1 No HW this week so...

- Start thinking about whether any questions you've had could be addressed in final projects (tentatively)
- Get started with Emergent tutorials: build your own networks etc.
- Linked on class website: `http://ski.cog.brown.edu/cogsim.html`

2 Sequential & Temporally-Delayed Learning

1. The Problem.
2. Sequential Learning & Context.
3. Temporally-delayed Learning & Reinforcement.

3 The Problem

Error-driven + Hebbian: Solve tasks, learn systematic representations, generalize to new stimuli.
What's left?...

Time!

Currently: networks learn *immediate* consequence of a given input.

- What if current input only makes sense as part of a *sequence* of inputs (e.g., language, social interactions)?
- What if the consequence of this input comes *later* (e.g., school/work, life)?

4 Sequence Learning

How do we do it?
For example:

My favorite color is purple.
Purple my color favorite is.
Is my purple color favorite.
Is purple my color favorite.

The girl picked up the pen.
The pig raced around the pen.

We represent the *context*, not just the current input.

in language, social interactions, driving (who goes at a 4-way stop?)

5 Representing Context for Sequence Learning

How does the brain do it?
How would we get our models to do it?

Add layers to keep track of context (prefrontal cortex; hippocampus...).

6 An Example Task

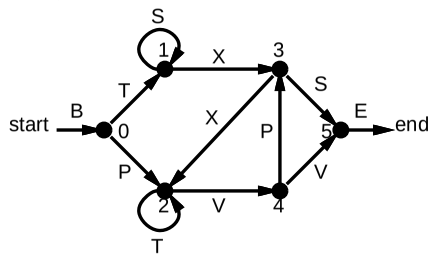
BTXSE
BPVPSE
BTSXXTVVE
BPTVPSE

Which of the following sequences are allowed?:
BTXXTTVVE
TSXSE
VVSXE
BSSXSE

7

An Example Task

BTXSE, BPVPSE, BTSXXTVVE, BPTVPSE, BTXXTTVVE
TSXSE, VVSXE, BSSXSE



We implicitly learn such grammars (e.g., pressing buttons faster to letters that follow grammar).

8

Time & Sequences

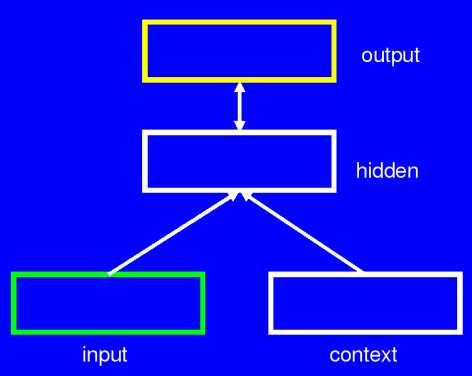
Currently: networks learn *immediate* consequence of a given input.

What if current input only makes sense as part of a temporally-extended *sequence* of inputs? (*context*)

What if the consequence of this input comes *later* in time? (next week)

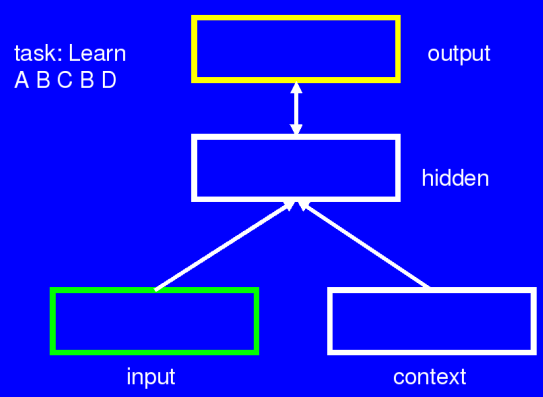
9

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



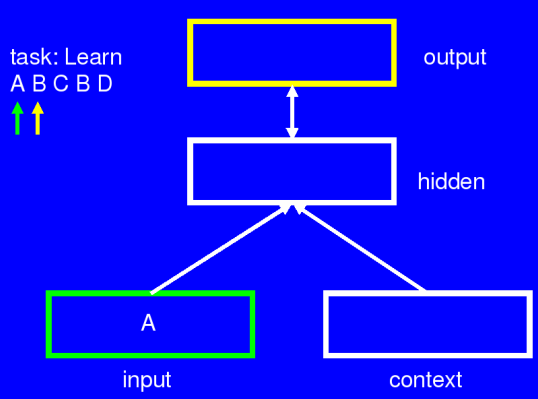
10

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



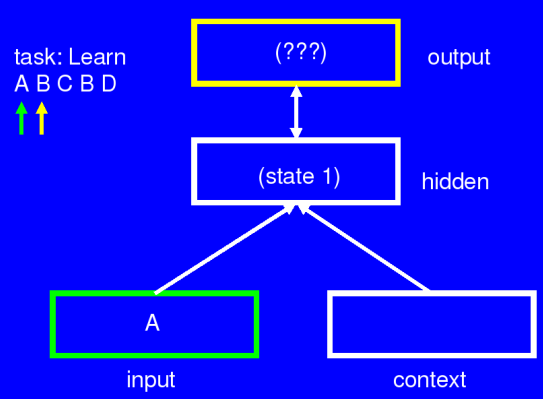
11

Simple Recurrent Network (SRN): An Architecture for Sequence Learning

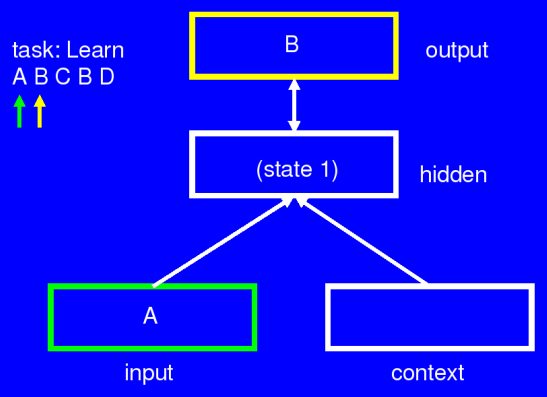


12

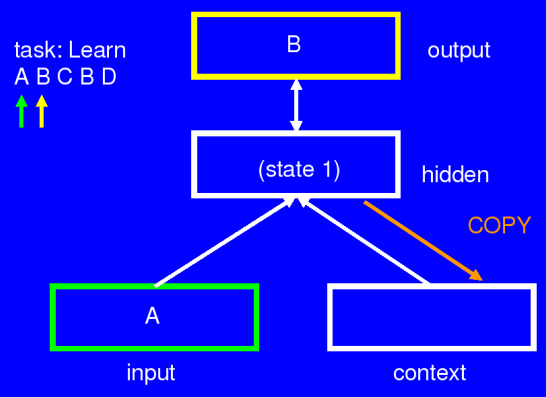
Simple Recurrent Network (SRN): An Architecture for Sequence Learning



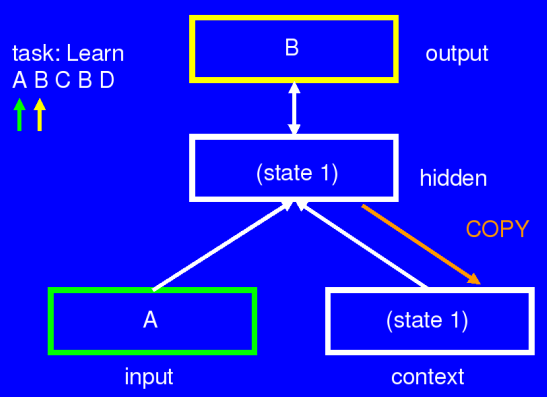
13 Simple Recurrent Network (SRN): An Architecture for Sequence Learning



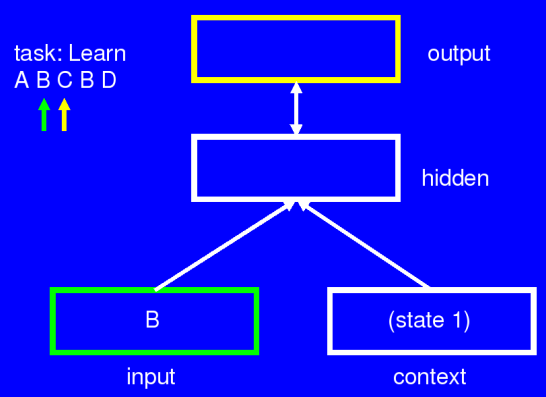
14 Simple Recurrent Network (SRN): An Architecture for Sequence Learning



