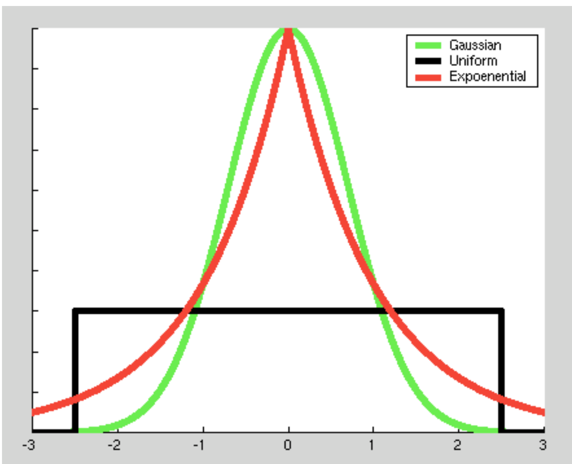
Notes on Independent Component Analysis

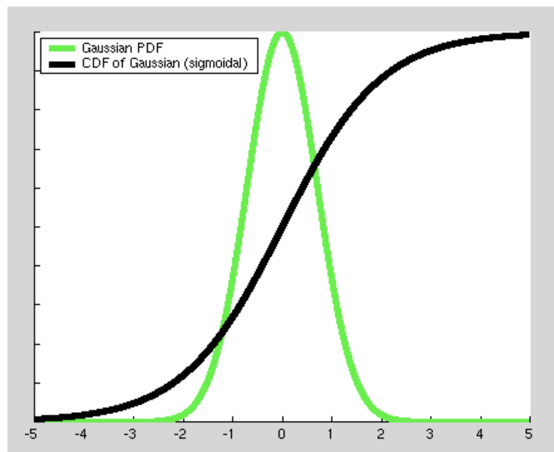
Jon Shlens
5 August 2002

II. Review: *pdf, cdf* and Entropy

a. Probability Density Functions (*pdf*) and Cumulative Density Functions (*cdf*)

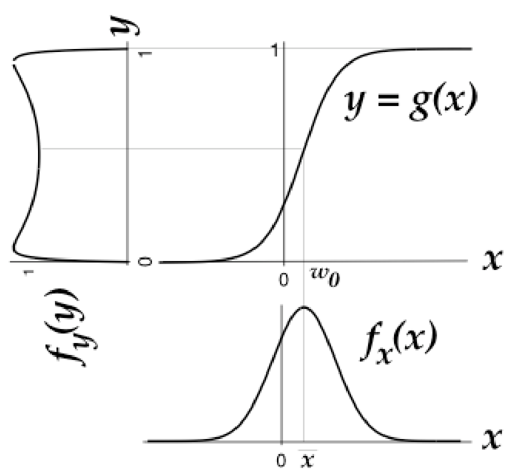
- Abandon knowledge of the temporal / presentation order in time series data
- 3 *pdf's* of interest: exponential, Gaussian, uniform
- *cdf* is the integral of the *pdf*





Note: Technically *pdf* of exponential distribution is only defined for x>0

b. Applying a function *g(x)* to a *pdf P(x)* produces a new *pdf* $P(y) = \dfrac{P(x)}{\partial g(x)/\partial x}$



[4]

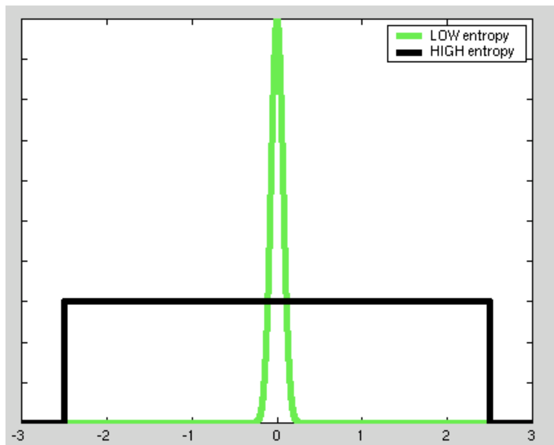
c. Entropy (*H*)

- A function of the *pdf*

$$H_{cont}(x) = \int p(x)\log(p(x))dx$$

$$H_{disc}(x) = \sum_{i} p(x_i)\log(p(x_i))$$

- Main idea: More certainty about a value → lower entropy (e.g. delta function)
  Less certainty about a value → higher entropy (e.g. uniform *pdf*)



II. The Goal of Independent Component Analysis

a. Preliminary: PCA

- The goal of PCA is to find a basis which maximizes the variance and along which the data has no covariance
- If *a* and *b* are row vectors and projections of the data along two principal component axes, then $ab^T = 0$
- Only concerned with second-order (variance) statistics

b. Goal of ICA

- Find a basis along which the data (when projected) is statistically independent
- Formally, if *(x,y)* are two "independent" components (bases), then

$$P[x,y] = P[x]P[y]$$

where *P[x], P[x]* are the distributions along *x* and *y*
P[x,y] is the joint distribution

