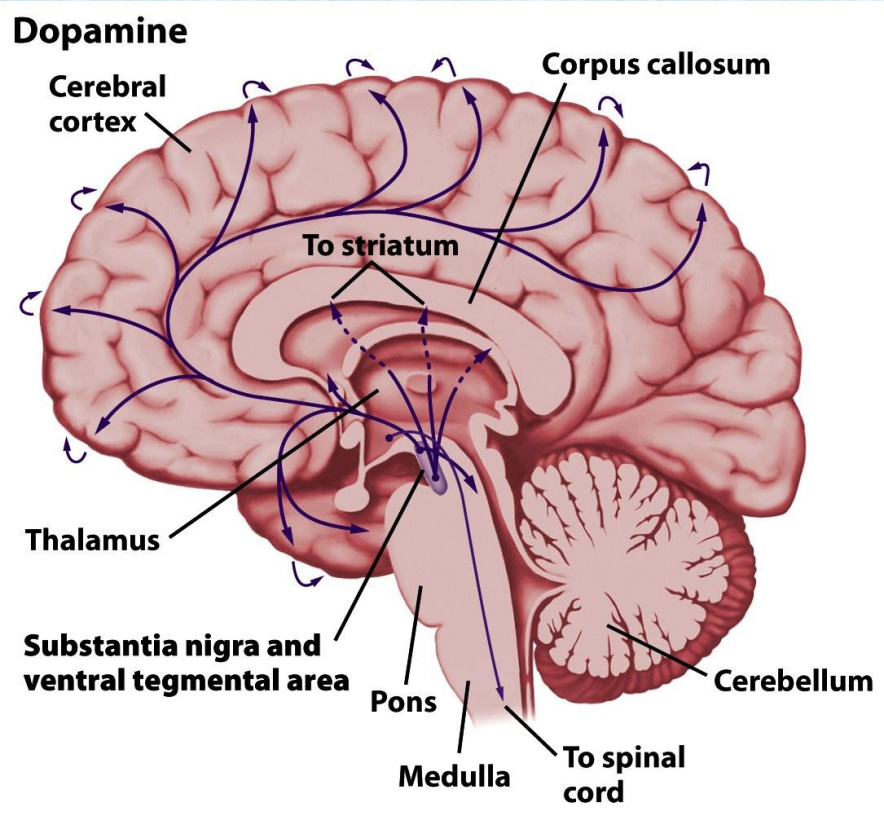
CN510: Principles and Methods of Cognitive and Neural Modeling

Dopaminergic Systems, Reinforcement, and Predicting Reward

Lecture 22

Instructor: Anatoli Gorchetchnikov <anatoli@bu.edu>
Teaching Fellow: Rob Law <nosimpler@gmail.com>

Dopaminergic System



Originate from dopaminergic cells in the substantia nigra pars compacta (SNc) and ventral tegmental area of the midbrain

Project to striatum (part of basal ganglia), parietal and frontal lobes of cerebral cortex

Classical and Instrumental Conditioning

Classical:

– Reward is delivered no matter what the animal does

Instrumental:

– Reward delivery requires action from the animal

Pavlovian		CS →R	CS →R’
Extinction	CS →R	CS →.	CS →.’
Partial		CS →R CS →.	CS →αR’
Blocking	CS_1 →R	CS_1+CS_2 →R	CS_1 →R’ CS_2 →.’
Inhibitory		CS_1+CS_2 →. CS_1→R	CS_1 →R’ CS_2 →-R’
Overshadow		CS_1+CS_2 →R	CS_1 →α_1R’ CS_2 →α_2R’
Secondary	CS_1 →R	CS_2 →CS_1	CS_2 →R’

Reinforcement Learning

Learning based solely on reward and punishment

Concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward

Input is not a teaching signal telling what the output shall be but rather a positive or negative reinforcement signal

DA signal in classic interpretation fits perfectly as this reinforcement signal

The one of the dominant theories of reinforcement learning is based on Temporal Difference rule (*Barto et al., 1983*)

Rescorla-Wagner Rule

We would consider a simple neuron based on

$$y(t) = w(t)x(t)$$

so prediction is neuronal output for given input *x(t)*

We want the weights to change so that they compensate for errors:

$$\Delta w(t) = \eta x(t)\delta(t) = \eta x(t)\left(R(t) - y(t)\right)$$

The equilibrium value of weight ensures $y_\infty = \langle R \rangle$

Therefore, the output of a neuron *y* learns to predict reward

The dynamics of R-W is the same as presynaptically gated decay (or leaky integrator if input is binary):

$$\Delta w(t) = \eta x(t)\left(R(t) - x(t)w(t)\right) = \eta x(t)R(t) - \eta x^2(t)w(t)$$

Compare with $\dot{w} = \eta x(t)\left(y(t) - w(t)\right)$

Dopamine Neurons as Reward Predictors

Schultz et al. (1997)

