Optimal, Unsupervised Learning in Invariant Object Recognition
Wallis and Baddeley, Neural Comp. 9, 883-894, (1997)

Richard Turner (turner@gatsby.ucl.ac.uk)

Gatsby Computational Neuroscience Unit, 02/12/2005

Outline

- Objects persist in the environment longer than the sampling time of our visual sensors

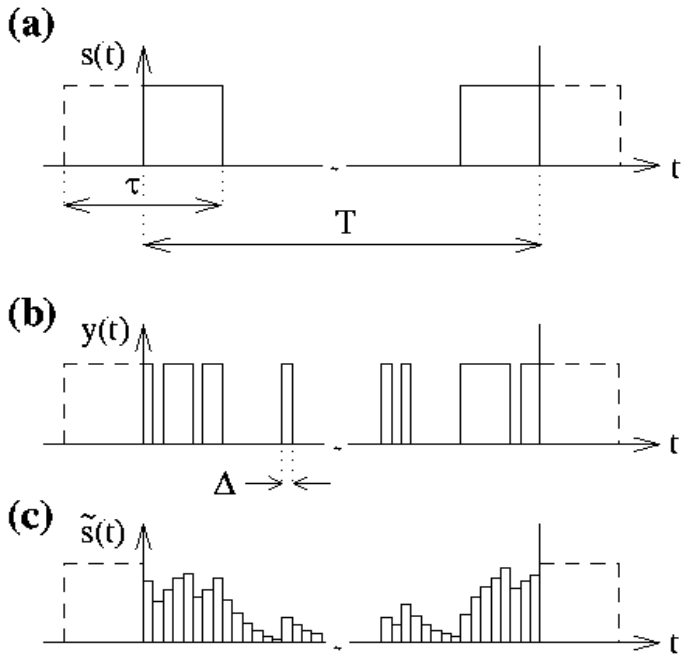
- This temporal slowness prior will help invariant object recognition (Hinton and Földiák)

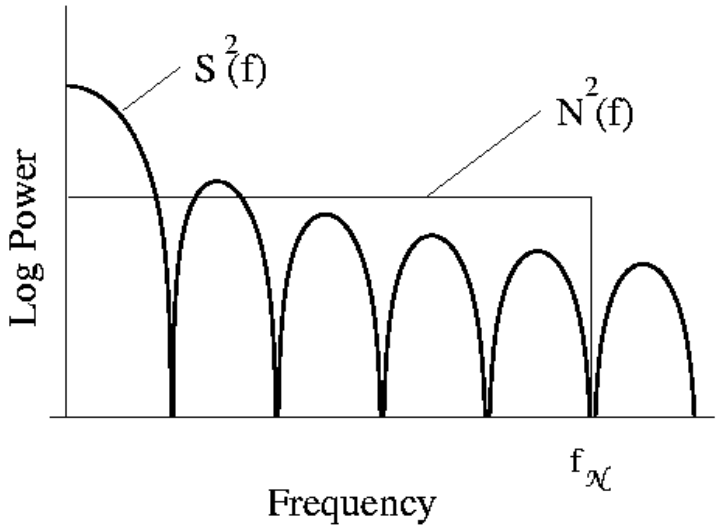
GRAND IDEA:
1. Give them the probability distribution of object persistance times
2. They'll give you the optimal local training rule

- Actually look at a simple toy example, and apply 'hand wavey' theory

- Provokes one to think how a Gatsbyite might approach the same problem

Toy problem 1 - Setup





- a. Objects persist for a fixed length of time τ in the visual field
- b. A single neural output signal
- c. An attempt to retrieve $s(t)$ from $y(t)$ by taking a weighted sum of the outputs at previous times

