Learning for Control of Synthetic and Cyborg

PI's: Pieter Abbeel (PI), Ron Fearing (co-PI),
Students: Nimbus Goehausen, Woody Hoburg,

Summary

Objective

Development of learning and adaptation capabilities that will enable operation of synthetic and cyborg insects in complicated environments, such as collapsed buildings.

Platforms

Cyborg Beetle

Synthetic Crawler

Year 1 Results

Cyborg beetle:
Setup, flight initiation, cessation

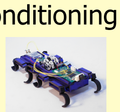
Synthetic crawler: 
On-board electronics: video, gyro, accel.

Learning and adaptation:
Faster-learning policy gradient method

Year 2 Results

Cyborg beetle:

Wing-folding muscle, Conditioning

Synthetic crawler: 

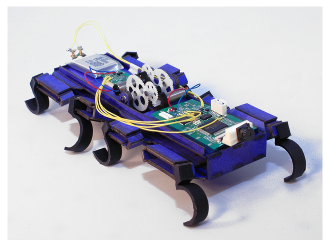
Complete crawler, Preliminary control

Learning and adaptation:

Risk-sensitive reinforcement learning

Synthetic Crawler

OctoRoACH

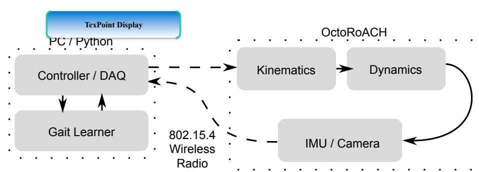


Specs

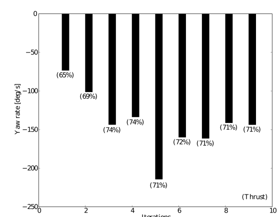
Sensing:
- Inertial Measurements Unit
- Back-EMF
- Cell-phone camera

Control:
- 40 MHz dsPIC microcontroller
- Differential drive-train
- Manually-tuned PID controllers
- 802.15.4 Wireless Radio

Learning to turn – Policy Search



Preliminary results were obtained using the policy search algorithm PEGASUS [Ng & Jordan 2000] for getting the optimal policy for turning on carpet.



From running gait at 10% thrust, robot turned fastest when other leg was run at 71% thrust.

Risk-Sensitive RL (1)

Motivation

- Expected return is often optimized
- Reasons: Law of large numbers + Mathematical convenience
- Not necessarily good criterion when only getting to execute once
- Motivating example: consider buying tickets to fly to a very important meeting, for which it would be disastrous if you arrived late. Some flights arrive on time more often than others, and these delays might be amplified if you miss connecting flights. With these risks in mind, would you rather take a route with an expected travel time of 13:48:09 and no further guarantees, would you choose a route that takes 13:53:28 on average with a standard deviation of 0:58:28, or would you prefer a route that takes less than 17:32:04 with 99% probability?

Preliminaries

- Markov Decision Process
  - State space S
  - Action space A
  - Stochastic dynamics
  - Stochastic reward function R
- Return from time t onwards

$$V^t(\pi) = \sum_{i=t}^{h-1} R^i(S_i, \pi^i(S_i), S_{i+1})$$

- "Optimize" $V(\pi) := V^0(\pi)$ (random var.)
- Traditionally: optimize expected return:

$$\max_{\pi} E_s\left[V(\pi)\right] \quad \text{where} \quad E_s[\ \diamond\ ] := E[\ \diamond \mid S_0 = s]$$

Insects in Uncertain Dynamic Environments

Michel Maharbiz (co-PI)
Amol Jadhav, Svet Kolev, Hirotaka Sato, Jie Tang

Cyborg Beetle

•All beetles were handled according to long-established institutional and scientific guidelines for experimenting with animals
• All insects in our lab are treated with respect and care
• Animals are NOT devices.

Goal

Create a reliable micro air vehicle (MAV) from a live insect

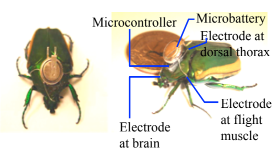
Minimum control functions:
- On/Off
- Throttle
- Turning

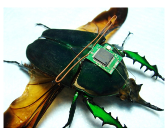
Why Beetles?

