Neural Adaptation of the Kalman-Gain

Gábor Szirtes*, Barnabás Póczos† and András Lőrincz‡

Department of Information Systems

Eötvös University, Budapest, Hungary

Abstract. *Anticipating* future events seems to be a crucial function of the central nervous system and may be realized by Kalman-filter like mechanisms which are optimal for predicting *linear* dynamical systems. However, a *connectionist* representation of such mechanisms using local Hebbian learning rules has not been derived yet. We show that the recursive prediction error method offers a connectionist form for the dynamic adaptation of the Kalman-gain. We also provide a biologically plausible mapping of the derived architecture onto the hippocampal-entorhinal loop. Our mapping suggests testable predictions.

Keywords: Kalman-filter, hippocampus, entorhinal cortex, neural prediction

* gszirtes@inf.elte.hu
† barn@ludens.elte.hu
‡ corresponding author: András Lőrincz
Email: lorincz@inf.elte.hu
Address: Eötvös Loránd University, Pázmány Péter sétány 1/C
Budapest, Hungary, H-1117
Phone: +36 1 209 0555 ext. 8437, Fax: +36 1 381 2140

1. Introduction

Linear dynamical systems (LDS) are well studied and widely applied tools in state estimation and control tasks. The so-called Kalman-filter (KF) recursion makes the inference in LDS simple, the resulting estimations are unbiased and have minimized covariance. Because of the great success in different engineering applications, Kalman-filters have been proposed as the underlying mechanism in neurobiological modelling, too. Kalman-filters (i) may support sensory processing (Rao and Ballard, 1997), (ii) may complement the memory organization in the hippocampus (Bousquet et al., 1999), or, (iii) directly shape control architectures (Todorov and Jordan, 2002). In this contribution we provide an approximation of the Kalman-filter. The approximation is based on the recursive prediction method (Ljung and Soderstrom, 1993) (Section 2), which (i) can be realized in neuronal form, (ii) is able to preserve the efficacy of the original Kalman-filter method and (iii), can be mapped onto the loop of the entorhinal cortex and the hippocampus (EC-HC) (Section 3). Such functional mapping is of great importance, as the EC-HC loop is supposed to have a central role in memory encoding, storage and retrieval (Scoville and Milner, 1957; Eichenbaum, 2001). Section 4 lists open questions as well as the predictions of the proposed model.

2. Kalman-filter and the recursive prediction error (RPE) method

Consider the following linear dynamical system (LDS):

$$\mathbf{y}^t = \mathbf{H}\mathbf{x}^t + \mathbf{n}^t \quad \text{observation process} \tag{1}$$

$$\mathbf{x}^{t+1} = \mathbf{F}\mathbf{x}^t + \mathbf{m}^t \quad \text{hidden process} \tag{2}$$

where variables $\mathbf{m}^t \propto \mathcal{N}(0, \Pi)$, $\mathbf{n}^t \propto \mathcal{N}(0, \Sigma)$ are independent Gaussian noise processes with expectation value 0 and covariance matrices Π and Σ, respectively. The task is to estimate the hidden variable $\mathbf{x}^t \in \mathbf{R}^n$ (that is the state of the observed system) given the series of observations $\mathbf{y}^\tau \in \mathbf{R}^p$, $\tau \leq t$. For Euclidean norm in the cost, the optimal solution was derived by Rudolf Kalman (Kalman, 1960; Bagchi, 1993): Let E and Cov denote the expectation value and the covariance matrix operators, respectively. Let us introduce the following notations: $\hat{\mathbf{x}}^{(t|\tau)} = E(\mathbf{x}^t \mid \mathbf{y}^1, \dots \mathbf{y}^\tau)$, $\mathbf{N}^t = Cov(\mathbf{x}^t \mid \mathbf{y}^1, \dots \mathbf{y}^t)$, and $\mathbf{M}^t = Cov(\mathbf{x}^t \mid \mathbf{y}^1, \dots \mathbf{y}^{t-1})$. Let us suppose $\hat{\mathbf{x}}^{(t-1|t-1)}$ and $\mathbf{N}^{t-1}$ have been already determined. Then

$$\hat{\mathbf{x}}^{(t|t-1)} = \mathbf{F}\hat{\mathbf{x}}^{(t-1|t-1)} \tag{3a}$$

$$\mathbf{M}^t = \mathbf{F}\mathbf{N}^{t-1}\mathbf{F}^T + \Pi \tag{3b}$$

$$\mathbf{K}^t = \mathbf{M}^t\mathbf{H}^T(\mathbf{H}\mathbf{M}^t\mathbf{H}^T + \Sigma)^{-1} \tag{3c}$$

