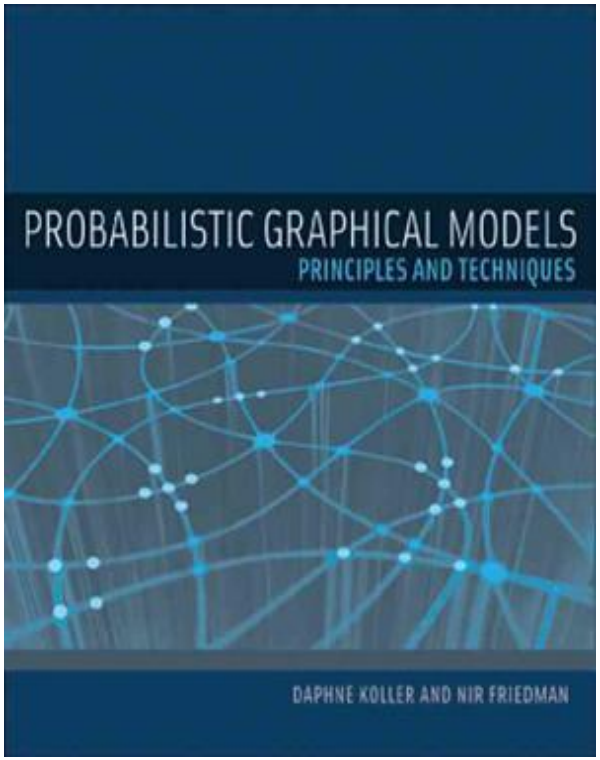
Hidden Markov Model

-- Probabilistic Graphical Model Perspective

Rui Li

Resources

- **Textbook and Tutorial**



A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Although initially introduced and studied in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years. There are two strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications. In this paper we attempt to carefully and methodically review the theoretical aspects of this type of statistical modeling and show how they have been applied to selected problems in machine recognition of speech.

I. INTRODUCTION

Real-world processes generally produce observable outputs which can be characterized as signals. The signals can be discrete in nature (e.g., characters from a finite alphabet, quantized vectors from a codebook, etc.), or continuous in nature (e.g., speech samples, temperature measurements, music, etc.). The signal source can be stationary (i.e., its statistical properties do not vary with time), or nonstationary (i.e., the signal properties vary over time). The signals can be pure (i.e., coming strictly from a single source), or can be corrupted from other signal sources (e.g., noise) or by transmission distortions, reverberation, etc.

A problem of fundamental interest is characterizing such real-world signals in terms of signal models. There are several reasons why one is interested in applying signal models. First of all, a signal model can provide the basis for a theoretical description of a signal processing system which can be used to process the signal so as to provide a desired output. For example if we are interested in enhancing a speech signal corrupted by noise and transmission distortion, we can use the signal model to design a system which will optimally remove the noise and undo the transmission distortion. A second reason why signal models are important is that they are potentially capable of letting us learn a great deal about the signal source (i.e., the real-world process which produced the signal) without having to have the source available. This property is especially important when the cost of getting signals from the actual source is high. In this case, with a good signal model, we can simulate the source and learn as much as possible via simulations. Finally, the most important reason why signal models are important is that they often work extremely well in practice, and enable us to realize important practical systems—e.g., prediction systems, recognition systems, identification systems, etc., in a very efficient manner.

These are several possible choices for what type of signal model is used for characterizing the properties of a given signal. Broadly one can dichotomize the types of signal models into the class of deterministic models, and the class of statistical models. Deterministic models generally exploit some known specific properties of the signal, e.g., that the signal is a sine wave, or a sum of exponentials, etc. In these cases, specification of the signal model is generally straightforward; all that is required is to determine (estimate) values of the parameters of the signal model (e.g., amplitude, frequency, phase of a sine wave, amplitudes and rates of exponentials, etc.). The second broad class of signal models is the set of statistical models in which one tries to characterize only the statistical properties of the signal. Examples of such statistical models include Gaussian processes, Poisson processes, Markov processes, and hidden Markov processes, among others. The underlying assumption of the statistical model is that the signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined (estimated) in a precise, well-defined manner.

For the applications of interest, namely speech processing, both deterministic and stochastic signal models have had good success. In this paper we will concern ourselves strictly with one type of stochastic signal model, namely the hidden Markov model (HMM). (These models are referred to as Markov sources or probabilistic functions of Markov chains in the communications literature.) We will first review the theory of Markov chains and then extend the ideas to the class of hidden Markov models using several simple examples. We will then focus our attention on the three fundamental problems[1] for HMM design, namely: the

Manuscript received January 15, 1988; revised October 4, 1988.
The author is with AT&T Bell Laboratories, Murray Hill, NJ 07974-2070, USA.
IEEE Log Number 8825949.

[1]The idea of characterizing the theoretical aspects of hidden Markov modeling in terms of solving three fundamental problems is due to Jack Ferguson of IDA (Institute for Defense Analysis) who introduced it in lectures and writing.

