15. LOCAL LINEAR MAPPINGS

In the treatment of robot control tasks, we have seen that often the use of matrices, *i.e.*, linear mappings, as output values provides a useful extension of output learning maps. In the simulation of visuomotor coordination, the network learned the transformation between the visual image of the target point and the joint angles for the required arm position. Each neuron represented this nonlinear transformation for a neighborhood of a grid point. To this end, it had a matrix available that gave the linear part of the expansion of the transformation about the relevant grid point. In this way, the required transformation was approximated by an adaptive superposition of many linear mappings, each one valid only locally. Compared to the use of fixed output values, this yields a considerably higher accuracy with the same number of neurons.

Another interesting possibility was demonstrated for ballistic movements in Chapter 13. There, an output quantity (torque amplitude), varying as a function of further parameters (components of the target velocity), was assigned by means of an array to each input signal (arm position) which describes a linear relationship between torque amplitudes and velocity components. Such linear relationship represented by a matrix eliminates the necessity of representing the further parameters (*e.g.*, velocities) in the map as well, and, hence, the dimension of the space to be projected onto the lattice can be significantly reduced.

A precondition for such a strategy is a splitting of the input variables $v_1, v_2, \ldots, v_d$ into two (not necessarily disjoint) sets $\{v'_1, v'_2, \ldots, v'_a\}$ and $\{v''_1, v''_2, \ldots, v''_b\}$ such that

$$\{v_1, \ldots, v_d\} \quad = \quad \{v'_1, \ldots, v'_a\} \cup \{v''_1, \ldots, v''_b\},$$

and such that that output quantities $\mathbf{f}$ locally depend only linearly on one of the sets, *i.e.*,

$$\mathbf{f} = \mathbf{A}(\mathbf{v}')\mathbf{v}''. \tag{15.1}$$

Here, we have put $\mathbf{v}' := (v'_1, \ldots, v'_a)$, $\mathbf{v}'' := (v''_1, \ldots, v''_b)$. All those parameters not represented in the map themselves are included in $\mathbf{v}''$.
Such a splitting is possible in many cases. For example, in Chapter 11 the vector $\mathbf{v}'$ consisted of the coordinates $\mathbf{u}$ of the target point in the two camera fields of view. The vector $\mathbf{f}$ was the change of the joint angle under a small shift $\mathbf{v}''$ of the location of the end effector in the camera fields of view. In Chapter 13, $\mathbf{v}'$ stood for the joint angles of the arm, $\mathbf{f}$ was the torque amplitude in the joints, and $\mathbf{v}''$ was the resulting velocity to which the end effector was accelerated under the action of $\mathbf{f}$.

15.1 The Learning Algorithm for Local Linear Mappings

In this section we formulate the general version of the learning algorithm already derived for the special discussed in Chapters 11, 12, and 13.
We assume as before that the only available information is a sequence of n-tuples $(\mathbf{v}', \mathbf{v}'', \mathbf{f})$ satisfying (15.1). These are created during the learning phase by the reaction of the system to, *e.g.*, pseudo-randomly selected targets. In visuomotor coordination (Chapter 11), for example, $\mathbf{v}'$ was the position in the field of view of the target point and $\mathbf{v}''$ and $\mathbf{f}$ were the changes of the position in the field of view of the end effector and the joint angles during fine positioning. In the ballistic movements of Chapter 13, the $\mathbf{v}'$ were the arm joint angles, and $\mathbf{v}''$ was the velocity of the end effector due to an acceleration with torque amplitudes $\mathbf{f}$.
The task of the network is to learn the matrix $\mathbf{A}(\mathbf{v}')$ of Eq. (15.1) for each $\mathbf{v}'$. As was shown in Chapter 11, this can occur by means of a linear error correction rule. Together with the principle of neighborhood cooperation in Kohonen's original model, this leads to the following learning algorithm:

1. Register the next input signal $(\mathbf{v}', \mathbf{v}'', \mathbf{f})$.

2. Determine the lattice site $\mathbf{s} := \phi_{\mathbf{w}}(\mathbf{v}')$, assigned to $\mathbf{v}'$ in the map.

3. Compute an improved estimate $\mathbf{A}^*$ for the linear mapping $\mathbf{A}^{old}_{\mathbf{s}}$ of the chosen lattice site $\mathbf{s}$

$$\mathbf{A}^* = \mathbf{A}^{old}_{\mathbf{s}} + \delta \cdot \Big(\mathbf{f} - \mathbf{A}^{old}_{\mathbf{s}} \mathbf{v}''\Big)(\mathbf{v}'')^T \tag{15.2}$$

4. Carry out a learning step
$$\mathbf{A}_{\mathbf{r}}^{new} = \mathbf{A}_{\mathbf{r}}^{old} + \epsilon h_{\mathbf{rs}}\Big(\mathbf{A}^* - \mathbf{A}_{\mathbf{r}}^{old}\Big) \tag{15.3}$$
for the assignment of linear mappings $\mathbf{A}_{\mathbf{r}}$.

5. Carry out a learning step
$$\mathbf{w}_{\mathbf{r}}^{new} = \mathbf{w}_{\mathbf{r}}^{old} + \epsilon' h'_{\mathbf{rs}}\Big(\mathbf{v}' - \mathbf{w}_{\mathbf{r}}^{old}\Big) \tag{15.4}$$
for the synaptic strengths $\mathbf{w}_{\mathbf{r}}$, and continue with Step 1

We encountered this algorithm in its application to the control of a robot arm by means of computer simulations. In the following, we analyze the algorithm mathematically in more detail. We are mainly interested in the question of convergence of the linear mappings $\mathbf{A}_{\mathbf{r}}$ to their correct values. For this, we first discuss the convergence behavior of the matrices $\mathbf{A}_{\mathbf{r}}$ in the absence of the lateral interaction, *i.e.*, for $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$. Building on this, we then investigate the important influence of lateral interaction.

