УДК 004.032.26

Б.А. СКОРОХОД

ДИФФУЗНЫЕ АЛГОРИТМЫ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ ПРЯМОГО РАСПРОСТРАНЕНИЯ

Ключевые слова: *нейронные сети прямого распространения, алгоритмы обучения, расширенный фильтр Калмана.*

ВВЕДЕНИЕ

Нейронные сети прямого распространения (НСПР) широко используются в разных приложениях, связанных с такими задачами, как прогнозирование временных рядов, классификация данных, идентификация и управление нелинейными объектами. Обзоры по методам их решения приведены в [1–3].

Обучение НСПР может рассматриваться как задача минимизации среднеквадратической ошибки относительно неизвестных параметров, входящих в ее описание (весов и смещений) при заданном обучающем множестве. Алгоритм обратного распространения (АОР) разработан в [4] и успешно применяется для их обучения. Вместе с тем известна присущая АОР медленная скорость сходимости, что существенно затрудняет или делает практически невозможным использование алгоритма в сложных задачах.

В многочисленных публикациях, посвященных НСПР, предложены различные алгоритмы обучения, превосходящие АОР по скорости сходимости и получаемой точности аппроксимации. В [5, 6] применяются методы Левенберга–Марквардта и квазиньютоновские, использующие информацию о матрице производных второго порядка критерия качества. В [7–9] алгоритмы обучения основываются на расширенном фильтре Калмана (РФК), использующем аппроксимационную ковариационную матрицу ошибки оценивания. Для НСПР с линейной функцией активации (ФА) в выходном слое в [10–12] представлены алгоритмы, учитывающие сепарабельную структуру сети [13]. В соответствии с VP-алгоритмом (variable projection) [10] исходная задача оптимизации преобразуется к эквивалентной относительно нелинейно входящих параметров (весов и смещений скрытого слоя). При этом уменьшается как размерность задачи, так и ее обусловленность, что позволяет сократить количество итераций для получения решения. Вместе с тем такой подход может использоваться только в пакетном режиме, и даже при аддитивно входящих ошибках измерений выходов НСПР в преобразованный критерий они входят нелинейно. Кроме того, существенно усложняется процедура определения частных производных критерия по параметрам. В [11, 12] предлагается ELM (extreme learning machine) алгоритм, с помощью которого обучаются только линейно входящие параметры (веса выходного слоя), а нелинейные выбираются случайно, без учета обучающей выборки, что сокращает время обучения, но может приводить к невысокой точности аппроксимации.

В данной работе предлагаются и исследуются новые алгоритмы обучения двухслойных НСПР с нелинейными ФА в скрытом слое и линейными — в выходном. Известно, что такие нейронные сети могут обладать универсальными аппроксимирующими свойствами (для сигмоидных функций доказательство приведено в [14]). Алгоритмы основываются на РФК и учете сепарабельной структуры НСПР, как в VP- и ELM-алгоритмах, но при этом одновременно обучаются все нейроны. Более точно, линейно входящие веса интерпретируются как диффузные — случайные величины, имеющие нулевое математическое ожида-

ние и матрицу ковариации, пропорциональную произвольно большому параметру λ [15]. Находятся асимптотические представления РФК при $\lambda \to \infty$, которые мы называем диффузными алгоритмами обучения (ДАО). Показано, что они, в отличие от их прототипа РФК с большим, но конечным λ, обладают свойством робастности по отношению к накоплению ошибок округления, и из ДАО при определенных упрощающих предположениях следует ELM-алгоритм. Приведен численный пример, показывающий, что ДАО могут превосходить ELM-алгоритм по точности аппроксимации.

1. ОСНОВНЫЕ ДОПУЩЕНИЯ И СООТНОШЕНИЯ ДЛЯ РФК

Рассмотрим двухслойную НСПР с нелинейными ФА в скрытом слое и линейными в выходном:

$$y_{it} = \sum_{k=1}^{m} w_{ik} \sigma \left(\sum_{j=1}^{n-1} a_{kj} z_{jt} + b_k \right), \ i = 1, \ldots, r, \ t = 1, \ldots, N. \tag{1}$$

Здесь z_{jt}, $j = 1, \ldots, n$, — входы, y_{it}, $i = 1, \ldots, r$, — выходы, a_{kj}, b_k, $k = 1, \ldots, m$, $j = 1, \ldots, n-1$, — веса и смещения скрытого слоя, w_{ik}, $i = 1, 2, \ldots, r$, $k = 1, 2, \ldots, m$, — веса выходного слоя, $\sigma(x) \in C^1(R)$ — ФА, $C^1(R)$ — пространство дифференцируемых функций на действительной прямой, или, в более компактной форме, используются векторно-матричные обозначения

$$y_t = W(\sigma(\upsilon_1 z_t), \ldots, \sigma(\upsilon_m z_t))^{\mathrm{T}} = f(z_t, \beta, \alpha) \ \ t = 1, \ldots, N, \tag{2}$$

где $y_t = (y_{1t}, \ldots, y_{rt})^{\mathrm{T}} \in R^r$, $z_t = (z_{1t}, \ldots, z_{n-1t}, 1)^{\mathrm{T}} \in R^n$, $W = (w_1^{\mathrm{T}}, \ldots, w_r^{\mathrm{T}})^{\mathrm{T}} \in \in R^{r \times m}$, $w_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, $i = 1, 2, \ldots, r$, $\upsilon_k = (\upsilon_{k1}, \ldots, \upsilon_{kn})$, $k = 1, \ldots, m$, $\beta = (\upsilon_1^{\mathrm{T}}, \ldots, \upsilon_m^{\mathrm{T}})^{\mathrm{T}} \in R^{mn}$, $\alpha = (w_1^{\mathrm{T}}, \ldots, w_r^{\mathrm{T}})^{\mathrm{T}} \in R^{rm}$, R^l, $R^{l \times p}$ — пространства векторов и матриц размерностей l и $l \times p$ соответственно, $(\cdot)^{\mathrm{T}}$ — операция транспонирования соответствующей матрицы.

Обучение нейронных сетей при последовательной обработке и обучающем множестве $\{z_t, y_t\}_{t=1}^N$ рассматривается как задача оценки состояния нелинейной динамической системы

$$\beta_{t+1} = \beta_t, \ \alpha_{t+1} = \alpha_t,$$

$$y_t = f(z_t, \beta_t, \alpha_t) + \xi_t, \ t = 1, \ldots, N, \tag{3}$$

где $\xi_t \in R^r$ — случайный процесс с некоррелированными значениями, нулевым математическим ожиданием и матрицей ковариации $E[\xi_t \xi_t^{\mathrm{T}}] = V_t$, характеризующий ошибки измерения выхода. Начальное состояние (3) удовлетворяет следующим условиям:

A1) векторы β_1, α_1 случайны и некоррелированы между собой и ξ_t, $t = 1, \ldots, N$;

A2) задана выборка весов и смещений скрытого слоя нейронных сетей $m_\beta = (\overline{\beta}_1, \ldots, \overline{\beta}_{mn})^{\mathrm{T}}$ из непрерывного распределения с матрицей ковариации S_β, определяемая характером решаемой задачи и используемой ФА, при этом $\beta_1 = m_\beta$;

A3) априорная информация относительно весов выходного слоя α отсутствует, и они интерпретируются как случайные величины, имеющие нулевое математическое ожидание и матрицу ковариации, пропорциональную большему параметру $\lambda > 0$, т.е.

$$E(\alpha) = 0, \ E(\alpha\alpha^{\mathrm{T}}) = \lambda I_{rm}, \tag{4}$$

где $I_{rm} \in R^{rm \times rm}$ — единичная матрица.

