Self-Consistent Learning of the Environment

Kukjin Kang
kkang@brain.riken.jp
Shun-ichi Amari
Amari@brain.riken.go.jp
BSI, RIKEN, Hirosawa, Wako, Saitama, 351-0198, Japan

We study the Bayesian process to estimate the features of the environment. We focus on two aspects of the Bayesian process: how estimation error depends on the prior distribution of features and how the prior distribution can be learned from experience. The accuracy of the perception is underestimated when each feature of the environment is considered independently because many different features of the environment are usually highly correlated and the estimation error greatly depends on the correlations. The self-consistent learning process renews the prior distribution of correlated features jointly with the estimation of the environment. Here, maximum a posteriori probability (MAP) estimation decreases the effective dimensions of the feature vector. There are critical noise levels in self-consistent learning with MAP estimation, that cause hysteresis behaviors in learning. The self-consistent learning process with stochastic Bayesian estimation (SBE) makes the presumed distribution of environmental features converge to the true distribution for any level of channel noise. However, SBE is less accurate than MAP estimation. We also discuss another stochastic method of estimation, SBE2, which has a smaller estimation error than SBE without hysteresis.

1 Introduction

Perception of the surrounding environment is a basic function of the brain. Noisy and high-dimensional sensory signals from sensory organs are processed to perceive the surrounding environment, and the perception is updated every moment in the brain. There has been a great deal of effort to understand the process. The correlations between sensory stimuli and neural activities in the brain have been evaluated with various information measures (Abbott, Rolls, & Tovee, 1996; Ringach, Shapley, & Hawken, 2002; Kang & Sompolinsky, 2001; Kang, Shapley, & Sompolinsky, 2004). The statistical features of natural scenes were studied in terms of efficient coding (Simoncelli & Olshausen, 2001). Merging the various sensory information, such as visual, audio, and haptic signals, into a robust percept, has been studied (Alais & Burr, 2004; Ernst & Banks, 2002, Ernst & Bulthoff, 2004).
It is known that human perception can be explained with Bayesian estimation (Körding & Wolpert, 2004; Deneve, Latham, & Pouget, 2001; Ma, Beck, Latham, & Pouget, 2006; Bahrami et al., 2010; Oram, Foldiak, Perrett, & Sengpiel, 1998; Berniker, Voss, & Körding, 2010). For example, Weiss, Simoncelli, and Adelson (2002) considered visual information channels with independent gaussian noise. They could explain the human perception of direction of velocity by assuming that the prior distribution of movement velocity favored slow velocities.

Human predictions in everyday life were also explained by the Bayesian inference model (Griffiths & Tenenbaum, 2006). Learning of Bayesian priors for depth perception was studied by Knill (2007). Prior information makes us perceive the environment in a biased way. For example, we often have light-from-above prior and feel depth in two-dimensional images accordingly (Adams, Graf, & Ernst, 2004).

The accuracy of a perception depends strongly on the statistical regularity of the environment. When the prior distribution of an environmental feature is highly peaked, the feature can be perceived accurately with a very noisy sensory signal. The correlations between different environmental features in prior distribution are very important in the perception process because the prior probability of observing some combination of features could be much higher than other combinations. For example, the positions of a moving object at different times are strongly correlated, and the visual and acoustic features of an object are also highly correlated. Correlations in natural images should not be ignored for efficient coding (Simoncelli & Olshausen, 2001). A sensory signal may be interpreted incorrectly with erroneous prior distribution.

The difficulty of learning prior distribution has been underestimated. In addition, how erroneous prior distribution causes perception error has not been studied systematically. It has been often assumed in previous work that correct prior distribution is a given. Studies on images of natural scenes have assumed that the brain has already learned the correct distribution through adaptation or evolution (Simoncelli & Olshausen, 2001). Causal inference in multisensory perception has been studied assuming that the correct prior distribution is available to an ideal observer (Körding et al., 2007). The effect of a correlation in prior distribution has been ignored in many previous studies because each feature of the environment, such as the frequency of sound, orientation, or contrast of a visual stimulus, has been considered separately (Abbott et al., 1996; Ringach et al., 2002; Kang & Sompolinsky, 2001; Kang et al., 2004).

Learning of prior distribution is not a trivial task and should be carefully studied. Noise in the sensory information channel and distribution in the environment are inseparably mixed in the perception process. It is fundamentally impossible for parts of the nervous system to have true access to the distribution in the environment. The sensory signal always comes through information channels with noise, and any estimate of the environment
depends on the estimate of the channel property, and vice versa. Our results show that the estimated distribution often does not converge to the true one and that the perception of the environment depends on the update rule of the prior distribution.

Here, we study the estimation of the state of a complex environment through multiple sensory channels, assuming that the decoder of sensory information has no information other than the sensory signals. Since the true distribution of an environmental feature is not available, the decoder needs to make assumptions about the statistical structure of the environment. We study how an estimation error depends on the assumed prior distribution and the true one for three different estimation rules. Then we consider the cumulative learning of prior distribution. We call it self-consistent learning because learning stops when the distribution of decoded features is the same as the distribution of features assumed in the decoding process. The self-consistent learning process with maximum a posteriori (MAP) estimation decreases the effective dimensions of the feature vector and minimizes the computational cost of estimating features of the environment. There are critical noise levels in self-consistent learning, and hysteresis behavior exists because of that. The self-consistent learning process with stochastic Bayesion estimation (SBE) makes the presumed distribution of environmental features converge to the true distribution for any level of channel noise, but it is slow in learning and has a larger perceptional error because the estimation is more stochastic. We also studied another stochastic method of estimation, SBE2, which has a smaller estimation error than SBE does and without hysteresis.

2 Environment with Correlated Features

Consider sensory information channels sending a signal, \( R_i \) (\( i = 1, \ldots, N \)), about the outside world. The \( R_i \) may be the firing rate of a neuron at a specified moment or the mean firing rate of a population of neurons. We consider a decoder receiving sensory signals from \( N \) such channels.

