CSE 312

Foundations of Computing II

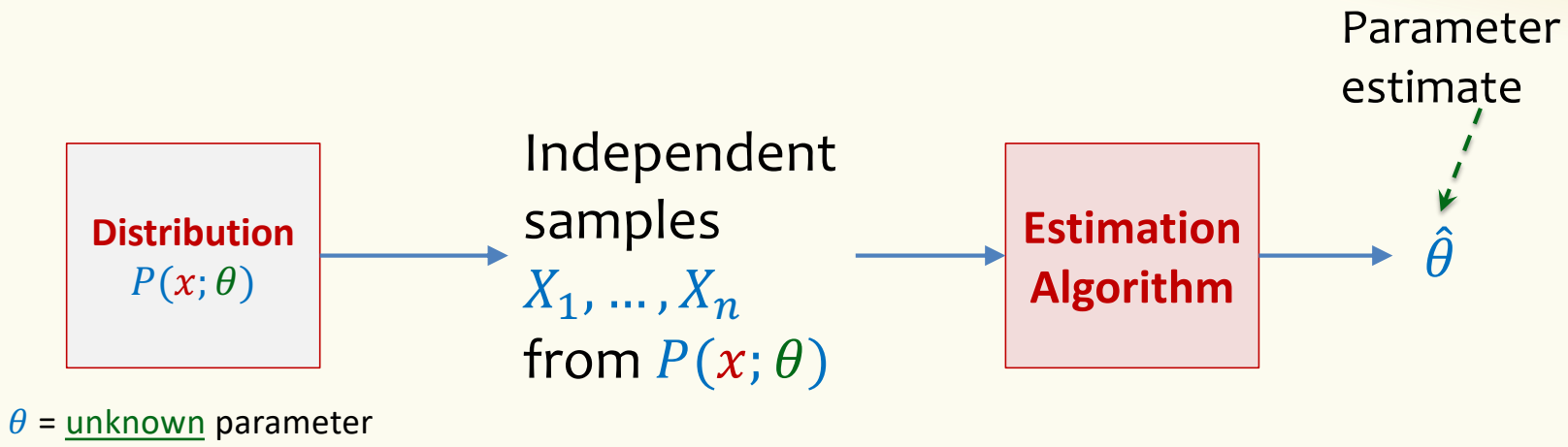
Lecture 24: Finish MLE, Start Markov Chains

www.slido.com/1692973

Agenda

- Wrap up MLE
  - Unbiased and Consistent Estimators
  - Intuition and Bigger Picture
- Markov Chains

Statistics: Parameter Estimation – Workflow



θ = unknown parameter

Example: samples from a normal distribution with unknown mean θ and variance 1

Observation: . $x_1 = 0.33,\ x_2 = -5.8,\ \ldots\ ,\ x_n = 4.5$

Likelihood of Different Observations

(Discrete case)

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} P(x_i; \theta)$$

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ such that $\mathcal{L}(x_1, \ldots, x_n \mid \hat{\theta})$ is maximized!

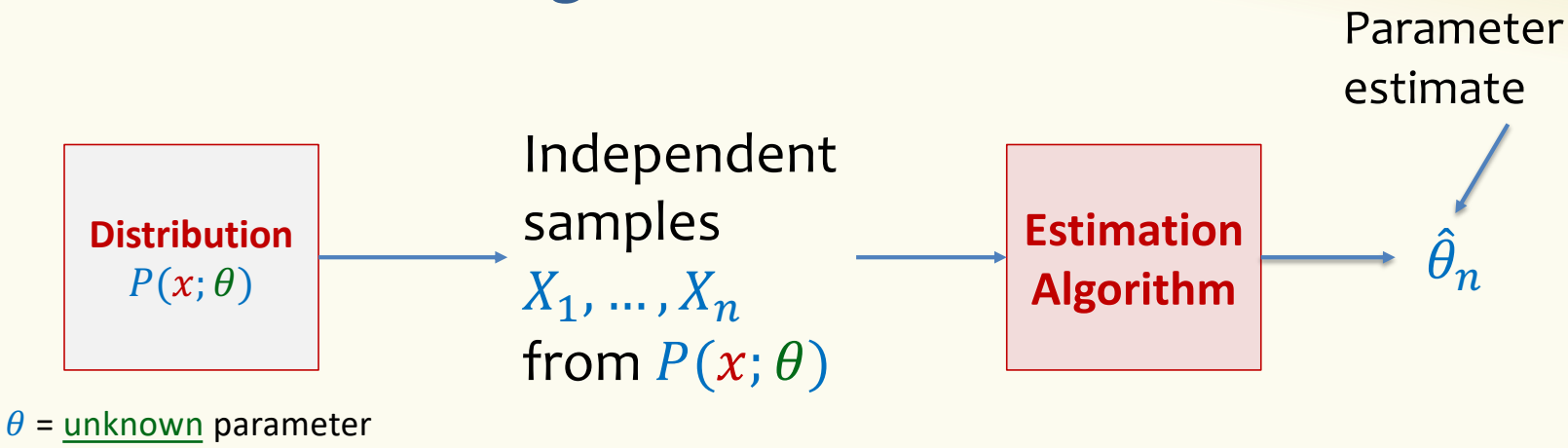
$$\hat{\theta} = \operatorname*{argmax}_{\theta} \mathcal{L}(x_1, \ldots, x_n \mid \theta)$$