Оценка состояния (3) $x_t = (\beta_t^{\mathrm{T}}, \alpha_t^{\mathrm{T}})^{\mathrm{T}}$ с помощью РФК удовлетворяет нелинейному разностному уравнению [16]

$$\hat{x}_{t+1} = \hat{x}_t + K_t(y_t - \tilde{f}(z_t, \hat{x}_t),\ \hat{x}_1 = (m_\beta^{\mathrm{T}}, 0_{rm}^{\mathrm{T}})^{\mathrm{T}},\ t = 1, \ldots, N, \tag{5}$$

где $\tilde{f}(z_t, x_t) = f(z_t, \beta_t, \alpha_t)$, $0_{rm} \in R^{rm}$ — вектор с нулевыми элементами,

$$K_t = P_t C_t^{\mathrm{T}} N_t^{-1},\ N_t = C_t P_t C_t^{\mathrm{T}} + V_t, \tag{6}$$

$$P_{t+1} = P_t - P_t C_t^{\mathrm{T}} N_t^{-1} C_t P_t,\ P_1 = \operatorname{block\,diag}(S_\beta, \lambda I_{rm}), \tag{7}$$

$$C_t = ((C_t^\beta)^{\mathrm{T}}, (C_t^\alpha)^{\mathrm{T}})^{\mathrm{T}},\ C_t^\beta = \left.\frac{\partial \tilde{f}(z_t, x_t)}{\partial \beta_t}\right|_{x_t = \hat{x}_t} \in R^{r \times mn},$$
$$C_t^\alpha = \left.\frac{\partial \tilde{f}(z_t, x_t)}{\partial \alpha_t}\right|_{x_t = \hat{x}_t} \in R^{r \times rm}, \tag{8}$$

матрица C_t определяется по выражениям

$$\frac{\partial y_{it}}{\partial v_k} = w_{ik}\sigma^{(1)}(v_k z_t) z_t,\ i = 1, \ldots, r,\ k = 1, \ldots, m, \tag{9}$$

$$\frac{\partial y_{it}}{\partial w_l} = (\sigma(v_1 z_t), \sigma(v_2 z_t), \ldots, \sigma(v_m z_t))^{\mathrm{T}},\ i = 1, \ldots, r, \tag{10}$$

$\sigma^{(1)}(x)$ — производная $\sigma(x)$.

При пакетной обработке и обучающем множестве $\{z_t, y_t\}_{t=1}^N$ обучение рассматривается как задача оценки состояния нелинейной динамической системы

$$\beta_{t+1} = \beta_t,\ \alpha_{t+1} = \alpha_t,\quad Y_t = F(Z, \beta_t, \alpha_t) + \xi_t,\ t = 1, \ldots, M, \tag{11}$$

где $F(Z, \beta_t, \alpha_t) = (f^{\mathrm{T}}(z_1, \beta_t, \alpha_t), \ldots, f^{\mathrm{T}}(z_N, \beta_t, \alpha_t))^{\mathrm{T}}$, $Z = (z_1^{\mathrm{T}}, \ldots, z_N^{\mathrm{T}})^{\mathrm{T}}$, M — количество итераций. Предполагается, что начальное состояние (11), как и при последовательном обучении, удовлетворяет условиям А1)–А3).

Оценка состояния (11) с помощью РФК удовлетворяет нелинейному разностному уравнению

$$\hat{x}_{t+1} = \hat{x}_t + K_t(Y - \tilde{F}(Z, \hat{x}_t),\ \hat{x}_1 = (\bar{\beta}^{\mathrm{T}}, 0_{rm}^{\mathrm{T}})^{\mathrm{T}},\ t = 1, \ldots, M, \tag{12}$$

где $\tilde{F}(Z, \hat{x}_t) = F(Z, \beta_t, \alpha_t)$, $Y = (y_1^{\mathrm{T}}, \ldots, y_N^{\mathrm{T}})^{\mathrm{T}}$,

$$K_t = P_t C_t^{\mathrm{T}} N_t^{-1},\ N_t = C_t P_t C_t^{\mathrm{T}} + V_t, \tag{13}$$

$$P_{t+1} = P_t - P_t C_t^{\mathrm{T}} N_t^{-1} C_t P_t,\ P_1 = \operatorname{block\,diag}(S_\beta, \lambda I_{rm}), \tag{14}$$

$$C_t = ((C_t^\beta)^{\mathrm{T}}, (C_t^\alpha)^{\mathrm{T}})^{\mathrm{T}},\ C_t^\beta = \left.\frac{\partial \tilde{F}(Z, \hat{x}_t)}{\partial \beta_t}\right|_{x_t = \hat{x}_t} \in R^{Nr \times mn}, \tag{15}$$

$$C_t^\alpha = \left.\frac{\partial \tilde{F}(Z, \hat{x}_t)}{\partial \alpha_t}\right|_{x_t = \hat{x}_t} \in R^{Nr \times rm},$$

матрица C_t определяется по выражениям (9), (10).

2. АСИМПТОТИКА РФК

Рассмотрим несколько вспомогательных утверждений.

Лемма 1. При $t = 1, \ldots, N$ справедливы представления

$$P_t = \widetilde{S}_t + \widetilde{R}_t M_t^{-1} \widetilde{R}_t^{\mathrm{T}}, \tag{16}$$

$$K_t = (\widetilde{S}_t + \widetilde{R}_{t+1} M_{t+1}^{-1} \widetilde{R}_t^{\mathrm{T}}) C_t^{\mathrm{T}} N_{1t}^{-1}, \tag{17}$$

где

$$\widetilde{S}_t = \begin{pmatrix} S_t & 0_{mn \times rm} \\ 0_{rm \times mn} & 0_{rm \times rm} \end{pmatrix}, \ \widetilde{R}_t = \begin{pmatrix} R_t \\ I_{rm} \end{pmatrix}, \tag{18}$$

$$S_{t+1} = S_t - S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} C_t^{\beta} S_t, \ N_{1t} = C_t^{\beta} S_t (C_t^{\beta})^{\mathrm{T}} + V_t, \ S_1 = S_{\beta}, \tag{19}$$

$$R_{t+1} = (I - S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} C_t^{\beta}) R_t - S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} C_t^{\alpha}, \ R_1 = 0_{mn \times rm}, \tag{20}$$

$$M_{t+1} = M_t + (C_t^{\beta} R_t + C_t^{\alpha})^{\mathrm{T}} N_{1t}^{-1} (C_t^{\beta} R_t + C_t^{\alpha}), \ M_1 = \frac{I_{rm}}{\lambda}, \ \lambda > 0. \tag{21}$$

Доказательство. Пусть $\widetilde{P}_t$ и $\overline{P}_t$ — два произвольных решения (7). Тогда [17] $Q_t = \widetilde{P}_t - \overline{P}_t$, $t = 1, \ldots, N$, где

$$Q_{t+1} = A_t Q_t A_t^{\mathrm{T}} - A_t Q_t C_t^{\mathrm{T}} N_{2t}^{-1} C_t Q_t A_t, \ Q_1 = \widetilde{P}_1 - \overline{P}_1, \tag{22}$$

$$A_t = I - \overline{P}_t C_t^{\mathrm{T}} N_{3t}^{-1} C_t, \ N_{2t} = C_t \widetilde{P}_t C_t^{\mathrm{T}} + V_t, \ N_{3t} = C_t \overline{P}_t C_t^{\mathrm{T}} + V_t.$$

Положим $\widetilde{P}_1 = P_1$, $\overline{P}_1 = \operatorname{block\,diag}(S_{\beta}, 0)$. Тогда $\widetilde{P}_t = P_t$, $\overline{P}_t = \widetilde{S}_t$ и $P_t = \widetilde{S}_t + Q_t$. Покажем, что $Q_t = \widetilde{R}_t M_t^{-1} \widetilde{R}_t^{\mathrm{T}}$. Так как

$$\widetilde{R}_{t+1} = A_t \widetilde{R}_t, \ \widetilde{R}_1 = (e_p, \ldots, e_q), \tag{23}$$

