CSE 312

Foundations of Computing II

Lecture 24: Wrap up discussion of estimators, Markov chains

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Anna R. Karlin

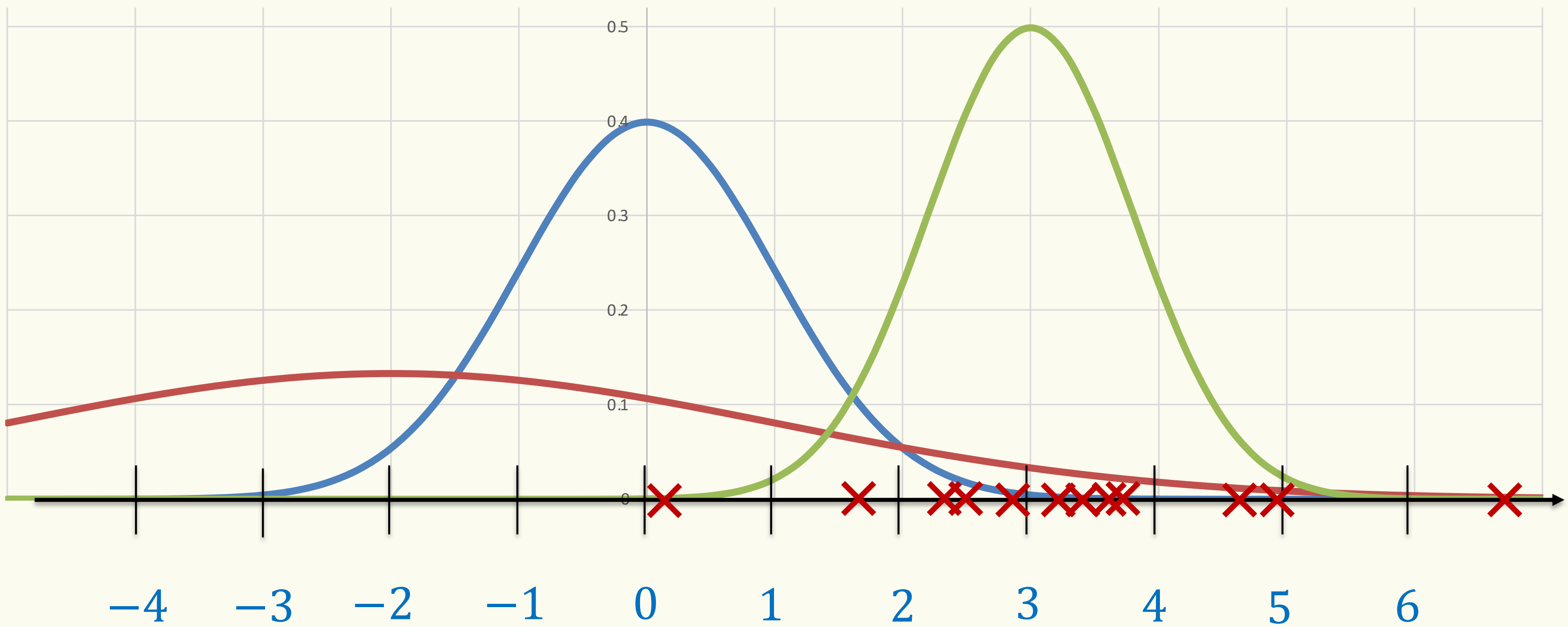
Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Ryan O'Donnell, Alex Tsun, Rachel Lin, Hunter Schafer & myself ☺

MLE Recipe

1. **Input** Given n iid 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 $\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \Pr(x_i ; \theta)$
   - 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.

