Selection Criteria for Neuromanifolds of Stochastic Dynamics

Nihat Ay
Guido Montúfar
Johannes Rauh

SFI WORKING PAPER: 2011-09-040

SANTA FE INSTITUTE

Selection Criteria for Neuromanifolds of Stochastic Dynamics

Nihat Ay[1,2], Guido Montúfar[1], Johannes Rauh[1]
[1]Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22, D-04103 Leipzig,Germany
[2]Santa Fe Institute, 1399 Hyde Park Road,
Santa Fe, New Mexico 87501, USA
{nay,montufar,rauh}@mis.mpg.de

Abstract

We present ways of defining neuromanifolds – models of stochastic matrices – that are compatible with the maximization of an objective function such as the expected reward in reinforcement learning theory. Our approach is based on information geometry and aims at the reduction of model parameters with the hope to improve gradient learning processes.

1 Introduction

Within many formal models of neural networks the dynamics of the whole system can be described as a stochastic transition in each time step, mathematically formalized in terms of a stochastic matrix. Well-known models of this kind are Boltzmann machines [2], their generalizations [5], and policy matrices within reinforcement learning [7]. It is helpful to consider not only one stochastic matrix but a parametrized family of matrices, which forms a geometric object, referred to as a neuromanifold within information geometry [1, 2]. This information geometric view point suggests to select appropriate neuromanifolds and to define corresponding learning processes as gradient flows on these manifolds. The natural gradient method, developed by Amari and co-workers (see for example [1]), proved the efficiency of the geometric approach to learning. The study of learning systems should further address the interplay between geometric properties and the quality of learning. In this paper we study criteria for the selection of neuromanifolds. We do not only focus on manifolds that are directly induced by neuronal models, but also study more general geometric objects that satisfy natural optimality conditions. Therefore, in the following we will talk about *models* instead of neuromanifolds.

We assume that learning maximizes an objective function $f : \mathcal{C} \to \mathbb{R}$ defined on the set $\mathcal{C}$ of stochastic matrices. A model $\mathcal{N} \subseteq \mathcal{C}$ is consistent with f, if the set of maximizers of f can be reached through the learning. This implies that the maximizers should be contained in the closure of $\mathcal{N}$. If f is convex on $\mathcal{C}$, then each locally maximal value is attained at an extreme point (vertex) of $\mathcal{C}$, and therefore corresponds to a

deterministic function. We refer to the following three examples in which optimal systems also turn out to be close to deterministic functions:

1. Optimal policies in reinforcement learning [6],
2. dynamics with maximal predictive information as considered in robotics [8], and
3. dynamics of neural networks with maximal network information flow [3].

This suggests to consider parametrizations that can approximate all extreme points of $\mathcal{C}$, the deterministic functions. In this paper we concentrate on the first example to illustrate the main idea.

2 The main geometric idea

We first consider general convex sets and return to stochastic matrices in Section 3. The convex hull of a finite set $\xi^{(1)}, \ldots, \xi^{(n)}$ in $\mathbb{R}^d$ is defined as

$$\mathcal{C} := \left\{ \sum_{i=1}^{n} p(i)\, \xi^{(i)} \;:\; p(i) \geq 0 \;\forall i \text{ and } \sum_{i=1}^{n} p(i) = 1 \right\}.$$

The set of extreme points of this polytope $\mathcal{C}$ is a subset of $\{\xi^{(1)}, \ldots, \xi^{(n)}\}$. In general, there are many ways to represent a point $x \in \mathcal{C}$ as a convex combination in terms of a probability distribution p. Here, we are interested in convex combinations obtained from an exponential family. To be more precise, denote $\mathcal{P}_n$ the set of probability measures $p = (p(1), \ldots, p(n)) \in \mathbb{R}^n$ and consider the map

$$\mathsf{m} : \mathcal{P}_n \to \mathcal{C}, \quad p \mapsto \sum_{i=1}^{n} p(i)\, \xi^{(i)}.$$

For a family of functions $\phi = (\phi_1, \ldots, \phi_l)$ on $\{1, \ldots, n\}$, we consider the exponential family $\mathcal{E}_\phi$ consisting of all $p \in \mathcal{P}_n$ of the form

$$p(i) = \frac{e^{\sum_{k=1}^{l} \lambda_k\, \phi_k(i)}}{\sum_{j=1}^{n} e^{\sum_{k=1}^{l} \lambda_k\, \phi_k(j)}}, \qquad i = 1, \ldots, n.$$