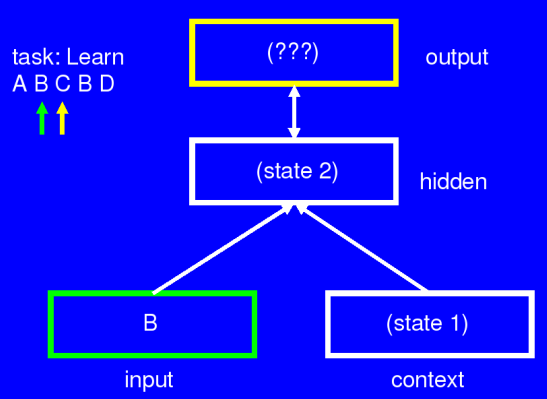
15 Simple Recurrent Network (SRN): An Architecture for Sequence Learning



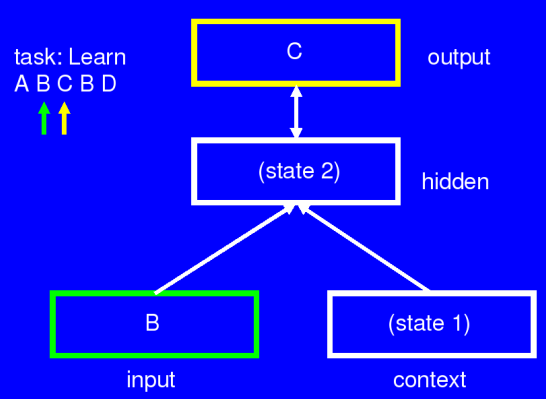
16 Simple Recurrent Network (SRN): An Architecture for Sequence Learning



17 Simple Recurrent Network (SRN): An Architecture for Sequence Learning

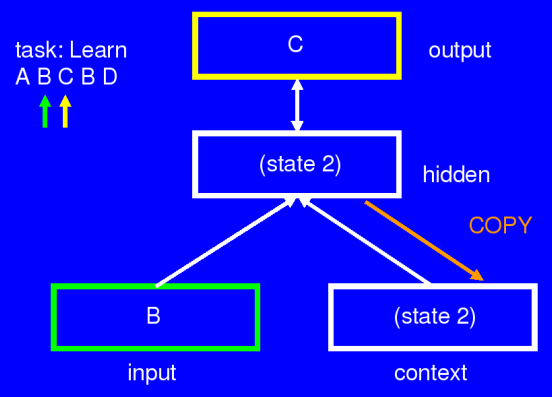


18 Simple Recurrent Network (SRN): An Architecture for Sequence Learning



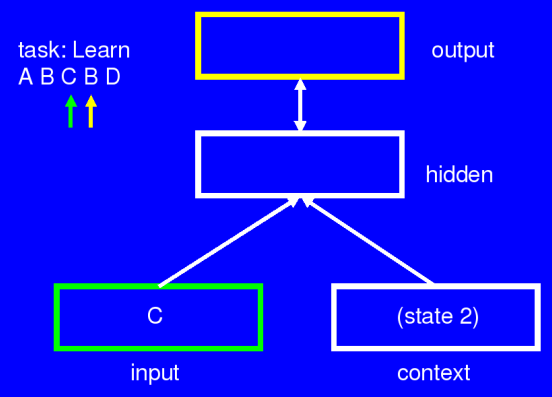
19

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



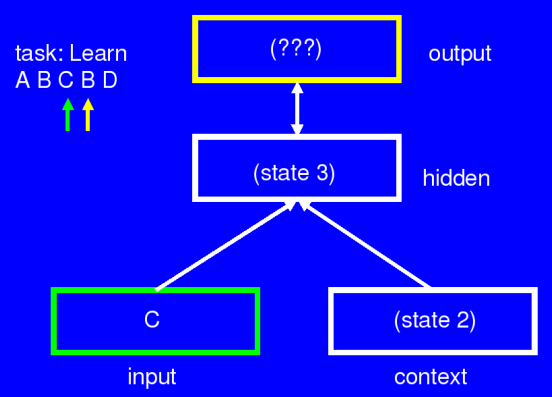
20

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



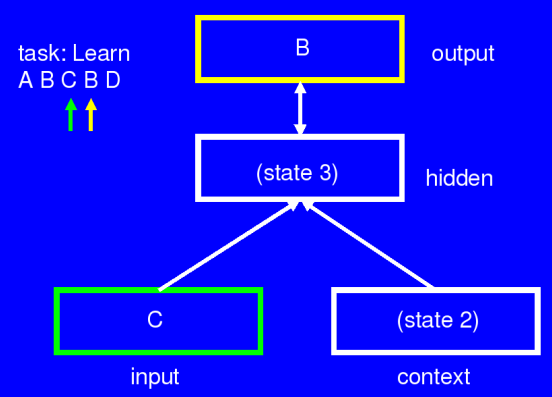
21

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



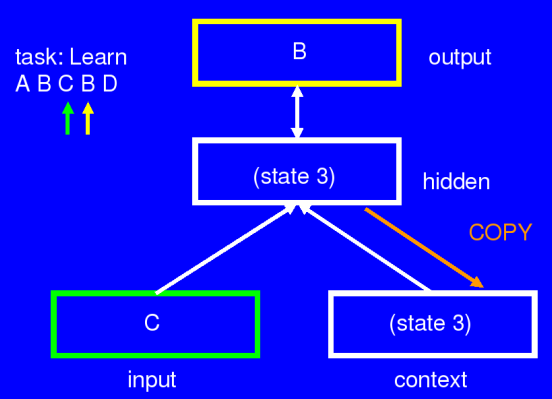
22

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



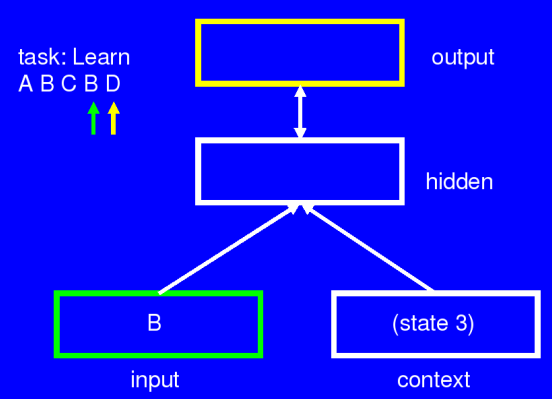
23

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



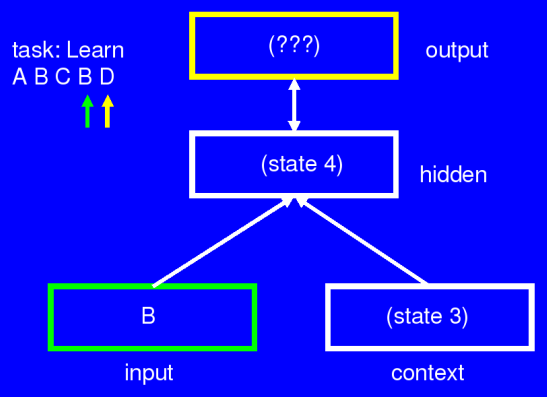
24

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



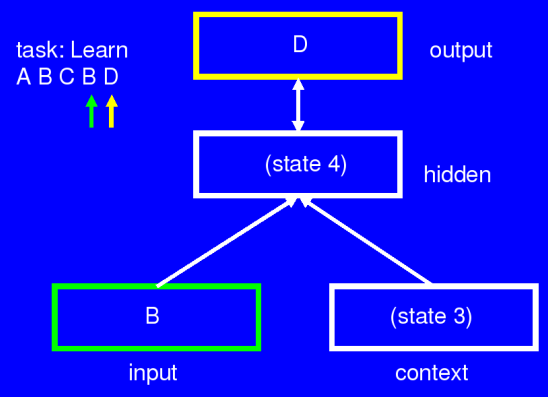
25

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



26

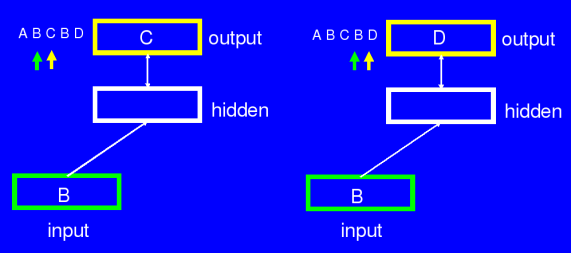
Simple Recurrent Network (SRN): An Architecture for Sequence Learning



27

Simple Recurrent Network: Summary

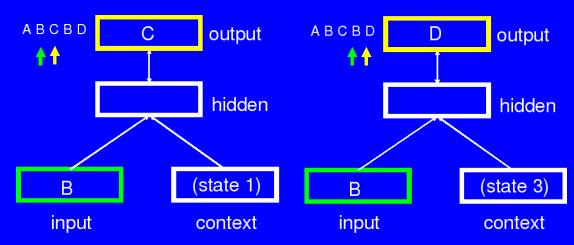
- Carries forward information by means of a **context** layer that contains the hidden representation from the previous time step.
- This hidden representation serves to disambiguate the input



28

Simple Recurrent Network: Summary

- Carries forward information by means of a **context** layer that contains the hidden representation from the previous time step.
- This hidden representation serves to disambiguate the input

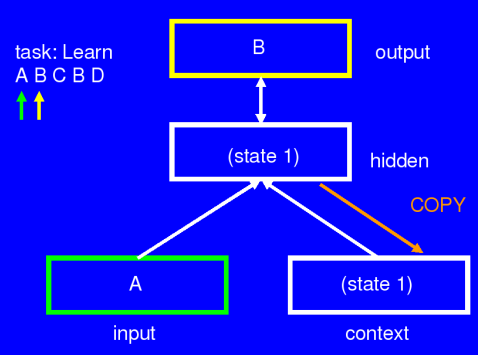


29

Why Copy the Hidden Representation?