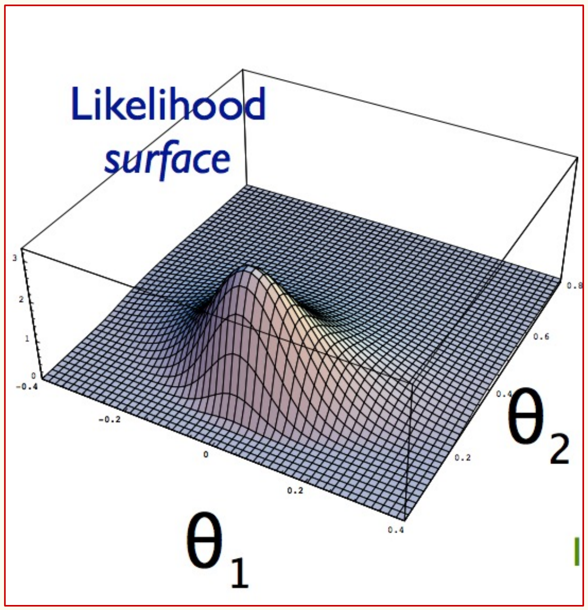
n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$. Most likely μ and σ^2?



Two-parameter optimization

Normal outcomes $x_1, \dots, x_n$

Goal: estimate $\theta_1 = \mu$ = expectation and $\theta_2 = \sigma^2$ = variance



$$L(x_1, \dots, x_n | \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - \theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) =$$

$$= -n \frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

Two-parameter estimation

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) = -n \frac{\ln(2\pi\, \theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

We need to find a solution $\hat{\theta}_1, \hat{\theta}_2$ to

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \dots, x_n | \theta_1, \theta_2) = 0$$
$$\frac{\partial}{\partial \theta_2} \ln L(x_1, \dots, x_n | \theta_1, \theta_2) = 0$$

MLE estimates for mean and variance.

Normal outcomes $x_1, \ldots, x_n$

$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$

MLE estimator for **expectation**

$$\hat{\theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_\mu\right)^2$$

MLE estimator for **variance**

MLE Recipe

1. **Input** Given n iid samples $x_1, \ldots, x_n$ from parametric model with multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)$
2. **Likelihood** Define your likelihood function $\mathcal{L}(x_1, \ldots, x_n \mid \boldsymbol{\theta})$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(x_i\,;\boldsymbol{\theta})$
   - For continuous $\mathcal{L}(x_1, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i\,;\,\boldsymbol{\theta})$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots, x_n|\boldsymbol{\theta})$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta_i} \ln \mathcal{L}(x_1, \ldots, x_n \mid \boldsymbol{\theta})$ for each i
5. **Solve for $\widehat{\boldsymbol{\theta}}$** by setting derivatives to 0 and solving system of equations.

Generally, you need to verify that you've found a maximum, but we won't ask you to do that in CSE 312.

Agenda

- Properties of estimators
- Markov chains

When is an estimator good?

Distribution $p(x|\theta)$ → samples $X_1, \dots, X_n$ from $p(x|\theta)$ → **Algorithm** → $\hat{\theta}_n$

Parameter estimate "The model"

θ = unknown parameter

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

$$\mathbb{E}(\hat{\theta}_n) = \theta.$$

Example – Consistency

Normal outcomes $x_1, \dots, x_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

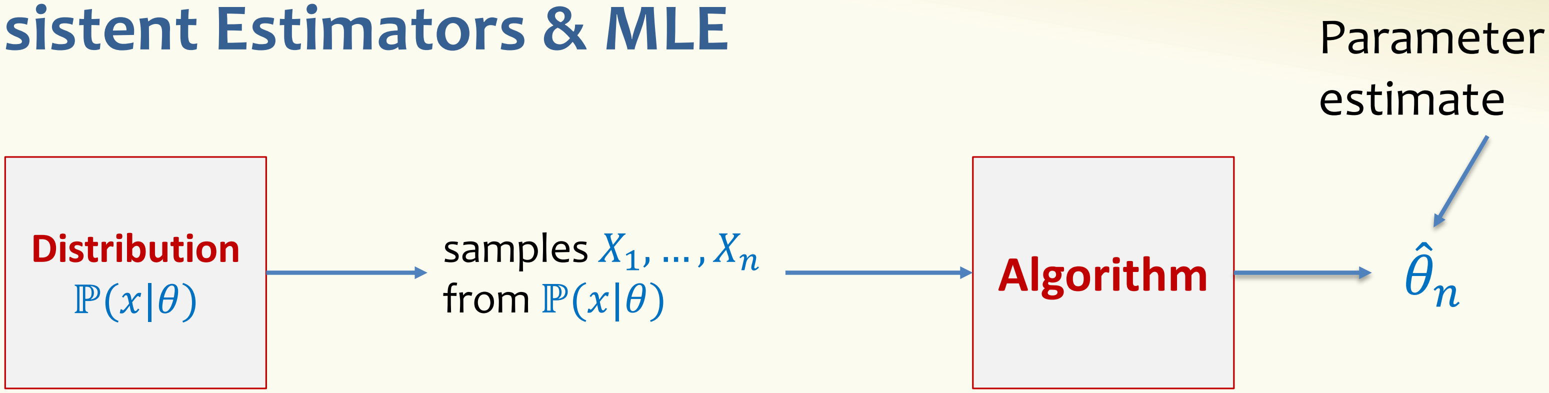
$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$

