

УДК 519.6
DOI 10.17223/19988621/55/3

MSC 65K05, 90C30, 82C32

В.Н. Крутиков, Н.С. Самойленко

О СКОРОСТИ СХОДИМОСТИ СУБГРАДИЕНТНОГО МЕТОДА С ИЗМЕНЕНИЕМ МЕТРИКИ И ЕГО ПРИЛОЖЕНИЯ В СХЕМАХ НЕЙРОСЕТЕВЫХ ПРИБЛИЖЕНИЙ

Исследуется релаксационный субградиентный метод с двухгранговой коррекцией матриц метрики. Доказано, что на сильновыпуклых функциях, в случае существования линейного преобразования координат, уменьшающего степень обусловленности задачи, метод имеет линейную скорость сходимости, соответствующую этой степени обусловленности. Экспериментально установлено, что скорости сходимости квазиньютоновского и изучаемого методов на гладких функциях практически эквивалентны. Вычислительные возможности метода используются для построения эффективных алгоритмов обучения нейронных сетей.

Ключевые слова: *метод, субградиент, минимизация, скорость сходимости, нейронные сети, регуляризация.*

В задачах обучения по прецедентам (см., например, [1]) при небольших по размеру обучающих выборках и неизвестном виде математической модели возникает необходимость поиска соответствующего описания в виде искусственной нейронной сети (ИНС) [1–4]. При этом структура модели должна быть достаточно сложной для качественного описания данных и достаточно простой для обеспечения хороших обобщающих свойств [1]. Подобные проблемы возникают в различных практических приложениях аппарата ИНС [5–7]. Для устранения избыточного описания нейросети используют различные способы регуляризации [1, 8–12]. В задачах обучения ИНС при небольших обучающих выборках [1–4, 6, 7, 13] используют, как правило, сети с небольшим числом слоев, а в качестве методов обучения применяют методы сопряженных градиентов (МСГ) [13], квазиньютоновские (КНМ) [14, 15] и Левенберга – Марквардта (ЛМ) [16]. Учитывая неприменимость этих методов для решения негладких задач, слабую устойчивость методов ЛМ [13] и МСГ в условиях плохой обусловленности и росте размерности задачи представляется актуальным исследование методов обучения ИНС, имеющих высокую скорость сходимости как на гладких, так и негладких овражных функциях, в том числе и невыпуклых.

К числу методов, обладающих возможностями минимизации негладких и, в том числе, невыпуклых функций, относятся релаксационные субградиентные методы (РСМ). Свойствами скорости сходимости, близкими свойствам метода сопряженных градиентов, обладают РСМ, предложенные в работах [17–22]. Существенного повышения эффективности РСМ удалось достичь в результате создания методов негладкой оптимизации с изменением метрики пространства [17, 23, 24]. В данной работе теоретически и экспериментально рассмотрен релаксационный субградиентный метод с двухгранговой коррекцией матриц метрики (СМДМ) [23], который при исключении операции сжатия пространства эквивалентен г-алгоритму Н.З. Шора [17]. В статье установлено, что на сильновыпуклых функци-

ях с липшицевым градиентом [14] метод сходится линейно. В силу инвариантных свойств алгоритма, полученная оценка справедлива и в системе координат с наилучшими для оценки скорости сходимости пропорциями констант сильной выпуклости и Липшица. Проведенный вычислительный эксперимент подтверждает близость свойств методов СМДМ и КНМ на квадратичных функциях и эффективность метода при минимизации негладких функций с высокой степенью вытянутости поверхностей уровня.

Использование ИНС при небольших по размеру обучающих выборках наталкивается на проблемы выбора хорошего начального приближения и быстро наступающего переобучения в случае излишнего числа нейронов. Предложен новый эффективный способ выбора начального приближения ИНС. Использование регуляризации позволило исключить эффекты переобучения и эффективно удалять малозначимые нейроны и связи внутри нейронов. Возможности эффективного решения подобных задач обеспечены методом СМДМ. В статье приведены примеры решения задач обучения ИНС.

2. О скорости сходимости субградиентного метода с изменением метрики

Рассматривается задача минимизации дифференцируемой функции $f(x)$, $x \in R^n$, где R^n – конечномерное евклидово пространство. Обозначим (x, y) – скалярное произведение векторов, $\|x\| = \sqrt{(x, x)}$ – норму вектора. Для произвольной симметричной строго положительно определенной матрицы H размера $n \times n$ будем использовать обозначение $H > 0$.

Условие А. Будем предполагать, что функция $f(x)$, $x \in R^n$, дифференцируема и сильновыпукла с константой $l > 0$ [14]:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - l\lambda(1-\lambda)\|x-y\|^2/2, \quad 0 \leq \lambda \leq 1, \quad (1)$$

а ее градиент удовлетворяет условию Липшица с константой $L > 0$

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \quad (2)$$

Последовательные приближения алгоритма СМДМ [23] на некоторой k -й итерации при точном одномерном спуске строятся по формулам

$$x_{k+1} = x_k - \gamma_k s_k, \quad s_k = H_k g_k; \quad (3)$$

$$\gamma_k = \arg \min_{\gamma > 0} f(x_k - \gamma s_k). \quad (4)$$

Здесь x_0 – заданная начальная точка, $H_0 > 0$ – заданная начальная матрица, а матрицы $H_k > 0$ вычисляются по формулам

$$H_{k+1} = H_k - (1 - \frac{1}{\alpha^2}) \frac{H_k y_k y_k^T H_k^T}{(y_k, H_k y_k)} - (1 - \frac{1}{\beta^2}) \frac{H_k p_k p_k^T H_k^T}{(p_k, H_k p_k)}, \quad (5)$$

$$y_k = g_{k+1} - g_k, \quad p_k = g_{k+1} + t_k y_k, \quad t_k = -\frac{(y_k, H_k g_{k+1})}{(y_k, H_k y_k)},$$

$$\alpha > 1, \quad \beta \in (0, 1], \quad \alpha \cdot \beta > 1, \quad (6)$$

где коэффициент t_k вычисляется из условия ортогональности $(y_k, H_k p_k) = 0$.

Здесь и далее g , $g(x)$ – некоторый субградиент из субградиентного множества $\partial f(x)$ функции $f(x)$, $g_k = g(x_k)$.

