Chapter 7

Memory

When you think of memory, you probably think of **episodic memory** — memory for specific episodes or events. Maybe you can remember some special times from childhood (birthdays, family trips, etc), or some traumatic times (ever get lost in a supermarket, or get left behind on a hike or other family outing?). Probably you can remember what you had for dinner last night, and who you ate with? Although this aspect of memory is the most salient for us, it is just one of many different types of memory.

One broad division in memory in mechanistic, computational terms is between **weight-based** and **activation-based** forms of memory. Weight based memory is a result of synaptic plasticity, and is generally relatively long lasting (at least several tens of minutes, and sometimes several decades, up to a whole lifetime). Activation-based memory is supported by ongoing neural activity, and is thus much more transient and fleeting, but also more flexible. Because weight-based memory exists at every modifiable synapse in the brain, it can manifest in innumerable ways. In this chapter, we focus on some of the most prominent types of memory studied by psychologists, starting with episodic memory, then looking at familiarity-based recognition memory, followed by weight-based priming, and activation-based priming. We'll look at more robust forms of activation-based memory, including working memory, in the *Executive Function* Chapter.

Probably most people have heard of the **hippocampus** and its critical role in episodic memory — the movie *Memento* for example does a great job of portraying what it is like to not have a functional hippocampus. We'll find out through our computational models *why* the hippocampus is so good at episodic memory — it has highly *sparse* patterns of neural activity (relatively few neurons active at a time), which allows even relatively similar memories to have very different, non-overlapping neural representations. These distinct neural patterns dramatically reduce *interference*, which is the primary nemesis of memory. Indeed, the highly distributed, overlapping representations in the neocortex — while useful for reasons outlined in the first half of this book — by themselves produce **catastrophic interference** when they are driven to learn too rapidly. But it is this rapid *one-shot* learning that is required for episodic memory! Instead, it seems that the brain leverages two specialized, **complementary learning systems** — the hippocampus for rapid encoding of new episodic memories, and the neocortex for slow acquisition of rich webs of semantic knowledge, which benefit considerably from the overlapping distributed learning and slower learning rates, as we'll see.

Countering the seemingly ever-present urge to oversimplify and modularize the brain, it is critical to appreciate that memory is a highly distributed phenomena, with billions of synapses throughout the brain being tweaked by any given experience. Several studies have shown preserved learning of new memories of relatively specific information in people with significant hippocampal damage — but it is critical to consider how these memories are cued. This is an essential aspect to remember about memory in general: whether a given memory can actually be retrieved depends critically on how the system is probed. We've probably all had the experience of a flood of memories coming back as a result of visiting an old haunt — the myriad of cues available enable (seemingly spontaneous) recall of memories that otherwise are not quite strong enough to rise to the surface. The memories encoded without the benefit of the hippocampus are weaker and more vague, but they do exist.

In addition to being highly distributed, memory in the brain is also highly interactive. Information that is initially encoded in one part of the brain can appear to "spread" to other parts of the brain, if those memories

are reactivated and these other brain areas get further opportunities to learn them. A classic example is that episodic memories initially encoded in the hippocampus can be strengthened in the surrounding neocortical areas through repeated retrieval of those memories. This can even happen while we are sleeping, when patterns of memories experienced during the day have shown to be re-activated! Furthermore, influences of the prefrontal cortex system, and affective states, can significantly influence the encoding and retrieval of memory. Thus, far from the static "hard drive" metaphor from computers, memory in the brain is a highly dynamic, constantly evolving process that reflects the complexity and interactions present across all the areas of the brain.

7.1 Episodic Memory

We begin with episodic memory, because it is such a major part of our conscious lives, and really of our identities. For example, the movie *Total Recall,* loosely based on the Philip K. Dick novel We Can Remember it for You Wholesale (Wikipedia link), explores this connection between episodic memories and our sense of self. All people with a functioning hippocampus have this remarkable "tape recorder" constantly encoding everything that happens during our waking lives — we don't have to exert particular effort to recall what happened 20 minutes or a few hours ago — it is just automatically there. Most people end up forgetting the vast majority of the daily flux of our lives, retaining only the particularly salient or meaningful events.

However, a tiny percentage of otherwise seemingly "normal" people are able to remember an abnormally large number of experiences in vivid detail (this is called hyperthymesia (Wikipedia link)). Interestingly, it is not the hippocampus itself that differentiates these people — instead they are characterized by the obsessive rehearsal and retrieval of episodic memories, with areas of the basal ganglia apparently enlarged (which is associated with obsessive compulsive disorder (OCD)). As we'll see in the *Executive Function* Chapter, the basal ganglia participate not only in motor control and reinforcement learning, but also the reinforcement of updating and maintenance of active memory. This suggests that in normal human brains, the hippocampus has the raw ability to encode and remember every day of our lives, but most people just don't bother to rehearse these memories to the point where they can all be reliably retrieved. Indeed, a major complaint that people with hyperthymesia have is that they are unable to forget all the unpleasant stuff in life that most people just let go.