$$\hat{\mathbf{x}}^{(t|t)} = \hat{\mathbf{x}}^{(t|t-1)} + \mathbf{K}^t(\mathbf{y}^t - \mathbf{H}\mathbf{F}\hat{\mathbf{x}}^{(t-1|t-1)}) \tag{3d}$$

$$\mathbf{N}^t \;=\; (\mathbf{I} - \mathbf{K}^t\mathbf{H})\mathbf{M}^t \tag{3e}$$

where $\mathbf{I}$ denotes the identity matrix and superscript T stands for transposition. $\mathbf{K}^t$ is an auxiliary matrix. The *prediction equation* estimates $\hat{\mathbf{x}}$ *before* the $(t+1)^{th}$ measurement:

$$\hat{\mathbf{x}}^{(t+1|t)} \;= \mathbf{F}\hat{\mathbf{x}}^{(t|t-1)} + \mathbf{K}^{p,t}(\mathbf{y}^t - \mathbf{H}\hat{\mathbf{x}}^{(t|t-1)}) = \mathbf{F}\hat{\mathbf{x}}^{(t|t)} \tag{4}$$

where $\mathbf{K}^{p,t} = \mathbf{F}\mathbf{K}^t$ is called the Kalman-gain, superscript p stands for 'prediction'. The difference in the second term of the right hand side can be identified as the *reconstruction error*: $\mathbf{e}^t = \mathbf{y}^t - \mathbf{H}\hat{\mathbf{x}}^{(t|t-1)} = \mathbf{y}^t - \hat{\mathbf{y}}^t$, where $\hat{\mathbf{y}}^t = \mathbf{H}\hat{\mathbf{x}}^{(t|t-1)}$ is the so called *reconstructed input* as it should match the input in squared norm. Elements of the reconstruction error are stochastic variables of $\mathcal{N}(0, \lambda_k)$.

In spite of its salient statistical and computational features, Kalman-filter based models have some serious drawbacks. The first problem is that the covariance matrices of measurement and observation noise (Π and Σ) are generally assumed to be known. The second problem is that parameters may also change in time. The third one is that to ensure dynamic adaptation of the control system through the Kalman-gain, the algorithm requires the calculation of a matrix inversion (Eq. 3c), which is hard to interpret in neurobiological terms. To the best of our knowledge, all previously proposed networks models proposing

Kalman-filters for a given brain function computed directly this matrix inversion.

In this section we present a new algorithm for approximating the Kalman-gain which eliminates the first and the third problems. The second problem of changing model parameters ($\mathbf{F}$, $\mathbf{H}$) is not addressed here (but see, (Lőrincz and Buzsáki, 2000)). Our proposal is to use an iterative estimation of the Kalman-gain which preserves the relevant features of the standard Kalman-filters, but without leaning on 'global' knowledge. In other words, tuning the elements in any transformation should require only local information and interaction. It is worth noting that on-line estimation is better suited to 'track' or predict the changing world as compared to estimations that use *all* past observations. In addition, it is known (Whittle, 1982) that under rather mild conditions, the Kalman-gain, $\mathbf{K}^{p,t}$, converges with geometrical speed to the optimal $\mathbf{K}^p$ in stationary environment, so on-line estimation of $\mathbf{K}^p$ is sufficient, fast and asymptotically optimal. In what follows, we derive an approximation of equations (3a-3e) by applying the RPE method (Ljung and Soderstrom, 1993). We also discuss the different constraints that the resulting construction may pose on a possible biological mapping.

2.1. KF APPROXIMATION

Let $\mathbf{K}^t\mathbf{z} \approx \theta^t . * \mathbf{K}\mathbf{z}$ with $\theta^t \in \mathbf{R}^p$ denote an arbitrary parametrization of $\mathbf{K}^t$, where $.*$ denotes element-wise multiplication. Other linear parametrization schemes can be applied as well. For instance, $\theta^t \in \mathbf{R}^1$ or $\theta^t \in \mathbf{R}^{p\times p}$ could also be used. The first case would correspond to a scalar scaling of the non-optimal gain (see, Póczos and Lőrincz (2003)), while the latter case would result in an element by element tuning of matrix $\mathbf{K}$. In the following, we provide the RPE approximation of the Kalman-filter and its further simplifications. A slight modification of the original parametrization shall also be presented.