Итерационный процесс (3) – (6) является частным случаем алгоритма минимизации из [23]. В [23] для преобразования (5) необходимо из множества $\partial f(x_{k+1})$ выбирать субградиент $g(x_{k+1})$, удовлетворяющий условию

$$(s_k, g(x_{k+1})) = (H_k g(x_k), g(x_{k+1})) \leq 0.$$

В силу точного одномерного спуска (4) для дифференцируемой функции это условие выполняется $(s_k, g(x_{k+1})) = 0$. В качестве нового агрегированного субградиента для формирования направления спуска в работе [23] предложено выбирать вектор на отрезке двух векторов p_k, g_{k+1} . В этой работе, как следует из (3), для формирования направления спуска используется только вектор g_{k+1} этого отрезка, что определяет алгоритм (3) – (6) как частный случай метода минимизации из [23].

Обозначим через x^* точку минимума функции $f(x)$, $f^* = f(x^*)$, $f_k = f(x_k)$, $\mu_k = f_k - f^*$, $A_k = H_k^{-1}$. Для матрицы H_{k+1} , полученной в результате (5) при параметрах α, β , удовлетворяющих (6), при условии $H_k > 0$ в [23] показано, что матрица $H_{k+1} > 0$ и для нее выполняются равенства

$$A_{k+1} = A_k + (\alpha^2 - 1) \frac{y_k y_k^T}{(y_k, H_k y_k)} + (\beta^2 - 1) \frac{p_k p_k^T}{(p_k, H_k p_k)}; \quad (7)$$

$$Sp(A_{k+1}) = Sp(A_k) + (\alpha^2 - 1) \frac{(y_k, y_k)}{(y_k, H_k y_k)} + (\beta^2 - 1) \frac{(p_k, p_k)}{(p_k, H_k p_k)}; \quad (8)$$

$$\det(H_{k+1}) = \frac{\det(H_k)}{\alpha^2 \beta^2}. \quad \det(A_{k+1}) = \alpha^2 \beta^2 \det A_k. \quad (9)$$

В следующей теореме показано, что наличие движения в результате итераций метода (3) – (5) приводит к уменьшению функции.

Теорема 1. Пусть функция $f(x)$ удовлетворяет условию А. Тогда для последовательности $\{f_k\}$, $k = 0, 1, 2, \dots$, заданной процессом (3), (4), имеет место оценка:

$$\mu_{k+1} \leq \mu_0 \exp \left[-\frac{l^2}{L^2} \sum_{i=0}^k \frac{\|y_i\|^2}{\|g_i\|^2} \right]. \quad (10)$$

Доказательство. Для сильновыпуклой функции выполняются неравенства [14]

$$l \|x - x_*\|^2 / 2 \leq f(x) - f_* \leq \|g(x)\|^2 / 2l. \quad (11)$$

Согласно определению μ_k , с учетом правого из неравенств в (11), получим

$$\begin{aligned} \mu_{k+1} &= \mu_k - (f_k - f_{k+1}) = \mu_k (1 - (f_k - f_{k+1}) / \mu_k) \leq \\ &\leq \mu_k (1 - 2l(f_k - f_{k+1}) / \|g(x)\|^2). \end{aligned} \quad (12)$$

Левое из неравенств в (11) справедливо и для одномерной функции

$\varphi(t) = f(x_k - tr_k / \|r_k\|)$. Отсюда, с учетом точного одномерного поиска (4) и условия Липшица (2), следует оценка

$$f_k - f_{k+1} \geq l \|x_k - x_{k+1}\|^2 / 2 \geq l \|y_k\|^2 / 2L^2.$$

Преобразуем (12), используя последнее соотношение и неравенство $\exp(-c) \geq 1 - c$ при $c \geq 0$:

$$\mu_{k+1} \leq \mu_k \left(1 - \frac{l^2 \|y_k\|^2}{L^2 \|g_k\|^2} \right) \leq \mu_k \exp \left(- \frac{l^2 \|y_k\|^2}{L^2 \|g_k\|^2} \right).$$

Рекуррентное использование последнего неравенства приводит к оценке (10). Теорема доказана.

В следующей теореме обосновывается линейная скорость сходимости метода СМДМ.

Теорема 2. Пусть функция $f(x)$ удовлетворяет условию A . Тогда для последовательности $\{f_k\}$, $k = 0, 1, 2, \dots$, заданной процессом (3) – (6), с ограниченной начальной матрицей H_0

$$m_0 \leq (H_0 z, z) / (z, z) \leq M_0, \quad (13)$$

имеет место оценка

$$\mu_{k+1} \leq \mu_0 \exp \left\{ - \frac{l^2}{L^2} \left[\frac{2(k+1) \ln(\alpha\beta)}{n(\alpha^2 - 1)} + \frac{\ln(m_0 / M_0)}{(\alpha^2 - 1)} \right] \right\}. \quad (14)$$

Доказательство. Исходя из (8), учитывая неравенство $\beta^2 - 1 \leq 0$, получим оценку следа матриц A_k :

$$\text{Sp}(A_{k+1}) \leq \text{Sp}(A_k) \left[1 + \frac{(\alpha^2 - 1)(y_k, y_k)}{\text{Sp}(A_k)(H_k y_k, y_k)} \right]. \quad (15)$$

В силу точного одномерного спуска (4) выполняется условие

$$(s_k, g(x_{k+1})) = (H_k g(x_k), g(x_{k+1})) = 0,$$

что вместе с положительной определенностью матриц H_k доказывает неравенство

$$(H_k y_k, y_k) = (H_k g(x_k), g(x_k)) + (H_k g(x_{k+1}), g(x_{k+1})) - 2(H_k g(x_k), g(x_{k+1})) \geq (H_k g(x_k), g(x_k)).$$

Отсюда, с учетом неравенства $\text{Sp}(A_k) \geq M_k$, где M_k – максимальное собственное значение матрицы A_k , получим

$$\begin{aligned} \text{Sp}(A_k)(H_k y_k, y_k) &\geq \text{Sp}(A_k)(H_k g(x_k), g(x_k)) \geq \\ &\geq \frac{\text{Sp}(A_k)}{M_k} (g(x_k), g(x_k)) \geq (g(x_k), g(x_k)). \end{aligned}$$

Неравенство (15) на основании последней оценки преобразуется к виду

$$\text{Sp}(A_{k+1}) \leq \text{Sp}(A_k) \left[1 + (\alpha^2 - 1) \frac{\|y_k\|^2}{\|g(x_k)\|^2} \right]. \quad (16)$$