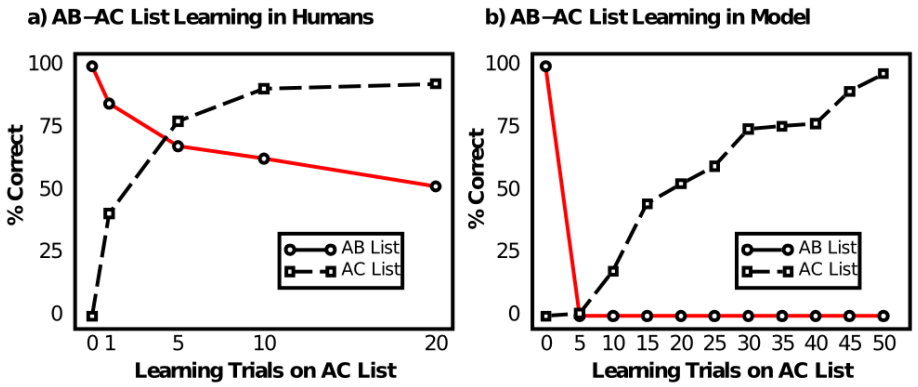


Figure 7.1: Data from humans (a) and a generic (cortical) neural network model (b) on the classic AB-AC list learning task, which generates considerable interference by re-pairing the A list items with new associates in the AC list after having first learned the AB list. People's performance on the AB items after learning the AC list definitely degrades (red line), but nowhere near as catastrophically as in the neural network model. Data reproduced from McCloskey and Cohen (1989).

So what exactly makes the hippocampus such an exceptionally good episodic memory system? Our investigation begins with failure. Specifically, the failure of a "generic" cortical neural network model of the sort we've been exploring in this textbook to exhibit any kind of useful episodic memory ability. This failure was first documented by (McCloskey & Cohen, 1989), using a generic backpropagation network trained on the

AB-AC paired associate list learning task (Figure 7.1). This task involves learning an initial list of arbitrary word pairs, called the AB list — for example:

- locomotive - dishtowel
- window - reason
- bicycle - tree
- ...

People are tested on their ability to recall the B associate for each A item, and training on the AB list ends when they achieve perfect recall. Then, they start learning the AC list, which involves new associates for the previous A items:

- locomotive - cloud
- window - book
- bicycle - couch
- ...

After 1, 5, 10, and 20 iterations of learning this AC list, people are tested on their ability to recall the original AB items, without any additional training on those items. The plot on the left shows that there is a significant amount of interference on the AB list as a result of learning the AC items, due to the considerable overlap between the two lists, but even after 20 iterations through the AC items, people can still recall about 50% of the AB list. In contrast, the plot on the right shows that the network model exhibited **catastrophic interference** — performance on the AB list went to 0% immediately. They concluded that this invalidated all neural network models of human cognition, because obviously people have much better episodic memory abilities.

But we'll see that this kind of whole-sale abandonment of neural networks is unjustified (indeed, the brain is a massive neural network, so there must be some neural network description of any phenomenon, and we take these kind of challenges as informative opportunities to identify the relevant mechanisms). Indeed, in the following exploration we will see that there are certain network parameters that reduce the levels of interference. The most important manipulation required is to increase the level of inhibition so that fewer neurons are active, which reduces the overlap between the internal representation of the AB and AC list items, thus allowing the system to learn AC without overwriting the prior AB memories. We'll then see that the hippocampal system exploits this trick to an extreme degree (along with a few others), making it an exceptionally good episodic memory system.

7.1.1 Exploration of Catastrophic Interference

Run the `abac` simulation from CCN Sims.

7.2 The Hippocampus and Pattern Separation / Pattern Completion

The hippocampus is specifically optimized to rapidly record episodic memories using highly **sparse** representations (i.e., having relatively few neurons active) that minimize overlap (through **pattern separation**) and thus interference. This idea is consistent with such a large quantity of data, that it is close to established fact (a rarity in cognitive neuroscience). This data includes the basic finding of episodic memory impairments (and particularly in pattern separation) that result from selective hippocampal lesions, the unique features of the hippocampal anatomy, which are so distinctive relative to other brain areas that they cry out for an explanation, and the vast repertoire of neural recording data from different hippocampal areas. We start with an overview of hippocampal anatomy, followed by the neural recording data and an understanding of how relatively sparse neural activity levels also results in pattern separation, which minimizes interference.

7.2.1 Hippocampal Anatomy

The anatomy of the hippocampus proper and the areas that feed into it is shown in Figure 7.2. The hippocampus represents one of two "summits" on top of the hierarchy of interconnected cortical areas (where

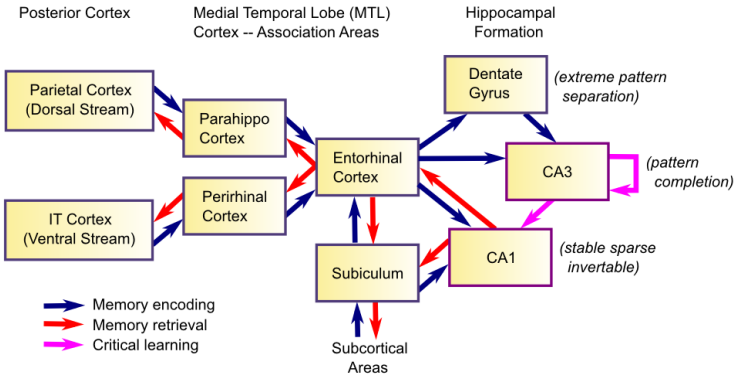


Figure 7.2: The hippocampus sits on "top" of the cortical hierarchy and can encode information from all over the brain, binding it together into an episodic memory. Dorsal (parahippocampal) and Ventral (perirhinal) pathways from posterior cortex converge into the entorhinal cortex, which is then the input and output pathway of the hippocampus proper, consisting of the dentate gyrus (DG) and areas of "ammon's horn" (cornu ammonis, CA) — CA3 and CA1. CA3 represents the primary "engram" for the episodic memory, while CA1 is an invertible encoding of EC, such that subsequent recall of the CA3 engram can activate CA1 and then EC, to reactivate the full episodic memory out into the cortex.

