TÜBİTAK

TÜRKİYE BİLİMSEL VE
TEKNİK ARAŞTIRMA KURUMU

THE SCIENTIFIC AND TECHNICAL
RESEARCH COUNCIL OF TURKEY

BİLİŞSEL SÜREÇLERİN YAPAY SİNİR AĞLARI İLE MODELLENMESİ

PROJE NO: EEEAG-126

Elektrik, Elektronik ve Enformatik Araştırma Grubu

Electric, Electronics and Informatics Research
Grant Committee

BİLİŞSEL SÜREÇLERİN YAPAY SİNİR AĞLARI İLE MODELLENMESİ

PROJE NO: EEEAG-126

UGUR HALICI
ASLI GÜLÖKSÜZ
UMUR TALASLI

EKİM 1997
ANKARA

ÖNSÖZ

Bilişsel süreçlerin Yapay Sinir Ağları ile modellenmesini amaçlayan bu proje ODTÜ Elektrik ve Elektronik Bölümü'nde ODTÜ Psikoloji Bölümü'nün danışmanlığı altında yürütülmüş ve proje TÜBİTAK tarafından EEEAG-126 proje kodu altında desteklenmiştir.

Projede önerilen ve kullanılan modeller, canlılarda gözlenen bazı davranışsal ve nörofizyolojik bilgilerden esinlenilerek ve psikoloji deneylerinde elde edilen verilere uygun düşecek biçimde tasarlanmışlardır. Bu modellerin gerçek nöral ağların nörobiyoljik yapılarını açıklama iddiası yoktur; modellemeye çalışılan bu sistemlerin işlemsel yönüdür ve doğal öğrenmeyle ilgili bazı bilgilerin akıllı sistemlerin geliştirilmesinde kullanılmasını amaçlamaktadır. Bu projede önerilen öğrenme biçimlerinin otomatik kontrol, navigasyon, robotik gibi konularda kullanılabileceği bir çok gerçek uygulama alanı mevcuttur.

ŞEKİLLER VE TABLOLAR

ÖZ

Bilişssel süreçlerin Yapay Sinir Ağları ile modellenmesini amaçlayan bu projede canlılardaki Çağrışımsal öğrenmenin iki temel çeşidi olan Koşullu Öğrenme ve Pekiştirimli Öğrenme süreçleri incelenmiştir.

Koşullu öğrenmeyi modellemek üzere kullanılmakta olan Geri Döngülü Çağrışımsal Dipol (READ) devresi üzerinde öğrenmenin sönümünü modellemek üzere değişiklikler yapılarak devrenin birincil ve ikincil koşullandırma altında çalışması gözlenmiş, ayrıca birden fazla READ birimin birarada çalışması incelenmiştir.

Pekiştirimli öğrenmenin modellenmesi amacıyla Rassal Sinir Ağları (RSA) kullanılmıştır. Tekli karar adımları ve zincirleme karar adımları için pekiştirimli öğrenme stratejileri önerilmiştir. Ödül, ceza ve ödül beklentisinin göz önüne alındığı durumlar için öğrenme kuralları geliştirilerek öğrenmenin sönümü ve sistemin değişen çevre koşullarına uyumu incelenmiştir.

Anahtar Kelimeler:
Bilişsel Süreçler, Yapay Sinir Ağları, Yapay Zeka, Koşullu Öğrenme, Pekiştirimli Öğrenme, Öğrenmenin Sönümü, Çağrışımsal Geridöngülü Dipol (READ), Rassal Sinir Ağları

ABSTRACT

In this project that aims to model cognitive processes by using Artificial Neural Networks, the two basic paradigm of associative learning, which are the Conditioned Learning and the Reinforcement Learning are examined.

The Recurrent Associative Dipole (READ) circuit which is used for modelling Conditioned Learning is modified to handle extinction and its operation is observed for primary and secondary conditioning. Furthermore, the operation of multiple READ units together is examined.

For modelling Reinforcement learning, the Random Neural Networks (RNN) are used. Learning strategies are proposed for single and cascaded decision steps. The learning rules for cases of reward, punishment and expectation of reward are developed. Adaptation of the system to changing environmental conditions and extinction of learning in the system are examined.

Keywords: Cognitive Processes, Artificial Neural Networks, Artificial Intelligence, Conditioned Learning, Reinforcement Learning, Extinction, Recurrent Associative Dipole (READ), Random Neural Networks

İÇİNDEKİLER

1. GİRİŞ

Bilişsel süreçler, kişilerin bilgi toplama, plan yapma ve problem çözmek üzere kullandıkları algılama, bellek ve bilgi işleme ile ilgili zihinsel süreçlerdir. Bilişsel psikoloji davranışların anlaşılmasında zihinsel süreçlerin rolünü vurgulayan genel bir yaklaşımdır. Bilişsel psikoloji davranışları açıklarken bu davranışların zihinsel seviyedeki temsillerini ve bu temsillerin üzerinde çalışarak sonuç üretilmesini sağlayan zihinsel süreçleri kullanır. Bu yaklaşım sadece düşünce veya bilgi üzerindeki çalışmalarla kısıtlı değildir; ilk zamanlarda sadece düşünce ve bilgi konudaki çalışmalar bilişsel psikoloji olarak adlandırılmasına rağmen, son yıllarda bu yaklaşım psikolojinin tüm alanlarına genellenmiştir. 1970'li yıllarda insanların bilgiyi nasıl topladığı ve organize ettiği konusunda yoğunlaşılmış ve bilişsel süreçleri anlamak üzere ortaya çıkan yeni bilim alanına bilişim denilmiştir. Psikolojiye ek olarak bilişim alanına giren konular içinde nöroloji, dilbilimi, felsefe ve bilgisayar bulunmaktadır. Bilgisayar alanında özellikle yapay zeka konusu bilişim ile yakından ilgilidir. Yapay zeka, bilgisayar bilimi ile bilişsel psikolojiyi birleştiren bir araştırma dalıdır. Hem bilgisayar kullanılarak insan düşünce süreçlerinin benzetimlerinin yapılması, hem de zekice davranan ve değişen koşullara uyum gösteren bilgisayar yöntemlerinin geliştirilmesi yapay zeka alanına giren konulardır (Atkinson et al 1985).

Yapay zekada temel olarak sembolist (symbolist) ve bağlantıcı (connectionist) olarak adlandırılan iki yaklaşım bulunmaktadır. Sembolist yaklaşımda kavramların, olayların sembollerle temsil edildiği zeki sistemler üzerinde çalışılırken, bağlantıcı yaklşımda nöron adı verilen temel işleme birimleri ve bu birimlerin birbirine yoğun bir biçimde bağlanmasından oluşan yapay sinir ağları kullanılmaktadır (Halıcı, Akman, Leloğlu 1993)

Yapay Sinir Ağları ya da Nöral Ağlar belirli giriş değerlerine karşılık gelen çıkış değerlerinin eşlendirilmesiyle eğitilmektedirler. Bazı uygulamalarda bir çıkış değeri bulunmamakta, giriş değerleri arasındaki benzerliklere göre yapılan gruplamalar nöral ağ tarafından kendiliğinden ortaya çıkarılmaktadır. Yapay sinir ağlarında bilgi, semboller yerine nöronlar arasındaki bağlantıların kuvvetleri ile temsil edilmektedir. Nöron ağlarının eğitilmesi, bağlantı kuvvetlerinin değiştirilmesiyle sağlanmaktadır.

Nöro-işleme (neurocomputing) konusunun başlangıcı 1943'te Mc Culloch ve Walter Pitts tarafından yazılan bir makaleye dayanmaktadır. Bu makalede, çok basit bir nöron ağının bile, prensipte herhangi bir matematik ya da mantık fonksiyonunu hesaplayabileceği gösterilmiştir.

1949 yılında Donald Hebb, Davranışın Organizasyonu (The Organization of the Behaviour) adlı kitabında, klasik psikolojide incelenen koşullu davranış konusunun, nöronların özelliklerinin doğal bir sonucu olarak ortaya çıktığını söylemiştir. Hebb konuyu kendinden öncekilerin ilerisine taşıyarak, koşullu davranışı nöronlar arası snaps adı verilen bağlantı noktalarındaki kuvvetlerin değişmesiyle açıklayarak, nöronlar arası bağlantı kuvvetlerinin ne şekilde düzenleneceğini gösteren yeni bir öğrenme kuralı öne sürmüştür. Hebb daha sonra bu öğrenme kuralını kullanarak psikolojinin bazı deneysel sonuçlarına niteliksel açıklamalar getirmiştir.

İlk Nöral bilgisayarlar 1950'li yıllarda Marvin Minsky tarafından geliştirilen "Snark" ve Frank Rosenblatt, Charles Wightman tarafından geliştirilen "Mark-I Perceptron"dur.

Bu gün yapay sinir ağları, örüntü tanıma ve optimizasyon alanlarında yaygın olarak kullanılmaktadır. Ancak, en yaygın kullanılan haliyle çok katmanlı bir yapıya ve geriyayılım algoritmalarına dayalı olan yapay sinir ağları işlevsel olarak sınırlı kalmakta; canlıların çevrelerine uyumunu sağlayan bilişsel süreçleri modellemekte yeterli olmamaktadır.

Başlıcaları algılama, öğrenme, seçici dikkat, hedef belirleme olarak sayılabilecek olan bu bilişsel süreçler, hem ortamdaki nesneler, hem de sistemin gereksinimleri ve içsel durumu tarafından belirlenirler. Bu süreçlerin anlaşılması ve modellenmesi, nörobiyoloji, psikoloji, matematik, elektrik mühendisliği veya bilgisayar mühendisliği gibi değişik alanların ortak konusudur. Bu süreçlerden esinlenen yapay sinir ağı sistemlerinde ise amaç, çok genel olarak, sistemin ortamda bulunan birçok nesne ve uyaranın arasında kendi gereksinimlerine ve içsel durumuna göre bazılarını öncelikle algılaması, bunların arasında bağlantılar kurması, beklentiler oluşturması, iç ve dış uyaranların durumuna göre hedefler belirleyip bunlarla ilgili eylemlerde bulunması ve eylemlerin ortam üzerindeki sonuçlarına göre yeni hedefler belirlemesidir.

Bu projede Çağrışımlı Öğrenmenin Nöral Ağlarla modellenmesi üzerinde çalışılmıştır. Çağrışımlı öğrenme olaylar arasındaki ilişkilerin öğrenilmesiyle ilgilidir. Çağrışımlı öğrenmenin iki temel biçimi *klasik koşullandırma* ve *işlemsel koşullandırma (Pekiştirimli Öğrenme)* incelenmiştir. Klasik koşullandırmada bir organizma bir olayın diğer bir olayı takip edeceğini öğrenirken, işlemsel koşullandırmada organizma belirli bir sonuca ulaşmayı sağlayan eylemi yapmayı öğrenir.

Koşullu öğrenmeyi (Pavlov koşullandırması) modelleyen sinir ağlarının arasında, daha sonraki teori ve pratikte en çok kullanılanı, Grossberg'in *kapılı dipol* (gated dipole) modeli olmuştur (Grossberg 1991, Grossberg ve Schmajuk 1987a, 1987b, Grossberg ve Schmajuk ve Levine 1992, Baloch ve Waxman 1991, Buanomato, Baxter ve Byre 1990, Raymond, Baxter ve Buonomana 1992 ve 1994). Grossberg, Pavlov koşullandırmasını, sürekli ateşlenen sinir hücrelerinde iletken maddelerin azalması gibi fiziksel bir özellikten de yararlanarak, birbiriyle rekabet eden iki kutuptan oluşan bir devre ile modellemiştir. Daha sonra Grossberg ve Schmajuk bu modeli ikincil koşullandırmayı da içerecek biçimde değiştirerek *Geridöngülü Dipol* (Recurrent Associative Dipole: READ) devresini oluşturmuşlardır. (Grossberg ve Schmajuk 1987a, 1987b)

READ devresi, bilişsel süreçleri modelleyen bir yapay sinir sisteminde temel yapı taşlarından biri olarak kullanılabilir. Ancak bu modelde, Pavlov koşullanması bir kez gerçekleşince, daha sonraki deneylerde beklentilerin yerine gelmemesi durumunda bile koşullandırmanın unutularak sönümü mümkün olmamakta; bu özellik modelin yeni durumlara uyum sağlama yeteneğini azaltmaktadır. Her ne kadar READ devresine ek olarak kullanılabilen. uyaranlar arası rekabete dayalı sinir ağları ile seçici unutma gerçekleştirilebilse de bu tür devrelere gerek duyulması sistemi karmaşıklaştırarak modellenmesi amaçlanan canlı yapılardan uzaklaştırmakta ve ayrıca kullanılabilecek READ devresi sayısını sınırlamaktadır. Bu tür bir çalışma, MIT Lincoln Laboratuarında geliştirilen MAVIN isimli robotta deneme amacı ile kullanılmıştır (Baloch ve Waxman 1991). Bu çalışmada yalnızca üç objenin bulunduğu bir ortam yaratılmış ve robotun bu objeler ile ilgili toplam altı tepkisi incelenmiştir. Anlamlı sayıda uyaranın bulunduğu bir ortamda işlev görebilecek bir yapay sinir sistemi tasarlamak için, yapı taşlarını daha çok işlevli fakat daha homojen bir biçimde tasarlamak yararlı olacaktır.

Bu projedeki klasik koşullandırma ile ilgili çalışmalarda doğal öğrenmede çok iyi bilinen bir özellik olan ve değişen şartlara uyumu sağladığından yaşamın sürdürülmesi açısından büyük önem taşıyan öğrenmenin sönümü üzerinde yoğunlaşılmıştır. Bağıntıların sönümü, sistemin artık geçerli olmayan bağıntıların unutularak yeni bağıntıların öğrenilmesine olanak tanımaktadır. Öğrenmenin sönümünü sağlamak üzere yukarıda sözü edilen READ devresi, sönümü modelleyecek biçimde değiştirimiş ve değiştirien READ deversi üzerinde birincil ve ikincil koşulandırmanın etkilerini görmek üzere benzetim deneyleri yapılmıştır. Daha sonra birden fazla READ devresinin bir arada kullanıldığı ağın ne şekilde çalıştığını incelemek üzere üç birimli bir ağ üzerinde benzetim çalışmaları yapılmıştır.

Pekiştirimli öğrenme, canlılarda görülen diğer bir çağrışımlı öğrenme biçimidir ve bu öğrenmenin makina öğrenmesi alanında uygulanmasına yönelik olarak bir çok çalışma yapılmaktadır (Tsetlin 1973, Klopf 1974, Sutton 1984, Sutton 1988, Barto ve Sutton 1989, Narendra ve Thathacher 1989). Bu projedeki Pekiştirimli öğrenme ile ilgili çalışmalarda ise, oldukça yeni bir yapay sinir ağı modeli olan Rassal Sinir Ağları'nın kullanımıştır. E. Gelenbe tarafından önerilen Rassal Sinir Ağı (RSA) modelinde sinyaller sabit voltaj seviyeleri yerine voltaj vurumları (pulse) ile temsil edilmektedir (Gelenbe 1989, Gelenbe 1990). Sinyallerin sabit voltaj seviyeleri ile temsil edildiği diğer YSA modelleri ile karşılaştırıldığında RSA modeli biyofisiksel nöronlardaki gerçek sinyal iletimini daha iyi temsil etmektedir. RSA için daha önce önerilen Geriiletim türü bir ögrenme algoritması bulunmaktadır (Gelenbe 1993). Bu projedeki çalışmalarda ise RSA modeli için ödüle dayalı yeni bir pekiştirimli öğrenme stratejisi geliştirilmiş ve bu strateji zincirleme karar aşamaları içeren labirent öğrenme üzerinde denenmiştir. Bu öğrenme stratejisinin kullanıldığında durağan (stationary) çevrelerde başarılı sonuçlar alınmıştır. Ancak çevrenin durağan olmadığı durumlarda sistem daha önce öğrenilen eyleme takılmakta ve öğrenmede sönüm mümkün olmamaktadır. Çalışmanın bir sonraki aşanmasında, ödül ile öğrenme için daha önce önerilen öğrenme stratejisi ceza için genelleştirilmiştir. Çalışmanın bu kısımında ayrıca pekiştirim beklentisinin gözönüne alındığı bir öğrenme stratejisi de önerilmiştir. Böyle bir strateji bir yandan öğrenmenin sönümüne olanak tanırken diğer yandan en çok ödül getiren davranışa bir çok durumda tam yakınsama da sağlayabilmektedir.

Raporun bundan sonra yer alan 2. bölümünde çağrışımsal öğrenmenin temel iki biçimi olan koşullu öğrenme ve pekiştirimli öğrenme ile ilgili bilgi verilmiştir. 3. bölümde orjinal READ devresi ve bu devre üzerinde tarafımızdan yapılan değişiklikler açıklandıktan sonra birincil ve ikincil koşullandırma için elde edilen benzetim sonuçları ve birden fazla READ devresi kullanılan sistem için elde edilen benzetim sonuçları verilmiştir. Raporun 4. Bölümünde RSA modeli açıklandıktan sonra bu model için önerilen ödül ile öğrenme kuralı ve bunu kullanan bir pekiştirimli öğrenme stratejisi önerilmiş ve labirentler için elde edilen benzetim sonuçları verilmiştir. Bu bölümde daha sonra ceza ve ödül beklentisi ile öğrenme için stratejiler önerilmiş ve benzetim sonuçları verilmiştir. 5. Bölümde projenin tümü hakkında varılan sonuçlar açıklanmış ve değerlendirilmiştir.

2. CANLILARDA ÖĞRENME

Bu bölümde canlı organizmalarda görülen öğrenme ile ilgili temel bilgiler verilmektedir. Bu amaçla Çağrışımsal Öğrenmede iki temel yaklaşım olan Klasik Koşullandırma ve İşlemsel Koşullandırma (Pekiştirimli Öğrenme) açıklanmakta ve ayrıca konuyla ilgili olarak öğrenmenin sönümü, öğrenmede son-zaman etkisi, labirent öğrenme ve bilişsel haritalar anlatılmaktadır.

2.1. KLASİK KOŞULLANDIRMA İLE ÖĞRENME

2.1. 1. Pavlov Deneyleri

Klasik koşullandırma, Pavlov'un köpeklerle yaptığı çok bilinen deneydeki öğrenme biçimidir (Hulse, Egeth and Deese 1980). Bu deneyde, bir zil sesinden sonra köpeğe bir parça et verilmektedir. Bu işlem belli bir süre yinelendikten sonra köpek zil sesinden sonra yiyecek geleceğini öğrenmekte ve zil sesini duyunca salya akıtmaya başlamakta, böylece et parçasına verdiği tepkiyi zil sesine vermeye başlamaktadır.

2.1.2. Klasik Koşullandırmanın Öğeleri

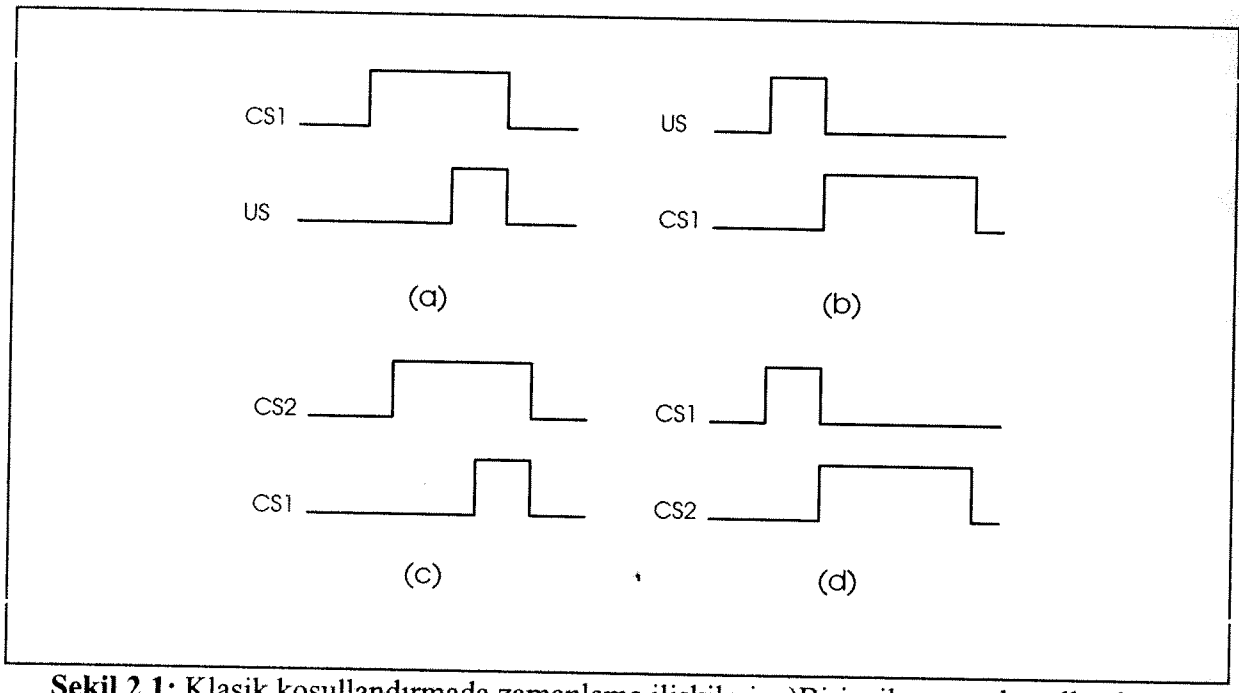
Klasik koşullandırmada, öğretilmeye gerek olmadan bir tepkiye yol açan uyaranlara *koşulsuz uyaran (unconditioned stimulus: US)* denir (Hulse, Egeth and Deese 1980, Klein 1996) (bkz. Şekil 2.1a). Pavlov'un yukarıda anlatılan deneyinde koşulsuz uyaran, salya salgılanmasına neden olan et parçasıdır. Koşulsuz uyaranın yol açtığı tepkiye koşulsuz tepki *(unconditioned response: UR)* denir. Pavlov'un deneyinde salya salgılanması koşulsuz tepkidir. *Öğretilmiş uyaran (conditioned stimulus: CS)* başlangıçta sözkonusu tepki konusunda etkisiz olduğu halde, US ile uygun aralıklarla eşlenerek bir tepkiye yol açan uyarandır. *Öğretilmiş tepki (conditioned response: CR)* ise CS'nin yol açtığı tepkidir. Pavlov'un deneyinde CS, zil sesidir. Öğretilmiş tepki, genellikle koşulsuz tepkiye çok benzemekle birlikte, bunlar arasında büyüklük farkı olabilir.

2.1.3. Tepkinin Ölçülmesi

Klasik koşullandırma sonucunda temel bir güdünün karşılanması suretiyle koşullanan uyaran için de temel güdünün karşılanmasını sağlayan koşulsuz uyarana gösterilen tepkiye benzer bir tepki gösterilmeye başlar. Örneğin Pavlov deneyinde bu tepki zil sesine karşı salya salgılanması şeklinde ortaya çıkmaktadır. Bu tepkinin büyüklüğü salgılanan salya miktarının ölçülmesiyle anlaşılabilir.

2.1.4. Birincil ve İkincil Koşullandırma

Klasik koşullandırma yoluyla bir tepkiye yol açmayı öğrenen öğretilmiş uyaran, daha sonra başka bir uyaranın koşullandırılmasında kullanılabilir (bkz. Şekil 2.1c ve d). Örneğin Pavlov'un deneylerinde, zil sesi ile bir ışık eşleştirilirse, köpek, ışığı görünce de salya salgılamaya başlamaktadır. Bu örnekte zil ile et parçası arasındaki ilişkinin öğretilmesi birincil koşullandırma, ışık ile zil arasındaki ise ikincil koşullandırmadır. Birincil ve ikincil koşullandırma arasında, koşullandırmanın etkisi açısından bir fark saptanmamıştır.



Şekil 2.1: Klasik koşullandırmada zamanlama ilişkileri: a)Birincil uyarıcı koşullandırma b)Birincil ketleyici koşullandırma c)İkincil uyarıcı koşullandırma d)İkincil ketleyici koşullandırma.

2.1.5. Uyarıcı ve Ketleyici Koşullandırma

Bu proje kapsamında kullanıldığı anlamıyla uyarıcı koşullandırma, öğretilmiş uyaran ile koşulsuz uyaran arasında bir ilişki kurularak, koşulsuz tepkinin CS'ye öğretilmesidir. Ketleyici koşullandırma ise, US'nin ortadan kalkması ile CS arasında bir ilişki kurulması; böylece US'nin ortadan kalkmasının yol açtığı tepkinin CS'ye öğretilmesidir (Bkz. Şekil 2.1b-d) (Levine 1991).

2.2 PEKİŞTİRİMLİ ÖĞRENME

2.2.1. İşlemsel Koşullandırma

Klasik koşullandırmada, koşullanan tepki tipik olarak koşulsuz uyarana verilen normal tepkinin (örneğin bir köpeğin yemeğe karşı verdiği normal salya tepkisi) benzeridir. Ancak bir köpeğe örneğin oturması veya takla atması öğretilmek istendiğinde koşullu öğrenme yöntemlerinin bir faydası olmayacaktır. Salya gibi koşulsuz tepkilerin köpeğin takla atması için bir kullanım yolunu bulmak mümkün değildir. Köpeğe bir eylemi öğretmek için izlenecek yol, köpeğin istenen eylemi yapamasına olanak tanıdıktan sonra onu, örneğin yiyecek vererek, ödüllendirmektir. Köpek arzu edilen her eylemi tekrarladığında ödüllendirilmeye devam edilirse zamanla arzu edilen eylemi yapmayı öğrenecektir.

Gerçek hayattaki davranışlar da buna benzer bir şekilde oluşmaktadır: tepkiler çevre üzerinde yaptıkları etkiler sonucunda öğrenilmektedirler. Bu tür öğrenme *işlemsel koşullandırma (operant conditioning)* ile öğrenme veya *Pekiştirimli Öğrenme (Reinforcement Learning)* olarak adlandırılmaktadır. Bir organizma belirli bir davranışı yaptığında, bu davranışın daha sonra tekrar edilmesine olan eğilim bu davranışın doğurduğu sonuçlara göre ortaya çıkmaktadır. Eğer bir davranış iyi bir sonuç doğuruyorsa, örneğin bir köpeğin bir topu yakalaması yiyecekle ödüllendiriliyorsa köpeğin bu davranışı yapmaya olan eğilim artacaktır.

2.2.2 Ödül ve Ceza ile Pekiştirim

İşlemsel koşullandırmada bir uyaran ve bir eylem arasındaki ilişkinin bir organizma tarafından öğrenilmesini amaçlar. Bu amaçla, organizmanın davranışlarını, o davranışın sonucunda oluşacak sonuçlara göre ayarlamasına izin verir. Eger organizmanın bir davranışını, o organizma için *arzulanan sonuçlar (favorable consequences)* izlediğinde organizma bu davranışı daha sık yapmaya eğilim göstermekte; organizmanın davranışını *arzulanmayan sonuçlar (unfavorable consequences)* izlediğinde ise, organizma bu davranışı daha seyrekleştirme eğilimi göstermektedir. Arzulanan sonuçlar genel olarak *ödül (reward)* ve arzulanmayan sonuçlar ise *ceza (punishment)* olarak adlandırılmaktadır (Carlson 1977, Szepesvari 1995, Zhang and Canu 1995).

2.2.3 Thorndike Deneyleri

İşlemsel koşullandırma ile ilgili ilk çalışmalar 1898 yılında E. L. Thorndike tarafından tasarlanan bir dizi deneylerle başlamıştır. Yapılan tipik deneylerde aç bir kedi kapısında basit bir sürgü bulunan bir kafese konulmakta ve kafesin hemen dışında bir balık parçası bulunmaktadır. Başlangıçta kedi kafesin parmaklıkları arasından kolunu uzatarak balığa ulaşmaya çalışmaktadır. Bu davranış bir sonuca ulaşamadığında, kedi kafesin içinde hareket etmekte ve bir çok değişik davranışlar yapmaktadır. Eğer kazara kapının sürgüsüne değerek onu açarsa serbest kalmakta ve balığı yiyebilmektedir. Deneyin bundan sonraki aşamasında kedi kafese geri konulmakta ve dışarıya yeni bir balık parçası yerleştirilmektedir. Kedi, bir takım hareketler yaptıktan sonra tekrar sürgüye çarptığında yine serbest kalarak balığı yiyebilmektedir. Bu işlemler bir çok kez tekrarlandığında, kedi ilgisiz davranışları eleyerek ve kafese konulur konulmaz etkin bir biçimde sürgüyü açmaya yönelmektedir. Böylece kedi yiyeceğe ulaşabilmek için kafesin sürgüsünü açmayı öğrenmektedir. Böyle bir öğrenmede kedinin deneme yanılma yoluyla yaptığı davranışlardan birini bir ödül takip ettiğinde, o davranışın kuvvetlendiği gözlenmektedir.

2.2.4 Skinner Deneyleri

İşlemsel öğrenme ile ilgili kavramlar, B.F. Skinner tarafından düzenlenen deneylerle daha da geliştirilmiştir. Skinner deneylerinde aç bir hayvan, genellikle bir fare veya bir güvercin, özel bir kutuya konulmaktadır. Skinner kutusu olarak adlandırılan bu kutuda, kutu dışındaki bir hazneden kutuya bir parça yem düşmesini sağlayan bir pedal bulunmaktadır. Fare kutuya bırakıldığında kutu içinde hareket etmekte, kutuyu araştırmaktadır. Kazara kutudaki pedalı bulup bastığında kutuya bir parça yem düşmekte ve kutuya yem düşmesi fare pedala her bastığında tekrarlanmaktadır. Başlangıçta fare pedala kazara basmakta ama zaman içerisinde fare yemi yer yemez pedala hemen basmayı öğrenmektedir. Yiyecek gelmesi pedala basma eylemini pekiştirmekte ve pedala basma hızı belirgin bir şekilde artmaktadır.

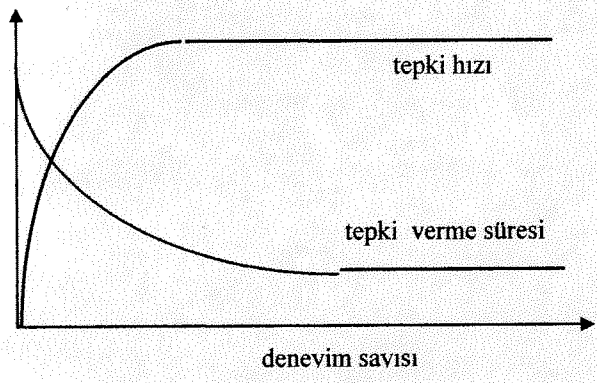
2.2.5 Tepkinin Ölçülmesi

İşlemsel koşullandırma sonucunda genellikle temel bir güdünün karşılanması suretiyle pekiştirilen davranışların yapılmasına eğilim artar. Deney düzenleyen kişiler,işlemsel koşullandırılmış tepkilerin kuvvetlerini çeşitli şekillerde ölçmektedirler. Skinner kutusunda pedal sürekli bulunmasına rağmen fare kendi seçimine göre pedala çok sık veya seyrek basabilir; dolayısıyla organizmanın tepki hızı tepki kuvvetinin ölçülmesinde kullanılabilecek bir ölçüttür. Bu durumda belli bir zaman aralığında bir tepkinin daha sık gözükmesi tepkinin daha kuvvetli olduğunun bir göstergesidir. Tepki kuvveti ölçmek için başka bir olası ölçü sönüm sırasında, deney düzenleyen kişi ödül vermemesine rağmen farenin pedala hala fazla sayıda basması tepkinin kuvvetinin daha büyük olduğunun göstergesidir.

2.2.6. Öğrenme Eğrileri

Psikolojide öğrenme eğrileri bir organizma bir şey öğrendiği sırada olup bitenleri göstermek amacıyla kullanılmaktadır (Hulse et al 1980). Şekil 2.2.de tepki hızının öğrenme süreci içinde nasıl değiştiğini gösteren tipik bir öğrenme eğrisi verilmiştir. Başlangıçta tepki hızı düşüktür, ancak organizma süreç hakkında daha fazla öğrendikçe bu hız zamanla artarak bir doyum

noktasına ulaşmaktadır. Deneyim sayısı arttıkça tepki hızının artmasından öğrenmenin ilerlediği çıkarılabilir. Şekil 2.2'de ayrıca tepki verme süresine göre çizilmiş tipik bir öğrenme eğrisi de verilmiştir. Burada başlangıçta büyük olan olan tepki verme süresi öğrenme ilerledikçe azalmakta ve bir minimuma ulaşmaktadır.



Şekil 2.2: Tipik öğrenme eğrileri

2.2.7. Hemen ve Geciktirilmiş Pekiştirim

Klasik koşullandırmada olduğu gibi, deneme sırasındaki olaylar arasındaki zamansal ilişkiler önem taşımaktadır. Yapılan eylemin hemen ardından verilen pekiştirim *(immediate reinforcement) geciktirilmiş pekiştirimden (delayed reinforcement)* çok daha etkilidir; eylem ile pekiştirim arasındaki zaman farkı arttıkça pekiştirimin etkisi azalmakta, öğrenme zorlaşmaktadır.

2.2.8 Ustalık ve Keşif Çatışması

Pekiştirimli öğrenmede, eğitilen sistem yaptığı çeşitli eylem denemeleri sonucunda elde ettiği pekiştirimin miktarından yaptığı eylemlerden hangisinin ne kadar iyi olduğunu kavrayarak en iyi eylemi öğrenir. Böyle bir sisteme hangi eylemi yapması gerektiği doğrudan öğretilmemekte, sistemin kendisinin en yüksek ödül getiren eylemi keşfetmesi gerekmektedir. Deneme yanılma yoluyla en iyiyi arama, pekiştirimle öğrenmenin en ayırtedici özelliğidir. Ancak bu durum pekiştirimle öğrenmede *ustalık ve keşif (exploitation and exploration)* arasındaki bir çatışmanın ortaya çıkmasına sebep olmaktadır (Narendra, Thathachar, 1989):

1. **Ustalık:** Sistemin yapabilecegi eylemlerin göreceli iyilikleri hakkkında eldeki hazır bilginin kullanılarak bilinen en iyi eylemin yapılması
2. **Keşif:** Gelecekte daha iyi seçebilme yeteneğine sahip olmak üzere çeşitli eylemlerin sonuçları hakkında daha iyi bilgi edinmek için değişik eylemlerin denenmesi.

