Письма в ЖЭТФ, том 41, вып. 8, стр. 328 – 331 *25 апреля 1985 г.*

ОБ ОДНОМ КЛАССЕ НЕЛИНЕЙНЫХ СИСТЕМ С ПАМЯТЬЮ

А.А.Веденов, Е.Б.Левченко

Показано, что из простого лагранжиана вытекают динамические уравнения, описывающие явления, которые могут быть интерпретированы как обучение, распознавание образов, бистабильность восприятия, забывание, выработку прототипа.

Рассмотрим динамическую систему, элементы которой ("нейроны") взаимодействуют между собой и с термостатом. Состояние i-го нейрона в системе ($i = 1 \dots N$) описывается переменной $\sigma_i(t)$ (t – время), изменяющейся в интервале $-\sigma_0 \leqslant \sigma_i(t) \leqslant +\sigma_0$; состояние с $\sigma_i = +\sigma_0$ отвечает "возбужденному", а с $\sigma_i = -\sigma_0$ – "невозбужденному" нейрону. Будем рассматривать величину

$$L = -E = \frac{1}{2}\lambda \sum_{ij} T_{ij}\sigma_i\sigma_j \qquad (1)$$

($\lambda > 0$ – параметр) как лагранжиан взаимодействия системы, являющийся функционалом независимых переменных σ и T. Учитывая взаимодействие нейронов с термостатом, приводящее к появлению "сил трения" $(\dot{\sigma}/\gamma_1, \dot{T}/\gamma_2)$, из (1) получим динамические уравнения

для σ и T

$$\frac{1}{\gamma_1}\dot{\sigma}_i - f_i(\sigma) = \frac{\partial L}{\partial \sigma_i} = \lambda \sum_j T_{ij}\sigma_j, \qquad (2)$$

$$\frac{1}{\gamma_2}\dot{T}_{ij} - F_{ij}(T) = \frac{\partial L}{\partial T_{ij}} = \frac{1}{2}\lambda\sigma_i\sigma_j. \qquad (3)$$

В эти уравнения добавлены слагаемые (f, F), препятствующие неограниченному возрастанию абсолютных величин переменных σ и T; в рамках лагранжевой схемы они могут быть включены в (1) в виде потенциалов, быстро возрастающих вблизи $|\sigma_i| = \sigma_0$, $|T_{ij}| = T_0$.

В работах [1-6] постулировались два независимых уравнения, одно из которых описывало линейную трансформацию "поля активности" нейронов σ под действием "оператора памяти" T, а второе описывало изменение самого оператора T, квадратичное по σ. В [6] рассматривался процесс релаксации начального значения σ к одному из углов (с координатами $\pm 1, \ldots, \pm 1$) N-мерного куба, а в [1] в задаче о такой релаксации с заданной матрицей T было введено понятие энергии системы, которая уменьшалась в процессе релаксации.

В настоящей работе мы используем лагранжев подход с целью получить единым образом как уравнение релаксации σ (2), так и уравнение изменения памяти (3), а также с целью указать аналогию с различными физическими системами.

Динамика системы, описываемой уравнениями (2), (3) существенно зависит от ее предыстории. Исследуем сначала эволюцию переменных σ и T в специальных условиях – в процессе, имеющем характер "обучения". "Обучение" заключается в том, что в (2) включается сильное внешнее поле, действующее в течение времени t_1, в результате чего вектор σ принимает стационарное значение φ^1, соответствующее "образу" с компонентами $\pm \sigma_0$. В результате "обучения" матрица T_{ij}, в соответствии с (3), изменяется на величину $\Delta T_{ij} = \gamma_2 \lambda t_1 \varphi_i^1 \varphi_j^1$ (предполагаем, что t_1 значительно больше времени релаксации во внешнем поле вектора $\sigma(t)$ к своему стационарному значению φ^1). Процедуру "обучения" можно повторить многократно, используя "образы" φ^s, $s = 1 \ldots s_0$. Считая для простоты, что до начала обучения $T_{ij} = 0$, после окончания процедуры получим

$$T_{ij} = \sum_s \mu^s \varphi_i^s \varphi_j^s, \qquad (4)$$

где коэффициенты μ^s зависят от длительности обучения.

Из (2), (3) (с учетом (4)) следует, что в отсутствие внешнего поля характерные времена релаксации переменных $\sigma(\tau_\sigma \sim (\gamma_1 \lambda N s_0)^{-1})$ и $T(\tau_T \sim (\gamma_2 \lambda s_0)^{-1}$ существенно различаются: $\tau_\sigma \ll \tau_T$. Динамику системы (2), (3) на временах порядка τ_σ можно интерпретировать как процесс "распознавания образов". Процесс "распознавания" состоит в том, что "исходный стимул" – вектор, выбранный в качестве начального условия релаксирует в соответствии с (2) к одному из векторов "памяти" φ^s.