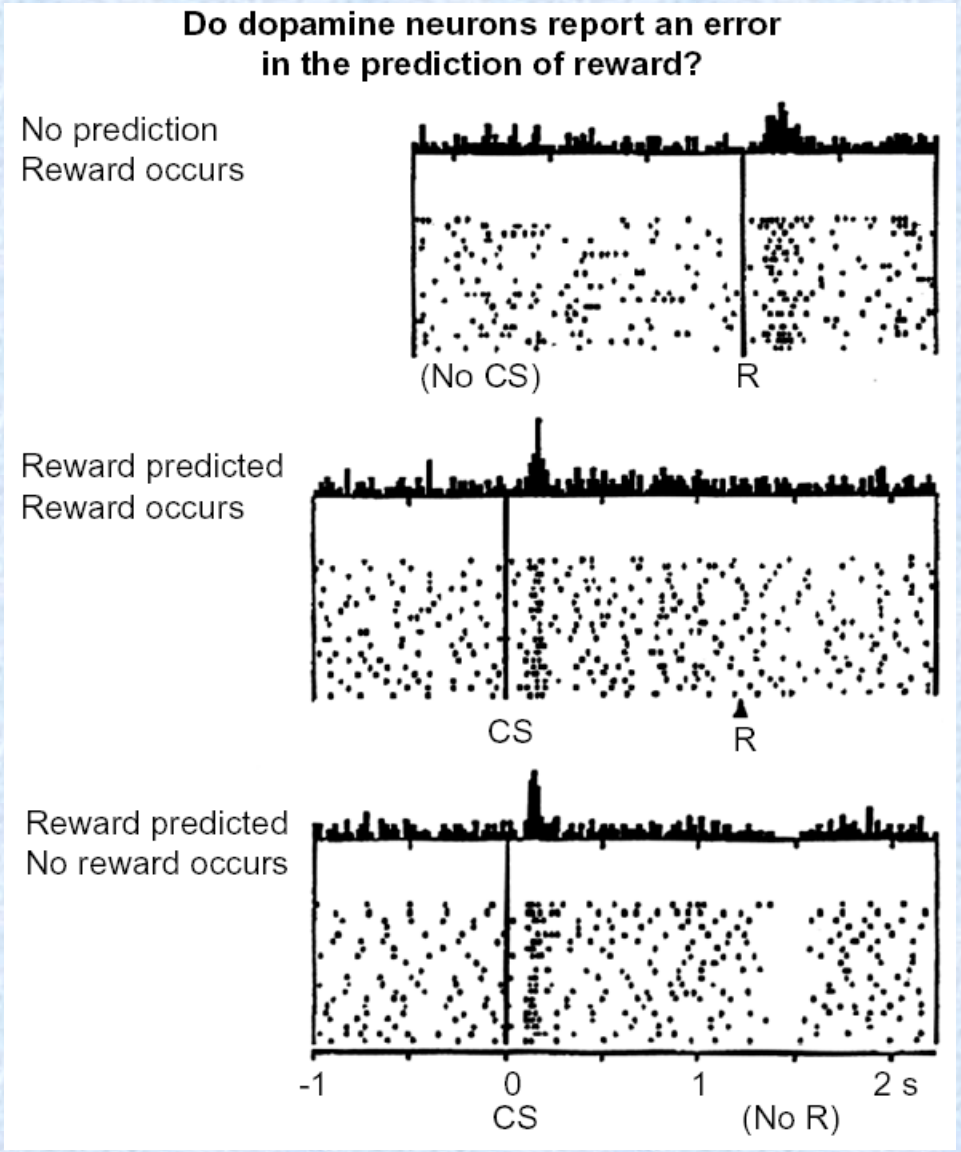
Summarized findings from multiple dopamine studies in the substantia nigra pars compacta (SNc)

Panel 1 – beginning of conditioning

Panel 2 – end of conditioning

Panel 3 – test trial with no US

New research suggests that DA signals a desire for a reward in addition to expectation



Rescorla-Wagner Rule: Blocking

What if we have multiple stimuli?

$$y(t) = \sum_i w_i(t) x_i(t)$$

so prediction is neuronal output for all given inputs $x_i(t)$

The learning rule becomes:

$$\Delta w_i(t) = \eta x_i(t) \delta(t) = \eta x_i(t) \left(R(t) - y(t) \right)$$

Note that $\delta(t)$ does not change, this allows to account for blocking:

- Associating stimulus 1 with reward makes w_1 such that $\delta(t)=0$
- Consecutive pairing of stimuli 1 and 2 with reward carries no learning as stimulus 1 predicts it fully and $\delta(t)$ is still 0

Rescorla-Wagner Rule: Summary

Inhibitory conditioning and overshadowing follow from the same property: sharing the weights to get cumulative reward prediction

Overshadowing caveat: experimental stimuli usually have different saliencies, learning rates have to account for that

Pavlovian		$CS \rightarrow R$	$CS \rightarrow R'$	√
Extinction	$CS \rightarrow R$	$CS \rightarrow .$	$CS \rightarrow .'$	√
Partial		$CS \rightarrow R$ $CS \rightarrow .$	$CS \rightarrow \alpha R'$	√
Blocking	$CS_1 \rightarrow R$	$CS_1+CS_2 \rightarrow R$	$CS_1 \rightarrow R'$ $CS_2 \rightarrow .'$	√
Inhibitory		$CS_1+CS_2 \rightarrow .$ $CS_1 \rightarrow R$	$CS_1 \rightarrow R'$ $CS_2 \rightarrow -R'$	√
Overshadow		$CS_1+CS_2 \rightarrow R$	$CS_1 \rightarrow \alpha_1 R'$ $CS_2 \rightarrow \alpha_2 R'$	√
Secondary	$CS_1 \rightarrow R$	$CS_2 \rightarrow CS_1$	$CS_2 \rightarrow R'$	X

Rescorla-Wagner Rule: Problems

$$\Delta w_i(t) = \eta x_i(t)\delta(t) = \eta x_i(t)\big(R(t) - y(t)\big)$$

Requires stimulus and reward to be simultaneously present: no trace conditioning

(Recall that Grossberg solved it by having leaky dynamics of a stimulus presentation, but this is not good either as neuronal time constants are too short for trace delays)

Secondary conditioning is impossible,

- No stimulus to stimulus associations, rather competition
- Temporal aspect is missing

