7. Motor Control and Reinforcement Learning
Outline

A. Action Selection and Reinforcement
B. Temporal Difference Reinforcement Learning
C. PVLV Model
D. Cerebellum and Error-driven Learning
Sensory-Motor Loop

- Why animals have nervous systems but plants do not: *animals move*
  — a nervous system is needed to coordinate the movement of an animal’s body
  — movement is fundamental to understanding cognition
- Perception conditions action
- Action conditions perception
  — profound effect of action on structuring perception is often neglected
Overview

• Subcortical areas:
  o basal ganglia
    ➢ reinforcement learning (reward/punishment)
    ➢ connections to “what” pathway
  o cerebellum
    ➢ error-driven learning
    ➢ connections to “how” pathway
  o disinhibitory output dynamic

• Cortical areas:
  o frontal cortex
    ➢ connections to basal ganglia & cerebellum
  o parietal cortex
    ➢ maps sensory information to motor outputs
    ➢ connections to cerebellum
## Learning Rules Across the Brain

<table>
<thead>
<tr>
<th>Area</th>
<th>Reward</th>
<th>Error</th>
<th>Self Org</th>
<th>Separator</th>
<th>Integrator</th>
<th>Attractor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primitive</strong></td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Basal Ganglia</td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>+</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>-</td>
<td>+++</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>+</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>-</td>
<td>+++</td>
</tr>
<tr>
<td>Neocortex</td>
<td>++</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td>+++</td>
<td>+++</td>
</tr>
</tbody>
</table>

+ = has to some extent
- = not likely to have

... +++ = defining characteristic – definitely has
... - - - = definitely does not have
## Primitive, Basic Learning...

<table>
<thead>
<tr>
<th>Area</th>
<th>Reward</th>
<th>Error</th>
<th>Self Org</th>
<th>Separator</th>
<th>Integrator</th>
<th>Attractor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primitive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basal Ganglia</td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- **Reward & Error** = most basic learning signals (self organized learning is a luxury…)
- **Simplest general solution to any learning problem** is a *lookup table* = separator dynamics

---

2/23/18

COSC 494/594 CCN

(slides < O'Reilly)
A. Action Selection and Reinforcement
Anatomy of Basal Ganglia

Basal Ganglia and Action Selection

Cerebral cortex

Possible responses

Basal ganglia

Most active response
Basal Ganglia: Action Selection

- Parallel circuits select motor actions and “cognitive” actions across frontal areas.

Motor actions

Oculomotor

Prefrontal strategies & plans

Orbitofrontal future rewards

Cingulate costs

2/23/18 COSC 494/594 CCN

(slide based on O’Reilly)
Release from Inhibition
Motor Loop Pathways

- **Direct**: striatum inhibits GPi (and SNr)
- **Indirect**: striatum inhibits GPe, which inhibits GPi (and SNr)
- **Hyperdirect**: cortex excites STN, which diffusely excites GPi (and SNr)
- **GPi** inhibits thalamus, which opens motor loops
Basal Ganglia System

- **Striatum**
  - matrix clusters (inhib.)
    - direct (Go) pathway → GPi
    - indirect (NoGo) path → GPe
  - patch clusters
    - to dopaminergic system

- **Globus pallidus, int. segment (GPi)**
  - tonically active
  - inhibit thalamic cells

- **Globus pallidus, ext. segment (GPe)**
  - tonically active
  - inhibits corresponding GPi neurons

- **Thalamus**
  - cells fire when both:
    - excited (cortex)
    - disinhibited (GPi)
  - disinhibits FC deep layers

- **Substantia nigra pars compacta (SNC)**
  - releases dopamine (DA) into striatum
  - excites D1 receptors (Go)
  - inhibits D2 receptors (NoGo)

- **Subthalamic nucleus (STN)**
  - hyperdirect pathway
  - input from cortex
  - diffuse excitatory output to GPi
  - global NoGo delays decision

*and substantia nigra pars reticulata (SNr)
*and superior colliculus (SC)
What is Dopamine Doing?

Dopamine carries the brain’s reward signal

Wise & Romper, 89

Schultz et. al, 98
• Feedforward, modulatory (disinhibition) on cortex/motor (same as cerebellum)
• Co-opted for higher level cognitive control → PFC
Basal Ganglia Architecture: Cortically-based Loops

Functional territories

- Limbic
- Associative
- Sensory
- Motor

- Cerebral cortex
- Striatum
- Pallidum/nigra
- Thalamus
Fronto-basal Ganglia Circuits in Motivation, Action, & Cognition
ChR2-mediated excitation of direct- and indirect-pathway MSNs in vivo drives activity in basal ganglia circuitry.
Human Probabilistic Reinforcement Learning

- Patients with Parkinson’s disease (PD) are impaired in cognitive tasks that require learning from positive and negative feedback.
- Likely due to depleted dopamine.
- But dopamine medication actually worsens performance in some cognitive tasks, despite improving it in others.