<u>Unbiased</u>

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

<u>Biased!</u>

Consistent Estimators & MLE



θ = <u>unknown</u> parameter

Definition. An estimator is **unbiased** if $\mathbb{E}\left(\hat{\theta}_n\right) = \theta$ for all $n \geq 1$.

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

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

$\widehat{\Theta}_{\sigma^2}$ is biased, but consistent.

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

$$\mathbb{E}\left(\widehat{\Theta}_{\sigma^2}\right) \overset{\text{linearity}}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_1\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]$$

$$\ldots$$

$$= \left(1 - \frac{1}{n}\right)\sigma^2 \quad = \frac{n-1}{n}\sigma^2$$

$\widehat{\Theta}_{\sigma^2}$ converges to σ^2, as $n \to \infty$.

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

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

Sample variance – Unbiased!

Agenda

- Properties of estimators
- Markov chains

So far, a single-shot random process

Random Process → **Outcome Distribution** D

So far, a single-shot random process

Random Process → **Outcome Distribution** D

Many-step random process

Random Process 1 → **Outcome Distribution** D_1 → Random Process 2 → **Outcome Distribution** D_2 → Random Process 3 → **Outcome Distribution** D_3 …

So far, a single-shot random process

Random Process → **Outcome Distribution** D

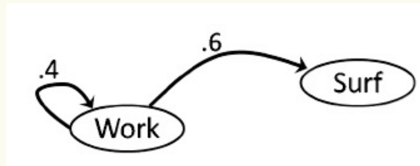
Today:
see a very special type of DTSP
Called a **Markov Chain**

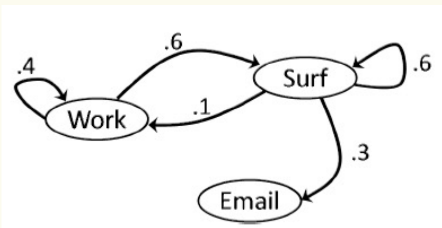
Many-step random process

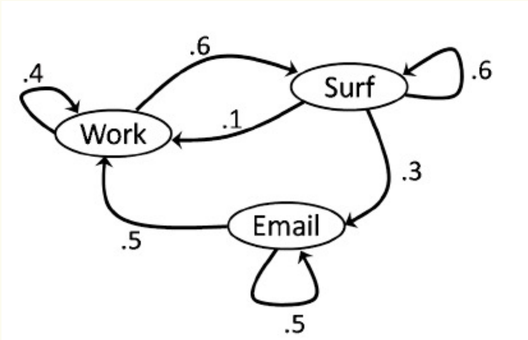
Random Process 1 → **Outcome Distribution** D_1 → Random Process 2 → **Outcome Distribution** D_2 → Random Process 3 → **Outcome Distribution** D_3 …

Definition: A **discrete-time stochastic process** (DTSP) is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$ where $X^{(t)}$ is the value at time t.

