ISRN SICS-R--97/04--SE

**Some Results on Activation and Scaling
of Sparse Distributed Memory**

by

Jan Kristoferson

Some Results on Activation and Scaling of Sparse Distributed Memory

Jan Kristoferson

RWCP¹ Theoretical Foundation SICS² Laboratory
Box 1263, SE-164 29 Kista, Sweden
Email: janke@sics.se

October 1997

Abstract

It has been suggested that in certain situations it would make sense to use different activation probabilities for writing and reading in SDM (Sparse Distributed Memory). However, here we model such a situation and find that, at least approximately, it is optimal to use the same probabilities for writing and reading.

We also investigate the scaling up of SDM, in connection with some observations made by Sjödin, see [4]. It is shown that the original SDM (here in Jaekel's version) does not scale up if the reading address is disturbed, but that this can be remedied by using a kind of sparse SDM.

Keywords: Sparse distributed memory, SDM.

Contents

1	Introduction	1
2	Different activation probabilities for writing and reading?	1
3	Scaling up the memory	5

¹Real World Computing Partnership

²Swedish Institute of Computer Science

1 Introduction

We consider a Sparse Distributed Memory of Jaekel's selected-coordinates design, see Jaekel [1]. Let us give a short explanation of the concept:

An *address* is a binary string (vector) of 1s and 0s. A *datum* is a binary string (vector) of 1s and -1 s. Let N be the length (dimension) of the addresses and U the length (dimension) of the data. A *hard location* is a place in the memory where to store data. The *content* of a hard location is a U -dimensional vector of integers. The coordinates of the content vectors are called *positions*. In conventional neural-net terms, hard locations are hidden units, their input weights define the position, and their output weights define the contents. A hard location is given by a *mask* (a subset of K coordinates out of the N coordinates in the address strings) and one bit for each coordinate in the mask, so there are totally $\binom{N}{K}2^K$ possible hard locations. An address X *activates* a hard location h if all bits in X lying in h 's mask match the corresponding bits in h . *Storing a datum at the address X* means adding (as a vector) the datum to the contents of all hard locations activated by X . *Reading at the address X* means calculating by some method, from the contents of the hard locations activated by X , a datum to be read at X . Here we just consider the standard reading method: we sum the contents and choose, for each position, the reading 1 if the sum is ≥ 0 , and -1 if the sum is < 0 . Let M be the number of hard locations, and $p = 2^{-K}$ the *probability of activation*³.

2 Different activation probabilities for writing and reading?

There has been some discussion whether something could be gained by using different activation probabilities p and r for writing and reading in the memory, respectively. In particular, it has been suggested in [2] that it could be reasonable to take $r > p$ in the situation when many noisy copies of each datum are written at noisy addresses.

Let us fix a model for this situation. Let us say that a binary vector A is ϵ -*disturbed* from another binary vector B if, for every bit in A independently, the probability of this bit being different from the corresponding bit in B is equal to ϵ . To begin with, assume $r \geq p$ and look at the following stochastic model:

- a) Choose randomly (and independently) the "ideal" addresses X_t , $0 \leq t \leq T$.
- b) Choose Y and $X_{t,k}$, $0 \leq t \leq T, 1 \leq k \leq n$ independently, with Y ϵ -disturbed from X_0 , and $X_{t,k}$ η -disturbed from X_t .

³The probability that a randomly chosen hard location is activated by a randomly chosen address.

c) Choose the M hard locations randomly, where each hard location is given by a mask with K bits. The writing activation is as usual with probability $p = 2^{-K}$. For the reading activation a submask of the given (writing) mask with $L \leq K$ bits is randomly chosen for each hard location, giving the activation probability $r = 2^{-L} \geq p$.

As usual, we consider the signal-to-noise ratio ρ (see below) for a single content position. For this position,

d) Choose randomly the “ideal” datum bits B_t , $1 \leq t \leq T$. It is assumed that the B_t s are independent and also independent of all activation figures, i.e., stochastic variables ($L_{t,k,h}$ below) determined by which hard locations are activated by which addresses. The bit B_0 to be retrieved from the memory is assumed to be 1.

e) Choose $B_{t,k}$, $0 \leq t \leq T, 1 \leq k \leq n$ randomly and independently, with $B_{t,k}$ β -disturbed from B_t . The $B_{t,k}$ s will then also be independent of all activation figures. $B_{t,k}$ is written at $X_{t,k}$ ⁴, $0 \leq t \leq T, 1 \leq k \leq n$.

