# Reinforcement Learning(强化学习)

# tutorial

Sam Gershaman

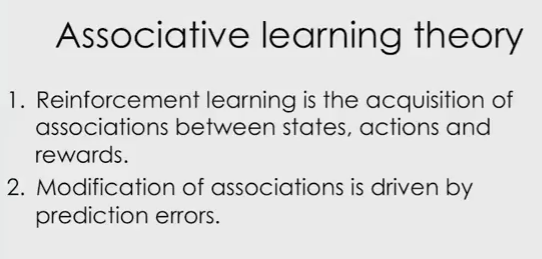
Harvard University

1. 巴甫洛夫条件反射(经典条件反射)

是指一个刺激和另一个带有奖赏或惩罚的无条件刺激多次联结，可使个体学会在单独呈现该一刺激时，也能引发类似无条件反应的条件反应。经典[条件反射](https://baike.baidu.com/item/%E6%9D%A1%E4%BB%B6%E5%8F%8D%E5%B0%84)具有获得、消退、恢复、泛化四个特征



1. an anticipation of upcoming shock（对即将到来的电击的预期）
2. 联想学习理论（是一种刺激——反应的心理联系或联想，或有机体在所受到的刺激与所作出的反应之间建立联系。）



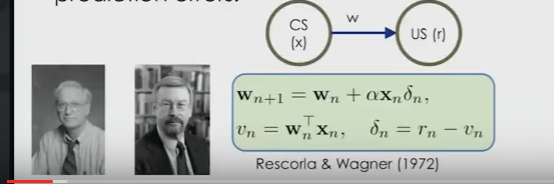
1. 强化学习是指states状态，行为和奖励之间的联系
2. 关联的更改是由预测错误驱动的

1、exactly a state—there's a mathematically precise notion of state and reinforcement learning theory （一个数学上的精确的状态和强化学习理论）

现在假设一个状态是刺激的它是用来预测即时回报的有用的东西

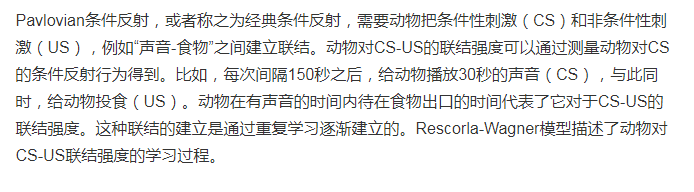
2、 what the animal expected to get and what it actually got

这些是动物期望得到的和实际得到的结果之间的差异我会详细解释这些预测误差



**Wagner模型** （“ **RW** ”）是[经典条件](https://translate.googleusercontent.com/translate_c?depth=1&hl=zh-CN&prev=search&rurl=translate.google.com.hk&sl=en&sp=nmt4&u=https://en.wikipedia.org/wiki/Classical_conditioning&usg=ALkJrhjP68Xwpoxe6eiSCjFQPO_fZ_F_3Q" \o "古典条件)模型

其中学习是根据条件（CS）和无条件（US）刺激之间的关联来概念化的



假设有一系列的线索条件刺激( a collection of  cues conditioned stimuli) CS(x)

这些线索表明存在或不在或在某些情况下这些线索的强度可以由X的矢量编码

这里的N个索引表示的是试号

这将是一种试验水平模式（a trial level model）

we're going to imagine that you have a collection of  cues conditioned stimuli so those cues that the presence or absence or in some cases the intensity of those cues can be encoded by a vector which I'm representing by X and the N index here denotes the trial number right so this is going to be a trial level model I'm going to describe things as though things are happening exactly at the same time you trial by trial and out and I'll come back to some of the limitations of that .

Q激活或存在或缺失指示器的矢量(this vector of Q activations or presence or absence indicators)

每一个线索都有一个相关的权重和重量囊括了q和期望的奖励之间的联系

权重矢量在这里用W表示

基本的想法是在观察了一个奖赏之后这个动物就会计算一个预测误差Delta n, 这就是它实际观察到的和它的期望之间的差异这是V N表示的, 这个预测错误表明当动物得到的奖励比预期的要多

积极的预测误差会导致关联权重的加强

(我得到了一个负的预测误差所以你得到的比你预期的要少你应该削弱这些联系强度)

这是由这个方程所体现的所以我们在这里匹配的是你的期望是由线索的线性组合编码的

本质上，如果我们认为这些是存在的

这个内积表示，我将把当前活动队列的所有权重加起来这会导致我的即时回报预期

(my immediate reward expectations )

