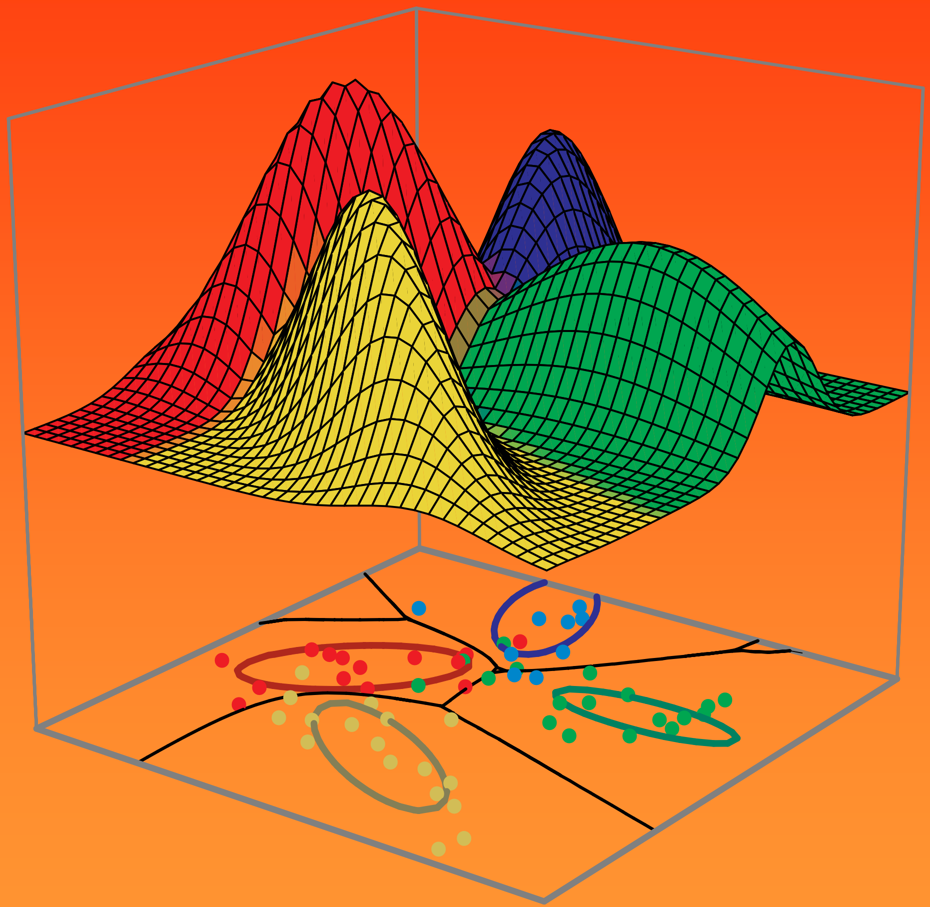


# Pattern Classification

All materials in these slides were taken from
*Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher

# Chapter 3 (Part 3): Maximum-Likelihood and Bayesian Parameter Estimation (Section 3.10)

- Hidden Markov Model: Extension of Markov Chains

- Hidden Markov Model (HMM)

  - Interaction of the visible states with the hidden states

    $\sum b_{jk} = 1$ *for all j where* $b_{jk} = P(V_k(t) \mid \omega_j(t))$.

  - 3 problems are associated with this model

    - The evaluation problem
    - The decoding problem
    - The learning problem

- The evaluation problem

It is the probability that the model produces a sequence $V^T$ of visible states. It is:

$$P(V^T) = \sum_{r=1}^{r_{max}} P(V^T \mid \omega_r^T) P(\omega_r^T)$$

where each r indexes a particular sequence $\omega_r^T = \{\omega(1), \omega(2), \ldots, \omega(T)\}$ of T hidden states.

$$(1) \qquad P(V^T \mid \omega_r^T) = \prod_{t=1}^{t=T} P(v(t) \mid \omega(t))$$

$$(2) \qquad P(\omega_r^T) = \prod_{t=1}^{t=T} P(\omega(t) \mid \omega(t-1))$$

Using equations (1) and (2), we can write:

$$P(V^T) = \sum_{r=1}^{r_{max}} \prod_{t=1}^{t=T} P(v(t) \mid \omega(t))\, P(\omega(t) \mid \omega(t-1))$$

Interpretation: The probability that we observe the particular sequence of T visible states $V^T$ is equal to the sum over all $r_{max}$ possible sequences of hidden states of the conditional probability that the system has made a particular transition multiplied by the probability that it then emitted the visible symbol in our target sequence.