In general: timing is missing from the rule

Dopamine Neurons as Reward Predictors

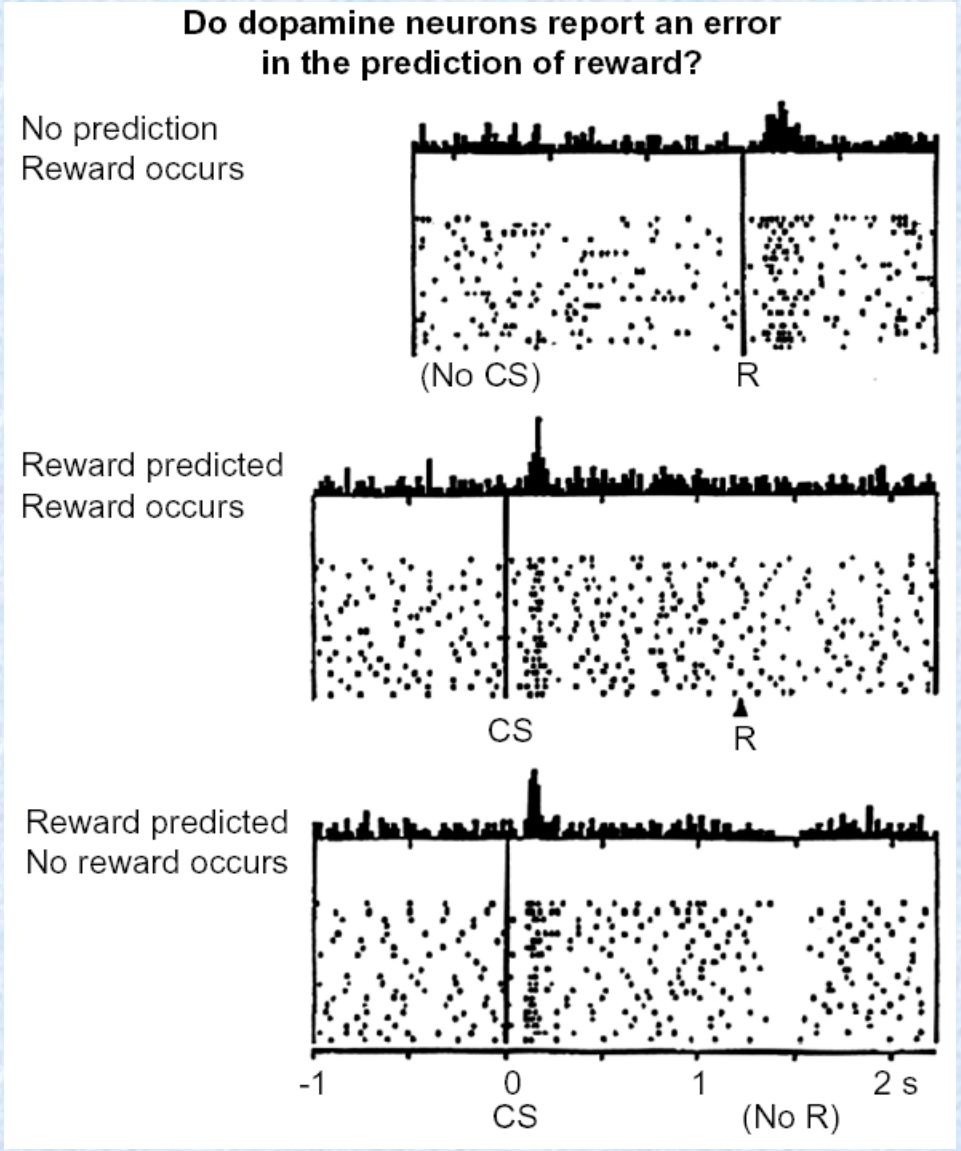
Schultz et al. (1997)

Panel 1 – beginning of conditioning

Panel 2 – end of conditioning

Panel 3 – test trial with no US

Note how in panel 3 the absence of reward causes dopamine neuron to decrease firing at the precise time when reward was expected



What Does Dopamine Signal Mean?

Suggestion 1:

