8 The Adaptive-Subspace Self-Organizing Map (ASSOM)

Teuvo Kohonen, Samuel Kaski, and Harri Lappalainen

A long-standing goal in our research has been to find out how certain invariant-feature filters may emerge in learning processes. This problem was recently solved by one of the authors [1-3]. The key insight was that if input patterns must be recognizable invariantly to certain transformations, the members in *natural sequences of* such patterns must also be produced from each other by the same transformations. If the sequences are relative short, one may think that a particular transformation predominates in them, and the successive patterns then belong to some linear subspace that corresponds to this transformations. Such signal subspaces can be learned by the architecture delineated in Fig. 6.

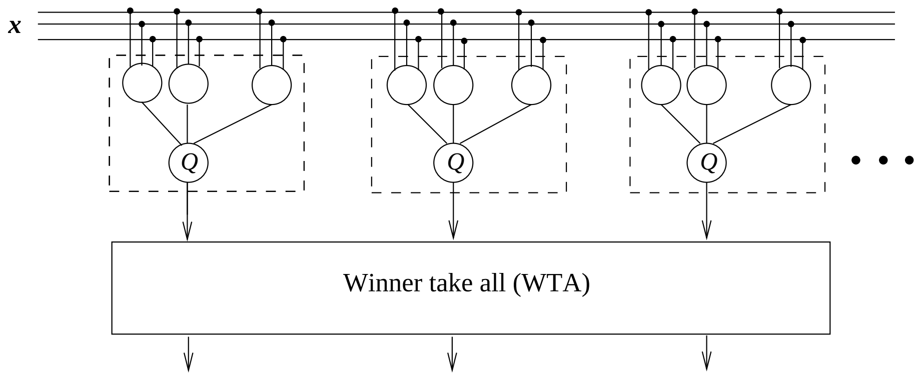


Figure 6: The ASSOM architecture.

Each dotted line in Fig. 6 distinguishes a module, a processing unit in a special SOM array. The first-layer neurons are linear and they output the sums of dot products of $\mathbf{x}$ with the various synaptic input weight vectors. The second-layer neurons (Q) form quadratic functions of the first-layer neuron outputs. If the weight vectors of the linear layer are orthonormalized, the neurons of the output layer shall form sums of squares of their inputs. The circuit represented by Fig. 6 can then be shown to compare the input pattern $\mathbf{x}$ with the linear subspaces spanned by the weight vectors of the first layer. If the weight vectors can be defined in a way in which their linear combinations represent some transformation groups, then matching becomes invariant with respect to these groups. Below it will be shown that such weight vectors emerge in an unsupervised learning process.

The outputs from the modules shall further be compared by a winner-take-all (WTA) function, which in Fig. 6 is shown as a separate operation. The WTA function specifies the "winner" module, indexed by c, as defined below; module c and its neighboring modules in the array will be updated in proportion to the so-called neighborhood function h_{ci} like in a usual SOM. In a physical network the WTA function may be integrated with the modules, for instance by their lateral interaction.

Let us denote the input weight vector, indexed by h of module i by $\mathbf{b}_h^{(i)}$. The $\mathbf{b}_h^{(i)}$ of the same module are now assumed orthonormal; at least they can be orthonormalized easily. They can then be regarded as the *orthonormal basis vectors of some linear subspace* $\mathcal{L}^{(i)}$, or a set of vectors $\mathbf{x}$, where every x can be expressed as a general linear combination of the $\mathbf{b}_h^{(i)}$.

Let $\hat{\mathbf{x}}^{(i)}$ be the *orthogonal projection* of $\mathbf{x}$ on $\mathcal{L}^{(i)}$, or

$$\hat{\mathbf{x}}^{(i)} = \sum_h \mathbf{b}_h^{(i)\mathrm{T}} \mathbf{x} \; . \tag{63}$$

For an arbitrary $\mathbf{x}$ that need not belong to $\mathcal{L}^{(i)}$ one can define its *distance* d from $\mathcal{L}^{(i)}$, defined by

$$d^2 = d^2(\mathbf{x}, \mathcal{L}^{(i)}) = \|\mathbf{x}\|^2 - \|\hat{\mathbf{x}}^{(i)}\|^2 \; . \tag{64}$$