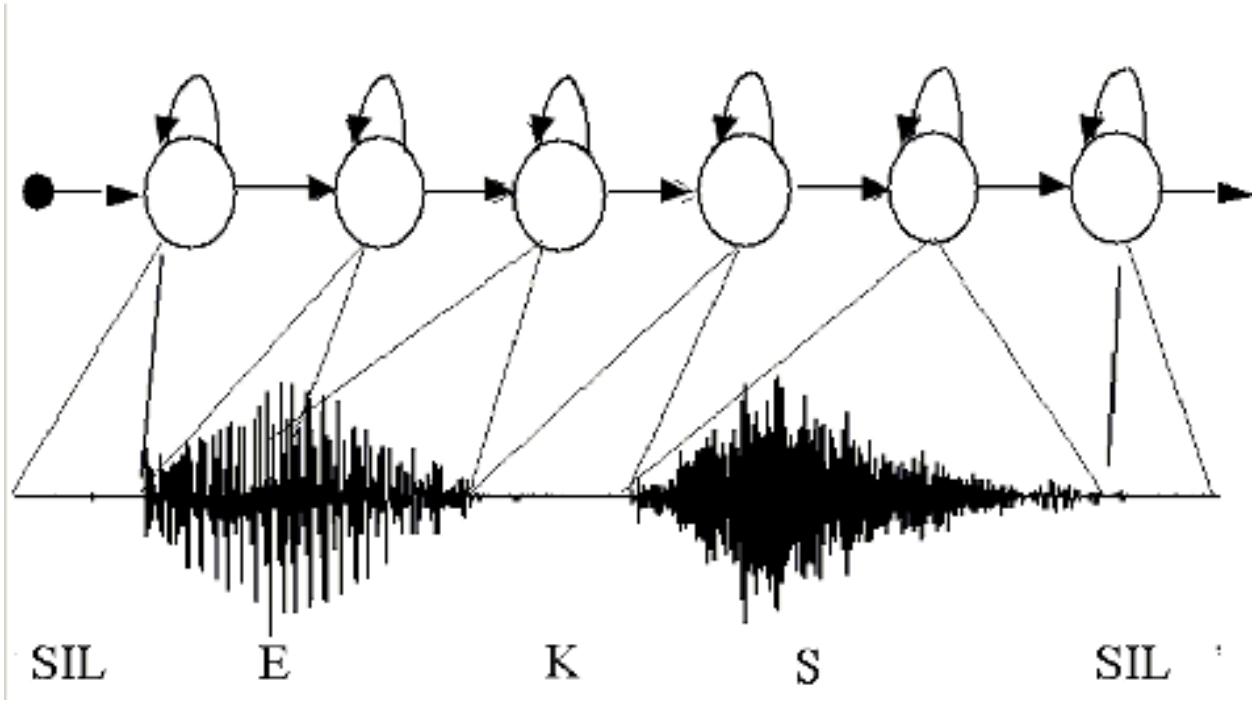
0018-9219/89/0200-0257$01.00 © 1989 IEEE

Resources

- **Software**
  - Hidden Markov Model (HMM) Matlab Toolbox
    - By Kevin Murphy
  - GraphLab
    - By CMU
  - Hidden Markov Model Toolkit (HTK)
    - C Libraries