General Recipe

1. **Input** Given n i.i.d. samples $x_1, \ldots, x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \ldots., x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \ldots., x_n | \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots., x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots., x_n | \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots., x_n | \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

When is an estimator good?



θ = unknown parameter

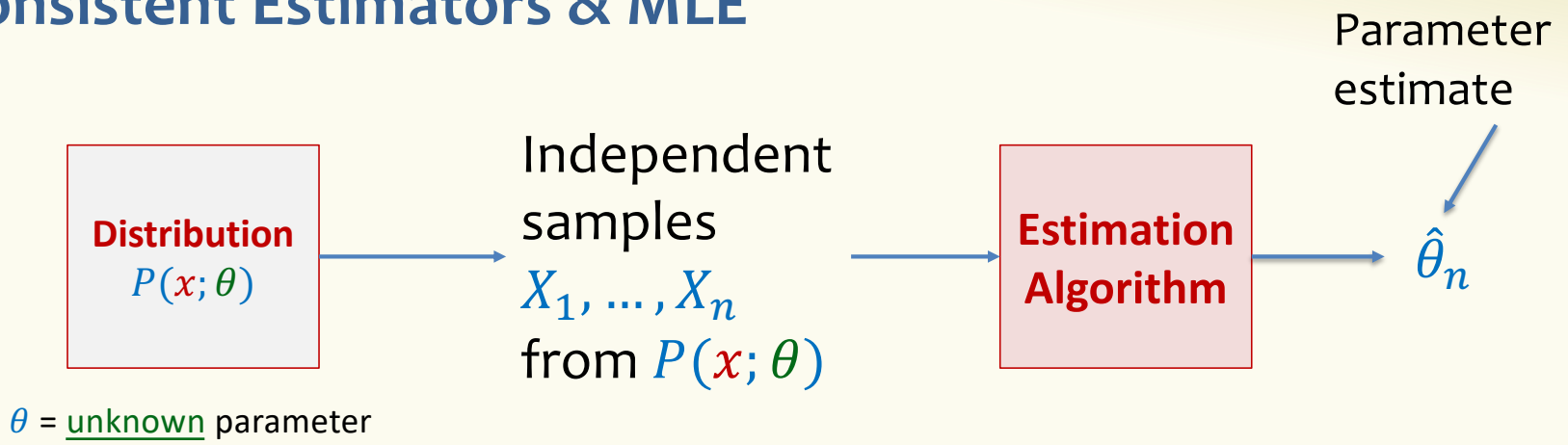
Definition. An estimator of parameter θ is an **unbiased estimator** if

$$\mathbb{E}[\hat{\theta}_n] = \theta.$$

Note: This expectation is over the samples $X_1, \ldots, X_n$

Three samples from $U(0, \theta)$

Consistent Estimators & MLE



θ = unknown parameter

Definition. An estimator is **unbiased** if $\mathbb{E}[\hat{\theta}_n] = \theta$ for all $n \geq 1$.

Example: samples from a normal distribution with unknown mean θ and variance 1

Example – Consistency

Normal outcomes $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Population variance – Biased!

Definition. An estimator is **consistent** if $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_n\right] = \theta.$

$\widehat{\Theta}_{\sigma^2}$ is "consistent"

Example – Consistency

Normal outcomes $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Population variance – Biased!

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Sample variance – Unbiased!

$\widehat{\Theta}_{\sigma^2}$ converges to same value as S_n^2, i.e., σ^2, as $n \to \infty$.

$\widehat{\Theta}_{\sigma^2}$ is "consistent"

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Why does it matter?

- When statisticians are estimating a variance from a sample, they usually divide by $n-1$ instead of n.