the bottom are sensory input areas, e.g., primary visual cortex) — the other such summit is the prefrontal cortex explored in the *Executive Function* Chapter. Thus, it possesses a critical feature for an episodic memory system: access to a very high-level summary of everything of note going on in your brain at the moment. This information, organized along the dual-pathway dorsal vs. ventral pathways explored in the *Perception and Attention* Chapter, converges on the **parahippocampal (PHC)** (dorsal) and **perirhinal (PRC)** (ventral) areas, which then feed into the **entorhinal cortex (EC)**, and then into the hippocampus proper. The major hippocampal areas include the **dentate gyrus (DG)** and the areas of "ammon's horn" (cornu ammonis (CA) in latin), **CA3** and **CA1** (what happened to CA2? turns out it is basically the same as CA3 so we just use that label). All of these strange names have to do with the shapes of these areas, including the term "hippocampus" itself, which refers to the seahorse shape it has in the human brain (hippocampus is Greek for seahorse).

The basic episodic memory encoding story in terms of this anatomy goes like this. The high-level summary of everything in the brain is activated in EC, which then drives the DG and CA3 areas via the **perforant pathway** — the end result of this is a highly sparse, distinct pattern of neural firing in CA3, which represents the main "engram" of the hippocampus. The EC also drives activity in CA1, which has the critical feature of being able to then re-activate this same EC pattern all by itself (i.e., an *invertible mapping* or *auto-encoder* relationship between CA1 and EC). These patterns of activity then drive synaptic plasticity (learning) in all the interconnected synapses, with the most important being the synaptic connections among CA3 neurons (in the CA3 recurrent pathway), and the connections between CA3 and CA1 (the **Schaffer collateral** pathway). These plastic changes effectively "glue together" the different neurons in the CA3 engram, and associate them with the CA1 invertible pattern, so that subsequent retrieval of the CA3 engram can then activate the CA1, then EC, and back out to the cortex. Thus, the primary function of the hippocampus is to bind together all the disparate elements of an episode, and then be able to retrieve this *conjunctive memory* and reinstate it out into the cortex during recall. This is how a memory can come "flooding back" — it floods back from CA3 to CA1 to EC to cortex, reactivating something approximating the original brain pattern at the time the memory was encoded.

As noted in the introduction, every attempt to simplify and modularize memory in this fashion is inaccurate, and in fact memory encoding is distributed among all the neurons that are active at the time of the episode. For example, learning in the perforant pathway is important for reactivating the CA3 engram from the EC inputs (especially when they represent only a partial memory retrieval cue). In addition, learning all the way through the cortical pathways into and out of the hippocampus "greases" the retrieval process. Indeed, if a memory pattern is reactivated frequently, then these cortical connections can be strong enough

to drive reactivation of the full memory, without the benefit of the hippocampus at all. We discuss this *consolidation* process in detail later. Finally, the retrieval process can be enhanced by controlled retrieval of memory using top-down strategies using the prefrontal cortex. We don't consider this aspect of controlled retrieval here, but it depends on a combination of activation and weight based memory analogous to some features we will explore in *Executive Function* Chapter.

7.2.2 Properties of Hippocampal Neurons: Sparseness, Pattern Separation

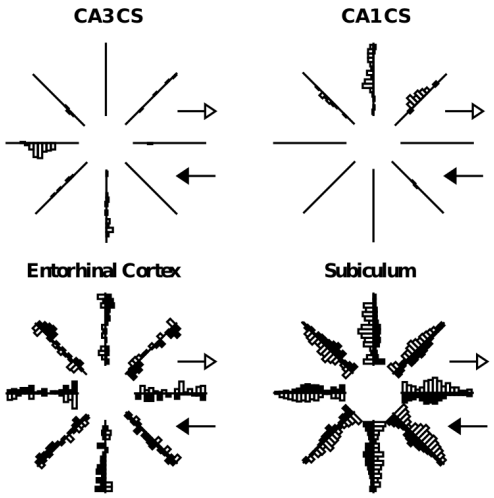


Figure 7.3: Comparison of activity patterns across different areas of the hippocampus, showing that CA fields (CA3, CA1) are much more sparse and selective than the cortical input areas (Entorhinal cortex (EC) and subiculum). This sparse, pattern separated encoding within the hippocampus enables it to rapidly learn new episodes while suffering minimal interference. Activation of sample neurons within each area are shown for a rat running on an 8 arm radial maze, with the bars along each arm indicating how much the neuron fired for each direction of motion along the arm. The CA3 neuron fires only for one direction in one arm, while EC has activity in all arms (i.e., a much more overlapping, distributed representation).

A representative picture of a critical difference between the hippocampus (CA3, CA1) and cortex is shown in Figure 7.3, where it is clear that CA3 and CA1 neurons fire much less often than those in the cortex (entorhinal cortex and subiculum). This is what we mean by *sparseness* in the hippocampal representation — for any given episode, only relatively few neurons are firing, and conversely, each neuron only fires under a very specific circumstance. In rats, these circumstances tend to be identifiable as spatial locations, i.e., *place cells*, but this is not generally true of primate hippocampus. This sparseness is thought to result from high levels of GABA inhibition in these areas, keeping many neurons below threshold, and requiring active neurons to receive a relatively high level of excitatory input to overcome this inhibition. The direct benefit of this sparseness is that the engrams for different episodes will overlap less, just from basic probabilities (Figure 7.4). For example, if the probability of a neuron being active for a given episode is 1% (typical of the DG), then the probability for any two random episodes is that value squared, which is .01% (a very small number). In comparison, if the probability is higher, e.g., 25% (typical of cortex), then there is a 6.25% chance of overlap for two episodes. David Marr appears to have been the first one to point out this *pattern separation* property of sparse representations, in an influential paper (D. Marr, 1971).