2.1.1. 1^{st} *approximation*

The following equation approximates Eq. 4:

$$\hat{x}_i^{t+1} = \sum_j F_{ij}\hat{x}_j^t + \theta_i^t \sum_l K_{il} e_l^{t,\theta} \tag{5}$$

in which $\hat{\mathbf{x}}^{t+1}$ is a shorthand for $\hat{\mathbf{x}}^{(t+1|t)}$ and $\hat{x}_i^{t+1}$ denotes the i^{th} component of $\hat{\mathbf{x}}^{t+1}$. For simplicity, the notation of the explicit dependence of the error on θ has been dropped. Let us use the following notation for the transformed error signal: $\epsilon_k^t = \sum_l K_{kl} e_l^t$. Let us suppose that a suboptimal matrix $\mathbf{K}(\theta^0)$ is given at time $t = 0$. Our goal is to tune parameter θ^t in order to minimize $J_k(\theta_k) = \frac{1}{2}E[(\epsilon_k^t)^2]$ with respect to θ_k, where $E[.]$ is the expectation operator. This task can be considered

as a maximum likelihood estimation. By estimating the expectation value with sample averaging, stochastic gradient approximations can be applied. Taken the negative gradient of J_k with respect to θ_k^t, the k^{th} component of θ^t, the parameter upgrade is as follows:

$$\theta_k^{t+1} \;=\; \theta_k^t + \alpha^t \sum_{lj} K_{kl} H_{lj} W_{jk} \epsilon_k^t \tag{6}$$

where $W_{ik}^t = \frac{\partial \hat{x}_i^t}{\partial \theta_k}$ is an auxiliary matrix and α^t denotes the possibly time-dependent learning rate. The iterative upgrade of the elements of $\mathbf{W}$ can be given by taking the derivative of both sides of Eq. 5:

$$W_{ik}^{t+1} \;=\; \sum_j F_{ij} W_{jk}^t \xi_k - \theta_i^t \sum_{lj} K_{il} H_{lj} W_{jk}^t \xi_k + \delta_{ik} \epsilon_k^t \tag{7}$$

in which δ denotes Kronecker's delta symbol and an additional auxiliary vector, $\xi = (\xi_1, \ldots, \xi_n)^T$, has been introduced to provide an equation in conventional neuronal form. This vector can be regarded as to a sparse, internally generated *noise* and its role will be discussed later. The resulting model will be referred as 'O1' (first online KF model).

In order to simplify the complexity of the iteration scheme, it can be supposed that the system is near optimal in the sense that adaptation is as fast as possible and the outer world to be modelled changes smoothly. This assumption is fulfilled if $\mathbf{K} \approx \mathbf{H}^{-1}$, because in the absence of noise and for the case of equality, there would be no need to adapt the prediction-observation equilibrium. For biologically sound mapping, such reduction is most attractive because directed informa-

tion flow through three different connection matrices (the second term of the right hand side of Eq. 7) seems unfeasible. The simplified update equations of $\mathbf{W}$ and θ (model 'O2') are:

$$W_{ik}^{t+1} \approx \sum_j F_{ij} W_{jk}^t \xi_k - \theta_i^t W_{ik}^t \xi_k + \delta_{ik} \epsilon_k^t \tag{8}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{9}$$

Interestingly, only the diagonal elements of $\mathbf{W}$ appear in the tuning θ (Eq. 9). It implies that another simplification can be made by neglecting the off-diagonal elements of matrix $\mathbf{W}$:

$$W_{ii}^{t+1} \approx F_{ii} W_{ii}^t \xi_i - \theta_i^t W_{ii}^t \xi_i + \epsilon_i^t \tag{10}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{11}$$

This model will be referred as 'O3'.

An important consequence is that matrix $\mathbf{F}$, which is responsible for the temporal evolution of the hidden variable, is now barely involved in the tuning of $\mathbf{W}$ and only the diagonal elements of $\mathbf{F}$ are involved in this adaptation process. Numerical simulations revealed (see 2.2) that even further simplifications can be made by neglecting this self-excitatory contributions of the updates of the diagonal elements of matrix $\mathbf{W}$. The resulting approximations (Eqs. 12-13, model 'O4') and its *incremental form* (Eqs. 14-15, model 'O5'), which resembles the more conventional

forms used in the artificial neural network literature, are as follows:

$$W_{ii}^{t+1} \approx -\theta_i^t W_{ii}^t \xi_i + \epsilon_i^t \tag{12}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{13}$$

$$W_{ii}^{t+1} \approx W_{ii}^t + \gamma\{-\theta_i^t W_{ii}^t \xi_i + \epsilon_i^t\} \tag{14}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{15}$$

2.1.2. 2^{nd} *approximation*

The scheme also allows to introduce an additional matrix (**N**) that will change our graphical representation resulting in a more 'convenient' mapping (see later):

$$\hat{x}_m^{t+1} = \sum_j F_{mj} \hat{x}_j^t + \sum_i N_{mi}\, \theta_i^t \sum_l K_{il} e_l^t \tag{16}$$

The upgrade of parameter θ remains as in Eq. 6. The upgrade of matrix **W**, however, is different:

$$W_{mk}^{t+1} = \sum_j F_{mj} W_{jk}^t \xi_k + \sum_i N_{mi} \frac{\partial}{\partial_{\theta_k}} [\theta_i^t \sum_l K_{il} e_l^t] \tag{17}$$