где $p = mn + 1$, $q = mn + rm$, $e_i \in R^q$ — i-единичный вектор, то

$$\widetilde{R}_{t+1} M_{t+1}^{-1} \widetilde{R}_{t+1}^{\mathrm{T}} = \widetilde{R}_{t+1} M_t^{-1} \widetilde{R}_{t+1}^{\mathrm{T}} - \widetilde{R}_{t+1} M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{2t}^{-1} C_t \widetilde{R}_t M_t^{-1} \widetilde{R}_{t+1}^{\mathrm{T}}.$$

Это равенство выполняется, при условии, что M_t^{-1} удовлетворяет разностному уравнению

$$M_{t+1}^{-1} = M_t^{-1} - M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{2t}^{-1} C_t \widetilde{R}_t M_t^{-1}. \tag{24}$$

Преобразовывая его с помощью тождества [15]

$$(P^{-1} + H^{\mathrm{T}} R^{-1} H)^{-1} = P - P H^{\mathrm{T}} (H P H^{\mathrm{T}} + R)^{-1} H P,$$

с $P = M_t^{-1}$, $H = C_t \widetilde{R}_t$, $R = N_{2t}$, получим

$$M_{t+1}^{-1} = (M_t + \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{2t}^{-1} C_t \widetilde{R}_t)^{-1}, \tag{25}$$

отсюда следуют (16).

Докажем (17). Сначала покажем, что

$$K_t = P_t C_t^{\mathrm{T}} N_t^{-1} = \widetilde{S}_t C_t^{\mathrm{T}} N_{3t}^{-1} + A_t Q_t C_t^{\mathrm{T}} N_{2t}^{-1}.$$

Так как $P_t = \widetilde{S}_t + Q_t$, то должно быть $(\widetilde{S}_t + Q_t) C_t^{\mathrm{T}} = \widetilde{S}_t C_t^{\mathrm{T}} N_{3t}^{-1} N_{2t} + A_t Q_t C_t^{\mathrm{T}}$. Преобразовывая правую часть этого выражения, устанавливаем, что

$$\widetilde{S}_t C_t^{\mathrm{T}} N_{3t}^{-1} N_{2t} + A_t Q_t C_t^{\mathrm{T}} = \widetilde{S}_t C_t^{\mathrm{T}} (I_m + N_{3t}^{-1} C_t Q_t C_t^{\mathrm{T}}) +$$

$$+ (I - \widetilde{S}_t C_t^{\mathrm{T}} N_{3t}^{-1} C_t) Q_t C_t^{\mathrm{T}} = (\widetilde{S}_t + Q_t) C_t^{\mathrm{T}}.$$

Используя (23), (25), находим

$$A_t Q_t C_t^{\mathrm{T}} N_{2t}^{-1} = A_t \widetilde{R}_t M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{2t}^{-1} =$$

$$= \widetilde{R}_{t+1} M_{t+1}^{-1} (M_t + \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{1t}^{-1} C_t \widetilde{R}_t) M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{2t}^{-1} =$$

$$= \widetilde{R}_{t+1} M_{t+1}^{-1} (\widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} + \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{1t}^{-1} C_t \widetilde{R}_t M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}}) N_{2t}^{-1} =$$

$$= \widetilde{R}_{t+1} M_{t+1}^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{1t}^{-1} (N_{1t} + C_t \widetilde{R}_t M_t^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}}) N_{2t}^{-1} = \widetilde{R}_{t+1} M_{t+1}^{-1} \widetilde{R}_t^{\mathrm{T}} C_t^{\mathrm{T}} N_{1t}^{-1},$$

отсюда следует (17).

Лемма доказана.

Лемма 2. Для любых $(m \times n)$-матриц F_t, $t = 1, \ldots, N$, справедливы равенства

$$(I_n - \Xi_t \Xi_t^+) F_t^{\mathrm{T}} = 0, \ t = 1, \ldots, N, \tag{26}$$

где $\Xi_t = \sum_{s=1}^{t} (F_s)^{\mathrm{T}} F_s$, $(\cdot)^+$ — псевдообратная матрица соответствующей матрицы.

Доказательство. Без ограничения общности будем полагать, что $F_t \neq 0$. Обозначим $\widetilde{F}_t = (F_1^{\mathrm{T}}, \ldots, F_t^{\mathrm{T}})$. Пусть $l_{1,t}, \ldots, l_{k(t),t}$ — любые линейно независимые столбцы матрицы $\widetilde{F}_t$ такие, что последний $l_{k(t),t}$ из них совпадает с последним ненулевым столбцом матрицы $\widetilde{F}_t$. Использование скелетного разложения для $\widetilde{F}_t$ дает $\widetilde{F}_t = L_t \Gamma_t$, где $L_t = (l_{1,t}, \ldots, l_{k(t),t})$, $\Gamma_t = (\Gamma_{1t}, \ldots, \Gamma_{tt})$ — $(n \times k(t))$-, $(k(t) \times mt)$-матрицы ранга $k(t)$, Γ_{it}, $i = 1, \ldots, t$ — $(k(t) \times m)$-матрицы. Покажем вначале, что $I_n - \Xi_t \Xi_t^+ = I - L_t (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}$. Имеем $\Xi_t = \widetilde{F}_t \widetilde{F}_t^{\mathrm{T}} = L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}}$, где $\widetilde{\Gamma}_t = \Gamma_t \Gamma_t^{\mathrm{T}}$. Так как $\widetilde{\Gamma}_t > 0$ — матрица Грамма, построенная по линейно независимым строкам матрицы Γ_t, $\operatorname{rank}(L_t) = \operatorname{rank}(\widetilde{\Gamma}_t L_t^{\mathrm{T}})$, и L_t — матрица полного ранга по столбцам, то $\Xi_t^+ = (L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}})^+ = (L_t^{\mathrm{T}})^+ \widetilde{\Gamma}_t^{-1} L_t^+$, $L_t^+ = (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}$. Таким образом,

$$I_n - \Xi_t \Xi_t^+ = I_n - L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}} (L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}})^+ = I_n - L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}} (L_t^{\mathrm{T}})^+ \widetilde{\Gamma}_t^{-1} L_t^+ =$$

$$= I_n - L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}} L_t (L_t^{\mathrm{T}} L_t)^{-1} \widetilde{\Gamma}_t^{-1} (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}} = I_n - L_t (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}.$$

Поскольку $F_t^{\mathrm{T}} = L_{\bar{t}} \Gamma_t$ для некоторой матрицы $\overline{\Gamma}_t$, то

$$(I_n - \Xi_t \Xi_t^+) F_t^{\mathrm{T}} = (I_{n-q} - L_t (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}) L_t \overline{\Gamma}_t = 0.$$

Далее будет использоваться обозначение $O(\lambda^{-k})$ для функций $\Phi(\lambda, y_t, \hat{x}_t)$, удовлетворяющих условию $P(\| \Phi(\lambda, y_t, \hat{x}_t) \| / \lambda^{-k} < \delta_1) > 1 - \delta_2$ при $\lambda \to \infty$, $k \geq 0$, некоторых постоянных $\delta_1, \delta_2 > 0$ и t, принадлежащих ограниченному множеству, где $\| \cdot \|$ — евклидова норма соответствующей матрицы.