We denote the image of $\mathcal{E}_\phi$ under m by $\mathcal{C}_\phi$. With the choice

$$\phi_k^*(i) := \xi_k^{(i)}, \qquad i = 1, \ldots, n, \;\; k = 1, \ldots, d,$$

the closure of the exponential family $\mathcal{E}_\phi$ can be identified with the polytope $\mathcal{C}$. This allows to define natural geometric structures on $\mathcal{C}$, such as a Fisher metric, by using the natural structures on the simplex $\mathcal{P}_n$. In the context of stochastic matrices this leads to a Fisher metric that has been studied by Lebanon [4] based on an approach by Čencov. The above construction also motivates the following definition: We call a family $\mathcal{C}_\phi$ an *exponential family* in $\mathcal{C}$ if the vectors ϕ_k, $k = 1, \ldots, l$, are contained in the linear span of the vectors ϕ_k^*, $k = 1, \ldots, d$.

In general, the families $\mathcal{C}_\phi$ are not exponential families but projections of exponential families. In this paper the models $\mathcal{C}_\phi$ will play the role of neuromanifolds. We are mainly interested in models that are compatible with the maximization of a given function $f : \mathcal{C} \to \mathbb{R}$ in the sense that the closure of $\mathcal{C}_\phi$ should contain the maximizers

of f. This is clearly not the only consistency condition, but here we focus on this assumption only.

As stated above, in many cases the local maximizers of f are elements of the set $\{\xi^{(1)},\dots,\xi^{(n)}\}$, and hence the problem stated above reduces to finding a family $\phi=(\phi_1,\dots,\phi_l)$ of functions such that $\mathcal{C}_\phi$ contains that set in its closure. This is always possible with only two functions ϕ_1,ϕ_2. One such family can be constructed as follows: Consider a one-to-one map φ of the n points $\xi^{(1)},\dots,\xi^{(n)}$ into $\mathbb{R}$, for instance $\xi^{(i)}\mapsto i$, $i=1,\dots,n$, and the following family of distributions:

$$\begin{aligned} p_{\alpha,\beta}(i) &= \frac{e^{-\beta\left(\varphi(\xi^{(i)})-\alpha\right)^2}}{\sum_{j=1}^n e^{-\beta\left(\varphi(\xi^{(j)})-\alpha\right)^2}} \\ &= \frac{e^{\lambda_1\,\phi_1(i)+\lambda_2\,\phi_2(i)}}{\sum_{j=1}^n e^{\lambda_1\,\phi_1(j)+\lambda_2\,\phi_2(j)}}, \end{aligned} \tag{1}$$

where $\phi_1(i):=\varphi(\xi^{(i)})$, $\phi_2(i):=\varphi^2(\xi^{(i)})$, and $\lambda_1:=2\,\alpha\,\beta$, $\lambda_2:=-\beta$. It is easy to see that for $\alpha=\varphi(\xi^{(i)})$ and $\beta\to\infty$, the distribution $p_{\alpha,\beta}$ converges to the point measure concentrated in i. The convex combination $\sum_{j=1}^n p_{\alpha,\beta}(i)\,\xi^{(i)}$ therefore converges to the point $\xi^{(i)}$. This proves that the closure of this two-dimensional family in $\mathcal{C}$ contains all the points $\xi^{(i)}$, $i=1,\dots,n$. In general, the geometric properties of this family strongly depend on φ, as we discuss in the following section.

3 Application to reward maximization

Given non-empty finite sets $\mathscr{X}$ and $\mathscr{Y}$, the stochastic matrices from $\mathscr{X}$ to $\mathscr{Y}$ are maps $(x,y)\mapsto\pi(x;y)$ satisfying

$$\pi(x;y)\geq 0 \;\text{ for all } x\in\mathscr{X}, y\in\mathscr{Y}, \quad \text{and}$$

$$\sum_{y\in\mathscr{Y}}\pi(x;y)=1 \;\text{ for all } x\in\mathscr{X}.$$

The set of stochastic matrices is denoted by $\mathcal{C}:=\mathcal{C}(\mathscr{X};\mathscr{Y})$. Stochastic matrices are very general objects and can serve as models for individual neurons, neural networks, and policies. Each extreme point of this convex set corresponds to a deterministic function $f:\mathscr{X}\to\mathscr{Y}$ and is given as

$$\pi^{(f)}(x;y)=\begin{cases}1, & \text{if } y=f(x),\\ 0, & \text{else.}\end{cases}$$

