Manifold Embeddings for Model-Based Reinforcement Learning of Neurostimulation Policies

Keith Bush KBUSH@CS.MCGILL.CA
Joelle Pineau JPINEAU@CS.MCGILL.CA
Massimo Avoli MASSIMO.AVOLI@MCGILL.CA

Abstract

Real-world reinforcement learning problems often exhibit nonlinear, continuous-valued, noisy, partially-observable state-spaces that are prohibitively expensive to explore. The formal reinforcement learning framework, unfortunately, has not been successfully demonstrated in a real-world domain having all of these constraints. We approach this domain with a two-part solution. First, we overcome continuous-valued, partially observable state-spaces by constructing manifold embeddings of the system's underlying dynamics, which substitute as a complete state-space representation. We then define a generative model over this manifold to learn a policy off-line. The model-based approach is preferred because it enables simplification of the learning problem by domain knowledge. In this work we formally integrate manifold embeddings into the reinforcement learning framework, summarize a spectral method for estimating embedding parameters, and demonstrate the model-based approach in a complex domain—adaptive seizure suppression of an epileptic neural system.

1. Introduction

A driving force in reinforcement learning research is the growing need for intelligent, autonomous control strategies that operate in real-world domains. Interesting real-world problems, however, often exhibit nonlinear, continuous-valued state-spaces that are only partially observable through signals containing some degree of noise. In a fully observable state representation the first two domain characteristics may be addressed by turning to a rich literature of function approximation. The third characteristic, partial observability, however, makes this class of problem arguably more complex because the functional dependence of the state-space must first be guaranteed before approximation may proceed.

Partial observability is not unique to reinforcement learning. Predictive modeling and dynamic systems research have long identified and addressed partial observability (Sauer et al., 1991) from a geometrical perspective through the method of delayed embeddings, known as Takens Theorem (Takens, 1981). In this approach, sequences of partial observations are grouped (i.e. embedded) onto a higher-dimensional embedding to characterize the phase space of the underlying system. This method is the basis for many nonlinear noise filtering and context-based predictive modeling approaches (Parlitz & Merkwirth, 2000; Huke, March 2006).

The appropriate embedding parameters of a system, unfortunately, are generally unknown *a priori* and must be determined empirically. Assuming that fixed-policy data is available we can estimate these parameters via spectral analysis (Galka, 2000). Moreover, the dynamic preserving characteristics of manifold embeddings allow us to take an additional step and directly model the system. This is an ideal domain for model-based reinforcement learning.

The primary contribution of this paper is to demonstrate reinforcement learning applied to a embedding-based generative model. For our example we choose a challenging problem in neuroscience research, adaptive seizure suppression via neurostimulation of epileptic brain tissue.

2. Background

Our research lies at the intersection of disparate research threads. To provide a single mathematical formalism, in this section we first work through the math-

Appearing in Proceedings of the ICML/UAI/COLT Workshop on Abstraction in Reinforcement Learning, Montreal, Canada, 2009.

ematics of reinforcement learning, partial observability, and embedding theory. We then detail the spectral method used to construct manifold embeddings in practice and motivate why this method is successful.

2.1. Reinforcement Learning

Reinforcement learning (RL) is a class of problems in which an agent learns an optimal solution to a multi-step decision task by interaction with its environment (Sutton & Barto, 1998). We focus on the Q-learning formulation.

Consider a system (i.e. the environment), which evolves according to nonlinear discrete dynamic system g,

$$\mathbf{s}(t+1) \quad = \quad g(\mathbf{s}(t), a(t)). \qquad (1)$$

This function maps the state of the world, $\mathbf{s}(t)$, at the current time, t, and action, $a(t)$ (i.e. decision), onto the state of world one time-step into the future. The environment also includes a reward function $r(t) = h(\mathbf{s}(t))$, which is a scalar measure of the *goodness* of taking the previous action with respect to the goal of the multi-step decision task. The agent selects actions according to the policy, π,

$$a(t) \quad = \quad \pi(\mathbf{s}(t)). \qquad (2)$$

We define reinforcement learning as the process of learning the policy function that maximizes the expected sum of future rewards. The optimal sequence of actions is the optimal policy, π^*, and the maximum expected sum of future rewards is termed the action-value function or Q-function, $Q^*(\mathbf{s}(t), \mathbf{a}(t))$, defined as

$$Q^*(\mathbf{s}(t), a(t)) = r(t+1) + \gamma Q^*(\mathbf{s}(t+1), a(t+1)), \quad (3)$$

where γ is the discount factor. We can then specify the optimal policy, π^*, in terms of the Q-function,

$$\pi^*(\mathbf{s}(t)) \quad = \quad \underset{a}{\operatorname{argmax}}\, Q^*(\mathbf{s}(t), a). \qquad (4)$$

Equations 3 and 4 assume that Q^* is known. Without *a priori* knowledge of Q^*, an approximation, Q, must be constructed iteratively, using temporal difference error (Sutton & Barto, 1998).