The connection between activity levels and pattern separation can also be observed within the hippocampus itself, by comparing the firing properties of DG vs. CA3 neurons, where DG neurons have the sparsest activity levels, even compared to the somewhat less sparse CA3 (roughly 2-5% activity level). Figure 7.5 from a study by (Leutgeb, Leutgeb, Moser, & Moser, 2007) shows that the DG exhibits more pattern separation than the CA3, as a function of systematic morphing of an environment from a square to a circle and back again. The DG neurons exhibit a greater variety of neural firing as a function of this environmental change, suggesting that they separate these different environments to a greater extent than the CA3. There are many

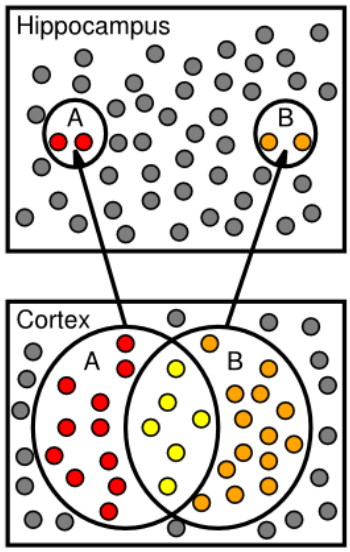


Figure 7.4: Schematic illustration of how more sparse activity levels can produce pattern separation, just because the odds of overlapping are that much lower. Graphically, this is evident in that the smaller circles (sparser activation) in the hippocampus are less likely to overlap than the larger ones in the cortex.

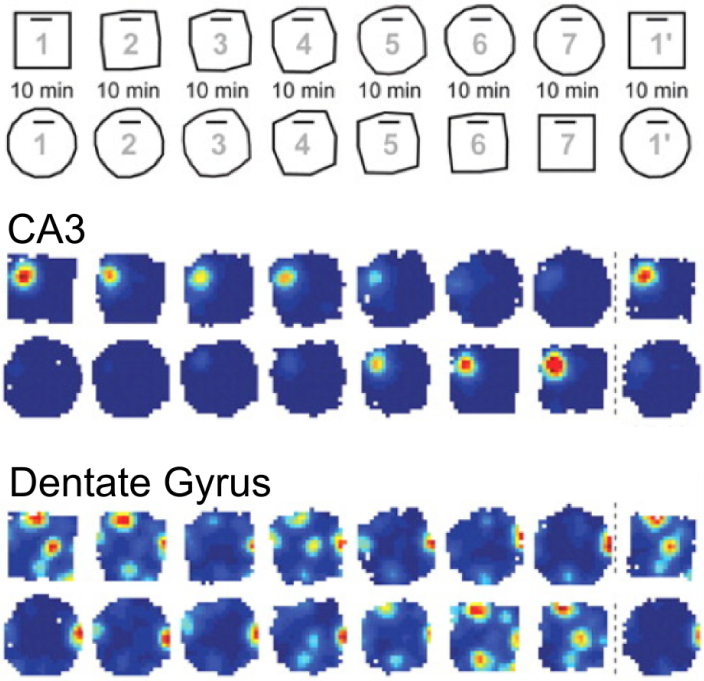


Figure 7.5: Pattern separation in the CA3 and Dentate Gyrus of the hippocampus, as a function of a rat's location in an environment that morphs gradually from a square into a circle, and vice-versa (indicated in top panel). The CA3 neuron shown here has two distinct "place cell" firing patterns, one for the square and one for the circle. In contrast, The DG neuron exhibits somewhat greater pattern separation by responding differentially in the middle of the morph sequence. Data from Leutgeb et al. (2007).

other compelling demonstrations of pattern separation in various hippocampal areas relative to cortex, and in particular in DG relative to other areas (see e.g., the extensive work of Kesner on this topic).

Another factor that contributes to effective pattern separation is the broad and diffuse connectivity from EC to DG and CA3, via the perforant pathway. This allows many different features in EC to be randomly combined in DG and CA3, enabling them to be sensitive to combinations or *conjunctions* of inputs. Because of the high inhibitory threshold associated with sparse activations, this means a given neuron in these areas must receive significant excitation from multiple of these diffuse input sources. In other words, these neurons have **conjunctive representations**.

Pattern separation is important for enabling the hippocampus to rapidly encode novel episodes with a minimum of interference on prior learning, because the patterns of neurons involved overlap relatively little.

7.2.3 Pattern Completion: Cued Recall

While pattern separation is important for encoding new memories, this encoding would be useless unless these memories can be subsequently recalled. This recall process is also known as **pattern completion**, where a partial retrieval cue triggers the completion of the full original pattern associated with the memory. For example, if I cue you with the question: "did you go to summer camp as a kid?" you can pattern complete from this to memories of summer camp, or not, as the case may be. The amazing thing about human memory is that it is **content addressable memory** — any sufficiently specific subset of information can serve as a retrieval cue, enabling recovery of previously-encoded episodic memories. In contrast, memory in a computer is accessed by a memory address or a variable pointer, which has no relationship to the actual content stored in that memory. The modern web search engines like Google demonstrate the importance of content addressability, and function much like the human memory system, taking search terms as retrieval cues to find relevant "memories" (web pages) with related information. As you probably know from searching the web, the more specific you can make your query, the more likely you will retrieve relevant information — the same principle applies to human memory as well.