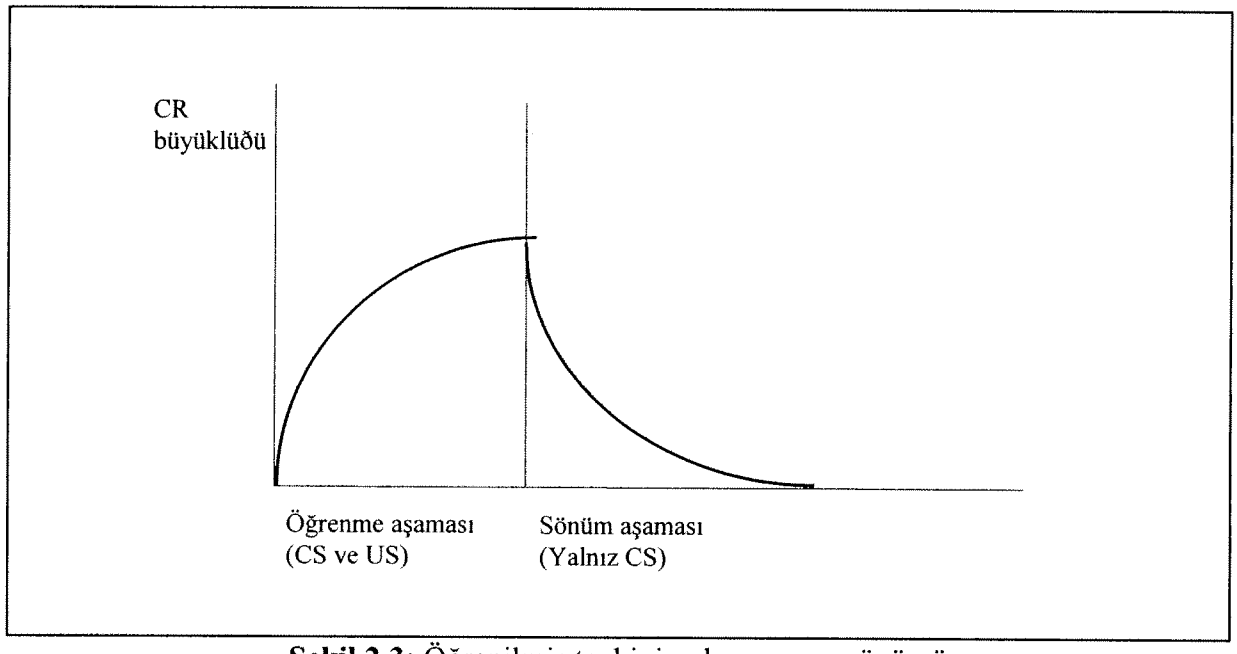
Sadece ödül üzerine dayalı pekiştirimli öğrenme daha önce bilinen en iyi hareketin yapılması yoluyla ustalığı ön plana çıkarırken, diğer eylemlerinin iyiliğinin denenmesini önlemektedir. Çevre şartlarının değiştiği durumlarda, ustalığın yanı sıra keşifin de önemi belirginleşmektedir.

2.3. ÖĞRENMENİN SÖNÜMÜ

Klasik koşullandırma ile ilgili olarak en yaygın kabul gören teori, bu koşullandırmada etken olan ilişkinin bir *öngörürlük ilişkisi* olduğudur (Klein 1996). Başka bir deyişle, bir uyaranın koşullandırma yoluyla öğretilmiş uyaran durumuna gelmesinin nedeni, tutarlı bir biçimde koşulsuz uyaran tarafından izlenmesidir. Öğretilmiş uyaran ile koşulsuz uyaranın eşlendiği öğrenme aşamasından sonra, koşulsuz uyaran verilmeden yalnız öğretilmiş uyaran deneğe gösterilir ve bu işlem belli bir süre yinelenirse, deneğin CS-US bağlantısını unuttuğu gözlenmektedir (Şekil 2.3).

Öğrenmenin sönümü doğal öğrenmede çok iyi bilinen bir özelliktir ve bu özellik değişen şartlara uyumu sağladığından yaşamın sürdürülmesi açısından büyük önem taşımaktadır. Eğer bir organizmanın davranışı ödüllendiriliyorsa ve bu ödüllendirme daha sonra örneğin yiyecek sağlayan mekanizmanın ortadan kaldırılması yoluyla kesiliyorsa, organizma bir süre daha ödül beklentisiyle eylemi sürdürecek ancak zamanla bu eylemler azalarak sönecektir. Öğrenilmiş beklentinin gerçekleşmemesi durumunda tepkilerin bu şekilde zamanla azalmasını tarif etmek üzere psikolojide *sönüm (extinction)* deyimi kullanılmaktadır (Hulse et al. 1980). Bağıntıların sönümü, sistemin artık geçerli olmayan bağıntıların unutularak yeni bağıntıların öğrenilmesine olanak tanımaktadır.

Pekiştirimli öğrenmede, Skinner deneylerinde gözlendiği gibi fare pedala bastıkça kafese yiyecek gelmesi pedala basma eylemini pekiştirmekte ve pedala basma hızı belirgin bir şekilde artmaktadır. Diğer yandan eğer öğrenmenin bir aşamasında yem haznesi ile pedal arasındaki bağlantı kaldırılırsa, bundan sonra pedala basılması yem gelmesini sağlayamamaktadır ve zamanla pedala basma eylemi seyrekleşip yok olmaktadır. Böylece öğrenilen işlemsel koşullandırmalı tepki sönüme uğramaktadır.

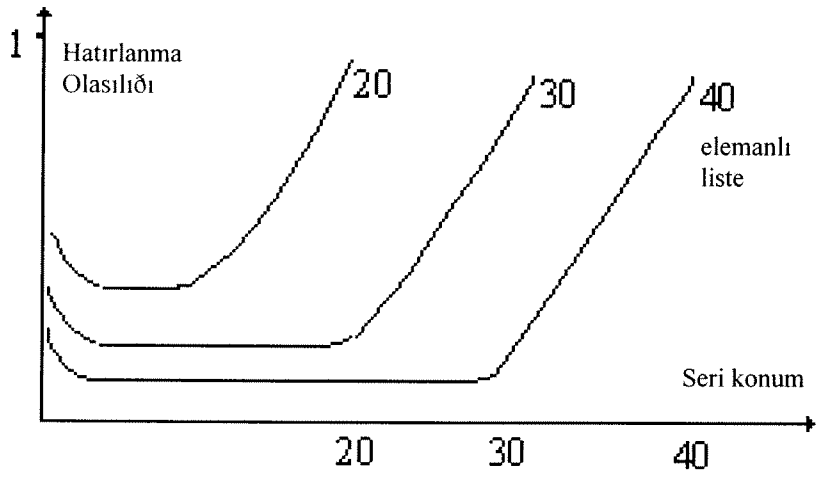


Şekil 2.3: Öğrenilmiş tepkinin oluşması ve sönümü

2.4 LİSTE ÖĞRENMEDE SON_ZAMAN ETKİSİ

Kısa süreli belleğin çalışmasını gözlemlemek amacıyla insanlar üzerinde uygulanan bir yöntem verilen uzunca bir listeyi öğrenerek listedeki kelimelerden mümkün olduğunca fazlasının hatırlanmasını istemektir. Deneylerde listenin en sonundaki bir kaç kelimenin çok iyi hatırlandığı gözlenmiş (Şekil 2.4) ve bu durum *son_zaman etkisi (recency effect)* olarak adlandırılmıştır (Baddaley 1982). Liste öğrenmede önemli diğer bir nokta ise listenin uzunluğudur. Liste uzunluğunun değişmesi, hangi sıradaki kelimenin hangi ölçüde hatırlandığını gösteren hatırlama eğrisinin son_zaman kısmında bir değişik yaratmazken başlangıç ve orta kısımlardaki kelimeleri hatırlanmasının etkilendiği gözlenmiştir. Liste uzadıkça, Şekil 2.4'te

görüldüğü gibi ilk ve orta kısımlardaki kellimelerin hatırlanma olasılığı azalmaktadır (Reynolds and Flagg, 1977)



Şekil 2.4 : Liste öğrenmede kelimelerin listedeki konumlarına göre hatırlanma olasılığı

2.5. LABİRENTLERİN ÖĞRENİLMESİ VE BİLİŞSEL HARİTALAR

Psikolojide davranışcı (behaviorist) yaklaşımı benimseyen araştırmacılar öğrenmeyi sadece gözlenen davranışlarla açıklamak üzere kısıtlarken, bilişsel yaklaşımı benimseyen araştırmacılar yüksek seviyedeki organizmaların, davranışcı yaklaşımda düşünüldüğünden daha zeki yaratıklar olduğunu öne sürmektedirler. Bilişsel yaklaşıma göre bu organizmaların gerçek dünyayı zihinlerinde temsil edebilme yetnekleri vardır ve gerçek dünya yerine gerçek dünyanın zihinsel temsilleri üzerinde işlem yaparak davranmaktadırlar (Atkinson et al, 1990)

Bilişsel yaklaşımın ilk savunucularından Tollmann araştırmalarında karmaşık labirentlerde yolunu bulmayı öğrenen fareler üzerine çalışmıştır. Tollmann bu çalışmalarının sonucunda karmaşık bir labirentte koşan farelerin labirentin yerleşim planını temsil eden bir bir çeşit *bilişsel harita (cognitive map)* geliştirdiğini öne sürmüştür. Bilişsel harita terimi, deney hayvanı tarafından labirentte yolunu bulmayı öğrendikçe oluşturulduğu varysayılan nöral yapıyı tarif etmek amacıyla kullanılmaktadır.

O'Keefe ve Nadel tarafından yapılan yapılan deneylerde fareler labirentleri öğrenirken hippocampustaki nöronal aktiviteler incelenmiş ve elde edilen bulgular ışığında beyinin bu

3. KLASİK KOŞULLANDIRMANIN YAPAY SİNİR AĞLARIYLA MODELLENMESİ

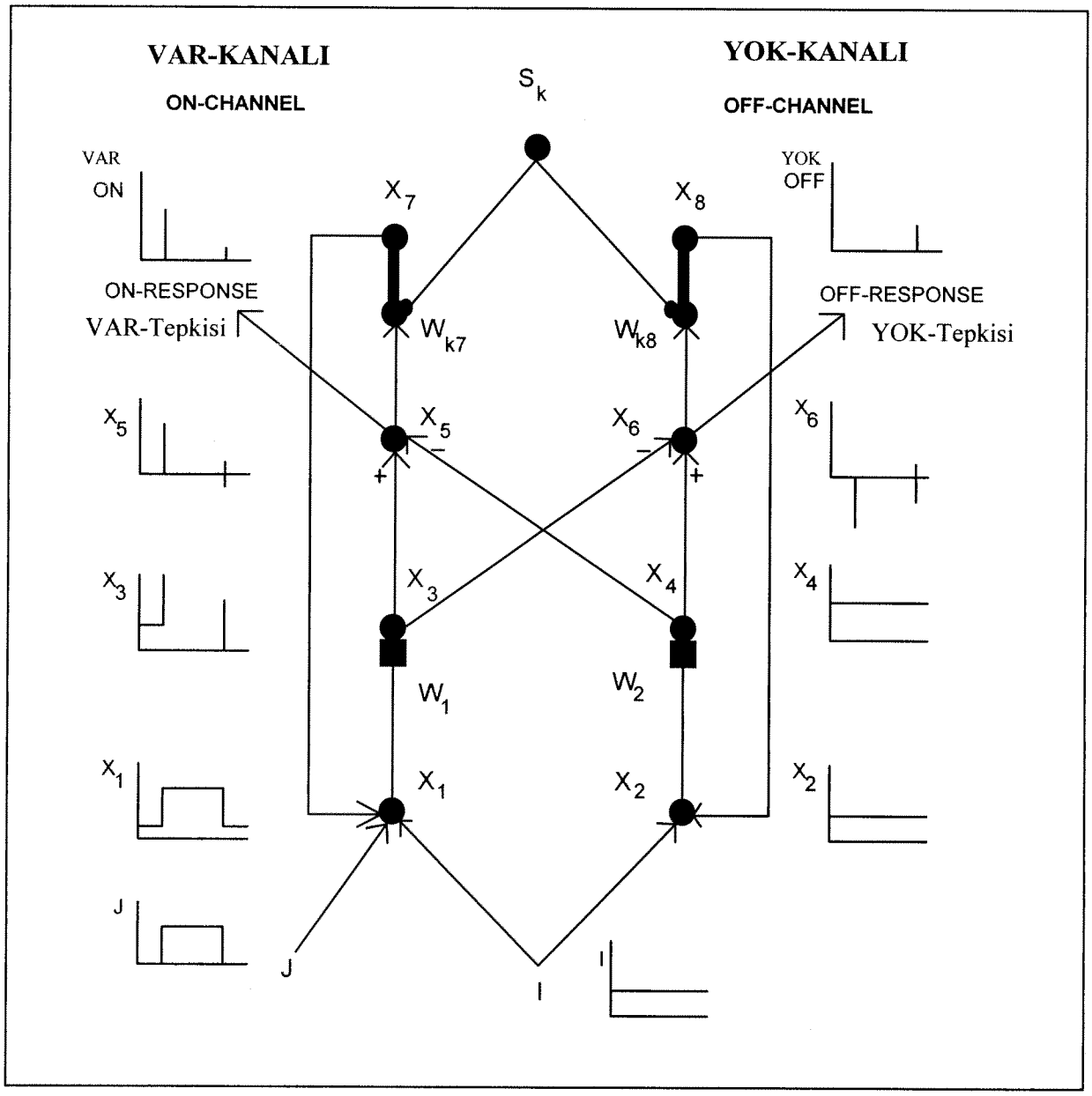
Bu bölümde birincil ve ikincil koşullandırma, uyarıcı ve ketleyici koşullandırma süreçlerini modelleyen READ devresi tanıtıldıktan sonra, READ devresinde, orjinal devrenin olanak tanımadığı edilgen sönümü modelleyen bir değişiklik önerilmekte, ve her iki devre için yapılan simülasyon sonuçları karşılaştırılmaktadır. Bu bölümün son kısımında ise birden fazla READ elemanı içeren ağ ile ilgili yapı ve benzetim sonuçları verilmektedir.

3.1. GERİ DÖNGÜLÜ ÇAĞRIŞIMSAL DİPOL (READ) DEVRESİ

Klasik koşullandırmayı modellemek amacıyla Grossberg ve Schmajuk tarafından tasarlanan Geridöngülü Çağrışımsal Dipol (READ) devresi (Grossberg and Schmajuk 1987.a), iki kanaldan oluşmaktadır. Bunlardan biri, devrenin karşılık geldiği koşulsuz uyaranın neden olduğu tepkiyi (VAR-tepkisi) diğeri ise koşulsuz uyaranın ortadan kalkmasının yol açtığı tepkiyi (YOK-tepkisi) koşullandırmaya yarar. Bu iki kanal, rekabetçi bir ilişki, ya da karşıtlık ilişkisi içinde olduğundan, belirli bir anda bu iki tepkiden yalnız biri etkindir (Bkz.Şekil 3.1).

Şekil 3.1'te görülen iki kanalın girdileri, etkin durumdayken yavaş yavaş azalan, edilgen durumda ise yeniden artan nöro-iletkenler tarafından denetlenmektedir (w_1 ve w_2). Bu durum, kanallardan birinin etkinliğinden sonra diğerinin etkinliğinde bir sıçramaya neden olmakta, böylece YOK-tepkisinin koşullandırılmasını sağlamaktadır (Grossberg 1991).

Şekil 3.1'de J ile gösterilen girdi, koşulsuz uyaran, S_k'ler ise öğrenilmesi olanaklı uyaranlardır. Koşullandırma deneylerinden sonra, koşullu uyaranın yokluğunda bile, öğretilmiş uyaran, öğretilmiş tepkiye yol açmak için yeterli olmaktadır.



Şekil 3.1: Geridöndülü Çağrışımsal Dipol (READ)

READ devresinin işleyişini tanımlayan diferansiyel denklemler aşağıda verilmiştir (Grossberg and Schmajuk 1987.b)

Baz + US + Geridöngü Var-aktivasyonu:

$$\frac{dx_1}{dt} = -ax_1 + I + J + f(x_7) \qquad (3.1)$$

I:Baz input J:VAR-kanalındaki girdi (US)

Baz + Geridöngü Yok-aktivasyonu:

$$\frac{dx_2}{dt} = -ax_2 + I + f(x_2) \qquad (3.2)$$

Azalan var ve yok iletkenleri:

$$\frac{dw_1}{dt} = b(1-w_1) - cg(x_1)w_1 \qquad (3.3)$$

$$\frac{dw_2}{dt} = b(1-w_2) - cg(x_2)w_2 \qquad (3.4)$$

Kapılı var ve yok aktivasyonları:

$$\frac{dx_3}{dt} = -ax_3 + eg(x_1)w_1 \qquad (3.5)$$

$$\frac{dx_4}{dt} = -ax_4 + eg(x_2)w_2 \qquad (3.6)$$

Normalleştirilmiş rekabetçi var ve yok aktivasyonları:

$$\frac{dx_5}{dt} = -ax_5 + (h-x_5)x_3 - (x_5+k)x_4 \qquad (3.7)$$

$$\frac{dx_6}{dt} = -ax_6 + (h-x_6)x_4 - (x_6+k)x_3 \qquad (3.8)$$

Toplam var ve yok aktivasyonları:

$$\frac{dx_7}{dt} = ax_7 + m[x_5]^+ - p\sum S_k w_{k7} \qquad (3.9)$$

$$\frac{dx_8}{dt} = -ax_8 + m[x_6]^+ - p\sum S_k w_{k8} \qquad (3.10)$$

Var-koşullandırılmış ve yok-koşullandırılmış öğrenme:

$$\frac{dw_{k7}}{dt} = S_k\,(-qw_{k7} + r[x_5]^+) \qquad (3.11)$$

$$\frac{dw_{k8}}{dt} = S_k\,(-qw_{k8} + r[x_6]^+) \qquad (3.12)$$

Var-tepkisi ve Yok-tepkisi:

$$\text{VAR} = [x_5]^+ \qquad (3.13)$$

$$\text{YOK} = [x_6]^+ \qquad (3.14)$$

Denklem 3.13 ve 3.14'ten görüleceği gibi, öğrenme, Şekil 3.1'de w_{k7} ve w_{k8} olarak gösterilen, öğretilmiş uyaranlar ile READ devresi arasındaki ağırlıkların artıp azalması yoluyla gerçekleşmektedir. w_1 ve w_2 iletkenlerin azalıp artmasını, x_5 ve x_6 ise kanallar arasındaki karşıtlık ilişkisini modellemektedir. İki kanaldaki geridöngü, hem öğretilmiş uyaranın kanalı harekete geçirmesini sağlamakta, hem de bu yolla ikincil koşullandırmaya olanak tanımaktadır.

READ devresi, birincil ve ikincil uyarıcı ve ketleyici koşulandırmayı modellemekle birlikte, edilgen sönüme olanak tanımamaktadır. Bu devre kullanılarak öğretilen uyaran-tepki

ilişkileri, CS-US ilişkisinin ortadan kalkması durumunda bile sürmektedir. Değişen çevre koşullarına uyum sağlama gücünü azaltan bu durumu ortadan kaldırmak için, READ devresinde aşağıdaki bölümde anlatılacak olan değişiklik önerilmiştir.

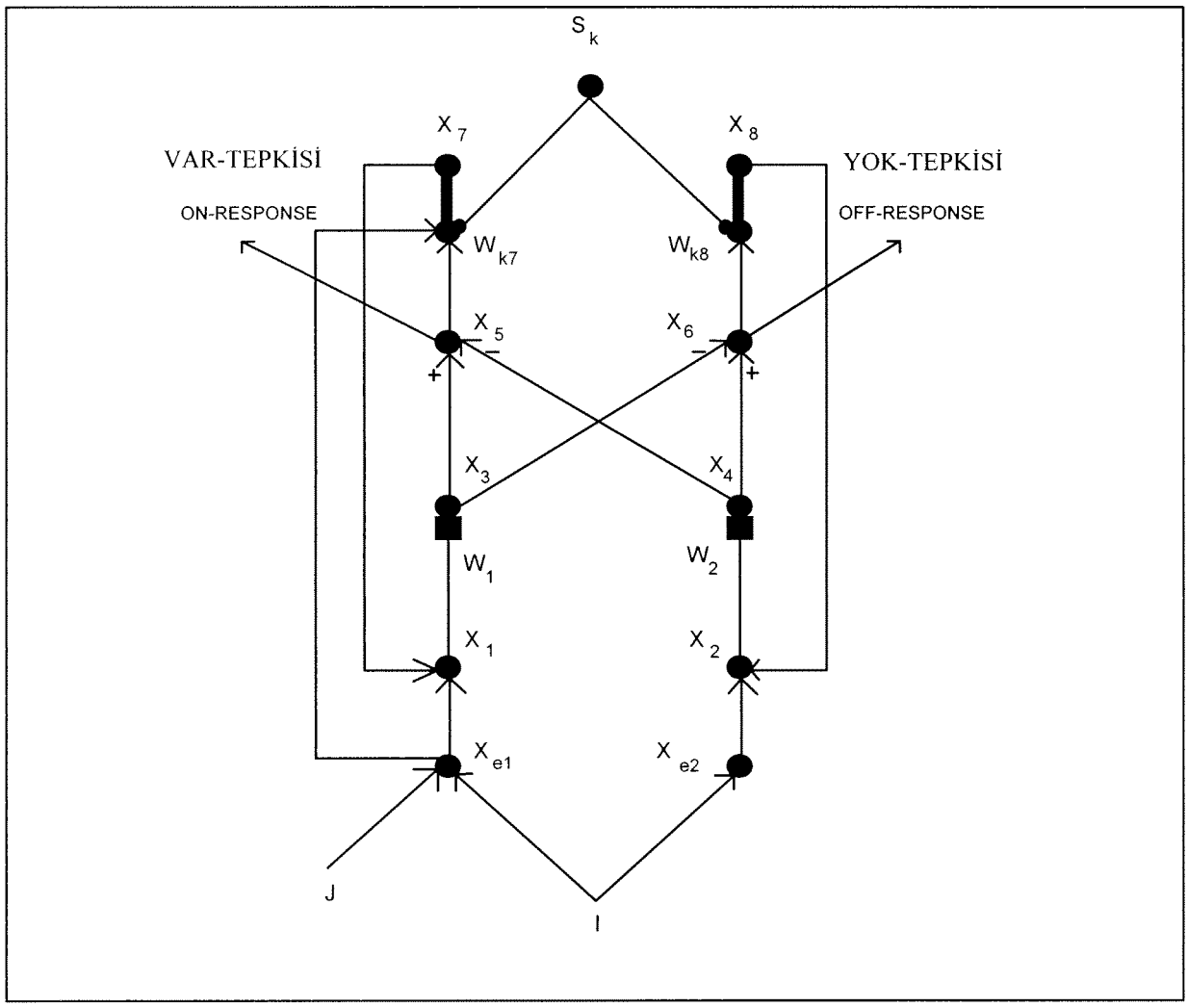
3.2. READ DEVRESİNDE SÖNÜMÜN MODELLENMESİ

Bölüm 3.1'de anlatılan READ devresinde, bir kez öğrenilen uyarı-tepki ilişkileri, öğrenilmiş ve koşulsuz uyaranlar arasındaki öngörürlük ilişkisi ortadan kalksa da sürmektedir. Oysa hayvanlarda, öğrenilmiş uyaran koşulsuz uyaran tarafından izlenmediği, dolayısıyla beklenti gerçekleşmediği zaman, bu bağlantı unutulmaktadır. Örneğin Pavlov'un deneyinde zil-et ikilisi ile koşullandırılan köpeğe, defalarca zil sesinin dinletilip ardından et verilmemesi durumunda, köpek artık zil sesine karşılık salya salgılamamaya başlamaktadır. Öğrenmenin sönümü, canlının davranışlarının geçerliliğini yitiren bağlantılardan etkilenmemesini ve yeni ilişkiler öğrenebilmesini sağladığı için yararlı bir özelliktir. READ devresi, uyarıcı koşullandırmada da, ketleyici koşullandırmada da unutmaya olanak sağlamamaktadır. Ketleyici koşullandırmada öngörürlük ilişkisi bulunmadığından, hayvanlarda da sönüm görülmemektedir. Ancak uyarıcı koşullandırmada, beklentinin gerçekleşmemesi durumunda, daha önce Şekil 2.3'te gösterildiği biçimde sönüm gerçekleşmektedir.

3.2.1 Değiştirilmiş READ Devresi

READ devresinde unutmanın gerçekleşmemesinin nedeni, koşullandırmadan sonra CS'in devreyi aynı US gibi aktive etmesidir. Tarafımızdan değiştirilen READ devresinde (Şekil 3.2) ikincil koşullandırmayı sağlayan geridöngü ile koşulsuz uyaranın verildiği düğüm birbirlerinden ayrılarak, edilgen sönüme olanak sağlanmıştır (Gulöksüz ve Halıcı 1996). Bu devreye eklenen düğümlerin işleyişini tanımlayan diferansiyel denklemler ve orjinal devreden değiştirilerek aktarılan denklemler Şekil 3.2'nin ardından verilmiştir. Bu devrenin öğrenme aşamasındaki işleyişi, Şekil 3.1'deki READ devresininkiyle hemen hemen aynıdır. Unutma aşamasında ise, x_{e1}'in aktivasyonu baz girdinin biraz üstündeki bir eşiğin altına düştüğünde,

başka bir deyişle CS, US tarafından izlenmediğinde, w_{k7} azalmakta, böylece unutma gerçekleşmektedir.



Şekil 3.2: Edilgen sönüme olanak sağlayan değiştirilmiş READ devresi

Eklenen düğümler:

$$\frac{dx_{e1}}{dt} = -ax_{e1} + I + J \qquad (3.15)$$

$$\frac{dx_{e2}}{dt} = -ax_{e2} + I \qquad (3.16)$$

Değiştirilen denklemler:

$$\frac{dx_1}{dt} = -ax_1 + x_{e1} + f(x_7) \qquad (3.17)$$

$$\frac{dx_2}{dt} = -ax_2 + x_{e2} + f(x_8) \qquad (3.18)$$

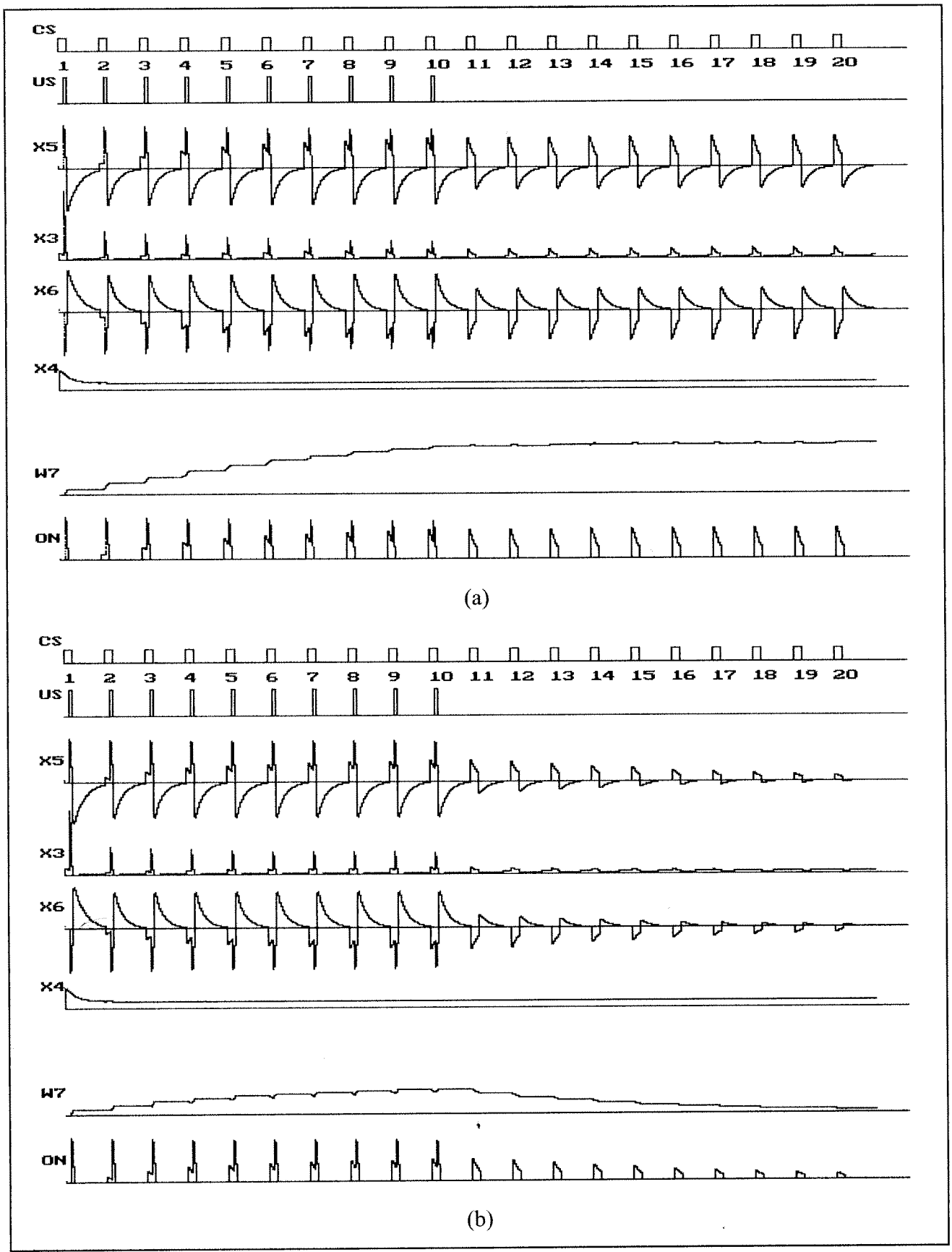
$$\frac{dw_{k7}}{dt} = S_k(-qw_{k7} + r[x_5 q(x_{e1} - I)]^+) \qquad (3.19)$$

3.2.2. Birincil Koşullandırma için Benzetim Sonuçları

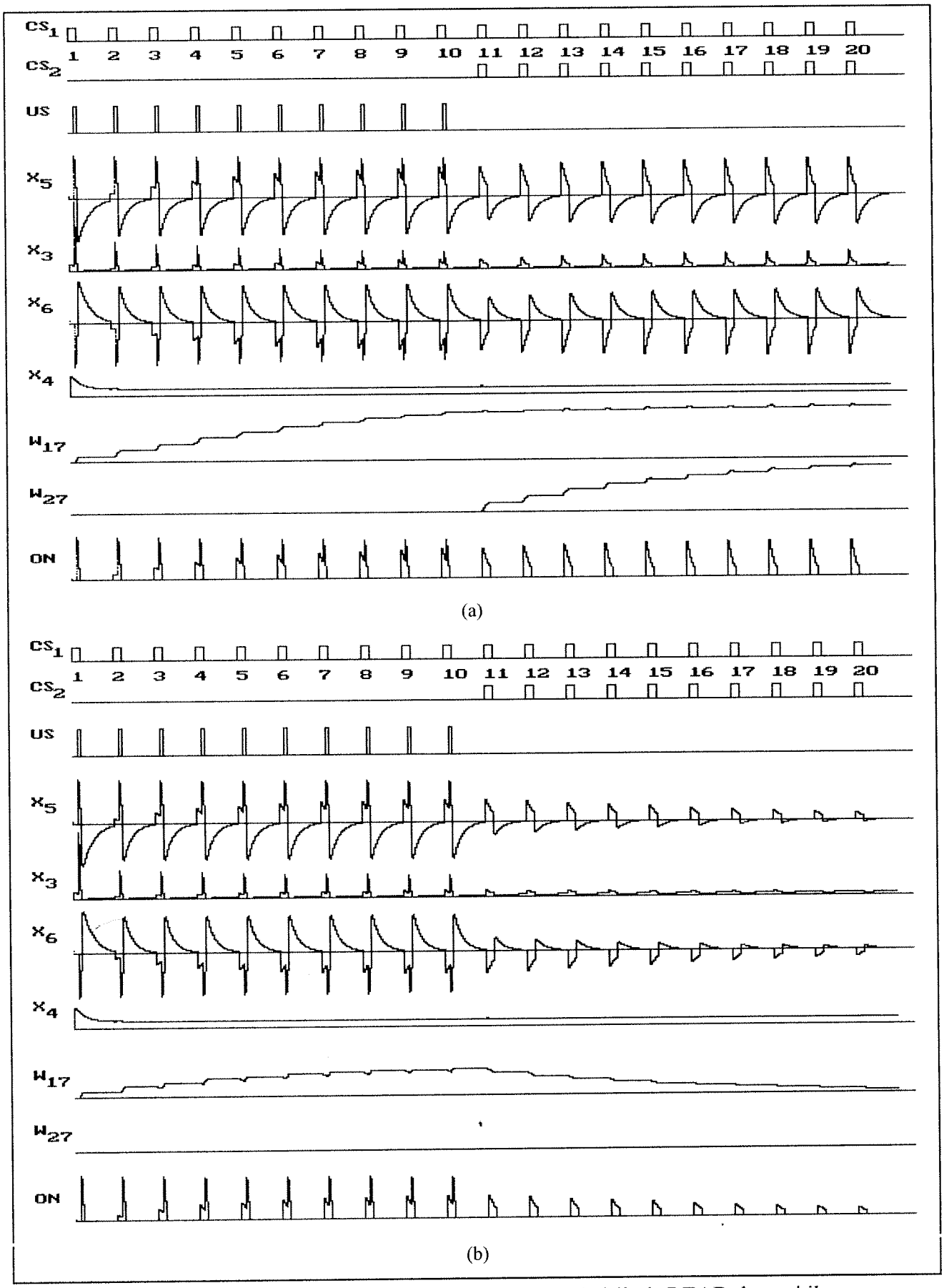
Orjinal READ devresi ve değiştirilmiş READ devresi ile birincil uyarıcı koşullandırma için yapılan benzetim sonuçları Şekil 3.3 a ve b'de görülmektedir. Her iki devre için de öğrenme aşamasında 10 deneyim yapılmış, bu deneyimlerin her birinin ilk 200 zaman birimi boyunca devreye CS, son 40 zaman birimi boyunca da US verilmiştir. Denemeler ilerledikçe, CS'e karşılık VAR-tepkisinin büyüdüğü gözlenmektedir. Değiştirilmiş READ devresi ile yapılan simülasyonda, CS'in verilmesi ile US'in verilmesi arasında geçen zamanda, devre beklentinin gerçekleşmemesini yaşadığı için, bu süre içinde w_7 biraz azalmaktadır. Ancak bu, öğrenme hızını ancak önemsiz derecede azaltmaktadır. Yapılan benzetimlerin sönüm aşamasında yer alan 11-20 numaralı deneyimlerde ise, CS devreye verilmiş, ancak US tarafından izlenmemiştir. Orjinal READ devresi ile yapılan benzetim sonuçlarında, unutma aşaması olması gereken bu denemeler boyunca VAR-tepkisinin büyüklüğü aynı kalmış, sönüm gerçekleşmemiştir. Şekil 3.3.a'da görüldüğü gibi, w_7'da bir küçülme yoktur. Değiştirilmiş READ devresi ile yapılan benzetimde ise, sönüm deneyimleri ilerledikçe VAR-tepkisi küçülmüş ve 10 deneyimin sonunda sıfıra yaklaşmıştır.