We assume that the decoder has no other information about the outside world than the sensory signals. The decoder introduces an assumption that the sensory signals are evoked by the features of the environment, and the sensory signals are contaminated by channel noise. That is, \( R_i \) includes information about feature \( \theta_i \) of a given environment, and the total features of the surrounding environment are given by a feature vector, \( \vec{\theta} = \{\theta_i\} (i = 1, \ldots, N) \). A response vector, \( \vec{R} = \{R_i\} \), is stochastically determined by \( \vec{\theta} \) because the sensory information channels have finite reliability. Equation 2.1 shows the expression for the marginalized distribution of the sensory signals:

\[
P(\vec{R}) = \int d\vec{\theta} P(\vec{R}|\vec{\theta})P(\vec{\theta}).
\]
Here, $P(\bar{R})$ is the distribution of $\bar{R}$. The $P(\bar{R}|\bar{\theta})$ is the distribution of $\bar{R}$ for a given $\bar{\theta}$. The $P(\tilde{\theta})$ is the true distribution of the environmental feature vector. Many previous studies on the stimulus-response relation in the brain have focused on $P(\bar{R}|\bar{\theta})$ (a distribution of responses for a given set of stimuli) or the mean of $\bar{R}$ for a given $\bar{\theta}$. This is the viewpoint of outsiders with additional knowledge $P(\tilde{\theta})$ about $\bar{\theta}$, and most researchers find themselves in this situation. However, from the viewpoint of insiders without such knowledge, equation 2.1 should be considered a subjective assumption. The assumption in equation 2.1 can be introduced because the sensory signal is very noisy and better explained by a simpler model in terms of computational cost.

There is a degree of freedom in making an assumption about the outside world. For example, consider a channel sending signal $R = \theta + a + b$. The $a$ and $b$ are gaussian random noise. This situation can be described in many different but equivalent ways. We can think of $a + b$ as channel noise and $\theta$ as a constant environmental feature. We can also think of $b$ as channel noise and $\theta + a$ as a stochastic environmental feature with a gaussian distribution. Two decoders of sensory signals in an identical situation may have different subjective interpretations of the sensory signals unless they have the same model of the outside world or sensory channels. We need to remember this degree of freedom in studying estimation error and learning of $P(\tilde{\theta})$.

For a given model of reliability of the sensory channel, $P(\bar{R}|\bar{\theta})$, a decoder of the sensory signal produces an estimator, $\tilde{\theta}_0$, of $\tilde{\theta}$ from $\bar{R}$, and the accuracy of the estimator, $\tilde{\theta}_0$, sensitively depends on $P(\tilde{\theta})$. Using the degree of freedom previously explained, the decoder can assume that his or her knowledge about channel noise is correct. However, the correct $P(\tilde{\theta})$ is not directly available to the decoder and should be learned from experience. The decoder assumes $P_0(\tilde{\theta})$ as the prior distribution of $\tilde{\theta}$ instead of $P(\tilde{\theta})$ because the true model of the environment is not available. The error in estimation is caused by the difference between $P_0(\tilde{\theta})$ and $P(\tilde{\theta})$ as well as channel noise in signals. The learning process describes how to update $P_0(\tilde{\theta})$ after each experience of estimating environmental features.

The next section explains how estimation error depends on $P(\bar{R}|\bar{\theta})$, $P_0(\tilde{\theta})$, and $P(\tilde{\theta})$. After that, we consider the learning processes of $P(\tilde{\theta})$ and explain their consequences.

To consider concrete and analytically tractable cases, we assume that $P(\bar{R}|\bar{\theta})$ and $P_0(\tilde{\theta})$ are gaussian distributions:

$$P(\bar{R}|\bar{\theta}) = \frac{1}{\sqrt{2\pi |K|}} \exp \left( -\frac{1}{2} (\bar{R} - \bar{\theta})^T K^{-1} (\bar{R} - \bar{\theta}) \right),$$  
(2.2)

$$P_0(\tilde{\theta}) = \frac{1}{\sqrt{(2\pi)^N |C_0|}} \exp \left( -\frac{1}{2} (\tilde{\theta} - \tilde{\theta}_0)^T C_0^{-1} (\tilde{\theta} - \tilde{\theta}_0) \right).$$  
(2.3)
The decoder assumes that the $\theta_i$s are correlated with each other, creating a gaussian distribution, $P_0(\vec{\theta})$. The $C_0$ is the correlation matrix for $\vec{\theta}$. The $\vec{h}_0$ is the center of the gaussian distribution. We also assume that $P(\vec{\theta})$ coincides with $P_0(\vec{\theta})$ when $C_0 = C$ and $\vec{h}_0 = \vec{h}$. Learning in this situation can be called learning in an achievable reality. It is possible to think of the case where learning is unachievable, such that $P_0(\vec{\theta})$ cannot attain $P(\vec{\theta})$ by parameter tuning. However, we consider only achievable cases in this letter.

### 3 Estimation of Environmental Features

The most probable estimator of $\vec{\theta}$ is given by MAP estimation, or by maximizing $P_0(\vec{\theta} \mid \vec{R})$, or

$$
\ln P_0(\vec{\theta} \mid \vec{R}) = \ln P_0(\vec{\theta}) + \ln P(\vec{R} \mid \vec{\theta}) + O,
$$

where $O$ is independent of $\vec{\theta}$.

The maximization of equation 3.1 yields MAP estimation:

$$
\vec{\theta}_0^{\text{MAP}} = U(\vec{R} - \vec{h}_0) + \vec{h}_0,
$$

$$
U = C_0(C_0 + K)^{-1}.
$$

Note that equation 2.3 can be considered a result of the second-order approximation of $\ln P_0(\vec{\theta})$ in equation 3.1. For most $\vec{R}$, the MAP estimation value of $\vec{\theta}$ is in a local area of $\vec{\theta}$ space. $\ln P_0(\vec{\theta})$ may have one local minimum, that is, $\vec{h}_0$ in the local area, and be approximated by a series expansion: $\ln P_0(\vec{\theta}) \approx \ln P_0(\vec{h}_0) - 1/2(\vec{\theta} - \vec{h}_0)^T C^{-1}(\vec{\theta} - \vec{h}_0)$. This approximation is improved by including higher-order terms.