Frank, Seeberger & O’Reilly (2004)

([Train/Test scenarios with choices A, B, C, D, E, F with probability ratios])

Choose A?

A > CDEF

Avoid B?

B < CDEF

(slides based on Frank)
Testing the Model: Parkinson’s and Medication Effects

Choose A
Avoid B

Test Condition
50
60
70
80
90
100

Percent Accuracy

Probabilistic Selection
Test Performance

Seniors
PD OFF
PD ON

Frank, Seeberger & O’Reilly (2004)

(See also: Cools et al, 06, Frank et al 07, Moustafa et al 08, B´odi et al 09, Palminteri et al, 09, Voon et al 10, etc)
(A) The corticostriato-thalamo-cortical loops, including the direct (Go) and indirect (NoGo) pathways of the basal ganglia.
(B) The Frank (in press) neural network model of this circuit.
(C) Predictions from the model for the probabilistic selection task

Michael J. Frank et al. Science 2004;306:1940-1943

Published by AAAS
emergent Demonstration: BG

A simplified model compared to Frank, Seeberger, & O'Reilly (2004)
PFC-STN provides an override mechanism
Subthalamic Nucleus: Dynamic Modulation of Decision Threshold

Conflict (entropy) in choice prob ⇒ delay decision!

Bayesian approach to dynamic learning
• Learning from individual noisy outcomes should depend on uncertainty
• For choice tasks, uncertainty in $A > B$ (overlap)
  e.g., Yu & Dayan 05; Behrens et al 2007; Nassar et al 2010; Mathys et al 2011
B. Temporal Difference Reinforcement Learning
Reinforcement Learning: Dopamine

Rescorla-Wagner / Delta Rule:

\[
\delta = r - \hat{r}
\]

\[
\delta = r - \sum xw
\]

But no CS-onset firing – need to anticipate the future!

\[
\delta = (r + f) - \hat{r}
\]

CS-onset = future reward = \( f \)
Temporal Differences Learning

- $V(t) = r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2)\ldots$

- $\hat{V}(t) = r(t) + \gamma \hat{V}(t+1)$

- $0 = \left( r(t) + \hat{V}(t+1) \right) - \hat{V}(t)$

- $\delta = \left( r(t) + \hat{V}(t+1) \right) - \hat{V}(t)$

- $f = \gamma \hat{V}(t+1)$

← this is the future!
Network Implementation
The RL-cond Model

- **ExtRew**: external reward $r(t)$ (based on input)
- **TDRewPred**: learns to predict reward value
  - minus phase = prediction $V(t)$ from previous trial
  - plus phase = predicted $V(t+1)$ based on Input
- **TDRewInteg**: Integrates ExtRew and TDRewPred
  - minus phase = $V(t)$ from previous trial
  - plus phase = $V(t+1) + r(t)$
- **TD**: computes temporal dif. delta value $\approx$ dopamine signal
  - compute plus – minus from TDRewInteg
Classical Conditioning

- **Forward conditioning**
  - unconditioned stimulus (US): doesn’t depend on experience
  - leads to unconditioned response (UR)
  - preceding conditioned stimulus (CS) becomes associated with US
  - leads to conditioned response (CR)

- **Extinction**
  - after CS established, CS is presented repeatedly without US
  - CR frequency falls to pre-conditioning levels

- **Second-order conditioning**
  - CS1 associated with US through conditioning
  - CS2 associated with CS1 through conditioning, leads to CR
CSC Experiment

• A serial-compound stimulus has a series of distinguishable components

• A complete serial-compound (CSC) stimulus has a component for every small segment of time before, during, and after the US

• RL-cond.proj implements this form of conditioning
  — somewhat unrealistic, since the stimulus or some trace of it must persist until the US
RL-cond.proj
emergent Demonstration: RL

A simplified model of temporal difference reinforcement learning
Actor - Critic

a) Dopamine Burst

b) Dopamine Dip
Opponent-Actor Learning (OpAL)

- Actor has independent G and N weights
- Scaled by dopamine (DA) levels during choice
- Choice based on relative activation levels
- Low DA: costs amplified, benefits diminished ⇒ choice 1
- High DA: benefits amplified, costs diminished ⇒ choice 3
- Moderate DA ⇒ choice 2
- Accounts for differing costs & benefits

2/23/18

COSC 494/594 CCN
C. PVLV Model of DA Biology

A model of dopamine firing in the brain
Brain Areas Involved in Reward Prediction

- **Lateral hypothalamus (LHA):** provides a primary reward signal for basic rewards like food, water etc.