На основе соотношения между среднеарифметическим и среднегеометрическим собственных значений матрицы $A > 0$ имеем $\text{Sp}(A)/n \geq [\det(A)]^{1/n}$. Отсюда и из (16) получим

$$\frac{\text{Sp}(A_0)}{n} \prod_{i=0}^k \left[1 + (\alpha^2 - 1) \frac{\|y_i\|^2}{\|g(x_i)\|^2} \right] \geq \frac{\text{Sp}(A_{k+1})}{n} \geq (\det(A_{k+1}))^{1/n} = [(\alpha^2 \beta^2)^{k+1} \det(A_0)]^{1/n}.$$

Последнее неравенство на основе соотношения $1 + p \leq \exp(p)$ преобразуем к виду

$$\frac{\text{Sp}(A_0)}{n} \exp \left[(\alpha^2 - 1) \sum_{i=0}^k \frac{\|y_i\|^2}{\|g(x_i)\|^2} \right] \geq (\alpha^2 \beta^2)^{(k+1)/n} (\det(A_0))^{1/n}. \quad (17)$$

В силу условия (13)

$$\text{Sp}(A_0)/n \leq 1/m_0, \quad (\det(A_0))^{1/n} \geq 1/M_0.$$

Логарифмируя (17), с учетом последних неравенств, найдем

$$\sum_{i=0}^k \frac{\|y_i\|^2}{\|g(x_i)\|^2} \geq \frac{2(k+1) \ln(\alpha\beta)}{n(\alpha^2 - 1)} + \frac{\ln(m_0/M_0)}{(\alpha^2 - 1)},$$

что вместе с (10) доказывает (14). Теорема доказана.

Полученные оценки скорости сходимости не объясняют факт высокой скорости сходимости метода СМДМ, например, на квадратичных функциях. Для обоснования наличия ускоряющих свойств у метода нам необходимо показать его инвариантность относительно линейного преобразования координат, а затем использовать оценку (14) в системе координат, в которой отношение l/L максимально. Подобная возможность существует, например, в случае квадратичных функций, где это отношение будет равно 1.

Пусть задано линейное преобразование координат $\bar{x} = Px$, $\bar{x}, x \in R^n$, где \bar{x} – переменные новой системы координат, P – невырожденная матрица размера $n \times n$. Образует функцию $\bar{f}(\bar{x}) = f(P^{-1}\bar{x}) = f(x)$. Здесь и далее черта сверху – признак принадлежности одноименной переменной новой системе координат. Обозначим $P^{-T} = (P^T)^{-1}$. Установим соответствие между характеристиками процесса (3) – (5), применяемого для минимизации функций $\bar{f}(\bar{x})$ и $f(x)$.

Лемма 1. Пусть начальные условия процесса (3) – (5), применяемого для минимизации функций $\bar{f}(\bar{x})$ и $f(x)$, связаны равенствами

$$\bar{x}_0 = Px_0, \quad \bar{H}_0 = PH_0P^T. \quad (18)$$

Тогда характеристики этих процессов связаны соотношениями

$$\bar{f}(\bar{x}_k) = f(x_k), \quad \bar{x}_k = Px_k, \quad \bar{g}(\bar{x}_k) = P^{-T}g(x_k), \quad \bar{H}_k = PH_kP^T \quad (k = 0, 1, 2, \dots). \quad (19)$$

Доказательство. Для производных функций $\bar{f}(\bar{x})$ и $f(x)$ справедлива взаимосвязь $\bar{g}(\bar{x}) = P^{-T}g(x)$. Отсюда и предположения (18) следует (19) при $k = 0$. Предположим, что равенства (19) выполнены при всех $k = 0, 1, \dots, i$. Покажем их выполнимость при $k = i + 1$. Из (3) при $k = i$ после умножения на P слева с учетом доказанных равенств (19) получим

$$Px_{i+1} = Px_i - \gamma_i PH_i P^T P^{-T} g(x_i) = \bar{x}_i - \gamma_i \bar{H}_i \bar{g}(\bar{x}_i). \quad (20)$$

Отсюда, согласно определению функции \bar{f} , на этапе одномерной минимизации (4) выполняется равенство $\gamma_i = \bar{\gamma}_i$. Поэтому правая часть (20) – реализация шага (3) в новой системе координат. Следовательно:

$$Px_{i+1} = \bar{x}_{i+1}, \quad \bar{g}(\bar{x}_{i+1}) = P^{-T}g(x_{i+1}) \quad \text{и} \quad \bar{y}_i = \bar{g}(\bar{x}_{i+1}) - \bar{g}(\bar{x}_i) = P^{-T}y_i. \quad (21)$$

Помножая (5) слева на P , а справа на P^T , с учетом (21) получим

$$\begin{aligned} PH_{i+1}P^T &= PH_iP^T - (1 - \frac{1}{\alpha^2}) \frac{PH_iP^TP^{-T}y_iy_i^TP^{-1}PH_i^TP^T}{(y_i, P^{-1}PH_iP^TP^{-T}y_i)} - \\ &\quad - (1 - \frac{1}{\beta^2}) \frac{PH_iP^TP^{-T}p_ip_i^TP^{-1}PH_i^TP^T}{(p_i, P^{-1}PH_iP^TP^{-T}p_i)} = \\ &= \bar{H}_i - (1 - \frac{1}{\alpha^2}) \frac{\bar{H}_i\bar{y}_i\bar{y}_i^T\bar{H}_i^T}{(\bar{H}_i\bar{y}_i, \bar{y}_i)} - (1 - \frac{1}{\beta^2}) \frac{\bar{H}_i\bar{p}_i\bar{p}_i^T\bar{H}_i^T}{(\bar{H}_i\bar{p}_i, \bar{p}_i)}, \end{aligned}$$

где правая часть есть реализация формулы (5) в новой системе координат. Поэтому $PH_{i+1}P^T = \bar{H}_{i+1}$. Следовательно, равенства (19) будут справедливы и при $k = i + 1$. Продолжая процесс индукции, получим доказательство леммы.

Обозначим через l_p, L_p соответственно константы сильной выпуклости и Липшица для функции $\bar{f}(\bar{x})$. Введем функцию $K(P) = l_p / L_p$. Обозначим V матрицу преобразования координат такую, что $K(V) \geq K(P)$ для произвольных невырожденных матриц P .

Теорема 3. Пусть функция $f(x)$ удовлетворяет условию A . Тогда для последовательности $\{f_k\}, k = 0, 1, 2, \dots$, заданной процессом (3) – (6), с ограниченной начальной матрицей H_0 (13) имеет место оценка

$$\mu_{k+1} \leq \mu_0 \exp \left\{ -\frac{l_V^2}{L_V^2} \left[\frac{2(k+1)\ln(\alpha\beta)}{n(\alpha^2-1)} + \frac{\ln(m/M)}{(\alpha^2-1)} \right] \right\}, \quad (22)$$

где m и M – соответственно минимальное и максимальные собственные значения матрицы $\bar{H}_0 = VH_0V^T$.

