Imperial College of Science, Technology and Medicine
Faculty of Natural Sciences
Department of Physics

Implementing spike-based tempering on the BrainScaleS-1 mixed-signal neuromorphic system

Madison Cotteret

Under the supervision of
Dr. Sebastian Schmitt and Ákos Kungl
at the Kirchhoff-Institute for Physics, July 2019

Abstract

Stochastic sampling networks approximating Boltzmann machines are a brain-inspired machine learning paradigm that has been used extensively for its Bayesian inference capabilities. However, due to their inherently inhomogeneous probability landscape, the network may get trapped in a single mode by low probability transition states, which pose an effective energy barrier, prevent mixing between modes and harm the generative capabilities of the network. Mixing has been facilitated in simulation by modulating the rate of noise to sampling neurons, effecting a flattening of the probability landscape. In this work we implement this tempering on the BrainScaleS-1 mixed-signal physical model system. We implemented tempering via noise *weight* modulation, with the addition of high weight low frequency noise resulting in a temperature increase. We verified the linear scaling of the associated abstract BM weights in spiking networks, and showed that mixing in a simple two-mode system is indeed facilitated. Contrary to simulation, tempering via noise *rate* modulation was not successful. Despite this, it still improved mixing in the simplified example, by distorting the network in a non-trivial manner. Weight modulated noise should thus be used to improve the performance of generative sampling networks.

Zusammenfassung

Netzwerke von spikenden Neuronen, die von Boltzmann Verteilungen sampeln, sind ein vom Gehirn inspiriertes maschinelles Lernverfahren, die vor allem wegen ihrer bayes'schen Inferenzfähigkeiten benutzt werden. Wegen ihrer inhomogenen Wahrscheinlichkeitsverteilung, werden Moden mit einer hohen Wahrscheinlichkeit voneinander durch effektiven Energiebarriere, bestehend aus Zuständen niedriger Wahrscheinlickheiten, getrennt. Dies verursacht Probleme mit dem Mixing zwischen diesen Moden. In Simulation war eine passende Modulation der Frequenz des Hintergrund-Poisson-Rauschens eine Lösung mit der die Energie-Landschaft abgeflacht wird. Wir streben hier an, dieses Tempering auf der BrainScaleS-1 Hardwareplatform zu implimentieren.Wir haben Tempering durch die Modulation der Rauschen*gewichte* implementiert, worin der Zusatz des hoch-gewichteten Rauschens mit einer niedrigen Frequenz einen Temperaturanstieg verursacht hat. Wir haben auch bestätigt, dass die entsprechenden BM-Gewichte dadurch linear skalieren, und wir haben auch gezeigt, dass Mixing in einem einfachen zwei-Moden System verbessert wird. Im Widerspruch zu Ergebnissen in Simulation wurde Tempering durch die Modulation der Rauschen*frequenz* nicht erreicht. Trotzdem wurde verbessertes Mixing in diesem zwei-Moden System beobachtet, indem das Netzwerk nicht trivial verdreht wird. Die Rauschengewichte sollten folglich moduliert werden, um die Leistungsfähigkeit der generativen Samplingnetzwerke zu verbessern.

Contents

Abstract **i**

Preface **1**

1 Introduction **2**

2 Theoretical background **4**

2.1 Single neuron dynamics 4

2.1.1 The leaky integrate-and-fire neuron model with conductance based synapses 5

2.1.2 High conductance state 6

2.1.3 Activation functions 7

2.2 Sampling theory 8

2.2.1 Boltzmann machines 8

2.2.2 LIF Networks implement Gibbs sampling 10

2.2.3 Training and the mixing problem 11

2.2.4 Temperature modulation 12

2.2.5 Noise networks 13

3 Hardware details **14**

4 Experiments **16**

4.1 Creating a modulated noise network 16

4.1.1 Initial considerations 16

4.1.2 Noise network optimisation 17

4.1.3 Spike loss on hardware 19

4.1.4 Modulating the noise network on hardware 23

4.2 On the refractory period 25

4.2.1 Accounting for spike loss 28

4.2.2 Refractory periods on Wafer 33 28

4.2.3 Refractory periods on Wafer 30 28

4.3 Limits of Poisson input noise on the HICANNv4 chip 29

4.3.1 Limits of Poisson noise weight 32

4.3.2 Membrane potential distributions with varying noise rates 38

4.4 Widening of single neuron activation functions on the HICANNv4 chip 39

4.4.1 Widening via noise rate modulation 39

4.4.2 Widening via noise weight modulation 48

4.5 State probability distribution changes under noise weight modulation 50

4.6 Mixing aided by noise modulation 51

4.6.1 Mixing with tempering via weight modulation 51

4.6.2 Mixing with noise rate modulation 53

5 Discussion **56**

6 Conclusions **60**

Bibliography **60**

Appendix **64**

Acknowledgements **66**

Preface

All experiment implementations within the Experiments section both on hardware and in simulation are my own work, along with all experiment-specific data analysis code. If experiments are referred to within the Experiments section which are not of my own making, I will make it explicitly clear. I made extensive use of the widely available Python packages matplotlib, NumPy, SciPy, and Elephant [Yeg+15]. I did not make any contributions to the theory, apart from coarsely modelling spike loss as a convolution, and the majority of the theory is taken from [Pet15]. I did not make any contribution to the BrainScaleS system itself [Sch+10]. For running on the hardware and in simulation, I used the PyNN network specification framework [Dav+10], to which I also did not make any contributions.

Chapter 1

Introduction

The human brain is able to perform pattern recognition and certain computational tasks far better than even the most cutting edge machine learning algorithms [Alc+19] [Sze+13]. Compared to conventional computing, it benefits from a low power consumption of about 20W, and is able to perform complex inference and decision making tasks in real time given noisy, imperfect and ambiguous data [Dru00]. The neuromorphic computing subdivision of the Human Brain Project thus aims to harness these desirable characteristics for computation, where novel brain-inspired computing frameworks are explored [Mar12]. For certain tasks, such analog neuromorphic implementations outperform their conventional software counterparts [Wun+19].

Novel computing architectures require correspondingly novel and suitable algorithms; and for that the brain provides inspiration. There is evidence that the brain's computing capability stems from constantly performing Bayesian inference, where the brain changes its internal representation to build up a prior of past experiences, against which it compares new input [Ber+11]. One such brain-inspired neural network architecture capable of this type of probabilistic inference is the *Boltzmann machine* (BM), with origins in statistical physics, where a network of connected binary random variables may be trained to sample from an underlying state distribution (the prior) [AHS85]. The resulting BM is then capable of performing simultaneous generation, classification and pattern completion tasks [SH09] [Esl+14].

Using the biologically plausible leaky integrate-and-fire (LIF) neuron model, networks of spiking LIF neurons may also be trained to approximate sampling from an underlying Boltzmann distribution, similarly to a BM driven by a suitable sampling mechanism. Such networks have been used extensively in simulation for a wide range of machine learning tasks [Dol+18] [Pet+16] [Sch+17] [Len+18] as well as the modelling of physical systems [Bau16].

Networks of LIF neurons are implemented on the BrainScaleS-1 neuromorphic mixed-signal hardware, where individual neurons are physically emulated by a dedicated capacitor and transistor circuit. This physical emulation benefits from a $10^4\times$ acceleration with respect to biological timescales [Sch+10]. Motivated to harness this hardware speedup, LIF network sampling on hardware was first successfully implemented on the BrainScaleS-1 system in [Kun16], where a small-scale LIF network was trained to sample from an arbitrary BM. LIF networks on hardware have since been used to perform accelerated Bayesian inference [Kun+18].

When such networks have been successfully trained to a target BM, the problem arises, both in simulation and on the hardware, that regions of high probability states may be separated by low

probability "energy barriers", and so hinder the ability of the network to explore the state space which it is to represent; the network gets trapped in sub-optimal local minima. The ability of the network to switch between high probability modes, and surpass said energy barriers, is known as *mixing*. Since there is growing evidence that decision making in the brain is also sampling based [Fis+10], the brain may be overcoming this issue with oscillations in neural activity, where the network is less constrained by the energy/probability landscape during periods of high activity. Among other solutions to mediate mixing in BMs [Des+] [Sal10], we consider here effecting a levelling of the probability landscape by increasing the temperature in the abstract BM, *tempering*, by increasing the strength of noise input to neurons within a LIF sampling network.

This was performed in simulation, and found indeed to improve sampling performance [Kor17] for tempering effected by modulating the rate of noise to sampling neurons. A definite relation between temperature and both the rate *and* weight of input noise was also found [Bau16].

The main aim of this project is thus to implement this tempering on the BrainScaleS neuromorphic mixed-signal physical model system [Sch+10]. Since the hardware is fundamentally a physical emulation, there are multiple hardware-specific distortive mechanisms and limitations to be considered, and so an intermediate aim is to probe the hardware-imposed limits of spike based tempering.

We will find that for a given input rate, there is limit to how high the weight of the noise can be set before saturation effects dominate, which we then use as a guide when maximising the noise strength. We find, contrary to theory and results found in simulation, that activation function widening and thus tempering via noise rate modulation is not successful. We do find however, that widening and thus tempering via modulation of noise *weight* is successful, and we achieve a doubling of temperature in the BM regime, a flattening of the state probability distribution, and a facilitation of mixing in a two-mode sampling network. Despite tempering not being achieved with noise rate modulation, it will nonetheless also be found to facilitate mixing.

Due to a limitation of the bandwidth of external input to the chip, and since neurons in sampling networks conventionally[1] each require a private source of externally generated noise, a separate network of inhibitorily connected neurons is often used as a source of this stochasticity [Jor+15]. Here we also aim to improve upon their design, by allowing the output network noise frequency to be modulated, for eventual use as a noise source in large scale tempering experiments.

[1] Though there are works where neural computation has been performed without an explicit source of stochasticity [Jor+17]

Chapter 2

Theoretical background

A brief theoretical overview is here given, so that the rather abstract statement that "LIF networks sample from an arbitrary Boltzmann machine" may be better understood, and also that an intuition may be given for the subsequent mixing problem that arises, and its tempering-based proposed solution. For a more rigorous mathematical treatment of this chain of logic, see [Pet15]. In this section I present the fundamental single neuron dynamics, and then the most important derived results of consequence to this work.

2.1 Single neuron dynamics

In order to capture the relevant dynamics of networks of spiking neurons in the brain, each neuron is formulated as an abstract ordinary differential equation (ODE), each with a time varying membrane potential u, and has the ability to spike. Although the following neuron models are biologically inspired, their biological plausibility is not discussed here, and the reader may be directed to [Pet15], [D+03], [GSD12] for such discussion. Neurons are connected to other neurons via *synapses*, which may be uni or bidirectional, and have an associated *weight*. If a neuron spikes, then the spike signal is transported by a synapse *from* the *presynaptic* spiking neuron *to* the *postsynaptic* neuron. The arrival of a spike at the postsynaptic neuron may then cause a change in the postsynaptic neuron's potential u. The change in a neuron's potential as a result of an incoming spike from one of its presynaptic partners is known as a *post synaptic potential* (PSP). Synapses are split into two groups: an *excitatory* synapse causes the membrane potential u of its postsynaptic neuron to increase upon arrival of a spike, and an *inhibitory* synapse causes u to decrease upon arrival of a spike. In the absence of synaptic input, a neuron's potential falls to a constant rest value. When the potential u of a neuron is high enough, the neuron *spikes*, and transmits this spike to all of its postsynaptic partners. For a finite time period after spiking τ_{ref}, the *refractory period*, the neuron enters a *refractory* state, where its potential is clamped to a lower potential V_{reset} and the neuron may not spike. These dynamics are captured by the leaky integrate-and-fire (LIF) neuron model [BV07], which I first present. The single neuron model will naturally lead to the emergence of a simplified behaviour in a so called high conductance state, where the membrane potential will become a linear transformation of its synaptic input, a necessary condition for sampling.

2.1.1 The leaky integrate-and-fire neuron model with conductance based synapses

The general LIF neuron model is given by

$$C_m \frac{\mathrm{d}u}{\mathrm{d}t} = g_l(E_l - u) + I^{syn} + I^{ext} \tag{2.1}$$

where C_m is the capacitance of the neuron's membrane, g_l the leak conductance, E_l the leak/rest potential, I^{syn} a the sum of the currents due to synaptic interactions from other neurons, and I^{ext} an externally applied current. The "leaky" part of the LIF equation is that in the absence of synaptic or external interaction ($I^{syn} = I^{ext} = 0$), the membrane potential u decays exponentially to the leak potential E_l with a characteristic time constant $\tau_m = \frac{C_m}{g_l}$, the *membrane time constant*. The leak potential E_l is thus often referred to as the *rest* potential. I_{ext} is omitted in all further equations, as no external current is applied throughout this work. The synaptic current I^{syn} is specified by the choice of synapse model, where here we use the conductance-based neuron model (COBA), as it is this model which is implemented on the BrainScaleS-1 system (Section 3).

In the COBA model, the effect of incoming spikes is modelled by an increase in conductance towards one of two reversal potentials, E_e^{rev} and E_i^{rev}, for excitatory and inhibitory synapses respectively. They are set above and below the rest potential E_l respectively. The synaptic current I^{syn} for a COBA neuron is thus given by

$$I^{syn} = g_e^{syn}(t)(E_e^{rev} - u) + g_i^{syn}(t)(E_i^{rev} - u) \tag{2.2}$$

where $g_e^{syn}(t)$ and $g_i^{syn}(t)$ are conductances which linearly sum up incoming spikes, for excitatory and inhibitory synapses respectively. Since a spike occurs at a single discrete point in time, it must first be convolved with an appropriate continuous and finite valued function, a *synaptic interaction kernel*. The form of the synaptic kernel is in general a modeling decision; on the BrainScaleS-1 system a decaying exponential is implemented with characteristic time constants τ_e^{syn} and τ_i^{syn} for excitatory and inhibitory synaptic interactions respectively. However, for the sake of simplicity, here we take $\tau_e^{syn} = \tau_i^{syn} = \tau^{syn}$. The synaptic conductances for a neuron are thus given by

$$g_x^{syn}(t) = \sum_{\text{syn } k} \sum_{\text{spk } s} w_k \Theta(t - t_s) \exp\left(-\frac{t - t_s}{\tau^{syn}}\right), \quad x \in \{e, i\} \tag{2.3}$$

where the first summation is over all synapses of type x, the second over all spikes arriving at said synapse, with spike times t_s, w_k is the weight associated with a given synapse, and Θ is the Heaviside step function. An incoming spike thus causes an exponentially decaying jump in the conductance towards one of the reversal potentials, which in turn enacts a change in u towards one of these reversal potentials $E_{e/i}^{rev}$. When the potential u passes a specified threshold V_{thresh}, the neuron spikes. The potential is then clamped to a sub-threshold reset value V_{reset} for a refractory time τ_{ref}, and is said to be in a refractory state.

2.1.2 High conductance state

When the neuron is subject to sufficiently high synaptic input, the neuron enters a so-called high conductance state (HCS), which is required for sampling [Pet+15]. It is useful to define the total conductance g^{tot} as

$$g^{\text{tot}} = g_l + \sum_k g_k^{\text{syn}} \tag{2.4}$$

where the summation is over all synapses. The distinction between whether the summation is over all synapses, each with a separate associated g_k, or whether it is a summation over $\{e, i\}$ and g_k is then the total conductance from all synapses of that synapse type, need not be made, since at every step the individual convoluted conductances from incoming spikes are summed linearly. The HCS is then defined as when there is sufficiently high synaptic input that the total conductance is dominated by the synaptic conductance, i.e. $\sum_k g_k^{\text{syn}} >> g_l$. By dividing the full COBA-LIF equation (Equations 2.1 and 2.2) by g^{tot}, the COBA-LIF equation may be reformulated as

$$\tau_{\text{eff}} \frac{\mathrm{d}u}{\mathrm{d}t} = u_{\text{eff}} - u \tag{2.5}$$

with the *effective membrane time constant* τ_{eff} defined by

$$\tau_{\text{eff}} = \frac{C_m}{g^{\text{tot}}} \tag{2.6}$$

and the *effective membrane potential* u_{eff} defined by

$$u_{\text{eff}} = \frac{g_l E_l + \sum_k g_k^{\text{syn}} E_k^{\text{rev}}}{g^{\text{tot}}} \tag{2.7}$$

If we assume that we are sufficiently far into the HCS, which is the case in sampling where neurons are bombarded with high frequency excitatory and inhibitory noise, then we assume that the expected value of g^{tot} approaches infinity, i.e. $\langle g^{\text{tot}} \rangle \to \infty$. As a direct consequence of this, the effective time constant τ_{eff} approaches 0, i.e. $\langle \tau_{\text{eff}} \rangle \to 0$. From Equation 2.5, this means that the potential u at all times (when not refractory) decays almost instantly to the effective membrane potential u_{eff}, and so effectively follows it instantly. Under the assumption of a perfect HCS, the effect of a single spike upon the total conductance may be treated as small and perturbative, and allows the expression for u_{eff} to be greatly simplified to

$$u_{\text{eff}}(t) = u_{\text{eff}}^0 + \frac{\sum_k g_k^{\text{syn}}(t) \Big(E_k^{\text{rev}} - \langle u_{\text{eff}} \rangle \Big)}{\langle g^{\text{tot}} \rangle} \tag{2.8}$$

where u_{eff}^0 is a constant offset in the effective membrane potential, as a result of our perturbative treatment. The effective membrane potential u_{eff}, and due to a very fast τ_{eff} also the "true"

membrane potential u, is thus a linear transformation of its convolved synaptic inputs, which will be of gross importance when considering sampling.

2.1.3 Activation functions

When the neuron is bombarded by high frequency Poisson noise, it can be shown that the effective membrane potential u_{eff} follows an Ornstein-Uhlenbeck (OU) [Gar09] process, a random walk. Since we assume that we are in a perfect HCS, this means that until u_{eff} passes the spiking threshold V_{thresh}, the membrane potential u below the spiking threshold also follows a random walk. If u_{eff} passes V_{thresh}, then the neuron spikes and is clamped to V_{reset} for time τ_{ref} in a refractory state. After time τ_{ref}, u then decays instantly to u_{eff}, possibly spiking again immediately. A simulated trace for u performing this random walk and spiking process is shown in Figure 2.1. That up until spiking, u follows a random walk process, means that whether the neuron spikes or not is an inherently stochastic process, with an associated probability of spiking and thus being in a refractory state. Furthermore, this means that in the absence of spiking behaviour, u has a Gaussian PDF with an associated mean μ, and variance σ^2 given by

$$\sigma^2 = \text{Var}[u] = \frac{\sum_k \nu_k \Big[w_k \big(E_k^{\text{rev}} - \mu \big) \Big]^2 \tau^{\text{syn}}}{\langle g^{\text{tot}} \rangle^2} \tag{2.9}$$

where ν_k is the frequency of Poisson noise to a synapse and w_k the weight of the synapse. A useful (but incomplete) geometric interpretation of the spiking probability is then that it may be proportional to the probability mass of u above the spiking threshold. If the neuron is then biased towards higher potentials, such that the mean membrane potential is increased, then the spiking probability will also increase. Again, this may be thought of as increasing the probability mass beyond V_{thresh}. This gives rise to the *activation function* of a neuron, where the spiking probability is found for different mean membrane potentials, or in practise, as the neuron is subject to varying bias (Figure 2.3). It can be shown that in the HCS, activation functions strongly resemble logistic (sigmoid) functions, which is of fundamental importance for sampling [Pet+15].

Continuing with the geometric interpretation of the activation function[1], then if its width is increased (as achievable initially through an increase in Poisson noise rate or weight as per Equation 2.9), then for an equivalent shift in the mean, there will be a more gradual change in the probability mass beyond V_{thresh} than if the Gaussian were slimmer. Thus the effect of a change in the mean potential upon the spiking probability, in practise brought about by the PSPs of other spiking neurons within a sampling network, will be lessened. This mechanism serves as the basis of implementing spike based tempering.

From Equation 2.9 however, the width of the membrane potential distribution increases only initially with increasing Poisson noise rate ν_k, before peaking and subsequently dropping, as seen in Figure 2.2, due to the fact that the squared g^{tot} implicitly also contains ν_k. There is no such limit on widening by increasing the weights. However, despite the fact that the membrane potential thins again with increasing rate, and thus from our geometric interpretation of the

[1]Since there does not exist an easy expression for the explicit relation between ν,w and alpha.

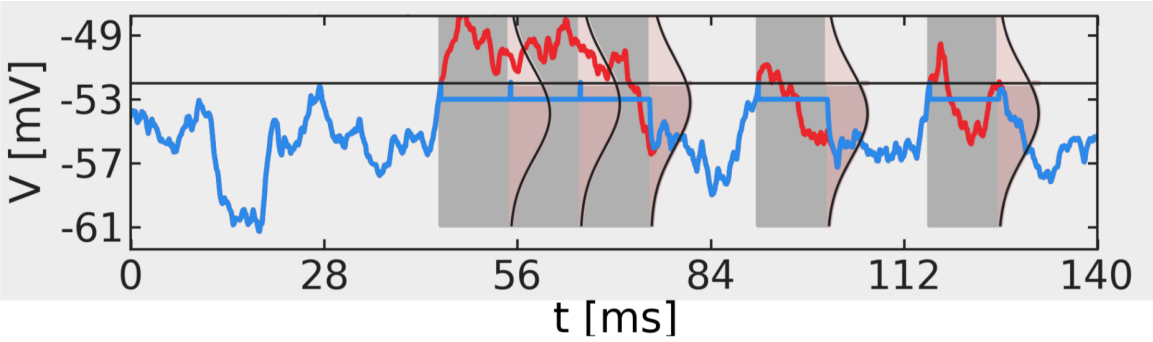

Figure 2.1: Membrane potential u (blue line) of a neuron in simulation subject to Poisson bombardment, leading to the effective membrane potential u_{eff} (blue and red lines) to perform an OU random walk. The membrane potential u follows u_{eff} until the spiking threshold V_{thresh} (black horizontal line) is reached, at which point the neuron spikes and is refractory for τ_{ref}. The pink Gaussians show the PDF of u_{eff}, and thus also u if spiking behaviour were removed. Figure adapted from [Pet15].

activation function we would expect the activation functions to thin also, this was found to not be the case. The activation function width α was instead found to obey

$$\alpha \propto \sqrt{w^2 \nu} \tag{2.10}$$

even when ν was much larger than the frequency at which σ^2 peaks in Figure 2.2 [Bau16]. Thus a widening of the membrane potential is deemed not necessary to achieve widening of the activation function. This also exemplifies the limits of the geometric interpretation of the relation between activation functions and the Gaussian distribution of the membrane potential.

2.2 Sampling theory

2.2.1 Boltzmann machines

A Boltzmann machine (BM) is a network of stochastic binary units with symmetric connections [HSA84]. The state of the network may thus be represented by a column vector $\boldsymbol{z}$, where $z_k \in \{0, 1\}$ represents the state of the k-th unit in the network, where 0 and 1 correspond to being in the *off* or *on* state respectively. For every network state $\boldsymbol{z}$, there is an associated energy $E(\boldsymbol{z})$ given by

$$E(\boldsymbol{z}) = -\frac{1}{2}\boldsymbol{z}^T \boldsymbol{W} \boldsymbol{z} - \boldsymbol{z}^T \boldsymbol{b} \tag{2.11}$$

where $\boldsymbol{W}$ is the weight matrix of the connections between units, is symmetric with zeroes along the diagonal, and $\boldsymbol{b}$ a generic bias vector. If each possible network state is then treated as a single microstate in a canonical ensemble, then the probability of a state is given by

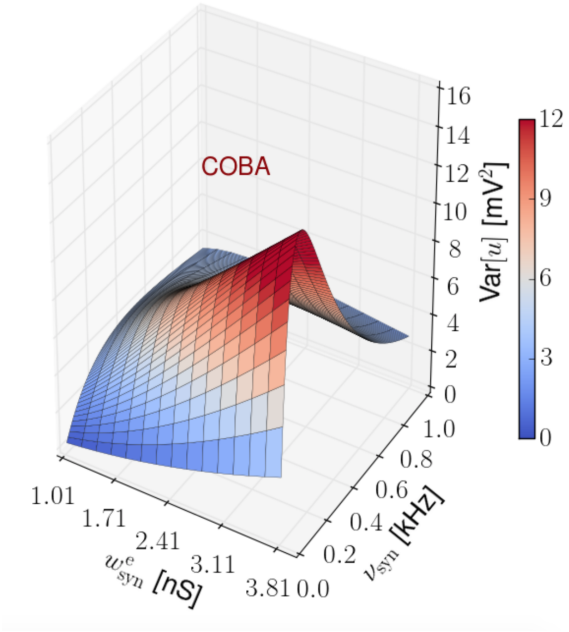

Figure 2.2: The variance of the membrane potential distribution for COBA-LIF neurons in simulation, for varying rate and weight of Poisson noise input. In agreement with Equation 2.9, there is an initial increase in the variance with increasing rate, however the variances peaks and drops off again. Figure adapted from [Pet15]. Despite this lack of membrane potential widening with increasing rate, the activation function width α was found to increase with increasing input rate well beyond the peak seen here [Bau16].

