Good Afternoon, Colleagues
Good Afternoon, Colleagues

Are there any questions?
Logistics

- Next week’s readings: adversarial search
Logistics

- Next week’s readings: adversarial search
Continuous Local Search to learn fast walk

Goal: Enable an Aibo to walk as fast as possible
Continuous Local Search to learn fast walk

Goal: Enable an Aibo to walk as fast as possible

- Start with a parameterized walk
- Learn fastest possible parameters
Continuous Local Search to learn fast walk

Goal: Enable an Aibo to walk as fast as possible

- Start with a parameterized walk
- Learn fastest possible parameters
- No simulator available:
  - Learn entirely on robots
  - Minimal human intervention
Walking Aibos

- Walks that “come with” Aibo are slow
- RoboCup soccer: 25+ Aibo teams internationally
  - Motivates faster walks
Walking Aibos

- Walks that “come with” Aibo are **slow**

- **RoboCup** soccer: **25+ Aibo teams** internationally
  - Motivates faster walks

<table>
<thead>
<tr>
<th>Hand-tuned gaits (2003)</th>
<th>Learned gaits</th>
</tr>
</thead>
<tbody>
<tr>
<td>230 mm/s</td>
<td>245</td>
</tr>
</tbody>
</table>
A Parameterized Walk

● Developed from scratch as part of **UT Austin Villa 2003**

● **Trot gait** with elliptical locus on each leg
Locus Parameters

- Ellipse length
- Ellipse height
- Position on $x$ axis
- Position on $y$ axis
- Body height
- Timing values

12 continuous parameters
Locus Parameters

- Ellipse length
- Ellipse height
- Position on $x$ axis
- Position on $y$ axis
- Body height
- Timing values

12 continuous parameters

- Hand tuning by April, ’03: 140 mm/s
- Hand tuning by July, ’03: 245 mm/s
## Parameters To Learn

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front ellipse:</td>
<td></td>
</tr>
<tr>
<td>(height)</td>
<td>4.2</td>
</tr>
<tr>
<td>(x offset)</td>
<td>2.8</td>
</tr>
<tr>
<td>(y offset)</td>
<td>4.9</td>
</tr>
<tr>
<td>Rear ellipse:</td>
<td></td>
</tr>
<tr>
<td>(height)</td>
<td>5.6</td>
</tr>
<tr>
<td>(x offset)</td>
<td>0.0</td>
</tr>
<tr>
<td>(y offset)</td>
<td>-2.8</td>
</tr>
<tr>
<td>Ellipse length</td>
<td>4.893</td>
</tr>
<tr>
<td>Ellipse skew multiplier</td>
<td>0.035</td>
</tr>
<tr>
<td>Front height</td>
<td>7.7</td>
</tr>
<tr>
<td>Rear height</td>
<td>11.2</td>
</tr>
<tr>
<td>Time to move through locus</td>
<td>0.704</td>
</tr>
<tr>
<td>Time on ground</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Experimental Setup

- Policy $\pi = \{\theta_1, \ldots, \theta_{12}\}$, $V(\pi) =$ walk speed when using $\pi$
Experimental Setup

- Policy $\pi = \{\theta_1, \ldots, \theta_{12}\}$, $V(\pi) =$ walk speed when using $\pi$

- Training Scenario
  - Robots time themselves traversing fixed distance
  - Multiple traversals (3) per policy to account for noise
Experimental Setup

• Policy $\pi = \{\theta_1, \ldots, \theta_{12}\}$, $V(\pi) =$ walk speed when using $\pi$

• Training Scenario
  - Robots time themselves traversing fixed distance
  - Multiple traversals (3) per policy to account for noise
  - Multiple robots evaluate policies simultaneously
  - Off-board computer collects results, assigns policies
Experimental Setup

- Policy $\pi = \{\theta_1, \ldots, \theta_{12}\}$, $V(\pi) =$ walk speed when using $\pi$

- Training Scenario
  - Robots **time themselves** traversing fixed distance
  - Multiple traversals (3) per policy to account for **noise**
  - **Multiple robots** evaluate policies simultaneously
  - Off-board computer **collects results**, assigns policies

