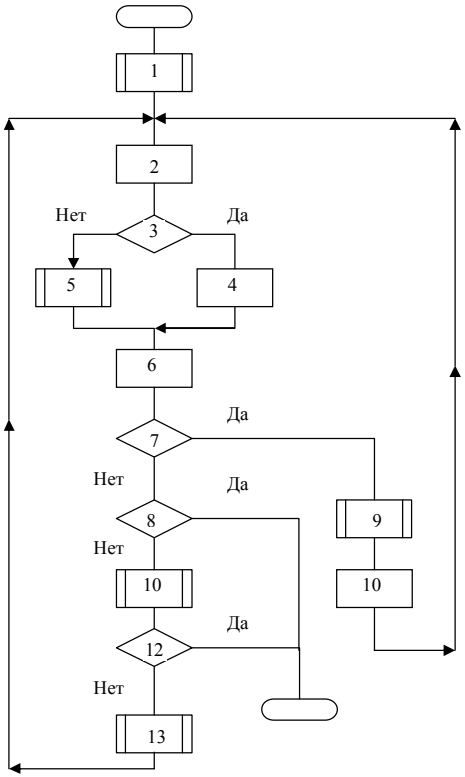

Рисунок 2.26 – Граф-схема алгоритма Розена

Блок № 1 выполняет выбор начальной точки X, удовлетворяющей системе ограничений задачи, отыскивает активные ограничения и разложение ограничений с неравенствами (2.107).

В блоке № 2 конструируется матрица $M^T = [A_1{}^T H^T]$, которая проверяется на равенство нулю всех её элементов условием блока № 3: $M = 0?$, если матрица вся нулевая (выход «да» блока № 3), то проецирующая матрица полагается равной единичной (блок № 4).

В противном случае (выход «нет» блока № 3), выполняется расчёт проецирующей матрицы в блоке № 5 по формуле (2.108).

После определения элементов проецирующей матрицы, рассчитывается возможное направление спуска $S = -P \times \nabla f(X)$ в блоке № 6.

Блок № 7 выполняет проверку: $S \neq 0?$, то есть, не является ли текущая точка точкой Куна-Таккера.

Если все компоненты вектора возможного направления нулевые (выход «нет» блока № 7), то это, возможно, искомое решение. Чтобы убедиться в этом, проверяется равенство нулю всех компонентов блочной матрицы M (блок № 8, $M = 0?$), и если это так (выход «да» блока № 8), то решение успешно заканчивается.

Если среди компонентов матрицы M есть ненулевые элементы (выход «нет» блока № 8), то ищется разложение $W^T = [U^T V^T]$, блок № 10.

Если компоненты подматрицы U неотрицательны, что проверяется блоком № 12 ($\forall_i u_i \geq 0?$), то X является точкой Куна-Таккера (выход «да»).

Если компоненты U отрицательны (выход «нет» блока № 12), то переопределяется матрицы $A_1 = \widetilde{A}_1$ и блочная матрица $\widetilde{M}^T = [\widetilde{A}_1^T H^T]$ (блок № 13).

Если не все элементы вектора возможного направления нулевые (выход «да» блока № 7), рассчитывается новая текущая точка. Сперва выбирается параметр λ решением задачи одномерной минимизации Коши (блок № 9): $\min\limits_{0 \leq \lambda \leq \lambda_{\max}} f(X + \lambda S)$, в которой граничное значение параметра λ определяется, как и в методе Зойтендейка:

$$\lambda_{\max} = \begin{cases} \min\limits_k \left\{ \dfrac{\widetilde{b}_k}{\widetilde{s}_k} \middle| \widetilde{s}_k > 0 \right\}, & \exists_k \widetilde{s}_k > 0, \\ \infty, & \forall_k \widetilde{s}_k \leq 0. \end{cases} \quad \text{где } \widetilde{b} = b_2 - A_2 X, \qquad \widetilde{S} = A_2 S.$$

Значение очередной текущей точки рассчитывается в блоке № 11 по рекуррентной формуле $X = X + \lambda \times S$.

2.5.6. Методы штрафных функций

Методы штрафных функций основаны на практике перехода от задачи ***условной минимизации*** к задаче ***безусловной минимизации*** путём построения специальной штрафной функции.

