Hamiltonian synaptic sampling in a model for reward-gated network plasticity

Zhaofei Yu∗
Department of Automation
Tsinghua University
100084 Beijing, China
yuzf12@mails.tsinghua.edu.cn

David Kappel∗
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz, Austria
kappel@igi.tugraz.at

Robert Legenstein∗
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz, Austria
legi@igi.tugraz.at

Sen Song
Department of Biomedical Engineering
Tsinghua University
100084 Beijing, China
songsen@tsinghua.edu.cn

Feng Chen
Department of Automation
Tsinghua University
100084 Beijing, China
chenfeng@mail.tsinghua.edu.cn

Wolfgang Maass
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz, Austria
maass@igi.tugraz.at

Abstract

Experimental data show that synaptic connections are subject to stochastic processes, and that neural codes drift on larger time scales. These data suggest to consider besides maximum likelihood learning also sampling models for network plasticity (synaptic sampling), where the current network connectivity and parameter values are viewed as a sample from a Markov chain, whose stationary distribution captures the invariant properties of network plasticity. However convergence to this stationary distribution may be rather slow if synaptic sampling carries out Langevin sampling. We show here that data on the molecular basis of synaptic plasticity, specifically on the role of CaMKII in its activated form, support a substantially more efficient Hamiltonian sampling of network configurations. We apply this new conceptual and mathematical framework to the analysis of reward-gated network plasticity, and show in a concrete example based on experimental data that Hamiltonian sampling speeds up the convergence to well-functioning network configurations. We also show that a regulation of the temperature of the sampling process provides a link between reinforcement learning and global network optimization through simulated annealing.

1 Biological basis and mathematical framework for the model

Numerous experimental data show that synaptic connections are subject to a continuous coming and going of dendritic spines [1, 2] and axonal sprouting [3], even in the adult cortex. These processes are likely to be governed by activity-dependent plasticity rules only when a synaptic connection has been formed on a sufficiently large spine, but spine dynamics appears to be inherently stochastic and independent of pre- or postsynaptic firing in the absence of a functional synaptic connection [1].

∗These authors contributed equally.
We introduce a second dynamic variable $\Gamma_{ki}$ while via autophosphorylation among its 12 subunits. Furthermore CaMKII triggers in its activated state a short term memory or low pass filter for calcium influx with a time constant in the range of 100 s (see e.g. ch. 15 in [8], Fig. 1c in [9], Fig. 3F in [10]). Calcium influx autophosphorylated (active) state a short term memory or low pass filter for calcium influx with a time constant in the range of 100 s (see e.g. ch. 15 in [8], Fig. 1c in [9], Fig. 3F in [10]). Calcium influx is a typical feature of the induction of longterm plasticity via NMDA receptors. More specifically, incoming calcium transforms CaMKII via calmodulin into its active state, which is maintained for a while via autophosphorylation among its 12 subunits. Furthermore CaMKII triggers in its activated state changes of synaptic efficacy through the phosphorylation of AMPA receptors and the anchoring of additional AMPA receptors in the postsynaptic density, dopamin-gated stabilization of spines (see e.g. Fig. 3, S5, S11 in [10]). It is also conjectured to contribute to the formation of temporally stable NMDA-CaMKII complexes on the much larger time scale of consolidation (i.e., days) that we do not consider in our model [11, 12]. We focus here on the previously sketched transient role of CaMKII as a low pass filter in the induction of synaptic plasticity and spine enlargement.

We introduce a second dynamic variable $\Gamma_{ki}$ for each synapse. $\Gamma_{ki}$ serves as a low pass filter for signals that may lead to synaptic changes, similarly as activated CaMKII is reported to serve as a low pass filter for calcium influx through NMDA receptors. We use $\theta$ and $\Gamma$ to represent the sets of all synaptic parameters $\theta_{ki}$ and $\Gamma_{ki}$ respectively.

The interaction between the parameters $\theta_{ki}$ and $\Gamma_{ki}$ gives rise to Hamiltonian dynamics. $\Gamma_{ki}$ acts as a momentum variable that dampens fast synaptic changes. As a results, synaptic plasticity is stabilized and synaptic weight changes are likely to follow the same direction (either potentiation or depression) for longer times. In addition we include a source of noise to the momentum variable $\Gamma_{ki}$. This allows us to model the permanently ongoing dendritic spine dynamics [11,12] and at the same time facilitates exploration of the parameter space. We therefore model the interaction between the variables $\theta_{ki}$ and
\( \Gamma_{ki} \) by the following set of stochastic differential equations (SDEs)

\[
d\theta_{ki} = a \Gamma_{ki} \, dt, \quad d\Gamma_{ki} = \left( a \frac{\partial}{\partial \theta_{ki}} \log p^*(\theta) - b \Gamma_{ki} \right) dt + \sqrt{2Tb} \, dW_{t,ki},
\]

where \( a, b > 0 \) are scaling constants and \( dW_{t,ki} \) denotes infinitesimally small changes of a Wiener process \( W_{t,ki} \).

