4. Оптимизация параметров нейросетевой модели

Предположим, что в результате проведения эксперимента и предварительной обработки данных получено некоторое множество

$$Z^N = \left\{ \left[u(t), y(t) \right],\ t = \overline{1, N} \right\}, \qquad (2.42)$$

где $u(t)$, $y(t)$ – соответственно входы и выходы системы, N – число дискретных отсчетов. Допустим также, что выбрана некоторая модельная структура

$$y(t) = \hat{y}(t|\theta) + e(t) = g(t, \theta) + e(t). \qquad (2.43)$$

В соответствии с общей схемой реализации процедуры идентификации следующим этапом является оценка параметров выбранной модельной структуры. При использовании нейросетевых модельных структур этот этап представляет собой настройку весовых коэффициентов сети в результате реализации процедуры обучения на множестве примеров. Обучение представляет собой отображение множества экспериментальных данных на множество параметров нейросетевой модели

$$Z^N \to \hat{\theta} \qquad (2.44)$$

с целью получения оптимального, в силу некоторого критерия, прогноза выходного сигнала $\hat{y}$. Традиционно используемым критерием [9, 5] является среднеквадратичная ошибка прогнозирования.

$$V_N(\theta, Z^N) = \frac{1}{2N} \sum_{t=1}^{N} (y(t) - \hat{y}(t|\theta))^2 = \frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t, \theta). \qquad (2.45)$$

Данный подход относится к классу методов ошибки прогнозирования (МОП) [9], так как основной задачей является минимизация суммарной нормы ошибки прогнозирования $\varepsilon = y(t) - \hat{y}(t|\theta)$. В некоторых случаях рассматриваются нормы, отличные от квадратичной, которые являются оптимальными при негауссовом распределении возмущений $e(t)$. При использовании критерия в виде (2.45) МОП соответствует оценке методом максимального правдоподобия при условии нормального распределения возмущений $e(t)$.

Наиболее привлекательной чертой метода является достаточно простой алгоритм оценки параметров (весовых коэффициентов) НС и независимость от возмущений (при условии их нормального распределения). В ряде случаев данный критерий не является абсолютно оптимальным [9], но в практических приложениях обычно приводит к наилучшей модели.

В разделе 2.4.1 представлены методы оптимизации с использованием критерия (2.45). В разделе 2.4.2 обсуждаются практические аспекты применения МПО к обучению нейронных сетей.

4.1. Метод ошибки прогнозирования

При использовании МОП основная задача состоит в нахождении параметров модели посредством минимизации функционала

$$\hat{\theta} = \arg \min_{\theta} V_N(\theta, Z^N). \qquad (2.46)$$

При условии квадратичности критерия рассматривается частный случай безусловной оптимизации – нелинейная задача о наименьших квадратах [7]. Существует ряд методик решения этой проблемы; дан-

ный раздел посвящен обсуждению алгоритмов, имеющих непосредственное отношение к обучению нейронных сетей.

Процедура поиска минимума. Разложение критерия в ряд Тейлора (до 2-го порядка включительно) в окрестности точки θ^* имеет вид:

$$V_N(\theta, Z^N) = V_N(\theta^*, Z^N) + (\theta - \theta^*)^{\mathrm{T}} V_N'(\theta^*, Z^N) + \\ + \frac{1}{2}(\theta - \theta^*)^{\mathrm{T}} V_N''(\theta^*, Z^N)(\theta - \theta^*), \tag{2.47}$$

где градиент определяется как

$$G(\theta^*) = V_N'(\theta^*, Z^N) = \left.\frac{dV_N(\theta, Z^N)}{d\theta}\right|_{\theta = \theta^*}, \tag{2.48}$$

а матрица вторых производных – гессиан, матрица Гессе:

$$H(\theta^*) = V_N''(\theta^*, Z^N) = \left.\frac{d^2 V_N(\theta, Z^N)}{d^2\theta}\right|_{\theta=\theta^*}. \tag{2.49}$$

Достаточными условиями минимума функции являются равенство нулю градиента (2.48) и положительная определенность гессиана (2.49):

$$G(\theta^*) = 0, \tag{2.50}$$

$$H(\theta^*) > 0. \tag{2.51}$$

В случае, когда критерий (2.45) имеет сложную нелинейную структуру, аналитическое нахождение минимума не представляется возможным, что приводит к использованию итеративных методов. В об-

щем случае итеративный алгоритм поиска минимума может быть представлен в следующем виде:

$$\theta^{(i+1)} = \theta^{(i)} + \mu^{(i)} f^{(i)}, \qquad (2.52)$$

где $\theta^{(i)}$ определяет значение параметров на текущей итерации (i), $f^{(i)}$ определяет направление поиска, а $\mu^{(i)}$ – шаг алгоритма на текущей итерации.

В общем случае критерий имеет более одного минимума, но, к сожалению, итеративные методы поиска не обеспечивают сходимости к глобальному минимуму. Проблема «локальных минимумов» непосредственно связана с выбором начальных значений параметров $\theta^{(0)}$.

Градиентный метод. В основе градиентного метода, или метода наискорейшего спуска, лежит определение направления поиска как противоположного направлению градиента, т.е.

$$\theta^{(i+1)} = \theta^{(i)} - \mu^{(i)} G(\theta^{(i)}). \qquad (2.53)$$

Сходимость метода существенно зависит от выбора шага $\mu^{(i)}$: при достаточно малом шаге обеспечивается уменьшение критерия на каждой итерации: $V_N(\theta^{(i+1)}, Z^N) \le V_N(\theta^{(i)}, Z^N)$. Применение метода к обучению нейронных сетей дает возможность организовать вычисления таким образом, чтобы рационально использовать структуру конкретной НС. В этом случае метод называется методом обратного распространения (ошибки), или обобщенным дельта-правилом.

Для выбора шага алгоритма, определяющего скорость сходимости, могут применяться различные методы, в том числе и адаптивные, хо-

тя во многих приложениях используются методы с постоянным шагом $\mu^{(i)}$.

Независимо от выбора шага, градиентный метод может обеспечить только линейную сходимость, т.е. $\left|\theta^{(i+1)}-\theta^{*}\right| \le c\left|\theta^{(i+1)}-\theta^{*}\right|$, $c \in [0,1)$. Недостаточно высокая скорость сходимости алгоритма делает невозможным применение метода для решения задач в режиме реального времени. Тем не менее метод может быть эффективно использован в нейросетевых приложениях благодаря значительной простоте реализации, скромным требованиям к оперативной памяти и возможности использования естественной параллельности алгоритма при наличии специализированного аппаратного обеспечения.

Метод Ньютона. Метод Ньютона является методом 2-го порядка, т.е. основан на следующем представлении критерия (2.45) рядом Тейлора (в окрестности текущей итерации):

$$\tilde{V}_N(\theta, Z^N) = V_N(\theta^{(i)}, Z^N) + \left(\theta-\theta^{(i)}\right)^{\mathrm{T}} G(\theta^{(i)}) + \\ + \frac{1}{2}\left(\theta-\theta^{(i)}\right)^{\mathrm{T}} H(\theta^{(i)})\left(\theta-\theta^{(i)}\right). \qquad (2.54)$$

Введя обозначение

$$\psi(t,\theta) = \frac{d\hat{y}(t(\theta)}{d\theta}, \qquad (2.55)$$

получим выражения для градиента (2.48) и гессиана (2.49) критерия наименьших квадратов:

$$G(\theta) = V_N'(\theta, Z^N) = \frac{1}{N}\sum_{i=1}^{N}\psi(t,\theta)\left(y(t)-\hat{y}(t\,|\,\theta)\right), \qquad (2.56)$$

$$H(\theta) = V_N''(\theta, Z^N) = \frac{1}{N}\sum_{i=1}^{N}\psi(t,\theta)\psi^{\mathrm{T}}(t,\theta) - \frac{1}{N}\sum_{i=1}^{N}\psi'(t,\theta)\varepsilon(t,\theta). \quad (2.57)$$

Минимум функции (2.54) находится в точке $\tilde{V}_N'(\theta, Z^N) = 0$. В силу симметричности гессиана имеем:

$$\begin{aligned} 0 &= G(\theta^{(i)}) + \frac{1}{2}(2H(\theta^{(i)})\theta - H(\theta^{(i)})\theta^{(i)} - H(\theta^{(i)})\theta^{(i)} = \\ &= G(\theta^{(i)}) + H(\theta^{(i)})(\theta - \theta^{(i)}). \end{aligned} \quad (2.58)$$

Анализ соотношения (2.58) приводит к следующему итеративному правилу настройки параметров:

$$\theta^{(i+1)} = \theta^{(i)} - \left[H(\theta^{(i)})\right]^{-1} G(\theta^{(i)}). \quad (2.59)$$

Очевидно, это правило соответствует шагу алгоритма $\mu^{(i)} = 1$ и направлению поиска, определяемому решением системы линейных уравнений

$$H(\theta^{(i)}) f^{(i)} = -G(\theta^{(i)}). \quad (2.60)$$

Направление поиска $f^{(i)}$ обычно называют ньютоновским направлением [7].

