Cognitive Neuroscience II
Lecture 4

Prof. Dr. Andreas Wendemuth

Lehrstuhl Kognitive Systeme

Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik

Fakultät für Elektrotechnik und Informationstechnik
Otto-von-Guericke Universität Magdeburg

http://iesk.et.uni-magdeburg.de/ko/

Resumé of previous lecture 3

- A simple autoassociative network with 2 neurons was considered,
  partially connected = 3 of 4 weights
- Depending on the values of the weights, this network can have 1 or 5 fixed points where the dynamics comes to a halt. I.e. at these points, the firing rates **v** are constant.
- The fixed point **v**=(0,0) (no firing) is stable.
- There are other fixed points **v**≠(0,0) which are unstable, i.e. we have seen that small deviations from this point will not vanish in time, but add up to large amounts.
- The behaviour can be simulated in discrete steps. The role of Δt and τ must be discussed.

4. Hard delimiters

- We now look at feedforward networks
- We are interested in the *information processing abilities* of such networks, rather than in the individual neuron‘s performance
- Hence, we simplify the firing rate equation further, using *hard delimiters*. (or *threshold activation functions*)

Hard delimiters

- Instead of using tanh, one can use a similar but hard delimiter: sign.
- Note that $\tanh(\gamma x) \xrightarrow{\gamma\to\infty} sign(x)$ and

$$\tau \frac{d\mathbf{v}}{dt} = -\mathbf{v} + \mathbf{tanh}(\gamma^{-1}\mathbf{M}^*\gamma\mathbf{v}) \cong -\mathbf{v} + \mathbf{sign}(\gamma^{-1}\mathbf{M}^*\mathbf{v}) \quad \text{or}$$

$$\frac{\tau}{\Delta t}(\mathbf{v}^{t+1} - \mathbf{v}^{t}) = -\mathbf{v}^{t} + \mathbf{sign}(\gamma^{-1}\mathbf{M}^*\mathbf{v}^{t}) \quad \text{if} \quad Ord(\gamma^{-1}\mathbf{M}) = 1$$

„Hard“ feedforward networks

- Write $\gamma^{-1}\mathbf{M} = \mathbf{W}$ and choose Δt= τ
- Then for feedforward networks with input **u**, output **v:** $\mathbf{v}^{t+1} = \mathbf{sign(W * u}^{t}\mathbf{)}$
- Regard 1 time step as 1 forward processing:

  $$\mathbf{v} = \mathbf{sign(W * u)}$$

- Note that due to sign, (u,v) can only attain values ±1.

An information processing question

- Let different sets μ of inputs to a neuron be given as $\mathbf{u}^{\mu}$, and corresponding outputs $\mathbf{v}^{\mu}$.
- How many (p) different sets of these input-output-relations can be realized with the same set of weights **W**? I.e. is the neuron able to handle p „tasks“ correctly?
- If dim(**u**)=N, α=p/N is called the „information capacity“ or simply, *capacity*.
- This already supposes that p scales $p \propto N$.

Capacities

- In a <u>single</u> feedforward neuron, this asks to simultaneously satisfy the p equations

$$v^{\mu} = sign(\mathbf{w} * \mathbf{u}^{\mu})$$

- Since (u,v) =±1, this is equivalent to

$$1 = sign(\mathbf{w} * v^{\mu} \mathbf{u}^{\mu})$$

- Defining a vector $v^{\mu}\mathbf{u}^{\mu} = \mathbf{x}^{\mu}$ (x=±1) we have

$$1 = sign(\mathbf{w} * \mathbf{x}^{\mu}) \quad \text{or} \quad \mathbf{w} * \mathbf{x}^{\mu} > 0 \quad ; \quad \mu = 1 \ldots p$$

Refresh: Binomials

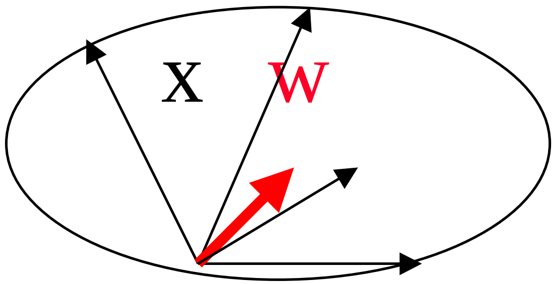
- The number of ways of arranging k items within N, without regarding the order, is

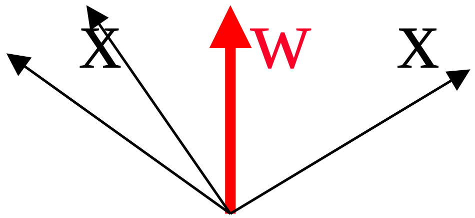
$$\binom{N}{k} = \frac{N!}{k!(N-k)!} = \frac{N(N-1)...(N-k+1)}{(N-k)(N-k-1)...1}$$

- Example: 3 students out of 5 can be selected in 10 ways.
- Further: $0! = 1$ $\binom{0}{0} = 1$

Geometric interpretation 1: Wrapping flower bouquets

- Not just for ladies ....