3.2.3. İkincil Koşullandırma için Benzetim Sonuçları

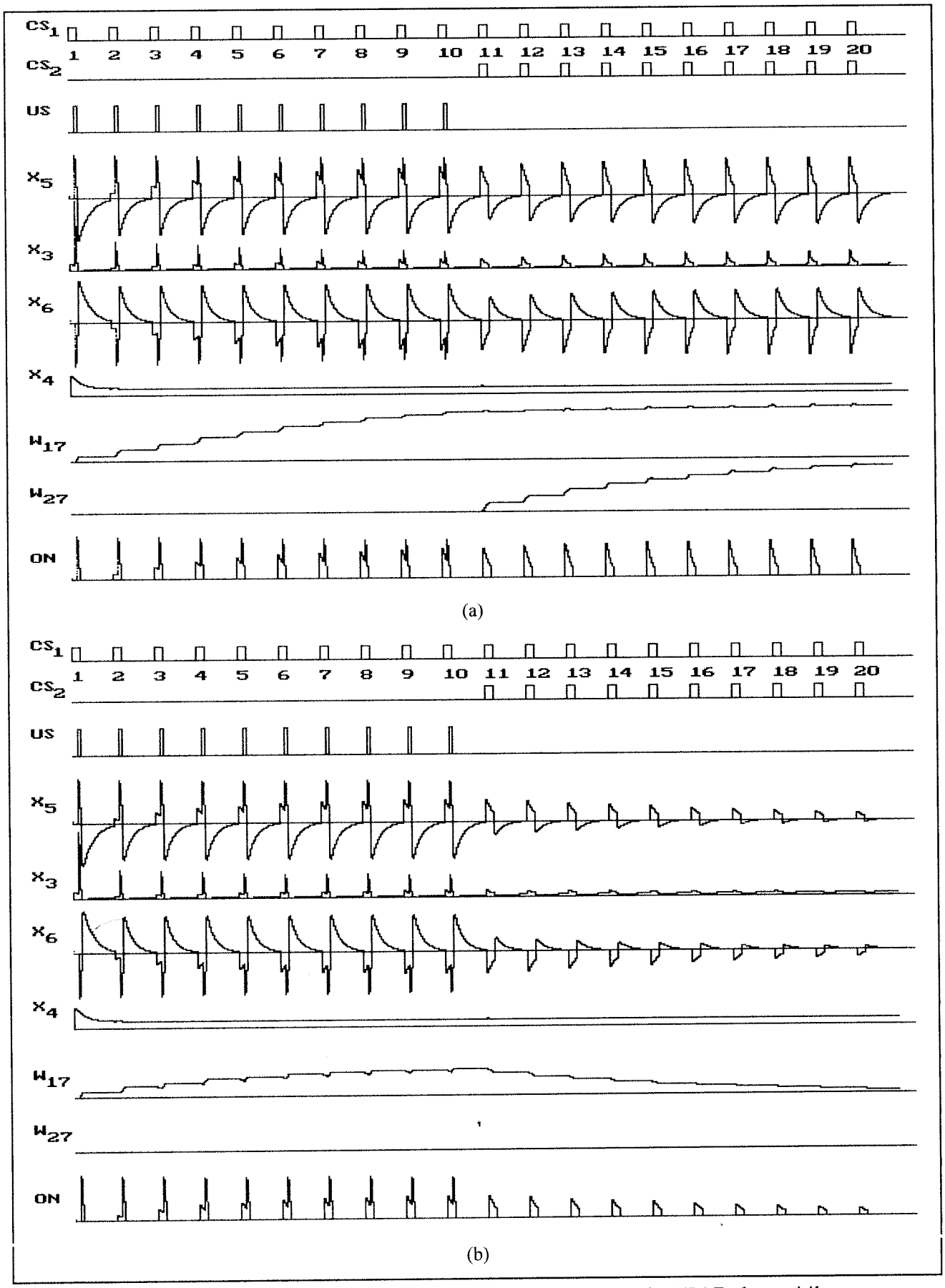
Sönümün uyum sağlama yeteneği açısından en önemli sonuçlarından biri, geçerliliğini yitirmiş bağlantıların, artık yeni ikincil bağlantıların kurulmasına neden olmamasıdır. Örneğin Pavlov'un deneyindeki ikincil koşullandırmada köpek açısından önemli olan ışık-zil bağlantısı değil, ışık-et bağlantısıdır. Şekil 3.4'da, her iki READ devresi ile ikincil koşullandırma için yapılan simülasyon sonuçları görülmektedir. Şekil 3.4.a ve b'de, w17 CS1 ile READ devresi arasında, w27 ise CS2 ile READ devresi arasındaki bağlantı ağırlığıdır. Şekil 3.4.a'da görüldüğü gibi, orjinal READ devresinde, geçerliliği kalmamasına rağmen CS1-US bağlantısından dolayı, CS2-CS1 bağlantısı kurulmaktadır. Değiştirilmiş READ devresi ile yapılan benzetimlerde ise, sönüme uğrayan CS1-US bağlantısından dolayı ortaya çıkan VAR-tepkisinin, ikincil koşullandırmaya yol açmadığı görülmüştür. Bu, sistemin değişen çevre koşullarına uyum sağlama yeteneği açısından olumlu bir sonuçtur.



Şekil 3.3: READ devresi ile birincil uyarıcı koşullandırma. Parametreler: a=1, b=.005, c=.00125, e=20, h=20, k=20, m=.5, q=.005, r=.025, p=20



Şekil 3.4: İkincil Koşullandırma a) READ devresi ile, b) Değiştirilmiş READ devresi ile



Şekil 3.4: İkincil Koşullandırma a) READ devresi ile, b) Değiştirilmiş READ devresi ile

3.3 ÇOK BİRİM İÇEREN AĞ

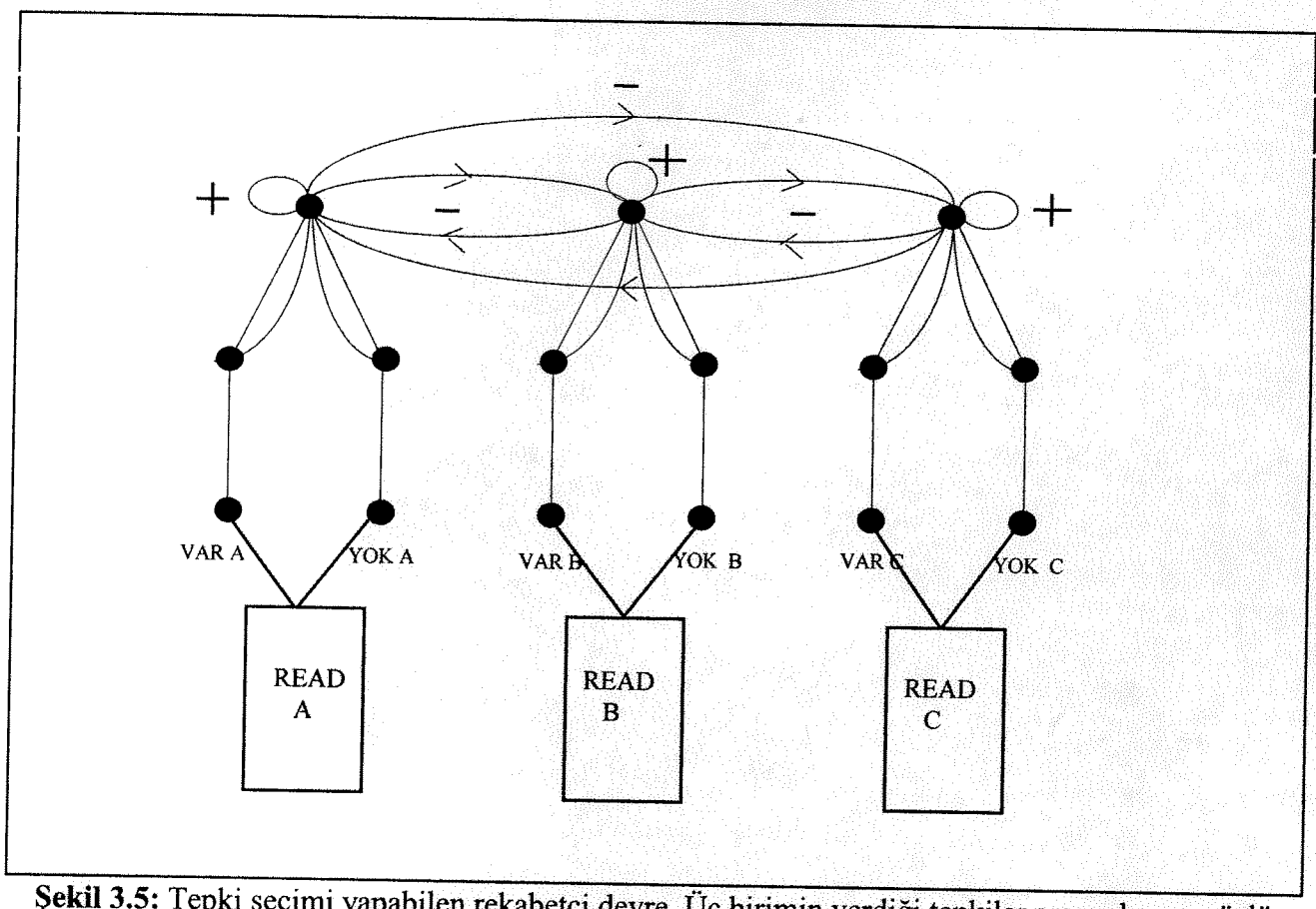
Hayvanlarda, türlere özgü bazı yaşamsal refleksler doğuştan başlayarak gözlenmektedir. Yapay sistemlere de bazı önemli tepkiler öğrenmeye gerek bırakmayacak biçimde sisteme eklenerek, sistemin daha sonra rastladığı uyaranlara karşı tepkisi bu temel uyaran-tepki ilişkileri tarafından koşullandırılabilir. Bu bölümde, Bölüm 3.2'de tanıtılan birim kullanılarak, başlangıçta üç adet koşulsuz uyaran-tepki ilişkisi bulunan bir sisteme, yenilerinin öğretilmesi için tasarlanan bir devre anlatılmaktadır. Bu sistemdeki üç nöral birimin herbiri, bir koşulsuz uyaran-koşulsuz tepki ikilisine karşılık gelmektedir.

3.3.1. Birimler Arasında Yarışmayı Modelleyen Bir Devre

Üç birimden oluşan bu sisteme, iki farklı öğretilmiş uyaran verilmektedir. Bir birimde koşullanmanın yer alıp almayacağı, Şekil 3.1 ve 3.2'te I ile gösterilen baz sinyalinin varlığı veya yokluğu ile ilintilidir. Birden fazla birim işler durumda olduğunda ise sistemin tepkisinin ne olacağı, birimler arasındaki rekabetçi bir devre tarafından belirlenmektedir.

Nöral birimlerin karşılık geldikleri tepkiler, çatışan, dolayısıyla aynı anda verilemeyecek davranışlar olabilir. Bu durumda, sistemin, herhangi bir anda yalnızca bir tepki göstermesi olanaklıdır. Şekil 3.5'de görülen devre, birden fazla birimin VAR ya da YOK kanalında bir çıktı bulunması durumunda, sistemin tepkisinin bunların hangisi olacağını belirlemektedir. Bunun için, Şekil 3.1 ve 3.2'de, eşiği sıfır olan VAR ve YOk tepki düğümlerinden sonra, eşiği daha yüksek olan birer düğüm kullanılmaktadır. Bu düğümlere, rekabetçi bir devreden girdi gelmektedir.

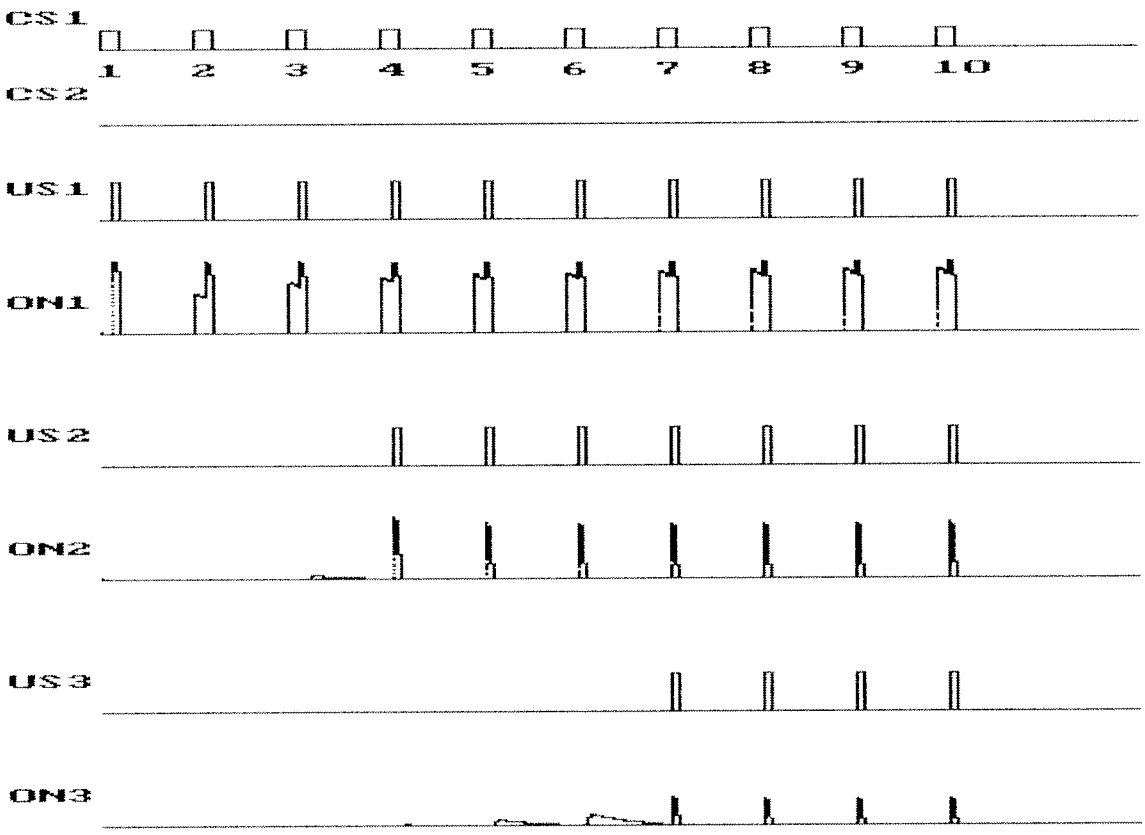
Aynı birimin içindeki VAR ve YOK kanalları arasında zaten bir karşıtlık ilişkisi bulunduğundan, bunların arasında ayrıca bir rekabetçi bağlantıya gerek yoktur. Bu nedenle, rekebatçi devrede, her birim için tek düğüm kullanılmaktadır. Bu düğümler, birbirlerini ketlemekte, kendilerine ise uyarıcı etki yapmaktadırlar.



Şekil 3.5: Tepki seçimi yapabilen rekabetçi devre. Üç birimin verdiği tepkiler arasından en güçlü olanı diğerlerin bastırmaktadır.

3.3.2. Üç Nöral Birimli Sistemin Benzetim Sonuçları

Üç nöral birim içeren ağ için elde edilen benzetim sonuçları Şekil. 3.6 da verilmiştir. Şekilden görüldüğü gibi, birinci READ devresi koşullandırılmaya daha erken başladığı için CS_1 ile arasındaki bağlantı (w_{17}) kuvvetlenmiştir. Dolayısı ile rekabetçi devrede diğer READ'lerin çıktıları etkin olamamaktadır.



Şekil 3.6: Üç nöral birimli sistemde birincil uyarıcı koşullandırma

4. PEKİŞTİRİMLİ ÖĞRENMENİN YAPAY SİNİRAĞLARI İLE MODELLENMESİ

Projenin bu kısımında, bir yapay sinir ağı modeli olan *Rassal Sinir Ağları'nın (RSA)* kullanıldığı bir pekiştirimli öğrenme stratejisi geliştirilmiş ve bu strateji ardışıl karar aşamaları içeren labirent öğrenme üzerinde denemiştir. Daha sonra RSA için ceza ile öğrenme ve beklenti ile öğrenmeye yönelik çalışmalar yapılmış ve benzetim sonuçları verilmiştir (Halıcı 1995, 1996, 1997)

4.1. RASSAL SİNİR AĞI MODELİ

RSA modelinde sinyaller sabit voltaj seviyeleri yerine voltaj *vurumları (pulse)* ile temsil edilmektedir. (Gelenbe, 1989, Gelenbe, 1990). Sinyallerin sabit voltaj seviyeleri ile temsil edildiği diğer YSA modelleri ile karşılaştırıldığında RSA modeli biyofisiksel nöronlardaki gerçek sinyal iletimini daha iyi temsil etmektedir. RSA modelinde aralarında *pozitif* ve *negatif* vurumların iletildiği n nöron bulunmaktadır. Her bir i nöronu kendine gelen vurumları biriktirmektedir. Eğer belirli bir andaki sinyal vurumlarının toplamı pozitif bir değer taşıyorsa, o an için nöron *ateşleme (firing)* yapabilmektedir. Ateşleme sırasında sinyal vurumları rassal zamanlarda üretilmekte, iki vurum arasındaki zaman *r(i) hızındaki (rate)* bir *üssel dağılım (exponential distribution)* ile tanımlanmaktadır. Nöronlar tarafından üretilen vurumlar ağdaki diğer nöronlara veya ağ dışına gönderilmektedir. RSA'da bir nörona gelen vurumlar dışsal kaynaklardan veya ağdaki diğer nöronlardan gelebilir. Ağdaki her bir i nöronu icin, o nörona gelen vurumların birikmesiyle elde edilen potansiyel $k_i(t)$ ile gösterilecek olursa, bu nörona varan her bir pozitif vurum potansiyel seviyesini 1 artırmaktadır. Negatif bir vurum ise, $k_i(t)>0$ olduğu durumda potansiyeli 1 azaltmakta, ancak $k_i(t)=0$ olduğu durumda ise bir değişiklik yapmamaktadır. Bu biyofiziksel nörondaki davranışın basitleştirilmiş bir halidir. Eğer $k_i(t)>0$ ise nöron ateşleyebilmekte ve üretilen her bir sinyal, $k_i(t)$ değerini 1 azaltarak nörondan ayrılmaktadır. i nöronundan ayrılan bir sinyal $p(i,j)=p^+(i,j)+ p^-(i,j)$ olasılığı ile j nöronuna iletilmektedir. Burada, i nöronundan ayrılan bir nöronun j nöronuna pozitif bir sinyal vurumu olarak ulaşması olasılığı $p^+(i,j)$ *ve* negatif bir sinyal vurumu olarak ulaşması olasılığı ise $p^-(i,j)$ ile gösterilmiştir. Üretilen sinyal $d(i)$ olasılığı ile hiç bir nörona iletilmeyip yitebilir. Her bir nöron için

$$\Sigma_j\ p(i,j) + d(i) = 1,\ \ 1 \le i \le n \qquad (4.1)$$

$0 \leq p^+(i,j) \leq 1$ ve $0 \leq p^-(i,j) \leq 1$ (4.2)

bağıntıları geçerlidir.

Nöronlara pozitif ve negatif olmak üzere iki dış kaynaktan gelen vurumlar $\Lambda(i)$ *ve* $\lambda(i)$ hızındaki Poisson süreçlerle temsil edilmektedir.

Nöronlar arası iletişim olasılıkları, YSA literatüründeki nöronlar arası bağlantı kuvvetleri cinsinden ifade edilmek istendiğinde

$$w^+(i,j) = r(i)p^+(i,j) \geq 0, \quad w^-(i,j) = r(i)p^-(i,j) \geq 0, \tag{4.3}$$

olarak yazılabilir, burada

$$r(i) = \sum_j (w^+(i,j) + w^-(j,i)) + d(i), \tag{4.4}$$

bağıntısı ile belirlenmiştir.

4.2 RASSAL SİNİR AĞI İÇİN PEKİŞTİRİMLİ ÖĞRENME

Bu bölümde rassal sinir ağı için önerdiğimiz öğrenme stratejisi önce tek karar adımı içeren sistemler için açıklanmakta ve sonra zincirleme karar adımları içeren daha karmaşık sistemler için genelleştirilmektedir.

4.2.1. Tek Basamaklı Karar Adımları İçin Hemen Pekiştirimli Öğrenme

Bir çevre (environment) ile etkileşim içinde öğrenen bir sistem gözönüne alalım. Bu sistem, her bir n zamanında, $n=0,1,2..$, sonlu sayıdaki eylem çeşidi arasından seçtiği bir a_n eylemini yapıyor olsun. Her bir a_n eylemi yapıldığında sistem bunun çevre ile etkileşimi sonucunda çevreden bir R^e_n pekiştirimi alsın. Dışarıdan alınan pekiştirim miktarı çevre tarafından, a_n eylemini sağlayan sonduruma bağlı olarak belirlenen bir rassal değişkendir. Genelde pekiştirim, pozitif (ödül) veya negatif (ceza) olabilmektedir. Ancak biz bu kısımda sadece ödüllendirici pekiştirimi gözönüne alacağız. Böyle bir sistemdeki pekiştirimli öğrenmenin amacı dıştan gelen pekiştirimi en çok yaparken bu bu pekiştirim için gerekli gideri (cost) en az yapan eylemi bulmaktır.

$$0 \le p^{+}(i,j) \le 1 \text{ ve } 0 \le p^{-}(i,j) \le 1 \qquad (4.2)$$

bağıntıları geçerlidir.

Nöronlara pozitif ve negatif olmak üzere iki dış kaynaktan gelen vurumlar $\Lambda(i)$ *ve* $\lambda(i)$ hızındaki Poisson süreçlerle temsil edilmektedir.

Nöronlar arası iletişim olasılıkları, YSA literatüründeki nöronlar arası bağlantı kuvvetleri cinsinden ifade edilmek istendiğinde

$$w^{+}(i,j) = r(i)p^{+}(i,j) \ge 0, \quad w^{-}(i,j) = r(i)p^{-}(i,j) \ge 0, \qquad (4.3)$$

olarak yazılabilir, burada

$$r(i) = \sum_{j} (w^{+}(i,j) + w^{-}(j,i)) + d(i), \qquad (4.4)$$

bağıntısı ile belirlenmiştir.

4.2 RASSAL SİNİR AĞI İÇİN PEKİŞTİRİMLİ ÖĞRENME

Bu bölümde rassal sinir ağı için önerdiğimiz öğrenme stratejisi önce tek karar adımı içeren sistemler için açıklanmakta ve sonra zincirleme karar adımları içeren daha karmaşık sistemler için genelleştirilmektedir.

4.2.1. Tek Basamaklı Karar Adımları İçin Hemen Pekiştirimli Öğrenme

Bir çevre (environment) ile etkileşim içinde öğrenen bir sistem gözönüne alalım. Bu sistem, her bir n zamanında, $n=0,1,2..$, sonlu sayıdaki eylem çeşidi arasından seçtiği bir a_n eylemini yapıyor olsun. Her bir a_n eylemi yapıldığında sistem bunun çevre ile etkileşimi sonucunda çevreden bir R^{e}_{n} pekiştirimi alsın. Dışarıdan alınan pekiştirim miktarı çevre tarafından, a_n eylemini sağlayan sonduruma bağlı olarak belirlenen bir rassal değişkendir. Genelde pekiştirim, pozitif (ödül) veya negatif (ceza) olabilmektedir. Ancak biz bu kısımda sadece ödüllendirici pekiştirimi gözönüne alacağız. Böyle bir sistemdeki pekiştirimli öğrenmenin amacı dıştan gelen pekiştirimi en çok yaparken bu bu pekiştirim için gerekli gideri (cost) en az yapan eylemi bulmaktır.

$$0 \le p^{+}(i,j) \le 1 \text{ ve } 0 \le p^{-}(i,j) \le 1 \qquad (4.2)$$

bağıntıları geçerlidir.

Nöronlara pozitif ve negatif olmak üzere iki dış kaynaktan gelen vurumlar $\Lambda(i)$ *ve* $\lambda(i)$ hızındaki Poisson süreçlerle temsil edilmektedir.

Nöronlar arası iletişim olasılıkları, YSA literatüründeki nöronlar arası bağlantı kuvvetleri cinsinden ifade edilmek istendiğinde

$$w^{+}(i,j) = r(i)p^{+}(i,j) \ge 0, \quad w^{-}(i,j) = r(i)p^{-}(i,j) \ge 0, \qquad (4.3)$$

olarak yazılabilir, burada

$$r(i) = \sum_{j} (w^{+}(i,j) + w^{-}(j,i)) + d(i), \qquad (4.4)$$

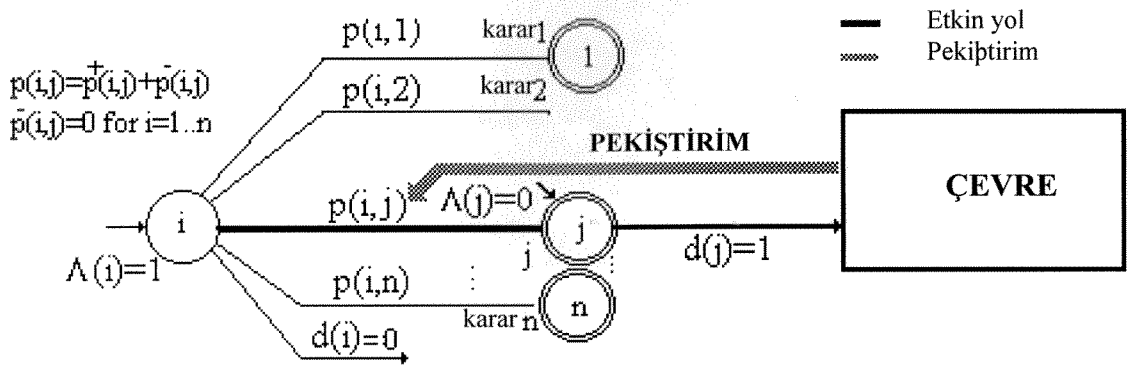
bağıntısı ile belirlenmiştir.

4.2 RASSAL SİNİR AĞI İÇİN PEKİŞTİRİMLİ ÖĞRENME

Bu bölümde rassal sinir ağı için önerdiğimiz öğrenme stratejisi önce tek karar adımı içeren sistemler için açıklanmakta ve sonra zincirleme karar adımları içeren daha karmaşık sistemler için genelleştirilmektedir.

4.2.1. Tek Basamaklı Karar Adımları İçin Hemen Pekiştirimli Öğrenme

Bir çevre (environment) ile etkileşim içinde öğrenen bir sistem gözönüne alalım. Bu sistem, her bir n zamanında, $n=0,1,2..$, sonlu sayıdaki eylem çeşidi arasından seçtiği bir a_n eylemini yapıyor olsun. Her bir a_n eylemi yapıldığında sistem bunun çevre ile etkileşimi sonucunda çevreden bir R^{e}_{n} pekiştirimi alsın. Dışarıdan alınan pekiştirim miktarı çevre tarafından, a_n eylemini sağlayan sonduruma bağlı olarak belirlenen bir rassal değişkendir. Genelde pekiştirim, pozitif (ödül) veya negatif (ceza) olabilmektedir. Ancak biz bu kısımda sadece ödüllendirici pekiştirimi gözönüne alacağız. Böyle bir sistemdeki pekiştirimli öğrenmenin amacı dıştan gelen pekiştirimi en çok yaparken bu bu pekiştirim için gerekli gideri (cost) en az yapan eylemi bulmaktır.



Şekil 4.1: Tek karar adımlı sistem

Aşağıda, bu tip bir öğrenmeyi sağlamak üzere RSA modelinin kullanıldığı bir öğrenme stratejisi önerilmiştir. Şekil 4.1'de *i* ile işaretlenen bir nöron, karardan hemen önceki durumu göstermek üzere kullanılmıştır. Bu nörondan her biri değişik bir kararı temsil eden N çıkış bağlantısı bulunmaktadır. *i* ile işaretlenen nöronu başlangıç nöronu ve her bir karardan sonra varılan *j=1..N* nöronları varış nöronları olarak adlandırılmıştır. Ağ üzerinde sadece başlangıç nöronuna dışsal giriş $\Lambda(i)=\Lambda$, ile beslenmiştir, diğer tüm nöronlar için $\Lambda(j)$ sıfırdır. Ayrıca tüm nöronlar için $\lambda(k)$ sıfırdır. Böylece, tüm sinyal vurumları başlangıç nöronunda yaratılmaktadır. *j* ile gösterilen varış nöronları için yitim parametresi $d(j)=1$ olarak ayarlanmış, başlangıç nöronu *i* içinse, $d(i)=0$ olarak seçilmiştir. Dolayısıyla, başlangıç nöronunda üretilen herhangi bir vurum varış nöronlarından herhangi birine ulaşarak orada yok olur. Vurumun hangi nörona ulaşacağı $p(i,j)=p^+(i,j)$ olasılığı ile belirlenmiştir. Bunun sonucu olarak büyük bağlantı değeri $w^+(i,j)=r(i)p^+(i,j)$'ye sahip olan varış nöronu *j* bu vurumu almak bakımından daha büyük şansa sahiptir. Vurum bir varış nöronuna ulaştığında, burada $d(i,j)=1$ olduğundan vurum yitime uğrar, bu arada çevreyi uyararak çevreden dışsal R^e_n pekiştiriminin gelmesini sağlar. Bir vurumun başlangıç nöronunda yaratıldıktan, varış nöronunda yitinceye kadar olan tüm hareketlerini bir *deneyim (trial)* olarak adlandırmaktayız. Bu çalışmada nöral ağın bir deneyim sırasında seçilen baglantıyı *kısa süreli bellekte (short term memory)* hatırlayabildiği varsayılmakta ve seçilen bağlantıya *etkin bağlantı* denilmektedir. Eğitim sırasında sadece etkin yol üzerindeki nöronlar eğitildiğinden etkin yolun hatırlanması önem taşımaktadır. Bir sonraki adımda bağlantı değerlerinin değiştirilmesi bilginin *uzun süreli bellekte (long term memory)* saklanması anlamını taşımaktadır.

Başlangıçta kararları temsil eden varış nöronlarına ulaşmayı sağlayan tüm bağlantılara eşit olasılık verilmekte ve ateşleme hızları $r(j)=1$ yapılmaktadır, ancak öğrenme ilerledikçe olasılıklar aşağıda çıkardığımız bağlantı kuvveti değiştirme kuralına göre değiştirilmektedir.

Deneyim n'de seçilen a_n eylemi k nöronuna ulaşmayı sağlayan eylem olsun ve bu eyleme verilen dışsal pekiştirim R^e_n olsun. En basit haliyle, n deneyiminde a_n eylemi için i nöronu üzerindeki içsel pekiştirim $R_n(i)=\varphi(R^e_n,L_n)$ olarak formüle edilebilir, burada L_n etkin karar için olan gider, φ *ise* L_n artıkça azalan rassal değerli bir fonksiyondur. Sadece ödüllendilmenin gözönüne aldındığı sistemler için pekiştirim fonksiyonu genellemeden birşey kaybetmeksizin $0\leq R_n(i)\leq 1$ olarak yazılabilir. Deneyim n sırasında seçilen bağlantıyı $R_n(i)$ oranında daha arzulanır yapmak için bağlantı kuvveti $w^+(i,k)$'yı $\Delta w_n(i,k)=\eta\, R_n(i)$ miktarı kadar artırılabilir, burada η öğrenme hızını temsil etmektedir. Ancak böyle bir bağlantı kuvveti değiştirme kuralı bağlantı kuvvetlerinin ve ateşleme hızının sınırsız bir biçimde sürekli artmasına sebep olur. Bu problemin üstesinden gelebilmek için, tüm bağlantı kuvvetleri $(\Sigma_m\, w_n^+(i,m)+\eta R_n(i))$ miktarına bölünmek suretiyle *düzgenlendiğinde (normalization)* ateşleme hızı $r(i)=1$ olacak şekilde sabitleştirilebilir. Bunun sonucunda aşağıdaki bağlantı kuvveti değiştirme kuralı elde edilmektedir.

$$w^+_{n+1}(i,j)=\begin{cases}(w^+_n(i,j)+\eta R_n(i))/(\Sigma_m\, w_n^+(i,m)+\eta R_n(i)) & j=k \text{ için}\\ w^+_n(i,j)/(\Sigma_m\, w_n^+(i,m)+\eta R_n(i)) & j\neq k \text{ için}\end{cases} \qquad (4.5)$$

Yukarıdaki bağlantı kuvveti değiştirme kuralının

$$\Sigma_m\, w_{n+1}^+(i,m)=r_{n+1}(i)=1 \qquad n=1,2... \qquad (4.6)$$

şartını sağladığı kolaylıkla gösterilebilir

Bu durumda $w_n^+(i,j)=r_n(i)p_n^+(i,j)=p_n^+(i,j)$ olur ve bağlantı kuvveti değiştirme kuralı bağlantı olasılıkarı cinsinden aşağıdaki gibi yazılabilir.

$$p^+_{n+1}(i,j)=\begin{cases}(p^+_n(i,j)+\eta R_n(i))/(1+\eta R_n(i)) & j=k \text{ için}\\ p^+_n(i,j)/(1+\eta R_n(i)) & j\neq k \text{ için}\end{cases} \qquad (4.7)$$

Biraz daha basitleştirme yapıldığında

$$p^{+}_{n+1}(i,j) = \begin{cases} p^{+}_{n}(i,j) + \eta R_n(i)(1 - p^{+}_{n}(i,j)) & j=k \text{ için} \\ p^{+}_{n}(i,j) + \eta R_n(i)(-p^{+}_{n}(i,j)) & j \neq k \text{ için} \end{cases} \quad (4.8)$$

elde edilir. Bu kurallın kullanıldığı ödülle öğrenme stratejisi aşağıda özetlenmiştir:

Ödülle Öğrenme

Başlangıç:

$j=1..N_i$ için dışsal ödül dağılımı belirle

$p^{+}_{0}(i,j)=1/N_i$ for $j=1..N_i$, $n=1$

Öğrenme:

1 n'inci deneme için bağlantı olasılıklarını göz önüne alarak bağlantılardan birini seç,

2 seçilen (i,k) bağlantısı için $R_n(k)= R_n$ ödülüne göre öğrenme kuralını uygulayarak bağlantı olasılığı $p^{+}_{n}(i,k)$ değerini değiştir

3 öğrenmeyi bağlantı olasılıkları kararlı bir değere ulaşana kadar $n+1$ için tekrarla

En az giderle en fazla pekiştirim almayı sağlayan karara bir yakınsama sağlanabilmesi için yukarıdaki formülde öğrenme hızı η mümkün olabildiğince küçük seçilmelidir. Diğer yandan η değeri yakınsama sağlamak için gerekli adım sayısını etkilemekte, η küçüldükçe öğrenme yavaşlamaktadır.

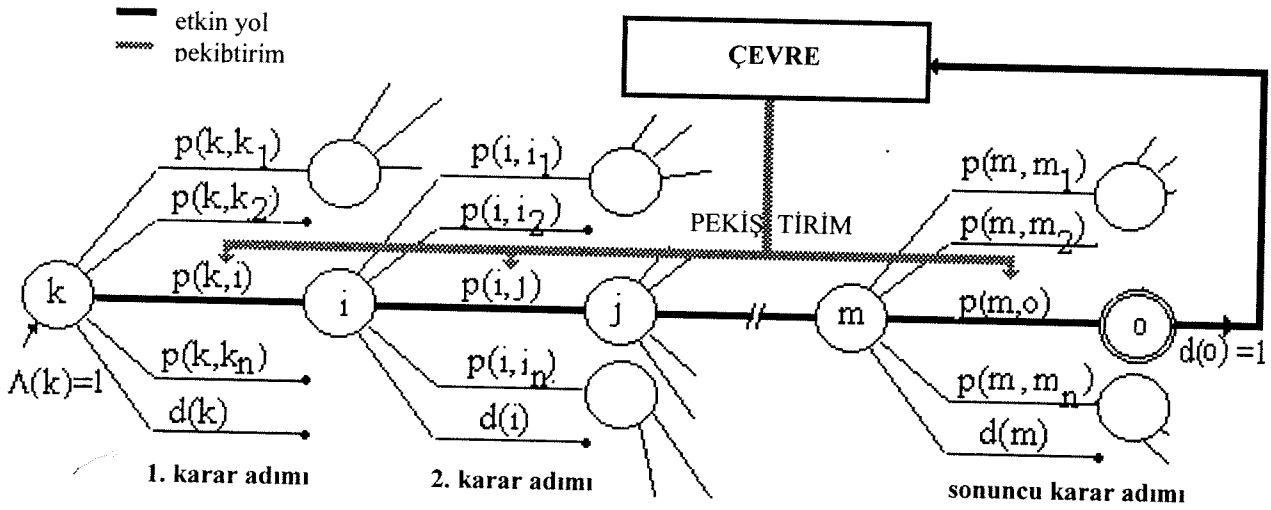
4.2.2. Zincirleme Karar Adımları İçeren Geciktirilmiş Pekiştirimle Öğrenme

Kararların tek adımda alındığı sistemlerde, pekiştirim eylemden hemen sonra elde edildiğinden dolayı en iyi başarı sağlayan eylemin bulunması fazla zor değildir. Ancak birkaç zincirleme karar adımının sonunda ve bu eylemlerin tümünün bir bileşkesi olarak pekiştirimin elde edildiği sistemlerde en iyi adımlar zincirinin bulunması oldukça karmaşık bir problemdir.