Although the number of these extreme points is $|\mathscr{Y}|^{|\mathscr{X}|}$, according to Section 2 there always exists a two-dimensional manifold that reaches all of them. Note that in the particular case of N binary neurons we have $\mathscr{X}=\mathscr{Y}=\{0,1\}^N$ and therefore $(2^N)^{(2^N)}$ extreme points.

To illustrate the geometric idea we consider the example $\mathscr{X}=\{1,2,3\}$ and $\mathscr{Y}=\{1,2\}$. This can, for instance, serve as a model for policies with three states and two actions. In this case $\mathcal{C}$ is a subset of $\mathbb{R}^{\mathscr{X}\times\mathscr{Y}}\cong\mathbb{R}^6$ which can be identified with the hypercube $[0,1]^3$ through the following parametrization (see Fig. 1 A):

$$[0,1]^3 \ni (r,s,t) \;\; \mapsto \;\; \begin{pmatrix} r & 1-r \\ s & 1-s \\ t & 1-t \end{pmatrix}.$$

To test the properties of that family with respect to the optimization of a function, we consider a map $(s,a) \mapsto \mathcal{R}^a_s$, which we interpret as *reward* that an agent receives if it performs action a after having seen state s. The policy of the agent is described by a stochastic matrix $\pi(s;a)$. The expected reward can be written as

$$f(\pi) = \sum_s p^\pi(s) \sum_a \pi(s;a)\, \mathcal{R}^a_s.$$

In reinforcement learning, there are several choices of p^π (see [7]). Here we simplify our study by assuming p^π to be the uniform measure.

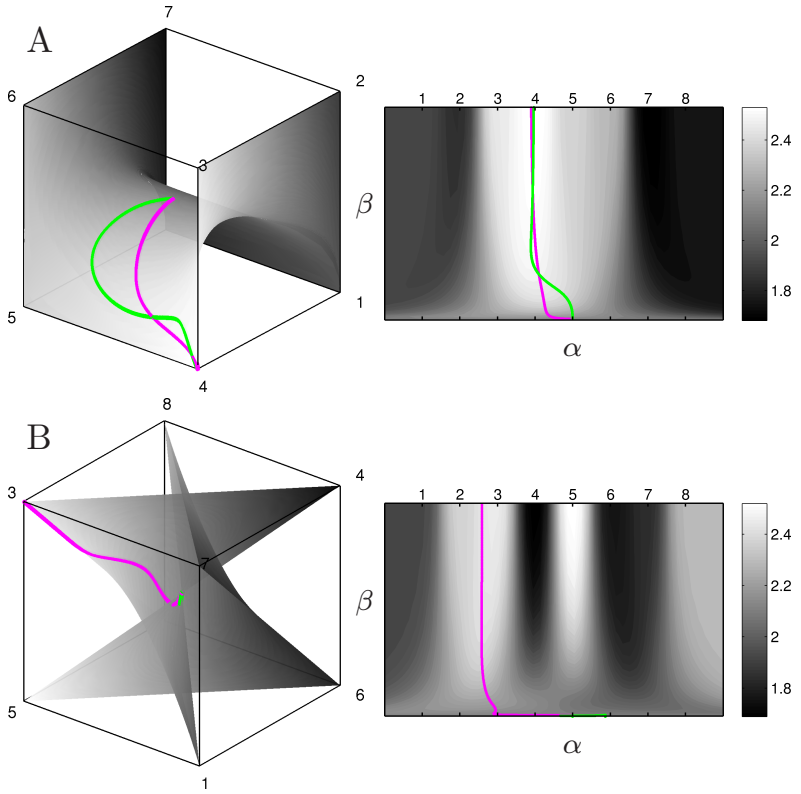

Figure 1: Optimization with ordinary (green) and natural (magenta) gradient on the model $\mathcal{C}_\phi$ for two different choices of φ.
A: A Hamilton path $\varphi = (1,2,3,4,5,6,7,8)$.
B: An arbitrary map $\varphi = (1,7,3,5,2,8,4,6)$.