2.2. Embedding Foundations

The formulation of Q-learning presented above relies on an assumption of state observability. That is, the complete environment state, $\mathbf{s}$, is an injective function of the observable state, $\tilde{\mathbf{s}}$. In real world domains this is often untrue. A major topic in dynamic systems, signal processing, and time-series analysis is the reconstruction of complete state from incomplete observations. One such reconstruction process, the method of delayed embeddings, may be defined formally by applying Takens Theorem (Takens, 1981). Here we will present the key points utilizing the notation of Huke (Huke, March 2006).

Consider the state-space, $\mathbf{s}$, of the environment is an M-dimensional, real-valued vector space and a is a real-valued action input to the environment. We substitute Equation 2 into Equation 1 and compose a new function, ϕ,

$$\begin{aligned} \mathbf{s}(t+1) &= g(\mathbf{s}(t), \pi(\mathbf{s}(t))), \\ &= \phi(\mathbf{s}(t)), \end{aligned}$$

which specifies the discrete time evolution of our combined system of agent and environment. Assume that this system is partially observable via observation function, y, such that

$$\tilde{s}(t) \quad = \quad y(\mathbf{s}(t)),$$

where $y : \mathbb{R}^M \rightarrow \mathbb{R}$ and y is noise free and represented with infinite floating point precision. If ϕ is invertible, and ϕ, ϕ^{-1}, and y are differentiable we may apply Takens Theorem to reconstruct the dynamics of system $\phi(\mathbf{s}(t))$ using the observables $\tilde{s}(t)$. For each $\tilde{s}(t)$, we construct a vector $\mathbf{s}_E(t)$,

$$\mathbf{s}_E(t) \quad = \quad [\tilde{s}(t), \tilde{s}(t-1), ..., \tilde{s}(t-E)]. \qquad (5)$$

If $E > 2M$ then the vectors $\mathbf{s}_E(t)$ lie on a subset of $\mathbb{R}^E$ which is an embedding of our original system (Huke, March 2006). This embedding forms a new dynamic system which preserves the structure of the original system. Thus, not only does there exist a vector $\mathbf{s}_E(t)$ for each observation $\tilde{\mathbf{s}}$, but the dynamics governing the evolution of these vectors in time is preserved, such that there exists a function, ψ,

$$\mathbf{s}_E(t+1) \quad = \quad \psi(\mathbf{s}_E(t)). \qquad (6)$$

In the context of reinforcement learning, the vectors $\mathbf{s}_E$(t) may be substituted into Equations 3 and 4 as replacements for complete state $\mathbf{s}(t)$ without loss of generality. In this same context, however, embeddings exhibit a serious limitation. Because we roll the policy into the definition of the state transition, each element of the reconstructed state space is policy dependent. Therefore, an embedding is only explicitly valid for the policy π under which the time-series was observed. By changing the policy slowly, and by carefully choosing a robust function approximation, these effects can be minimized.

2.3. Spectral Embedding Method

Embedding theory does not tell us how to select the embedding parameters. In general, the intrinsic dimension of the system, M, is unknown and not easily determined. Therefore, we need a reliable method for selecting parameters that yield a high-quality embedding. In practice we utilize a spectral method (Galka, 2000) employing the singular value decomposition (SVD).

A summary of this method follows. Consider a partially-observable, discretely-sampled time-series, $\tilde{\mathbf{s}}$, of length $\tilde{S}$ that we desire to embed. We choose a *sufficiently large* fixed embedding dimension, E. Sufficiently large refers to a cardinality of dimension which is certain to be greater than the dimension in which the actual state-space resides.