15.2 Convergence Behavior without Lateral Interaction

Without lateral interaction, each lattice site learns its linear mapping isolated from all the others. We can then consider the evolution of the matrix of a single lattice site in our treatment of convergence. We further assume that the correspondence between lattice sites and values $\mathbf{v}'$ given by the vectors $\mathbf{w}_{\mathbf{r}}$ has already formed and no longer changes significantly in the course of the learning phase. To each vector $\mathbf{v}'$ is assigned a fixed lattice site $\mathbf{s}$ and thus a matrix $\mathbf{A}_{\mathbf{s}}$. We emphasize this in the following by writing $\mathbf{A}(\mathbf{v}', t)$ instead of $\mathbf{A}_{\mathbf{s}}$, where t gives the number of learning steps after which lattice site $\mathbf{s}$ was chosen in step 2 of the algorithm. With $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$ and equations (15.2) and (15.3), one then has
$$\mathbf{A}(\mathbf{v}', t+1) = \mathbf{A}(\mathbf{v}', t) + \delta \cdot \Big(\mathbf{A}(\mathbf{v}') - \mathbf{A}(\mathbf{v}', t)\Big)\mathbf{v}''(\mathbf{v}'')^T \tag{15.5}$$
Here, we have absorbed the product $\epsilon \cdot \delta$ into the single constant δ. Denoting by $\mathbf{D}(\mathbf{v}', t) := \mathbf{A}(\mathbf{v}', t) - \mathbf{A}(\mathbf{v}')$ the deviation from the exact matrix $\mathbf{A}(\mathbf{v}')$,

we obtain for the change of the Euclidean matrix norm $\|\mathbf{D}\| = (\mathrm{Tr}\ \mathbf{D}^T\mathbf{D})^{1/2}$ during one step (15.5

$$\begin{aligned} \Delta\|\mathbf{D}\|^2 &= 2\mathrm{Tr}\ \mathbf{D}^T\Delta\mathbf{D} + \mathrm{Tr}\ \Delta\mathbf{D}^T\Delta\mathbf{D} \\ &= -\delta(2-\delta\|\mathbf{v}''\|^2)\|\mathbf{D}\mathbf{v}''\|^2 \end{aligned} \tag{15.6}$$

If $0 < \delta < 2/\|\mathbf{v}''\|^2$, the norms $\|\mathbf{D}(\mathbf{v}',t)\|$ thus constitute a monotonically decreasing sequence. For a nonsingular correlation matrix $\langle \mathbf{v}''(\mathbf{v}'')^T\rangle$, this guarantees the convergence $\lim_{t\to\infty}\mathbf{A}(\mathbf{v}',t) = \mathbf{A}(\mathbf{v}')$. [1]
The preceding treatment of convergence assumes that the correlation matrix $\langle \mathbf{v}''(\mathbf{v}'')^T\rangle$ is independent of $\mathbf{A}(\mathbf{v}',t)$. However, this is often not satisfied because the values of $\mathbf{v}''$ are generated by the system itself, *i.e.*, the system tries to learn *from its own reactions.* At each learning step, the system receives a *target* $\mathbf{v}''_{targ}$ for $\mathbf{v}''$. For ballistic movements, this is the target velocity of the end effector, in visuomotor coordination, it is the residual difference between achieved and prescribed end effector position in the two camera fields of view after coarse positioning. In order to reach the target, the system determines its output quantity $\mathbf{f}$ by (15.1), but instead of the correct matrix $\mathbf{A}(\mathbf{v}')$ it uses the matrix $\mathbf{A}(\mathbf{v}',t)$ which deviates more or less from $\mathbf{A}(\mathbf{v}')$. Thus,

$$\mathbf{f} = \mathbf{A}(\mathbf{v}',t)\mathbf{v}''_{targ}. \tag{15.7}$$

By (15.1), this leads to

$$\mathbf{v}'' = \mathbf{A}(\mathbf{v}')^{-1}\mathbf{A}(\mathbf{v}',t)\mathbf{v}''_{targ}. \tag{15.8}$$

Hence, a nonsingular correlation matrix $\langle \mathbf{v}''_{targ}(\mathbf{v}''_{targ})^T\rangle$ of the target is not enough to guarantee convergence, because if $\mathbf{A}(\mathbf{v}',t)$ evolves "unfavorably" during learning, $\langle \mathbf{v}''(\mathbf{v}'')^T\rangle$ can still become singular, and the learning process can get stuck. This was the reason why in Chapter 11 and 13 we obtained convergence only for a fraction of the lattice sites without neigborhood cooperation between the neurons (see Figs. 11.7 and 13.5). We now analyse this behavior mathematically in more detail.
We neglect the slight variation of $\mathbf{v}'$ within the "parcels" of the particular lattice site $\mathbf{s}$ chosen and thus write $\mathbf{v}$ in place of $\mathbf{v}''$ and $\mathbf{A}(t)$ or $\mathbf{A}$ in

[1] For singular $\langle \mathbf{v}''(\mathbf{v}'')^T\rangle$ $\mathbf{D}$ can converge to a nonvanishing value from the null-space $\langle \mathbf{v}''(\mathbf{v}'')^T\rangle$, but even in this case the mean squared error Tr $\mathbf{D}\langle \mathbf{v}''(\mathbf{v}'')^T\rangle\mathbf{D}^T$ goes to zero.

place of $\mathbf{A}(t, \mathbf{v}')$ or $\mathbf{A}(\mathbf{v}')$, respectively. In order to investigate convergence, we consider the matrix $\mathbf{B}(t) = \mathbf{A}^{-1}\mathbf{A}(t) - \mathbf{1}$. We obtain for the change $\Delta\mathbf{B} := \mathbf{B}(t+1) - \mathbf{B}(t)$ of $\mathbf{B}$ under a learning step the expression

$$\Delta\mathbf{B} = \; -\delta \cdot \mathbf{B}\Big(\mathbf{1} + \mathbf{B}\Big)\mathbf{u}\mathbf{u}^T\Big(\mathbf{1} + \mathbf{B}\Big)^T. \tag{15.9}$$

Similar to (15.6), the change of $\|\mathbf{B}\|^2$ under a learning step (15.5) becomes

$$\Delta\|\mathbf{B}\|^2 = -\delta\Big(2 - \delta\|\mathbf{v}\|^2\Big)\|\mathbf{B}(\mathbf{1} + \mathbf{B})\mathbf{u}\|^2. \tag{15.10}$$

Hence, a monotonically decreasing sequence again arises for $\|\mathbf{B}(t)\|$, provided $0 < \delta < 2/\|\mathbf{v}\|^2$ holds. Maximization of the decrease per learning step occurs by means of the choice $\delta = 1/\|\mathbf{v}\|^2$. If δ is to be chosen independently of $\mathbf{v}$, then the condition $0 < \delta < 2/\alpha(1 + \|\mathbf{B}(0)\|)^2$ with $\alpha = \sup\|\mathbf{u}\|^2$ is sufficient for $\Delta\|\mathbf{B}\|^2 < 0$. Every possible stationary value for $\|\mathbf{B}(t)\|$ requires $\mathbf{B}^2 = -\mathbf{B}$. Since for $\|\mathbf{B}(0)\| < 1$ solutions $\mathbf{B} \neq 0$ with $\mathbf{B}^2 = -\mathbf{B}$ can no longer be reached, we obtain the convergence statement For $\|\mathbf{B}(0)\| < 1$ and $0 < \delta < 2/\|\mathbf{v}\|^2$ holds $\lim_{t\to\infty} \mathbf{B}(t) = 0$, *i.e.*, $\lim_{t\to\infty} \mathbf{A}(t) = \mathbf{A}$.

