The Abdus Salam International Centre for Theoretical Physics

IAEA International Atomic Energy Agency

unesco United Nations Educational, Scientific and Cultural Organization

ICTP 1964–2024

Junior Scientists Workshop on Recent Advances in Theoretical Neuroscience | (SMR 3943)

03 Jun 2024 - 07 Jun 2024
ICTP, Trieste, Italy

T01 - CATONI Josefina

Uncertainty representations in variational inference models of low-level visual perception

T02 - CHADWICK Angus Mathew

Rotational Dynamics Enables Noise Robust Working Memory

T03 - ECKMANN Samuel

Structured inhibition-stabilized supralinear networks

T04 - GOEDEKE Sven Ole

Dynamics of drifting cell assemblies

T05 - GRANT Erin Marie

Use case determines the validity of neural systems comparisons

T06 - GURNANI Harsha

Feedback controllability constrains learning timescales during motor adaptation

T07 - HELSON Pascal Max Baptiste

Mean Field Analysis of a Stochastic STDP model

T08 - MARIN VARGAS Alessandro

Modeling the sensorimotor system with task-driven modeling

T09 - NEJATBAKHSHESFAHANI Mohammadamin

Estimating Noise Correlations Across Continuous Conditions With Wishart Processes

T10 - PROCA Alexandra Maria

How context representations emerge during training: a linear network perspective

T11 - SCHÖNSBERG Francesca

A unifying neural network model shows perceptual biases emergence from Hebbian plasticity

T12 - SHAO Yuxiu

Identifying the impact of local connectivity features on network dynamics

T01

Uncertainty representations in variational inference models of low-level visual perception

Josefina Catoni[1], **Enzo Ferrante**[1], **Diego H. Milone**[1] **and Rodrigo Echeveste**[1]

[1]*sinc(i), CONICET-Universidad Nacional del Litoral, Santa Fe, Argentina*

Bayes rule provides an optimal way to perform inference in probabilistic scenarios, and it is hence a natural tool to understand perception in the context of uncertainty. Indeed, increasing evidence indicates the brain is able to represent and operate with probability distributions to (approximately) perform probabilistic inference in several scenarios [1]. A popular choice to approximate the process is variational inference. Variational Autoencoders (VAEs) [2] are a useful tool to learn internal probabilistic representations in an unsupervised fashion. This procedure can be useful when modeling an inference process where the generative model is unknown, since in VAEs the encoder and the decoder are simultaneously learned from the data. This architecture provides a means to model inference in the cortex by learning from the statistics of stimuli. Indeed, previous work has shown that classical receptive fields emerge when training sparse VAEs [3].

Here we studied the properties of the posterior distributions of those VAEs, finding a counterintuitive behavior. While the signal mean and signal variance in the latent representations increase with the contrast of the images, as expected since the images and orientations present in them become more and more distinguishable (cf Fig. 1**a** and) first column), the reported uncertainty (noise variance) grows. This is counterintuitive as the uncertainty would be expected to decrease as contrast increases, with a blank zero contrast image being maximally uninformative. Taking inspiration from the Gaussian Scale Mixture (GSM) model [4], we incorporate a global multiplicative contrast variable to the generative model of the VAE. The GSM has been shown to capture basic properties of natural image statistics, and has been used as a model of cortical visual processing [5]. We call this model explaining-away VAE (EA-VAE) alluding to the explaining-away phenomenon observed in the GSM. Our model fixes the aforementioned problems showing decreasing uncertainty with contrast. Importantly, posteriors converge the prior for zero contrast, which in turn matches the average posterior (Fig. 1**b**).

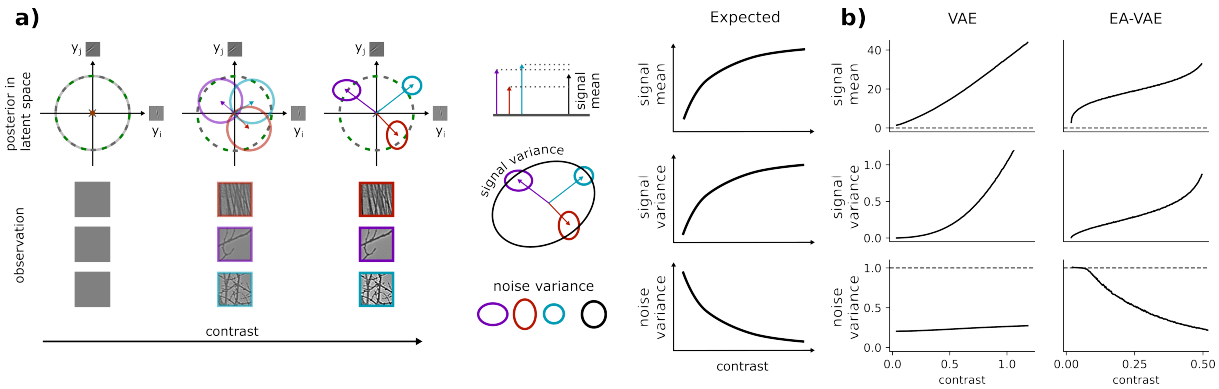


Figure 1: **a)** Sketch of the expected behavior of an inference model for natural images. **b)** Comparison of the behavior of inferred posteriors in a VAE with the inferred by our EA-VAE.