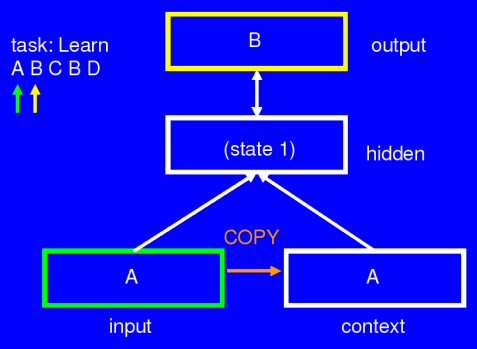
30

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



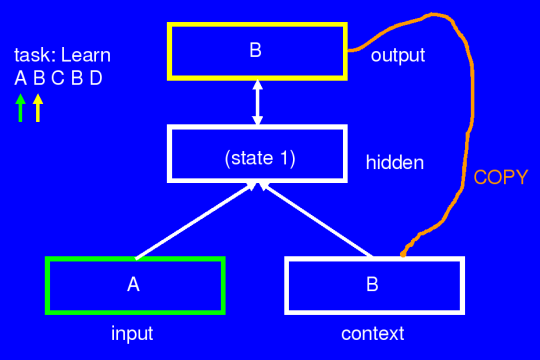
31

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



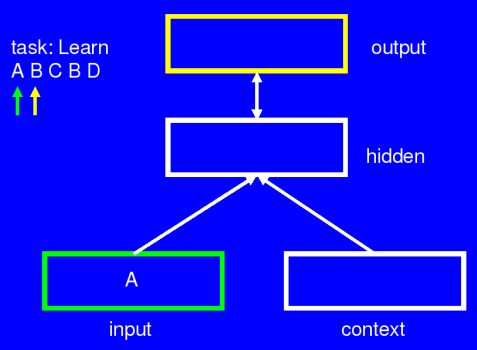
32

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



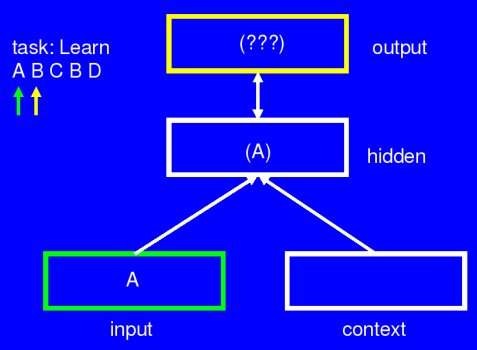
33

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



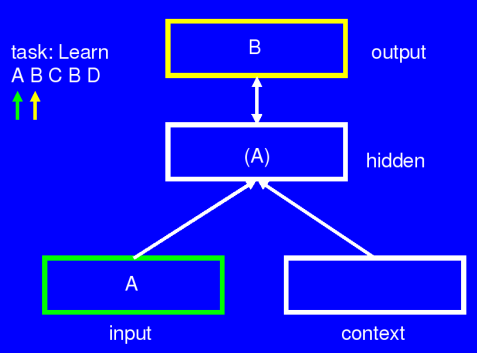
34

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



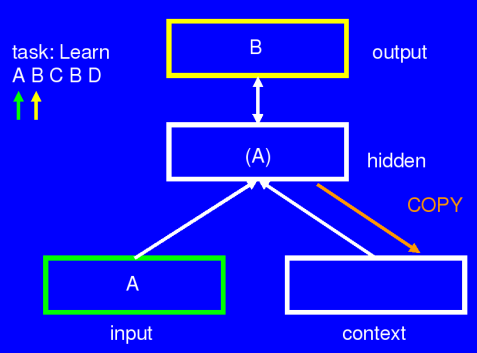
35

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



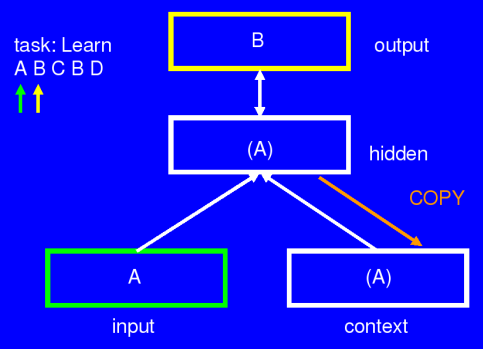
36

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



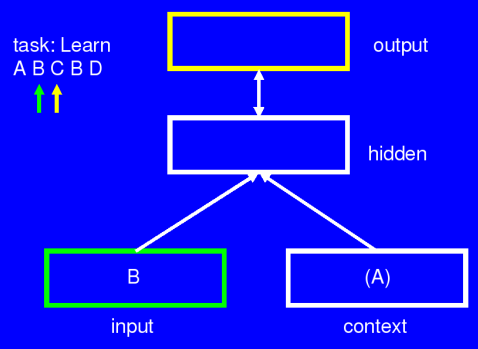
37

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



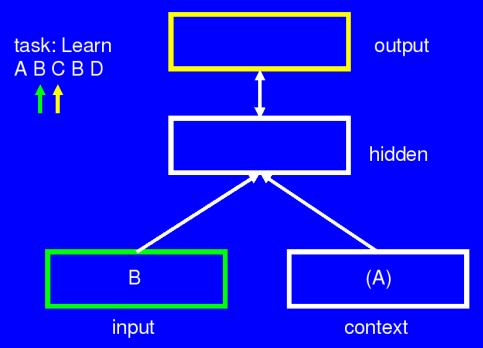
38

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



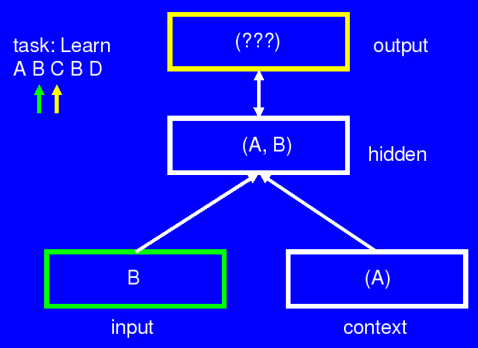
39

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



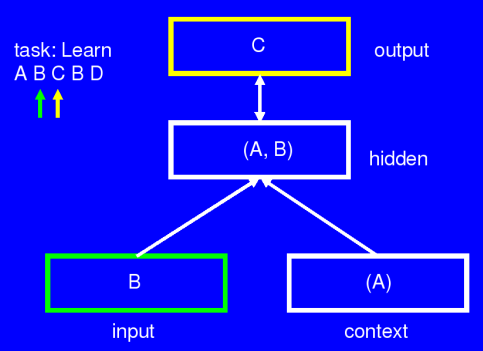
40

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



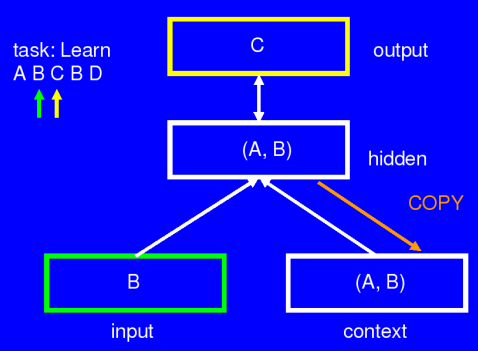
41

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



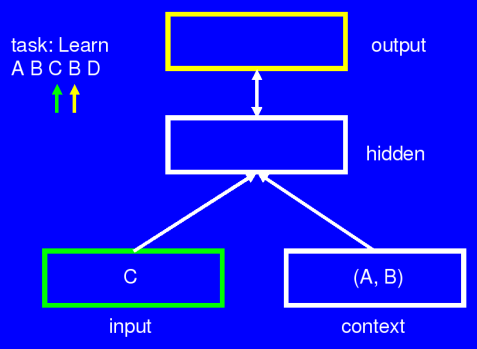
42

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



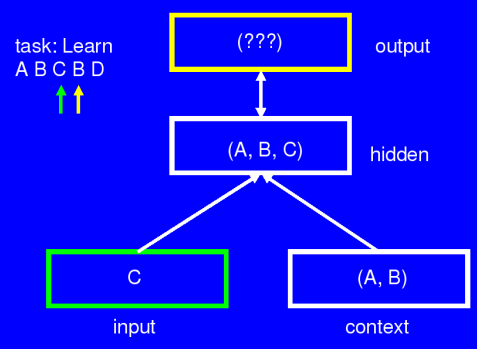
43

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



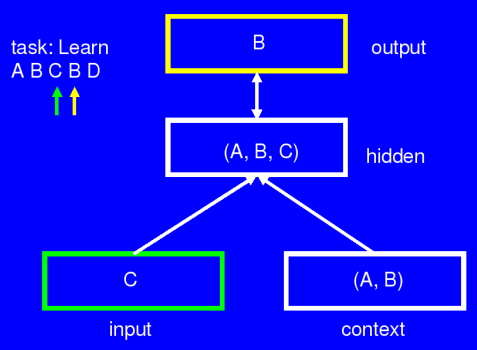
44

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



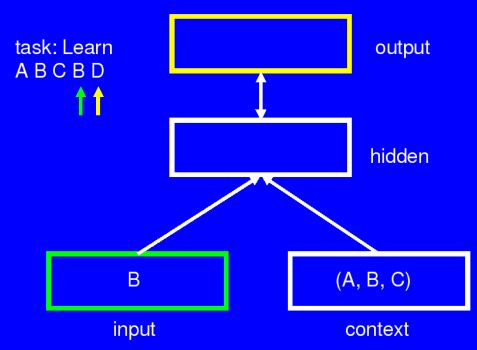
45

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



46

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



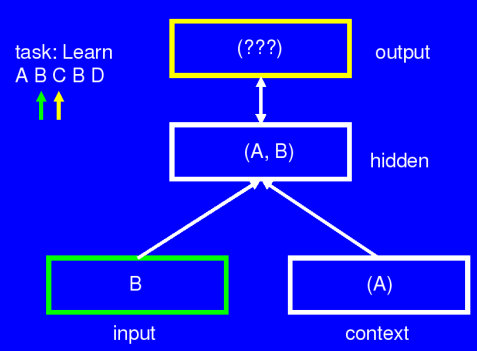
47

Why Copy the Hidden Representation?

- Copying input or output only lets the network hold on to one previous item
- Copying the hidden layer lets the network hold on to an arbitrarily large number of items – **even though it is always just copying last hidden state at time t-1.**
- The network *learns* how strongly to hold on to past items

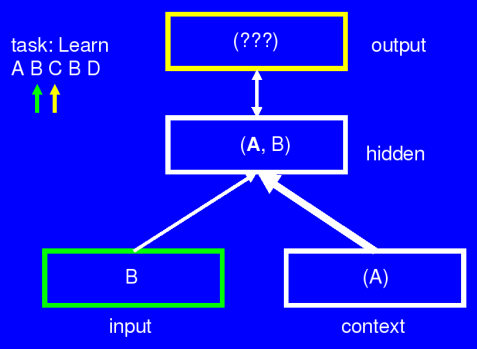
48

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



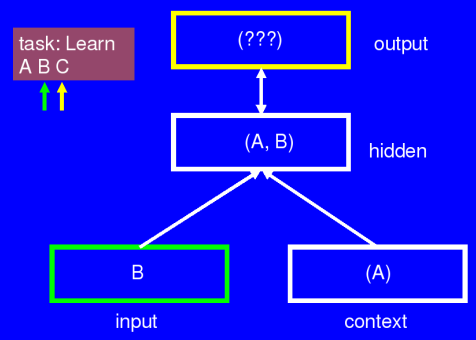
49

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



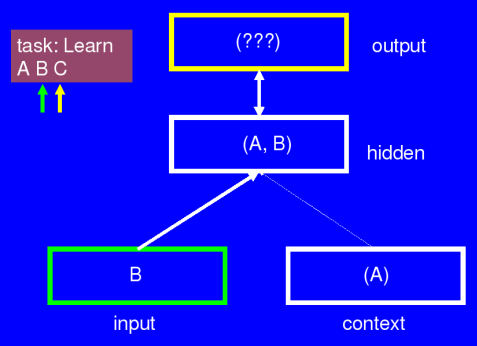
50

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