The spectral method we propose works by manipulating $\tilde{\mathbf{s}}$ into a form usable in Equation 5. To do this we define the embedding window size, T_{min}. We then utilize a desample rate, $\tau = T_{min}/(E-1)$, to uniformly select elements of $\tilde{\mathbf{s}}$ from the window T_{min}, according to the rule,

$$\mathbf{s}_E(t) \quad = \quad [\tilde{s}(t), \tilde{s}(t-\tau), ..., \tilde{s}(t-(E-1)\tau)]. \quad (7)$$

We assume a range $[T_{min}^{low}, T_{min}^{high}]$ and interval, ΔT_{min}, over which to explore, $T_{min}(i) = T_{min}^{low} + i\Delta T_{min}$, $i \in 1, 2, ..., N_{min}$ where $N_{min} = T_{min}^{high}/\Delta T_{min}$. A good upper bound, T_{min}^{high}, is the fundamental period of the system if estimable.

For each i, we construct the vectors $\mathbf{s}_E(t)$, $t \in 1, ..., N_E(i)$, where $N_E(i) = \tilde{S} - T_{min}(i)$, and concatenate them as rows of a matrix, $\mathbf{S}_E(i)$, of size $N_E \times E$. We compute the SVD of $\mathbf{S}_E(i)$,

$$\mathbf{S}_E(i) \quad = \quad \mathbf{U}(i)\mathbf{\Sigma}(i)\mathbf{V}^T(i)$$

where $\mathbf{\Sigma}(i)$ is a diagonal matrix of loadings, (i.e. the energy contained in the dimensions of the new space) and $\mathbf{V}^T(i)$ is the basis of the SVD coordinate space. We store the diagonal of each $\Sigma(i)$ for analysis as the ith row of matrix χ, of size $N_{min} \times E$.

Why is spectral method effective for identifying high-quality embedding parameters? In the SVD decomposition, the right singular vectors, $\mathbf{V}^T$ are the Eigen vectors of the covariance matrix of the embedding, $\mathbf{C}_E = \mathbf{S}_E^T\mathbf{S}_E$, if matrix S_E is mean centered. Therefore, columns of $\mathbf{U}$ define the principal components of $\mathbf{S}_E$ (Kirby, 2001).

In the limits of this embedding technique the covariance matrix, $\mathbf{C}_E$, will take on two forms. For very small T_{min} the rows of $\mathbf{S}_E$ are redundant, which drives elements of $\mathbf{C}_E$ toward a uniform constant value. Thus, $\mathbf{C}_E$ is rank one and all of the variance is in the first component. This is the time-series itself. In the second case, for very long T_{min}, the rows of $\mathbf{S}_E$ become uncorrelated. Therefore, diagonal values of the covariance matrix, $\mathbf{C}_E$, will trend toward constant values while off-diagonal elements tend to zero. In this case $\mathbf{C}_E$ is full rank and is indistinguishable from noise.

Suitable embedding parameters can be found by analysis of the principal components at intermediate lengths of T_{min}. Good embedding parameters are found when a subset (preferably small) of the eigenvalues stored in the i_{opt} row of χ simultaneously exhibit a local optima. The value $T_{min}(i_{opt})$ identifies the appropriate embedding window length, which also determines the desample rate $\tau(i_{opt})$. The subset of eigenvalues with non-trivial values defines the appropriate embedding dimension, E'.

3. Methods

Our approach combines best practices from both nonlinear dynamic analysis and reinforcement learning to identify high-quality policies in partially observable, real-world domains that are prohibitively expensive to explore. This practice incorporates the following steps: 1) record a partially observable system under the control of a random policy or some other policy or set of policies known to be near the desired optimal policy; 2) perform spectral embedding; 3) identify good candidate parameters for embedding; 4) choose a local function approximator well-suited to the demands of the embedding; 5) construct an integrable model of the system's dynamics; and 6) learn a desired policy on this model via reinforcement learning.

4. Case Study: Adaptive Neurostimulation

Epilepsy afflicts approximately 1% of the world's population (The Epilepsy Foundation, 2009). Of those suffering from this disease 30% do not respond to currently available anticonvulsant treatments or are not candidates for surgical resection. Development of new treatments, therefore, is a priority for epilepsy research.

Neurostimulation shows promise as an epilepsy treatment. *In vitro* studies indicate that fixed-frequency external electrical stimulation applied to substructures within the hippocampus can effectively suppress seizures (Durand & Bikson, 2001). Fixed-frequency policies, however, do have limitations. *In vitro* neurostimulation experiments suggest that the efficacy of fixed-frequency stimulation varies across epileptic neural systems. The recency of this technology also raises questions about long-term impacts such as stimulation induced tissue damage.