Reward delivered at certain time

Not really –

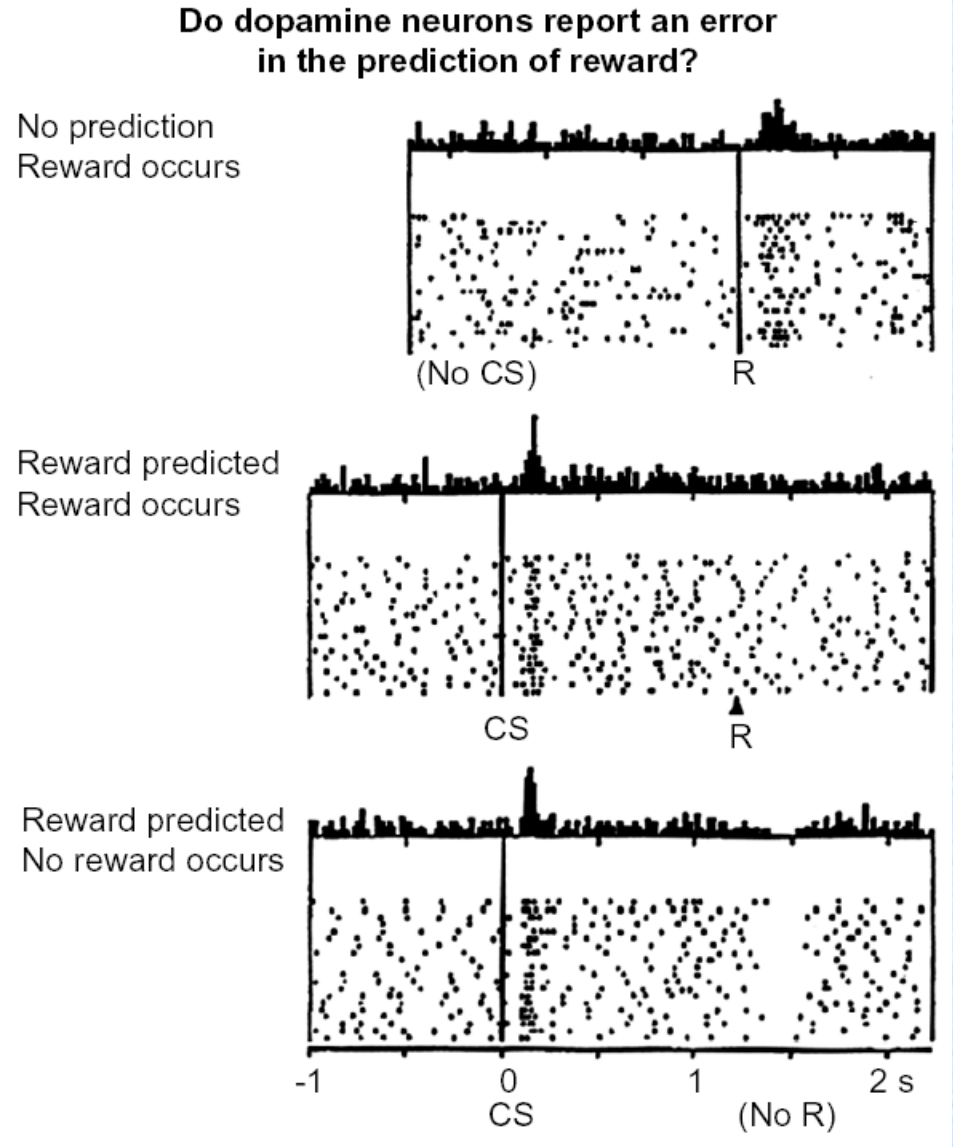
- panel 2 has no timing

Suggestion 2:

Cumulative current and future reward

Possibly –

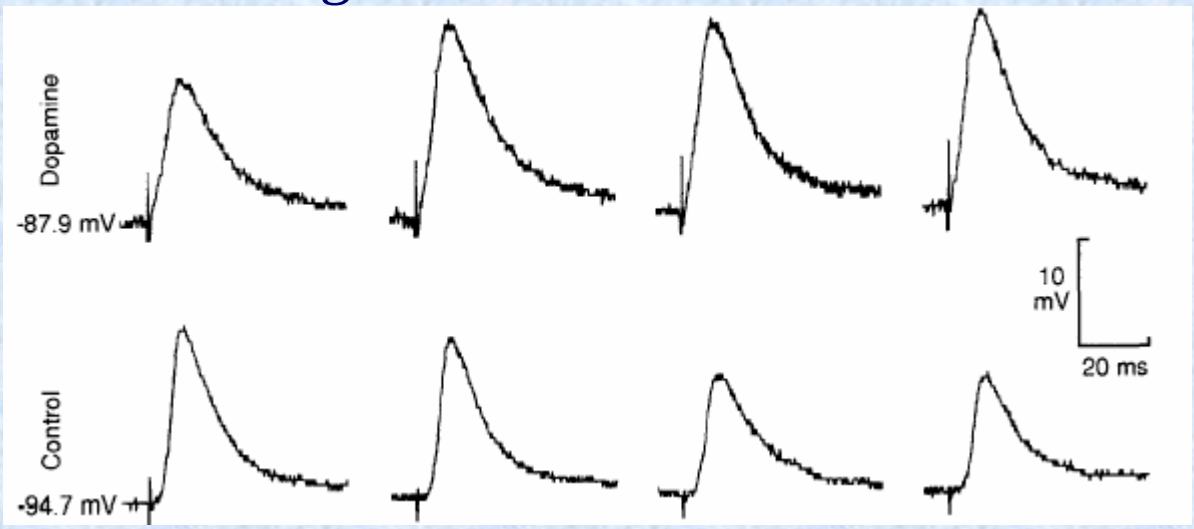
- panel 1 burst means “now is good”,
- panel 2 means “it will be good soon”,
- panel 3 means “it will be good soon”, and then “oops, I was wrong”

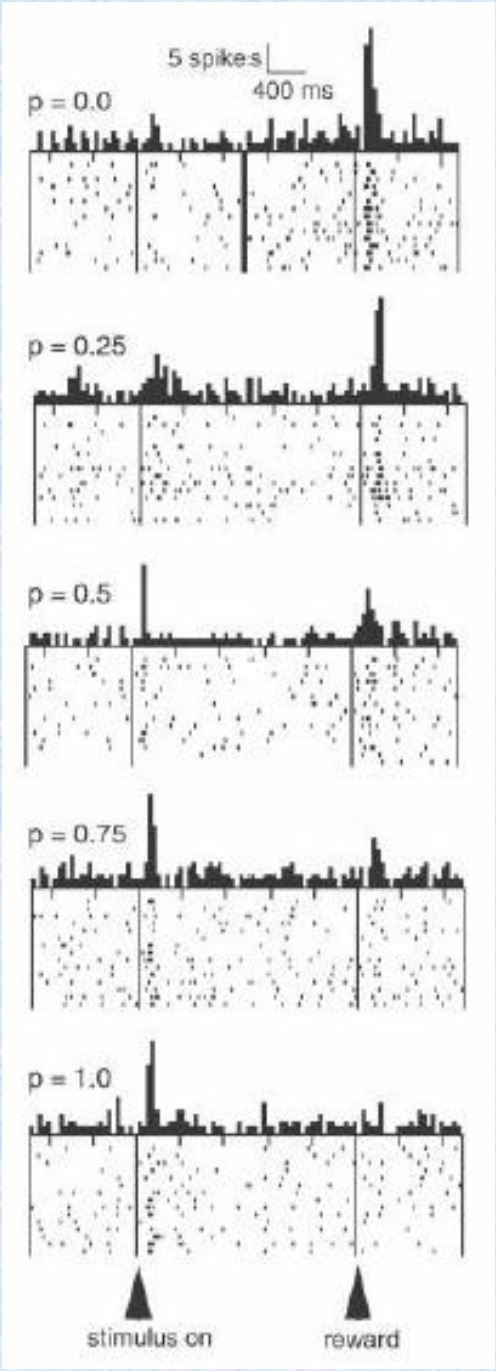


In the case of probabilistic reward delivery:

- High probability of prediction – early burst
- Low predictive level of the stimulus – burst at reward delivery

Dopamine enhances cortico-striatal learning





Temporal Difference Rule: Cumulative Reward

Start with the same neuron as for Rescorla-Wagner

$$y(t) = w(t)x(t)$$

Still using a delta rule:

$$\Delta w(t) = \eta\delta(t) = \eta\left(R(t) - y(t)\right)$$

But now *R(t)* is a cumulative future reward

We don't know it yet!