Despite the differences between eqs. (1) and eq. (2), one can show that they give rise to the same stationary distribution \( p^*_T(\theta) \). In Sec. 4 we derive the stationary distribution of the SDEs eq. (2) and show, that \( p^*_T(\theta) = \frac{1}{Z} p^*(\theta)^{\frac{1}{T}} \) is the marginal of this stationary distribution with regard to the variables \( \theta_{ki} \) for some normalizing constant \( Z \). In fact, the dynamics (1) and (2) are special cases of a more general range of possible dynamics for synaptic sampling (see Sec. 4 and the Supplement), that all have the same stationary distribution \( p^*_T(\theta) \). Hence, these different dynamics do not differ in their stationary properties, however, they do differ in the way how these are approached. The dynamics (2) introduces a hidden synaptic state \( \Gamma_{ki} \) for each synapse that filters the gradient on a slower time scale than single pairings of pre- and postsynaptic activity (defined by parameter \( b \)) and synaptic parameters are updated indirectly based on this hidden synaptic state. We propose that this gradient filtering is performed on the time scale of CaMKII dynamics, i.e., on the time scale of seconds to minutes [8, 9]. In artificial neural network learning, such gradient filtering has been proven advantageous and is commonly known as gradient descent with momentum. Here however, due to the added noise, the parameter dynamics do not perform gradient descent but sampling. In this context, the stochastic dynamics is known as Hamiltonian sampling [14, 15] based on the same idea of a momentum that biases a physical mass (i.e., the state variable) to keep its direction of movement. Compared to Langevin sampling implemented by the synaptic sampling dynamics (1), Hamiltonian sampling is known to drastically speed up the sampling process in many situations [14].

Whereas previous work on synaptic sampling [5, 6] had only considered the case of unsupervised learning, we apply it here to reward-based learning, and show that previously proposed deterministic rules for reward-gated synaptic plasticity arise as special cases if stochastic spine dynamics is ignored. It was shown, however, that rewards also have an important impact on spine dynamics: Dopamine promotes spine enlargement during a narrow time window (0.3 to 2 s) after the glutaminergic inputs [10]. For modelling reward-gated network plasticity we set \( p^*(\theta) = p_N(\theta | R = 1) \propto p_S(\theta) p_N(\theta | R = 1 | \theta) \), where \( p_S(\theta) \) denotes a prior over network parameters and \( p_N(\theta | R = 1 | \theta) \) is the probability that the neural network \( N \) receives a reward for the given parameters. \( R \) is here an abstract binary variable that denotes the presence of a global reward signal. Then the preceding mathematical results imply for a flat prior that the stochastic dynamics of \( \theta_{ki} \) and \( \Gamma_{ki} \) according to (2) enables the network to sample on the long run from the distribution \( p^*_T(\theta) = \frac{1}{Z_T} p_N(\theta | R = 1)^{\frac{1}{T}} \) that has most of its mass on network configurations that provide attractive compromises between the prior and reward-maximization.

In fact, this mathematical framework allows us to create a link from reinforcement learning to optimization theory, which establishes conditions under which a neural circuit could attain not only functionally attractive locally optimal network configurations, but in principle even a global optimum. In particular, for an uninformative prior, when the network has reached the stationary distribution \( p^*_T(\theta) \), the expected reward at temperature \( T \), \( E[R_T] \), is given by (see Sec. 2.2)

\[
E[R_T] = \frac{1}{Z_T} \int p_N(R = 1 | \theta)^{1 + \frac{1}{T}} \, d\theta.
\]

Hence, sampling from \( p^*_T(\theta) \) with decreased temperature \( T < 1 \) concentrates parameter samples at values that lead to large rewards and therefore increases the expected reward of the network. For small temperatures, the posterior is concentrated at the global optima of the reward landscape. Getting close to a globally optimal network configuration is practically encumbered by the need for a suitable cooling schedule, and the need to converge for each temperature \( T \) within a reasonable time length to the associated stationary distribution. While some data suggest that the genetic program for developmental learning has some features that are reminiscent of a cooling schedule [16], a Hamiltonian sampling dynamics is likely to improve the convergence speed for each temperature.