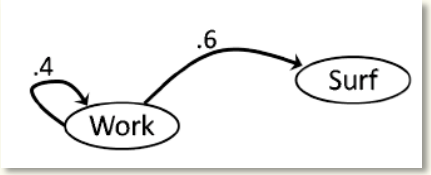
- They and we not only want good estimators (unbiased, consistent)
  - They/we also want **confidence bounds**
    - Upper bounds on the probability that these estimators are far the truth about the underlying distributions
  - Confidence bounds are just like what we wanted for our polling problems, but CLT is usually not the best thing to use to get them (unless the variance is known)

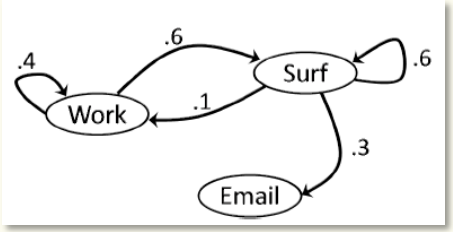
Agenda

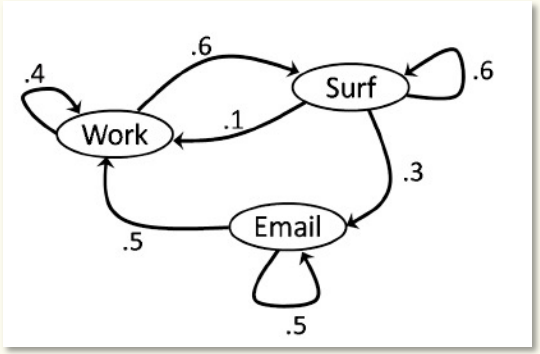
- Wrap up MLE
  - Unbiased and Consistent Estimators
  - Intuition and Bigger Picture
- Markov Chains

A typical day in my life….

time $t = 0$

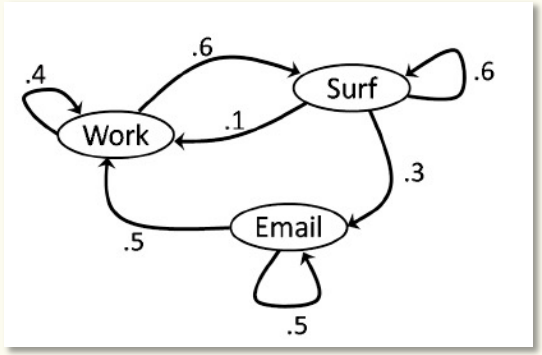






A typical day in my life

How do we interpret this diagram?



This kind of random process is called a **Markov Chain**

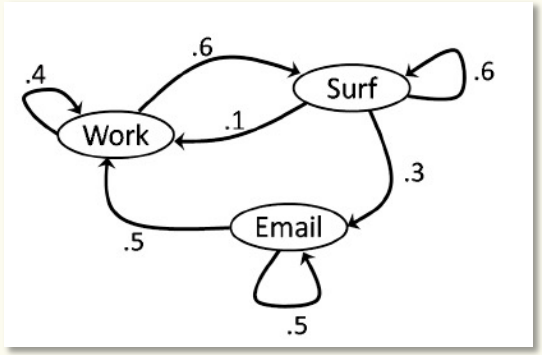
At *each* time step t

- I can be in one of 3 **states**
  - Work, Surf, Email
- If I am in some state s at time t
  - the **labels of out-edges** of s **give the probabilities** of my moving to each of the states at time $t + 1$ (as well as staying the same)
    - so **labels on out-edges sum to** 1

e.g. If I am in Email, there is a 50-50 chance I will be in each of Work or Email at the next time step, but I will never be in state Surf in the next step.

This diagram looks vaguely familiar if you took CSE 311 …

Markov chains are a special kind of *probabilistic (finite) automaton*

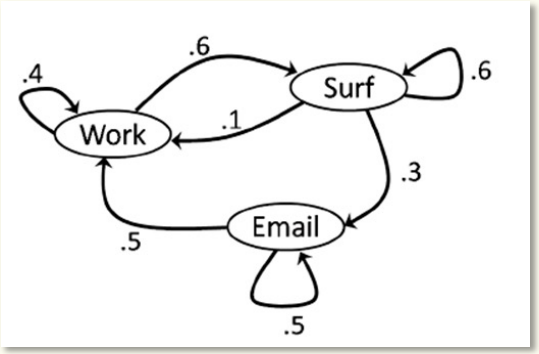


The diagrams look a bit like those of Deterministic Finite Automata (DFAs) you saw in 311 except that…

- There are no input symbols on the edges
  - Think of there being only one kind of input symbol "another tick of the clock" so no need to mark it on the edge
- They have multiple out-edges like an NFA, except that they come with probabilities

But just like DFAs, the only thing they remember about the past is the state they are currently in.