We investigate the influence of the map φ and compare the natural gradient flow (gradient with respect to the Fisher metric, see [1]) with the ordinary gradient. For the experiments we drew a random reward matrix $\mathcal{R}$ and applied gradient ascent

(with fixed step size) on $f(\pi)$ restricted to our model and several choices of φ (see Figs. 1 A/B for typical outcomes). The optimization results strongly depend on φ. We restricted ourselves to the case that φ maps the vertices of $\mathcal{C}$ onto the numbers $\{1,\dots,n\}$. Such a map is equivalent to an ordering of the vertices. We recorded the best results when φ corresponds to a Hamilton path on the graph of the polytope $\mathcal{C}$, i.e. a closed path along the edges of the polytope that visits each vertex exactly once. This way φ preserves the locality in $\mathcal{C}$, and the resulting model $\mathcal{C}_\phi$ is a smooth manifold. In Fig. 1 A, both methods reach the global optimum $\begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$. In Fig. 1 B, φ is 'unordered'. We see that the landscape $f(\pi_{\alpha,\beta})$ is more intricate and contains several local maxima. The natural gradient method only converged to a local but not global optimum, and the ordinary gradient method failed. In Figs. 1 A/B every vertex ξ of the cube is labeled by $\varphi(\xi)$ for the corresponding φ.

4 Towards a construction of neuromanifolds

Here we approach implementations of policies π in the context of neural networks. We start with the case of two input neurons and one output neuron (Fig. 2, left). All neurons are considered to be binary with values 0 and 1.

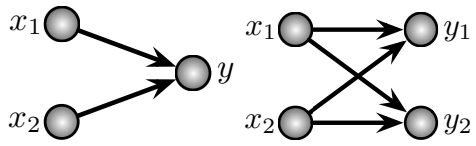

Figure 2: Two simple neural networks.

The input-output mapping is modelled in terms of a stochastic matrix π. The set of such 4×2-matrices forms a four-dimensional cube. A prominent neuronal model assumes synaptic weights w_1 and w_2 assigned to the directed edges and a bias b. The probability for the output 1, which corresponds to the spiking of the neuron, is then given as

$$\pi(x_1, x_2; 1) \quad = \quad \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2 - b)}}. \tag{2}$$

This defines a three-dimensional model in the four-dimensional cube, see Fig. 3. Some extreme points are not contained in this model, e.g. the matrix $\pi(0,0;1) = \pi(1,1;1) = 0$, $\pi(0,1;1) = \pi(1,0;1) = 1$. This corresponds to the well-known fact that the standard model cannot represent the XOR-function. On the other hand, it is possible to reach all extreme points, including the XOR-function, with the two-dimensional models introduced above. However, there are various drawbacks of our models in comparison with the standard model. They are not exponential families but only projections. Moreover, we do not have a neurophysiological interpretation of the parameters.

We now discuss models for the case of one additional output neuron. The system is modelled by stochastic 4×4 matrices, which form the 12-dimensional polytope

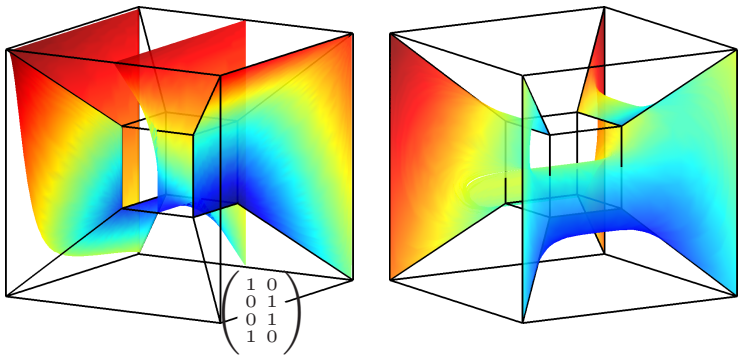

Figure 3: The standard model given in equation (2) for three values of the bias parameter b (left) and the new model (right) introduced in Section 2.