Sonlu sayıda durum, X, içeren ayrık zamanlı bir çevre ile etkileşim içinde öğrenen bir sistemi gözönüne alalım. $t=0,1,2$ zaman adımlarında çevre $x(t) \in X$ durumunda olsun. t zaman adımında, $x(t)$ durumu için muhtemel seçenekler gözlendikten sonra sistem eylemlerden birini, $a(t)$, gerçekleştirir. $a(t)$ eylemi çevre ile etkileşim yaratarak, *belirlenmiş (determistic)* ve geçmişten bağımsız bir şekilde $x(t)=x_i$ durumundan yeni bir $x(t+1)=x_j$ durumuna geçilmesini sağlar. Bir çok $a(t)$ eylemi yapıldıktan sonra eğer bir *varış durumuna (final state)* ulaşıldıysa

öğrenen sistem çevreden R^e_n pekiştirimini alır. R^e_n pekiştirimin miktarı, son durum x_{fn} ve tüm eylemlerin $\mathbf{a}=<a(t),\ t=0..t_f>$ bileşkesine bağlı rassal bir fonksiyondur. Buradaki pekiştirimli öğrenmenin amacı, öğrenme ilerledikçe n'inci denemede beklenen dışsal pekişrimi en çok yaparken tüm eylemlerin toplam giderini en az yapan eylemler zinciri $\mathbf{a_n}=<a_n(t),\ t_n=1..t_{fn}>$ bulmaktır.

Zincirleme kararlar içeren durumu modellemek üzere kullanılan RSA Şekil 4.2'de gösterilmiştir. Sistemdeki her bir nöron mümkün olan değişik bir duruma karşılık gelmektedir. Nöronlardan dışarıya doğru olan her bir bağlantı değişik bir kararı temsil etmektedir. Nöronlardan biri başlangıç nöronu olarak, diğer bazıları ise varış nöronları olarak gösterilmiştir, bunlar dışındakiler ara nöronlardır. Başlangıç nöronu henüz hiç bir kararın alınmadığı duruma karşılık gelirken, varış nöronları sistemin dışşal pekiştirim ile ödüllendirildiği durumlara karşılık gelmektedir. Başlangıç ve varış nöronları için RSA parametreleri daha önce olduğu gibi ayarlanmakta, ara nöronlar içinse $\Lambda(k)=0,\ \lambda(k)=0,\ d(k)=0$ yapılmaktadır.



Şekil 4.2: Zincirleme karar adımları içeren sistem

Ağ içinde sadece başlangıç nöronuna dışşal giriş verilerek $\Lambda(i)=\Lambda$ yapılmıştır, diğer tüm nöronlar için $\Lambda(j)$ sıfırdır. Böylece tüm vurumlar başlangıç nöronunda yaratılmaktadır, ve bu vurumlar varış durumlarından birine ulaşarak orada yok olmaktadır. Öğrenme sırasında Λ çok küçük bir değere ayarlanarak, herhangi bir zamada ağ içinde tek bir vurumun dolaşıyor olması sağlanmaktadır. Eğer bir önceki vurum yitmeden yeni bir vurum yaratılmış olursa, bir önceki

vurumun etkisini gözönüne alınmayarak ihmal edilmektedir. Böylece, bir vurum yarıtıldıktan sonra henüz yok olmadan önce ağ içerisinde sadece bir nöron için $k(i)=1$ olacak ve karar zincirleri içerisinde nerede olunduğunu göstermektedir. Herhangi bir i nöronunda verilen bir karar sonucu, buna karşı gelen j nöronu vurumu almakta ve $k(i)=0$ olurken $k(j)=1$ olmaktadır. Komşuluk içindeki hangi nöronun seçileceğine geçiş olasılığı $p(i,j)=p^{+}(i,j)$ gözönüne alınarak karar verilmektedir. Bu komşular içinde en büyük kuvvete $w^{+}(i,j)=p^{+}(i,j)$ sahip olan bağlantı seçilme şansı en yüksek olandır. Varış nöronlarından birine ulaşıldığında , $d(i)=1$ olduğundan, vurum burada çevreyi uyararak yitmekte, çevreden buna cevap olarak o deneme sırasında gerçeklenen tüm eylemlere bağlı olarak bir pekiştirim gelmektedir. Tek adımlı karar durumunda olduğu gibi, başlangıç nöronunda yaratılan bir vurumun yitime uğradığı varış nöronlarından birine ulaşıncaya kadar olan hareketi bir deneyime karşılık gelmektedir. Bir deneyim sırasında, sistemin tüm etkin bağlantıları kısa süreli bellekte tutarak hatırladığı düşünülmekte ve bağlantı kuvvetlerinin değiştirilmesi sırasında sadece etkin yol üzerindeki noronlar için değişiklik yapılmaktadır.

Başlangıçta i nöronunda alınabilecek tüm kararlara karşılık gelen bağlantılara eşit olasılık verilmektedir. Karar ağacının tümünü başlangıçtan itibaren bilmeye ihtiyaç yoktur. Nöronlar ilk defa ziyaret edildiğinde, bu nöronlar için olabilecek kararlar görülür hale gelecek ve ilgili bağlantılar kurularak eşit olasılıklar verilecektir. Öğrenme ilerledikçe bağlantı kuvvetleri aşağıdaki formule göre değiştirilmektedir:

$$p^{+}_{n+1}(i,j)=\begin{cases} p^{+}_{n}(i,j)+\eta R_n(i)(1-p^{+}_{n}(i,j)) & \text{eğer } i \text{ etkin yol üzerinde ve } (i,j) \text{ etkinse} \\ p^{+}_{n}(i,j)+\eta R_n(i)(-p^{+}_{n}(i,j)) & \text{eğer } i \text{ etkin yol üzerinde ancak } (i,j) \text{ etkin değilse} \\ \text{değişiklik yok} & \text{diğer durumlarda} \end{cases} \quad (4.9)$$

Deneyim n sırasında, eğer bir varış nöronuna ulaşıldıysa çevreden dışsal R^{e}_{n} pekiştirimi alınmaktadır. Zincirleme karar adımları içeren sistemde, bu karar zicirini en iyilemek üzere kullanılacak içsel pekiştirim fonksiyonu $R_n(i)=\varphi(R^{e}_{n},L_n,l_n(i))$ olarak tanımlanabilir, burada L_n deneyim n sırasında seçilen etkin karar yolunun toplam gideri, $l_n(i)$ toplam giderin i nöronundan varış nöronuna kadar olan kararlarla ilgi bölümü, φ ise L_n veya $l_n(i)$ arttıkça azalan rassal değerli bir fonksiyondur ve $0 \leq R_n(i) \leq 1$ şartını sağlamaktadır.

vurumun etkisini gözönüne alınmayarak ihmal edilmektedir. Böylece, bir vurum yarıtıldıktan sonra henüz yok olmadan önce ağ içerisinde sadece bir nöron için $k(i)=1$ olacak ve karar zincirleri içerisinde nerede olunduğunu göstermektedir. Herhangi bir i nöronunda verilen bir karar sonucu, buna karşı gelen j nöronu vurumu almakta ve $k(i)=0$ olurken $k(j)=1$ olmaktadır. Komşuluk içindeki hangi nöronun seçileceğine geçiş olasılığı $p(i,j)=p^{+}(i,j)$ gözönüne alınarak karar verilmektedir. Bu komşular içinde en büyük kuvvete $w^{+}(i,j)=p^{+}(i,j)$ sahip olan bağlantı seçilme şansı en yüksek olandır. Varış nöronlarından birine ulaşıldığında , $d(i)=1$ olduğundan, vurum burada çevreyi uyararak yitmekte, çevreden buna cevap olarak o deneme sırasında gerçeklenen tüm eylemlere bağlı olarak bir pekiştirim gelmektedir. Tek adımlı karar durumunda olduğu gibi, başlangıç nöronunda yaratılan bir vurumun yitime uğradığı varış nöronlarından birine ulaşıncaya kadar olan hareketi bir deneyime karşılık gelmektedir. Bir deneyim sırasında, sistemin tüm etkin bağlantıları kısa süreli bellekte tutarak hatırladığı düşünülmekte ve bağlantı kuvvetlerinin değiştirilmesi sırasında sadece etkin yol üzerindeki noronlar için değişiklik yapılmaktadır.

Başlangıçta i nöronunda alınabilecek tüm kararlara karşılık gelen bağlantılara eşit olasılık verilmektedir. Karar ağacının tümünü başlangıçtan itibaren bilmeye ihtiyaç yoktur. Nöronlar ilk defa ziyaret edildiğinde, bu nöronlar için olabilecek kararlar görülür hale gelecek ve ilgili bağlantılar kurularak eşit olasılıklar verilecektir. Öğrenme ilerledikçe bağlantı kuvvetleri aşağıdaki formule göre değiştirilmektedir:

$$p^{+}_{n+1}(i,j)=\begin{cases} p^{+}_{n}(i,j)+\eta R_n(i)(1-p^{+}_{n}(i,j)) & \text{eğer } i \text{ etkin yol üzerinde ve } (i,j) \text{ etkinse} \\ p^{+}_{n}(i,j)+\eta R_n(i)(-p^{+}_{n}(i,j)) & \text{eğer } i \text{ etkin yol üzerinde ancak } (i,j) \text{ etkin değilse} \\ \textit{değişiklik yok} & \text{diğer durumlarda} \end{cases} \quad (4.9)$$

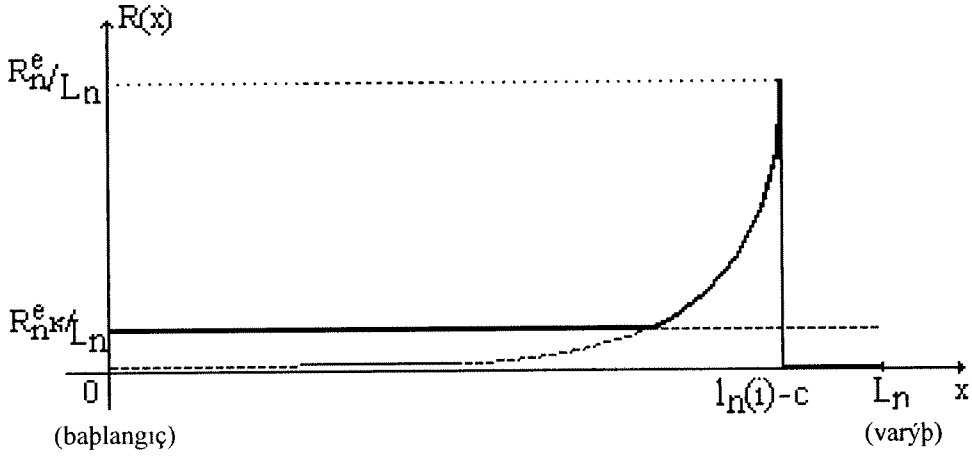
Deneyim n sırasında, eğer bir varış nöronuna ulaşıldıysa çevreden dışsal R^{e}_{n} pekiştirimi alınmaktadır. Zincirleme karar adımları içeren sistemde, bu karar zicirini en iyilemek üzere kullanılacak içsel pekiştirim fonksiyonu $R_n(i)=\varphi(R^{e}_{n},L_n,l_n(i))$ olarak tanımlanabilir, burada L_n deneyim n sırasında seçilen etkin karar yolunun toplam gideri, $l_n(i)$ toplam giderin i nöronundan varış nöronuna kadar olan kararlarla ilgi bölümü, φ ise L_n veya $l_n(i)$ arttıkça azalan rassal değerli bir fonksiyondur ve $0 \leq R_n(i) \leq 1$ şartını sağlamaktadır.

4.2.3. Son_Zaman Etkisi Taşıyan bir Pekiştirim Fonksiyonu

Labirentler üzerine kurduğumuz benzetim deneylerimizde pekiştirim fonksiyonunun son_zaman (recency) etkisini yansıttığında daha iyi sonuç alındığını gördük. Son_zaman etkisi daha önce 2. bölümde açıklanmıştı. Son-zaman etkisini kısmen içermek üzere aşağıdaki gibi bir pekiştirim fonksiyonu $R_n(i)=\varphi(R^e_n, L_n, l_n(i))$ seçilebilir (Şekil 4.3):

$$R_n(i)=\begin{cases} (R^e_n/L_n)(\max(\kappa,\ 1/(l_n(i)-c)) & c < l_n \text{ için} \\ 0 & c \geq l_n(i) \text{için} \end{cases} \qquad (4.10)$$

burada κ, 0 ile 1 arasında değer alan bir sabittir ve son_zaman etkisinin fonksiyona katkısını ayarlamakta kullanılmaktadır.

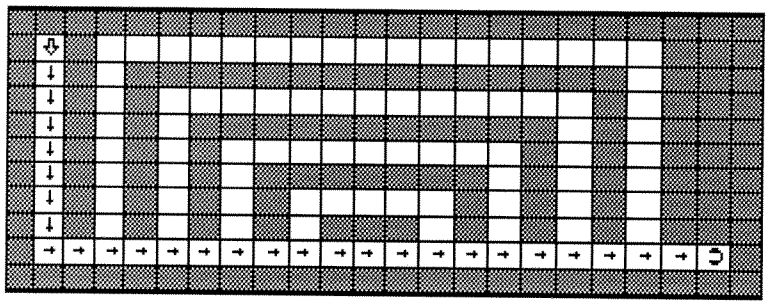


Şekil 4.3. Deneylerde kullanılan pekiştirim fonksiyonu

Pekiştirim fonksiyonunda son_zaman etkisine ek olarak fonksiyonunu sola kaydırmaya yarayan bir c parametresi gözönüne alınmıştır. Bu parametre, etkin yolun son tarafındaki kesinlikle öğrenilen kısımın giderine karşılık gelmektedir. Etkin yolun son tarafında tüm bağlantı kuvvetlerinin tümü 1'e çok yakın değerde olan bir kısım varsa bu kısım için kesin öğrenilmiş kısım diyoruz. Pekiştirim fonksiyonunda eğer κ=1 seçilirse $R_n(i)= R^e_n /L_n$ olmakta ve son_zaman etkisi tamamiyle ihmal edilmektedir; bu durumda etkin yoldaki tüm nöronlar varış nöronuna yakınlığına bakılmaksızın aynı şekilde öğrenmektedirler. Diğer yandan eğer κ=0 seçilirse $R_n(i)=(R^e_n/L_n)(1/(l_n(i)-c))$ olmakta ve son_zaman etkisi baskın bir hale gelmektedir; bu durumda sadece kesinlikle öğrenilmiş kısıma yakın nöronlar öğrenme için pekiştirim almaktadır.

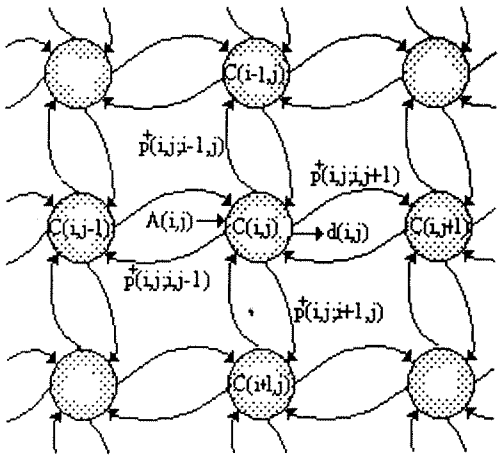
4.2.4. Labirentler İçin Elde Edilen Benzetim Sonuçları

En iyi zincirleme karar adımlarını öğrenmek üzere önerilen pekiştirimli öğrenme algoritması Şekil 4.4'de gösterilen 11x23 büyüklüğündeki labirentte denenmiştir (Madenoglu, 1994). Şekilde karanlık hücreler duvarlara, aydınlık hücreler ise geçitlere karşılık gelmektedir. Burada amaç başlangıç hücresini varış hücresine bağlayan bir yolun bulunması ve denemeler ilerledikçe mümkün olduğunca kısa bir yolun öğrenilmesidir (Halıcı ve Yaranlı, 1992). Başlangıç hücresinden varış hücresine uzanan en kısa yol şekilde işaretlenmiştir.



Şekil 4.4: Bir labirenti temsil eden hücresel dizin ve başlangıç hücresini varışa bağlayan en kısa yol

Labirenti modelleyen RSA'da labirentteki herbir hücreye bir nöron karşılık gelmekte, komşu hücrelerde her iki yöde bağlantı bulunmaktadır (Şekil 4.4)



Şekil 4.5: Labirenti temsil etmek üzere kullanılan RSA nöronları ve aralarındaki bağlantılar

RSA parametreler 4.2.2. kısımda açıklandığı şekilde ayarlanmıştır. RSA çeşitli η ve κ değerleri için herbirinde 100 değişik rassal tohum ile denenmiştir. Herbir tohum için aşağıdaki işlemler tekrarlanmaktadır:

Zincirleme Karar Adımları İçin Öğrenme

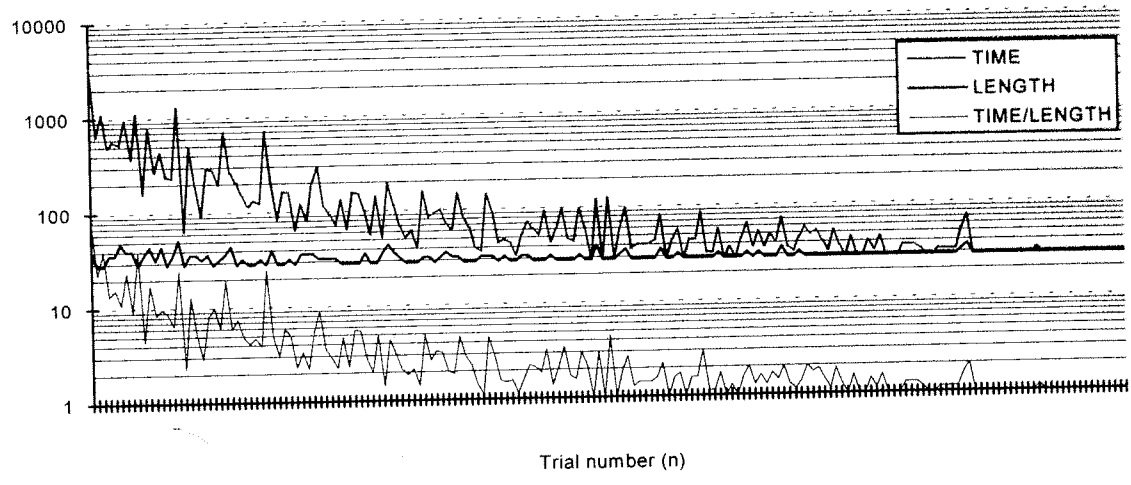
Başlangıç:

Her bir varış nöronu için dışsal ödül dağılımını belirle,
$p^{+}_{0}(i,j)=1/N_i$ for $j=1..N_i$, $n=1$

Öğrenme:

1. Başlangıç nodunda bir vurum yaratılana kadar bekle
2. Vurumun izlediği rassal etkin yolu takip et
3. Etkin yol üzerinde seçilen her bir (i,k) bağlantısı için $R_n(k)= \varphi(R^e_n, L_n, l_n(k))$ pekiştirimine göre bağlantı olasılığı $p^{+}_{n}(i,k)$ değerini değiştir
4. Öğrenmeyi etkin yol üzerindeki tüm bağlantılar kesinlikle öğrenilinceye kadar $n+1$ için tekrarla

Benzetim deneylerimizde tipik öğrenme eğrileri elde edilmiştir. Deneme sayısı ilerledikçe o deneme için harcanan toplam zaman öğrenilen yolun uzunluğuna yaklaşmaya başlaması başlangıç nöronunda yaratılan vurumun varış nöronuna doğru geri dönmeksizin doğrudan ilerlediğini göstermektedir (Şekil 4.6)



Şekil 4.6: Labirent için elde edilen öğrenme eğrisi: Varış nöronuna varıncaya kadar harcanan zaman, aktif yolun uzunluğu, ve öğrenme ilerledikçe harcanan zamanın yol uzunluğuna oranı

Tablo 1'de öğrenilen yolların yol uzunluğuna göre dağılımı çeşitli κ ve η parametreleri için verilmiştir.

TABLO I. Öğrenilen yolların yol uzunlukları üzerindeki dağılımı

		Ortalama	ÖĞRENİLEN YOL UZUNLUĞU DAĞILIMI				
κ	η	Deneyim	28	32	36	40	44
1.000	0.75/28	266	87	13	0	0	0
1.000	1.00/28	204	78	19	2	0	1
1.000	1.25/28	155	72	26	2	0	0
1.000	1.50/28	133	72	24	4	0	0
1.000	1.75/28	117	64	24	11	1	0
1.000	2.00/28	103	69	22	8	1	0
1.000	2.25/28	90	66	29	4	0	1
1.000	2.50/28	86	56	27	15	2	0
1.000	2.75/28	74	58	25	13	4	0
0.000	6.00/28	283	100	0	0	0	0
0.000	8.00/28	211	100	0	0	0	0
0.000	10.00/28	161	98	2	0	0	0
0.000	12.00/28	134	98	2	0	0	0
0.000	14.00/28	113	100	0	0	0	0
0.000	16.00/28	96	98	2	0	0	0
0.000	18.00/28	86	98	2	0	0	0
0.000	20.00/28	75	96	4	0	0	0
0.125	4.00/28	304	97	3	0	0	0
0.200	4.00/28	218	92	8	0	0	0
0.400	4.00/28	121	72	24	3	1	0
0.500	4.00/28	98	63	27	9	1	0
0.600	4.00/28	86	63	22	15	0	0
0.800	4.00/28	67	53	3	11	5	1

Bu tabloda ilk ve ikinci kolonlar seçilen κ ve η değerlerini, üçüncü kolon ise yolu kesinlikle öğrenmek için gerekli ortalama deneme sayısını göstermektedir. Diğer kolonlar 100 değişik tohumla tekrarlanan deneylerden kaçında kolon başlığında belirtilen uzunluktaki yolun öğrenildiğini göstermektedir.

Tablodan görüldüğü gibi, κ parametresinin değeri öğrenme üzrinde çok önemli bir etki yapmaktadır. Tablonun yukarı bölümü son_zaman etkisinin tamamiyle gözardı edildiği bir pekiştirim fonksiyonuna karşılık gelirken ikinci kısım son_zaman etkisinin baskın olduğu pekiştirim fonksiyonuna karşılık gelmektedir. Tablodaki son kısım ise bu iki uç noktanın arasındaki durumları gözönüne almaktadır. Bu tablodan, en kısa yolun son_zaman etkisinin baskın olduğu ve öğrenme katsayısının düşük olduğu durumda en iyi öğrenildiği ortaya çıkmaktadır.

4.3. DEĞİŞEN ÇEVRE KOŞULLARINA UYUM

Değişen şartlara uyumu sağladığından yaşamın sürdürülmesi açısından büyük önem taşıyan öğrenmenin sönümü daha önce 2. bölümde açıklanmış ve 3. bölümde ise koşullu öğrenmede sönümün modellenebilmesi için READ devresi üzerinde değişiklikler yapılarak incelenmişti. Bu kısımda pekiştirimli öğrenmenin sönümünü sağlamak üzere yapılan çalışmalar sunulacaktır.

RSA için 4.2. kısımda önerilen pekiştirimli öğrenmeyi kullanan sistem durağan ortamlarda iyi başarı göstermesine rağmen, sistemin sadece ödülle öğrenmesi çevrede keşif yapmasını engellemektedir. Eğer ortam durağan değilse sistem daha önce öğrenilen eylemler sönüme uğramamakta ve böylece sistemin değişen durumlara uyum sağlayamaması sistemin değişken değişken ortamlarda başarısız kalmasına neden olmaktadır. Çalışmamızın bu kısımında RSA özelliklerini bozmayacak şekilde ceza ile bağlantı kuvveti değiştirme kuralları önerilerek daha önce önerilen pekiştirimle öğrenme stratejisi ceza ile öğrenmeye yönelik olarak genelleştirilmektedir. Daha sonra pekiştirim beklentisinin gözönüne alındığı bir öğrenme stratejisi önerilmektedir. Bu stratejiye göre gelen ödül beklenenden az olmadıkça ödülle öğrenme kuralları uygulanmakta, ustalık ön plana çıkmaktadır. Diğer durumda ise ceza kurallarına göre öğrenen sistem diğer olasılıkların keşfedilmesi yoluna gitmektedir.

4.3.1. Ödüle/Ceza ile Pekiştirim

Şekil 4.1'de verilen tek karar adımlı sistemin ödülle öğrenmesi için gerekli bağlantı kuvveti değiştirme kuralı daha önce bölüm 4.2.1 de verilmişti. Aşağıda bu kuralı gösterimdeki bir kaç küçük değişiklikle tekrar veriyoruz:

$$p^+_{n+1}(i,j)= \begin{cases} p^+_n(i,k)+\eta^+ R^+_n(k)(1-p^+_n(i,k)) & j=k \\ p^+_n(i,j)-\eta^+ R^+_n(k)(p^+_n(i,j)) & j\neq k,\ j=1..N_i \end{cases} \qquad (4.12)$$

burada n deneme numarası, $R^+_n(k)$ bu denemede k nöronu için elde edilen ödül, η^+ ödülle öğrenme hızıdır.

Aşağıda ceza ile öğrenmeye yönelik olarak önerdiğimiz bağlantı kuvveti değiştirme kuralı verilmiştir:

$$p^{+}_{n+1}(i,j)= \begin{cases} p^{+}_{n}(i,k)-\eta^{-}R^{-}_{n}(k)p^{+}_{n}(i,k)) & j=k \text{ için} \\ p^{+}_{n}(i,j)+ \eta^{-}R^{-}_{n}(k)(p^{+}_{n}(i,k)/(1- p^{+}_{n}(i,k))\, p^{+}_{n}(i,j) & j\neq k, j=1..N_i \text{ için} \end{cases} \quad (4.12)$$

burada $R^{-}_{n}(k)$ bu denemede k nöronu için verilen ceza, η^{-} cezayla öğrenme hızıdır. Ceza ile öğrenme stratejisi aşağıdadır:

Ceza ile Öğrenme

Başlangıç:
$j=1..N_i$ için dışsal ceza dağılımı belirle
$p^{+}_{0}(i,j)=1/N_i$ for $j=1..N_i$, $n=1$
Öğrenme:
1 n'inci deneme için bağlantı olasılıklarını gözönünde alarak bağlantılardan birini seç,
2 seçilen (i,k) bağlantısı için $R^{-}_{n}(k)= R^{-}_{n}$ cezasına göre öğrenme kuralını uygulayarak bağlantı olasılığı $p^{+}_{n}(i,k)$ değerini değiştir
3 öğrenmeyi bağlantı olasılıkları kararlı bir değere ulaşana kadar $n+1$ için tekrarla

Yukarıdaki (4.11) ve (4.12) bağıntılarında genellemeden bir şey kaybetmeksizin pekiştirim fonksiyonun $0\leq R^{+}_{n}(i),\ R^{-}_{n}(i) \leq 1$ olduğu varsayılabilir. Ayrıca η^{+} ve η^{-} öğrenme katsayalılarıdır ve iyi bir yakınsama için mümkün olduğunca küçük seçilmeleri gerekmektedir. Burada gözlenmesi gereken önemli bir nokta $0\leq R^{+}_{n}(i),\ R^{-}_{n}(i) \leq 1$ *ve* $0\leq \eta^{+}, \eta^{-} \leq 1$ olduğunda eğer RSA özellikleri başlangıçta ağ için korunuyorsa, ödül veya ceza ile bağlantı kuvvetleri değiştirildikten sonra da bu özelliklerin korunmaya devam ettiğidir. Buradaki ödülle öğrenme kuralı, Bush and Mosteliert (1958) tarafından önerilen ve (Narendra and Thathachar, 1990) tarafından Öğrenen Otomata sisteminde kullanılan bağlantı değiştirme kuralına benzemektedir, ancak ceza ile öğrenme kuralı oldukça farklıdır. Öğrenen Otomata sisteminde kullanılan öğrenme kuralını, RSA için uyarladığımızda aşağıdaki öğrenme kuralı elde edilmektedir:

$$p^{+}_{n+1}(i,j)= \begin{cases} p^{+}_{n}(i,k)-\eta^{-}R^{-}_{n}(k)p^{+}_{n}(i,k)) & j=k \\ p^{+}_{n}(i,j)+ \eta^{-}R^{-}_{n}(k)/(N_i-1_i) - \eta^{-}R^{-}_{n}(i)\, p^{+}_{n}(i,j) & j\neq k,\ j=1..N_i \end{cases} \quad (4.13)$$

yukarıda verilen öğrenme kuralında gözlenmesi gereken önemli bir nokta, bu kurala göre bağlantı kuvvetleri değiştirildiğinde, ceza görmek üzere seçilmiş olmamalarına rağmen bazı bağlantı kuvvetlerinde azalmaya sebep olabilmesidir. Böyle bir anormallik, öğrenen otomata ceza kuralına göre eğitilen sistemin bizim önerdiğimiz kurala göre eğiltildiğinden daha kötü bir başarı sergilemesine neden olmaktadır.

4.3.2. Beklentinin Önemi

Daha önce bahsettiğimiz gibi, ödülle öğrenen sistem durağan ortamlarda çok başarılı olmasına rağmen, çevre şartları değiştiğinde yeni şartlara uyum gösteremiyordu. Çevre şartları değiştiğinde ortaya çıkan bu problemin üstesinden gelmek üzere ödül beklentisini gözönüne alan bir öğrenme stratejisi aşağıda önerilmiştir. Bu stratejide bağlantı kuvvetleri gelen ödül beklenenden kötü olduğunda ceza ile öğrenme kurallarına göre, diğer durumlarda ise ödülle öğrenme kurallarına göre değiştirilmektedir.

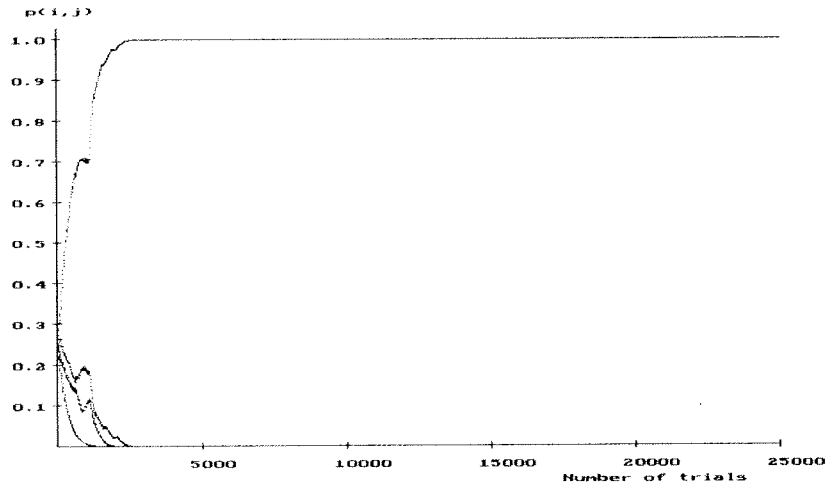
Ödül Beklentisiyle Öğrenme

Başlangıç:
j=1..N_i için ödüllerin dışsal ödül dağılımı belirle
$p^+_0(i,j)=1/N_i$ for j=1..Ni, $\underline{R}^+_0(j)=0$
Öğrenme:
1 n'inci deneme için bağlantı kuvvetlerini gözönüne alarak bağlantılardan birini seç,
2 seçilen (i,k) bağlantısı için eğer $R^+_n > \underline{R}^+_0(j)$ *ise* ödülle öğrenme kuralını uygulayarak $R^+_n(k)= R^+_n$ ödülüne göre bağlantı olasılığı $p^+_n(i,k)$ değerini değiştir, diğer durumda cezayla öğrenme kuralını uygulayarak $R^-_n(k)= \underline{R}^+_n(k)-R^+_n$ cezasına göre bağlantı olasılığı değerini değiştir, burada $\underline{R}^+_n(k)$ *ödül beklentisidir.*
3 ödül beklentisini aşağıdaki bağıntıya göre değiştir:
$\underline{R}^+_n(k)=(1-\beta)\underline{R}^+_n(k)+\beta R^+_n \quad 0<\beta<<1 \quad (\beta=0.01)$
4. öğrenmeyi n+1 için tekrarla

4.3.3. Benzetim Sonuçları

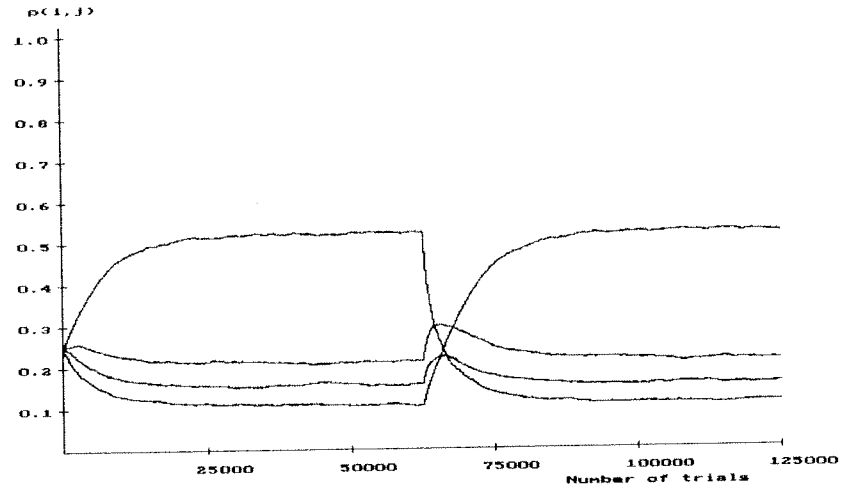
Ödül ile öğrenmenin değişen çevre şartlarına uyumunu incelemek üzere yapılan benzetim deneylerinde elde edilen sonuçlar Şekil 4.7'de verilmiştir. Başlangıçta 1-4 varış nöronlarına 0.1, 0.4, 0.6, 0.9 ortalama değerine sahip ödül uygulanmış, eğitimin sönüm fazında, 1 ve 4 varış nöronlarına uygulanan ortalama değerler yer değiştirilerek öğrenmenin sönümüne ihtiyaç duyulan çevre şartları değişikliği yaratılmaya çalışılmıştır. Ödülle eğitilen RSA'da en fazla ödül alan eylem için bağlantı olasılığı öğrenme fazının sonunda sonunda 1 değerine yaklaşmaktadır, dolayısı ile belirli bir aşamadan sonra sürekli olarak en fazla ödül verilen eylem seçilerek ustalık sergilenmektedir. Ancak şekilde görüldüğü gibi, sönüm fazında çevre

şartları değişmesine ve seçilen eyleme gelen ödül miktarı azalmasına rağmen öğrenilen eylem sönüme uğramamakta ve yapılmaya devam edilmektedir.