Many interesting questions about Markov Chains



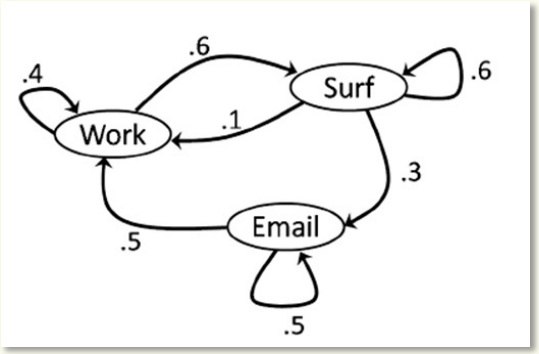
1. **What is the probability that I am in state s at time 1?**
2. **What is the probability that I am in state s at time 2?**

Define variable $X^{(t)}$ to be state I am in at time t

Given: In state Work at time $t = 0$

t	0	1	2
$P(X^{(t)} = \text{Work})$	**1**		
$P(X^{(t)} = \text{Surf})$	**0**		
$P(X^{(t)} = \text{Email})$	**0**		

Many interesting questions about Markov Chains



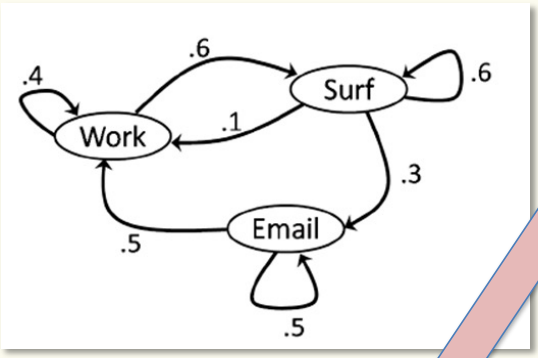
1. **What is the probability that I am in state s at time 1?**
2. **What is the probability that I am in state s at time 2?**

Define variable $X^{(t)}$ to be state I am in at time t

Given: In state Work at time $t = 0$

t	0	1	2
$q_W^{(t)} = P(X^{(t)} = \text{Work})$	**1**	**0.4**	$= \mathbf{0.4} \cdot 0.4 + \mathbf{0.6} \cdot 0.1 = 0.16 + 0.06 = \mathbf{0.22}$
$q_S^{(t)} = P(X^{(t)} = \text{Surf})$	**0**	**0.6**	$= \mathbf{0.4} \cdot 0.6 + \mathbf{0.6} \cdot 0.6 = 0.24 + 0.36 = \mathbf{0.60}$
$q_E^{(t)} = P(X^{(t)} = \text{Email})$	**0**	**0**	$= \mathbf{0.4} \cdot 0 + \mathbf{0.6} \cdot 0.3 = 0 + 0.18 = \mathbf{0.18}$

An organized way to understand the distribution of $X^{(t)}$

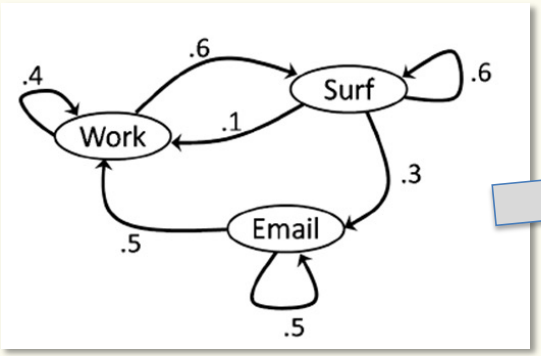


Write as a tuple $(q_W^{(t)}, q_S^{(t)}, q_E^{(t)})$ a.k.a. a row vector:

$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$$

t	0	1	2
$q_W^{(t)} = P(X^{(t)} = \text{Work})$	**1**	**0.4**	$= \mathbf{0.4} \cdot 0.4 + \mathbf{0.6} \cdot 0.1 = 0.16 + 0.06 = \mathbf{0.22}$
$q_S^{(t)} = P(X^{(t)} = \text{Surf})$	**0**	**0.6**	$= \mathbf{0.4} \cdot 0.6 + \mathbf{0.6} \cdot 0.6 = 0.24 + 0.36 = \mathbf{0.60}$
$q_E^{(t)} = P(X^{(t)} = \text{Email})$	**0**	**0**	$= \mathbf{0.4} \cdot 0 + \mathbf{0.6} \cdot 0.3 = 0 + 0.18 = \mathbf{0.18}$

An organized way to understand the distribution of $X^{(t)}$



$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$

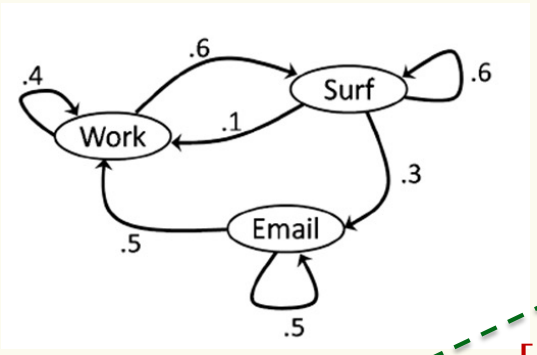
$$\boldsymbol{M} = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Write as a "transition probability matrix" $\boldsymbol{M}$