[1] A. Pouget et al., Nat. Neurosci. **16** (9), 1170-1178 (2013).
[2] D.P. Kingma, M. Welling, arXiv:1312.6114 (2022).
[3] F. Csikor et al., arXiv:2206.00436 (2022).
[4] M.J. Wainwright, E. Simoncelli, Adv. Neural Inf. Process. Syst. **12** (1999).
[5] R. Echeveste et al., Nat. Neurosci. **23** (9), 1138-1149 (2020).

T02

Rotational dynamics enables noise robust working memory

Laura Ritter[1] **and Angus Chadwick**[1]

[1] *University of Edinburgh*

Working memory is fundamental to higher-order cognitive function, yet the circuit mechanisms through which memoranda are maintained in neural activity after removal of sensory input remain subject to vigorous debate [1,2]. Prominent theories propose that stimuli are encoded in either stable and persistent activity patterns configured through recurrent attractor dynamics or dynamic and time-varying patterns of population activity brought about through non-normal or feedforward network architectures [2,3]. However, the optimal dynamics for working memory, particularly when faced with ongoing neuronal noise, has not been resolved.

Here, we address this question within the analytically tractable setting of linear recurrent neural networks. First, we develop a novel method to optimise continuous-time linear RNNs driven by Gaussian noise to solve working memory tasks. Our method employs exact analytical expressions to perform gradient descent with respect to the average loss over infinitely many trials, without requiring forward-simulation or backpropagation-through-time. Application of this optimisation method yields a novel and previously overlooked mechanism for working memory maintainence combining both non-normal and rotational dynamics. To test whether these dynamics are a consequence of our optimisation method, we took two approaches: first we confirmed analytically that these non-normal rotational dynamics substantially outperform both perstistent/attactor and non-normal/sequential/feedforward mechanisms; second, we derived analytical expressions for the updates generated by backpropagation-through-time (again over an infinitely large batch size), which upon implementation produced near-identical learning dynamics to those produced by our method.

We next asked whether the non-normal rotational dynamics we identified capture experimentally observed features of neural population activity during working memory tasks. Indeed, we found that the optimised networks replicated several core features of neural population activity in prefrontal cortex, including "dynamic coding" (as quantified by both cross-temporal decoding analysis and switching of single-neuron neuronal selectivity over the delay period) despite stable representational geometry [4].

Taken together, our findings suggest that memoranda are stored and maintained during working memory using combination of non-normal and rotational dynamics, which support a stable and optimally noise-robust representation of working memory contents within a time-varying and dynamic population code.

[1] J Zylberberg, BW Strowbridge, Annu. Rev. Neurosci 40, 603-627 (2017).
[2] M Lundqvist, P Herman, EK Miller, J. Neurosci 38, 7013-7019 (2018).
[3] M Goldman, Neuron 61(4): 621-634 (2009).
[4] E Spaak, K Watanabe, S Funahashi, MG Stokes, J. Neurosci, 37 (27) 6503-6516 (2017).

T03

Structured inhibition-stabilized supralinear networks

Samuel Eckmann[1], **Yashar Ahmadian**[1], **and Máté Lengyel**[1,2]

[1] *Computational and Biological Learning Lab, University of Cambridge, UK*
[2] *Center for Cognitive Computation, Department of Cognitive Science, Central European University, Hungary*

Recurrent network models of excitatory (E) and inhibitory (I) neurons with supralinear activation functions have successfully explained several cortical phenomena [1, 2]. However, the scope of these networks remained limited as their connectivity needed to be random, designed by hand, or fitted by complex machine learning algorithms to yield stable activity and computations [3, 4, 5, 6, 7]. Here we present a general method to efficiently construct stable recurrent E-I networks, without hand-tuning or numerical optimization. We employ a local, synapse-type-specific competitive Hebbian learning rule at all recurrent synapses, while total synaptic weights of each synapse type (EE, EI, IE, II) are constrained to remain constant, reflecting synaptic competition for a limited amount of synaptic resources [8]. When the network's activity is dominated by feedforward inputs, we can solve for the steady-state weight matrix analytically. The matrix reflects the covariance structure of the stimulus statistics, with synaptic weight norms as free parameters that define the general computational properties of the network [9].