Note that I dropped the gating of the learning by presynaptic signal for simplicity

What Awaits Us in the Future?

All future rewards consist of a sum of rewards at every moment in the future:

$$R(t) = r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots = \sum_{i=0} \gamma^i r(t+i)$$

It can be rewritten as

$$R(t) = r(t) + \gamma\left(r(t+1) + \gamma r(t+2) + \ldots\right) =$$

$$= r(t) + \gamma \sum_{i=1} \gamma^{i-1} r(t+i) = r(t) + \gamma R(t+1)$$

so the reward at every moment is the sum of reward at this moment and cumulative reward from here on

$$R(t) = r(t) + \gamma R(t+1)$$

Temporal Difference Rule

Using

$$\Delta w(t) = \eta\big(R(t) - y(t)\big) = \eta\big(R(t) - x(t)w(t)\big)$$

we can compute a new value of weight as

$$w(t+1) = w(t) + \eta\delta(t) = w(t) + \eta R(t) - \eta x(t) w(t) =$$

$$= \big(1 - \eta x(t)\big) w(t) + \eta R(t)$$

We can substitute our expression for reward here:

$$w(t+1) = \big(1 - \eta x(t)\big) w(t) + \eta\big(r(t) + \gamma R(t+1)\big)$$

But we still don't know *R(t+1)*

If You Don't Know the Future – Predict It!

$$w(t+1) = (1-\eta x(t))w(t) + \eta(r(t+1) + \gamma R(t+1))$$

For that to work we need to wait for all predicted rewards to happen so

we replace the actual future reward with its prediction:

$$w(t+1) = (1-\eta x(t))w(t) + \eta(r(t) + \gamma y(t))$$

For the case with presynaptic gating we will have

$$w(t+1) = (1-\eta x^2(t))w(t) + \eta x(t)(r(t) + \gamma y(t))$$

Temporal Difference Rule

$$w(t+1) = \left(1 - \eta x^2(t)\right) w(t) + \eta x(t)\left(r(t) + \gamma y(t)\right)$$

Intuitively we use future predictions to correct current predictions

What makes future predictions better?

Over the course of many trials later predictions tend to become accurate faster (reward at the end of experiment)

Closely related to Markov decision processes with prediction being a value function and update rule corresponding to dynamic programming

Temporal Difference Rule: Where Is the Timing?

$$w(t+1) = \left(1 - \eta x^2(t)\right) w(t) + \eta x(t)\left(r(t) + \gamma y(t)\right)$$

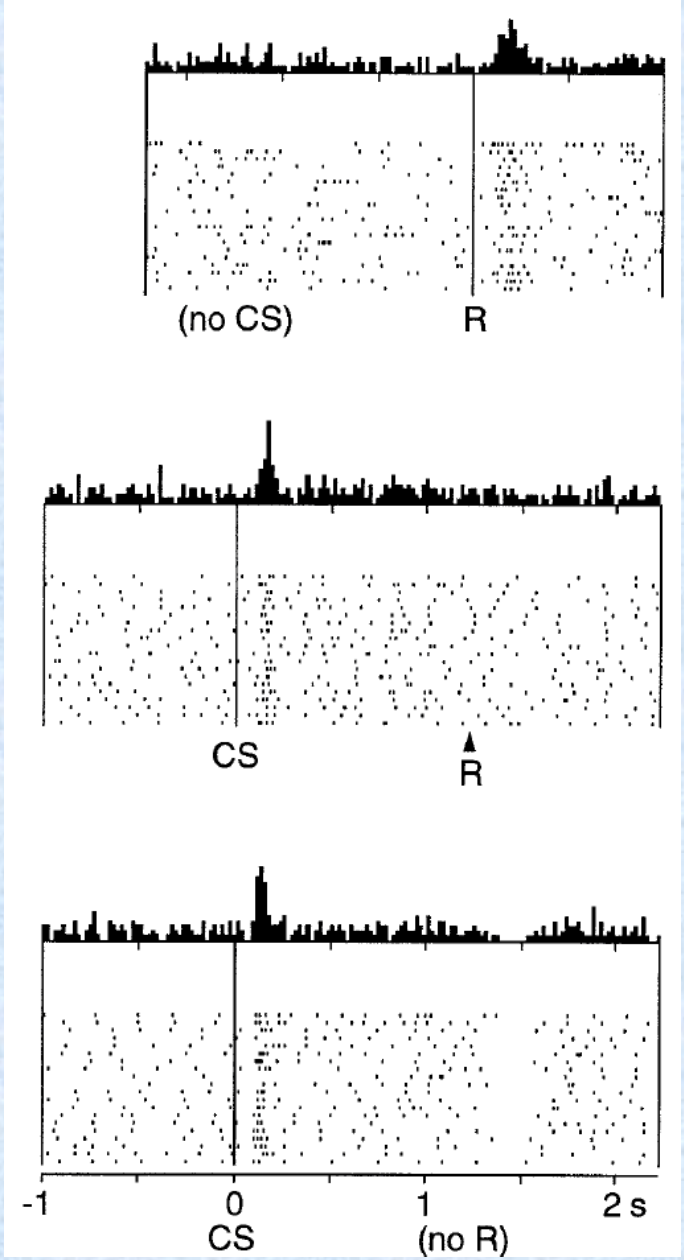
can learn to associate stimulus *x* with reward, but it still will not be able to learn the timing of the predicted reward

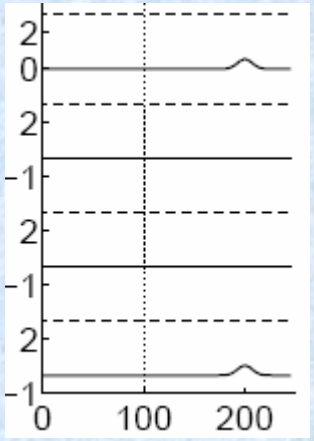
To accomplish this we need to represent each stimulus as a vector through time and build multiple weights:

$$y(t) = \sum_{\tau=0}^{t} w_\tau(t)\, x_\tau(t)$$

$$w_\tau(t+1) = \left(1 - \eta x_\tau^2(t)\right) w_\tau(t) + \eta x_\tau(t)\left(r(t) + \gamma y(t)\right)$$