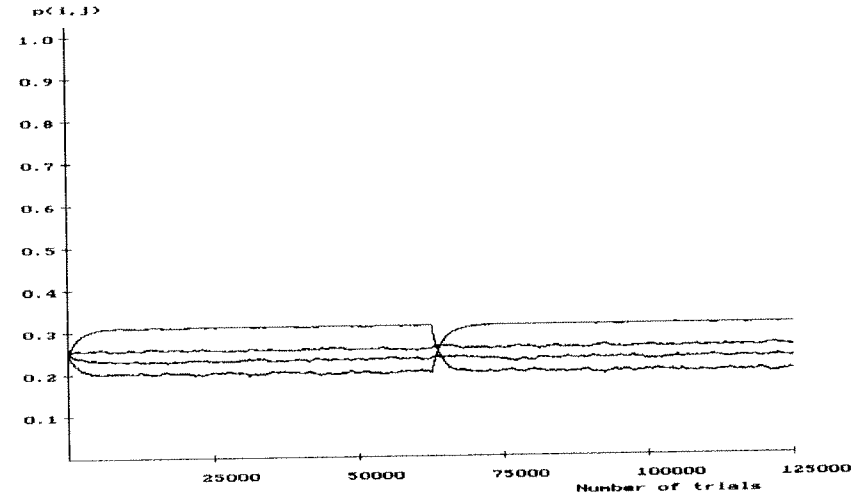


Şekil 4.7: Ödül ile öğrenme stratejisinde öğrenme ve sönüm fazları

Ceza ile öğrenmenin benzetim deneylerinde elde edilen sonuçlar Şekil 4.8'de verilmiştir. Başlangıçta 1-4 varış nöronlarına 0.1, 0.4, 0.6, 0.9 ortalama değerine sahip ceza uygulanmış, eğitimin ikinci fazında, 1 ve 4 varış nöronlarına uygulanan ortalama değerler yer değiştirilmştir. Şekil 8.a da görüldüğü gibi, RSA'ya ceza ile öğrenme stratejisi uygulandığında, öğrenme fazında eylemlerin aldıkları cezalar öğrenilmekte ve en az ceza alan eylem en çok, en fazla ceza alan eylem ise en az yapılır hale gelmektedir. Ancak, ceza ile öğrenme kuralına göre eğitilen sistemde en az ceza gören eylem eğitim sonunda en yüksek bağlantı olasılığına sahip olmasına rağmen bu bağlantı olasılığı 1 yerine daha düşük bir değere yakınsamaktadır. Bunun bir sonucu olarak, ceza ile öğrenmede en az ceza gören eylem diğerlerinde daha fazla sergilenmekte, ancak zaman içerisinde diğer eylemlerinde denenmesine devam edilmektedir. Diğer yandan, sönüm fazında en fazla seçilen seçilen eyleme gelen ceza miktarı arttığında en çok seçilen eylem sönüme uğramakta ve bu durumda en az ceza alan yeni eylem öğrenilerek daha fazla yapılmaktadır.



a)

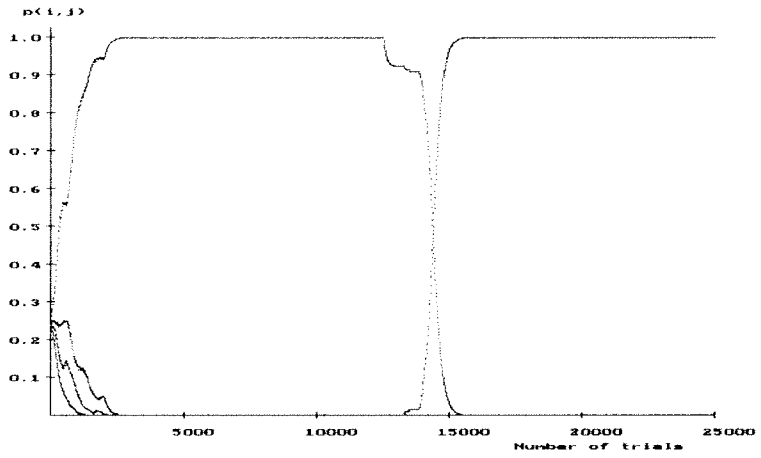


b)

Şekil 4.8. Ceza ile öğrenme stratejisinde öğrenme ve sönüm fazları
a) RSA için burada önerilen öğrenme kuralı b) Öğrenen otomatadaki öğrenme kuralı

Şekil 4.8.b'de öğrenen otomata için önerilen cezayla öğrenme kuralına göre RSA eğitildiğinde elde edilen sonuçlar verilmiştir. Burada önerilen ceza kuralına göre eğitilen RSA (Şekil 4.8.a) ile öğrenen otomata ceza kuralına göre eğitilen RSA (Şekil 4.8.b) karşılaştırıldığında burada önerilen kurala göre eğitilen sistemin başarısının daha yüksek olduğu görülmektedir. Şekil 4.8.a'da en az ceza gören eylemin seçilme olasılığı 4.8.b'dekinden daha yüksektir. Dolayısı ile her iki sistemde yapılan eylemler sonucunda alınan toplam cezanın beklenen değeri karşılaştırıldığında burada önerilen kurala göre eğitilen sistemin beklenen toplam ceza değeri diğerinden daha düşük olmaktadır.

Şekil 4.9'da ödül beklentisi kullanılarak eğitilen RSA için elde edilen sonuç görülmektedir. Ödül ile öğrenme ve ceza ile öğreme stratejileri değişen cevre şartların uyum gösterme açısından incelendiğinde, ödülle öğrenmede değişen çevre şartlarından haberdar olunmadığı, ceza ile öğrenmede işe değişen şartlara göre bağlantı kuvvetlerinin uyum gösterdiği gözlemiştir. Sistemin normal şartlarda en eyi eyleme tam bir yakınsama ve şartlar değiştiğinde yeni şartlara uyum göstermesini sağlamak amacıyla önerilen ödül beklentisiyle öğrenme stratejisinin sonuçları arzu edilen başarıyı göstermiştir.



Şekil 4.9: Beklenti ile öğrenme stratejisinde öğrenme ve sönüm fazları

5. SONUÇLAR VE DEĞERLENDİRME

Bilişsel süreçlerin Yapay Sinir Ağları ile modellenmesini amaçlayan bu projede canlılardaki Çağrışımlı Öğrenmenin iki temel çeşidi olan Koşullu Öğrenme ve Pekiştirimli Öğrenme süreçleri incelenmiştir. Çalışmalar ODTÜ Elektrik ve Elektronik Mühendisliği Bölümü'nde, ODTÜ Psikoloji Bölümü'nün danışmanlığı altında yürütülmüştür Proje önerisinde sadece koşullu öğrenme için yapılan çalışmalar kapsanmıştı. Öneride belirtilenlerden "sönüme uğrayan tepkilerin tekrar öğrenilmesi" dışındaki tümü proje çalışmaları sırasında gözönüne alınmıştır. Projenin araştırma safhasının sonunda, unutulan davranışların tekrar öğrenilmesinin proje açısından fazla bir şey kazandırmayacağı gözlenmiş, bunun yerine Pekiştirimli Öğrenme konusu üzerinde çalışmanın, çağrışımlı öğrenmedeki ikinci temel yaklaşım olması dolayısıyla çok daha faydalı olacağı kanısına varılmıştır. Dolayısıyla pekiştirimli öğrenme ile ilgili konular da proje kapsamına alınarak üzerinde çalışılmıştır.

Koşullu öğrenme ile ilgili çalışmalarda, Koşullu öğrenmeyi modellemek üzere kullanılan Geri Döngülü Çağrışımsal Dipol (READ) devresi üzerinde öğrenmenin sönümünü modellemek üzere değişiklikler yapılmıştır. Öğrenmenin sönümü, daha önceden öğrenilen ancak artık faydası olmayan tepkilerin bırakılmasını sağlaması sebebiyle, canlıların değişen çevre koşullarına uyum sağlamasında çok önemli bir özelliktir. Sönümün modellendiği READ devresinin birincil ve ikincil koşullandırma altında çalışması incelenmiş ve her iki durumda da devrenin başarı ile çalıştığı gözlenmiştir. Birincil koşullandırma ile ilgili benzetimlerde, sistemin öğrenme fazında başarıyla öğrendiği ve sönüm fazında orjinal READ devresinde görülmeyen sönümün değiştirilmiş devrede sağlandığı görülmüştür. İkincil koşullandırma ile ilgili yapılan benzetimlerde, orjinal READ devresinde geçerliği kalmayan birincil koşullamaların ikicil koşullamaya yolaçtığı gözlenirken, değiştirilen READ devresinde ikincil koşulllanmanın önlenebilmesi sağlanmıştır. Koşullu öğrenme ile ilgili çalışmalarda birden fazla READ birimin birarada kullanılabileceği bir yapı önerilmiş, ayrıca bu yapı içerisinde birbiriyle çelişen davranışların aynı anda gösterilmesinin önlenebildiği gözlenmiştir.

Pekiştirimli öğrenmenin modellenmesi ile ilgili çalışmalarda Rassal Sinir Ağları (RSA) kullanılmıştır. İlk önce tekli karar adımları içeren sistemler için bir öğrenme stratejisi önerilmiş ve bu strateji bir amaca ulaşmak için zincirleme karar adımlarından geçilmesini

gerektiren daha karmaşık sistemler için genelleştirilmiştir. Son_zaman etkisini gözönüne alan bir pekiştirim fonksiyonu önerilerek öğrenme stratejisi labirent öğrenme için denenmiştir. Labirent öğrenme için elde edilen sonuçlar öğrenmenin varlığının gösterilmesi açısından tatmin edicidir. Böyle bir öğrenme stratejisinde en kısa yolun öğrenileceği garanti edilememektedir, ancak benzetim sonuçlarından son_zaman etkisi gözönüne alındığında ve öğrenme hızı küçük tutulduğunda kısa yolların tercih edildiği gözlenmiştir. Öğrenme hızı artırıldıkça yolların öğrenilmesi için gerekli yakınsama zamanı düşmektedir, ancak hızlı bir öğrenme, en kısa yol dışındaki yolların da öğrenilmesine sebep olabilmektedir. Sadece ödül gözönüne alınarak önerilen pekiştirimli öğrenme stratejisi durağan çevrelerde başarıyla çalışmasına karşın çevrenin durağan olmadığı durumlarda sistem daha önce öğrediği eyleme takılmakta ve öğrenmede sönüm mümkün olmamaktadır. Çalışmanın bir sonraki kısmında ödül ile öğrenme için daha önce önerilen öğrenme stratejisi ceza için genelleştirilmiştir. Bu çalışmada önerilen ceza kuralıyla eğitilen sistemin başarısı Öğrenen Otomata kuralıyla eğitilen sisteminkiyle karşılaştırılmış ve daha başarılı olduğu gözlenmiştir. Çalışmanın bu kısımında ayrıca pekiştirim beklentisinin gözönüne alındığı bir öğrenme stratejisi önerilmiştir. Bu strateji ile sisteme, gelen ödülün beklenenden az olduğu durumlarda diğer olanakların araştırılması imkanı tanınmış olmaktadır. Böyle bir strateji öğrenim sönümüne olanak tanırken bir çok durumda en çok ödül getiren davranışa tam yakınsama sağlayabilmektedir.

Projede önerilen ve kullanılan modeller, canlılarda gözlenen bazı davranışsal ve nörofizyolojik bilgilerden esinlenilerek ve psikoloji deneylerinde elde edilen verilere uygun düşecek biçimde tasarlanmışlardır. Bu modellerin gerçek nöral ağların nörobiyoljik yapılarını açıklama iddiası yoktur; modellemeye çalışılan bu sistemlerin işlemsel yönüdür ve doğal öğrenmeyle ilgili bazı bilgilerin akıllı sistemlerin geliştirilmesinde kullanılmasını amaçlamaktadır. Bu projedeki çalışmalar ilk aşamada teorik nitelikte olmakla birlikte pratik uygulamalar için potansiyel taşımaktadır. Bu projede önerilen öğrenme biçimlerinin otomatik kontrol, navigasyon, robotik gibi konularda kullanılabileceği bir çok gerçek uygulama alanı mevcuttur. Klasik Koşullandırma ve İşlemsel Koşullandırma, bilişsel süreçlerinin uyarı-tepki bağlantılarını ve uyaranlar arası ilişkileri, ve eylem-etki ilişkilerini öğrenebilen yapay sinir ağlarının tasarlanmasında kullanılması, endüstride değişen ortamlara uyum sağlayabilen sistemlerin geliştirilmesinde yararlı olacaktır.

KAYNAKLAR

Atkinson R.L. Atkinson R.C., Smith E.E., Bem D.J., Hilgard E.R, 1990, Introduction to Psychology, HBJ

Baddaley, A., 1982, *Your Memory*, Pelican Book, U.K

Baloch, A.A. and Waxman A.M., 1991, "Visual learning, adaptive expectations and behavioural conditioning of the mobile robot MAVIN" *Neural Networks* 4, pp.271-302, 1991

Barto A.G., Sutton R.S. and Watkins C.J.C.H., 1989, *Learning and Sequential Decision Making*, COINS Technical Report

Buonomano, D.V., Baxter, D.A.,. and Byrne, J.H., 1990, "Small networks of empirically-derived elements simulate some higher-order features of classical conditioning", *Neural Networks* Vol.3, pp.507-523

Carlson N.R., 1991, *Physiolology of Behavior*, Allyn and Bacon

Gaudiano, P., Surmeli, D. and Wilson, F.D.M., 1994, "Gated dipoles for operant conditioning", Letter to the editor, *Neural Networks* Vol.7, pp.405-406.

Gelenbe, E., 1989, "Random Neural Networks with Negative and Positive Signals and Product form Solution", *Neural Computation*, 1, 502-510

Gelenbe, E., 1990, Stability of the Random Neural Network Model, *Neural Computation*, 2, No.2, 239,

Gelenbe E., 1993, Learning in the Recurrent Random Neural Network, *Neural Computation*, 5, 154

Grossberg, S., "A neural network architecture for Pavlovian conditioning: reinforcement, attention, forgetting, timing" in *Neural Network Models of Conditioning and Action,* Lawrence Erlbaum Assoc., 1991

Grossberg, S. and Schmajuk, N.A.,1987.a "Neural dynamics of attentionally-modulated Pavlovian conditioning; Conditioned reinforcement, inhibition and opponent processing", *Psychobiology* 15, 195-240

Grossberg, S. and Schmajuk, N.A., 1987.b., "Neural dynamics of attentionally-modulated Pavlovian conditioning; blocking, interstimulus interval and secondary reinforcement", *Applied Optics* 26, 5015-5030,

Grossberg, S, Schmajuk, N. and Levine D.S., 1992, "Associative learning and selective forgetting in a neural network regulated by reinforcement and attentive feedback", *in Motivation, Emotion and Goal Direction in Neural Networks,* Lawrence Erlbaum Assoc.

Halýcý U., 1995 (Abstract) "Reinforcement Learning in Random Neural Networks for Mazes", *Symposium on Neuronal Coding,* Prague

Halýcý U., 1996, (Extended Abstract) Reward, Punishment and Expectation in Reinforcement Learning for random Neural Networks, *Workshop on Biologically Inspired Autonomous Systems: Computation, Cognition and Control,* Duke University, Durham, North Caroline, USA

Halıcı U., 1997, "Reinforcement Learning in Random Neural Networks for Cascaded Decisions, *Journal of Biosystems,* Elsevier, Vol 40 No 1-2, pp 83-91

Halıcı U., Akman V., Leloğlu U., 1993, "Yapay Zekada Felsefi Sorunlar" , *Proc of Symposium on Ethical, Sociolagical and Philosophical Aspects of Artificial Intelligence,* Editörler: Halici U., Ucoluk G.

Halıcı U., Yaranlı U., 1992, Neural Networks in Mazes, *IEEE-INNS International Joint Conference on Neural Networks*, Beijing, China, November, 1992, Vol-II, pp 711-716

Gülöksüz A., Halýcý U., 1994, "Artificial Neural Networks for Conditioned Learning", in Hints From Life to Artificial Intelligence, Editor: U. Halýcý, METU pp 28-40

Gulöksüz A., Halıcı U., 1996, "A Neural Circuit to Handle Passive Extinction in Conditioned Reinforcement Learning", *Proceedings of Thirteenth European Meeting on Cybernetics and System Research*, Vienn, Austria

Hulse S.H., Egeth H. and Deese J., 1980, *The Psychology of Learning*, fifth edition, McGraw-Hıll,

Klein, S.B., 1996, *Learning: Principles and Applications,* 3rd edition, McGraw-Hill

Klopf A.H., 1974, Brain Function and Adaptive Systems, *Proc. of IEEE International Conference on Systems, Man Cybernetics.*

Levine D.S., 1989, "Neural network principles for theoretical psychology", *Behaviour Research Methods, Instruments and Computers*, Vol.21, No:2, pp.213-224

Levine, D.S., 1991, *Neural and Cognitive Modeling,* Lawrence Erlbaum Assoc.

Madenoglu A., 1994, *Simulation results for Connectionist Maze Learning*, B.Sc. Project Report, Dept. of EE, METU.

Maki, W.S. and Abunamass, A.M., 1991, "A connectionist approach to conditional discriminations; learning, short-term memory and attention", in *Neural Network Models of Conditioning and Action,* Lawrence Erlbaum Assoc., 1991

Narendra, K., Thathachar M.A.L., 1989, *Learning Automata: An Introduction*, Prentice Hall, Englewood Cliffs, NJ,

O'Keefe J., Nadel L., 1978, *The Hippocampus as a Cognitive Map,* Oxford University Press

O'Keefe John, 1989, "Computation the Hippocampus Might Perform", *in Neural Connections, Mental Computation,* Editors: Nadel L. Et al, MIT Press, pp 225-284

Raymond, J.L., Baxter, D.A., Buonomano, D.V. and Byrne, J.H., 1992, "A learning rule based on empirically-derived activity-dependent neuromodulation supports operant conditioning in a small network", *Neural Networks* Vol.5, pp.789-803

Raymond, J.L., Baxter, D.A., Buonomano, D.V. and Byrne, J.H., 1994, Response to "Gated dipoles for operant conditioning" by Gaudiano et.al., *Neural Networks* Vol.7, pp.405-406

Reynolds A. G. and P. W. Flagg, 1977, *Cognitive Psychology*, Wintrop Publishers,

Schmajuk, N.A. and DiCarlo J.J., 1991, "Neural dynamics and hippocampal modulation of classical conditioning", in *Neural Network Models of Conditioning and Action,* Lawrence Erlbaum Assoc.

Sutton, R.S., 1984, "Temporal credit assignment in reinforcement learning", *Doctoral Dissertation, University of Massachusetts,* Amherst

Sutton R.S., 1988, "Learning to predict by the methods of temporal Difference", *Machine Learning* 3:9-44.

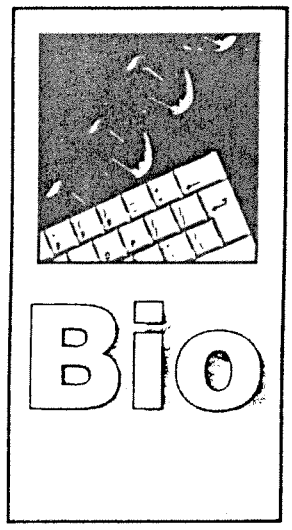
Szepesvary C., 1995, A general framework for reinforcement learning, In Proc. of *International Conference on Artificial neural Networks*, Paris, pp II-165-170

Tsetlin, M.L., 1973, *Automaton Theory and Modelling Biological Systems*, Academic Press, Newyork.

Zhang P. and Canu S., 1995, Indirect adaptive explorations in Entropy-based reinforcement learning, In *International Conference on Artificial neural Networks*, Paris, pp II-171-176

EKLER:

PROJE İLE İLGİLİ YAYINLAR VE TEBLİĞLER



BioSystems

Journal of Biological and Information Processing Sciences

Volume 40 (1997)

ELSEVIER

Amsterdam — Lausanne — New York — Oxford — Shannon — Tokyo

ELSEVIER

BioSystems 40 (1997) 83–91

BioSystems

Reinforcement learning in random neural networks for cascaded decisions

Ugur Halici

Department of Electrical and Electronics Engineering, 06531, METU, Ankara, Turkey

Abstract

The Random Neural Network (RNN) model, in which signals travel as voltage spikes rather than as fixed signal levels, represents more closely the manner in which signals are transmitted in biophysical neural networks. In this paper a reinforcement learning strategy is proposed to make a sequence of cascaded decisions to achieve a goal while aiming to optimize the total cost of the cascaded decisions. For this purpose, RANs are used to model the system and a weight update rule together with a reinforcement function is provided. The performance of the learning strategy is analysed by applying it to the maze learning problem. The simulation results show that the performance of the system is highly dependent on the chosen reinforcement function and quite satisfactory results are obtained when the reinforcement function takes the recency effect into consideration.

Keywords: Random neural networks; Reinforcement learning; Mazes; Recency effect

1. Introduction

One of the major classes of stimulus-response learning is instrumental conditioning which involves association between an action and a stimulus, and permits an organism to adjust its behaviour according to the consequences of that behaviour. That is, when a behaviour is followed by favorable consequences, it tends to occur more frequently; when it is followed by unfavorable consequences, it tends to occur less frequently. Favorable consequences are generally referred to as reward, and unfavorable consequences are referred to as punishment (Carlson, 1977; Szepesvari, 1995; Zhang and Canu, 1995).

In this paper a reinforcement learning strategy using RNNs is proposed and it is tested for mazes. The RNN model is introduced in (Gelenbe, 1989), and extended in (Gelenbe, 1990). In the RNN model signals travel as voltage spikes. This model represents more closely the manner in which signals are transmitted in a biophysical neural network than widely used artificial neuron models in which signals are represented by fixed signal lev-

* Corresponding author, e-mail: halici@rorqual.cc.nmetu.edu.tr

PII S0303-2647(96)01633-4

els. A backpropagation type learning algorithm for recurrent RNN model is introduced in (Gelenbe, 1993).

The reinforcement learning strategy proposed in this paper is inspired by some behavioural and neurophysiological phenomena only as far as this is useful for the particular artificial intelligence problem. The aim here is not to model the neurophysiology of natural learning, but to make use of some natural phenomena in developing a machine learning strategy applicable to cascaded decisions. Although the proposed weight update rule for learning resembles the learning automata (Narendra and Thatchar 1989; Tsetlin, 1973), the way reinforcement is applied in this work differentiates them. The one proposed in this paper is a form of associative reinforcement learning, since it associates different actions with different levels of reinforcement, while learning automata is based on non-associative reinforcement. The method presented here also differs from Q-learning (Watkins and Dayan, 1992) and temporal difference learning (Sutton, 1984; Barto, Sutton and Watkins, 1989), since it does not explicitly maintain estimates of reinforcements for each state-action pair. Such an estimate is somehow implicit in the connection weights. An earlier version of the learning strategy, particular to the maze problem is proposed in (Halici and Yaranli, 1992). While training the network for cascaded decisions, a reinforcement function which takes recency effect into consideration is utilized in this paper. An alternative way of representing the recency effect is introduced as eligibility trace in (Klopf, 1974) and is used in conjunction with temporal difference method in (Sutton, 1988).

The organization of the paper is as follows: a brief explanation of RNN is provided in Section 2. A reinforcement learning strategy for systems requiring a single decision step is proposed in Section 3 and then in Section 4 it is generalized to a more complicated case in which the system makes cascaded decisions to be able to reach a goal. A reinforcement function taking the recency effect into consideration is proposed in Section 5. The simulation results obtained through maze learning are provided in Section 6. Finally Section 7 concludes the study.

2. The random neural network model

In the RNN model (Gelenbe, 1989), n neurons exchange positive and negative impulse signals. Each neuron i accumulates signals as they arrive, and fires if its total signal count at a given instant of time is positive. Firing occurs at random according to an exponential distribution of constant rate $r(i)$ and signals are sent out to other neurons or to the outside of the network. Each neuron i of the network is represented at time t by its input signal potential $k_i(t)$. A negative signal reduces by 1 the potential of the neuron to which it arrives or has no effect on the signal potential if it is already zero, while an arriving positive signal adds 1 to the neuron potential. This is a simplified representation of biophysical neural behaviour. In RNN, signals arrive to a neuron from the outside of the network (exogenous signals) or from other neurons. Each time a neuron fires, a signal leaves it depleting its total input potential. A signal leaving neuron i heads for neuron j with probability $p^+(i,j)$ as a positive signal, or as a negative signal with probability $p^-(i,j)$ or it departs from the network with probability $d(i)$. $p(i,j)=p^+(i,j)+p^-(i,j)$ is the transition probability of a Markov chain representing the movement of signals between neurons satisfying

$$\Sigma_j p(i,j) + d(i) = 1 \qquad \text{for } 1 \le i \le n \tag{2.1}$$

where

$$0 \le p^+(i,j) \le 1 \quad \text{and} \quad 0 \le p^-(i,j) \le 1 \tag{2.2}$$

Exogenous inputs to each neuron i of the network are provided by stationary Poisson processes of rate $\Lambda(i)$, and $\lambda(i)$. For notational convenience the connection weights are defined as

$$w^+(i,j) = r(i)p^+(i,j) \ge 0,$$
$$w^-(i,j) = r(i)p^-(i,j) \ge 0 \tag{2.3}$$

and, therefore

$$r(i) = \Sigma_j (w^+(i,j) + w^-(i,j)) + d(i). \tag{2.4}$$

In (Gelenbe, 1989) it is shown that the steady state probability distribution of the network can be written as the product of probabilities of the states of neurons. However, those equations describing the steady state behaviour is out of the scope of this paper.

3. The weight update rule for a single decision step

Consider a learning system interacting with an environment such that it performs an action a_n selected from a finite set of actions at each time step $n = 0, 1, 2$. After each action a_n is taken, the learning system receives an external reinforcement R_n^e as a result of the interaction with the environment. The amount of the external reinforcement is determined by the environment in some random manner depending on the final state considering the action a_n. Although the reinforcement in general can be positive (reward) or negative (punishment), in this study we will consider only the rewarding case. The objective of reinforcement learning in such a system is to find the action that maximizes the expected external reinforcement while minimizing the cost related with the action necessary to achieve this reinforcement.

In the following, a learning strategy is proposed for handling this type of learning by using the RNN model. As shown in Fig. 1, a neuron, labeled as i in the figure, is used to demonstrate the state just before the decision. This neuron has N outward connections, each representing a different decision. A neuron j is assigned to each possible decision to represent the states achieved by performing the decided action. Call the neuron i the initial node, and the neurons, $j = 1 \ldots N$, achieved after each possible decision the final nodes.

In the network only the initial node is connected to exogenous input, $\Lambda(i) = \Lambda$, and for the other nodes $\Lambda(j)$ is zero, and also $\lambda(k)$ is zero for any neuron. Therefore, all the pulses are originally created at the initial node. The dissipation is set as $d(k) = 1$ if neuron k is a final node, and $d(k) = 0$ otherwise. Since there is no inhibitory connection in the network, and $d(i)$ is zero for the initial node, any signal generated at the initial node reaches one of the final nodes. Which final node it reaches is determined according to the transition probabilities $p(i,j) = p^+(i,j)$. Therefore, the larger the connection weight $w^+(i,j) = r(i)p^+(i,j)$ for a final node j, the more probable node j to receive the signal. When the final cell is reached, since $d(i,j) = 1$, the pulse dissipates thereby exciting the environment, which in turn produces the external reinforcement R_n^e. We call the travel of the pulse from the initial cell, where it is generated, to a final cell, where it is dissipated, a trial. Within a trial, the network is assumed to be able to remember the selected connection in its short term memory, so the selected connection is said to be activated. The active connections are important since only the connection weights of the neurons on the active path are updated in training the network. Updating con-

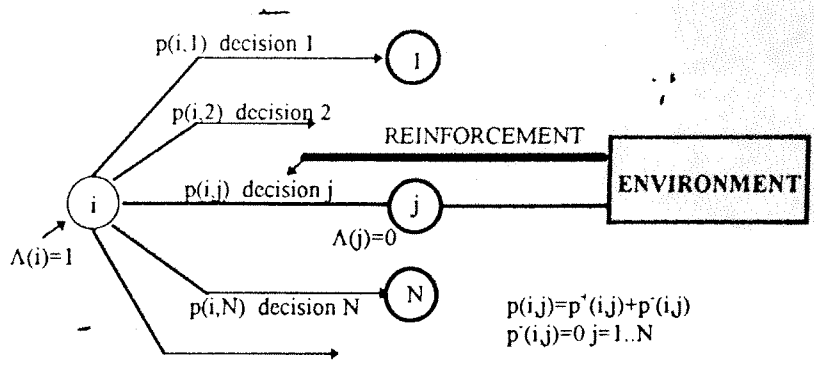


Fig. 1. The system having a single decision step.

nection weights corresponds to storing information into long term memory.

Initially connections for all decisions are assigned an equal probability, all firing rates are set to 1, however, as learning progresses the probabilities are updated according to the weight update rule that we are going to derive in the following.

Let a_n, the action selected at trial n, be the one corresponding to the connection reaching to the node k and let this selection be externally rewarded by amount R_n^e. In its simplest form, the internal reinforcement $R_n(i)$ on node i at trial n for action a_n can be formulated as $R_n(i) = \varphi(R_n^e, L_n)$ where L_n is the cost of the active decision, φ is a function (possibly random valued) that varies inversely with L_n. In systems where reinforcements are the only rewards, the reinforcement function may be assumed to take values in the region $0 \leq R_n(i) \leq 1$ with no loss of generality.

To favour connection selected at trial n in proportion to reward $R_n(i)$, the connection weight $w^+(i,k)$ can be incremented with an amount $\Delta w_n(i,k) = \eta R_n(i)$ where η is the learning rate. However such a weight update rule will cause connection weights and the firing rate to increase in an unlimited manner. To overcome the problem, the value of the firing rate can be fixed as $r(i) = 1$ by normalizing all the weights by amount $(\Sigma_m w_n^+(i,m) + \eta R_n(i))$. Therefore, we obtain the weight update rule below:

$$w_{n+1}^+(i,j) = \begin{cases} (w_n^+(i,j) + \eta R_n(i))/(\Sigma_m w_n^+(i,m) + \eta R_n(i)) & \text{for } j = k \\ w_n^+(i,j)/(\Sigma_m w_n^+(i,m) + \eta R_n(i)) & \text{for } j \neq k \end{cases} \quad (3.1)$$

It can be easily shown that this weight update rule guarantees that

$$\Sigma_m w_{n+1}^+(i,m) = r_{n+1}(i) = 1 \qquad \text{for } n = 1, 2 \ldots \quad (3.2)$$

Since we have $w_n^+(i,j) = r_n(i) p_n^+(i,j) = p_n^+(i,j)$, the weight update rule can be reformulated in terms of connection probabilities as:

$$p_{n+1}^+(i,j) = \begin{cases} (p_n^+(i,j) + \eta R_n(i))/(1 + \eta R_n(i)) & \text{for } j = k \\ p_n^+(i,j)/(1 + \eta R_n(i)) & \text{for } j \neq k \end{cases} \quad (3.3)$$

which can be further simplified as:

$$p_{n+1}^+(i,j) = \begin{cases} p_n^+(i,j) + \eta R_n(i)(1 - p_n^+(i,j)) & \text{for } j = k \\ p_n^+(i,j) + \eta R_n(i)(-p_n^+(i,j)) & \text{for } j \neq k \end{cases} \quad (3.4)$$

In this formula the learning rate η should be chosen as small as possible to provide convergence to the decision with maximum expectation of reinforcement at a minimum cost. On the other hand, the value of η is critical on the number of trials necessary for convergence.

4. Systems with cascaded decision steps

It is not hard to optimize action when the system makes its decision at a single step and the reinforcement is obtained immediately after the action performed. However, it is not that easy a task when the system is required to make a decision comprising different steps and the reinforcement is obtained only at the end as a consequence of the overall action performed at each step.

Consider a learning system interacting with an environment as described by a discrete-time dynamical process with a finite set of states X. At time steps $t = 0, 1, 2 \ldots$ the environment is in state $x(t)$, where $x(t) \in X$. At time step t, after observing the possible choices at state $x(t)$, the learning system decides to perform one of the actions: $a(t)$. The action $a(t)$ affects the environ-

ment, causing it to make a transition from state $x(t) = x_i$ to a new state $x(t+1) = x_j$ in a deterministic manner and independently of its past history. After several actions $a(t)$ are taken, if a final state is reached the learning system receives an external reinforcement R_n^e which is determined in some random manner depending on the final state x_{fn} and the collection of actions, $\mathbf{a} = < a(t),\ t = 0..t_f >$. The objective of reinforcement learning here is to find an optimal sequence of actions $\mathbf{a}_n = < a_n(t),\ t_n = 1..t_{fn} >$ for trial n that maximizes the expected external reinforcement while minimizing the total cost of the sequence of actions necessary to achieve this external reinforcement, as learning progresses.

The RNN to handle the situation involving sequential decisions is demonstrated in Fig. 2. Each neuron in the system represents a possible state, and each outward connection corresponds to the decision for a possible action. One of the neurons is labeled as the initial node, some of them are final nodes and the others are intermediate nodes. While the initial node corresponds to a state at which no decision is made yet, the final nodes correspond to states at which the system is granted an external reward. The parameters of the network are set as previously depicted for the initial and final nodes. However, for the intermediate nodes we have $\Lambda(k) = 0$, $\lambda(k) = 0$, $d(k) = 0$.

In the network only the initial node is connected to the exogenous input, $\Lambda(i) = \Lambda$, for the other nodes $\Lambda(j)$ is zero. All the pulses originate at the initial node. Any pulse created at the initial node reaches the final node at the end and dissipates there. During the learning phase we set Λ to a very small value, so that only one pulse travels at a time. If a new pulse happens to be generated before the previous one is dissipated, we simply ignore the effect of the previous one. Therefore, after the pulse is created and before it is dissipated, exactly one neuron has potential value $k(i) = 1$ indicating where we are in the decision sequence. According to the decision made at neuron i, the corresponding neuron, say neuron j, will receive the signal, and so will have $k(j) = 1$ indicating the current state. Such a neuron j in the neighborhood is decided in accordance with the transition probabilities $p(i,j) = p^+(i,j)$. Therefore, the one among the neighbours with the largest connection weight $w^+(i,j) = p^+(i,j)$ has the highest probability of being selected. When the final cell is reached, since $d(i) = 1$, the pulse dissipates there by exciting the environment, which in turn produces reinforcement depending on all the actions decided. As in

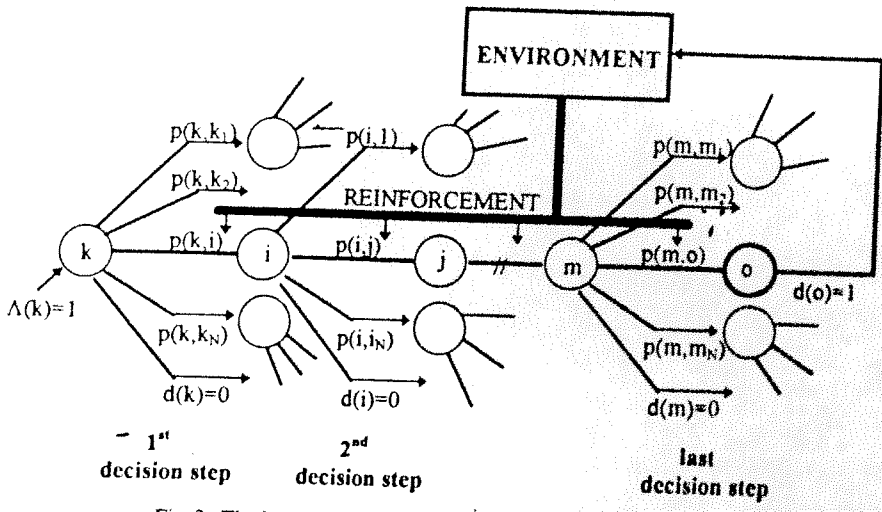


Fig. 2. The learning system requiring cascaded decision steps.

the single step-decision case, the travel of the pulse from the initial cell, where it is generated, to a final cell, where it is dissipated, corresponds to a trial. Within a trial, the network is assumed to be able to remember all of the activated path in its short term memory. Only the connection probabilities of the neurons on the active path are updated.