N-dimensional cones

- With N-dimensional **x**, **w** is the centre of an N-dimensional *cone* which wraps the flowers $\mathbf{x}^{\mu}$.
- The p conditions $\mathbf{w} * \mathbf{x}^{\mu} > 0$ are the wrapping conditions: as long as they are satisfied for one **w**, the bouquet is „wrappable“.
- The capacity is the amount of flowers wrappable per dimension.

Example: flower wrapping

- Since x=±1, we have in general 2^N flowers.
- In N=2 dimensions, we have 4 possible flowers.
- p=3 of these 4 are (at most) wrappable (almost).
- We now take all S= $\binom{2^N}{p}$ selections of p flowers.
- The *capacity* is reached if for half of these S many, the bouquet is wrappable (probabilistic definition).
- For N=2, p=3, all S= $\binom{4}{3}$= 4 selections are wrappable.
- For N=2, p=4, the S=1 selection is not wrappable
- Hence, for N=2, we have α=p/N=3/2= 1.5

Limitations

- All S= $\binom{2^N}{p}$ selections of p flowers have to be considered for the flower wrapping equations.
- Since $p = \alpha N$,

$$S = \binom{2^N}{\alpha N} = \frac{(2^N)(2^N - 1)\ldots(2^N - \alpha N)}{(\alpha N)!} \xrightarrow{\text{N large}} \frac{2^{(\alpha N^2)}}{(\alpha N)!} \propto 2^{(\alpha N^2)}$$

- This becomes prohibitively large even for moderate N. So a different approach is needed.

Ex4: Capacities in 3,N dimensions

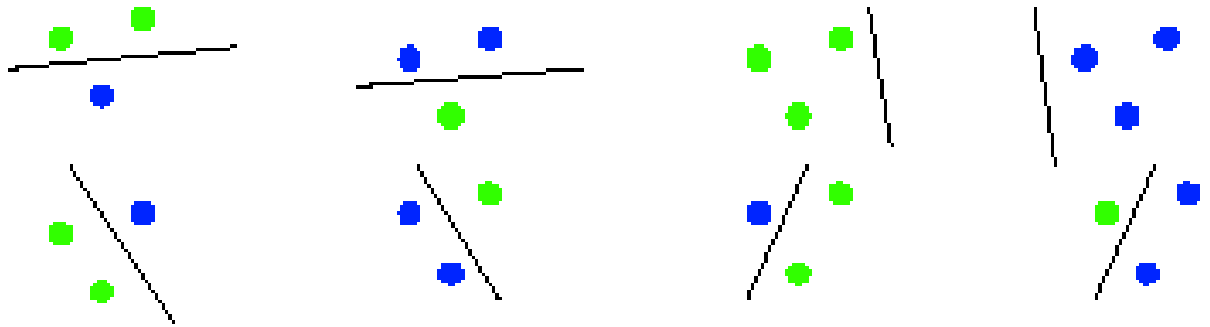
- What is the wrapping capacity for N=3?
- [Ladies only: for N=3, is that what you always/normally/sometimes/never get?
  If there is a difference: why?]
- Do you have an idea/educated guess what the capacity would be for large N?

Geometric Interpretation 2: Arranging Sets in Dichotomies

- Regard again the p eqns. $v^{\mu} = sign(\mathbf{w} * \mathbf{u}^{\mu})$
- For *any* given set of p inputs, there are 2^p ways of assigning the outputs. One such way is called a *dichotomy*.
- Of these, C dichotomies can be realized by the neuron by arranging **w**. These dichotomies are *linear separations* of the 2 classes of data. **w*x**=0 defines a *linear separating hyperplane*.
- The capacity is reached at $C = ½ \, 2^p$,
  i.e. half of all assignments can be processed by the neuron / can be linearly separated.

Dichotomies für p=3, N=2:

- $8 = 2^p$ Dichotomies with linear separating hyperplanes:



General position

- p vectors in N dimensions are said to be in „general position“ if no subset of $M \leq N$ vectors are linearly dependent.
- Example 1: in N=3 dimensions, a subset M=3 of p vectors are situated on 1 line. Then the p vectors are not in general position.
- Example 2: in N=9 dimensions, a subset M=7 of p vectors are situated in a 4d-subspace. Then the p vectors are not in general position.

