

Chapter 3

Projection Pursuit

3.1 Introduction

One of the central themes of this work is the alteration of neuronal properties via the alteration of the environment. This has allowed us to simulate biological experiments, and has yielded some predictions arising from the parameter dependence. In this chapter we introduce projection pursuit (Friedman, 1987), which looks at the other side of the coin of the problem: the environment. It allows us to understand which properties of the environment which influence the neuron, in order to give the specific modification equations we have presented. It then offers a straightforward way of introducing other learning rules starting from the properties in the environment which influence them.

In the projection pursuit formalism, we think of the process of learning in neurons as an optimization process. The synapses in the neuron modify in order to find structure, for example edges, in the inputs. The modification of synapses then can thus be seen as the maximization of a cost function¹ which, in some way, is a measure of the structure in the inputs. Though it is somewhat irregular to speak of maximizing cost (as opposed to minimizing it) I use this phrase because it is convenient for the particular cost functions we will consider.

Diaconis and Freedman (1984) show that for most high-dimensional clouds (of points), most low-dimensional projections are approximately Gaussian. This finding suggests that important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. The definition of “far from Gaussian” is non-unique: it depends on the structure that one wants find. Two examples, shown in Figure 3.1, are directions of high kurtosis or multi-modality.

In the language of the rest of this work, the “data” are the inputs to the neuron, \mathbf{x} , and the direction onto which the data is projected is simply the weight vector, \mathbf{w} . The projection value is the output of the neuron, y , which may or may not be passed through a rectifying sigmoid. We introduced

¹The cost function is often called an energy function. This can be misleading because the functions about which we speak do not satisfy conservation laws as the term “energy” would imply.

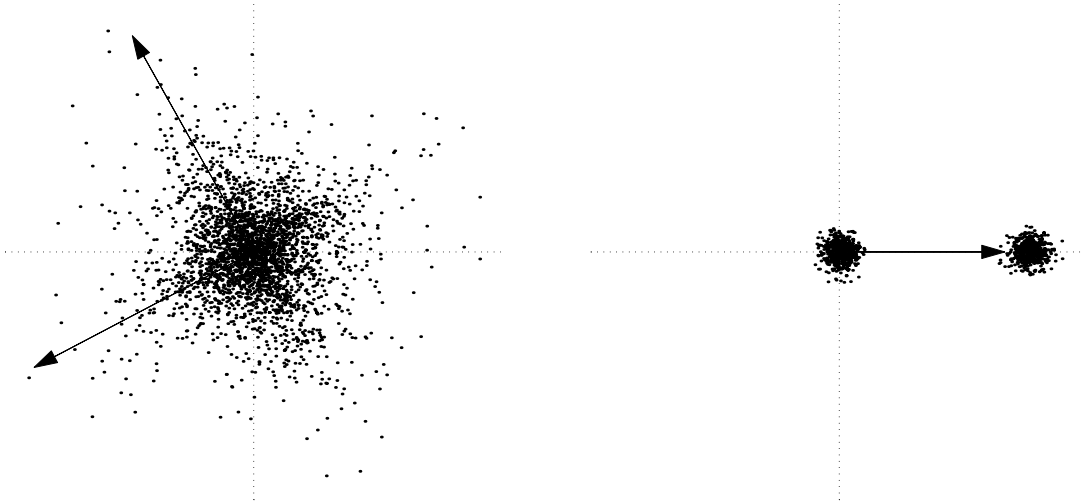


Figure 3.1: Example directions of significant structure. Directions of high kurtosis (left) and multimodality (right) are two examples of “interesting” directions in the space.

the sigmoid earlier for biological plausibility, but its function here depends on the form of the cost function, as we shall see. The cost function is then a function of the statistics of the output, y . The synaptic modification equation is achieved by doing a gradient ascent of this cost function with respect to the weights: the neuron seeks directions which maximize the cost function.

3.2 BCM and PCA

Intrator and Cooper (1992) introduced a cost function to measure the deviation from a Gaussian distribution in the form of multi-modality. The motivation for this functional form is given in Section A.5.2.

$$R_{\mathbf{w}}(\mathbf{x}) = \frac{1}{3}E[(\mathbf{x} \cdot \mathbf{w})^3] - \frac{1}{4}E^2[(\mathbf{x} \cdot \mathbf{w})^2] \quad (3.1)$$

This type of deviation, which is measured by the first three moments of the distribution, is particularly useful for finding clusters in high dimensional data through the search for multi-modality in the *projected distribution* rather than in the original high dimensional space. It is thus useful for classification or recognition tasks.

The synaptic modification equations which arise from this cost function are achieved using gradient ascent with respect to the weights, giving

$$\begin{aligned} \frac{d\mathbf{w}}{dt} = \frac{\partial}{\partial \mathbf{w}} R_{\mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{3}E[(\mathbf{x} \cdot \mathbf{w})^3] - \frac{1}{4}E^2[(\mathbf{x} \cdot \mathbf{w})^2] \right) \\ &= E[(\mathbf{x} \cdot \mathbf{w})^2 \mathbf{x}] - \underbrace{E[(\mathbf{x} \cdot \mathbf{w})^2]}_{\theta} E[(\mathbf{x} \cdot \mathbf{w}) \mathbf{x}] \\ &= E[(y^2 - y\theta) \mathbf{x}] \end{aligned}$$

$$\boxed{\frac{d\mathbf{w}}{dt} = E[\phi(y, \theta)\mathbf{x}]} \quad (3.2)$$

which is almost the same as the BCM equation (Equation 1.4). The difference is that Equation 3.2 is no longer stochastic: the random variable \mathbf{x} is averaged over. This equation has the same fixed points as Equation 1.4, and similar dynamics under certain conditions (see Intrator and Cooper (1992), Appendix A). The deterministic equation, however, can make analysis much easier in many cases. From it we can determine both the fixed points and the stability of those fixed points in many types of environments. The details of the fixed points for an environment made of linearly independent vectors is given in Section A.5.4.

Therefore, we can think of the BCM learning rule, introduced earlier, as an equation allowing the neuron to find structure in the inputs in the form of multi-modality. This very intuitive perspective on synaptic modification helps us understand some of what the neuron is communicating to other neurons; what the neuron considers to be important. It also gives us a clue about some of the important aspects of the environment, to which neurons could be optimized.

Consider the following cost function

$$R_{\mathbf{w}}(\mathbf{x}) = E \left[(\mathbf{x} \cdot \mathbf{w})^2 \left(1 - \frac{1}{2} \mathbf{w}^2 \right) \right] \quad (3.3)$$

This gives the learning rule

$$\begin{aligned} \frac{d\mathbf{w}}{dt} = \frac{\partial}{\partial \mathbf{w}} R_{\mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} E \left[(\mathbf{x} \cdot \mathbf{w})^2 \left(1 - \frac{1}{2} \mathbf{w}^2 \right) \right] \\ &= E \left[y\mathbf{x} \left(2 - \mathbf{w}^2 \right) - y^2 \mathbf{w} \right] \end{aligned}$$

This rule has the same fixed points as the PCA equation (Equation 1.3), so we can think of Equation 3.3 as a cost function for PCA. The dynamics of this equation would be different than the PCA rule, but we are not concerned with that right now.

Looking at the cost functions in Equations 3.1 and 3.3, makes explicit a statement made much earlier in Section 1.6, that the PCA rule is trying to maximize the *average squared activity* and the BCM rule is trying to maximize *bi-modality*. We now have the technique to address these questions explicitly. We can propose alternative cost functions which can attain orientation selectivity, and ask what is necessary for the development of orientation selectivity. This leads us naturally into a discussion of the structure of natural scenes.

3.2.1 Example with two dimensional model

The concept of projection pursuit can be seen in a simple two dimensional example, shown in Figure 3.2. Here we have chosen the input patterns so that both the solutions for BCM and for PCA fall on a unit circle. We can then look at the two cost functions as a function of the angle around this circle. From Figure 3.2, it is clear that the BCM cost function is maximized when the weight is equal to a solution

of the original BCM equation (Equation 1.4). The different weight vectors which maximize the PCA cost are the solutions of the original PCA equation (Equation 1.3).

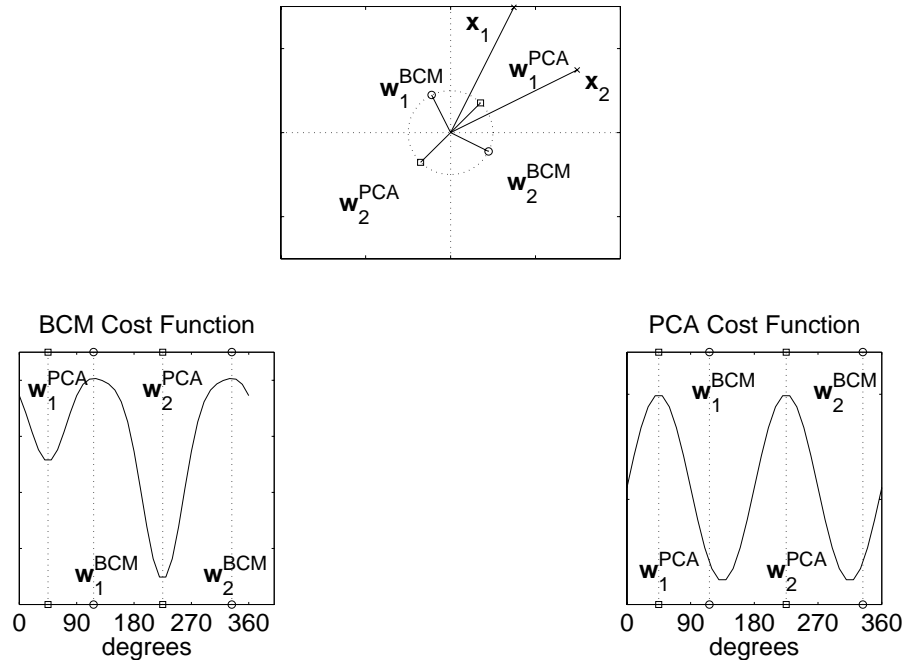


Figure 3.2: Example Cost Maximization in 2D. Shown (above) are sample input patterns, \mathbf{x}_1 and \mathbf{x}_2 , chosen so both the solutions for BCM and for PCA, $\mathbf{w}_i^{\text{BCM}}$ (circles) and $\mathbf{w}_i^{\text{PCA}}$ (squares), fall on a unit circle. Also shown (below) are the BCM cost function (left) and PCA cost function (right) as a function of angle around that unit circle. The angles where both sets of solutions fall are labeled on both cost function graphs. It is clear that the BCM cost function is maximized when the weight is equal to a solution of the original BCM equation (Equation 1.4). The different weight vectors which maximize the PCA cost are the solutions of the original PCA equation (Equation 1.3).