However, the condition $\mathbf{B}^2 = -\mathbf{B}$ has, in contrast to the previous situation described by (15.6), in addition to $\mathbf{B} = 0$ a whole manifold M of undesired stationary solutions. As we will show, a subset of M possesses an attractive neighborhood. Hence, there are initial values with the property $\|\mathbf{B}(0)\| > 1$ that evolve toward M under the learning rule and thus do not lead to the desired limit $\mathbf{A}$. For such initial values, the learning procedure converges to the wrong value.

This behavior can be illustrated well if one restricts to the one-dimensional case. In this case, $\mathbf{u}$ and $\mathbf{B}$ are scalar quantities, and (15.9) simplifies to

$$\dot{B} = -\delta \cdot B \cdot (B+1)^2 u^2. \tag{15.11}$$

For sufficiently small learning step lengths δ, one can neglect statistical fluctuations due to the random variables u and replace u^2 by its average. Without loss of generality, we assume $\langle u^2 \rangle = 1$. This yields

$$\dot{B} = -\delta \cdot B \cdot (B+1)^2. \tag{15.12}$$

We can interpret B as the position coordinate of a mass point in a viscous medium, *e.g.*, a small, not too heavy sphere in a jar of honey. The equation of motion for viscous motion is

$$\frac{m}{\gamma}\ddot{B} + \dot{B} = -\frac{d}{dB}V(B). \tag{15.13}$$

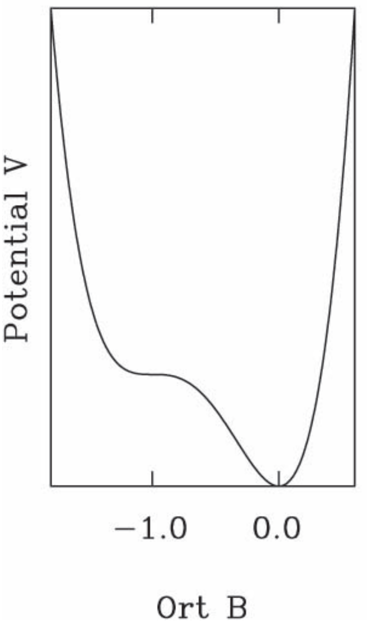

Abb. 15.1: The shape of the potential $V(B)$. The absolute minimum lies at $B = 0$, which corresponds to the correctly learned matrix. The "force" acting at the position $B = -1$ on B also vanishes. Hence, for an unfavorable initial value, B gets "stuck" on this plateau.

Here $V(B)$ is a potential in which the sphere moves. In the case of a small mass m and large viscosity γ, *i.e.*, $m/\gamma \ll 1$, the acceleration term with $\ddot{B}$ can be neglected, and the velocity $\dot{B}$ is proportional to the force $-V'(B)$. Equation (15.13) goes over to (15.12) in this limit for

$$V(B) = \frac{1}{4}B^4 + \frac{2}{3}B^3 + \frac{1}{2}B^2. \tag{15.14}$$

Figure 15.1 presents the shape of $V(B)$. The global minimum lies at $B = 0$, the value to be learned. The finite attractive neighborhood of this minimum extends from $B > -1$ to $B = \infty$. Any initial value of B within this interval converges to the desired value during the learning process described by (15.12). The condition $\|\mathbf{B}(0)\| < 1$ assumed in the above convergence statement thus implies that we are located within the basin of attraction of the minimum at $B = 0$. In the one-dimensional case, the submanifold M of "false" stationary solutions consists of just the isolated point $B = -1$. Figure 15.2 shows that the attractive region of M is given by the interval $]-\infty, -1]$. Since the motion in the potential surface $V(B)$ is "infinitely" viscous, any initial value within the interval $]-\infty, -1]$ is pushed towards the point M and gets stuck there. However, an arbitrarily small disturbance in the positive direction suffices for leaving M in favor of the desired minimum $B = 0$. In higher dimensions, one has in addition undesired stationary solutions which are no longer unstable, and in this case M even has points where a small perturbation no longer leaves M. We show this in the remainder of this section.

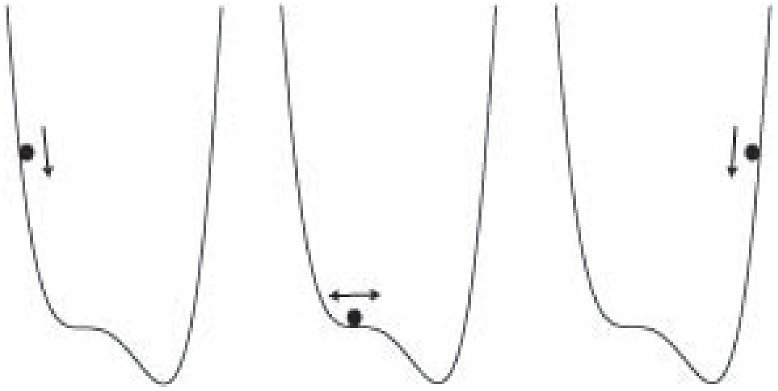

Abb. 15.2: An illustration of the convergence behavior of B. The sphere with initial value less than -1 (left) will get stuck on the plateau at $B = -1$ (middle). Any sphere to the right of the plateau will roll as desired into the global minimum at $B = 0$ (right).