The mean square of the estimation error vector, $\delta\vec{\theta}_0^{\text{MAP}} = \vec{\theta}_0^{\text{MAP}} - \vec{\theta}$, is

$$
\varepsilon^{\text{MAP}} = \langle ||\delta\vec{\theta}_0^{\text{MAP}}||^2 \rangle_R = \text{Tr}(U K U^T + (I - U) C (I - U)^T + (I - U) \Delta\vec{h} \Delta\vec{h}^T (I - U)^T)
$$

where $\Delta\vec{h} = \vec{h} - \vec{h}_0$. The $U^T$ is the transpose of $U$. The $\text{Tr}V$ is the trace of matrix $V$. The $\langle \cdot \cdot \rangle_R$ is the average with respect to $P(\vec{R})$. The first term of the estimation error in equation 3.4 is due to the channel noise in $\vec{R}$. The second term is due to the variability of $\vec{\theta}$. The last term is due to error in estimating the mean of $\vec{\theta}$. 

Sometimes it is important to calculate the error of $\tilde{\theta}_0$ in a direction $\tilde{a}$, that is, the error projected to a normalized vector, $\tilde{a}$:

$$
\langle (\tilde{a}^T \delta \tilde{\theta}_0^{\text{MAP}})^2 \rangle_R = \tilde{a}^T (UKU^T + (I - U)C(I - U)^T + (I - U)\Delta \tilde{h}\Delta \tilde{h}^T (I - U)^T) \tilde{a}.
$$

(3.5)

Another method of estimation is SBE. An estimator of $\tilde{\theta}$ in SBE, $\tilde{\theta}_0^{\text{SBE}}$, is a randomly generated subject to the posterior distribution of $\theta$ given $\vec{R}$, $P_0(\tilde{\theta}|\vec{R})$. It is easy to show that $P_0(\tilde{\theta}|\vec{R})$ is a gaussian distribution:

$$
P_0(\tilde{\theta}|\vec{R}) = \frac{1}{\sqrt{(2\pi)^N |A|}} \exp \left( -\frac{1}{2} (\tilde{\theta} - \tilde{\theta}_0^{\text{MAP}})^T A^{-1} (\tilde{\theta} - \tilde{\theta}_0^{\text{MAP}}) \right),
$$

(3.6)

where $A = (C_0^{-1} + K^{-1})^{-1}$. The mean of SBE of $\tilde{\theta}_0^{\text{SBE}}$ is $\tilde{\theta}_0^{\text{MAP}}$. The $\tilde{\theta}_0^{\text{SBE}}$ is $\tilde{\theta}_0^{\text{MAP}}$ plus some gaussian noise.

The total and projected errors of SBE are

$$
\varepsilon^{\text{SBE}} = \langle |\delta \tilde{\theta}_0^{\text{SBE}}|^2 \rangle_R = \varepsilon^{\text{MAP}} + \text{Tr}(A),
$$

(3.7)

$$
\langle (\tilde{a}^T \delta \tilde{\theta}_0^{\text{SBE}})^2 \rangle_R = \langle (\tilde{a}^T \delta \tilde{\theta}_0^{\text{MAP}})^2 \rangle_R + \tilde{a}^T A\tilde{a}.
$$

(3.8)

Finally, SBE2 is the last method of estimation considered in this letter:

$$
\theta_0^{\text{SBE2}} = \theta_0^{\text{MAP}} + \vec{\alpha},
$$

(3.9)

where $\vec{\alpha}$ is a gaussian random noise vector whose correlation matrix is $\hat{\sigma}^2 I$. The $I$ is an identity matrix. The $\hat{\sigma}^2$ is the variance of $\alpha_i$. SBE is MAP estimation with additional gaussian fluctuations. SBE2 is the same as SBE in this sense; however, the covariance of $\theta_0^{\text{SBE2}}$ for a given $\vec{R}$ is independent of $C_0$.

The total and projected errors of SBE2 are

$$
\varepsilon^{\text{SBE2}} = \langle |\delta \tilde{\theta}_0^{\text{SBE2}}|^2 \rangle_R = \varepsilon^{\text{MAP}} + N\hat{\sigma}^2,
$$

(3.10)

$$
\langle (\tilde{a}^T \delta \tilde{\theta}_0^{\text{SBE2}})^2 \rangle_R = \langle (\tilde{a}^T \delta \tilde{\theta}_0^{\text{MAP}})^2 \rangle_R + \hat{\sigma}^2.
$$

(3.11)

When $K$ and $C_0$ commute with each other such that $KC_0 = C_0K$, $K$ and $C_0$ can be diagonalized by the same set of eigenvectors. The meaning of the $U$ matrix for MAP estimation in equation 3.3 is easier to understand in this

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/NECO_a_00371 by guest on 14 May 2021
where \( \vec{u}_i \) is the \( i \)th eigenvector of \( C_0 \) with eigenvalue \( \lambda_{0i} \). The \( \sigma_i^2 \) is the \( i \)th eigenvalue of \( K \). The \( \vec{u}_{0i}\vec{u}_{0i}^T \) is a projection operator to the direction of \( \vec{u}_{0i} \). For \( \lambda_{0i} \gg \sigma_i^2 \), the coefficient of the projection operator, \( \vec{u}_{0i}\vec{u}_{0i}^T \), in equation 3.12 is close to 1. For \( \lambda_{0i} \ll \sigma_i^2 \), the coefficient of the projection operator, \( \vec{u}_{0i}\vec{u}_{0i}^T \), is close to 0. Equations 3.2 and 3.12 indicate that \( U \) filters out the components of \( \vec{R} - \vec{h}_0 \) with eigenvalues smaller than \( \sigma_i^2 \) in MAP estimation. Such components may have originated from noise in the sensory channel.

The estimation error in equation 3.4 is

\[
\epsilon^{\text{MAP}} = \sum_{i=1}^{N} \left\{ \left( (\vec{u}_{0i}^T (\theta_0^{\text{MAP}} - \theta))^2 \right) \right\}_R
\]

\[
= \sum_{i=1}^{N} \sigma_i^2 \frac{\lambda_{0i}^2 + \sigma_i^2 \hat{\lambda}_i}{(\sigma_i^2 + \lambda_{0i})^2} + \left( \frac{\sigma_i^2}{\sigma_i^2 + \lambda_{0i}} \right)^2 (\vec{u}_i^T \Delta \vec{h})^2, \tag{3.13}
\]

\[
\hat{\lambda}_i = \sum_{j=1}^{N} \lambda_j (\vec{u}_{0i}^T \vec{u}_j)^2, \tag{3.14}
\]

where \( \vec{u}_i \) is the \( i \)th eigenvector of \( C \) with eigenvalue \( \lambda_i \). Since \( \sum_{j=1}^{N} (\vec{u}_{0i}^T \vec{u}_j)^2 = 1 \), \( \hat{\lambda}_i \) can be considered the weighted average of \( \lambda_i \) (\( i = 1, \ldots, N \)). The error due to incorrect estimation of \( \vec{h} \) is trivial and is set to 0 in this section.