	Beetle	Moth	Locust	Fly
Payload capacity /gram	3.0	1.0	0.5	0.001
Adult life span / days	180	10	65	30
Survival under starving /days	30	-	-	3

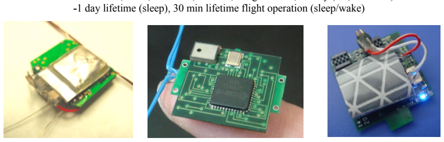
Variety, size range: 1mm to 10cm
Strong flyers
  maximum velocities: 7 – 14 km/hr
  flight durations: 10 min – 3 hours
Easy to rear and breed
Generally harmless to humans

Hardware

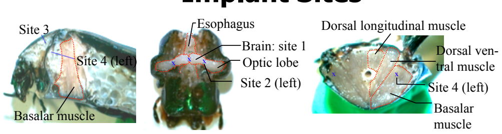




V3.0 specs: -1.6 x 1.6 cm, 1.22 grams (0.67 g board + 0.55 g battery)
- controller, radio, oscillator, antenna, wingbeat mic, battery (4 v, 8.5 mAh)
-1 day lifetime (sleep), 30 min lifetime flight operation (sleep/wake)



Implant Sites

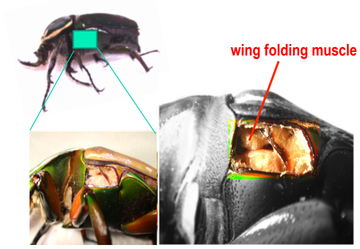


Flight Initiation, Cessation

- •Initiation success rate: **97% (N= 30)**
- •Cessation success rate: **100% (N=100)**
- •Mean response time, cessation: **73 ms**
- •Mean response time, initiation: **540 ms**

Wing-Folding Muscle

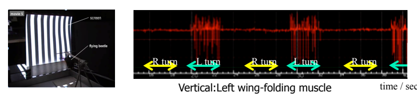
Anatomy



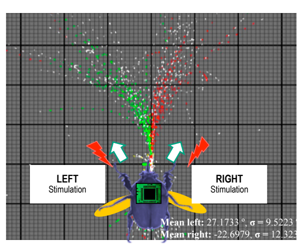
RE Snodgrass, Principles of Insect Morphology (1935), FW Darwin & JWS Pringle, Proc RSL (1959), K Ikeda et al., J Ins Physiol (1965)

Findings

The beetles missing wing folding muscles still folded and unfolded the wings but were losing steerage.
→ Hypothesis: "wing-folding muscle" could be key to steering rather than wing-folding



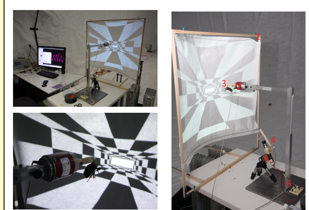
Train of the EMG spikes appeared only during the turn in the ipsilateral direction.



left stimulation   right stimulation

Conditioning

Setup



1. Oval screen
2. Torque meter
3. Beetle attachment point
4. Servo shutter
5. 1000 mW infrared laser
6. Control signals + power

Findings

- Beetles seem to try and center a bright pattern when provided with the visual-torque feedback loop
- Attempts at overriding natural preferences (e.g., reaction to optical flow from striped patterns) through conditioning have not been a success

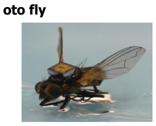
Plans

- Set up virtual environment with tunnel-like surroundings for beetle.
- Hypothesis: beetle will center bright end of tunnel
- Set up splits in the tunnel, with the two sides of the split naturally being equally good for the beetle. Condition the beetle to prefer one type.
- Move to stimuli that can be "mounted" onto the beetle itself

Where We're Headed

Low-level feedback
- Graded turning control based on either or both of wing-folding muscle stimulation and conditioning

Miniaturization oto fly



Risk-Sensitive RL (2)

Proposed Objective

- The Chernoff functional:

$$C_s^\delta[V(\pi)] = \sup_{\theta>0} \theta^{-1}\left(\log\delta - \log E_s\left[e^{-\theta V(\pi)}\right]\right)$$

Theorem 1. *Let X be a random variable which has a moment generating function, that is $E[\exp(\theta X)]$ is finite in a neighborhood of $\theta = 0$, and $\delta \in [0,1]$. Then, the Chernoff functional of this random variable, $C^\delta[X]$, is well defined and has the following properties:*

(i) $P(X \le C^\delta[X]) \le \delta$

(ii) $C^1[X] = E[X]$

(iii) $\lim_{\delta\to 0} C^\delta[X] = \inf\{x : P\{X \le x\} > 0\}$ *which could be* $-\infty$.