The remainder of the paper is organized as follows: In Sec. 2 we first derive the reward-modulated synaptic plasticity rule for a network of spiking neurons. In Sec. 3 we analyze the behavior of this network for a motor learning experiment. Finally in Sec. 4 we present a general framework of synaptic parameter dynamics and rewiring.
Figure 1: Emergence of a suitable network architecture and parameters through reward-based learning in a model for the task of [13]. A: Illustration of the network architecture and learning task. B: Spiking activity of the network at the beginning and after learning. Activities of random subsets of neurons from all populations are shown. Lever position inferred from the neural activity is shown at the bottom. Rewards are indicated by red bars. Vertical blue lines indicate cue onsets. C: Trial-averaged network activity (top) and lever movements (bottom) before and after learning. Arrowheads indicate cue onset. Average lever movement (black) and 10 individual movements (gray) are shown. Both network activity and resulting lever presses become more efficient and stereotypical through learning. D: Illustration of the model for structural plasticity. Values of synaptic parameters $\theta_{ki}$ are mapped to synaptic weights $w_{ki}$. Negative values of $\theta_{ki}$ indicate non-functional (disconnected) synapses. E: Snapshots of four example synapses at three different time points throughout learning. Synaptic parameters are represented by spine head size. Synapses show different behaviors (blue: transient decaying, green: transient emerging, gray: stable) qualitatively similar to experimental data [2, 1]. F: The synaptic turnover throughout learning. Mean values over 5 runs are shown. Bars indicate SEM. G: Comparison of learning curves for synaptic sampling with (blue) Hamiltonian dynamics according to equ. (2), and without according to equ. (1) in green. Average performance is measured here as the average time from cue onset to reward. Shaded areas indicate SEM. Means and SEM values are computed over 5 independent simulation runs. H: Traces of the synaptic parameter $\theta_{ki}$, the momentum $\Gamma_{ki}$ and the eligibility trace $e_{ki}$ for one example synapse. I: PCA of a random subset of the parameters $\theta_{ki}$. The plot suggests continuing dynamics in task-irrelevant dimensions after the learning goal has been reached. Line color: task performance.
2 Application to reward-gated reconfiguration and synaptic plasticity in a network of spiking neurons

Here, we apply the Hamiltonian synaptic sampling framework to a spiking neural network. We first describe the network model and then derive the gradient \( \frac{\partial}{\partial \theta_{ki}} \log p^*(\theta) \) which is necessary to define concrete plasticity rules for parameter sampling in a reward-based learning context.

2.1 Network model

We considered a network \( N \) of \( K \) neurons \( z_1, \ldots, z_K \) with potentially asymmetric recurrent connections, where \( w_{ki} \) denotes the weight from presynaptic neuron \( z_i \) to network neuron \( z_k \). Network neurons were modeled by a standard stochastic variant of the spike response model [17].

\[
\begin{align*}
  u_k(t) &= \sum_{i \neq k} y_i(t) w_{ki} + \varphi_k(t),
  \end{align*}
\]

where \( \varphi_k(t) \) denotes the slowly changing bias potential of neuron \( z_k \), used to ensure that the output rate of each neuron stays within finite bounds (see Supplement). \( y_i(t) \) denotes the trace of the (unweighted) postsynaptic potentials (PSPs) from neuron \( z_i \) at time \( t \). We used standard PSP kernels with a brief finite rise and exponential decay, but any other PSP shape may be used in principle. We denote the output spike train of neuron \( z_k \) by \( z_k(t) \) defined as a sum of Dirac delta pulses positioned at the spike times \( t^{(1)}_k, t^{(2)}_k, \ldots \), i.e., \( z_k(t) = \sum \delta(t - t^{(l)}_k) \). The firing probability of neuron \( k \) at time \( t \) is defined as \( f_k(t) = f(u_k(t), \rho_k(t)) \), where \( \rho_k(t) \) denotes a refractory variable that is given by the time elapsed since the last spike of neuron \( z_k \). In this article, we set \( f(u_k, \rho_k) = \sigma(u_k)\Theta(\rho_k - \tau_{ref}) \), where \( \sigma(u_k) \) is a sigmoid activation function \( \sigma(u_k) = \frac{1}{1 + e^{-u_k}} \) and \( \Theta(\cdot) \) denotes the Heaviside step function, i.e. \( \Theta(x) = 1 \) for \( x \geq 0 \) and 0 otherwise.