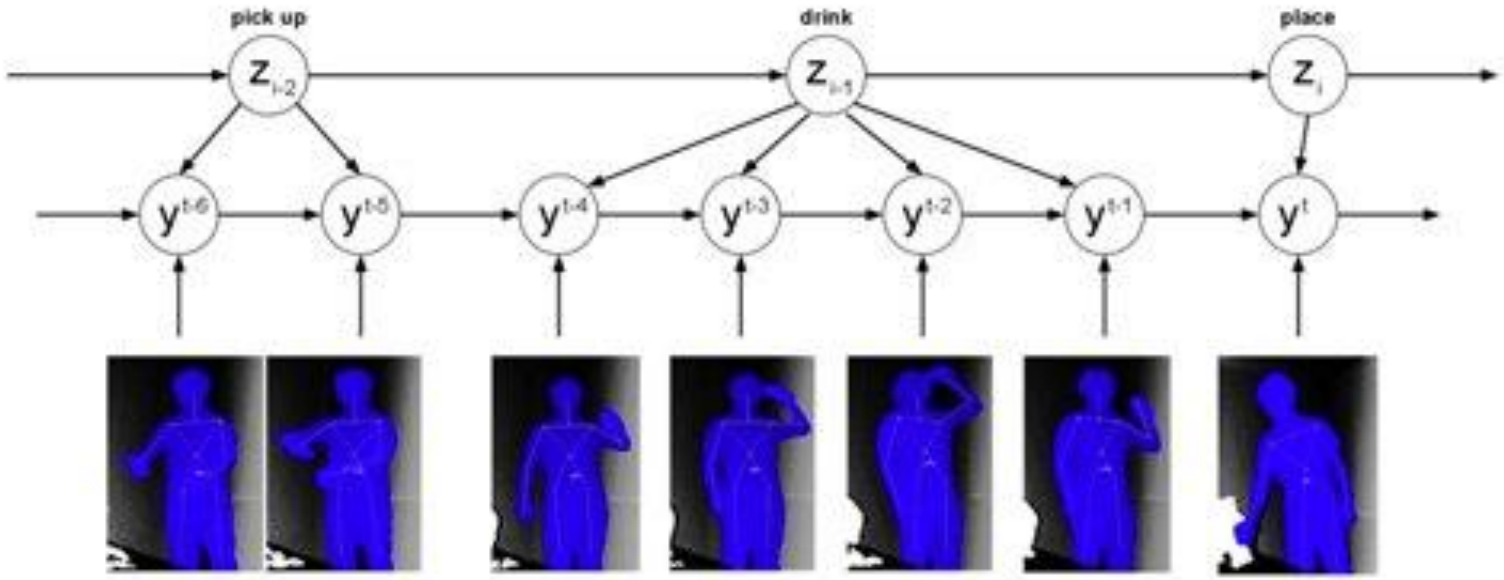
Dynamic Phenomena

- **Speech Recognition**



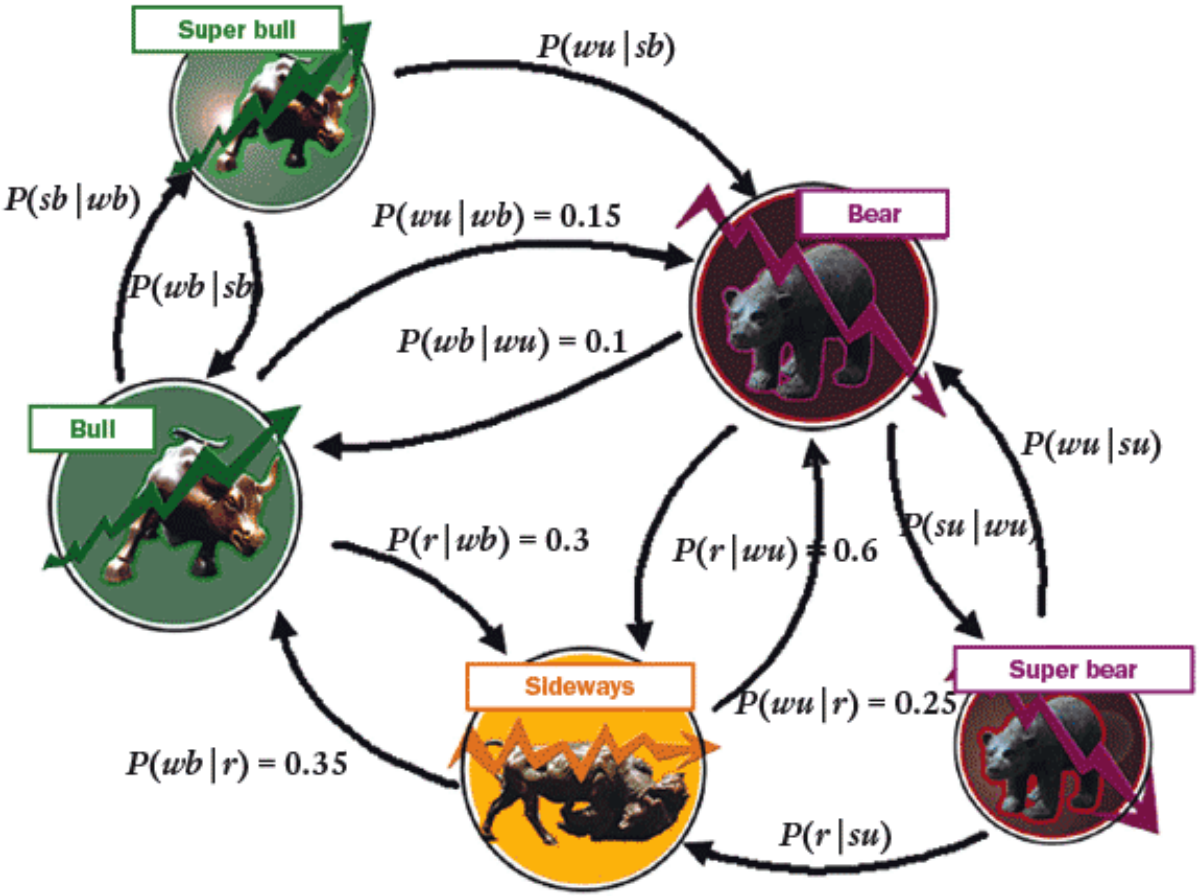
Dynamic Phenomena

- **Body Motion Tracking**



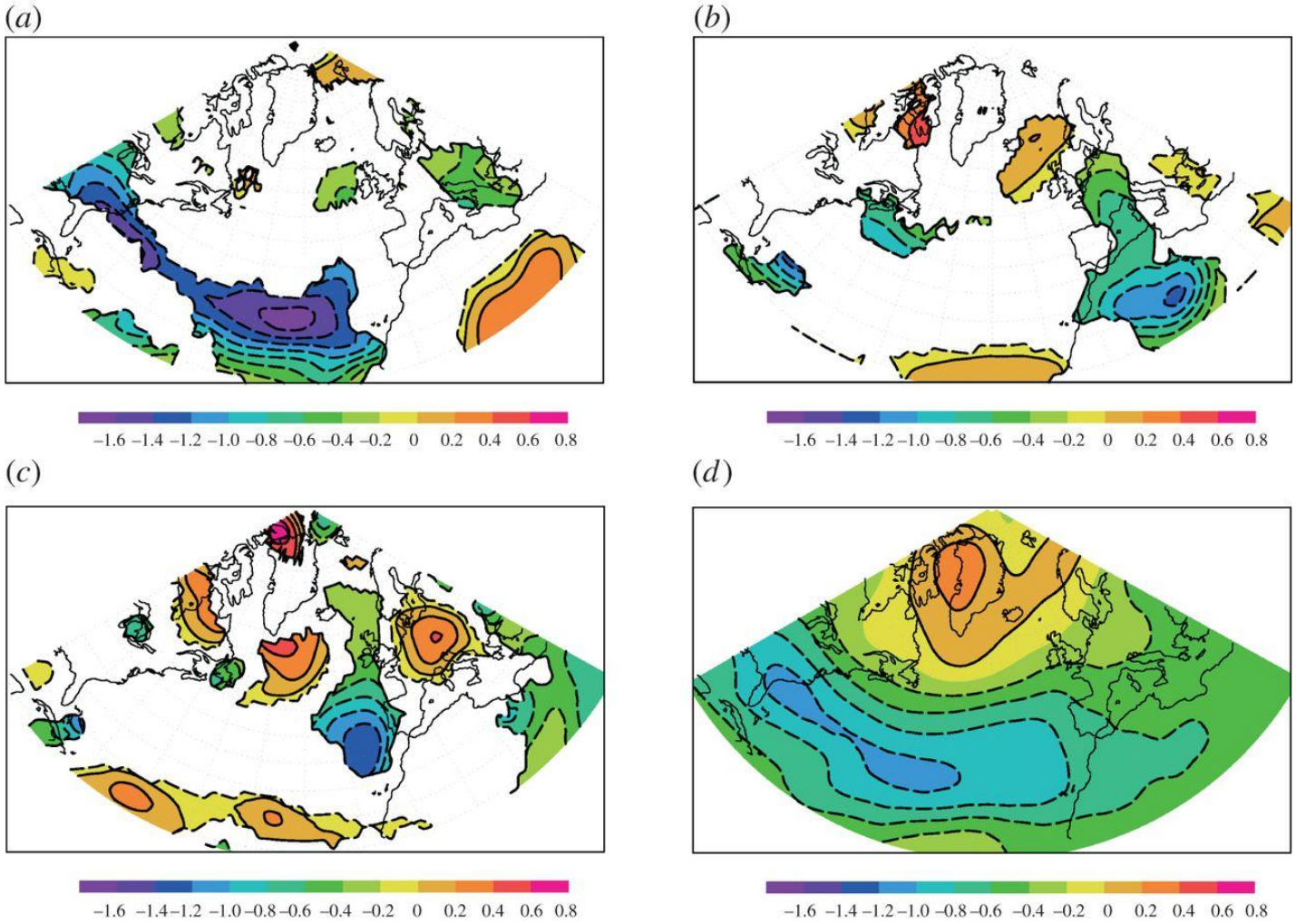
Dynamic Phenomena

- **Stock Prediction**



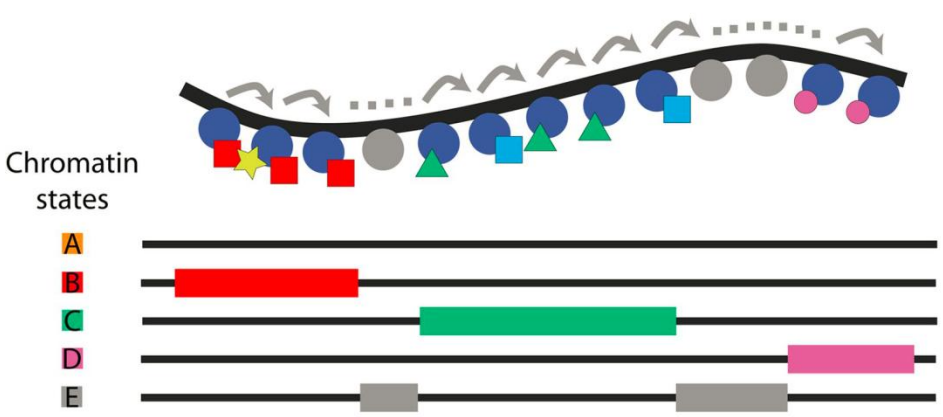
Dynamic Phenomena

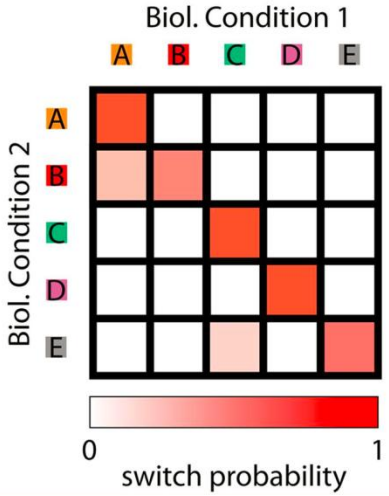
- **Climate Change**

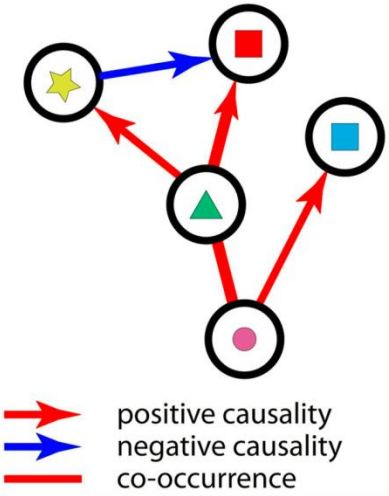


Bioinformatics

- **DNA Sequences**





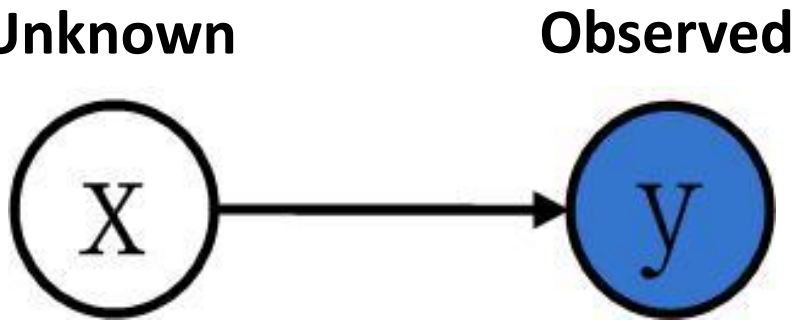


Outline

- **Lecture One**
  - HMMs as Probabilistic Graphical Models
    - Motivation
    - Algebraic representation
    - Graphical representation
- **Lecture Two**
  - HMMs with Inference and Learning
    - Message Passing (Forward-Backward)
    - Expectation-Maximization (Baum-Welch)
    - Application Demos