На практике метод должен дополняться линейным [17] поиском, так как выражение (2.54) представляет собой лишь аппроксимацию критерия (2.45). Аппроксимация действует только в некоторой окрестности текущей итерации, что может привести к существенной разнице между реальным и прогнозируемым (полученным в результате аппроксимации) значением. В случае, когда метод Ньютона дополняется линейным поиском, алгоритм носит название модифицированного (демпфированного) метода Ньютона. Эта модификация не обеспе-

чивает абсолютной сходимости, поэтому обычно используется для увеличения скорости сходимости в окрестности точки минимума, тогда как для первоначального приближения используется градиентный метод.

Рассмотрим аппроксимацию в окрестности минимума θ^*:

$$\tilde{V}_N(\theta, Z^N) = V_N(\theta^*, Z^N) + \frac{1}{2}(\theta - \theta^*)^{\mathrm{T}} H(\theta^*)(\theta - \theta^*). \qquad (2.61)$$

Несмотря на то, что гессиан положительно определен, он может быть плохо обусловлен. Очевидно, что лишний параметр (весовой коэффициент НС) приводит к бесконечному множеству параметров θ^*, удовлетворяющих условию 2.61, определяя сингулярное число гессиана. Однако гессиан в принципе не может быть сингулярным в точке минимума, что объясняется наличием возмущений в реальной системе. Сингулярность гессиана приводит к проблемам вычислительного характера при определении направления поиска. Проблема может быть решена путем добавления к гессиану диагональной матрицы перед решением системы (2.60) с целью улучшения обусловленности.

С точки зрения вычислительных затрат определение гессиана и направления поиска является достаточно трудоемкой процедурой. Несмотря на квадратичную скорость сходимости в окрестности минимума, реализация алгоритма требует даже больших временных затрат, чем для методов первого порядка. Эта проблема разрешается при использовании аппроксимаций гессиана, используемых в квазиньютоновских методах [7].

Наиболее удачной схемой аппроксимации гессиана является алгоритм Бройдена - Флетчера - Гольдфарба - Шанно [7]. Алгоритм дает

положительно определенную аппроксимацию гессиана на основе значений на предыдущих итерациях и соответствующих градиентов. Существует ряд модификаций алгоритма [7], хотя наиболее часто используется вариант непосредственного получения матрицы, обратной гессиану. В этом случае итеративная процедура поиска минимума принимает вид:

$$\theta^{(i+1)} = \theta^{(i)} - \mu^{(i)} B(\theta^{(i)}) G(\theta^{(i)}), \tag{2.62}$$

где $B(\theta^{(i)}) \approx \left[H(\theta^{(i)}) \right]^{-1}$ – аппроксимация обращенного гессиана, модифицируемая в соответствии со следующим выражением:

$$B(\theta^{(i)}) = \left(I - \frac{\Delta\theta^{(i)} (\Delta G^{(i)})^{\mathrm{T}}}{(\Delta G^{(i)})^{\mathrm{T}} \Delta\theta^{(i)}} \right) B(\theta^{(i-1)}) \left(I - \frac{\Delta G^{(i)} (\Delta\theta^{(i)})^{\mathrm{T}}}{(\Delta G^{(i)})^{\mathrm{T}} \Delta\theta^{(i)}} \right) + \frac{\Delta\theta^{(i)} (\Delta\theta^{(i)})^{\mathrm{T}}}{(\Delta G^{(i)})^{\mathrm{T}} \Delta\theta^{(i)}}, \tag{2.63}$$

где

$$\Delta\theta^{(i+1)} \equiv \theta^{(i)} - \theta^{(i-1)}, \tag{2.64}$$

$$\Delta G^{(i)} \equiv G(\theta^{(i)}) - G(\theta^{(i-1)}). \tag{2.65}$$

Положительная определенность гессиана обусловлена выполнением следующего условия:

$$(\theta^{(i+1)})^{\mathrm{T}} \Delta G^{(i+1)} > 0 . \tag{2.66}$$

Несмотря на теоретическую возможность применения метода к обучению нейронных сетей, практическая реализация обычно не дает желаемых результатов по причине малой начальной скорости сходимости алгоритма. Поэтому для решения проблемы оптимизации нели-

нейных квадратичных критериев наиболее применимы специально разработанные алгоритмы, основанные на семействе методов Гаусса - Ньютона.

Метод Гаусса - Ньютона. В методе Гаусса - Ньютона используется линейная аппроксимация ошибки прогнозирования $\varepsilon(t,\theta) = y(t) - \hat{y}(t|\theta)$

$$\begin{aligned}\tilde{\varepsilon}_N(t,\theta) &= \varepsilon(t,\theta^{(i)}) + (\varepsilon'(t,\theta^{(i)}))^{\mathrm{T}}(\theta-\theta^{(i)}) = \\ &= \varepsilon(t,\theta^{(i)}) - (\psi(t,\theta^{(i)}))^{\mathrm{T}}(\theta-\theta^{(i)})^{\mathrm{T}}.\end{aligned} \qquad (2.67)$$

Модифицированный критерий (2.45) для *i*-ой итерации имеет вид:

$$V_N(\theta, Z^N) \approx L^{(i)}(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(\tilde{\varepsilon}(t,\theta))^2 . \qquad (2.68)$$

Выражение для градиента является аналогом метода Ньютона:

$$G(\theta^{(i)}) = L^{(i)}(\theta^{(i)}, Z^N) = \frac{1}{N}\sum_{i=1}^{N}\psi(t,\theta^{(i)})(y(t) - \hat{y}(t(\theta^{(i)})) . \qquad (2.69)$$

Выражение для определения гессиана претерпевает следующие изменения:

$$R(\theta) = R(\theta^{(i)}) = \frac{1}{N}\sum_{i=1}^{N}\psi(t,\theta^{(i)})\psi^{\mathrm{T}}(t,\theta^{(i)}), \qquad (2.70)$$

где $R(\theta)$ – гессиан Гаусса - Ньютона (является положительно полуопределенным). Для нахождения гессиана требуется определение только первых производных, что дает значительные преимущества в вычислительном плане.

По аналогии с методом Ньютона итеративная процедура минимизации критерия (2.68) принимает вид:

$$\theta^{(i+1)} = \theta^{(i)} - \left[R(\theta^{(i)}) \right]^{-1} G(\theta^{(i)}) . \qquad (2.71)$$

На практике направление поиска Гаусса - Ньютона вычисляется на основе решения системы уравнений (2.72)

$$R(\theta^{(i)}) f^{(i)} = -G(\theta^{(i)}) . \qquad (2.72)$$

В случае, когда для определения шага алгоритма используется линейный поиск, алгоритм носит название модифицированного (демпфированного) метода Гаусса - Ньютона [7].

Очевидно, что при равенстве нулю матрицы вторых производных (2.73), (2.74) метод Гаусса - Ньютона идентичен методу Ньютона:

$$H(\theta^{(i)}) = R(\theta^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \psi'(t, \theta^{(i)}) \varepsilon(t, \theta^{(i)}) , \qquad (2.73)$$

$$\psi'(t, \theta) = \frac{d^2 \hat{y}(t(\theta)}{d\theta^2} = 0 . \qquad (2.74)$$

Помимо этого частного случая, локальная сходимость метода линейна. При нулевых или близких к нулю невязках, т.е. при небольших значениях ошибки прогнозирования в окрестности минимума, применение метода Гаусса - Ньютона не дает желаемых результатов. Несмотря на теоретически медленную локальную сходимость метода, в практических приложениях он приводит к лучшим результатам, чем метод Ньютона или квазиньютоновский метод.

Псевдоньютоновский метод. Псевдоньютоновский метод получил широкое распространение в технологии обучения нейронных сетей, так как представляет собой значительное упрощение метода Га-

усса - Ньютона. В алгоритме используется аппроксимация гессиана, полученная путем удаления недиагональных элементов.

Итеративная процедура минимизации определяется следующим выражением:

$$\theta_k^{(i+1)} = \theta_k^{(i)} - \mu^{(i)} G_k(\theta^{(i)}) / R_{kk}(\theta^{(i)}) . \qquad (2.75)$$

Преимущество метода заключается в отсутствии необходимости решать систему уравнений большой размерности при определении направления поиска и значительной экономией оперативной памяти, обусловленной использованием только диагональных элементов гессиана. Однако практическая сходимость алгоритма ниже, чем при использовании метода Гаусса - Ньютона.