3.3 Output Distribution

We can see that BCM maximizes multi-modality by looking at the distribution of the projection values (outputs) on the final direction found (weight vector). In the two dimensional environment (Section 1.6), where we had linearly independent inputs, we did see that the direction found by BCM had a non-zero projection for only one of the input patterns, thus making the output distribution bi-modal. In fact, as shown in Appendix A.5.4, in an environment made of N linearly independent inputs, BCM finds a direction which has non-zero projection for only one of the input patterns. Natural scenes, however, are clearly *not* an environment of linearly independent inputs, so we need to explore what structure the BCM learning rule finds in natural scenes when it attains oriented receptive fields.

One way to see this is to look at the output distribution, as in Figure 3.3. Shown is the output distribution of a neuron, with a particular receptive field, over the natural scene environment, or in other words, the probability of finding a pattern to elicit a particular response. The distributions shown

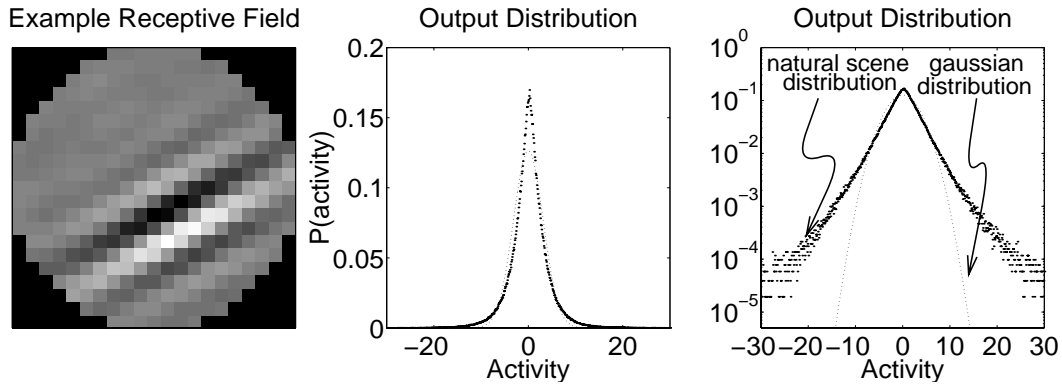


Figure 3.3: Output Distribution From Natural Scenes. Shown are an example receptive field (left) trained with BCM, the output distribution of a neuron with this receptive field on a linear scale (center) and a log scale (right). The distribution is achieved by performing a histogram of the output values (pre-sigmoid), over the entire environment. Also shown, with a dotted line, is a Gaussian distribution with the same variance.

are for the output values *before* the sigmoid, or simply $\mathbf{w} \cdot \mathbf{x}$. It is immediately apparent that this distribution is not bi-modal! It is, however, a distribution of a selective neuron because there are a few patterns that have very high response (best seen on the log scale), yet most of the patterns have near zero response.

In fact, the distribution looks very similar to a double exponential, or Laplace, distribution.

$$P(\text{activity}) = \frac{1}{2\lambda} e^{-|\text{activity}|/\lambda}$$

We might expect a non-multimodal distribution if there are no clusters in the data, but we don't know exactly what BCM is optimizing for in such an environment. In order to determine this, we introduce other cost functions, which then lead to synaptic modification rules to which we can compare BCM.

3.4 Other Cost Functions

Since the cost functions for BCM (Equation 3.1) and for PCA (Equation 3.3) depend on various moments of the output, it makes some sense to consider other measures based on these moments. Also, it is apparent from the output distribution that the important information lies in the *tails* of the distribution: the important events are rare and have high activity. We introduce four cost functions, inspired from statistics (Kendall and Stuart, 1977):

$$\text{Skewness 1: } S_1 = E[y^3]/E^{1.5}[y^2] \quad (3.4)$$

$$\text{Skewness 2: } S_2 = E[y^3] - E^{1.5}[y^2] \quad (3.5)$$

$$\text{Kurtosis 1: } K_1 = E[y^4]/E^2[y^2] - 3 \quad (3.6)$$

$$\text{Kurtosis 2: } K_2 = E[y^4] - 3E^2[y^2] \quad (3.7)$$

The measures of kurtosis depend primarily on the fourth moment of the data, and emphasize the tails of the distribution. The skewness measures depend primarily on the third moment of the data. Since the third moment is zero for any symmetric distribution, and the natural scene distribution is symmetric, any rule based on a cost function dependent strongly on the third moment must use a rectified output value, $y = \sigma(\mathbf{w} \cdot \mathbf{x})$, to break the symmetry. These rules include the BCM cost function and the skewness measures.

Since the BCM rule was originally proposed by specifying some general properties of BCM synaptic modification functions (Bienenstock et al., 1982), we will refer from now on to the specific form introduced by Intrator and Cooper (1992) as the quadratic form of BCM. BCM synaptic modification functions are characterized by a negative region for small activity, a positive region for large activity, and a threshold which moves and switches between the two regions. Several of the rules we consider have these properties, as we shall see, though they vary somewhat in their behavior.

In order to attain learning rules from the measures presented, we do a gradient ascent of those measures. We note that the rules fall into two classes based on the form of the modification function:

$$\text{Class 1:} \quad \frac{d\mathbf{w}}{dt} = \phi(y)\mathbf{x} \quad (\text{stable}) \quad (3.8)$$

$$\text{Class 2:} \quad \frac{d\mathbf{w}}{dt} = \phi(y)(\mathbf{x} - y\mathbf{w}) \quad (\text{unstable without decay term}) \quad (3.9)$$

where $\phi(y)$ is some function of the output of the cell. We will often write $\phi(y, \Theta_M)$ to make more explicit the dependence on the threshold.

The weight decay term, $-\phi(y)y\mathbf{w}$, in the second class of learning rules (Equation 3.9) is used only when the learning rule is not stable otherwise. This occurs when the rule either doesn't have a sliding threshold, or where the threshold does not depend strongly enough on the activity of the cell to enforce the stability.

We used the following learning rules, shown graphically in Figure 3.4.

Class 1

Quadratic BCM (Intrator and Cooper, 1992)

$$R_{\text{QBCM}} = \frac{1}{3}E[y^3] - \frac{1}{4}E^2[y^2] \quad (3.10)$$

$$\nabla_{\mathbf{w}} R_{\text{QBCM}} \equiv \frac{d\mathbf{w}}{dt} = \phi(y, \Theta_M)\mathbf{x} \quad (3.11)$$

$$\phi(y, \Theta_M) = y(y - \Theta_M)$$

$$\Theta_M \equiv E[y^2]$$

where $E[\cdot]$ is an average over the entire input environment.

Skewness 1

$$S_1 = E[y^3]/E^{1.5}[y^2] \quad (3.12)$$

$$\nabla_{\mathbf{w}} S_1 \equiv \frac{d\mathbf{w}}{dt} = \phi(y, \Theta_M) \mathbf{x} \quad (3.13)$$

$$\phi(y, \Theta_M) = y(y - \Theta_M)/\theta_S \quad (3.14)$$

$$\Theta_M \equiv E[y^3]/E[y^2]$$

$$\theta_S \equiv E^{1.5}[y^2]$$

where θ_S scales the overall modification equation, but does not affect any of the fixed points.

Kurtosis 1

$$K_1 = E[y^4]/E^2[y^2] \quad (3.15)$$

$$\nabla_{\mathbf{w}} K_1 \equiv \frac{d\mathbf{w}}{dt} = \phi(y, \Theta_M) \mathbf{x} \quad (3.16)$$

$$\phi(y, \Theta_M) = y(y^2 - \Theta_M)/\theta_S \quad (3.17)$$

$$\Theta_M \equiv E[y^4]/E[y^2]$$

$$\theta_S \equiv E^2[y^2]$$

Class 2

For the learning rules in Class 2 we *need* the decay term, $-\phi(y)y\mathbf{w}$, for stability. We saw this instability when we originally introduced the Hebb rule (Equation 1.2) and solved the instability by introducing a stabilizing term, which had the effect of normalizing the weights. The result was Oja's stabilized Hebb rule, or the PCA rule (Equation 1.3). One way to motivate the particular form of the decay term to normalize the weights is as follows. We write down the possibly unstable learning rule in *discrete* form

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n \quad (3.18)$$

where η is a small learning rate and $\phi(y_n)$ is some function of the activity of the cell at time n . For the PCA rule, $\phi(y_n)$ is simply y_n . We assume that n is large and that \mathbf{w}_n is very nearly normalized. We then *explicitly* normalize the weight \mathbf{w}_{n+1} , and expand to first order in η .

$$\mathbf{w}_{n+1} = \frac{\mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n}{\sqrt{(\mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n)^2}} \quad (3.19)$$

$$\begin{aligned} &= \frac{\mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n}{\sqrt{1 + 2\eta\phi(y_n)(\mathbf{w}_n \cdot \mathbf{x}_n) + O(\eta^2)}} \\ &\approx (\mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n)(1 - \eta\phi(y_n)y_n) \\ &= \mathbf{w}_n + \eta\phi(y_n)\mathbf{x}_n - \eta\phi(y_n)y_n\mathbf{w}_n + O(\eta^2) \end{aligned} \quad (3.20)$$

which leads us to the continuous case

$$\frac{d\mathbf{w}}{dt} = \phi(y)\mathbf{x} - \phi(y)y\mathbf{w} \quad (3.21)$$

For example, the Oja rule has $\phi(y) = y$, which gives us

$$\dot{\mathbf{w}} = y\mathbf{x} - y^2\mathbf{w} \quad (3.22)$$

The other rules follow in the same way.