This form of temporal representation is what *Sutton and Barto* call a complete serial-compound stimulus



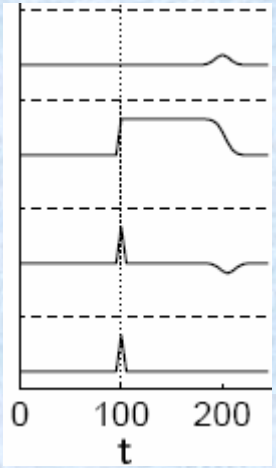


r

y

Δy

$\delta = r + \Delta y$



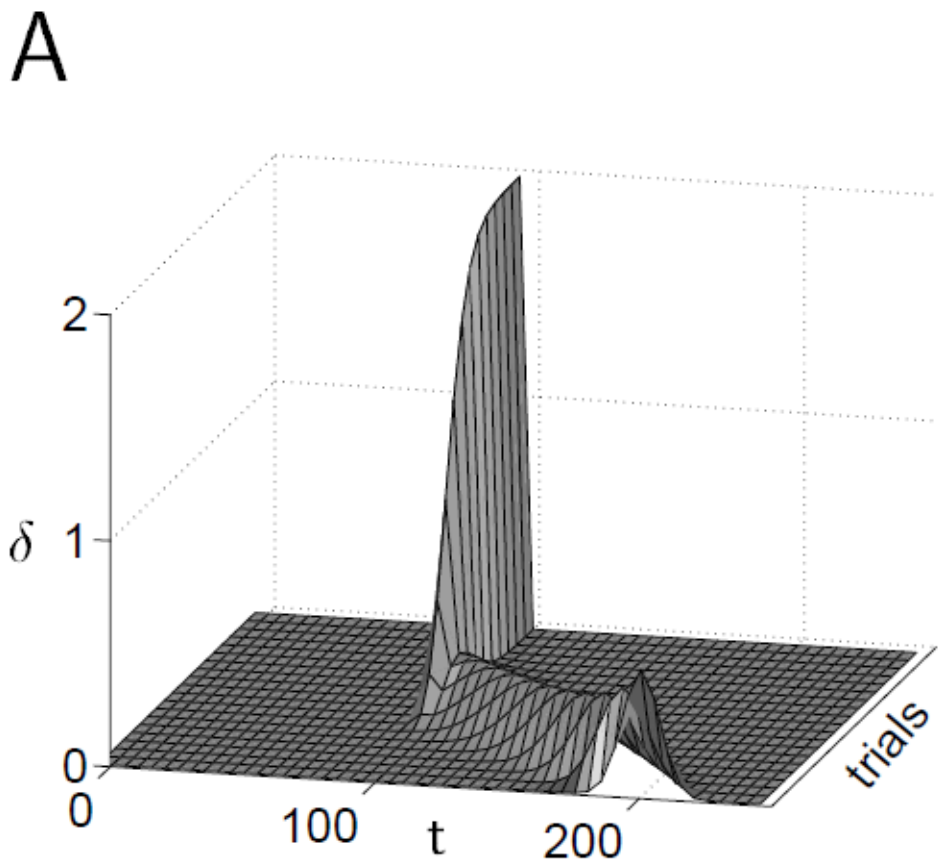
r

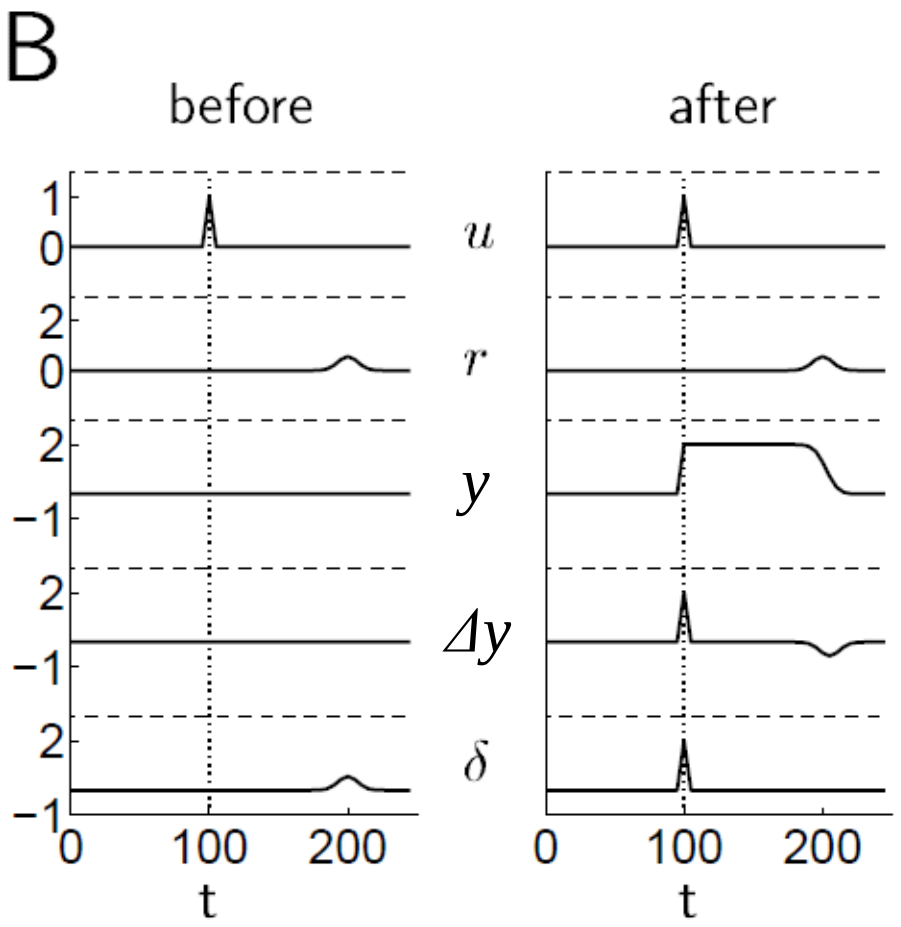
y

Δy

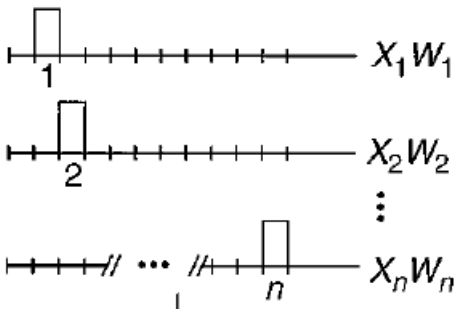
$\delta = r + \Delta y$

How will the bottom panel look?





Note how prediction stays on for the whole interval between stimulus and reward



Temporal Difference Rule: Major Problem

We keep complete stimulus history at all times

What if we have many stimuli?