Метод Левенберга - Маркардта. Направление поиска в методе Гаусса **-** Ньютона (Ньютона) не является абсолютно оптимальным, так как определяется по аппроксимации критерия $L^{(i)}(\theta)$ в некоторой окрестности текущей итерации. Вследствие того что минимум $L^{(i)}(\theta)$ в общем случае может находиться вне заданной окрестности, выбор направления поиска может оказаться некорректным. Можно предположить целесообразность поиска минимума $L^{(i)}(\theta)$ только в некоторой окрестности текущей итерации. Выбрав в качестве окрестности сферу радиуса $\delta^{(i)}$, можно сформулировать проблему оптимизации следующим образом:

$$\hat{\theta} = \arg\min_{\theta} L^{(i)}(\theta); \left|\theta - \theta^{(i)}\right| \le \delta^{(i)}. \qquad (2.76)$$

Итеративная процедура поиска минимума при наличии ограничений (2.76) может быть представлена следующим образом:

$$\theta^{(i+1)} = \theta^{(i)} + f^{(i)}, \qquad (2.77)$$

$$\left[R(\theta^{(i)}) + \lambda^{(i)} I \right] f^{(i)} = -G(\theta^{(i)}), \qquad (2.78)$$

где параметр $\lambda^{(i)}$ определяет окрестность $\delta^{(i)}$. Данный метод основан на работах [61, 65] и известен как метод Левенберга - Маркардта. Гиперсфера радиуса $\delta^{(i)}$ интерпретируется как окрестность $\theta^{(i)}$, в пределах которой $L^{(i)}(\theta)$ может рассматриваться как адекватная аппроксимация критерия $V_N(\theta, Z^N)$. Этот же принцип, используемый при решении нелинейной задачи наименьших квадратов, носит название подхода «модель – доверительная область» [7].

В отличие от рассмотренных ранее методов, идеология линейного поиска противоречит концепции алгоритма Левенберга - Маркардта, где шаг алгоритма выбирается автоматически в окрестности $\delta^{(i)}$. Взаимосвязь между $\delta^{(i)}$ и параметром $\lambda^{(i)}$ может быть установлена в результате эвристической трактовки влияния $\lambda^{(i)}$ на направление поиска. В случае, когда $\left[R(\theta^{(i)}) + \lambda^{(i)} I \right]$ заменяется диагональной матрицей, направление поиска представляет собой направление антиградиента критерия. Очевидно, что при $\lambda \to \infty$ диагональная матрица доминирует над $R(\theta)$, что приводит к градиентному методу поиска. С другой стороны, установка $\lambda = 0$ приводит к методу Гаусса – Ньютона. Направления поиска при промежуточных значениях $\lambda^{(i)}$ представлены на рис. 2.8. Очевидно наличие взаимосвязи между уменьшением радиуса $\delta^{(i)}$ и увеличением параметра $\lambda^{(i)}$ (и наоборот). К сожалению, явных выражений, определяющих $\lambda^{(i)}$ для конкретного значения

$\delta^{(i)}$, не существует. Это приводит к двум различным методам подстройки $\delta^{(i)}$ – прямым и косвенным. В прямых методах настройка $\delta^{(i)}$ производится непосредственно, после чего применяются итеративные процедуры для определения соответствующего значения $\lambda^{(i)}$ [7]. Косвенные методы представляют собой более простую схему, аналогичную предложенной в оригинальной работе [65]. При использовании косвенных методов происходит непосредственная настройка параметра $\lambda^{(i)}$, причем определение действительного размера доверительной области не производится. Для обеспечения сходимости метода предлагается последовательное увеличение $\lambda^{(i)}$ до тех пор, пока не произойдет уменьшение критерия $L^{(i)}(\theta)$, после чего итерация завершается. Значение параметра λ для следующей итерации уменьшается. Применение метода к обучению НС представлено в работе [43].

Другой косвенный метод представлен в работе [35]. Метод превосходит схему Маркардта, особенно с точки зрения простоты применения. Основная идея метода заключается в сопоставлении реального уменьшения критерия и уменьшения, прогнозируемого на основе аппроксимации $L^{(i)}(\theta)$. В качестве меры точности аппроксимации рассматривается коэффициент

$$r^{(i)} = \frac{V_N(\theta^{(i)}, Z^N) - V_N(\theta^{(i)} + f^{(i)}, Z^N)}{V_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)})}. \qquad (2.79)$$

В случае близости коэффициента к единице, $L^{(i)}(\theta)$ является адекватной аппроксимацией $V_N(\theta, Z^N)$ и значение λ уменьшается (что соответствует увеличению $\delta^{(i)}$). С другой стороны, небольшие или от-

рицательные значения коэффициента приводят к необходимости увеличения λ. Общая схема реализации алгоритма может быть представлена следующим образом:

Шаг 1. Выбрать начальные значения вектора настраиваемых параметров $\theta(0)$ и коэффициента $\lambda(0)$.

Шаг 2. Определить направление поиска из системы уравнений $\left[R(\theta^{(i)})+\lambda^{(i)}I\right]f^{(i)}=-G(\theta^{(i)})$.

Шаг 3. Если $r^{(i)}>0,75 \;\Rightarrow\; \lambda^{(i)}=\lambda^{(i)}/2.$

Шаг 4. Если $r^{(i)}<0,25 \;\Rightarrow\; \lambda^{(i)}=2\lambda^{(i)}.$

Шаг 5. Если $V_N(\theta^{(i)}+f^{(i)},Z^N)<V_N(\theta^{(i)},Z^N)$, то принять $\theta^{(i+1)}=\theta^{(i)}+f^{(i)}$ как новую итерацию и установить $\lambda^{(i+1)}=\lambda^{(i)}$.

Шаг 6. Если критерий останова не достигнут, перейти на шаг 2.

Значения констант (0,25, 0,75 и 2) выбраны произвольно и могут быть изменены без потери сходимости алгоритма. Критерий останова (шаг 6) обсуждается в разделе 2.4.3.

Значение минимизируемого критерия $L^{(i)}\left(\theta^{(i)}+f\right)$ может быть представлено в следующем виде:

$$L^{(i)}(\theta^{(i)}+f)=V_N(\theta^{(i)},Z^N)+f^{\mathrm{T}}G(\theta^{(i)})+\frac{1}{2}f^{\mathrm{T}}R(\theta^{(i)})f. \qquad (2.80)$$

Подставляя в (2.80) значение выражения для определения направления поиска, полученное из соотношения

$$R(\theta^{(i)})f^{(i)}=-G(\theta^{(i)})-\lambda f^{(i)}, \qquad (2.81)$$

получим:

$$V_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)}) = \frac{1}{2}\left(-(f^{(i)})^{\mathrm{T}} G(\theta^{(i)}) + \lambda^{(i)} \left| f^{(i)} \right|^2 \right). \quad (2.82)$$

Соотношение (2.82) позволяет достаточно просто определять на шаге 3 и 4 данного алгоритма коэффициент $r^{(i)}$.

Значение выражения $[V_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)})]$ всегда неотрицательно. Таким образом, если направление поиска, определенное на шаге 2 алгоритма, не приводит к уменьшению критерия, то значение коэффициента $r^{(i)}$ отрицательно и удовлетворяет неравенству, используемому на шаге 4. Следовательно, значение λ увеличивается до тех пор, пока не будет достигнуто уменьшение критерия.

Сходимость метода Левенберга - Маркардта приблизительно такая же, как и у метода Гаусса - Ньютона с демпфированием. Дополнительным преимуществом является хорошая обусловленность гессиана, получаемая за счет добавления диагональной матрицы (2.78). Данный подход является оптимальным для реализации процедуры обучения нейронных сетей, так как обеспечивает быструю сходимость и вычислительную робастность. Основным недостатком метода является необходимость вычисления направления поиска при изменении значения λ вне зависимости от того, производилось изменение весовых коэффициентов или нет.

Очевидно, что выбор доверительной области как сферы в окрестности текущей итерации не является оптимальным в случае, если значения настраиваемых параметров значительно различаются. Это может привести к снижению скорости сходимости метода. По этой причине иногда целесообразно выбирать доверительную область, вводя матрицу масштабирования $D^{(i)}$:

$$\left| D^{(i)}(\theta - \theta^{(i)}) \right| \le \delta^{(i)} . \qquad (2.83)$$

При использовании матрицы масштабирования направление поиска определяется следующим соотношением:

$$\left[R(\theta^{(i)}) - \lambda^{(i)} D^{(i)\mathrm{T}} D^{(i)} \right] f^{(i)} = -G(\theta^{(i)}) . \qquad (2.84)$$

В случае, когда нейронные сети используются в качестве модельных структур для решения задачи идентификации и экспериментальные данные предварительно масштабированы, значительные различия весовых коэффициентов обычно не составляют проблемы.

Рекуррентные методы. Рассмотренные ранее методы оптимизации относятся к классу нерекуррентных методов, или методов пакетной (групповой) обработки. Термин «пакетная обработка» подразумевает использование всего множества экспериментальных данных Z^N на каждой итерации алгоритма оптимизации выбранной модельной структуры. Однако в ряде случаев необходимо идентифицировать систему в режиме реального времени по мере поступления измерений. Типичным примером являются адаптивные системы, в которых на каждом шаге синтеза сигнала управления необходимо иметь адекватную модель реального объекта [9]. Методы идентификации, пригодные для использования в реальном масштабе времени для адаптивного оценивания параметров модели по текущим данным, носят название рекуррентных. Традиционный критерий оптимизации параметров (2.45) не может быть использован в рекуррентных алгоритмах (в случае нестационарности системы). При рекуррентной оптимизации на каждой итерации для настройки параметров используется

только одна входо-выходная пара $[\varphi(t), y(t)]$ из множества экспериментальных данных:

$$\theta(t) = \theta(t-1) + \mu(t) f(t). \qquad (2.85)$$

Следует отметить, что в выражении (2.85) индекс $^{(i)}$ заменен на аргумент t (время).