Motivation

- **A simple graphical model**

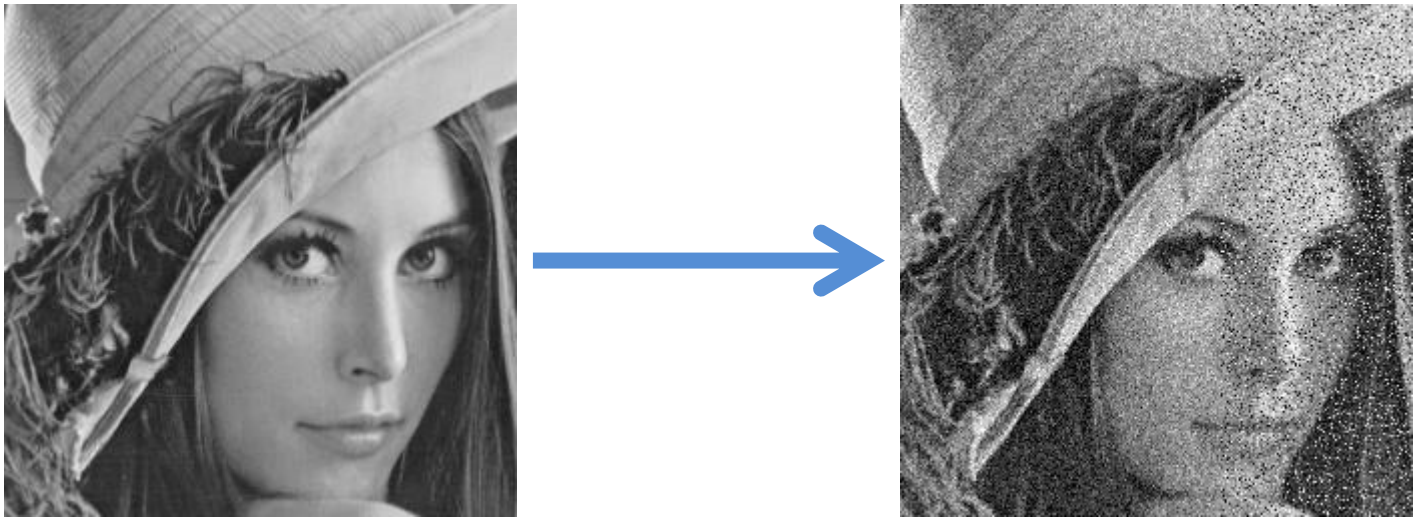


$$P(Y \mid X)$$

Motivation

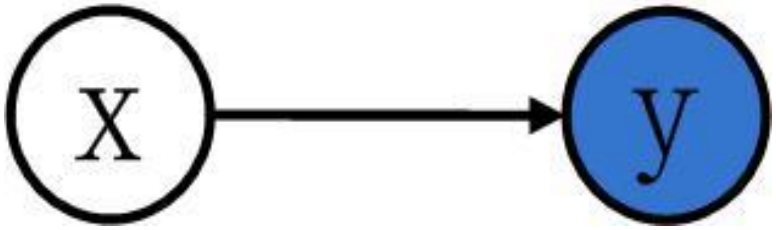
- **An Example**

x **y**



Motivation

- **Inference**



Posterior probability

Prior probability

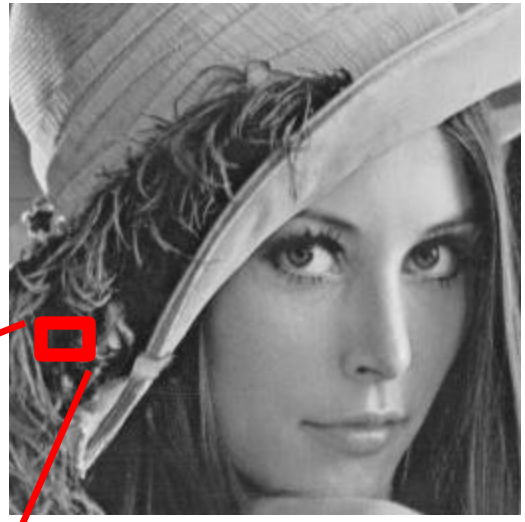
Noise model

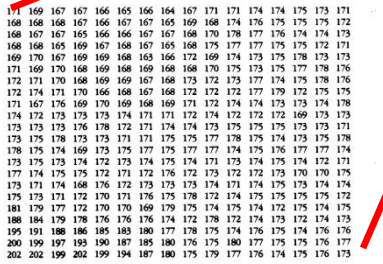
$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y \mid X)}{P(Y)}$$

