Dimensionality reduction (cont.)

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

April 25th, 2007

Lower dimensional projections

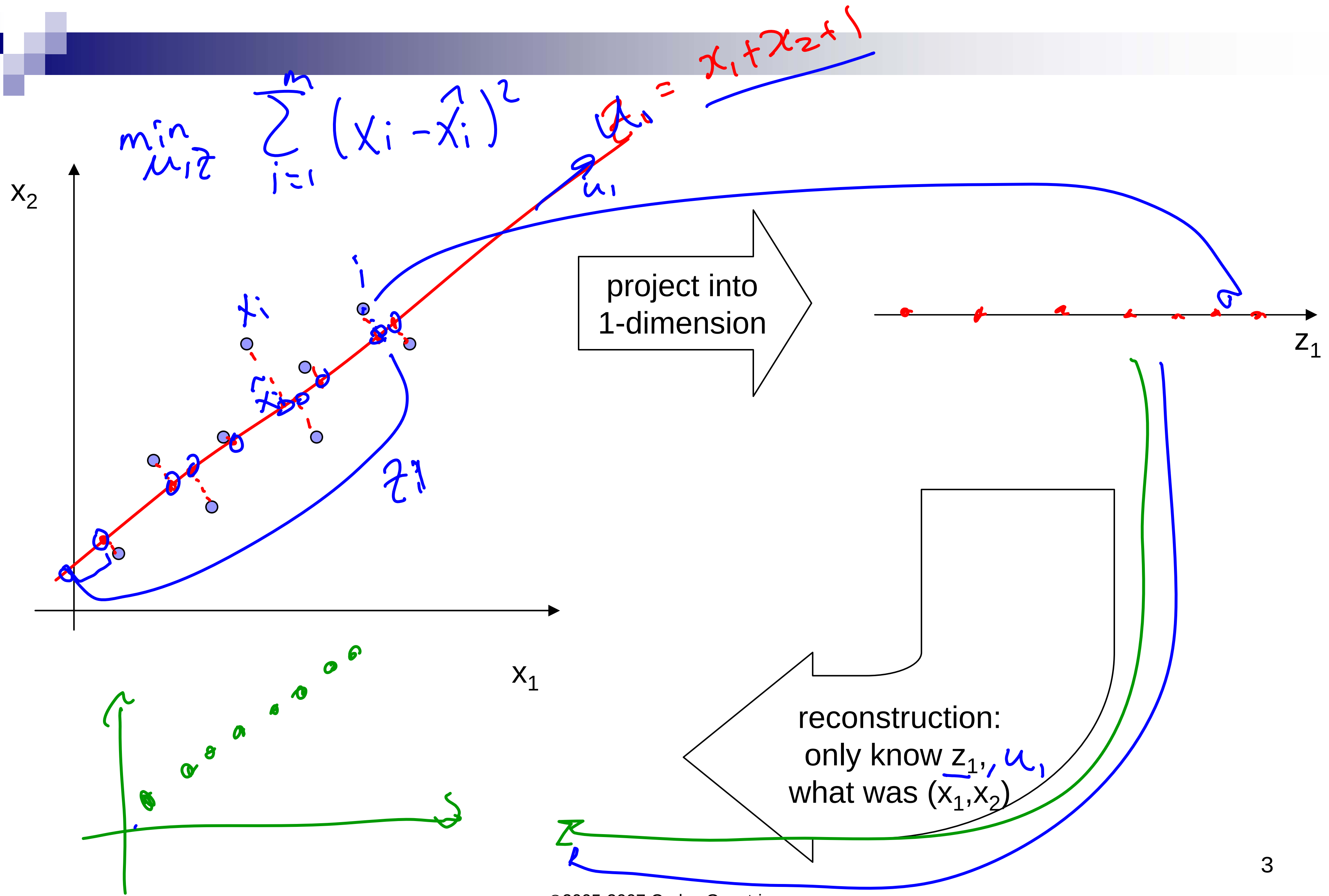
- Rather than picking a subset of the features, we can new features that are combinations of existing features

x_1, x_2, x_3
$x_2^2, x_1 x_3^3 \ldots$
Select some of these

projection:
new feature, e.g.,
$\boxed{X_{new}} = 0.5 X_1 - 0.75 X_2 + 0.92 X_3 \ldots$
new feature
$X_{new} = g(X)$

- Let's see this in the unsupervised setting
  - just **X**, but no Y

Linear projection and reconstruction

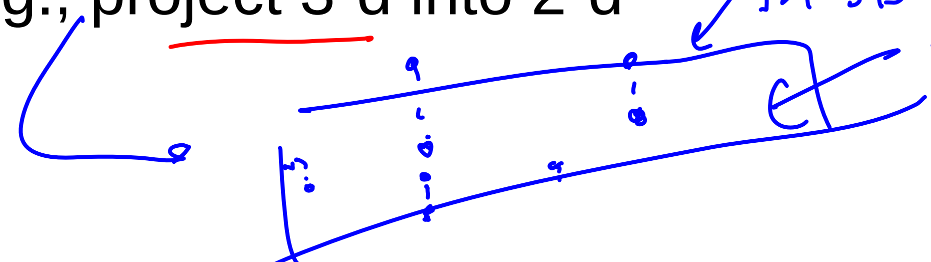


Principal component analysis – basic idea

- Project n-dimensional data into k-dimensional space while preserving information:
  - e.g., project space of 10000 words into 3-dimensions
  - e.g., project 3-d into 2-d

in 3D points

2D plane

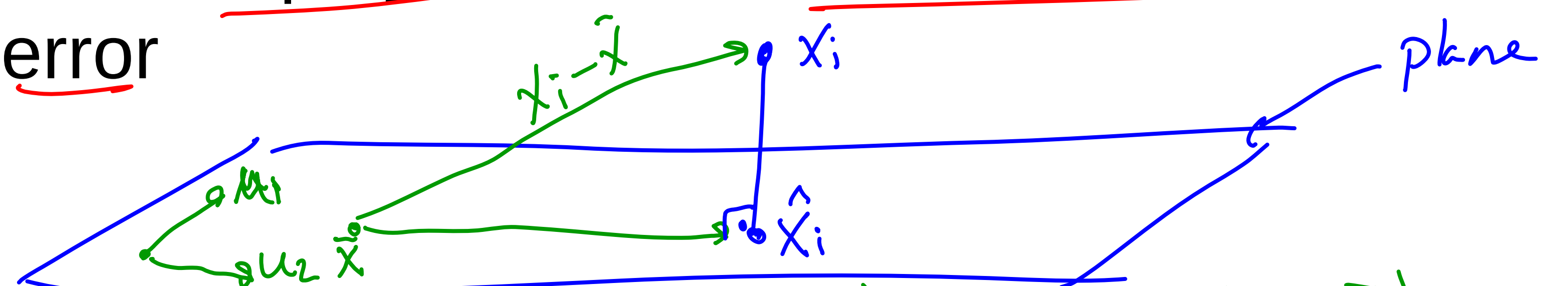


$||u_i|| = 1$

$u_1 \perp u_2$: $u_1 \cdot u_2 = 0$

orthonormal basis

- Choose projection with minimum reconstruction error



$$\hat{x}_i = \bar{x} + \underbrace{(x_i - \bar{x})\cdot u_1}_{z_1^i}\, u_1 + \underbrace{(x_i - \bar{x})\cdot u_2}_{z_2^i}\, u_2$$

Linear projections, a review

- Project a point into a (lower dimensional) space:
  - **point**: $\mathbf{x} = (x_1,\ldots,x_n)$ n-Dim
  - **select a basis** – set of basis vectors – $(\mathbf{u}_1,\ldots,\mathbf{u}_k)$ k-dim
    - we consider orthonormal basis:
      - $\mathbf{u}_i \cdot \mathbf{u}_i=1$, and $\mathbf{u}_i \cdot \mathbf{u}_j=0$ for $i\neq j$
  - **select a center** – $\bar{\mathbf{x}}$, defines offset of space
  - **best coordinates** in lower dimensional space defined by dot-products: $(z_1,\ldots,z_k)$, $z_i = (\mathbf{x}-\bar{\mathbf{x}}) \cdot \mathbf{u}_i$ $\quad z_i = (x-\bar{x})\cdot u_i$
    - minimum squared error

for any x: given $u_1 \ldots u_k$

$$\min_{z} \left(x - \left(\bar{x} + \sum_{i=1}^{k} z_i \cdot u_i\right)\right)^2$$