Большинство рекуррентных алгоритмов [9, 5, 15,19] разработаны для оценки достаточно простых линейных моделей. В случае, если модельная структура содержит большое число настраиваемых параметров, использование рекуррентных алгоритмов в режиме реального времени становится проблематичным. Тем не менее рекуррентные алгоритмы могут быть эффективно использованы и для оценки моделей на всем множестве экспериментальных данных Z^N. Последовательность обработки данных при использовании рекуррентных и пакетных методов обработки представлена на рис. 2.9. Рекуррентные методы имеют следующие преимущества:

- достаточно простая реализация;
- скромные требования к использованию оперативной памяти ЭВМ;
- эффективное использование избыточности множества экспериментальных данных для получения высокой скорости сходимости алгоритма.

Рис. 2.8. Направление поиска в алгоритме Левенберга - Маркардта

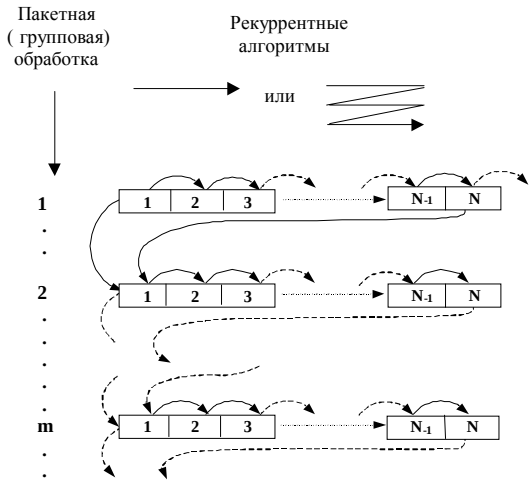

Рис. 2.9. Последовательность обработки экспериментальных данных

при использовании рекуррентных и групповых методов оптимизации параметров модельных структур:

N – число входо-выходных соответствий в экспериментальном

множестве,

m – число «проходов»

Особенности применения рекуррентных алгоритмов к оптимизации параметров нейронных сетей рассмотрены в работе [32].

Рекуррентный метод Гаусса - Ньютона. Алгоритм основан на последовательном включении пар вход-выход в множество данных, используемых для оптимизации модели. При использовании априори полученного множества экспериментальных данных Z^N: $u(t) = u(t+N) + u(t+2N) = \ldots$, $y(t) = y(t+N) + y(t+2N) = \ldots$, критерий, оптимизируемый в каждый момент времени t, определяется следующим выражением:

$$V_t(\theta, Z^t) = \frac{1}{2t}\sum_{k=1}^{t}\varepsilon^2(k,\theta). \qquad (2.86)$$

Значения настраиваемых параметров вычисляются по формуле

$$\theta(t) = \theta(t-1) + R^{-1}(t)V_t'(\theta(t-1), Z^t), \qquad (2.87)$$

где градиент определяется как

$$V_t'(\theta, Z^t) = -\frac{1}{t}\sum_{k=1}^{t}\psi(k,\theta)\varepsilon(k,\theta) = -\frac{1}{t}\sum_{k=1}^{t-1}\psi(k,\theta)\varepsilon(k,\theta) - \\ -\frac{1}{t}\psi(t,\theta)\varepsilon(t,\theta) = \frac{t-1}{t}V_t'\left(\theta, Z^{t-1}\right) - \frac{1}{t}\psi(t,\theta)\varepsilon(t,\theta). \qquad (2.88)$$

Предположим, что вектор параметров модели $\theta(t-1)$ доставляет минимум критерию (2.86) в момент времени $t-1$, т.е. $V_{t-1}'(\theta(t-1), Z^{t-1}) = 0$, что приводит к выражению

$$V_t'(\theta(t-1), Z^t) = -\frac{1}{t}\psi(t,\theta(t-1))\varepsilon(t,\theta(t-1)). \qquad (2.89)$$

Выражение (2.70) для определения гессиана Гаусса **-** Ньютона может быть переписано в следующем виде:

$$R(t) = \frac{1}{t}\sum_{k=1}^{t}\psi(k,\theta)\psi^{\mathrm{T}}(k,\theta) = R(t-1) + \frac{1}{t}(\psi(t,\theta)\psi^{\mathrm{T}}(t,\theta) - R(t-1)). \quad (2.90)$$

Таким образом, получаем следующие выражения для настройки параметров модели на текущей итерации:

$$\theta(t) = \theta(t-1) + \frac{1}{t}R^{-1}(t)\psi(t)\left(y(t) - \hat{y}(t\left(\theta(t-1)\right)\right), \qquad (2.91)$$

$$R(t) = R(t-1) + \frac{1}{t}(\psi(t,\theta)\psi^{\mathrm{T}}(t,\theta) - R(t-1)). \qquad (2.92)$$

Чтобы избежать обращения гессиана Гаусса **-** Ньютона можно применить следующее выражение (лемма об обращении матриц):

$$(A^{-1} + BCD)^{-1} = A - AB(C^{-1} + DAB)^{-1} \qquad (2.93)$$

непосредственно к матрице ковариации $P(t) = \frac{1}{t}R^{-1}(t)$:

$$P(t) = \frac{1}{t}R^{-1}(t) = P(t-1) - \frac{P(t-1)\psi(t)\psi^{\mathrm{T}}(t)P(t-1)}{1+\psi^{\mathrm{T}}(t)P(t-1)\psi(t)}. \qquad (2.94)$$

Начальное значение обычно выбирается как $P(0) = cI$, где c **–** достаточно «большое» число, обычно 10^4 **–** 10^8.

Для модельных структур типа ARX алгоритм представляет собой традиционный рекуррентный метод наименьших квадратов (РМНК) [9]. Причиной использования адаптивных методов и рекур-

рентной идентификации на практике является то, что свойства системы могут изменяться во времени, а алгоритмы идентификации должны отслеживать эти изменения. Это достигается путем взвешивания экспериментальных данных, причем меньшие веса назначаются более старым измерениям, которые мало информативны. Для адаптивной оценки нестационарных систем могут быть применены различные модификации рекуррентных алгоритмов.

Алгоритм экспоненциального затухания. Одним из способов удаления устаревшей информации из множества экспериментальных данных является введение в критерий (2.86) фактора затухания λ:

$$V_t(\theta, Z^t) = \frac{1}{2t}\sum_{k=1}^{t}\lambda^{t-k}\varepsilon^{\mathrm{T}}(k,\theta)\varepsilon(k,\theta). \qquad (2.95)$$

Оптимизационная процедура может быть представлена следующим образом:

$$K(t) = \frac{P(t-1)\psi(t)}{1+\psi^{\mathrm{T}}(t)P(t-1)\psi(t)},$$

$$\theta(t) = \theta(t-1) + K(t)\left(y(t) - \hat{y}(t(\theta(t-1))\right), \qquad (2.96)$$

$$P(t) = \left(P(t-1) - K(t)\psi^{\mathrm{T}}(t)P(t-1)\right)/\lambda,$$

где λ – фактор затухания – выбирается в интервале $[0,1]$. В случае, если в некотором направлении пространства параметров затухание происходит быстрее, чем появляются новые данные, собственные значения матрицы ковариации стремительно возрастают. Эта проблема может быть решена путем введения ограничений на собствен-

ные значения матрицы ковариации. Алгоритм с ограничениями может быть представлен в следующем виде:

$$K(t) = \alpha P(t-1)\psi\left(1+\psi^{\mathrm{T}}(t)P(t-1)\psi(t)\right)^{-1},$$

$$\theta(t) = \theta(t-1) + K(t)\left(y(t) - \hat{y}(t\,|\,\theta(t-1))\right), \qquad (2.97)$$

$$P(t) = \frac{1}{\lambda}P(t-1) - K(t)\psi^{\mathrm{T}}(t)P(t-1) + \beta I - \delta P^2(t-1),$$

где $\alpha,\beta,\delta,\lambda$ – параметры, настраиваемые с учетом следующих ограничений:

$$\begin{cases} 0 < \gamma < \alpha < 1, \\ (\gamma-\alpha)^2 + 4\beta\delta < (1-\alpha)^2, \\ \beta > 0, \delta > 0. \end{cases} \qquad (2.98)$$

В неравенствах (2.98) значение $\gamma \equiv (1-\lambda)/\lambda$. Минимальные и максимальные значения матрицы ковариации выбираются на основе следующих выражений:

$$\alpha_{\min} = \left(\frac{\alpha-\gamma}{2\delta}\right)\left(\sqrt{1+\frac{4\beta\delta}{(\alpha-\gamma)^2}}-1\right), \qquad (2.99)$$

$$\alpha_{\max} = \frac{\gamma}{2\delta}\left(\sqrt{1+\frac{4\beta\delta}{\gamma^2}}+1\right). \qquad (2.100)$$

Рекуррентный градиентный метод. Оптимизация параметров в рекуррентной модификации градиентного метода реализуется путем подстановки $\frac{1}{t}R^{-1} = \mu I$ в выражение (2.91). В теории нейронных сетей

подход получил название рекуррентного метода обратного распространения ошибки [24].