We now resume the discussion of the general case. M consists of all matrices $\mathbf{B} \neq 0$ satisfying the condition $\|\mathbf{B}(\mathbf{B} + \mathbf{1})\| = 0$. Hence, we can take the quantity $d(\mathbf{B}) = \|\mathbf{B}(\mathbf{B}+\mathbf{1})\|^2 = \mathrm{Tr}\ \mathbf{B}(\mathbf{B}+\mathbf{1})(\mathbf{B}+\mathbf{1})^T\mathbf{B}^T$ as a measure of the distance from $\mathbf{B}$ to M. If δ is small enough to justify neglect of the terms of quadratic order, the learning step (15.5) leads to the change

$$\begin{aligned}\Delta d(\mathbf{B}) &= -2\delta \mathrm{Tr}\ \mathbf{B}(\mathbf{1}+\mathbf{B}) \\ &\times \Big[\mathbf{B}\mathbf{u}\mathbf{u}^T(\mathbf{1}+\mathbf{B})^T + \mathbf{u}\mathbf{u}^T(\mathbf{1}+\mathbf{B})^T(\mathbf{1}+\mathbf{B})\Big] \\ &\times (\mathbf{1}+\mathbf{B})^T\mathbf{B}^T. \qquad (15.15)\end{aligned}$$

This expression is not particularly accessible to further manipulation. Hence, we restrict ourselves as above to the case in which (15.15) can be averaged over the target vector $\mathbf{u}$. This is consistent with the assumption of small learning step lengths δ. We further assume for $\mathbf{u}$ an isotropic distribution independently in each component, so that (perhaps after appropriate rescaling) $\langle \mathbf{u}\mathbf{u}^T \rangle = \mathbf{1}$ holds. This leads to

$$\begin{aligned}\langle \Delta d(\mathbf{B}) \rangle &= -2\delta \cdot \mathrm{Tr}\ \mathbf{B}(\mathbf{1}+\mathbf{B})\Big(\mathbf{1} + 2\mathbf{B} + \mathbf{B}^T + \mathbf{B}\mathbf{B}^T + \mathbf{B}^T\mathbf{B}\Big) \\ &\times (\mathbf{1}+\mathbf{B})^T\mathbf{B}^T \\ &= -2\delta \cdot \mathrm{Tr}\ \mathbf{B}(\mathbf{1}+\mathbf{B})\Big(\mathbf{1} + \frac{3}{2}\mathbf{B} + \frac{3}{2}\mathbf{B}^T + \mathbf{B}\mathbf{B}^T + \mathbf{B}^T\mathbf{B}\Big) \\ &\times (\mathbf{1}+\mathbf{B})^T\mathbf{B}^T \\ &= -2\delta \cdot \mathrm{Tr}\ \mathbf{B}(\mathbf{1}+\mathbf{B})\mathbf{H}(\mathbf{B})(\mathbf{1}+\mathbf{B})^T\mathbf{B}^T, \qquad (15.16)\end{aligned}$$

where the matrix $\mathbf{H}(\mathbf{B})$ is defined by

$$\mathbf{H}(\mathbf{B}) = \mathbf{1} + \frac{3}{2}\mathbf{B} + \frac{3}{2}\mathbf{B}^T + \mathbf{B}\mathbf{B}^T + \mathbf{B}^T\mathbf{B}. \qquad (15.17)$$

For all regions of M for which $\mathbf{H}$ is strictly positive, one has $\langle \Delta d(\mathbf{B}) \rangle < 0$. Thus, any point $\mathbf{B}$ located sufficiently close to such a region is drawn farther toward M on the average. A condition for this to occur results from the following

Theorem 1. Let $\mathbf{B}_0 := \sum_{i=1,n} \mathbf{p}_i\mathbf{q}_i^T$, where $\mathbf{p}_i, \mathbf{q}_i$ are $2n$ vectors, whose scalar products satisfy the conditions

$$\mathbf{p}_i \cdot \mathbf{p}_j = 0, \quad \mathbf{q}_i \cdot \mathbf{q}_j = 0, \quad (i \neq j);$$

together with $\|\mathbf{p}_i\| \cdot \|\mathbf{q}_i\| \geq 3/2,\ \ i = 1, \ldots, n$. For every $\mathbf{B}$ sufficiently close to $\mathbf{B}_0$, one then has $\langle \Delta d(\mathbf{B}) \rangle < 0$.

Proof: For $i = 1, \dots, n$, define

$$\begin{aligned}\alpha_i : &= \|\mathbf{q}_i\|;\\ \beta_i : &= \frac{3}{2\|\mathbf{q}_i\|} \leq \|\mathbf{p}_i\|;\\ \mathbf{w}_i : &= \alpha_i \mathbf{p}_i + \beta_i \mathbf{q}_i;\end{aligned}$$

This yields

$$\mathbf{H}(\mathbf{B}_0) = \mathbf{1} + \sum_{i=1..n} \mathbf{w}_i \mathbf{w}_i^T + \sum_{i=1..n} (\|\mathbf{p}_i\|^2 - \beta_i^2)\mathbf{q}_i \mathbf{q}_i^T. \tag{15.18}$$

Therefore, $\mathbf{H}(\mathbf{B}_0)$ is strictly positive. Since $\mathbf{H}$ depends continuously on its argument, this holds throughout a whole neighborhood of $\mathbf{B}_0$ and implies $\langle d(\mathbf{B})\rangle < 0$ there.

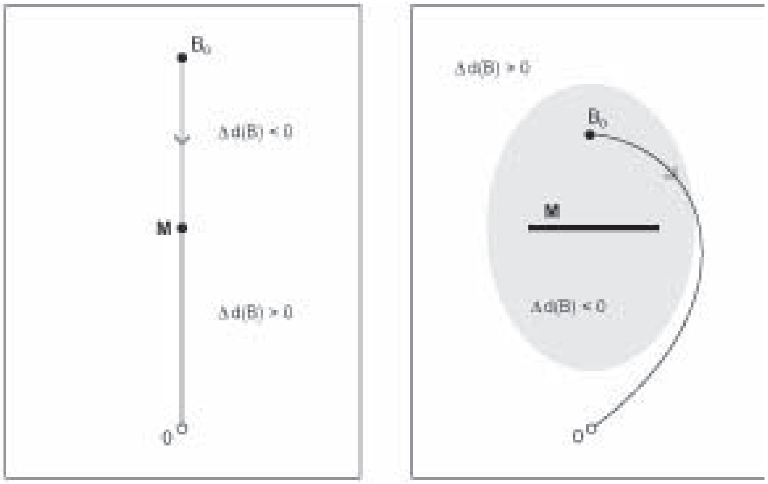

Abb. 15.3: Difference between the one-dimensional and the multidimensional case. *Left:* In the one-dimensional case, the desired solution $\mathbf{B} = 0$ cannot be reached if the undesired fixed point M separates the initial value $\mathbf{B}_0$ from the origin. *Right:* In the multidimensional case, on the other hand, it is possible to avoid the manifold M of undesired fixed points. It may even be possible to reach the desired solution $\mathbf{B} = 0$ if the initial value $\mathbf{B}_0$ lies in the (shaded) neigborhood of M within which $d(\mathbf{B})$, the distance to M, is everywhere decreasing.