2.2 Reward-based learning with stochastic plasticity and rewiring

In the reward-based framework we set \( p^*(\theta) = p_N(\theta | R = 1) \propto p_S(\theta)p_N(R = 1 | \theta) \) as described above, where \( p_S(\theta) \) is a Gaussian prior with mean \( \mu \) and variance \( \sigma^2 \) for each parameter \( \theta_{ki} \). The derivative that appears in (2) and (1) can in this case be decomposed into a contribution from the prior and a contribution from the reward likelihood:

\[
\begin{align*}
  \frac{\partial}{\partial \theta_{ki}} \log p_N(\theta | R = 1) &= \frac{\partial}{\partial \theta_{ki}} \log p_N(\theta) + \frac{\partial}{\partial \theta_{ki}} \log p_N(R = 1 | \theta).
  \end{align*}
\]

The contribution of the Gaussian prior is given by \( \frac{\partial}{\partial \theta_{ki}} \log p_S(\theta) = \frac{1}{\sigma^2} (\mu - \theta_{ki}) \). In the model network we denote by \( r(t) \) the current reward at time \( t \). This reward function encodes the probability that the global binary reward \( R \) is acquired, given the network activity up to time \( t \). Then, the gradient of the reward likelihood can be estimated as (see Supplement for details):

\[
\begin{align*}
  \frac{\partial}{\partial \theta_{ki}} \log p_N(\theta | R = 1) &= \int_{t=0}^{\infty} r(t) \int_{\tau=0}^{t} \lambda^\tau w_{ki} y_i(t-\tau)(z_k(t-\tau) - f_k(t-\tau)) \, d\tau \, dt,
  \end{align*}
\]

where \( \lambda < 1 \) is a discount factor that exponentially discounts the contribution of previous pre-post spike pairs to the current reward. In the network model, eq. (5) can be realized by a plasticity mechanism that uses an eligibility trace \( e_{ki}(t) \), which is updated online:

\[
\begin{align*}
  de_{ki}(t) &= (\log \lambda \, e_{ki}(t) + w_{ki} y_i(t)(z_k(t) - f_k(t))) \, dt,
  \end{align*}
\]

Learning rules of this from were found previously in the context of reward-based learning [18, 19], but these models did not include stochastic rewiring and Hamiltonian dynamics and were not applied for reward-based learning in recurrent networks.

Eligibility traces of the form (6) work on time scales of seconds and are known to lead to reliable structural plasticity if they are paired with reward [10]. We used a time constant of 1 s for the eligibility trace in our experiments. Using these derivatives, reward-based learning can be implemented through both Hamiltonian [2] and classical [11] synaptic sampling. In the simulation we used a time constant of 100 s for the momentum variable \( \Gamma_{ki} \), and a fixed temperature of \( T = 0.1 \).
The temperature $T$ plays a critical role in the parameter dynamics, as it shapes the stationary distribution $p_T^*(\theta)$. Denote by $R_T$ the Bernoulli random variable that indicates reward at temperature $T$. Under the assumption that the network has reached the stationary distribution $p_T^*(\theta)$, the expectation of $R_T$ (i.e., the expected reward at temperature $T$) is then

$$E[R_T] = \sum_{q \in \{0,1\}} q \cdot p_N(R_T = q) = p_N(R_T = 1) = \int p_N(R = 1 \mid \theta) p_T^*(\theta) d\theta =$$

$$= \frac{1}{Z} \int p_N(R = 1 \mid \theta) p^*(\theta) d\theta = \frac{1}{Z} \int p_N(R = 1 \mid \theta) \frac{1}{2} \int p_s(\theta) \frac{1}{2} d\theta = \frac{1}{Z} \int p_N(R = 1 \mid \theta)^{\frac{1}{2} + \frac{1}{2}} d\theta,$$

where we assumed an uninformative prior $p_s(\theta)$ in the last step. In the limit $T \to 0$, the stationary distribution $p_T^*(\theta)$ converges to the uniform distribution over optimal parameter settings (with other parameter settings assuming zero probability) and $E[R_T]$ also assumes its global optimum. One attempt to attain such an optimum is to start with a large temperature and reduce it slowly towards 0. Such an annealing procedure is used in simulated annealing, a non-linear optimization technique. In order to be able to apply a reasonably fast annealing schedules however, fast convergence to the stationary distribution is necessary at each temperature. Therefore, the faster sampling dynamics of Hamiltonian synaptic sampling may turn out crucial. In summary the plasticity dynamics that implements on-line reward-based Hamiltonian synaptic sampling are given by

$$d\theta_{ki} = a \Gamma_{ki} dt, \quad d\Gamma_{ki} = \left( a \left( \frac{1}{\sigma^2} (\mu - \theta_{ki}) + r(t) e_{ki}(t) \right) - b \Gamma_{ki} \right) dt + \sqrt{2TB} dW_{\Gamma_{ki}}. \quad (8)$$