Because,

$$\frac{\partial}{\partial_{\theta_k}} [\theta_i^t \sum_l K_{il} e_l^t] = \delta_{ik} \sum_l K_{kl} e_l^t + \theta_i \sum_l K_{il} \frac{\partial}{\partial_{\theta_k}} e_l^t \tag{18}$$

$$\frac{\partial}{\partial_{\theta_k}} e_l = \frac{\partial}{\partial_{\theta_k}} (y_l^t - \sum_j H_{lj} x_j^t) = -\sum_j H_{lj} W_{jk}^t \tag{19}$$

we have

$$W_{mk}^{t+1} = \sum_j F_{mj} W_{jk}^t \xi_k$$

$$+ \sum_i N_{mi}(-\theta_i^t \sum_{lj} K_{il} H_{lj} W_{jk}^t \xi_k + \delta_{ik}\epsilon_k^t) \tag{20}$$

Now, we can proceed with the approximations as before. First, we take $\mathbf{KH} \approx \mathbf{I}$. Then

$$W_{mk}^{t+1} = \sum_j F_{mj} W_{jk}^t \xi_k + \sum_i N_{mi}(-\theta_i^t W_{ik}^t \xi_k + \delta_{ik}\epsilon_k^t) \tag{21}$$

Neglecting the off-diagonal elements of $\mathbf{W}$, one has

$$W_{mm}^{t+1} \approx F_{mm} W_{mm}^t \xi_m + N_{mm}(-\theta_m^t W_{mm}^t \xi_m + \epsilon_m^t) \tag{22}$$

Dropping the recurrent feedback connections, i.e., setting $F_{mm} = 0$, the last approximation is:

$$W_{mm}^{t+1} \approx N_{mm}(-\theta_m^t W_{mm}^t \xi_m + \epsilon_m^t) \tag{23}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{24}$$

which, apart from constant factors (N_{mm}, $m = 1, \ldots, p$) are identical to Eqs. 12 and 13, respectively. The constant factors may modify the relaxation speed of the different components of θ^t. Naturally, if all diagonal elements of $\mathbf{N}$ are constrained to be one, then we get back to our previous approximation.

2.2. Simulations

Figure 1 depicts the simulation results for the first approximation scheme. The linear dynamical system has been made of two-dimensional

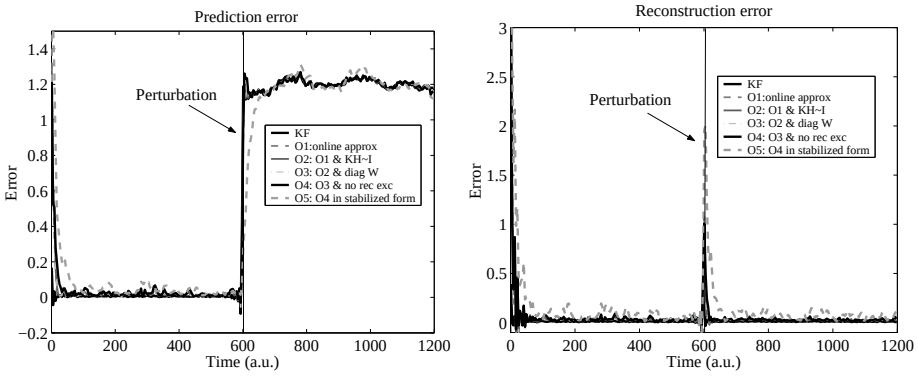

Figure 1. **Comparison of direct and approximated Kalman filters for a simple dynamical system. (First approximation scheme)** The 'prediction error' ($\mathbf{e_{pr}} = |\mathbf{x} - \hat{\mathbf{x}}|$) and the 'reconstruction error' ($\mathbf{e_{rec}} = |\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}|$) are plotted as a function of time. 'KF' stands for the optimal direct Kalman filter. 'O1' stands for the first approximation when no simplifying condition is assumed. 'O2' is the first simplification, when $\mathbf{KH} \approx \mathbf{I}$. 'O3' denoted the model in which off-diagonal elements of $\mathbf{W}$ are ignored. And finally 'O4' is the simplest form, in which recurrent self-excitation is also ignored. The stabilized form (delta rule) of the simplest form is denoted by 'O5'. 'Perturbation' denotes the time when model parameters($\mathbf{H}$) have been changed. For better visualization, data have been resampled at a lower rate after low-pass filtering. The negative error of the Kalman filter around the model change is an artefact of the Chebyshev filter design. The angles belonging to the rotational matrices: $\alpha_F = 10^\circ$, $\alpha_H = 50^\circ$ and after perturbation: $\alpha_{H_2} = 20^\circ$ Signal to noise ratio of the hidden process: $SNR = 59dB$. Learning rate is 0.01 for all simulations. Signal to noise ratio of the observation process: $SNR = 51dB$.