$C(p,N) = 2^p$ ($\alpha \geq 1$) for $p \leq N$.

- If $p \leq N$, and we can always use the flower bouquet inequalities $\mathbf{w} * \mathbf{x}^{\mu} > 0$ and demand an even tighter condition, $\mathbf{w} * \mathbf{x}^{\mu} = 1$
- We then have $p(\leq N)$ linear equations in N variables **w** which, if the patterns are in general position, can always be solved.
- This works for all assignments of outputs, so $C(p,N) = 2^p$ for $p \leq N$, hence $\alpha \geq 1$.

Counting Theorem (Cover 1965)

- Compute the number of available dichotomies C(p,N) recursively (Thomas Cover, 1965):
- p patterns in N dimensions are assigned to C(p,N) many dichotomies. *Now add one more pattern*. Without changing anything in the C different weight vectors **w**, the new pattern will be assigned according to $v^{\mu} = sign(\mathbf{w} * \mathbf{u}^{\mu})$.
- This yields the same number of dichotomies $C(p+1,N)_1 = C(p,N)$

Counting (Ctd.)

- If one wants to have the other assignment for the new pattern, with all previous assignments fixed, it is necessary to modify **w**.
- For doing so, one needs one degree of freedom. This is given if the assignment of the previous p patterns could as well have been achieved in N-1 dimensions. So this works in
  $C(p+1,N)_2 = C(p,N-1)$ cases.

Counting (Ctd.)

- If the linear separating hyperplanes must include the origin, i.e. **w***x=0 , the set of patterns including the origin must be in general position. If not, i.e. **w***x**-**T=0 with a free threshold T, only the set of patterns must be in general position.
- Cover showed that this is necessary and sufficient.

Recursive Counting

- In summary, we get
  C(p+1,N) = C(p,N) + C(p,N-1)
- Since C(p=1, .) = 2, we can solve the recursion:

$$C(p, N) = 2 * \sum_{i=0}^{N-1} \binom{p-1}{i}$$

- Note that this is still valid if p ≤N, since

$$\binom{p-1}{p-1+i} = 0$$ for i > 0.

Ex 5: Counting Dichotomies (1)

- In the following, use linear separating hyperplanes which include the origin. Draw 4 points, where these 4 points and the origin are in general position.
  With Cover, this gives $C(4,2) = 2 * \sum_{i=0}^{1} \binom{3}{i} = 8$ dichotomies.
- Use C(p=1, .) = 2, and Cover‘s recursion formula C(p+1,N) = C(p,N) + C(p,N-1), to arrive at the same number 8.
- Draw these dichotomies.
- Draw 4 patterns, where exactly 2 patterns and the origin are on one line. Draw the linearly separable dichotomies which include the origin. What happens? Why?
- Draw 4 patterns such that only 2 linearly separable dichotomies are possible. Can this be derived from Cover‘s formula?

Ex 5: Counting Dichotomies (2)

- Allow for linear separating hyperplanes which do not include the origin. This is equivalent to adding a *threshold dimension*. With this extra dimension, drawing 4 points on a sheet of paper gives $C(4,2+1) = 2 * \sum_{i=0}^{2} \binom{3}{i} = 14$ dichotomies.
- Use C(p=1, .) = 2, and Cover's recursion formula C(p+1,N) = C(p,N) + C(p,N-1), to arrive at the same number 14.
- Draw these dichotomies, and the remaining 2 which are not linearly separable.
- Draw 4 patterns which are not in general position, and the linear separable (<14) dichotomies.

Ex 5: Counting Dichotomies (3)

- Use C(p=1, .) = 2, and Cover‘s recursion formula C(p+1,N) = C(p,N) + C(p,N-1).
- Draw a 2-dim. tabloid in (p,N) with p>N. Count the ways of arriving, with Cover‘s recursion, at some point (p,N), and by this show, both for p>N and for p<N , that

$$C(p, N) = 2 * \sum_{i=0}^{N-1} \binom{p-1}{i}$$

Refresh 2: Evaluating the binomials

- Binomials stem from

$$(a+b)^{p-1} = \sum_{i=0}^{p-1} \binom{p-1}{i} a^i b^{p-1-i}$$

- The sum of binomials is

$$\sum_{i=0}^{p-1} \binom{p-1}{i} = 2^{p-1}, \quad \sum_{i=0}^{p/2\ -1} \binom{p-1}{i} = 2^{p-2} \quad \text{(p even)}$$

Evaluating Cover‘s formula

- Let p even, then for N = p/2:

$$C(p,N) = 2 * \sum_{i=0}^{p/2-1} \binom{p-1}{i} = 2 * (2^{p-2}) = \frac{1}{2} 2^{p}$$

- At p=2N (α=2), the capacity condition is met.
- For p >2N, the capacity condition will be violated.