Consider the stochastic variables $L_{t,k,h}$ defined thus:

$L_{t,k,h} = 1$, if hard location no. h is activated by both Y and $X_{t,k}$ ⁵, and $= 0$ otherwise.

The sum read at the considered position is

$$Z = \sum_{t=0}^T \sum_{k=1}^n \sum_{h=1}^M L_{t,k,h} B_{t,k} \quad (1)$$

Granted that Z is approximately normally distributed (this assumption will be discussed below), the probability of getting the wrong reading at the given position is $\approx \Phi(-\rho)$, where Φ is the standard normal distribution function, and $\rho = \frac{E(Z)}{\sigma(Z)}$ the signal-to-noise ratio. So we want to calculate

$$\rho^2 = \frac{E^2(Z)}{\text{Var}(Z)} = \frac{E^2(Z)}{E(Z^2) - E^2(Z)} \quad (2)$$

We need the following expected values (see proofs below):

$$E(B_{0,k}) = 1 - 2\beta \quad (3)$$

$$E(B_{t,k}) = 0, \quad t \geq 1 \quad (4)$$

⁴with activation probability p

⁵with activation probabilities r and p , respectively

$$E(B_{t,k}^2) = 1 \quad (5)$$

$$E(B_{t,k}B_{t,j}) = (1 - 2\beta)^2, \quad k \neq j \quad (6)$$

$$E(B_{t,k}B_{s,j}) = 0, \quad t \neq s \quad (7)$$

$$E(L_{0,k,h}) = pr^\gamma \quad (8)$$

where $\gamma = -\log_2[(1 - \epsilon)(1 - \eta) + \epsilon\eta]$.

$$E(L_{t,k,h}) = pr, \quad t \geq 1 \quad (9)$$

$$E(L_{t,k,h}^2) = E(L_{t,k,h}) \quad (10)$$

$$E(L_{0,k,h}L_{0,j,h}) = p^{1+\lambda}r^{\kappa-\lambda}, \quad k \neq j \quad (11)$$

where $\lambda = -\log_2[(1 - \eta)^2 + \eta^2]$ and $\kappa = -\log_2[(1 - \epsilon)(1 - \eta)^2 + \epsilon\eta^2]$

$$E(L_{t,k,h}L_{t,j,h}) = p^{1+\lambda}r, \quad t \geq 1, \quad k \neq j \quad (12)$$

$$E(L_{t,k,h}L_{t,j,g}) \approx E(L_{t,k,h})E(L_{t,j,g}), \quad h \neq g \quad (13)$$

(3) holds since $B_{0,k} = B_0 = 1$ with probability $1 - \beta$, and $B_{0,k} = -1$ with probability β . (4) is also easily derived. (5) and (10) come from $B_{t,k}^2 = 1$ and $L_{t,k,h}^2 = L_{t,k,h}$. To get (6), observe that $E(B_{t,k}B_{t,j}) = \Pr(B_{t,k} = B_{t,j}) \cdot 1 + \Pr(B_{t,k} \neq B_{t,j}) \cdot (-1) =$

$(1 - \beta)^2 + \beta^2 + 2(1 - \beta)\beta \cdot (-1) = (1 - 2\beta)^2$. (7) comes from the independence of $B_{t,k}$ and $B_{s,j}$.

We have $E(L_{t,k,h}) = \Pr(L_{t,k,h} = 1) = \Pr(h \text{ is activated by both } Y \text{ and } X_{t,k})$. If $t = 0$, this equals $[\frac{1}{2}(1 - \epsilon)(1 - \eta) + \frac{1}{2}\epsilon\eta]^L \cdot [\frac{1}{2}]^{K-L} = [\frac{1}{2}]^K \cdot r^\gamma = pr^\gamma$, and we get (8). If $t \geq 1$ it equals $[\frac{1}{2} \cdot \frac{1}{2}]^L \cdot [\frac{1}{2}]^{K-L} = pr$, and we get (9).