Лемма 3. Пусть матрица Ω_t определена выражением

$$\Omega_t = \frac{\Omega_1}{\lambda} + \sum_{s=1}^{t-1} F(y_s, \hat{x}_s)^{\mathrm{T}} F(y_s, \hat{x}_s), \ t = 2, \ldots, N, \tag{27}$$

где $\Omega_1 > 0$, $m \times n$-матрица $F(y_t, \hat{x}_t)$ удовлетворяют условию $F(y_t, \hat{x}_t) = O(1)$ при $\lambda \to \infty$ и $t = 2, \ldots, N$. Тогда

$$\Omega_t^{-1} = \Omega_1^{-1} (I_n - \Xi_t \Xi_t^+) \lambda + \Xi_t^+ + O(\lambda^{-1}), \ t = 2, \ldots, N, \tag{28}$$

при $\lambda \to \infty$, где $\Xi_t = \sum_{s=1}^{t-1} F(y_s, \hat{x}_s)^{\mathrm{T}} F(y_s, \hat{x}_s)$.

Доказательство. Имеем $\Omega_t = \Omega_1 (I_n + \lambda \Omega_a^{-1} \Xi_t)/\lambda$. Отсюда с помощью формулы обращения возмущенных матриц [18] получим

$$(I_n + \lambda V)^{-1} = (I_n - VV^+) + V^+ (V^+ + \lambda I_n)^{-1}, \tag{29}$$

где V — произвольная симметрическая матрица, находим

$$\Omega_t^{-1} = (I_n + \lambda \Omega_1^{-1} \Xi_t)^{-1} \Omega_1^{-1} \lambda =$$

$$= [(I_n - \Omega_1^{-1} \Xi_t \Xi_t^+ \Omega_1) \Omega_1^{-1} + \Xi_t^+ \Omega_1 (\Xi_t^+ \Omega_1 + \lambda I_n)^{-1} \Omega_1^{-1}] \lambda .$$

Представим матрицу $\Omega_1^{-1} \Xi_t$ в виде $\Omega_1^{-1} \Xi_t = T_t^{\mathrm{T}} D_t T_t$, где T_t и D_t — соответственно ортогональная и диагональная матрицы. Тогда

$$(\Xi_t^+ \Omega_1 + \lambda I_n)^{-1} \lambda = T_t (D_t^+ + \lambda I_n)^{-1} T_t^{\mathrm{T}} \lambda =$$

$$= I_n - \Xi_t^+ \Omega_1 \lambda^{-1} + O(\lambda^{-2}), \ \lambda \to \infty . \tag{30}$$

Отсюда следует (28).

Теорема 1. 1. При ограниченном N справедливы асимптотические представления

$$P_t = \widetilde{R}_t (I_{rm} - W_t W_t^+) \widetilde{R}_t^{\mathrm{T}} \lambda + \widetilde{S}_t + \widetilde{R}_t W_t^+ \widetilde{R}_t^{\mathrm{T}} + O(\lambda^{-1}), \ t = 2, \ldots, N, \tag{31}$$

$$K_t = K_t^{dif} + O(\lambda^{-1}), \ C_t \neq 0, \ t = 1, \ldots, N, \ \lambda \to \infty , \tag{32}$$

где

$$W_{t+1} = W_t + (C_t^{\beta} R_t + C_t^{\alpha})^{\mathrm{T}} N_{1t}^{-1} (C_t^{\beta} R_t + C_t^{\alpha}), \ W_1 = 0_{mn}, \tag{33}$$

$$K_t^{dif} = (\widetilde{S}_t + \widetilde{R}_{t+1} W_{t+1}^+ \widetilde{R}_t^{\mathrm{T}}) C_t^{\mathrm{T}} N_{1t}^{-1} . \tag{34}$$

2. Если дополнительно $\sigma(x) \in C^2(R)$, то

$$\hat{x}_t = \hat{x}_t^{dif} + O(\lambda^{-1}), \ t = 2, \ldots, N, \ \lambda \to \infty , \tag{35}$$

где

$$\hat{x}_{t+1}^{dif} = \hat{x}_t^{dif} + K_t^{dif} (y_t - \widetilde{f}(z_t, \hat{x}_t^{dif})), \ \hat{x}_1^{dif} = (\overline{\beta}^{\mathrm{T}}, 0_{rm}^{\mathrm{T}})^{\mathrm{T}} . \tag{36}$$

Доказательство. Покажем, что $\hat{x}_t = O(1)$ при $\lambda \to \infty$, $t = 2, \ldots, N$. Отсюда и леммы 3, если положить в (28) $F(y_t, \hat{x}_t) = N_{1t}^{-1/2} (C_t^{\beta} R_t + C_t^{\alpha})$, будем иметь

$$M_{t+1}^{-1} = (I_n - \Xi_t \Xi_t^+) \lambda + \Xi_t^+ + O(\lambda^{-1}), \ t = 1, \ldots, N, \ \lambda \to \infty . \tag{37}$$

Подставляя (37) в (16), получаем (31). Воспользуемся индукцией. Так как C_1 не зависит от λ, то (31) справедливо при $t = 1$ и $\hat{x}_2 = O(1)$, $\lambda \to \infty$. Предположим, что $\hat{x}_i = O(1)$ при $\lambda \to \infty$ и каждом $i \in Z = \{3, \ldots, t-1\}$. В силу ограниченности Z и $\hat{x}_2 = O(1)$, $\lambda \to \infty$ эта оценка будет выполняться равномерно на $\{1, \ldots, t-1\}$, что влечет $C_i = O(1)$ на Z и $\hat{x}_t = O(1)$ при $\lambda \to \infty$, $t = 2, \ldots, N$.

В результате подстановки (37) в (17) и в силу леммы 2 получаем (32).

Воспользуемся индукцией для доказательства (35). Обозначим $e_t = \hat{x}_t - \hat{x}_t^{dif}$. Из (5) и (36) следует

$$e_{t+1} = e_t + (K_t(\hat{x}_t) - K_t^{dif} (\hat{x}_t - e_t)) y_t - K_t(\hat{x}_t) \widetilde{f}(z_t, \hat{x}_t) +$$

$$+ K_t^{dif} (\hat{x}_t - e_t) \widetilde{f}(z_t, \hat{x}_t - e_t), \ t = 1, \ldots, N. \tag{38}$$

Поскольку $e_1 = 0$, $\widetilde{f}(z_t, \hat{x}_1) = 0$, $K_1 = K_1^{dif} + O(\lambda^{-1})$ (без ограничения общности полагаем, что $C_1 \neq 0$), то из (38) следует оценка $e_2 = O(\lambda^{-1})$, $\lambda \to \infty$. Пусть t произвольно и $e_t = O(\lambda^{-1})$, $\lambda \to \infty$. Если $C_t = 0$, то $e_{t+1} = e_t = O(\lambda^{-1})$. При

$C_t \neq 0$ воспользуемся оценкой $K_t = K_t^{dif} + O(\lambda^{-1})$ и теоремой о среднем

$$K_t^{dif}(\hat{x}_t - e_t) = K_t^{dif}(\hat{x}_t) + O(\lambda^{-1}),\ \widetilde{f}(z_t, \hat{x}_t - e_t) = \widetilde{f}(z_t, \hat{x}_t) + O(\lambda^{-1}),\ \lambda \to \infty.$$

Подстановка этих выражений в (38) дает $e_{t+1} = O(\lambda^{-1})$, $\lambda \to \infty$. Отсюда в силу ограниченности множества $t = 1, \ldots, N$ следует оценка (35).

Соотношения (36), (33), (34), (19), (20) определяют ДАО.

Следствие 1. Диффузная составляющая $P_t^{dif} = \widetilde{R}_t (I_{n-q} - W_t W_t^+) \widetilde{R}_t^{\mathrm{T}} \lambda$ — компонента в разложении P_t, пропорциональная большому параметру, равна нулю, начиная с $t_{tr} = \min\limits_t \{t:\ W_t > 0,\ t = 2, \ldots, N\}$.