Initially connections for all decisions possible at state i are assigned equal probability. It should be noted that there is initially no need to know the complete structure of the decision tree. As the nodes are visited for the first time, the connections for the visited node become known and all the outgoing connections are assigned equal probability for the node just visited. As learning progresses the connection probabilities are updated according to the following weight update rule:

$$p_{n+1}^{+}(i,j) = \begin{cases} p_n^{+}(i,j) + \eta R_n(i) & \text{if } i \text{ is on the active path and } (i,j) \text{ is activated} \\ (1 - p_n^{+}(i,j)) & \\ p_n^{+}(i,j) + \eta R_n(i) & \text{if } i \text{ is on the active path but } (i,j) \text{ not activated} \\ (-p_n^{+}(i,j)) & \\ \text{no change} & \text{otherwise} \end{cases} \tag{4.1}$$

At a trial η, if a final state is reached then an external reinforcement R_n^e is obtained. For the system with cascaded decision steps, an internal reinforcement function $R_n(i)$, for optimizing these cascaded decisions may be defined as $R_n(i) = \varphi(R_n^e, L_n, l_n(i))$ where L_n is the total cost of the active decision path at trial n, $l_n(i)$ is the cost of the portion after decision at neuron i, φ is a function (possibly random valued) that varies inversely with L_n and $l_n(i)$ such that $0 \leq R_n(i) \leq 1$. In our experiments on mazes, we observed better performance when the reinforcement function reflected the recency effect, which is explained in the next section.

5. A reinforcement function with recency effect

One method of observing short term memory is asking people to study a long list of items and to recall as many of these words as possible. The tendency of the last few items to be well recalled is known as recency effect because it reflects the recall of the most recent items (Baddaley, 1983). Another important point in list learning is the length of the list. Manipulations of list length do not seem to affect the recency section of the curve. However, the early and middle sections are affected. The longer the list, the lower the probability that an item from one of these positions will be recalled (Reynolds and Flagg, 1977)

A reinforcement function $R_n(i) = \varphi(R_n^e, L_n, l_n(i))$ that partially reflects the recency effect in list learning might be chosen as (Fig. 3):

$$R_n(i) = \begin{cases} (R_n^e/L_n)(\max(\kappa, 1/(l_n(i) - c)) & \text{for } c < l_n(i) \\ 0 & \text{for } c \geq l_n(i) \end{cases} \tag{5.1}$$

where κ is a constant, having value between 0 and 1, and used to adjust contribution of the recency effect. Here in addition to recency effect another parameter, namely c, is taken into consideration, according to which, the reinforcement function is shifted to left. Here c corresponds to the cost of the actions at the certainly learned portion of the active path. We say that the last portion of the path is certainly learned if all the connection probabilities belonging to this portion are larger than a predetermined threshold value very close to 1.

Notice that if $\kappa = 1$ then we have $R_n(i) = R_n^e/L_n$ so the recency effect is completely ignored, all the nodes are reinforced to learn in the same way whether they are close or not to the final node. On the other extreme if $\kappa = 0$ then $R_n(i) = (R_n^e/L_n)\ (l/(l_n(i) - c))$, so the recency effect is dominant, only the nodes close to the certainly learned portion are reinforced to lear

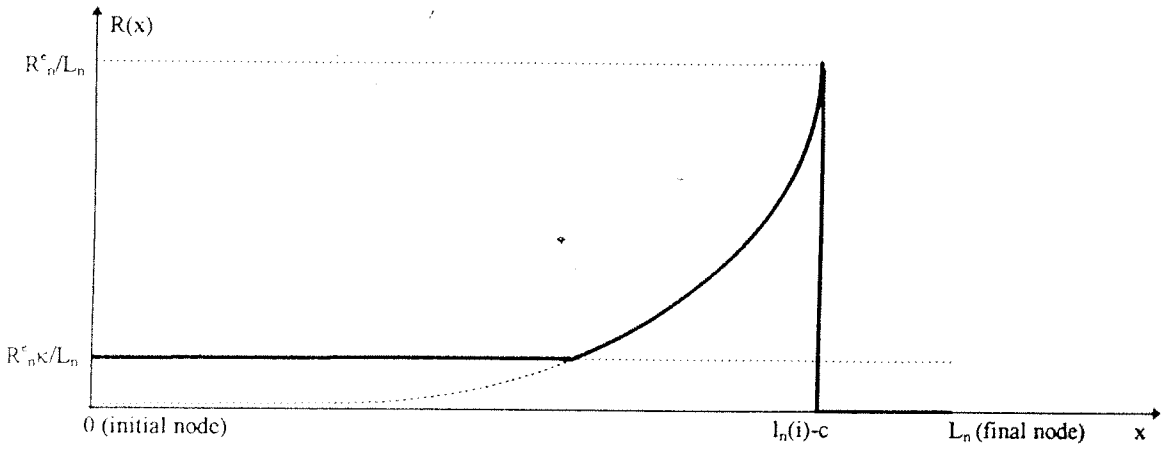


Fig. 3. The reinforcement function used in the experiments.

6. Simulation results

The reinforcement learning algorithm proposed in this paper for systems with cascaded decision steps is tested for a maze of size 11×23 shown in Fig. 4, where the dark cells correspond to the walls and the others correspond to the passages (Madenoglu, 1994). The aim is to find out a path in the passages connecting the initial cell to the final one, and meanwhile to learn a reasonably short path as the experiment on the maze is repeated several times. The cells placed on the shortest path from the initial cell to the final one are marked with arrows in the figure.

The maze pattern is represented by RNN by assigning a neuron for each cell in the passages of the maze, and establishing connections between neighbouring cells in both directions. The network parameters are set in accordance with Section 4. The network is tested for 100 different random seeds for several different values of learning rate η and κ in the reinforcement function.

In our experiment we obtained typical learning curves, that is, as the number of trials increases, the total elapsed time for a trial converges to the length of the learned path. This means that a pulse that progresses in a path does not turn

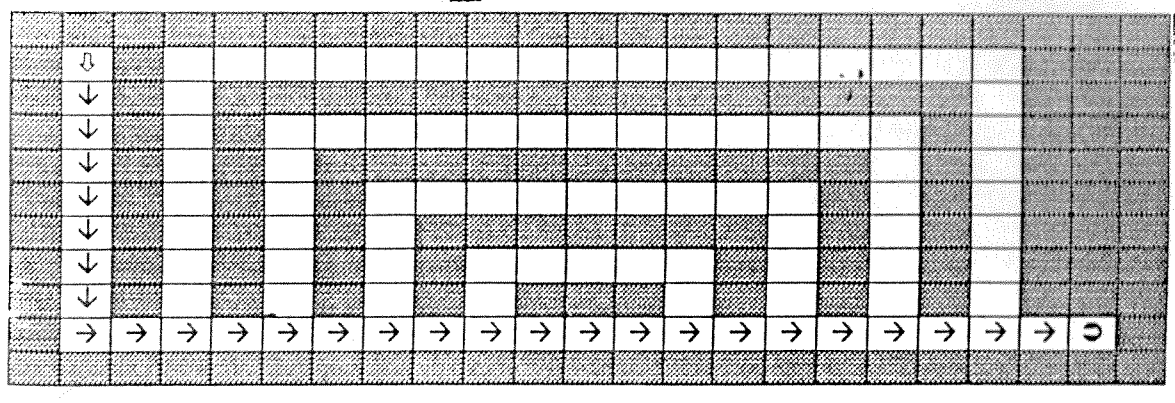


Fig. 4. The cellular array representing a maze and the shortest path connecting the initial cell to the final.

back, it goes from the initial node to the final node directly.

Table 1 demonstrates the distribution of the number of paths learned on path length. In this table, the first and second columns show the values set for κ and η. The third column shows the average numbers of trials required for certainly learning the path. The rest of the columns indicates how many times the path having the mentioned length at the column heading is learned when the network is tested with 100 different random number generation seeds.

Notice that the value of κ is quite important in learning the shortest path. In the table the top portion corresponds to a reinforcement function in which the recency effect is completely ignored, while the second portion corresponds to a reinforcement function in which the recency effect is dominant. The last portion correspond to intermediate cases. Within the first and the second portion the learning rate is changed from a smaller value to a larger one. Notice that the best distribution favouring the shortest path is obtained when the recency effect is dominant and the learning rate is small.

7. Conclusion

In this paper a reinforcement learning strategy is proposed for RNNs and it is tested for mazes. First a reinforcement learning strategy is provided for systems requiring a single decision step; and then it is generalized to a more complicated case, in which the system makes cascaded decisions to reach a goal. Meanwhile, a reinforcement function taking the recency effect into consideration is proposed and the learning strategy is applied for maze learning.

The experimental results on maze learning are quite satisfactory, indicating the existence of

Table 1
Distribution of learned paths on various lengths

κ	η	AvgTrial	Distribution of the learned paths (%)				
			28	32	36	40	44
1.000	0.75/28	266	87	13	0	0	0
1.000	1.00/28	204	78	19	2	0	1
1.000	1.25/28	155	72	26	2	0	0
1.000	1.50/28	133	72	24	4	0	0
1.000	1.75/28	117	64	24	11	1	0
1.000	2.00/28	103	69	22	8	1	0
1.000	2.25/28	90	66	29	4	0	1
1.000	2.50/28	86	56	27	15	2	0
1.000	2.75/28	74	58	25	13	4	0
0.000	6.00/28	283	100	0	0	0	0
0.000	8.00/28	211	100	0	0	0	0
0.000	10.00/28	161	98	2	0	0	0
0.000	12.00/28	134	98	2	0	0	0
0.000	14.00/28	113	100	0	0	0	0
0.000	16.00/28	96	98	2	0	0	0
0.000	18.00/28	86	98	2	0	0	0
0.000	20.00/28	75	96	4	0	0	0
0.125	4.00/28	304	97	3	0	0	0
0.200	4.00/28	218	92	8	0	0	0
0.400	4.00/28	121	72	24	3	1	0
0.500	4.00/28	98	63	27	9	1	0
0.600	4.00/28	86	63	22	15	0	0
0.800	4.00/28	67	53	3	11	5	1

learning. Such a schema does not guarantee that the shortest path will be learned, but it favors shorter paths when the recency effect is dominant in the reinforcement function. Convergence time decreases with increasing learning rate, however a fast learning rate may cause paths other than the shortest one to be learned as well.

Acknowledgements

This work is being partially supported by Scientific and Technical Council of Turkey under grant EEEAG-126 Project: Modelling Cognitive Processes by Artificial Neural Networks.

References

Baddaley, A., 1982, Your Memory, Pelican Book, UK

Barto, A.G., Sutton, R.S. and Watkins, C.J.C.H., 1989, Learning and Sequential Decision Making, COINS Technical Report.

Carlson, N.R., 1977, Physiology of Behaviour, Allyn and Bacon.

Gelenbe, E., 1989, Random neural networks with negative and positive signals and product form solution. Neural Comput. 1(4), 502–510.

Gelenbe, E., 1990, Stability of the random neural network model. Neural Comput. 2(2), 239–247.

Gelenbe, E., 1993, Learning in the recurrent random neural network. Neural Computation 5(1), 154–164.

Halici, U. and Yaranli, U., 1992, Neural Networks in Mazes, in: Proc. of IEEE-INNS International Joint Conference on Neural Networks. Beijing, China, November, Vol-II, 711–716.

Klopf, A.H., 1974, Brain Function and Adaptive Systems, in: Proc. of IEEE International Conference on Systems, Man Cybernetics.

Madenoglu, A., 1994, Simulation results for Connectionist Maze Learning, B.Sc. Project Report, Dept. of Electrical Engineering, METU.

Narendra, K. and Thathachar M.A.L., 1989, Learning Automata: An Introduction, Prentice Hall, Englewood Cliffs.

Reynolds, A.G. and Flagg, P.W., 1977, Cognitive Psychology, Wintrop Publishers.

Szepesvary, C., 1995, A general framework for reinforcement learning, in: Proc. of International Conference on Artificial neural Networks, Paris, II, 165–170.

Sutton, R.S., 1984, Temporal Credit Assignment in Reinforcement Learning. Doctoral Dissertation, University of Massachusetts, Amherst.

Sutton, R.S., 1988, Learning to predict by the methods of temporal difference. Machine Learn. 3, 9–44.

Tsetlin, M.L., 1973, Automaton Theory and Modelling Biological Systems, Academic Press.

Watkins, C. and Dayan, P., 1992, Technical Note: Q learning. Machine Learn. 8, 55–68.

Zhang, P. and Canu, S., 1995, Indirect Adaptive Explorations in Entropy-based Reinforcement Learning, in: Proc. of International Conference on Artificial Neural Networks, Paris, II, 171–176.

CYBERNETICS AND SYSTEMS '96

volume 2

ROBERT TRAPPL

editor

Austrian Society for Cybernetic Studies

A Neural Circuit to Handle Passive Extinction in Conditioned Reinforcement Learning

Aslı Gülöksüz **Uğur Halıcı**
Department of Electrical and Electronics Engineering
METU, 06531 Ankara, TURKEY
{guloksuz, halici}@rorqual.cc.metu.edu.tr

Abstract

The concepts of classical conditioning can be used in designing neural networks for association and expectation learning, and behavioural conditioning in artificial adaptive systems. Classical conditioning concepts such as excitatory and inhibitory primary and secondary conditioning have been modelled by the Recurrent Associative Dipole (READ) [Grossberg and Schmajuk, 1987] . However, this circuit does not satisfy the behavioural data on extinction in case of the nonoccurence of an expected event. In this work, the circuit is modified so that it will model passive extinction as well. The changes in the performance of the READ circuit introduced by this modification are then explored.

Keywords: Artificial neural networks, reinforcement learning, classical conditioning, behavioural conditioning, extinction

1. Introduction

The concepts of classical conditioning in psychology can be used to design neural networks that can learn associations between different stimuli, and between stimuli and responses. Such neural networks can be incorporated in any adaptive system.

As some associations between stimuli and responses seem to be inborn in animals; such as the association between the sight of food and salivation for a dog [3]; some vital stimulus-response pairs can be hard-wired or initially set by software in artificial systems. Then, these initial associations can be used to acquire new ones.

This work is being supported by The Scientific and Technical Council of Turkey under grant EEEAG-126.

The Recurrent Associative Dipole (READ) [Grossberg and Schmajuk, 1987] is a neural circuit that models some classical conditioning concepts such as excitatory and inhibitory conditioning, primary and secondary conditioning, opponent extinction and habituation. A number of these circuits are used as part of the cognitive system of the mobile robot MAVIN [Baloch and Waxman, 1991] developed at the MIT Lincoln Laboratory.

The operation of the READ circuit in itself does not conform with the psychological phenomenon of extinction in the case of unconfirmed expectations (passive extinction). However, the extinction of associations is a useful property for the adaptive capability of a system, in that it allows the system to learn new associations after forgetting ones that are no longer valid, besides preventing it from learning further associations based on ones that are no longer valid. We have made a modification to the READ circuit and to the set of differential equations defining it, in order to model extinction. In Section 4, these modifications are described, and simulation results of the READ circuit and its modified version are given in Section 5.

2. Classical Conditioning

Classical or Pavlovian conditioning is the type of learning in Pavlov's well known experiment in which a dog is repeatedly presented with the sound of a bell before being given food, and learns to salivate at the sound of the bell alone. In this section, we will give some basic definitions on classical conditioning that will be used in the following sections [Baloch and Waxman, 1991].

In Pavlov's experiment, food is the *unconditioned stimulus* (US); i.e., a stimulus that is assumed to be already related to an *unconditioned response* (UR); namely salivation. The bell, on the other hand, is called *conditioned stimulus* (CS). When a CS is conditioned to a response, the response becomes its *conditioned response* (CR).

The type of conditioning in which the CS is conditioned to the response related to the US, as in Pavlov's experiment, is called *excitatory conditioning*. In *inhibitory conditioning*, the CS is repeatedly presented after the offset of the US, and is thus conditioned to the response related to this offset.

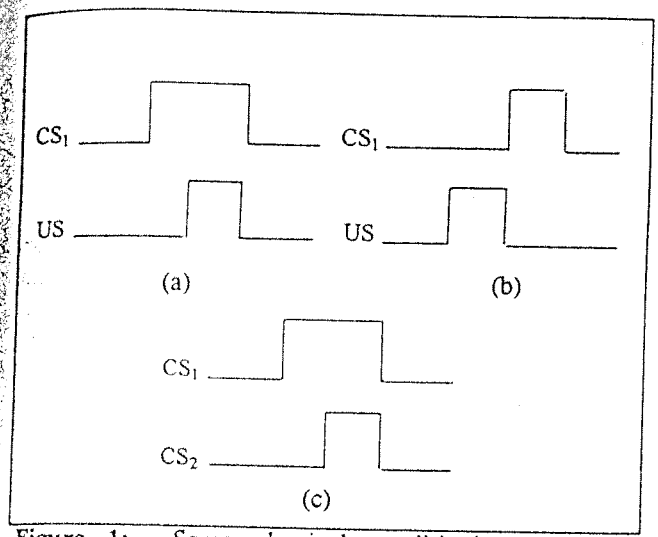


Figure 1: Some classical conditioning schedules (a)Primary excitatory conditioning (b)Primary inhibitory conditioning (c)Secondary excitatory conditioning

Once a conditioned stimulus CS_1 has been conditioned by using a US, it can in turn be used in conditioning another stimulus CS_2. In Pavlov's experiment, for instance, after the bell has been paired with food for a sufficient number of times, repeatedly pairing a light with the bell will cause the dog to salivate at the sight of the light, even when food is not used as an US. This phenomenon is called *secondary conditioning*. Schematic diagrams of these types of conditioning are given in Figure 1.

3. The Recurrent Associative Dipole (READ)

The Recurrent Associative Dipole [Grossberg and Schmajuk, 1987] is a gated dipole with the addition of a feedback path to allow secondary conditioning. It consists of two channels: one related to the *on-response* of a particular stimulus US, and the other to the *off-response* of the US in question.

The on-channel and off-channel of the READ circuit in Fig. 2a are the columns on the left and the right respectively. . Both channels have modifiable synapses (w_{k7} and w_{k8}) with nodes pertaining to other stimuli (CS's). There is a competitive, or opponent interaction between the on-channel and the off-channel. The negative connections x_3-x_6 and x_4-x_5 provide opponent extinction between the two channels, i.e., the activation of one channel inhibits the opposite response.

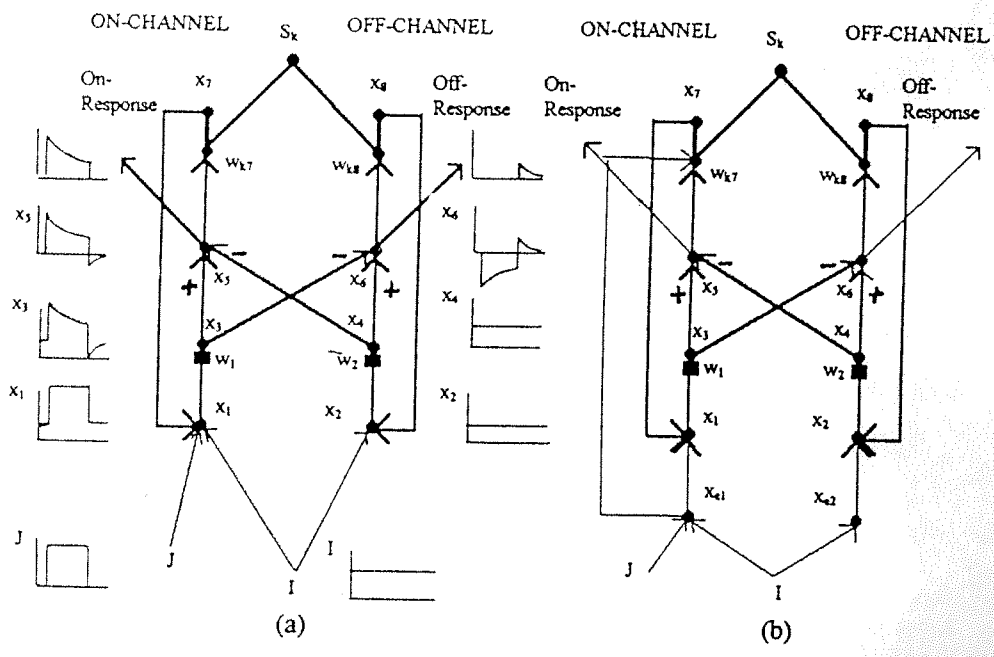


Figure 2: (a) The Recurrent Associative Dipole (READ) circuit (b) Modified READ that handles passive extinction. The responses of the nodes of the READ circuit in part (a) to a pulse of J are plotted beside each node. The responses of the nodes in part (b) are similar if the US is presented at J. See the simulation results in Figure 3 for the case in which US is not applied.

Signals to both channels are gated by slowly habituating and recovering transmitters. A sustained arousal input (I in Figure 2) which is equal for both channels provides the energy necessary to generate a rebound at the offset of an input. When the US is applied to the on-channel, w_1 will be depleted according to eqn.3. Thus, just after the offset of the US, the off channel will win the competition. This phenomenon makes inhibitory conditioning possible.

The operation of the READ circuit in Figure 2a is defined by equations (1)-(14) given below. Primary excitatory conditioning takes place when a CS is presented at a node S_k, and the US pertaining to the READ circuit is presented at x_1 as shown in Fig. 2a, causing an increase in the weight w_{k7} according to eqn. 11. The squares adjacent to the nodes x_3 and x_4 indicate that these have synapses with habituating and recovering transmitters. For example, w_1 decreases when x_1 is active, and recovers when it is not, according to eqn. 3. This causes the activations x_3 and x_5 to decay. Since the bias input I is equal for both channels, the habituation of w_1 results in a *rebound* in the off-channel after the offset of the US. If the CS is presented during this rebound, the weight w_{k8} increases according to equation (12), and the CS is conditioned to the off-response, i.e., primary inhibitory conditioning takes place.

Arousal + US + Feedback On-Activation:

$$\frac{dx_1}{dt} = -ax_1 + I + J + f(x_7) \quad (1)$$

I:Arousal input J:Input to the On-channel (US)

Arousal + Feedback Off-Activation:

$$\frac{dx_2}{dt} = -ax_2 + I + f(x_2) \quad (2)$$

Depletable On and Off Transmitters:

$$\frac{dw_1}{dt} = b(1-w_1) - cg(x_1)w_1 \quad (3)$$

$$\frac{dw_2}{dt} = b(1-w_2) - cg(x_2)w_2 \quad (4)$$

Gated On and Off Activations:

$$\frac{dx_3}{dt} = -ax_3 + eg(x_1)w_1 \quad (5)$$

$$\frac{dx_4}{dt} = -ax_4 + eg(x_2)w_2 \quad (6)$$

Normalized Opponent On and Off Activations:

$$\frac{dx_5}{dt} = -ax_5 + (h-x_5)x_3 - (x_5+k)x_4 \quad (7)$$

$$\frac{dx_6}{dt} = -ax_6 + (h-x_6)x_4 - (x_6+k)x_3 \quad (8)$$

Total On and Off Activations:

$$\frac{dx_7}{dt} = -ax_7 + m[x_5]^+ - p\sum S_k w_{k7} \quad (9)$$

$$\frac{dx_8}{dt} = -ax_8 + m[x_6]^+ - p\sum S_k w_{k8} \quad (10)$$

On -conditioned and Off-conditioned Reinforcer Learning:

$$\frac{dw_{k7}}{dt} = S_k(-qw_{k7} + r[x_5]^+) \quad (11)$$

$$\frac{dw_{k8}}{dt} = S_k(-qw_{k8} + r[x_6]^+) \quad (12)$$

On and Off Responses:

$$ON = [x_5]^+ \quad (13)$$

$$OFF = [x_6]^+ \quad (14)$$

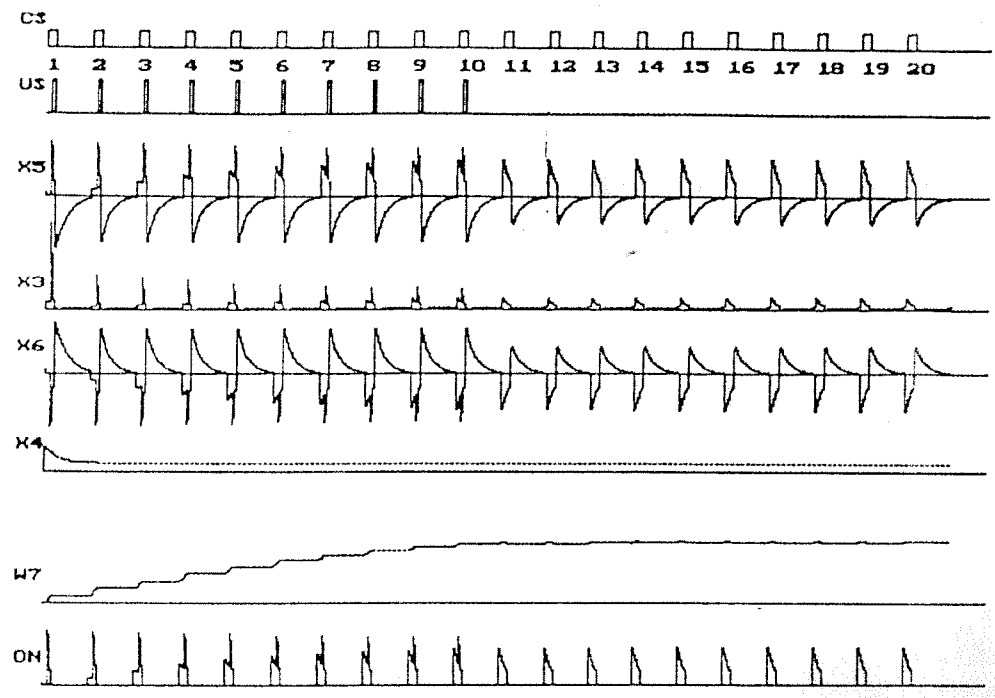
After conditioning, the feedback path from x_7 (x_8) to x_1 (x_2) allows the CS alone to activate x_1 (x_2) and thus generate the on-response (off-response). This makes secondary conditioning possible.

In the original READ circuit, once excitatory conditioning has taken place, the weight w_{k7} does not decay even if the US never arrives after the CS; i.e., even if the expectation that the US will arrive is not confirmed. This can be noticed by observing equation (11): in order for w_{k7} to decay, the CS must be active while x_5 is not. However, after conditioning, the CS alone is sufficient to activate the on-channel. Therefore this decay never takes place. However, behavioural data obtained by cognitive psychologists suggests that the repeated nonoccurrence of the US after the CS results in the extinction of conditioned learning [Hulse et.al.,1980]. In the next section, this phenomenon is examined in some more detail, and a modification is made in the READ circuit to support it.

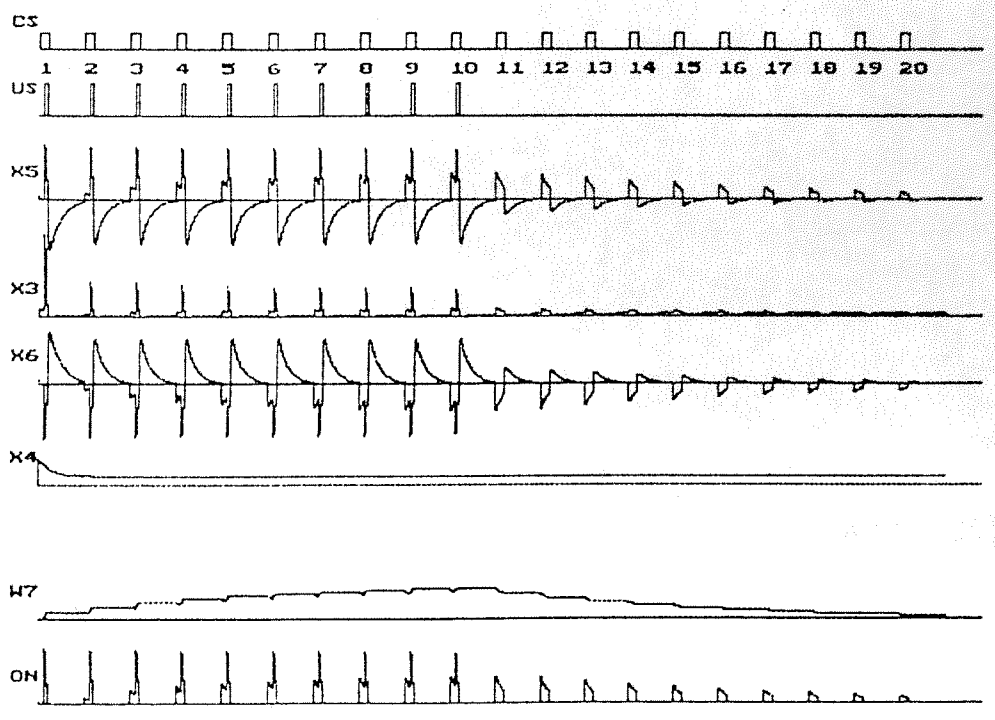
4. A Modification to the READ Circuit to Handle Passive Extinction

In animals, learned responses are dropped if they are not reinforced [Hulse *et.al.*, 1980]. For example, if after being conditioned by using the bell-food pair, the dog in Pavlov's experiment is repeatedly presented with a bell but no food appears, it will start to salivate less and less in response to the bell, and eventually it will not salivate at all. This phenomenon is called *extinction*.

The READ circuit supports extinction in neither excitatory nor inhibitory conditioning. Inhibitory conditioning, as defined in the introduction, does not involve expectation learning. Therefore one cannot talk about the nonoccurrence of an expected event, and extinction is not supposed to occur. However, associations learned by excitatory conditioning are subject to extinction.As explained in the previous section, the reason for extinction not to take place in the READ circuit is that, after conditioning, CS activates the on-channel in exactly the same way that the US does (only with a slightly smaller activation.) Therefore it is not possible to differentiate between the US and a previously conditioned CS. This suggests that another level of neurons is needed to make this differentiation.

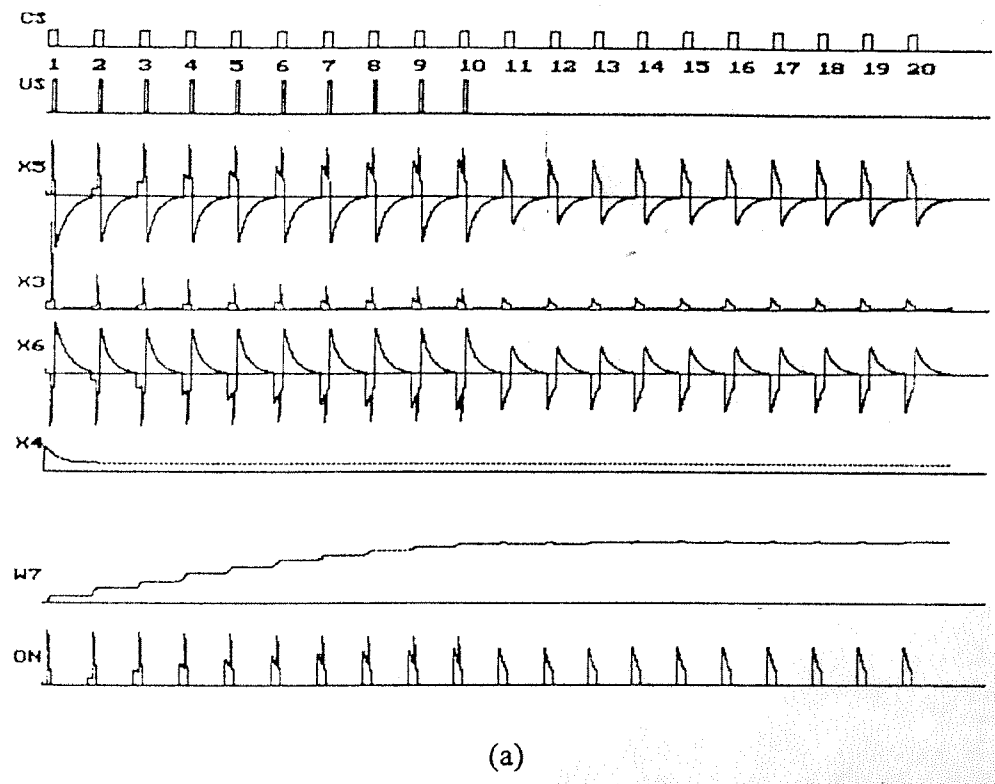


(a)

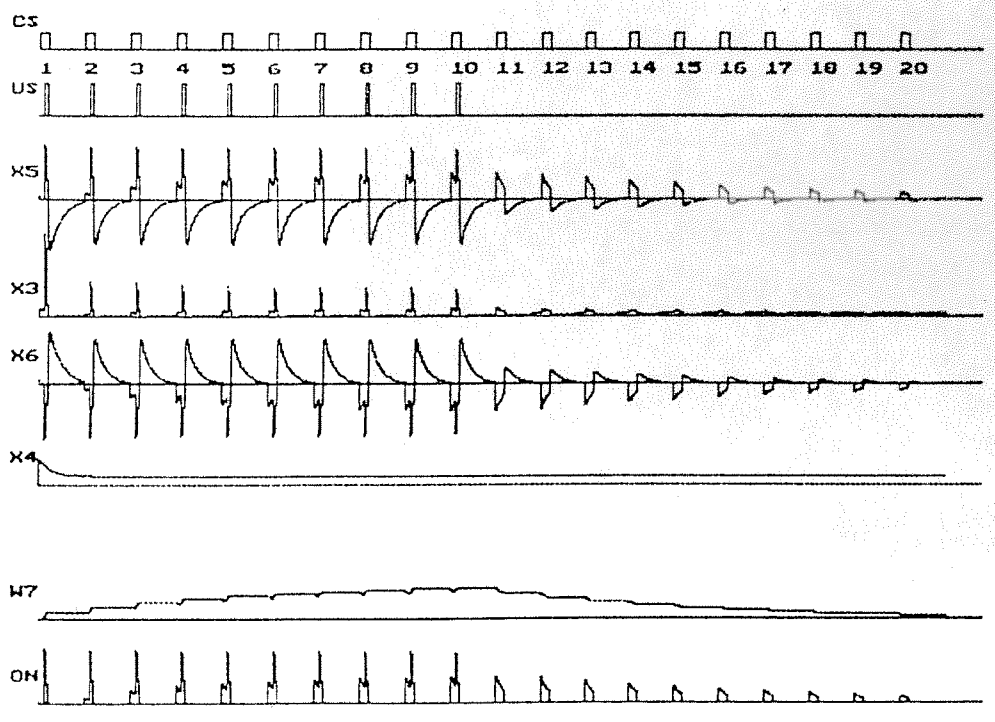


(b)

Figure 3: Simulation results for primary excitatory conditioning (a)with the READ circuit in Fig. 2a. (b) with the modified READ circuit in Fig.2b. In the conditioning phase (intervals 1-10) the US is paired with the CS In the second phase (intervals 11-20), presenting the CS alone results in an on-response. (Parameters are a=1, b=.005, c=.00125, e=20, h=20, k=20, m=.5, q=.005, r=.025, p=20)



(a)



(b)

Figure 3: Simulation results for primary excitatory conditioning (a)with the READ circuit in Fig. 2a. (b) with the modified READ circuit in Fig.2b. In the conditioning phase (intervals 1-10) the US is paired with the CS In the second phase (intervals 11-20), presenting the CS alone results in an on-response. (Parameters are a=1, b=.005, c=.00125, e=20, h=20, k=20, m=.5, q=.005, r=.025, p=20)

A modified version of the READ circuit is given in Fig. 2b, in which node x_{e1} provides the possibility to make this differentiation. Node x_{e2} has been added solely to preserve the symmetry of the circuit. The differential equations defining the operation of the modified circuit are given below:

Newly added nodes:

$$\frac{dx_{e1}}{dt} = -ax_{e1} + I + J \qquad (15)$$

$$\frac{dx_{e2}}{dt} = -ax_{e2} + I \qquad (16)$$

Modified equations:

$$\frac{dx_1}{dt} = -ax_1 + x_{e1} + f(x_7) \qquad (17)$$

$$\frac{dx_2}{dt} = -ax_2 + x_{e2} + f(x_8) \qquad (18)$$

$$\frac{dw_{k7}}{dt} = S_k(-qw_{k7} + r[x_5q(x_{e1}-I)]^+) \qquad (19)$$

According to the modified equations above, during the conditioning phase, the operation of the circuit is identical to that of the original READ circuit. However, w_{k7} decays each time the activation of x_{e1} is below a threshold slightly higher than the bias input; i.e., w_{k7} decays when the CS is present and the US is not. This results in extinction in case of the nonoccurrence of the expected US.