The \( C_0 \) generally does not need to have the same eigenvectors with \( C \) to minimize MAP estimation error, \( \epsilon^{\text{MAP}} \). For example, if \( \lambda_j = \lambda \) such that all the eigenvalues of \( C \) are the same, \( \hat{\lambda}_i = \lambda \) for any set of eigenvectors and \( \epsilon^{\text{MAP}} \) is independent of \( \vec{u}_i \).

Each term in the sum of equation 3.13 is an estimation error in the feature vector, that is, \( \vec{\theta} \) is projected to the \( \vec{u}_{0i} \) direction. Equation 3.13 without the \( \Delta \vec{h} \) term can be written as

\[
\epsilon^{\text{MAP}} = \sum_{i=1}^{N} \sigma_i^2 f(x_i, y_i), \tag{3.15}
\]

where \( f(x, y) = (x^2 + y^)/(1 + x^2) \), \( x_i = \lambda_{0i}/\sigma_i^2 \), and \( y_i = \hat{\lambda}_i/\sigma_i^2 \) are normalized eigenvalues. The \( f(x, y) \) as a function of \( x \), has the unique minimum value of \( f_{\text{min}} = y/(1 + y) \) at \( x = y \) or \( \lambda_{0i} = \hat{\lambda}_i \). Since \( \sigma_i^2 f_{\text{min}} = \sigma_i^2 \lambda_i/ \)
\((\sigma_i^2 + \lambda_i) \approx \sigma_i^2\) for \(\lambda_i \gg \sigma_i^2\), the minimum estimation error for \(\bar{\theta}\) projected to an eigenvector with a large eigenvalue is determined by the corresponding eigenvalue of \(K\).

The \(\epsilon_{MAP}\) has a minimum for \(C_0 = C\). This is not true for SBE. The SBE error is

\[
\epsilon_{SBE} = \sum_{i=1}^{N} (\lambda_{0i} \sigma_i^2 (\theta_0 - \theta))^2_R = \sum_{i=1}^{N} \frac{\lambda_{0i} \sigma_i^2}{\lambda_{0i} + \sigma_i^2} + \epsilon_{MAP}.
\] (3.16)

Equation 3.16 can be rewritten as

\[
\epsilon_{SBE} = \sum_{i=1}^{N} \sigma_i^2 g(x_i, y_i),
\] (3.17)

where \(g(x, y) = x/(1 + x) + f(x, y)\). The \(g(x, y)\) does not have a minimum at \(x = y\). The \(g(x, y)\), as a function of \(x\), has the unique minimum value of \(g_{\text{min}} = (2y - 1/4)/(1 + y)\) at \(x = (2y - 1)/3\) as far as \(y > 0.5\). For \(y < 0.5\), \(g_{\text{min}} = y\) at \(x = 0\). The stochastic process to generate \(\theta_0\) produces additional estimation error, and it is better to use a narrower distribution of \(\theta_0\) than the true variability of \(\theta\) in \(P(\theta)\) to minimize SBE error.

It should be emphasized that a small number of eigenvalues of \(C\) can be very large even when the pairwise correlations are small. It is known that natural stimuli such as natural scenes have such properties (Simoncelli & Olshausen, 2001). However, each property of the stimulus, such as the orientation of a bar or contrast, is often considered separately. This may lead to substantial underestimates of the capacity of the sensory system.

For a concrete example, consider a case of a uniform correlation where \(C_{ij} = z(\delta_{ij} + q(1 - \delta_{ij}))\) and \(C_{0ij} = z(\delta_{ij} + q_0(1 - \delta_{ij}))\) and \(C_0\) and \(C\) are diagonalized by the same set of orthogonal eigenvectors. The \(\lambda_1 = z(1 + (N - 1)q)\) and \(\lambda_2 = z(1 - q)\) with \(N - 1\) degeneracy. The \(\lambda_{01} = z(1 + (N - 1)q_0)\) and \(\lambda_{02} = z(1 - q_0)\) with \(N - 1\) degeneracy.

Note that a small error in estimating the correlations can be greatly amplified because \(\lambda_1\) and \(\lambda_{01}\) can be very large for a large \(N\). To measure the error caused by assuming an incorrect value of \(q_0\), we define the ratio of estimation error to the minimum value, \(T_R = f(x, y)/f_{\text{min}}:\)

\[
T_R = \frac{f(x, y)}{f_{\text{min}}} = 1 + \frac{(x - y)^2}{y(1 + x)^2}.
\] (3.18)

Figure 1A shows \(T_R\) as a function of \(\delta q_0 = q_0 - q\) for \(q = 0.1, 0.5, \) and 0.8 for \(\lambda_1\). The \(T_R\) for \(\delta q_0 = 0\) is 1. We found that underestimating \(q\) made the estimator inaccurate and that it was better to assume that all environmental features were strongly correlated to achieve a small estimation error. For
Figure 1: Plots of $T_R$ for $N = 100$. $\sigma_i^2 = z = z_0 = 1$. (A) $T_R$ for $\lambda_1$ as a function of $q_0 - q$ for $q = 0.1, 0.5,$ and $0.8$. (B) $q_0 - q$ is within the interval of $-q \leq q_0 - q < 1 - q$. $T_R$ for $\lambda_2$ as a function of $q_0 - q$ for $q = 0.1, 0.5,$ and $0.8$. (C) $T_R$ for $\lambda_1$ (dashed line) and $\lambda_2$ (solid line) as a function of $q$. $q_0 = 0$. (D) $T_R$ for $\lambda_1$ (dashed line) and $\lambda_2$ (solid line) as a function of $q$. $q_0 = 1$.