Эволюция динамической системы (2), (3) была исследована методом дискретного стохастического моделирования: с помощью процедуры Метрополиса и др. [7] (модель системы в термостате) и с помощью алгоритма, предложенного Хопфилдом [1] [1)]. При этом, вместо непрерывных переменных σ_i вводятся дискретные, принимающие значения $\sigma_{i0} = \pm 1$, а динамика системы задается с помощью вероятности "переворота спина" либо в соответствии со знаком со знаком "молекулярного поля" $T\sigma$ [1], либо в соответствии с изменением энергии системы (1) при заданной температуре [7].

[1)] Численные расчеты проведены А.А.Ежовым, Л.А.Книжниковой и Ю.Г.Черновым, которым авторы выражают глубокую благодарность.

Покажем, что в дискретной модели состояние с $\sigma = \varphi^s$ реализует локальный минимум энергии системы (которая уменьшается в процессе эволюции σ) и устойчиво, если число "образов" s_0 невелико. Положим $\sigma^s = \varphi^s + \delta^s$ и рассмотрим вариацию энергии (1) ($\sigma_0 \equiv \equiv 1$, $\mu^r \equiv 1$) :

$$\frac{1}{\lambda}\delta E = -N\varphi^s\delta^s - \sum_{r \neq s}(\varphi^r\varphi^s)(\varphi^r\delta^s) - \frac{1}{2}\sum_r(\varphi^r\delta^s)^2, \qquad (5)$$

где $\varphi\psi = \Sigma\varphi_i\psi_i$. В случае, когда различные φ^s приближенно ортогональны друг другу, второе слагаемое в (5) мало по сравнению с первым, и для вариации δ^s, соответствующей изменению одного из символов ± 1 в образе φ^s на противоположный, первое слагаемое равно $2N$, а третье $-2s_0$. Таким образом, $\delta E > 0$ – образ устойчив – при $s_0 \lesssim N$. Эта оценка качественно согласуется с результатами анализа ошибок распознавания с помощью матрицы (4), проведенного в [1].

Отметим, что при дискретном моделировании устойчивыми состояниями системы (2), (3) могут являться векторы, не совпадающие с φ^s [1, 8]. Так, если в матрице T (4) записана группа образов, получающаяся при небольших случайных искажениях некоторого вектора φ_0 (но не содержащая φ_0), то, как показывает анализ, аналогичный (5), φ_0 может являться устойчивым состоянием системы, т. е. иметь смысл "прототипа" [2] (см. [9]). При моделировании с помощью алгоритма [1] устойчивыми могут оказаться также "ложные образы" φ^* (например, при $s_0 = 3$ φ^* устойчив, если величины $p_i = \varphi^*\varphi^i > 0$ удовлетворяют неравенствам треугольника: $p_1 \leqslant p_2 + p_3$, $p_2 \leqslant p_1 + p_3$, $p_3 \leqslant p_1 + p_2$), которые исчезают при повышении температуры в методе Метрополиса и др. [7].

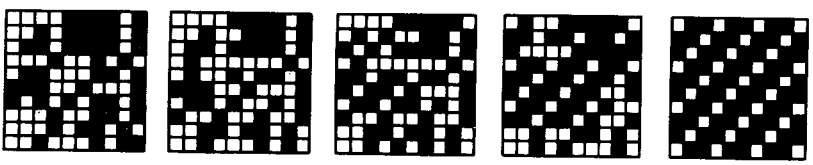

Рис. 1

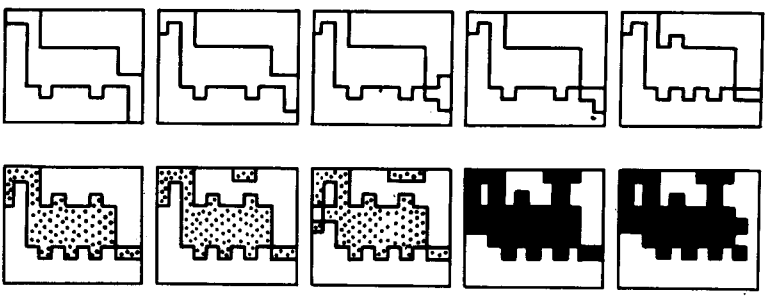

Рис. 2

Рис. 1. Последовательность состояний системы нейтронов, возникающая в процессе релаксации "исходного стимула" (левая картинка) в "образ" (правая картинка), полученная при дискретном моделировании уравнения (2) с помощью алгоритма [1]

Рис. 2. Последовательность "стимулов", предъявлявшихся в качестве начального условия при дискретном моделировании (2), (3). В отсутствие обучения ($\gamma_2 = \epsilon = 0$) смена устойчивого состояния происходит после предъявления 5-й картинки. При достаточно большой модификации матрицы T ($\epsilon = 0{,}025$) возникает гистерезис: смена образов при направлении от "ящера" к "паровозу" и наоборот происходит на 8-й и 3-й картинках соответственно