In the hippocampus, pattern completion is facilitated by the recurrent connections among CA3 neurons, which glues them together during encoding, such that a subset of CA3 neurons can trigger recall of the remainder. In addition, the synaptic changes during encoding in the perforant pathway make it more likely that the original DG and CA3 neurons will become reactivated by a partial retrieval cue.

Interestingly, there is a direct tension or tradeoff between pattern separation and pattern completion, and the detailed parameters of the hippocampal anatomy can be seen as optimizing this tradeoff (O'Reilly & McClelland, 1994). Pattern separation makes it more likely that the system will treat the retrieval cue like a novel stimulus, and thus encode a new distinct engram pattern in CA3, instead of completing to the old one. Likewise, if the system is too good at pattern completion, it will reactivate old memories instead of encoding new pattern separated ones, for truly novel episodes. Although the anatomical parameters in our model do help to find a good balance between these different forces of completion and separation, it is also likely that the hippocampus benefits from strategic influences from other brain areas, e.g., prefrontal cortex executive control areas, to emphasize either completion or separation depending on whether the current demands require recall or encoding, respectively. We will explore this issue further in the *Executive Function* Chapter.

7.2.4 Exploration

To explore how the hippocampus encodes and recalls memories, using the AB-AC task, run the `hip` simulation in CCN Sims.

7.3 Complementary Learning Systems

As noted earlier, when McCloskey & Cohen first discovered the phenomenon of catastrophic interference, they concluded that neural networks are fatally flawed and should not be considered viable models of human cognition. This is the same thing that happened with (Minsky & Papert, 1969), in the context of networks that lack a hidden layer and thus cannot learn more difficult mappings such as XOR (see the *Learning* Chapter for more details). In both cases, there are ready solutions to these problems, but people seem all too

willing to seize upon an excuse to discount the neural network model of the mind. Perhaps it is just too reductionistic or otherwise scary to think that everything that goes on in your brain could really boil down to mere neurons... However, this problem may not be unique to neural networks — researchers often discount various theories of the mind, including Bayesian models for example, when they don't accord with some pattern of data. The trick is to identify when any given theory is fundamentally flawed given challenging data; the devil is in the details, and oftentimes there are ways to reconcile or refine an existing theory without "throwing out the baby with the bathwater".

Complementary Learning Systems

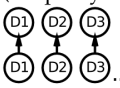
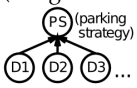
Goals:	Remember Specifics	Extract Generalities
Example:	Where is car parked?	Best parking strategy?
Need to:	Avoid interference	Accumulate experience
	Solution:	
1.	Separate reps (keep days separate) 	Overlapping reps (integrate over days) 
2.	Fast learning (encode immediately)	Slow learning (integrate over days)
3.	Learn automatically (encode everything)	Task-driven learning (extract relevant stuff)
	These are incompatible, need two different systems:	
System:	Hippocampus	Neocortex

Figure 7.6: Summary of the Complementary Learning Systems perspective on functional roles of Hippocampus vs. Neocortex, in context of memory about parking spaces. The hippocampus can rapidly encode in a relatively interference-free way where you parked your car today, as distinct from previous days. The neocortex in contrast can integrate across many different experiences (using a slow learning rate) to extract an overall parking strategy that reflects effects of many different factors on likelihood of finding parking in a given lot. The functional demands for these two different kinds of learning are in direct conflict, so the best overall functionality can be achieved by having two complementary learning systems, each separately optimized for these different functions.

Such musings aside, there are (at least) two possible solutions to the catastrophic interference problem. One would be to somehow improve the performance of a generic neural network model in episodic memory tasks, inevitably by reducing overlap in one way or another among the representations that form. The other would be to introduce a specialized episodic memory system, i.e., the hippocampus, which has parameters that are specifically optimized for low-interference rapid learning through pattern separation, while retaining the generic neural network functionality as a model of neocortical learning. The advantage of this latter perspective, known as the **complementary learning systems (CLS)** framework (J. L. McClelland et al., 1995; Norman & O'Reilly, 2003), is that the things you do to make the generic neural model better at episodic memory actually interfere with its ability to be a good model of neocortex.

Specifically, neocortical learning for things like object recognition (as we saw in the *Perception and Attention* Chapter), and semantic inference (as we'll see in the *Language* Chapter) really benefit from highly overlapping distributed representations, and slow interleaved learning. These overlapping distributed representations enable patterns of neural activity to encode complex, high-dimensional similarity structures among items (objects, words, etc), which is critical for obtaining a "common sense" understanding of the world. Figure 7.6 summarizes this fundamental tradeoff between statistical or semantic learning (associated with the neocortex) and episodic memory (associated with the hippocampus). Figure 7.7 shows the effects of learning rate on integrating across discrete experiences to compute an overall probability of a given outcome — only a slow learning rate is able to properly do this integration, instead of just reflecting the very recent past as occurs with faster learning rates.

Consistent with this basic tradeoff, people with exceptional episodic memory abilities (as discussed earlier) often suffer from a commensurate difficulty with generalizing knowledge across episodes. Even more extreme, autistic memory savants, who can memorize all manner of detailed information on various topics,

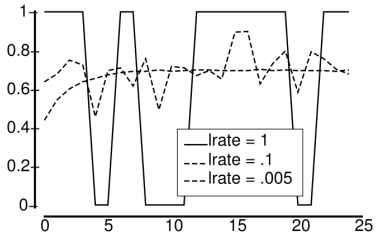


Figure 7.7: Effect of learning rate on ability to integrate across discrete events to compute an overall probability. A simple model experienced an event with a .66 probability as an output unit being active or not. With a very fast learning rate (lrate = 1) it only memorizes the very last outcome, shown by the line alternating between 1 and 0. With learning rate of .005, it is able to integrate across experiences to accurately and stably compute the overall probability. lrate = .1 shows an intermediate case with a lot of variability and thus uncertainty about the true probability.

generally show an even more profound lack of common sense reasoning and general ability to get by in the real world. In these cases, it was speculated that the neocortex also functions much more like a hippocampus, with sparser activity patterns, resulting in overall greater capacity for memorizing specifics, but correspondingly poor abilities to generalize across experiences to produce common sense reasoning (J. L. McClelland, 2000).