example, $T_R$ for $q_0 = 0$ can be very large (see the dashed line in Figure 1C). For $q = 1$ and $q_0 = 0$, $T_R \cong N/4 + 1/2$. However, the overestimation of $q$ does not increase the estimation error much because the $T_R$ for $\lambda_1$ is a flat function of $\delta q_0 > 0$ (see Figure 1A). The $T_R$ is not very different from 1 even when $q_0 = 1$ (see the dashed line in Figure 1D). For $q = 0$ and $q_0 = 1$, $T_R \cong 2 - 4/N$ and $T_R$ decreases to 1 rapidly as $q$ increases. Equation 3.18 is not symmetric with respect to $x$ and $y$. The $T_R$ for $q = 1$ and $q_0 = 0$ is very different from $T_R$ for $q = 0$ and $q_0 = 1$. The $T_R$ for $\lambda_2$ has a more symmetric shape as a function of $\delta q_0$ (see Figure 1B). For $q = 1$ and $q_0 \cong 0$, $T_R \cong 1/4q_0$ (see Figure 1C). For $q = 0$ and $q_0 = 1$, $T_R = 2$ (see Figure 1D).

The plots of $T_{RSE}(x, y) = g(x, y)/g_{\min}$ in the same cases are qualitatively similar to those in Figure 1.

4 Learning About the Environment

The previous section introduced several methods for estimation and explained how the estimation of given environmental features ($\vec{\theta}$) depended on the correlation structure of the real world (C) and our assumption about
it \( C_0 \). A crucial question is how the experience of the environment can be used to learn its statistical characteristics.

Here, we consider a learning strategy where prior distribution is updated according to the past observations. Since gaussian distribution is determined by mean and correlation, they are updated with the observed values of environmental features. \( C_0 \) is updated to \( C_0^{\text{new}} \) whenever the sensory signal, \( \vec{R} \), is decoded to produce \( \vec{\theta}_0 \) in the following way:

\[
C_0^{\text{new}} = \eta (\vec{\theta}_0 - h_0)(\vec{\theta}_0 - h_0)^T + (1 - \eta)C_0
\]
\[
h_0^{\text{new}} = \eta \vec{\theta}_0 + (1 - \eta)h_0
\]

where \( \eta \) is a small mixing parameter.

If the update process of prior distribution converges to a stationary distribution, we call it self-consistent because the distribution of observed \( \vec{\theta}_0 \) is the same as the distribution of \( \vec{\theta} \) assumed in the estimation process. The cumulative update of prior distribution is just one of many other possible learning rules (see section 4.4).

We assume that \( \eta \) is very small for stable updates. For small \( \eta \), equations 4.1 and 4.2 can be written as a differential equation by taking the average over \( P(\vec{R}) \),

\[
\frac{d}{dt}C_0 = \frac{1}{\eta} \langle C_0^{\text{new}} - C_0 \rangle_R = U_0 (K + C + (h - h_0)(h - h_0)^T)U_0^T - C_0,
\]
\[
\frac{d}{dt}h_0 = \frac{1}{\eta} \langle h_0^{\text{new}} - h_0 \rangle_R = U_0 (h - h_0).
\]

The true value of \( \vec{\theta} \) is not available for accumulating experience. What is available is its accurate estimator, such as \( \vec{\theta}_0^{\text{MAP}} \). This update rule means that when we perceive the environment as \( \vec{\theta}_0^{\text{MAP}} \), \( (C_0)_{ij} \) is updated by using \( (\vec{\theta}_0^{\text{MAP}} - \vec{h}_0)(\vec{\theta}_0^{\text{MAP}} - \vec{h}_0) \). The decoder assumes that the mean of \( \theta_i \) is \( h_{0i} \) and that the correlation between \( \theta_i \) and \( \theta_j \) is \( C_{0ij} \) to produce \( \vec{\theta}_0^{\text{MAP}} \). The self-consistency of learning requires the mean of \( \theta_i^{\text{MAP}} \) to be \( h_{0i} \) and the correlation between \( \theta_i^{\text{MAP}} \) and \( \theta_j^{\text{MAP}} \) to be \( C_{0ij} \).

Consider a simple case of \( N = 1 \). When \( h_0 = h \), equation 4.3 can be written in terms of \( x = \lambda_0 \sigma^2 \) and \( y = \lambda \sigma^2 \), where \( C_0 = \lambda_0 \), \( K = \sigma^2 \) and \( C = \lambda \):

\[
\frac{dx}{dt} = \frac{x^2(1 + y)}{(x + 1)^2} - x.
\]

Figure 2 shows the plots of equation 4.5 for \( y = 2, 3, \) and \( 4 \). The stationary equation, \( dx/dt = 0 \), has only one stable real solution, \( x = 0 \) for \( y < 3 \). Note that \( x = 0 \) corresponds to a sharply peaked distribution
Figure 2: Plots of $dx/dt$ in equation 4.5 for $y = 2$, 3, and 4.

of $\theta$ like the delta function. The estimation of features of the environment is a constant in this case because the variance is 0. When $y = \lambda/\sigma^2 > 3$, the results of learning depend on the initial condition. There are three stationary solutions: $x = 0$, $x^C$, and $x^*$ for $y > 3$ (see Figure 2):

$$x^C = \frac{1}{2} \left( y - 1 - \sqrt{y^2 - 2y - 3} \right), \tag{4.6}$$

$$x^* = \frac{1}{2} \left( y - 1 + \sqrt{y^2 - 2y - 3} \right), \tag{4.7}$$

where $x = 0$ and $x = x^*$ are stable solutions and $x = x^C$ is unstable. When the initial value of $x$ is smaller than $x^C$, $x$ will evolve to $x = 0$. When the initial value of $x$ is larger than $x^C$, $x$ will evolve to $x = x^*$. $x^* = y - 1 - 7/4y + O(1/y^2)$ and $x^C = 7/4y + O(1/y^2)$ for large $y$.

There are hysteresis phenomena in self-consistent learning because $x = 0$ is always a stable solution to the stationary condition equation. Suppose that $y$ is slowly changing over time (see Figure 3). This may be because the statistical characteristics of the environment are changing or the sensory channel noise level is changing. Initially, $y$ is above the critical value of 3 and $x$ converges to a nonzero value. In the case in Figure 3, $y$ is 5 at $t = 0$ and the nonzero stable solution to equation 4.5 is $x = 3.73$. As $y$ decreases slowly, $x$ decreases slowly as far as $y > 3$. At $y = 3$, a sudden change appears in $x$. The nonzero solution disappears, and $x$ suddenly jumps from 1 to 0 as $y$ becomes smaller than 3. Later, $y$ may increase and become larger than 3 and then become the initial value of 5. However, the stationary $x$ does not
recover to the previous nonzero value because $x = 0$ is a stable solution to any value of $y$.