Следствие 2. Матрица K_t не зависит от диффузной составляющей и в отличие от матрицы P_t как функция λ равномерно ограничена по норме при $t = 1, \ldots, N$ и $\lambda \to \infty$.

Следствие 3. Ошибки численной реализации могут приводить к расходимости РФК при больших λ. Действительно, пусть δW_t^+ — ошибка, связанная с вычислением псевдообратной матрицы W_t^+. Тогда использование (17) и леммы 3 дает

$$K_t = (\widetilde{S}_t + \widetilde{R}_{t+1} M_{t+1}^{-1} \widetilde{R}_t^{\mathrm{T}}) C_t^{\mathrm{T}} N_{1t}^{-1} =$$

$$= [\widetilde{S}_t + \widetilde{R}_{t+1}((I_{rm} - W_t(W_t^+ + \delta W_t^+)\lambda + O(1))\widetilde{R}_t^{\mathrm{T}}] C_t^{\mathrm{T}} N_{1t}^{-1},\ t = 1, \ldots, N,\ \lambda \to \infty. \quad (39)$$

При $\delta W_t^+ \neq 0$ матрица K_t становится зависимой от диффузной составляющей. Поскольку операция псевдообращения не является непрерывной, то вызванное ею изменение K_t может быть существенным и приводить к расходимости. Более того, поскольку матрица K_t пропорциональна большому параметру λ, то даже если и выполнено условие непрерывности операции псевдообращения $\operatorname{rank}(W_t) = \operatorname{rank}(W_t + \delta W_t)$, для произвольных, достаточно малых по норме матриц δW_t, диффузная составляющая по-прежнему может приводить к потере точности. Кроме того, этот эффект может усиливаться при поступлении измерений высокой точности и малой по норме матрице S_t. ДАО при численной реализации не имеют указанных особенностей. Об этом свидетельствует отсутствие в его конструкции диффузных компонент — величин, пропорциональных большому параметру, причем характеристики предельного алгоритма не зависят от неизвестной априорной информации о начальном векторе состояния.

Замечание 1. Теорема 1 и следствия 1–3 справедливы для алгоритма обучения в пакетном режиме, описываемого соотношениями (12)–(15).

3. СВЯЗЬ С ELM-АЛГОРИТМОМ

Полагая в ДАО $S_\beta = 0$, получаем

$$\hat{\alpha}_{t+1} = \hat{\alpha}_t + K_t^{\alpha, dif}(y_t - C_t^\alpha \hat{\alpha}_t),\ \hat{\alpha}_1 = 0_{rm},\ \hat{\beta}_t = m_\beta^{\mathrm{T}}, \quad (40)$$

$$K_t^{\alpha, dif} = W_{t+1}^+ (C_t^\alpha)^{\mathrm{T}} V_t^{-1},\ W_{t+1} = W_t + (C_t^\alpha)^{\mathrm{T}} V_t^{-1} C_t^\alpha,\ W_a = 0_{rn \times rn}, \quad (41)$$

т.е., как и в ELM-алгоритме, оцениваются только веса выходного слоя, а веса и смещения скрытого слоя в процессе обучения не изменяются. Покажем, что оценки, получаемые с использованием ДАО и ELM, совпадают.

Рассмотрим задачу минимизации взвешенной суммы квадратов

$$J_t(\alpha) = \sum_{i=1}^{t} (y_i - C_i^\alpha \alpha)^{\mathrm{T}} V_i^{-1} (y_i - C_i^\alpha \alpha)^{\mathrm{T}} \quad (42)$$

при фиксированном $\beta = \overline{\beta}$. Пусть $\operatorname{rank}(\widetilde{C}_t) = rm$, где $\widetilde{C}_t = ((C_1^\alpha)^{\mathrm{T}}, \ldots, (C_t^\alpha)^{\mathrm{T}})$.

Тогда

$$\alpha_t^{\min} = \left(\sum_{i=1}^{t} (C_i^{\alpha})^{\mathrm{T}} V_i^{-1} C_i^{\alpha} \right)^{-1} \sum_{i=1}^{t} (C_i^{\alpha})^{\mathrm{T}} V_i^{-1} y_i. \tag{43}$$

Теорема 2. Оценка ДАО удовлетворяет условию $\hat{\alpha}_{t+1} = \alpha_t^{\min}$ при любом $\hat{\alpha}_1 = \overline{\alpha}$ и $t_{tr} \geq t$.

Доказательство. Имеем

$$\hat{\alpha}_{t+1} = \Phi_{t+1,1}\overline{\alpha} + \sum_{s=1}^{t} \Phi_{t+1,s+1} K_s^{\alpha,dif} y_s, \tag{44}$$

где переходная матрица $\Phi_{t,s}$ удовлетворяет уравнению

$$\Phi_{t+1,s} = (I_{rm} - K_t^{\alpha,dif} C_t^{\alpha}) \Phi_{t,s} = A_t \Phi_{t,s},\ \Phi_{s,s} = I_{rm},\ t = s, s+1, \ldots.$$

Покажем, что

$$\Phi_{t,s} = W_t^{-1} W_s,\ t \geq tr,\ 1 \leq s \leq t, \tag{45}$$

но вначале подтвердим, что решения матричных уравнений

$$G_{t,s} = (I_{rn} - K_t^{\alpha,dif} C_t^{\alpha})^{\mathrm{T}} G_{t+1,s},\ G_{s,s} = I_{rn},$$

$$Z_{t,s} = Z_{t+1,s} - (C_t^{\alpha})^{\mathrm{T}} V_t^{-1} C_t^{\alpha} W_s^{-1},\ Z_{s,s} = I_{rn}$$

совпадают, где $s \geq R = \min\limits_{s} \{s : W_s > 0, s = 2, \ldots, N\}$, $t = s-1, \ldots, 1$. Для этого должно выполняться условие $(K_t^{\alpha,dif})^{\mathrm{T}} Z_{t+1,s} = V_t^{-1} C_t^{\alpha} W_s^{-1}$. Имеем

$$Z_{t,s} = I_{rm} - \sum_{i=t}^{s-1} (C_i^{\alpha})^{\mathrm{T}} V_i^{-1} C_i^{\alpha} W_s^{-1},$$

$$(K_t^{\alpha,dif})^{\mathrm{T}} Z_{t+1,s} = (W_{t+1}^{-1} (C_t^{\alpha})^{\mathrm{T}} V_t^{-1})^{t} \left(I_{rm} - \sum_{i=t+1}^{s-1} (C_i^{\alpha})^{\mathrm{T}} V_i^{-1} C_i^{\alpha} W_s^{-1} \right) =$$

$$= (W_{t+1}^{-1} (C_t^{\alpha})^{\mathrm{T}} V_t^{-1})^{\mathrm{T}} \left(W_s - \sum_{i=t+1}^{s-1} (C_i^{\alpha})^{\mathrm{T}} V_i^{-1} C_i^{\alpha} \right) W_s^{-1} = V_t^{-1} C_t^{\alpha} W_s^{-1}.$$

Так как $G_{s,t} = A_s^{\mathrm{T}} A_{s+1}^{\mathrm{T}} \ldots A_{t-1}^{\mathrm{T}}$, $\Phi_{t,s} = A_{t-1} A_{t-2} \ldots A_s^{\mathrm{T}}$, то выполняется (45). Подстановка (45) в (44) дает

$$\hat{\alpha}_{t+1} = W_{t+1}^{-1} \sum_{s=1}^{t} W_{s+1} W_{s+1}^{+} (C_s^{\alpha})^{\mathrm{T}} V_s^{-1} y_s. \tag{46}$$