- **Patch-like neurons in ventral striatum (VS-patch)**
  - have direct inhibitory connections onto dopamine neurons in VTA and SNc
  - likely role in canceling influence of primary reward signals when they’re successfully predicted

- **Central nucleus of amygdala (CNA)**
  - important for driving dopamine firing at the onset of conditioned stimuli
  - receives input broadly from cortex
  - projects directly and indirectly to the VTA and SNc (DA neurons)
  - neurons in the CNA exhibit CS-related firing
PVLV Model of Dopamine Firing

- Two distinct systems: Primary Value (PV) and Learned Value (LV)
- DA signal at time of external reward (US):
  \[ \delta_{pv} = PV_e - PV_i = r - \hat{r} \]
- DA signal for LV when PV not present/expected:
  \[ \delta_{lv} = LV_e - LV_i \]
- \( LV_e \) is excitatory drive from CNA responding to CS (eventually canceled by \( LV_i \))
- \( LV_e \) and \( LV_i \) values learned from \( PV_e \) when rewards present/expected
- Hence, CS (or some trace) must still be present when US occurs
- CNA supports 1\textsuperscript{st} order conditioning, but not 2\textsuperscript{nd} order (that’s in BLA)
Biology of Dopamine Firing
More Detailed Description of PVLV

- Major issue: Which of PV/LV systems should be in charge of overall dopamine system?
- PV and LV learning occur when PV present or expected (indicated by $PV_r > \Theta_{pv}$)
- PVr system learns: $\delta w_{pvr} = r_{\text{present}} - PV_r$ (improves prediction)
- Recall alternative DA signals:
  $\delta_{pv} = PV_e - PV_i$, $\delta_{lv} = LV_e - LV_i$
- Novelty Value (NV) signal reflects stimulus novelty
- Overall dopamine signal:
  $$\delta = \begin{cases} 
  \delta_{pv}(t) - \delta_{pv}(t - 1) & \text{if } PV_r > \Theta_{pv} \\
  [\delta_{lv}(t) - \delta_{lv}(t - 1)] + [NV(t) - NV(t - 1)] & \text{otherwise}
  \end{cases}$$
- Note DA burst is phasic (ceases after CS onset)
More Detailed Description (ctu’d)

• Learning PV\textsubscript{i} weights:
  \[ \delta w\textsubscript{pv} = \varepsilon (PV\textsubscript{e} - PV\textsubscript{i})x \]

• Learning LV weights is conditional on PV filter:
  \[ \delta w\textsubscript{lv} = \begin{cases} 
  \varepsilon (PV\textsubscript{e} - LV\textsubscript{e})x & \text{if } PV\textsubscript{r} > \Theta_{pv} \\
  0 & \text{otherwise}
\end{cases} \]
PVLV.proj Model

- PV in Ventral Striatum system
- LV in Amygdala system
- VTA$_1$ and VS adapt to US$^+$
- Eventually VTA$_1$ bursts for CS onset
- LHB+RMTg and VS adapt to US$^-$
- VTA$_m$ and VS adapt to US$^-$
- Eventually DA dip for CS
emergent Demonstration: PVLV
D. Cerebellum and Error-driven Learning

“The blessing of dimensionality”
Functions of Cerebellum

• Maintenance of equilibrium and posture
• Timing of learned, skilled motor movement
  — any motor movement that improves with practice
  — timing, fluency, rhythm, coordination
  — involved in cognitive processes too
• Correction of errors during the execution of movements
  — error-driven learning
• Many inputs from cortical motor and sensory areas
• Influences cortical motor outputs to spinal chord
Lookup Table & Pattern Separation

f(x)

Lookup Table -- store learned input/output pairs

Cerebellum

Cortex

A

B

A

B

2/23/18

COSC 494/594 CCN

(slide < O’Reilly)
Cerebellum

- Inputs from parietal cortex and motor areas of frontal cortex
- Three layers, very many cortical maps
- Single basic circuit replicated throughout
- 200 million mossy fiber inputs (each to 500 granule cells)
  - projection of input into hyperdimensional space
  - separator learning and dynamics
- 40 billion granule cells (input from 4–5 mossy fibers)
- 15 million Purkinje cells (input from 200,000 granule cells)
  - matrix organization
  - enormous integration and cross connection
- Climbing fibers (one per Purkinje, from inferior olive)
Cerebellar Error-driven Learning