Compared to the synaptic sampling equations eq. (1), Eqs. (8) updates the parameters $\theta_{ki}$ according to the average gradient $\Gamma_{ki}(t)$ instead of the current gradient $\frac{1}{\sigma^2} (\mu - \theta_{ki}) + r(t) e_{ki}(t)$.

### 3 A model for experimental data on reward-based learning in motor cortex

Changes of network activity and spine turnover in motor cortex were monitored in [13] through calcium imaging over 2 weeks, while mice acquired a forelimb lever-press task through reward-based learning. A reward was given when the lever press crossed two thresholds within a given time window marked by an auditory cue. We examined to what extent a simple model based on the previously described framework for network plasticity would be able to reproduce the observed changes in neural activity, the observed transient turnover in spine dynamics, and the learning of the task.

We adapted the learning task [13] in the following way for our model (see Fig. 1A and B). Trial duration was indicated through the presentation of a cue pattern (a fixed, randomly generated rate pattern for all 200 input neurons, see Supplement). Reward was delivered when the lever crossed the threshold +5 after first crossing a lower threshold -5 (see black lines in Fig. 1B) within 10 s after cue onset. After each trial a brief holding phase of random length was inserted, during which input neurons were set to a background input rate of 2 Hz.

We trained a simple generic recurrent network consisting of 70 excitatory and 20 inhibitory neurons to solve this task. One pool $D$ of 10 excitatory neurons within this recurrent network was randomly selected to cause downwards movements of the lever, and another pool $U$ of 10 neurons for upwards movements. All synaptic connections from the external input (cue) and between the 70 excitatory neurons (including those in the pools $D$ and $U$) in the network were subjected to the rule (8) for reward-based Hamiltonian synaptic sampling. The constants $b$ where chosen to amount to a time constant of 100 s for the dynamics of $\Gamma_{ki}$ (see Supplement for additional parameters). We also allowed for multiple synaptic connections between neurons, which commonly exist according to experimental data [20], see Supplement. Synaptic connections and weights from and to inhibitory neurons (black arrows in Fig. 1A) were randomly chosen and fixed. Thus the network had to learn without any guidance, except for the reward in response to good performance, to create after the onset of the cue first higher firing in pool $D$ and then higher firing in pool $U$ (see Supplement for details).

We used the model for spine motility from [5][6], that is illustrated in Fig. 1D. Three example spines at different sizes are shown. Corresponding values of synaptic parameters $\theta_{ki}$ and synaptic efficacies $w_{ki}$ are shown next to the spines. Negative synaptic parameters $\theta_{ki}$ denote nonfunctional (disconnected)
synapses with \( w_{ki} = 0 \). For \( \theta_{ki} > 0 \) the weight \( w_{ki} \) of the synapse is given by \( \exp(\theta_{ki} - \theta_i) \). Spine
heads are represented in Fig. 1D,E by circular volumes with diameters proportional to \( \sqrt{w_{ki}} \). Spine
neck dimensions are scaled accordingly to facilitate the illustration. Four example spines are shown
at three different time points during learning in Fig. 1E. Different patterns of behavior emerge for the
spines, analogously as in recorded data \([21]\). Many synapses are transient (either decay or emerge
over time) whereas some synapses show stable behavior (persistently functional or non-functional).

Network responses before and after learning are shown in Fig. 1B. Initially the rewarded goal is
only reached occasionally. After learning for 2 hours the network is able to solve the task in most
of the trials and the average trial duration (time to reward) decreases significantly (Fig. 1H). This
effect is accompanied by more stereotyped network activity and lever movement patterns as in the
experimental data of \([13]\), compare our Fig. 1C with Fig. 1b and Fig. 2 of \([13]\). In Fig. 1C we
shows the trial-averaged activity of the 70 excitatory neurons before and after learning for 8 hours.
The neurons are sorted by their maximum activity peak after cue onset (arrowheads). The left and
center plots are sorted according to the activity at the beginning of learning. The right plot was
produced by resorting the center plot according to the activity after learning. The role of different
neurons is restructured throughout learning. Also, the neural activity peaks become narrower and
less background noise is observed after learning. Lower panels in Fig. 1C show the average lever
movement and 10 individual movement traces at the beginning and the end of learning. The lever
movement trajectories become more stereotyped after learning. Note that this learning is nontrivial
because of the credit assignment problem which the network has to solve: the network does not
"know" which neurons belong to the movement-triggering pools D, U, and synapses do not "know"
whether they connect hidden neurons, neurons within a pool, hidden neurons and pool-neurons, or
input neurons with other neurons. Furthermore the plasticity of all these different synapses is gated
by the same global reward signal.