(iv) As $\delta \to 1$, $C^\delta[X] \approx E[X] - \sqrt{2\log(1/\delta)Var[X]}$

(v) $C^\delta[X] = E[X] - \sqrt{2\log(1/\delta)Var[X]}$ *if X is Gaussian.*

(vi) $C^\delta[X]$ *is a smooth, increasing function of* δ.

Cumulant Generating Function Iteration

Theorem 2. *Consider an MDP with a finite number of states and actions, and finite horizon, such that rewards are independent conditional on states, $R^t(S_t, \pi^t(S_t), S_{t+1}) \perp\!\!\!\perp V^{t+1}(\pi) \mid S_t, S_{t+1}$, and they have finite conditional moment generating functions, $E\exp(\theta R^t_{s,a,s'})$, in a neighborhood of $\theta = 0$. Then we can find $\pi^* = \operatorname{argmin}_\pi \log E_s[\exp(-\theta V(\pi))]$ by the iteration below, and the moment generating function $E_s[\exp(-\theta V(\pi^*))]$ is finite in a neighborhood of $\theta = 0$.*

$$q^t(s,a) := \log\sum_{s'} \exp\left(\log p_{s,a,s'} + \log E\left[e^{-\theta R^t_{s,a,s'}}\right] + u^{t+1}(s')\right) \tag{5}$$

$$u^t(s) := \min_a q^t(s,a) = \min_\pi \log E\left[e^{-\theta V^t(\pi)} \mid S_t = s\right], \qquad \pi^t(s) := \operatorname*{argmin}_a q^t(s,a)$$

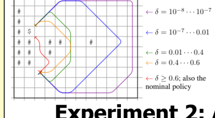
Finding the Global Optimum

$$\max_\pi C_s^\delta[V(\pi)] = \max_\pi \sup_{\theta>0} \frac{\log\delta - \log E_s[e^{-\theta V(\pi)}]}{\theta} = \sup_{\theta>0}\frac{\log\delta - \min_\pi \log E_s[e^{-\theta V(\pi)}]}{\theta}$$

The problem is not always convex in θ, but, we will see that it becomes convex after a change of variables to $z = \theta^{-1}$. Recall the notation $f(\theta) := \min_\pi \log E_s[e^{-\theta V(\pi)}]$:

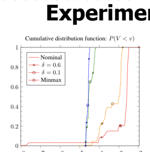
$$\max_\pi C_s^\delta[V(\pi)] = \sup_{\theta>0}\theta^{-1}(\log\delta - f(\theta)) = -\inf_{z>0}(-z\log\delta + zf(z^{-1}))$$

Experiment 1: Gridworld



Each state corresponds to a square in the grid, and the actions, {N, NE, E, SE, S, SW, W, NW}, cause a move in the respective direction. In unmarked squares, the actor's intention is executed with probability .93. Each of the seven remaining actions might be executed instead, each with probability 0.01. Squares marked with # and $ are absorbing states. The latter gives a reward of 35 when entered, and the former gives a penalty of 30. Any other state transitions cost 1. The horizon is 35.

Experiment 2: Air Travel Planning



Nominal policy ($\delta = 1$):
```
10:24 FAT - LAS 11:45
13:40 LAS - AMA 17:45
```
Chernoff policy, $\delta = 0.6$:
```
09:30 FAT - PHX 12:06
12:40 PHX - LAS 12:50
13:40 LAS - AMA 17:45
```
Chernoff policy, $\delta = 0.1$:
```
07:55 FAT - SFO 09:05
09:35 SFO - SAN 11:05
11:40 SAN - IAH 16:48
18:05 IAH - AMA 19:54
```
Minmax policy ($\delta = 0$):
```
07:55 FAT - SFO 09:05
12:50 SFO - DEN 16:20
18:30 DEN - AMA 20:40
```

- We assume that, in the case of a missed connecting flight due to delays, the airline will re-issue a ticket for the route of your choice leading to the original destination.
- We use historical data for February 2011, from the Office of Airline Information, Bureau of Trans- portation Statistics (BTS), available at http://www.transtats.bts.gov/.
- In general, we observed that lowering δ will produce policies with higher expected travel time, but lower standard deviation.