rotational matrices $(\mathbf{F}, \mathbf{H})$ and $\mathbf{K} \approx \mathbf{H}^{-1}$. It can be seen that the simplified models converge to the optimal solution. The figure also depicts the effect of changes of the basic conditions on the system; at a given time instant observation matrix $\mathbf{H}$ was modified, whereas the evolution of hidden variable $\mathbf{x}$ was not changed (matrix $\mathbf{F}$ was kept). Change of the observation matrix can in interpreted in different ways, such as the learning (or the tuning) of this matrix, or alternatively, a change in the environment. In our special case, the angle belonging to the rotational matrix was modified. The figure also demonstrates that

even for the case $\mathbf{KH} \neq \mathbf{I}$ the online models can provide acceptable predictions. To ensure that the experienced convergence properties of the online models are not specific to the arbitrary parameter set, we studied the models' behavior (without model switch) in the whole angle space. Figure 2 depicts the results. Parameter sets (α_F and $\alpha_H + \alpha_K$) belonging to convergent models are displayed. We considered a given model 'convergent' if the norm of the belonging $\mathbf{W}$ matrix remained below an arbitrary limit ($\|\mathbf{W}\| < 50$) through the whole simulation *and* the total reconstruction error remained below an arbitrary limit ($\sum_{t=1}^{t=1000} \|\mathbf{e}_{rec}^t\| < 1000$). Because of the rotational periodicity, the parameter values $\alpha_F = 0^\circ - 180^\circ$ and $\alpha_H + \alpha_K = -180^\circ - 0^\circ$ have been studied. The angle space was measured on a grid with grid points separated by 10°. To characterize the convergent parameter regions, we introduced the 'ratio of convergent set' as the number of convergent sets/ number of studied sets. The ratios are: 0.77, 0.36, 0.53, 0.45 and 0.71 for cases O1, O2, O3, O4 and O5, respectively. Surprisingly, the simplest model (O5) is almost as good as the full online KF approximation (i.e., O1).

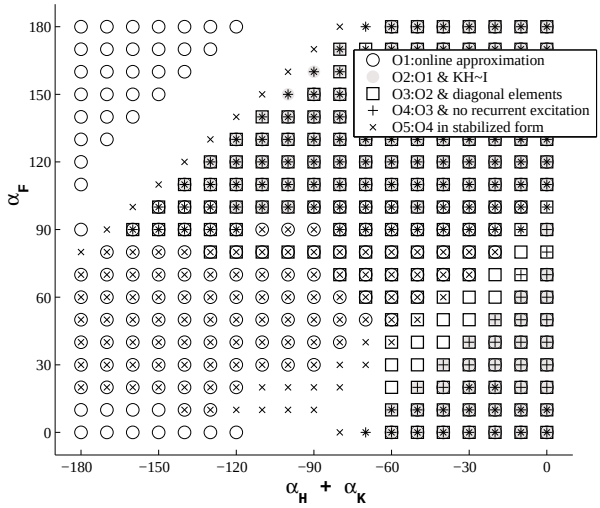

Figure 2. Convergent models. The parameter sets of different models resulting in convergent behavior are depicted. It can be seen that the the full online model (O1) and its simplest approximation (O5) are both convergent at many parameter settings. Noise and learning rate are as in Fig. 1.

3. Mapping onto the hippocampal-entorhinal loop

In the previous section we have shown that the Kalman-filter method can be approximated without the need of direct matrix inversion and the use of global knowledge. This section is intended to map the architecture onto the hippocampal formation of the mammalian brain, which contains the CA3, CA1 subfields of the hippocampus, the dentate gyrus, and the entorhinal cortex (EC) (see, e.g., (Amaral and Witter, 1989)). This mapping is motivated by striking similarities between the mathematical construction and the biological architecture (see Fig. 3**A** and 3**C**) and can be considered as an extension of the previously