For $k \neq j$ we have $E(L_{t,k,h}L_{t,j,h}) = \Pr(h \text{ is activated by } Y, X_{t,k} \text{ and } X_{t,j})$. If $t = 0$, this equals $[\frac{1}{2}(1 - \epsilon)(1 - \eta)^2 + \frac{1}{2}\epsilon\eta^2]^L \cdot [\frac{1}{2}(1 - \eta)^2 + \frac{1}{2}\eta^2]^{K-L} = [\frac{1}{2}]^K \cdot r^\kappa \cdot p^\lambda \cdot r^{-\lambda} = p^{1+\lambda} \cdot r^{\kappa-\lambda}$, and we get (11). If $t \geq 1$, it equals $[\frac{1}{2} \cdot \frac{1}{2} \cdot ((1 - \eta)^2 + \eta^2)]^L \cdot [\frac{1}{2} \cdot ((1 - \eta)^2 + \eta^2)]^{K-L} = p^{1+\lambda}r$, and we get (12).

With the model given by a), b) and c) above, we face the problem that the $L_{t,k,h}$ s are not even pairwise independent for different h s. To remedy this we replace a) and b) by the simplifying approximation that Y , X_t and $X_{t,k}$ are all fixed, but with mutual distances being such that the probability for any two of these addresses to have the same bit in a randomly chosen coordinate is approximately the same as in the original model. Observe that by a) and b) Y will be $\frac{1}{2}$ -disturbed from X_t , $t \geq 1$, since $\epsilon\frac{1}{2} + (1 - \epsilon)\frac{1}{2} = \frac{1}{2}$, and in fact from all $X_{t,k}$. Thus, let us assume that $d(Y, X_t) \approx d(Y, X_{t,k}) \approx \frac{N}{2}$ for $1 \leq t \leq T$, $1 \leq k \leq n$ (N is the address length). Now we get (13). By the law of large numbers, this approximation will be better the bigger N (and the memory) is.

Using (3) – (13) we find

$$\rho^2 \approx f(p, r) = \frac{\text{num}}{\text{denom}} \quad (14)$$

where $\text{num} = nMpr^\gamma(1 - 2\beta)^2$ and $\text{denom} = 1 + (n - 1)(1 - 2\beta)^2p^\lambda r^{\kappa-\lambda-\gamma} + [M - 1 + (1 - 2\beta)^2(1 - n - M)]pr^\gamma + Tr^{1-\gamma}[1 + (n - 1)(1 - 2\beta)^2p^\lambda + (M - 1 + (1 - 2\beta)^2(n - 1)(M - 1))pr]$.

Now fix any value of r between 0 and 1. Then $f(p, r)$ is a function of p , where $0 < p \leq r$. If we divide num and denom by p , we get a constant numerator and a denominator of the form $ap^{-1} + bp^{\lambda-1} + c$, where a, b, c are constants, a and b positive. Since $(1 - \eta)^2 + \eta^2 \geq \frac{1}{2}$, we have $\lambda \leq 1$, and hence this denominator is a decreasing function of p .

Thus, assuming $0 < p \leq r$, we find that $p = r$ when $f(p, r)$ is maximal. Let us now consider the case $0 < r \leq p$. Then $K \leq L$, and the writing masks are submasks of the reading masks. The equations (1) – (13) still hold, with two modifications: $p^\gamma r$ replaces pr^γ in (8), and $p^\kappa r$ replaces $p^{1+\lambda}r^{\kappa-\lambda}$ in (11). From this we get (14) with $\text{num} = nMp^\gamma r(1 - 2\beta)^2$ and $\text{denom} = 1 + (n - 1)(1 - 2\beta)^2p^{\kappa-\gamma} + [M - 1 + (1 - 2\beta)^2(1 - n - M)]p^\gamma r + Tp^{1-\gamma}[1 + (n - 1)(1 - 2\beta)^2p^\lambda + (M - 1 + (1 - 2\beta)^2(n - 1)(M - 1))pr]$.

Now fix any value of p between 0 and 1. Then $f(p, r)$ is a function of r , where $0 < r \leq p$. If we divide num and denom by r , we get a constant numerator and a denominator of the form $ar^{-1} + b$, where a is a positive constant and b a constant.