7.3.1 Amnesia: Anterograde vs. Retrograde

Having seen how the intact hippocampus functions, you may be wondering what goes wrong to produce **amnesia.** The hollywood version of amnesia involves getting hit on the head, followed by a complete forgetting of everything you know (e.g., your spouse becomes a stranger). Then of course another good whack restores those memories, but not before many zany hijinks have ensued. In reality, there are many different sources of amnesia, and memory researchers typically focus on the kind that is caused by direct damage to the hippocampus and related structures, known as *hippocampal amnesia.* The most celebrated case of this is a person known to science as H.M. (Henry Molaison), who had his hippocampus removed to prevent otherwise intractable epilepsy. He then developed the inability to learn new episodic information (**anterograde amnesia**), as well as some degree of forgetting of previously learned knowledge (**retrograde amnesia**). But he remembered how to talk, the meanings of different words and objects, how to ride a bike, and could learn all manner of new motor skills. This was a clear indication that the hippocampus is critical for learning only some kinds of new knowledge.

More careful studies with H.M. showed that he could also learn new semantic information, but that this occurred relatively slowly, and the learned knowledge was more brittle in the way it could be accessed, compared to neurologically intact people. This further clarifies that the hippocampus is critical for episodic, but not semantic learning. However, for most people semantic information can be learned initially via the hippocampus, and then more slowly acquired by the neocortex over time. One indication that this process occurs is that H.M. lost his most recent memories prior to the surgery, more than older memories (i.e., a temporally-graded retrograde gradient, also know as a Ribot gradient). Thus, the older memories had somehow become *consolidated* outside of the hippocampus, suggesting that this gradual process of the neocortex learning information that is initially encoded in the hippocampus, is actually taking place. We discuss this process in the next section.

Certain drugs can cause a selective case of anterograde amnesia. For example, the benzodiazepines (including the widely-studied drug *midazolam*) activate GABA inhibitory neurons throughout the brain, but benzodiazepene (GABA-A) receptors are densely expressed in the hippocampus, and because of the high levels of inhibition, it is very sensitive to this. At the right dosage, this inhibition is sufficient to prevent synaptic plasticity from occurring within the hippocampus, to form new memories, but previously-learned memories can still be reactivated. This then gives rise to a more pure case of anterograde, without retrograde, amnesia. Experimentally, midazolam impairs hippocampal-dependent rapid memory encoding but spares other forms of integrative learning such as reinforcement learning (Michael J. Frank, O'Reilly, & Curran, 2006; Hirshman, Passannante, & Arndt, 2001).

Another source of amnesia comes from Korsakoff's syndrome, typically resulting from lack of vitamin B1

due to long-term alcoholism. This apparently affects parts of the thalamus and the mammillary bodies, which in turn influence the hippocampus via various neuromodulatory pathways, including GABA innervation from the medial septum, which can then influence learning and recall dynamics in the hippocampus.

7.3.2 Memory Consolidation from Hippocampus to Neocortex

Why do we dream? Is there something useful happening in our brains while we sleep, or is it just random noise and jumbled nonsensical associations? Can you actually learn a foreign language while sleeping? Our enduring fascination with the mysteries of sleep and dreaming may explain the excitement surrounding the idea that memories can somehow migrate from the hippocampus to the neocortex while we sleep. This process, known as **memory consolidation**, was initially motivated by the observation that more recent memories were more likely to be lost when people suffer from acquired amnesia, as in the case of H.M. discussed above. More recently, neural recordings in the hippocampus during wakefulness and sleep have revealed that patterns of activity that occur while a rat is running a maze seem to also be reactivated when the animal is then asleep. However, the measured levels of reactivation are relatively weak compared to the patterns that were active during the actual behavior, so it is not clear how strong of a learning signal could be generated from this. Furthermore, there is considerable controversy over the presence of the temporally-graded retrograde gradients in well-controlled animal studies, raising some doubts about the existence of the consolidation phenomenon in the first place. Nevertheless, on balance it seems safe to conclude that this process does occur at least to some extent, in at least some situations, even if not fully ubiquitous. In humans, slow wave oscillations during non-REM sleep are thought to be associated with memory consolidation. Indeed, one recent study showed that external induction of slow wave oscillations during sleep actually resulted in enhanced subsequent hippocampal-dependent memories for items encoded just prior to sleep (Marshall, Helgadóttir, Mölle, & Born, 2006).

One prediction from the complementary learning systems perspective regarding this consolidation process is that the information encoded in the neocortex will be of a different character to that initially encoded by the hippocampus, due to the very different nature of the learning and representations in these two systems. Thus, to the extent that episodic memories can be encoded in the neocortex, they will become more "semanticized" and generalized, integrating with other existing memories, as compared to the more distinct and crisp pattern separated representations originally encoded in the hippocampus. Available evidence appears to support this idea, for example by comparing the nature of the intact memories from hippocampal amnesics to neurologically intact controls.