3.1 Analysis of rewiring and the Hamiltonian dynamics of synaptic sampling

The total synaptic turnover (number of emerging and decaying synapses in time windows of 10
minutes) is shown in Fig. 1F for every phase of learning. The initial phase of learning causes a large
increase of synaptic turnover. This phase of network rewiring lasts for about 1 h, during which the
largest performance increase is observed (see Fig. 1G). This result is compatible with experimental
data, which show that learning of a new task is accompanied by increased spine turnover \([21]\). After
this brief initial phase of strong turnover, the level of rewiring is reduced to a significantly lower
but nonzero level of continuously ongoing connectivity change, which persists throughout the total
learning session. The connectivity patterns that emerge through learning are sparse. Only about 15% of
the total number of synaptic connection opportunities are functional at the end of the learning
session.

Despite this permanent rewiring the network is able to maintain its performance throughout the 8
hour session. Fig. 1G shows the mean time from cue onset to reward (used as measure for learning
progress in the following). The network’s random activity allows it to solve the task on average after
5 seconds at the beginning of training. The network performance is significantly enhanced during
the first 2 hours of learning and then this performance is stably maintained throughout prolonged
learning. Fig. 1G also shows that the learning performance was significantly worse if the synaptic
sampling learning rule \((1)\) was used instead of the Hamiltonian dynamics \((2)\) (mean trial duration
was 4.6 without and 2.1 with Hamiltonian dynamics at the end of training).

In Fig. 1H we analyze the evolution of the eligibility trace \( e_{ki} \), the momentum variable \( \Gamma_{ki} \), and the
synaptic parameter \( \theta_{ki} \) for one example synapse over a time span of 20 minutes during learning. The
momentum variable acts as a low passing filter that stabilizes learning signals. Note that \( \Gamma_{ki} \) can also
assume negative values, and then enhances LTD. This feature is consistent with experimental data
which show that also LTD requires the activated form of CaMKII, and that the switch between LTP
and LTD is implemented through other mechanisms \([22, 23]\). The time constant of 100 seconds used
for the momentum variable \( \Gamma_{ki} \) is consistent with biological time constants of CaMKII \([8, 9, 10]\).

A structural difference between stochastic learning models such as synaptic sampling and other
learning models that focus on convergence of parameters to a (locally) optimal setting becomes
apparent through Fig. 1F. I. Although performance did no longer improve after 2 h, both network
connectivity and parameters kept changing in task-irrelevant dimensions, as often observed in
experimental data, see e.g. \([24]\). For Fig. 1I we randomly selected 1% of the roughly 90000
parameters \( \theta_{ki} \) and plotted the first 3 principal components of their dynamics. After about 2 hours of learning the parameter vector reaches a lower dimensional manifold where high performance is achieved. This transition is accompanied by a clearly visible change in movement direction in the PCA space.

4 Relating synaptic plasticity rules to the stationary distribution of network configurations

Here we present the general mathematical framework of synaptic parameter dynamics and derive the emerging stationary distribution of network configurations that results from this dynamics. For different parameter settings, this framework includes various types of dynamics as special cases, of which two have been considered in this article: Hamiltonian synaptic sampling \( (2) \) and synaptic sampling \( (1) \). The generalized model is given by the following set of SDEs:

\[
\begin{align*}
d\theta_{ki} &= \left(-a \frac{\partial \log p^*(\Gamma)}{\partial \theta_{ki}} + c \frac{\partial \log p^*(\theta)}{\partial \theta_{ki}}\right) dt + \sqrt{2Tc} dW_{\theta_{ki}} \\
d\Gamma_{ki} &= \left(a \frac{\partial \log p^*(\theta)}{\partial \Gamma_{ki}} + b \frac{\partial \log p^*(\Gamma)}{\partial \Gamma_{ki}}\right) dt + \sqrt{2Tb} dW_{\Gamma_{ki}}.
\end{align*}
\]

