where η is the learning-rate parameter. The average values $\bar{x}$ and $\bar{y}$ constitute respective presynaptic and postsynaptic thresholds, which determine the sign of synaptic modification. In particular, the covariance hypothesis allows for the following:

- convergence to a nontrivial state, which is reached when $x_k = \bar{x}$ or $y_j = \bar{y}$;
- prediction of both synaptic *potentiation* (i.e., increase in synaptic strength) and synaptic *depression* (i.e., decrease in synaptic strength).

Figure A illustrates the difference between Hebb's hypothesis and the covariance hypothesis. In both cases, the dependence of Δw_{kj} on y_k is linear; however, the intercept with the y_k-axis in Hebb's hypothesis is at the origin, whereas in the covariance hypothesis it is at $y_k = \bar{y}$.

We make the following important observations from Eq. (A):

1. The synaptic weight w_{kj} is enhanced if there are sufficient levels of presynaptic and postsynaptic activities—that is, the conditions $x_j > \bar{x}$ and $y_k > \bar{y}$ are both satisfied.
2. The synaptic weight w_{kj} is depressed if there is either
   - a presynaptic activation (i.e., $x_j > \bar{x}$) in the absence of sufficient postsynaptic activation (i.e., $y_k < \bar{y}$), or
   - a postsynaptic activation (i.e., $y_k > \bar{y}$) in the absence of sufficient presynaptic activation (i.e., $x_j < \bar{x}$).

This behavior may be regarded as a form of temporal competition between the incoming patterns.

4. *Historical Note.* Long before the publication of Sanger's GHA in 1989, Karhunen and Oja (1982) published a conference paper that described a new algorithm, called the *stochastic gradient algorithm* (*SGA*), derived for the purpose of computing the eigenvectors of PCA. It turns out that the SGA is very close in its composition to the GHA.
5. *Wavelets.* The preface to the book by Mallat (1998) states the following:

> *Wavelets are based not on a "bright new idea," but on concepts that already existed under various forms in many different fields. The formalization and emergence of this "wavelet theory" is the result of a multidisciplinary effort that brought together mathematicians, physicists and engineers, who recognized that they were independently*

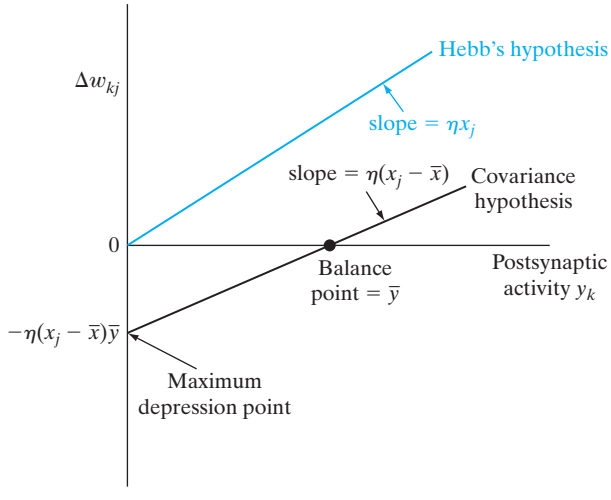

FIGURE A Illustration of Hebb's hypothesis and the covariance hypothesis.

developing similar ideas. For signal processing, this connection has created a flow of ideas that goes well beyond the construction of new bases or transforms.

Let $\psi(t)$ denote a function of zero mean, as shown by

$$\int_{-\infty}^{\infty} \psi(t)dt = 0$$

The function $\psi(t)$ represents the impulse response of a band-pass filter; such a function is called a *wavelet*. The wavelet is *dilated* with a *scale* parameter s and *shifted* in position by a *time* parameter u; we thus write

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$$

Given a real-valued signal $g(t)$ with Fourier transform $G(f)$, the *continuous wavelet transform* of $g(t)$ is defined by the inner product in integral form:

$$\begin{aligned} W_g(u,s) &= \langle \psi_{u,s}(t), g(t) \rangle \\ &= \int_{-\infty}^{\infty} g(t)\psi_{u,s}(t)dt \end{aligned}$$