Intermediate Resumé: Capacity of hard neurons

- Our model of „hard“ neurons **v = sign(W *u)** with N inputs has an information capacity of α=2, i.e. it can handle p= αN = 2 N uncorrelated „tasks“ correctly (by assigning its synaptic efficiacies).
- If the tasks are highly correlated, the assignments will be grouped „close“ together, and α>2. This is the case for „ordinary“ flower bouquets. Or, in Cover‘s model, less than half of all assignments must be processed by the neuron, i.e. only $C < \frac{1}{2} 2^p$ is required.

Ex 6: visualize C(p,N)

- Write a Matlab program which computes for various p and for fixed N:

$$C(p, N)/2^{p} = 2^{1-p} * \sum_{i=0}^{N-1} \binom{p-1}{i}$$

- Draw the left hand side (LHS) as a function of α.
- Now use the same program to draw more lines (in the same figure), each with different N. What happens? Why?

Ex 7: Analysis of C(p,N)

- Let lb be the binary logarithm, i.e. basis=2.
- Use the program from the previous exercise and draw for various N and $\alpha > 2$

$$H_N(\alpha) = \frac{\mathrm{lb}\, C(p, N)}{p}$$

- Draw in the same figure for $\alpha > 2$

$$H(\alpha) = -\frac{1}{\alpha}\mathrm{lb}(\frac{1}{\alpha}) - (1-\frac{1}{\alpha})\mathrm{lb}(1-\frac{1}{\alpha})$$

- When is this a good approximation? Why? Is there an „information processing“ interpretation of H (α)?

Processing content?

- Note that high correlation means less information content: is it useful to be able to process more data (α>2) which contains fewer information (due to correlation)?
- This can only be computed in the flower bouquet model.
  [Get famous if you do it with Cover.]

Biased patterns

- In autoassociative networks, consider biased patterns with bias $m = \frac{1}{N}\sum_{i=1}^{N} u_i^\mu$ or $< u_i^\mu > = m$
- This invokes a correlation

$$< \mathbf{u}^\mu \mathbf{u}^\nu > = \frac{1}{N}\sum_{i=1}^{N} < u_i^\mu u_i^\nu > = \frac{1}{N}\sum_{i=1}^{N} < u_i^\mu > < u_i^\nu > = m^2$$

- Bias is one (not the only) way of invoking correlation

Capacity for biased / correlated patterns

- One can show with methods from Statistical Physics for the flower bouquet model with large N: [Gardner 1988]

α	m
2.0	0
2.0527	0.2
2.6675	0.6
6.0792	0.9
∞	± 1

Information Content

- The information content of p=αN patterns with N bits each is N^2 I, with I given (Shannon) by

$$I(\alpha, m) = -\alpha\left[\frac{1+m}{2}\,\mathrm{lb}\left(\frac{1+m}{2}\right) + \frac{1-m}{2}\,\mathrm{lb}\left(\frac{1-m}{2}\right)\right]$$

- Note I(α ,m=0) = α and I(α ,|m|=1)=0.
- The factor [] decreases with increasing |m|

Information content with bias

- Now take the α values for bias!

m	α	$I(\alpha,m)$
0	2	2
0.2	2.0527	1.9938
0.6	2.6675	1.9257
0.9	6.0792	1.7411
±1	∞	0

- Clearly, a network can handle almost the same information content with correlated (biased) patterns, as long as $|m|$ is noticeably < 1.

Benefits of correlation

- Hence one may spread a given information content into many correlated / biased patterns (*redundancy)* and have a neural network learn that.
- That may have advantages (robust information processing).
- We will show later that it doesn‘t take much *longer* to learn more but redundant data.

Resumé: Processing with hard neurons

- Hard neurons have *sign* as transfer function
- Capacity α is an important information processing feature.
- Capacity can be calculated by flower wrapping or dichotomy counting.
- Hard neurons can handle p= αN = 2 N uncorrelated „tasks“ correctly (by assigning their synaptic efficiacies), and many more correlated ones.
- The processable information content I is almost 2 even in correlated cases, which can be used for redundancy and hence, robustness.