Известны [3, 4, 24, 33, 40, 56] ***параметрические*** и ***непараметрические*** методы построения штрафных функций в виде полинома

При использовании ***параметрических*** методов, штрафной полином конструируется с использованием выражений, описывающих ограничения, в качестве параметров других функций (функционалов) и весовых коэффициентов.

В *непараметрических* методах функция рассматривается в качестве дополнительного ограничения, которое постоянно усиливается в процессе решения.

По характеру перемещения точки к оптимуму различают:

- методы внутренней точки,
- методы внешней точки и
- комбинированные методы.

При использовании методов ***внутренней точки***, последовательные приближения к оптимуму производятся ***внутри области***, определяемой ограничениями, благодаря специальной функции штрафа, называемой ***барьерной***.

Когда поиск оптимума осуществляется по методу ***внешней точки***, текущее решение находится ***за пределами области*** ограничений, попадая вовнутрь её на последнем шаге итераций.

Комбинированные методы используют, когда большинство ограничений задачи имеют вид равенства, и, в процессе решения, попеременно, одни ограничения выполняются, а другие нет.

Искомое решение получается, при удовлетворении заданных условий, в пределах отведённого допуска.

Пусть условия задачи имеют вид

$$\min f(X);$$
$$\begin{cases} h_i(X) = 0, & i = \overline{1,m}; \\ g_i(X) \geq 0, & i = \overline{m+1,l}. \end{cases}$$

Пользуясь параметрическим методом, по этим условиям можно построить следующую фунцию без ограничений:

$$P(X,\rho) = f(X) + \sum_{i=1}^{m} \rho_i H[h_i(X)] + \sum_{i=m+1}^{l} \rho_i G[g_i(X)],$$

где $P(X,\rho)$ – штрафная функция; ρ_i, $i = \overline{1,l}$ – весовые коэффициенты, отражающие значимость соблюдения того или иного ограничения; $H[h_i(X)]$ и $G[g_i(X)]$ – некоторые функционалы.

Рассмотрим, какими свойствами должны обладать эти функционалы.

Функционал $H[h_i(X)]$ должен сделать невыгодным любое отклонение аргумента X от поверхности $h_i(X) = 0$, то есть

$$\lim_{h_i(X)\to 0} H[h_i(X)] = 0, \qquad i = \overline{1,m}.$$

Поэтому в качестве функционала выбирают чётную степенную функцию

$$H[y] = y^p, \ p = 2,4,... \text{ либо } H[y] = |y|^p, \ p = 1,2,...$$

Функционал $G[g_i(X)]$ зависим от местоположения текущей точки в процессе решения.

- **Метод внутренней точки**:

$$\lim_{g_i(X)\to 0^+} G[g_i(X)] = \infty, \qquad i = \overline{m+1,l} \text{ для } g_i(X) > 0.$$

- **Метод внешней точки**:

$$\lim_{g_i(X)\to 0^-} G[g_i(X)] = 0, \qquad i = \overline{m+1,l} \text{ для } g_i(X) < 0.$$

- **Комбинированный метод**:

$$G[g_i(X)] > 0 \text{ для } g_i(X) < 0, \ i = \overline{m+1,l},$$
$$G[g_i(X)] = 0 \text{ для } g_i(X) = 0, \ i = \overline{m+1,l}.$$

При этом, не зависимо от траектории движения точки, в процессе решения задачи необходимо обеспечение выполнения условий:

$$\lim_{k\to\infty} \sum_{i=1}^{m} \rho_{i,k} H[h_i(X_k)] = 0,$$

$$\lim_{k\to\infty} \sum_{i=m+1}^{l} \rho_{i,k} G[g_i(X_k)] = 0,$$
$$\lim_{k\to\infty} \left| P(X_k,\rho_k) - f(X_k) \right| = 0,$$

где k – номер итерации.

Особо подчеркнём, что штрафная функция должна быть построена таким образом, что ***невыполнение*** какого либо их ограничений должно приводить ***к резкому возрастанию*** её в целом, и, по мере приближению к оптимуму, ***влияние*** штрафных добавок ***уменьшается***.