接下来我要做的是更新权重与预测误差的比例但只针对目前的提示

所以这是这个理论的一个重要部分即你只是更新了对当前线索的关联这实际上是这个理论的一个弱点我稍后会展示

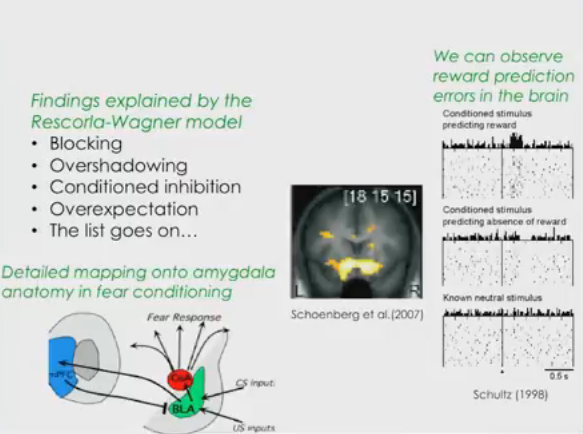
we're going to imagine this vector of Q activations or presence or absence indicators and for each of those cues there's a Associated weight and the weight encapsulate the association between that q and the expected reward that weight vector is represented here by W and basic idea is that after observing a reward RN the animal is computing  a prediction error Delta n so that's the discrepancy between what it actually observed and its expectation which denoted here by V N and so this prediction error says that when the animal got more reward than it was expected so it was a pleasant surprise that positive prediction error should lead to strengthening of the associate  theassociated weights and I got a negative prediction errror so you got less than you expected than you should weaken those associative strength and that's embodied by this equation here so what  we're matching here is that value your expectation is encoded by a linear combination of the cue in essence if we think about these as being present for absent this inner product just says that I'm going to sum up all the weights of the currently active queues and that leads to my immediate reward expectations

then what I'm going to do is I'm going to update the weights in proportion to the prediction errors but only for the cues that were present so that's an important part of this theory which is that you're only updating the associations for present cues and that turns out to be actually a weakness of this theory

Question: it's well in some sense you can say the RN is coming from the environment right it's the shock let's say in a pebble in fear conditioning storm but  actually you kown more generally our n it should index something like the animals hedonic state right so that could be subject to certain kinds of internal variables not necessarily the objective reward from the environment

从某种意义上说，你可以说RN是从环境中来的, 在恐惧条件反射风暴中，这是一种冲击（ in a pebble in fear conditioning storm）但实际上，我们更普遍地说我们的n应该是像动物享乐状态那样的指数（hedonic state）所以这可能会受到某些内部变量的影响不一定是环境的客观回报

1. 从科学的角度来看为什么这是一个如此重要的想法

恐惧条件下的杏仁核解剖图

我们可以观察到大脑中的奖励预测错误

1、首先，修复的瓦格纳mdoel(Wagner mdoel)能够解释经典条件下大量的经验现象, 事实上，这些现象中有一些是由模型预测的并随后进行了测试

first of all the restoral Wagner mdoel was able to explain a huge number of empirical phenomenon in classical conditioning and in fact some of these phenomena were predicted by the model and subsequently tested

1. 他们基本上绘制出了这个细节的结构这是如何通过杏仁核（杏仁体）和内侧前额叶皮质的相互作用来实现的

(杏仁核，附着在海马的末端，呈杏仁状，是边缘系统的一部分。是产生情绪，识别情绪和调节情绪，控制学习和记忆的脑部组织，而且研究发现，幼儿自闭症似乎也与扩大的杏仁核有关。)前额叶皮质与情绪之间有着密切的关系

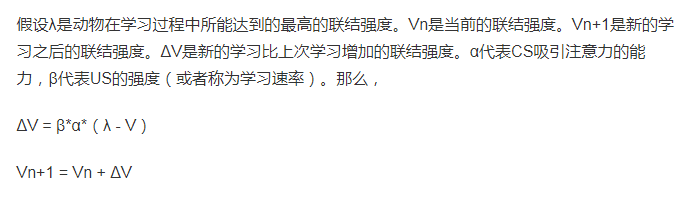
scientists studying fear conditioning being heavily on this formal framework ,I mean not so much explicitly but kind of implicitly it represented and general theoretical framework in which they are operating and they basically mapped out the detailed structure of how this these coputations could be implemented by amygdala circuitry in as well as interactions with medial prefrontal cortex