PCA

$$R_{\text{PCA}} = \frac{1}{2}E[y^2] \quad (3.23)$$

$$\nabla_{\mathbf{w}}R_{\text{PCA}} \equiv \phi(y)\mathbf{x} \quad (3.24)$$

$$\frac{d\mathbf{w}}{dt} = \phi(y)(\mathbf{x} - y\mathbf{w}) \quad (3.25)$$

$$\phi(y) = y$$

Skewness 2

$$S_2 = E[y^3] - E^{1.5}[y^2] \quad (3.26)$$

$$\nabla_{\mathbf{w}}S_2 \equiv \phi(y)\mathbf{x} \quad (3.27)$$

$$\frac{d\mathbf{w}}{dt} = \phi(y, \Theta_M)(\mathbf{x} - y\mathbf{w}) \quad (3.28)$$

$$\phi(y, \Theta_M) = y(y - \Theta_M)$$

$$\Theta_M \equiv E^{0.5}[y^2]$$

Kurtosis 2

$$K_2 = E[y^4] - 3E^2[y^2] \quad (3.29)$$

$$\nabla_{\mathbf{w}}K_2 \equiv \phi(y)\mathbf{x} \quad (3.30)$$

$$\frac{d\mathbf{w}}{dt} = \phi(y, \Theta_M)(\mathbf{x} - y\mathbf{w}) \quad (3.31)$$

$$\phi(y, \Theta_M) = y(y^2 - \Theta_M)$$

$$\Theta_M \equiv 3E[y^2]$$

In all of the above learning rules the value of the activity of the cell is given by

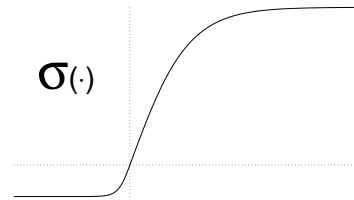
$$y = \sigma(\mathbf{w} \cdot \mathbf{x}) \quad (3.32)$$

where $\sigma(\cdot)$ is a non-linear function of its argument. For PCA we use, here, a non-linear variant, where the “sigmoid” is a polynomial function, such as $\sigma(\mathbf{w} \cdot \mathbf{x}) = (\mathbf{w} \cdot \mathbf{x})^3$. The non-linear PCA rule is not

biologically plausible but has some nice mathematical properties(Oja, 1995). One of those properties is that it depends on more than just second order correlations.

All of the other rules use the more realistic function

$$\sigma(\mathbf{w} \cdot \mathbf{x}) = \begin{cases} \sigma^+ \tanh(\mathbf{w} \cdot \mathbf{x}/\sigma^+) & \text{for } \mathbf{w} \cdot \mathbf{x} \geq 0 \\ \sigma^- \tanh(\mathbf{w} \cdot \mathbf{x}/\sigma^-) & \text{for } \mathbf{w} \cdot \mathbf{x} < 0 \end{cases} \quad \mathbf{\Sigma}(\cdot) \quad (3.33)$$



where σ^+ and σ^- set the maximum and minimum activity levels, respectively. Since $y = 0$ represents spontaneous activity, and a cell can fire much farther above spontaneous than it can below, $|\sigma^-| < |\sigma^+|$. In the simulations we take $\sigma^- = -1$ and $\sigma^+ = 50$. For all rules the ensemble average activity is approximated by the temporal average:

$$E[y^n(t)] \approx \frac{1}{\tau} \int_{-\infty}^t y^n(t') e^{-(t-t')/\tau} dt'$$

3.5 The Effect of Noise on Monocular Deprivation

In this section we show the effect of noise on the dynamics of monocular deprivation, for the different learning rules presented in Section 3.4. Some of the results generalize to the other deprivation paradigms, but there are complications which make the picture less simple. Monocular deprivation gives robust results for all of the learning rules, and can be understood with some analysis (Section 3.6).

The half-times are measured as before, and the only variable explored is the noise level. Since the different learning rules have different half-times in the different parameter regimes, we set the half-time to 1 for unit noise variance in all cases. Thus, we look at the relative changes when increasing or decreasing the noise level. All of the half-times presented were averaged over several simulations with the same learning rule, at the same noise level. The results are shown in Figure 3.5.

Shown is the half-time for the loss of response to the closed eye in monocular deprivation, as a function of the closed eye noise level, for the different learning rules. The half-times are scaled so that the half-time is set to 1 at a noise level of unit variance, $\sigma = 1$. The natural scene input also has a unit variance. It is readily apparent that the Class 1 learning rules have a *faster* loss of response to the closed eye, for *more noise* into the closed eye. The Class 2 learning rules have the qualitatively *opposite* behavior, though much less pronounced. In order to achieve similar changes in the Class 2 rules, one needs to use much higher noise levels, $\sigma \geq 5$, which is over five times the variance of the environment. At these unrealistically high noise levels two of the Class 1 rules, K_1 and S_1 , also show increases in the MD half-time.

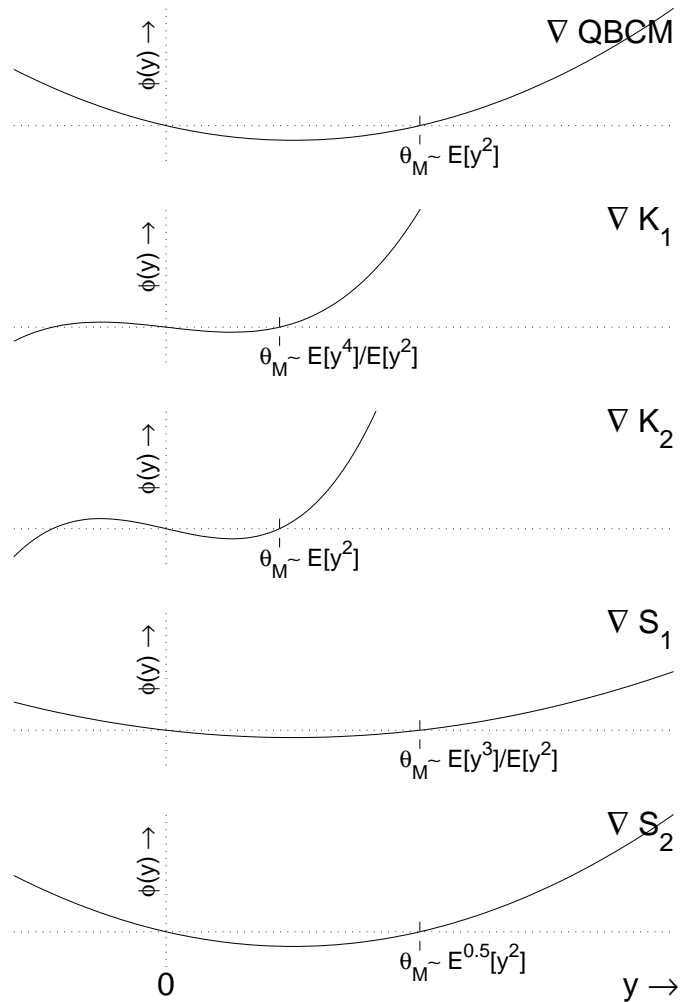


Figure 3.4: Example learning rules, as a function of the output of the cell. Shown are Quadratic BCM, Kurtosis 1 and 2, and Skewness 1 and 2.

3.5.1 Experimental Verification

We have seen that these two classes of learning rules predict the opposite, qualitative dependence on the noise into the closed eye. One method for testing this would be to experimentally alter the levels of activity in the closed eye.² Injecting TTX into the eye (which silences the neuronal activity), using a translucent patch or a dark patch would result in different activity levels in the closed eye.

An experiment was recently performed (Rittenhouse et al., 1998), where monocular deprivation was performed on two groups of kittens. One group using lid suture to deprive the eye, and the other group using TTX in the retina to eliminate the retinal activity. Rules from Class 1, which include BCM, would predict that a *larger* ocular dominance shift will occur in the lid suture animals than the TTX animals, over the same duration, because the lid suture would yield a larger activity variance from the closed eye. Rules from Class 2 would predict a smaller OD shift in the lid suture animals over the

²It is presumed that altering the levels of the spontaneous activity in the closed eye will also alter the activity levels in LGN.

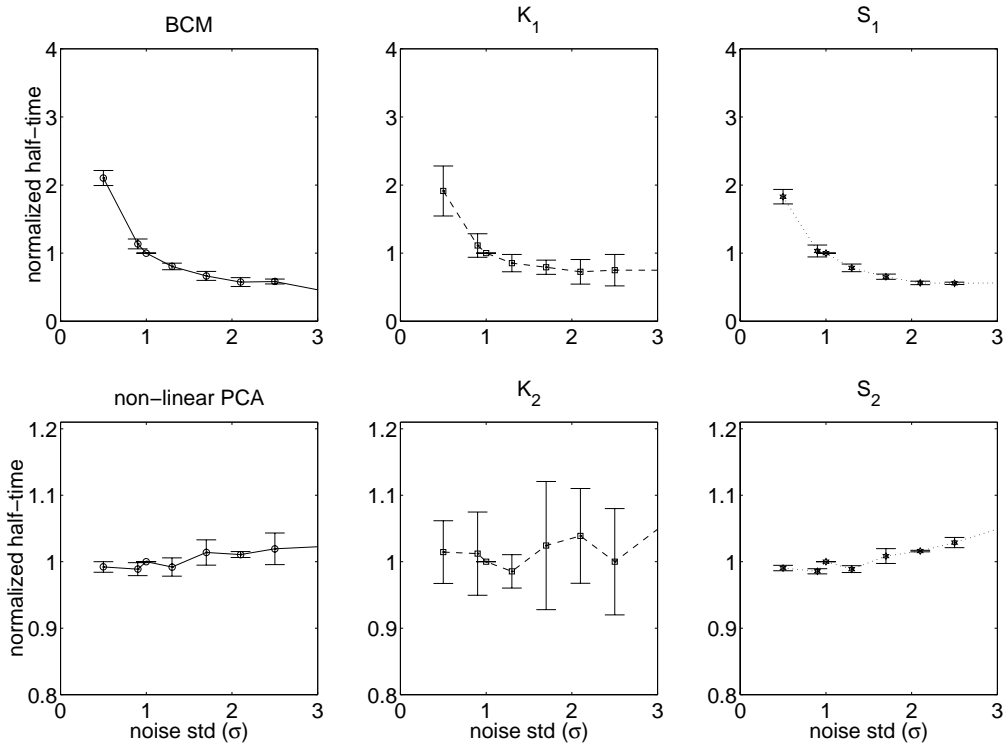


Figure 3.5: The half-time for the loss of response to the closed eye in monocular deprivation, as a function of the closed eye noise level (structured input variance=1). The half-times are scaled so that the half-time is set to 1 at a noise level of unit variance, $\sigma = 1$. The rules in Class 1 (QBCM, K_1 , and S_1) are shown above, and the rules in Class 2 (non-linear PCA, K_2 , and S_2) are shown below. The Class 1 learning rules have a *faster* loss of response to the closed eye, for *more noise* into the closed eye. The Class 2 learning rules have the opposite behavior, though much less pronounced.

same duration. The observation from the experiment was an enhanced ocular dominance shift in the lid suture animals, consistent with the Class 1 learning rules.