Доказательство. Согласно результатам леммы 1, мы можем выбрать произвольную систему координат для оценки скорости сходимости процесса минимизации (3) – (5). Поэтому используем оценку (14) в системе координат с матрицей $P = V$, получим оценку (22).

Для метода скорейшего спуска (схема (3), (4) при $H_k = I$) на функциях, удовлетворяющих условию A , порядок скорости сходимости определяется выражением $\mu_k \leq \mu_0 \exp(-kl/L)$ [14, 25]. При условии $l_V^2 / L_V^2 \gg l/L$ оценка (22) оказывается предпочтительнее. Такая ситуация возникает, например, при минимизации квадратичных функций, матрицы вторых производных которых имеют большой разброс собственных значений. Второе слагаемое оценки (22) характеризует этап настройки матрицы СМДМ-алгоритма. При больших значениях α настройка матрицы протекает интенсивнее. Таким образом, при конечных значениях параметра

растяжения пространства алгоритм СМДМ на сильно выпуклых функциях, без предположения существования вторых производных, обладает ускоряющими свойствами сравнительно с методом скорейшего спуска.

3. Результаты вычислительного эксперимента

1. **Исследование скорости сходимости алгоритма СМДМ.** Предварительный вычислительный эксперимент имеет целью сравнить скорости сходимости квазиньютоновского метода Бroyдена – Флетчера – Гольдфарба – Шанно (BFGS) [15] и СМДМ и на квадратичных функциях с высокой степенью обусловленности ($\mu = 10^{10}$). Вторая часть эксперимента состоит в соотношении скорости сходимости СМДМ на квадратичных и негладких функциях с равной степенью вытянутости функции:

$$f_1(x) = \sum_{i=1}^n x_i^2 \cdot (1 + (i-1)(10^5 - 1)/(n-1))^2, \quad x_{0,i} = 1, \quad x_i^* = 0, \quad i = 0, 1, 2, \dots, n,$$
$$f_2(x) = \sum_{i=1}^n |x_i| \cdot (1 + (i-1)(10^5 - 1)/(n-1)), \quad x_{0,i} = 1, \quad x_i^* = 0, \quad i = 0, 1, 2, \dots, n.$$

Обозначим it – число итераций метода, а nfg – количество вычислений функции и градиента, требуемые для достижения заданной точности $f_k - f^* \leq \varepsilon$. Результаты для методов приведены в табл. 1.

Таблица 1

Результаты сходимости методов на сложных функциях

| Функция | $f_1 (\varepsilon = 10^{-10})$ | $f_2 (\varepsilon = 10^{-5})$ | | $f_3 = f_1 + f_2 (\varepsilon = 10^{-5})$ |
|------------|--------------------------------|-------------------------------|---------------|-------------------------------------------|
| Методы | BFGS (it/nfg) | СМДМ (it/nfg) | СМДМ (it/nfg) | СМДМ (it/nfg) |
| $n = 500$ | 543 / 993 | 755 / 1267 | 2747 / 4883 | 2596 / 4600 |
| $n = 1000$ | 1049 / 1974 | 1330 / 2144 | 5451 / 9159 | 5178 / 8305 |
| $n = 5000$ | 5100 / 9873 | 5411 / 8321 | 20073 / 29607 | 18328 / 26514 |

Метод BFGS является конечным на квадратичных функциях с числом итераций, равным размерности задачи. Стратегии методов BFGS и СМДМ различные. В квазиньютоновских методах важна точность одномерного поиска, а в релаксационных субградиентных алгоритмах наоборот, чем больше окрестность поиска, тем эффективнее выбор направления для последующего выхода из этой окрестности. Результаты табл. 1 свидетельствуют о практической идентичности методов на квадратичных функциях с высокой степенью обусловленности и высокой эффективности СМДМ при минимизации сложных негладких функций f_2 и f_3 . В представленных ниже сценариях обучения ИНС требуется высокая точность решения задач негладкой минимизации. Данный эксперимент дает определенные гарантии возможности СМДМ решать подобные задачи.

2. **Задача аппроксимации двухслойной ИНС сигмоидального типа.** ИНС представляют собой мощный инструмент аппроксимации и находят применение в различных областях, в том числе и при решении уравнений математической физики [6, 7]. Требования к аппарату приближения – это надежность и качество приближения. Ниже будут изложены способы решения подобных проблем, где важ-

ную роль играет исследованный выше релаксационный субградиентный метод с двухранговой коррекцией матриц метрики СМДМ.

Рассмотрим задачу аппроксимации

$$w^* = \arg \min_w E(\alpha, w, D),$$

$$E(\alpha, w, D) = \sum_{x, y \in D} (y - f(x, w))^2 + \sum_{i=1}^k \alpha_i R_i(w), \quad (23)$$

где $D = \{(x^i, y_i) | x^i \in R^p, y_i \in R^1\}$, $i = 1, \dots, N$ – данные наблюдения, $R_i(w)$ – различные виды регуляризаторов; α_i – параметры регуляризации, $f(x, w)$ – аппроксимирующая функция; $x \in R^p$ – вектор данных; $w \in R^n$ – вектор настраиваемых параметров, p и n – их размерности. В качестве регуляризаторов можно использовать следующие:

$$R_2(w) = \sum_{i=1}^n w_i^2 \quad [8], \quad R_1(w) = \sum_{i=1}^n |w_i| \quad [26],$$

$$R_\gamma(w) = \sum_{i=1}^n (|w_i| + \varepsilon)^\gamma \quad (\varepsilon = 10^{-6}, \gamma = 0.7) \quad [9].$$

Использование R_2 приводит к подавлению в большей мере больших компонент вектора w , R_1 – больших и малых, а R_γ – преимущественно малых. Подобное свойство R_γ позволяет сводить к нулю слабые компоненты, несущественные для описания данных. В задачах приближения ИНС в отсутствие помех мы будем использовать регуляризатор R_γ .