A day in my life

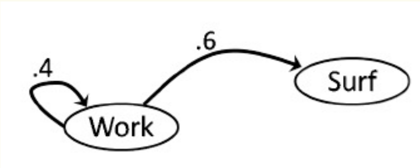


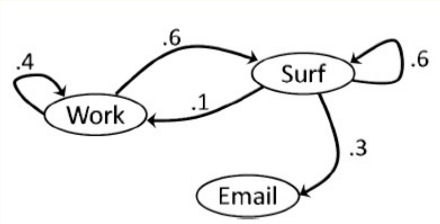


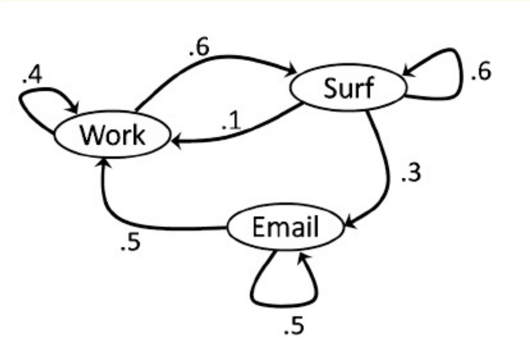


A day in my life

t = 0







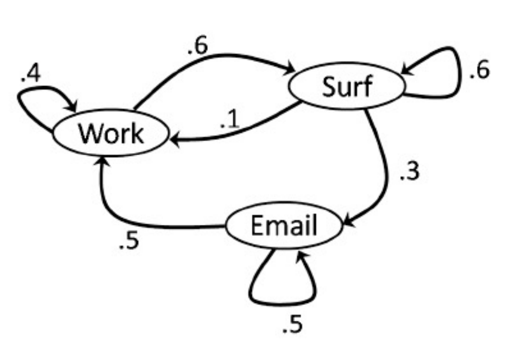
This type of probabilistic finite automaton is called a **Markov Chain**

The next state depends only on the current state and not on the history

For ANY $t \geq 0$,
if I was working at time t, then at t+1
with probability 0.4 I continue working
with probability 0.6, I switch to surfing, and
with probability 0, I switch to emailing

This is called History Independent (similar to memoryless)

A day in my life



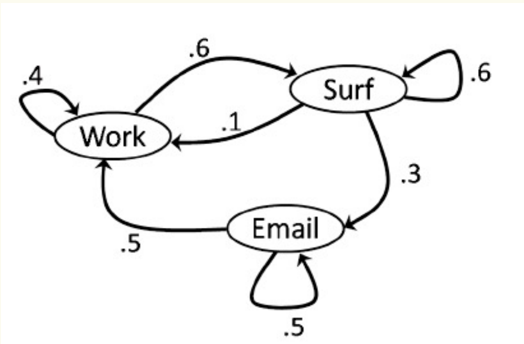
Many interesting questions.

1. What is the probability that I work at time 1?
2. What is the probability that I work at time 2?

$X^{(t)}$ state I'm in at time t (random variable)

t	0	1	2
$q_W^{(t)} = \Pr(X^{(t)} = \text{work})$			
$q_S^{(t)} = \Pr(X^{(t)} = \text{surf})$			
$q_E^{(t)} = \Pr(X^{(t)} = \text{email})$			

A day in my life

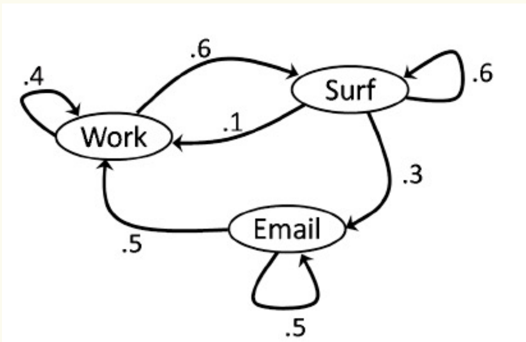


Many interesting questions

1. What is the probability that I work at time 1?
2. What is the probability that I work at time 2?
3. What is the probability that I work at time t=100?
4. What is the probability that I'm working at some random time far in the future?