Покажем, что

$$W_{s+1} W_{s+1}^{+} (C_s^{\alpha})^{\mathrm{T}} = (C_s^{\alpha})^{\mathrm{T}},\ s = 1, \ldots, t. \tag{47}$$

Используя скелетное разложения $\widetilde{C}_{s+1} = L_{s+1} \Gamma_{s+1}$, получаем

$$W_{s+1} = \widetilde{C}_{s+1} \widetilde{C}_{s+1}^{\mathrm{T}} = L_{s+1}^{\mathrm{T}} \widetilde{\Gamma}_{s+1} L_{s+1},\ W_{s+1}^{+} = (L_{s+1}^{\mathrm{T}})^{+} \widetilde{\Gamma}_{s+1}^{-1} L_{s+1}^{+},\ C_{s+1} = L_{s+1} \overline{\Gamma}_{s+1},$$

где $\widetilde{\Gamma}_{s+1} = \Gamma_{s=1} \Gamma_{s+1}^{\mathrm{T}}$, $\overline{\Gamma}_{s+1}$ — некоторая матрица. Поскольку $L_{s=1}$ — матрица полного ранга по столбцам, то $L_{s+1}^{+} = (L_{s+1}^{\mathrm{T}} L_{s+1})^{-1} L_{s+1}^{\mathrm{T}}$ и $W_{s+1} W_{s+1}^{+} = L_{s+1} (L_{s+1}^{\mathrm{T}} L_{s+1})^{-1} L_{s+1}^{\mathrm{T}}$. Отсюда следует (47), подстановка которого в (46) дает $\hat{\alpha}_{t+1} = \alpha_t^{\min}$.

Замечание 2. Соотношения (40), (41) представляют собой новую, рекуррентную версию (диффузную) ELM-алгоритма.

4. ДВУХЭТАПНЫЕ АЛГОРИТМЫ ОБУЧЕНИЯ

Получим упрощенные ДАО, используя двухэтапную процедуру оценивания [19]. Фиксируя последовательно в (3) $\hat{\beta}_t$ и $\hat{\alpha}_t$ и используя РФК, находим:

$$\hat{\beta}_{t+1} = \hat{\beta}_t + K_t^{\beta}(y_t - f(z_t, \hat{\beta}_t, \hat{\alpha}_t)),\ \hat{\beta}_1 = m_{\beta}^{\mathrm{T}}, \tag{48}$$

$$\hat{\alpha}_{t+1} = \hat{\alpha}_t + K_t^{\alpha}(y_t - f(z_t, \hat{\beta}_t, \hat{\alpha}_t)),\ \hat{\alpha}_1 = 0_{rm}, \tag{49}$$

$$K_t^{\beta} = S_t (C_t^{\beta})^{\mathrm{T}} N_t^{-1},\ K_t^{\alpha} = P_t (C_t^{\alpha})^{\mathrm{T}} N_t^{-1}, \tag{50}$$

$$S_{t+1} = S_t - S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} C_t^{\beta} S_t,\ N_{1t} = C_t^{\beta} S_t (C_t^{\beta})^{\mathrm{T}} + V_t,\ S_1 = S_{\beta}, \tag{51}$$

$$P_{t+1} = P_t - P_t (C_t^{\alpha})^{\mathrm{T}} N_{2t}^{-1} C_t^{\alpha} P_t,\ N_{2t} = C_t^{\alpha} P_t (C_t^{\alpha})^{\mathrm{T}} + V_t,\ P_1 = \lambda I_{rm}. \tag{52}$$

Теорема 3. 1. При ограниченном N справедливы асимптотические представления

$$P_t = (I_r - W_t W_t^{+})\lambda + W_t^{+} + O(\lambda^{-1}),\ t = 2, \ldots, N, \tag{53}$$

$$K_t^{\alpha} = K_t^{\alpha, dif} + O(\lambda^{-1}),\ C_t^{\alpha} \neq 0,\ t = 1, \ldots, N,\ \lambda \to \infty, \tag{54}$$

где

$$W_{t+1} = W_t + (C_t^{\alpha})^{T} V_t^{-1} C_t^{\alpha},\ W_1 = 0_{mn}, \tag{55}$$

$$K_t^{\alpha, dif} = W_{t+1}^{+} (C_t^{\alpha})^{\mathrm{T}} V_t^{-1}. \tag{56}$$

2. Если дополнительно $\sigma(x) \in C^2(R)$, то

$$\hat{\beta}_t = \hat{\beta}_t^{dif} + O(\lambda^{-1}),\ \hat{\alpha}_t = \hat{\alpha}_t^{dif} + O(\lambda^{-1}),\ t = 1, \ldots, N,\ \lambda \to \infty, \tag{57}$$

где

$$\hat{\beta}_{t+1}^{dif} = \hat{\beta}_t^{dif} + K_t^{\beta}\delta_t,\ \hat{\alpha}_{t+1}^{dif} = \hat{\alpha}_t^{dif} + K_t^{\alpha, dif}\delta_t, \tag{58}$$

$$\delta_t = y_t - f(z_t, \hat{\beta}_t^{dif}, \hat{\alpha}_t^{dif}),\ \hat{\beta}_1^{dif} = \overline{m}\beta,\ \hat{\alpha}_a^{dif} = 0_{rm}. \tag{59}$$

Доказательство. Доказательство того, что $\hat{x}_t = (\hat{\alpha}_t^{\mathrm{T}}, \hat{\beta}_t^{\mathrm{T}})^{\mathrm{T}} = O(1)$ при $\lambda \to \infty$, $t = 2, \ldots, N$, проводится так же, как и в теореме 1, поэтому его опускаем. Поскольку

$$P_{t+1}^{-1} = P_t^{-1} + (C_t^{\alpha})^{\mathrm{T}} V_t^{-1} C_t^{\alpha},\ P_1^{-1} = I_{rm}/\lambda,\ t = 1, \ldots, N,$$

то из леммы 3, если положить в (27) $F(y_t, \hat{x}_t) = V_t^{-1/2} C_t^{\alpha}$, будет следовать (53). Имеем

$$K_t^{\alpha} = P_t (C_t^{\alpha})^{\mathrm{T}} N_{2t}^{-1} = P_{t+1} P_{t+1}^{-1} P_t (C_t^{\alpha})^{\mathrm{T}} N_{2t}^{-1} =$$

$$= P_{t+1}(P_t^{-1} + (C_t^{\alpha})^{\mathrm{T}} V_t^{-1} C_t^{\alpha}) P_t (C_t^{\alpha})^{\mathrm{T}} N_{2t}^{-1} = P_{t+1}(C_t^{\alpha})^{\mathrm{T}} V_t^{-1}. \tag{60}$$

В результате подстановки (53) в (60) и в силу леммы 2 получаем (54).