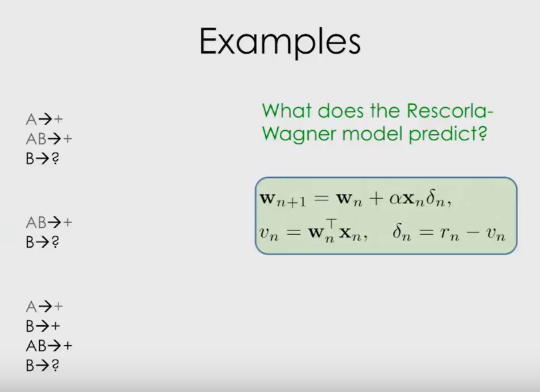
1. 在食欲方面这通常是这样的这些预测错误经常与多巴胺信号联系在一起

（多巴胺是一种神经传导物质，用来帮助细胞传送脉冲的化学物质。这种脑内分泌物和人的情欲、感觉有关，它传递兴奋及开心的信息。另外，多巴胺也与各种上瘾行为有关）

in appetitive domain this has often been so these kinds of prediction errors have often been associated with dopamine signaling

1. 试着弄清楚这个模型的预测对于某些非常简单的动物学习范例



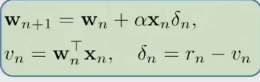


just to familiarize you with this kind of occasion labeling the different cues by A B and so on the arrow means that it's predicting its leading up to some outcome so plus indicate a reward or a shock it donesn't matter and a minus indicates no shock or no reward although in this case we're not using any minus sign here

使用A B来标记不同的cues线索，所以在箭头上意味着它预示着将会出现一些结果，所以加表示一个奖励或一个电击，减号表示没有电击，也没有奖励尽管在这种情况下我们没有使用任何减号

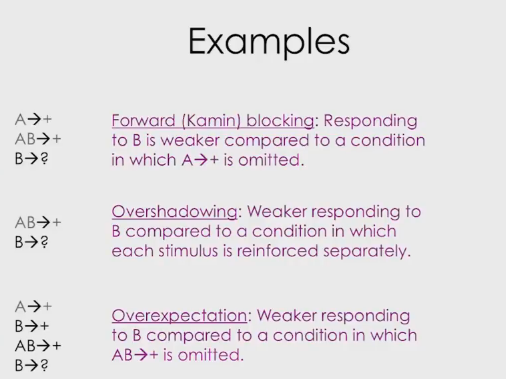
so for the purposes of mapping this on to some simple math let's imagine that plus didn't  a reward value of one

为了把这个映射到一些简单的数学上，让我们想象一下，加上一个没有奖励值的值，让我们想象一下这个公式每次有一个+我们就有一个奖励值，假设所有的权重都被初始化为0，

 the X is here the X vector denotes the stimuli taht are present or absent on the in the current trial X在这里X矢量表示刺激在当前的试验中是存在的或不存在的so it's one when the stimulus is present a zero when the stimulus is (noppers) 所以当刺激是零的时候刺激是（）

 which is known as forward blocking or caming blocking there's unfortunately a tendency in the animal learning literature to name phenomena in these very contentious way which actually prefers a theoretical explanation for the phenomenon这就是所谓的正向阻断或凸块, 不幸的是，动物学习文献中有一种倾向以一种很有争议的方式命名现象,实际上更倾向于对这种现象进行理论解释或者是逻辑上的使用

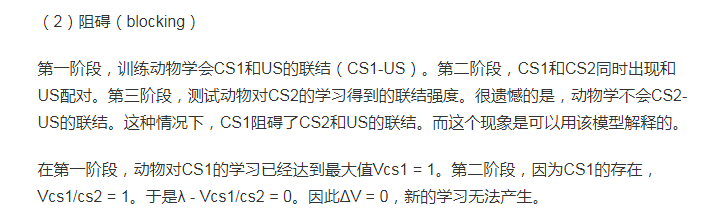
 the very first phase where we have a associated with reward第一阶段，我们与奖励有关联