PCA finds projection that minimizes reconstruction error

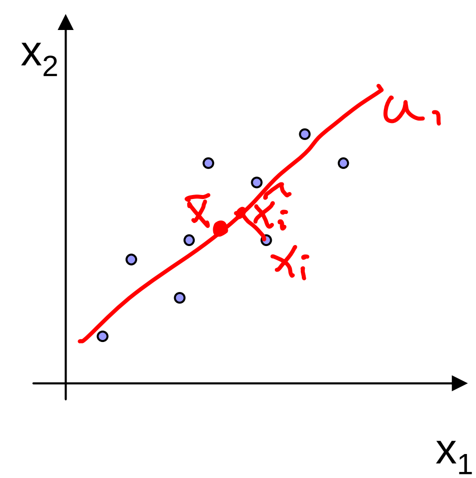
n-D

- Given m data points: $\mathbf{x}^i = \overbrace{(x_1^i,\ldots,x_n^i)}^{n\text{-D}}$, i=1…m
- Will represent each point as a projection:
  - $\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$ where: $\underbrace{\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}^i}_{\text{mean over points}}$ and $\boxed{z_j^i = (\mathbf{x}^i - \bar{\mathbf{x}}) \cdot \mathbf{u}_j}$

3 40 000

- PCA:
  - Given k≤n, find ($\mathbf{u}_1,\ldots,\mathbf{u}_k$) minimizing reconstruction error:

$$\boxed{error_k = \sum_{i=1}^{m} (\mathbf{x}^i - \hat{\mathbf{x}}^i)^2}$$



Understanding the reconstruction error

n-dim. space can be represented exactly by n-basis

- Note that $\mathbf{x}^i$ can be represented exactly by n-dimensional projection:

$$\mathbf{x}^i = \bar{\mathbf{x}} + \sum_{j=1}^{n} z_j^i \mathbf{u}_j$$

- Rewriting error:

$$error_k = \sum_{i=1}^{m} (x^i - \hat{x}^i)^2 = \sum_{i=1}^{m} \left[\bar{x} + \sum_{j=1}^{n} z_j^i u_j - \left(\bar{x} + \sum_{j=1}^{k} z_j^i u_j \right) \right]^2 = \sum_{i=1}^{m} \left(\sum_{j=k+1}^{n} z_j^i u_j \right)^2$$

$$= \sum_{i=1}^{m} \left[\sum_{j=k+1}^{n} z_j^i \underbrace{u_j \cdot u_j}_{1} z_j^i + \sum_{r=k+1}^{n} \sum_{\substack{s=k+1 \\ s \neq r}}^{n} z_s^i \cancel{u_s \cdot u_r}^{\,0} z_r^i \right]$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} (z_j^i)^2$$

basis is orthonormal

$u_j \cdot u_j = 1$

$u_s \cdot u_r = 0$ $s \neq r$

$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$$

$$z_j^i = (\mathbf{x}^i - \bar{\mathbf{x}}) \cdot \mathbf{u}_j$$

- Given $k \leq n$, find $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ minimizing reconstruction error:

$$error_k = \sum_{i=1}^{m} (\mathbf{x}^i - \hat{\mathbf{x}}^i)^2$$

Reconstruction error and covariance matrix

keep ignore

$$error_k = \sum_{i=1}^{m} \sum_{j=k+1}^{n} [\overbrace{\mathbf{u}_j \cdot (\mathbf{x}^i - \bar{\mathbf{x}})}^{z_j^i}]^2 \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^i - \bar{\mathbf{x}})(\mathbf{x}^i - \bar{\mathbf{x}})^T$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} [u_j \cdot (x^i - \bar{x})][(x^i - \bar{x}) \cdot u_j]$$

$$= \sum_{j=k+1}^{n} u_j^T \Big[\underbrace{\sum_{i=1}^{m} (x^i - \bar{x})(x^i - \bar{x})^T} \Big] u_j$$

$$= m \sum_{j=k+1}^{n} u_j^T \, \Sigma \, u_j = error_k$$

Minimizing reconstruction error and eigen vectors

if $u_1 \ldots u_n$ are eigen vectors of Σ, then $u_j^T \Sigma u_j = \lambda_j \leftarrow j$'th eigen value

- Minimizing reconstruction error equivalent to picking orthonormal basis ($\mathbf{u}_1,\ldots,\mathbf{u}_n$) minimizing:

$$error_k = m \sum_{j=k+1}^{n} \mathbf{u}_j^T \Sigma \mathbf{u}_j$$

- Eigen vector:

u: $\Sigma u = \lambda u$ (λ ← eigen value)

multiply both sides by u →

$$u^T \Sigma u = \lambda \underbrace{u^T u}_{1} = \lambda$$

- Minimizing reconstruction error equivalent to picking ($\mathbf{u}_{k+1},\ldots,\mathbf{u}_n$) to be eigen vectors with smallest eigen values

$$error_k = \sum_{j=k+1}^{n} \lambda_j$$

Basic PCA algoritm

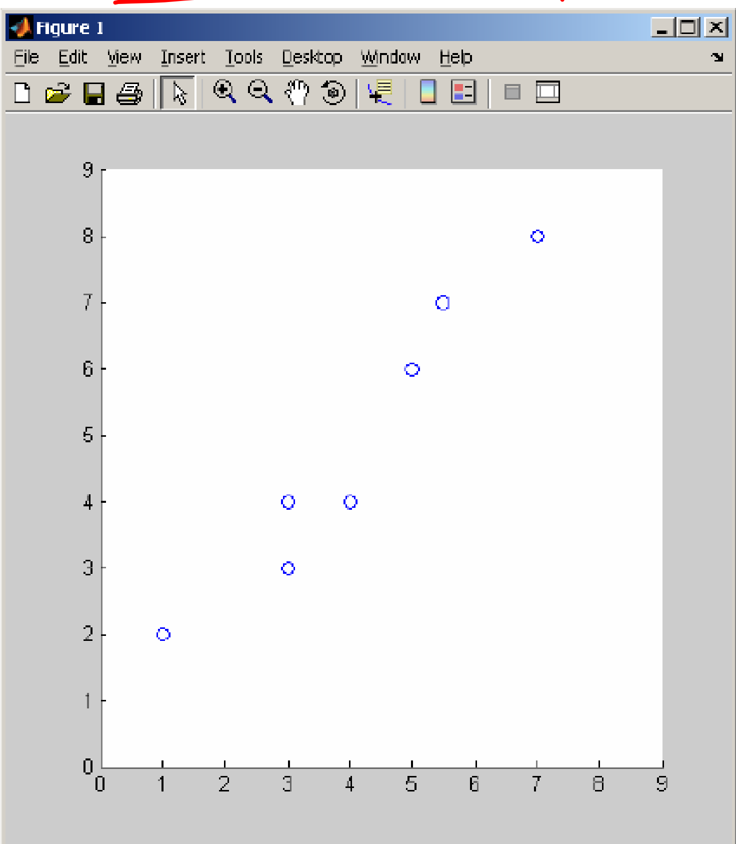
X = m data points, n features, x^i

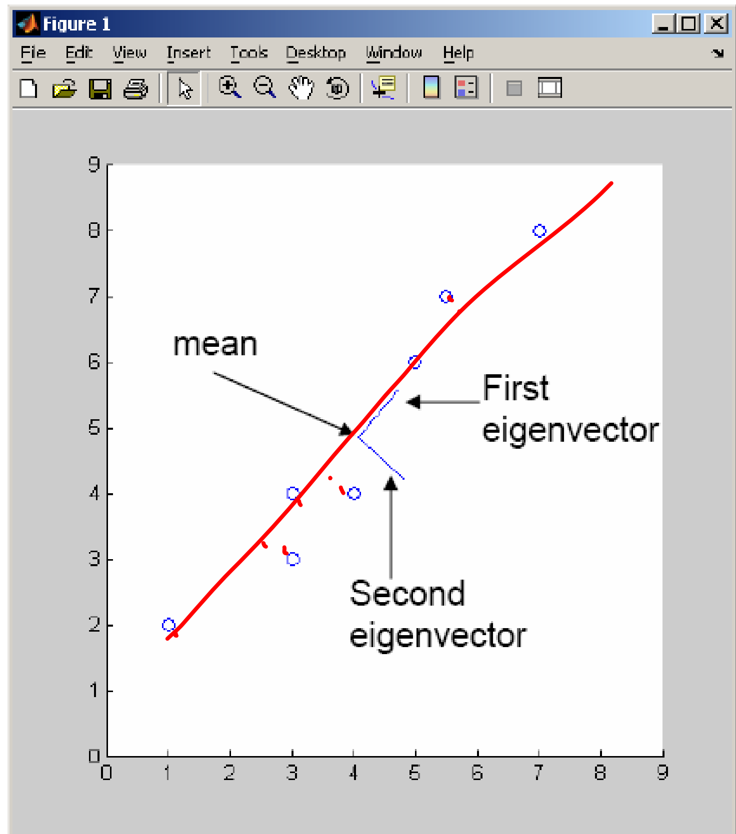
- Start from m by n data matrix **X**
- **Recenter**: subtract mean from each row of **X**
  - $\mathbf{X_c} \leftarrow \mathbf{X} - \overline{\mathbf{X}}$
- **Compute covariance matrix**:
  - ${}_n\Sigma^n \leftarrow 1/m\ \mathbf{X_c}^T\ \mathbf{X_c}$
- Find **eigen vectors and values** of Σ
- **Principal components:** k eigen vectors with highest eigen values

eigs in Matlab ... :)