This deserves two remarks. First, there are matrices $\mathbf{B}_0$ for which the above theorem holds, but which are located so far from the manifold M, *i.e.*, for

which $\|\mathbf{B}_0(\mathbf{1}+\mathbf{B}_0)\|$ is so large, that the matrices are attracted to the desired solution $\mathbf{B} = 0$ before reaching M. For these initial values, the above theorem does not necessarily imply convergence to M, since the learning steps (15.5) might induce the system to leave the neighborhood of the initial value within which this property exists, even if they decrease $\|\mathbf{B}(\mathbf{1} + \mathbf{B})\|$ on the average. This is shown in Fig. 15.3 on the right. A sufficient condition for $\mathbf{B}_0 \in M$ is for example $\mathbf{p}_i \cdot \mathbf{q}_j = -\delta_{ij}$.
Secondly, M possesses points for which $\langle \Delta d(\mathbf{B})\rangle < 0$ can not even be guaranteed within an entire neighborhood. Near these points, it is no longer possible to guarantee convergence to M. An example of such a point is $\mathbf{B} = -\mathbf{1}$. As we have seen, in the one-dimensional case M consists only of this one point.
Thus, we have shown that under the learning rule (15.5) $\mathbf{A}(t)$ converges to the desired value $\mathbf{A}$, provided the initial value $\mathbf{A}(0)$ is not "too badly" chosen. The basin of attraction for the desired $\mathbf{A}$ contains the region $\|\mathbf{A}^{-1}\mathbf{A}(0)-\mathbf{1}\| < 1$. Moreover, there is a whole manifold of undesired fixed points which can be reached for bad initial values. This unfortunate property led in Chapters 11 and 13 to poor results in the computer simulations whenever there was no neighborhood cooperation between the neurons. With sufficient neighborhood cooperation between the neurons, on the other hand, convergence to the desired state occured. In the following, we show how this improvement through neighborhood cooperation arises.

15.3 Improvement of Convergence through Neighborhood Cooperation

We now investigate the effects of neighborhood cooperation due to the lateral interaction $h_{\mathbf{rs}}$. A significant consequence of neighborhood cooperation is that none of the adaptation steps is restricted to the particular lattice site $\mathbf{s}$, but rather all of the adjacent lattice sites participate in the adaptation step as well. The degree of participation decreases according to $h_{\mathbf{rs}}$ with increasing distance from $\mathbf{s}$. In the following, we will show that this offers at least two advantages. First, the effective rate of convergence is improved, and secondly the robustness of the system with respect to unfavorable initial values of the linear mappings $\mathbf{A}_\mathbf{r}$ is increased. Even for initial values for which, in the absence of lateral interaction, not all the mappings $\mathbf{A}_\mathbf{r}$ would converge as desired, convergence of all $\mathbf{A}_\mathbf{r}$ to the correct matrices is ensured in the presence of lateral interaction.

To make the following investigation feasible, we make a few additional simplifying assumptions. First, we suppose that the adaptation of the synaptic strengths $\mathbf{w}_\mathbf{r}$ is already finished and has attained an asymptotic distribution such that each lattice site is selected in step 2 of the algorithm with the same probability. As shown in Chapter 5, it is just this (approximate) creation of such a state which forms an essential property of the algorithm. Secondly, we restrict ourselves to the case in which the correct mapping $\mathbf{A}(\mathbf{v}')$ is independent of $\mathbf{v}'$ and thus the same for every lattice site. This assumption will not significantly influence the results in all cases where the change in $\mathbf{A}(\mathbf{v}')$ is small over the range of the function $h_{\mathbf{rs}}$. We further suppose that the step lengths δ are small enough to allow one to neglect terms of quadratic and higher order. Under these assumptions, we can summarize steps 1–4 for the matrices $\mathbf{B}_\mathbf{r} = \mathbf{A}(\mathbf{w}_\mathbf{r})^{-1}\mathbf{A}_\mathbf{r}(t) - \mathbf{1}$ as follows:

1. Choose $\mathbf{s} = \phi_\mathbf{w}(\mathbf{v}')$.

2. Set
$$\mathbf{B}^* = \mathbf{B}_\mathbf{s}(t) + \Delta_L\Big(\mathbf{B}_\mathbf{s}(t)\Big), \tag{15.19}$$
where $\Delta_L\Big(\mathbf{B}_\mathbf{s}(t)\Big) = -\delta\mathbf{B}_\mathbf{s}(t)(\mathbf{1}+\mathbf{B}_\mathbf{s}(t))\mathbf{u}\mathbf{u}^T(\mathbf{1}+\mathbf{B}_\mathbf{s}(t))^T$ is the change of $\mathbf{B}_\mathbf{s}(t)$ under the learning rule (15.5).

3. Improve the matrices $\mathbf{B}_\mathbf{r}(t)$ according to
$$\mathbf{B}_\mathbf{r}(t+1) = \mathbf{B}_\mathbf{r}(t) + \epsilon h_{\mathbf{rs}}\Big(\mathbf{B}^* - \mathbf{B}_\mathbf{r}(t)\Big), \tag{15.20}$$
and begin again at step 1.

With (15.19) and (15.20), we obtain for the average time rate of change $\dot{\mathbf{B}}_\mathbf{r}$ of the matrix $\mathbf{B}_\mathbf{r}$ in the presence of additional neighborhood cooperation

$$\dot{\mathbf{B}}_\mathbf{r} = \sum_\mathbf{s} h_{\mathbf{rs}}(\mathbf{B}_\mathbf{s} - \mathbf{B}_\mathbf{r}) - \delta \cdot \sum_\mathbf{s} h_{\mathbf{rs}}\mathbf{B}_\mathbf{s}(\mathbf{B}_\mathbf{s}+\mathbf{1})(\mathbf{B}_\mathbf{s}+\mathbf{1})^T. \tag{15.21}$$

Here, we have again replaced $\mathbf{u}\mathbf{u}^T$ by its mean, and we have assumed as before $\langle\mathbf{u}\mathbf{u}^T\rangle = \mathbf{1}$. A multiplicative factor ϵ/N has been normalized to unity by an appropriate scaling of the time constant.