Constant

Curse of Dimensionality

x



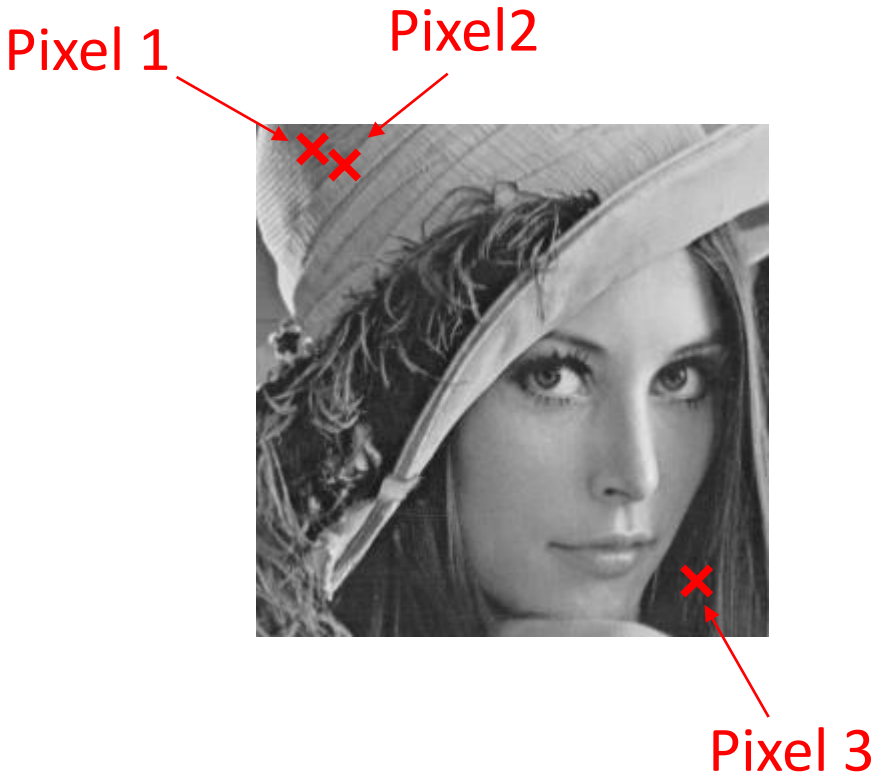


Size of the lookup table of $P(X)$

$$|P(X)| = 256^{100\times100} = 2^{80000}$$

Probabilistic Graphical Model

- **The basic idea**
  - $P(X)$ has some locality properties encoded by graphs

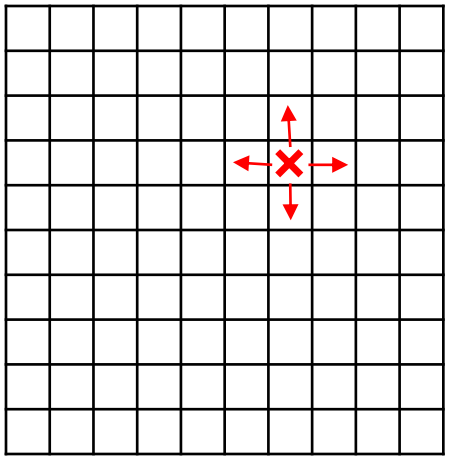


Object Tracking

$X = \{x_1, x_2, ..., x_{T-1}, x_T\}$ x_t location at time *t*

$Y = \{y_1, y_2, ..., y_{T-1}, y_T\}$ y_t sensor measurement at time *t*

$T = 1000$



Inference: $P(X \mid Y)$

Computation Complexity:

$$| P(X \mid Y) | = (10 \times 10)^{1000}$$

Probabilistic Graphical Model

- **PDF Representation**
  - $P(X, Y)$
- **Inference**
  - Given Y'
  - Use $P(X \mid Y')$ to solve problems
- **Learning**
  - Given $\{(x_t, y_t)\}$
  - Fit $P(X \mid Y)$

Hidden Markov Models

- **Definition**
  - $\{x_t, y_t\}_{1...T}$ are a HMM, if
    - X is a Markov process

$$P(x_t \mid x_1, x_2 ..., x_{t-1}, x_{t+1}, ..., x_{T-1}, x_T) = P(x_t \mid x_{t-1})$$

    - y_t only depends on x_t

$$P(y_t \mid x_1, x_2, ..., x_{t-1}, x_t, x_{t+1}, ..., x_{T-1}, x_T) = P(y_t \mid x_t)$$

Hidden Markov Models

- **Representation**
  - Claim: for $\{x_t, y_t\}_{1...T}$ as a HMM

$$P(X,Y) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}) P(y_t \mid x_t)$$

Proof: $P(X,Y) = P(X)P(Y \mid X)$

$$\begin{aligned}
P(X) &= P(x_1, x_2, ..., x_T) \\
&= P(x_T \mid x_{T-1}, x_{T-2}, ...x_1) P(x_{T-1} \mid x_{T-2}, x_{T-3}, ..., x_1) ... P(x_2 \mid x_1) P(x_1) \\
&= P(x_T \mid x_{T-1}) P(x_{T-1} \mid x_{T-2}) ... P(x_2 \mid x_1) P(x_1) \\
&= \prod_{t=1}^{T} P(x_t \mid x_{t-1})
\end{aligned}$$

$$\begin{aligned}
P(Y \mid X) &= P(y_1, y_2, ...y_T \mid X) \\
&= P(y_T \mid y_{T-1}, y_{T-2}, ...y_1, X) P(y_{T-1} \mid y_{T-2}, y_{T-3}, ..., y_1, X) ... P(y_2 \mid y_1, X) P(y_1 \mid X) \\
&= P(y_T \mid x_T) P(y_{T-1} \mid x_{T-1}) ... P(y_2 \mid x_2) P(y_1 \mid x_1) \\
&= \prod_{t=1}^{T} P(y_t \mid x_t)
\end{aligned}$$