Toy Problem 1 - Their Analysis

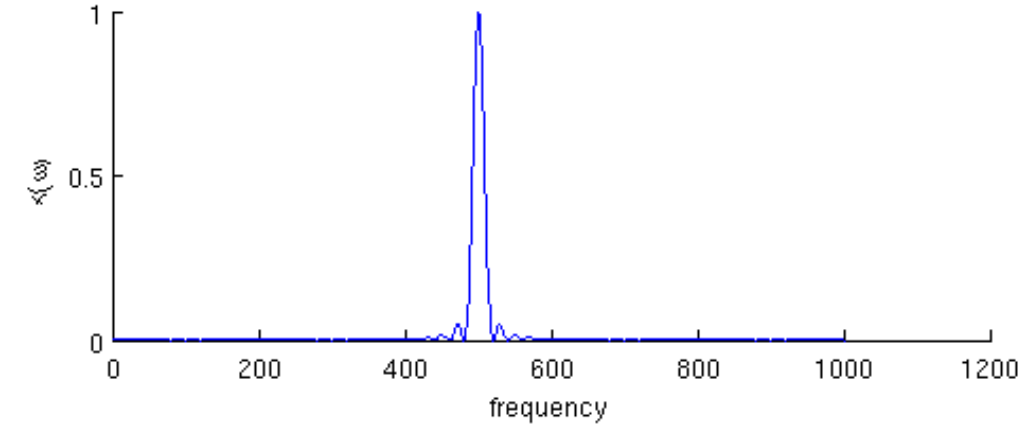
- Let: $y(t) = s(t) + n(t)$
- Recover the signal by a linear filter: $\hat{s}(t) = \int dt' \Phi(t') y(t - t')$
- Wiener tells us the optimal linear filter is:

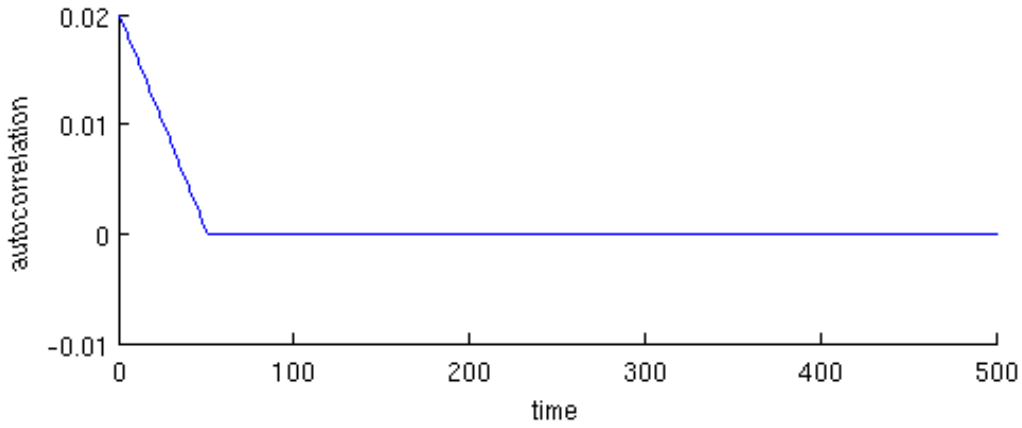
$$\tilde{\Phi}(\omega) = \frac{|\tilde{s}(\omega)|^2}{|\tilde{s}(\omega)|^2 + |\tilde{n}(\omega)|^2} \tag{1}$$

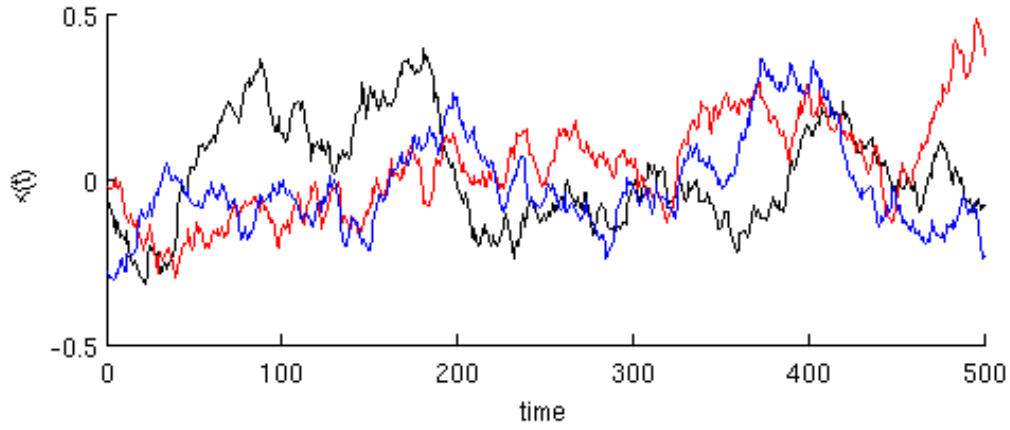
- Specturm of the signal is a sinc function: $\tilde{C}(\omega) = |\tilde{s}(\omega\tau)|^2 \propto \sin^2(\omega\tau)/\omega^2$
- Assume the noise is white: $|\tilde{n}(\omega)| = \rho$