A day in my life

What is the probability I'm in each state at time t, as a function of the probability distribution over states at time t-1



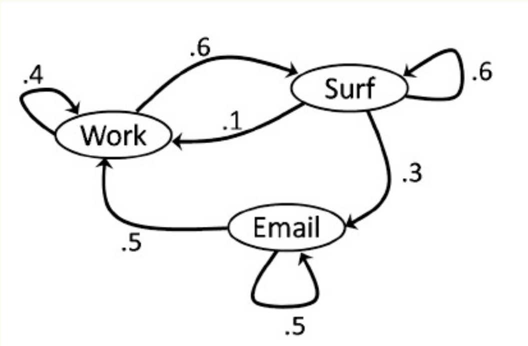
$X^{(t)}$ state I'm in at time t (random variable)

t-1	t
$q_W^{(t-1)}$ = Pr($X^{(t-1)}$= work)	$q_W^{(t)}$ =
$q_S^{(t-1)}$ = Pr($X^{(t-1)}$= surf)	$q_S^{(t)}$ =
$q_E^{(t-1)}$ = Pr($X^{(t-1)}$= email)	$q_E^{(t)}$ =

$$\left(q_W^{(t)},\quad q_S^{(t)},\quad q_E^{(t)}\right) = \left(q_W^{(t-1)},\quad q_S^{(t-1)},\quad q_E^{(t-1)}\right)\begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$

Transition Probability Matrix

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



➔ $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t-1)}\,\boldsymbol{P}$

$$\boldsymbol{q}^{(t)} = \left(q_W^{(t)},\quad q_S^{(t)},\quad q_E^{(t)}\right)$$

Apply $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t-1)}\ \boldsymbol{P}$ inductively.

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

$\Rightarrow \boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\ \boldsymbol{P}^{t}$

The t-step walk $\boldsymbol{P}^t$

Recall $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\ \boldsymbol{P}^t$

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

$$P^2 = \begin{array}{cc} & \begin{matrix} W & S & E \end{matrix} \\ \begin{matrix} W \\ S \\ E \end{matrix} & \begin{pmatrix} .22 & .6 & .18 \\ .25 & .42 & .33 \\ .45 & .3 & .25 \end{pmatrix} \end{array}$$

$$P^3 = \begin{array}{cc} & \begin{matrix} W & S & E \end{matrix} \\ \begin{matrix} W \\ S \\ E \end{matrix} & \begin{pmatrix} .238 & .492 & .270 \\ .307 & .402 & .291 \\ .335 & .450 & .215 \end{pmatrix} \end{array}$$

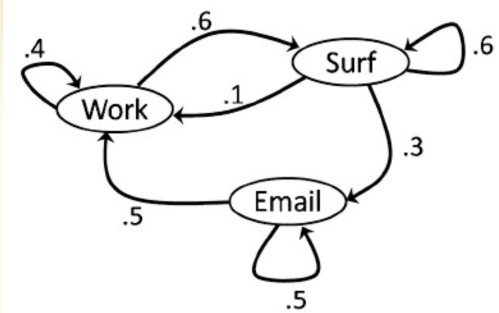
$$P^{10} \approx \begin{array}{cc} & \begin{matrix} W & S & E \end{matrix} \\ \begin{matrix} W \\ S \\ E \end{matrix} & \begin{pmatrix} .2940 & .4413 & .2648 \\ .2942 & .4411 & .2648 \\ .2942 & .4413 & .2648 \end{pmatrix} \end{array}$$

$$P^{30} \approx \begin{array}{cc} & \begin{matrix} W & S & E \end{matrix} \\ \begin{matrix} W \\ S \\ E \end{matrix} & \begin{pmatrix} .29411764705 & .44117647059 & .26470588235 \\ .29411764706 & .44117647058 & .26470588235 \\ .29411764706 & .44117647059 & .26470588235 \end{pmatrix} \end{array}$$

$$P^{60} \approx \begin{array}{cc} & \begin{matrix} W & S & E \end{matrix} \\ \begin{matrix} W \\ S \\ E \end{matrix} & \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

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

Observation



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

Called a "stationary distribution" and has a special name

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

Solution to $\boldsymbol{\pi} = \boldsymbol{\pi} \boldsymbol{P}$