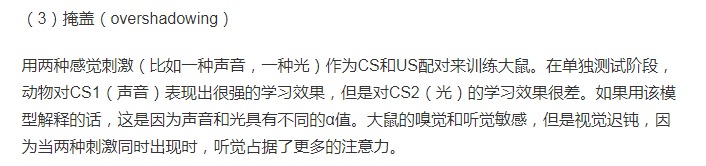


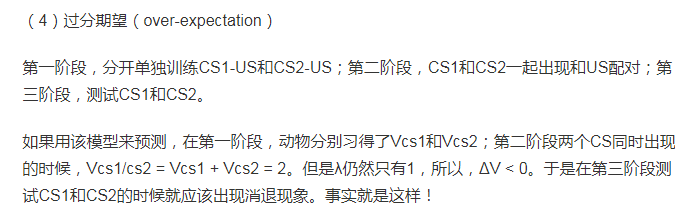
与 被省略的条件相比，对B的响应要弱一些（forward blocking）

对B的反应较弱而每个刺激都被单独加强（overshadowing）

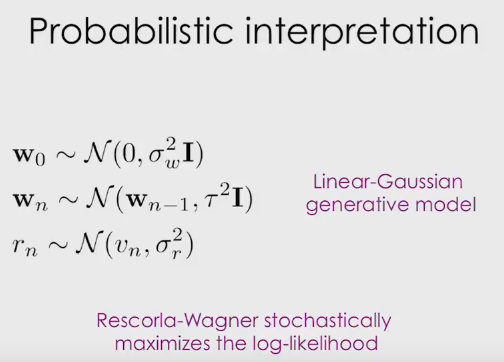
与 被省略的条件相比，对B的响应要弱一些(overexpectation)







1. 概率解释

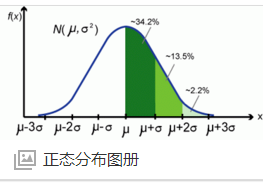
高斯产生的模型

Linear-Gaussian生成模型

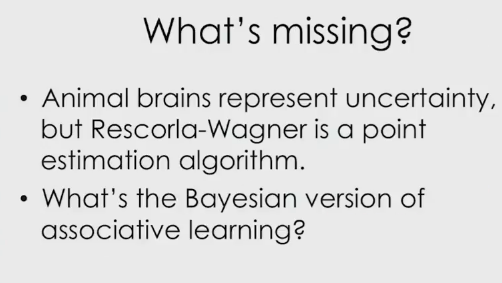
rescorla-Wagener随机地将对数可能性最大化

1、N表示正态分布，里面的变量是一些均值和一些协方差矩阵，所以假设该初始权重是从绘制一些正态分布的平均0和一些协方差矩阵，其实只是缩放单位矩阵仅仅是一个所有的矩阵除了沿对角线那些0，有一些数是沿对角线的协方差矩阵