The *vectorial projection error* is the residual

$$\tilde{\mathbf{x}}^{(i)} = x - \hat{\mathbf{x}}^{(i)} \; . \tag{65}$$

For an arbitrary $\mathbf{x}$, its *minimum projection error* can be defined as the distance of $\mathbf{x}$ *from the closest subspace* $\mathcal{L}^{(i)}$, and the "winner subspace" with index c is defined by

$$\|\tilde{\mathbf{x}}^{(c)}\| \;=\; \min_i\{\|\tilde{\mathbf{x}}^{(i)}\|\} \; , \text{ or} \tag{66}$$

$$\|\hat{\mathbf{x}}^{(c)}\| \;=\; \max_i\{\|\hat{\mathbf{x}}^{(i)}\|\} \; . \tag{67}$$

Our goal is to let all the modules of Fig. 6 approximate $\mathbf{x}$ by its different projections, and always select the module that produces the best approximation over the array. The objective function that defines the *average expected spatially weighted normalized squared projection error* is

$$E_1 = \int \sum_i h_{ci} \frac{\|\tilde{\mathbf{x}}^{(i)}\|^2}{\|\mathbf{x}\|^2} p(\mathbf{x}) d\mathbf{x} \; , \tag{68}$$

where h_{ci} is the neighborhood function that defines the interaction of modules c and i like in a usual SOM, and c is the index of the winner subspace $\mathcal{L}^{(i)} = \mathcal{L}^{(c)}$. Notice that c is a function of $\mathbf{x}$ and all the basis vectors $\mathbf{b}_h^{(i)}$.

Minimization of (68), i.e., selection of the basis vectors $\mathbf{b}_h^{(i)}$ for all subspaces $\mathcal{L}^{(i)}$ such that the *average expected distance of* $\mathbf{x}$ *from the closest subspace is minimized*, is a rather complicated process [1-3]. Some extra problems are caused by the stability of the recursion by which E_1 is minimized. Without quoting all the details it may be mentioned that if the Robbins-Monro stochastic approximation process [4] is used, the optimal values of the $\mathbf{b}_h^{(i)}$ are obtained in the recursion [5].

$$\mathbf{b}_h^{(i)}(t+1) = \mathbf{b}_h^{(i)}(t) + \lambda(t) h_{ci}(t) \frac{\mathbf{x}(t) x^{\mathrm{T}}(t)}{\|\mathbf{x}(t)\|^2} \mathbf{b}_h^{(i)}(t) \; . \tag{69}$$

Consider now an "episode" $\mathcal{S}$ that consist of a finite set of successive sampling times t_p; denote $\mathcal{S} = \{t_p\}$. The set of samples $X = \{\mathbf{x}(t_p) | t_p \in \mathcal{S}\}$ has to be recognized as one class, such that any member of X and even an arbitrary linear combination of the $\mathbf{x}(t_p), t_p \in \mathcal{S}$ shall be decoded by the same module of Fig. 6 (subspace $\mathcal{L}^{(i)}$). In learning, the vector set X, defined as the Cartesian product of the $\mathbf{x}(t_p), t_p \in \mathcal{S}$, must be taken as one batch, instead of optimizing the error using single patterns $\mathbf{x}(t_p)$ one at a time. The error minimization problem will now be modified by defining the new objective function in terms of the *average expected spatially weighted normalized squared projection error over the episodes*:

$$E_2 = \int \sum_{t_p \in \mathcal{S}} \sum_i h_{ci} \frac{\|\tilde{\mathbf{x}}^{(i)}(t_p)\|^2}{\|\mathbf{x}(t_p)\|^2} p(X) dX \ . \tag{70}$$

Here $p(X)$ is the joint probability density for the samples $\mathbf{x}(t_p), t_p \in \mathcal{S}$ that produce the Cartesian product set X, and dX is a shorthand notation meaning a volume differential in the Cartesian product space of the $\mathbf{x}(t_p)$.
Minimization of (70) defines the basis vectors $\mathbf{b}_h^{(i)}$ and a set of analyzers that are *optimally invariant to the transformations that occur in the input signal patterns.*
The Robbins-Monro stochastic approximation is applicable to the minimization of E_2, too, when the gradient step is made to consist of the whole episode $\mathcal{S}$. The learning phase is then desribed by the following equation:

$$\mathbf{b}_h^{(i)}(t+1) = \mathbf{b}_h^{(i)}(t) + \lambda(t) h_{c_r}^{(i)} \sum_{t_p \in \mathcal{S}(t)} \frac{\mathbf{x}(t_p)\mathbf{x}^{\mathrm{T}}(t_p)}{\|\mathbf{x}(t_p)\|^2} \mathbf{b}_h^{(i)}(t) \ . \tag{71}$$