Greuel and Singer (1987) have also performed this experiment, but without observing significant differences. It is worthwhile to point out that their experiment used very few animals, and *all* of the differences that are observed are consistent with the Rittenhouse experiment, but not at statistically significant levels.

A related experiment was performed (Chapman et al., 1986; Greuel et al., 1987), where several groups of kittens were compared. One group was monocularly deprived with lid suture. The other two groups had *both* eyes deprived, but in different ways. One of these had TTX in one eye, and the other eye was sutured, while the other group had TTX in one eye, and was kept in the dark. The result was an observed ocular dominance shift towards the more active eye. The strongest shift was observed for the lid suture monocular deprivation group, and the smallest shift was observed for the TTX-dark group. Such an observation would seem to be inconsistent with Rittenhouse (1998), Greuel and Singer (1987) who reported no significant differences between the groups, and also Blakemore (1976) and Wiesel and Hubel (1965) mentioned in the introduction. If the noise is not completely uncorrelated, then there may be other implications which may at least partially explain these differences. This work is left for the future.

In the next section, we present some analysis which explains the origin of the difference between two rules, one from each class. First, however, we need a quick note about the other deprivation experiments.

3.5.2 Binocular Deprivation and Reverse Suture

It turns out that monocular deprivation is the most robust of the deprivation experiments. It is for this reason that experiments with monocular deprivation are some of the best ways to examine the properties of the learning rules. Binocular deprivation, for instance, can have some theoretical properties which make it not as valuable a tool. In Section 3.6.4 shows how binocular deprivation is particularly sensitive on the sigmoid on the output, which reflects the assumption of activity measured above spontaneous. This type of sensitivity is not useful, when one wishes to propose experiments.

Reverse suture, similarly, is sensitive both on the sigmoid and on the length of monocular deprivation preceding it. The dynamics of reverse suture and binocular deprivation, don't depend primarily on the learning rule but on the other details of the model. They are therefore not as useful in determining which types of learning rules are experimentally verified. These sensitivities to other parts of the model may allow the formulation of different experiments to test the effects of the deprivation, but that must be left for the future.

3.6 A Simpler Environment

Many of the deprivation effects described so far can be seen in a simple model of the input environment. The primary motivation for the simple model comes from the observation that these learning rules find directions of high kurtosis. Essentially, the procedure will be

- *assume* a distribution on the inputs which has an analytically simple form. Although we are going to present low dimensional distributions, we are not really interested in these low dimensional inputs directly, but are concerned with *projections* with these distributions. Thus, the low dimensional distributions represent the projections along particular directions in a high dimensional space, and therefore can be used to represent a higher dimensional environment such as the natural scene environment. To model the natural environment, we will use the Laplace (or double exponential) distribution

$$f(x) = \frac{1}{2\lambda} e^{-|x|/\lambda}$$

which has a positive kurtosis. For the noise environment we will either use a Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$$

which has zero kurtosis, or a uniform distribution

$$f(x) = \frac{1}{2a} \text{ in the range } [-a..a]$$

which has negative kurtosis.

- calculate the distribution on the *output* of the neuron
- from the output distribution, calculate the moments and thus the various cost functions (K_1 , R_{QBCM} , etc.)
- calculate the fixed points and dynamics for the weight vector

We will explore two rules, BCM and K_2 , because these are both analytically tractable, depend on higher order statistics, and are examples of each class of learning rule introduced in Section 3.4.

3.6.1 One Dimensional Example: Laplace

As an example, we take a one dimensional neuron (a neuron with one synapse), and assume that the input distribution is a Laplace (or double exponential) distribution. Again, the distributions presented here represent the distributions of projections in a higher dimensional space.

$$f_x(x) = \frac{1}{2\lambda} e^{-|x|/\lambda} \tag{3.34}$$

The output of the neuron is simply

$$y = x \cdot w \text{ (note: no sigmoid)} \quad (3.35)$$

$$z \equiv \sigma(y) \equiv \begin{cases} y & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.36)$$

where we have introduced the notation y for the *pre-sigmoid* output, and z as the *post-sigmoid* output. The sigmoid we are using here has a lower value of 0 and no upper limit. This choice is purely to make the analysis tractable, and becomes inappropriate in some cases (see Section 3.6.4), but does lead to a significantly better understanding of deprivation.

To calculate the distribution of the outputs, y and z , we use some theorems found in many textbooks on statistics (see Appendix A.9.1).

$$f_y(y) = \frac{1}{2\lambda|w|} e^{-|y/w|/\lambda} \quad (3.37)$$

$$f_z(z) = \begin{cases} f_y(z) & \text{if } z > 0 \\ \frac{1}{2}\delta(z) & \text{if } z = 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (3.38)$$

We will assume that w is positive, so we don't have to carry the absolute value through the calculations. Note that that any solution we find, say $w = w_o$, *must* be positive, because it represents $|w| = w_o$, but that the solution will really be $w = \pm w_o$.

To calculate the values of, say, R_{QBCM} and K_2 , we calculate the moments of the distribution $f_z(z)$.

$$\begin{aligned} E[z^2] &= \int_{-\infty}^{\infty} z^2 f_z(z) dz \\ &= \frac{1}{2\lambda w} \int_0^{\infty} z^2 e^{-z/w\lambda} dz \\ &= w^2 \lambda^2 \end{aligned} \quad (3.39)$$

$$\begin{aligned} E[z^3] &= \int_{-\infty}^{\infty} z^3 f_z(z) dz \\ &= \frac{1}{2\lambda w} \int_0^{\infty} z^3 e^{-z/w\lambda} dz \\ &= 3w^3 \lambda^3 \end{aligned} \quad (3.40)$$

$$\begin{aligned} E[z^4] &= \int_{-\infty}^{\infty} z^4 f_z(z) dz \\ &= \frac{1}{2\lambda w} \int_0^{\infty} z^4 e^{-z/w\lambda} dz \\ &= 12w^4 \lambda^4 \end{aligned} \quad (3.41)$$

which gives us

$$\begin{aligned} R_{\text{QBCM}} &= \frac{1}{3}E[z^3] - \frac{1}{4}E^2[z^2] \\ &= w^3\lambda^3 - \frac{1}{4}w^4\lambda^4 \end{aligned} \quad (3.42)$$

$$\begin{aligned} K_2 &= E[z^4] - 3E[z^2] \\ &= 9w^4\lambda^4 \end{aligned} \quad (3.43)$$

Clearly, K_2 is unstable (as we saw in Section 3.4) so we have to add the constraint that $|w|^2 = 1$. We have already seen how this constraint can be added in a local fashion, so we do not require a local version of the constraint here, where we simply restrict the weights to the unit sphere. With this constraint, the one dimensional solution is trivially $w = 1$.

The equation for R_{QBCM} can be maximized with respect to the weights,

$$\begin{aligned} \frac{\partial R_{\text{QBCM}}}{\partial w} &= 3w^2\lambda^3 - w^3\lambda^4 \\ &= w^2\lambda^3(3 - w\lambda) = 0 \end{aligned} \quad (3.44)$$

$$\Rightarrow w = 0, \frac{3}{\lambda} \quad (3.45)$$

and we can evaluate the stability of any fixed points found.

$$\begin{aligned} \frac{\partial^2 R_{\text{QBCM}}}{\partial w^2} &= 6w\lambda^3 - 3w^2\lambda^4 \\ &= 3w\lambda^3(2 - w\lambda) \end{aligned} \quad (3.46)$$

$$= \begin{cases} \frac{\partial^2 R_{\text{QBCM}}}{\partial w^2}|_{w=0} = 0 & \text{unstable} \\ \frac{\partial^2 R_{\text{QBCM}}}{\partial w^2}|_{w=3/\lambda} < 0 & \text{stable} \end{cases} \quad (3.47)$$

There are a few things to notice about these solutions. Although the BCM cost function was introduced to find bi-modality, or clusters, the cost function finds kurtotic projections when no clusters exist. This is more striking in the case of strabismus, covered later in Section 3.6.6. The length of the weight vector also gives extra information about the distribution, namely the inverse of the environmental parameter λ . This is a common occurrence for the BCM learning. In the one dimensional case presented in Section 1.2, the fixed point weight was large when the input was the small, in order to keep the total output fixed (and thus the threshold, θ , fixed).

We can also get some dynamics by setting dw/dt equal to the gradient of the cost function.

$$\begin{aligned} \frac{dw}{dt} &= \frac{\partial R_{\text{QBCM}}}{\partial w} \\ &= 3w^2\lambda^3 - w^3\lambda^4 \end{aligned} \quad (3.48)$$

$$\Rightarrow \frac{1}{9\lambda^3 w} - \frac{1}{9\lambda^2} \log\left(\frac{\lambda w - 3}{w}\right) = t + \text{const} \quad (3.49)$$

Though Equation 3.49 gives us the dynamics in this one dimensional case, it is pretty opaque. With some approximations it might give us some information, but we do not concern ourselves with this further, because we mean it merely as an example of *how* to get dynamics information from this procedure.