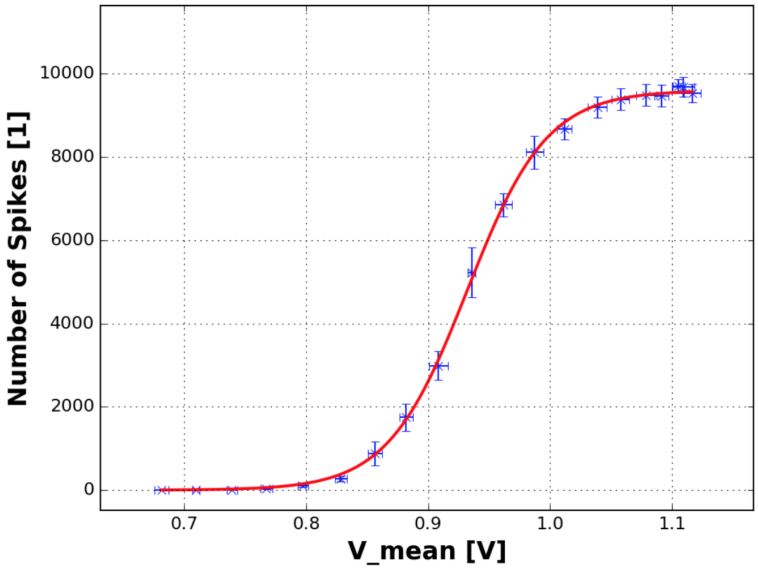

Figure 2.3: Activation function of a neuron on the HICANNv4 chip. The number of spikes produced is a proxy for the spiking rate, and so also the spiking probability. Figure adapted from [Kun16].

$$p(\boldsymbol{z}) = \frac{1}{Z}\exp\{-E(\boldsymbol{z})\} = \frac{1}{Z}\exp\left\{\frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{b}\right\} \tag{2.12}$$

where Z is the partition function given by $\sum_{\boldsymbol{z}}\exp\{-E(\boldsymbol{z})\}$. If from this we now consider the conditional probability of the k-th unit being in an on state, given the state of the rest of the network $\boldsymbol{z}_{\setminus k}$, we get

$$p(z_k = 1 \mid \boldsymbol{z}_{\setminus k}) = \frac{1}{1+\exp\{-u_k\}} = S(u_k) \tag{2.13}$$

where u_k is a weighted linear transformation of the current state of the rest of the network, as well as a fixed bias, and $S()$ is the logistic (sigmoid) function. The u_k, which we will refer to as a "potential" is given by

$$u_k = b_k + \sum_{i \neq k} W_{ik} z_i \tag{2.14}$$

and so is solely responsible for determining the conditional probability of unit k being in an on state. If this conditional probability is re-purposed as an update rule, such that all units within the network are sequentially set to an on or off state, with probabilities determined by the conditional probability, then the evolution of the network state $\boldsymbol{z}$ forms a *Markov chain Monte Carlo* (MCMC) sampler. This particular case of sequentially updating the network state using the conditional probabilities is known as *Gibbs sampling* [GG87], and it can be shown that for an arbitrarily large number of update steps, the distribution of observed network states $\boldsymbol{z}$ converges on the exact target distribution of the BM given by Equation 2.12.

2.2.2 LIF Networks implement Gibbs sampling

A general overview of the link between BMs and LIF networks is given here in order to justify the statement that LIF networks sample from BMs, however for a more precise mathematical treatment, especially in the transition from discrete time-step updates to continuous updating, the reader is again directed to [Pet15].

Since every neuron in a LIF network may be in either a refractory state or not, each neuron may be treated as a two state binary system, with 1 being that the neuron is refractory, and 0 not. As was seen in Section 2.1.3, neurons in the 0 state may transition to the 1 state according to the spiking probability given by the neuron's activation function. Since the activation functions in LIF networks are found to be logistic in shape, they thus bear direct similarity to the logistic update rule found in the BM regime.

The input field for logistic activation function is the mean membrane potential u_{eff}, which was found in the HCS to be a linear sum of its synaptic input 2.8. If the synaptic time constants τ^{syn} are set such that $\tau_e^{syn} = \tau_i^{syn} = \tau_{ref}$, then the membrane potential u_{eff} is affected by a spiking neuron for approximately as long as the neuron is refractory. Thus, u_{eff} becomes a linear transformation of the state of the rest of the network. This bears striking resemblance to logistic update rule's input field variable u_k in the BM regime, which per Equation 2.14 is

also a linear transformation of the state of the rest of the network $\boldsymbol{z}_{\setminus k}$. If the neurons within a LIF network are fully connected, the weights are made symmetric, and self-connections are disallowed, then the weight matrix in the BM regime $\boldsymbol{W}$ directly mirrors the synaptic weights in the LIF regime[2].

The largest dissimilarity between the two regimes however, is that in the BM regime, it is implicitly assumed that spikes are convoluted with a rectangular synaptic interaction kernel of exact length τ_{ref}, since per Equation 2.14, the abstract potential u_k is uniformly affected for as long as the presynaptic-equivalent unit is refractory/on [Bue+11]. This is in contrast to the LIF regime, where neurons interact approximately via decaying exponential PSPs.

Due to the strong similarities discussed above and despite the mismatch in PSP shapes, under suitable conditions neurons in a LIF network implement Gibbs sampling, and thus the observed network states $\boldsymbol{z}$ sample from an arbitrary BM [Bue+11]. This result serves as the basis for all sampling experiments. LIF sampling networks were first implemented on the HICANNv4 chip in [Kun16], where it was found that a minimum of 300Hz excitatory and inhibitory Poisson noise was required in order to achieve the HCS required for sampling.

2.2.3 Training and the mixing problem

Since no training of LIF networks is performed within this work, a thorough detailed procedure of the training algorithms is not given. It is however to useful to appreciate the general result of training, such that the mixing problem central to this work may be understood. The reader is directed to [AHS85] and [Bre15] for a detailed procedure of the implementation of such training algorithms.

The training algorithm most used in related works here is *Contrastive Divergence* (CD). By observing the correlations between spiking neurons $\langle z_i z_j \rangle$ and individual neuron spiking probabilities $\langle z_i \rangle$, the weights and biases respectively are updated to move the LIF network state distribution towards a target distribution defined by a target BM. In this way, the spiking LIF network gains an internal representation of the data/task for which it is trained.

Using the example of a classification problem, after training the LIF network will have converged on a target BM, where states corresponding to the input training data/classifications are given a high probability. As a corollary to this, states which do not represent the trained data-set are assigned a low probability. These states thus have respectively low and high associated energies $E(\boldsymbol{z})$ in the BM regime. As a result of successful training, these energy differences will be maximised, and so the state energy/probability landscape will become increasingly inhomogeneous. Though this would not be a problem if states were sampled from this state distribution $p(\boldsymbol{z})$ directly (Equation 2.12), we know that the evolution of the network state $\boldsymbol{z}$ instead implements Gibbs sampling (and indeed most useful sampling algorithms use some kind of MCMC and have the mixing problem if the energy landscape has deep wells). In order to move from one high probability state to another, as is required to accurately represent the underlying distribution, the network will need to move through the aforementioned low probability transition states. These states thus pose an effective energy barrier, and mean that

[2]Though the translation from abstract weights to synaptic weights is non-trivial

the network state $\boldsymbol{z}$ may be confined[3] to a sub-optimal local minimum in energy, corresponding to a poor classification. The movement of the network state $\boldsymbol{z}$ past these high energy barriers so that it may explore the network more freely is known as *mixing*. One solution to this problem is to effect an increase in temperature in the corresponding BM, such that energy landscape is flattened, and is known as *tempering*. This was achieved for *current based* LIF (CUBA-LIF) neurons in simulation in [Kor17].

2.2.4 Temperature modulation

Until now, the notion of temperature in a BM has been omitted, since the temperature scale may be defined arbitrarily. If it is re-added, we get an altered probability distribution given by

$$p(\boldsymbol{z}) = \frac{1}{Z}\exp\left\{-\frac{E(\boldsymbol{z})}{T}\right\} = \frac{1}{Z}\exp\left\{\frac{\frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{b}}{T}\right\} \tag{2.15}$$

and an altered update rule given by

$$p(z_k = 1 \mid \boldsymbol{z}_{\setminus k}) = \frac{1}{1+\exp\{-\frac{u_k}{T}\}} = S(\frac{u_k}{T}) \tag{2.16}$$

where T is the temperature, and the partition function Z has also been appropriately amended. From these it can be seen that if the temperature T were defined to be different from 1, the energies E and potentials u_k, and thus weights and biases $\boldsymbol{W}$ and $\boldsymbol{b}$ could simply be scaled such that there the distribution $p(\boldsymbol{z})$ is invariant. It is thus only useful to consider the notion of temperature when considering a temperature *change* while the weights and biases are kept fixed. Thus an increase in temperature leads to a flattening of the energy landscape, and effectively manifests as a scaling down of all weights and biases in the BM regime. Due to the equivalence between the update rule in the BM regime and the activation function in the LIF regime, an increase in T in the BM regime is synonymous with an increase in the activation function width α in the LIF regime[4]. Due to the linearity of the potential in both regimes, this is in keeping with the idea that a temperature increase is equivalent to a linear scaling down of all weights and biases, as a scaling down of the input field of a function (the logistic function) is exactly equivalent to a widening of said function. This temperature modulation was performed for CUBA-LIF neurons in simulation in [Kor17], and by varying the rate of Poisson noise input to sampling neurons the relationship

$$\alpha \propto T \propto \sigma \propto \sqrt{\nu} \tag{2.17}$$

was found, where α is the width of activation functions in the LIF regime, T the temperature in the abstract BM regime, σ the standard deviation/width of the individual neuron membrane potential distributions, and ν the rate of Poisson noise input to sampling neurons [Kor17]. This

[3]Or rather, it may take a very long time before it escapes, and so it requires a very long time for the LIF network distribution to resemble the BM distribution $p(\boldsymbol{z})$ and thus implement effective sampling as desired.

[4]It is also only useful to define α with respect to it changing, and so the unit α, like T is also defined arbitrarily.

was also tested for COBA-LIF neurons, which are the focus of this work, where it is known from Figure 2.2 that σ does not continue to increase with increasing ν. The noise rate *and* weight was varied, and the relationship

$$\alpha \propto T \propto \sqrt{w^2 \nu} \tag{2.18}$$

was found, where w is the weight of Poisson noise input to a sampling neuron [Bau16]. These results are all from simulation, and they serve as the theoretical basis to implement a temperature change on hardware.

2.2.5 Noise networks

Sampling theory assumes that each neuron has a source of perfectly uncorrelated, private Poisson noise, such that each neuron's OU random walk is also uncorrelated. Due to input bandwidth constraints, there is thus an upper limit on the number of neurons for which Poisson noise can be externally privately generated. If, in order to circumvent this limitation, the privately generated noise is replaced by noise shared between multiple neurons, the resulting shared-noise correlations lead to a significant reduction in sampling performance. It was found that sampling performance could be regained, if the shared Poisson noise was instead replaced by noise produced from a dedicated network of inhibitorily connected neurons, *noise neurons*. The noise neurons are set up with $V_{thresh} < E_l$ to ensure constant spiking, and by exploiting the decorrelating effect of inhibitory noise, act as a source of slightly anti-correlated noise, which counteracts the positive correlation brought about by sampling neurons sharing noise neurons [Jor+15]. Any slightly remaining correlation in the noise may be accounted for during training by an appropriate sampling weight change, since from a training perspective, there is no difference between a correlation brought about due to shared noise sources and a correlation due to a causal link between the neurons [Dol+18]. However, it is sill favourable to rid of unwanted correlations as much as possible, and to produce noise which bears as much similarity to Poisson generated noise as possible. We here wish to further this paradigm, by altering the noise network such that it may have a modulatable output frequency, so that it may be a source of noise for large scale tempering.

Chapter 3

Hardware details

PyNN [Dav+10] is a high-level simulator-agnostic modelling language which may be used to specify networks of spiking neurons. The NEST [GD07] simulator backend was used for the initial simulations of the noise network, and then all subsequent experiments were run using the Heidelberg BrainScaleS-1 [Sch+10] physical hardware emulation backend with the HICANNv4 chip, since tempering has already been extensively shown to work in simulation [Kor17] [Bau16]. When we refer to a given experiment being run "on hardware" (as opposed to in simulation), it is this physical emulation to which we are referring. For further details on the hardware implementation, see [Kun16] and [Sch+10].

The BrainScaleS system is physical neuron emulation system, with each wafer being composed of 384 *High Input Count Analog Neural Network* chips (HICANNs), each of which is capable of the analogue emulation of 512 COBA-LIF[1] neurons, where each neuron is essentially emulated by a discrete capacitor and accompanying neuron circuitry. The HICANNv4 chip is used here. Each wafer may simultaneously emulate up to 196,608 neurons, with over 44 million synapses. As a result of this physical emulation, the hardware emulates neurons with a 10^4 speedup over their biological counterparts. There is thus a distinction in time scales when dealing with physical emulation on the hardware, whether the times refer to hardware run times or biological times. Throughout this work, all times are given as the biological time. Similarly, any parameters are given in biological units $(mV, \mu S, ms)$, which are then scaled to appropriate physical values on the hardware. The exception to this is where the explicitly digital hardware parameters are manually set for a greater degree of control, and the parameters are often unitless. When transitioning from digital simulation to analogue emulation on the hardware, there are some detrimental effects to be avoided or handled as necessary. The main hardware-specific distortive effects considered within this work are listed.

Fixed pattern noise

Due to the imperfect manufacturing of the wafer, there is a fixed mismatch in the transistors governing individual neuron behaviour. Since this mismatch does not vary over time, it is known as a *fixed-pattern noise*. Though these fixed differences form neuron to neuron may

[1]The neuron model used is actually the adaptive exponential integrate-and-fire neuron model [BG05], but the adaptive behaviour may be switched off to effectively implement COBA-LIF neurons

be calibrated away, by sweeping the settable digital hardware parameters and measuring the realised biological parameters (e.g. E_l, τ_{ref}) for each neuron individually [Sch13], there is still a degree of systematic variation from neuron to neuron. Also, not all parameters utilise this individual calibration.

Floating gate variations

The digital parameters which are then sent to the individual neuron circuits at runtime are stored on so called *floating gates* (FG), which are analogue units responsible for storing the neuron parameters. Due to a limited precision of the setting of the FG voltages, the realised voltage (and thus neuron characteristics) vary from trial to trial. To avoid going into to much detail about the neuron circuit at a transistor level, we will simply treat this effect as an additional noise upon setting the neuron parameters, which vary from trial to trial [Pet+14]. This effect is know as *floating gate variation*.

Spike loss

Spikes are recorded by a digital spike recorder on a per reticle (grouping of 8 HICANNs) basis. If the local neuron spiking rate is too high, then there is the possibility for spike events to go unrecorded [Klä17]. We later argue that the spikes are lost at a much greater rate upon readout to the host computer than they are between neurons on chip.

Synapse loss

Since there are a fixed number of synapses available on the hardware, and indeed not enough that every neuron may be connected to every other, beyond a certain number of synapses, requested connections must fail to be realised on the hardware. This effect is known as *synapse loss*, and in actuality begins to occur well before the maximum number of synapses is reached, due to the sparsity in available routing resources and the complexity of the mapping algorithm. [Pet+14].

OTA saturation

The synaptic interaction current is governed by a set of *Operational transconductance amplifiers* (OTAs), which essentially ideally provide a current which is proportional to the linear sum of all convolved input spikes with their associated synaptic weight, as given by g^{syn} in the theory. However, since the OTAs are physical components and so have associated physical limits, they deviate from theory by having a limit on the total output current. The reaching of this limit, where the OTA may not output any more current despite receiving further input spikes is know as *OTA saturation*, after which point the hardware will fail to accurately emulate a COBA-LIF neuron [Mil12]. This is of gross importance in sampling, where it is fundamental that the synaptic conductance g^{syn} is linear in the HCS.

Chapter 4

Experiments

4.1 Creating a modulated noise network

Since there is limited bandwidth available for inputting external spikes to neurons on the hardware [Kar14], a network of inhibitorily connected neurons is placed on the wafer to serve as a (slightly anticorrelated) noise source, as has been found to work as a functional replacement for each neuron having a private source of externally generated Poisson noise [Jor+15]. Since we wish later to vary the input noise rate to sampling neurons to implement tempering, we thus wish to alter this noise network so that its noise output frequency may be modulated. This modulation will be effected by subjecting the neurons within the noise network ("noise neurons") to an excitatory bias, such that they are biased to spike faster.

4.1.1 Initial considerations

For sampling on the hardware, a minimum frequency of 300Hz excitatory and inhibitory input is required [Kun16]. To avoid OTA saturation (discussed in Section 4.3.1), this 300Hz baseline frequency must actually be split among at least 8 sources. Furthermore, it is highly desirable for the frequency of each of these sources to be similar (again, to avoid OTA saturation as will be seen). We aim here for an achievable noise frequency range of 300Hz $\rightarrow$ 1200Hz, such that as per Equation 2.18, a 2x widening in activation function widths should be achievable. To this end, some general statements about the noise network's parameters can be made:

On controlling the spiking rate range

In order to be achieve a 4x increase in spiking rate, the inter-spike-interval (ISI) of the neurons must be reduced by a factor of 4. The ISI of spiking neuron is in general composed of two time periods: the fixed refractory time τ_{ref} and an arbitrary dynamic rise time towards the spiking threshold t_{arb}. The addition of excitatory stimuli reduces the dynamics time t_{arb}, but since τ_{ref} cannot be dynamically changed, there is thus a hard limit on the possible frequency

speedup of $\frac{ISI_{slow}}{ISI_{fast}} = \frac{t_{arb}+\tau_{ref}}{\tau_{ref}}$. Thus τ_{ref} is initially set to 0, to rid of this limit[1]. This then poses the issue that, without a refractory time as a hard limit, the neurons may now spike too quickly, as $\frac{300}{8}$Hz is the maximum baseline spike rate desired from the network. In order to replace the delay which the refractory time would have provided, and indeed with delays that may be overridden by external excitatory input, many of the other neuron parameters are thus determined: a large membrane time constant τ_m, a large membrane capacitance C_m (limited to two values on the hardware), and a large potential gap $V_{thresh} - V_{reset}$ to be traversed.

On ensuring spiking

The neuron rest potential E_l is placed above V_{thresh}, such that in the absence of external inhibition, each neuron will spike at a constant rate. In order to slow down the spike rate as desired above, the exponential decay behaviour of the membrane potential could be exploited by placing the E_l arbitrarily close to but still above V_{thresh}. Though this would work in simulation, due to the inaccuracy of setting the potentials on the hardware [Kok17], the target E_l must be set sufficiently far above V_{thresh} to ensure that all neurons are, without inhibition, in a constant spiking state.

On the synaptic time constants

The mechanism for ensuring that the noise produced by the network is decorrelated is that the noise neurons are inhibitorily connected to each other. The inhibitory synaptic time constant τ_i^{syn} thus plays a large role in the noise network dynamics. One noise neuron spiking should have the effect that its post synaptic noise neurons may not spike for a short time, their spiking is temporarily delayed. We propose here that only a short τ_i^{syn} compared to the mean ISI is required to fulfill this role. A large τ_i^{syn} may have the effect, especially when the network activity is increased, that a neuron's spiking behaviour may be delayed indefinitely, and as such does not ever spike, which is to be avoided.

On the other hand, a long excitatory synaptic time constant τ_e^{syn} is preferred. This is so that the controlling external excitatory bias spikes may better resemble a constant bias current. If a constant frequency excitatory input is considerably time varying (as with a short τ_e^{syn}), then there is the possibility that a spike in the excitatory bias may trigger many noise neurons to fire simultaneously, which is entirely contrary to the aim of producing anti/decorrelated noise. This also implies that the excitatory stimulus weight should be very low. If these criteria are not enforced strongly enough, the produced noise spike trains may become entirely synchronised.

4.1.2 Noise network optimisation

The noise network is configured as shown in Figure 4.1. It is initially run in simulation, in order to better understand the behaviour of the network before implementing it on the hardware, where the hardware-specific distortion mechanisms described in Chapter 3 may pose additional difficulties. Each noise neuron is randomly connected to N_{pre} other noise neurons inhibitorily, at

[1]There are also other hardware imposed limits on the minimum possible ISI, such as the hardware clock speed, however we will find that a minimum in t_{arb} is met before this becomes relevant.

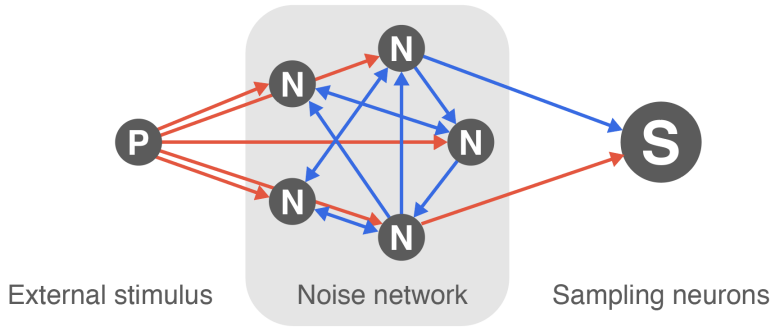

Figure 4.1: Network setup for using a modulated noise network as a noise source during sampling. The P, N, and S circles represent Poisson input sources, Noise neurons, and Sampling neurons respectively. The red and blue arrows show the directed synaptic connections, excitatory and inhibitory respectively. Each noise neuron is randomly connected to N_{pre} other presynaptic noise neurons, at fixed inhibitory weight. In the experiments studying the dynamics of the noise network only, the sampling neurons are omitted.

a fixed weight w_{inh}. Every noise neuron is connected excitatorily to a single source of externally generated Poisson noise spikes, at a weight low enough to avoid any synchronisation effects as previously discussed. The individual neuron parameters are set as discussed in the initial considerations, but are also varied slightly from neuron to neuron, to better reflect the effect of fixed pattern noise and FG variations on neuron parameters when the network is implemented on the hardware.

For each noise neuron, the relevant statistics to characterise its produced noise are the frequency ν and the coefficient of variation (CV) of the ISIs. The CV is used since a perfect Poisson noise source, by definition, would produce spikes with an exponential ISI distribution, which has a CV of 1. Thus, the CV is used as a measure of similarity to the ideal case of Poisson noise, and thus the "health" of the noise, since sampling theory presupposes Poisson generated noise, and it is not *a priori* known what effect this deviation from theory will have. Due to a nonzero rise time, there is a lower limit on the possible ISIs for an individual neuron, whereas no such limit exists for Poisson noise, where a spike may be followed immediately by another without delay. Thus, a CV of 1 for an individual neuron is somewhat unachievable. However, when the spike trains from many neurons are superimposed (as will be done when a sampling neuron is connected to multiple noise neurons), this minimum ISI limit is removed, and so we expect the CV of the resulting composite spike trains to be improved. It must also be noted that a CV of exactly 1 would not necessarily imply that the noise is suitable for sampling, as it does not give any measure of correlation between spike trains.

For arbitrarily chosen initial N_{pre} and w_{inh} values in simulation, the CV/ν plot is shown in Figure 4.2. There is an apparent trade-off between CV and ν, manifesting in a "banana" shape, and this shape is preserved when the external excitatory stimulus is applied. This trade-off was found to not be inherent to the dynamics of the noise network, but rather the result of applying homogeneous w_{inh} values to an inhomogeneous neuron population, as a simple iterative algorithm to alter the individual w_{inh} values to move the neurons towards more favourable CV/nu values was found to be successful, as shown in Figure 4.3. Due to spike loss making this difficult to implement on the hardware (as will be discussed), this technique was not explored or used further, but still represents a possible improvement in future to improve the health of noise produced, and decrease the variance in noise neuron frequency. Despite this, it

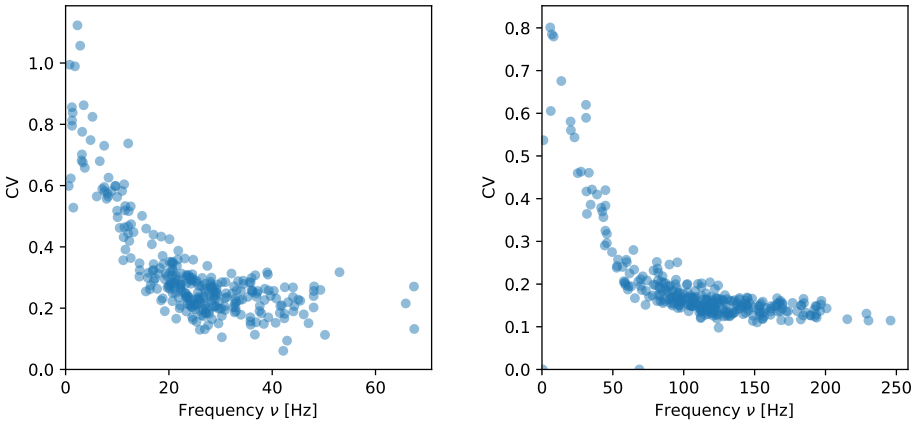

Figure 4.2: CV and frequency of the spikes produced from individual noise neurons in simulation, using initial guesses for N_{pre} and w_{inh} values. **Left** is without any external stimulus, **right** is with 1200Hz excitatory Poisson noise stimulus. The characteristic "banana" shape can be seen here, representing the apparent trade-off between CV and ν. From the left to right, the frequencies have been successfully increased as a result of external excitatory stimulus. A CV closer to 1 indicates the noise better resembles Poisson noise. The CV values also change from left to right, and so present the problem of optimising N_{pre} and w_{inh} to maximise the CV in both cases.

still remains to find values of N_{pre} and w_{inh} to be applied to the entire noise network, for which healthy (Poisson-like) noise is produced in both the stimulated and unstimulated cases. The problem then arises of how to clarify the "health" of the network to be maximised. However, from CV/ν plots with purposefully too much inhibition, some features to be avoided can be identified, as shown in an extreme case in Figure 4.4. There are two undesirable characteristics of note: due to too much inhibition, and particularly by being connected to a hyperactive presynaptic neuron, some neurons have an almost negligible spiking rate, and are denoted as being "silent". On the other hand, some neurons receive too little inhibition (in particular due to relying on the silent neurons for inhibition), and so have a CV close to 0 and a very high frequency, and are denoted as being "hyperactive". Quantifying these effects by looking at the variance of frequencies or CVs did not provide enough differentiation between comparatively healthy noise. Instead, viewing the neuron hyperactivity as a consequence of the existence of silent neurons, the number of silent neurons is chosen as an additional measure of the health of a noise network. The mean CV, mean ν, and number of silent neurons (classified as having a frequency less than 2Hz) were plotted for a sweep of N_{pre} and w_{inh} values for both 0Hz and 1200Hz excitatory stimulus in simulation, and are shown in Figure 4.5. Though no obvious optimal point appears, it is observed that for similarly achieved mean CV values, there are far more silent neurons for high N_{pre} networks. When implementing on hardware, a low N_{pre} ($\approx$ 3) is thus used, and w_{inh} will be increased until adverse effects are encountered.