When $\lambda(t)$ is small, (71) is equivalent with the following learning process in which the basis vectors are formed by a product of elementary projection operators, each one corresponding to one pattern $\mathbf{x}(t_p), t_p \in \mathcal{S}$:

$$\mathbf{b}_h^{'(i)} = \prod_{t_p \in \mathcal{S}} \left[I + \alpha(t_p) h_{c_r}^{(i)} \frac{\mathbf{x}(t_p)\mathbf{x}^{\mathrm{T}}(t_p)}{\|\hat{\mathbf{x}}^{(i)}(t_p)\| \, \|\mathbf{x}(t_p)\|} \right] b_h^{(i)}(t_p) \ . \tag{72}$$

The special learning-rate factor $\lambda = \alpha(t_p)\|\mathbf{x}(t_p)\| / \|\hat{\mathbf{x}}^{(i)}(t_p)\|$ in (72) has been chosen for stability reasons.
There are several other minor details in the process that improve the algorithm [3,5]. We have produced various ASSOM filters for very different input data [1,2]. Here a simple demonstration, illustrating the basic idea, is shown.
Over the input field we generated patterns consisting of colored noise (white noise, low-pass filtered by a second-order Butterworth filter with cut-off frequency of 0.6 times the Nyquist frequency of the sampling lattice). The input episodes for learning were formed by taking samples from this data field. The mean of the samples was always subtracted from the pattern vector.
In the *translation-invariant filter* experiment, the episodes were formed by shifting the receptive field randomly into five nearby locations, the average shift thereby being ± 2 pixels in both dimensions. Fig. 8 shows the basis vectors $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$, similar to Gabor filters, in a gray scale at each array point of a two-dimensional ASSOM. One should notice that the spatial frequencies of the basis vectors of the same unit are the same, but the $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$ are mutually 90 degrees out of phase. (The absolute phase of $\mathbf{b}_{i1}$ can be zero or 180 degrees, though.)
The episodes for the *rotation filters* were formed by rotating the input field at random five times in the range of zero to 60 degrees, the rotation center coinciding with the

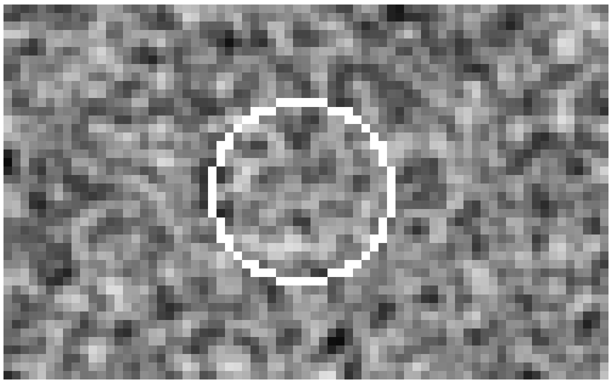
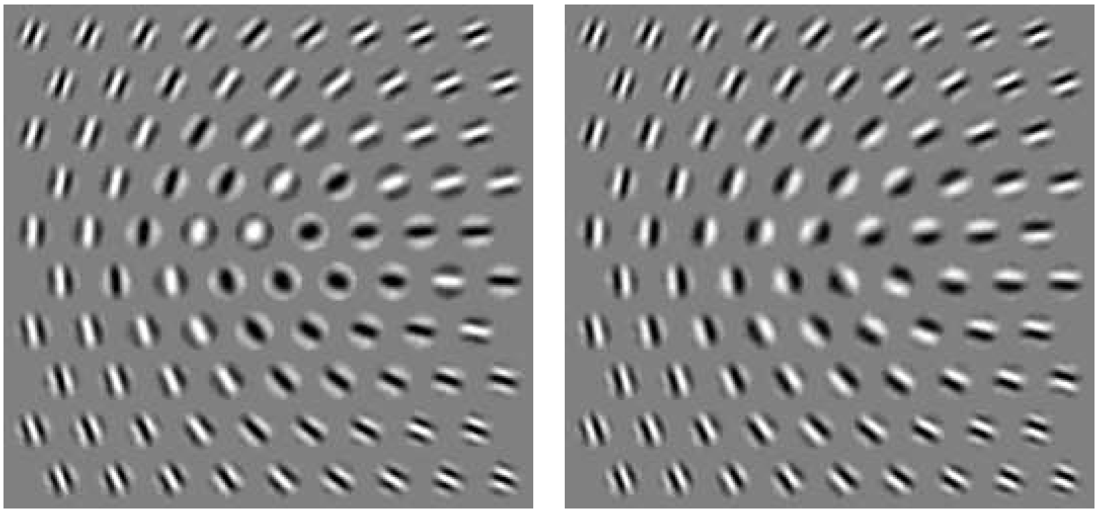


Figure 7: Colored noise (second-order Butterworth-filtered white noise with cut-off frequency of 0.6 times the Nyquist frequency of the lattice) used as input data. The receptive field is demarcated by the white circle.



