Andrei Andreyevich Markov
1856–1922

1878 gold medal at St Petersburg University: *On the integration of differential equations by means of continued fractions*.

1880 Master's: *On the binary quadratic forms with positive determinant*.

1884 doctorate: *On certain applications of continued fractions*.

After 1900, Markov applied continued fractions, pioneered by Chebyshev, to probability theory.

## Axioms of Probability

Definition: A sample space is a set S together with a function P on subsets A of S, called events, such that:

- $P(A) \geq 0$, for events $A \subseteq S$,
- $P(S) = 1$,
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, for countably many pairwise disjoint events $A_1, A_2, \ldots$.

## Using Axioms to Prove Propositions

Propostion: For any events $A$ and $B$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof: Notice $A \cup B = A \cup (A^c \cap B)$.

Since $A$ and $A^c \cap B$ are disjoint, the axioms give

$$P(A \cup B) = P(A) + P(A^c \cap B). \qquad (1)$$

Notice also $B = (A \cap B) \cup (A^c \cap B)$.

Since $A \cap B$ and $A^c \cap B$ are disjoint, we have

$$P(B) = P(A \cap B) + P(A^c \cap B) \qquad (2)$$

The Proposition follows from (1) and (2).

Note: $P(A \cup B) = P(A) + P(B)$, if $A \cap B = \emptyset$.

# Markov Model of Kitten Behavior

When we find a kitten napping we check back every fifteen minutes to record whether she is exploring, hunting or napping. If we check back 4 times, we might record

$$n, n, h, e, n.$$

A probability distribution on the sample space $\{e, h, n\}^5$ is defined using a *transition matrix* and an *initial distribution*:

$$A = \begin{array}{c} e \\ h \\ n \end{array} \begin{pmatrix} .4 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .8 \end{pmatrix} \qquad \pi = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\begin{aligned} P(nnhen) &= P(n)P(n|n)P(h|n)P(e|h)P(n|e) \\ &= \ \ 1 \ \cdot \ .8 \ \cdot \ .1 \ \cdot \ .2 \ \cdot \ .3 \\ &= .0048 \end{aligned}$$

What fraction of the time does she explore, hunt and nap?

$$\lim_{t \to \infty} A^t \pi = ?$$

| | $t=0$ | $t=1$ | $t=2$ | $t=3$ | $t=4$ |
|---|---|---|---|---|---|
| explore | 0 | .1 | .14 | .159 | .1690 |
| hunt | 0 | .1 | .17 | .213 | .2383 |
| nap | 1 | .8 | .69 | .628 | .5927 |

# Hidden Markov Models

Lynne Butler
Haverford College

**Def:** Let $N, M > 0$. A *Hidden Markov Model* is a triple $(A, B, \pi)$, where

$$\begin{aligned} A &= (a_{rq})_{\substack{r\in Q\\ q\in Q}} \text{ is an } N \times N \text{ matrix,} \\ B &= (b_{vq})_{\substack{v\in V\\ q\in Q}} \text{ is an } M \times N \text{ matrix,} \\ \pi &= (\pi_q)_{q\in Q} \text{ is an } N \times 1 \text{ vector.} \end{aligned}$$

All entries are nonnegative, and in each column the entries sum to 1.

These models are used in speech recognition, cryptology, finance, and genomics.

Elements of $V$ are *observations*, and elements of $Q$ are *states*.

The model defines a probability distribution on the sample space $Q^T \times V^T$, where $T > 0$. The *probability* of a state sequence $(s_0, s_1, \ldots, s_{T-1})$ and observation sequence $(o_0, o_1, \ldots, o_{T-1})$ is:

$$p(s, o) = \pi_{s_0} b_{o_0 s_0} \prod_{t=1}^{T-1} a_{s_t s_{t-1}} b_{o_t s_t}.$$

At any time $t$, we find $b_{vq} = P(O_t = v \mid S_t = q)$ and $a_{rq} = P(S_{t+1} = r \mid S_t = q)$. So, when in state $q$, the probability of observing $v$ is $b_{vq}$ and the probability of transiting to state $r$ is $a_{rq}$.

**Ex:** For sequences of characters observed in English text, let $V = \{a, b, \ldots, z, \_\}$. A 2-state model, obtained from a long training text, has:

$$A = \begin{pmatrix} .27 & .71 \\ .73 & .29 \end{pmatrix} \qquad B = \begin{array}{c} a \\ b \\ \vdots \\ e \\ f \\ g \\ \vdots \\ z \\ \_ \end{array} \begin{pmatrix} .136 & .000 \\ .000 & .024 \\ \vdots & \vdots \\ .202 & .000 \\ .000 & .032 \\ .007 & .023 \\ \vdots & \vdots \\ .000 & .001 \\ .336 & .015 \end{pmatrix}.$$

How would you interpret the states of this model?

What does the model say about English text?

**Ex:** For sequences of % changes in S&P, let $V = \{d, dm, n, um, u\}$ for 1/3/50 to 2/24/07:

$$v = \begin{cases} d & \text{if} \qquad\qquad\qquad\quad \% \leq -.0054004 \\ dm & \text{if} \; -.0054004 < \% \leq -.0010955 \\ n & \text{if} \; -.0010955 < \% \leq \phantom{-}.0019845 \\ um & \text{if} \quad\; .0019845 < \% \leq \phantom{-}.0060816 \\ u & \text{if} \quad\; .0060816 < \% \end{cases} .$$

Bins are defined to capture equal numbers of observations.

How might you expect to interpret the states in a 2-state HMM?

Using code written by Alden Walker, Haverford math major Sumana Shrestha found:

$$A = \begin{pmatrix} .9933 & .0075 \\ .0067 & .9925 \end{pmatrix} \qquad B = \begin{array}{c} d \\ dm \\ n \\ um \\ u \end{array} \begin{pmatrix} .125 & .283 \\ .228 & .169 \\ .264 & .129 \\ .243 & .153 \\ .140 & .266 \end{pmatrix}.$$

To find the 2-state model that best fits given data, use the EM algorithm (also known as the Baum-Welch or forward-backward algorithm).

See "A tutorial on hidden Markov models and selected applications in speech recognition" by Rabiner, published in Proceedings of the IEEE.