Hidden Markov Models

- **Representation**
  - Claim: for $\{x_t, y_t\}_{1...T}$ as a HMM

$$P(X,Y) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}) P(y_t \mid x_t)$$

Computational Complexity:

before claim: $|P(X,Y)| = 100^{1000} \times 100^{1000}$

after claim: $|P(X,Y)| = 2000 \times 100^2$

Hidden Markov Models

- **Representation**
  - Claim: for $\{x_t, y_t\}_{1...T}$ as a HMM

$$P(X,Y) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}) P(y_t \mid x_t)$$

Statistical queries:

$$P(x_t \mid x_{t-1}, x_{t-2}) = P(x_t \mid x_{t-1})$$

$$P(x_t \mid x_{t-2}, x_{t-3})$$

$$= \sum_{x_{t-1}} P(x_t, x_{t-1} \mid x_{t-2}, x_{t-3})$$

$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}, x_{t-2}, x_{t-3}) P(x_{t-1} \mid x_{t-2}, x_{t-3})$$

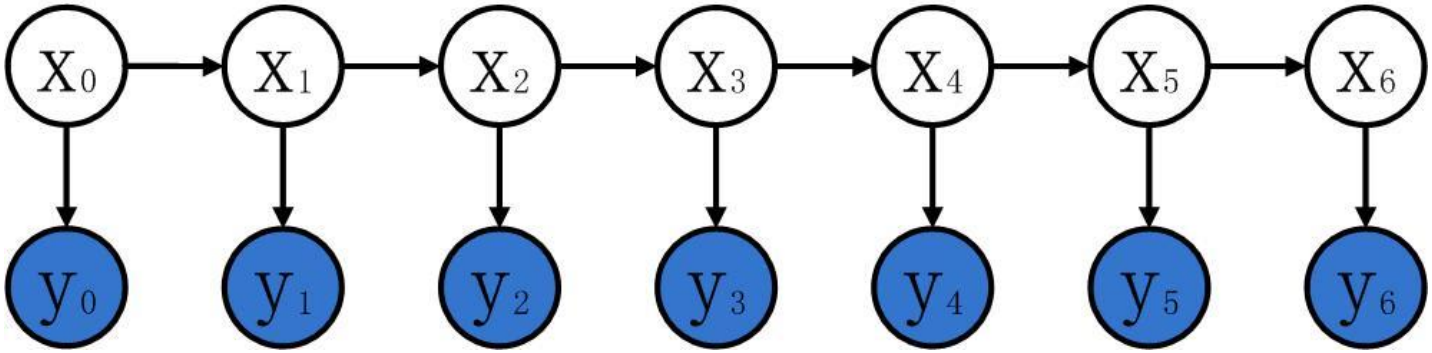
$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid x_{t-2})$$

$$= \sum_{x_{t-1}} P(x_t, x_{t-1} \mid x_{t-2})$$

$$= P(x_t \mid x_{t-2})$$

HMMs and Graphical Models

- **Definition**
  - The graph represents a HMM is
    - a chain of x_t
    - y_t connects to x_t