Probabilistic interpretation of their method

- $P[s(t)] = \frac{1}{Z_1} \exp\left[-\frac{1}{2}\int dt dt' s(t) C(t-t') s(t')\right]$
- $P[y(t)|s(t)] = \frac{1}{Z_2} \exp\left[-\frac{1}{2\rho^2}\int dt dt' [y(t)-s(t)][y(t')-s(t')]\right]$
- $P[s(t)|y(t)] = \frac{1}{Z_2} \exp\left[-\frac{1}{2}\int dt dt' [s(t)-\hat{s}(t)]\Sigma^{-1}(t,t')[s(t')-\hat{s}(t')]\right]$
- Clearly the **Gaussian process prior** on $s(t)$ is a terrible choice for **binary variables**.
- Typical samples are not top-hats (mixture of high frequency (fast on/offset of top-hat) and low frequency components from the constant portion.
- As the expected value of a binary $s(t)$ is continuous their model won't look too bad.

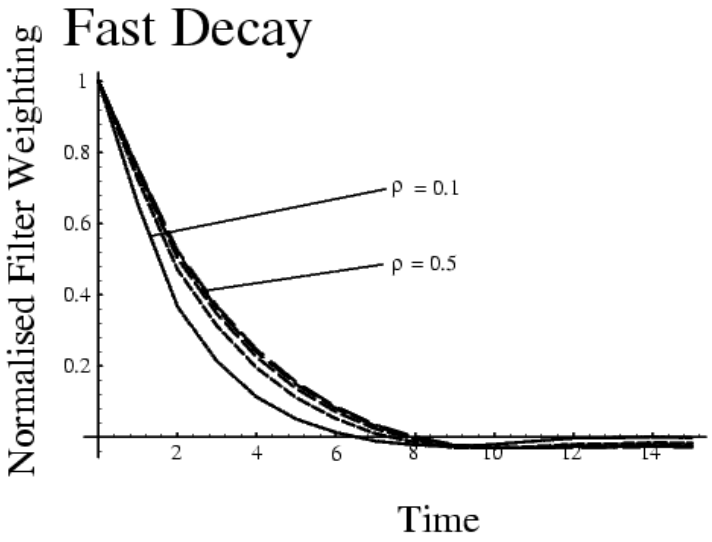

1
0.5
0
<(ω)
0
200
400
600
800
1000
1200
frequency


0.02
0.01
0
-0.01
autocorrelation
0
100
200
300
400
500
time


0.5
0
-0.5
<(t)
0
100
200
300
400
500
time

Their extension to variable presentation lengths

- **Propose presentation times are Jeffrey's distributed**: $P(\tau) = 1/\tau$
- Colloqially: 'I have no idea of the scale of τ'
- Then **average over the optimal filters**: $\langle \Phi(t) \rangle = \int d\tau P(\tau) \Phi_\tau(t)$



The learning rule - trace learning

- They state, rather than derive, a 'compatible' learning rule
- **Normal Hebb rule** says: 'If I'm high and an input j is high - let's strengthen the weight between us'.
- **New trace rule** says: 'If the object to which the output responds is present, and input j is high then we should strenthen the weights between us.'

$$\Delta w_{i,j}^{(t)} = \alpha \bar{y}_i^{(t)} \tag{2}$$

$$\sum_j w_{i,j}^2 = 1 \quad \forall \quad i \tag{3}$$

$$\bar{y}_i^{(t)} = (1-\eta) y_i^{(t)} + \eta \bar{y}_i^{(t-1)} \tag{4}$$

- They show that the recursion 4 can form filters like the 'optimal' filter derived previously.