- one row/col per state. Row=now, Col=next
- each row sums to 1

t	0	1	2
$q_W^{(t)} = P(X^{(t)} = \text{Work})$	$\mathbf{1}$	$\mathbf{0.4}$	$= \mathbf{0.4} \cdot 0.4 + \mathbf{0.6} \cdot 0.1 = 0.16 + 0.06 = \mathbf{0.22}$
$q_S^{(t)} = P(X^{(t)} = \text{Surf})$	$\mathbf{0}$	$\mathbf{0.6}$	$= \mathbf{0.4} \cdot 0.6 + \mathbf{0.6} \cdot 0.6 = 0.24 + 0.36 = \mathbf{0.60}$
$q_E^{(t)} = P(X^{(t)} = \text{Email})$	$\mathbf{0}$	$\mathbf{0}$	$= \mathbf{0.4} \cdot 0 + \mathbf{0.6} \cdot 0.3 = 0 + 0.18 = \mathbf{0.18}$

An organized way to understand the distribution of $X^{(t)}$



$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \quad \overset{\boldsymbol{M}}{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

$$[0.4, 0.6, 0] \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} = [0.22, 0.60, 0.18]$$

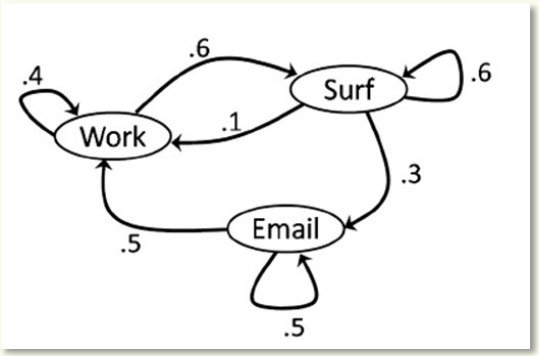
Vector-matrix multiplication

$$q_W^{(1)} = \mathbf{0.4} \qquad q_W^{(2)} = \mathbf{0.4} \cdot 0.4 + \mathbf{0.6} \cdot 0.1 = 0.16 + 0.06 = \mathbf{0.22}$$

$$q_S^{(1)} = \mathbf{0.6} \qquad q_S^{(2)} = \mathbf{0.4} \cdot 0.6 + \mathbf{0.6} \cdot 0.6 = 0.24 + 0.36 = \mathbf{0.60}$$

$$q_E^{(1)} = \mathbf{0} \qquad q_E^{(2)} = \mathbf{0.4} \cdot 0 \quad + \mathbf{0.6} \cdot 0.3 = 0 \quad + 0.18 = \mathbf{0.18}$$

An organized way to understand the distribution of $X^{(t)}$



$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \underbrace{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}}_{\boldsymbol{M}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

Vector-matrix multiplication

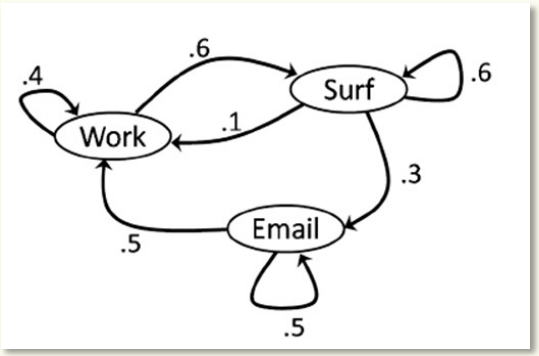
$$q_W^{(t)} \cdot 0.4 + q_S^{(t)} \cdot 0.1 + q_E^{(t)} \cdot 0.5 = q_W^{(t+1)}$$

$$q_W^{(t)} \cdot 0.6 + q_S^{(t)} \cdot 0.6 + q_E^{(t)} \cdot 0 = q_S^{(t+1)}$$

$$q_W^{(t)} \cdot 0 + q_S^{(t)} \cdot 0.3 + q_E^{(t)} \cdot 0.5 = q_E^{(t+1)}$$

Write $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ Then for all $t \geq 0$, $\boldsymbol{q}^{(t)}\boldsymbol{M} = \boldsymbol{q}^{(t+1)}$

An organized way to understand the distribution of $X^{(t)}$



$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \underbrace{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}}_{\boldsymbol{M}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

Vector-matrix multiplication

$$q_W^{(t)} \cdot 0.4 + q_S^{(t)} \cdot 0.1 + q_E^{(t)} \cdot 0.5 = q_W^{(t+1)}$$

$$q_W^{(t)} \cdot 0.6 + q_S^{(t)} \cdot 0.6 + q_E^{(t)} \cdot 0 = q_S^{(t+1)}$$

$$q_W^{(t)} \cdot 0 + q_S^{(t)} \cdot 0.3 + q_E^{(t)} \cdot 0.5 = q_E^{(t+1)}$$

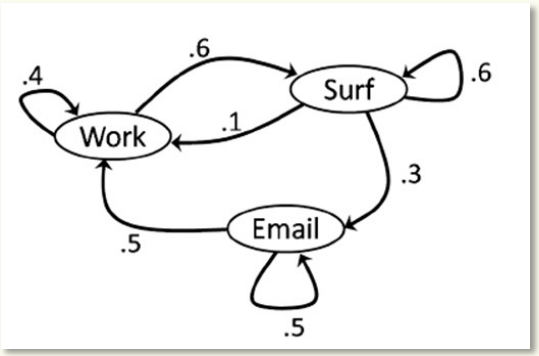
Write $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ Then for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