4.2. Регуляризация и концепция обобщения

В разделе 2.4.1 рассматривались методы отображения множества экспериментальных данных на некоторую модельную структуру с целью получения оптимальной, в силу среднеквадратичного критерия, оценки. В настоящем разделе рассматриваются методы регуляризации, применяемые к нейросетевым моделям с целью улучшения их рабочих характеристик (в частности, свойств к обобщению) [59, 76].

Предположим, что система может быть представлена некоторой функцией f от предыдущих значений экспериментальных данных Z^{t-1} с аддитивными помехами типа белого шума $e(t)$:

$$y(t) = f(Z^{t-1}) + e(t)\,. \qquad (2.101)$$

Данная реальная система может быть представлена с некоторой степенью точности нейронной сетью с конечным числом настраиваемых параметров (весовых коэффициентов). Тем не менее, можно предположить, что множество данных генерируется абсолютно оптимальной нейросетевой моделью g_0:

$$y(t) = g_0(\varphi(t, \theta_0), \theta_0) + e(t)\,. \qquad (2.102)$$

Принцип получения модели, рассмотренный в разделе 2.4.1, состоит в отображении множества экспериментальных данных Z^N на модельную структуру M, содержащую p настраиваемых параметров:

$$\hat{y}(t\,|\,\theta) = g(\varphi(t, \theta), \theta),\ \theta \in D_M \subset R^p\,. \qquad (2.103)$$

Настраиваемые параметры определяются в соответствии со следующим выражением:

$$\hat{\theta} = \arg \min_{\theta} V_N(\theta, Z^N) = \frac{1}{2N} \sum_{t=1}^{N} \left(y(t) - \hat{y}(t|\theta) \right)^2 . \qquad (2.104)$$

Очевидно, что более оптимальным, чем V_N, критерием является математическое ожидание ошибки прогнозирования, называемое ошибкой обобщения:

$$\bar{V}(\theta) = \frac{1}{2} E \left\{ \left(y(t) - \hat{y}(t|\theta) \right)^2 \right\} . \qquad (2.105)$$

Оценка критерия (2.105) практически невозможна, тем не менее при наличии соответствующих условий [9] выполняется следующее соотношение:

$$\lim_{N \to \infty} V_N(\theta, Z^N) = \bar{V}(\theta) . \qquad (2.106)$$

Таким образом, $\hat{\theta} \to \theta^*$ при $N \to \infty$, где набор параметров θ^* доставляет минимум ошибке обобщения (2.105). Если реальная система, представленная описанием (2.102), действительно входит в модельную структуру $S \in M$, оценка параметров также является состоятельной: $\theta^* = \theta_0$. В действительности множество экспериментальных данных всегда конечно. В работе [9] рассматриваются вопросы сходимости $\hat{\theta}$ к θ^*. В частности, показано, что оценка $\hat{\theta}$ асимптотически нормальна со средним значением θ^* и матрицей ковариации P_θ:

$$\hat{\theta} \in As\ N\left(\theta^*, \frac{1}{N} P_\theta \right) . \qquad (2.107)$$

При условии, что $S \in M$, асимптотическая матрица ковариации определяется соотношением

$$P_\theta = \sigma_e^2 \left[E\left\{ \psi(t,\theta_0)\psi^{\mathrm{T}}(t,\theta_0) \right\} \right]^{-1} \approx 2V_N(\hat{\theta}, Z^N) \left[\frac{1}{N} \sum_{t=1}^{N} \psi(t,\hat{\theta})\psi^{\mathrm{T}}(t,\hat{\theta}) \right]^{-1}. \quad (2.108)$$

Ошибка обобщения может быть получена путем оценки обученной нейросетевой модели на тестовом множестве данных Z^{T}, не используемых при обучении нейросети. В случае, если значение критериальной функции $\bar{V}(\hat{\theta}) \approx V_T(\hat{\theta}, Z^{\mathrm{T}})$ близко к $V_N(\hat{\theta}, Z^N)$, значения параметров $\hat{\theta}$ близки к θ^*, то обученная нейросетевая модель - удовлетворительна. Если в силу некоторых причин проверка на множестве тестовых данных невозможна, оценка ошибки обобщения достаточно затруднительна. Другое ограничение на оценку обобщения связано с ее зависимостью от множества Z^N через вектор параметров $\hat{\theta}$. Следовательно, ошибка обобщения не показывает, насколько хорошо выбрана модельная структура, т.е. как будет вести себя конкретная модель, обученная на множестве Z^N, при предъявлении сигналов, не вошедших в это множество. Таким образом, целесообразно ввести среднюю ошибку обобщения как меру качества модели:

$$V_M = E\left\{ \bar{V}(\hat{\theta}) \right\}, \quad (2.109)$$

где $E(*)$ – математическое ожидание критерия по множеству данных размера N. Одним из примеров является оценка финальной ошибки прогнозирования (ФОП) Акайке [9, 21], состоятельная при условии принадлежности реальной системы выбранной модельной структуре $S \in M$:

$$\hat{V}_M = \frac{1}{2}\sigma_e^2\left(1+\frac{p}{N}\right). \qquad (2.110)$$

Минимальным значением ошибки прогнозирования является половина дисперсии шума $\frac{1}{2}\sigma_e^2$, но по причине конечности множества данных реальное значение всегда больше. При анализе причин, влияющих на увеличение ошибки прогнозирования, целесообразно выделить две следующие составляющие:

- смещение: составляющая ошибки, обусловленная недостаточностью модельной структуры ($S \notin M$). Если нейронная сеть не содержит достаточного числа настраиваемых параметров, то при $N \to \infty$ значения весовых коэффициентов сходятся к θ^*, отличным от θ_0;
- дисперсионная составляющая: обусловлена обучением нейросети на недостаточно большом множестве «зашумленных» данных.

Если предположить, что невязки системы $g_0(\varphi(t),\theta_0) - g(\varphi(t),\theta^*)$ являются по сути белым шумом, то

$$V_M = E\left\{\overline{V}(\hat{\theta})\right\} = E\left\{\left|g_0(\varphi(t),\theta_0) - g(\varphi(t),\hat{\theta})\right|^2\right\} + \sigma_e^2 \approx$$
$$\approx \underbrace{E\left\{\left|g_0(\varphi(t),\theta_0) - g(\varphi(t),\theta^*)\right|^2\right\}}_{\text{смещение}} + \underbrace{E\left\{\left|g(\varphi(t),\theta^*) - g(\varphi(t),\hat{\theta})\right|^2\right\}}_{\text{дисперсия}} + \sigma_e^2. \qquad (2.111)$$

Как было отмечено ранее, принадлежность реальной системы выбранной модельной структуре практически недостижима, следовательно, всегда существует некоторое смещение. Очевидно, что смещение уменьшается по мере роста числа настраиваемых параметров нейросетевой модели. Тем не менее увеличение числа параметров приводит к увеличению дисперсионной составляющей ошибки. Это

явление носит название дилеммы смещения / дисперсии [38], что может быть проиллюстрирована следующим практическим примером:

Десять различных модельных структур типа NNARX (2, 2, 1) обучаются на множестве зашумленных экспериментальных данных Z^N, где $N = 500$. Простейшая модельная структура содержит один нейрон в скрытом слое. Структуры последовательно наращиваются на один нейрон, что соответствует увеличению числа настраиваемых параметров на 6 единиц. Обученная нейросеть проверяется на тестовом множестве Z^{T}, содержащем 2000 входо-выходных пар. Критерий $V_T(\hat{\theta}, Z^{\mathrm{T}})$ интерпретируется как оценка средней ошибки обобщения. В связи с возможностью существования локальных минимумов каждая нейросетевая структура обучается 5 раз при различных начальных значениях весовых коэффициентов. Результаты эксперимента представлены на рис. 2.10. По результатам эксперимента можно определить, что компромисс между ошибкой смещения и дисперсии достигается при использовании 4 – 6 нейронов в скрытом слое. При увеличении числа нейронов доминирует дисперсионная составляющая ошибки смещения. Это явление объясняется избыточностью структуры нейросети, т.е. обученная модель включает не только признаки исследуемой системы, но и нежелательные возмущения, содержащиеся в обучающем множестве. Недостаточность модельной структуры, напротив, приводит к доминированию ошибки смещения.