2、生成模型是权重，慢慢随着时间变化，这里是一种非常简单的高斯随机



1. 是在前一实验的中心分布进行的，基本上随着时间的推移略有扩散，假设奖励绘制，通过这种线性Wagner模型这个内积组合
2. 少了什么

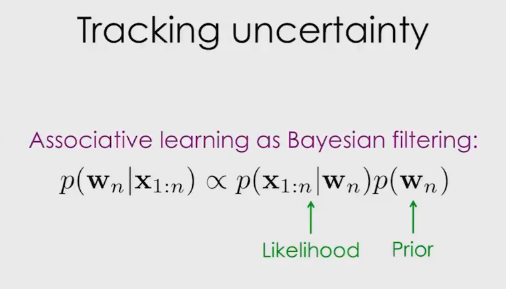


动物的大脑代表着不确定性，但是rescorla-wagner是一个点估算算法

什么是联想学习的贝叶斯版本

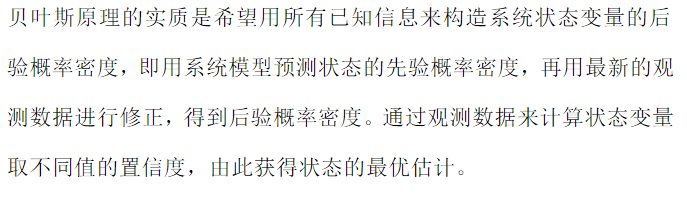


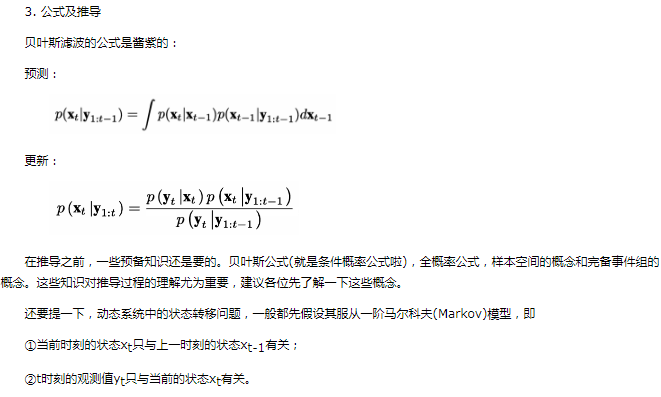
1. 跟踪不确定性



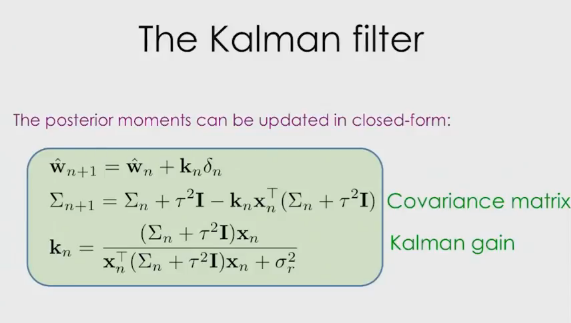
联想学习的贝叶斯滤波 prior优先的 likelihood可能性





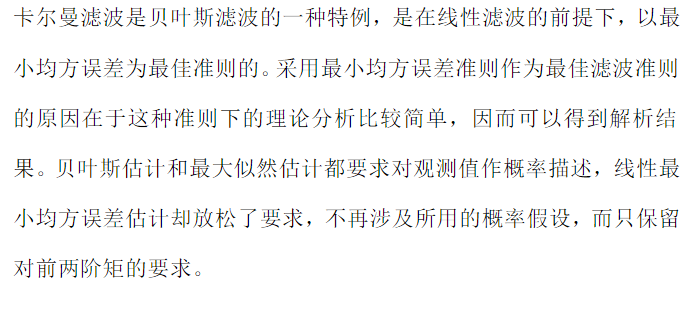