Extending the result for the one-dimensional case to the multidimensional case is straightforward when $C$ commutes with $K$ such that they can be diagonalized with the same set of eigenvectors. Here, $\vec{h}_0 = \vec{h}$ and $C_0 = C_0^*$ commuting with $C$ is the solution to equations 4.3 and 4.4. Then each eigenvector direction can be considered separately, as we did for $N = 1$. For the $i$th eigenvector, $y_i = \lambda_i / \sigma_i^2$ has a critical value of 3, determining the results of learning. The $\lambda_i$ and $\sigma_i^2$ correspond to the $i$th eigenvalue of $C$ and $K$. When $y_i < 3$, the component of $\vec{R} - \vec{h}$ in that direction will be ignored, and the estimation of features of the environment in that direction will be a constant. The stationary results of learning depend on the initial conditions, as they did for $N = 1$. When $C$ and $K$ did not commute, we did numerical studies (we discuss these in section 5.6).

The method of self-consistent learning may overestimate the correlation between features of the environment. Let us consider the case of $C_{ij} = z(\delta_{ij} + q(1 - \delta_{ij}))$ again as a concrete example. $\lambda_1 = z(1 + (N - 1)q)$ and $\lambda_2 = z(1 - q)$ with $N - 1$ degeneracy. $K = \sigma^2 I$. If $\lambda_2 < 3\sigma^2 < \lambda_1$, $\lambda_{02} = 0$, and $\lambda_{01} = \lambda_{01}^* \neq 0$. Note that if $q_0 = 1$ and $z_0 = \lambda_{01}^* / N$, the eigenvalues of $C_{0ij} = z_0(\delta_{ij} + q_0(1 - \delta_{ij}))$ are 0 and $\lambda_{01}^*$. Therefore, self-consistent learning makes $q_0$ converge to 1 in this case. The decoder believes that the correlation between features is strong such that the environmental features are effectively one-dimensional.

Figure 4 shows examples of self-consistent learning for various noise levels. The three columns in Figure 4 have examples of the evolution of estimation error; squared errors of $C_0$ and $\vec{h}_0$ for $\sigma^2 = 1, 3$, and 4; and...
Figure 4: Change in estimation error due to accumulated experience. Each column corresponds to the results of simulation for $\sigma^2 = 1$, 3, and 4. $N = 10$. One of the eigenvalues of $C$ is 110. The others are 10.

$N = 10$. Let us consider the case of uniform correlation with $\lambda_1 = 110$ and $\lambda_2 = 10$. For $\lambda_2 = 10$, the critical value of $\sigma^2$ is $10/3$. The estimation error monotonically decreases in the first column of Figure 4 ($\sigma^2 = 1$). However, the second column of the figure shows a nonmonotonic change in estimation error. The difference between $C_0$ and $C$ increases compared with $\sigma^2 = 1$ while $\vec{h}_0$ still converges to $\vec{h}$. The eigenvalues of $C_0$ converge to some nonzero values because $\sigma^2 = 3 < 10/3$. However, the estimation error worsens after $\vec{h}_0$ converges to the correct mean $\vec{h}$. The third column reveals an estimation error diverging to a large value. The channel noise is larger than the critical value for $\lambda_2 = 10$ ($\sigma^2 = 4 > 10/3$), and the $N - 1$ eigenvalues of $C_0$ converge to 0. We add an identity matrix with a very small coefficient to $C_0$ to make $C_0$ an invertible matrix.

Self-consistent learning with MAP estimation is a dimension-reduction process, and $C_0$ can be very different from $C$ depending on channel noise. We considered two other methods of estimation to see how the results of self-consistent learning depended on the method of estimation. Consider self-consistent learning with SBE. The equation for the change of variance in SBE for $N = 1$ is

$$
\frac{dx}{dt} = \frac{x}{1 + x} + \frac{x^2(1 + y)}{(x + 1)^2} - x. \quad (4.8)
$$
Self-consistent learning with SBE makes $x$ converge to $y$ for any values of $y$. $dx/dt = 0$ has $x = y$ as a unique stable solution and $x = 0$ as an unstable solution for any $y$ (see Figure 5). In fact, it can be analytically demonstrated that the self-consistent condition is satisfied in the SBE case for any $P(\theta)$, not just for the gaussian case.

The self-consistency equation for SBE2 is

$$\frac{dx}{dt} = \gamma + x^2(1+y)(x+1)^2 - x,$$

where $\gamma$ is $\hat{\sigma}^2/\sigma^2$. This is the same as equation 4.5 except for constant $\gamma$. For $\gamma \geq 1/8$, $dx/dt = 0$ has a unique nonzero stable solution, and there is no hysteresis behavior.

The MAP estimation error is minimized for $C_0 = C$, but self-consistent learning produces $C_0 \neq C$. SBE error is not minimized for $C_0 = C$, but self-consistent learning produces $C_0 = C$. This suggests that another stochastic learning method between these two such as SBE2 can be advantageous. The error in SBE2 is minimized for $C_0 = C$ because $\epsilon^{SBE2} = \epsilon^{MAP} + \gamma\sigma^2$. Figure 6 plots the estimation errors for three different methods. For MAP estimation, the estimation error is smaller than the others as far as $y > 3$. However, for $y < 3$, estimation error is very large because $P_0(\theta)$ is highly peaked. For SBE,
Figure 6: Estimation errors for MAP estimation, SBE, and SBE2 with a self-consistent solution. The dashed-dotted line is for MAP estimation. The dashed line is for SBE. The dotted line is for SBE2 with $\gamma = 1/4$. The thick solid line is for MAP estimation with correct model of environment.

$C_0$ converges to $C$, but SBE has additional variability in estimation, and the error is larger than MAP estimation. For SBE2, $C_0$ does not converge to $C$, but there is no transition to $x = 0$ for $\gamma > 1/8$. SBE2 performs well in terms of errors over a wide range of $y$.