4.1.3 Spike loss on hardware

Guided by Figure 4.5, the noise network is then run on the hardware, where it is expected for the optimal values to need reconsideration, as the behaviour of the network more often than not changes considerably when moving to the hardware, for reasons such as a lack of calibration on some neuron parameters ($\tau_m, \tau_{e/i}^{syn}$ and other limitations as discussed in Chapter 3).

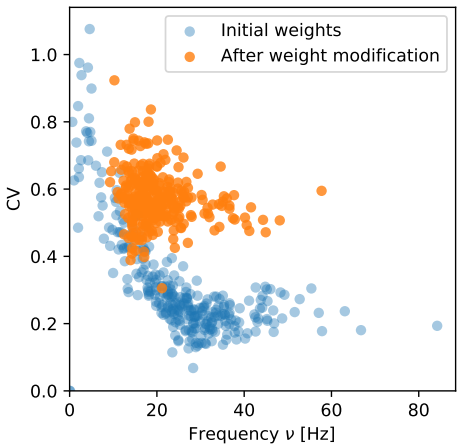

Figure 4.3: Demonstration in simulation that the banana shaped CV/ν trade-off is not inherent to the noise network, and may be overridden by iterative modification of individual w_{inh} values. On each iteration, the noise network was run with its current set of weights, and the CV/ν found for every neuron. Neurons were classified as being "silent", i.e. receiving too much inhibition and having a negligible spiking frequency, or hyperactive, i.e. having a poor CV value and a considerable spiking frequency. The weights from these neurons' presynaptic partners were then decreased or increased respectively. 10 iterations were performed. The CV for almost all neurons improved, and the variance in neuron frequency was reduced. The improvements were retained when Poisson noise was subsequently added. Unfortunately due to spike loss issues on the hardware, this technique was not implemented or explored further.

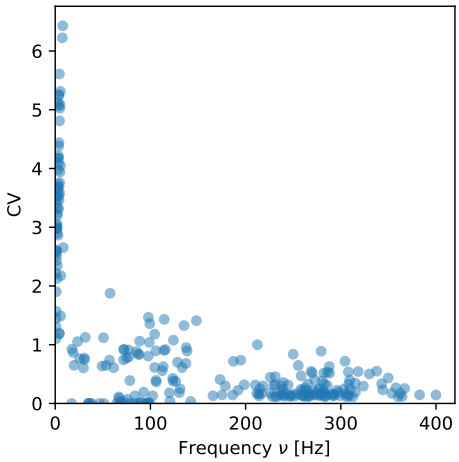

Figure 4.4: Example of a noise network in simulation with highly undesirable features. The number of presynaptic inhibitory partners N_{pre} is very low (2), and the connected weight w_{inh} is relatively high. This results in the noise neurons splitting into two distinct groups: those which receive too much inhibition, barely spike at all and are "silent", and those that consequently receive barely any inhibition and are "hyperactive" with a very high spiking frequency.

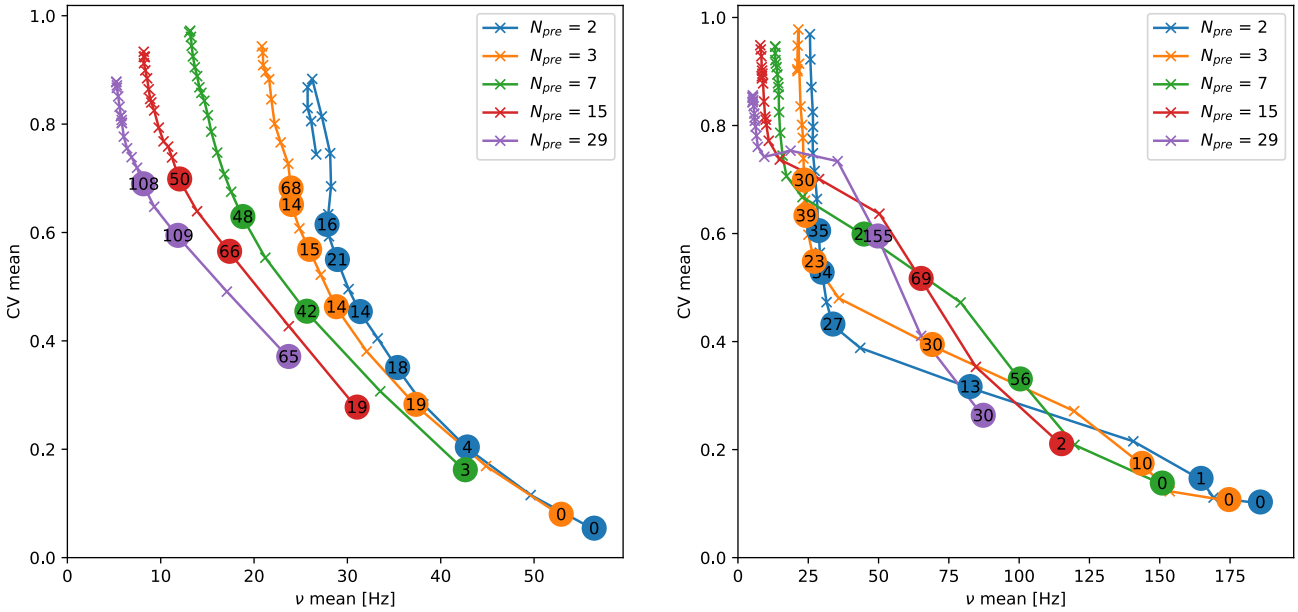

Figure 4.5: Sweeping N_{pre} and w_{inh} values in simulation to find optimal values to use in the noise network. **Left** is the noise network without any external stimulus, and **right** is with 1200Hz of external excitatory Poisson stimulus. The optimal values will be those which produce "healthy" noise in both cases, and which allow the required change in output frequency. Each plotted point (circle or cross) represents a separate run of the noise network with different N_{pre} and w_{inh} values. Each line is a different N_{pre} value as labelled, and each point (circle or cross) along a line is a new w_{inh} value, increasing from the bottom at $0.001\mu S$ up to $0.050\mu S$ at the top in 20 *equally spaced* intervals. For each run, the mean CV and mean frequency ν of all the neurons is plotted, while the number of silent neurons (those with $\nu < 2$Hz) is plotted intermittently. The optimal point would be where the CV is high in both cases, the frequency is considerably increased, and there are few silent neurons. Though there is no obvious optimal point to pick, a few general observations can be made. Increasing N_{pre} and w_{inh} both in general result in an increase of CV and a decrease in ν, as is to be expected from increasing the amount of inhibition within the network. For similar CV values around 0.6, the number of silent neurons is vastly greater for higher N_{pre} values, indicating a large variance in neuron frequency, which is to be avoided. Also, for high w_{inh} values, there is almost no change in neuron frequency between unstimulated and stimulated. This figure and these observations are then used as a guide when implementing the noise network on hardware.

However, when run on the hardware with similar values as in simulation, the CV/ν plots yielded suspicious results, with vastly improved CV values, but simultaneously with a large fraction of silent neurons. A possible cause of these anomalous results was spike loss, and the noise network was altered to test this hypothesis. N_{pre} was set to 0, such that there was no inter-neuron connectivity or inhibitory behaviour. The noise neurons should thus spike each at a constant rate, and so have a single sharp spike in their ISI distributions, and thus have CV ≈ 0. This was not the case, and the neurons appeared to have CVs ≈ 0.5 still. When viewing the individual ISI distributions, there was the expected sharp peak, but followed by subsequent lower peaks at integer multiples of the first peak's ISI, as in Figure 4.6.

These results can easily be explained by spike loss, as the loss of a spike upon readout would lead to an observed ISI consisting of two of the true ISIs, and so would cause an observed ISI peak at twice the true ISI's value. To extend this intuition, we make the IID (independent and identically distributed) assumption with spike loss events, that is that there is a fixed probability γ for any given spike to be lost upon readout. This is a strong assumption, because spike losses are in reality more likely to occur simultaneously, since they occur when the local spiking rate is too high for all spikes to read out. However, this assumption is shown to adequately model most of the effects of spike loss upon the ISI distributions here. If the true ISI distribution is given by $p^*(t)$ and is normalised, then the distribution of ISIs for which exactly one spike is lost $p_1(t)$ may be given by the joint probability of two true ISI distributions, marginalised over all values for which the sum of the individual true ISIs is constant, or

$$p_1(t) = \int_0^t p^*(t-\tau)p^*(\tau)d\tau = \int_{-\infty}^{+\infty} p^*(t-\tau)p^*(\tau)d\tau = p^* \circledast p^*(t) \tag{4.1}$$

where the integral limits may be extended to the infinities since $p^*(t) = 0 \ \forall t < 0$, and $\circledast$ represents a convolution. As a corollary to this, the ISI distribution resulting from 2 consecutive spike losses $p_2(t)$ is given by $p_2(t) = p^* \circledast p^* \circledast p^*(t)$ and so forth. From the IID assumption, the probability of an n-spike loss event is $\gamma^n(1-\gamma)$, and so the observed ISI distribution $p(t)$ may be given by

$$p(t) = (1-\gamma)p^* \circledast \Big[\delta(\tau) + \gamma p^* + \gamma^2 p^* \circledast p^* + \gamma^3 p^* \circledast p^* \circledast p^* + \dots\Big](t) \tag{4.2}$$

where $\delta(\tau)$ is the Dirac delta distribution. In order to test the validity of this treatment of spike loss, the true ISI distributions of the neurons were recorded by sequentially recording the analogue membrane potential trace of individual neurons (due to bandwidth constraints, only 2 may be recorded simultaneously), and inferring spike events therefrom. The true (analogue recorded) ISI distribution may then be compared with that which was digitally recorded (and thus subject to spike loss), as well as the ISI predicted to be digitally observed as per Equation 4.2, and is shown in Figure 4.6. We see that spikes obtained from the analogue read indeed differ from those recorded digitally, and so that spike loss is indeed occurring. The degree of spike loss was larger than expected, as for a network of 100 neurons on one HICANN spiking at $\approx$ 20Hz each, the spike loss rate varies from 0.1% for the 100th neuron to 80% for the 1st neuron[2]. When the spike rate is increased to $\approx$ 80Hz per neuron, the spike loss rate increases to over 99% for the first 30 neurons. As well as affecting the recorded ν values in a linear manner,

[2]The spike loss rate being dependent upon the index of the neuron is to be expected, as the digital spike encoder prioritises neurons in the same order.

the recorded CV values are disproportionately affected even for low spike loss rates, with a 9% spike loss rate artificially inflating the CV by a factor of 8 in Figure 4.6. As such, digital spike recording may not be used to reliably infer the CV/ν values of neurons within the noise network. For this reason, the aforementioned iterative weight modification algorithm shown in Figure 4.3 could not easily be implemented on hardware. The predicted ISI distributions obtained from the convolutional treatment of spike loss showed good agreement with the digitally recorded distributions, and predict artificially inflated CV values of similar values to those recorded, justifying the convolutional treatment and the assumptions made.

Although some algorithms were designed to try to undo the effects of the infinite self convolution upon the true ISI distribution, in order to try to recover $p^*(t)$ from $p(t)$, in most cases they suffered from the fact that the spike loss rate γ had to be known in advance exactly. However, for the simple case of $p(t)$ consisting of consecutive peaks advanced by multiples of the first peak's ISI, a simplistic treatment was found to be adequate: all ISI values from the n-th peak were divided by n, to bring it in line with the first peak. This was useful for situations where the true ISI distribution $p^*(t)$ had a single well defined value, as for the case of neurons spiking at a constant rate. However, this approach could not be applied with the noise network, where we are inherently aiming for a large spread in ISI. Thus when wanting to reliably measure the noise network's activity, the spikes must be inferred from sequentially analogue recording the neurons' potentials. This however has the downside of being slow, as each[3] neuron requires that the network be run again for it to be recorded. More importantly, this means that the inferred spikes from each neuron are not recorded simultaneously, and so we cannot check for unwanted spike train correlations. All subsequent noise network runs on the hardware use this analogue spike reading method.

It must be noted that we are implicitly assuming that the digital spikes are being lost on readout *only*, and are still being communicated to their target neurons. This hypothesis is however supported by the CV/ν plots of the noise networks where spike loss was at its worst. For the case of 1200Hz input noise, the first 30% of neurons had almost 100% spike loss. For our low value of $N_{pre} = 3$, if those spike were also lost between neurons and not just on readout, it would be overwhelmingly likely for at least 1 neuron in the population to have all 3 of its presynaptic partners be from this 30%, and so to receive no inhibition and have a CV of ≈ 0, but no such neurons were observed.

4.1.4 Modulating the noise network on hardware

The true CV/ν plots for the stimulated and unstimulated noise networks on hardware are shown in Figure 4.7. They show that undesirable noise network characteristics have successfully been avoided (silent and hyperactive neurons). It is then to be determined, how many noise neurons a sampling neuron should be connected to. Figure 4.8 shows the CV/ν of the composite spike trains arising from combining the noise from a varying number of noise sources N. A total noise frequency of 300Hz for no network stimulus is desirable, as it is the minimum required frequency for sampling. It is also desirable for N to be a multiple of 8, as the noise input may then be distributed equally among a sampling neuron's OTAs. A higher N is also desirable, as it is met with an increase in the resulting CV, though this is naturally a diminishing effect. Too high an N will however be problematic, since this requires more synapses to be drawn between

[3]Or rather, every other, since 2 neurons may be analogue recorded simultaneously.

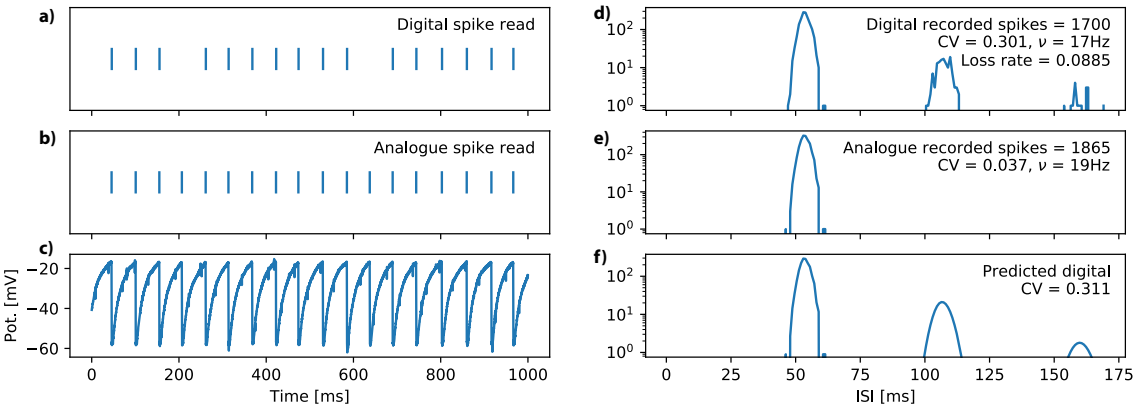

Figure 4.6: Spike loss and its effect upon the observed ISI of a neuron spiking at a constant frequency on hardware. **a)** The spikes received by the digital spike recorder. **b)** The spikes inferred from the analogue membrane trace, thus are loss free. **c)** The analogue membrane potential trace, from which the lossless true spike train and true ISI distribution is inferred. **d)** The observed ISI distribution $p(t)$ of the digitally recorded spikes, showing the recorded CV/ν values, as well as the spike loss rate when compared to e). **e)** The true ISI distribution $p^*(t)$ inferred from the analogue recorded spikes in b), as well as the true CV value. **f)** The ISI distribution predicted to be digitally observed from e), treating spike loss as a convolution as described by Equation 4.2. Spikes are indeed shown to be lost from b) to a), and the effect of this spike loss is evident in d), where the initial true peak is followed by consecutive smaller peaks at integer multiples of the 1st peak's ISI. The predicted ISI distribution to be observed in f) shows good agreement with d), and predicts a CV similar to that observed in d), justifying our convolutional treatment of spike loss. To be noted is that the spike loss rate γ here of $\approx 9\%$ has resulted in disproportionately large 8x increase in the observed CV. The 50th neuron in a population of 100 neurons on a single HICANN all spiking at $\approx$ 20Hz is shown here. This is an example of a case where the true ISI distribution $p^*(t)$ may be reconstructed from the digitally observed lossy distribution $p(t)$ by dividing the n-th peak's ISI values by n to shift them back to the true peak.

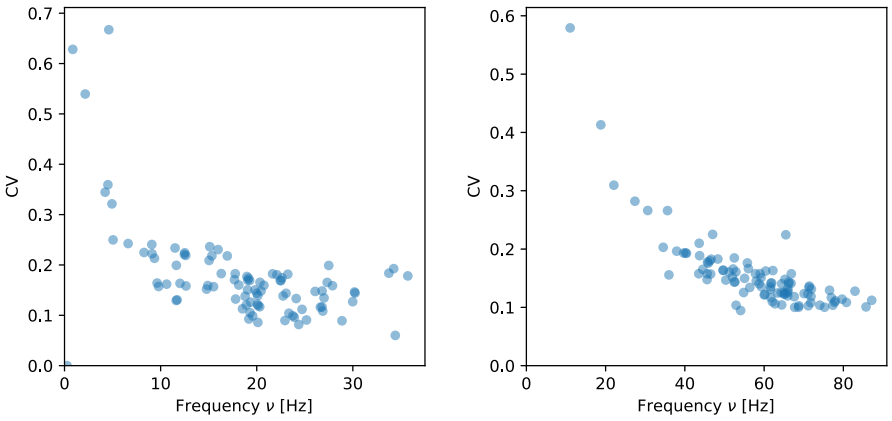

Figure 4.7: CV/ν plots for the noise network implemented on the hardware, without stimulation **(left)** and with 500Hz excitatory Poisson stimulus **(right)**. The spikes were recorded by sequentially recording the analogue membrane trace for each neuron, as described in Section 4.1.3, in order to counteract spike loss. Although the individual neuron CVs are not as good as found in simulation, this will be rectified when combining the spike trains from many noise neurons. The desired increase in noise neuron spiking frequency can be seen here, and the possible undesirable characteristics of a CV/ν plot, as shown in Figure 4.4, have been avoided.

the noise network and each sampling neuron, and so may increase the chance of synapse loss. For a finite number of noise neurons to sample from, it will also result in sampling neurons sharing more noise sources, and so will receive correlated noise. An N of 16 was chosen as a sufficient compromise among these constraints. The output noise frequency of the noise network is plotted in Figure 4.9 for various excitatory stimulus frequencies. We see that the noise network has an output frequency range of $\approx$ 300Hz $\rightarrow$ 1200Hz, at a consistently high (>0.9) CV value, and that there is very little variance in the noise frequency for different samples of 16 neurons. This figure may then be used as a calibration curve when wishing to use this noise network as a source of modulated noise in sampling experiments.

4.2 On the refractory period

Since sampling theory assumes that all sampling neurons have the same refractory period, it is thus favourable to ensure that the sampling neurons on the hardware also have a small spread in their refractory periods. Furthermore, since the state of a neuron (whether it is refractory or not) at time t will eventually be inferred by whether a spike has been recorded between t and $t - \tau_{ref}$, if a constant global τ_{ref} value is used rather than finding the "true" τ_{ref} for each neuron individually (and after every FG resetting), a large spread in the neuron population's refractory period values would cause the recorded network state to deviate from its "true" state. Though in simulation these issues can be avoided by simply setting all neurons to have a similar τ_{ref} value, we expect that on the hardware the previously discussed fixed pattern noise as well as FG variations will introduced an unwanted spreading of the τ_{ref} values. The optimal τ_{ref} value would thus be sufficiently large to minimise the effects of synaptic delays on the recorded

[4]Or, more likely, their excitatory OTA circuits are saturating.

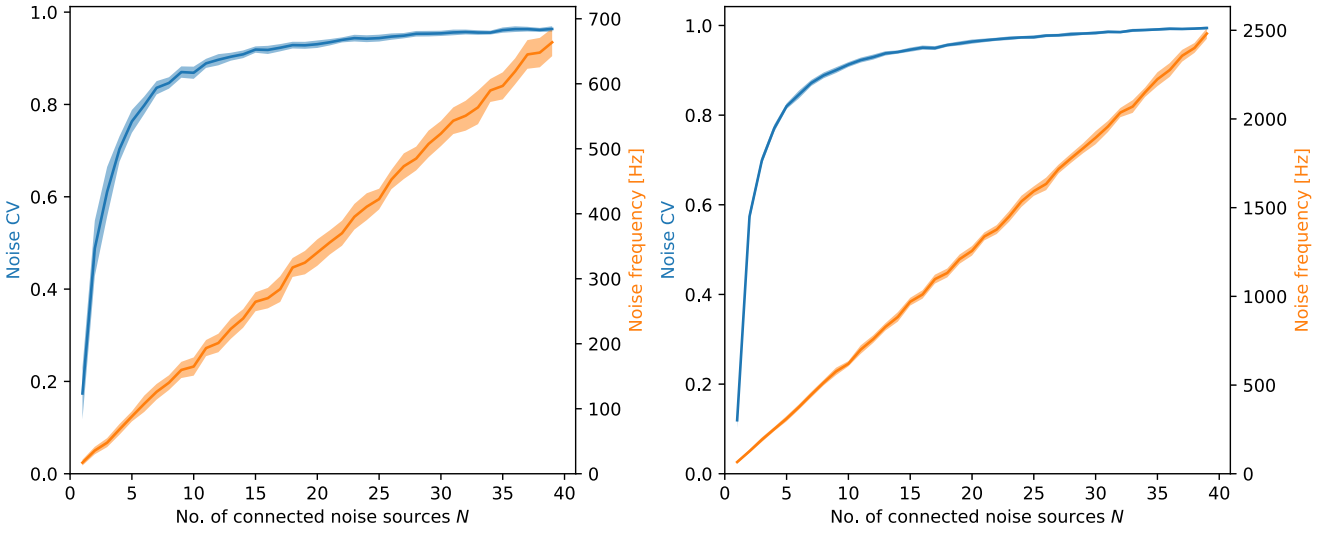

Figure 4.8: The CV and ν of the resulting spike trains when noise from N individual noise neurons (on hardware) is combined. The blue and orange lines are the CV and frequency respectively. **Left** is without any external stimulation, **right** is with 500Hz of excitatory Poisson stimulus. The coloured regions show the standard deviation of the respective variable between different samples of N noise neurons from the noise network. We see that although the CVs of the individual neurons were ≈ 0.2 (Figure 4.7), the CV of the composite spike train is greatly increased as desired. Since 300Hz was found to be the minimum required frequency for sampling [Kun16], we will thus connect each sampling neuron to 16 noise neurons, such that without stimulus, the noise network will provide $\approx$ 300Hz noise with a high CV. 16 is also a favourable number of noise neurons to use, as it means that each OTA on a sampling neuron with be connected to 2 noise neurons, and so each OTA should receive a similar amount of noise, the importance of which will become apparent when reviewing OTA saturation in Section 4.3.1.