- Granule cells = high-dimensional encoding (separation)
- Purkinje/Olive = delta-rule error-driven learning
Cerebellum is Feed Forward

Feedforward circuit:

Input (PN) $\rightarrow$ granules $\rightarrow$ Purkinje $\rightarrow$ Output (DCN)

Inhibitory interactions – no attractor dynamics

**Key idea:** does delta-rule learning bridging small temporal gap:

$S(t-100) \rightarrow R(t)$

$\uparrow$ Error$(t+100)$
Mesostructure

• Microzone: defined by group of adjacent PCs contacted by CFs with same receptive profiles
  — comprises hundreds of PCs and several hundreds of thousands of other neurons
  — shaped as narrow strips a few PCs wide and several dozens of PCs in length
  — a fraction of a millimeter in width and several millimeters in length
  — parallel fibers (PFs) extend for several millimeters, crossing width of microzone and extending into neighbors
  — estimated that cat has about 5000 microzones

• Multizonal micro-complexes (MZMCs): basic functional units of cerebellar cortex
  — each comprises several microzones receiving common CF input and delivering their PC output to the same region of the cerebellar nuclei
  — seem to have an integrated function
  — constituent microzones may be in different regions of the cortex, which receive different MF input and may be associated with different aspects of motor control
  — MZMCs may provide for parallel processing and integration of inputs
Properties of Hyperdimensional Spaces

- Hyperdimensional spaces = spaces of very high dimension
- Consider vectors of 10,000 bits
  - measure distance by Hamming distance (HD)
  - or normalized Hamming distance (NHD)
- Mean HD = 5000, SD = 50 (binomial distribution)
- < 10^{-9} of space closer than NHD = 0.47 or farther than 0.53 (±300 = ±6 SD)
- Therefore random vectors almost surely have NHD = 0.5±0.03
- Vectors with < 3000 changed bits still accurately recognized
- Ref: Pentti Kanerva (2009), Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, *Cognitive Computation*, 1(2)
Orthogonality of Random Hyperdimensional Bipolar Vectors

- 99.99% probability of being within 4\(\sigma\) of mean

\[|u \cdot v| < 4\sigma\]

- It is 99.99% probable that random \(n\)-dimensional vectors will be within \(\varepsilon = 4/\sqrt{n}\) orthogonal

\[\text{iff } \|u\| \|v\| |\cos \theta| < 4\sqrt{n}\]

\[\text{iff } n|\cos \theta| < 4\sqrt{n}\]

\[\text{iff } |\cos \theta| < 4/\sqrt{n} = \varepsilon\]

- \(\varepsilon = 4\%\) for \(n = 10,000\)

- Probability of being less orthogonal than \(\varepsilon\) decreases exponentially with \(n\)

\[\Pr \{|\cos \theta| > \varepsilon\} = \text{erfc} \left(\frac{\varepsilon \sqrt{n}}{\sqrt{2}}\right)\]

\[\approx \frac{1}{6} \exp(-\varepsilon^2 n / 2) + \frac{1}{2} \exp(-2\varepsilon^2 n / 3)\]

- The brain gets approximate orthogonality by assigning random high-dimensional vectors

2/23/18
Hyperdimensional Pattern Associator

- Suppose $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P$ are a set of random hyperdimensional bipolar vectors (inputs).
- Let $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_P$ be arbitrary bipolar vectors (outputs).
- Define Hebbian linear associator matrix
  \[
  M = \sum_{k=1}^{P} \mathbf{q}_k \mathbf{p}_k^T
  \]
  - Then $M\mathbf{p}_k \approx \mathbf{q}_k$ (table lookup).
- To encode a sequence of random vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P$:
  \[
  M = \sum_{k=1}^{P-1} \mathbf{p}_{k+1} \mathbf{p}_k^T
  \]
  - Then $M\mathbf{p}_k = \mathbf{p}_{k+1}$ (sequence readout).
BG + Cerebellum Capacities

• Learn what satisfies basic needs, and what to avoid (BG reward learning)
  — And what information to maintain in working memory (PFC) to support successful behavior
• Learn basic Sensory $\rightarrow$ Motor mappings accurately (Cerebellum error-driven learning)
  — Sensory $\rightarrow$ Sensory mappings? (what is going to happen next)
BG + Cerebellum Incapacities

- Generalize knowledge to novel situations
  - Lookup tables don’t generalize well...
- Learn abstract semantics
  - Statistical regularities, higher-order categories, etc
- Encode episodic memories (specific events)
  - Useful for instance-based reasoning
- Plan, anticipate, simulate, etc...
  - Requires robust working memory
emergent Demonstration: Cereb