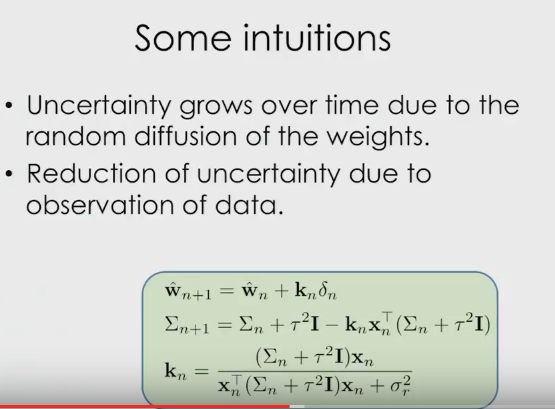
1. 卡尔曼滤波器



后段可以以闭合形式进行更新

协方差矩阵卡尔曼增益



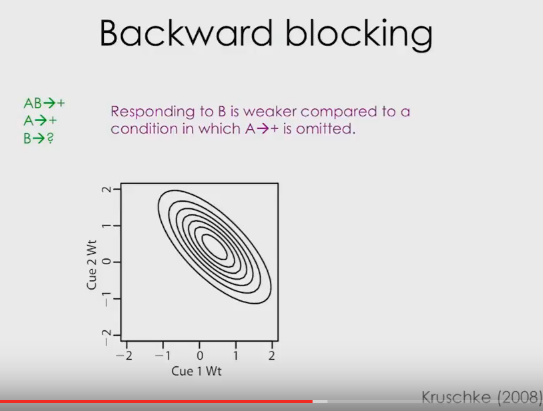


一些直觉

由于权重的随机扩散，不确定性随时间而增长

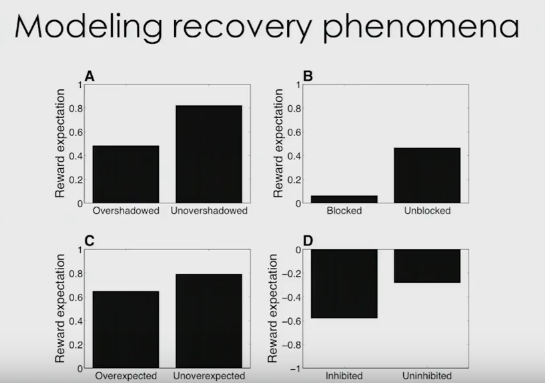
数据观测的不确定性减少

1. 向后阻断

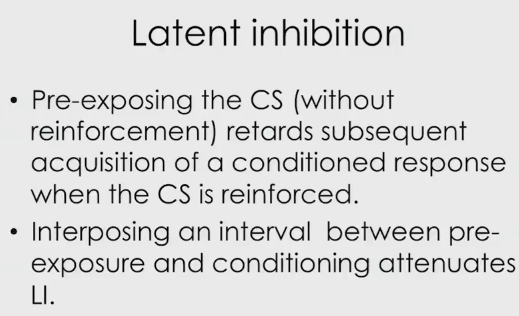


与 被省略的条件相比，对B的响应要弱一些

1. 建模复苏现象



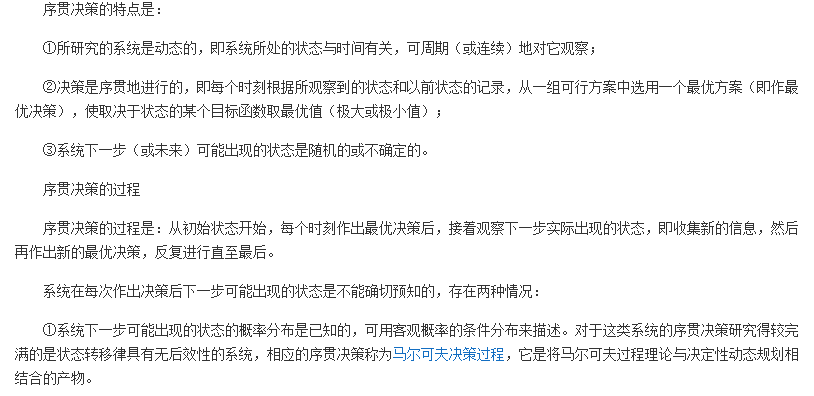
十一、潜伏抑制

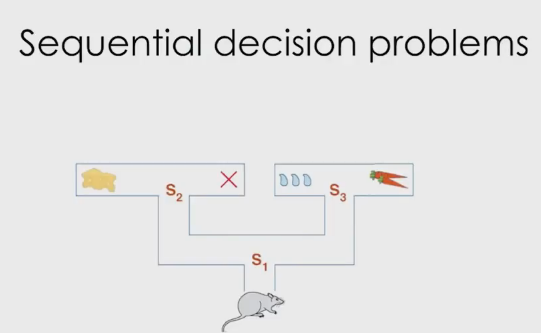
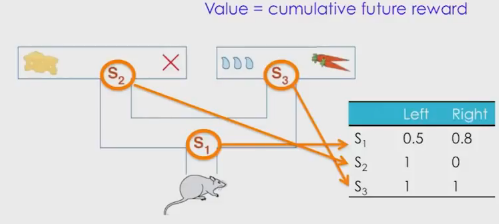