No human intervention except battery changes
Policy Gradient RL

- From $\pi$, want to move in direction of gradient of $V(\pi)$
Policy Gradient RL

- From $\pi$ want to move in direction of gradient of $V(\pi)$
  - Can’t compute $\frac{\partial V(\pi)}{\partial \theta_i}$ directly: estimate empirically
Policy Gradient RL

- From $\pi$ want to move in direction of gradient of $V(\pi)$
  - Can’t compute $\frac{\partial V(\pi)}{\partial \theta_i}$ directly: estimate empirically

- Evaluate neighboring policies to estimate gradient

- Each trial randomly varies every parameter
Policy Gradient RL

- From $\pi$ want to move in direction of gradient of $V(\pi)$
  - Can’t compute $\frac{\partial V(\pi)}{\partial \theta_i}$ directly: estimate empirically

- Evaluate neighboring policies to estimate gradient

- Each trial randomly varies every parameter
## Gradient Estimation

<table>
<thead>
<tr>
<th>$-\epsilon_1$</th>
<th>$\pi_1$</th>
<th>$\pi_2 - \pi_N$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1 - \epsilon_1$</td>
<td>...</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>$\theta_1 - \epsilon_1$</td>
<td>...</td>
<td>12.7</td>
<td></td>
</tr>
</tbody>
</table>

$\Rightarrow$ Average: 12.1

<table>
<thead>
<tr>
<th>$+0$</th>
<th>$\pi_1$</th>
<th>$\pi_2 - \pi_N$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1 + 0$</td>
<td>...</td>
<td>12.3</td>
<td></td>
</tr>
<tr>
<td>$\theta_1 + 0$</td>
<td>...</td>
<td>13.7</td>
<td></td>
</tr>
</tbody>
</table>

$\Rightarrow$ Average: 13.2

<table>
<thead>
<tr>
<th>$+\epsilon_1$</th>
<th>$\pi_1$</th>
<th>$\pi_2 - \pi_N$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1 + \epsilon_1$</td>
<td>...</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>$\theta_1 + \epsilon_1$</td>
<td>...</td>
<td>14.7</td>
<td></td>
</tr>
</tbody>
</table>

$\Rightarrow$ Average: 14.9
Taking a step
Taking a step

\[ A_i = \begin{cases} 
0 & \text{if } Avg_{+0,i} > Avg_{+\epsilon,i} \text{ and } Avg_{+0,i} > Avg_{-\epsilon,i} \\
Avg_{+\epsilon,i} - Avg_{-\epsilon,i} & \text{otherwise}
\end{cases} \] (1)
Taking a step

\[ A_i = \begin{cases} 
0 & \text{if } \text{Avg}_{+0,i} > \text{Avg}_{+\epsilon,i} \text{ and } \text{Avg}_{+0,i} > \text{Avg}_{-\epsilon,i} \\
\text{Avg}_{+\epsilon,i} - \text{Avg}_{-\epsilon,i} & \text{otherwise}
\end{cases} \]  

(1)

- Normalize \( A \), multiply by scalar step-size \( \eta \)
- \( \pi = \pi + \eta A \)
Experiments

• Started from **stable**, but fairly slow gait

• Used **3 robots** simultaneously

• Each iteration takes 45 traversals, $7\frac{1}{2}$ minutes
Experiments

- Started from **stable**, but fairly slow gait
- Used **3 robots** simultaneously
- Each iteration takes 45 traversals, $7\frac{1}{2}$ minutes

Before learning

![Before learning image]

After learning

![After learning image]

- 24 iterations = **1080 field traversals**, $\approx 3$ hours
Results

Velocity of Learned Gait during Training

Learned Gait (UT Austin Villa)
Learned Gait (UNSW)
Hand–tuned Gait (UNSW)
Hand–tuned Gait (UT Austin Villa)
Hand–tuned Gait (German Team)