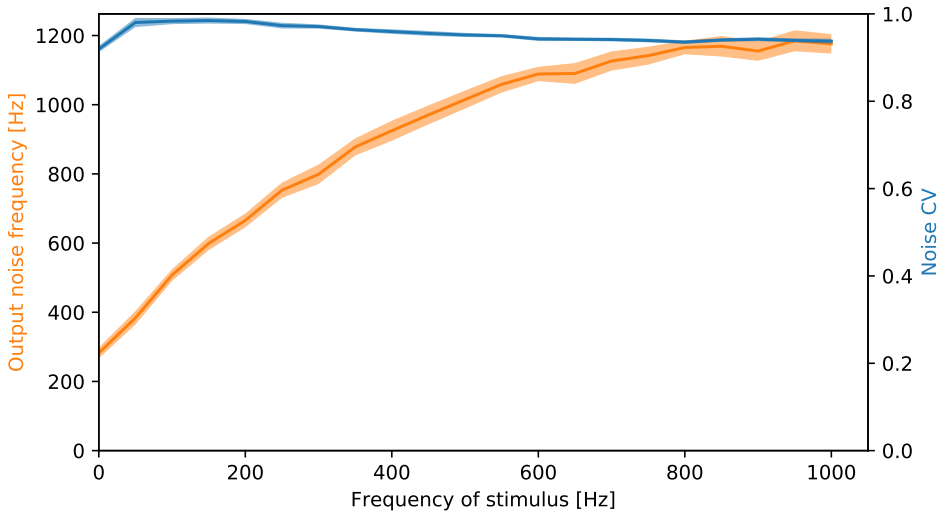

Figure 4.9: Frequency of noise produced by the noise network on hardware when stimulated with varying rates of excitatory Poisson noise input. The CV and ν is taken from the composite spike train of noise produced by 16 noise neurons, as determined in Figure 4.8. As desired, the network has $\approx$ 300Hz noise output when unstimulated, and may rise to $\approx$ 1200Hz. Increasing the noise frequency is met with a slight decrease in CV, as is it be expected when the neurons are being pushed to their maximum spiking frequency[4], and so individually have worsening CVs. This also explains the plateauing of the resulting output noise frequency. This figure may be used as a calibration curve for determining the required stimulus frequency in order to achieve a target noise network output frequency.

states and imperfect HCS [Kun16], but also have a small spread in its realised value across the sampling neuron population.

It is also favourable to perform this fixing of the sampling neuron τ_{ref} value early on, as this also fixes the synaptic time constants $\tau_{e/i}^{syn} = \tau_{ref}$ as dictated by sampling theory.

In order to measure the realised refractory period distribution, a network of unconnected, bursting neurons was set up in accordance with [Sch14]. The spike threshold potential was set to be lower than the rest potential, such that each neuron was bursting at frequency that was constant over time, and the remaining parameters (reset potential, membrane time constant) were set as to minimise the time taken for the neuron to spike again after it has finished being refractory, the *rise time*. Assuming that the rise time does not vary considerably between neurons, is not affected by the set $\tau_{ref,set}$ value, and that the ISI for a neuron set to $\tau_{ref,set} = 0$ consists purely of the rise time, the measured refractory time of the neuron is given by

$$\tau_{ref} = ISI_{mean} - ISI_0 \tag{4.3}$$

where τ_{ref} is the measured refractory time of the neuron, ISI_{mean} the mean ISI of the neuron (here disregarding spike loss), and ISI_0 the mean ISI of the neuron at $\tau_{ref,set} = 0$, corresponding to the rise time [Sch14].

4.2.1 Accounting for spike loss

With the effect of spike loss upon the ISI distributions better understood as per Section 4.1.3, precautions were taken to ensure it did not affect the measured refractory times. Since there are no synaptic connections in the desired network, the neurons can be spread out across many HICANNs without incurring the usual negative effects which would need to be considered (e.g. synapse loss). Thus, by placing few neurons on each HICANN, it can be ensured that the total spike rate per reticle (grouping of 8 HICANNs) is low enough such that spike loss is minimised, or rather can be easily accounted for.

When viewing the ISI distributions of *an individual neuron*, the same repeated but shifted multiple peaks were found as seen in Figure 4.6, and so the same treatment was applied to "undo" the convolution that spike loss had applied. That is, ISI values from the n-th peak were divided by n to scale them back to the first peak. Since the resulting ISI distribution was very sharp (individual neurons' ISIs do not vary considerably over time), and to reduce the risk of outliers affecting the recorded ISI, the median of each neuron's ISI distribution was taken to be the "true" fixed value for that neuron. The distribution of median ISIs across all neurons is shown in Figures 4.11 and 4.10.

4.2.2 Refractory periods on Wafer 33

The refractory period measurement experiment was run first on wafer 33, where *average calibration* for the refractory periods was active, meaning that the $\tau_{ref,set}$ space has been swept by manual setting of the digital I_{pl} parameter (which is responsible for the setting of the refractory period at a transistor level) and the refractory periods measured as per the procedure described above. When the refractory period is averaged over all neurons, a mapping between I_{pl} values and τ_{ref} is then realised, which is then used for all neurons indiscriminately. This calibration method therefore does not account for the variation due to fixed pattern noise present between neurons.

The results are shown in Figure 4.10. The distribution obtained for $\tau_{ref,set} = 0$ was as expected, with a sharp spike at 0.8ms with a width of 0.1ms. When viewing the individual membrane potential traces, this was found to indeed correspond to the rise time, confirming the assumption that the rise time can be subtracted off of the measured ISI as a constant. However, for all non-zero $\tau_{ref,set}$ values, the measured τ_{ref} was much larger than requested, and more importantly had a width larger than the requested value. To highlight the consequences of this, if a neuron were set to $\tau_{ref,set} = 10$ms, the neuron spiking every 50ms would be indistinguishable from the neuron being refractory 100% or 20% of the time. Wafer 33 was thus deeded unsuitable for sampling with the currently available calibration.

4.2.3 Refractory periods on Wafer 30

The experiment was thus moved to wafer 30, where *individual neuron calibration* was active, meaning that a similar calibration procedure had been performed as in the average calibration, except that a mapping between the set I_{pl} value and τ_{ref} was found for every neuron individually, thus reducing the neuron to neuron variability. The results are shown in Figure 4.11. Here the

distributions are much thinner, with widths being approximately $\frac{1}{10}$ of their corresponding τ_{ref}. The relative sizes of ISI_{mean} and their corresponding standard deviation σ were concordant with previous calibration studies [Kug18].

However, another feature of note is that while the τ_{ref} values are within one or two standard deviations of the corresponding $\tau_{ref,set}$, the difference between the τ_{ref} and the measured ISI_{mean} is much larger than that on wafer 33, corresponding to a larger rise time. This contradicts the findings for $\tau_{ref,set} = 0$ on wafer 33, where the rise time was found to be a fixed value of about 0.8ms, which did not change considerably from neuron to neuron. In comparison, here on wafer 30, $\tau_{ref,set} = 0$ yields a much larger ISI_{mean} of 2.5ms, with a not-insignificant spread of 0.5ms. It was found that for $\tau_{ref,set} = 0$, the digital I_{pl} parameter was not being set to its maximum value of 1023 as it should in order to yield the lowest possible refractory time, but was instead being set to values as low as 100. Therefore the $\tau_{ref,set} = 0$ ISI distribution is not representative of the lowest possible τ_{ref} value, and so does not constitute the rise time only, and instead is erroneously increased by a finite refractory time, which also accounts for the increased spread. If this ISI was erroneously taken to be ISI_0 during calibration, it would thus manifest as a constant decrease in the measured τ_{ref} during calibration, meaning that I_{pl} values would thus be incorrectly mapped to τ_{ref} values lower than the true realised refractory period. Thus when requesting a certain refractory period, the set I_{pl} will correspond to a true refractory period greater than that requested by the same fixed offset, as can be seen by the ISI_{mean} values being considerably larger than their $\tau_{ref,set}$ counterparts, by more than the true rise time of approximately 0.8ms. To support this hypothesis, when the I_{pl} values were manually forced to 0, the ISI plot closely resembled that from wafer 33. The offset also did not vary with neuron size significantly.

As a measure of the spread to be minimised, the coefficient of variation CV = $\frac{\sigma}{\tau_{ref}}$ was used, as it more meaningfully gives the proportion of the refractory time that is indeterminate. As per the previous reasoning for the need for a long τ_{ref}, as well as in keeping with values used in other works and to minimise the CV, $\tau_{ref} = 10$ms was selected. In order to account for the above discussed offset, a value of $\tau_{ref,set} = 8.5$ms is instead requested. All subsequent sampling neurons use this value[5] of τ_{ref} and $\tau_{e/i}^{syn} = 10$ms, and are conducted on wafer 30.

4.3 Limits of Poisson input noise on the HICANNv4 chip

The underlying mechanism to implement tempering is to effect a widening of the neurons' activation functions, i.e. increase the associated temperature, by increasing the strength[6] of noise input to said neurons.

As introduced in Section 2.1.3, there is a limited but intuitive geometric interpretation of the link between the width of individual neuron membrane potential distributions, and the width of the corresponding activation functions. The main limitation is that as the noise rate is increased, the membrane potential width peaks and then drops off, whereas the activation

[5] Synaptic time constant calibration was not yet available

[6] We are here purposefully ambiguous with the word strength, as we mean either the weight or rate of the Poisson noise.

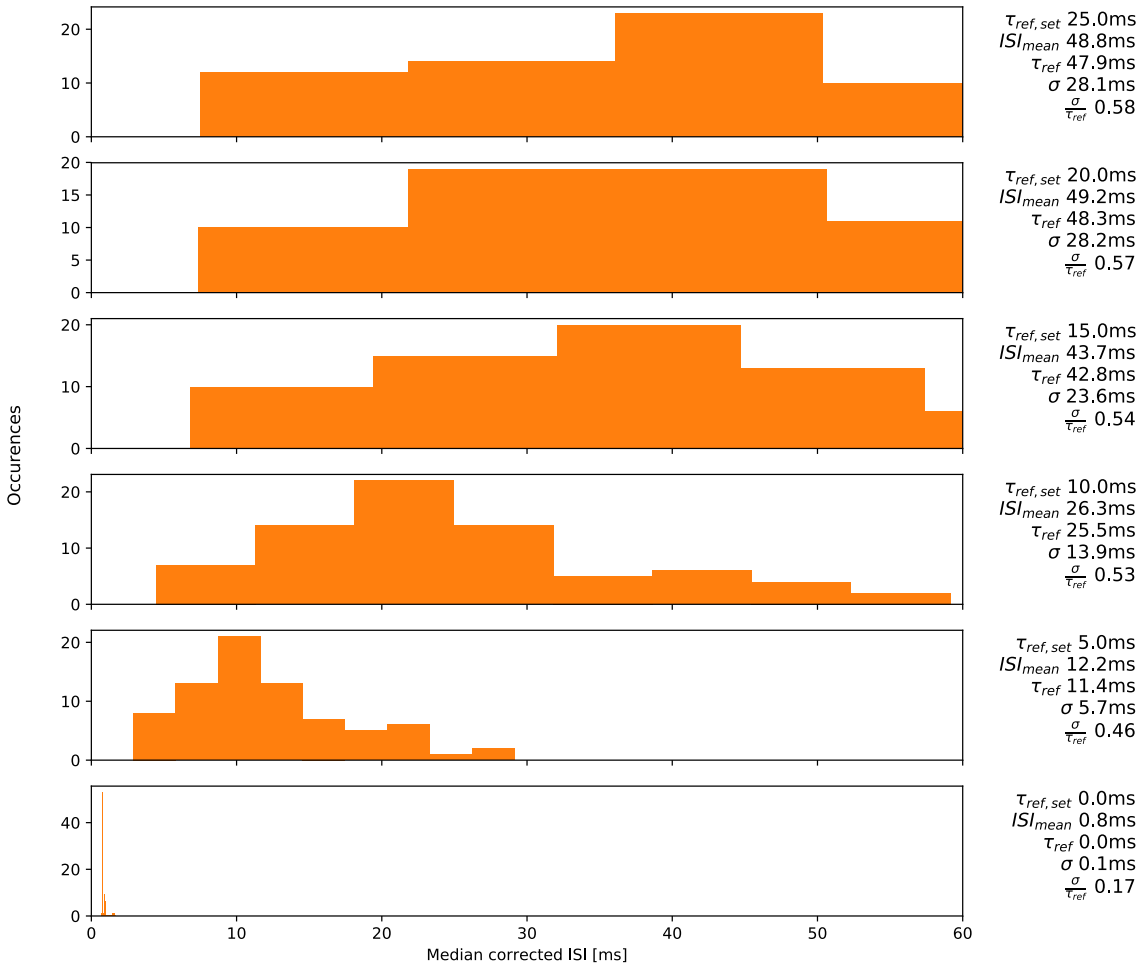

Figure 4.10: Refractory period distribution for varying requested refractory period $\tau_{ref,set}$ on wafer 33. Each binned data point is the median ISI of a single neuron, corrected for spike loss. ISI_{mean}, τ_{ref}, σ and $\frac{\sigma}{\tau_{ref}}$ are the mean of the ISI distribution, the subsequent refractory period as determined by Equation 4.3, the standard deviation / width of the distribution, and the coefficient of variation. Since for non-zero $\tau_{ref,set}$ the width σ is of similar size to the corresponding ISI_{mean}, this wafer is deemed unsuitable for sampling with the currently available calibration in favour of wafer 30. The ISI for $\tau_{ref,set} = 0$ is fittingly very small and almost single valued, corresponding to the ISI being due to the rise time of the neurons only, and being affected very little by both FG variation and fixed pattern noise. To minimise spike loss, only 85 neurons were used with only 5 neurons on a single reticle, and so large bin sizes are here required. The figure has been limited prematurely on the right to enable comparison with Figure 4.11. Within the population, 2 outliers have been removed, having ISIs of $\approx$ 600ms.

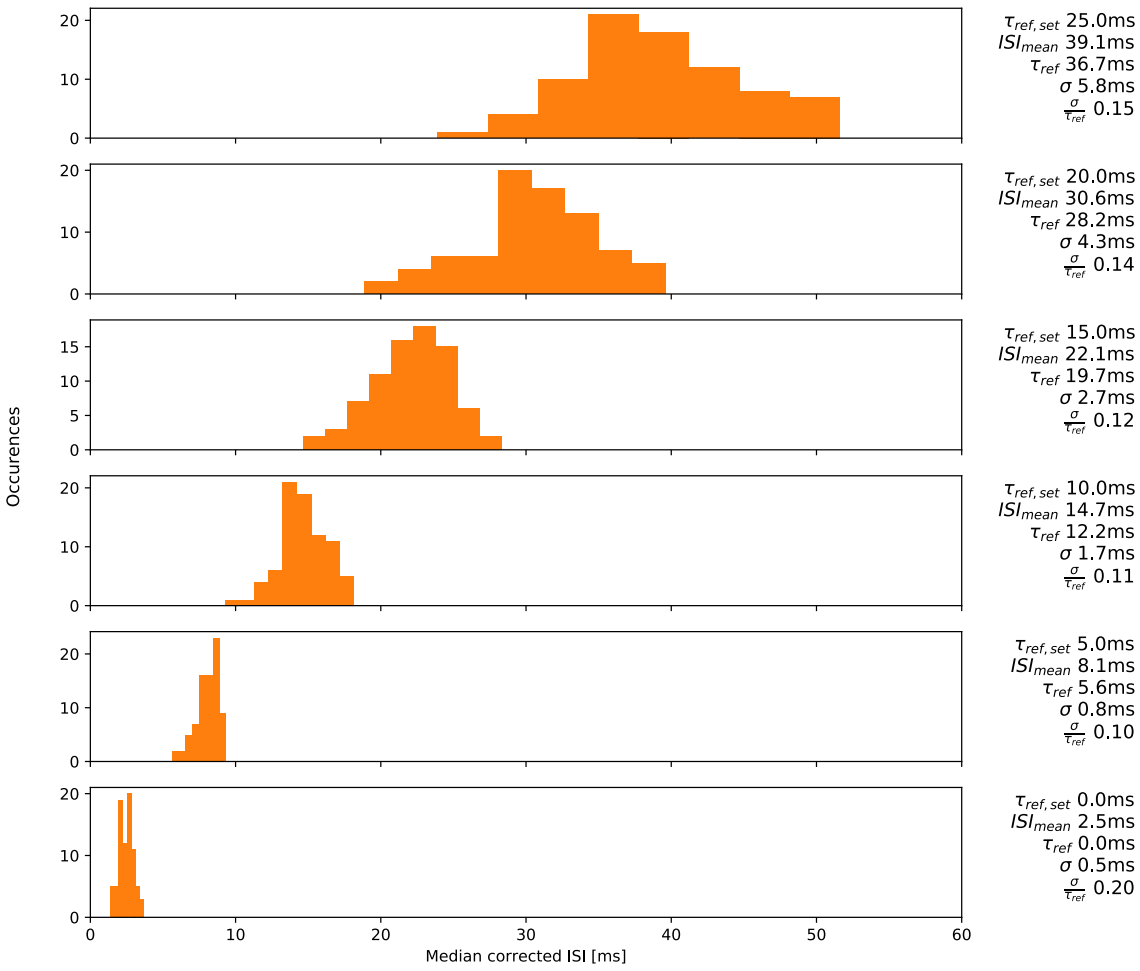

Figure 4.11: Refractory period distribution for varying requested refractory period $\tau_{ref,set}$ on wafer 30. Figure elements have the same meaning as Figure 4.10.
The τ_{ref} values are much closer to their corresponding $\tau_{ref,set}$ values, and both the widths and coefficient of variations are much smaller, making this wafer much more suitable for sampling. Note that for $\tau_{ref,set} = 0$, the mean is much greater and not single valued as on wafer 33, caused by the controlling I_{pl} digital parameter not being set to 1023. Since the ISI_{mean} from $\tau_{ref,set}$ sets the value of the apparent rise time ISI_0, this erroneously high rise time could have had a knock on effect on all subsequent recorded τ_{ref} values. $\tau_{ref,set} = 8.5$ms is used in all subsequent experiments, for a realised τ_{ref} value of ≈ 10ms.

function continues to widen. They are nonetheless related phenomena, and so it is worthwhile to check that the membrane potential widths are responding on the hardware as expected as the noise strength is varied. In simulation, the activation functions widened with both increasing weight and rate, each over a much larger range of values than covered here [Bau16]. We however expect to encounter hardware-specific limitations upon the noise strength and thus achievable activation function widths before then.

The major limitation that is discussed here is the saturation of the OTA circuits, as introduced in Chapter 3. Spikes received at the postsynaptic neuron are convoluted with a decaying exponential kernel with time constant $\tau_{e/i}^{syn}$, and are then passed to the OTA, which would ideally produce a synaptic current proportional to the linear sum of all the incoming spikes. However, due to the fact that the OTA will have an upper limit on the maximum current output, it is expected that a saturation regime may be reached, whereby the OTA reaches this maximum output current. Such a regime is to be avoided, as saturating at a constant output current would mark a complete loss of the stochasticity upon which sampling is reliant. Furthermore, saturation would cause the neuron to become unresponsive to any further incoming spikes, which would naturally be detrimental to sampling. Since reaching a constant maximum in the OTA output current would mark a complete loss of stochasticity, we expect the stochasticity of the neuron dynamics (characterised by the free membrane potential distribution width σ) to suffer before reaching complete saturation.

In order to detect OTA saturation occurring, we make two statements about the expected behaviour of the membrane potential, if there is no OTA saturation.

1. With increasing noise weight, it has been found in simulation (Figure 2.2) that the membrane potentials continue to widen monotonically.
2. If the neuron is exposed to only excitatory or inhibitory, then increasing either the noise rate or weight should result in the membrane potential moving closer to the respective reversal potential.

Though there is the possibility that the behaviour of the membrane widths may in fact be more complex, if the means stop shifting towards the respective reversal potential (and the mean is not close to it) then this is a definite sign of OTA saturation occurring. This thus motivates the sweeping of the rate and weight Poisson noise input to a non-spiking neuron, to check for saturation effects.

4.3.1 Limits of Poisson noise weight

The weight of input noise to a sampling neuron was first swept for a fixed Poisson input frequency. The weight was varied by varying the parameters $w \in [0, 15]$ and $gmax \in [0, 1023]$, the 4-bit digital weight of the individual connection and a wafer-wide[7] scale factor on the current respectively, whereby the synaptic current I_{syn} produced by a PSP should obey $I_{syn} \propto w \cdot \frac{gmax}{g_{div}}$, where g_{div} is another digital scale factor, is set to 2 and due to redundancy with $gmax$ is not varied [Sch+10]. The setup for the experiment is shown in Figure 4.12 a): a sampling neuron

[7] gmax may actually take multiple values on a single wafer, however for the sake of simplicity it is here treated as a wafer-wide constant.

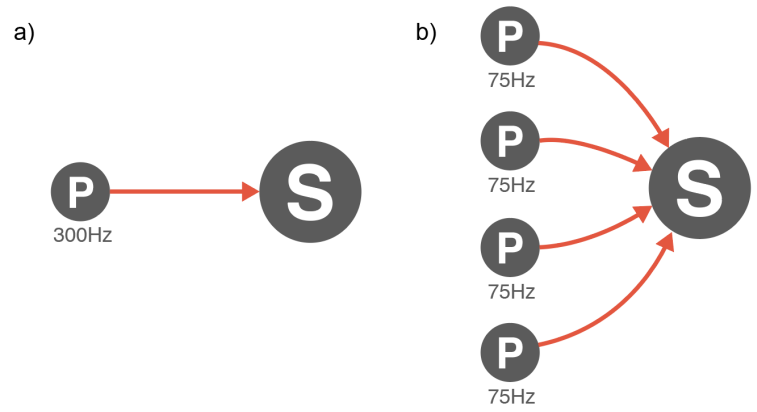

Figure 4.12: The three network setups for determining the limits of noise rate and weight input to a sampling neuron. Sampling neurons and Poisson input neurons are denoted S and P respectively. Red arrows denote an excitatory connection. All setups were also run with inhibitory connections, later denoted as blue arrows. **a)** The initial setup for probing the limits, with no regard for OTA saturation. **b)** In order to limit the load on individual OTAs, the input noise is split up into 8 (4 pictured for the sake of clutter) separate Poisson sources, with total input rate 300Hz still. In doing so, saturation effects are delayed and a larger membrane potential distribution is reached.

is exposed to a single source of Poisson noise either excitatorily or inhibitorily, initially at the minimum frequency required for sampling of 300Hz. The spiking threshold V_{thresh} was set high enough such that no spiking behaviour may occur, and the remaining parameters were set to the values which will be used during sampling. In particular, the reversal potentials E_i^{rev} and E_e^{rev} were set to their respective extremes of -100mV and +45mV, such that none of the saturation effects could be confused with effects associated with the membrane potential nearing the reversal potentials.

For a varying noise weight at a fixed input rate of 300Hz, the membrane potential distributions are shown in Figures 4.13 and 4.14 for excitatory and inhibitory input noise respectively.

In both figures, with increasing gmax and digital weight there is an initial mean shift towards their respective reversal potentials, as well as a widening of the distributions. But at high gmax and weight values, saturation behaviour begins to occur, whereby the mean does not shift any further, and the width diminishes to the readout noise width. For comparison, Figure 4.15 shows a neuron under 900Hz excitatory noise, where saturation effects occur at much smaller weight values. In order to better quantify these two effects, the same experiments were run again with many more gmax and digital weight samples, and the mean and standard deviation from every gmax/weight pair were plotted as in Figures 4.16 and 4.17 for 300Hz and 900Hz excitatory noise respectively. The mean plateaus at much lower values of gmax and digital weight for the 900Hz plot compared to the 300Hz, as is to be expected if the OTA is indeed entering a saturation regime. Their inhibitory counterparts showed identical behaviour, except with the mean potentials falling rather than rising to a plateau, and as such have been omitted. In order to reduce the effects of OTA saturation, and consequently to allow greater noise weights to be safely used, the setup was altered to resemble Figure 4.12 b). In this configuration, the 300Hz input noise is split up instead into 8 separate Poisson sources, each with $\frac{1}{8}$ of the total desired input frequency. This thus allowed the total noise input to be shared among multiple OTAs, reducing the load on any single OTA. The results from this setup are shown in Figures

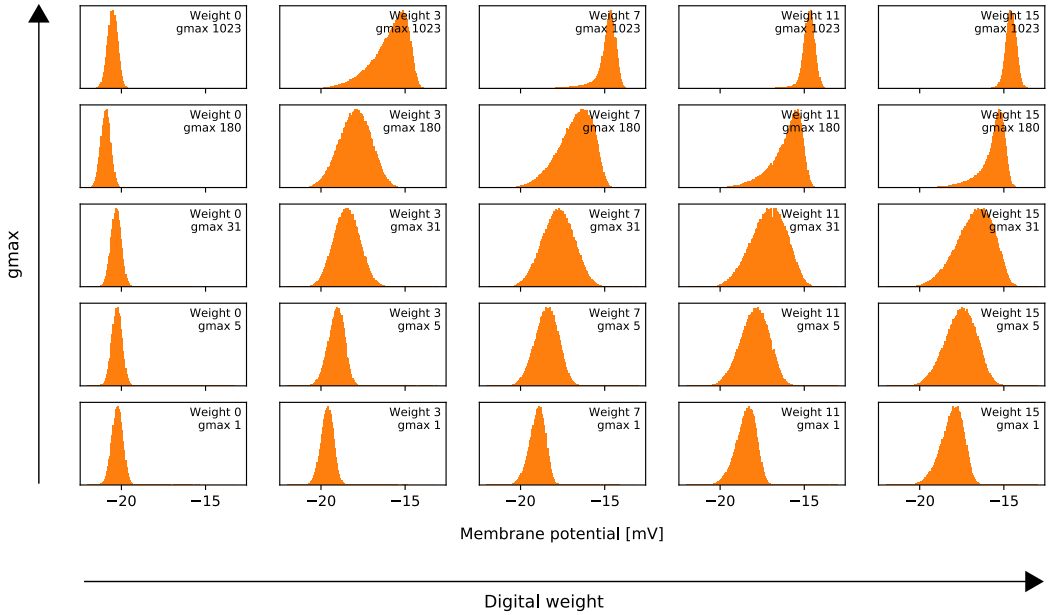

Figure 4.13: Free membrane potential of a sampling neuron under 300Hz *excitatory* Poisson noise from a single source, with varying weight strength. With increasing gmax and digital weight, the width increases as expected, and the mean also shifts upwards towards E_e^{syn}. However, at very high gmax and weight values, a maximum mean shift is reached, and the distribution thins to the readout noise width (as in weight 0, gmax 1), signifying OTA saturation. Setup described in Figure 4.12 a).