5. Simulation Results

Figure 3 contains the simulation results of primary excitatory conditioning with the READ circuit in Figure 2a and b. Note that the on-response is identical to the positive portion of x_5. In either part, 10 trials are made in which the CS is presented for 200 time units, and the US is applied as well in the last 40 of these. One can observe the growth in the on-response *ON* at the presentation of CS at each trial. In the second phase of Figure 3a, at each of the 10 trials in which only the CS is presented, the on-response has the same amplitude; in fact, there is no change in w_7 either.

Figure 3b contains the results of the same experiment performed by using the modified READ circuit in Figure 2b. During the conditioning phase, the circuit experiences the nonoccurrence of an expectation at the beginning of each presentation of CS until US is applied, and therefore w_7 decays slightly, but when US is eventually applied, the weight recovers from this decay. This slows down the learning process slightly, but not significantly. In the second phase of Figure 3b, the on-response generated by the CS alone decreases each time the US does not arrive, and approaches zero at the end of 10 trials. Thus extinction takes place.

6. Conclusion

Extinction is an important property in animal learning, since it increases the adaptive capabilities of the animal by allowing it to drop responses that are no longer useful. Simulation results have shown that the modification made to the READ circuit in this paper handles passive extinction as well as all the phenomena modelled by the READ circuit.

One phenomenon that must be mentioned here is secondary conditioning. The effect of the modifications on secondary conditioning will be as follows: Assume that CS_1 has been conditioned by using US. When CS_1 and CS_2 are then presented as in Figure 1c, secondary conditioning will take place to a lesser and lesser degree each time CS_1 is not followed by US. Then, the extinction of the weight corresponding to CS_2 will be dependent on the occurrence of US and not of CS_1. This is a desirable property, since, for instance in the case of Pavlov's experiment, the useful expectation to be learned is that the light will be followed by food and not that it will be followed by the sound of a bell. Therefore this circuit implements secondary conditioning in a way that conforms with behavioural data.

References

[Grossberg and Schmajuk, 1987] Grossberg, S. and Schmajuk, N.A. "Neural dynamics of attentionally modulated Pavlovian conditioning: conditioned reinforcement, inhibition, and opponent processing." *Psychobiology* 15, pp.95-240, 1987

[Baloch and Waxman, 1991] Baloch A.A. and Waxman A.M. "Visual learning, adaptive expectations, and behavioural conditioning of the mobile robot MAVIN", *Neural Networks*, Vol.4, pp.271-302, 1991

[Hulse *et.al.*, 1980] Hulse, H.H., Egeth,H., and Deese,J.*The Psychology of Learning*, McGraw-Hill,1980

[Grossberg, 1991] Grossberg, S. "A neural network architecture for Pavlovian conditioning: reinforcement, attention, forgetting, timing ", in *Neural Network Models of Conditioning and Action*, ed. Commons,M.L., Grossberg, S., Staddon, J.E.R., Lawrence Erlbaum Assoc., 1991

[Levine, 1991] Levine,D.S., *Neural and Cognitive Modelling*, Lawrence Erlbaum Assoc., 1991

Workshop Biologically Inspired Autonomous Systems: Computation, Cognition and Control, Duke University, Durham, North Caroline, USA, 4-5 March, 1996

EXTENDED ABSTRACT

REWARD, PUNISHMENT AND EXPECTATION IN REINFORCEMENT LEARNING FOR RANDOM NEURAL NETWORKS[1]

Ugur HALICI
Department of Electrical and Electronics Engineering, 06531, METU, Ankara, Turkey,
email: halici@rorqual.cc.metu.edu.tr

The reinforcement learning strategy proposed in (Halici 1995) for random neural networks is based on reward and performs well for a stationary environment. However when the environment is not stationary it suffers from getting stuck to the previously learned action and extinction is not possible. In this paper the reinforcement learning strategy is extended by introducing a weight update rule for punishment preserving the properties peculiar to RNNs. The performance obtained by the weight update rule with punishment proposed in this paper is better than that of the rule used in learning automata. Furthermore, in this paper a learning strategy which takes into consideration the expectation of reinforcement is introduced. With the proposed strategy, the system behaves as in learning with reward when the reward for the learned action is not below the expectation, otherwise it behaves as in learning with punishment so that other possibilities can be explored. Such a strategy has made extinction possible while resulting in a total convergence, in most cases to the most rewarding action.

Keywords: *random neural networks, reinforcement learning, punishment, extinction, expectation*

Introduction

One of the major classes of stimulus-response learning is instrumental conditioning (also called operant conditioning) where the organism is allowed to have an active role in the learning situation. This kind of learning permits an organism to adjust its behaviour according to the consequences of that behaviour. That is, when a behaviour is followed by favourable consequences, the behaviour tends to occur more frequently; when it is followed by unfavourable consequences, it tends to occur less frequently. Collectively favourable consequences are referred to as *reward*, and unfavourable consequences are referred to as *punishment*. (Carlson 1977, Hulse et al 1980)

In an artificial neural network it is the weight values assigned to interconnections that allows the network to learn and remember.. In the beginning, the weights in the network are assigned initial values and in the training phase, these connection weights are updated iteratively, by applying the training rules to each sample presented from a predefined training set. Neural network adaptation always takes place in accordance with a training strategy. That is the network subjected to particular schedule to achieve desired end result. These training strategies can, at the most fundamental level, be divided into three categories: supervised training, reinforcement (or graded) training and self-organisation.

Reinforcement training is similar to supervised training except that, instead of being given the correct output at each individual training trial, the network receives only a grade that tells it how well it has done over a sequence of multiple training trails. Therefore the learner in reinforcement learning is not told which action to take, but instead must discover which actions yield the highest reward by trying them. Trial-and-error search is the most important distinguishing feature of reinforcement learning. However this results in a deficiency in reinforcement because of the phenomenon known as the conflict between exploitation and exploration, which is also known as identification and control (Narendra, Thathachar, 1989). There is always a conflict between the following two factors: The desire to use knowledge already available about the relative merits of actions taken by the system; and the desire to acquire more knowledge about the consequences of actions so as to make a better selection in the future.

Reinforcement learning based on only reward prevents exploration while supporting exploitation of the previously known best action. It suffers from getting stuck to the previously learned action, thereby loosing its ability to change. A solution to the problem is inclusion of punishment in the weight update rule so that it somehow supports exploration, however such a scheme prevents a total convergence to the best action even when the environment is stationary.

[1] *This work is being partially supported by Scientific and Technical Council of Turkey under grant EEEAG-126 Project: Modelling Cognitive Processes by Artificial Neural Networks*

Extinction is a well known property in animal learning, and it is extremely important for the survival of living organism since it provides adaptation to changing conditions. If an animal's behaviour is reinforced, and if the reinforcement is then terminated, for example by disconnecting the food dispensing mechanism, the animal will respond for a while and then gradually cease to respond. In other words, the behaviour extinguishes, but not immediately. Extinction refers to the decline of nonreinforced responses, (Hulse et al. 1980). The extinction of associations allows the system to learn new associations after forgetting ones that are no longer valid, and also prevents it from learning further associations based on those that are not valid. An attempt to model passive extinction in Neural Networks is provided in (Guloksuz and Halici, 1996) where an extent'on is made to the Recurrent Associative Dipole (READ) circuit proposed in (Grossberg and Schmajuk, 1987) which models classical conditioning concepts such as excitatory/inhibitory conditioning and opponent extinction.

In (Halici 1995) a reinforcement learning strategy is proposed to make a sequence of cascaded decisions to achieve a goal while aiming to optimise the total cost of the cascaded decisions. There, random neural networks (RNN), are used to model the system. The RNN model that represents more closely the manner in which signals are transmitted in a biophysical neural network where they travel as voltage spikes rather than as fixed signal levels, is introduced in (Gelenbe, 1989) and extended in (Gelenbe, 1990). A *backpropagation* type learning algorithm for Recurrent RNN model using gradient descent of quadratic error function is introduced in (Gelenbe, 1993)

In (Halici 1995) a reward based weight update rule along with a reinforcement function is provided .It is shown that the properties peculiar to RNNs are preserved after the application of the weight update rule. The performance of the learning strategy is analysed by applying it to the maze learning problem. The simulation results show that the performance of the system is highly dependent on the choice of reinforcement function and quite satisfactory results are obtained when the reinforcement function takes the recency effect into consideration. Although the weight update rule proposed there resembles the learning automata L_{R-I} scheme (Narendra and Thathachar, 1989), the way the reinforcement applied differentiates them. The learning scheme proposed in (Narendra and Thathachar, 1989) uses a form of *nonassociative* reinforcement learning which has the task of selecting a single optimal action rather than to associate different actions with different stimuli. Nonassociative reinforcement is therefore inadequate for problems where a sequence of actions affects the resultant reinforcement. The learning strategy proposed in (Halici, 1995) is a form of associative reinforcement learning, since it associates different actions with different levels of reinforcement, and therefore performe- well for the maze problem in which the sequence of actions affects the length of the path, which in turn affects the resultant reinforcement. The method presented in (Halici 1995) also differs from Q-learning (Watson, 1989) and Temporal Difference (Sutton 1984, Barto, Sutton and Watkins 1989) which are well known examples of associative learning. In these methods the learning system maintains estimates of reinforcements for all state-action pairs and makes use of these estimates to select actions. However, the learning system proposed in (Halici 1995) does not explicitly maintain estimates of the reinforcements for each state action pair. Such an estimate being implicit in the connection weights. An earlier version of the learning strategy, particular to the maze problem was proposed in (Halici and Yaranli, 1992).

Although the learning system proposed in (Halici 1995) performs well for the stationary environment, it suffers from getting stuck on to the previously learned action thereby losing its ability to adapt in changing conditions, because of the reinforcement based on only reward prevents exploration. In this paper we extend the reinforcement learning strategy to handle punishment by introducing a weight update rule preserving the properties peculiar to RNNs. Then we propose a learning scheme which takes the expectation of reinforcement into consideration such that it behaves as reward learning as long as the reward for the learned action is not below the expectation. Otherwise it behaves as punishment learning the learned action continues being not worse than expected, however it behaves as punishment learning so that other possibilities are explored.

Random Neural Network Model:

In the RNN model (Gelenbe, 1989), n neurons exchange *positive and negative* impulse signals. Each neuron accumulates signals as they arrive. If the total signal count at a given instant of time is positive, firing occurs at random according to an exponential distribution of constant rate, and signals are sent out to other neurons or outside the network. Each neuron i of the network is represented at time t by its input signal potential $k_i(t)$, constituted only by positive signals which have accumulated and which have not yet been cancelled by negative signals, and which have not yet

been sent out by the neuron as it fires. Positive signals represent excitation, while negative signals represent inhibition. A positive signal *adds 1* to the neuron potential, while a negative signal *reduces it by 1* or has no effect on the signal potential if it is already zero. This is a simplified representation of biophysical neural behaviour. In RNN, signals arrive to a neuron from outside the network (exogenous signals) or from other neurons. Each time a neuron fires, a signal leaves it, depleting its total input potential. A signal leaving neuron i heads for neuron j as a positive signal with probability $p^+(i,j)$, or as a negative signal with probability $p^-(i,j)$, or it departs from the network with probability $d(i)$.

$p(i,j) = p^+(i,j)+p^-(i,j)$ is the transition probability of a Markov chain representing the movement of signals between neurons satisfying

$$\Sigma_j\, p(i,j) + d(i) = 1 \text{ for } 1 \le i \le n \qquad (1)$$

where

$$0 \le p^+(i,j) \le 1 \text{ and } 0 \le p^-(i,j) \le 1 \qquad (2)$$

Positive and negative exogenous inputs to each neuron i of the network are provided by stationary Poisson processes of rate $\Lambda(i)$, and $\lambda(i)$, respectively. A neuron is capable of firing and emitting signals if its potential is strictly positive, and firing times are modelled by exponential neuron firing times with rate $r(i)$, at neuron i. In (Gelenbe, 1989) it is shown that the network's steady state probability distribution can be written as the product of the marginal probabilities of the state of each neuron. However the equations describing the steady state behaviour is out of the scope of this paper.

Reinforcement Learning for RNN with Reward/Punishment/ Expectation

Consider a learning system interacting with an environment such that it performs an action a_n at time steps n=0,1,2 selected from a finite set of actions. After each action a_n is taken, the learning system receives reinforcement R_n as a result of the interaction with the environment. The amount of the reinforcement is determined in some random manner depending on the action a_n. The can be reward R^+_n or punishment R^-_n The objective of reinforcement learning in such a system is to find a strategy for selecting the action that is maximises the expected reward and/or minimises the expected punishment.

In the following we first provide the learning strategy proposed in (Halici 95) for handling the rewarding case by using the RNN model. We then extend the strategy to cover punishment as well. Finally we propose a new learning scheme that takes into consideration the expectation of reward instead of only reward or punishment.

In Figure 1, a neuron, labeled i is used to represent the state just before the decision, this neuron has Ni outward connections, each representing a different decision. A neuron j is assigned to each possible decision to represent the states achieved by performing the decided action. Call the neuron i the initial node, and the neurons, j=1..N, achieved after each possible decision as the final nodes. The random neural network parameters except for the connection probabilities are set as follows:

r(k)=1 for any neuron in the system,
$\Lambda(k)=\Lambda$ if k is the initial node,
$\Lambda(k)=0$ otherwise
$\lambda(k)=0$ for any neuron
d(k)=1 if neuron k is a final node,
d(k)=0 otherwise (5)

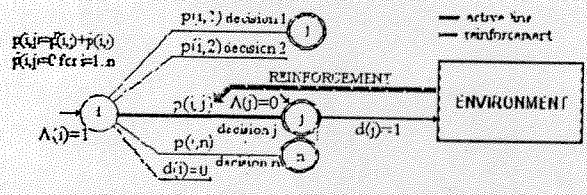


Figure 1: The system having a single decision step

In the network of Figure 1, only the initial node is connected to exogenous input, $\Lambda(i)=\Lambda$, for the other cells $\Lambda(i)=0$. Therefore all the pulses are originally created at the initial node. Since there is no inhibitory connection in the network, and d(i) is zero for this initial node any signal generated at the initial node reaches to one of the final node. Such a final node is determined in accordance with the transition probabilities p(i,j)=p+(i,j). Therefore the one among the final nodes having the largest connection probability p+(i,j) is more probable to receive the signal. When the final cell is reached, since d(i,j)=1, the pulse dissipates there by exciting the environment, which in turn produces reinforcement depending on the chosen action. We call, the travel of the pulse from the initial cell, where it is generated to the final cell, where it is dissipated, as a trial. During a trial, the network is assumed to be able to remember the activated connection in its short term memory. This active connection is important while training the network, since only the connection probabilities of the neurons on the active path are updated in training the network. Updating connection probabilities corresponds to storing information into long term memory.

Initially connections for all decisions are assigned equal probability, however as learning progress the

probabilities are updated according to the weight update rules given later on..

The rewarding case was examined previously in (Hal ici 1995) in which the following weight update rule is used:

$$p^+_{n+1}(i,j)= \begin{cases} p^+_n(i,k)+\eta^+R^+_n(k)(1-p^+_n(i,k)) & j=k \\ p^+_n(i,j)-\eta^+R^+_n(k)(p^+_n(i,j)) & j\neq k,\ j=1..N_i \end{cases} \quad (6)$$

where n is the trial number, R^+_n is the reward at trial n, η^+ is the learning rate for reward and N_i is the number of outward connections from i. When cascaded steps of decisions are to be used, the reward R^+_n in these equations becomes more complicated and it turns out to be $R^+_n(i)=\varphi(n,L_n,l_n(i))$ where n is the trial number, L_n is the total cost of the active decision path at trial n, $l_n(i)$ is the cost of the portion after decision at neuron i, φ is possibly a random valued function that varies inversely with L_n and $l_n(i)$. In our experiments on mazes, we observed better performance when the reinforcement function reflected the recency effect (Halici 1995). The learning strategy with reward is summarised below:

Learning Strategy with Reward
Initialisation:
set distribution of rewards for j=1..Ni,
$p^+_0(i,j)=1/N_i$ for j=1..Ni, n=1
Learning
1 select one of the connection, say k, with probability $p^+_n(i,k)$
2. apply weight update with reward $R^+_n(k)=R^+_n$
3. repeat learning with n+1 until probabilities are stabilised

In the following we propose the weight update rule to be used in the case of punishment:

$$p^+_{n+1}(i,j)= \quad (7)$$

$$\begin{cases} p^+_n(i,k)-\eta^-R^-_n(k)p^+_n(i,k)) & \text{for } j=k \\ p^+_n(i,j)+\eta^-R^-_n(k)(p^+_n(i,k)/(1-p^+_n(i,k))\,p^+_n(i,j) & \text{for } j\neq k,\ j=1..N_i \end{cases}$$

where R^-_n is the punishment at trial n and η^- is the learning rate for punishment. In the learning with punishment the learning step 2 is modified as follows:

Learning Strategy with Punishment
Initialisation:
set distribution of punishments for j=1..Ni,
set $p^+_0(i,j)=1/N_i$ for j=1..Ni, n=1
Learning
1.select one of the connection, say k, with probability $p^+_n(i,k)$
2. apply weight update with punishment $R^-_n(k)=R^-_n$
3. repeat learning with n+1 until, probabilities are stabilised

In equations (6) and (7) the reinforcement function may be assumed to be $0\leq R^+_n(i), R^-_n(i) \leq 1$ with no loss of generality. The symbol η in the formula is used to represent the learning speed and it should be chosen as small as possible to provide convergence to an optimum decision. On the other hand, the value of η is critical on the necessary number of trials to provide for convergence.

Theorem: If the properties of RNN are initially satisfied then they are not violated at any iteration for learning whenever $0\leq R^+_n(i), R^-_n(i) \leq 1$ *and* $0\leq \eta^+, \eta^- \leq 1$.

While the weight update rule for the rewarding case closely resembles the weight update rule proposed by (Bush and Mosteliert, 1958) and used in learning automata (Narendra and Thathachar) the weight update equation for punishment differs from the one used in learning automata of which the version adapted for RNN is given below:

$$p^+_{n+1}(i,j)= \quad (8)$$

$$\begin{cases} p^+_n(i,k)-\eta^-R^-_n(k)p^+_n(i,k)) & j=k \\ p^+_n(i,j)+\eta^-R^-_n(k)/(N-1_i)-\eta^-R^-_n(i)\,p^+_n(i,j) & j\neq k,\ j=1..N_i \end{cases}$$

Note that the weight update equation of learning automata (Eq. 8) may result in a decrease in some of the connection weights even though they are not selected and therefore not punished. Such an anomaly deficiency results in poor performance which is overcome by the weight update rule that we provide for RNN in Eq. 7.

To overcome the difficulties faced when the environment changes, we propose a learning strategy which take into consideration the expected reward and applies the weight update rule with punishment when the latest reward is worse than expected, otherwise applies the rule for the rewarding case. In the following we provide the algorithm where $\underline{R}^+_n(k)$ represents the expected reward R^+ when action k is selected at step n.

Learning Strategy with Expectation
Initialisation:
set distribution of reward for j=1..Ni
$p^+_0(i,j)=1/N_i$ for j=1..Ni, $\underline{R}^+_0(j)=0$
Learning
1 select one of the connection, say k, with probability $p^+_n(i,k)$
2. apply weight update such that if $R^+_n > \underline{R}^+_0(j)$ then apply weight update with reward $R^+_n(k)=R^+_n$ else apply weight update with punishment $R^-_n(k)=\underline{R}^+_n(k)-R^+_n$
3. update reward expectation as
$\underline{R}^+_n(k)=(1-\beta)\underline{R}^+_n(k)+\beta R^+_n \quad 0<\beta<<1 \quad (\beta=0.01)$
4. repeat learning with $n+1$ until probabilities are stabilised

The simulation results for learning with reward, punishment as proposed here, punishment as in learning automata, and expectation are provided in figure 2(a)-(d). Initially the rewards/punishments are 0.1, 0.4, 0.6, 0.9 for the final nodes 1-4 respectively, in the second phase the rewards for nodes 1 and 4 are swapped. Note that, the rewarding case is not able to handle extinction. The case with punishment able to handle extinction but there is not a total convergence to the action with minimum punishment, however it performs better than the punishment rule used for learning automata. The case with expectation, has a total convergence to the most rewarding action and furthermore the extinction is handled by the this learning strategy.

Acknowledgements

The author would like to thank Asli Guloksuz for her assistance in obtaining simulation results for the learning strategies proposed in this paper.

REFERENCES

Barto, A. Sutton R. and Watkins C.J., 1989, Learning and Sequential Decision Making, COINS Technical Report.

Bush, R.R. and F. Mosteller, Stochastic Models of Learning, John Wiley, 1958

Carlson N.R., 1977, Physiology of Behaviour, Allyn and Bacon

Gelenbe, E., 1989, Random Neural Networks with Negative and Positive Signals and Product form Solution, *Neural Computation*, 1, 502-510

Gelenbe, E., 1990, Stability of the Random Neural Network Model, *Neural Computation*, 2, No.2, 239,

Gelenbe E., 1993, Learning in the Recurrent Random Neural Network, *Neural Computation*, 5, 154

Guloksuz A. and Halici U., 1996, A Neural Circuit to Handle Passive Extinction in Conditioned Reinforcement Learning, Proceedings of 13th European Meeting an Cybernetics and System Res., April 9-12

Halici, U., 1995, Reinforcement Learning in Random Neural Networks for Mazes, International Workshop on Neuronal Coding, Prague, Czech Republic September 11-14

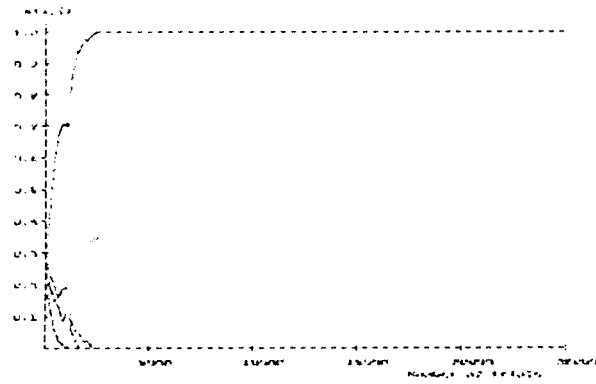
Halici U., Yaranli U., 1992, Neural Networks in Mazes, IEEE-INNS International Joint Conference on Neural Networks, Beijing, China, November, 1992, Vol-II, pp 711-716

Hulse H. S., H. Egeth and J. Deese, 1980, *The Psychology of Learning*, 5th Ed., Mc Graw Hill,

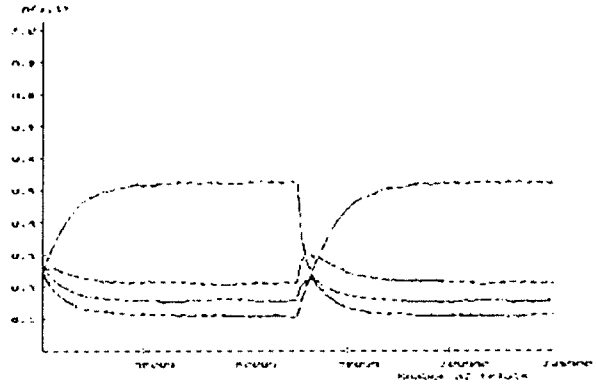
Narendra K., Thathachar M.A.L., 1989, Learning Automata: An Introduction , Prentice Hall, Englewood Cliffs

Sutton, R.S., 1984, 'Temporal credit assignment in reinforcement learning', Doctoral Dissertation, University of Massachussets, Amherst

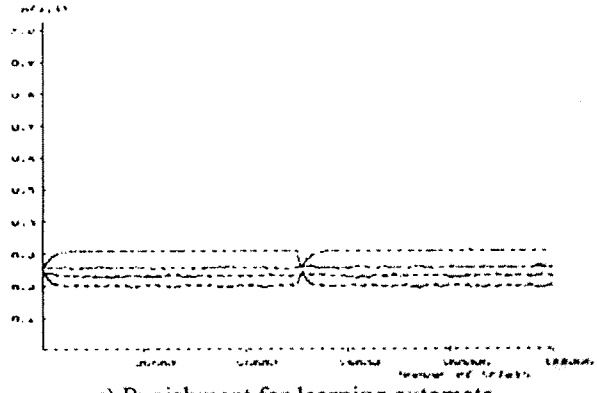
Sutton R.S. 1988, Learning to predict the methods of temporal difference, Machine Learning, 3:9-44



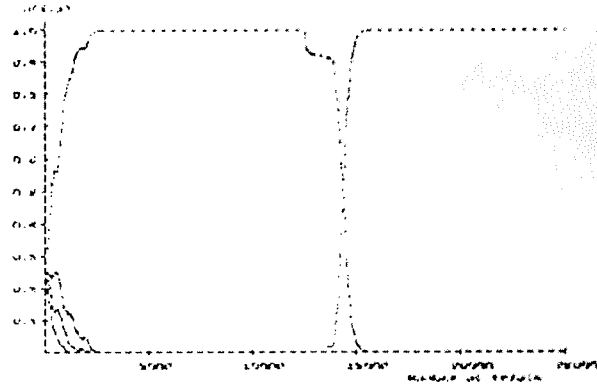
a) Reward



b) Punishment



c) Punishment for learning automata



d) Expectation

Figure 2. Evolution of connection probabilities during learning and extinction phases for various weight update strategies

Symposium on Neuronal Coding, Prague, September 1995

ABSTRACT

REINFORCEMENT LEARNING IN RANDOM NEURAL NETWORKS FOR MAZES

Ugur HALICI
Department of Electrical and Electronics Engineering,
06531, METU, Ankara, Turkey,
email: halici@rorqual.cc.metu.edu.tr

Abstract: The Random Neural Network model, where signals travel as voltage spikes rather than as fixed signal levels, represents more closely the manner in which signals are transmitted in biophysical neural networks. Until now several technical applications are developed on Random Neural and a learning algorithm for recurrent random network model that uses gradient descent of a quadratic error function is provided in the literature. In this paper a reinforcement learning strategy is proposed to find out a sequence of cascaded decisions to achieve a goal while aiming to optimize the total cost of the cascaded decisions. The purpose Random Neural Networks are used to model the system and a weight update rule together with a reinforcement function is provided. It is proved that the properties special to Random Neural Network are preserved after the application of the weight update rule.Furthermore, the performance of the learning strategy is analysed by applying it on maze learning problem. The experimental results are showed that the performance of the system is highly dependent on the chosen reinforcement function and quite satisfactory results are obtained when the reinforcement function takes the recency effect into consideration.

Keywords: random neural networks, reinforcement learning, mazes, cognitive map, recency effect,

[1] This work is being partially supported by Scientific and Technical Council of Turkey under grant EEEAG-126 Project: Modelling Cognitive Processes by Artificial Neural Networks

ARTIFICIAL NEURAL NETWORKS AND ARTIFICIAL LIFE SYMPOSIUM
MIDDLE EAST TECHNICAL UNIVERSITY ANKARA TURKEY 15/12/1994

HINTS FROM LIFE TO ARTIFICIAL INTELLIGENCE

Edited by
Uğur HALICI

METU • HALICI Co. • TÜBITAK • INNS Turkey SIG • IEEE CS TC • CAGRI

Hints from Life to AI, edited by Uğur Halıcı, METU, 1994

artificial neural circuits for conditioned learning[1]

Aslı Gülöksüz
Uğur Halıcı
Dept.of Electrical and Electronics Eng.
Middle East Technical University,
06531, Ankara, TURKEY
{guloksuz,halici}@rorqual.cc.metu.edu.tr

The concepts of classical conditioning can be used in designing neural networks for association and expectation learning, and behavioural conditioning in artificial systems. Classical conditioning concepts such as excitatory conditioning, inhibitory conditioning, secondary conditioning, opponent extinction and habituation have been modelled by the Recurrent Associative Dipole (READ). However, this circuit does not satisfy the experimental data on extinction in case of the nonoccurence of an expected event. In this work, a brief overview of the basic concepts in conditioned learning is made, the operation of the READ circuit is depicted, and the circuit is modified so that it will model extinction as well. The changes in the performance of the READ circuit introduced by this modification are then explored.

1. Introduction

An intelligent system using neural networks would include, among other things, networks for visual perception, pattern learning and object recognition, association learning, expectation learning, emotional states and behavioural actions [1]. Such neural networks can be incorporated in a robot, or in adaptive systems in any practical field.

[1] This work is being supported by TÜBİTAK under grant EEEAG-126, Project: Modelling Cognitive Processes by Artificial Neural Networks.

The scope of this paper is only the parts related to association and expectation learning, and behavioural conditioning. Therefore, it is assumed that the objects that act as stimuli in these types of learning have already been recognized, and are presented to the network at the conceptual level rather than as patterns.

The neural network models depicted here aim at using the concepts of classical conditioning in animals, to design neural networks that can learn associations between objects (stimuli) and between objects and responses. As some associations between stimuli and responses seem to be inborn in animals; such as the association between the sight of food and salivation for a dog; some vital stimulus-response pairs can be hard-wired or initially set by software in artificial systems. Then, these initial associations can be used to form new associations between stimuli and between stimuli and responses. Some classical conditioning concepts will be defined in Section 2.

The Recurrent Associative Dipole (READ) [3] is a neural circuit that models some of the classical conditioning concepts that will be described in Section 2. A number of these circuits are used as part of the cognitive system of a mobile robot called MAVIN [1] developed at the Lincoln Laboratory at the MIT. The operation of the READ circuit will be summarized in Section 3.

The operation of the READ circuit doesn't conform with the psychological data on extinction in the case of unconfirmed expectations. However, the extinction of associations allows the system to learn new associations after forgetting ones that are no longer valid, and also prevents it from learning further associations based on those that are not valid. We have made a to the READ circuit and to the differential equations defining it, in order to handle the concept of extinction. In Section 4, these modifications are described, and simulation results of the READ circuit and its modified version are given.

2. Conditioned Learning

In this section, some classical conditioning phenomena that are aimed to be modelled are described.

2. 1. Classical Conditioning

Classical or Pavlovian conditioning is the type of learning in Pavlov's well known experiment in which a dog is repeatedly presented with the sound of a tuning fork before being given food, and learns to salivate at the sound of the tuning fork alone [2]. In this section, the basic concepts of conditioned learning will be discussed.

The fundamental components of classical conditioning are the following:

- **US:** The unconditioned stimulus which triggers a response without prior training. In Pavlov's experiment, the US is food.

- **UR:** The unconditioned response triggered by the US; in Pavlov's experiment, the dog's salivation.

- **CS:** The conditioned stimulus which comes to trigger a response by being repeatedly paired with the US; in Pavlov's experiment, the sound of a tuning fork.

- **CR:** The conditioned response which arises at the occurrence of the CS, after the CS has been repeatedly paired with the US. The CR is not necessarily identical to the UR; it may vary in amplitude and latency. In Pavlov's experiment, the CR is salivation at the sound of the tuning fork alone.

The occurrence of the CR indicates that an association has been formed between the CS and the US. This *association learning* is meaningful in itself, regardless of the relationship between the UR and the CR [2].

2.2. Concepts in Classical Conditioning

In this subsection, some basic concepts of classical or Pavlovian conditioning are be defined.

2.2.1. Excitatory and Inhibitory Conditioning

The type of conditioning in which the CS is conditioned to the response related to the US, as in Pavlov's experiment, is called *excitatory conditioning*. In this type, firstly the CS is presented, and then the US is presented in the presence of the CS. This pairing is repeated a number of times.

In *inhibitory conditioning*, the CS is repeatedly presented after the offset of the US, and is thus conditioned to the response related to this offset.

The schedules for excitatory and inhibitory conditioning are demonstrated in Figures 1(a) and (b) respectively.

2.2.2. Primary and Secondary Conditioning

The conditioning of a CS by repeated pairing with a US is called primary conditioning. Once a conditioned stimulus CS_1 has been conditioned by using a US, it can in turn be used in conditioning another stimulus CS_2. In Pavlov's

experiment, for instance, after the sound has been paired with food for a sufficient number of times, repeatedly pairing a light with the sound will cause the dog to salivate at the sight of the light, even when food is not used as an US. This phenomenon is called *secondary conditioning*.

Schematic diagrams of *secondary excitatory conditioning* and *secondary inhibitory conditioning* are shown in Figure 1(c) and (d) respectively.

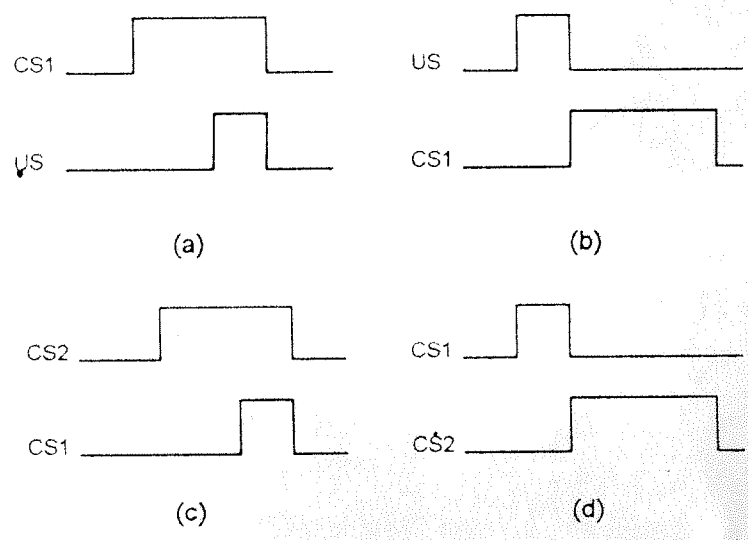


Figure 1: Some classical conditioning schedules (a)primary excitatory conditioning (b)primary inhibitory conditioning (c)secondary excitatory conditioning (d)secondary inhibitory conditioning

2.2.3. Blocking and Overshadowing

Assume that a stimulus CS_1 is repeatedly paired with a US until conditioning takes place; and then CS_1 and a neutral unconditioned stimulus CS_2 are presented together and paired with the US. When CS_2 is presented alone afterwards, it does not result in the CR. This phenomenon of a previously conditioned stimulus preventing the conditioning of a neutral stimulus is called *blocking*.The experimental stages of blocking are as below [2]:

1. $CS_1 \Rightarrow US$
   $CS_1 \Rightarrow CR$
2. $CS_1+CS_2 \Rightarrow US$
   $CS_1+CS_2 \Rightarrow CR$
3. $CS_2 \Rightarrow CR$

Assume that stimuli CS_1 and CS_2 which are initially both neutral are presented together and paired with US. It may be the case that conditioning occurs for the compound stimulus CS_1+CS_2 and for one of the stimuli; say CS_1, but does not occur for CS_2. This situation may arise because CS_1 is a more salient stimulus and therefore receives more attention than CS_2. This phenomenon is called *overshadowing*. Overshadowing can be demonstrated by the following stages [2]:

1. $CS_1+CS_2 \Rightarrow US$
   $CS_1+CS2 \Rightarrow CR$

2. $CS_2 \Rightarrow CR$

2.2.4. Timing Considerations in Conditioning

For conditioning to take place, the CS must be presented before the US, as shown in Figure 1(a). Otherwise, the US will attract more attention because of its relevance to drives and responses in the system, and will in a way overshadow the CS.