$\mathcal{C} := \mathcal{C}(\{0,1\}^2;\{0,1\}^2)$. A natural assumption is the independence of the outputs Y_1 and Y_2 given the input pair X_1, X_2. This is the case if and only if the input-output map of each neuron i is modelled by a separate stochastic matrix π_i, $i = 1, 2$. The stochastic matrix of the whole system is given by

$$\pi(x_1, x_2; y_1, y_2) \;=\; \pi_1(x_1, x_2; y_1) \cdot \pi_2(x_1, x_2; y_2).$$

This defines an 8-dimensional model $\mathcal{N}_{\text{product}}$ that contains all extreme points of $\mathcal{C}$. Furthermore, it contains the submodel $\mathcal{N}_{\text{standard}}$ given by the additional requirement that π_1 and π_2 are of the form (2). The model $\mathcal{N}_{\text{standard}}$ is an exponential family of dimension 6. However, as in the one-neuron case, $\mathcal{N}_{\text{standard}}$ does not reach all extreme points. Another submodel $\mathcal{N}_{\text{new}}$ of $\mathcal{N}_{\text{product}}$ is defined by modelling each of the stochastic matrices π_i in terms of two parameters as described above. The following table gives a synopsis:

model	dim.	exponential family	reaches all extreme points
$\mathcal{C}$	12	yes	yes
$\mathcal{N}_{\text{product}}$	8	yes	yes
$\mathcal{N}_{\text{standard}}$	6	yes	no
$\mathcal{N}_{\text{new}}$	4	no	yes

We conclude this section with the maximization of a reward function in the family $\mathcal{N}_{\text{new}}$, as in the previous section. Fig. 4 shows a histogram of the results for a fixed randomly chosen reward $\mathcal{R}$ after 500 steps for ordinary gradient and natural gradient methods. We chose a constant learning rate and 5,000 different initial values. Both methods find 3 local maxima. The natural gradient process tends to converge faster. Furthermore, it finds the global maximum for a majority of the initial values, which is not the case for the ordinary gradient.

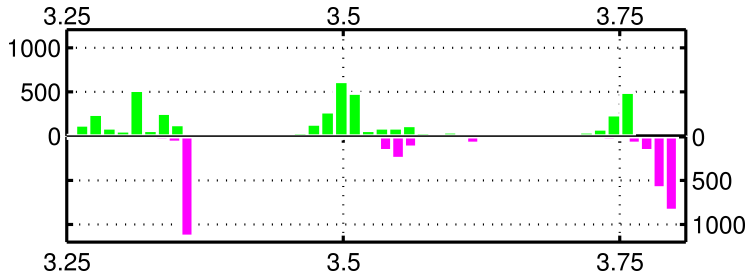

Figure 4: Histogram of the objective value $f(\pi)$ after 500 steps of gradient ascent in $\mathcal{N}_{\text{new}}$. Magenta: natural gradient. Green: ordinary gradient.

5 Conclusions

We proposed and studied models which contain all extreme points in the set of stochastic matrices (the global maximizers for a variety of optimization problems). These models have very few parameters, a rich geometric structure and allow a simple implementation of natural gradient methods. At this stage we do not suggest them for describing neural systems but as basis for extensions to more plausible models.

References

[1] S. Amari. *Natural Gradient Works Efficiently in Learning*. Neural Comput., 10(2) (1998) 251–276.

[2] S. Amari, K. Kurata, H. Nagaoka. *Information Geometry of Boltzmann machines*. IEEE T. Neural Networks, 3(2) (1992) 260–271.

[3] N. Ay, T. Wennekers. *Dynamical Properties of Strongly Interacting Markov Chains*. Neural Networks 16 (2003) 1483–1497.

[4] G. Lebanon. *Axiomatic geometry of conditional models*. IEEE Transactions on Information Theory, 51(4) (2005) 1283–1294.

[5] G. Montúfar, N. Ay. *Refinements of Universal Approximation Results for DBNs and RBMs*. Neural Comput. 23(5) (2011) 1306–1319.

[6] R. Sutton, A. Barto. *Reinforcement Learning*. MIT Press (1998).

[7] R. Sutton, D. McAllester, S. Singh, Y. Mansour. *Policy Gradient Methods for Reinforcement Learning with Function Approximation*. Adv. in NIPS 12 (2000) 1057 – 1063.

[8] K.G. Zahedi, N. Ay, R. Der. *Higher Coordination With Less Control – A Result of Information Maximization in the Sensorimotor Loop*. Adaptive Behavior 18 (3-4) (2010), 338 – 355.