These limitations motivate the search for stimulation policies that satisfy additional constraints. Ideally, a treatment policy should adapt to optimally suppress seizures in each unique patient while minimizing the number of stimulations necessary to do so. By posing the problem's components in this way, an agent (implant) that learns a policy (treatment) that maximizes rewards (maximum suppression using minimum stimulation constraints) by interacting with the environment (patient), we can recast neurostimulation treatment of epilepsy, formally, as a reinforcement learning problem (Sutton & Barto, 1998).

Recasting as a learning problem requires that these components be mathematically well-defined. The complex dynamics of neural systems, however, are typically observable only through low-dimensional time-series corrupted by noise (e.g. extracellular recording electrodes). Therefore, the objectives of this case study are twofold: 1) identify low-dimensional state-space and transition model from field potential recordings of neural systems that accurately reproduces observed neural dynamics and 2) learn a neurostimulation therapy in this model that minimizes both seizures and stimulations.

To fulfill our first objective we construct a state-space and transition model from previously recorded data under fixed policies. Our dataset is comprised of field potential recordings from five epileptic rat hippocampal slices were made under fixed-frequency stimulation policies of 0.2, 0.5, and 1.0 Hz as well as control (i.e., unstimulated). The dataset totals 4,639 seconds of recordings including 15 seizures (seizure labels are hand annotated).

From this dataset we desire to construct a high-quality state-space and transition model, using only the control data. We measure quality as the predictive accuracy of a generative model built from these components. Using the spectral embedding method, presented in Figure 1, we extract the manifold embedding parameters of the dataset ($E' = 5$, $T_{min} = 1.2$ seconds) and then embed the dataset, using Equation 5. This defines the state-space of our system in $\mathbb{R}^{E'}$. We define the transition model as the local time-derivative of the element of the dataset that is nearest the current state (i.e., a nearest neighbors derivative). Because the stimulation events are not logged in the dataset, we must find a reasonable mapping of actions into our domain. To do this, we define an action as an interictal-like derivative which is added to the current derivative of the system. We then define the generative model of the system as numerical integration over the embedding.

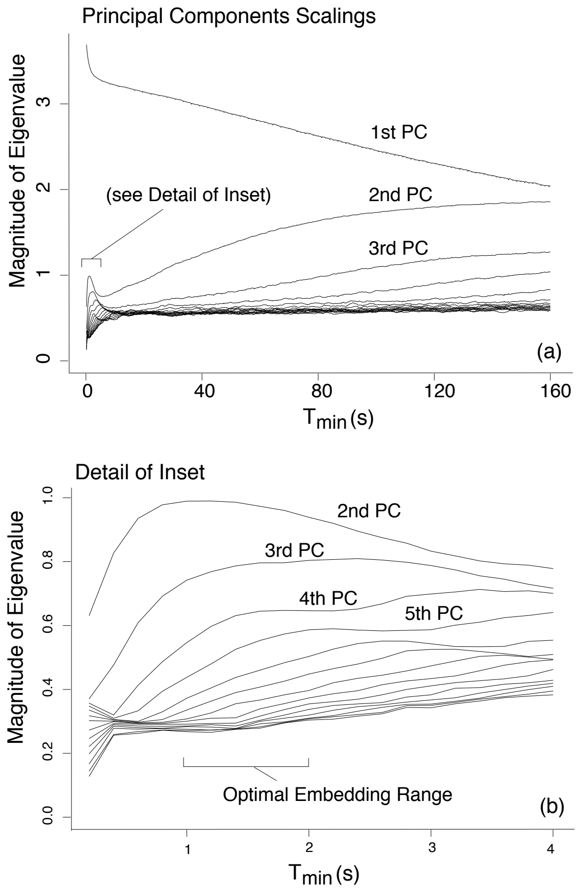

Figure 1. Selecting embedding parameters via PCA: (a) eigenvalues of the principal components of the embedding as a function of embedding window length T_{min} holding embedding dimension constant at $E' = 15$; (b) detail of principal components 2–15 from plot (a) for the embedding window range $T_{min} = [0.1, 4]$ seconds, highlighting the local maxima of components 2–5 in this range.

We simulate rat hippocampal dynamics under control and fixed-frequency stimulation policies, choosing the embedding parameters that yield the smallest simulation error with respect to analogous fixed-frequency policies in the dataset.

Using reinforcement learning, we find an optimal policy which maximizes seizure suppression with minimal stimulations. We define the Q-function approximation for each state in $\mathbb{R}^{E'}$ as the Q-value of the nearest element of the model. Actions take one of two forms, either *on* or *off*. We perform ϵ-greedy SARSA on the generative model subject to the reward function (-1 for

	Mean Frac. of Seiz. States	Mean Seizure Length (s)	Mean Seizure Interval (s)
model	0.19 ± 0.02	61.0 ± 4.6	251.7 ± 14.5
data	0.19 ± 0.03	64.4 ± 39.4	271.6 ± 121.3

Table 1. Comparison of summary statistics between the original dataset and 30 generative model simulations of 40,000 seconds.