PCA example

$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$$





PCA example – reconstruction

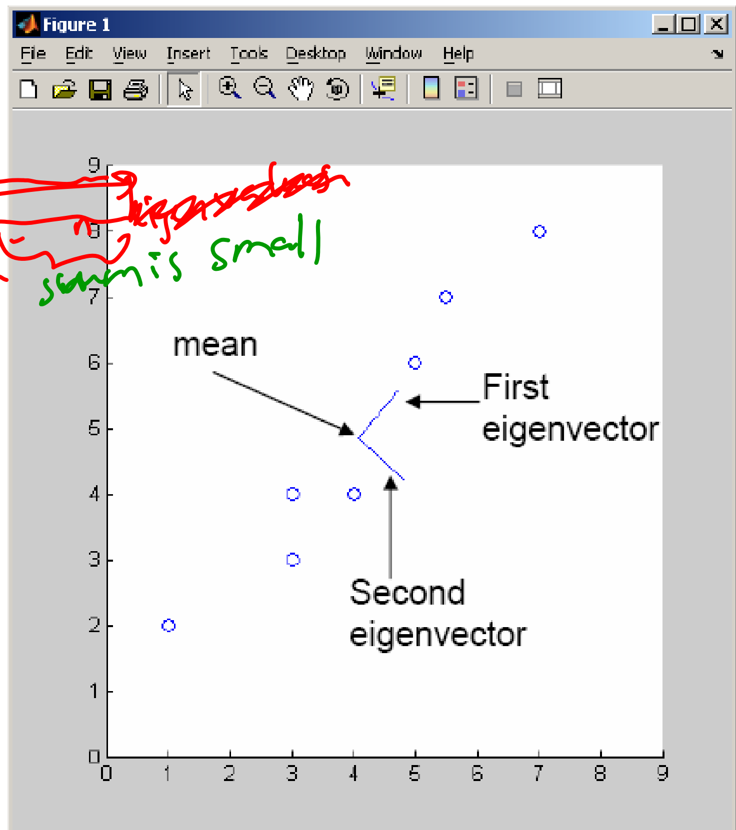
Simple way to pick (slow)

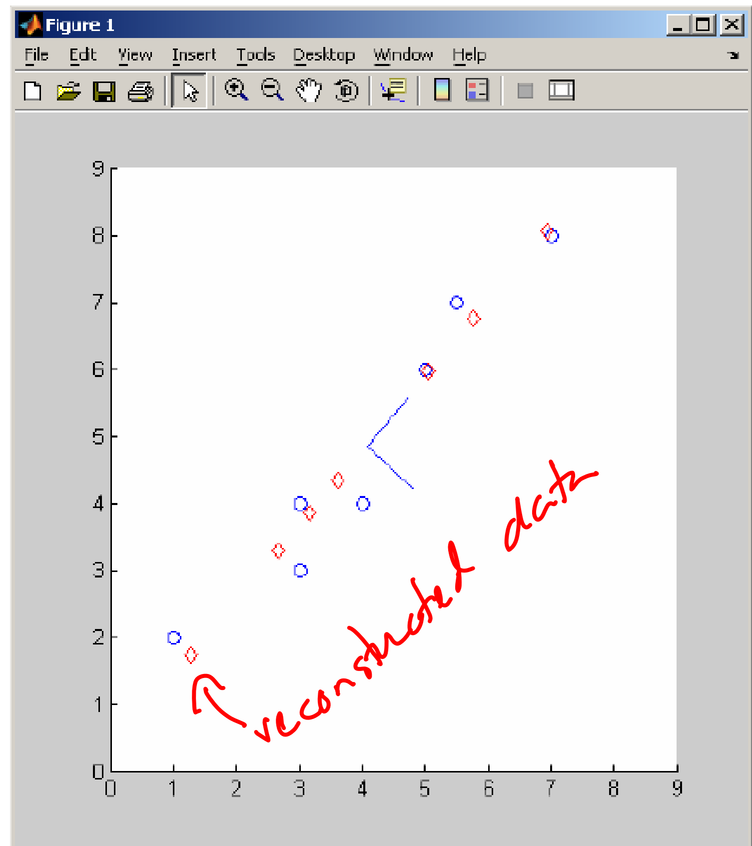
$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$$

only used first principal component

eigen values

sum is small



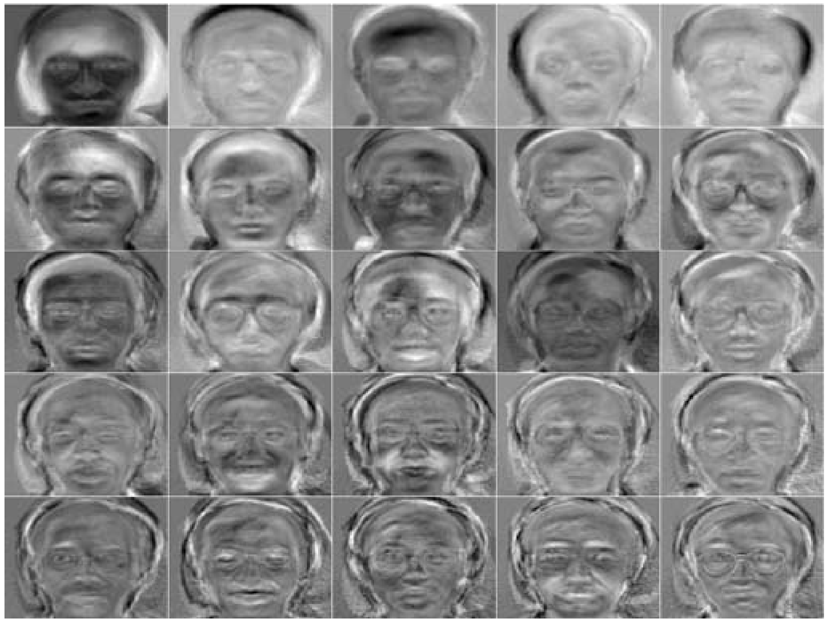


reconstructed data

Eigenfaces [Turk, Pentland '91]

- Input images:

- Principal components:



x^i = (pixels)

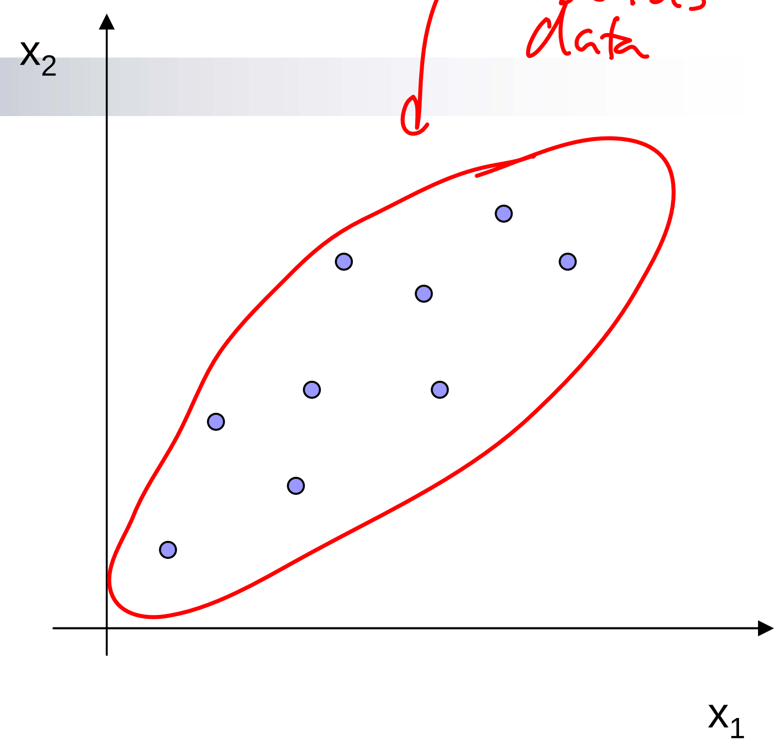
Eigenfaces reconstruction

- Each image corresponds to adding 8 principal components:

8 numbers describe face

Relationship to Gaussians

Gaussian describes data



- PCA assumes data is Gaussian
  - $\mathbf{x} \sim N(\bar{\mathbf{x}};\Sigma)$
- Equivalent to weighted sum of simple Gaussians:

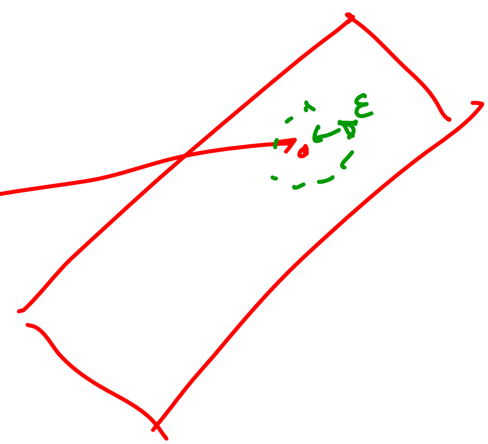
Gaussian R.V.

$$\mathbf{x} = \bar{\mathbf{x}} + \sum_{j=1}^{n} z_j \mathbf{u}_j; \quad z_j \sim N(0;\sigma_j^2)$$

scalar

- Selecting top k principal components equivalent to lower dimensional Gaussian approximation:

$$\mathbf{x} \approx \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j \mathbf{u}_j + \varepsilon; \quad z_j \sim N(0;\sigma_j^2)$$

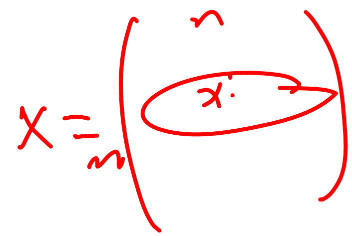


  - $\varepsilon \sim N(0;\sigma^2)$, where σ^2 is defined by error_k

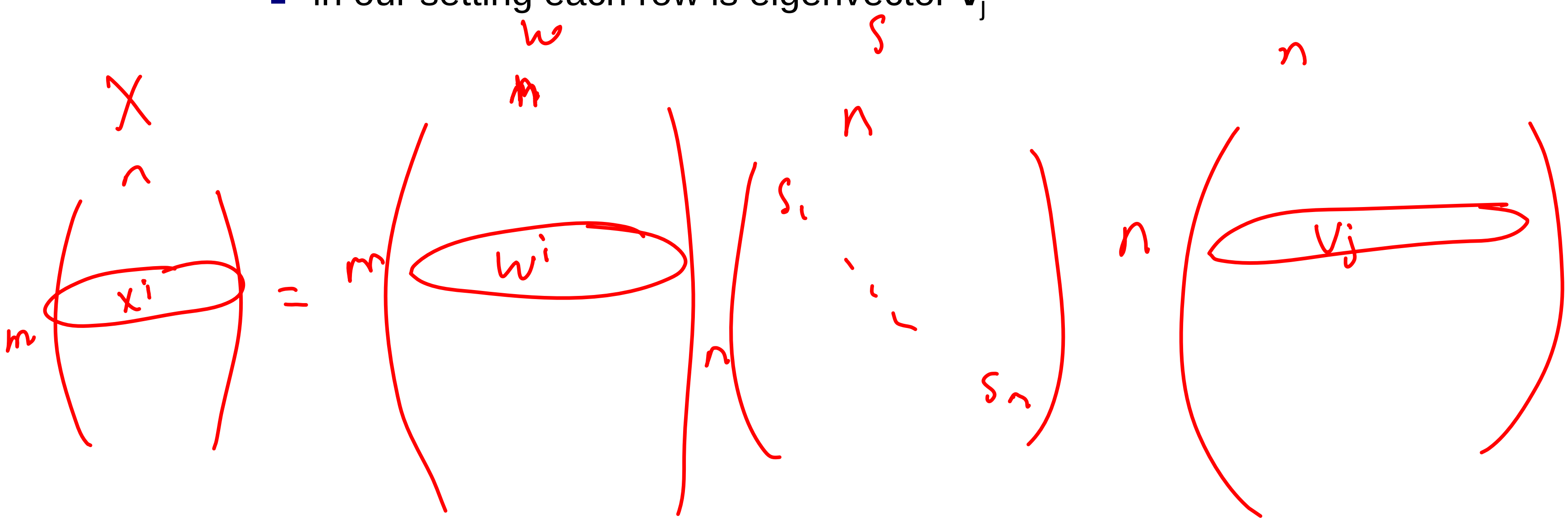
Scaling up

- Covariance matrix can be really big!
  - Σ is n by n
  - 10000 features $\to |\Sigma|$
  - finding eigenvectors is very slow…
- Use singular value decomposition (SVD)
  - finds to k eigenvectors
  - great implementations available, e.g., Matlab svd

SVD



- Write $\mathbf{X} = \mathbf{W}\,\mathbf{S}\,\mathbf{V}^T$
  - $\mathbf{X} \leftarrow$ data matrix, one row per datapoint
  - $\mathbf{W} \leftarrow$ weight matrix, one row per datapoint – coordinate of $\mathbf{x}^i$ in eigenspace
  - $\mathbf{S} \leftarrow$ singular value matrix, diagonal matrix
    - in our setting each entry is eigenvalue λ_j
  - $\mathbf{V}^T \leftarrow$ singular vector matrix
    - in our setting each row is eigenvector $\mathbf{v}_j$



PCA using SVD algoritm

- Start from m by n data matrix **X**
- **Recenter**: subtract mean from each row of **X**
  - $\mathbf{X_c} \leftarrow \mathbf{X} - \overline{\mathbf{X}}$
- Call SVD algorithm on $\mathbf{X_c}$ – ask for k singular vectors
- **Principal components:** k singular vectors with highest singular values (rows of $\mathbf{V}^T$)
  - **Coefficients** become:

$$z_j^i = (x^i - \bar{x}) \cdot v_j$$

or

from rows of W
scaled by S

Using PCA for dimensionality reduction in classification

- Want to learn $f:\mathbf{X} \mapsto Y$
  - $\mathbf{X}=<X_1,\ldots,X_n>$
  - but some features are more important than others
- **Approach**: Use PCA on **X** to select a few important features $u_1 \ldots u_k$

rather than learning $f(X) \mapsto Y$

$f(X_1, \ldots, X_n) \mapsto Y$

you learn $f(\underbrace{z_1, \ldots, z_k}) \longmapsto Y$

$z_i = (X - \bar{X}) \cdot u_i$

PCA for classification can lead to problems…

- Direction of maximum variation may be unrelated to "discriminative" directions:

- PCA often works very well, but sometimes must use more advanced methods
  - e.g., Fisher linear discriminant

What you need to know

- Dimensionality reduction
  - why and when it's important
- Simple feature selection
- Principal component analysis
  - minimizing reconstruction error
  - relationship to covariance matrix and eigenvectors
  - using SVD
  - problems with PCA

Announcements

- Homework 5:
  - Extension: Due Friday at 10:30am
  - Hand in to Monica, Wean 4619
- Project:
  - Poster session: Friday May 4th 2-5pm, NSH Atrium
    - please arrive a 15mins early to set up
  - Paper: Thursday May 10th by 2pm
    - electronic submission by email to instructors list
    - maximum of 8 pages, NIPS format
    - no late days allowed
- FCEs!!!!
  - Please, please, please, please, please, please give us your feedback, it helps us improve the class! ☺ 3
    - http://www.cmu.edu/fce

Markov Decision Processes (MDPs)

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 25th, ~~2006~~ 7

Thus far this semester

- Regression:
- Classification:
- Density estimation:

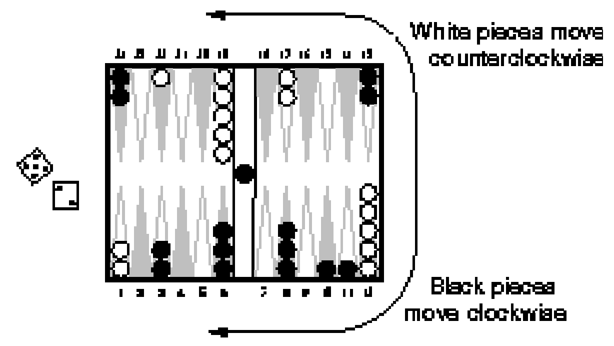
Learning to act

[Ng et al. ’05]

- Reinforcement learning
- An agent
  - Makes sensor observations
  - Must select action
  - Receives rewards
    - positive for “good” states
    - negative for “bad” states

Learning to play backgammon
[Tesauro '95]

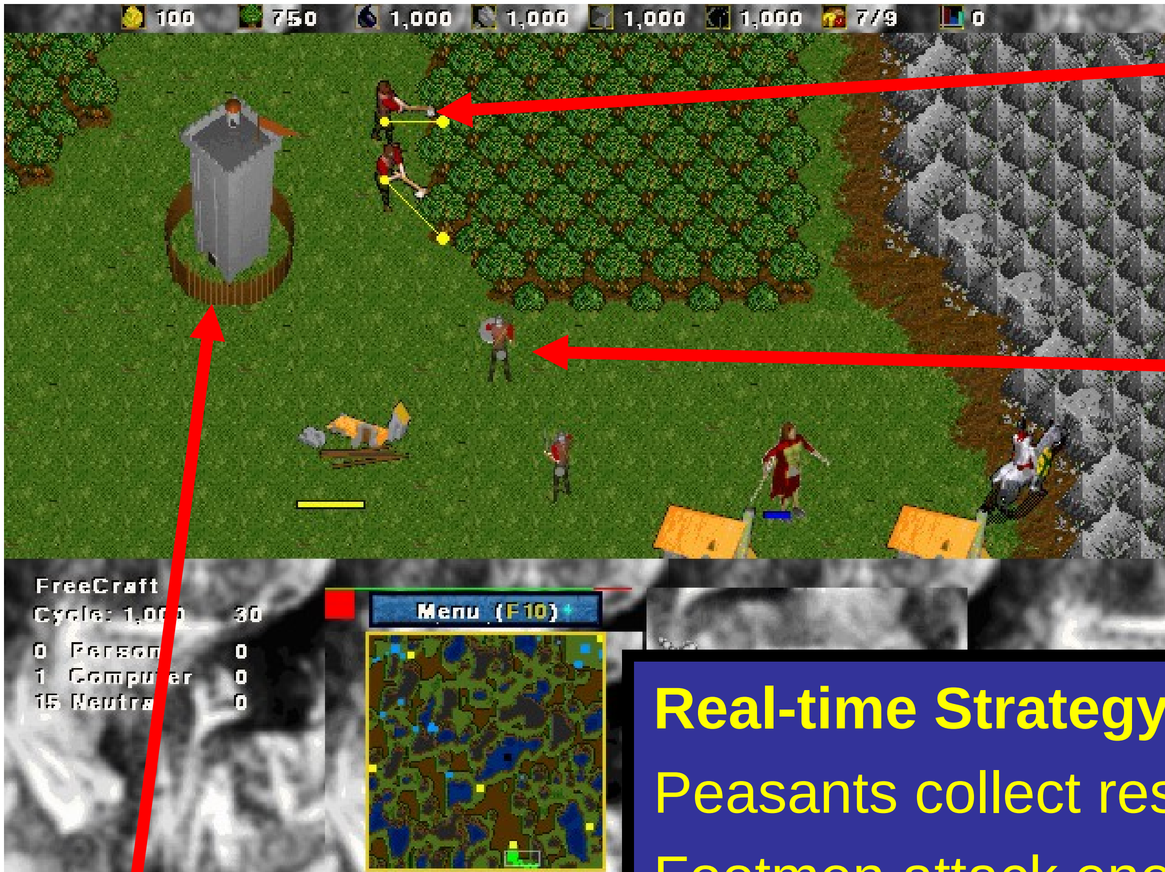
- Combines reinforcement learning with neural networks
- Played 300,000 games against itself
- Achieved grandmaster level!



Roadmap to learning about reinforcement learning

- When we learned about Bayes nets:
  - First talked about formal framework:
    - representation
    - inference
  - Then learning for BNs
- For reinforcement learning:
  - Formal framework
    - Markov decision processes
  - Then learning

FreeCraft



Real-time Strategy Game

Peasants collect resources and build

Footmen attack enemies

Buildings train peasants and footmen

States and actions

- State space:
  - Joint state **x** of entire system

- Action space:
  - Joint action **a**= {a_1,…, a_n} for all agents

States change over time

- Like an HMM, state changes over time
- Next state depends on current state and action selected
  - e.g., action="build castle" likely to lead to a state where you have a castle
- Transition model:
  - Dynamics of the entire system P(**x**'|**x**,**a**)

Some states and actions are better than others

- Each state **x** is associated with a reward
  - positive reward for successful attack
  - negative for loss

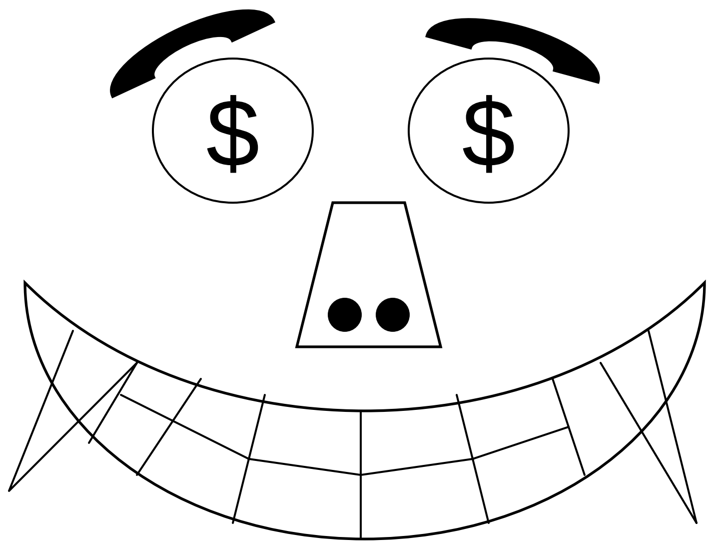
- Reward function:
  - Total reward R(**x**)

Discounted Rewards

An assistant professor gets paid, say, 20K per year.

How much, in total, will the A.P. earn in their life?

20 + 20 + 20 + 20 + 20 + … = Infinity



What's wrong with this argument?

Discounted Rewards

"A reward (payment) in the future is not worth quite as much as a reward now."

- Because of chance of obliteration
- Because of inflation

Example:

Being promised \$10,000 next year is worth only 90% as much as receiving \$10,000 right now.

Assuming payment *n* years in future is worth only $(0.9)^n$ of payment now, what is the AP's Future Discounted Sum of Rewards ?

Discount Factors

People in economics and probabilistic decision-making do this all the time.