- This is equivalent to saying: for a every data point, the knowledge of *x* in no way provides you with any information about *y*.
- In information theory, the mutual information between *P(x)* and *P(y)* is zero.

$$I(x,y) = 0 \quad \text{[short-hand]}$$

c. Why neuroscience?
- Several papers have conjectured that the goal of cortical processing is *redundancy reduction* [1,2]
- "the activation of each feature detector is supposed to be as statistically independent from the others as possible" [5]

III. Several Solutions to ICA

a. Expectation Maximization (EM) with Maximum Likelihood Estimation (MLE)
- Dayan and Abbott; difficult to understand.

b. Other methods
- Kurtosis maximization: http://www.cs.toronto.edu/~roweis/kica.html
- Projection pursuit: http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/

c. Information maximization
- The Bell and Sejnowski formulation

IV. Framework for ICA

a. Set-up (2 signal example)
- An *unknown* set of statistically independent signals: **S**
- An *unknown* mixing matrix: **A**

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} | & | \\ signal_1 & signal_2 \\ | & | \end{bmatrix}$$

- Assume the data we receive **X** is a mixture of the original signals

$$\mathbf{X} = \begin{bmatrix} | & | \\ mixed_1 & mixed_2 \\ | & | \end{bmatrix} = \mathbf{AS}$$

- Because **X** is a mixture of signals, the mixed components (e.g. $mixed_1$, $mixed_2$) are not statistically independent.

b. The ICA Question
- Can we recover **A** and **S** just from **X**?
- Mathematically, can we find a matrix **W** such that:

$$\mathbf{U} = \mathbf{WX}$$

$$\text{where } \mathbf{AW} = \mathbf{I} \text{ or equivalently, } \mathbf{U} = \mathbf{S}$$

- (Yes. By finding the basis vectors **A** that are statistically independent.)

c. The Discovery / Trick: Information Maximization

- <u>The Main Goal</u>: Maximize the joint entropy of $\mathbf{Y} = g(\mathbf{U})$ where g is a sigmoid function $g(x) = \frac{1}{1+e^{-x}}$
  1. Find a matrix **W** such that: $\max\{H(g(\mathbf{WX}))\}$
  2. The matrix **W** becomes the inverse of **A**. ($W = A^{-1}$)
  3. Side note: We are coincidentally maximizing the mutual information between **X** and **Y** if we assume our model does not magnify the entropy of the noise. [5]

- An intuitive algorithm to implement this:
  1. Define a surface $H(g(\mathbf{WX}))$
  2. Find the gradient $\frac{\partial}{\partial W} H(...)$ and <u>ascend</u> it!
  3. When the gradient is zero, done.

V. <u>Proof of Information Maximization: Why does this work?</u>

a. Notes
- This section is basically a rip off of a not-commonly-cited paper [3].
- The "Why" is not discussed (fully) in the original papers [4,5]
- Also, for my convenience I am being "fast and loose" with my notation. Namely, y_i could mean the "distribution of y_i" (technically denoted $P(y_i)$ or just the variable y_i.

b. <u>THE PROOF</u>

1. If the columns of **U** are statistically independent, applying an invertible transformation $\mathbf{Y} = g(\mathbf{U})$ can <u>not</u> make them dependent.

   $I(u_i, u_j) = 0 \Rightarrow I(y_i, y_j) = 0$ where u_i and y_j are the columns of **U** and **Y**

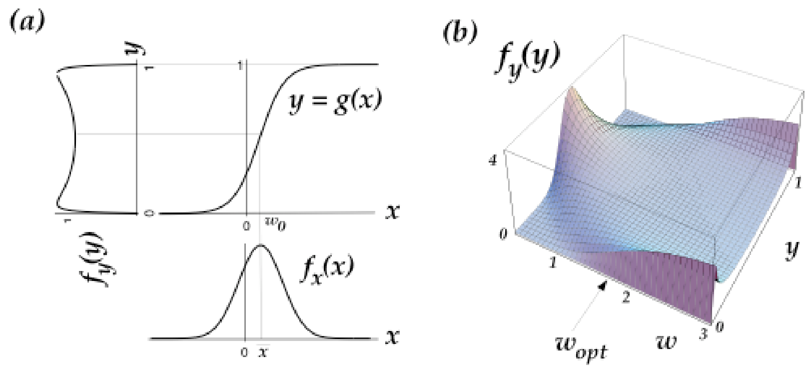
2. The individual entropies $H(y_i)$ are maximized when the distribution of y_i is the *cdf* of the distribution of u_i (the *pdf*).

   - $H(y_i)$ is maximum when $y_i = g(u_i)$ is the uniform distribution
   - By review section (b), we can equate:

$$P(y_i) = \frac{P(u_i)}{\partial y_i / \partial u_i}$$

$$1 = \frac{P(u_i)}{\partial y_i / \partial u_i}$$

$$y_i = \int P(u_i) du$$



[4]

Note: At w_{opt} the output distribution becomes uniform and the sigmoid aligns to become the *cdf* of the input distribution.