7.3.3 Role of Space in the Hippocampus

A large amount of research on the hippocampus takes place in rats, and spatial navigation is one of the most important behavioral functions for a rat. Thus, it is perhaps not too surprising that the rat hippocampus exhibits robust **place cell** firing (as shown in Figure 7.3), where individual DG, CA3 and CA1 neurons respond to a particular location in space. A given neuron will have a different place cell location in different environments, and there does not appear to be any kind of topography or other systematic organization to these place cells. This is consistent with the random, diffuse nature of the perforant pathway projections into these areas, and the effects of pattern separation.

More recently, spatial coding in the entorhinal cortex has been discovered, in the form of **grid cells**. These grid cells form a regular hexagonal lattice or grid over space, and appear to depend on various forms of oscillations. These grid cells may then provide the raw spatial information that gets integrated into the place cells within the hippocampus proper. In addition, **head direction cells** have been found in a number of different areas that project into the hippocampus, and these cells provide a nice *dead reckoning* signal about where the rat is facing based on the accumulation of recent movements.

The combination of all these cell types provides a solid basis for spatial navigation in the rat, and various computational models have been developed that show how these different signals can work together to support navigation behavior. An exploration model of this domain will be available in a future edition.

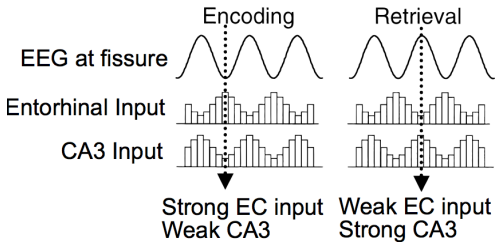


Figure 7.8: Different areas of the hippocampal system fire out of phase with respect to the overall theta rhythm, producing dynamics that optimize encoding vs. retrieval. We consider the strength of the EC and CA3 inputs to CA1. When the EC input is strong and CA3 is weak, CA1 can learn to encode the EC inputs. This serves as a plus phase for an error-driven learning dynamic in the Leabra framework. When CA3 is strong and EC is weak, the system recalls information driven by prior CA3 -> CA1 learning. This serves as a minus phase for Leabra error-driven learning, relative to the plus phase encoding state. (adapted from Hasselmo et al, 2002)

7.3.4 Theta Waves

An important property of the hippocampus is an overall oscillation in the rate of neural firing, in the so-called *theta* frequency band in rats, which ranges from about 8-12 times per second. As shown in Figure 7.8, different areas of the hippocampus are out of phase with each other with respect to this theta oscillation, and this raises the possibility that these phase differences may enable the hippocampus to learn more effectively. Hasselmo and colleagues argued that this theta phase relationship enables the system to alternate between encoding of new information vs. recall of existing information (Hasselmo, Bodelon, & Wyble, 2002). This is an appealing idea, because as we discussed earlier, there can be a benefit by altering the hippocampal parameters to optimize encoding or retrieval based on various other kinds of demands.

The Emergent software now supports an extension to this basic theta encoding vs. retrieval idea that enables Leabra error-driven learning to shape two different pathways of learning in the hippocampus, all within one standard trial of processing (Ketz, Morkonda, & O'Reilly, 2013). Each pathway has an effective minus and plus phase activation state (although in fact they share the same plus phase). The main pathway, trained on the standard minus to plus phase difference, involves CA3-driven recall of the corresponding CA1 activity pattern, which can then reactivate the EC and so on. The second pathway, trained using a special initial phase of settling within the minus phase, is the CA1 <-> EC invertible auto-encoder, which ensures that CA1 can actually reactivate the EC if it is correctly recalled. In our standard hippocampal model explored previously, this auto-encoder pathway is trained in advance on all possible sub-patterns within a single subgroup of EC and CA1 units (which we call a "slot"). This new model suggests how this auto-encoder can instead be learned via the theta phase cycle. See the Appendix *Hippocampus Theta Phase* for details on this *theta phase* version of the hippocampus.

Theta oscillations are also thought to play a critical role in the grid cell activations in the EC layers, and perhaps may also serve to encode temporal sequence information, because place field activity firing shows a *theta phase procession*, with different place fields firing at different points within the unfolding theta wave. We will cover these topics in greater detail in a subsequent revision.

7.3.5 The Function of the Subiculum

The subiculum is often neglected in theories of hippocampal function, and yet it likely plays various important roles. Anatomically, it is situated in a similar location as the entorhinal cortex (EC) relative to the other hippocampal areas, but instead of being interconnected with neocortical areas, it is interconnected more directly with subcortical areas (Figure 7.2). Thus, by analogy to the EC, we can think of it as the input/output pathway for subcortical information to/from the hippocampus. One very important function that the subiculum may perform is computing the relative novelty of a given situation, and communicating this to the midbrain dopamine systems and thence to basal ganglia, to modulate behavior appropriately (Lisman & Grace, 2005). Novelty can have complex affective consequences, being both anxiogenic (anxiety producing) and motivational for driving further exploration, and generally increases overall arousal levels. The hippocampus is uniquely capable of determining how novel a situation is, taking into account the

full *conjunction* of the relevant spatial and other contextual information. The subiculum could potentially compute novelty by comparing CA1 and EC states during the recall phase of the theta oscillation, for example, but this is purely conjecture at this point. Incorporating this novelty signal is an important goal for future computational models.

7.4 Familiarity and Recognition Memory

Stepping back now from the specific memory contributions of the hippocampus, we consider a broader perspective of how the hippocampal system fits into the larger space of human memory capacities. One of the most important questions that researchers have focused on here is whether the neocortex can contribute anything at all to single trial episodic memory. Does a single exposure to a given stimulus leave a big enough trace anywhere in the cortex so as to influence overt behavior? As noted previously, we feel confident that synapses throughout the brain are likely to be affected by every learning experience, but is neocortical learning simply too slow, and the representations too overlapping, to produce a behaviorally significant change from a single experience?