Одним из способов решения проблемы смещения / дисперсии является расширение критерия $V_N(\hat{\theta}, Z^N)$ регуляризующим компонентом (коэффициентом сложности модельной структуры). Этот компонент

может быть введен как коэффициент затухания весовых коэффициентов нейросетевой модели:

$$W_N(\theta, Z^N) = \frac{1}{2N}\sum_{t=1}^{N}\left(y(t) - \hat{y}(t|\theta)\right)^2 + \frac{1}{2N}\theta^{\mathrm{T}} D\theta, \qquad (2.112)$$

где D – диагональная матрица, определяемая соотношением $D = \alpha I$; α – коэффициент затухания весов. В некоторых случаях различные значения коэффициента затухания используются для весов входного - скрытого и скрытого - выходного слоев соответственно, иногда собственные значения коэффициентов устанавливаются для каждой структурной связи. Очевидно, что введение регуляризующего компонента уменьшает ошибку смещения. Оценка ФОП для нейросетевых моделей, обученных в соответствии с критерием (2.112), при условии, что $D = \alpha I$ и $S \in M$, может быть получена следующим образом [76]:

$$\hat{V}_M = \frac{1}{2}\left(\sigma_e^2\left(1 + \frac{p_1}{N}\right) + \gamma\right), \qquad (2.113)$$

где

$$p_1 = \mathrm{tr}\left[R\left(R + \frac{\alpha}{N}I\right)^{-1} R\left(R + \frac{\alpha}{N}I\right)^{-1}\right], \qquad (2.114)$$

$$\gamma = \frac{\alpha^2}{N^2}\theta_0^{\mathrm{T}}\left(R + \frac{\alpha}{N}I\right)^{-1} R\left(R + \frac{\alpha}{N}I\right)^{-1}\theta_0 \le \frac{\alpha}{4N}|\theta|^2, \qquad (2.115)$$

$$R = E\left\{\psi(t,\theta_0)\psi^{\mathrm{T}}(t,\theta_0)\right\} \approx \frac{1}{N}\sum_{t=1}^{N}\psi(t,\hat{\theta})\psi^{\mathrm{T}}(t,\hat{\theta}). \qquad (2.116)$$

Так как след матрицы равен сумме собственных значений, число настраиваемых параметров нейросетевой модели p_1 определяется следующим соотношением:

$$p_1 = \sum_{i=1}^{p} \frac{\delta_i^2}{(\delta_i + \alpha / N)^2}, \qquad (2.117)$$

где δ_i является i – собственным числом гессиана R. В разделе 2.4.1 было показано, что излишняя связь (весовой коэффициент) приводит к нулевому значению собственного числа гессиана. На практике ни один из весовых коэффициентов не может быть излишним, так как нейросетевая структура не может быть избыточной. Таким образом, гессиан всегда положительно определен. Однако малосущественные весовые коэффициенты приводят к небольшим собственным значениям гессиана и наоборот. Это явление может быть объяснено путем рассмотрения производных выходных сигналов по вектору входов ($\psi(t, \theta)$) как матрицы чувствительности. Если весовой коэффициент i несуществен, его производная будет мала при всех значениях t. В этом случае все диагональные элементы, так же как и элементы строки i (столбца i) матрицы R, будут малы, что приводит к небольшим собственным значениям гессиана. Для более существенных весовых коэффициентов наблюдается противоположный эффект. Следовательно, можно разделить собственные числа гессиана на две группы: группу, соответствующую весовым коэффициентам с небольшой значимостью, и группу более значимых весовых коэффициентов. Если предположить, что значение α / N больше минимального собственного числа гессиана и меньше максимального, то число p_1 можно рас-

сматривать как число эффективных (значимых) весовых коэффициентов нейросетевой модели. В такой интерпретации (при условии, что значением γ можно пренебречь) оценка ФОП для регуляризованного критерия совпадает с оценкой для нерегуляризованного критерия. Настраивая параметр затухания весовых коэффициентов, можно определить эффективный размер нейросетевой структуры. Основной проблемой является выбор затухания весов, минимизирующего среднюю ошибку обобщения.

Следует отметить, что оценки типа (2.110), (2.113) не могут быть вычислены непосредственно при отсутствии информации о статистических характеристиках шумов. В разделе 2.4.4. рассматриваются варианты решения этой проблемы.

Эффект регуляризации может быть также достигнут путем останова оптимизационной процедуры до момента достижения минимума. Этот факт может быть проиллюстрирован следующим практическим примером:

Рассмотрим модельную структуру типа NNARX, содержащую 20 нейронов в скрытом слое (случай чрезмерной параметризации). Нейронная сеть обучается по методу Левенберга - Маркардта. На каждой итерации производится оценка ошибки обучения и обобщения (рис. 2.11). Ошибка обучения является монотонно убывающей функцией от номера итерации (вследствие применения метода Левенберга - Маркардта). Ошибка обобщения убывает только в начале процедуры обучения, а затем, после достижения минимума, увеличивается. Это объясняется тем, что в начале процедуры нейросетевая модель

обучается на характерные признаки системы, после чего идет подстройка под возмущения, отраженные в обучающем множестве.

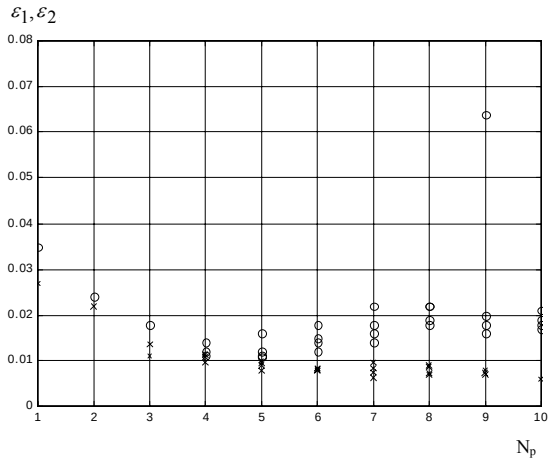

Рис. 2.10. Результаты пятикратного обучения (с различными начальными условиями) 10-ти различных модельных структур на множестве данных Z^N:

«x» – ошибка обучения (ε_1); *«o» – оценка ошибки прогнозирования на тестовом множестве* (ε_2); *N_p – число нейронов в скрытом слое*

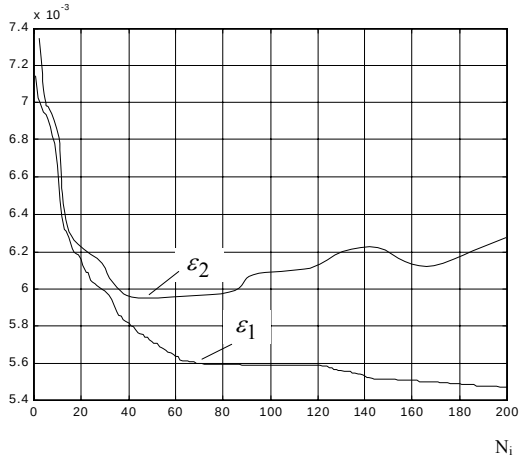

Рис. 2.11. Изменения ошибки обучения (ε_1) и оценки ошибки прогнозирования на тестовом множестве (ε_2) в ходе процедуры оптимизации нейросетевой структуры методом Левенберга – Маркардта: N_i – номер итерации оптимизационной процедуры

В работе [76] показано, что эффект предварительного останова не только аналогичен эффекту регуляризации, но и имеет с ним много общего по сути. Тем не менее рекомендуется на практике отдавать предпочтение именно прямым методам регуляризации, а не предварительному останову, так как большинство методов структурной оптимизации и подтверждения модели предполагают достижение оптимизационной процедурой точки минимума.

4.3. Особенности оптимизации параметров нейросетевых модельных структур

В настоящем разделе рассматриваются аспекты практического применения методов безусловной оптимизации к обучению нейросетевых моделей.

Обучение с использованием затухания весовых коэффициентов. В разделе 2.4.2 рассмотрены методы обучения нейросетевых моделей на основе регуляризованного критерия $W_N(\theta^{(i)}, Z^N)$. Регуляризация осуществляется путем добавления дополнительных компонентов к градиенту и гессиану:

$$G(\theta) = W_N'(\theta, Z^N) = V_N'(\theta, Z^N) + \frac{1}{N} D\theta, \qquad (2.118)$$

$$H(\theta) = W_N''\left(\theta, Z^N\right) = V_N''\left(\theta, Z^N\right) + \frac{1}{N} D. \qquad (2.119)$$

Реализация метода Левенберга – Маркардта требует некоторых дополнительных модификаций. Аппроксимация Гаусса – Ньютона критерия $W_N(\theta^{(i)}, Z^N)$ может быть представлена в следующем виде:

$$W_N(\theta, Z^N) \approx L^{(i)}(\theta) = \frac{1}{2N}\left(\sum_{t=1}^{N} (\tilde{\varepsilon}(t, \theta))^2 + \theta^{\mathrm{T}} D\theta\right), \qquad (2.120)$$

где гессиан определяется следующим выражением:

$$R(\theta) = L^{(i)''}(\theta^{(i)}) = \frac{1}{N}\left(\sum_{t=1}^{N} \psi(t, \theta^{(i)})\psi^{\mathrm{T}}(t, \theta^{(i)}) + D\right). \qquad (2.121)$$

Показатель, используемый для подстройки параметра Левенберга – Маркардта (λ), может быть найден как

$$r^{(i)} = \frac{W_N(\theta^{(i)}, Z^N) - W_N(\theta^{(i)} + f^{(i)}, Z^N)}{W_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)})}. \qquad (2.122)$$

Знаменатель $W_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)})$ в выражении (2.122) может быть определен непосредственно из следующих матричных преобразований:

$$W_N(\theta^{(i)}, Z^N) - L^{(i)}(\theta^{(i)} + f^{(i)}) = \\ = \frac{1}{2}\left((f^{(i)})^{\mathrm{T}} \left(G(\theta^{(i)}) + \left(\lambda^{(i)} I + \frac{1}{N} D \right) f^{(i)} \right) \right). \qquad (2.123)$$

Вычисление градиентов. За исключением метода Ньютона, требующего вычисления вторых производных, единственным определяющим компонентом процедуры оптимизации является производная прогноза нейросетевой модели по вектору настраиваемых параметров (весовых коэффициентов НС) $\psi(t, \theta)$.