Пусть $e_t^{\beta} = \hat{\beta}_t - \hat{\beta}_t^{dif}$, $e_t^{\alpha} = \hat{\alpha}_t - \hat{\alpha}_t^{dif}$. Тогда

$$e_{t+1}^{\beta} = e_t^{\beta} + (K_t^{\beta}(\hat{\beta}_t, \hat{\alpha}_t) - K_t^{\beta}(\hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}))y_t - K_t^{\beta}(\hat{\beta}_t, \hat{\alpha}_t) f(z_t, \hat{\beta}_t, \hat{\alpha}_t) +$$

$$+ K_t^{\beta}(\hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}) f(z_t, \hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}), \tag{61}$$

$$e_{t+1}^{\alpha} = e_t^{\alpha} + (K_t^{\alpha}(\hat{\beta}_t, \hat{\alpha}_t) - K_t^{\alpha, dif}(\hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}))y_t - K_t^{\beta}(\hat{\beta}_t, \hat{\alpha}_t) f(z_t, \hat{\beta}_t, \hat{\alpha}_t) +$$

$$+ K_t^{\alpha, dif}(\hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}) f(z_t, \hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}),\ t = 1, \ldots, N. \tag{62}$$

Так как $e_1^{\beta} = 0$, $e_1^{\alpha} = 0$, $f(u_t, \hat{\beta}_1, \hat{\alpha}_1) = 0$, $K_1^{\alpha} = K_1^{\alpha, dif} + O(\lambda^{-1})$ (без ограничения общности полагаем, что $C_t^{\alpha} \neq 0$), то из (61), (62) следует $e_2^{\beta} = 0$, $e_2^{\alpha} = O(\lambda^{-1})$. $\lambda \to \infty$. Пусть t произвольно и $e_t = O(\lambda^{-1})$, $\lambda \to \infty$. Если $C_t^{\alpha} = 0$, то $e_{t+1} = e_t = O(\lambda^{-1})$. При $C_t^{\alpha} \neq 0$ используем оценку $K_t^{\alpha} = K_t^{\alpha, dif} + O(\lambda^{-1})$ и теорему о среднем

$$f(z_t, \hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}) = f(z_t, \hat{\beta}_t, \hat{\alpha}_t) + O(\lambda^{-1}),$$

$$K_t^{\alpha, dif}(\hat{\beta}_t - e_t^{\beta}, \hat{\alpha}_t - e_t^{\alpha}) = K_t^{\alpha, dif}(\hat{\beta}_t, \hat{\alpha}_t) + O(\lambda^{-1}),\ \lambda \to \infty.$$

Подстановка этих выражений в (61), (62) дает $e_{t+1} = O(\lambda^{-1})$, $\lambda \to \infty$. Отсюда в силу ограниченности множества $t = 1, \ldots, N$ следуют оценки (57).

Замечание 3. Теорема 3 остается справедливой для двухэтапного ДАО в пакетном режиме, соотношения для которого нетрудно выписать, используя (11).

5. АСИМПТОТИЧЕСКИ ТОЧНЫЕ ИЗМЕРЕНИЯ

Рассмотрим поведение ДАО при малых шумах измерения, матрица интенсивности которых V_t входит в алгоритмы через обратную матрицу $N_{1t}^{-1} = (C_t^{\beta} S_t (C_t^{\beta})^{\mathrm{T}} + V_t)^{-1}$. С учетом этого важно понимать, как будут вести себя ДАО при $V_t \to 0$.

Теорема 4. При ограниченном N и $V_t = I_{mn}/\lambda$, $\lambda \to \infty$ и $S_1 > 0$

$$\widetilde{K}_t = \lim_{\lambda \to \infty} K_t^{dif} = \begin{pmatrix} Q_{t+1}(C_t^{\beta})^{\mathrm{T}} \\ 0_{r \times rm} \end{pmatrix} + \begin{pmatrix} G_{t+1} \\ I_{rm} \end{pmatrix} L_{t+1}^{+} \begin{pmatrix} G_t \\ I_{rm} \end{pmatrix} \begin{pmatrix} (C_t^{\beta})^{\mathrm{T}} \\ (C_t^{\alpha})^{\mathrm{T}} \end{pmatrix} \overline{H}_t, \tag{63}$$

где

$$Q_{t+1} = Q_t + (C_t^{\beta})^{\mathrm{T}} C_t^{\beta},\ Q_1 = 0_{mn \times mn}, \tag{64}$$

$$G_{t+1} = (I_{mn} - Q_{t+1}^{+}(C_t^{\beta})^{\mathrm{T}} C_t^{\beta}) G_t - Q_{t+1} C_t^{\alpha},\ G_1 = 0_{mn \times r}, \tag{65}$$

$$L_{t+1} = L_t + (C_t^{\beta} G_t + C_t^{\alpha})^{\mathrm{T}} \overline{H}_t (C_t^{\beta} G_t + C_t^{\alpha}),\ L_1 = 0_{rm \times rm}, \tag{66}$$

$$\begin{gathered} \overline{H}_t = (I_r - \widetilde{H}_t \widetilde{H}_t^{+})(I_r + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})^{-1}, \\ \widetilde{H}_t = (I_r + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})^{-1} C_t^{\beta} H_t (C_t^{\beta})^{\mathrm{T}}, \end{gathered} \tag{67}$$

$$H_t = S_1 (I_{mn} - Q_t Q_t^{+}). \tag{68}$$

Доказательство. Так как

$$S_t^{-1} = \left(\frac{S_1^{-1}}{\lambda} + \sum_{s=a}^{t-1} (C_t^{\beta})^T C_t^{\beta} \right) \lambda,\ t = 2, \ldots, N,$$

то из леммы 3, если положить в (27) $\Omega_1 = S_1^{-1}$, $F(y_t, \hat{x}_t) = C_t^{\beta}$, будет следовать

$$S_t = S_1 (I_{mn} - Q_t Q_t^{+}) + Q_t^{+}/\lambda + O(1/\lambda^2),\ t = 2, \ldots, N. \tag{69}$$

Поскольку $K_t^{\beta} = S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} = S_{t+1} (C_t^{\beta})^{\mathrm{T}} V_t^{-1}$, то из (69) и леммы 2 имеем

$$\lim_{\lambda \to \infty} S_t (C_t^{\beta})^{\mathrm{T}} N_{1t}^{-1} = Q_{t+1}^{+} (C_t^{\beta})^{\mathrm{T}}, \tag{70}$$

но так как $R_{t+1} = (I_{mn} - K_t^{\beta} C_t^{\beta}) R_t - K_t^{\beta} C_t^{\alpha}$, $R_1 = 0_{mn}$, то отсюда вытекает (65).

Преобразуем N_{1t} с помощью (69):

$$N_{1t} = C_t^{\beta} S_t (C_t^{\beta})^{\mathrm{T}} + V_t = C_t^{\beta} H_t (C_t^{\beta})^{\mathrm{T}} + (I_{mn} + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})/\lambda + O(1/\lambda^2) =$$

$$= (I_{mn} + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})(\widetilde{H}_t + I_{mn}/\lambda + O(1/\lambda^2)).$$

Использование леммы 3 дает

$$N_{1t}^{-1} = (\widetilde{H}_t + I_r / \lambda)^{-1}(I + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})^{-1} + O(1/\lambda^2) =$$

$$= \lambda(I_r - \widetilde{H}_t \widetilde{H}_t^{+})(I + C_t^{\beta} Q_t^{+} (C_t^{\beta})^{\mathrm{T}})^{-1} + O(1) = \lambda \overline{H}_t + O(1).$$

Имеем

$$W_{t+1} = W_t + \lambda(C_t^{\beta} R_t + C_t^{\alpha})^{\mathrm{T}} (\overline{H}_t + O(1/\lambda))(C_t^{\beta} R_t + C_t^{\alpha}),\ W_a = 0_{rm \times rm}.$$

Отсюда следует, что

$$\lim_{\lambda \to \infty} W_t / \lambda = \sum_{s=a}^{t-1} (C_s^{\beta} G_s + C_s^{\alpha})^{\mathrm{T}} \overline{H}_s (C_s^{\beta} G_s + C_s^{\alpha}),$$

$$\lim_{\lambda \to \infty} \begin{pmatrix} R_{t+1} \\ I_{rm} \end{pmatrix} W_{t+1}^{+} \begin{pmatrix} R_t \\ I_{rm} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} (C_t^{\beta})^{\mathrm{T}} \\ (C_t^{\alpha})^{\mathrm{T}} \end{pmatrix} N_{1t}^{-1} = \begin{pmatrix} G_{t+1} \\ I_{rm} \end{pmatrix} L_{t+1}^{+} \begin{pmatrix} G_t \\ I_{rm} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} (C_t^{\beta})^{\mathrm{T}} \\ (C_t^{\alpha})^{\mathrm{T}} \end{pmatrix} \overline{H}_t.$$