3.6.2 Two Dimensions

We introduce a low dimensional, two-eye model, by having one dimension for each eye. This can cause BCM to have some peculiar behavior, depending on the distributions used. More specifically, if the output distributions are not sparse, or in other words, the neuron is not selective, then the fixed points are non-physiological. Figure 3.6 shows an example of this. Shown is a sample taken from a 1D-2 eye input distribution where both eyes receive the same input, corrupted by noise. In the situations where the neuron cannot find a selective fixed point (i.e. those fixed points where *most* responses are around zero, and a few responses have significant non-zero value), the fixed points found are non-physiological: the cell becomes responsive to an odd combination of left and right eyes. In the cases where the neuron can become selective, the fixed points are physiological: both eyes respond equally.

Since we are working with Laplace distributions, which are sparse, we do not have this difficulty. Therefore, we will continue to use the 1D-each eye model for simplicity, keeping in mind the exceptions where it becomes problematic.

Our goal is to look at the dynamics of the two eye case, using the simple model of the structured and deprived environments. For two eyes we have the following

$$y = w_1 x^l + w_2 x^r \quad (3.50)$$

which, if the eyes see the same thing, reduces to the following

$$y = (w_1 + w_2)x \quad (3.51)$$

$$= w^{\text{eff}}x \quad (3.52)$$

We notice that we can use our solutions for the 1D case to find w^{eff} , but that the solution is not uniquely defined. *Any* solution satisfying $w_1 + w_2 = w^{\text{eff}}$ would work, so we have a continuum of solutions. If we look at the change in the *difference* between the left and right eye weights

$$\frac{dw_1}{dt} - \frac{dw_2}{dt} = E \left[\phi(y)(x^l - x^r) \right] \quad (3.53)$$

$$= 0 \quad (3.54)$$

it is clear that the difference in the weights does not change, so if our initial weights are small, $|\mathbf{w}_o| \ll 1$, the final weights for left and right eye will be approximately the same.

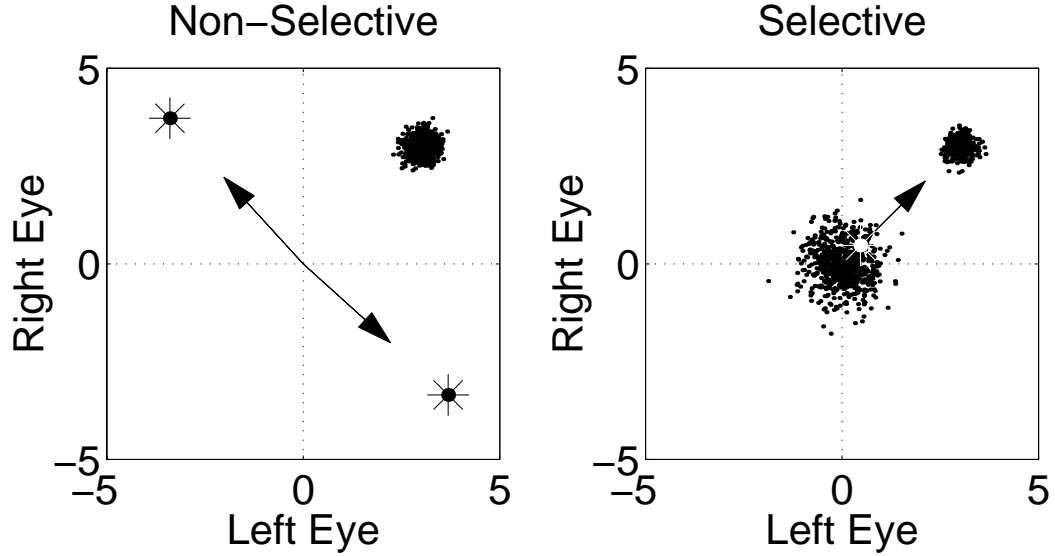


Figure 3.6: Odd low dimensional behavior of BCM. Shown is a sample taken from a 1D-2 eye input distribution where both eyes receive the same input, corrupted by noise. Input points are shown as small dots, the BCM fixed points are shown with asterisks and their direction shown with an arrow for clarity. All of the input falls along the (left eye)=(right eye) line. On the left is shown a situation where the neuron can find no selective fixed points (i.e. those fixed points where *most* responses are around zero, and a few responses have significant non-zero value). In these cases, the fixed points found (shown in asterisks) are non-physiological: the cell becomes responsive to an odd combination of left and right eyes. The one to the right shows a case where the neuron becomes selective, and the fixed point is physiological: both eyes respond equally.

Thus, for normal rearing using BCM in the Laplace environment, we have

$$w_{\text{eff}} \equiv w_1 + w_2 \quad (3.55)$$

$$= \pm 3/\lambda \quad (3.56)$$

$$\begin{aligned} \mathbf{w} &\equiv \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \\ &= \begin{pmatrix} \pm 3/2\lambda - \frac{\delta}{2} \\ \pm 3/2\lambda + \frac{\delta}{2} \end{pmatrix} \end{aligned} \quad (3.57)$$

where $\delta \equiv (w_2)_o - (w_1)_o$ is the difference between the initial conditions. This is shown in Figure 3.7A.

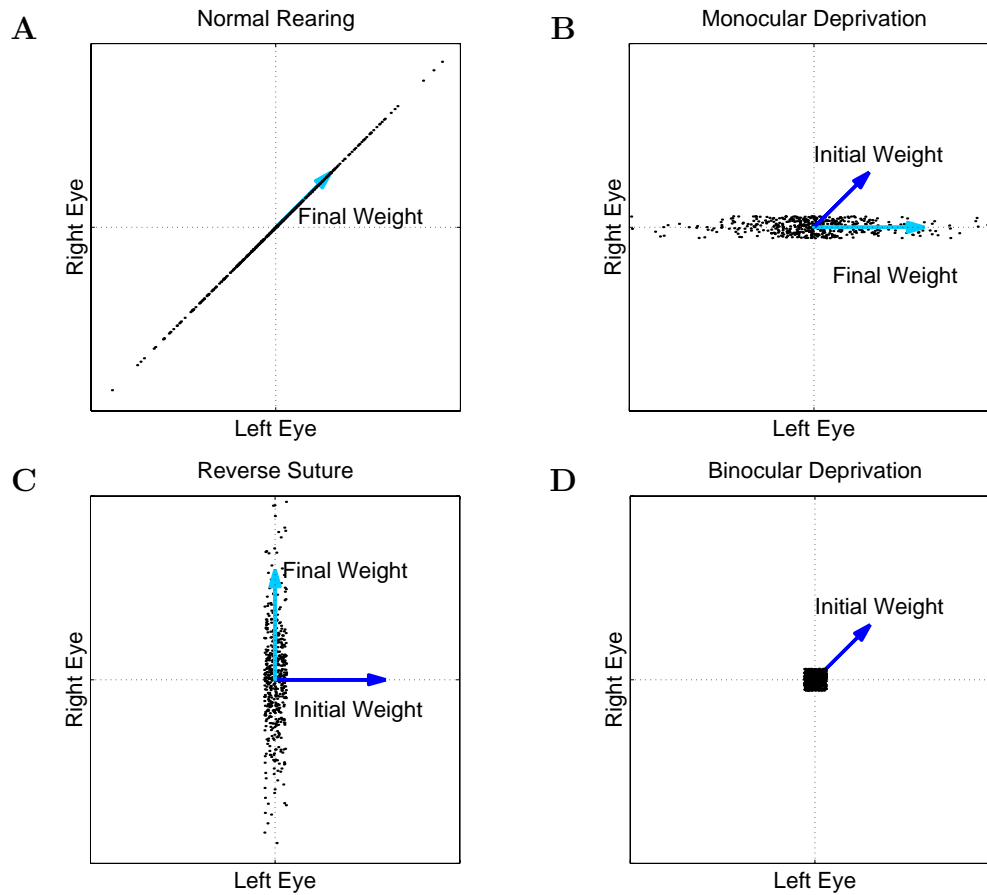


Figure 3.7: 2D Deprivation. Normal rearing (A) is modeled with numbers chosen from a Laplace distribution presented to each eye. The eyes see identical inputs, so the input distribution (samples shown with dots) lies along the 45° line. The initial weights are small, and the final weights are equal for each eye (Equation 3.57 for BCM and Equation 3.58 for kurtosis). Monocular deprivation (B) is modeled with Laplace numbers to one eye, representing the structure from the natural environment, and uniform (or Gaussian) numbers to the other eye, representing the noise of activity from the closed eye. The initial weight is from normal rearing, and the final weight is in the direction of the open eye: the cell comes to respond only to the open eye. Reverse suture (C) is modeled with Laplace numbers to one eye and uniform (or Gaussian) numbers to the other eye, following monocular deprivation. Binocular deprivation (D) is modeled using uniform (or Gaussian) noise presented to both eyes, following normal rearing.

For K_2 , with the normalization condition, the normal rearing fixed point is simply

$$\begin{aligned} \mathbf{w} &\equiv \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \\ &= \begin{pmatrix} \pm \frac{\sqrt{2}}{2} - \frac{\delta}{2} \\ \pm \frac{\sqrt{2}}{2} + \frac{\delta}{2} \end{pmatrix} \end{aligned} \quad (3.58)$$

Deprivation

In order to model, say, monocular deprivation we start from the fixed point of normal rearing and alter the distribution, presenting uniform (or Gaussian) noise to one of the eyes, as shown in Figure 3.7B. Since both eyes don't see the same thing, the weights from both eyes need not remain equal: the response of the cell to one, or the other eye can change. We will see that the neuron comes to respond to that eye which has structure, or in our case the one with the highly kurtotic input distribution.

The basic idea, setting aside the math for a second, is that the deprivation effects occur because of a competition between the input distributions. When presented with two different distributions, the neuron chooses to respond more strongly to the more kurtotic distribution and respond more weakly to the less kurtotic distribution (see Figure 3.7). This is why deprivation doesn't depend exclusively on the variance of the input (see Section 2.3.3), and why binocular deprivation has a slower time course than monocular deprivation.