So $\boldsymbol{q}^{(1)} = \boldsymbol{q}^{(0)}\boldsymbol{M}$

$\boldsymbol{q}^{(2)} =$

…

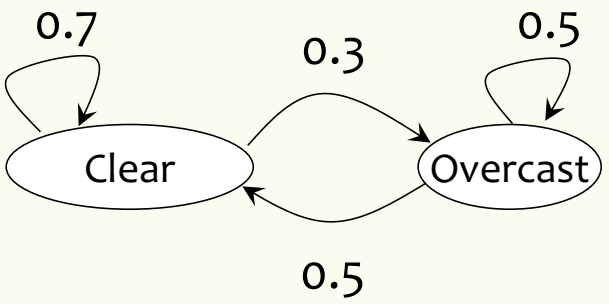
By induction … we can derive



$$\boldsymbol{M} = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$$\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^{t} \text{ for all } t \geq 0$$

Another example:



Suppose that $\boldsymbol{q}^{(0)} = \left[q_C^{(0)}, q_O^{(0)}\right] = [0,1]$

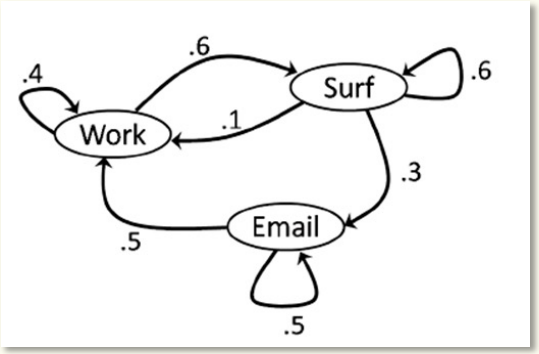
We have $\boldsymbol{M} = \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix}$

Poll: **www.slido.com/1692973**

What is $\boldsymbol{q}^{(2)}$?

a. $[0.3, 0.7]$
b. $[0.6, 0.4]$
c. $[0.7, 0.3]$
d. $[0.5, 0.5]$
e. $[0.4, 0.6]$

Many interesting questions about Markov Chains



Given: In state Work at time $t = 0$

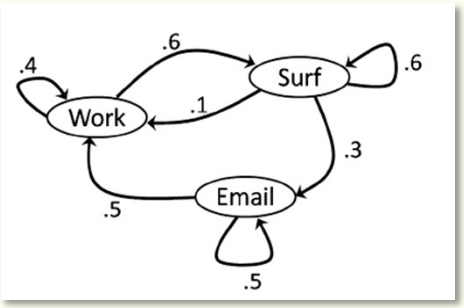
1. **What is the probability that I am in state s at time 1?**
2. **What is the probability that I am in state s at time 2?**
3. **What is the probability that I am in state s at some time t far in the future?**

$$\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t \text{ for all } t \geq 0$$

What does $\boldsymbol{M}^t$ look like for really big t ?

M^t as t grows

$$q^{(t)} = q^{(0)} M^t \text{ for all } t \geq 0$$



M

$$\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

M^2

$$\begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .22 & .6 & .18 \\ .25 & .42 & .33 \\ .45 & .3 & .25 \end{pmatrix} \end{array}$$

M^3

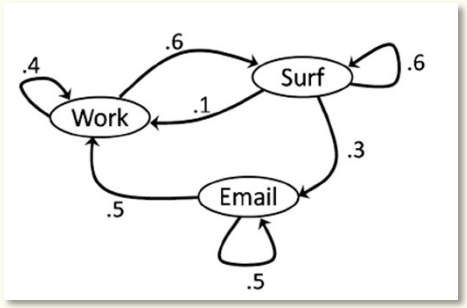
$$\begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .238 & .492 & .270 \\ .307 & .402 & .291 \\ .335 & .450 & .215 \end{pmatrix} \end{array}$$

M^{10}

$$\begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .2940 & .4413 & .2648 \\ .2942 & .4411 & .2648 \\ .2942 & .4413 & .2648 \end{pmatrix} \end{array}$$

M^t as t grows

$$q^{(t)} = q^{(0)} M^t \text{ for all } t \geq 0$$



M

$$\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

M^2

	W	S	E
W	.22	.6	.18
S	.25	.42	.33
E	.45	.3	.25

M^3

	W	S	E
W	.238	.492	.270
S	.307	.402	.291
E	.335	.450	.215

M^{10}

	W	S	E
W	.2940	.4413	.2648
S	.2942	.4411	.2648
E	.2942	.4413	.2648

M^{30}

	W	S	E
W	.29411764705	.44117647059	.26470588235
S	.29411764706	.44117647058	.26470588235
E	.29411764706	.44117647059	.26470588235