3. The joint entropy of two sigmoidally-transformed outputs $H(y_1, y_2)$ is maximal when y_1, y_2 are statistically independent and g is the *cdf* of u_1, u_2.
   - Remember the definition of joint entropy

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2)$$

   - The mutual information is minimized (equal to zero) when y_1, y_2 are statistically independent.
   - Via step 2, we know that that $H(y_i)$ is maximized when the sigmoid is the *cdf* of u_i

4. When one combines 2 non-Gaussian *pdf's*, the new *pdf* is more Gaussian.
   - This is the Central Limit Theorem.

5. <u>FINAL POINT</u>: The big one!

   - Reconsider the joint entropy for two variables: $H(y_1, y_2)$
   - This quantity is maximized when y_1, y_2 are statistically independent and when g is the *cdf* of u_i. By Step 1, this statement is equivalent to requiring u_1, u_2 be statistically independent. Hence, it must be that $u_i = signal_i$!
   - If there is any deviation from this causing a mixing of the signals, then:
     i. There exists a statistical dependency between the u_i. This increases $I(u_i, u_j)$, increasing $I(g(u_i), g(u_j))$ and thus decreasing the joint entropy.
     ii. There is a decrease in the individual entropies $H(y_i)$ as the individual distributions y_i deviate from a uniform distribution (via Step 4 and Step 2). This also decreases the joint entropy.
   - Therefore, maximizing the joint entropy is equivalent to $u_i = signal_i$.

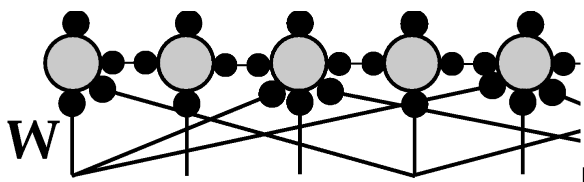
c. Problems with Information Maximization

1. The *cdf* must be able to "match" the *pdf* of the signal distributions ($signal_i$).
    - One must use a judicious choice of non-linearity
    - The sigmoid works well in practice for super-Gaussian (or positive kurtosis) distributions. Surprisingly, in practice the distribution of most real-world sensory input has a high kurtosis. [See figure 5 in 5]

2. There can <u>not</u> be more than one Gaussian source. There is no statistical information to "pull" these distributions apart because of the Central Limit Theorem.

VI. <u>The ICA Learning Rule</u>

a. Use a simple, single layer network set-up which implements $\mathbf{Y} = g(\mathbf{WX})$



[6]

b. Perform gradient ascent on the joint entropy.

c. Here is the learning rule for a single input and output $y = g(wx)$. (Following [5])

- The joint entropy (only one variable) is defined as:

$$
\begin{aligned}
H(y) &= -\int P(y)\ln P(y)\,dy \\
&= -\langle \ln P(y) \rangle \\
&= -\left\langle \ln \frac{P(x)}{\partial y/\partial x} \right\rangle \\
H(y) &= \left\langle \ln \left| \frac{\partial y}{\partial x} \right| \right\rangle - \langle \ln P(x) \rangle
\end{aligned}
$$

- We can now change our weights according to gradient ascent: $\Delta w \propto \dfrac{\partial H(y)}{\partial w}$
- All that is left is to evaluate the $\partial H(y)/\partial w$. Through a little calculus and using the sigmoid function, one finds:

$$\Delta w \propto \frac{1}{w} + x(1-2y)$$

d. This learning rule can be generalized to multiply outputs and sped up (with the natural gradient trick [7]), producing:

$$\Delta W \propto (\mathbf{I} + \hat{\mathbf{y}}\mathbf{u}^T)W \quad \text{where } \hat{y}_i = \frac{\partial}{\partial u_i}\ln\frac{\partial y_i}{\partial u_i} \text{ and } \mathbf{I} \text{ is the identity matrix}$$

e. General comments
   - A competition between an “anti-Hebbian” (first) term and a “Hebbian” (second) term.
   - The learning rule is global (not local) making it not biologically plausible in its current mathematical form.
   - … although many believe ICA is begin performed somewhere (e.g. primary visual cortex [5]) but using a different mathematical form.

VII. References

[1] Barlow, H (1989) “Unsupervised Learning” *Neural Computation 1*, 295-311.

[2] Atick JJ (1992) “Could information theory provide an ecological thery of sensory processing?” *Network 3*, 213-251.

[3] Bell A. and Sejnowski (1995) “Fast blind separation based on information theory.” *Proceedings of the International Symposium on Nonlinear Theory and Applications.* [Available at ftp://ftp.cnl.salk.edu/pub/tony/nolta3.ps.Z]

[4] Bell A.J. and Sejnowski T.J. (1995). An information maximization approach to blind separation and blind deconvolution, *Neural Computation*, 7, 6, 1129-1159.

[5] Bell A.J. and Sejnowski T.J. (1996). The `Independent Components' of natural scenes are edge filters, to appear in Vision Research

[6] Dayan, Abbott (1999) *Theoretical Neuroscience*: MIT Press.

[7] Amari et al (1996) “A new learning algorithm for blind signal separation.” *Advances in neural information processing systems (Vol 8)*. Cambridge, MA: MIT Press.