The main issue here is combinatorial explosion as the number of stimuli, time interval between CS and US, and temporal grid precision go up

The system can be redone as a simple avalanche, where the source cell represents CS and the external input to the border cells represents reward

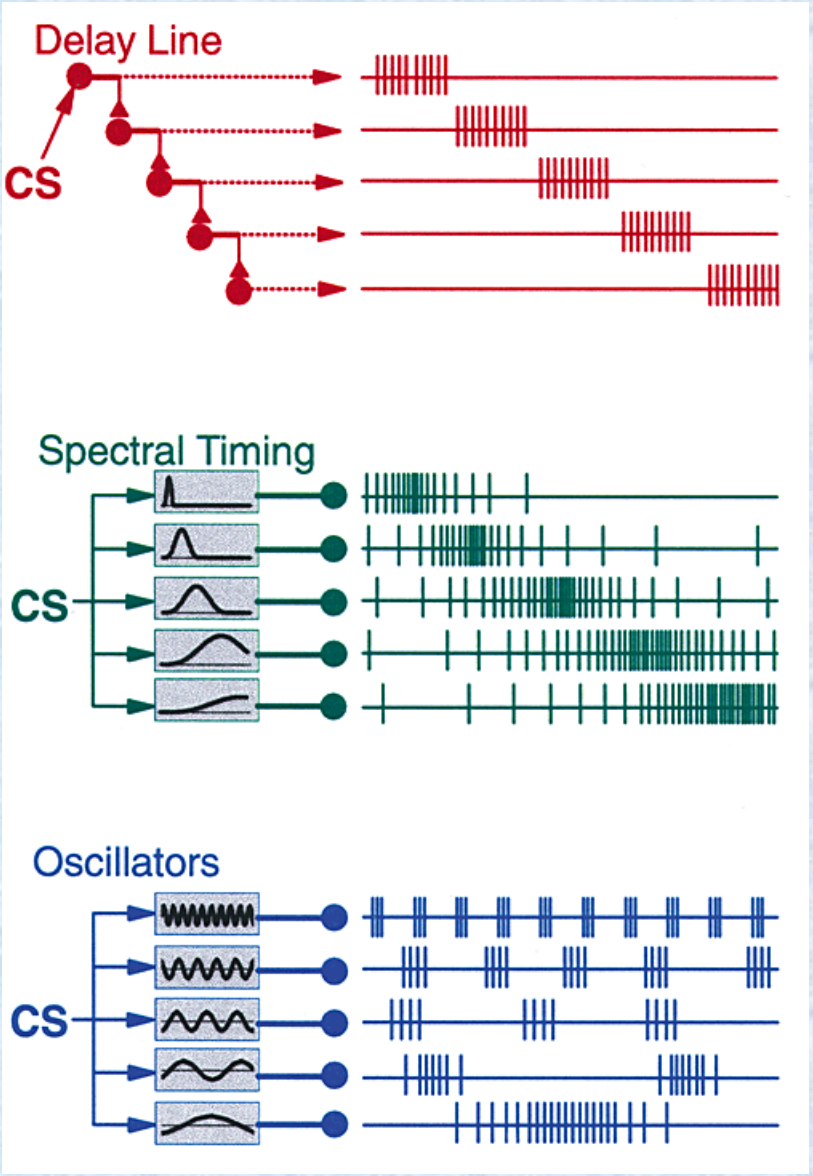
Avalanche-style architecture makes savings in terms of number of cells: same matrix but stored as weights

More Elaborate Solutions

Grossberg's Spectral Timing model

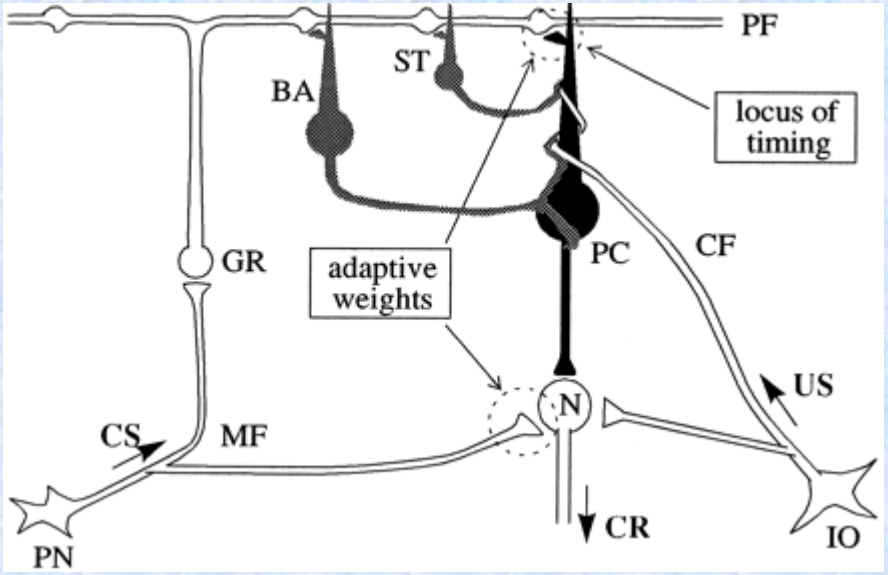
- learns a combination of stimulus and timing through a gradient of synaptic time constants rather then axonal delays
- is more scalable