We decompose the summation over all lattice points $\mathbf{s}$ into sums over nearest neighbors of $\mathbf{r}$, next nearest neighbors, etc. This yields the expression

$$\dot{\mathbf{B}}_\mathbf{r} \;\; = \;\; h\sum_{\langle\mathbf{s}\rangle}(\mathbf{B}_\mathbf{s} - \mathbf{B}_\mathbf{r}) + h^2\sum_{\langle\langle\mathbf{s}\rangle\rangle}(\mathbf{B}_\mathbf{s} - \mathbf{B}_\mathbf{r}) + \ldots$$

$$\begin{aligned} &- \quad \delta \cdot \mathbf{B_r}(\mathbf{B_r}+\mathbf{1})(\mathbf{B_r}+\mathbf{1})^T \\ &- \quad \delta \cdot h \sum_{\langle \mathbf{s} \rangle} \mathbf{B_s}(\mathbf{B_s}+\mathbf{1})(\mathbf{B_s}+\mathbf{1})^T + \ldots \ , \end{aligned} \tag{15.22}$$

where $\langle \mathbf{s} \rangle$ is to be understood as a summation over nearest neighbors and $\langle\langle \mathbf{s} \rangle\rangle$ as a sum over next nearest neighbors. The factor h is the fall-off of the Gaussian $h_{\mathbf{rs}}$ from the center of the excitation $\mathbf{s}$ to the nearest lattice points, *i.e.*, $h = \exp(-1/2\sigma^2)$. The fall-off up to the next nearest neighbors then has the value h^2 etc.

Three cases can be discussed on the basis of Eq. (15.22). First, the limit $h \approx 1$ and $\delta \ll h$. This corresponds to a very-long-range neighborhood interaction and (relative to this) a negligible length δ of the improvement step, as present at the beginning of the learning phase. For this extreme case, we can again give a potential for the viscous motion $\dot{\mathbf{B}}_\mathbf{r}$, namely

$$V = \frac{h}{4} \sum_{\mathbf{r}} \sum_{\langle \mathbf{s} \rangle} (\mathbf{B_s} - \mathbf{B_r})^2 + \frac{h^2}{4} \sum_{\mathbf{r}} \sum_{\langle\langle \mathbf{s} \rangle\rangle} (\mathbf{B_s} - \mathbf{B_r})^2 + \ldots \ , \tag{15.23}$$

which corresponds to the simple situation of coupled springs with spring constants depending on the lattice spacing. In this potential, the matrices $\mathbf{B_r}$ try to take the same value at every lattice site. This is important in the initial phase of learning, because "deviants" in the initial values are "tamed" by all other neighbors, and each $\mathbf{B_r}$ settles down to an average over all initial values. This average need not lie at the desired $\mathbf{B_r} = 0$; this can be seen from the fact that the above potential is translationally invariant, and thus every value for $\mathbf{B_r}$ which is equal at all lattice sites minimizes V. Hence, we need an additional term favoring $\mathbf{B_r} = 0$.

We obtain the opposite case at the end of the learning phase, when $h \ll \delta$. The neighborhood interaction then falls off very rapidly and in the extreme case is negligible compared to the learning step length δ. Evidently, the time rate of change $\dot{\mathbf{B}}_\mathbf{r}$ in this approximation is given by the expression

$$\dot{\mathbf{B}}_\mathbf{r} = -\delta \cdot \mathbf{B_r}(\mathbf{B_r}+\mathbf{1})(\mathbf{B_r}+\mathbf{1})^T, \tag{15.24}$$

which we have already discussed thoroughly. By itself, this expression produced unsatisfactory convergence of the system as a whole, to the degree that initial values could lie in the wrong region of attraction. However, now neighborhood cooperation can pull the values of all $\mathbf{B_r}$ into the potential well at $\mathbf{B_r} = 0$ before entering the final phase of the learning process, which allows it to be completed successfully. The manner in which this occurs is shown by consideration of the intermediate learning phase.

The intermediate learning phase is characterized as a state lying between the two previous extreme cases, *i.e.*, a state for which $h \approx \delta$ and $h, \delta \ll 1$ hold simultaneously. If we neglect the terms of quadratic and higher order in these factors in (15.22), one obtains

$$\dot{\mathbf{B}}_{\mathbf{r}} = h \sum_{\langle \mathbf{s} \rangle} (\mathbf{B}_{\mathbf{s}} - \mathbf{B}_{\mathbf{r}}) - \delta \cdot \mathbf{B}_{\mathbf{r}}(\mathbf{B}_{\mathbf{r}} + \mathbf{1})(\mathbf{B}_{\mathbf{r}} + \mathbf{1})^T. \tag{15.25}$$

15.3.1 One-Dimensional Case

If we again discuss this approximation for the one-dimensional case, a very interesting situation occurs. Here, it is again possible to state a potential for $\dot{B} = -dV/dB$, namely

$$V = \frac{h}{4} \sum_{\mathbf{r}} \sum_{\langle \mathbf{s} \rangle} (B_{\mathbf{s}} - B_{\mathbf{r}})^2 + \delta \sum_{\mathbf{r}} \Big(\frac{1}{4} B_{\mathbf{r}}^4 + \frac{2}{3} B_{\mathbf{r}}^3 + \frac{1}{2} B_{\mathbf{r}}^2 \Big). \tag{15.26}$$

Our “spheres in honey” again move in the potential whose shape is shown in Fig. 15.1, but now the “spheres” of each lattice site are coupled via springs to the nearest neighbors. In contrast to the potential (15.23), one now has the necessary additional term favoring the desired value $B_{\mathbf{r}} = 0$. Figure 15.4 presents this new situation.

“Spheres” that are stuck on the undesired plateau at $B_{\mathbf{r}} = -1$ can now be “pulled” or “pushed” off the plateaus by a neighbor located inside the well at $B_{\mathbf{r}} = 0$.

In principle, the system as a whole can still of course remain stuck outside the desired state. For example, this is the case when all initial values without exception lie in the interval $[-\infty, -1]$. All matrices then converge simultaneously to the value $B_{\mathbf{r}} = -1$, and the coupling via springs may even accelerate this convergence. However, this situation becomes more and more unlikely as the number N of lattice points increases: The probability of such an occurrence decreases exponentially like α^N, where $\alpha < 1$ gives the probability that the initial value of $B_{\mathbf{r}}$ lies to the left of the plateau $B_{\mathbf{r}} = -1$.

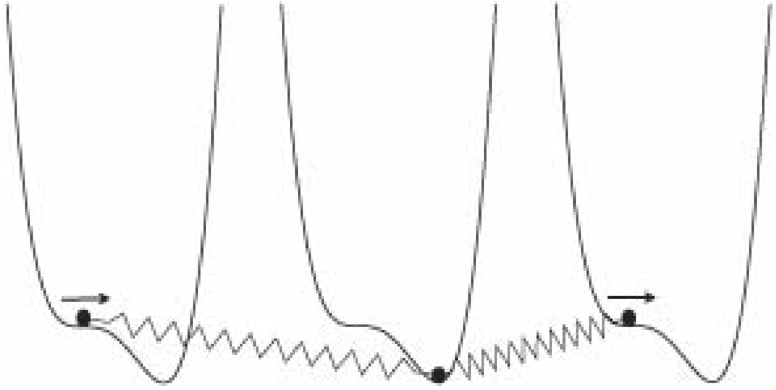

Abb. 15.4: An illustration of the convergence behavior of B in the presence of additional coupling by means of springs. The spheres that are "stuck" on the plateau (left and right) are pulled or pushed into the potential well at $B = 0$ by the springs from the adjacent sphere (middle) and are thus able to assume the desired position in the global minimum.