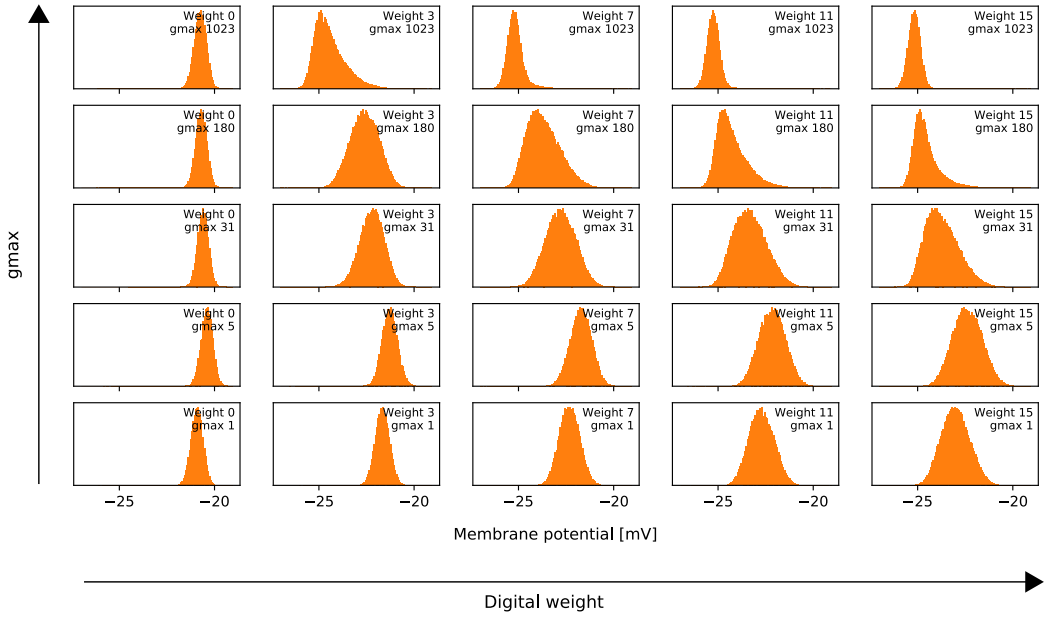

Figure 4.14: Free membrane potential of a sampling neuron on hardware under 300Hz *inhibitory* Poisson noise from a single source, with varying weight strength. The same initial widening then saturation behaviour is observed here for the inhibitory case as for the excitatory, as seen in Figure 4.13

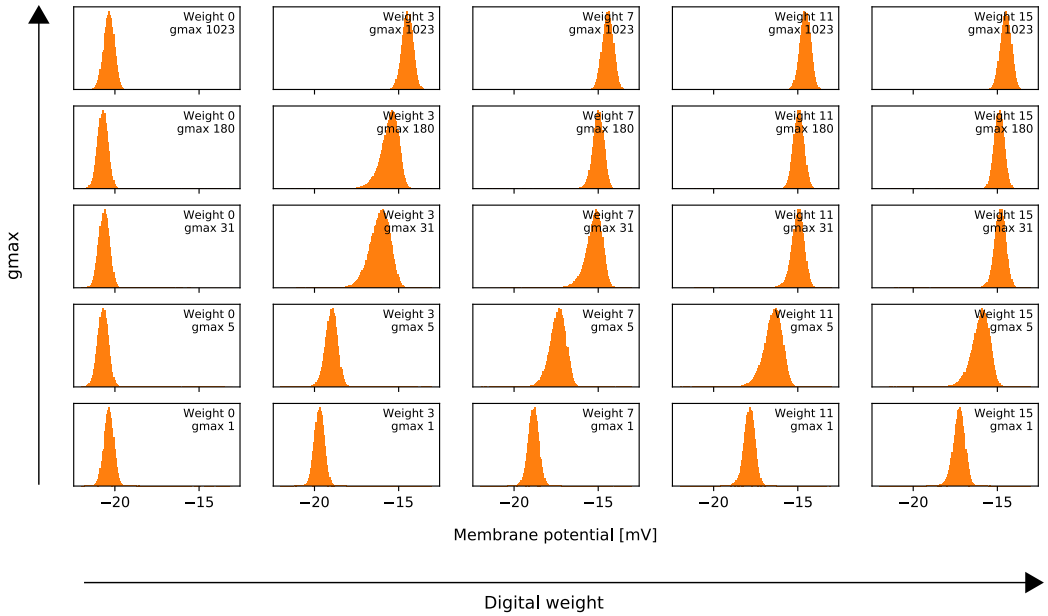

Figure 4.15: Free membrane potential of a sampling neuron on hardware under 900Hz excitatory Poisson noise from a single source, with varying weight strength. The widening here is less noticeable, but what is of importance is that the saturation effects (reduction of width, reaching a maximum in the shift of the mean) occurs much earlier than as seen for 300Hz in Figures 4.13 and 4.14.

4.18 and 4.19 for 300Hz and 900Hz input rate respectively. In both cases, the weights at which saturation effects occurred were increased significantly, and the maximum membrane potential width in both cases increased by 2-3×. Unless specified otherwise, all subsequent experiments will ensure that noise inputs are split among the OTAs in a similar manner.

Though the gmax and digital weight values at which saturation occur have been increased, we still wish to robustly identify, for a range of frequencies, the exact values at which saturation becomes significant, such that it may be avoided. From the mean and standard deviation plots, two such identifiers of saturation thresholds were considered. First, as in Figure 4.15, the points at which the mean begins to plateau was considered as an identifier, as plateauing is a definite consequence of OTA saturation, whereas the behaviour of the width may be more complex. However, this method raised multiple issues, such as how the maximum is defined (especially in the low frequency cases), and at what distance to the maximum should saturation be said to occur (especially since the standard deviation is not singly determined), and thus was deemed not robust enough. Instead, the saturation points were chosen to be the points at which the slope of the standard deviation plot, in the positive gmax-weight direction, become negative. Though the falling of the standard deviation may not be as decisive an indicator of saturation as the mean, in all cases the standard deviation begins to drop close to but before an obvious plateauing of the mean is reached, and so provides a robust identifier that errs on the side of caution. The results are shown Figure 4.20, and show that for increasing frequency, the maximum weight values to which the noise may be set decrease.

The gmax and digital weight value for noise in rate modulation based tempering experiments is then chosen as follows: for an intermediate noise weight value of 7 (such that sampling

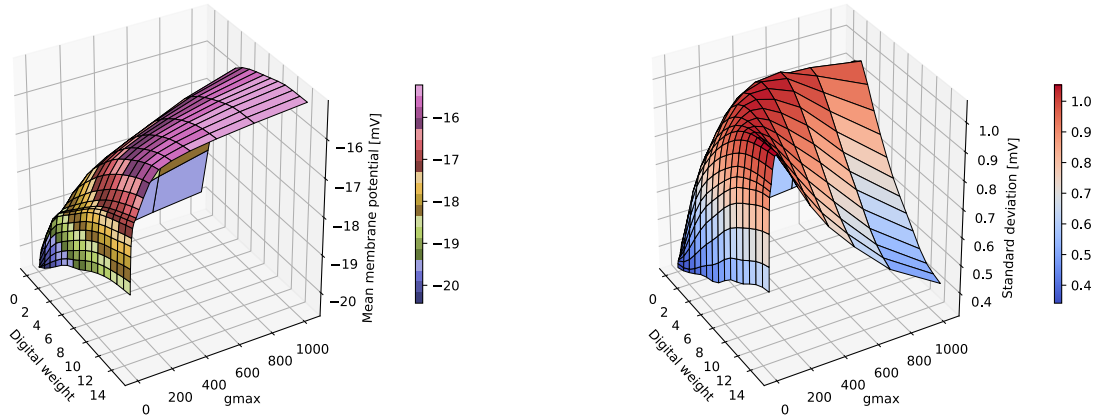

Figure 4.16: The mean and standard deviation of the membrane potential distribution for a sampling neuron on hardware subject to 300Hz excitatory Poisson noise from a single source, with varying noise weight. As in Figure 4.13, saturation effects can be seen in the plateauing of the mean, as well as the the simultaneous drop in the standard deviation.

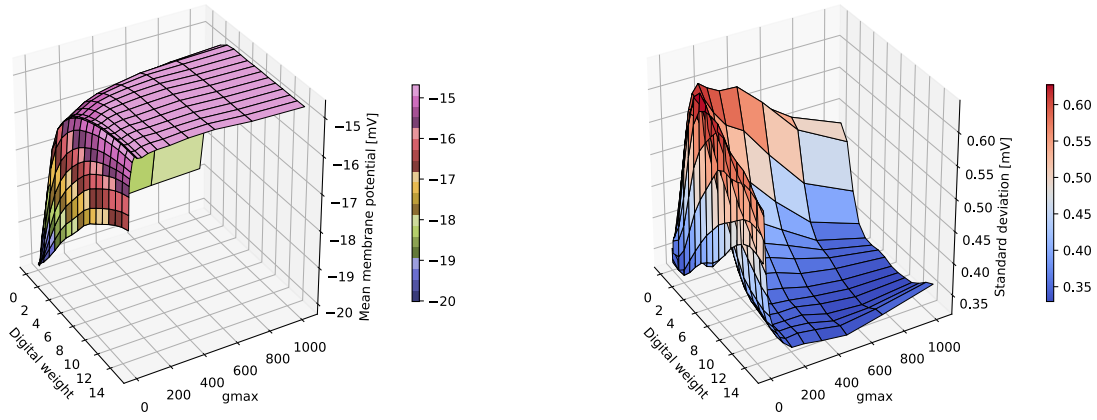

Figure 4.17: The mean and standard deviation of the membrane potential distribution for a sampling neuron on hardware subject to 900Hz excitatory Poisson noise from a single source, with varying noise weight. Compared to the 300Hz noise input in 4.16, the saturation effects occur at much lower weights and are considerably more noticeable, with the mean becoming almost completely flat, and the standard deviation dropping to a minimum almost immediately.

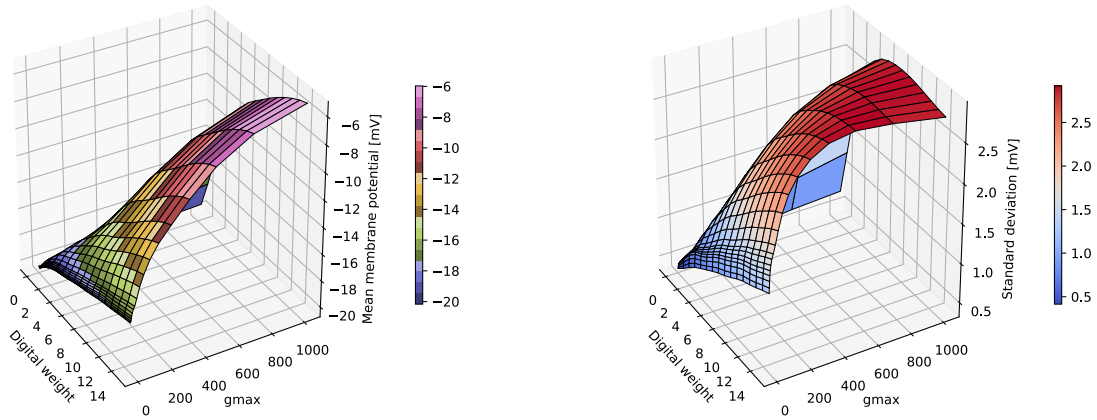

Figure 4.18: The mean and standard deviation of the membrane potential distribution for a sampling neuron on hardware subject to 300Hz excitatory Poisson noise, spread among 8 sources, with varying noise weight. Compared with Figure 4.16, where the only difference is that the noise input is split equally among all available OTAs, here the maximum shift in the mean is much greater, with little sign of plateauing. Similarly, the maximum achieved standard deviation is much greater, and does not decrease again significantly. OTA saturation effects have thus been prevented for all but the most extreme weight and gmax values.

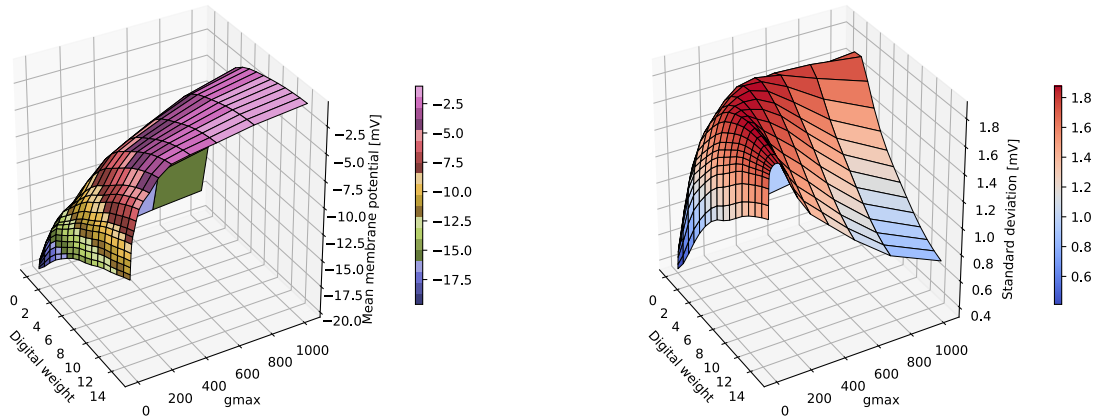

Figure 4.19: The mean and standard deviation of the membrane potential distribution for a sampling neuron on hardware subject to 900Hz excitatory Poisson noise, spread among 8 sources, with varying noise weight. Again compared with Figure 4.17, where the only difference is that the noise input is split equally among all available OTAs, though there is still visible saturation behaviour, it has been delayed to intermediate gmax and digital weight values, rather than occurring immediately. Thus by splitting the input among multiple OTAs, using 900Hz input noise for sampling should be possible as long as the weight parameters are set low enough as to avoid the saturation regions.

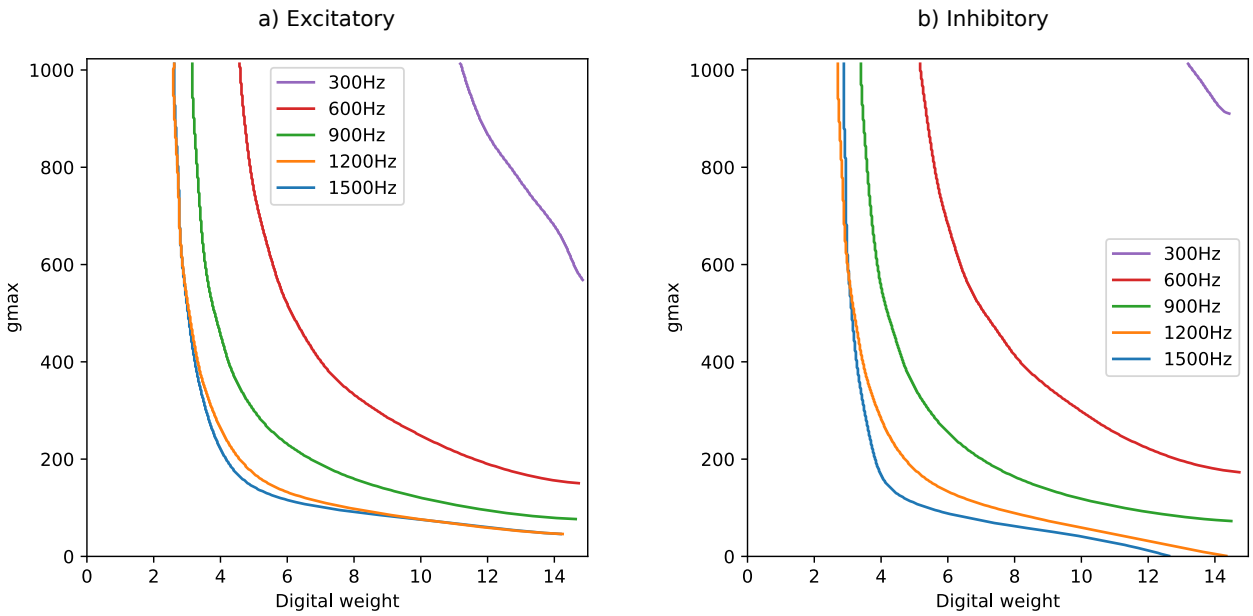

Figure 4.20: Noise weight values at which OTA saturation begins to occur for different frequencies, for purely excitatory and purely inhibitory input in **a)** and **b)** respectively. Setup is shown in Figure 4.12 b). Saturation occurring is characterised by a decrease in membrane potential width with increasing gmax and digital weight. The gmax value is then set wafer wide to the maximum value for which a digital weight of 7 remains below the curves of the desired frequency range. The strange behaviour in the bottom right corner of **b)** is due to unwanted artifacting arising due to the low sampling resolution of the digital weights sweep (only integers values) and subsequent processing. The gmax range was swept with 20 logarithmically spaced samples, explaining the kinks in in the 300Hz curves.

neurons may have weights weaker and stronger than the noise), the highest gmax value is picked for which saturation is not reached for the desired maximum frequency. For the following experiments, gmax was set to 150, such that a tripling of the frequency from 300Hz to 900Hz can be safely made, but saturation effects are expected to occur for higher frequencies.

4.3.2 Membrane potential distributions with varying noise rates

Using the allowed weight values determined in the previous section, the dependence of the membrane potential width upon the noise input rate was examined. For the sake of confirming the findings from the previous section, the same experiment was also run with a lower gmax value. The results are shown in Figures 4.21 and 4.22 for gmax values 150 and 50 respectively. In the former, the membrane width maximises at 300Hz and then decreases, while the latter maximises around 600Hz and then decreases. Since OTA saturation has been avoided up to at least 900Hz in both plots, supported by the fact that the widths do increase from gmax 50 to 150 up to 900Hz as enforced and justified in Section 4.3.1, the lack of widening cannot be attributed to OTA saturation. The lack of widening also cannot be attributed to spike loss, as in all cases the distributions of the individual excitatory or inhibitory plots continue to shift towards their respective reversal potentials with increasing noise rate. That the lack of widening is due to the theoretical effect described in Section 2.1.3, thus becomes a more probable possibility. In order

to better compare with the behaviour of the width in simulation (found to be in accordance with theory [Pet15]), the weight and input frequency were swept simultaneously to replicate Figure 2.2 on the hardware. The results for gmax 150 and 50 are shown in Figure 4.23, and largely mirror the behaviour found in simulation, with a notable exception that the peak in width for an increasing input rate is followed by a plateau at an intermediate value, rather than dropping to 0. This deviation from theory is especially prominent for the values used in the previous rate variation figures 4.21 and 4.22, where the peak at approximately 300Hz is followed by an almost immediate plateau. Despite this deviation from theory, if the lack of widening is indeed a theoretical effect only, then it should not pose a restriction on the widening of the activation functions, as discussed in Section 2.1.3.

4.4 Widening of single neuron activation functions on the HICANNv4 chip

Knowing now at what noise strengths OTA saturation occurs, the activation functions are found when varying either the weight or rate of Poisson input noise. The theoretical requirement for sampling is that the activation functions should be approximately sigmoidal, with the input variable being the mean membrane potential. Though in simulation the mean membrane potential can easily be swept directly by varying the rest potential E_l, a change in mean membrane potential was instead effected by connecting the sampling neuron to a set of bias neurons spiking at a constant rate, and then varying the connection weight ("bias weight") between maximum excitatory and maximum inhibitory (denoted as a weight sweep from 15 $\rightarrow$ -15). This thus emulates the bias that a sampling neuron would be subject to from other neurons within a connected sampling network. Though the addition of bias input could alter the frequencies/weights at which OTA saturation occurs, since the majority of the activation function dynamics occur at small bias weights, any possible bias-induced saturation is here disregarded.

4.4.1 Widening via noise rate modulation

The first attempted method to induce a widening of the activation functions was to modulate the rate of Poisson input noise to the sampling neurons, as successfully shown in [Kor17] for CUBA-LIF neurons in simulation, where the activation function widths α were found to obey $\alpha \propto \sigma \propto \sqrt{\nu}$ where σ is the standard deviation (width) of the free membrane potential distribution and ν the rate of Poisson noise input. Though the relation $\sigma \propto \sqrt{\nu}$ holds only for CUBA-LIF neurons, since a peak in width is reached at around 300Hz input rate (Figure 4.23) for COBA-LIF neurons, this relationship was found in simulation to not be necessary in order for $\alpha \propto \sqrt{\nu}$ to hold [Bau16] for COBA-LIF neurons, as is here desired. Modulating the noise rate rather than weight to change the temperature has the advantage that a continuous possible temperature range may be achieved, and that the modulated Poisson noise may instead be replaced by noise produced from a modulated noise network as described in Section 4.1. The network setup is described in Figure 4.24.

The spiking threshold V_{thresh} was set such that a zero bias corresponded approximately to a spike rate of half the neuron's maximum rate, so that an unbiased neuron does not have any

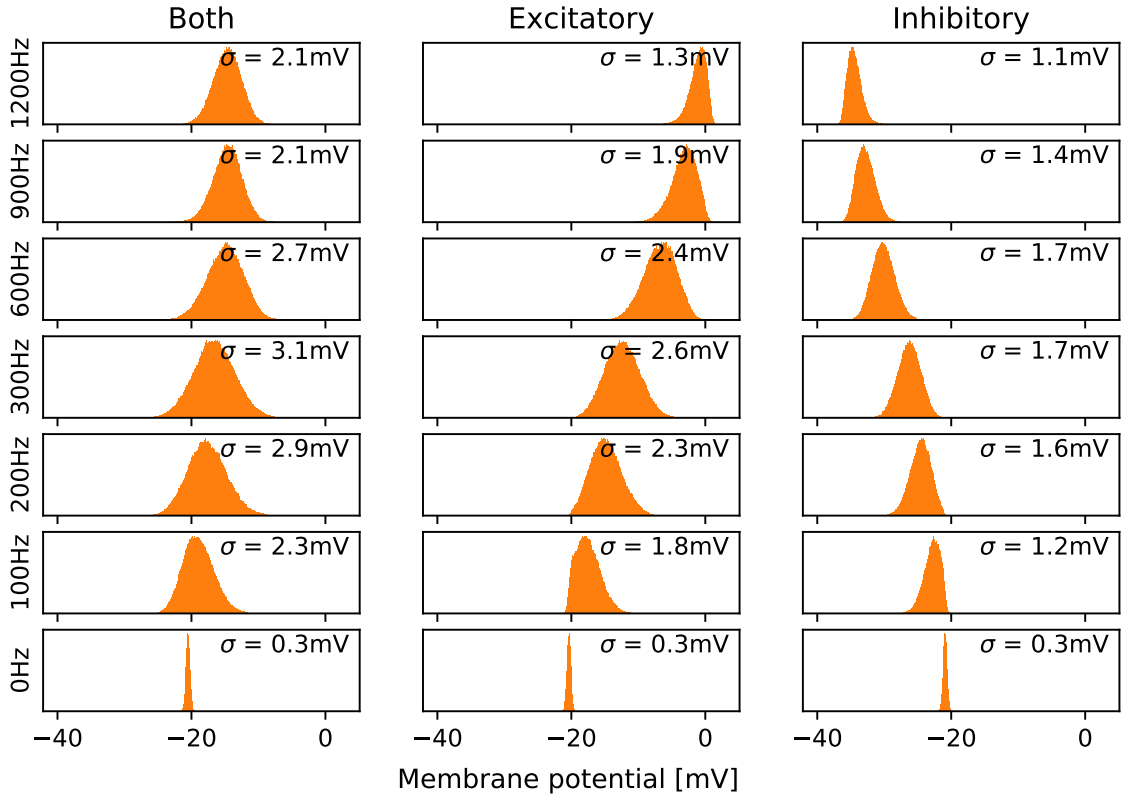

Figure 4.21: Membrane distributions for different noise input rates (varying from row to row), for either excitatory and inhibitory, or only excitatory or inhibitory (for columns 1,2,3 respectively), at gmax 150 digital weight 7 as determined in Section 4.3.1. The weight values were chosen such that the OTA saturation regime has not been entered, at least for rates up to 900Hz. Despite the rate of noise input increasing, past 300Hz (the minimum frequency required for sampling), the membrane distributions do not widen, and instead become thinner, indicating that the width may already have reached its maximum value as theorised in [Pet15] at 300Hz. The 1200Hz row should be interpreted with caution, as it is entering the saturation regime as shown in Figure 4.20. The lack of widening past 300Hz should not be attributed to spike loss, as both the excitatory and inhibitory only columns show that the distributions continue to be shifted further towards their respective reverse potentials with increasing noise up to at least 900Hz, showing that the spikes are at least being received.