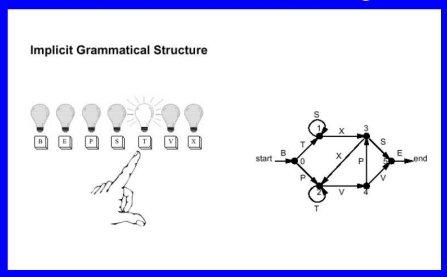
51

Simple Recurrent Network (SRN): An Architecture for Sequence Learning



52

Artificial Grammar Learning

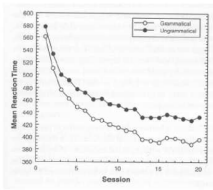


Task: Press buttons that implicitly follow the grammar

Note: Subjects do not know about the grammar!

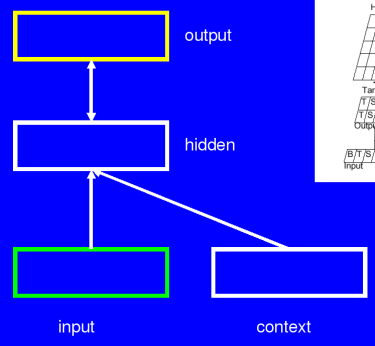
53

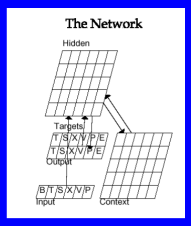
Evidence For Implicit Learning



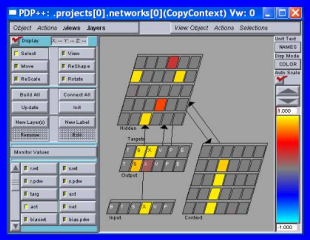
- Response times are lower for grammatical sequences.
- Learners report no knowledge of a sequential pattern.
- Learners are at chance when asked to predict the next light.

54





55



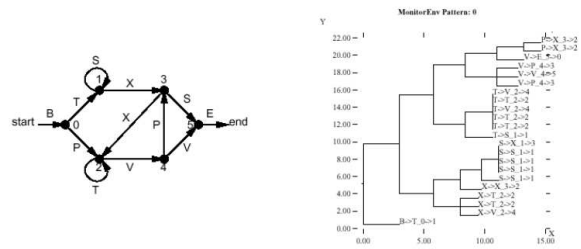
Some details:

- Input patterns: Generated (randomly) using the finite state grammar
- Sometimes there are multiple "correct" answers for what comes next. kWTA forces the network to guess one answer

56

Hidden Activity Reflects State Information

Grammatical state information, along with information concerning the last input, is available in the pattern of hidden layer activity.



57 Simple SRN story is not flawless

- How is hidden→ "copy" function implemented biologically?
- During settling, context must be *actively maintained* (ongoing hidden activity has no effect on context).
- Assumes all context is relevant: What if distracting information presented in middle of sequence? Want to only hold on to *relevant* context.

→ Stay tuned for specialized biological/computational mechanisms for *updating / gating* vs. robust *maintenance* of context.

58 Motivating Motivation

Why does anyone go to grad school?

(or, why do we ever do anything besides eat, sleep, have sex, etc)?

e.g., Why am I here today, instead of lying on a beach in Mexico, drinking mojitos and reading a good book?

Challenge: make a neural network model that decides to go to grad school!

59 The Motivational Bootstrap

Puzzle: Motivations must be built-in (else we would likely die)
But if purely fixed, there would be no art, science, etc..

Need a motivational bootstrap: *learning* on top of built-in's.

Culture & social drives provide cumulative shaping of learning.

So, why does anyone go to grad school?

- Socially-mediated standards of success.
- Strong built-in desire to be accepted by others.
- Strong built-in desire to share w/others.
- Strong built-in desire to learn (dopamine?)