The "Discounted sum of future rewards" using discount factor γ" is

(reward now) +

γ (reward in 1 time step) +

γ^2 (reward in 2 time steps) +

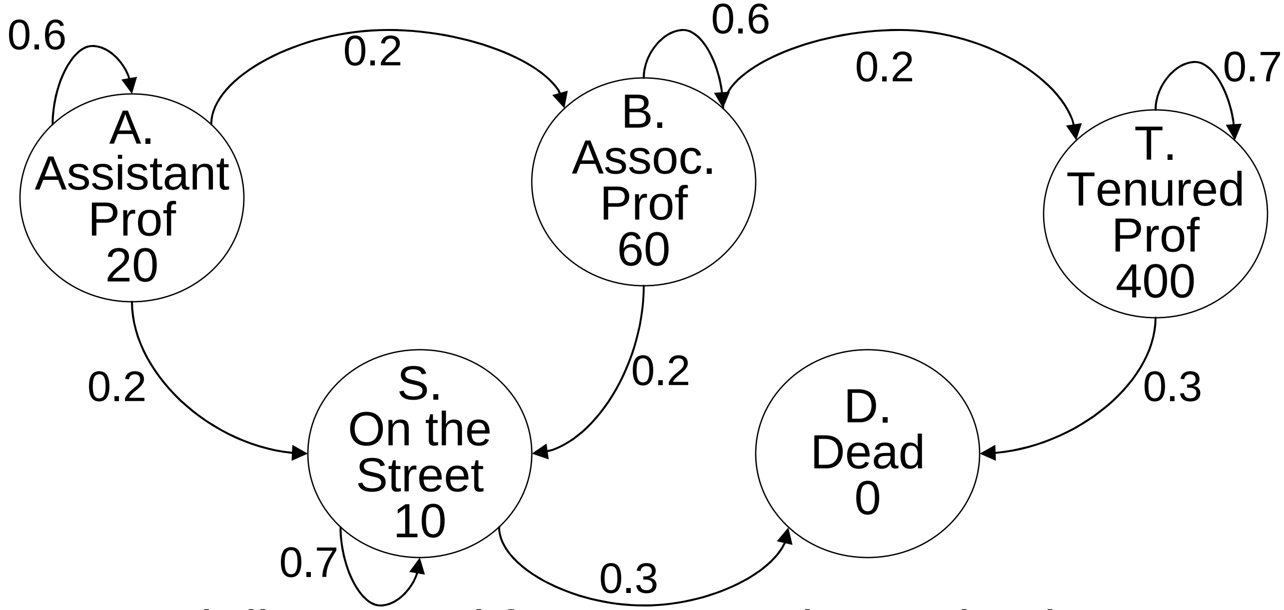
γ^3 (reward in 3 time steps) +

:

: (infinite sum)

The Academic Life

Assume Discount Factor $\gamma = 0.9$



Define:

V_A = Expected discounted future rewards starting in state A

V_B = Expected discounted future rewards starting in state B

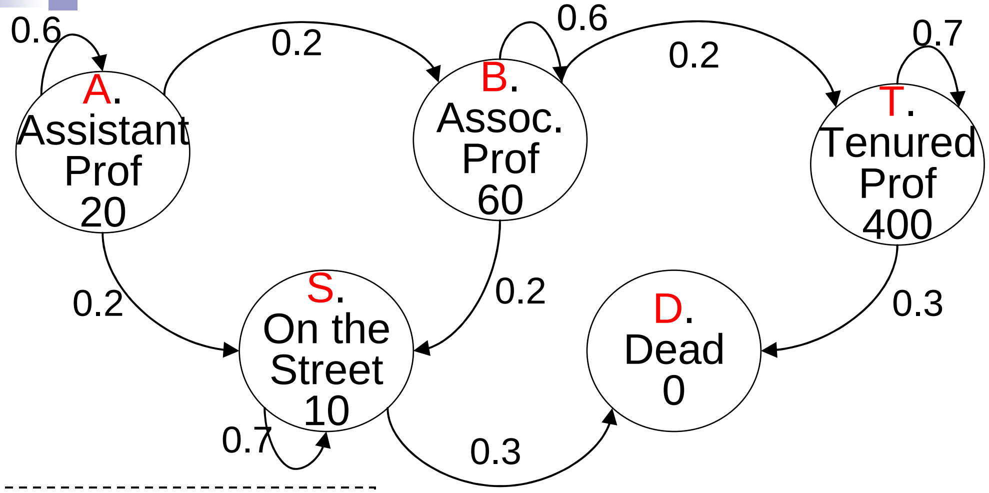
V_T = " " " " " " " T

V_S = " " " " " " " S

V_D = " " " " " " " D

How do we compute V_A, V_B, V_T, V_S, V_D ?

Computing the Future Rewards of an Academic



Assume Discount Factor $\gamma = 0.9$

Joint Decision Space

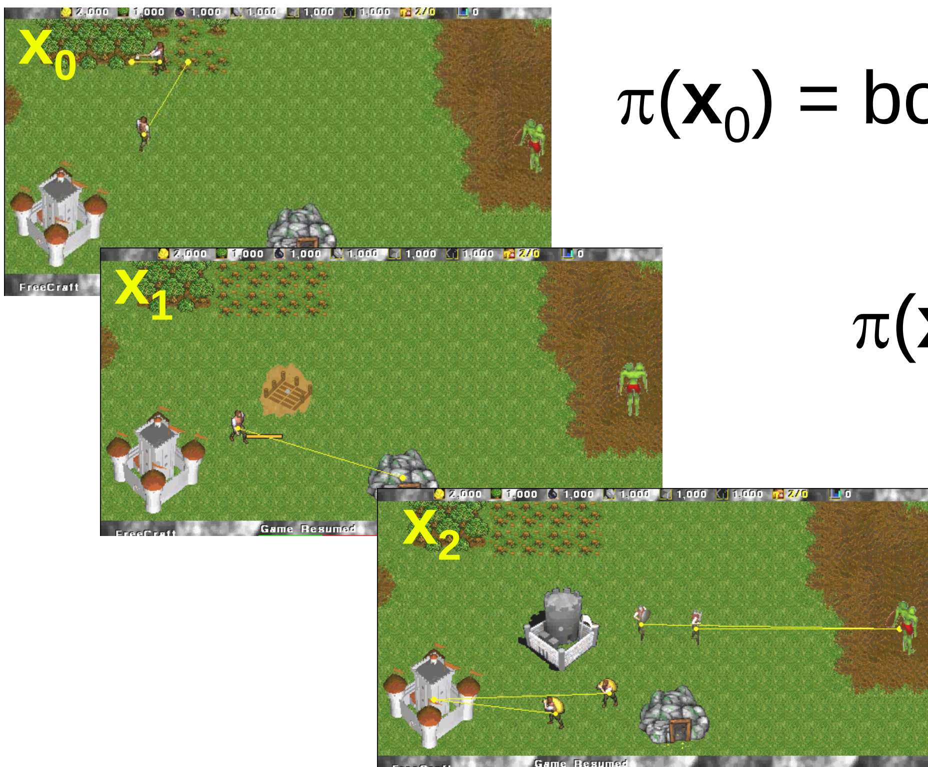
Markov Decision Process (MDP) Representation:

- State space:
  - Joint state **x** of entire system
- Action space:
  - Joint action **a**= $\{a_1,\ldots, a_n\}$ for all agents
- Reward function:
  - Total reward R(**x**,**a**)
    - sometimes reward can depend on action
- Transition model:
  - Dynamics of the entire system P(**x**'|**x**,**a**)