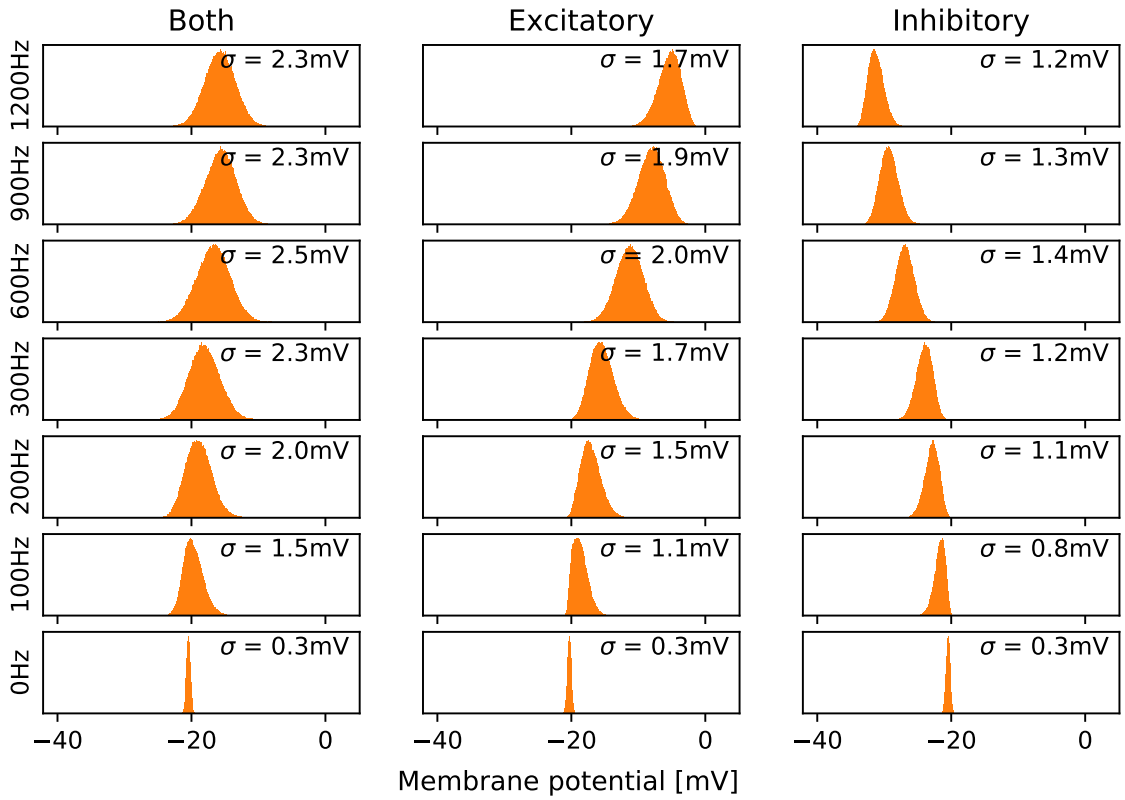

Figure 4.22: Figure construction is identical to 4.21, except with a lower gmax value of 50 (weight 7 remains unchanged). Due to the lower gmax value, and according to the saturation regions defined by Figure 4.20, saturation effects have been avoided for all frequencies shown. In agreement with 4.20, and rather as the desired consequence of the chosen width-based indicator of OTA saturation, the widths of all the excitatory or inhibitory distributions up to 900Hz increase from this figure to Figure 4.21, as should be expected from an increase in noise weight. However for 1200Hz, for which OTA saturation has been determined to occur for gmax 150, the widths instead *decrease* from gmax 50 to 150. Similar to the gmax 150 case, except that the width maximises at 600Hz rather than 300Hz, there is overall very little widening, and indeed a decrease in width from 600Hz to 1200Hz.

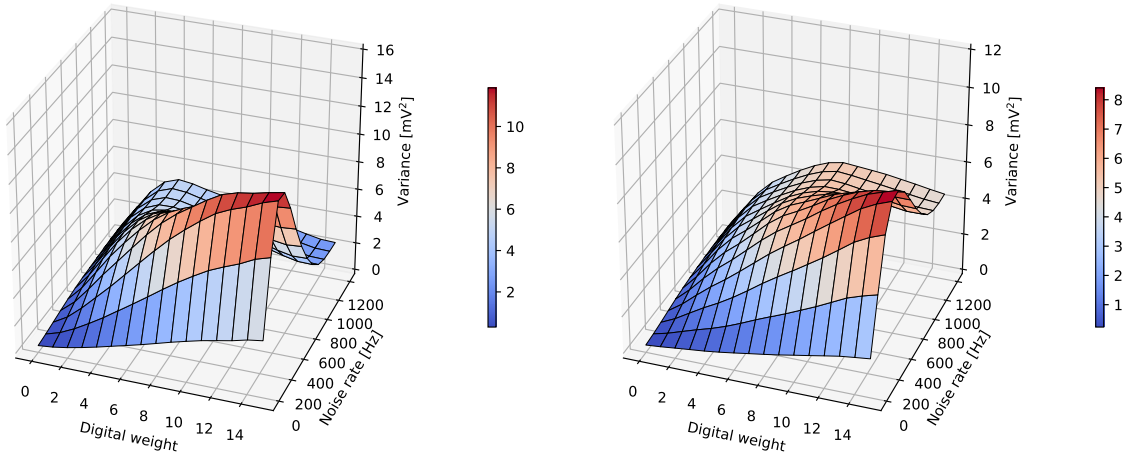

Figure 4.23: The width of the membrane potential distribution on hardware subject to excitatory and inhibitory Poisson noise at different weights and rates, at gmax 150 (**left**) and gmax 50 (**right**). A sample is taken every 100Hz and at every integer digital weight. Figures 4.21 and 4.22 thus represent a slice at digital weight 7 from the left and right figures respectively. Using Figure 4.20, it is known that for gmax 150, saturation effects are avoided for 900Hz approximately up to weight 9, and for 1200Hz up to weight 5, and thus the behaviour of the width beyond these values should be disregarded. Similarly, for gmax 50 (right), saturation may be disregarded for all but the highest digital weight values at 1200Hz. Comparing with the same plot from simulation (Figure 2.2), the overall behaviour is largely similar, with the width initially increasing increasing with both rate and weight, but reaching a peak at a fixed rate. Notably different however is that past the peak, the width plateaus at an intermediate value rather than going to zero with large input rate. This occurs in regions where saturation effects may be ignored. Besides which, saturation would be associated with a *decrease* in width. Also, the width does not become independent of weight at high input rate, though this could be a direct consequence of the fact that the width does not drop to 0 (or rather the voltage readout noise width).

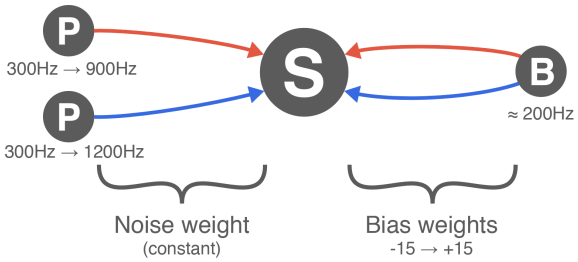

Figure 4.24: Setup for finding single neuron activation functions, where activation function widening is to be achieved by modulation of Poisson noise input rate. Diagram elements have the same meaning as in Figure 4.12, with the addition that neurons labelled B are bias neurons, with $E_l > V_{thresh}$ to have a constant spiking rate. The actual implemented network had all Poisson inputs and bias neurons instead split into 8 separate inputs, each with an $\frac{1}{8}$ of the total frequency shown, as described in Figure 4.12 b). The rate of excitatory and inhibitory input noise was increased simultaneously, with shift compensation being coarsely implemented by increasing the excitatory less than the inhibitory as shown. For every frequency value, the bias weights are swept from -15 $\rightarrow$ +15, corresponding to 15 $\rightarrow$ 0 inhibitory with 0 excitatory, then 1 $\rightarrow$ 15 excitatory with 0 inhibitory.

inherent preference to be in a refractory "1" or non-refractory "0" state, corresponding to a spiking probability of 0.5, and thus to set the interesting dynamics regime in the middle of the available bias values. Due to both the fixed pattern noise and FG variations, this was very weakly enforced, and so the appropriate V_{thresh} value was only heuristically selected.

Due to an imbalance in the relative strength of the excitatory and inhibitory noise, when noise rates are increased past the 300Hz minimum, the bias corresponding to a 0.5 spiking probability shifts. In order to try to emulate that only the temperature in the underlying Boltzmann distribution is being increased, the shift should be minimised as best possible. This can easily be seen from equation 2.16, where the input corresponding to a spiking probability of 0.5 would have a 0 in the exponent (though not necessarily corresponding to 0 bias input in our activation functions). Since a temperature increase in the abstract domain would result in linear scaling in all of the exponents, the spiking probability at this point would be invariant, and thus the 0.5 spiking point should not shift.

In order to implement this so called "shift compensation", in simulation the tuning of the inhibitory noise rate was chosen as a means of counteracting this shift [Kor17]. However, since the 0.5 point was found most often to be shifting towards greater inhibitory bias, and to avoid increasing the rate of either noise input rate lest OTA saturation regimes accidentally be entered, here the excitatory noise rate was chosen as a means of tuning/modulation instead. Again, due to the fixed pattern noise and FG variations, as well as an apparent non linearity in the shift at different rates, this was in practise difficult to achieve or optimise. Thus a simplistic, heuristically found solution was employed, whereby the excitatory noise rates ν_{ex} are a linear function of the inhibitory noise rates ν_{in}, with 300Hz inhibitory mapping to 300Hz excitatory, and 900Hz inhibitory to 700Hz excitatory. The results from a shift compensated activation function with varying noise input rate are shown in Figure 4.25 and 4.26 for gmax 150.

No widening of the magnitude expected occurs, and remains to be explained. In order to try to remedy this, multiple parameter variations were explored, however none yielded the expected widening. This included using a gmax value of 50 such that OTA saturation effects may be completely eliminated, or using doubled or halved synaptic time constants (as permitted for sampling by [Kun16]). Two parameters variations which may shed light upon the lack of widening however were when V_{thresh} was raised or lowered to confine the sigmoid dynamics to either the excitatory or inhibitory bias region respectively. The results are shown in Figures 4.27 and 4.28 respectively. When the sigmoid dynamics are confined to the excitatory bias regions, a thinning is observed, corresponding to the neuron paradoxically becoming more deterministic with increasing noise rate, even well before OTA saturation is expected to occur. With the sigmoid dynamics confined to the inhibitory bias regions however, the expected widening occurs. This is highly contrary to that expected from theoretical considerations, as a lowering of V_{thresh} should be (ignoring the slight change in distance to the reversal potentials and assuming a degree of translation invariance) identical to an increase in E_l (and vice versa), which corresponds simply to a direct way to effect a change in the mean membrane potential, as per Equation 2.8. A potential source of the asymmetry could be the asymmetry in the strength of inhibitory/excitatory synapses (as seen, for example, in the need shift compensation), however even in the worst case of nonlinear asymmetry, this would result in a nonlinear horizontal stretching of the activation function from the left to the right. A complete change in the widening behaviour however, is unexpected and yet to be explained.

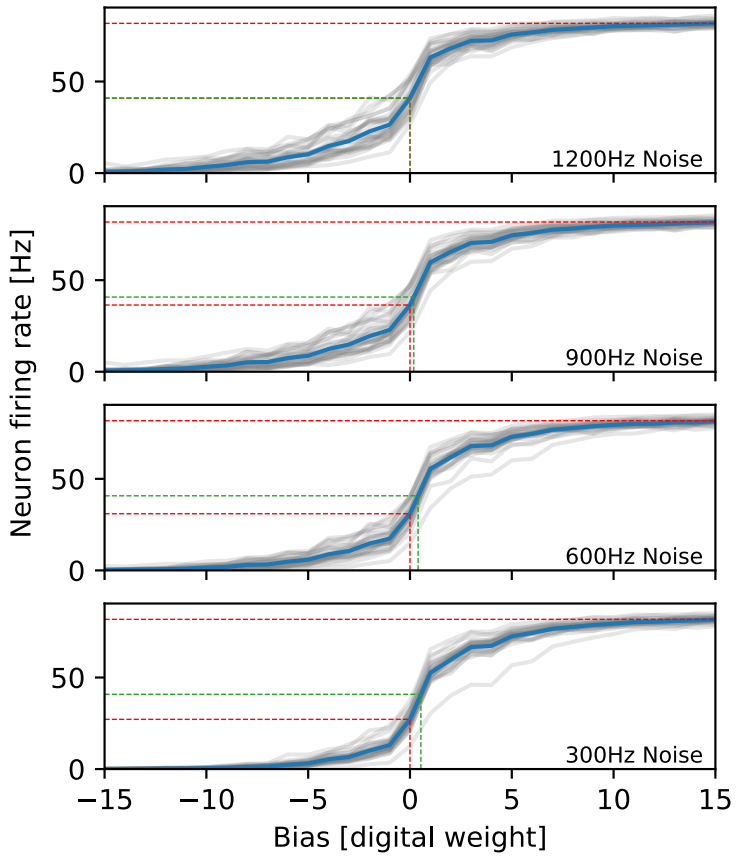

Figure 4.25: Shift compensated activation functions on hardware with modulated Poisson noise input rate. The network setup is described in Figure 4.24. The denoted rates are the inhibitory rates. If shift compensation has been implemented (as in this figure), the excitatory rates are linearly scaled like the rate pairs (ν_{in}, ν_{ex}) (300Hz, 300Hz) and (1200Hz, 900Hz). All data in this figure is from the same HICANN and same neuron, so the fixed pattern noise of no consequence here. The particular neuron was selected for its "good" sigmoidal shape. Each grey line (30 in total) represents a completely distinct run on the hardware, and so each involves a new setting of the floating gates. Within each grey line run, all the biases and weights were swept multiple (5) times, in order to find an associated error in the spiking rate arising purely from the inherently stochastic nature of spiking here. However the variance arising therefrom was found to be negligible compared to the variance arising due to FG variations, and as such is not shown. Thus the variation between grey lines is purely a consequence of FG variations. Each grey line is the mean over these digital re-sweeps (which have negligible variation), and the blue line is the average over all grey lines, and so is an average over FG variations. The two red dashed lines show the spike rate at 0 bias and the maximum measured spike rate. The green dashed line shows the bias at 50% of the maximum spike rate, and so corresponds to a 0.5 spiking probability. Ideally, the 0 bias red dashed line and green dashed line would be superimposed upon each other. This figure is for gmax 150, and so saturation has been avoided at least up to 900Hz. We thus expect to see a $\sqrt{3} \approx 1.7$ times increase in the activation function width. Apart from a slight lifting for the inhibitory biases, no such widening is observed, and remains to be explained.

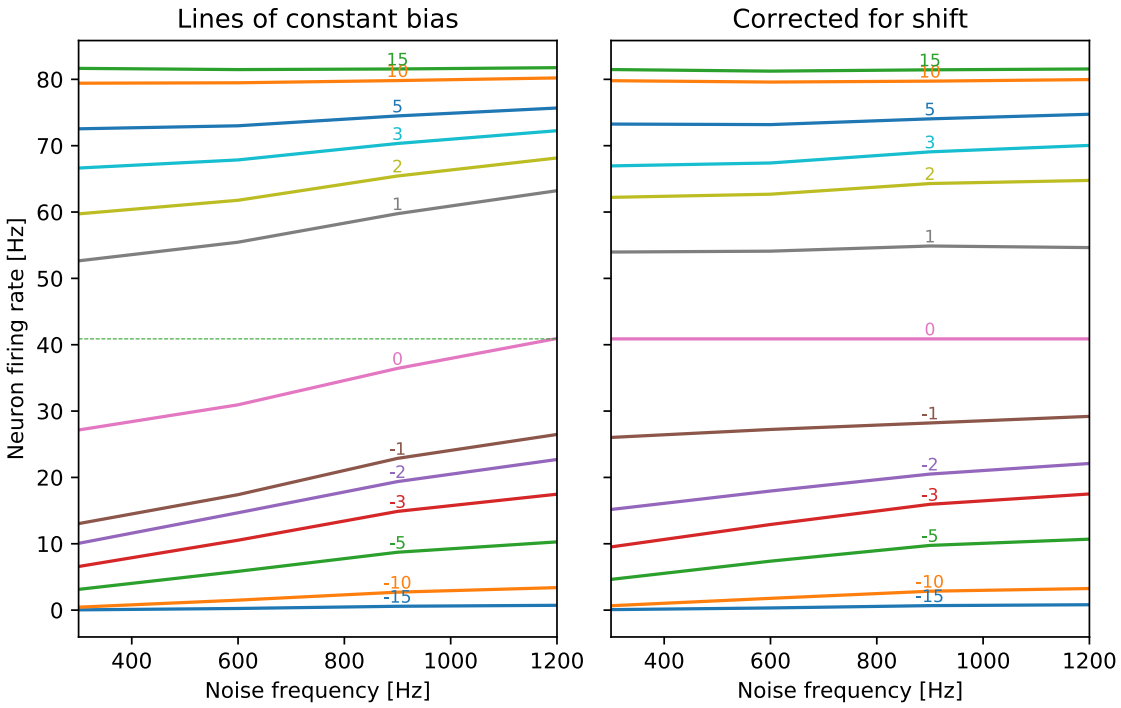

Figure 4.26: An alternate form of the activation function with rate modulation showed in Figure 4.24. **Left** is the spiking rate for lines of constant bias (bias weight values labelled on the lines) with increasing noise rate. A widening of the activation functions would result in a convergence of these lines, corresponding to different biases (the deterministic element) having a lessened impact on the spiking rate. Ideally, the 0 bias line would be at 0.5 spiking probability (shown by the dashed green line), and thus with increasing noise rate, if shift compensation were perfectly implemented, would have a spiking rate invariant of noise frequency. **Right** is the same as the left, except where it is falsely imposed that the ideal shift compensation has been realised. That is, 0 bias is defined to be at 0.5 spiking probability (for each frequency), and all other biases are defined with respect to this false 0 bias point. Any unwanted shifting of the activation function is therefore removed, and any widening should be easily visible in the convergence of the (false) bias lines towards 0.5 spiking probability. Since there is very little appreciable convergence (except again a slight lifting of the inhibitory bias lines), this figure more clearly shows that the expected widening for increasing noise input rate is not occurring.

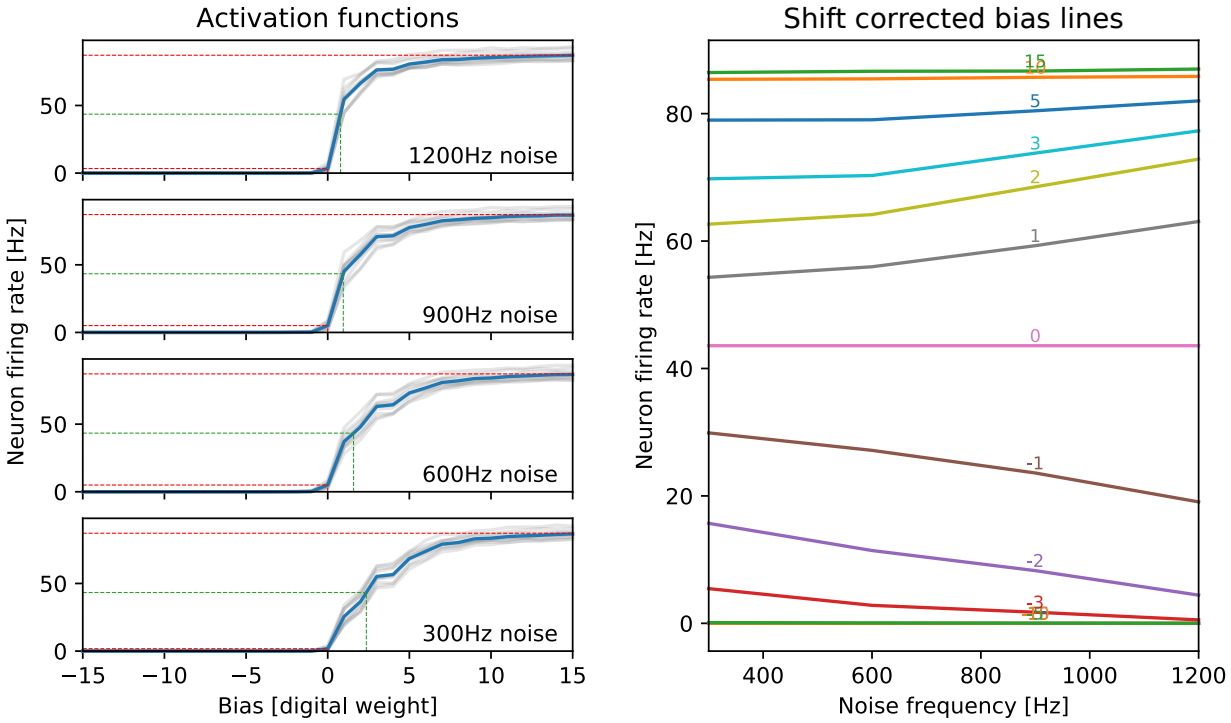

Figure 4.27: Activation function on hardware with a modulated Poisson noise input rate, where the sigmoid dynamics have been confined to the excitatory bias range by raising V_{thresh}. The figure construction is the same as in Figure 4.25 and 4.26. Besides the lack of shift compensation, the altered V_{thresh} value and that the bias neurons here spike at approximately double the rate, the setup is identical to 4.25, where the sigmoid is centered at 0 bias, and no widening is observed. Here the activation function paradoxically thins with increasing input noise rate, as evidenced by the divergence of the shift corrected bias lines.

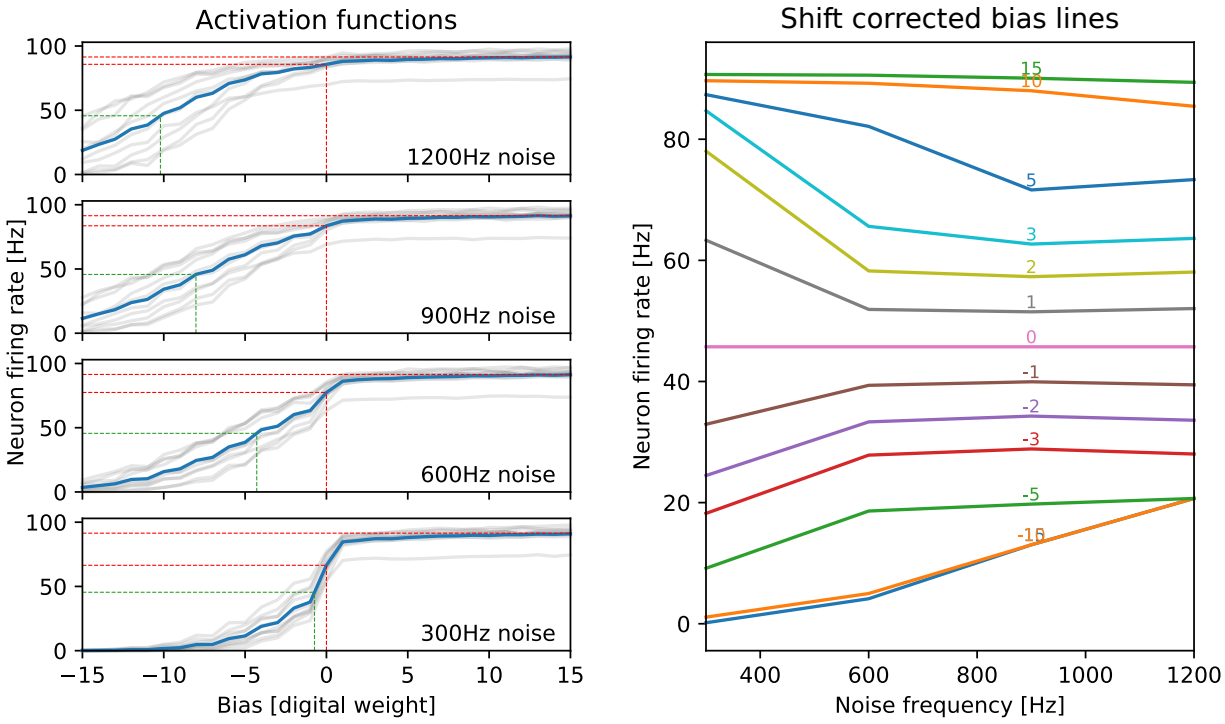

Figure 4.28: Activation function on hardware with a modulated Poisson noise input rate, where the sigmoid dynamics have been confined to the *inhibitory* bias range by lowering V_{thresh}. Figure construction and network setup is completely identical to Figure 4.27, except with a lower V_{thresh}. Here, as expected, the activation function widens with increasing noise rate, and is further evidenced by the strong converging of the shift corrected bias lines. The strange behaviour of the strong inhibitory bias lines at high frequency lines is due to the premature clipping of the sigmoid shape.

