Письма в ЖЭТФ, том 44, вып. 3, стр. 148 – 150 *10 августа 1986 г.*

АСИММЕТРИЯ ВЗАИМОДЕЙСТВИЙ И ИЕРАРХИЯ ОБРАЗОВ В МОДЕЛЯХ АССОЦИАТИВНОЙ ПАМЯТИ

Л.Б.Иоффе, М.В.Фейгельман

Предложены и исследованы модели ассоциативной памяти, допускающие несимметрию межнейронных связей, а также запоминание скоррелированных образов, образующих иерархические структуры наподобие иерархической структуры спинового стекла.

1. В последнее время в физической литературе появилось много работ, в которых изучаются методами статистической физики физические свойства несколько необычного объекта – нейтронной сети. Дело в том, что число элементов такой сети $\sim 10^{10}$, поэтому ее статистическое описание является, по-видимому, единственно возможным. Более того, свойства нейронных сетей во многом аналогичны свойствам спиновых стекол, что позволяет при их исследовании пользоваться методами теории спиновых стекол. В наиболее популярной модели Хопфилда [1,2] нейроны заменяются на элементы (σ_i), принимающие в основном два значения (± 1) ("мягкие" изинговские спины); запоминаемые образы в этой модели – это набор $\{\xi_i^{(p)}\}$ $(\xi_i^{(p)} = \pm 1 \quad 1 \leqslant i \leqslant N, \ 1 \leqslant p \leqslant k)$, индекс i нумерует различные нейроны, а индекс p – образы. Принимается, что уравнение движения переменных σ_i совпадает с чисто релаксационной ланжевеновской динамикой спинов с энергией $H_t\{\sigma_i\}$:

$$\frac{\partial \sigma_i}{\partial t} = -\frac{\partial H_t}{\partial \sigma_i} + f_i(t), \qquad \langle f_i(t) f_j(t') \rangle = 2T\delta_{ij}\delta(t-t'), \qquad H_t = H_0 + H$$

$$H_0 = \sum_i \lambda(\sigma_i^2 - 1)^2, \qquad H = -\frac{1}{2}\sum_{i,j} J_{ij}^{(0)} \sigma_i \sigma_j, \qquad \lambda \gg 1. \tag{1}$$

Матрица $J_{ij}^{(0)}$ должна быть выбрана таким образом, чтобы стационарные решения уравнения (1) совпадали с набором запоминаемых образов $\{\xi_i^{(p)}\}$. Если это условие выполнено, то решение $\sigma_i(t)$ с начальными условиями $\sigma_i(0)$, достаточно близкими к одному из образов $\xi_i^{(p)}$, релаксирует к этому образу, т. к. происходит процесс "вспоминания". В модели Хопфилда $J_{ij}^{(0)}$ выбирается в простейшем виде: $J_{ij}^{(0)} = \frac{1}{N}\sum_p \xi_i^{(p)} \xi_j^{(p)}$. Эта модель в настоящее время детально исследована методами статистической физики [3-5].

Модель Хопфилда обладает двумя серьезными недостатками: во-первых, она хорошо работает (т. е. число записываемых образов велико) лишь при записи нескоррелированных образов $\xi_i^{(p)}$ и, во-вторых, в ней заложена симметрия матрицы J_{ij} – предположение совершен-

но неестественное с точки зрения реальных нейронных сетей. В данной статье мы обсудим, как можно устранить эти ограничения (подробные вычисления будут опубликованы в другом месте). Начнем с несимметрии матрицы J_{ij}.

2. Исследуем простейшее обобщение уравнения (1) с несимметричной матрицей J_{ij}:

$$\frac{\partial \sigma_i}{\partial t} = -\frac{\partial H_0}{\partial \sigma_i} + \sum_i J_{ij}\sigma_j + f_i(t), \qquad \langle f_i(t) f_j(t') \rangle = 2T\delta_{ij}\delta(t-t'). \tag{2}$$

Предположим, что не все связи (i, j) могут осуществляться, т. е. выберем J_{ij} в виде $J_{ij} = (1 + \epsilon_{ij})J_{ij}^{(0)}$, где ϵ_{ij} принимает значения ± 1 с равными вероятностями и для всех связей (i, j), а $J_{ij}^{(0)} = \frac{1}{N}\sum_{p=1}^{k} \xi_i^{(p)} \xi_j^{(p)}$. Для исследования решения уравнения (2) используем метод динамического производящего функционала (см., например [6]), получим после усреднения по ϵ_{ij}:

$$\langle \sigma_i(t)\sigma_i(t') \rangle = \int \sigma_i(t)\sigma_i(t') \exp(iS\{\sigma_i(t), \psi_i(t)\}) D\sigma_i(t) D\psi_i(t),$$

$$S = \int dt \Big\{ \sum_i \psi_i(t)\Big(\dot{\sigma}_i + \frac{\partial H_0}{\partial t} + \sum_j J_{ij}^{(0)}\sigma_j\Big) + \sum_i T\psi_i^2(t) \Big\} + \tilde{S}, \tag{3}$$

$$\tilde{S} = \frac{\alpha q}{2}\sum_i \Big(\int \psi_i(t)dt\Big)^2 + \frac{\alpha}{2}\sum_i \iint dt dt' \psi_i(t)\psi_i(t') D(t-t').$$