	Mean Fraction of Seizure States			
policy	0.2 Hz	0.5 Hz	1.0 Hz	Adaptive*
model	0.15±0.02	0.14±0.02	0.12±0.02	0.02±0.01
data	0.16±0.23	0.12±0.16	0.08±0.09	—

Table 2. Comparison of summary statistics between the original dataset and 30 generative model simulations of 40,000 seconds under fixed-frequency neurostimulation of 0.2, 0.5, and 1.0 Hz and adaptive neurostimulation (*effective frequency = 0.02 Hz).

each stimulation and -20 for visiting an element of the dataset that is labeled seizure). We restart the simulation from a random initial element of the dataset every 1,500 steps (5 min. simulation time) until 90% of the model states have been visited during simulation more than 20 times.

Performance validation of the simulation is summarized in Tables 1 and 2 for both control and fixed frequency stimulation policies. The model's seizure dynamics and suppression predictions fall within confidence intervals of the original dataset. Our learned stimulation policy achieves 0.02 ± 0.01 fraction of seizure states using an effective mean stimulation frequency of 0.02Hz, more than an order of magnitude better than the best fixed-frequency policy.

The learned policy also provides useful qualitative knowledge of the domain. For each simulation timestep, we calculate the nearest dataset element of the current state, the seizure label of this element, and the action requested. From this data we construct the agent's policy graph, Figure 2(a), projected onto the first two principle components. This graph is formed by 1) drawing edges between elements of the dataset that request stimulation events less than 4 seconds apart and 2) scaling each element's size by the proportion of stimulations requested by the agent at that element.

This graph unmasks two distinct classes in the learned policy. The first class consists of pulse trains requested when the simulation operates in the post-seizure region of the manifold, Figure 2(b). These pulse trains are visually identifiable as cliques in the policy graph formed by the edges. The second policy class consists of individual, well-timed stimulations requested when the simulation operates in the dynamically normal region of the manifold, Figure 2(c)—note the lack of edges, indicating individual stimulation requests.

Without on-line verification of this policy, these results are of limited value. Qualitatively, however, these results agree well with neurostimulation results found in the literature (Durand & Bikson, 2001). Prior studies have observed 1) avoidance of full seizure development using well-timed stimulations immediately at seizure onset (i.e., at the interface between normal and seizure dynamics) (Durand & Warman, 1994) and 2) early termination of fully developed seizures via low-frequency stimulation (D'Arcangelo et al., 2005). The learned neurostimulation policy incorporates both of these policy classes, but applies them under different dynamic regimes, seemingly to maximize the suppression benefits of each class. The literature does not claim that well-timed stimulations can abort fully developed seizures. Moreover, there exists evidence suggesting that fixed-frequency policies below 0.2 Hz exhibit little or no seizure suppression. Therefore, it makes sense that a well-timed stimulation policy in the normal regime, if possible, would be the best available choice.

5. Discussion

The reinforcement learning community has long been aware of the need to find low-dimensional representations to capture complex domains. Approaches for efficient function approximation, basis function construction, and discovery of embeddings have been the topic of significant investigations (Bowling et al., 2005; Keller et al., 2006; Smart, 2004; Mahadevan & Maggioni, 2007). However most of this work has been limited to the fully observable (MDP) case and has not been shown to extend to partially observable environments. The question of state space representation in partially observable domains was tackled under the POMDP framework (McCallum, 1996) and more recently in the PSR framework (Singh et al., 2003). While these methods tackle a similar problem they have primarily been limited to discrete action and observation spaces.