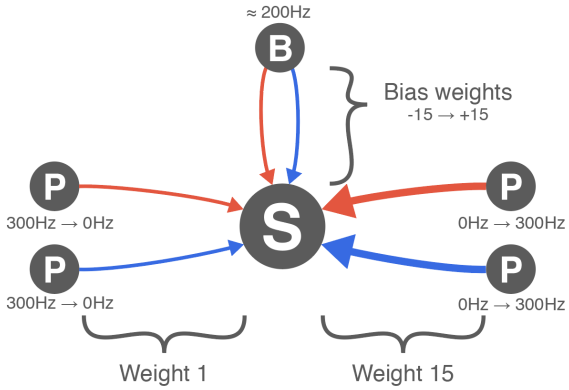

Figure 4.29: Setup for finding single neuron activation functions, where activation function widening is to be achieved by modulation of Poisson noise input *weight*. Diagram elements have the same meaning as in Figure 4.24, and input splitting to avoid OTA saturation was implemented here also (not shown). The total noise input rate to the sampling neuron was kept constant at 300Hz, but with a varying proportion of the 300Hz coming from Poisson sources connected at digital weight 1 or 15. The frequencies were thus swept in opposite directions as shown. For each frequency value, the bias weights were swept as described in Figure 4.24.

4.4.2 Widening via noise weight modulation

Since an activation function widening was not achieved on the hardware as the noise rate was modulated, we thus look to modulating the noise weight instead. In order to maximise the allowed weights, we use the 300Hz minimum required noise frequency for sampling [Kun16]. The saturation regions of Figure 4.20 are then revisited, and the highest gmax value for which digital weight 15 does not saturate either synaptic input was chosen. Although a simple way to modulate the noise weight would be to simply reconfigure the digital noise weight values on chip within a run using the digital configurator, performing a digital reconfiguration whenever a temperature change is required was deemed too slow, especially if high-frequency temperature changes (e.g./ as a solution to mixing) are sought. Instead, the network was set up as in Figure 4.29, whereby a sampling neuron is connected to *two* Poisson noise pairs (disregarding splitting), where one pair (excitatory and inhibitory) is connected permanently at weight 15, and the other at weight 1. Weight modulation was then achieved by only ever having one of the Poisson noise source pairs active at 300Hz, while the other is at 0Hz. In actuality, a range of frequencies was swept such that the total input frequency was 300Hz, such that the 300Hz transitions from completely at weight 1 to completely at weight 15, in an attempt to regain a degree of continuousness of the settable temperature without requiring separate noise source pairs at intermediate weights. The results are shown in Figure 4.30, where a considerable widening is achieved. To quantify the widening, we do not measure the activation function widths α directly, as many neurons showed an asymmetry in sigmoidal shape from the left inhibitorily biased flank to the right excitatory. Furthermore, our x-axis independent variable is not the mean membrane potential, as is the case in theory, but rather the bias to the neuron, which is only a proxy therefor. We will instead measure the network temperature change in the BM regime more directly, by observing how the weights and biases scale, as will be seen in Section 4.5.

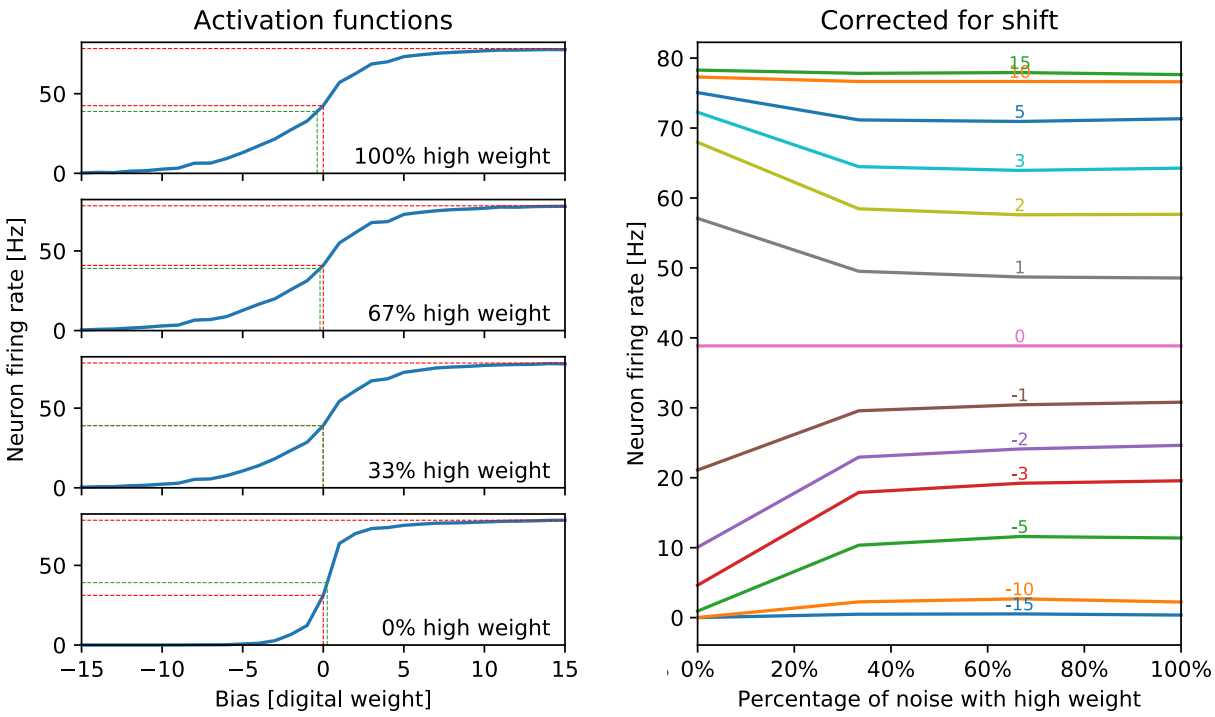

Figure 4.30: Activation function with weight modulation. Figure construction is the same as in Figure 4.25. Network setup is described in Figure 4.29. The weight is modulated by varying the frequency of noise between two Poisson source pairs at weight 1 and weight 15, such that the total input frequency is a constant 300Hz. The listed percentages are thus the percentage of the 300Hz noise that originates from a Poisson source connected at weight 15. There is considerable widening of the activation function from all 0% to 100%, as is to both expected and desired, and is made clear by the convergence of the shift-corrected bias lines. It should be noted that the maximum widening is largely reached when the percentage of high weight noise is very low, and so lends itself to the idea that activation function widening is dominated by the weight of input spikes, or rather the largest weight among the input spikes.

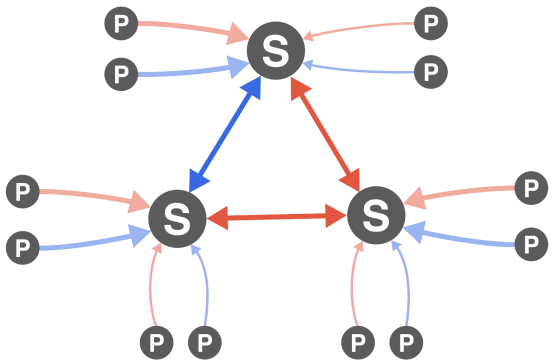

Figure 4.31: Setup for recording the state probability distribution of a fully connected randomly initialised sampling network, as the weight of Poisson noise is modulated. Symbols have the same meaning as in Figure 4.12, with the addition that thick arrows denote a connection of weight 15, and thin of 1. The weight is modulated by having only one pair of Poisson noise sources input spikes active at 300Hz at any one time, for each sampling neuron. The connections between sampling neurons are symmetric and randomly selected to be either inhibitory or excitatory. 6 sampling neurons were used (3 shown here). The noise input connections have been dimmed to avoid visual clutter.

4.5 State probability distribution changes under noise weight modulation

We wish to see how the state distribution changes as the noise weight is modulated, as an increase in noise strength should result in an increase in the temperature in the abstract Boltzmann regime. Rather than training, a sampling network was set up with randomised weights as described in Figure 4.31, and where each sampling neuron may receive Poisson noise at either weight 1 or 15. The neurons to use for sampling were picked based on their activation functions, picking those with similar maximum spike rates ν_{max}, so that a mean effective refractory time of $\frac{1}{\nu_{max}}$ could be reliably used to infer the network state. If the temperature in the corresponding Boltzmann regime is increased, as a direct result of Equation 2.15 we expect to see a flattening in the state probability distribution. The resulting state probability distributions are shown in Figure 4.32. From the "cold" noise weight 1 distribution to the "hot" noise weight 15 distribution a definite flattening occurs, indicating that tempering has been achieved to a degree. Although flatter, the hot distribution resembles the cold (the states remain roughly sorted in probability), indicating that the sampling distribution may be transforming as would be expected from an increase in temperature in the abstract regime, since as per Equation 2.15 the order of states by probability would remain the same, since the associated energies of the states $E(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}^T\boldsymbol{z}$ are temperature invariant. Boltzmann machines were then fitted to both distributions separately, and the resulting fitted parameters are shown in Figure 4.33. We see that the fitted weights transform by scaling linearly from the cold to the hot distribution, as a temperature change in a Boltzmann machine should manifest. Though there are only few fitted biases (6, one per neuron), they do not transform in a clear manner, with some inhibitory biases becoming excitatory and vice versa. This was to be expected, as no reliable shift compensation was applied, and any shift in the 0.5 spiking probability point (as in Figure 4.30) would manifest as a change in the neuron's internal bias.

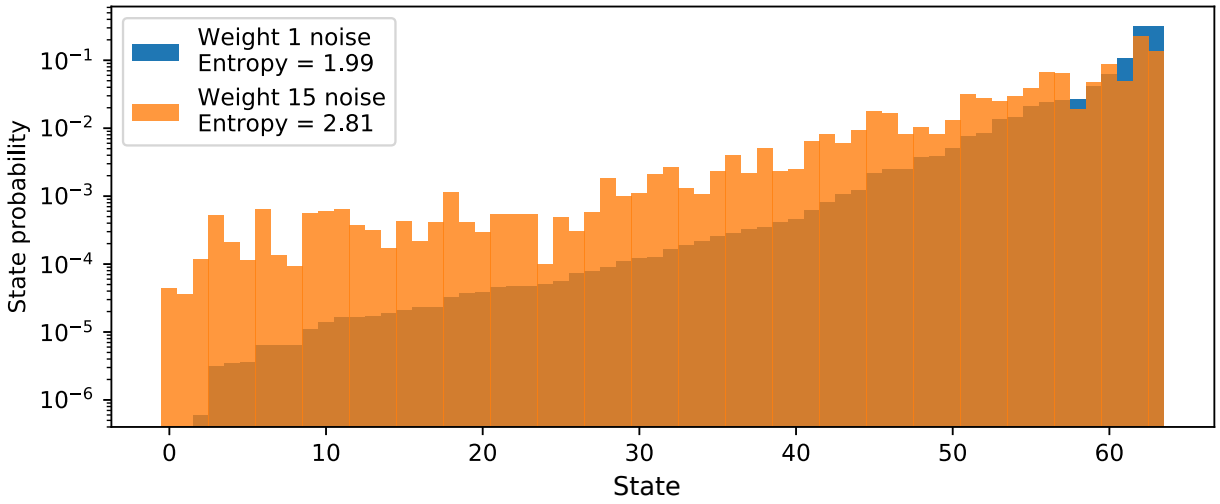

Figure 4.32: Sampling network state probability distributions for two different Poisson noise input weights on hardware. The binary network state vector $\boldsymbol{z}$ at time t is inferred by which neurons have spiked since $t = \tau_{ref}$. The states have been sorted with respect to the "cold" weight 1 distribution. A definite flattening of the distribution occurs when changing to the "hot" weight 15 distribution, as is evident from all states apart from the 6 most probable receiving an increase in probability, as well as the Gibbs entropy (where each state is an individual microstate) increasing from cold to hot. The hot state distribution resembles the cold, in that there is still a general slope in the same direction, meaning that the sampling distribution is transforming at least somewhat as is expected from a temperature change.

4.6 Mixing aided by noise modulation

4.6.1 Mixing with tempering via weight modulation

To simulate a mixing problem to be solved by tempering, a network was created as described in Figure 4.34, consisting of two discrete clusters of neurons. Within each cluster, there are only excitatory connections, but between clusters the connections are only inhibitory. Either cluster may thus be active (and very stable) with a high probability, but due to the strong inhibitory connections between the clusters, the activities of the neuron clusters should be mutually exclusive. This thus poses a mixing problem, since exclusively either cluster may be active with a high probability, but to change to a network state where the other cluster is active involves transitioning through states where both are partially active, and so have very low associated probabilities and high energies. To quantify the activity of each neuron cluster, the mean binary neuron state across each cluster is taken. We thus look at the activity of each cluster as the noise weight is varied as described in Section 4.4.2. The results are shown in Figure 4.35. When the network is subject only to the "cold" weight 1 noise, only one neuron cluster is active, while the other does not spike at all, and no mixing occurs. When some (16%, 33%) "hot" weight 15 noise is introduced, mixing between the two modes is allowed to occur, while the clusters still retain a degree of mutual exclusivity. When the network is subject purely to weight 15 noise, the mutual exclusivity is lost, and the activity of the clusters fluctuates freely without any obvious anticorrelation. A weight variation scheme is then employed whereby mixing is facilitated by intermittently switching to high weight noise to allow the the activities

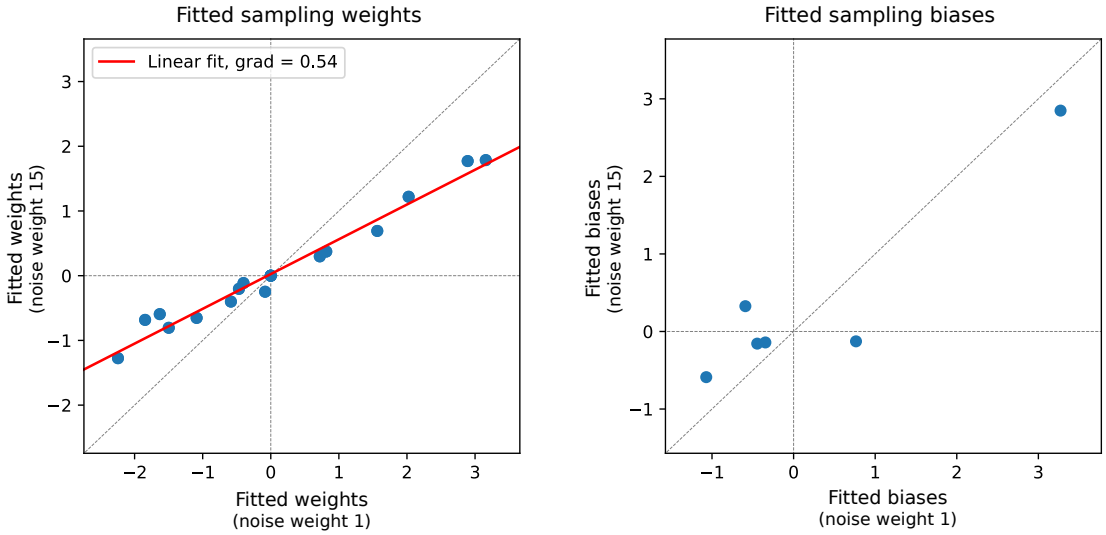

Figure 4.33: Fitted abstract Boltzmann parameters for a sampling network state distribution for two different Poisson noise input weights on hardware. The parameters are fitted to the "cold" noise weight 1 and "hot" noise weight 15 state probability distributions from Figure 4.32 separately. The fitting minimises the square loss between the measured probabilities and probabilities determined by Equation 2.12, with the weight matrix $\boldsymbol{W}$ and bias vector $\boldsymbol{b}$ to be fitted, where $\boldsymbol{W}$ is symmetric with zeros along the diagonal. Here, the fitted abstract weights in the hot regime are plotted against those in the cold **(left)**, as well as the fitted biases **(right)**. We see a definite linear scaling in the fitted weights, which matches how the weights should transform under a temperature increase. Though there are not enough bias values to draw a clear conclusion, they do not seem to transform in any clear manner. This was to be expected, since no shift compensation was applied, and thus different biases were introduced by the modulated noise weight.

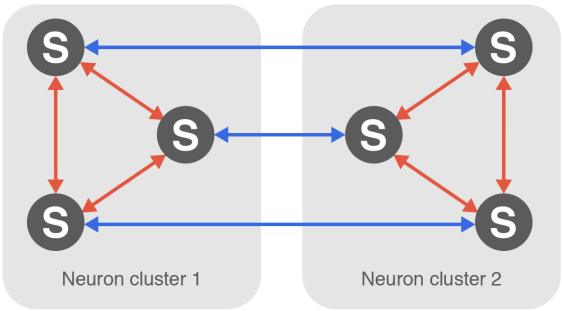

Figure 4.34: The network setup for simulating a mixing problem to be solved via noise weight modulation. Symbols have the same meaning as in Figure 4.12. To each sampling neuron, private Poisson noise is generated and connected as in Figure 4.29. The neurons are connected to realise two neuron clusters whose activity should be mutually exclusive, but the probability of exclusively one of the clusters firing is very high and approximately equal. Each neuron cluster contained 8 sampling neurons.

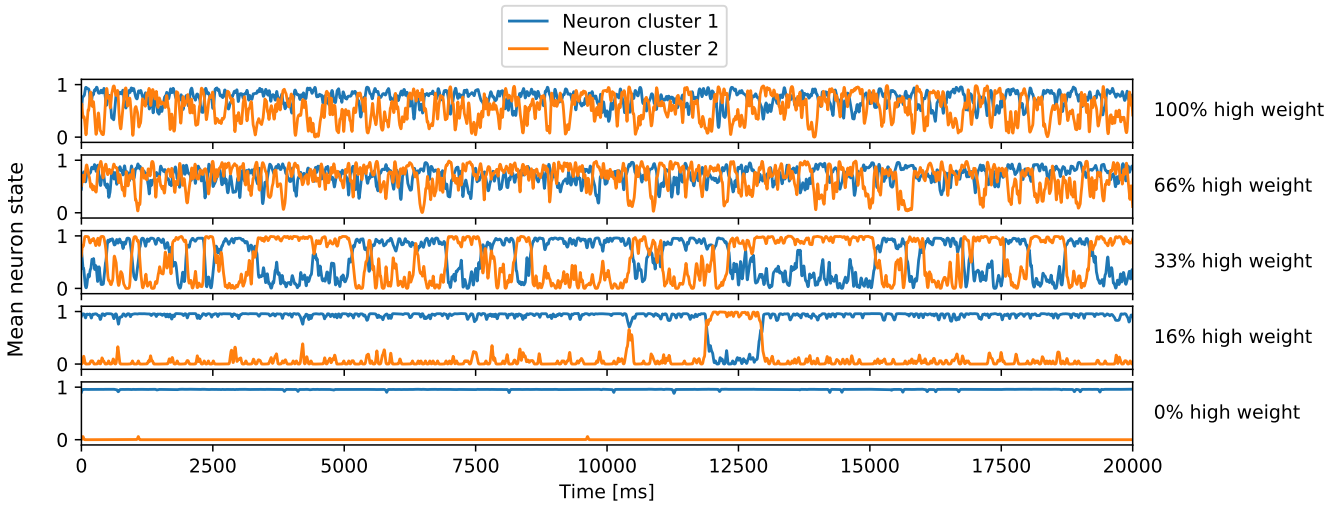

Figure 4.35: Mixing between two high probability sets of states, corresponding to exclusively one of two neuron clusters spiking, for different Poisson noise input weights. The network setup is described in Figure 4.34. The blue and orange lines denote the mean binary neuron state within each cluster, and hence the cluster's spiking activity. For 0% high weight, corresponding to 300Hz at weight 1, there is no mixing between the two modes, corresponding to the active cluster never switching. With increasing noise weight however, the network is able to transition between the two sets of states, as evidenced by a swapping in which neuron cluster is active. For weights 66% and 100% high weight, the inter-sampling neuron connections are overridden by the Poisson noise, as the neuron clusters' activities cease to be mutually exclusive. For the 100% weight scenario, there is the possibility that saturation effects are occurring for cluster 1, or that too great an accidental excitatory bias is being induced by the lack of shift compensation.

of the neuron clusters to fluctuate freely. The results are shown in Figure 4.36. Within any cold region, the network remains stuck in one of its two modes of only one of the clusters being active. Within a hot region, the activities of the two clusters fluctuate freely as expected. After a hot region, the network collapses to one of the two modes, and indeed does not switch to the same mode after every hot region. Thus the tempering induced by weight modulation has successfully enabled mixing.

4.6.2 Mixing with noise rate modulation

Despite the fact that widening of the activation functions was not achieved with noise rate modulation, we apply it to a sampling network and mixing problem regardless. This is motivated by the fact that although the weight modulated activation functions stop widening past approximately 30% of noise being high weight (Figure 4.30), the mixing between the two modes in Figure 4.35 becomes increasingly erratic with increasing weight still. There is thus the possibility that mixing may be aided without the need for the activation functions to widen. Two networks were set up, where the sampling neurons were connected to one pair of Poisson input sources[8] as in Figure 4.12, and then the sampling neurons were connected to each-other as in Figure 4.31 or 4.34, for seeing how the sampling distribution changes and whether mixing is

[8]As always, this is disregarding that the Poisson noise sources are actually split up into 8 sources to avoid OTA saturation.

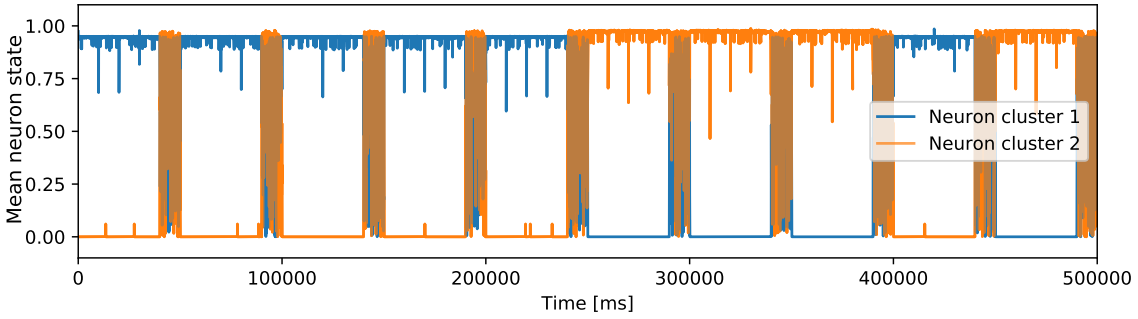

Figure 4.36: Mixing between two modes on hardware, facilitated by an intermittently high noise weight. The network setup is shown in Figure 4.34. The neurons are exposed to "cold" Poisson noise at weight 1 for 40,000ms, and then "hot" Poisson noise at weight 15 for 10,000ms. Within the cold regions, the activity of the two clusters is mutually exclusive. Within the hot regions, the mutual exclusivity is lost (as in Figure 4.35), and the activity of both networks fluctuates freely (as shown by the intermittent high frequency fluctuations, which appear as the thick vertical bars). After the some of the hot regions, the active cluster has swapped, indicating that mixing has successfully been facilitated. Within each 40,000ms cold region, the spike trains were actually generated as 4 spike trains for 10,000ms each. The thin periodic dips in activity thus occur when switching to a new spike train being generated. Although this should have no effect, since the spike trains are generated in software before running, the dips may thus be attributed to possible edge effects during spike train generation, where the spikes may momentarily be generated with too high or low a frequency.

facilitated respectively. The results are shown in Figures 4.37 and 4.38 respetively. The results show that the state distribution does not flatten as with a temperature change, but is instead altered to a very different BM. Despite this, mixing between the two modes is still aided. The explanation for this, and the implications thereof, are discussed in Chapter 5.

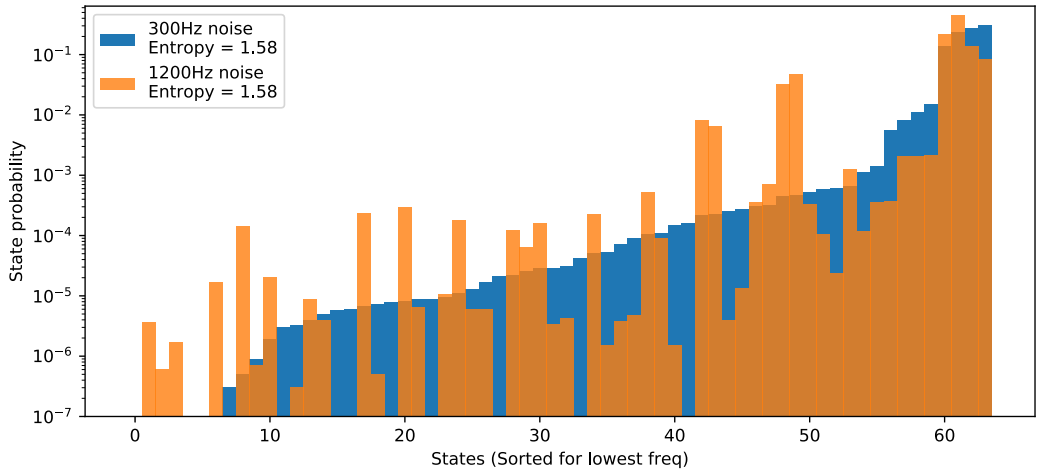

Figure 4.37: The state probability distribution of an arbitrary sampling network on hardware as the noise rate is modulated between 300Hz and 1200Hz. The states are sorted according to the 300Hz states. Since a widening of the activation function via noise rate modulation was not achieved, we do not expect to see a coherent flattening associated with a temperature increase here. Indeed we see only that the states are arbitrarily distorted, with no strong resemblance to what would be the "cold" 300Hz distribution. Furthermore, the Gibbs entropy of the distributions does not increase, indicating that despite the distortion, the temperature in the 1200Hz case has not increased.