Здесь $\alpha = k/N$ – относительное число образов, q – параметр Едвардса – Андерсона, $D(t)$ – коррелятор спинов: $\overline{\langle \sigma_i(0)\sigma_i(t) \rangle} = q + D(t)$ ($\lim D(t) = 0$), член $\tilde{S}$ в S появился в результате усреднения по ϵ_{ij}. Последнее слагаемое в $\tilde{S}$ отвечает наличию собственного шума $\eta_i(t)$ с коррелятором $\langle \eta_i(t)\eta_j(t') \rangle = \alpha\delta_{ij}D(t-t')$, что означает нарушение ФДТ [7]. Уравнения (3) являются фактически уравнениями самосогласования для величины $D(t)$. Мы ограничимся их исследованием в случае $\alpha \ll 1$. В этом случае для состояния системы, близкого к одному из образов $\xi_i^{(p)}$ (скажем $\xi_i^{(1)}$), удобно сделать замену переменных $\sigma_i \to \sigma_i\xi_i^{(1)}$, $\psi_i \to \psi_i\xi_i^{(1)}$, $J_{ij}^{(0)} \to J_{ij}^{(0)}\xi_i^{(1)}\xi_i^{(1)}$, не меняющую вида (3), после чего разбить $J_{ij}^{(0)}$ на две части: $J_{ij}^{(0)} = 1/N + \tilde{J}_{ij}$, $\tilde{J}_{ij} = \frac{1}{N}\sum_{p=1}^{k} \xi_i^{(1)}\xi_i^{(p)}\xi_j^{(1)}\xi_j^{(p)}$. Условие $\alpha \ll 1$ позволяет считать $\tilde{J}_{ij}$ случайными и независимыми величинами. Усредним поэтому уравнение (3) по $\tilde{J}_{ij}$ и будем действовать далее так же, как и при исследовании динамики спинового стекла [8]. В результате получим уравнения движения спина во внешнем постоянном поле h и переменном $\eta(t)$. При температуре $T \gg \exp(-1/4\alpha)$ можно перейти к глауберовским уравнениям на $\varphi = p_+ - p_-$, где $p_\pm$ – вероятность найти спин в состоянии ± 1 в данном поле $h + \eta(t)$:

$$\dot{\varphi} = \operatorname{th}[(h + \eta(t))/T] - \varphi. \tag{4}$$

В этой температурной области постоянные поля h имеют гауссово распределение со средним значением $\overline{h} = m$, $m \simeq 1$, $\overline{(h-\overline{h})^2} = \alpha r$ $r \simeq 2$. Выражая коррелятор $\langle \sigma(0)\sigma(t) \rangle$ через решения уравнения (4) и усредняя по шуму $\eta(t)$, получим выражение для $D(0)$, определяющее интенсивность нетеплового шума:

$$1 - D(0) = \int \frac{dh}{\sqrt{2\pi\alpha r}} \exp\Big(-\frac{(h-m)^2}{2\alpha r}\Big)\Big[\int \operatorname{th}\Big(\frac{h+\eta}{T}\Big)\exp\Big(-\frac{\eta^2}{2\alpha D(0)}\Big)\frac{d\eta}{\sqrt{2\pi\alpha D(0)}}\Big]^2 \tag{5}$$

откуда следует, что при $\exp(-1/4\alpha) \ll T \ll \alpha$ $D(0) \simeq \exp(-1/4\alpha)$, т. е. в этой температурной области несимметрия матрицы J_{ij} приводит лишь к добавочному, малому по сравнению с тепловым, шуму, оставляющему без изменения стационарные состояния системы, а, следовательно, не ухудшающему работу модели.