15.3.2 Multi-Dimensional Case

For the further investigation of convergence properties in the multidimensional case, we consider the quantity

$$S(t) := \sum_{\mathbf{r}} \|\mathbf{B}_{\mathbf{r}}(t)\|. \tag{15.27}$$

For each iteration 1–3, one has

$$\begin{aligned}
\Delta\|\mathbf{B}_{\mathbf{r}}(t)\|^2 &= 2\mathrm{Tr}\ \Delta\mathbf{B}_{\mathbf{r}}(t)\mathbf{B}_{\mathbf{r}}(t)^T \\
&= 2h_{\mathbf{rs}}\mathrm{Tr}\ \Big[\Big(\mathbf{B}^* - \mathbf{B}_{\mathbf{r}}(t)\Big)\mathbf{B}_{\mathbf{r}}(t)^T\Big] \\
&\leq 2h_{\mathbf{rs}}\Big(\|\mathbf{B}^*\| - \|\mathbf{B}_{\mathbf{r}}(t)\|\Big)\|\mathbf{B}_{\mathbf{r}}(t)\| \\
&= 2h_{\mathbf{rs}}\Big(\Delta_L\|\mathbf{B}_{\mathbf{s}}(t)\| + \|\mathbf{B}_{\mathbf{s}}(t)\| - \|\mathbf{B}_{\mathbf{r}}(t)\|\Big)\|\mathbf{B}_{\mathbf{r}}(t)\|,
\end{aligned} \tag{15.28}$$

where we have written $\|\mathbf{B}^*\| - \|\mathbf{B}_{\mathbf{s}}(t)\| =: \Delta_L\|\mathbf{B}_{\mathbf{s}}(t)\|$. Inequality (15.28) yields

$$\Delta\|\mathbf{B}_{\mathbf{r}}(t)\| \leq h_{\mathbf{rs}}\Big(\Delta_L\|\mathbf{B}_{\mathbf{s}}(t)\| + \|\mathbf{B}_{\mathbf{s}}(t)\| - \|\mathbf{B}_{\mathbf{r}}(t)\|\Big), \tag{15.29}$$

where we recall that $\Delta_L\|\mathbf{B}_{\mathbf{s}}(t)\|$ also depends on the target vector $\mathbf{u}$, which as before is assumed to be a random variable with $\langle\mathbf{u}\mathbf{u}^T\rangle = \mathbf{1}$. For the change of the quantity $S(t)$ with an iteration step, averaged over $\mathbf{u}$ and lattice sites $\mathbf{s}$ by taking into account the symmetry of $h_{\mathbf{rs}}$ and equation (15.29), one obtains

$$\begin{aligned}\langle \Delta S(t)\rangle_{\mathbf{s},\mathbf{u}} &\leq \frac{1}{N}\sum_{\mathbf{r},\mathbf{s}} h_{\mathbf{rs}}\Big(\|\mathbf{B}_{\mathbf{s}}(t)\| - \|\mathbf{B}_{\mathbf{r}}(t)\| + \Delta_L\|\mathbf{B}_{\mathbf{s}}(t)\|\Big)\\ &= \frac{h}{N}\sum_{\mathbf{s}}\langle \Delta_L \|\mathbf{B}_{\mathbf{s}}(t)\|\rangle_{\mathbf{u}} \leq 0, \qquad (15.30)\end{aligned}$$

where N is the number of lattice sites, and $h = \sum_{\mathbf{r}} h_{\mathbf{rs}}$. If we ignore boundary effects, h is independent of $\mathbf{s}$. Without lateral interaction, *i.e.*, $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$, we would have obtained (15.30) with $h = 1$. Hence, because of lateral interaction, the convergence rate is raised by a factor of h. Since h is a measure for the size of the neighborhood region participating in a learning step, this region should be chosen as large as possible consistent with the requirement of small variations of $\mathbf{A}_{\mathbf{r}}$ and $\mathbf{B}_{\mathbf{r}}$.

This result concerning the convergence rate is still quite general, since we have not yet used special properties of the learning rule for $\Delta_L \mathbf{B}$. This will done in the remainder of this section, where we will show that lateral interaction leads to an effective enlargement of the attraction region about the desired fixed point of the learning rule (15.5), thus raising the robustness of the algorithm against poorly chosen initial values.

To this end, we first show two lemmas.

Lemma 1: Let $h_{\mathbf{rs}}$ be nonnegative, symmetric with respect to commutation of $\mathbf{r}$ and $\mathbf{s}$ and nonvanishing at least for all nearest neighbor pairs $\mathbf{r}$ and $\mathbf{s}$ of the lattice. Let $Q(t) := \sum_{\mathbf{r}} \|\mathbf{B}_{\mathbf{r}}(t)\|^2$. Then the mean change $\langle \Delta Q\rangle$ per learning step vanishes only if all norms $\|\mathbf{B}_{\mathbf{r}}(t)\|$ are equal.

Proof: From (15.28) and $\Delta_L \|\mathbf{B}_{\mathbf{s}}(t)\| \leq 0$, we obtain

$$\Delta Q \leq 2\sum_{\mathbf{r}} h_{\mathbf{rs}}\Big(\|\mathbf{B}_{\mathbf{s}}(t)\| - \|\mathbf{B}_{\mathbf{r}}(t)\|\Big)\|\mathbf{B}_{\mathbf{r}}(t)\|. \qquad (15.31)$$

Averaging over $\mathbf{s}$ and taking into account the symmetry of $h_{\mathbf{rs}}$ yields

$$\langle \Delta Q\rangle_{\mathbf{s}} \leq -\frac{1}{N}\sum_{\mathbf{r},\mathbf{s}} h_{\mathbf{rs}}\Big(\|\mathbf{B}_{\mathbf{s}}(t)\| - \|\mathbf{B}_{\mathbf{r}}(t)\|\Big)^2. \qquad (15.32)$$

Together with $h_{\mathbf{rs}} > 0$ for all nearest-neighbor pairs $\mathbf{r}$, $\mathbf{s}$, this proves the claim.