Для модельных структур типа NNARX значение $\psi(t, \theta)$ определяется следующим выражением:

$$\psi(t, \theta) = \frac{d\hat{y}(t|\theta)}{d\theta} = \frac{\partial \hat{y}(t|\theta)}{\partial \theta} = \phi(t). \qquad (2.124)$$

Значение $\psi(t, \theta)$ для NNARMAX моделей определяется как

$$\psi(t, \theta) = \frac{d\hat{y}(t|\theta)}{d\theta} = \frac{\partial \hat{y}(t|\theta)}{\partial \theta} - \frac{\partial \hat{y}(t|\theta)}{\partial \varepsilon(t-1, \theta)} \frac{d\hat{y}(t-1|\theta)}{d\theta} - \cdots - \\ - \frac{\partial \hat{y}(t|\theta)}{\partial \varepsilon(t-k, \theta)} \frac{d\hat{y}(t-k|\theta)}{d\theta} = \\ = \phi(t) - c_1(t)\psi(t-1, \theta) - \cdots - c_k \psi(t-k, \theta), \qquad (2.125)$$

или, при введении $C(t, q^{-1}) = 1 + c_1(t) q^{-1} + \cdots + c_k(t) q^{-k}$,

$$\psi(t,\theta) = \frac{1}{C(t,q^{-1})}\phi(t)\,. \qquad (2.126)$$

Зависимость регрессионного вектора от весовых коэффициентов нейронной сети прослеживается при сопоставлении (2.126) и (2.124). При использовании NNARMAX моделей градиент получается в результате «временной линейной фильтрации» частной производной $\phi(t)$. Очевидно, что это может привести к проблемам с устойчивостью алгоритма, особенно при начальной инициализации весовых коэффициентов случайными числами. Таким образом, для получения приемлемого решения иногда приходится проводить повторное обучение НС.

Вычисление градиента для моделей типа «обновления пространства состояний» (NNSSIF) представляет более трудоемкую процедуру, чем для рассмотренных ранее случаев. Полагая, что все компоненты вектора состояний могут быть оценены, получаем следующее выражение для определения градиента:

$$\psi(t,\theta) = \frac{d\hat{x}(t|\theta)}{d\theta}C^{\mathrm{T}} = \psi_x(t,\theta)C^{\mathrm{T}}, \qquad (2.127)$$

где

$$\begin{aligned}\psi_x(t,\theta) = \frac{d\hat{x}^{\mathrm{T}}(t|\theta)}{d\theta} = \frac{\partial\hat{x}^{\mathrm{T}}(t|\theta)}{\partial\theta} + \frac{d\hat{x}^{\mathrm{T}}(t-1|\theta)}{d\theta}\frac{\partial\hat{x}^{\mathrm{T}}(t|\theta)}{\partial\hat{x}(t-1,\theta)} - \\ - \frac{d\hat{y}(t-r|\theta)}{d\theta}\frac{\partial\hat{x}^{\mathrm{T}}(t|\theta)}{\partial\varepsilon(t-1,\theta)} = \phi(t) + \psi_x(t-1,\theta)A(t) - \\ - \psi(t-r,\theta)K^{\mathrm{T}}(t) = \phi(t) + \psi_x(t-1,\theta)(A(t)-K(t))^{\mathrm{T}}.\end{aligned} \qquad (2.128)$$

В случае невозможности получения полной информации о векторе состояний необходимо использовать несколько нейронных сетей (аналог фильтра Калмана [9]). Однако при вычислении градиентов

можно рассматривать модель как единую НС. В этом случае имеет место следующее преобразование матрицы $A(t)$ для всех значений $j \in \{q_i\}$ и $k \in \{q_i + 1\}$:

$$\begin{cases} A_{j,k} = 1, \; j = k + 1, \\ A_{j,k} = 0, \; i \neq k + 1. \end{cases} \tag{2.129}$$

Определение остальных компонент матриц $A(t)$ и $K(t)$ осуществляется с использованием соотношения (2.133).

Частная производная $\phi(t)$ является производной прогнозируемого значения (выхода НС) по весовым коэффициентам нейронной сети в случае пренебрежения зависимостью регрессора от весовых коэффициентов.

Обозначим обобщенную выходную переменную НС модели как $\hat{z}_k$. Тогда для двухслойной НС с линейными активационными функциями нейронов выходного слоя и тангенциальными активационными функциями нейронов скрытого слоя

$$\begin{aligned} \hat{z}_k(t|\theta) = g_k(\varphi(t,\theta),\theta) = \sum_{j=1}^{n_h} W_{kj} \tanh\left(\sum_{l=1}^{n_\varphi} w_{ji} \varphi_l(t,\theta) + w_{j0} \right) + W_{i0} = \\ = \sum_{j=1}^{n_h} W_{kj} h_j(t,\theta) + W_{i0}. \end{aligned} \tag{2.130}$$

Частные производные вычисляются следующим образом:

$$\frac{\partial \hat{z}_k(t|\theta)}{\partial W_{ij}} = \begin{cases} h_j(t), \; j > 0, \; k = i, \\ 1, \quad j = 0, \; k = i, \\ 0, \quad i \neq k; \end{cases} \tag{2.131}$$

$$\frac{\partial \hat{z}_k(t|\theta)}{\partial w_{ji}} = \begin{cases} W_{kj}(1-h_j^2(t))\varphi_l(t,\theta), l>0, \\ W_{kj}(1-h_j^2(t)), \ l=0. \end{cases} \quad (2.132)$$

Количество строк матрицы $\phi(t)$ определяется числом весовых коэффициентов НС, число столбцов равно числу выходов НС.

Якобиан нейросетевой модели. Якобиан, или мгновенная матрица усиления, является производной выхода НС по входам для заданной пары «вход-выход».

Для двухслойной НС с линейными активационными функциями нейронов выходного слоя и тангенциальными активационными функциями нейронов скрытого слоя частная производная по произвольному входу определяется как

$$\frac{\hat{z}_k(t|\theta)}{\partial \varphi_l(t,\theta)} = \sum_{j=1}^{n_h} W_{kj} w_{jl} \left[1 - \tanh^2 \left(\sum_{l=1}^{n_\varphi} w_{ji}\varphi_l(t,\theta) + w_{j0} \right) \right] = \quad (2.133)$$
$$= \sum_{j=1}^{n_h} W_{kj} w_{jl} \left[1 - h_j^2(t,\theta) \right].$$

Метод обратного распространения (ошибки). В случае, когда нейронная сеть содержит более одного скрытого слоя с нелинейными функциями активации, выражения для определения значений $\phi(t)$, соответствующих элементам матрицы частных производных, очевидно, становятся более сложными. Алгоритм определения градиента минимизируемого критерия для сети с произвольным числом скрытых слоев и произвольным видом активационных функций, использующий особенности структуры НС, носит название метода обратного распространения (ошибки), или обобщенного дельта-правила [6]. Так как алгоритм рассчитан на обучение НС прямого действия, то он может

быть непосредственно применен только к модельным структурам типа NNARX. Тем не менее метод может быть модифицирован и для получения частных производных $\phi(t)$.

Градиент критерия наименьших квадратов (рассматривается общий случай НС с несколькими выходами) может быть представлен следующим образом:

$$G(\theta) = V_N'(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\frac{\partial \varepsilon^{\mathrm{T}}(t,\theta)}{\partial \varepsilon}\varepsilon(t,\theta) = \\ = -\frac{1}{N}\sum_{t=1}^{N}\frac{\partial \hat{y}^{\mathrm{T}}(t|\theta)}{\partial \theta}(y(t) - \hat{y}(t|\theta)). \qquad (2.134)$$

Прогнозируемое НС выходное значение вычисляется в соответствии со следующим выражением (для выхода k):

$$\hat{y}_k(t|\theta) = F_k\left(\sum_{j=0}^{n_h} W_{kj} h_j(t)\right) = F_k\left(\sum_{j=1}^{n_h} W_{kj} f_j\left(\sum_{l=0}^{n_\varphi} w_{jl}\varphi_l(t)\right) + W_{k0}\right), \quad (2.135)$$

где f_j – активационная функция нейрона скрытого слоя *j*, а F_k – активационная функция выхода k. Для простоты, нейронные смещения представлены как добавочные весовые коэффициенты, т.е. $h_0(t) = \varphi_0(t) = 1$.