In the following sections, we explore the noise dependence of deprivation using this simple environment. We demonstrate analytically the two classes that we have seen in the simulations. We also show a somewhat surprising result for the role of the sigmoid on binocular deprivation, which highlights one of the assumptions we have been making throughout this work.

3.6.3 Monocular Deprivation

In this section we present the simple environment model of monocular deprivation, for BCM and K_2 , and show how the noise dependence we have seen in the simulations arises from a competition between distributions.

In this model, we have the following input densities

$$f_{x_1}(x_1) = \frac{1}{2\lambda} e^{-|x_1|/\lambda} \quad (3.59)$$

$$f_{x_2}(x_2) = \frac{1}{2a} \text{ in the range } [-a..a] \quad (3.60)$$

where we are using x_1 and x_2 for the left and right eye inputs, respectively.

The output of the cell is broken up into a sum of two one dimensional outputs, each of which we

know the distribution.

$$\begin{aligned} y &= \mathbf{x} \cdot \mathbf{w} = x_1 w_1 + x_2 w_2 \\ &\equiv y_1 + y_2 \text{ (note: no sigmoid)} \\ z &\equiv \sigma(y) \end{aligned}$$

The densities of y_1 and y_2 are simply given by

$$f_{y_1}(y_1) = \frac{1}{2\lambda w_1} e^{-|y_1|/w_1\lambda} \quad (3.61)$$

$$f_{y_2}(y_2) = \frac{1}{2w_2 a} \text{ in the range } [-w_2 a..w_2 a] \quad (3.62)$$

from which we can calculate the total output density and the moments.

We are still using the assumption that $w_i = |w_i|$, so any solution for w_i that we find must be positive. We also have to remember that any positive solution found for w_i , the negative of it is also a solution. Since the fixed points will be intuitively clear, this is not much of a problem.

We calculate the distribution of the output y (and thus the distribution of z) by taking the convolution of the individual densities of y_1 and y_2 , since they are independent (see Appendix A.9.1).

$$\begin{aligned} f_y(y) &= \int_{-\infty}^{\infty} f_{y_1}(y - y_2) f_{y_2}(y_2) dy_2 \\ &= \frac{1}{4\lambda a w_1 w_2} \int_{-aw_2}^{aw_2} e^{-|y-y_2|/w_1\lambda} dy_2 \\ &\equiv \frac{1}{4\lambda a w_1 w_2} \int_{-aw_2-y}^{aw_2-y} e^{-|u|/w_1\lambda} du \end{aligned} \quad (3.63)$$

assume, for now, that $y > 0$.

$$\begin{aligned} \text{if } y < aw_2: f_y(y) &= \frac{1}{4\lambda a w_1 w_2} \left(\int_{-aw_2-y}^0 e^{u/w_1\lambda} du + \int_0^{aw_2-y} e^{-u/w_1\lambda} du \right) \\ &= \frac{2\lambda w_1 - \lambda w_1 e^{(-aw_2-y)/\lambda w_1} - \lambda w_1 e^{-(aw_2+y)/\lambda w_1}}{4\lambda a w_1 w_2} \end{aligned} \quad (3.64)$$

$$\begin{aligned} \text{if } y > aw_2: f_y(y) &= \frac{1}{4\lambda a w_1 w_2} \int_{-aw_2-y}^{aw_2-y} e^{u/w_1\lambda} du \\ &= \frac{\lambda w_1 e^{(aw_2-y)/\lambda w_1} - \lambda w_1 e^{-(aw_2+y)/\lambda w_1}}{4\lambda a w_1 w_2} \end{aligned} \quad (3.65)$$

which generalizes for the $y < 0$ case in the following

$$\text{if } |y| < aw_2: f_y(y) = \frac{1}{4aw_2} \left(2 - e^{(-aw_2+|y|)/\lambda w_1} - e^{(-aw_2-|y|)/\lambda w_1} \right) \quad (3.66)$$

$$\text{if } |y| > aw_2: f_y(y) = \frac{1}{4aw_2} \left(e^{(aw_2-|y|)/\lambda w_1} - e^{-(aw_2+|y|)/\lambda w_1} \right) \quad (3.67)$$

$$f_z(z) = \begin{cases} f_y(z) & \text{if } z > 0 \\ \frac{1}{2}\delta(z) & \text{if } z = 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (3.68)$$

With moments

$$E[z^2] = \frac{1}{6}a^2w_2^2 + \lambda^2w_1^2 \quad (3.69)$$

$$E[z^3] = \frac{1}{8aw_2} \left(12\lambda^2w_1^2a^2w_2^2 + a^4w_2^4 + 24\lambda^4w_1^4(1 - e^{-aw_2/\lambda w_1}) \right) \quad (3.70)$$

$$E[z^4] = \frac{1}{10}a^4w_2^4 + 12\lambda^4w_1^4 + 2\lambda^2a^2w_1^2w_2^2 \quad (3.71)$$

Kurtosis

Kurtosis is simply calculated from the moments, and converted to polar coordinates.

$$K_2 = \frac{1}{60}a^4w_2^4 + 9\lambda^4w_1^4 + \lambda^2a^2w_1^2w_2^2 \quad (3.72)$$

$$= \frac{1}{60}a^4R^4 + R^4 \cos^4(\theta) \left(9\lambda^4 - \lambda^2a^2 + \frac{1}{60}a^4 \right) + R^4 \cos^2(\theta) \left(-\frac{1}{30}a^4 + a^2\lambda^2 \right) \quad (3.73)$$

Although the subtractive form of kurtosis, K_2 , has the property

$$K_2(x_1 + x_2) = K_2(x_1) + K_2(x_2) \quad (3.74)$$

if x_1 and x_2 are *independent* random variables, we do not see this in Equation 3.72. It is clearly not the simple sum of a function of w_1 and a function of w_2 . This is due entirely by the sigmoid. We do not use this possible simplification here, in order to maintain a consistent procedure. We mention it only to avoid possible confusion.

We then take derivatives with respect to θ , evaluate at $R = 1$, and look at the fixed points.

$$\frac{\partial K_2}{\partial \theta} = \left(-36\lambda^4 + 4\lambda^2a^2 - \frac{1}{15}a^4 \right) \sin(\theta) \cos^3(\theta) + \left(\frac{1}{15}a^4 - 2\lambda^2a^2 \right) \sin(\theta) \cos(\theta) \quad (3.75)$$

which has fixed points at $\theta = 0$ and $\theta = \pi/2$. The $\theta = 0$ fixed point corresponds to the weight equal to zero for the input receiving uniform noise (w_2 , or the closed eye), or in other words, the cell losing responsiveness to the closed eye. This fixed point is stable for $a < \lambda 3\sqrt{2}$, or small noise compared to the structure in the environment, which makes physiological sense. We are using “structure” here as synonymous with “high λ ”, because the structure in the natural scenes arises from directions of high kurtosis. The $\theta = \pi/2$ is unstable in this noise regime, so the neuron will lose responsiveness to the closed eye starting from the initial conditions from normal rearing, $\mathbf{w} = (\sqrt{2}/2 \ \sqrt{2}/2)$, or in polar, $\theta = \pi/4$.

The main concern we have is with dynamics. How *quickly* does the cell lose responsiveness, and how does it depend on the *noise level*. To do this we set, as before, Equation 3.75 equal to $d\theta/dt$ to get the dynamics. We expand $\cos(\theta)$ and $\sin(\theta)$ in the resulting equation about the initial point $\theta = \pi/4$, drop high order terms in the expansion, and introduce a low noise approximation.

$$\begin{aligned}\cos(\pi/4 + x) &\approx \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}x \\ \sin(\pi/4 + x) &\approx \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}x \\ \frac{dx}{dt} &\approx \left(\frac{1}{30}a^4 + 18\lambda^4 - 2\lambda^2a^2\right)x + \left(-9\lambda^4 + \frac{1}{60}a^4\right)\end{aligned}\quad (3.76)$$

$$\equiv \xi_1x + \xi_2 \quad (3.77)$$

which has an exponential solution. Plugging in $x(0) = 0$, we get

$$x(t) = \frac{\xi_2}{\xi_1} (\exp(\xi_1 t) - 1) \quad (3.78)$$

The rate of the cutoff from the closed eye is determined by the value of ξ_1 . The larger the value of ξ_1 , the *faster* the cutoff of response to the closed eye. Since the noise to the closed eye is an experimentally controllable parameter, we want to know how the rate depends on the noise level, a . This is simply done by looking at the sign of the change in ξ_1 with respect to a .

$$\begin{aligned}\frac{\partial \xi_1}{\partial a} &= \frac{2a}{15}(a^2 - 30\lambda^2) \\ &< 0 \text{ if } a^2 < 30\lambda^2\end{aligned}\quad (3.79)$$

which is certainly true if $a < \lambda 3\sqrt{2}$, which was the requirement for stability of the biologically reasonable solution. This means that the noise tends to *slow down* the decay of response to the closed eye, as we have seen in the simulations.