Policy

Policy: $\pi(\mathbf{x}) = \mathbf{a}$

At state $\mathbf{x}$, action $\mathbf{a}$ for all agents



$\pi(\mathbf{x}_0)$ = both peasants get wood

$\pi(\mathbf{x}_1)$ = one peasant builds barrack, other gets gold

$\pi(\mathbf{x}_2)$ = peasants get gold, footmen attack

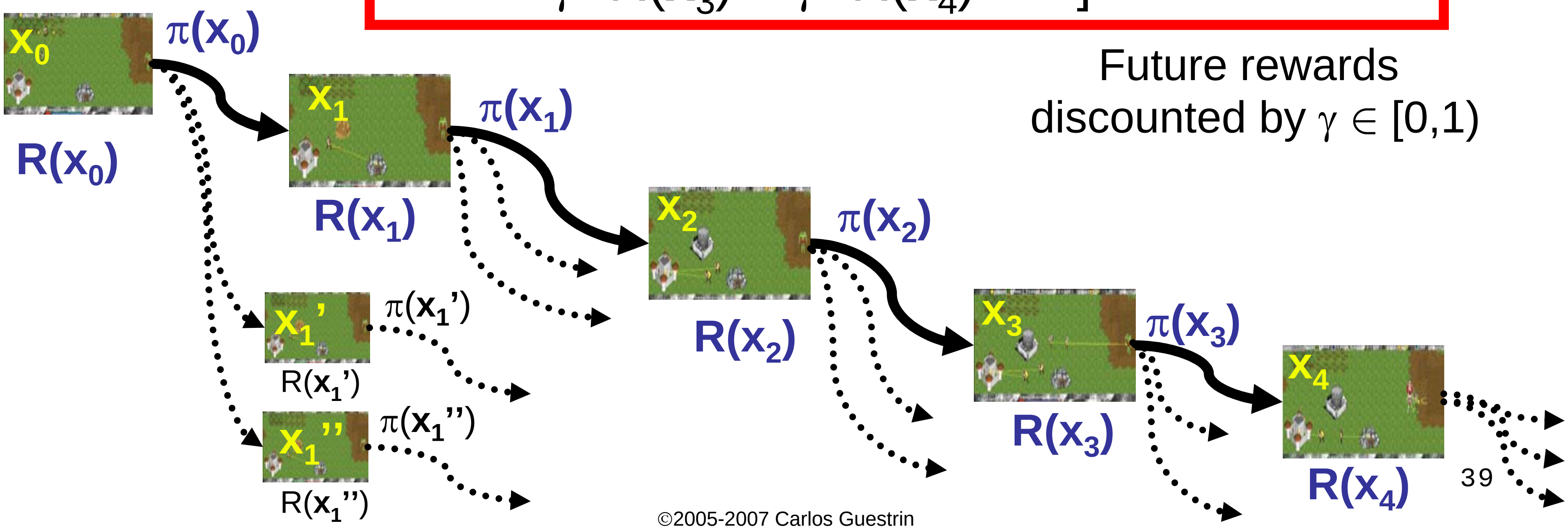
Value of Policy

Value: $V_\pi(\mathbf{x})$ ⇒ Expected long-term reward starting from **x**

$$V_\pi(\mathbf{x_0}) = \mathbf{E}_\pi[R(\mathbf{x}_0) + \gamma\, R(\mathbf{x}_1) + \gamma^2\, R(\mathbf{x}_2) + \gamma^3\, R(\mathbf{x}_3) + \gamma^4\, R(\mathbf{x}_4) + \quad]$$

Future rewards discounted by $\gamma \in [0,1)$

Start from $\mathbf{x_0}$



Computing the value of a policy

$$V_\pi(\mathbf{x_0}) = \mathbf{E}_\pi[R(\mathbf{x}_0) + \gamma\, R(\mathbf{x}_1) + \gamma^2\, R(\mathbf{x}_2) + \gamma^3\, R(\mathbf{x}_3) + \gamma^4\, R(\mathbf{x}_4) + \quad]$$

- Discounted value of a state:
  - value of starting from x_0 and continuing with policy π from then on

$$\begin{aligned} V_\pi(x_0) &= E_\pi[R(x_0) + \gamma R(x_1) + \gamma^2 R(x_2) + \gamma^3 R(x_3) + \cdots] \\ &= E_\pi[\sum_{t=0}^{\infty} \gamma^t R(x_t)] \end{aligned}$$

- A recursion!

Computing the value of a policy 1 – the matrix inversion approach

$$V_\pi(x) \;\; = \;\; R(x) + \gamma \sum_{x'} P(x' \mid x, a = \pi(x)) V_\pi(x')$$

- Solve by simple matrix inversion:

Computing the value of a policy 2 – iteratively

$$V_\pi(x) \;=\; R(x) + \gamma \sum_{x'} P(x' \mid x, a = \pi(x)) V_\pi(x')$$

- If you have 1000,000 states, inverting a 1000,000x1000,000 matrix is hard!
- Can solve using a simple convergent iterative approach: (a.k.a. dynamic programming)
  - Start with some guess V_0
  - Iteratively say:
    - $V_{t+1} = R + \gamma P_\pi V_t$
  - Stop when $||V_{t+1}-V_t||_\infty \leq \varepsilon$
    - means that $||V_\pi-V_{t+1}||_\infty \leq \varepsilon/(1-\gamma)$

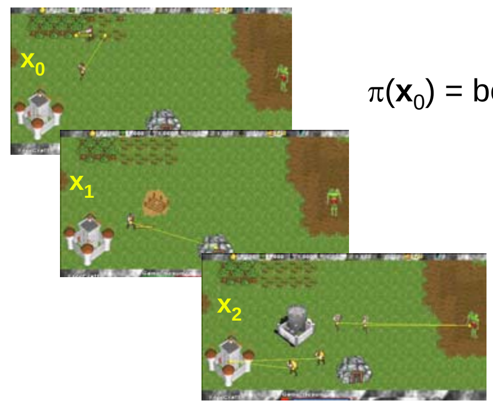
But we want to learn a **Policy**

- So far, told you how good a policy is…
- But how can we choose the best policy???

- Suppose there was only one time step:
  - world is about to end!!!
  - select action that maximizes reward!

Policy: $\pi(\mathbf{x}) = \mathbf{a}$

At state **x**, action **a** for all agents



$\pi(\mathbf{x}_0)$ = both peasants get wood

$\pi(\mathbf{x}_1)$ = one peasant builds barrack, other gets gold

$\pi(\mathbf{x}_2)$ = peasants get gold, footmen attack

Another recursion!

- Two time steps: address tradeoff
  - good reward now
  - better reward in the future

Unrolling the recursion

- Choose actions that lead to best value in the long run
  - Optimal value policy achieves optimal value V*

$$V^*(x_0) \;=\; \max_{a_0} R(x_0, a_0) + \gamma E_{a_0}[\max_{a_1} R(x_1) + \gamma^2 E_{a_1}[\max_{a_2} R(x_2) + \cdots]]$$

Bellman equation

- Evaluating policy π:

$$V_\pi(x) \;=\; R(x) + \gamma \sum_{x'} P(x' \mid x, a = \pi(x)) V_\pi(x')$$

- Computing the optimal value V* - Bellman equation

$$V^*(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' | \mathbf{x},\mathbf{a}) V^*(\mathbf{x}')$$

Optimal Long-term Plan

Optimal value function $V^*(\mathbf{x})$

Optimal Policy: $\pi^*(\mathbf{x})$

$$Q^*(\mathbf{x},\mathbf{a}) = R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V^*(\mathbf{x}')$$

Optimal policy:

$$\pi^*(\mathbf{x}) = \arg\max_{\mathbf{a}} Q^*(\mathbf{x},\mathbf{a})$$

Interesting fact – Unique value

$$V^*(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V^*(\mathbf{x}')$$

- *Slightly surprising fact*: There is only one V^* that solves Bellman equation!
  - there may be many optimal policies that achieve V^*
- *Surprising fact*: optimal policies are good everywhere!!!

$$V_{\pi^*}(x) \geq V_\pi(x), \ \forall x, \ \forall \pi$$

Solving an MDP

Solve Bellman equation → Optimal value $V^*(\mathbf{x})$ → Optimal policy $\pi^*(\mathbf{x})$

$$V^*(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V^*(\mathbf{x}')$$

Bellman equation is non-linear!!!

Many algorithms solve the Bellman equations:

- Policy iteration [Howard '60, Bellman '57]
- Value iteration [Bellman '57]
- Linear programming [Manne '60]
- …