According to this formula, the wavelet transform *correlates* the signal $g(t)$ with $\psi_{u,s}(t)$. Equivalently, we may write

$$\begin{aligned} W_g(u,s) &= \langle \Psi_{u,s}(f), G(f) \rangle \\ &= \int_{-\infty}^{\infty} G(f)\Psi_{u,s}^*(f)df \end{aligned}$$

where $\Psi_{u,s}(f)$ is the Fourier transform of $\psi_{u,s}(t)$ and the asterisk denotes complex conjugation. We thus see that the wavelet transform $W_g(u,s)$ depends on the values of the signal $g(t)$ and its Fourier transform $G(f)$ in the time–frequency domain, where the energy of $\psi_{u,s}(t)$ and that of its Fourier transform $\Psi_{u,s}(f)$ are concentrated.

For authoritative treatment of the wavelet transform, the reader is referred to the books by Mallat (1998) and Daubechies (1992, 1993). The introductory book by Meyer (1993) includes a historical perspective of the wavelet transform.

6. *Nonlinear PCA methods.*

These methods may be categorized into four classes:

(i) **Hebbian networks**, where the linear neurons in the generalized Hebbian algorithms, for example, are replaced with nonlinear neurons (Karhunen and Joutsensalo, 1995).

(ii) **Replicator networks** or **autoencoders**, which are built around multilayer perceptrons containing three hidden layers (Kramer, 1991):

- mapping layer;
- bottleneck layer;
- demapping layer.

Replicator networks were discussed in Chapter 4.

(iii) **Principal curves**, which are based on an iterative estimation of a curve or a surface capturing the structure of the input data (Hastie and Stuetzle, 1989). The self-organizing maps, studied in Chapter 9, may be viewed as a computational procedure for finding a discrete approximation of principal curves.

(iv) **Kernel PCA**, originated by Schölkopf et al. (1998), was studied in Section 8.8 of this chapter.

7. In Kim et al. (2005), the results of image-denoising experiments involving the KHA are also presented for the following scenarios:
   - superresolution and denoising of face (single-patch) images;
   - multipatch superresolution images of natural scenes.
8. The *median filter* is a filter (estimator) that minimizes the Bayes risk for the absolute error cost function

$$R(e(n)) = |e(n)|$$

   where $e(n)$ is the *error signal*, defined as the difference between a desired response and the actual response of the filter. It turns out that the result of this minimization is the median of the posterior probability density function—hence the name of the filter.
9. *Adaptive Wiener filters.* The Wiener filter was studied in Chapter 3. In the adaptive Wiener filter, the training sample $\{\mathbf{x}_i(n), \mathbf{d}_i(n)\}_{i=1}^{N}$ is split into a successive sequence of windowed batches of labeled data, and the filter parameters are computed using the normal equations (or discrete form of the Wiener–Hopf equation) on a batch-by-batch basis. In effect, within each batch, the data are viewed as pseudostationary, and the statistical variations in the training sample show up in the corresponding changes in the filter parameters as the computation proceeds from one batch to the next.
10. The *Sobolev space* is the space of all functions that contains all m derivatives in the space L^m and in which the mth derivative is absolutely integrable (Vapnik, 1998). The *Besov space* refines the condition for *smoothness* by including a third parameter for $m = 1$ and $m = \infty$.

PROBLEMS

Competition and Cooperation

8.1 In a self-organizing system that involves competition as well as cooperation, we find that competition precedes cooperation. Justify the rationale behind this statement.

Principal-Components Analysis: Constrained-Optimization Approach

8.2 In Section 8.4, we used perturbation theory to derive the PCA. In this problem, we address this same issue from the perspective of a constrained-optimization approach.

Let $\mathbf{x}$ denote an m-dimensional zero-mean data vector and $\mathbf{w}$ denote an adjustable parameter vector of the same dimension m. Let σ^2 denote the variance of the projection of the data vector $\mathbf{x}$ onto the parameter vector $\mathbf{w}$.

(a) Show that the Lagrangian for maximizing the variance σ^2, subject to the normalizing condition $\|\mathbf{w}\| = 1$, is defined by

$$J(\mathbf{w}) = \mathbf{w}^T\mathbf{R}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

where $\mathbf{R}$ is the correlation matrix of the data vector $\mathbf{x}$ and λ is the Lagrangian multiplier.