BCM

The BCM equations depend not only on the *direction* but the *magnitude* of the weights. We can't simply convert to polar, and reduce the dimensionality of the problem. The cost function is calculated from the moments of the Laplace-Uniform environment.

$$\begin{aligned}R_{\text{QBCM}} &= \frac{1}{2}\lambda^2w_1^2aw_2 + \frac{1}{24}a^3w_2^3 + \lambda^4w_1^4 \left(1 - e^{-aw_2/\lambda w_1}\right) / aw_2 \\ &\quad - \frac{1}{144}a^4w_2^4 - \frac{1}{12}a^2w_2^2\lambda^2w_1^2 - \frac{1}{4}\lambda^4w_1^4\end{aligned}\quad (3.80)$$

If $a \ll \lambda$, (and we assume that the weights themselves are not particularly large), then we can expand the exponential in the risk function (Equation 3.80) to powers of $(a/\lambda)^2$.

$$\begin{aligned}R_{\text{QBCM}} &= \frac{1}{2}\lambda^2w_1^2aw_2 + \frac{1}{24}a^3w_2^3 + \frac{\lambda^4w_1^4}{aw_2} - \lambda^4w_1^4 \left(1 - \frac{aw_2}{\lambda w_1} + \frac{a^2m_2^2}{2\lambda^2m_1^2}\right) / aw_2 \\ &\quad - \frac{1}{144}a^4w_2^4 - \frac{1}{12}a^2w_2^2\lambda^2w_1^2 - \frac{1}{4}\lambda^4w_1^4\end{aligned}\quad (3.81)$$

$$= \frac{1}{24}a^3w_2^3 + \lambda^3w_1^3 - \frac{1}{144}a^4w_2^4 - \frac{1}{12}a^2w_2^2\lambda^2w_1^2 - \frac{1}{4}\lambda^4w_1^4 \quad (3.82)$$

The modification equations are then

$$\frac{dw_1}{dt} \equiv \frac{\partial R_{\text{QBCM}}}{\partial w_1} = 3\lambda^3 w_1^2 - \frac{1}{6} a^2 \lambda^2 w_1 w_2^2 - \lambda^4 w_1^3 \quad (3.83)$$

$$\frac{dw_2}{dt} \equiv \frac{\partial R_{\text{QBCM}}}{\partial w_2} = \frac{1}{8} a^3 w_2^2 - \frac{1}{111} a^4 w_2^3 - \frac{1}{6} a^2 \lambda^2 w_2 w_1^2 \quad (3.84)$$

which have a single stable fixed point at $w_1 = 3/\lambda, w_2 = 0$, much like the case with kurtosis.

In the small noise approximation, the second term in Equation 3.83 is small, and we end up with an equation identical to the weight modification in the one dimensional Laplace environment (Equation 3.44). In other words, in small noise, the input with structure (the open eye) develops normally to the fixed point for the lower dimensional environment.

For the closed eye (w_2) the dynamics are quite different. If we keep only the last term in Equation 3.84, which is clearly larger than the others, and we assume that the other weight has converged quickly and is at the fixed point ($w_1 = 3/\lambda$), then we obtain a simple exponential decay of w_2 with $(3/2)a^2$ as the exponent of decay, where a is related to the variance of the noise with $\frac{3}{2}\sigma^2 = a^3$. Keeping further terms yields qualitatively similar forms. For BCM, then, more noise into the closed eye during MD yields a *faster* cutoff of response to the closed eye.

Conclusions about MD

We saw in Section 3.5 that we can divide many learning rules into two classes, based on their dependence on the noise level for monocular deprivation. The key difference between these two classes is not whether they depend on higher order statistics, nor on which moments of the data they are most dependent. It appears that the key difference is the presence, or lack, of a stabilization term which depends on the weight vector itself (heterosynaptic).

Though the presence of such a term could be measured experimentally (Section 1.7), it is not enough to differentiate between the rules. The BCM rule, for instance, could have such a term without significantly changing the noise dependence. One needs to know whether the heterosynaptic stabilization term is dominant. This, we have shown, can be done by examining the predictions for deprivation. The analysis we have presented provides a straightforward way of doing just that: comparing different learning rules, and different environments.

3.6.4 Binocular Deprivation

For binocular deprivation, we can do much the same. We are helped by the fact that we can use either uniform or Gaussian noise in the analysis. The model is shown in Figure 3.7D, where we follow NR by presenting noise to both eyes. We will begin with the Gaussian noise case, and then merely quote the result for the uniform case. We will see that there are some significant differences between these two cases, and that the results of the dynamics are somewhat peculiar.

Gaussian Noise

In this model, we have the following input densities

$$f_{x_i}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x_i^2/2\sigma^2} \quad (3.85)$$

and the outputs are defined as before.

$$\begin{aligned} y &= \mathbf{x} \cdot \mathbf{w} = x_1 w_1 + x_2 w_2 \\ &\equiv y_1 + y_2 \text{ (note: no sigmoid)} \\ z &\equiv \sigma(y) \end{aligned}$$

The densities of y_i are simply given by

$$f_{y_i}(y_i) = \frac{1}{w_i \sqrt{2\pi\sigma^2}} e^{-y_i^2/2\sigma^2 w_i^2} \quad (3.86)$$

from which we can calculate the total output density and the moments.

$$\begin{aligned} f_y(y) &= \int_{-\infty}^{\infty} f_{y_1}(y - y_2) f_{y_2}(y_2) dy_2 \\ &= \frac{1}{w_1 w_2} \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-(y-y_2)^2/2\sigma^2 w_1^2} e^{-y_2^2/2\sigma^2 w_2^2} dy_2 \\ &= \sqrt{\frac{1}{w_1^2 + w_2^2}} \sqrt{\frac{1}{2\pi\sigma^2}} e^{-y^2/2\sigma^2(w_1^2 + w_2^2)} \end{aligned} \quad (3.87)$$

With moments

$$E[z^2] = \frac{\sigma^2}{2} (w_1^2 + w_2^2) \quad (3.88)$$

$$E[z^3] = \sqrt{\frac{2}{\pi}} \sigma^3 (w_1^2 + w_2^2)^{3/2} \quad (3.89)$$

$$E[z^4] = \frac{3\sigma^4}{2} (w_1^2 + w_2^2)^2 \quad (3.90)$$

Kurtosis

Kurtosis is simply calculated from the moments, and converted to polar coordinates.

$$K_2 = \frac{3\sigma^4}{4} (w_1^2 + w_2^2)^2 \quad (3.91)$$

$$= \frac{3\sigma^4}{4} R^4 \quad (3.92)$$

which doesn't depend on the angle at all. Since we are starting from the $R = 1$ point, the weights do a random walk in the $R = 1$ circle. A picture of this is in Figure 3.8.

BCM

In this case, the BCM cost function can also be converted to polar, because it doesn't depend on the length of the weights.

$$R_{\text{QBCM}} = \frac{1}{3} \sqrt{\frac{2}{\pi}} \sigma^3 (w_1^2 + w_2^2)^{3/2} - \frac{\sigma^4}{16} (w_1^2 + w_2^2)^2 \quad (3.93)$$

$$= \frac{1}{3} \sqrt{\frac{2}{\pi}} \sigma^3 R^3 - \frac{\sigma^4}{16} R^4 \quad (3.94)$$

which has a fixed point at

$$\begin{aligned} (w_1^2 + w_2^2) &\equiv R \\ &= \pm 4 \sqrt{\frac{2}{\pi \sigma^2}} \end{aligned} \quad (3.95)$$

Once the weights reach this circle, they do a random walk around the circle (Figure 3.8).

Uniform Noise for BCM and Kurtosis

The moments for the uniform case are simply given as

$$E[z^2] = \frac{a^2}{6} (w_1^2 + w_2^2) \quad (3.96)$$

$$E[z^3] = \frac{a^3}{40w_2} (5w_2^4 + 10w_1^2w_2^2 + w_1^4) \quad (3.97)$$

$$E[z^4] = \frac{a^4}{30} (3w_1^4 + 10w_1^2w_2^2 + 3w_2^4) \quad (3.98)$$

The cost functions are then

$$K_2 = \frac{a^4 w_1^4}{60} + \frac{a^4 w_1^2 w_2^2}{6} + \frac{a^4 w_2^4}{60} \quad (3.99)$$

$$= \frac{a^4 R^4}{15} \left(\frac{1}{4} + 2 \cos^2(\theta) - 2 \cos^4(\theta) \right) \quad (3.100)$$

$$R_{\text{QBCM}} = \frac{a^3}{720} (30w_2^4 + 60w_1^2w_2^2 + 6w_1^4 - 5aw_2^5 - 10aw_2^3w_1^2 - 5aw_2w_1^4) \quad (3.101)$$

The kurtosis stable fixed point is

$$\theta = \frac{\pi}{4} \quad (3.102)$$

$$\mathbf{w} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad (3.103)$$

The BCM stable fixed point is

$$\mathbf{w} = \begin{pmatrix} \frac{18}{5a} \\ \frac{18}{5a} \end{pmatrix} \quad (3.104)$$

$$(3.105)$$

Conclusions about BD

One conclusion we make from these calculations, is that the cells remain binocular. Selectivity could be lost (if the weights decrease), but the cells remain, on average, binocular. The uniform distribution actually contains enough structure to have stable fixed points, so weights actually converge to the fixed point, and do not wander.

One peculiar consequence of the BCM Equations 3.95 and 3.104 is that, for small noise ($\sigma < \lambda$), the weights increase! This makes sense when we think of the behavior of BCM in the one dimensional case, because the lower the input, the larger the weight fixed point is in order to keep the output constant (Equation 1.11). This does not make sense biologically, so there must be something wrong here.

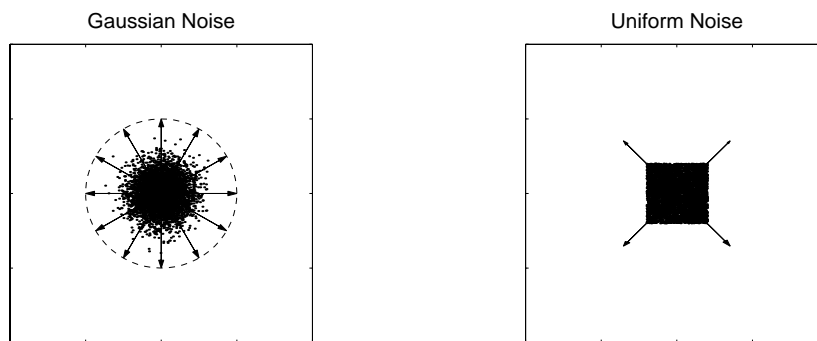


Figure 3.8: Binocular Deprivation Fixed Points in a low dimensional environment. For the Gaussian case (left), the fixed points for both K_2 and R_{BCM} fall on a circle. For the uniform case (right), they point in the direction of the corners of the distribution.

The reason why we have this behavior in these examples, and not in the simulations, is that the sigmoid in this simple analysis has a lower cutoff of zero. This sigmoid defines what we mean by spontaneous activity. This is the first time where the definition of spontaneous activity becomes crucial to the model. We can see where this comes from by looking at two extremes, low and high noise, rectified by a sigmoid with non-zero minimum.