A large body of data suggests that indeed the neocortex can support episodic memory traces, but that they have very different properties compared to those supported by the hippocampus. Specifically, it seems that the perirhinal cortex can produce a useful *familiarity* signal, that indicates in a very coarse manner whether a given stimulus was experienced recently or not. This familiarity signal can be contrasted with the *recollective* memory signal provided by the hippocampus: a full explicit recall of the details of the previous episode when the item was last experienced. The familiarity signal is instead more like a single graded value that varies in intensity depending on how strongly familiar the item is. One hypothesis about the neural basis for this signal is the *sharpness* of the representations in perirhinal cortex — single trials of learning in a generic cortical model leave measurable traces on the overall pattern of neural activity, such that the contrast between strongly active and more weakly active neurons is enhanced (Norman & O'Reilly, 2003). This results from the basic self-organizing learning dynamic we observed in the *Learning* Chapter, where the most strongly activated neurons strengthen their synaptic connections, and thus are better able to out-compete other neurons.

Interestingly, people can obtain subjective conscious access to this familiarity signal, and use it to make overt, conscious evaluations of how familiar they think an item is. The neural mechanism for this explicit readout of a sharpness signal has not been identified. This main challenge here is identifying why signals in perirhinal cortex are consciously accessible, while similar such signals in other neocortical areas do not appear to be accessible to consciousness (as we discuss in the next section).

This combination of hippocampal recall and perirhinal familiarity memory systems is called a *dual process* model of recognition memory, and after many years of controversy, it is now widely accepted in the field. Some of the data consistent with this dual process model include preserved familiarity signals in people with substantial hippocampal lesions, and a variety of neuroimaging and behavioral studies that have been able to distinguish between these two memory signals in various ways.

7.5 Priming: Weight and Activation-Based

Moving further afield from the hippocampus and surrounding cortical areas (e.g., the familiarity signal in the perirhinal cortex), can perceptual and other association cortex areas make useful memory contributions based on single or small numbers of exposures? The answer here is also in the affirmative, but unlike the familiarity signal, these memory traces remain almost entirely below the radar of conscious awareness — scientists can measure memory effects in terms of various behavioral measures, but we are not subjectively aware of having these memories. The general term for this form of memory is **priming,** because the main behavioral manifestation is a speedup in reaction time, or an increased probability of making a particular behavioral response — as if the "pump is being primed" by these memory traces. Indeed, we think of the slow incremental neocortical learning effects as doing exactly this pump priming level of tweaking to the underlying neural representations. Only sustained changes over many experiences can truly reshape these more stable neural representations in more dramatic ways. And as we get older, it seems that perhaps the

learning rate gets slower, making it even more difficult to fundamentally reshape the most basic neocortical representations.

In addition to the subtle effects of slow learning changes, priming can also result from residual activation — neural firing that persists from previously processed information. Thus, we can distinguish between weight-based priming and activation-based priming. As might be expected, activation-based priming is very short-lived, disappearing as soon as the neural firing dissipates. By contrast, weight-based priming can be remarkably persistent, with some cases of priming lasting a year or more, from a single exposure! This kind of behavioral result puts strong constraints on the stability of synaptic plasticity — various computational models introduce forms of synaptic weight decay, but this seems inconsistent with the extreme durability of priming, and of our long-term memories more generally.

One behavioral paradigm used to reveal priming effects is called *stem completion*. Here, the first letters of a word are presented, and the participant is asked to complete the stem with the first word that comes to mind. For example, you might see stems like this:

- win____
- let____

and respond with words like "window" or "winter", "letter" or "lettuce". The priming effect is revealed by first exposing people to one of the possible words for these stems, often in a fairly disguised, incidental manner, and then comparing how much this influences the subsequent likelihood of completing the stem with it. By randomizing which of the different words people are exposed to, you can isolate the effects of prior exposure relative to whatever baseline preferences people might otherwise have. We know that those priming effects are not due to learning in the hippocampus, because they remain intact in people with hippocampal lesions.

7.5.1 Exploration

You can explore both weight-based and activation-based priming on a simple stem-completion like task, using a very generic cortical learning model, in the `priming` simulation in CCN Sims.

7.6 Appendix

- *Hippocampus Theta Phase:* theta phase learning version of the hippocampus.

7.6.1 Hippocampus Theta Phase

This appendix provides more information about the theta phase hippocampus implementation (Ketz et al., 2013), which is used in the `hip` exploration. See leabra/hip on github for the source code and more details about the current implementation.

Here are the three phases of activation dynamics in the network:

- First half of minus phase: the EC input layer drives CA1, which then drives the EC output layer, and CA1 is not influenced by CA3 (in the theta cycle, CA3 is inhibited, but we actually just set its effective weight scale for influencing CA1 to 0). This is a minus phase for training the EC <-> CA1 auto encoder pathway.
- Second half of the minus phase: CA3 now influences CA1 to drive recall, while EC input does not drive CA1. This state at the end of settling is the regular minus phase, which drives learning relative to the plus phase, to train the CA3 -> CA1 recall pathway.
- Plus phase: the EC output layer units are activated directly by the EC input layer activities, such that EC output learns to reproduce the EC input pattern, and CA1 is also in the same state as the first half of the minus phase, where it is being driven by EC input but not CA3. Thus, the target CA1 activity pattern for the CA3 -> CA1 recall connections is the state that properly recalls EC input on the EC output layer, and similarly this same pattern is the target for the EC <-> CA1 auto encoder.