(b) Using the result of part (a), show that maximization of the Lagrangian $J(\mathbf{w})$ with respect to $\mathbf{w}$ yields the defining *eigenequation*

$$\mathbf{R}\mathbf{w} = \lambda\mathbf{w}$$

Hence, show that $\sigma^2 = \mathbb{E}[(\mathbf{w}^T\mathbf{x})^2] = \lambda$. In the eigendecomposition terminology, $\mathbf{w}$ is the eigenvector and λ is the associated eigenvalue.

(c) Let the Lagrange multiplier λ_i represent the *normalizing condition* $\|\mathbf{w}_i\| = 1$ for the ith eigenvector, and let the Lagrange multiplier λ_{ij} represent the *orthogonality condition* $\mathbf{w}_i^T \mathbf{w}_j = 0$. Show that the Lagrangian now assumes the expanded form

$$J(\mathbf{w}_i) = \mathbf{w}_i^T \mathbf{R} \mathbf{w}_i - \lambda_{ii}(\mathbf{w}_i^T \mathbf{w} - 1) - \sum_{j=1}^{i-1} \lambda_{ij} \mathbf{w}_i^T \mathbf{w}_j, \qquad i = 1, 2, ..., m$$

Hence, show that the maximization of $J(\mathbf{w}_i)$ yields a set of m equations for which the optimal solution is the eigenvalue λ_i associated with the eigenvector $\mathbf{w}_i$.

8.3 Let the estimator of an m-dimensional zero-mean data vector $\mathbf{x}$ be defined by the expansion

$$\hat{\mathbf{x}}_l = \sum_{i=1}^{l} a_i \mathbf{q}_i, \qquad l \leq m$$

where $\mathbf{q}_i$ is the ith eigenvector of the correlation matrix

$$\mathbf{R} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$$

and $a_1, a_2, ..., a_l$ are the coefficients of the expansion, subject to the condition

$$\mathbf{q}_i^T \mathbf{q}_j = \begin{cases} 1 & \text{for } j = i \\ 0 & \text{otherwise} \end{cases}$$

Show that minimization of the mean-square error

$$J(\hat{\mathbf{x}}_i) = \mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}_i\|^2]$$

with respect to the adjustable coefficients $a_1, a_2, ..., a_l$ yields the defining formula

$$a_i = \mathbf{q}_i^T \mathbf{x}, \quad i = 1, 2, ..., l$$

as the ith principal component—that is, the projection of the data vector $\mathbf{x}$ onto the eigenvector $\mathbf{q}_i$.

8.4 Following on the constrained-optimization problem considered in Problem 8.2, consider the Lagrangian

$$J(\mathbf{w}) = (\mathbf{w}^T \mathbf{x})^2 - \lambda(\mathbf{w}^T \mathbf{w} - 1)$$

where $(\mathbf{w}^T \mathbf{x})^2$ denotes the instantaneous value of the variance of a zero-mean data vector $\mathbf{x}$ projected onto the weight vector $\mathbf{w}$.

(a) Evaluating the gradient of the Lagrangian $J(\mathbf{w})$ with respect to the adjustable weight vector $\mathbf{w}$, show that

$$\begin{aligned} g(\mathbf{w}) &= \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \\ &= 2(\mathbf{w}^T \mathbf{x})\mathbf{x} - 2\lambda \mathbf{w} \end{aligned}$$

(b) With the stochastic gradient ascent in mind for on-line learning, we may express the weight-update formula as

$$\hat{\mathbf{w}}(n+1) = \hat{\mathbf{w}}(n) + \frac{1}{2}\eta \mathbf{g}(\hat{\mathbf{w}}(n))$$

where η is the learning-rate parameter. Hence, derive the iterative equation

$$\hat{\mathbf{w}}(n+1) = \hat{\mathbf{w}}(n) + \eta[(\mathbf{x}(n)\mathbf{x}^T(n))\hat{\mathbf{w}}(n) - \hat{\mathbf{w}}^T(n)(\mathbf{x}(n)\mathbf{x}^T(n))\hat{\mathbf{w}}(n)\hat{\mathbf{w}}(n)]$$

which is a rewrite of Eq. (8.47) defining the evolution of the maximum eigenfilter across discrete time n, with $\hat{\mathbf{w}}(n)$ written in place of $\mathbf{w}(n)$.