Частные производные выходов НС по весовым коэффициентам определяются следующими соотношениями:

$$\phi_{jk}^{(W)} = \frac{\partial \hat{y}_k(t|\theta)}{\partial W_{kj}} = h_j(t) F_k'\left(\sum_{j=0}^{n_h} W_{kj} h_j(t)\right), \qquad (2.136)$$

$$\phi_{kjl}^{(w)} = \frac{\partial \hat{y}_k(t|\theta)}{\partial w_{jl}} = \varphi_l(t) f_j'\left(\sum_{j=0}^{n_\varphi} w_{jl}\varphi_l(t)\right) W_{kj} F_k'\left(\sum_{j=0}^{n_h} W_{kj} h_j(t)\right), \quad (2.137)$$

что приводит к следующему выражению для определения градиента скрытого-выходного слоя:

$$G(W_{kj}) = \sum_{t=1}^{N} h_j(t) F_k' \left(\sum_{j=0}^{n_h} W_{kj} h_j(t) \right) \left(y_k(t) - \hat{y}_k(t|\theta^{(i)}) \right) = \\ = \sum_{t=1}^{N} h_j(t) \delta_k^{(W)}(t), \tag{2.138}$$

где «ошибка» или «дельта» вводится как

$$\delta_k^{(W)}(t) = F_k' \left(\sum_{j=0}^{n_h} W_{kj} h_j(t,\theta) \right) \left(y_k(t) - \hat{y}_k(t|\theta) \right). \tag{2.139}$$

Градиент для входного-скрытого слоя определяется как

$$G(w_{jl}) = \sum_{t=1}^{N} \varphi_l(t) f_j' \left(\sum_{l=0}^{n_\varphi} w_{jl} \varphi_l(t) \right) \sum_{k=1}^{n_y} W_{kj}(t) \delta_k^{(W)}(t) = \\ = \sum_{t=1}^{N} \varphi_l(t) \delta_j^{(W)}(t), \tag{2.140}$$

где

$$\delta_j^{(w)}(t) = f_j' \left(\sum_{l=0}^{n_\varphi} w_{jl} \varphi_l(t) \right) \sum_{k=1}^{n_y} W_{kj}(t) \delta_k^{(W)}(t). \tag{2.141}$$

Аналогичным образом алгоритм обобщается на случай произвольного числа скрытых слоев НС. Процедура заключается в распространении по НС значения «дельта» (2.139) слой за слоем в обратном направлении.

При необходимости частные производные $\phi(t)$ могут быть получены из выражений (2.138) и (2.140). Из выражения (2.139) удаляется значение ошибки прогнозирования (невязки), затем отдельно для ка-

ждого выхода применяется метод обратного распространения ошибки.

Определение направления поиска. При использовании методов Ньютона, Гаусса - Ньютона и Левенберга - Маркардта для определения направления поиска необходимо решать системы линейных уравнений [7, 13, 14]. Методы, основанные на использовании таких свойств гессиана Гаусса - Ньютона, как положительная определенность и симметричность, рассмотрены в работе [40].

Многомерные системы. Системы, имеющие векторный вход и (или) выход, называются многомерными. Очевидно, что построение моделей таких систем является более сложной процедурой (по сравнению с одномерным случаем). Тем не менее при многомерном входе могут быть использованы рассмотренные в настоящей работе методы оптимизации. Модели с несколькими выходами имеют гораздо более сложную структуру, вследствие чего их параметризация нетривиальна. Существует несколько подходов к решению данной задачи. Наиболее простым (с точки зрения практического применения) является использование независимой модели для каждого выхода. В этом случае задача может быть решена без каких-либо дополнительных модификаций. Однако более естественно рассматривать систему в целом, тогда модифицированный критерий идентификации

$$V_N(\theta, Z^N) = \frac{1}{2N}\sum_{t=1}^{N}(y(t) - \hat{y}(t|\theta))^{\mathrm{T}}(y(t) - \hat{y}(t|\theta)) = \\ = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon^{\mathrm{T}}(t,\theta)\varepsilon(t,\theta). \qquad (2.142)$$

Критерий (2.142) не является достаточно эффективным при значительном различии дисперсии шумов по каждому входу и существенной взаимно-корреляционной функции. Вместо критерия (2.142) может быть использована следующая модификация:

$$V_N(\theta, Z^N) = \frac{1}{2N}\sum_{t=1}^{N} \varepsilon^{\mathrm{T}}(t,\theta)\Lambda^{-1}\varepsilon(t,\theta), \qquad (2.143)$$

где Λ – матрица ковариации шумов, $\Lambda = E\{e(t)e^{\mathrm{T}}(t)\}$. Оценка, полученная путем минимизации критерия (2.143), соответствует оценке максимального правдоподобия при условии нормального распределения шумов и известной матрице Λ. При реализации метода Левенберга - Маркардта используются следующие выражения для определения градиента и гессиана:

$$G(\theta) = V_N'(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \psi(t,\theta)\Lambda^{-1}(y(t) - \hat{y}(t|\theta)), \qquad (2.144)$$

$$R(\theta) = \frac{1}{N}\sum_{t=1}^{N} \psi(t,\theta)\Lambda^{-1}\psi(t,\theta), \qquad (2.145)$$

где матрица производных выхода по весовым коэффициентам определяется как

$$\psi(t,\theta) = \frac{\partial \hat{y}^{\mathrm{T}}(t|\theta)}{\partial\theta}. \qquad (2.146)$$

Алгоритм Гаусса – Ньютона (версия с экспоненциальным затуханием) модифицируется следующим образом:

$$K(t) = P(t-1)\psi(t)\left(\lambda\Lambda^{-1} + \psi^{\mathrm{T}}(t)P(t-1)\psi(t)\right)^{-1},$$

$$\theta(t) = \theta(t-1) + K(t)\Lambda^{-1}\left(y(t) - \hat{y}(t|\theta(t-1))\right), \qquad (2.147)$$

$$P(t) = \left(P(t-1) - K(t)\psi^{\mathrm{T}}(t)P(t-1)\right)/\lambda .$$

В случае, когда матрица ковариации неизвестна, возникает необходимость определения Λ одновременно с определением весовых коэффициентов нейросетевой модели. Данная проблема подробно рассмотрена в работах [9, 40].

Критерий останова. Критерий останова вводится для автоматического завершения процедуры обучения. Обычно критерием останова является некоторое условие, выполнение которого показывает, что дальнейшая оптимизация параметров модели неэффективна. Существенной проблемой при выборе критерия является невозможность определения адекватности модели на стадии обучения. Решением проблемы является введение сразу нескольких ограничений.

Первый вариант – задание максимального числа итераций процедуры обучения. Данный критерий не имеет физического (математического) обоснования и определяет лишь максимально допустимые временные затраты на процедуру обучения. Тем не менее критерий может быть использован в качестве основного в системах реального времени. Следует отметить, что для достижения минимума с требуемой степенью точности при использовании метода Левенберга - Маркардта обычно достаточно провести 500 – 600 итераций для нейросетевой модели, содержащей 100 – 500 весовых коэффициентов.

Другой способ задания критерия останова – введение ограничений на градиент $G(\theta)$. В точке минимума $\theta = \theta^*$ градиент должен быть равен нулю. При использовании методов последовательного приближе-

ния достижение равенства $G(\theta^*) = 0$ невозможно. Тем не менее в качестве критерия останова может быть задано следующее условие:

$$\left\| G(\theta^{(i)}) \right\| \le \varepsilon, \qquad (2.148)$$

где ε – некоторая константа. Несмотря на математическую обоснованность критерия (2.148), практическое применение является затруднительным. Этот факт объясняется сложностью выбора значения ε.

Критерий останова может быть получен на основе оценки изменения весовых коэффициентов между двумя итерациями. В случае, если максимальное изменение весовых коэффициентов $\max\limits_k \left\{ \theta_k^{(i+1)} - \theta_k^{(i)} \right\}$ меньше некоторой величины, процедура обучения завершается.

В некоторых нейросетевых приложениях в качестве критерия останова может быть использовано достижение некоторого значения ε минимизируемой функции $V_N(\theta, Z^N)$. Однако, с точки зрения нейросетевой реализации процедуры идентификации, использование этого подхода затруднительно, так как априорное определение ε невозможно в силу отсутствия информации о дисперсии возмущений.

При оптимизации нейросетевых моделей достаточно часто возникает ситуация, когда аппроксимация критерия оптимизации вблизи точки минимума становится неадекватной [20]. При использовании алгоритма Левенберга – Маркардта возникает риск недопустимого уменьшения доверительной области, что приводит к проблемам вычислительного плана. Коэффициент Маркардта λ (2.78), находясь в обратно пропорциональной зависимости от размера доверительной

области, постоянно возрастает по мере приближения к точке минимума. Таким образом, можно установить некоторое максимальное значение λ как критерий останова процедуры обучения нейросетевой модели.

В разделе 2.4.2 обсуждались вопросы предварительного (без достижения минимума) останова процедуры обучения с целью достижения эффекта регуляризации. При использовании данного подхода использование упомянутых ранее критериев останова становится невозможным. Вместо них используются результаты оценки модели на тестовом множестве: процедура обучения завершается по достижении минимума ошибки обобщения.

Оценка эффективности алгоритмов обучения НС. В результате алгоритмизации и компьютерного моделирования получен ряд оценок методов оптимизации параметров нейросетевых моделей (табл. 2).

Таблица 2

Сравнительная оценка алгоритмов обучения НС

Алгоритм обучения	Оценка алгоритма по пятибалльной шкале		
	Скорость сходимости	Вычислительная робастность	Требования к оперативной памяти
Обратное распространение ошиб-	1	4	2

ки			
Обратное распространение ошибки (рекуррентный)	1	4	5
Метод Гаусса – Ньютона	3	3	3
Метод Левенберга – Маркардта	5	5	1