В задаче аппроксимации сетью прямого распространения требуется по данным D обучить двухслойную сигмоидальную нейронную сеть следующего вида (оценить ее неизвестные параметры w):

$$f(x, w) = w_0^{(2)} + \sum_{i=1}^m w_i^{(2)} \varphi(s_i), \quad \varphi(s) = s / (1 + |s|),$$

$$s_i = w_{i0}^{(1)} + \sum_{j=1}^p x_j w_{ij}^{(1)}, \quad i = 1, 2, \dots, m, \quad (24)$$

где x_j – компоненты вектора $x \in R^p$, $w = ((w_i^{(2)}, i = 0, \dots, m), (w_{ij}^{(1)}, j = 0, \dots, p, i = 1, \dots, m))$ – набор неизвестных параметров, которые необходимо оценить методом наименьших квадратов (23), $\varphi(s)$ – функция активации нейрона, m – число нейронов. Для решения задач (23) используем субградиентный метод СМДМ.

3. Оптимизационный алгоритм нахождения начального приближения ИНС. Начальное приближение в задаче обучения ИНС играет решающую роль. В литературе по нейронным сетям [2, 3] предлагается задавать начальные значения параметров нейронов w случайным образом. Рассмотрим процесс задания начальных параметров сети, в котором каждому нейрону отводится зона активного приближения данных и при этом зоны нейронов покрывают область данных.

Рабочие области нейронов $\varphi(s)$ в (24) имеют характер активной зависимости только в некоторой окрестности значений $s = 0$, а при значительных отклонениях значений s от нуля значения $\varphi(s)$ близки к своим асимптотам, принимающим значения $\{-1, 1\}$. Важно иметь такие параметры нейронов w , которые обеспечивают для векторов области данных $x \in R^p$ принадлежность рабочей области хотя бы одного нейрона.

При произвольном задании начальных параметров в задаче минимизации (23) зачастую оказывается, что рабочие области нейронов охватывают только часть области аппроксимации либо выходят за ее пределы, образуя локальные минимумы, выход из которых нельзя осуществить приемами локальных изменений текущего приближения. Даже если предположить, что рабочие области нейронов расположены правильно, нельзя гарантировать, что их положение сохранится при дальнейшем решении задачи обучения. Расположение нейронов в точках с высокой концентрацией данных также не обеспечивает сохранения этого положения, поскольку при дальнейшем обучении нейроны могут покинуть изначально заданные области. Поэтому требуется дополнительная привязка рабочих областей нейронов посредством обучения нейросети при фиксированных центрах. В этом случае нейрон сможет покинуть свой регион только в случае, когда в этом регионе будет обеспечена уже имеющаяся точность приближения данных.

В следующем алгоритме предлагается найти приближение ИНС, т.е. параметры нейронов при фиксированном положении рабочих областей нейронов с помощью заданных центров $c_i \in R^p$, $i = 1, 2, \dots, m$, в области аппроксимации $x \in R^p$, определяемой данными. В этом случае в (24) будут использоваться выражения

$$s_i = \sum_{j=1}^p (x_j - c_{ij}) w_{ij}^{(1)}, \quad i = 1, 2, \dots, m, \\ w = ((w_i^{(2)}, i = 0, \dots, m), (w_{ij}^{(1)}, j = 1, \dots, p, i = 1, \dots, m)). \quad (25)$$

Центры c_i можно найти некоторым алгоритмом кластеризации данных $x^i \in R^p$, $i = 1, \dots, N$, что полезно и с точки зрения расположения нейронов в областях с высокой плотностью данных. В этой работе использовался максиминный алгоритм [27], в котором в качестве первых двух центров выбираются две максимально удаленные друг от друга точки данных. Каждый новый центр получается выбором точки данных x^i , расстояние от которой до ближайшего известного центра максимально.

Оптимизационный алгоритм нахождения начального приближения ИНС (ОНП).

1. Задать данные D , число нейронов $m < N$. Выбрать регуляризатор и его параметры в (23).
2. На данных D определить центры рабочих областей нейронов $c_i \in R^p$, $i = 1, 2, \dots, m$.
3. Выбрать начальное приближения параметров ИНС (24) в форме (25).
4. Для ИНС (24) в форме (25) найти неизвестные параметры посредством решения задачи (23).
5. Вернуться к исходному описанию сети в виде (24) посредством образования параметров

$$w_{i0}^{(1)} = -\sum_{j=1}^p c_{ij} w_{ij}^{(1)}, \quad i = 1, 2, \dots, m. \quad (26)$$

Пункт 4 алгоритма в определенной степени гарантирует, что область данных будет покрыта рабочими областями нейронов. В своей выделенной области каждый нейрон обеспечит некоторое качество приближения, которое при возврате (26) к виду (24) сохраняется, а при дальнейшем обучении может только улучшиться.

4. Алгоритм обучения ИНС в задачах аппроксимации. При малой размерности данных $x \in R^p$, например, при решении уравнений математической физики [6], можно обойтись без удаления переменных внутри нейронов, а сосредоточится на выборе оптимального числа нейронов, удаляя избыточные. В следующем алгоритме задается избыточное число нейронов, а регуляризация проводится только по параметрам $w_i^{(2)}$, $i = 1, \dots, m$, из (24) с использованием регуляризатора $R\gamma$.

Алгоритм обучения ИНС в задачах аппроксимации при отсутствии помех (A0).

1. Задать данные D , число нейронов $m < N$. Выбрать регуляризатор $R\gamma(w)$ и параметр α для алгоритма ОНП и для алгоритма обучения нейросети (23).

2. Найти начальное приближение ИНС W_0 , используя алгоритм ОНП.

3. Для $k=1, 2, \dots, m-1$ выполнить действия:

3.1. $w^k = \arg \min_w E_\Omega(\alpha, w^{k-1}, D)$. Вычислить величину среднеквадратичной погрешности

$$S_k = S(D, f_k) = \sum_{x, y \in D} (y - f(x, w^k))^2 / N. \quad (27)$$

3.2. Последовательно по одному удалить нейроны, обеспечивающие минимальное после удаления значение показателя $S(D, f)$, не превосходящее значение S_k более чем на заданное число процентов.

3.3. Если в пункте 3.2 не произошло удаления нейронов, то удалить один из нейронов, приводящий к наименьшему росту показателя $S(D, f)$.

4. В качестве окончательной модели аппроксимации выбрать ИНС $f(x, w^k)$ с числом параметров n , не превосходящим N , имеющую наименьшее значение показателя S_k .

Первоначально алгоритм, подобный A0, по аналогии с методом построения компактной линейной модели [9], не содержал пункта 2. Для получения качественной модели приходилось многократно применять алгоритм со случайным выбором начального приближения ИНС. При этом не всегда удавалось достигнуть необходимого качества. Сочетание первоначального равномерного покрытия области данных рабочими областями избыточного числа нейронов с последующим удалением избыточных нейронов средствами негладкой регуляризации позволило получать качественные приближения за один просчет алгоритма A0.