The PSR framework was recently extended to continuous (nonlinear) domains (Wingate & Singh, 2007). This method is significantly different from our work, both in terms of the class of representations it considers and in the criteria used to select the appropriate representation. Furthermore, it has not yet been applied to real-world domains. Nonetheless it could potentially be used to tackle the problems presented in

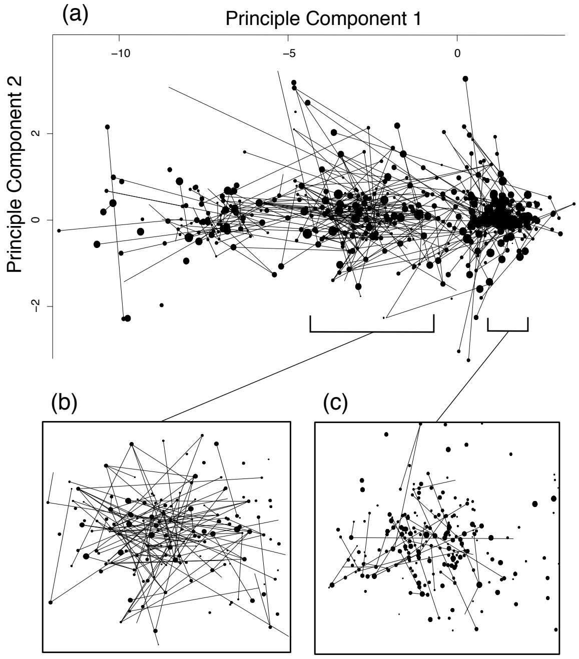

Figure 2. Policy analysis: (a) The learned *policy graph* plotted on the first two-principle components of the embedding manifold. (b) Detail of the policy graph in the post-seizure portion of the attractor. (c) Detail of the policy graph in the dynamically normal portion of the attractor.

Section 4; an empirical comparison with our approach is left for future consideration.

The implications of this technique for modeling poorly understood, partially observable domains are more concrete. We demonstrate a data-driven, generative model construction technique that accurately reproduces the dynamics of fixed-frequency electrical stimulation applied to a slice of epileptic neural tissue. Using reinforcement learning, we identify an optimal suppression policy that is, interestingly, two-class. This learned policy predicts a dependence between suppression efficacy of these two policy classes and the dynamic regime in which these policy classes are applied, which is supported, in part, by previous *in vitro* experiments.

References

Bowling, M., Ghodsi, A., & Wilkinson, D. (2005). Action respecting embedding. *Proceedings of ICML.*

D'Arcangelo, G., Panuccio, G., Tancredi, V., & Avoli, M. (2005). Repetitive low-frequency stimulation reduces epileptiform synchronization in limbic neuronal networks. *Neurobiology of Disease*, *19*, 119–128.

Durand, D., & Warman, E. (1994). Desynchronization of epileptiform activity by extracellular current pulses in rat hippocampal slices. *Journal of Physiology*, *71*, 2033–2045.

Durand, D. M., & Bikson, M. (2001). Suppression and control of epileptiform activity by electrical stimulation: A review. *Proceedings of the IEEE*, *89(7)*, 1065–1082.

Galka, A. (2000). *Topics in nonlinear time series analysis: with implications for eeg analysis.* World Scientific.

Huke, J. (March, 2006). *Embedding nonlinear dynamical systems: A guide to takens' theorem* (Technical Report). Manchester Institute for Mathematical Sciences, University of Manchester.

Keller, P., Mannor, S., & Precup, D. (2006). Automatic basis function construction for approximate dynamic programming and reinforcement learning. *Proceedings of ICML.*

Kirby, M. (2001). *Geometric data analysis: An empirical approach to the dimensionality reduction and the study of patterns.* Wiley and Sons.

Mahadevan, S., & Maggioni, M. (2007). Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, *8.*

McCallum, A. K. (1996). *Reinforcement learning with selective perception and hidden state.* Doctoral dissertation, University of Rochester.

Parlitz, U., & Merkwirth, C. (2000). Prediction of spatiotemporal time series based on reconstructed local states. *Physical Review Letters*, *84(9)*, 1890–1893.

Sauer, T., Yorke, J. A., & Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, *65:3/4*, 579–616.

Singh, S., Littman, M. L., Jong, N. K., Pardoe, D., & Stone, P. (2003). Learning predictive state representations. *Machine Learning: Proceedings of the 2003 International Conference (ICML)* (pp. 712–719).

Smart, W. (2004). Explicit manifold representations for value-functions in reinforcement learning. *Proceedings of the Int. Sympo. on AI and Math.*

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: The MIT Press.

Takens, F. (1981). Detecting strange attractors in turbulence, dynamical systems and turbulence. In D. A. R. . L. S. Young (Ed.), *Lecture notes in mathematics*, vol. 898, 366–381. Springer.

The Epilepsy Foundation (2009). Epilepsy and seizure statistics. Available at `http://www.epilepsyfoundation.org/about/statistics.cfm`.

Wingate, D., & Singh, S. (2007). On discovery and learning of models with predictive state representations of state for agents with continuous actions and observations. *Proceedings of AAMAS.*