(a) (b)

Figure 8: The ASSOM that has formed Gabor-type filters: (a) The $\mathbf{b}_{i1}$, (b) The $\mathbf{b}_{i2}$.

center of the receptive field. Fig. 9 shows the rotation filters thereby formed at the ASSOM units; clearly they are sensitive to azimuthal optic flow.

Scale-invariant filters were formed by zooming the input pattern field, with the center of the receptive field coinciding with the zooming center. The filters thereby formed, shown in Fig. 10, have clearly become sensitive to radial optic flow, corresponding to approaching or withdrawing objects.

References

[1] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg, 1995.

[2] T. Kohonen. Emergence of invariant-feature detectors in self-organization. In Palaniswami, M. et al. (eds.), *Computational Intelligence, A Dynamic System*

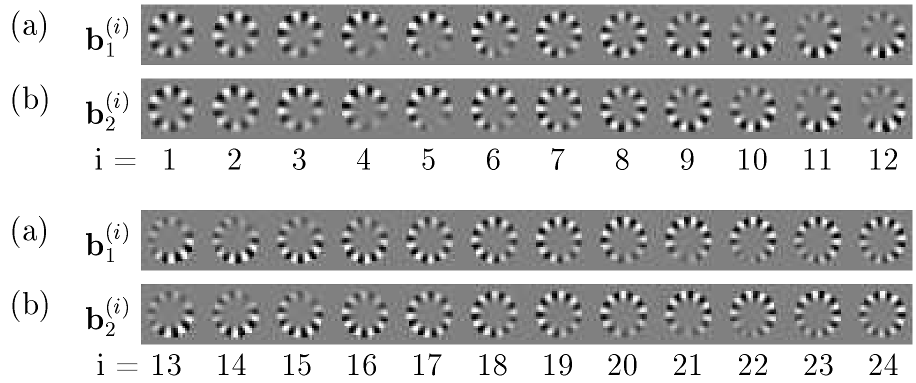


Figure 9: One-dimensional rotation-invariant ASSOM. (a) Cosine-type "azimuthal wavelets" ($\mathbf{b}_{i1}$), (b) Sine-type "azimuthal wavelets" ($\mathbf{b}_{i2}$). Notice that the linear array has been shown in two parts.

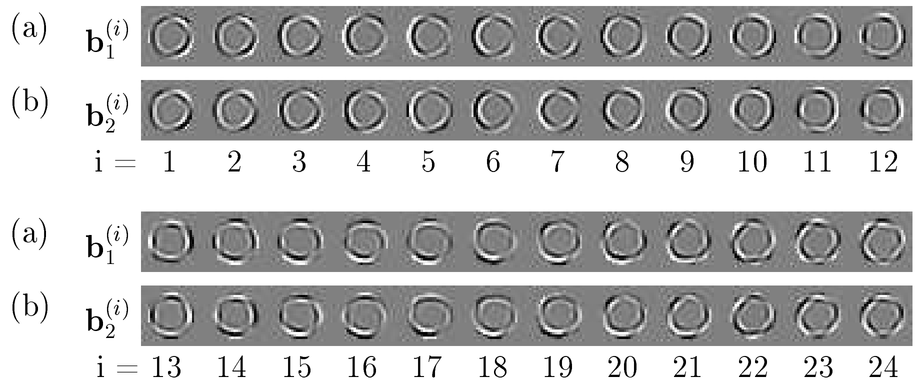


Figure 10: One-dimensional zoom-invariant ASSOM. (a) Cosine-type "radial wavelets" ($\mathbf{b}_{i1}$), (b) Sine-type "radial wavelets" ($\mathbf{b}_{i2}$). Notice that the linear array has been shown in two parts.

Perspective, pp. 17-31. IEEE Press, New York, 1995.

[3] T. Kohonen. Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75:281-291, 1996.

[4] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407, 1951.

[5] T. Kohonen, S. Kaski, and H. Lappalainen. Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, 9:1321-1344, 1997.