Figure 3.9 shows the unrectified and rectified Gaussian distributions, for both a high and a low noise case. In the low noise case the sigmoid doesn't change the distribution much, so we would only have a fixed point at $\mathbf{w} = 0$. In the high noise case the sigmoid introduces a significant skew, so our fixed point would have a non-zero value, but because it is the large noise case, the value would be small. Therefore, if the sigmoid has a minimum far enough from zero, we can keep the weights from growing in binocular deprivation. This point requires further examination, to determine if this is a fundamental property or merely an artifact of the model.

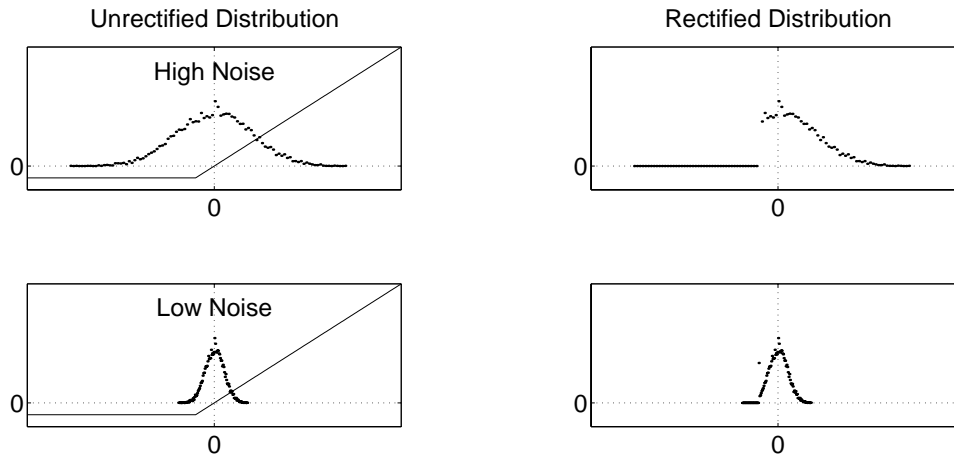


Figure 3.9: Shown are the unrectified (left) and rectified (right) Gaussian distributions, along with the rectifying sigmoid. In the high noise case (above) the sigmoid introduces a significant skew, whereas in the low noise case (below) the sigmoid doesn't change the distribution much.

3.6.5 Reverse Suture

The environment for reverse suture is identical to that for monocular deprivation (Figure 3.7C), so all of the results we derived for MD apply here as well. The only difference is the initial conditions. Essentially, if there is some component of the closed eye remaining in MD, then reverse suture will eventually lead to the proper fixed point: the direction of the newly open eye. The sigmoid plays a crucial role here, as it does with binocular deprivation. At the start of reverse suture, conditions are similar to binocular deprivation. We have small responses from each eye, because we have noise coming in from one eye, and a small weight (initially) for the other eye. If the responses are too small, and the sigmoid has a non-zero lower value, then reverse suture will not work for the same reasons that binocular deprivation *requires* such a sigmoid: all of the moments will vanish.

This apparent inconsistency can be resolved by making sure that there is some, somewhat significant, residual left from monocular deprivation and that the bottom value of the sigmoid is not so large (in magnitude) to make all of the moments equal to zero. The particular timing of the reverse suture depends, then, on the residual left from monocular deprivation and the lower level of the sigmoid.

3.6.6 Strabismus

Strabismus was described in Section 2.3.4 as the procedure where the eyes of the kitten were artificially misaligned. Simulations which demonstrate the resulting change from binocular to monocular cells are shown in Shouval et. al. (1995). This result can be understood very easily in this simple environment. Essentially, each eye is presented with structured input (numbers taken from a Laplace distribution), but the eyes are uncorrelated. The distribution does not lie on the 45° line, as it does for normal rearing (Figure 3.7A), but the two input directions are independent, as shown in Figure 3.10.

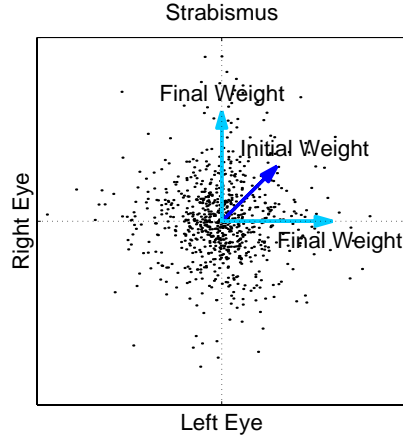


Figure 3.10: 2D Strabismus. Strabismus is modeled with numbers chosen from a Laplace distribution presented to each eye. The eyes have *independent* inputs (samples shown with dots). The initial weight is from normal rearing, and the final weight is in the direction of one of the eyes alone: the cell comes to respond only to only one eye.

In this case we have an input vector \mathbf{x} whose elements are chosen independently from a Laplace distribution.

$$f_{x_i}(x_i) = \frac{1}{2\lambda} e^{-|x_i|/\lambda} \quad (3.106)$$

We can see from Figure 3.10 that the tails of the distribution extend out in the \hat{x}_1 and \hat{x}_2 directions, which are the directions of highest kurtosis. Again we have

$$y = \mathbf{x} \cdot \mathbf{w} = x_1 w_1 + x_2 w_2 \quad (3.107)$$

$$\equiv y_1 + y_2 \text{ (note: no sigmoid)} \quad (3.108)$$

$$z \equiv \sigma(y) \quad (3.109)$$

$$f_{y_i}(y_i) = \frac{1}{2\lambda w_i} e^{-|y_i|/w_i \lambda} \quad (3.110)$$

and we calculate the distribution of the output y (and thus the distribution of z) by taking the convolution of the individual densities of y_1 and y_2 .

$$\begin{aligned} f_y(y) &= \int_{-\infty}^{-\infty} f_{y_1}(y - y_2) f_{y_2}(y_2) dy_2 \\ &= \frac{1}{4\lambda^2 w_1 w_2} \int_{-\infty}^{-\infty} e^{-\left|\frac{y-y_2}{w_1}\right|/\lambda} e^{-|y_2/w_2|/\lambda} dy_2 \end{aligned} \quad (3.111)$$

$$= \frac{1}{2\lambda(w_1 - w_2)(w_1 + w_2)} \left(w_1 e^{-|y|/w_1 \lambda} - w_2 e^{-|y|/w_2 \lambda} \right) \quad (3.112)$$

$$f_z(z) = \begin{cases} f_y(z) & \text{if } z > 0 \\ \frac{1}{2}\delta(z) & \text{if } z = 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (3.113)$$

The moments are calculated in a straightforward, though algebraically messy, fashion.

$$\begin{aligned} E[z^2] &= \frac{1}{2\lambda(w_1 - w_2)(w_1 + w_2)} (w_1 2\lambda^3 w_1^3 - w_2 2\lambda^3 w_2^3) \\ &= \lambda^2 (w_1^2 + w_2^2) \end{aligned} \quad (3.114)$$

$$\begin{aligned} E[z^3] &= \frac{1}{2\lambda(w_1 - w_2)(w_1 + w_2)} (6w_1^5 \lambda^4 - 6w_2^5 \lambda^4) \\ &= 3\lambda^3 \frac{w_1^5 - w_2^5}{w_1^2 - w_2^2} \end{aligned} \quad (3.115)$$

$$\begin{aligned} E[z^4] &= \frac{1}{2\lambda(w_1 - w_2)(w_1 + w_2)} (24\lambda^5 w_1^6 - 24\lambda^5 w_2^6) \\ &= 12\lambda^4 (w_1^4 + w_1^2 w_2^2 + w_2^4) \end{aligned} \quad (3.116)$$

Kurtosis

The value of K_2 is then simply

$$\begin{aligned} K_2 &= E[z^4] - 3E[z^2]^2 \\ &= 9\lambda^4 w_1^4 + 6\lambda^4 w_1^2 w_2^2 + 9\lambda^4 w_2^4 \end{aligned} \quad (3.117)$$

$$K_2(R, \theta) = \lambda^4 R^4 (12 \cos^4(\theta) - 12 \cos^2(\theta) + 9) \quad (3.118)$$

where we have converted K_2 to polar, as before.

This function has stable maxima at $\theta = 0$ and $\theta = \pi/2$, which are the monocular solutions.

BCM

The cost function is calculated from the moments, as before.

$$\begin{aligned} R_{\text{QBCM}} &= \frac{1}{3} E[z^3] - \frac{1}{4} E^2[z^2] \\ &= \lambda^3 \frac{w_1^5 - w_2^5}{w_1^2 - w_2^2} - \frac{\lambda^2}{4} (w_1^2 + w_2^2)^2 \end{aligned} \quad (3.119)$$

which has stable maxima

$$\mathbf{w} = \begin{pmatrix} 0 \\ 3/\lambda \end{pmatrix}, \begin{pmatrix} 3/\lambda \\ 0 \end{pmatrix} \quad (3.120)$$

which are also monocular solutions.

3.6.7 Conclusions about the Simple Environment

We have now seen how a simple two dimensional environment can be used to explain many of the results seen with the more realistic natural scene environment. The simple environment gives us an easy way

to think about the results of deprivation as the competition between distributions, and can also lead to the qualitative dependence on such things as the noise in deprivation. We saw how the correlations between the eyes can explain the results of normal rearing and strabismus.

We also saw how the environment highlighted the important role of the sigmoid in binocular deprivation and reverse suture. This role is central to the interpretation of the activity level of the cell above spontaneous. Current work is exploring how to relax this assumption, thereby yielding only positive activities, and possibly eliminating the difficulty with binocular deprivation.

More work needs to be done in exploring the statistics of natural scenes, which clearly determine much of the behavior of these models. The simple environment can provide a straightforward way to answer questions regarding the input statistics and different learning rules.

The simple environment generalizes nicely to higher dimensions with, say, each dimension having an input distribution chosen from independent Laplace numbers (in the case of structured input) or Gaussian/uniform numbers (in the case of noise). The inputs need not be independent, but can be at least linear mixtures of each other. In this case, both BCM and kurtosis can be seen to be performing independent component analysis (ICA)(Comon, 1994; Bell and Sejnowski, 1997; Hyvarinen and Oja, 1997) for which there is a growing literature. The fundamental interpretation of competing distributions remains, so we need not examine the higher dimensional cases for our purposes here.