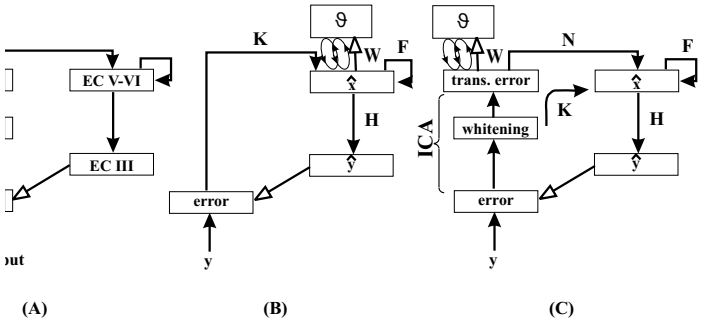

Figure 3. **(A)** The basic anatomical connections of the HC and its environment (without the EC III→CA1 'direct pathway'). The dentate gyrus (DG), the CA3 and CA1 subfields are three major regions of the hippocampus (HC). EC II and III and EC V-VI denote the superficial and deep layers of the entorhinal cortex, respectively. Solid (empty) arrows denote excitatory (inhibitory) connections. Local circuits (not shown) can be found in every layer with different modulatory effects. **(B)** The minimal structure that describes the proposed Kalman-filter. An additional functional unit has to be introduced for the modulatory θ. **(C)** Mapping of the KF model onto the HC-EC loop. Transformation **K** is realized in two consecutive processes. The first one, denoted by ICA, consists of two connection systems and separates the error signals into independent sources. The second one, denoted by **N**, may subserve different goals (not yet determined).

proposed basic model of memory organization (Lőrincz and Buzsáki, 2000; Lőrincz et al., 2002a; Lőrincz et al., 2002b).

The basic structure with the major subfields and connections is depicted on Fig. 3**A**. The proposed role of the dentate gyrus is temporal deconvolution (Lőrincz and Buzsáki, 2000) and is not specifically modelled now. The so called 'direct pathway' between EC III and CA1, the importance of which is emphasized by Sybirska et al. (2000) is not depicted, but its presumed role is noise filtering (Lőrincz et al., 2002b). Fig. 3**B** depicts the minimal architecture of the neural Kalman-filter. The functional mapping, which incorporates the results of our previous models, is shown on Fig. 3**C**.

The aforementioned basic model stated (i) that the role of internal representation is to minimize the reconstruction error between the experienced and the expected signals, and (ii) efficient information processing is based on signal separation that results in information transfer maximization (Indenpendent Component Analysis, ICA, see (Hyvärinen, 1999b) and references therein, (Lőrincz et al., 2002a)) and requires additional noise filtering, or 'sparsification' (Olshausen and Field, 1996; Hyvärinen, 1999a; Lőrincz et al., 2002b) and pattern completion mechanisms. Both processes (ICA and sparsification) can be easily realized in neuronal form. The ICA algorithm family, designed to minimize second and all higher order correlations, has recently received much attention and has also been proposed to play a role in neuronal signal processing (see, e.g., (Bell and Sejnowski, 1997)). Furthermore, it has been shown that ICA can be accelerated by a two-stage method (Amari et al., 1996). *Signal whitening* (∼normalization and decorrelation) can take place in the first stage and can also be realized in neural form (Cardoso and Laheld, 1996; Lőrincz and Buzsáki, 2000). The trade off is the requirement to maintain a separate whitening layer (Lőrincz and Buzsáki, 2000). The *second* step, the actual signal separation into independent components has also some direct effects on the mathematical construction. It implies that upon parameter optimization (i) mutual information between components of $\hat{\mathbf{x}}$ are minimized and (ii) the

'bottom up' and 'top down' processes (**K** and **H**) invert each other. This assumption, which is based on the emergent properties of the proposed reconstructing loop, is *exactly* what was made in Eq. 8 and Eq. 9. It then follows that on average the dependence of $\hat{x}_l$ on θ_k is minimal for all $k \neq l$ and, presumably, it can be neglected. This is the functional reasoning behind the simplification applied in Eq. 10 and Eq. 11.

The building blocks of the model are as follows. According to Eq. (4), (12) and (13), **F**, **H**, **K**, **N** and **W** are connection matrices determining the transition, the observation, the Kalman-gain and the auxiliary derivative matrices, respectively. Vector quantities $\hat{\mathbf{x}}$, $\mathbf{e}$, ϵ, ξ, $\hat{\theta}$ describe the internal model vector, the reconstruction error of the observations, the transformed error, the internally generated noise and the auxiliary parameter vector, respectively. As it was proposed in our previous models (Lőrincz and Buzsáki, 2000; Lőrincz et al., 2002a), CA1 seems to be a good candidate to hold the ICA transformed error signal while the internal model can be maintained by the recurrent collaterals of EC V-VI. Such persistent activity have been observed at this place (Egorov et al., 2002). It implies that functional 'micro-circuits representing 'θ may modify the CA1 output *and/or* the synaptic inputs of EC V. However, according to the order of the transformations prescribed in Eq. 16, the first case fits better and it requires fewer specific connections. This specific localization implies further that *noise* is also required to

originate at or flow through CA1 to tune $\mathbf{W}$ (Eq.12 or Eq. 24). In other words the (transformed) error and the intrinsic noise can be generated within the same subset of cells e.g., in CA1, but their generation should take place in *different* phases to avoid interference between these two terms when tuning $\mathbf{W}$. According to this view, error is propagated in one of the phases, whereas noise is propagated in in the other.