5 Discussion

5.1 Perceptual Error and Prior Distribution. The accuracy of the estimation or the perception of the surrounding environment depends on the distribution of the environmental feature $P(\theta)$. For example, it is important to know the structure of correlation between different environmental features to perceive the environment accurately because some combinations of features have much higher probabilities of being observed than others. Since the true distribution is not always available, our perception is often biased by the wrong prior. We studied how the perceptual error depends on the prior distribution assumed in the estimation process and the true distribution of environmental feature in this letter.

Estimation error can be greatly overestimated when each feature is considered separately. Even when the pairwise correlation is relatively small, ignoring the correlation can cause a large estimation error because the correlation matrix can have very large eigenvalues in a high-dimensional case and the variability in the environment may be well described by a few principal components with the largest eigenvalues (Osowski, Majkowski, &
Cichocki, 1997). Our study shows that finding such components of sensory signal is the essence of efficient estimation of environmental features.

The optimal prior distribution to minimize the perception error is not always the true distribution of the environmental features. When the true correlation matrix of environmental features is \( C \) and \( C_0 \) is assumed for decoding, the estimation error has a minimum value for \( C_0 = C_0^* \neq C \) for SBE because of the additional variability generated by the process of estimation. For MAP estimation and SBE2, the estimation error takes a minimum value when \( C_0 = C \). The estimation error as a function of \( C_0 \) has degeneracy such that the eigenvector and the eigenvalues do not need to be the same as \( C \) to minimize estimation error.

5.2 Fundamental Ambiguity in the Perception Process. There is a fundamental ambiguity in the perception process. When information about the outside world arrives only through sensory channels with finite reliability, a different choice of \( P(\vec{R} | \vec{θ}) \) and \( P(\vec{θ}) \) provides another valid explanation of sensory signal distribution, \( P(\vec{R}) \), as long as \( P(\vec{R}) = \int d\vec{θ} P(\vec{R} | \vec{θ}) P(\vec{θ}) \) is the same. It means that the outside world appears as a subjective model with an internal statistical structure to explain sensory data and that the \( P(\vec{R} | \vec{θ}) \) should be considered as the degree of subjective confidence in the sensory channels.

Studying each hierarchical level of information processing in the brain separately could be misleading in this sense because each level is the “environment” of the other levels. For example, studying the V1 and LGN of many different experimental subjects and creating an average picture of each component of visual-information processing could be wrong. The “average” V1 and the “average” LGN may not fit each other well, because two different people may develop different models to understand the outside world.

5.3 Dynamics of Prior Distribution. Prior distribution is changed by experience until it converges to a stationary distribution. The exact equations for the dynamics of prior distribution remain to be studied. Studying plausible principles underlying such learning processes is important and interesting. For example, Friston (2010) suggests that the minimization of surprise is the principle underlying perception and action. Other suggestions, such as Bayesian brain hypotheses, could be unified under such principles. Because the theory of evolution provides explanations about the functions and structures of various living forms, we may understand why the brain or any other biological system behaves in the observed way by thinking of perception as a result of a process to learn the statistical features of the surrounding environment.

We consider a simple cumulative update of prior distribution in this letter and studied how perception is determined by such dynamics. Since the prior distribution is updated with the current estimation of environmental
features in this case, the distribution of the estimated environmental feature becomes the same as the prior distribution used to generate the estimated value when the prior distribution converges to a stationary value. Therefore, we call the cumulative update self-consistency learning.

Self-consistent learning depends on the estimation method. We consider three decoding methods in this letter: MAP estimation, SBE, and SBE2. We found that a cumulative update of the internal model of the outside world for MAP estimation did not generate convergence to the true distribution. Self-consistent learning is a process to reduce dimensions in this case. Depending on the initial condition and noise level, the variability of environmental features to the directions of eigenvectors with small eigenvalues is ignored, and the environmental features are perceived as constants to that direction. Reducing dimensions has an advantage in terms of computational cost. $N$ could theoretically be the total number of sensory neurons in the brain. Tracking the state of the environment is not possible without reducing dimensions. However, it does not always produce wanted results. Hysteresis may make the decoder insensitive to important directions (see Figure 3) and may also produce overtraining (see Figure 4).

The self-consistent update of the internal model for SBE generates convergence to the true model. However, self-consistent learning with SBE may require more experience because the larger decoding error with SBE needs to be averaged out in the learning process. When the statistical structure of the environment is changing quickly, this could be disadvantageous. We also found that SBE error is not minimized when $C_0 = C$.

Learning with another method of estimation, SBE2, could produce $C_0$ closer to $C$ than MAP estimation. It does not exhibit hysteresis or reduce dimensions. The error with SBE2 is smaller than with SBE with a reasonable choice of parameters (see Figure 6).

5.4 Learning of Distribution in the Environment. We do not consider the cumulative update of prior distribution as the only possible learning rule of prior distribution. Cumulative update may be more biologically plausible, but batch learning, where all the available previous observations are used together, is also possible. We could calculate the marginal distribution of $\vec{R}$ by integrating over all possible environmental features and maximize $P(\vec{R})$ in terms of $C_0$.

We can think of a case where the change of the environmental state is modeled by a Markov process and consider the update rule of Kalman filter approach. Knill (2007) studied a learning rule without self-consistency. In general, a different learning rule has different advantages and disadvantages. A method that shows good performance in engineering applications may not be possible in the brain because of biological constraints. The issue of the true learning rule in the brain should be further studied experimentally.
The noise in the sensory information channel and the uncertainty in the environment are inseparably mixed in the perception process, and they cannot be separated without additional assumptions. However, the brain may have additional information or assumptions about noise sources. For example, if the sensory channel noise distribution or prior distribution of environmental features has known properties, this could be the case. It is also possible to assume that that one of the sensory signals provides information without noise. Adams et al. (2004) show that haptic information can change the light-from-above prior in visual information processing because any ambiguity is resolved by the haptic signal. Finally, in some cases, it may be assumed that we are in the same environment as before. We can estimate the distribution of noise in the sensory signal with multiple experiences of the same environment. However, such additional sources of information may not exist when complex and fast-changing environments are perceived in a short time period. We consider in this letter only the cases where such additional information is not available.