(9)

Here, \( T > 0 \) is the temperature parameter, \( W_{\theta_{ki}}, W_{\Gamma_{ki}} \) are independent one-dimensional Wiener processes, and \( a, b, c \) are positive constants. This dynamics describes a general noisy first-order interaction between visible synaptic parameters \( \theta_{ki} \), that determine the efficacy of the synapse, and hidden synaptic parameters \( \Gamma_{ki} \), that model the local concentration of CaMKII in its activated state. The dynamics can thus be seen as a generalization of standard gradient-based synaptic plasticity rules (e.g. for maximum likelihood learning) that includes both, hidden synaptic variables and stochastic plasticity. For the general dynamics, the joint distribution over the sets of all parameters \( \theta \) and \( \Gamma \) will converge after a while to \( p^*_T(\theta, \Gamma) = \frac{1}{Z_T} p^*(\theta)^\frac{1}{2} p^*(\Gamma)^\frac{1}{2} \) and produce samples from it. This result can be formalized in the following theorem:

**Theorem 4.1** Let \( p^*(\theta), p^*(\Gamma) \) be strictly positive, continuous probability distributions over parameters \( \theta \) and \( \Gamma \) respectively, twice continuously differentiable with respect to \( \theta \) and \( \Gamma \). Let \( a, b, c \) be positive constants. Then the set of stochastic differential equations (9) leaves the distribution \( p^*_T(\theta, \Gamma) = \frac{1}{Z_T} p^*(\theta)^\frac{1}{2} p^*(\Gamma)^\frac{1}{2} \) invariant. Furthermore, this is the unique stationary distribution of the sampling dynamics.

The proof of Theorem\( 4.1 \) follows directly by writing the SDEs in the corresponding Fokker-Planck equation (FPE) and then inserting the assumed stationary distribution \( p^*_T(\theta, \Gamma) \) as the stochastic equilibrium of the FPE. By marginalization over the hidden synaptic parameters \( \Gamma_{ki} \) it then follows that the stationary distribution over the visible synaptic parameters is given by \( p^*_T(\theta) = \int p^*_T(\theta, \Gamma) d\Gamma = \frac{1}{Z_T} p^*(\theta)^\frac{1}{2} \) (See the Supplement for a detailed proof).

Hamiltonian sampling and synaptic sampling are special cases of the more general parameter dynamics (9). Hamiltonian synaptic sampling as defined in (2) is obtained by choosing \( c = 0 \) and a Gaussian distribution for the hidden parameters \( p^*(\Gamma) \sim \text{NORMAL}(0, 1) \). Synaptic sampling as defined in (1) is obtained by choosing \( a = b = 0 \). We remark that various types of gradient descent can also be recovered from the generalized dynamics for \( T = 0 \), e.g. gradient descent with momentum for the noiseless Hamiltonian dynamics. Equation \( 9 \) can be seen as the continuous version of Hamiltonian sampling \( 13 \), where a Metropolis update is performed after simulating Hamiltonian dynamic. Hamiltonian sampling has been considered recently in \( 15 \) for neuronal dynamics, but not for parameter dynamics as we propose in this article. Equation \( 9 \) can also be seen as an extension of \( 13\) \( 25 \) to the case where the temperature \( T \) is used to shape the static distribution \( p^*_T(\theta) \).

5 Discussion

We have presented a theoretical framework for reward-based network plasticity that captures not only plasticity of weights, but in contrast to preceding models also network rewiring (spine dynamics). Since experimental data suggest that spine dynamics is inherently stochastic, one needs a stochastic learning framework for that. Reward-based network learning is modeled in our approach as a synaptic sampling process, whose stationary distribution of network configurations is determined by structural
priors (such as sparse connectivity) and the likelihood of receiving rewards. We have shown that meaningful local rules for reward-based local synaptic plasticity and spine dynamics can be derived from this general framework.

The underlying theory implies that reward-based network plasticity is in principle able to acquire through down-regulation of the temperature the full power of simulated annealing for optimizing the network for a specific task. It is essential for this approach that network plasticity converges for each temperature quite fast to its stationary distribution. Hence it is of interest to realize that biological data on the activation dynamics of the kinase CaMKII support the hypothesis that biological networks of neurons are able to approximate Hamiltonian sampling of network configurations, rather than only the slower Langevin sampling. We propose that this feature makes biological reward-based network reconfiguration and synaptic plasticity substantially more efficient. We have demonstrated the advantage of Hamiltonian sampling in a model for a concrete reward-based motor learning experiment in mice that was shown to involve a substantial turnover of synaptic connections.