60 What I'm Actually Talking About

Skinnerian learning

The basic stuff that every mammal has in common:

Neural mechanisms of Pavlovian conditioning
(from a computational perspective). No supervised target signal available: only good/bad outcomes
Enables *bootstrap* of new stimuli (CS's) onto built-in desires (US's):

CS (money) → US (food, etc) But what if consequence of given input comes *later* in time?

61 Temporally Extended Tasks With Sparse Reward

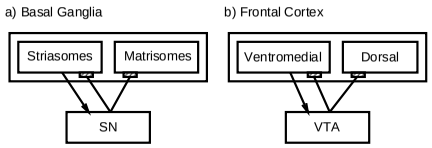
62 Temporally-delayed Learning & Reinforcement

Reinforcement often delayed from the action(s) that lead to it: need to "span the gap".

Key ideas: We want to predict rewards consistently over time. This process leads us to learn what events are associated with rewards, earlier and earlier back in time.

We use the Temporal Differences (TD) algorithm (Sutton & Barto, Dayan, Montague, ...).

63 Reinforcement Biology (somewhat outdated..)



Midbrain areas send dopamine (DA) to modulate cortex & basal ganglia: Substantia Nigra (SN) & Ventral-Tegmental Area (VTA).

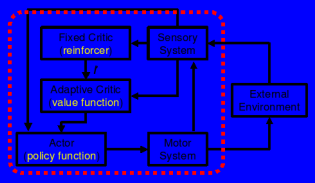
SN & VTA are controlled by specific cortical/BG areas.

These other areas are like an "Adaptive Critic" (AC), which evaluates stimuli & actions for their rewarding value..

The critic can then *train* an *Actor* system to execute rewarding actions (eg "matrisomes", dorsal striatum, dorsal prefrontal cortex).

Barto, Sutton, & Anderson (1983)

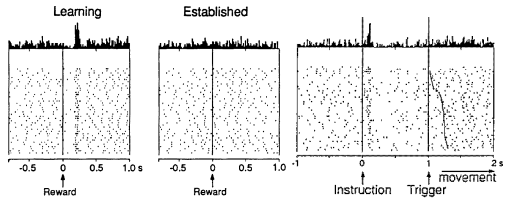
The Actor-Critic Framework



At every moment, a scalar reinforcement signal, *r(t)*, is available (though it may often be "zero").

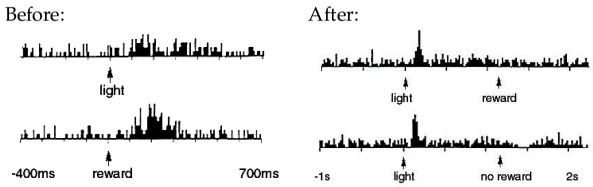
The Adaptive Critic (AC) evaluates the "goodness" of the current situation in terms of both current reward and the likelihood of future reward. The Actor selects the next action.

65 VTA dopamine neuron recordings



VTA firing moves from responding to reward to *anticipating* it at the instruction.

66 Basic Data: VTA DA Neural Firing in Conditioning



67 Temporal Difference Learning: Equations

Value function, sum of discounted future rewards:

$$V(t) = \langle \gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) ... \rangle \quad (1)$$

Recursive definition:

$$V(t) = \langle r(t) + \gamma V(t+1) \rangle \quad (2)$$

Error in predicted reward (from previous to next time-step):

$$\delta(t) = \left(r(t) + \gamma \hat{V}(t+1) \right) - \hat{V}(t) \quad (3)$$

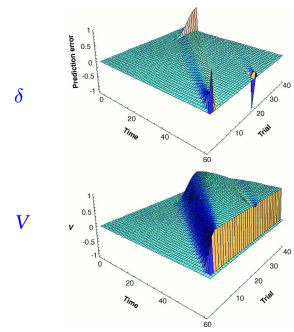
Update value estimate:

$$\hat{V}(t) \leftarrow \hat{V}(t) + \alpha \delta(t) \quad (4)$$

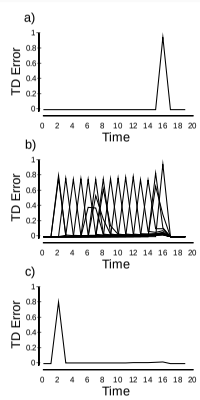
α = learning rate

68 TD and Dopamine Relationship

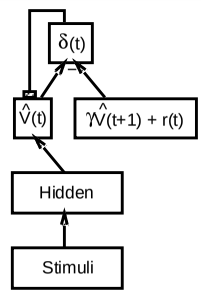
Schultz, Dayan & Montague, 1996, *Science*



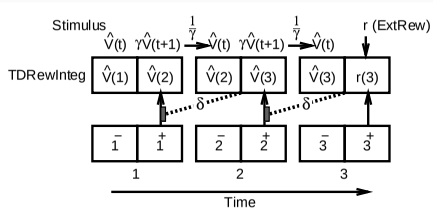
69 Model: CS at t=2, US at t=16



70 Network Implementation



71 Phase-based Implementation

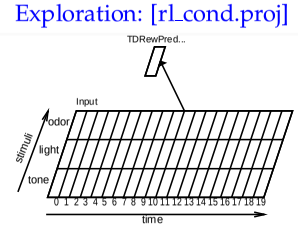


TDRewInteg = TDRewPred + ExtRew

Minus phase: TDRewInteg clamped to prev plus phase value.

Plus phase: TDRewInteg settles via weights
= expected reward at t+1, plus any ExtRew at time t. Learning signal δ (= "TD") trains prediction for *previous* time step.
(*eligibility traces* needed)

72 Exploration: [rl_cond.proj]



'Complete Serial Compound' (CSC) input representation:
unique unit for each stimulus at each time point
(used in Sutton & Barto, Montague et al, etc)

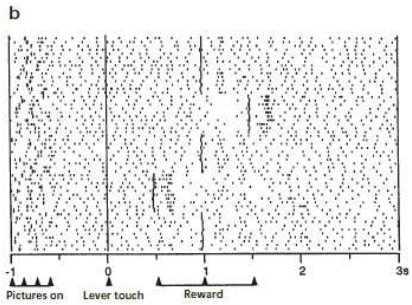
Not realistic, but good for demonstration. This assumption can be relaxed without changing core ideas (e.g. Ludvig et al, 2008).

73 Exploration: [rl_cond.proj]

Standard TD: $\hat{V}(t) = \sum_i w_i x_i(t)$ [x_i are inputs: tone, light]

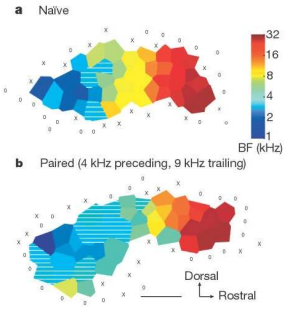
Here: passed thru activation function – has to surpass threshold, subject to inhibitory competition from other value reps

74 DA and Timing: Late and Early Rewards

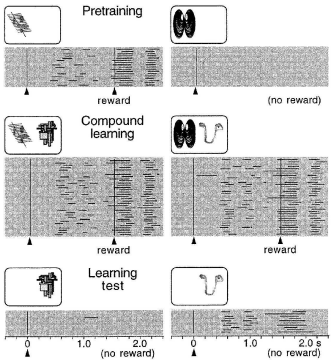


75 DA and Learning: Auditory Cortex

Bao et al, 2001, *Nature*

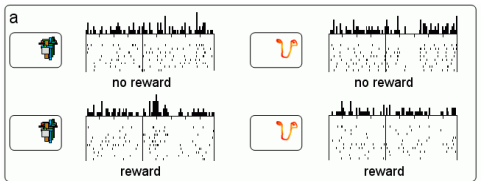


76 Learning Theory: Blocking (Behavior)



77 Learning Theory: Blocking (Dopamine)

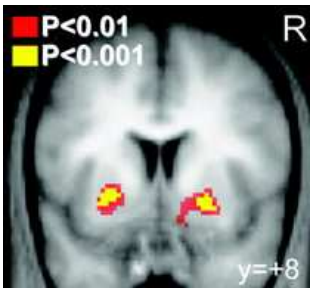
Waelti et al, 2001, *Nature*



Blocked stimulus Control (not blocked) stim

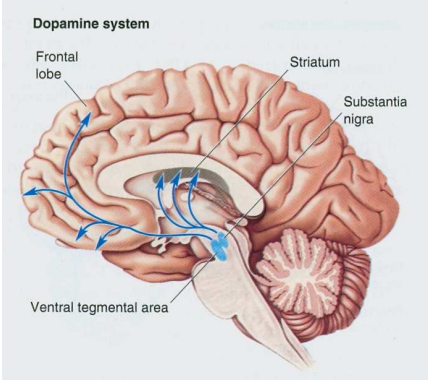
78 TD prediction error and human functional imaging