Thus, assuming $0 < r \leq p$, we find again that $p = r$ when $f(p, r)$ is maximal. Summing up, approximately nothing could be gained by allowing different probabilities of activation for writing and reading, and in the rest of this article we will just use one probability of activation $p = 2^{-K}$. We remark that we have considered only Jaeckel’s design of SDM and not Kanerva’s original design, for instance. However, the calculations are much more involved for the latter and, given other experience of the different designs, it would be surprising if the result were not the same in all cases. Besides, Jaeckel’s design, or the sparse version of it discussed below, is what is being used for further work, since it is easier to deal with in many ways, and also seems to be the biologically most plausible design.

Note the special case ($\beta = 0$) when the data are not noisy, but each datum is stored at many noisy addresses. An example could be storing a lot of handwritten letters, and then trying to identify another handwritten letter (the shape of a handwritten letter defining the address, and the name of the letter the datum).

We conclude this section by discussing the assumption that Z is approximately normally distributed. Writing $Z = \sum_t A_t$, where $A_t = \sum_{k=1}^n \sum_{h=1}^M L_{t,k,h} B_{t,k}$, we cannot claim that the A_t s are independent: an observed relatively large absolute value of a certain A_t indicates that more than expected hard locations have activated Y , giving an increased probability for large absolute values also for the other A_t s. However, the larger the memory, the more the distribution of the number of hard locations activating Y is concentrated around the expected value Mr , and so the normal approximation should work for large enough memories. There are also arguments using the independence relations concerning the $L_{t,k,h}$ s and $B_{t,k}$ s, but we don’t go into them here. This kind of somewhat sloppy reasoning, also used in the derivation of (13) above, has been standard in the literature on SDM. It has been justified by simulations showing good agreement with the theory for interesting values of the memory parameters. We plan to test the results of this article by simulations, too.

3 Scaling up the memory

We want to investigate what happens to the performance of the memory, i.e., the signal-to-noise ratio ρ , when the memory gets bigger and bigger. For instance, what happens to ρ if we keep the “load” $\tau = \frac{T}{M}$ constant and let $M \rightarrow \infty$? We will consider only the case $\eta = \beta = 0$. Putting $r = p$, $\lambda = \beta = 0$, $\kappa = \gamma = -\log_2(1 - \epsilon)$, $T = \tau M$ in equation (14) and simplifying, we get the following approximate equation⁶ which was already derived in [3] (assuming $n = 1$):

⁶From now on we will approximate $M - 1$ by M without further mention.

$$\rho^2 = \frac{Mp^{1+\gamma}}{1 - p^{1+\gamma} + \tau Mp^{1-\gamma} + \tau M^2 p^{3-\gamma}} \quad (15)$$

where $\gamma = -\log_2(1 - \epsilon)$.

The performance is studied in [3] for different values of M and τ , always using the optimal value of p , i.e., maximizing ρ . It is shown that the optimal p satisfies the following equation:

$$p^{3-\gamma} - \frac{\gamma}{(1-\gamma)M} p^{1-\gamma} - \frac{1+\gamma}{2(1-\gamma)\tau M^2} = 0 \quad (16)$$

For $\epsilon = \gamma = 0$ we get the wellknown formula

$$p = \sqrt[3]{\frac{1}{2\tau M^2}} \quad (17)$$

For $\epsilon > 0$, however, only numerical methods are used in [3], and no (approximate) explicit formulae for the functions $p(M, \tau)$ and $\rho(M, \tau)$ are given. Now we want to study the asymptotic behaviour of p and ρ when $M \rightarrow \infty$ (with τ constant, the problem of scaling up, or with τ varying in some sensible way)⁷. Then such formulae will be of interest.

To finish the case $\epsilon = 0$, we get $\rho^2 \approx \frac{1}{\tau(1+3\sqrt[3]{\frac{1}{4\tau^2 M}})}$ from (15) and (17). Now $4\tau^2 M \gg 1$ for “normal” values of M and τ , so that

$$\rho \approx \frac{1}{\sqrt{\tau}} \quad \text{if } \epsilon = 0 \quad (18)$$

Furthermore, it is clear that, for fixed τ ,

$$\lim_{M \rightarrow \infty} \rho = \frac{1}{\sqrt{\tau}} \quad \text{if } \epsilon = 0 \quad (19)$$

So the memory scales up when $\epsilon = 0$.