Обладание техникой размещения рабочих зон сигмоидальных нейронов в нужных областях данных и способом удаления избыточных нейронов позволяет построить другие разновидности алгоритма A0, например, с последовательным добавлением нейронов в областях данных с низким качеством приближения на предыдущих этапах. При этом негладкая регуляризация позволит исключить из модели малоинформативные нейроны.

5. Примеры решения задач аппроксимации. Обоснование эффективности изложенных алгоритмов проведем на примерах функций, для которых известны результаты аппроксимации ИНС [3]. Будем использовать следующую функцию активации нейрона $\phi(s) = 1/(1 + \exp(-s))$. Зададим параметр регуляризации $\alpha = 10^{-9}$. При решении задач алгоритмом A0 в качестве решения будем выбирать ИНС с наименьшим значением показателя (27) при условии $n < N$.

В [3, с. 149) на данных при $N = 625$, сформированных в области $\Omega = [-3, 3] \times [-3, 3]$ датчиком равномерных случайных чисел, аппроксимировалась функция

$$f_3(x_1, x_2) = 3(1 - x_1)^2 \exp(-x_1^2 - (x_2 + 1)) - \\ - 10(x_1 / 5 - x_1^3 - x_2^5) \exp(x_1^2 - x_2^2) - \exp(-(x_1 + 1) - x_2^2) / 3.$$

Максимальное уклонение построенной в [3] ИНС, основанной на радиальных базисных функциях (RBF), на проверочной выборке из 1000 данных составило $\Delta_{1000} = 0.06$ [3]. Функция f_3 – типичный пример удобной для аппроксимации сетью RBF функции. Тем не менее использование алгоритма А0 на меньшей по размеру выборке ($N = 600$) позволяет получить сигмоидальную сеть с меньшим уклонением $\Delta_{1000} = 0.0171$.

В той же работе [3, с. 162] в области $\Omega = [-1, 1] \times [-1, 1]$ аппроксимировалась функция $f_4(x_1, x_2) = \sin(\pi x_1^2) \sin(2\pi x_2) / 2$. На основе выборки с $N = 500$ получено $\Delta_{1000} = 0.15$ [3]. Алгоритм А0 при меньшем количестве данных $N = 150$ позволяет получить сигмоидальную ИНС, для которой максимальное уклонение почти на порядок меньше $\Delta_{1000} = 0.018$.

Отметим, что при аппроксимации функций f_3 и f_4 сигмоидальной нейронной сетью без использования оптимизационного алгоритма нахождения начального приближения не удавалось получить качество аппроксимации выше, чем в [3].

В работе [6] ИНС применялись для решения уравнений математической физики. Использовался метод доверительных областей [28], в котором накладываются ограничения на область изменения параметров ИНС при обучении. В силу сложности и низкой скорости сходимости используемого метода в [13] предпринята попытка найти наиболее подходящий алгоритм для обучения сетей RBF. Среди исследуемых алгоритмов в [13] присутствовали и эффективные методы обучения глубоких нейронных сетей [29, 30]. На функции $f_5(x_1, x_2) = x_1^2 + x_2^2$ в области $\Omega = [-3, 3] \times [-3, 3]$ на равномерно распределенных в области данных при $N = 100$ лучшим оказался метод Левенберга – Марквардта. При этом достигнутая величина среднеквадратичной погрешности на обучающей выборке составляет $S_{100} = 10^{-6}$ [13]. Для сигмоидальной ИНС, полученной алгоритмом А0, имеем на обучающей выборке $S_{100} = 1.55 \cdot 10^{-11}$, а на тестовой выборке – $S_{1000} = 5.3 \cdot 10^{-10}$, что на несколько порядков превосходит имевший место результат.

Приведенные результаты сведены в табл. 2, где m – число нейронов аппроксимирующей сети, m_0 – первоначальное число нейронов. Остальные обозначения введены ранее.

Т а б л и ц а 2

Результаты аппроксимации нейросетями

| Функция | Известные результаты | | | Полученные результаты | | | |
|---------|----------------------|-----|------------------------|-----------------------|-------|-----|----------------------------------------------------------------|
| | N | m | Результат | N | m_0 | m | Результат |
| f_3 | 625 | 36 | $\Delta_{1000} = 0.06$ | 600 | 70 | 64 | $\Delta_{1000} = 0.0171$ |
| f_4 | 500 | 41 | $\Delta_{1000} = 0.15$ | 150 | 70 | 48 | $\Delta_{1000} = 0.018$ |
| f_5 | 100 | 16 | $S_{100} = 10^{-6}$ | 100 | 30 | 16 | $S_{100} = 1.55 \cdot 10^{-11}; S_{1000} = 5.3 \cdot 10^{-10}$ |

Таким образом, оптимизационный алгоритм отыскания начального приближения сети вместе с процедурой подавления избыточных нейронов позволяет получать ИНС высокого качества. Высокая точность решения задачи минимизации и скорость сходимости метода СМДМ дают возможность эффективно реализовать этапы алгоритма А0.

Заключение

Доказано, что на сильно выпуклых функциях с липшицевым градиентом релаксационный субградиентный метод с двухранговой коррекцией матриц метрики сходится линейно, а преобразование метрики пространства в алгоритме обеспечивает его ускоряющие свойства. Вычислительный эксперимент устанавливает близость свойств скорости сходимости изучаемого алгоритма и квазиньютоновских методов на квадратичных функциях. Метод обладает высокой скоростью сходимости и на негладких функциях.

Предложен комплекс алгоритмов построения ИНС в условиях небольших по размеру обучающих выборок. Сюда входит оптимизационный алгоритм нахождения начального приближения ИНС, который состоит в закреплении рабочих областей нейронов в области данных посредством построения первоначальной сети с фиксированными центральными линиями сигмоидальных нейронов. В основной схеме построения ИНС используется негладкая регуляризация, необходимая для целей устранения эффектов переобучения и удаления малозначимых нейронов. Приводимые примеры решения задач построения ИНС позволяют сделать заключение об эффективности предложенных в работе алгоритмов. Высокая скорость сходимости на гладких и негладких функциях алгоритма минимизации СМДМ дает возможность эффективно решать задачи минимизации в схемах построения ИНС.

Авторы считают своим долгом выразить признательность анонимным рецензентам, замечания и комментарии которых позволили существенным образом улучшить изложение результатов.