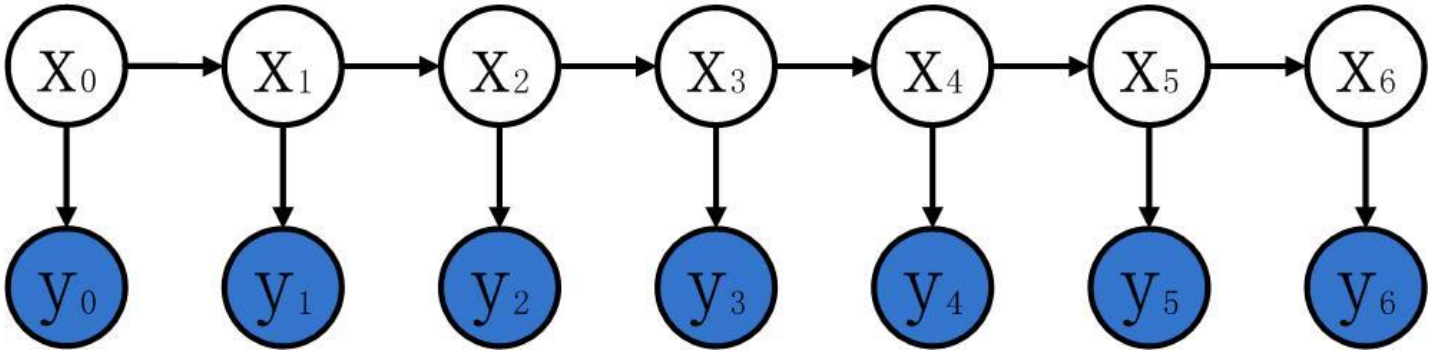


HMMs and Graphical Models

- **Theorem (Hammersley-Clifford)**
  - Given any random variables A, B and C

$$P(C \mid A, B) = P(C \mid B)$$

iff B separates A and C in the graph



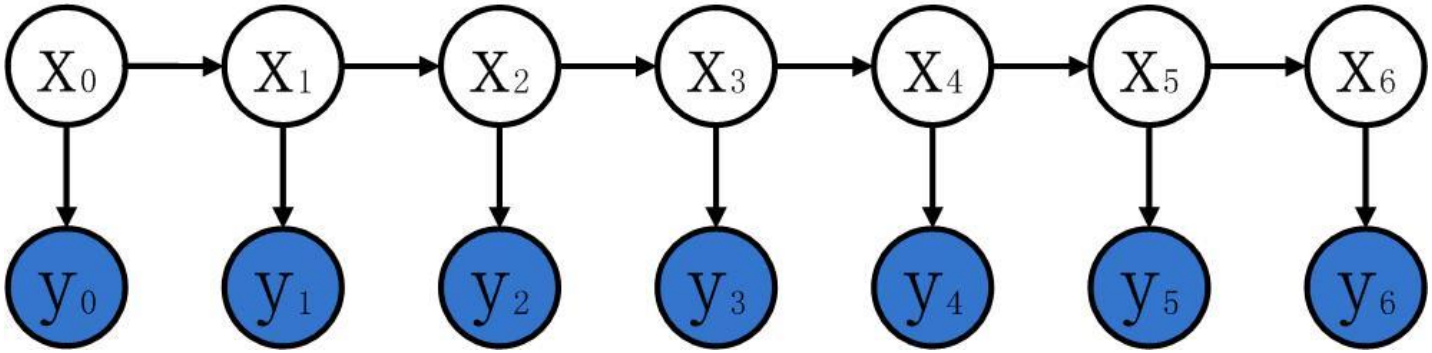
$$P(x_6 \mid x_1, x_2) = P(x_6 \mid x_2)$$

$$P(y_6 \mid y_4, y_5)$$

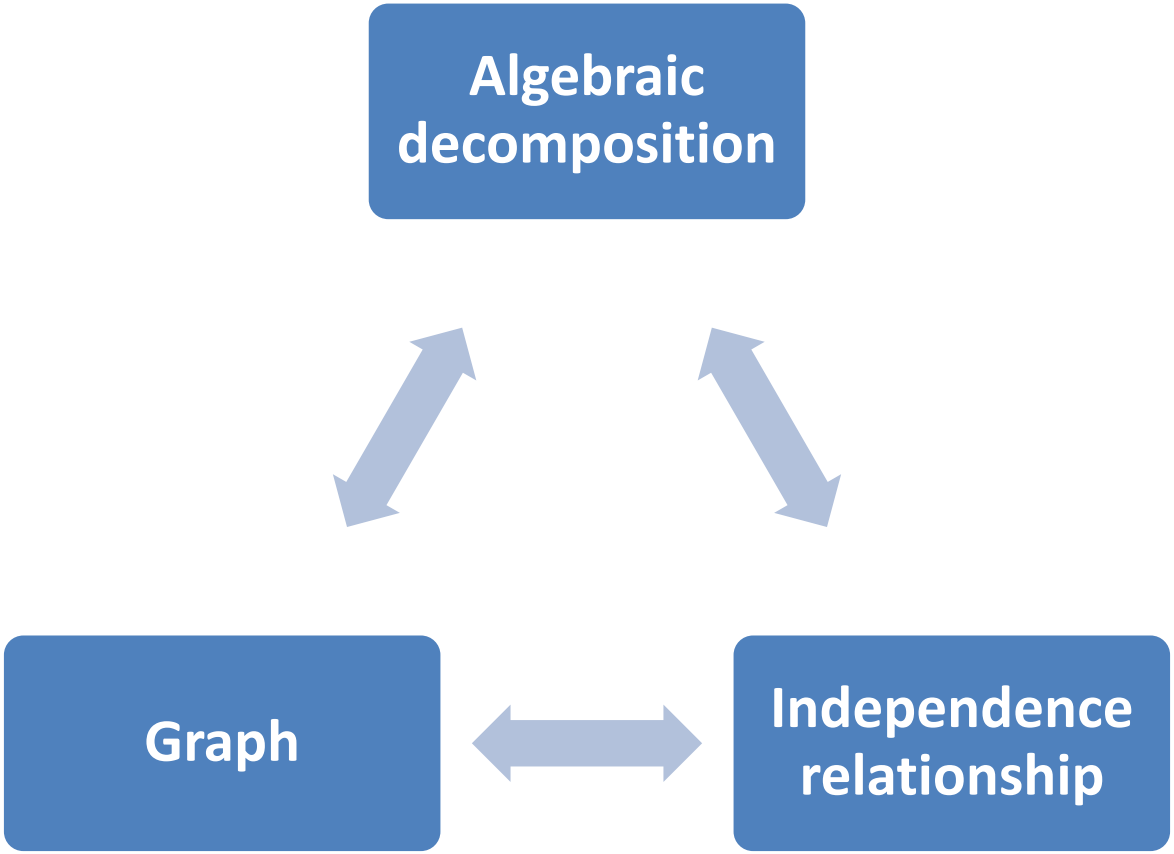
HMMs and Graphical Models

- **Graph & Factorization**

$$P(X,Y) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}) P(y_t \mid x_t)$$



HMMs and Graphical Models



Probabilistic Graphical Model

- **PDF Representation**
  - $P(X, Y)$
- **Inference**
  - Given Y'
  - Use $P(X \mid Y')$ to solve problems
- **Learning**
  - Given $\{(x_t, y_t)\}$
  - Fit $P(X \mid Y)$

Inference

- **MAP (Maximum A Posteriori)**

$$X^* = \arg\max_X P(X \mid Y')$$

- **Marginalization**

$$P(x_t \mid Y') \qquad \forall t$$

Henceforth, “inference”== marginalization