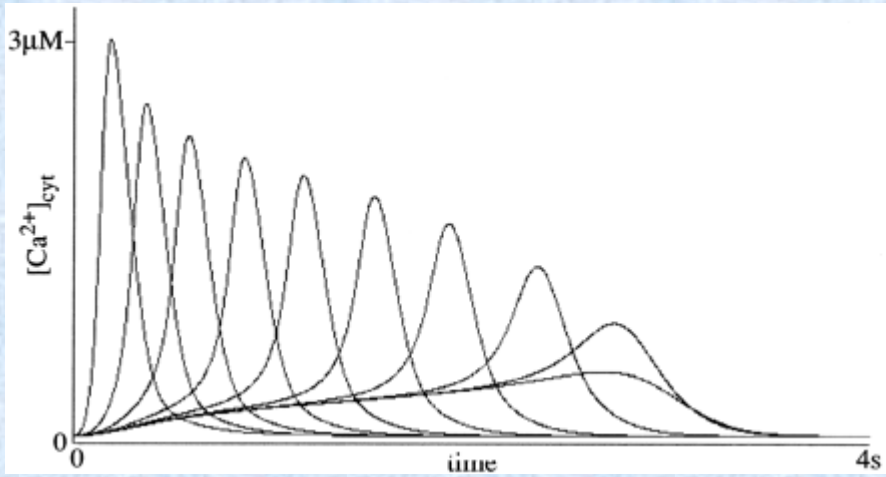
Bullock's Slide and Latch model – elaboration of spectral timing



Spectral Timing

- Calcium level in the dendrite builds slowly
- Positive feedback results in a rapid rise in calcium level
- But when Ca2+ level high enough, IP3R channels close again
- The speed depends on the number of *mGluR1* receptors in the synapse
- Different concentrations of *mGluR1* receptors produce different timing characteristics

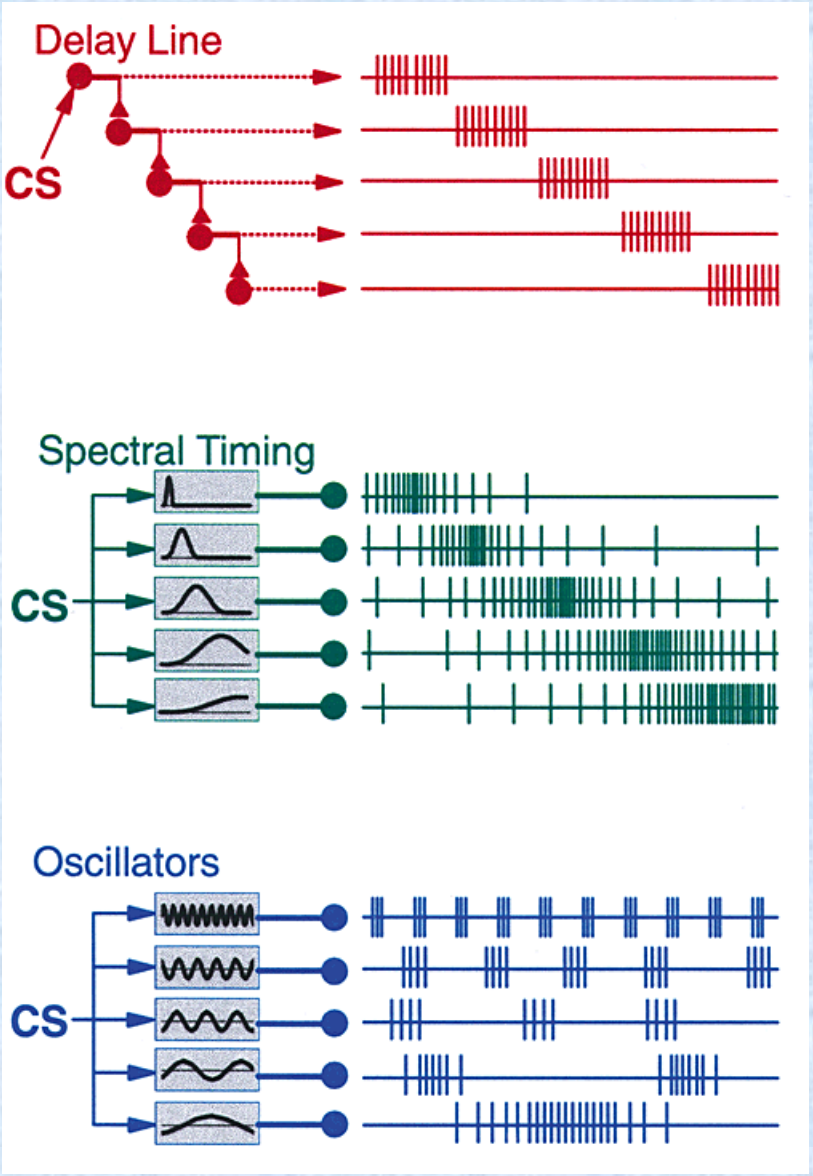




More Elaborate Solutions

BEATS model (second lecture on STDP) applied to time domain is even more flexible and even less expensive

- uses a combination of basis frequencies instead of all possible timings
- small number of bases allows to code a long range of intervals
- topography of hippocampal projections ensures that for long interval precision reduces



Next Time

Adaptive resonance theory is introduced by synthesizing the network building blocks discussed in the course into the ART architecture

Readings

- Appendix D of Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review*, **87**, pp. 1-51.