ЛИТЕРАТУРА

1. Воронцов К.В. Курс лекций «Математические методы обучения по прецедентам» URL: <http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf>
2. Хайкин С. Нейронные сети: полный курс. М.: Вильямс, 2006. 1104 с.
3. Осовский С. Нейронные сети для обработки информации. М.: Горячая линия – Телеком, 2016. 448 с.
4. Горбань А.Н. Обучение нейронных сетей. М.: Изд-во СССР – США СП «Параграф», 1990. 160 с.
5. Бурнаев Е.В., Приходько П.В. Об одной методике построения ансамблей регрессионных моделей // Автомат. и телемех. 2013. Вып. 10. С. 36–54.
6. Горбаченко В.И., Жуков М.В. Решение краевых задач математической физики с помощью сетей радиальных базисных функций // Журнал вычислительной математики и математической физики. 2017. Т. 57. № 1. С. 133–143.
7. Кретинин А.В. Метод взвешенных невязок на базе нейросетевых пробных функций для моделирования задач гидродинамики // Сиб. журн. вычисл. матем. 2006. Т. 9. № 1. С. 23–35.
8. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1986.
9. Крутиков В.Н., Арышев Д.В. Алгоритм последовательного отсева неинформативных переменных линейной модели // Вестник Кемеровского государственного университета. 2004. № 3(7). С. 124–129.

10. Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine // *Statistica Sinica*. V. 16. No. 2. P. 589–615.
11. Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective Support Vector Machines // *Proc. of the 19th Int. Conf. on Pattern Recognition*, Vol. 1–6, IEEE, ISBN 978-1-4244-2174-9. 2008. P. 2336–2339.
12. Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities // *Multiple Classifier Systems. Lecture Notes In Computer Science*. V. 5997. Berlin; Heidelberg: Springer-Verlag, 2010. P. 165–174.
13. Алкезуини М.М., Горбаченко В.И. Совершенствование алгоритмов обучения сетей радиальных базисных функций для решения задач аппроксимации // *Модели, системы, сети в экономике, технике, природе и обществе*. 2017. № 3 (23). С. 123–138.
14. Поляк Б.Т. Введение в оптимизацию. М.: Наука, 1983.
15. Дэннис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. М.: Мир, 1988.
16. Marquardt D.W. An algorithm for least-squares estimation of nonlinear parameters // *J. Society for Industrial and Applied Mathematics*. 1963. V. 11. No 2. P. 431–441.
17. Шор Н.З. Методы минимизации недифференцируемых функций и их приложения. Киев: Наукова думка, 1979.
18. Wolfe Ph. Note on a method of conjugate subgradients for minimizing nondifferentiable functions // *Math. Program*. 1974. V. 7. No. 1. P. 380–383.
19. Lemarechal C. An extension of Davidson methods to non-differentiable problems // *Math. Program. Study*. 1975. V. 3. P. 95–109.
20. Нурминский Е.А., Тьен Д. Метод сопряженных субградиентов с ограниченной памятью // *Автомат. и телемех.* 2014. № 4. P. 67–80; *Autom. Remote Control*. 2014. V. 75. No. 4. P. 646–656.
21. Крутиков В.Н., Вершинин Я.Н. Многошаговый субградиентный метод для решения негладких задач минимизации высокой размерности // *Вестник Томского государственного университета. Математика и механика*. 2014. № 3. С. 5–19.
22. Крутиков В.Н., Вершинин Я.Н. Субградиентный метод минимизации с коррекцией векторов спуска на основе пар обучающих соотношений // *Вестник Кемеровского государственного университета*. 2014. Т.1. № 1 (57). С. 46–54. DOI: <https://doi.org/10.21603/2078-8975-2014-1-46-54>
23. Крутиков В.Н., Горская Т.А. Семейство релаксационных субградиентных методов с двухранговой коррекцией матриц метрики // *Экономика и мат. методы*. 2009. Т. 45. Вып. 4. С. 37–80.
24. Крутиков В.Н., Петрова Т.В. Релаксационный метод минимизации с растяжением пространства в направлении субградиента // *Экономика и мат. методы*. 2003. Т. 39. Вып. 1. С. 106–119.
25. Карманов В.Г. Математическое программирование. М.: Наука, 1980. 256 с.
26. Tibshirani R.J. Regression shrinkage and selection via the lasso // *J. Royal Statistical Society. Series B (Methodological)*. 1996. V. 58. No. 1. P. 267–288.
27. Гу Дж., Гонсалес Р. Принципы распознавания образов. М.: Мир, 1978.
28. Conn A.R., Gould N.I.M., Toint P.L. Trust regions methods. Society for Industrial and Applied Mathematics, 2000. 959 p.
29. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. – М.: ДМК Пресс, 2017. 652 с.
30. Sutskever I., Martens J., Dahl G., Hinton G. On the importance of initialization and momentum in deep learning // *Proc. 30th Int. Conf. on Machine Learning*. V. 28. Atlanta, Georgia, 2013. P. 1139–1147.

Krutikov V.N., Samoylenko N.S. (2018) ON THE CONVERGENCE RATE OF THE SUBGRADIENT METHOD WITH METRIC VARIATION AND ITS APPLICATIONS IN NEURAL NETWORK APPROXIMATION SCHEMES *Vestnik Tomskogo gosudarstvennogo universiteta. Matematika i mekhanika* [Tomsk State University Journal of Mathematics and Mechanics]. 55. pp. 22–37

DOI 10.17223/19988621/55/3

Keywords: method, subgradient, minimization, rate of convergence, neural networks, regularization.

In this paper, the relaxation subgradient method with rank 2 correction of metric matrices is studied. It is proven that, on high-convex functions, in the case of the existence of a linear coordinate transformation reducing the degree of the task casualty, the method has a linear convergence rate corresponding to the casualty degree. The paper offers a new efficient tool for choosing the initial approximation of an artificial neural network. The use of regularization allowed excluding the overfitting effect and efficiently deleting low-significant neurons and intra-neural connections. The ability to efficiently solve such problems is ensured by the use of the subgradient method with metric matrix rank 2 correction. It has been experimentally proved that the convergence rate of the quasi-Newton method and that of the method under research are virtually equivalent on smooth functions. The method has a high convergence rate on non-smooth functions as well. The method's computing capabilities are used to build efficient neural network learning algorithms. The paper describes an artificial neural network learning algorithm which, together with the redundant neuron suppression, allows obtaining reliable approximations in one count.

AMS Mathematical Subject Classification: 65K05, 90C30, 82C32

KRUTIKOV Vladimir Nikolayevich (Doctor of Technical Science, Professor at the Department of Applied Mathematics, Kemerovo State University, Kemerovo, Russian Federation).