O'Doherty et al, 2004, *Science*

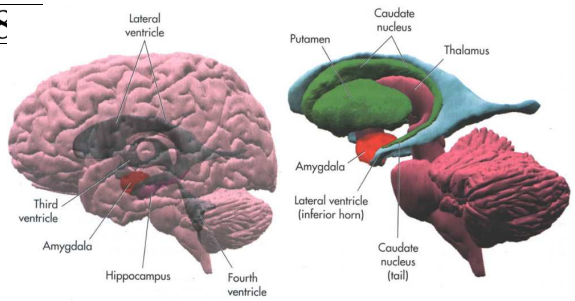


Ventral striatum = DA enriched, correlates with TD PE=Critic!
Dorsal str only activated when *Actor* involved (instrumental)

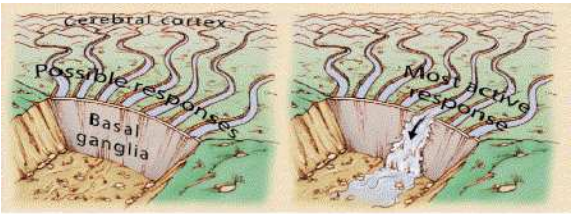
79



8



81 The Basal Ganglia and Action Selection



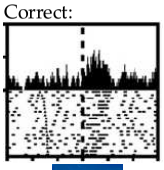
(figure borrowed from Ivry & Spencer, 2004)

- BG selectively facilitates one command while suppressing others (Mink, 1996; Frank et al, 2001; Gurney et al, 2001; Brown et al, 2004)
- Occurs in parallel for multiple BG-frontal circuits: from motor to cognitive actions (Frank, 2005; O'Reilly & Frank, 2006; Houk, 2005)

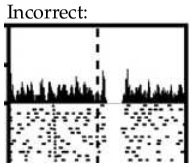
82 Reinforcement Learning and Dopamine

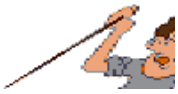
Satoh et al (2003): DA bursts/dips depending on monkey's *action* (not just Pavlovian)

Correct:

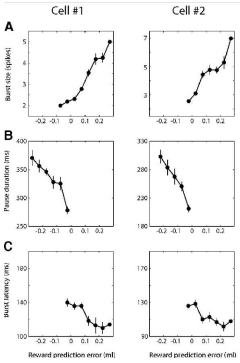


Incorrect:



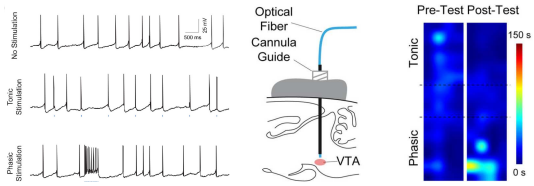


83 Burst/Pause correlations with Rew Prediction Errors



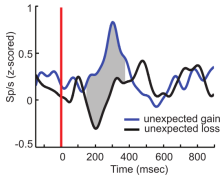
Bayer et al, 2007, JNeurophys

84 Optical phasic DA stimulation causally induces conditioning



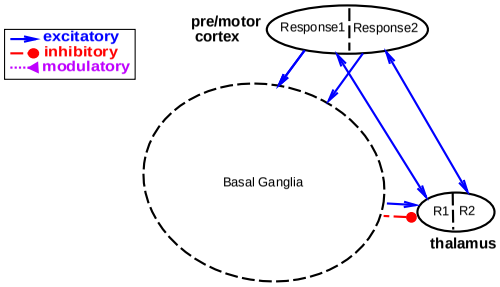
(Tsai et al, 2009, Science)

85 DA neuron spiking during reinforcement task in humans

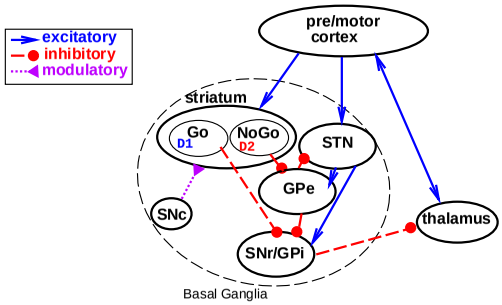


(Zaghoul et al, 2009, Science)

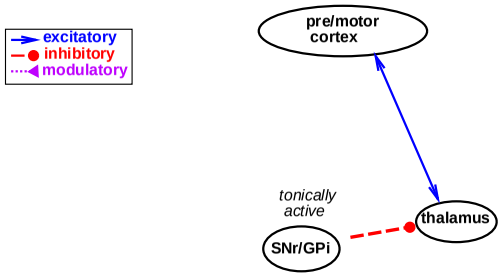
86 Striato-Cortical Functional Circuitry



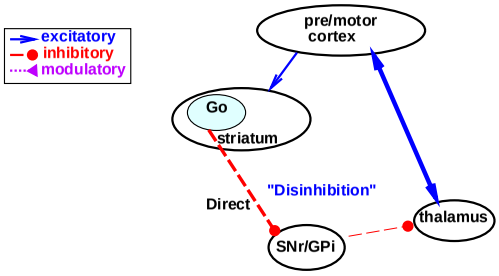
87 Striato-Cortical Functional Circuitry



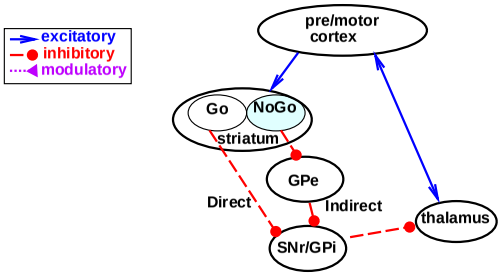
88 Striato-Cortical Functional Circuitry



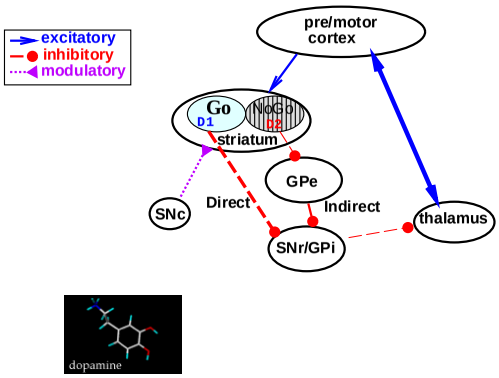
89 Striato-Cortical Functional Circuitry



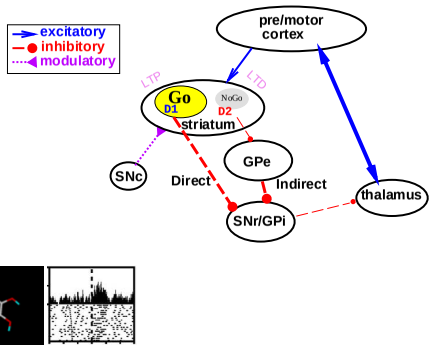
90 Striato-Cortical Functional Circuitry



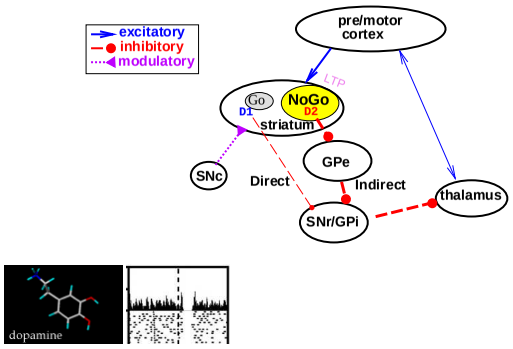
91 Striato-Cortical Functional Circuitry



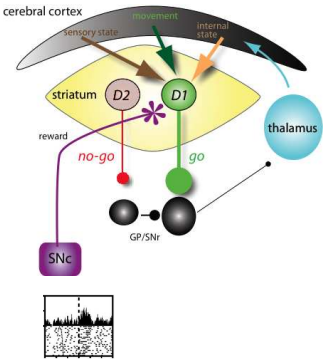
92 DA Feedback Effects on BG Learning



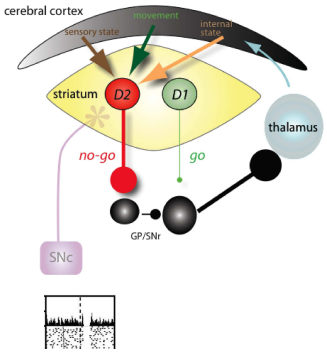
93 DA Feedback Effects on BG Learning



94 DA effects on BG Learning: Positive PE



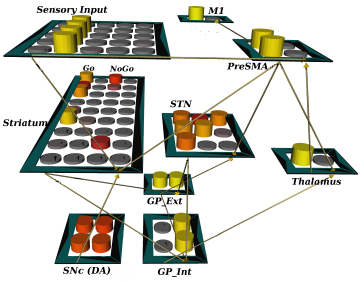
95 DA effects on BG Learning: Negative PE