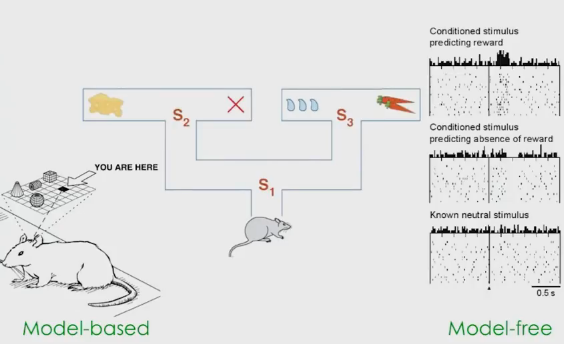
暴露CS(不加强化)在CS被强化时，会阻碍对条件反应的后续处理

预曝光和调节之间的间隔

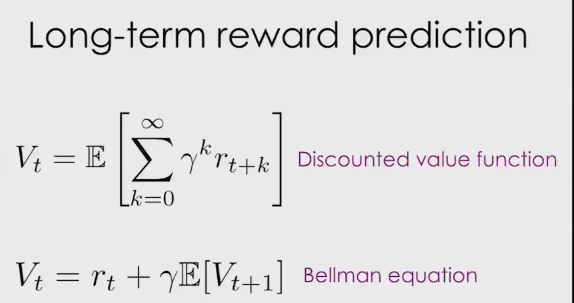
十二、序贯决策问题

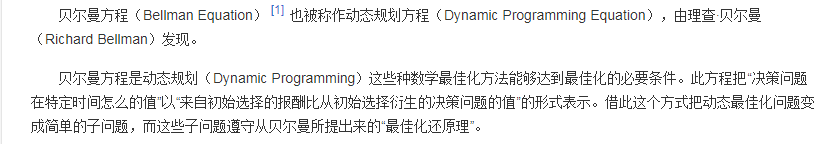


十三、长期回报预测

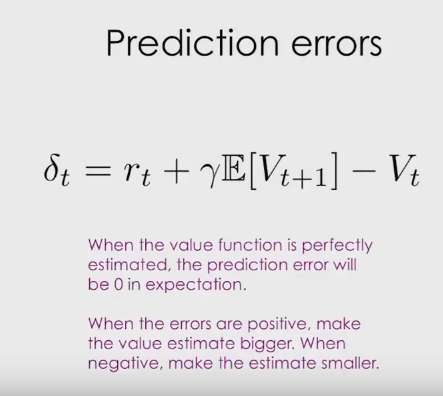


贴现值函数

贝尔曼方程

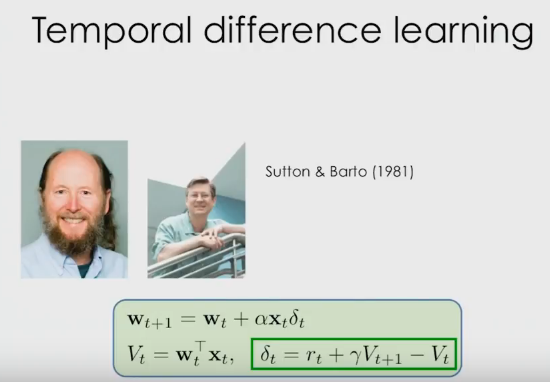


十四、预测错误



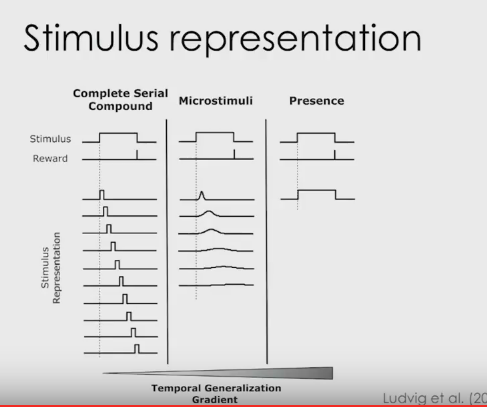
当值函数被完全估计时，预测误差将为0

当错误为正时，使估算值更大。当负的时候，使估算更小



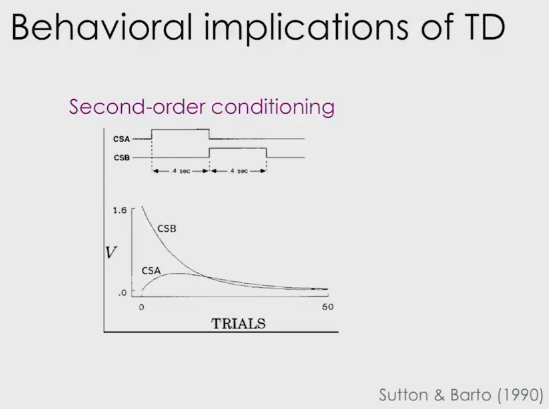
学习时态的区别

十五、刺激计划表示

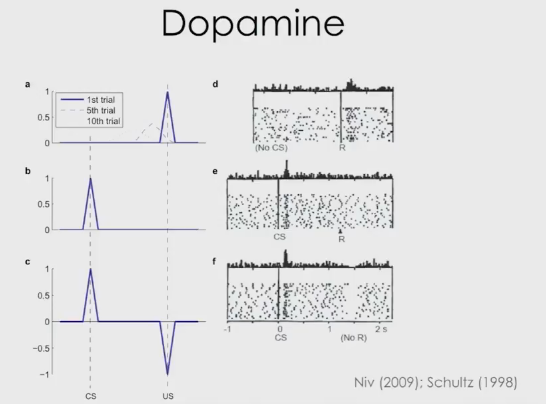


完整的系列化合物complete serial compound

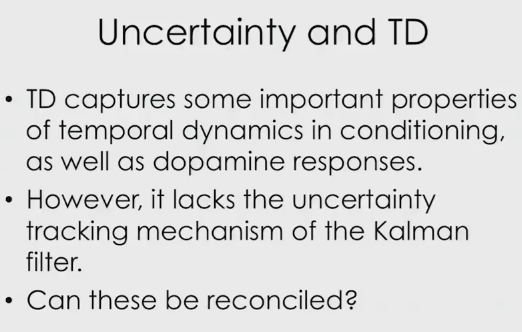
十六、TD行为的影响



十七、多巴胺（一种治脑神经病的药物）



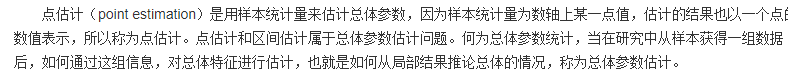
十八、不确定性和TD