SAMOYLENKO Natalya Sergeevna (Department of Applied Mathematics, Kemerovo State University, Kemerovo, Russian Federation).

REFERENCES

1. Vorontsov K.V. *Course of lectures "Mathematical methods of learning by precedents"* <http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf>
2. Khaikin S. (2006) *Neyronnyye seti: polnyy kurs* [Neural networks: a full course]. Moscow: Williams. 1104 p.
3. Osovski S. (2016) *Neyronnyye seti dlya obrabotki informatsii* [Neural networks for information processing]. Moscow: Goryachaya liniya – Telecom. 448 p.
4. Gorban A.N. (1990) *Obucheniye neyronnykh setey* [Training of neural networks]. Moscow: USSR – USA JV Paragraph. 160 p.
5. Burnaev E.V., Prikhodko P.V. (2013) On a method for constructing ensembles of regression models. *Autom. Remote Control*. 74(10). pp. 1630–1644. DOI: <https://doi.org/10.1134/S0005117913100044>.
6. Gorbachenko V.I., Zhukov M.V. (2017) Solving boundary value problems of mathematical physics using radial basis function networks. *Comput. Math. and Math. Phys.* 57(1). pp. 145–155. DOI: <https://doi.org/10.1134/S0965542517010079>.
7. Kretinin A.V. (2006) Metod vzveshennykh nevyazok na baze neyrosetevykh probnykh funktsiy dlya modelirovaniya zadach gidrodinamiki [The weighted residuals method based on neural net trial functions for simulation of hydrodynamics problems]. *Siberian J. Num. Math.* 9(1). pp. 23–35.
8. Tikhonov A.N., Arsenin V.Ya. (1986) *Metody resheniya nekorrektnykh zadach* [Methods for solving ill-posed problems]. Moscow: Nauka.

9. Krutikov V.N., Aryshev D.V. (2004) Algorithm posledovatel'nogo otseva neinformativnykh peremennykh lineynoy modeli [Algorithm of sequential screening of non-informative variables of a linear model]. *Bulletin of Kemerovo State University*. 3(7). pp. 124–129.
10. Li Wang, Ji Zhu, Hui Zou. (2006) The doubly regularized support vector machine. *Statistica Sinica*. 16. pp. 589–615.
11. Tatarchuk A., Mottl V., Eliseyev A., Windridge D. (2008) Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective support vector machines. *Proceedings of the 19th International Conference on Pattern Recognition*. Vol. 1–6. pp. 2336–2339.
12. Tatarchuk A.I., Urlov E., Mottl V., Windridge D. (2010) A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. *Multiple Classifier Systems. Lecture Notes In Computer Science*. 5997. Springer-Verlag: Berlin, Heidelberg. pp. 165–174.
13. Alkezuini M.M., Gorbachenko V.I.. (2017) Sovershenstvovaniye algoritmov obucheniya setey radial'nykh bazisnykh funktsiy dlya resheniya zadach approksimatsii [Improving the training algorithms for the networks of radial basis functions for solving approximation problems]. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve – Models, systems, networks in economics, engineering, nature and society*. 3(23). pp. 123–138.
14. Polyak B.T. (1983) *Vvedeniye v optimizatsiyu* [Introduction to optimization]. Moscow: Nauka.
15. Dennis J., Schnabel R. (1988) *Chislennyye metody bezuslovnoy optimizatsii i resheniya nelineynykh uravneniy* [Numerical Methods for Unconstrained Optimization and Nonlinear Equations]. Moscow: Mir.
16. Marquardt D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*. 11(2). pp. 431–441.
17. Shore N.Z. (1979) *Metody minimizatsii nedifferentsiruyemykh funktsiy i ikh prilozheniya* [Minimization Methods for Non-Differentiable Functions]. Kiev: Naukova Dumka.
18. Wolfe Ph. (1974) Note on a method of conjugate subgradients for minimizing nondifferentiable functions. *Math. Program*. 7(1). pp. 380–383.
19. Lemarechal C. (1975) An extension of Davidon methods to non differentiable problems. *Math. Program. Study*. 3. pp. 95–109.
20. Nurminskii E.A., Thien D. (2014) Method of conjugate subgradients with constrained memory. *Autom. Remote Control*. 75(4). pp. 646–656. DOI: <https://doi.org/10.1134/S0005117914040055>.
21. Krutikov V.N., Vershinin Ya.N. (2014) Mnogoshagovyy subgradiyentnyy metod dlya resheniya nekladkikh zadach minimizatsii vysokoy razmernosti [The subgradient multistep minimization method for nonsmooth high-dimensional problems]. *Vestnik Tomskogo gosudarstvennogo universiteta. Matematika i mekhanika – Tomsk State University Journal of Mathematics and Mechanics*. 3(29). pp. 5–19.
22. Krutikov V.N., Vershinin Ya.N. (2014) Subgradiyentnyy metod minimizatsii s korrektsiyey vektorov spuska na osnove par obuchayushchikh sootnosheniy [Subgradient minimization method with descent vectors correction based on pairs of training relations]. *Bulletin of Kemerovo State University*. 1-1(57). pp. 46–54. DOI: <https://doi.org/10.21603/2078-8975-2014-1-46-54>.
23. Krutikov V.N., Gorskaya T.A. (2009) Semeystvo relaksatsionnykh subgradiyentnykh metodov s dvukhrangovoy korrektsiyey matrits metriki [A family of subgradient relaxation methods with rank 2 correction of metric matrices]. *Ekonomika i mat. metody – Economy and math. methods*. 45(4). pp. 105–120.
24. Krutikov V.N., Petrova T.V. (2003) Relaksatsionnyy metod minimizatsii s rastyazheniyem prostranstva v napravlenii subgradiyenta [Relaxation method of minimization with space extension in the subgradient direction]. *Ekonomika i mat. metody – Economy and math. methods*. 39(1). pp. 106–119.
25. Karmanov V.G. (1980) *Matematicheskoye programmirovaniye* [Mathematical programming]. Moscow: Nauka.

26. Tibshirani R.J. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 58(1). pp. 267–288.
27. Tou J.T., Gonzalez R.C. (1974) *Pattern recognition principles*. Reading, MA: Addison-Wesley.
28. Conn A.R., Gould N.I.M., Toint P.L. (2000) *Trust region methods*. Philadelphia PA: Society for Industrial and Applied Mathematics (SIAM). 959 p.
29. Goodfellow J. et al. (2016) *Deep Learning*. MIT Publ.
30. Sutskever I. et al. (2013) On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on International Conference on Machine Learning*. 28. Atlanta, Georgia. pp. 1139–1147.