See Shen et al, 2008, *Science* for evidence of dual D1/D2 plasticity rules

96 Neural Model of BG and DA

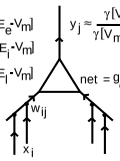
Integrates a wide range of physiological data into a single coherent framework; can then explore dynamics in learning & decision making



$$c\dot{V}_m = g_e\bar{g}_e[E_e - V_m] + g_l\bar{g}_l[E_l - V_m] + g_I\bar{g}_I[E_I - V_m] + \ldots$$

$$y_j \approx \frac{\gamma[V_m - \Theta]_+}{\gamma[V_m - \Theta]_+ + 1}$$

$$net = g_e \approx \langle x_i w_{ij} \rangle + \frac{\beta}{N}$$

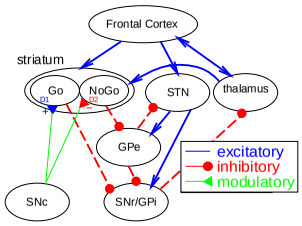


$$\Delta w_{ij} \approx (x_i^p y_j^p) - (x_i^t y_j^t)$$

Frank, 2005, 2006 *J Cog Neurosci, Neural Networks*

97

Some Supporting Evidence

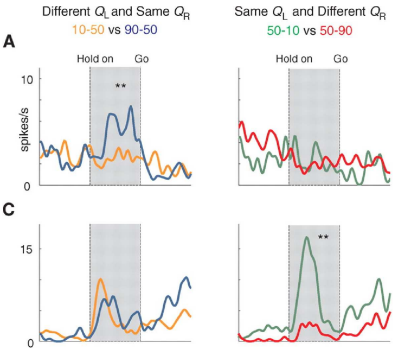


* SNr activity (Jiang, Stein & McHaffie, 2003):

"Go" "NoGo"

98

Samejima et al, 2005, Science



Striatal coding for action-values: Go and NoGo cells?

99

BG Model movies

Selection & Positive Feedback

Selection & Negative Feedback

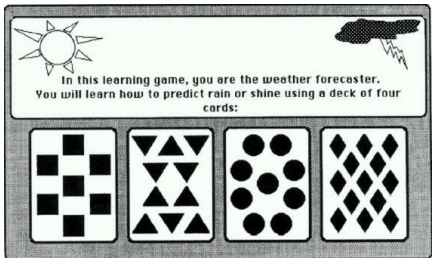
After more extended training: Cortical response selection

Response Sequencing Trial

See these movies and explanations at
`http://ski.cog.brown.edu/BGmodel_movies.html`

100

Application to probabilistic reinforcement learning: Weather Prediction Task



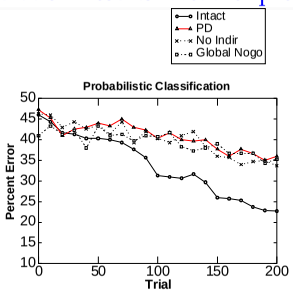
101

Application to probabilistic reinforcement learning: Weather Prediction Task

Probability Structure of the Weather Prediction Task

Cue	Frequency	Sunshine	Rain	P(Sun)
1 1 0 0	4	4	0	1.00
1 1 1 0	1	1	0	1.00
1 0 0 0	7	6	1	.86
1 0 1 0	4	3	1	.75
0 1 0 0	5	3	2	.60
1 0 0 1	2	1	1	.50
0 1 1 0	2	1	1	.50
1 0 1 1	2	1	1	.50
1 1 0 1	2	1	1	.50
0 0 1 0	5	2	3	.40
0 1 0 1	4	1	3	.25
0 0 0 1	7	1	6	.14
0 1 1 1	1	0	1	.00
0 0 1 1	4	0	4	.00
Total	50	25	25	

102 PD Simulation Results: Weather prediction



- Intact nets extracted probabilistic structure by resolving differences in Go/NoGo representations.
- PD nets were impaired due to reduced dynamic range of DA.

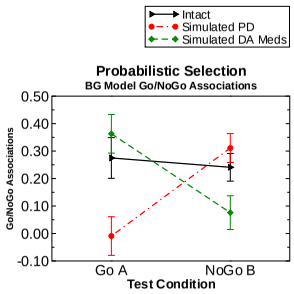
Frank, 2005, *J Cog Neurosci*

103 What about Dopaminergic Medication Effects in PD?

- Medication actually *causes* deficits in some tasks: Probabilistic Reversal (Cools et al, 2002).
- Hypothesis: medication blocks the effects of DA dips needed to learn "NoGo".

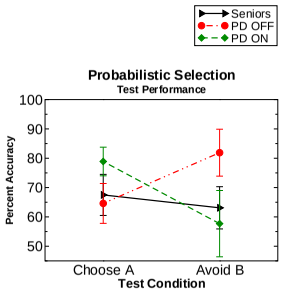
normal DA dip medication

104 BG Model Go/NoGo associations, Predictions for Parkinson's and Medication



PD ON: Go A PD OFF: NoGo B

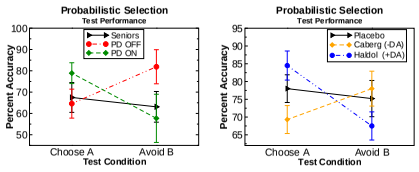
105 Parkinson's Behavioral Results: Probabilistic Selection



Frank, Seeberger & O'Reilly, 2004, *Science*

(See Cools et al, 2006 for similar findings)

106 Converging Evidence: PD & DA drugs in Healthy Participants



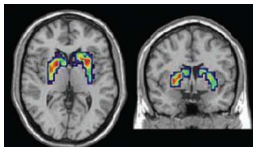
Frank et al 2004, *Science*

Frank & O'Reilly (2006), *Behavioral Neuroscience*

107 Testing the models: Striatal Genetic Contributions

Striatal Genes: DARPP-32 and DRD2

- *DARPP-32*: protein concentrated in striatum, activated by *D1* receptor stimulation, mediates DA effects on plasticity (Greengard et al, 1999).

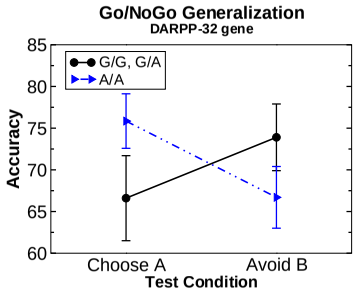


Meyer-Lindenberg et al, 2007

Model Prediction: *DARPP-32 gene modulates Go learning.*

108 DARPP-32 Modulates Relative Go learning

N=69



Frank, Moustafa, Haughey, Curran & Hutchison (2007), *PNAS*

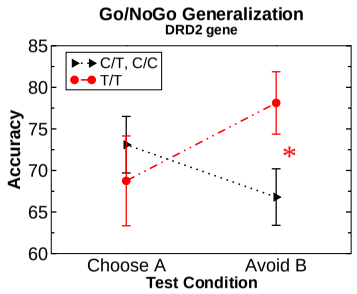
109 Source of Individual Differences

Striatal Genes: DARPP-32 and DRD2

- *DRD2 gene: D2 receptor in striatum* (Hirvonen et al, 2004, 2005)
  - slow integration of BG NoGo signals
  - supports avoidance of most aversive stimulus "on average"
    → *Model Prediction*: *Avoid-B test phase performance*

110 Sources of Individual Difference: DRD2 Gene

N=69



Frank, Moustafa, Haughey, Curran & Hutchison (2007), *PNAS*

(See Klein et al, 2007 for similar findings)

111 Want to learn more?

- Check out BasalGanglia.prt.pdf slides on cogsim website
- Check out `http://ski.cog.brown.edu/BG_Projects/` for a few Emergent simulations replicating published effects. (papers avail on my website). One possibility to explore for final project..
- Later: BG interactions with PFC in higher level cognition

112 Extra

The following slides describe a recently developed alternative to TD, called PVLV, which we think is more biologically plausible and computationally powerful. This material is optional for the course.

113 The Problem

Q: How do we learn to attach positive/negative valence to environmental stimuli?

A: The same way we learn lots of other stuff: the Delta Rule!