This is, however, not the case when $\epsilon > 0$, as first observed by Sjödin (cf. [4]). To prove it, consider a given $\epsilon > 0$, M and τ . Observe that

$$\rho_\epsilon = (1 - \epsilon)^{K_\epsilon} \rho_0^* = p_\epsilon^\gamma \rho_0^* \leq p_\epsilon^\gamma \rho_0 \quad (20)$$

⁷Observe that M shall be $\ll \binom{N}{K} 2^K$, the number of possible hard locations. This implies that N has to grow approximately like $\log M$ when $M \rightarrow \infty$.

where

p_ϵ and ρ_ϵ are the optimal values of p and ρ for the given $\epsilon > 0$, M and τ ,

K_ϵ is the corresponding mask-length, i.e., $p_\epsilon = 2^{-K_\epsilon}$,

ρ_0^* is the value of ρ we get if we change ϵ to 0 without changing $p = p_\epsilon$,

and ρ_0 is the optimal value for $\epsilon = 0$, M and τ .

Now fix τ and let $M \rightarrow \infty$. Then by equation (16) $p_\epsilon \rightarrow 0$, and by (19) $\rho_0 \rightarrow \frac{1}{\sqrt{\tau}}$. Thus for fixed τ

$$\lim_{M \rightarrow \infty} \rho = 0 \quad \text{if } \epsilon > 0 \quad (21)$$

and the memory does not scale up.

Let us examine more in detail how p_ϵ and ρ_ϵ depend on M and τ , if $\epsilon > 0$. Substituting $p_\epsilon = \frac{k}{\sqrt{M}}$ in (16) and simplifying, we get

$$\phi(k) = k^{1-\gamma} \left(k^2 - \frac{\gamma}{1-\gamma} \right) - \frac{1+\gamma}{2(1-\gamma)\tau} M^{-\frac{\gamma+1}{2}} = 0 \quad (22)$$

Observe that $\phi(\sqrt{\frac{\gamma}{1-\gamma}}) < 0$. Furthermore, $\phi(\sqrt{\frac{1}{1-\gamma}}) = \left(\frac{1}{1-\gamma}\right)^{\frac{1-\gamma}{2}} - \frac{1+\gamma}{2(1-\gamma)\tau} M^{-\frac{\gamma+1}{2}} > \frac{1}{(1-\gamma)^{\frac{1-\gamma}{2}}} \left(1 - \frac{M^{-\frac{\gamma+1}{2}}}{\tau(1-\gamma)^{\frac{1+\gamma}{2}}}\right)$, which is > 0 if $\tau^2 M > \frac{\tau^{\frac{2\gamma}{1+\gamma}}}{1-\gamma}$.

For “normal” values of M and τ , $\tau^2 M$ is large enough for this condition to be satisfied, and then we have $\phi(k) = 0$ for some $k = k_\epsilon = k_\epsilon(M, \tau)$ between $\sqrt{\frac{\gamma}{1-\gamma}}$ and $\sqrt{\frac{1}{1-\gamma}}$, i.e.,

$$p_\epsilon = \frac{k_\epsilon}{\sqrt{M}} \quad (23)$$

where $0 < \sqrt{\frac{\gamma}{1-\gamma}} < k_\epsilon < \sqrt{\frac{1}{1-\gamma}}$.

Substituting (23) in (15) we get, for $\tau^2 M$ large enough, $\rho_\epsilon^2 \approx \frac{k_\epsilon^{2\gamma}}{(1+k_\epsilon^2)\tau M^\gamma}$. We now observe that (22) implies that $k_\epsilon \rightarrow \sqrt{\frac{\gamma}{1-\gamma}}$ when $\tau^2 M \rightarrow \infty$, and thus for $\tau^2 M$ large enough,

$$\rho_\epsilon \approx \frac{c_\epsilon}{\sqrt{\tau M^\gamma}} \quad (24)$$

where $c_\epsilon = \sqrt{\gamma^\gamma(1-\gamma)^{1-\gamma}}$ lies between $\frac{1}{\sqrt{2}}$ and 1.