2.5.6.1 Метод барьерных поверхностей (МБП)

Данный алгоритм (рисунок 2.27) относится к группе методов внутренней точки, используется для решения задач вида

$$\min f(X);$$
$$\begin{cases} g_i(X) \ge 0, & i = \overline{1,m}; \\ x_j \ge 0, & j = \overline{1,n}. \end{cases}$$

По условиям задачи строится штрафная функция вида

$$P(X,r,\Omega) = f(X) + r \cdot \sum_{i=1}^{m} \omega_i G[g_i(X)]. \qquad (2.109)$$

В формуле (2.109) использованы: параметр $r > 0$, убывающий по мере вычислений и Ω – вектор положительных весовых коэффициентов, учитывающих важность (значимость) ограничений модели.

Функции барьера $G[g_i(X)]$ выбирается из альтернатив

$$G[Y] = 1/Y \text{ или } G[Y] = -\ln[Y].$$

Легко видеть, что $\lim_{Y\to 0^+} G[Y] = \infty$, то есть функция барьера, при приближении к нему изнутри области, неограниченно возрастает. Штрафная добавка к функции цели определяется формулой

$$\Delta = r \cdot \sum_{i=1}^{m} \omega_i G[g_i(X)]$$

Функционирование алгоритма представлено на рисунке 2.27, в его работе использована уменьшающая константа $0 < \beta < 1$, точность расчётов задаётся константой ε.

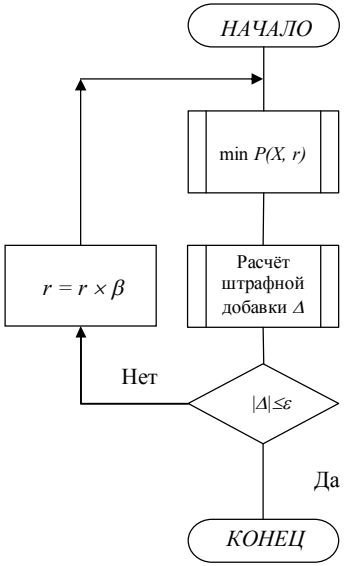

Рисунок 2.27 – Схема алгоритма МБП

Замечания:

- область ограничений должна быть непустой: $X_{opt} \in \{X: g_i(X) > 0, i = 1, m\}$;
- начальное значение X должно быть допустимой точкой;
- метод критичен по отношению к параметру r: если значение r мало, то будет отыскан оптимум функции, не совпадающий с истинным, а если значение r велико, то возрастёт число итераций;
- аналогично, близкие к единице значения β могут привести к преждевременному прекращению вычислений.

2.5.6.2 Метод внешней точки

Ориентируется на решение НП-задач в общей постановке

$$\min f(X);$$
$$\begin{cases} h_i(X) = 0, & i = \overline{1,m}; \\ g_i(X) \geq 0, & i = \overline{m+1,l}. \end{cases}$$

Штрафная функция описывается выражением

$$P(X, r) = f(X) + rL(X),$$

где r – коэффициент штрафа, а величина штрафа в текущей точке X есть

$$L(X) = \sum_{i=1}^{m} H[h_i(X)] + \sum_{i=m+1}^{l} G[g_i(X)].$$

Используются функционалы:

$$G[Y] = [\max\{0; -Y\}]^P, P > 0, \text{ целое},$$
$$H[Y] = |Y|^P, P > 0, \text{ целое}.$$

Легко видеть, что функционал $G[Y]$ обращается в нуль при попадании текущей точки в область ограничений.

Алгоритм метода полностью дублирует алгоритм метода барьерных поверхностей, показанный на рисунке 2.27, но, в отличие от него, используется параметр увеличения штрафа β больше единицы.

Если взять параметры r и β достаточно большИми, то минимизация штрафной функции будет выполняться за счёт функции цели и величины штрафа. Поэтому, в ходе расчётов, коэффициент штрафа будет постоянно возрастать, а штраф – убывать. Следовательно, положив r и β порядка десятков, сотен, а то и тысяч, можно получить решение задачи с приемлемой точностью.

Алгоритм, сам по себе, обладает хорошей сходимостью и устойчивостью. Очевидно, что вельми проблемным местом его функционирования будет процедура минимизации штрафной функции, что является нетривиальной задачей.