3. Рассмотрим (в рамках симметричной модели) запись коррелированных образов $\xi_i^{(p)} = \xi_i(1 - 2\beta_i^{(p)})$, где $\beta_i^{(p)}$ принимают значения 0 и 1 ($\mathrm{Prob}(\beta = 1) = c \ll 1$). При использовании обычного алгоритма записи [1] отдельные образы $\xi_i^{(p)}$ неразличимы, стационарное состояние системы отвечает "базовому" образу ξ_i. Чтобы добиться разрешения "тонкой структуры" образов, необходимо выбрать матрицу взаимодействия в виде

$$J_{ij}^{(0)} = \frac{1}{N}\xi_i\xi_j\left[1 + \frac{1}{\epsilon}\sum_{p=1}^{k}(\beta_i^{(p)} - c)(\beta_j^{(p)} - c)\right], \qquad (6)$$

где $\epsilon < 2c$. Мы исследовали свойства такой системы памяти при конечном α методами работы [4]. В отличие от [4] у нас имеется два типа параметров порядка:

$$m = \frac{1}{N}\sum_{i=1}^{N}\xi_i\langle\sigma_i\rangle; \qquad u^{(p)} = \frac{1}{N\epsilon}\sum_{i=1}^{N}(\beta_i^{(p)} - c)\langle\sigma_i\rangle\xi_i. \qquad (7)$$

Образу $\xi_i^{(p)}$ отвечает стационарное состояние с $m \neq 0$, $u^{(p)} \neq 0$; если $m \neq 0$, а $u^{(p)} = 0$ для всех p, то возникает базовый образ ξ_i. При $\epsilon = c$ оба типа состояний имеют одинаковую свободную энергию и реализуются при "температуре" $T < T_c(\alpha)$ где зависимость $T_c(\alpha)$ совпадает с найденной в [4], так же как максимальное относительное число воспроизводимых (при $T = 0$) образов $\alpha_c \simeq 0{,}14$. При $c < \epsilon < 2c$ имеется промежуточная область $\alpha_2 < \alpha < \alpha_1$, где устойчив только базовый образ ξ_i, причем $\alpha_2 < \alpha_c < \alpha_1$ при $\epsilon \to 2c$ $\alpha_2 \to 0$. При $\alpha < \alpha_2$ и $T < T_2(\alpha)$ разрешимы отдельные образы $\xi_i^{(p)}$. Интересно, что имеется промежуточная область "температур": $T_2(\alpha) < T < T_1(\alpha)$, где разрешим только базовый образ. Зависимости $\alpha_{1,2}(\epsilon)$ получаются численным решением уравнений среднего поля и будут опубликованы отдельно.

4. Алгоритм записи информации (6) допускает обобщение на случай памяти с иерархической структурой, имеющей k_1 базовых некоррелированных образов $\xi_i^{(\lambda)}$ с k_2 "сателлитами" $\xi_i^{(\lambda, p)}$ у каждого:

$$J_{ij}^{(0)} = \frac{1}{N}\sum_{\lambda=1}^{k_1}\xi_i^{(\lambda)}\xi_j^{(\lambda)}\left[1 + \frac{1}{\epsilon}\sum_{p=1}^{k_2}(\beta_i^{(\lambda, p)} - c)(\beta_j^{(\lambda, p)} - c)\right]. \qquad (8)$$

Максимальное полное число устойчиво воспроизводимых образов $k_1 k_2 \sim N$. Иерархическая организация памяти допускает быстрое распознавание базовых образов (классов) $\xi_i^{(\lambda)}$ при более высоком уровне шума T, с последующим разрешением отдельных образов $\xi_i^{(\lambda, p)}$ при уменьшении T. Возможно непосредственное обобщение (8) на случай иерархии с числом уровней $n > 2$, однако полное число образов $k = k_1 \cdot k_2 \ldots \cdot k_n$ остается порядка N. В работе [9] была предложена конструкция того же типа, что и (8) в частном случае $\epsilon = c$.

Литература

1. *Hopfield J.J.* Proc. Nat. Acad. Sci. USA, 1982, 79, 2554.
2. *Hopfield J.J.* Proc. Nat. Acad. Sci. USA, 1984, 81, 3088.
3. *Amit D.J., Gutfreund H., Sompolinsky H.* Phys. Rev., 1985, A32, 1007.
4. *Amit D.J., Gutfreund H., Sompolinsky H.* Phys. Rev. Lett., 1985, 55, 1530.
5. *Feigel'man M.V., Ioffe L.B.* Europhys. Lett., 1986, 1, 197.
6. *Sompolinsky H., Zippelius A.* Phys. Rev., 1982, B25, 6860.
7. *de Almeida J.R.L., Thouless D.J.* J. Phys., 1978, A11, 983.
8. *Sompolinsky H.* Phys. Rev. Lett., 1981, 47, 935.
9. *Parga N., Virasoro M.A.* Preprint Triest IC 1985-186.

Институт теоретической физики им. Л.Д.Ландау
Академии наук СССР

Поступила в редакцию
15 мая 1986г.