The time interval between the presentation of the CS and that of the US is called the *interstimulus interval ISI*. The synchronization problem in conditioning arises from the necessity that, although the ISI varies for the various repetitions of an experiment, the CS should become associated only with the US and not with a mixture of the US and the noise in the environment.

2.2.5. Extinction

In animals, learned responses are dropped if they are not reinforced [2]. For example, if after being conditioned with the tune-food pair, the dog in Pavlov's experiment is repeatedly presented with a tune but no food appears, it will start to salivate less and less, and eventually it will not salivate at all in reponse to the tune. This phenomenon is called *extinction*. In Figure 2, a rough curve representing the cumulative number of responses during the extinction process is given. This curve has been obtained by using the well known Skinner box, in which there is a rat or a pigeon in a box, receives a piece of food each time it presses a lever. If no more food is dropped after the lever press, the rat gives this response less and less frequently until it doesn't press it at all. The curve in Figure 2 is the number of total lever presses in the course of an hour after the reward of food is removed [2].

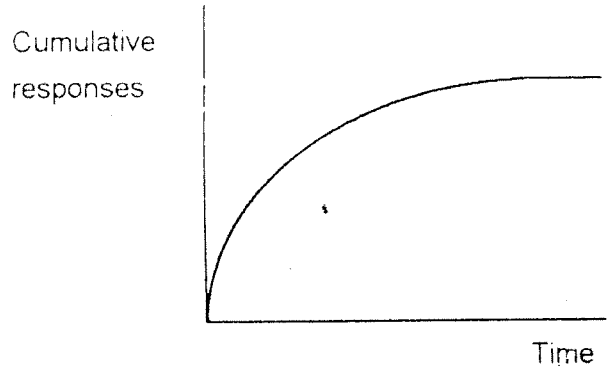


Figure 2: Rough extinction curve. Total number of responses versus time during the period in which the CS is not followed by the US, i.e., the expected event does not occur.

3. READ:A Neural Network that Models Conditioned Learning

In this section, the Recurrent Associative Dipole [3] that models most of the concepts defined in Section 1 will be depicted, and the simulation results for excitatory and inhibitory conditioning will be given.

3.1. The Recurrent Associative Dipole (READ)

The Recurrent Associative Dipole is a neural circuit that consists of two channels: one related to the *on-response* of a particular stimulus US, and the other to the *off-reponse* of the US in question. The on-channel and off-channel of the READ circuit in Figure 3 are the columns on the left and the right respectively.

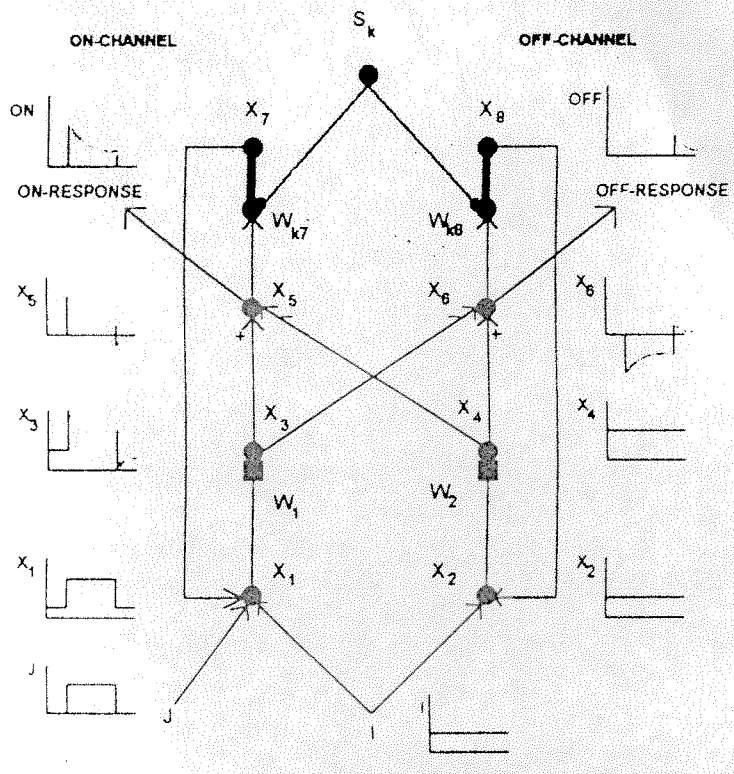


Figure 3: The Recurrent Associative Dipole (READ) circuit and the response of each node to the J input shown on the bottom left

Both channels have modifiable synapses with nodes pertaining to other stimuli (CS's). The READ circuit models primary and secondary excitatory and inhibitory conditioning, opponent extinction and habituation. The differential equations defining the operation of the circuit are given below:

Arousal + US + Feedback On-Activation:

$$\frac{dx_1}{dt} = -ax_1 + I + J + f(x_7) \quad (1)$$

I:Arousal input J:Input to the On-channel (US)

Arousal + Feedback Off-Activation:

$$\frac{dx_2}{dt} = -ax_2 + I + f(x_8) \quad (2)$$

Depletable On and Off Transmitters:

$$\frac{dw_1}{dt} = b(1-w_1) - cg(x_1)w_1 \quad (3)$$

$$\frac{dw_2}{dt} = b(1-w_2) - cg(x_2)w_2 \quad (4)$$

Gated On and Off Activations:

$$\frac{dx_3}{dt} = -ax_3 + eg(x_1)w_1 \quad (5)$$

$$\frac{dx_4}{dt} = -ax_4 + eg(x_2)w_2 \quad (6)$$

Normalized Opponent On and Off Activations:

$$\frac{dx_5}{dt} = -ax_5 + (h-x_5)x_3 - (x_5+k)x_4 \quad (7)$$

$$\frac{dx_6}{dt} = -ax_6 + (h-x_6)x_4 - (x_6+k)x_3 \quad (8)$$

Total On and Off Activations:

$$\frac{dx_7}{dt} = -ax_7 + m[x_5]^+ - p\Sigma S_k w_{k7} \quad (9)$$

$$\frac{dx_8}{dt} = -ax_8 + m[x_6]^+ - p\Sigma S_k w_{k8} \quad (10)$$

On -conditioned and Off-conditioned Reinforcer Learning:

$$\frac{dw_{k7}}{dt} = S_k(-qw_{k7} + r[x_5]^+) \quad (11)$$

$$\frac{dw_{k8}}{dt} = S_k(-qw_{k8} + r[x_6]^+) \quad (12)$$

On and Off Responses:

$$ON = [x_5]^+ \quad (13)$$

$$OFF = [x_6]^+ \quad (14)$$

3.2. Primary and Secondary Excitatory Conditioning of the READ Circuit

Primary excitatory conditioning takes place when a CS is presented at a node S_k, and the US pertaining to the READ circuit is presented at X_1 in the order shown in

Figure 1(a), causing an increase in the weight w_{k7} according to equation 11. The squares adjacent to the nodes x_3 and x_4 indicate that these have synapses with habituating and recovering transmitters. For example, w_1 decreases when x_1 is active and recovers when it is not, according to equation 3. This causes the activations x_3 and x_5 to decay as shown in Figure 3. Since the bias input I is equal for both channels, the habituation of w_1 results in a rebound in the off-channel after the offset of the US. If the CS is presented during this rebound, the weight w_{k8} increases according to equation 12, and the CS is conditioned to the off-response, i.e., primary inhibitory conditioning takes place.

After conditioning, the feedback path from $x_7(x_8)$ to $x_1(x_2)$ allows the CS alone to activate $x_1(x_2)$ and thus generate the on-response(off-response). This also makes secondary conditioning possible.

3.3. Extinction in the READ Circuit

In the READ circuit, once excitatory conditioning has taken place, the weight w_{k7} does not decay even if the US never arrives after the CS; i.e., if the expectation that the US will arrive is not confirmed. This can be noticed by observing equation 11: in order for w_{k7} to decay, the CS must be active while x_5 is not. However, after conditioning, the CS alone is sufficient to activate the on-channel. Therefore this decay never takes place. However, psychological data suggests that the repeated nonoccurrence of the US after the CS results in the extinction of conditioned learning [2]. In the next part, this phenomenon is examined in some more detail, and a modification is made in the READ circuit to support this phenomenon.

4. A Modification of the READ Circuit to Model Extinction

The READ circuit supports extinction in neither excitatory nor inhibitory conditioning. Inhibitory conditioning, as defined in Part I, does not involve expectation learning. Therefore one cannot talk about the nonoccurrence of an expected event, and extinction is not supposed to occur.

As explained in the previous section, the reason for extinction not to take place in the READ circuit is that, after conditioning, CS activates the on-channel in exactly the same way that the US does (only with a slightly smaller activation.) Therefore it is not possible to differentiate between the US and a previously conditioned CS. This suggests that another level of neurons is needed to make this differentiation.

A modified version of the READ circuit is given in Figure 6, in which node X_{e1} provides the possibility to make this differentiation. Node X_{e2} has been added solely to preserve the symmetry of the circuit.

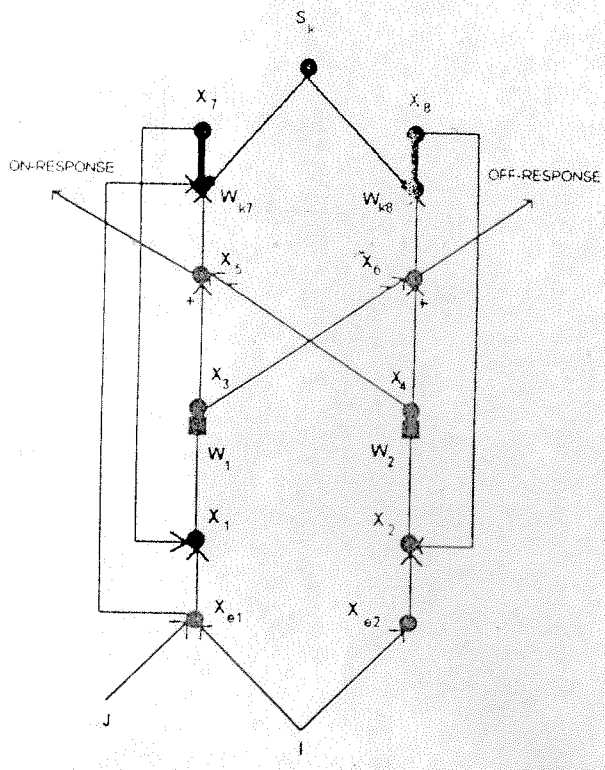


Figure 4: Modified READ circuit that handles extinction

The additional and modified differential equations defining the operation of th modified circuit are given below:

New nodes:

$$\frac{dx_{e1}}{dt} = -ax_{e1} + I + J \tag{15}$$

$$\frac{dx_{e2}}{dt} = -ax_{e2} + I \tag{16}$$

Modified equations:

$$\frac{dx_1}{dt} = -ax_1 + x_{e1} + f(x_7) \tag{17}$$

$$\frac{dx_2}{dt} = -ax_2 + xe_2 + f(x_8) \tag{18}$$

$$\frac{dw_{k7}}{dt} = S_k(-qw_{k7} + r[x_5 q(x_{e1} - I)]^+) \tag{19}$$

As can be observed from the modified equations for the circuit given below, during the conditioning phase, the operation of the circuit is identical to that of the original READ circuit. However, W_{k7} decays each time the activation of X_{e1} is below a threshold slightly higher than the bias input; i.e., W_{k7} decays when the CS

is present and the US is not. This results in extinction in case of the nonoccurrence of the expected US.

4.4. Experimental Results

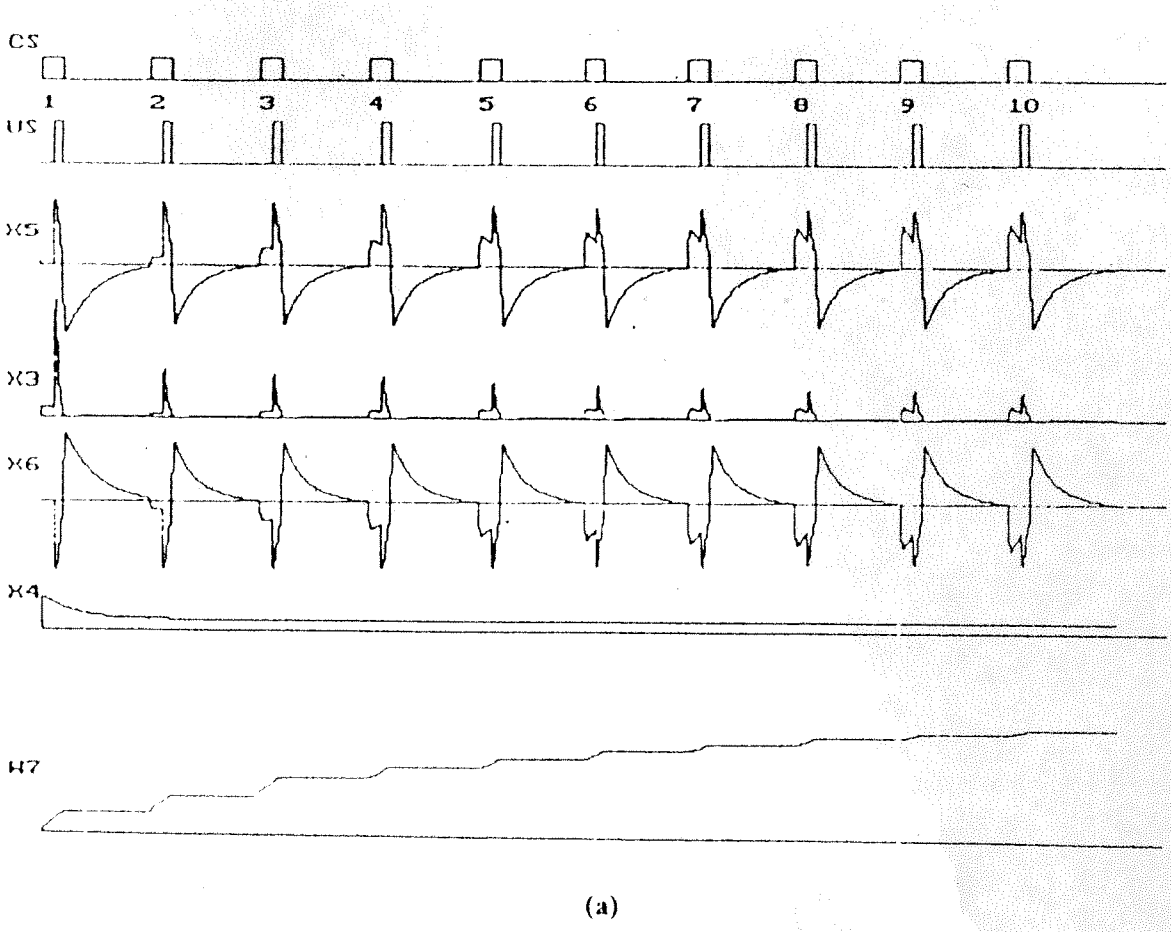
Figure 5 contains the simulation results of primary excitatory conditioning with the READ circuit in Figure 3. The on-response and off-response are identical to the positive portions of X_5 and X_6, and are therefore not separately plotted.

In part (a), 10 trials are made in which the CS is presented for 200 time units, and the US is also presented in the last 40 of these. One can observe the growth in the response of X_5 at the presentation of CS at each trial. Part (b) of the same figure shows the 10 following trials during which only the CS is presented. At each of the 10 trials the on-response has the same amplitude; in fact, there is no change in W_7 either.

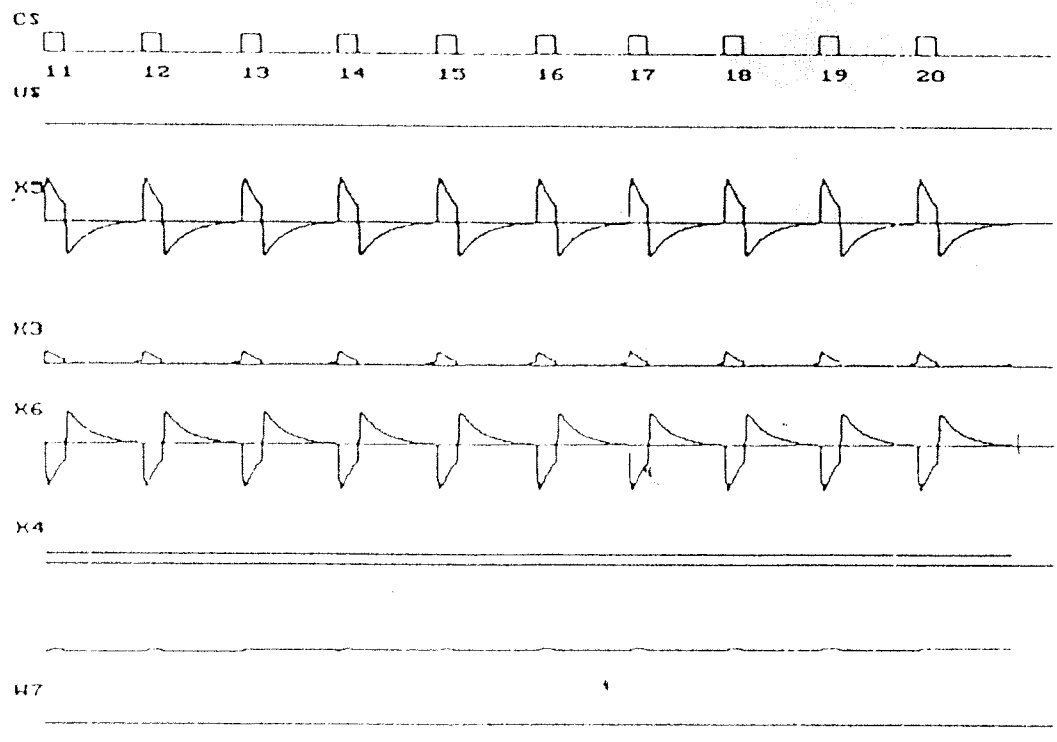
Figure 6 contains the results of the same experiment performed by using the modified READ circuit in Figure 4. During the conditioning phase, the circuit experiences the nonoccurrence of an expectation at the beginning of each presentation of CS, and therefore W_7 decays slightly, but when US is presented, the weight recovers from this decay. This slows down the learning process slightly, but not significantly. At the end of the learning phase, the on-response generated by the CS alone decreases each time The US does not arrive, and approaches zero at the end of 10 trials. Thus extinction takes place.

One phenomenon that must be mentioned here is secondary conditioning. The effect of the modifications on secondary conditioning will be as follows: Assume that CS_1 has been conditioned by using US. When CS_1 and CS_2 are then presented as in Figure 1(c), secondary conditioning will take place to a lesser and lesser degree each time CS_1 is not followed by US. Then, the extinction of the weight corresponding to CS_2 will be dependent on the occurrence of US and not of CS_1. This is a desirable property, since, for instance in the case of Pavlov's experiment, the useful expectation to be learned is that the light will be followed by food and not that it will be followed by the sound of a tune fork.

Therefore this circuit implements secondary conditioning in a much more practical way than the original READ circuit, since it results in the dog learning to salivate by using the light-tune pair as long as the light is followed by the sound of the tune which is mostly followed by food.

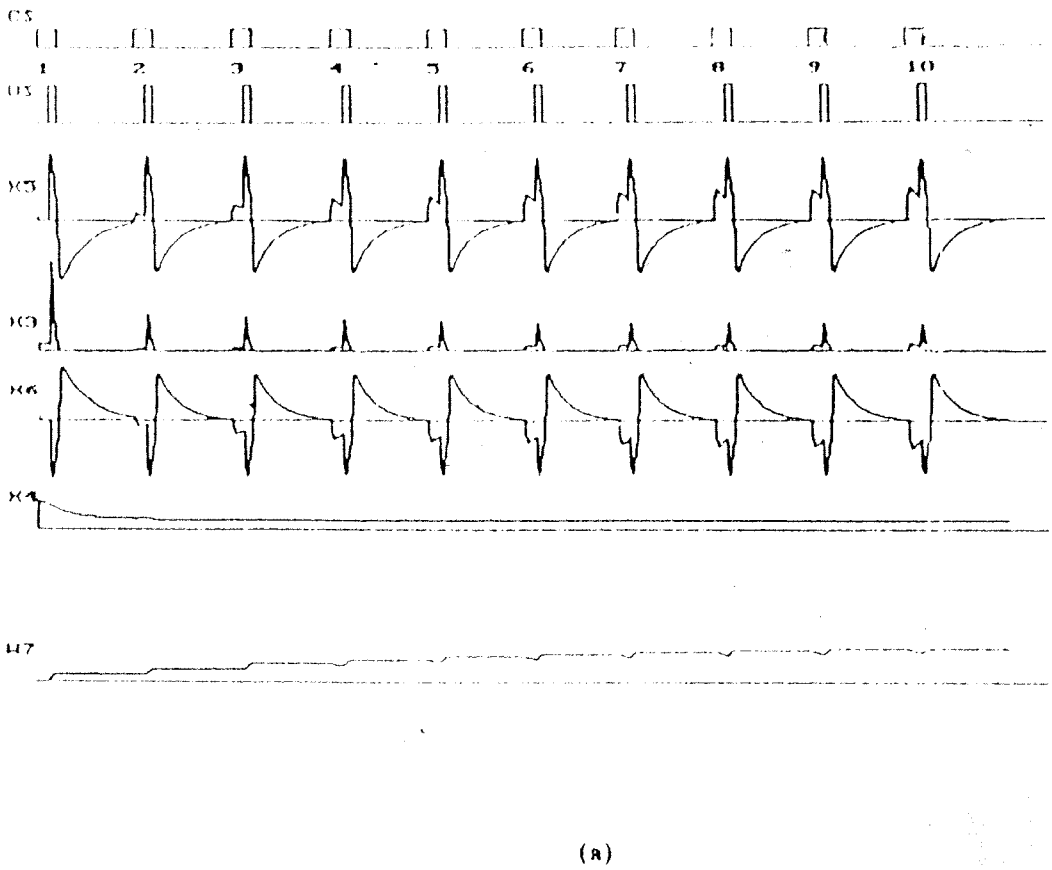


(a)

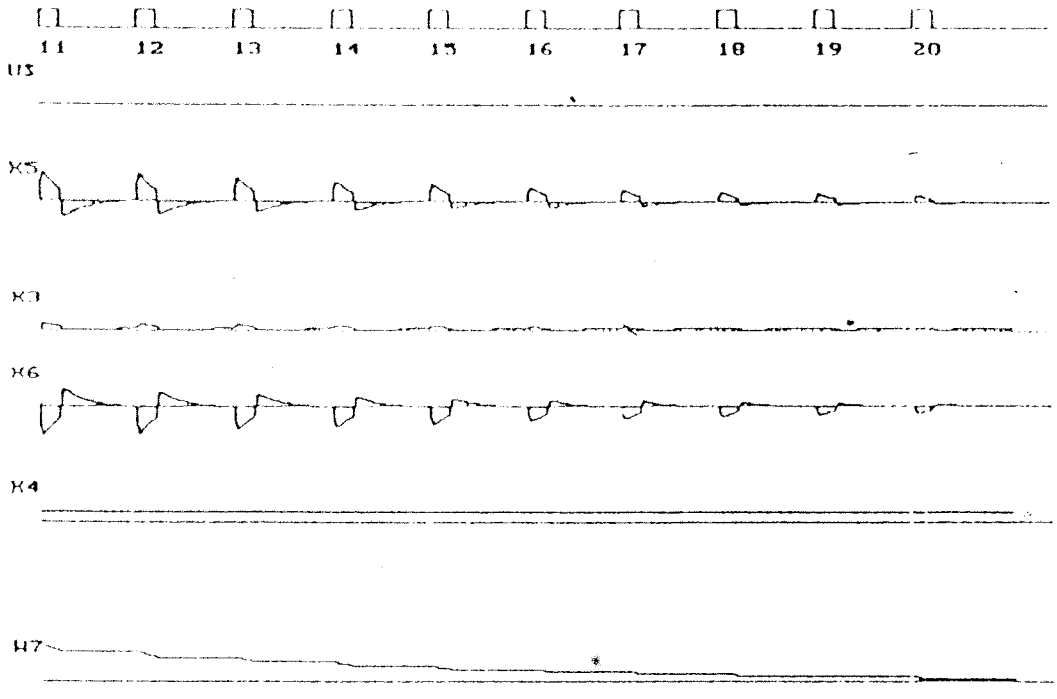


(b)

Figure 5: Simulation results for primary excitatory conditioning with the READ circuit in Figure 2. (a) In the conditioning phase the US is paired with the CS (b) In the second phase, presenting the CS alone results in an on-response from the READ circuit



(a)



(b)

Figure 6: Simulation results for primary excitatory conditioning with the modified READ circuit (a)In the conditioning phase the US is paired with the CS (b)In the extinction phase the CS is presented alone

5. Conclusion

Extinction is an important property in animal learning, since it increases the adaptive capabilities of the animal by allowing it to drop responses that are not useful any more. The modification made to the READ circuit in this paper handles extinction as well as all the phenomena modelled by the READ circuit.

While the modified circuit handles extinction, there is another psychological phenomenon that must be modelled; namely the phenomenon of *spontaneous recovery* [2]. In experiments performed with animals, if time elapses after an extinction session like the one in Figure 6(b), the strength of the conditioned response is found to have recovered. The amount of recovery increases with the time interval between extinction sessions. However, if more extinction sessions are carried out, spontaneous recovery decreases and finally dissapears. Modelling spontaneous recovery is the next goal on this subject.

References

1. Baloch A.A. and Waxman A.M., "Visual learning, adaptive expectations, and behavioural conditioning of the mobile robot MAVIN", *Neuarl Networks*, Vol.4, pp.271-302, 1991
2. Hulse, H.H., Egeth,H., and Deese,J., *The Psychology of Learning*, McGraw-Hill,1980
3. Grossberg, S. and Schmajuk, N.A., "Neural dynamics of attentionally modulated Pavlovian condtioning: conditioned reinforcement, inhibition, and opponent processing." *Psychobiology* 15, pp.95-240
4. Grossberg,S., "A neural network architecture for Pavlovian conditioning: reinforcement, attention, forgetting, timing ", in *Neural Network Models of Conditioning and Action*, ed. Commons,M.L., Grossberg, S., Staddon, J.E.R., Lawrence Erlbaum Assoc., 1991
5. Levine,D.S., *Neural and Cognitive Modelling*, Lawrence Erlbaum Assoc., 1991

BİBLİYOGRAFİK BİLGİ FORMU	
1. Proje No: EEEAG-126	**2. Rapor Tarihi:** 4.2.1997
3. Projenin Başlangıç ve Bitiş Tarihleri: 1.11.1994-1.11.1996	
4. Projenin Adı: Bilişsel Süreçlerin Yapay Sinir Ağları ile Modellenmesi	
5. Proje Yürütücüsü: Prof. Dr. Uğur Halıcı **Yardımcı Araştırıcılar:** Araş. Gör. Aslı Gülöksüz, Prof. Dr. Umur Talaslı	
6. Projenin Yürütüldüğü Kuruluş ve Adresi: Elektrik ve Elektronik Mühendisliği Bölümü, Orta Doğu Teknik Üniversitesi 06531, Ankara	
7. Destekleyen Kuruluş(ların) Adı ve Adresi: TÜBİTAK , ODTÜ	

8. Öz: Bilişssel süreçlerin Yapay Sinir Ağları ile modellenmesini amaçlayan bu projede canlılardaki Çağrışımsal öğrenmenin iki temel çeşidi olan Koşullu Öğrenme ve Pekiştirimli Öğrenme süreçleri incelenmiştir. Koşullu öğrenmeyi modellemek üzere kullanılmakta olan Geri Döngülü Çağrışımsal Dipol (READ) devresi üzerinde öğrenmenin sönümünü modellemek üzere değişiklikler yapılarak devrenin birincil ve ikincil koşullandırma altında çalışması gözlenmiş, ayrıca birden fazla READ birimin birarada çalışması incelenmiştir. Pekiştirimli öğrenmenin modellenmesi amacıyla Rassal Sinir Ağları (RSA) kullanılmıştır. Tekli karar adımları ve zincirleme karar adımları için pekiştirimli öğrenme stratejileri önerilmiştir. Ödül, ceza ve ödül beklentisinin göz önüne alındığı durumlar için öğrenme kuralları geliştirilerek öğrenmenin sönümü ve sistemin değişen çevre koşullarına uyumu incelenmiştir.

Anahtar Kelimeler: Bilişsel Süreçler, Yapay Sinir Ağları, Yapay Zeka, Koşullu Öğrenme, Pekiştirimli Öğrenme, Öğrenmenin Sönümü, Çağrışımsal Geridöngülü Dipol (READ), Rassal Sinir Ağları

Abstract: In this project that aims to model cognitive processes by using Artificial Neural Networks, the two basic paradigm of associative learning, which are the Conditioned Learning and the Reinforcement Learning are examined. The Recurrent Associative Dipole (READ) circuit which is used for modelling Conditioned Learning is modified to handle extinction and its operation is observed for primary and secondary conditioning. Furthermore the operation of multiple READ units together is examined. For modelling Reinforcement learning, the Random Neural Networks (RNN) are used. Learning strategies are proposed for single and cascaded decision steps. The learning rules for cases of reward, punishment and expectation of reward are developed. Adaptation of the system to changing environmental conditions and extinction of learning in the system are examined.

Keywords: Cognitive Processes, Artificial Neural Networks, Artificial Intelligence, Conditioned Learning, Reinforcement Learning, Extinction, Recurrent Associative Dipole (READ), Random Neural Networks

9. Proje ile ilgili Yayın/Tebliğlerle ilgili Bilgiler:

Halıcı U., "Reinforcement Learning in Random Neural Networks for Cascaded Decisions, *Journal of Biosystems,* Elsevier, Vol 40 No 1,2, January 1997 pp 83-91

Gulöksüz A., Halıcı U., 1996, "A Neural Circuit to Handle Passive Extinction in Conditioned Reinforcement Learning", *Proceedings of Thirteenth European Meeting on Cybernetics and System Research*, Vienn, Austria

Halýcý U., 1996, (Extended Abstract) Reward, Punishment and Expectation in Reinforcement Learning for random Neural Networks, *Workshop on Biologically Inspired Autonomous Systems: Computation, Cognition and Control,* Duke University, Durham, North Caroline, USA, 4-5 March,

Halýcý U., 1995, (Abstract) "Reinforcement Learning in Random Neural Networks for Mazes", *Symposium on Neuronal Coding,* Prague, September

Gülöksüz A., Halýcý U., 1994, "Artificial Neural Networks for Conditioned Learning", *in Hints From Life to Artificial Intelligence,* Editor: U. Halýcý, METU, December

10. Bilim Dalı: Bilgisayar **Doçentlik B. Dalı Kodu:** Akıllı Sistemler **Uzmanlık Alanı Kodu:**		**ISIC Kodu:**
11. Dağıtım:	☐ Sınırlı	☑ Sınırsız
12. Raporun Gizlilik durumu:	☐ Gizli	☑Gizli Değil

BİBLİYOGRAFİK BİLGİ FORMU	
1. Proje No: EEEAG-126	**2. Rapor Tarihi:** 4.2.1997
3. Projenin Başlangıç ve Bitiş Tarihleri: 1.11.1994-1.11.1996	
4. Projenin Adı: Bilişsel Süreçlerin Yapay Sinir Ağları ile Modellenmesi	
5. Proje Yürütücüsü: Prof. Dr. Uğur Halıcı **Yardımcı Araştırıcılar:** Araş. Gör. Aslı Gülöksüz, Prof. Dr. Umur Talaslı	
6. Projenin Yürütüldüğü Kuruluş ve Adresi: Elektrik ve Elektronik Mühendisliği Bölümü, Orta Doğu Teknik Üniversitesi 06531, Ankara	
7. Destekleyen Kuruluş(ların) Adı ve Adresi: TÜBİTAK , ODTÜ	

8. Öz: Bilişsel süreçlerin Yapay Sinir Ağları ile modellenmesini amaçlayan bu projede canlılardaki Çağrışımsal öğrenmenin iki temel çeşidi olan Koşullu Öğrenme ve Pekiştirimli Öğrenme süreçleri incelenmiştir. Koşullu öğrenmeyi modellemek üzere kullanılmakta olan Geri Döngülü Çağrışımsal Dipol (READ) devresi üzerinde öğrenmenin sönümünü modellemek üzere değişiklikler yapılarak devrenin birincil ve ikincil koşullandırma altında çalışması gözlenmiş, ayrıca birden fazla READ birimin birarada çalışması incelenmiştir. Pekiştirimli öğrenmenin modellenmesi amacıyla Rassal Sinir Ağları (RSA) kullanılmıştır. Tekli karar adımları ve zincirleme karar adımları için pekiştirimli öğrenme stratejileri önerilmiştir. Ödül, ceza ve ödül beklentisinin göz önüne alındığı durumlar için öğrenme kuralları geliştirilerek öğrenmenin sönümü ve sistemin değişen çevre koşullarına uyumu incelenmiştir.

Anahtar Kelimeler: Bilişsel Süreçler, Yapay Sinir Ağları, Yapay Zeka, Koşullu Öğrenme, Pekiştirimli Öğrenme, Öğrenmenin Sönümü, Çağrışımsal Geridöngülü Dipol (READ), Rassal Sinir Ağları

Abstract: In this project that aims to model cognitive processes by using Artificial Neural Networks, the two basic paradigm of associative learning, which are the Conditioned Learning and the Reinforcement Learning are examined. The Recurrent Associative Dipole (READ) circuit which is used for modelling Conditioned Learning is modified to handle extinction and its operation is observed for primary and secondary conditioning. Furthermore the operation of multiple READ units together is examined. For modelling Reinforcement learning, the Random Neural Networks (RNN) are used. Learning strategies are proposed for single and cascaded decision steps. The learning rules for cases of reward, punishment and expectation of reward are developed. Adaptation of the system to changing environmental conditions and extinction of learning in the system are examined.

Keywords: Cognitive Processes, Artificial Neural Networks, Artificial Intelligence, Conditioned Learning, Reinforcement Learning, Extinction, Recurrent Associative Dipole (READ), Random Neural Networks

9. Proje ile ilgili Yayın/Tebliğlerle ilgili Bilgiler:

Halıcı U., "Reinforcement Learning in Random Neural Networks for Cascaded Decisions, *Journal of Biosystems,* Elsevier, Vol 40 No 1,2, January 1997 pp 83-91

Gulöksüz A., Halıcı U., 1996, "A Neural Circuit to Handle Passive Extinction in Conditioned Reinforcement Learning", *Proceedings of Thirteenth European Meeting on Cybernetics and System Research*, Vienn, Austria

Halýcý U., 1996, (Extended Abstract) Reward, Punishment and Expectation in Reinforcement Learning for random Neural Networks, *Workshop on Biologically Inspired Autonomous Systems: Computation, Cognition and Control,* Duke University, Durham, North Caroline, USA, 4-5 March,

Halýcý U., 1995, (Abstract) "Reinforcement Learning in Random Neural Networks for Mazes", *Symposium on Neuronal Coding,* Prague, September

Gülöksüz A., Halýcý U., 1994, "Artificial Neural Networks for Conditioned Learning", *in Hints From Life to Artificial Intelligence,* Editor: U. Halýcý, METU, December

10. Bilim Dalı: Bilgisayar **Doçentlik B. Dalı Kodu:** Akıllı Sistemler **Uzmanlık Alanı Kodu:**		**ISIC Kodu:**
11. Dağıtım:	☐ Sınırlı	☑ Sınırsız
12. Raporun Gizlilik durumu:	☐ Gizli	☑Gizli Değil