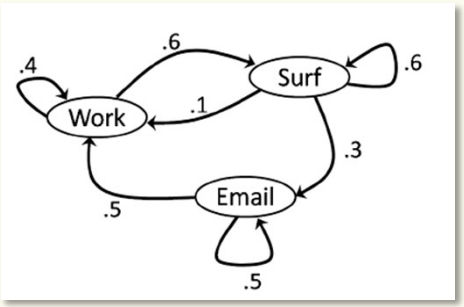
M^{60}

	W	S	E
W	.294117647058823	.441176470588235	.264705882352941
S	.294117647068823	.441176470588235	.264705882352941
E	.294117647068823	.441176470588235	.264705882352941

What does this say about $q^{(t)}$?

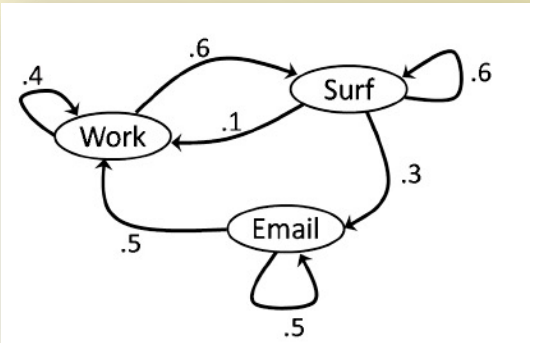
M^t as t grows

$$q^{(60)} = q^{(0)} M^{60}$$



$$[q_W^{(0)}, q_S^{(0)}, q_E^{(0)}] \cdot \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array} = [q_W^{(60)}, q_S^{(60)}, q_E^{(60)}]$$

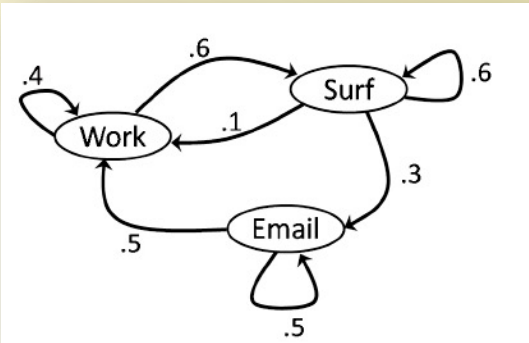
Observation



If $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$ then it will never change again!

Since for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

Observation



If $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$ then it will never change again!

Since for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

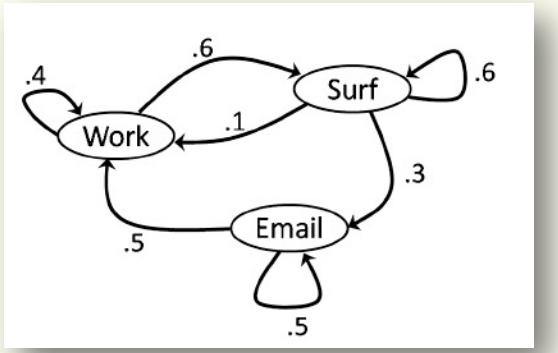
Called a **stationary distribution** and has a special name

$$\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$$

Solution to $\boldsymbol{\pi} = \boldsymbol{\pi}\,\boldsymbol{M}$

Solving for Stationary Distribution

$$\boldsymbol{M} = \begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$



Stationary Distribution satisfies

- $\boldsymbol{\pi} = \boldsymbol{\pi M}$, where $\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$
- $\pi_W + \pi_S + \pi_E = 1$

➔ $\pi_W = \frac{10}{34}$, $\pi_S = \frac{15}{34}$, $\pi_E = \frac{9}{34}$

As $t \to \infty$, $\boldsymbol{q}^{(t)} \to \boldsymbol{\pi}$ no matter what distribution $\boldsymbol{q}^{(0)}$ is !!

Markov Chains in general

- A set of n **states** $\{1, 2, 3, \ldots\ n\}$
- The state at time t is denoted by $X^{(t)}$
- A **transition matrix** $\boldsymbol{M}$, dimension $n \times n$

$$\boldsymbol{M}_{ij} = P\left(X^{(t+1)} = j \;\middle|\; X^{(t)} = i\right)$$

- $\boldsymbol{q}^{(t)} = (q_1^{(t)}, q_2^{(t)}, \ldots, q_n^{(t)})$ where $q_i^{(t)} = P(X^{(t)} = i)$
- Transition: LTP $\Rightarrow \boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\,\boldsymbol{M}$ so $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\,\boldsymbol{M}^t$

Stationary Distribution of a Markov Chain

Definition. The **stationary distribution of a Markov Chain** with n states is the n-dimensional row vector $\boldsymbol{\pi}$ (which must be a probability distribution; that is, it must be nonnegative and sum to 1) such that

$$\boldsymbol{\pi M = \pi}$$

Intuition: Distribution over states at next step is the same as the distribution over states at the current step

Fundamental Theorem of Markov Chains

Recall $\boldsymbol{q}^{(t)}$ is the distribution of being at each state at time t computed by $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t$. As t gets large $\boldsymbol{q}^{(t)} \approx \boldsymbol{q}^{(t+1)}$.