Пусть в T записано два ортогональных образа, хеммингово расстояние между которыми (число символов $+1$ или -1, отличающих один образ от другого) равно m, и пусть в качестве начального условия для системы (2), (3) выбираются элементы последовательности векторов (см. рис. 2), первый член которой совпадает с первым ("левым") образом, а пос-

[2] При дискретном моделировании с помощью алгоритма [1] возможность появления "прототипа" была показана А.А.Ежовым и Л.А.Книжниковой

ледний — со вторым ("правым") образом. Как показало моделирование на ЭВМ процесса распознавания с учетом обучения (после каждого акта распознавания стимула как искаженного образа φ^s к матрице T, в соответствии с (3), добавлялось слагаемое $\epsilon\varphi_i^s\varphi_j^s$) средние стимулы последовательности могут восприниматься либо как левый, либо как правый образ в зависимости от порядка, в котором предъявляются стимулы, т. е. возникает гистерезис. Отметим, что аналогичное явление — гистерезис зрительного восприятия — известно в психологии [10]. Оценку ширины области гистерезиса x получим, полагая, что на границе этой области энергия (1), соответствующая взаимодействию стимула с левым образом, увеличенная в $\left[1+\epsilon\left(\frac{m}{2}+\frac{x}{2}\right)\right]$ раз, равна исходной энергии, соответствующей взаимодействию стимула с правым образом:

$$\left[1+\epsilon\left(\frac{m}{2}+\frac{x}{2}\right)\right]\left[N-2\left(\frac{m}{2}+\frac{x}{2}\right)\right]^2=\left[N-2\left(\frac{m}{2}-\frac{x}{2}\right)\right]^2.$$

Отсюда при $x \ll N_1 \equiv N-m$ находим полуширину области гистерезиса $x=m/[(8/\epsilon N_1)-1]$. Область гистерезиса охватывает всю последовательность стимулов ($x=m$) при $\epsilon \geqslant 4/N_1$ и стягивается до одной фигуры (при уменьшении N_1 и постоянных ϵ, m) при $N_1=8/\epsilon m$.

Ограничение на величину T_{ij}, возникающее из нелинейного слагаемого в (3), приводит к тому, что вновь записываемые образы уменьшают относительный вес старых образов, а также частично искажают их (это "забывание" старых образов аналогично известному психологическому явлению [9]). Рассматривая увеличение x как расширение "воронки", в которую попадают стимулы, близкие к некоторому, например, левому образу, запишем приближенное уравнение для оценки ширины "воронки" левого образа при предъявлении случайных стимулов с частотой Ω: $\dot{x}=\epsilon\Omega(2x/m)$, где m — хеммингово расстояние до ближайшего образа. Таким образом, "воронка" экспоненциально расширяется со временем, что и приводит к вытеснению одного образа другим.

В заключение мы хотим отметить качественную аналогию между описанными явлениями и явлением модуляционной неустойчивости (коллапса) турбулентной плазмы [11] и с некоторыми явлениями в нелинейной оптике, которые описываются системой уравнений, подобной (2), (3).

Литература

1. *Hopfield J.J.* Proc. Nat. Acad. Sci USA, 1982, **79**, 2554; 1984, **81**, 3088; *Hopfield J.J., Feinstein D.J., Palmer R.G.* Nature 1983, **304**, 158.
2. *Ingber L.* Physica, 1982, **5D**, 83; Phys. Rev., 1983, **8 A**, 395; 1984, **9A**, 3346.
3. *Little W.A.* Math. Biosci, 1974, **19**, 101; *Little W.A., Show G.L.* ib, 1978, **39**, 281; *Levine D.S.* ib., 1983, **66**, 1.
4. *Peretto P.* Biol. Cybern., 1984, **50**, 51.
5. *Clark J.W., Winston J.V., Rafelski J.* Phys. Lett., 1984, **102 A**, 207.
6. *Anderson J.A., Silverstein J.W., Ritz S.A., Jones R.S.* Psych. Rev., 1977, **84**, 413; *Hinton G.E., Anderson J.A.* (Eds), Parallel Models of Associative memory, N.Y. 1981.
7. *Metropolis N. et al.* J. Chem. Phys., 1953, **21**, 1087; *Kirkpatrick S., Gellat C.D., Vecchi M.P.* Science, 1983, **220**, 671.
8. *Crick F.C., Mitchison G.* Nature, 1983, **304**, 111.
9. *Клацки Р.* Память человека: процессы и структуры, М.: Мир, 1978,
10. *Постон Т., Стюарт И.* Теория катастроф и ее приложения, М.: Мир, 1980.
11. *Goldman M.V.* Rev. Mod. Phys., 1984, **56**, №4.

Институт атомной энергии
им. И.В.Курчатова

Поступила в редакцию
27 февраля 1985 г.