Number of Iterations

0 5 10 15 20 25

Velocity (mm/s)
Results

- Additional iterations didn’t help
- Spikes: evaluation noise? large step size?
# Learned Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial Value</th>
<th>$\epsilon$</th>
<th>Best Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front ellipse:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(height)</td>
<td>4.2</td>
<td>0.35</td>
<td>4.081</td>
</tr>
<tr>
<td>(x offset)</td>
<td>2.8</td>
<td>0.35</td>
<td>0.574</td>
</tr>
<tr>
<td>(y offset)</td>
<td>4.9</td>
<td>0.35</td>
<td>5.152</td>
</tr>
<tr>
<td>Rear ellipse:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(height)</td>
<td>5.6</td>
<td>0.35</td>
<td>6.02</td>
</tr>
<tr>
<td>(x offset)</td>
<td>0.0</td>
<td>0.35</td>
<td>0.217</td>
</tr>
<tr>
<td>(y offset)</td>
<td>-2.8</td>
<td>0.35</td>
<td>-2.982</td>
</tr>
<tr>
<td>Ellipse length</td>
<td>4.893</td>
<td>0.35</td>
<td>5.285</td>
</tr>
<tr>
<td>Ellipse skew multiplier</td>
<td>0.035</td>
<td>0.175</td>
<td>0.049</td>
</tr>
<tr>
<td>Front height</td>
<td>7.7</td>
<td>0.35</td>
<td>7.483</td>
</tr>
<tr>
<td>Rear height</td>
<td>11.2</td>
<td>0.35</td>
<td>10.843</td>
</tr>
<tr>
<td>Time to move</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>through locus</td>
<td>0.704</td>
<td>0.016</td>
<td>0.679</td>
</tr>
<tr>
<td>Time on ground</td>
<td>0.5</td>
<td>0.05</td>
<td>0.430</td>
</tr>
</tbody>
</table>

Peter Stone
Algorithmic Comparison, Robot Port

Before learning

After learning
Summary

- Used policy gradient RL to learn fastest Aibo walk
- All learning done on real robots
- No human intervention (except battery changes)
Grasping the Ball

- **Three stages:** walk to ball; slow down; lower chin

- Head proprioception, IR chest sensor $\rightarrow$ ball distance

- Movement specified by 4 parameters
Grasping the Ball

- Three stages: walk to ball; slow down; lower chin
- Head proprioception, IR chest sensor $\rightarrow$ ball distance
- Movement specified by 4 parameters

Brittle!
Parameterization

- **slowdown_dist**: when to slow down
- **slowdown_factor**: how much to slow down
- **capture_angle**: when to stop turning
- **capture_dist**: when to put down head
Learning the Chin Pinch

- **Binary, noisy** reinforcement signal: multiple trials
- **Robot evaluates self:** no human intervention
Results

- Evaluation of policy gradient, hill climbing, amoeba
## What it learned

<table>
<thead>
<tr>
<th>Policy</th>
<th>slowdown dist</th>
<th>slowdown factor</th>
<th>capture angle</th>
<th>capture dist</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>200mm</td>
<td>0.7</td>
<td>15.0°</td>
<td>110mm</td>
<td>36%</td>
</tr>
<tr>
<td>Policy gradient</td>
<td>125mm</td>
<td>1</td>
<td>17.4°</td>
<td>152mm</td>
<td>64%</td>
</tr>
<tr>
<td>Amoeba</td>
<td>208mm</td>
<td>1</td>
<td>33.4°</td>
<td>162mm</td>
<td>69%</td>
</tr>
<tr>
<td>Hill climbing</td>
<td>240mm</td>
<td>1</td>
<td>35.0°</td>
<td>170mm</td>
<td>66%</td>
</tr>
</tbody>
</table>
Instance of Layered Learning

- For domains too complex for tractably mapping state features $S \rightarrow$ outputs $O$
- Hierarchical subtask decomposition given: $\{L_1, L_2, \ldots, L_n\}$
- Machine learning: exploit data to train, adapt
- Learning in one layer feeds into next layer
Relaxing the Assumptions

- Nondeterministic actions:
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations:
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations: Belief states
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations: Belief states
- Unknown environments:
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations: Belief states
- Unknown environments: Online search
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations: Belief states
- Unknown environments: Online search
- Adversaries:
Relaxing the Assumptions

- Nondeterministic actions: AND-OR search
- Partial observations: Belief states
- Unknown environments: Online search
- Adversaries: Next week....