Fundamental Theorem of Markov Chains : For a Markov Chain that is aperiodic* and irreducible*, with transition probabilities $\boldsymbol{M}$ and for any starting distribution $\boldsymbol{q}^{(0)}$ over the states

$$\lim_{t\to\infty} \boldsymbol{q}^{(0)}\boldsymbol{M}^t = \boldsymbol{\pi}$$

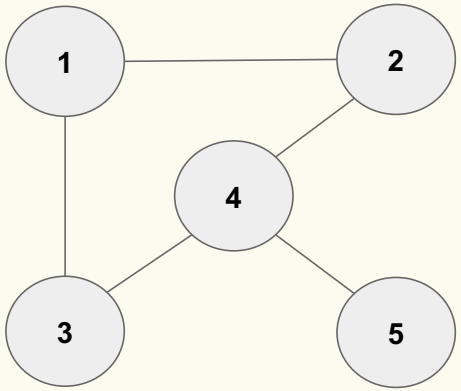
where $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{M}$ (i.e., $\boldsymbol{\pi M} = \boldsymbol{\pi}$)

**These concepts are beyond us but they turn out to cover a very large class of Markov chains of practical importance.*

Another Example: Random Walks

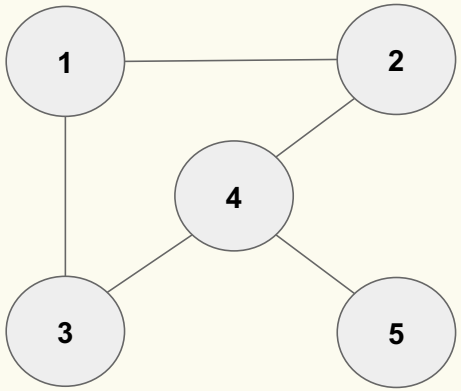
Suppose we start at node 1, and at each step transition to a neighboring node with equal probability.

This is called a "random walk" on this graph.



Example: Random Walks on an Undirected Graph

Start by defining transition probs.



	To **1**	To **2**	To **3**	To **4**	To **5**
From **1**					
From **2**					
From **3**					
From **4**					
From **5**					

$$\boldsymbol{M}_{ij} = P\left(X^{(t+1)} = j \mid X^{(t)} = i\right)$$

Brain Break

Markov Chain Monte Carlo

- Technique for sampling from high dimensional distributions
- Computational method for studying large, very complex sets.
- In some cases, a technique for getting good approximate solutions to complex optimization problems
- Used in every field of science and engineering
- "*To someone working in my part of the world, asking about applications of MCMC is like asking about applications of the quadratic formula. The results are really used in every aspect of scientific inquiry*" --- Persi Diaconis, Stanford

MCMC

- Idea: simulate a random walk that moves among possible configurations of a system and will converge to a useful distribution over these configurations.
- Example: Knapsack Problem (NP-complete)
  - Input: collection of n items, for each item
    - Value
    - Weight
  - Goal: output subset S of items of maximum total value, that has total weight < W.

MCMC for knapsack

- Define a Markov chain with states being possible solutions and transition probabilities that have higher probabilities on "good solutions"
- Simulate the Markov chain for many iterations until reach a "good" state.

MCMC for Knapsack Problem

Algorithm 1 MCMC for 0-1 Knapsack Problem

subset $\leftarrow$ vector of n zeros (indexed by 0 to $n-1$), where subset is always a binary vector in $\{0,1\}^n$ that represents whether or not we have each item. (This means that we initially start with an empty knapsack).
best_subset $\leftarrow$ subset
for $t = 1, \ldots,$ NUM_ITER **do**
    $k \leftarrow$ a uniformly random integer in $\{0, 1, \ldots, n-1\}$.
    new_subset $\leftarrow$ subset but with subset$[k]$ flipped ($0 \rightarrow 1$ or $1 \rightarrow 0$).
    $\Delta \leftarrow$ value(new_subset) $-$ value(subset)
    if new_subset satisfies weight constraint (total weight $\leqslant W$) **then**
        if $\Delta > 0$ OR ($T > 0$ AND Unif$(0,1) < e^{\Delta/T}$) **then**
            subset $\leftarrow$ new_subset
    if value(subset) $>$ value(best_subset) **then**
        best_subset $\leftarrow$ subset