It is worth noting that the first term of the aforementioned equations seems to be 'anti-Hebbian' resulting in decorrelation, while the second term has a presynaptic strengthening effect. Surprisingly, according to Komatsu and his colleagues (Komatsu, 1994; Komatsu, 1996; Komatsu and Iwakiri, 1993) this form is in good agreement with the experimentally found *inhibitory* synaptic changes: when the presynaptic inhibitory neuron is active, the inhibitory synapse is strengthened if it is co-activated with other inhibitory inputs to the same postsynaptic cell ($\sim$presynaptic tuning), and weakened if the postsynaptic cell is simultaneously active (the anti-Hebbian term). As a consequence, depending on the nature of the targeted θ-circuits or cells, signalling through the *inhibitory* $\mathbf{W}$ may result in inhibition or disinhibition. As far as the interference is considered, gamma-oscillation would be suitable for such temporal separation. During the phase of error signal propagation, the signal ($\mathbf{y}$ or $\mathbf{y} - \hat{\mathbf{y}}$) is whitened (at CA3) and separated (at CA1) and the transformed signal modifies the internal model,

the θ-circuits and the inhibitory connections of $\mathbf{W}$. The θ-modulated output of CA1 influences the internal model, which, in turn, is used to reconstruct the original input. This signal is conveyed back to EC III. The reconstructed input is then compared to the original input and the iteration goes on. The main advantage of processing the difference instead of the whole input is that sparse activity and higher speed can be achieved at each layer. During the second phase, error propagation stops and sparse random noise is generated (in CA3 or CA1) to modify $\mathbf{W}$.

4. Discussion

This paper is intended to extend the functional model of Lőrincz and Buzsáki (Lőrincz and Buzsáki, 2000; Lőrincz et al., 2002a) on the memory organization in the EC-HC loop. One of its predictions has been that specific mechanisms should exist at the model layer to ensure the temporal integration of incoming signals. This property has been recently found in the deep layers of the entorhinal cortex (Egorov et al., 2002), which – according to the mapping – is exactly the place that was assigned to contain internal representation. It is worth noting that at

the same time, the activity self-terminates in the superficial layers of the entorhinal cortex as it was demonstrated by the same work.

However, the internal dynamics needs to be constrained for a plausible mapping. Two examples illustrate the importance of temporal aspects: (i) it has been observed that theta and gamma oscillations regulate the specific sequence learning present at the network of place cells (O'Keefe and Dostrovsky, 1971) (ii) place fields are more elongated in the EC than in the HC, suggesting that these neurons may represent longer trajectories through the environment and fire in a "path-equivalent manner" (Frank et al., 2000). By incorporating a local, online, RPE-based KF like architecture, the joint model may easily account for such predictive coding mechanism. Interestingly, the constraints of the *original* model on signal separation have simplified the RPE equations considerably and provided a plausible and more specific mapping onto the EC-HC loop.

Furthermore, Kalman-filter, as a control architecture, allows for explicit incorporation of the actual goal via top-down modulation. The theoretical connection between Kalman-filter and reinforcement learning has been described elsewhere (Szita and Lőrincz, 2004). This connection might resolve the problem of learning in multiple environments (see e.g., (Burgess and O'Keefe, 2002)). In accord with the concepts expressed in (Gupta et al., 2000), our model predicts the exis-

tence of functional interneuronal micro-circuits in CA1 and/or EC deep layers, in which individual neurons may take part in several functions (e.g., 'θ-circuits' and noise propagation). In summary, we make the following conjectures:

- There exist a predictive modeling system in the deep layers of the EC.

- There should be specific local circuits in CA1 with at least 2 functionally separable cell types.

- Strict timing is needed to prevent parallel propagation through **W** (CA1$\rightarrow \theta$) and **N** (CA1$\rightarrow$EC V-VI) when *noise* is generated.

The issue of right timing and its corresponding time scale remains to be uncovered within our modelling scheme.

Acknowledgements

This work was partially supported by the Hungarian National Science Foundation, under Grant No. OTKA T-32487.

References

Amaral, D. and M. Witter: 1989, 'The three-dimensional organization of the hippocampal formation: A review of anatomical data'. *Neuroscience* **31**, 571–591.

Amari, S., A. Cichocki, and H. Yang: 1996, 'A new learning algorithm for blind signal separation.'. In: *Advances in Neural Information Processing Systems 8*. Cambridge MA: MIT Press., pp. 757–763.

Bagchi, A.: 1993, *Optimal control of stochastic systems*. New York: Prentice Hall.

Bell, A. J. and T. J. Sejnowski: 1997, 'The 'independent components' of natural scenes are edge filters'. *Vision Res.* **37**, 3327–3338.