TD捕捉了在条件反射和多巴胺反应中，时间动力学的重要特性

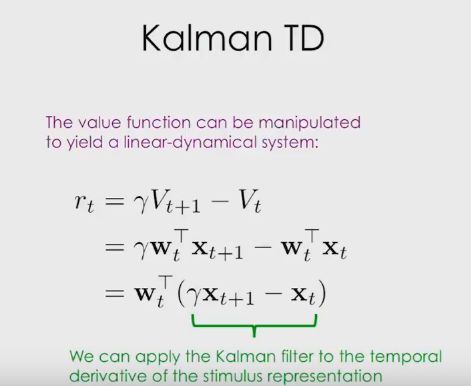
TD捕捉了在条件反射和多巴胺反应中，时间动力学的重要特性

这些可以被调节吗？

TD模型作为点估计算法

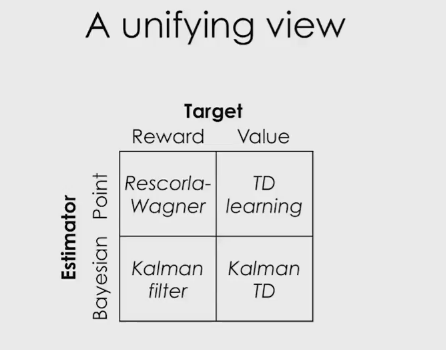


十九、卡尔曼TD

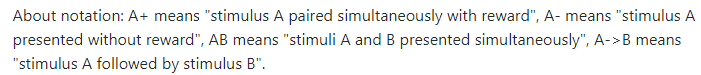


可以对值函数进行操作，从而产生一个线性动力系统。

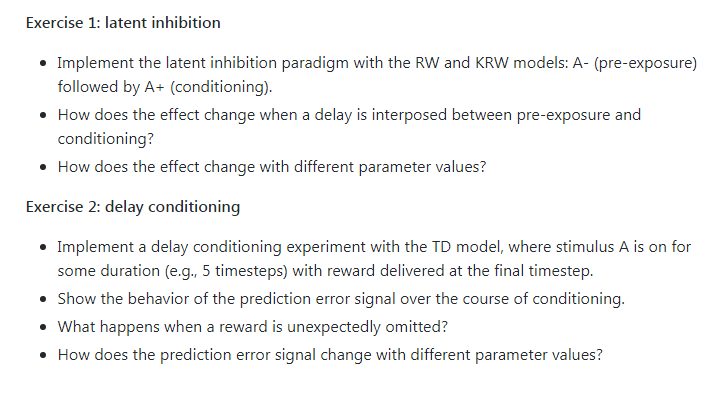
我们可以把Kalman的过滤器应用到刺激方案的时间导数上



一个统一的视图



关于符号:A+表示“刺激同时与奖励同时配对”，A表示“刺激A没有奖励”，AB表示“刺激A和B同时呈现”，A-B表示“刺激A后面跟着刺激B”。



这些练习可以通过模拟模拟模型中的例子来完成。可以在实践中找到解决方案。

练习1:潜在的抑制

用RW和KRW模型来实现潜在的抑制范式:A-(预曝光)后面跟着A+(条件反射)。

当一个延迟在预曝光和条件作用之间发生的时候，效果会如何变化?

如何使用不同的参数值来改变效果呢?

练习2:延迟条件

在TD模型中实现一个延迟条件反射试验，在这个模型中，刺激a在一定的时间内持续。5个时间步骤)，在最后的时间步骤中提供奖励。

在条件反射过程中，显示预测误差信号的行为。

当一个奖励意外被忽略时会发生什么?

预测误差信号如何随着不同参数值的变化而变化?