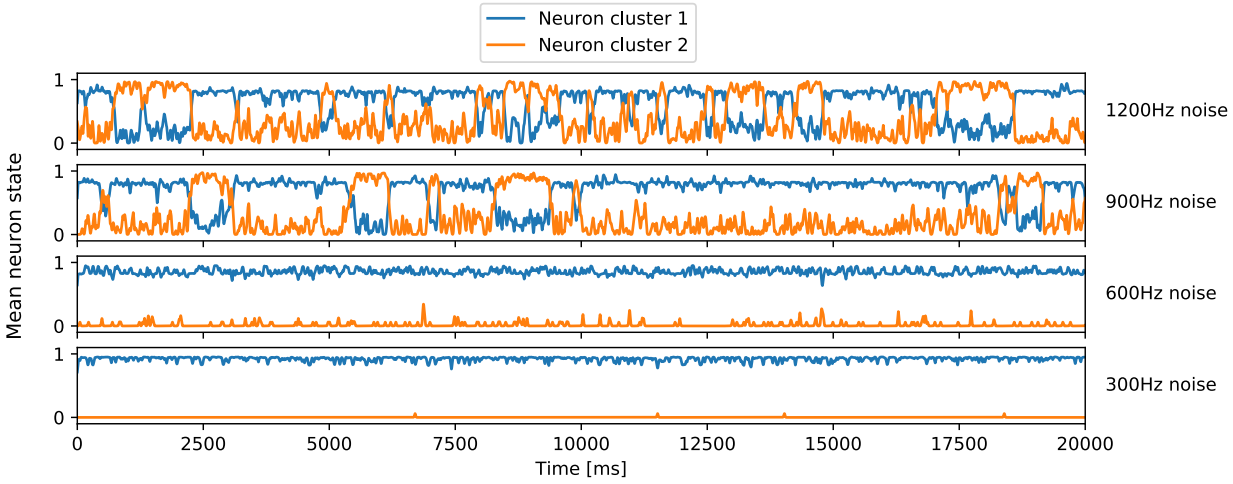

Figure 4.38: Mixing between two high probability modes on hardware, corresponding to two neuron clusters with mutually exclusive activity, for different Poisson noise *rates*. Although we have repeatedly seen that the higher rates are not associated with a higher temperature increase, mixing is still facilitated by the increased noise rate.

Chapter 5

Discussion

All tempering experiments were carried out with Poisson noise externally generated for each sampling neuron, as that it was the theory presumes for the OU (Ornstein-Uhlenbeck) processes to occur. This was done to avoid an additional layer of separation and thus lack of control of the sampling neurons.

We then presented a method for determining the limits of noise input to a hardware sampling neuron, where the most important hardware limitation was OTA saturation. We were thus able to determine, at each frequency, at what possible noise weights saturation could be safely avoided (Figure 4.20). Though this experiment was carried out for a single neuron only, and so did not take the so-called fixed pattern noise into account, it was nonetheless used as a guideline for all subsequent experiments.

On the lack of activation function widening due to rate modulation

In order to be able to make a direct comparison with a result from simulation, the width of the free membrane potential was plotted as the weight and rate of noise input was varied (Section 4.3.2. The results qualitatively agreed with those found in simulation, with the width peaking with increasing input rate. A certain marked deviation from the theory however occurred after this peak, whereby in simulation (in accordance with theory) the width then falls to 0, whereas we found instead that the width plateaus. Given the limited validity of the geometric interpretation of the link between membrane potential widths and activation function widening as discussed in Section 2.1.3, this is not of direct consequence to the subsequent lack of activation function widening. However, that both these deviations from simulation occur at similar rates could indicate that they are caused by a similar phenomenon.

Activation function widening is the critical mechanism by which a temperature change manifests in LIF sampling. Figure 4.25 shows that activation function widening by modulating the noise rate was found to not be successful on the hardware. Since in simulation the widening was found to obey a proportionality in $w^2\nu$ over a much larger frequency and weight range than here explored, we thus look to possible hardware-specific explanations. Due to the asymmetry in the implementation of the inhibitory and excitatory synapses on the hardware, and that the behaviour of the activation function varied greatly when it was confined to either excitatory

or inhibitory biases, a closer look at the responsible OTA circuits is thus warranted, which is beyond the scope of this work.

A possible interim experiment to isolate the issue further could be to remove the spiking behaviour of a neuron used in activation function experiments entirely, and to infer an activation function from its analogue membrane trace, by assuming that we are effectively observing a spiking neuron's effective membrane potential. This would have the immediate advantage of being able to measure the mean membrane potential directly, rather than inferring it from the amount of bias the neuron is subjected to (though this discrepancy could not cause a change in widening behaviour). In any case it may shed some light on the situation, for example to check whether the neuron is still following an OU process.

Successful tempering via noise weight modulation

We thus switched to effecting activation function widening by increasing the weight of Poisson noise to a sampling neuron (Figure 4.30). We found that the probability landscape flattened as desired, and the weights closely followed a 0.5x linear scaling, implying a doubling of temperature (Figure 4.33). The biases did not follow the same trend, which was to be expected since no robust shift compensation was applied to counteract the arbitrary unwanted shifting of the activation function as the input noise changes.

We then applied this tempering to a network with two distinct modes, where for low-weight noise no mixing between the states is observed (Figure 4.35). In order to try to regain a degree of continuity in the settable temperature without having separate Poisson sources at every digital weight, we substitute, for example, 300Hz of moderately weighted noise with 150Hz of high weight and 150Hz of low weight noise. Though this does not have any mathematical backing, it is still useful for observing what happens when transitioning between the two extremes. We see that with increasing noise weight, mixing between the two states is indeed facilitated (Figure 4.36).

Multiple mixing facilitation mechanisms

We note however that the activation functions stopped widening further when approximately 33% of the noise was high weight, however the mixing between the two modes continues to become increasingly erratic with higher weight. We thus propose that there are two mechanisms contributing to the increased mixing between modes:

1. An initial widening of the activation functions up to 33% high weight noise encompasses a temperature increase, causing the probability landscape to flatten as per the fundamental defining BM equations. This is what we refer to as tempering.

2. Any other distortive effects upon the sampling distribution. We may simplistically model this as a large arbitrary bias, here determined by the uncompensated arbitrary shifting of the activation functions, causing the underlying sampling distribution to change arbitrarily.

Since these other distortive effects effectively push the network to sample from a completely different arbitrary BM, and it is unlikely that two networks will have minima and maxima occurring in the same places, these distortive effects may allow the network to escape from probability wells present in the original BM. This has the downside compared to tempering however, that the subsequent network evolution is not determined by the original BM, and we are simply hoping that the local landscape in the altered BM is flatter than in the original.

This treatment of other distortive effects as an increasing bias exerted upon the network is validated in Figure 4.35, where past 33% high weight noise (approximately where a maximum in activation function width, and thus temperature occurs) the mutual exclusivity between the two modes that defined the network is completely lost. Since we believe that we are at most getting a 2x increase in temperature, the states corresponding to mutual exclusivity should still dominate. If we suppose that we are actually in a regime of infinite temperature, such that all states are equally likely, then each neuron cluster should oscillate around 0.5 probability. This is not the case, and instead one of the neuron clusters remains almost completely active, while the other oscillates freely, indicating that the sampling network has been greatly disturbed, but not in a manner expected from tempering.

This is especially evident when considering the same experiments, but with noise rate rather than weight modulation. Since we did not observe any activation function widening, we expect the tempering mechanism to not be present. In comparison to the flattening, but retaining of the overall ordering of states when tempering via weight modulation, the state distribution under rate modulation changes to not resemble the original (Figure 4.37). Furthermore, the associated entropy of the distribution does not increase, supporting that the distribution changes are due to non-tempering distortions only. When applied to the aforementioned dual-mode mixing problem, mixing was also facilitated as expected (Figure 4.38).

Though these distortion effects may be useful when tempering is not sufficient to allow the network to escape from especially deep probability wells, they should in general be avoided. This is due to the fact they effectively allow the network to temporarily sample from a (fixed) separate arbitrary BM. When switching back to the original BM, a systematic bias could be induced based upon where the network emerges from the 2^{nd} BM. An extreme example of this would be if the distortive bias were aligned with a particular mode, causing that mode to be over-represented during sampling.

From the hypothesis of the two mixing mechanisms, we may then conclude that in order to minimise these biasing distortion effects, when implementing weight modulation based tempering, a sampling neuron should only be exposed to as much high weight noise as is needed in order to reach the maximum activation function width, since any further high weight noise results in an increase in the non-tempering distortive effects *only.*

Since a maximum activation function widening was here achieved by using the maximum range of settable digital weights, the widening could possibly be increased by drawing connections from the high-weight noise to the sampling neurons multiple times, to increase the PSP effected per spike and thus the effective noise weight further. Alternatively, multiple gmax values could be set per HICANN to similarly increase the range of noise weights. Since a 2x temperature increase has been demonstrated, tempering could now be employed using a particular temperature variation scheme with trained networks, as done in simulation [Kor17], to see if a similar increase in generative performance can be achieved.

On spike loss

Though spike loss was not a focus of this project per se, and was instead encountered as a negative effect to be avoided, we found that even for moderately sized experiments (100 neurons on 1 HICANN), the spike loss for a significant proportion of these neurons was close to 100% (this did not affect the subsequent sampling experiments, as very few neurons were placed on the hardware). We heuristically argued that this spike loss must be predominantly occurring on readout, however it would be useful to check this, and quantify the degree of spike loss at different points in the hardware stack. One experiment to check that the spikes are not being lost on chip would be to check that the PSPs from sequentially connected bursting noise neurons are seen on a target analogue recorded sampling neuron.

A modulated noise network was created but unused

A noise network where the frequency could be modulated from 300Hz $\rightarrow$ 1200Hz noise output was created, with the purpose of using it in rate modulation induced tempering experiments. Since tempering was not achieved using rate modulated Poisson generated noise, the noise network was never used as a noise source for sampling neurons. If tempering via rate modulation is eventually shown to work on the hardware, then the Poisson generated noise may then be swapped out for noise generated from this noise network.

We assume that the noise network size N is large enough in comparison to N_{pre} that effects arising from a finite network size have been avoided, and thus that the network (barring synapse, spike loss etc.) may be scaled up for use in larger experiments, if rate modulation is eventually found to be effective. We have also assumed that the PSP due to individual excitatory stimulus spikes has been made low enough, such that the stimulus does not induce any correlation in the noise neurons. Due to spike loss, the noise spike trains could not be compared for correlations directly, however the fact that the max noise network output frequency of each neuron spiking at $\approx$ 80Hz required a comparatively fast $\approx$ 1000Hz stimulus, may indicate that the PSP was indeed made small enough and that the spike trains should indeed be uncorrelated. If this is not the case, then when used as high-frequency noise source during rate variation experiments, the noise correlation should manifest as an apparent increase in the excitatory weight between sampling neurons, relative to when Poisson noise is used.

A preliminary test was done to see if the noise network could be modified to work as a noise source in rate modulation experiments. V_{thresh} was set above E_l, so that in the absence of external stimulus, no neurons should spike, and thus the network could effectively be switched on and off. However, even when V_{thresh} was set 10mV above El, the noise network still output noise at $\approx$ 50Hz. Due to the fact that activation widening is achieved even when the amount of high-weight noise is comparatively low, the noise network was deemed unsuitable for weight modulation experiments.

Chapter 6

Conclusions

In this work we have successfully realised tempering in a Boltzmann machine inspired stochastic spiking network on the BrainScalesS-1 neuromorphic hardware system. In our setup, a relatively small proportion of the input noise frequency is replaced with noise of a higher weight, then by selecting neurons with a sigmoidal shape and that show widening behaviour, tempering is implemented for these neurons by changing between higher and the lower noise weight. Although the shift of the activation functions in the "hot" high weight regime should be minimised, this is only approximately achievable, and leads to the biases transforming arbitrarily when modulating the noise weight. On the other hand, the weights in the abstract Boltzmann regime are linearly scaled down as expected from a temperature increase. This resulted in a clear flattening of the network state distribution, and enabled mixing in a simplified two-mode sampling network.

Tempering via modulating the noise rate, in contrast to simulation, was not successful. Instead of the state distribution flattening, it was instead distorted non-trivially. This was however found to also facilitate mixing in the two-mode sampling network. As discussed in Chapter 5, this method should be employed with care, as there is the possibility for systematic biases to be induced due to the fixed arbitrary distortion.

As an outlook, tempering via weight modulation may now be used to improve the generative performance of trained sampling networks.

A noise network was designed for use in rate modulation experiments, with its output frequency able to be increased from the baseline required for sampling. If this network is to be adapted for tempering via weight modulation, it must be ensured that the noise network can be made completely inactive with certainty.

Bibliography

[AHS85] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. "A learning algorithm for Boltzmann machines". In: *Cognitive science* 9.1 (1985), pp. 147–169.

[Alc+19] Michael A Alcorn et al. "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4845–4854.

[Bau16] Andreas Baumbach. "Magnetic Phenomena in Spiking Neural Networks". Masterarbeit. Universität Heidelberg, July 2016.

[Ber+11] Pietro Berkes et al. "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment". In: *Science* 331.6013 (2011), pp. 83–87.

[Bre15] Oliver Breitwieser. "Towards a Neuromorphic Implementation of Spike-Based Expectation Maximization". Masterarbeit. Universität Heidelberg, 2015.

[BG05] Romain Brette and Wulfram Gerstner. "Adaptive exponential integrate-and-fire model as an effective description of neuronal activity". In: *Journal of neurophysiology* 94.5 (2005), pp. 3637–3642.

[BV07] Nicolas Brunel and Mark CW Van Rossum. "Lapicque's 1907 paper: from frogs to integrate-and-fire". In: *Biological cybernetics* 97.5-6 (2007), pp. 337–339.

[Bue+11] Lars Buesing et al. "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons". In: *PLoS computational biology* 7.11 (2011), e1002211.

[Dav+10] Andrew Davison et al. "A common language for neuronal networks in software and hardware". In: *The Neuromorphic Engineer* (2010).

[D+03] Peter Dayan, L Abbott, et al. "Theoretical neuroscience: computational and mathematical modeling of neural systems". In: *Journal of Cognitive Neuroscience* 15.1 (2003), pp. 154–155.

[Des+] Guillaume Desjardins et al. "Parallel tempering for training of restricted Boltzmann machines". In:

[Dol+18] Dominik Dold et al. "Stochasticity from function - why the Bayesian brain may need no noise". In: *arXiv* (2018). URL: `https://arxiv.org/abs/1809.08045`.

[Dru00] Daniel Drubach. *The brain explained*. Prentice Hall, 2000.

[Esl+14] SM Ali Eslami et al. "The shape boltzmann machine: a strong model of object shape". In: *International Journal of Computer Vision* 107.2 (2014), pp. 155–176.

[Fis+10] József Fiser et al. "Statistically optimal perception and learning: from behavior to neural representations". In: *Trends in cognitive sciences* 14.3 (2010), pp. 119–130.

[Gar09] Crispin Gardiner. *Stochastic methods*. Vol. 4. Springer Berlin, 2009.

[GG87] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *Readings in computer vision*. Elsevier, 1987, pp. 564–584.

[GSD12] Wulfram Gerstner, Henning Sprekeler, and Gustavo Deco. "Theory and simulation in neuroscience". In: *science* 338.6103 (2012), pp. 60–65.

[GD07] Marc-Oliver Gewaltig and Markus Diesmann. "Nest (neural simulation tool)". In: *Scholarpedia* 2.4 (2007), p. 1430.

[HSA84] Geoffrey E Hinton, Terrence J Sejnowski, and David H Ackley. *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, 1984.

[Jor+15] Jakob Jordan et al. "Deterministic neural networks as sources of uncorrelated noise for probabilistic computations". In: vol. 16. 2015, Suppl 1: P62.

[Jor+17] Jakob Jordan et al. "Stochastic neural computation without noise". In: *arXiv* 1710.04931 (Oct. 2017). URL: https://arxiv.org/abs/1710.04931.

[Kar14] Vitali Karasenko. "A communication infrastructure for a neuromorphic system". Masterarbeit. Universität Heidelberg, 2014.

[Klä17] Johann Klähn. "Training Functional Networks on Large-Scale Neuromorphic Hardware". Master. Universität Heidelberg, 2017.

[Kok17] Christoph Koke. "Device Variability in Synapses of Neuromorphic Circuits". PhD thesis. 2017. DOI: 10.11588/heidok.00022742. URL: http://archiv.ub.uni-heidelberg.de/volltextserver/22742/1/dissertation%7B%5C_%7D2017%7B%5C_%7D01%7B%5C_%7D02.pdf.

[Kor17] Agnes Korcsak-Gorzo. "Simulated Tempering in Spiking Neural Networks". Masterarbeit. Universität Heidelberg, 2017.

[Kug18] Alexander Kugele. "Solving the Constraint Satisfaction Problem Sudoku on Neuromorphic Hardware". Masterarbeit. Universität Heidelberg, 2018.

[Kun16] Akos Kungl. "Sampling with leaky integrate-and-fire neurons on the HICANNv4 neuromorphic chip". Masterarbeit. Universität Heidelberg, 2016.

[Kun+18] Akos F. Kungl et al. "Generative models on accelerated neuromorphic hardware". In: *CoRR* abs/1807.02389 (2018). arXiv: 1807.02389. URL: http://arxiv.org/abs/1807.02389.

[Len+18] Luziwei Leng et al. "Spiking neurons with short-term synaptic plasticity form superior generative networks". In: *Scientific reports* 8.1 (2018), p. 10651.

[Mar12] Henry Markram. "The human brain project". In: *Scientific American* 306.6 (2012), pp. 50–55.

[Mil12] Sebastian Millner. "Development of a multi-compartment neuron model emulation". PhD thesis. 2012.

[Pet15] Mihai A. Petrovici. "Form vs. Function - Theory and Models for Neuronal Substrates". PhD thesis. Universität Heidelberg, 2015.

[Pet+14] Mihai A Petrovici et al. "Characterization and compensation of network-level anomalies in mixed-signal neuromorphic modeling platforms". In: *PloS one* 9.10 (2014), e108590.

[Pet+15] Mihai A. Petrovici et al. “The high-conductance state enables neural sampling in networks of LIF neurons”. In: vol. 16. Dec. 2015, Suppl 1: O2.

[Pet+16] Mihai A. Petrovici et al. “Stochastic inference with spiking neurons in the high-conductance state”. In: *Physical Review E* 94.4 (Oct. 2016). DOI: 10.1103/PhysRevE.94.042312. URL: http://journals.aps.org/pre/abstract/10.1103/PhysRevE.94.042312.

[Sal10] Ruslan Salakhutdinov. “Learning deep Boltzmann machines using adaptive MCMC”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 943–950.

[SH09] Ruslan Salakhutdinov and Geoffrey Hinton. “Deep boltzmann machines”. In: *Artificial intelligence and statistics*. 2009, pp. 448–455.

[Sch+10] J. Schemmel et al. “A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling”. In: *Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (ISCAS”10)* (2010), pp. 1947–1950.

[Sch14] Dominik Schmidt. “Automated Characterization of a Wafer-Scale Neuromorphic Hardware System”. Masterarbeit. Universität Heidelberg, 2014.

[Sch+17] Sebastian Schmitt et al. “Neuromorphic Hardware In The Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System”. In: *Proceedings of the 2017 IEEE International Joint Conference on Neural Networks* (2017). DOI: 10.1109/IJCNN.2017.7966125. URL: http://ieeexplore.ieee.org/document/7966125/.

[Sch13] Marc-Olivier Schwartz. “Reproducing Biologically Realistic Regimes on a Highly-Accelerated Neuromorphic Hardware System”. PhD thesis. 2013.

[Sze+13] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).

[Wat94] Bill Watterson. *Homicidal Psycho Jungle Cat - A Calvin and Hobbes Collection*. Turtleback Scho. Andrews and McMeel, 1994. ISBN: 978-0-606-00099-4.

[Wun+19] Timo Wunderlich et al. “Demonstrating Advantages of Neuromorphic Computation: A Pilot Study”. In: *Frontiers in Neuroscience* (2019). DOI: 10.3389/fnins.2019.00260. URL: https://arxiv.org/abs/1807.02389.

[Yeg+15] Alper Yegenoglu et al. *Elephant–Open-Source Tool for the Analysis of Electrophysiological Data Sets*. Tech. rep. Computational and Systems Neuroscience, 2015.

Appendix

Noise neuron parameters		
Name	*Value*	*Description*
V_{gmax}	500	Digital current scale
C_m	1.0 nF	Membrane capacitance
V_{reset}	-60 mV	Reset potential
E_{leak}	-10 mV	Leak potential
V_{thresh}	-20 mV	Threshold potential
$E_{\mathrm{e}}^{\mathrm{syn}}$	60 mV	Excitatory reversal potential
$E_{\mathrm{i}}^{\mathrm{syn}}$	-100 mV	Inhibitory reversal potential
τ_{ref}	0 ms	Refractory time
τ_{m}	100 ms	Membrane time constant
$\tau_{\mathrm{e}}^{\mathrm{syn}}$	30 ms	Excitatory synaptic time constant
$\tau_{\mathrm{i}}^{\mathrm{syn}}$	4 ms	Inhibitory synaptic time constant
Noise network parameters		
Name	*Value*	*Description*
N	100	Number of neurons
N_{pre}	3	Number of presynaptic partners
w_{inh}	15	Inter neuron inhibition weight
w_{stim}	1	External excitatory stimulus weight

Table 6.1: Neuron and network parameters for the modulated noise network on hardware.

Sampling neuron parameters		
Name	*Value*	*Description*
V_{gmax}	500 (150)	Digital current scale
C_m	0.2 nF	Membrane capacitance
V_{reset}	-35 mV	Reset potential
E_{leak}	-30 mV	Leak potential
V_{thresh}	-25 mV	Threshold potential
$E_{\text{e}}^{\text{syn}}$	60 mV	Excitatory reversal potential
$E_{\text{i}}^{\text{syn}}$	-100 mV	Inhibitory reversal potential
τ_{ref}	10 ms	Refractory time
τ_{m}	1 ms	Membrane time constant
$\tau_{\text{e}}^{\text{syn}}$	10 ms	Excitatory synaptic time constant
$\tau_{\text{i}}^{\text{syn}}$	10 ms	Inhibitory synaptic time constant
Bias neuron parameters		
Name	*Value*	*Description*
V_{reset}	-40 mV	Reset potential
E_{leak}	0 mV	Leak potential
V_{thresh}	-30 mV	Threshold potential
τ_{ref}	20 ms	Refractory time
τ_{m}	20 ms	Membrane time constant
Activation function network parameters		
Name	*Value*	*Description*
N_{bias}	8	Number of connected bias neurons
N_{split}	8	No. sources total Poisson input split between
Sampling network parameters		
Name	*Value*	*Description*
N	6	Number of sampling neurons
w_{sample}	15	Sampling weights (ex. or in.)
Mixing network parameters		
Name	*Value*	*Description*
N	16	Number of sampling neurons
w_{sample}	5	Sampling weights (ex. or in.)

Table 6.2: Neuron and network parameters for sampling experiments on the hardware. Values with parentheses are the changed values for rate rather than weight noise modulation experiments.

Acknowledgements

I would like to thank:

Dr. Sebastian Schmitt for supervising me, but more importantly for his help taming certain aspects of the hardware.

Ákos Kungl for letting me constantly interrupt him to ask questions.

Andreas Baumbach for repeatedly explaining sampling to me until it made sense.

Julian Göltz for providing pizza in a time of need.

Dominik Dold for providing a source of stochasticity for office conversation.

And all of the above for their ideas and knowledge upon which I relied so heavily, as well as their time spent proofreading this work.

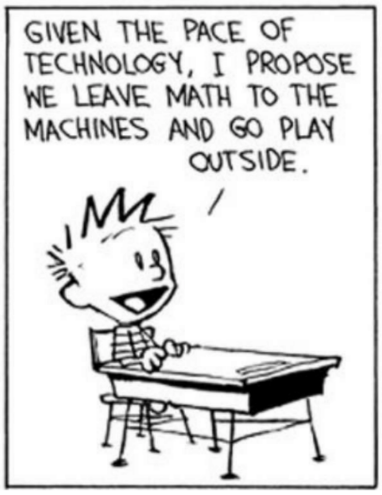

Figure 6.1: A cause to which I hope to have contributed [Wat94].