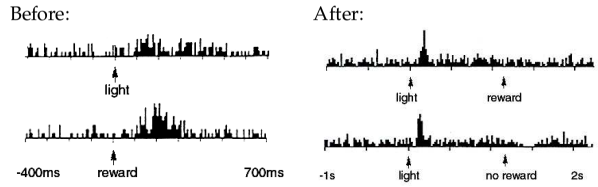
$$\delta_{pv} = r - \widehat{V}_{pv}$$

$\widehat{V}_{pv}$: expected reward based on prior associations
r: reward
δ_{pv}: learning signal

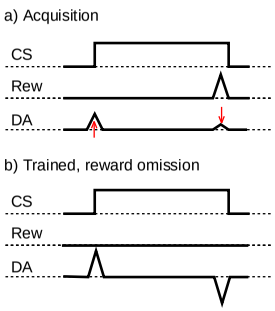
114 The Problem

Q: But what happens when environmental stimulus occurs *before* reward?

115 Basic Data: VTA DA Neural Firing in Conditioning



116 Basic Data: VTA DA Neural Firing in Conditioning



Dopamine spikes/dips are learning signals Delta rule fails to account for predictive DA spike!

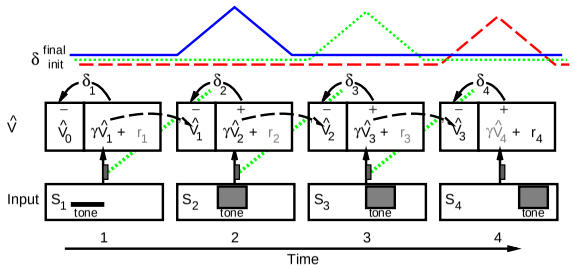
117 Standard Approach: TD

Predict all future rewards (discounted):
$V_t = \sum_{\tau=t+1}^{\tau=\infty} \gamma^{\tau-(t+1)} r_\tau$

Recursively:
$\hat{V}_{t-1} = r_t + \gamma \hat{V}_t$

Error = Temporal Difference = TD:
$DA = \delta_t = [r_t + \gamma \hat{V}_t] - \hat{V}_{t-1}$

118 TD Illustrated

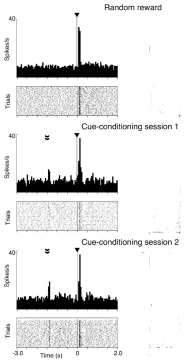


119 Problems with TD

- Great algorithm, developed in computer science / machine learning, but is this actually what the brain does?
- Even if so, doesn't specify how these signals are computed by systems upstream of DA... just predicts DA and δ but says nothing about V, etc.
- Current reward value is always relative to what happened just before. Too much temporal dependency?
- Chaining not seen in neural recordings.
- What determines "discount factor" γ, biologically?

120 Rat study: simultaneous CS and US DA spike

Pan et al, 2005, *Journal of Neuroscience*



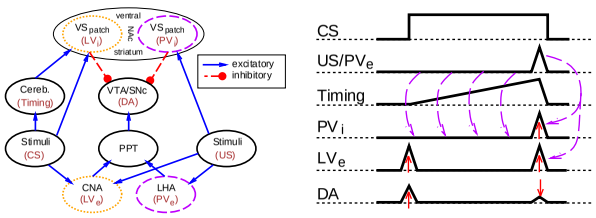
Inconsistent with standard TD!

121 The PVLV Alternative

PVLV = Primary Value, Learned Value
(O'Reilly, Frank, Hazy & Watz, 2007, Behav Neurosci)

- No reward predictions, just *associations*!
- No temporal dependencies: DA depends only on current state.
- Uses same basic delta-rule learning as TD (Rescorla-Wagner).

122 PVLV: Two Separate Mechanisms (PV, LV)



- PV (Primary Value): Primary rewards (US), canceled.
- LV (Learned Value): Learned associations (CS → DA).

123 PVLV: Two Separate Mechanisms (PV, LV)

PV: Primary Value

- Trained at each point in time on actual reward value present: $\delta_t = r_t - \hat{V}_t$
- This uses *immediate* prediction ($\hat{V}_t$) of current rew value (r_t)
- Accounts for canceling of DA spike @ rew, and DA dips when no rew received.
- But this doesn't account for predictive DA spikes... (actually results in predictive DA dips!)

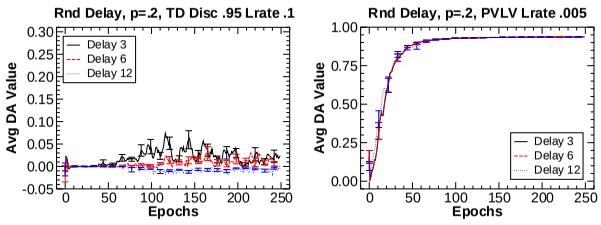
124 PVLV: Two Separate Mechanisms (PV, LV)

LV: Learned value

- Represents perceived values of stims **even when there is no current rew expectation**.
- Only gets training signal @ rew, or when PV expects some rew. (ie learning is *filtered* by primary PV system.)
- → Learns at time of rew, but not at CS onset.
- → Generalizes rew values to CS...
- → Accounts for DA spikes for stimuli that have previously been associated with reward!

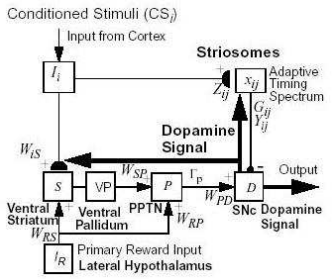
125 PVLV: Computationally Powerful

Comparison with TD on Random Delays (breaks TD chaining):



Enables working memory model to learn complex WM tasks.

126 Similar to Brown, Bullock & Grossberg, '99



Diffs: Anatomical (CNA vs. VS; Dorsal vs. Ventral Patch)
Functional (intrinsic timing? LV system cannot train itself).

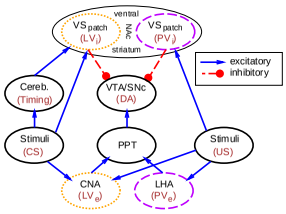
127 PVLV accounts for timing data better than TD!

- Data: during transient learning period, both rews and CS elicit activation.
- This accounted for by PV, LV systems operating in parallel.
- TD: predicts chaining back in time from rew to CS.

128 PVLV accounts for timing data better than TD!

- Data: delayed rewards cause dips @ usual time, then spikes
- This accounted for by both TD and PV.
- Data: early rewards cause spikes, then dips @ usual time
- This accounted for by PV (spike), PV (dip), but TD only accounts for spike.

129 More Key Predictions from PVLV



- CNA = Pavlovian conditioning (e.g., Killcross et al. '97).
- NAc (patch/shell) = Extinction (Ferry et al. '00; Annett et al., 89), Blocking (data?).
- NAc (matrix/core) = Basic actions (OR's, approach, avoid).
- CNA can't train itself: No 2nd order conditioning!
- BLA = 2nd order cond, uses DA-independent mechanisms (CNA/BLA double-dissoc).

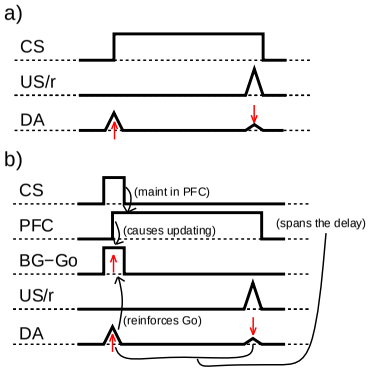
130 Conclusions

PVLV provides computationally motivated architecture that seems to fit with biology & behavioral data.

These learning mechanisms enable *arbitrary* stimuli/goals to be plugged into our fixed set of built-in motivational drives.

Something motivates every generated mental-state, always!

131 PVLV, WM, and DA



132 PVLV: Two Separate Mechanisms (PV, LV)

PV learning:

$\delta_{pv} = r - \hat{V}_{pv}$ – or – $\delta_{pv} = PV_e - PV_i$

$\Delta w_i = \epsilon x_i \delta_{pv}$

LV learning (filtered by PV):

$$\Delta w_i = \begin{cases} \epsilon(r_t - \hat{V}_{lv})x_i & \text{if } \hat{V}_{pv} > \theta_{pv} \text{ or } r_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Global DA (PV dominates):

$$\delta_t = \begin{cases} \delta_{pv} & \text{if } \hat{V}_{pv} > \theta_{pv} \text{ or } r_t > 0 \\ \delta_{lv} & \text{otherwise} \end{cases}$$

$\delta_{lv} = LV_e - LV_i$

133 LV Extras

- DA spikes only observed @ CS onset, don't continue throughout delay until reward. Problem for PV?
- Solution: PV system has synaptic depression, accommodates to constant sensory inputs; only perceives values of stims that were not present in last time step.
- This is also important for PFC learning.. (stay tuned)