This gives another proof of (21), and we also find the following generalization of (19):

if $\tau = \tau_0 M^{-\gamma}$, where τ_0 is a constant, then

$$\lim_{M \rightarrow \infty} \rho_\epsilon = \frac{c_\epsilon}{\sqrt{\tau_0}} \quad (25)$$

Thus we keep the performance approximately constant when $M \rightarrow \infty$, if we take $T = \tau M$ proportional to $M^{1-\gamma}$. Values of γ around 0.5 or even higher could occur in applications of SDM. Hence we would like to do better, i.e., to have scaling up also for $\epsilon > 0$.

Dropping the subscript ϵ , we get $\rho \leq (1-\epsilon)^K \rho_0$ from (20). Recall that $p = 2^{-K}$. Since $p \rightarrow 0$ by (16), we have $K \rightarrow \infty$, and thus $\rho \rightarrow 0$, when $M \rightarrow \infty$, τ constant. We would like to let K be (a small) constant, but still have optimal values of p . This can be achieved by using something like Jaekel's hyperplane design (see [2]) with sparse addresses, as suggested by Sjödin in [4]. Thus, let the hard locations be defined by randomly chosen masks with a fixed number K of selected coordinates, and let the addresses be chosen randomly, only that for each coordinate the probability of 1 be not $\frac{1}{2}$ as before, but some number α near 0. A hard location is activated by an address if the address has 1s in all selected coordinates. Then $p = \alpha^K$, and we assume that $\alpha = \alpha(M, \tau, \epsilon)$ is chosen so that p be optimized. Here ϵ is the probability that a bit being equal to one in the correct address turns into zero in the reading address (the corresponding probability for a zero turning into one probably being much less). Let us see how this works.

We find (cf. [3]) that the stochastic variable L_t = the number of hard locations activated by both Y and X_t , has the distribution $\text{Bin}(M, p^2)$ for $t \geq 1$, and the distribution $\text{Bin}(M, \delta p)$ for $t = 0$, where δ is the constant $(1-\epsilon)^K$. This implies (cf. equation (15))

$$\rho^2 = \frac{M\delta p}{1 - \delta p + \frac{\tau M p}{\delta} + \frac{\tau M^2 p^3}{\delta}} \quad (26)$$

By differentiation it is found that the optimal value of p is (cf equation (17))

$$p = \sqrt[3]{\frac{\delta}{2\tau M^2}} \quad (27)$$

(26) and (27) imply $\rho^2 \approx \frac{\delta^2}{\tau(1+3\sqrt[3]{\frac{\delta^2}{4\tau^2 M}})}$. Assuming $4\tau^2 M \gg 1$ we get

$$\rho \approx \frac{\delta}{\sqrt{\tau}} \quad (28)$$

and, for fixed τ ,

$$\lim_{M \rightarrow \infty} \rho = \frac{\delta}{\sqrt{\tau}} \quad (29)$$

So the memory scales up.⁸

We remark that the formulae (28) and (29) are not very sensitive to the value of p . They hold even if we use, e.g., $p = \sqrt[3]{\frac{1}{2\tau M^2}}$ as in [4].

References

- [1] L. A. Jaeckel. An Alternative Design for a Sparse Distributed Memory. RIACS Technical Report 89.28, Research Institute for Advanced Computer Science, NASA Ames Research Center, 1989.
- [2] P. Kanerva. Sparse Distributed Memory and Related Models. In Mohamad H. Hassoun, editor, *Associative Neural Memories*. Oxford University Press, 1993.
- [3] J. Kristoferson. Best Probability of Activation and Performance Comparisons for Several Designs of Sparse Distributed Memory. SICS Research Report R95:09, Swedish Institute of Computer Science, 1995.
- [4] G. Sjödin. The Sparchunk Code: A Method to Build Higher-Level Structures in a Sparsely Encoded SDM. 1997. Submitted to IJCNN 98.

⁸Since M must now be $\ll \binom{N}{K} \approx \frac{N^K}{K!}$ (the number of possible hard locations), N now has to grow like $M^{\frac{1}{K}}$. Cf footnote 7.