Value iteration (a.k.a. dynamic programming) – the simplest of all

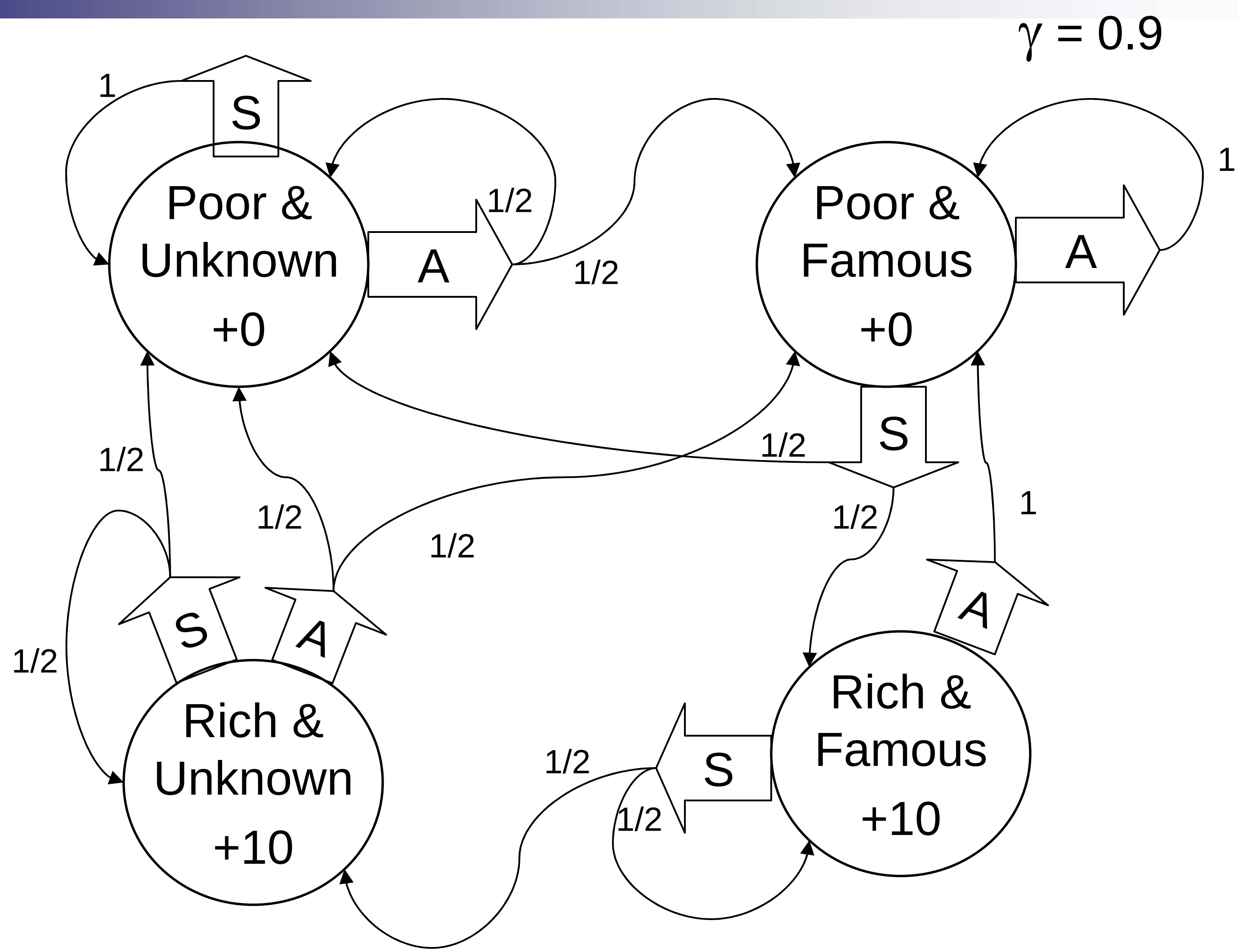
$$V^*(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V^*(\mathbf{x}')$$

- Start with some guess V_0
- Iteratively say:
  - $$V_{t+1}(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V_t(\mathbf{x}')$$
- Stop when $\|V_{t+1}-V_t\|_\infty \leq \varepsilon$
  - means that $\|V^*-V_{t+1}\|_\infty \leq \varepsilon/(1-\gamma)$

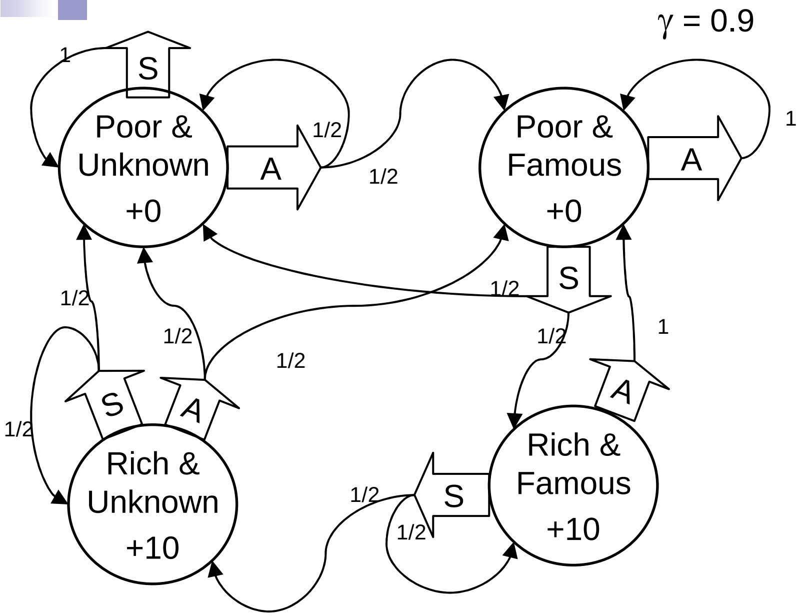
A simple example

You run a startup company.

In every state you must choose between Saving money or Advertising.



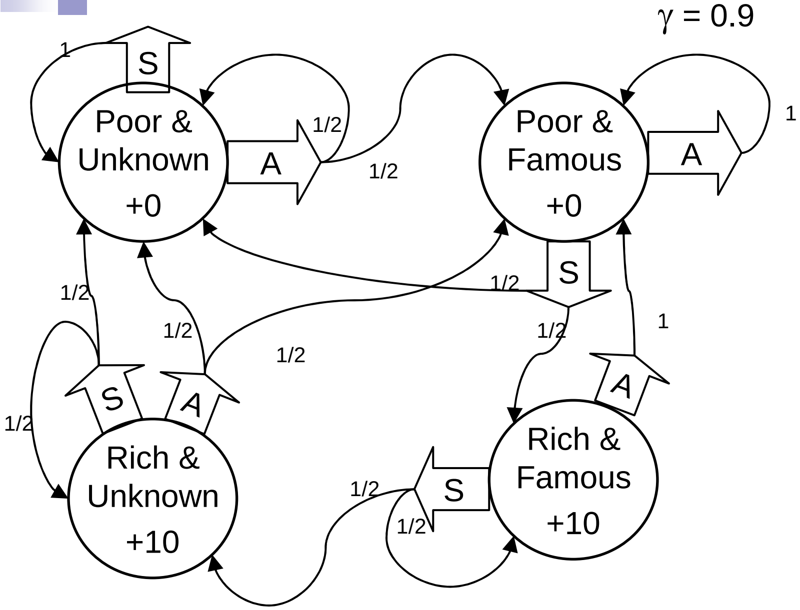
Let's compute $V_t(x)$ for our example



t	V_t(PU)	V_t(PF)	V_t(RU)	V_t(RF)
1				
2				
3				
4				
5				
6				

$$V_{t+1}(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V_t(\mathbf{x}')$$

Let's compute $V_t(x)$ for our example



t	$V_t(PU)$	$V_t(PF)$	$V_t(RU)$	$V_t(RF)$
1	0	0	10	10
2	0	4.5	14.5	19
3	2.03	6.53	25.08	18.55
4	3.852	12.20	29.63	19.26
5	7.22	15.07	32.00	20.40
6	10.03	17.65	33.58	22.43

$$V_{t+1}(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V_t(\mathbf{x}')$$

Policy iteration – Another approach for computing π^*

- Start with some guess for a policy π_0
- Iteratively say:
  - evaluate policy: $$V_t(\mathbf{x}) = R(\mathbf{x},\mathbf{a}=\pi_t(\mathbf{x})) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}=\pi_t(\mathbf{x})) V_t(\mathbf{x}')$$
  - improve policy: $$\pi_{t+1}(\mathbf{x}) = \max_{\mathbf{a}} R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V_t(\mathbf{x}')$$
- Stop when
  - policy stops changing
    - usually happens in about 10 iterations
  - or $||V_{t+1}-V_t||_\infty \leq \varepsilon$
    - means that $||V^*-V_{t+1}||_\infty \leq \varepsilon/(1-\gamma)$

Policy Iteration & Value Iteration: Which is best ???

It depends.

Lots of actions? Choose Policy Iteration

Already got a fair policy? Policy Iteration

Few actions, acyclic? Value Iteration

Best of Both Worlds:

Modified Policy Iteration [Puterman]

…a simple mix of value iteration and policy iteration

3rd Approach

Linear Programming

LP Solution to MDP

[Manne '60]

Value computed by linear programming:

$$\text{minimize}: \sum_{\mathbf{x}} V(\mathbf{x})$$

$$\text{subject to}: \begin{cases} V(\mathbf{x}) \geq \; R(\mathbf{x},\mathbf{a}) + \gamma \sum_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) V(\mathbf{x}') \\ \forall \mathbf{x}, \mathbf{a} \end{cases}$$

- One variable $V(\mathbf{x})$ for each state
- One constraint for each state **x** and action **a**
- **Polynomial time solution**

What you need to know

- What's a Markov decision process
  - state, actions, transitions, rewards
  - a policy
  - value function for a policy
    - computing V_π
- Optimal value function and optimal policy
  - Bellman equation
- Solving Bellman equation
  - with value iteration, policy iteration and linear programming

Acknowledgment

- This lecture contains some material from Andrew Moore's excellent collection of ML tutorials:
  - http://www.cs.cmu.edu/~awm/tutorials