Bousquet, O., K. Balakrishnan, and V. Honavar: 1999, 'Is the Hippocampus a Kalman Filter?'. In: *Proceedings of the Pacific Symposium on Biocomputing*. pp. 619–630.

Burgess, N. and J. O'Keefe: 2002, *The Handbook of Brain Theory and Neural Networks*, Chapt. Spatial Models of the Hippocampus. MIT press, Cambridge MA, 2nd edition.

Cardoso, J. and B. Laheld: 1996, 'Equivalent adaptive source separation.'. *IEEE Trans. on Signal Proc.* **44**, 3017–3030.

Egorov, A., B. Hamam, E. Fransén, M. Hasselmo, and A. Alonso: 2002, 'Graded persistent activity in entorhinal cortex neurons'. *Nature* **420**, 173–178.

Eichenbaum, H.: 2001, 'The hippocampus and declarative memory: Cognitive mechanisms and neural codes'. *Behavioural Brain Research* **127**, 199207.

Frank, L., E. Brown, and M. Wilson: 2000, 'Trajectory Encoding in the Hippocampus and Entorhinal Cortex'. *Neuron* **27**, 169178.

Gupta, A., Y. Wang, and H. Markram: 2000, 'Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex.'. *Science* **287**, 273278.

Hyvärinen, A.: 1999a, 'Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation'. *Neural Computation* **11**, 1739–1768.

Hyvärinen, A.: 1999b, 'Survey on independent component analysis'. *Neural Computing Surveys* **2**, 94–128.

Kalman, R. E.: 1960, 'A New Approach to Linear Filtering and Prediction Problems'. *Transactions of the ASME–Journal of Basic Engineering* **82**(Series D), 35–45.

Komatsu, Y.: 1994, 'Age-dependent long-term potentiation of inhibitory synaptic transmission in rat visual cortex.'. *J. Neurosci.* **14**, 6488–6499.

Komatsu, Y.: 1996, 'GABAB receptors, monoamine receptors, and postsynaptic inositol trisphosphate-induced Ca2+ release are involved in the induction of long-term potentiation at visual cortical inhibitory synapses.'. *J. Neurosci.* **16**, 6342–6352.

Komatsu, Y. and M. Iwakiri: 1993, 'Long-term modification of inhibitory synaptic transmission in developing visual cortex.'. *NeuroReport* **4**, 907–910.

Ljung, L. and T. Soderstrom: 1993, *Theory and practice of recursive identification.* Cambridge, MA: MIT Press.

Lőrincz, A. and G. Buzsáki: 2000, 'Two–phase computational model training long–term memories in the entorhinal–hippocampal region'. In: H. Scharfman, M. Witter, and R. Schwarz (eds.): *The parahippocampal region: Implications for neurological and psychiatric diseases*, Vol. 911. New York: NYAS, pp. 83–111.

Lőrincz, A., B. Szatmáry, and G. Szirtes: 2002a, 'Mystery of structure and function of sensory processing areas of the neocortex: A resolution'. *J. Comp. Neurosci.* **13**, 187–205.

Lőrincz, A., G. Szirtes, B. Takács, I. Biederman, and R. Vogels: 2002b, 'Relating priming and repetition suppression'. *Int. J. of Neural Systems* **12**, 187–202.

O'Keefe, J. and J. Dostrovsky: 1971, 'The hippocampus as a spatial map. preliminary evidence from unit activity in the freelymoving rat.'. *Brain Research* **34**, 171–175.

Olshausen, B. and D. Field: 1996, 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images'. *Nature* **381**, 607–609.

Póczos, B. and A. Lőrincz: 2003, 'Kalman-filter using local interaction'. http://arxiv.org/abs/cs.AI/0302039.

Rao, R. and D. Ballard: 1997, 'Dynamic model of visual recognition predicts neural response properties in the visual cortex'. *Neural Comp.* **9**, 721–763.

Scoville, W. and B. Milner: 1957, 'Loss of recent memory after bilateral hippocampal lesions.'. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21.

Sybirska, E., L. Davachi, and P. S. Goldman-Rakic: 2000, 'Prominence of Direct EntorhinalCA1 Pathway Activation in Sensorimotor and Cognitive Tasks Revealed by 2-DG Functional Mapping in Nonhuman Primate'. *J. Neurosci.* **20**, 5827–5834.

Szita, I. and A. Lőrincz: 2004, 'Kalman filter control embedded into the reinforcement learning framework'. *Neural Computation* **16**, 491–499.

Todorov, E. and M. Jordan: 2002, 'Optimal feedback control as a theory of motor coordination'. *Nature Neurosci.* **5**, 1226–1235.

Whittle, P.: 1982, *Optimization over time*, Vol. I. of *Dynamic Programming and stochastic control*. Wiley, Chichester.