Acknowledgments

Written under support of the European Union project #604102 The Human Brain Project (HBP).

References

[1] Yasumatsu N, Matsuzaki M, Miyazaki T, Noguchi J, and Kasai H. Principles of long-term dynamics of dendritic spines. The Journal of Neuroscience, 28(50):13592–13608, 2008.
[2] Holtmaat A and Svoboda K. Experience-dependent structural synaptic plasticity in the mammalian brain. Nature Reviews Neuroscience, 10(9):647–658, 2009.
[3] Stettler DD, Yamahachi H, Li W, Denk W, and Gilbert CD. Axons and synaptic boutons are highly dynamic in adult visual cortex. Neuron, 49:877–887, 2006.
[4] Ziv Y, Burns LD, Cocker ED, Hamel EO, Ghosh KK, Kitch LJ, Gamal AE, and Schnitzer MJ. Long-term dynamics of CA1 hippocampal place codes. Nature Neuroscience, 16(3):264–266, 2013.
[5] Kappel D, Habenschuss S, Legenstein R, and Maass W. Synaptic sampling: A Bayesian approach to neural network plasticity and rewiring. In Advances in Neural Information Processing Systems, pages 370–378, 2015.
[6] Kappel D, Habenschuss S, Legenstein R, and Maass W. Network plasticity as Bayesian inference. PLoS Comput Biol, 11(11):e1004485, 2015.
[7] Sheng M and Kim E. The postsynaptic organization of synapses. Cold Spring Harbor perspectives in biology, 3(12):a005678, 2011.
[8] Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, and Walter P. Molecular Biology of the Cell, 6th edition. Garland Science, 2014.
[9] Lisman J, Yasuda R, and Raghavachari S. Mechanisms of CaMKII action in long-term potentiation. Nature reviews neuroscience, 13(3):169–182, 2012.
[10] Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, and Kasai H. A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science, 345(6204):1616–1620, 2014.
[11] Lisman J and Raghavachari S. Biochemical principles underlying the stable maintenance of LTP by the CaMKII/NMDAR complex. Brain Research, 1621:51–61, 2015.
[12] Graupner M and Brunel N. STDP in a bistable synapse model based on CaMKII and associated signaling pathways. PLoS Computational Biology, 3(11):2299–2323, 2007.
[13] Peters AJ, Chen SX, and Komiyama T. Emergence of reproducible spatiotemporal activity during motor learning. Nature, 510(7504):263–267, 2014.
[14] Neal RM. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo, 2:113–162, 2011.
[15] Aitchison L and Lengyel M. The Hamiltonian brain. arXiv preprint:1407.0973, 2014.
[16] Gopnik A, Griffiths TL, and Lucas CG. When younger learners can be better (or at least more open-minded) than older ones. Current Directions in Psychological Science, 24(2):87–92, 2015.
[17] Gerstner W, Kistler WM, Naud R, and Paninski L. Neuronal dynamics: From single neurons to networks and models of cognition. Cambridge University Press, 2014.
[18] Urbanczik R and Senn W. Reinforcement learning in populations of spiking neurons. Nature neuroscience, 12(3):250–252, 2009.
[19] Brea J, Senn W, and Pfister JP. Matching recall and storage in sequence learning with spiking neural networks. *The Journal of Neuroscience*, 33(23):9565–9575, 2013.

[20] Feldmeyer D, Lübke J, Silver RA, and Sakmann B. Synaptic connections between layer 4 spiny neurone-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: physiology and anatomy of interlaminar signalling within a cortical column. *The Journal of physiology*, 538(3):803–822, 2002.

[21] Xu T, Yu X, Perlik AJ, Tobin WF, Zweig JA, Tennant K, Jones T, and Zuo Y. Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature*, 462(7275):915–919, 2009.

[22] Coultrap SJ, Freund RK, O’Leary H, Sanderson JL, Roche KW, Dell’Acqua ML, and Bayer KU. Autonomous CaMKII mediates both LTP and LTD using a mechanism for differential substrate site selection. *Cell Reports*, 6(3):431–437, 2014.

[23] Connor SA and Wang YT. A place at the table: LTD as a mediator of memory genesis. *The Neuroscientist*, pages 1–13, 2015.

[24] Todorov E and Jordan MI. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002.

[25] Ma YA, Chen T, and Fox E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2899–2907, 2015.