Теорема 5. При ограниченном N и $V_t = I_{mn} / \lambda$, $\lambda \to \infty$ и $S_1 > 0$ справедливо

$$\overline{K}_t^{\beta} = \lim_{\lambda \to \infty} K_t^{\beta} = Q_{t+1} (C_t^{\beta})^{\mathrm{T}},\ \overline{K}_t^{\alpha} = \lim_{\lambda \to \infty} K_t^{dif,\alpha} = L_{t+1} (C_t^{\alpha})^{\mathrm{T}},$$

где

$$Q_{t+1} = Q_t + (C_t^{\beta})^{\mathrm{T}} C_t^{\beta},\ Q_1 = 0_{mn \times mn},\ L_{t+1} = L_t + (C_t^{\alpha})^{\mathrm{T}} \overline{H}_t C_t^{\alpha}),\ L_1 = 0_{rm \times rm}.$$

Доказательство повторяет вывод формулы (70) в теореме 4.

6. ЧИСЛЕННЫЙ ПРИМЕР

Проиллюстрируем эффективность полученных в работе результатов на одном из стандартных примеров, используемых при тестировании алгоритмов [11]. Рассмотрим задачу аппроксимации функции

$$y(x) = \begin{cases} \sin(x)/x, & x \neq 0 \\ 1, & x = 0. \end{cases}$$

Обучающее и тестирующее множества (x_i, y_i) включают по 2000 точек, а значения x_i равномерно распределены на интервале $[-10, 10]$. К обучающим значениям y_i добавляется равномерно распределенный на интервале $[-0.2, 0.2]$ шум, а тестирующее множество предполагается незашумленным. Используется сигмоидная ФА и сравнивается точность аппроксимации ELM и ДАО с последовательной обработкой. В качестве меры точности используется оценка 90-й процентили ошибки аппроксимации по 500 выборкам. Веса и смещения скрытого слоя предполагаются равномерно распределенными случайными величинами на интервалах $[-1, 1]$, $[0, 1]$ соответственно, $V_t = 0.16/12$. Результаты моделирования приведены в табл. 1, для алгоритмов, следующих из теорем 1, 3 (ДАО1 и ДАО2 соответственно). Видно, что ДАО превосходят по точности ELM как на обучающем, так и на тестирующем множествах. ДАО1 и ДАО2 дают практически неразличимые результаты по точности, но при этом ДАО2 превосходит ДАО1 по быстродействию на 32 %. На рис. 1 приведены графики аппроксимирующих зависимостей *1*, *2* (ELM и ДАО1 соответственно) для пяти нейронов скрытого слоя и одного использованного обучающего множества.

Таблица 1

Количество нейронов скрытого слоя	Точность алгоритмов					
	при обучении			при тестировании		
	ELM	ДАО1	ДАО2	ELM	ДАО1	ДАО2
4	0.315	0.169	0.174	0.295	0.125	0.13
5	0.272	0.159	0.159	0.246	0.108	0.11
6	0.205	0.148	0.147	0.17	0.092	0.092

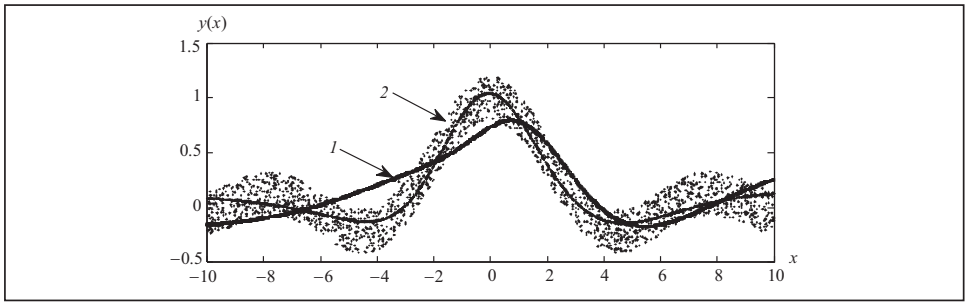

Рис. 1

ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрена задача обучения НСПР. Для ее решения предложены алгоритмы, основанные на асимптотическом анализе поведения РФК и сепарабельной структуре сети. Получение условий сходимости разработанных алгоритмов — одно из направлений дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Palit A., Popovic D. Computational intelligence in time series forecasting: Theory and engineering applications. — New York: Springer, 2005. — 372 p.
2. Devroye L., Gyorfi L., Lugosi G. Probabilistic theory of pattern recognition. — New York: Springer, 1996. — 636 p.
3. Neural systems for control / D.L Elliott, O.M. Omidvar (Eds). — Boston: Academ. Press, 1997. — 362 p.
4. Rumelhart D.E, Hinton G.E., Williams R.J. Learning internal representation by error propagation // Parallel Distributed Proces. — Cambridge, MA: MIT, 1986. — P. 318–362.
5. Battiti R. First and second order methods for learning: Between steepest descent and Newton's method // Neural Comput. — 1992. — **4**, N 2. — P. 141–166.
6. Hagan M.T., Menhaj M. Training multilayer networks with the Marquardt algorithm // IEEE Transact. on Neural Networks. — 1994. — **5**, N 6. — P. 989–993.
7. Singhal S., Wu I. Training multilayer perceptrons with the extended Kalman filter // Advances in Neural Inform. Proces. Systems. — 1989. — **1**. — P. 133–140.
8. Iiguni Y., Sakai H., Tokumaru H. A real time learning. Algorithm for a multilayered neural network based on the extended. Kalman filter // Signal Proces., IEEE Transact. — 1992. — **40**, N 4. — P. 959–966.
9. Bertsekas D.P. Incremental least squares methods and the extended Kalman filter // SIAM J. Optim. — 1996. — **6**. — P. 807–822.
10. Pereyra V., Scherer G., Wong F. Variable projections neural network training // Mathematics and Comput. in Simulat. — 2006. — **73**. — P. 231–243.
11. Huang G.B., Zhu Q.Y., Siew C.K. Extreme learning machine: Theory and applications // Neurocomputing. — 2006. — 7**0**, N 1–3. — P. 489–501.
12. Huang G.B., D.H. Wang D.H., Lan Y. Extreme learning machines: A survey // Intern. Journ. of Machine Leaning and Cybernetics. — 2011. — **2**, N 2. — P. 107–122.
13. Cybenko G. Aproximation by superpositions of a sigmoidal function // Mathematics of Control, Signals and Systems. — 1989. — **2**. — P. 303–314.
14. Golub G.H., Pereyra V. Separable nonlinear least squares: The variable projection method and its applications // Inverse Problems. — 2003. — **19**, N 2. — P. 1–26.
15. Ansley C.F., Kohn R. Estimation, filtering and smoothing in state space models with incompletely specified initial conditions // Ann. Statist. — 1985. — N 13. — P. 1286–1316.
16. Anderson B.D.O., Moore J.B. Optimal filtering. — New York: Prentice-Hall, 1979. — 19. — 368 p.
17. Wimmer H.R. Stabilizing and unmixed solutions of the discrete time algebraic Riccati equation // Proc. Workshop on the Riccati equation in Control, Systems, and Sygnals. — 1989. — Italy. — P. 95–98.
18. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. — М.: Наука, 1977. — 224 с.
19. Ljung L. System identification. Theory for the user. — New York: Prentice-Hall, 1999. — 603 p.

Поступила 06.06.2012