5.5 Neural Representation of Uncertainty in the Cortex. It is not yet clear how uncertainty or probability in the environment is represented in the cortex. Previous theoretical proposals to answer this question can be divided into two categories. Neural activities represent parameters of the probability distribution in one scheme and sensory variables themselves in the other (Fiser, Berkes, Orban, & Lengyel, 2010). It is important to remember that we need many features to describe the environment completely, and these features are correlated with each other. The number of neurons required for the first scheme may increase too fast as the dimension of the feature vector increases (Fiser et al., 2010). The other category assumes that probability distribution is estimated based on the multiple sampling of the feature variables. However, the number of samplings should be very large for correlated multidimensional feature for accurately estimating the probability distribution, and the response time should be very long. Berkes, Orban, Lengyel, and Fiser (2011) assume that the spontaneous neural activity represents the prior distribution in the environment and showed that the averaged stimulus-evoked neural activities are similar to the spontaneous one. This suggests that spontaneous activity plays an important role in information processing in the brain and that the processing is being done according to a Bayesian scheme. It also suggests that a self-consistent prior distribution is being used in adult ferrets. If the neural activity represented $\vec{\theta}_0$, the sample average of the neural activity would be the self-consistent distribution studied in this letter. If the neural activity represented $\vec{R}$, the estimated correlation matrix of prior distribution would be $C+K$.

5.6 Extension to Other Cases. We considered gaussian distributions where correlation matrices commute and can be diagonalized together.
There are many possible ways of extending our results to other cases. Here, we briefly comment on the three extensions of our study.

First, we considered a nongaussian case of \( P(\theta) \propto \exp(-\alpha \theta^4) \) and found a critical level of sensory channel noise. This indicated that such a transition was not a property of the gaussian distribution. In fact, there is a simple qualitative explanation to why the \( x = 0 \) point or delta function like \( P(\theta) \) is a self-consistent fixed point. When \( P(\theta) \) is highly peaked, the channel noise distribution is not important in the MAP estimation because the peak of \( P(\theta) \) determines the maximum point of \( P(\theta|R) \). Self-consistent learning makes the sharply peaked \( P(\theta) \) even more so. For low levels of sensory channel noise, there may be other self-consistent fixed points, and the results of learning depend on initial conditions.

Second, we considered the case of randomly generated \( K \) and \( C \) where they do not commute. We generated \( K_{ij} \) and \( C_{ij} (i \leq j) \) randomly with a gaussian random number generator and added an identity matrix with a multiplicative constant to the randomly generated matrices such that the resultant matrices were positive definite. For given \( K \) and \( C \), equations 4.3 and 4.4 were numerically integrated to find the stationary values of \( C_0 \) and \( h_0 \). In most of the cases we tried, many eigenvalues of \( C_0 \) converged to 0, and \( C_0 \) converged to different matrices depending on the initial value. We define a zero space of a matrix as a space spanned by the eigenvectors of the matrix whose eigenvalue is zero. We can demonstrate that a matrix with a given zero space is a fixed point of the self-consistent learning equation of \( C_0 \) in equation 4.3. We could not prove the stability of these fixed points analytically, but numerical results revealed that if \( C_0 \vec{a} = 0 \) initially, it remained so after self-consistent learning. Starting from many random initial values, \( C_0 \) converged to many different matrices with overlapping but not identical zero spaces.

Finally, we can also consider the case where our confidence in the sensory information channel is updated together. In this case, we update the channel noise \( K \) as well as \( C_0 \). Our study shows that when the estimation rule is MAP, there are two stationary points of the learning dynamics and either \( K \) or \( C_0 \) is zero. This means we have come to believe that the variability of a sensory signal stems from either sensory channel noise or the environment itself. We will report this work elsewhere.

5.7 Role of Noise in the Brain. Our study provides a new point of view about noise in the brain. First, noise in neural activities may play an important role in reducing the dimensions of signals. Considering the computational cost, it is crucial to reduce dimensions in information processing in the brain. The hysteresis we have discussed in this letter may be actively used in the brain by controlling the noise level. When one part of the brain is receiving input from many other parts of the brain, the most important features of the signals can be obtained with the hysteresis effect. With a slowly oscillating noise level, the self-consistent learning rule may
suppress less important components of the neural signal due to the hysteresis effect.

Second, neurons in the brain and many parts of the brain are “environments” of other neurons and other parts of the brain. If self-consistent learning with MAP estimation is what is happening in the brain, the parts of “the environment” with small variability should be neglected. Neurons with predictable behavior will be neglected by self-consistent learning. Once categorized as being neglected, it is difficult for them to escape the position. All parts of the brain need to be unpredictable or noisy for the brain to be ready to respond quickly.

5.8 Observation of Hysteresis. Our study suggests that hysteresis may exist (see Figure 3). The basic mechanism of hysteresis is the feedback between perception and prior distribution. Perception is made with the prior distribution, and the prior distribution is updated with the perception. If there are multiple stationary point in this process, hysteresis can be observed.

There is hysteresis in high cognitive levels, and we all experience it in our daily lives. Our prejudice against somebody often strengthens itself. Our political and religious opinions are made with such feedback. It is not clear yet, however, whether this happens at a low cognitive level. For example, Knill (2007) showed that the change of prior distribution can change depth perception. It is not clear whether the changed perception enhanced the change of prior distribution. A light-from-above prior may be an example of stationary self-consistent point (Adams et al., 2004).

Probably the length of time during which we are trapped in one self-consistent point is actually limited in reality. We may be able to study this in many different perceptual tasks such as the recognition of visual pattern or coherent movement. For visual pattern recognition, we can blur visual patterns to different degrees and test whether our subject can recognize a pattern in the blurred image. The performance of the subject may strongly depend on whether he or she was trained with blurred visual patterns or very clear patterns before the test. In the case of coherent movement, we can show many dots moving in the same direction to our subject and test whether the subject can recognize the direction of coherent movement. The task can be made more difficult by adding increasing numbers of incoherently moving dots. Again, the performance of the subject for a given level of coherence may strongly depend on the coherence level with which the subject was trained before the test.

Acknowledgments

We thank Taro Toyoizumi for the helpful discussions with him and his suggestions for the example of the non-Gaussian case.
References


Received January 30, 2011; accepted July 1, 2012.