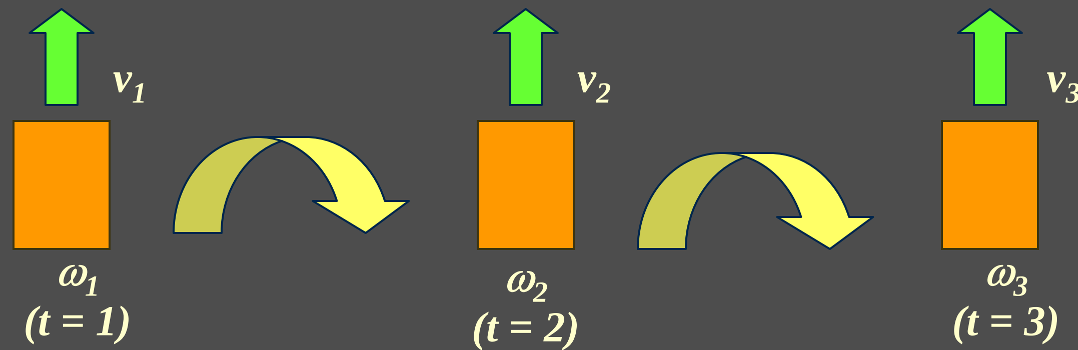
Example: Let $\omega_1, \omega_2, \omega_3$ be the hidden states; $v_1, v_2, v_3$ be the visible states

and $V^3 = \{v_1, v_2, v_3\}$ is the sequence of visible states

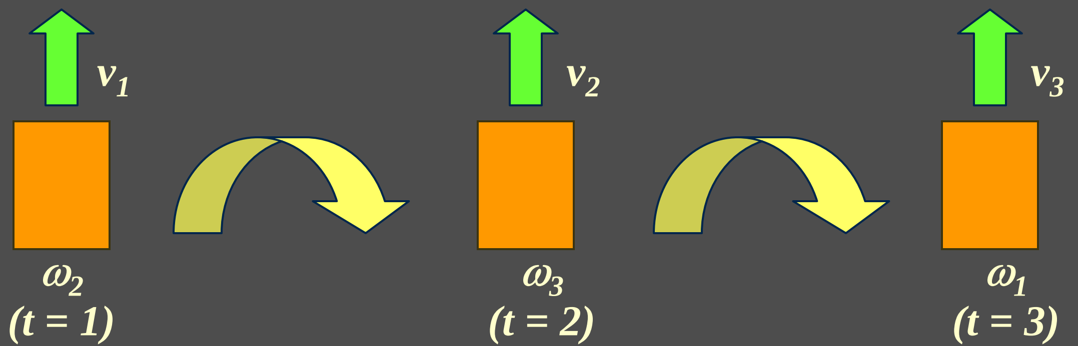
$$P(\{v_1, v_2, v_3\}) = P(\omega_1).P(v_1 \mid \omega_1).P(\omega_2 \mid \omega_1).P(v_2 \mid \omega_2).P(\omega_3 \mid \omega_2).P(v_3 \mid \omega_3)$$

+…+ (possible terms in the sum= all possible ($3^3$= 27) cases !)

First possibility:




Second Possibility:




$P(\{v_1, v_2, v_3\}) = P(\omega_2).P(v_1 \mid \omega_2).P(\omega_3 \mid \omega_2).P(v_2 \mid \omega_3).P(\omega_1 \mid \omega_3).P(v_3 \mid \omega_1) + \ldots +$

Therefore:

$$P(\{v_1, v_2, v_3\}) = \sum_{\substack{\text{possible sequence} \\ \text{of hidden states}}} \prod_{t=1}^{t=3} P(v(t) \mid \omega(t)).P(\omega(t) \mid \omega(t-1))$$

- The decoding problem (optimal state sequence)

Given a sequence of visible states $V^T$, the decoding problem is to find the most probable sequence of hidden states.

This problem can be expressed mathematically as:

*find the single "best" state sequence (hidden states)*

$$\hat{\omega}(1),\hat{\omega}(2),\ldots,\hat{\omega}(T)\ \text{such that:}$$

$$\hat{\omega}(1),\hat{\omega}(2),\ldots,\hat{\omega}(T) = \underset{\omega(1),\omega(2),\ldots,\omega(T)}{\arg\max}\ P\left[\omega(1),\omega(2),\ldots,\omega(T),v(1),v(2),\ldots,V(T)\,|\,\lambda\right]$$

Note that the summation disappeared, since we want to find Only one unique best case !

Where: $\lambda = [\pi, A, B]$

$\pi = P(\omega(1) = \omega)$ *(initial state probability)*

$A = a_{ij} = P(\omega(t+1) = j \mid \omega(t) = i)$

$B = b_{jk} = P(v(t) = k \mid \omega(t) = j)$

In the preceding example, this computation corresponds to the selection of the best path amongst:

$$\{\omega_1(t=1), \omega_2(t=2), \omega_3(t=3)\}, \{\omega_2(t=1), \omega_3(t=2), \omega_1(t=3)\}$$
$$\{\omega_3(t=1), \omega_1(t=2), \omega_2(t=3)\}, \{\omega_3(t=1), \omega_2(t=2), \omega_1(t=3)\}$$
$$\{\omega_2(t=1), \omega_1(t=2), \omega_3(t=3)\}$$

- The learning problem (parameter estimation)

This third problem consists of determining a method to adjust the model parameters $\lambda = [\pi, A, B]$ to satisfy a certain optimization criterion. We need to find the best model

$$\hat{\lambda} = [\hat{\pi}, \hat{A}, \hat{B}]$$

Such that to maximize the probability of the observation sequence:

$$\underset{\lambda}{Max}\, P(V^T \mid \lambda)$$

We use an iterative procedure such as Baum-Welch or Gradient to find this local optimum