With respect to convergence to the desired fixed point $\mathbf{B} = 0$, all matrices $\mathbf{B}_{\mathbf{r}}(t)$ share the same fate: either all of them converge to $\mathbf{B} = 0$, or else all of them tend to the manifold M of undesired fixed points. However, as soon as the mean of the $\|\mathbf{B_r}(t)\|$ of the lattice gets below the value unity, at least some of the $\|\mathbf{B_r}(t)\|$ must converge to $\mathbf{B} = 0$ by (15.30) and Theorem 1. But this induces the convergence of all the others to $\mathbf{B} = 0$, no matter how bad their initial values may have been. Without lateral interaction, *i.e.*, $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$, one does not have this result. In this case (15.32) does not imply a restriction on the norms $\|\mathbf{B_r}(t)\|$, and Lemma 1 no longer applies. Hence, lateral interaction enables those lattice sites with good initial values to extend the zone of convergence about the desired fixed point for all the other lattice sites. As a consequence, even if a considerable portion of the lattice sites has poor initial values, the common convergence of all matrices $\mathbf{B_r}(t)$ to the desired fixed point cannot be prevented.

It is even possible to improve the bound for the mean norm $\|\mathbf{B_r}(t)\|$ below which convergence is guaranteed. For this, we prove

Lemma 2: For sufficiently small step sizes δ, the expectation value $\langle d(\mathbf{B}(t))\rangle_{\mathbf{u}}$ of the function $d(\mathbf{B}) = \|\mathbf{B}(\mathbf{B}+\mathbf{1})\|^2$ obeying Eq. (15.12) satisfies the inequality

$$\langle d(\mathbf{B}(t))\rangle_{\mathbf{u}} \geq d(\mathbf{B}(0)) \cdot \exp(-2\delta\lambda t). \tag{15.33}$$

Here, λ is a constant upper bound for the matrix $\mathbf{H}$ of (15.17) over the complete time evolution, which is equivalent to

$$\lambda \geq \sup_{\mathbf{B}(t)} \|\mathbf{H}(\mathbf{B}(t))\|. \tag{15.34}$$

(Such an upper bound can always be determined, since $\|\mathbf{H}\|$ is bounded by some polynomial in $\|\mathbf{B}\|$, which itself is bounded). *Proof:* From (15.16) and $\mathrm{Tr}\ \mathbf{AB} \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$ one has

$$\frac{\langle \Delta d(\mathbf{B})\rangle_{\mathbf{u}}}{d(\mathbf{B})} \geq -2\delta\|\mathbf{H}(\mathbf{B})\| \geq -2\delta\lambda. \tag{15.35}$$

For sufficiently small δ, we can replace (15.35) by

$$\langle \Delta \ln d(\mathbf{B})\rangle_{\mathbf{u}} \geq -2\delta\lambda. \tag{15.36}$$

This yields

$$\begin{aligned} \langle d(\mathbf{B}(t))\rangle_{\mathbf{u}} &\geq \exp\Big(\langle \ln(d(\mathbf{B}(t))\rangle_{\mathbf{u}}\Big) \\ &\geq d(\mathbf{B}(0)) \cdot \exp(-2\delta\lambda t), \end{aligned} \tag{15.37}$$

which proves the claim.
One thus has

$$\begin{aligned}\langle \Delta_L \|\mathbf{B}_\mathbf{s}(t)\|\rangle_\mathbf{u} &= \frac{1}{2}\langle \Delta_L \|\mathbf{B}_\mathbf{s}(t)\|^2\rangle_\mathbf{u} / \|\mathbf{B}_\mathbf{s}(t)\| \\ &= -\frac{\delta}{2}\langle \|\mathbf{B}_\mathbf{s}(t)(\mathbf{B}_\mathbf{s}(t) + \mathbf{1})\mathbf{u}\|^2\rangle_\mathbf{u} / \|\mathbf{B}_\mathbf{s}(t)\| \\ &= -\frac{\delta}{2} d(\mathbf{B}_\mathbf{s}(t)) / \|\mathbf{B}_\mathbf{s}(t)\|. \qquad (15.38)\end{aligned}$$

Equations (15.31), (15.38) and Lemma 2 yield

$$\begin{aligned}\langle \Delta S(t)\rangle_{\mathbf{s},\mathbf{u}} &\leq -\frac{h\delta}{2N}\sum_\mathbf{s} \frac{d(\mathbf{B}_\mathbf{s}(t))}{\|\mathbf{B}_\mathbf{s}(t)\|} \\ &\leq -\frac{h\delta e^{-2\delta\lambda t}}{2N}\sum_\mathbf{r} \frac{d(\mathbf{B}_\mathbf{r}(0))}{\|\mathbf{B}_\mathbf{r}(t)\|}. \qquad (15.39)\end{aligned}$$

This shows that $\|\mathbf{B}_\mathbf{r}(t)\|$ decreases on the average. Hence the replacement of the denominator $\|\mathbf{B}_\mathbf{r}(t)\|$ with $\|\mathbf{B}_\mathbf{r}(0)\|$ should not destroy the inequality. It then follows that

$$\langle \Delta S(t)\rangle_{\mathbf{s},\mathbf{u}} \leq -\frac{h\delta e^{-2\delta\lambda t}}{2N}\sum_\mathbf{r} \frac{d(\mathbf{B}_\mathbf{r}(0))}{\|\mathbf{B}_\mathbf{r}(0)\|}. \qquad (15.40)$$

Integration of this equation gives the final result

$$\lim_{t\to\infty}\langle S(t)\rangle_{\mathbf{s},\mathbf{u}} \leq S(0) - \frac{h}{2\lambda}D_0, \qquad (15.41)$$

with

$$\begin{aligned}D_0 &= -\frac{1}{2N}\sum_\mathbf{r} \frac{d(\mathbf{B}_\mathbf{r}(0))}{\|\mathbf{B}_\mathbf{r}(0)\|} \\ &= -\frac{1}{N\delta}\sum_\mathbf{r}\langle \Delta_L \|\mathbf{B}_\mathbf{r}(0)\|\rangle_\mathbf{u}. \qquad (15.42)\end{aligned}$$

The quantity described by $-D_0$ can be interpreted as the average initial change of $\|\mathbf{B}\|$ of a lattice site due to the learning rule (15.5), but with respect to $\delta = 1$.
Equation (15.42) shows that on the average each $\|\mathbf{B}_\mathbf{r}(0)\|$ is shifted by at least $hD_0/2N\lambda$ towards the desired fixed point $\mathbf{B} = 0$. The bound of unity

given above for the critical value of the mean norm $\|\mathbf{B}_{\mathbf{r}}(t)\|$ leading to global convergence rises by this shift, which is proportional to the strength of the lateral interaction.

This concludes our theoretical discussion of the properties of the learning algorithm.