We demonstrate our approach by constructing image-computable cortical network models of >7,500 neurons and >55 million recurrent synapses, encoding the covariance structure of natural image datasets resembling the visual field of mice [10]. We chose synaptic weight norms that result in a cortex-like computational regime [2] and predict fine-grained differences in response normalization and center-surround suppression in neurons encoding the upper or lower region of the mouse visual field. In the hippocampus, theoretical considerations require a separation between phases of memory encoding and recall within the same neural circuit [11]. However, the mechanisms supporting such a separation remain unknown. We find that inhibition-stabilized supralinear networks [1, 2] provide a robust mechanism for this separation, whereby the strength of external input controls whether their dynamics are dominated by feedforward or recurrent connections. In addition, we show that continuous memories can be stored in the network by synapse-type-specific competitive Hebbian learning at both excitatory and inhibitory synapses [12]. As a result, for weak input, the cued memory is recalled, and neurons are strongly stabilized by inhibition. For strong input, the external cue is encoded, while inhibition stabilization is paradoxically weaker.

In summary, we present a general framework for the construction of E-I networks that meet key biological constraints. We predict fine-grained cortical computations depending on the sensory input statistics and reveal a novel mechanism for alternating between memory storage and recall within the same hippocampal circuit.

[1] Y. Ahmadian et al., *Neural computation* (2013).
[2] D. B. Rubin et al., *Neuron* (2015).
[3] A. J. Keller et al., *Neuron* (2020).
[4] R. Echeveste et al., *Nature neuroscience* (2020).
[5] D. P. Mossing et al., *bioRxiv* (2021).
[6] Y. Ahmadian et al., *Neuron* (2021).
[7] W. W. M. Soo et al., *bioRxiv* (2022).
[8] S. Eckmann et al., *bioRxiv* (2022).
[9] N. Kraynyukova et al., *PNAS* (2018).
[10] S. Eckmann et al., *Cosyne* (2023).
[11] M. E. Hasselmo et al., *Neural computation* (2002).
[12] S. Eckmann et al., *Cosyne* (2024).

T04

Dynamics of drifting cell assemblies

Sven Goedeke[1,2], **Yaroslav Felipe Kalle Kossio**[2], **Christian Klos**[2], **and Raoul-Martin Memmesheimer**[2]

[1] *(Presenting author underlined) Theoretical Systems Neuroscience, Bernstein Center Freiburg, University of Freiburg, Germany*
[2] *Neural Network Dynamics and Computation, Institute of Genetics, University of Bonn, Germany*

In a standard model, associative memories are represented by assemblies of strongly interconnected neurons. We have proposed a contrasting memory model with complete temporal remodeling of assemblies, based on experimentally observed changes in synapses and neural representations [1]. The assemblies drift freely as noisy autonomous network activity and spontaneous synaptic turnover drive neuron exchange. The gradual exchange allows activity-dependent and homeostatic plasticity to conserve the representational structure and keep inputs, outputs, and assemblies consistent, leading to persistent memory. This explains experimental findings of changing memory representations.

At the level of single neurons, assembly drift is reflected by characteristic dynamics: relatively long periods of stable assembly membership interspersed with fast transitions. How can we mechanistically understand these dynamics? Here we answer this question by proposing simplified, reduced models. We first constructed a random walk model for neuron transitions between assemblies based on the statistics of synaptic weight changes measured in simulations of plastic spiking neural networks exhibiting assembly drift. It shows that neuron transitions between assemblies can be understood as noise-activated switching between metastable states. The random walk's potential landscape and inhomogeneous noise strength induce metastability and thus support assembly maintenance in the presence of ongoing fluctuations. In a second step, we derive an effective random walk model from first principles. In this model, a neuron spikes at a fixed background rate and, depending on the neuron's coupling to the assembly, together with its current or another assembly. The model generates neuron transitions between assemblies as well as potentials and inhomogeneous noise similar to those observed in spiking network simulations. The approach can be applied generally to the dynamics of drifting assemblies, irrespective of the employed neuron and plasticity models.

[1] Y. F. Kalle Kossio, S. Goedeke, C. Klos, R.-M. Memmesheimer, Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation. PNAS **118**, e2023832118 (2021).

T05

Use case determines the validity of neural systems comparisons

Erin Grant[1], **Brian Cheung**[2], **Tomaso Poggio**[2], **Andrew Saxe**[1]

[1] *Gatsby Unit & Sainsbury Wellcome Centre, University College London*
[2] *Center for Brains, Minds and Machines, Massachusetts Institute of Technology*

Deep learning provides new data-driven tools to relate neural activity to perception and cognition, aiding neuroscientists and cognitive scientists in developing theories of neural computation that increasingly resemble biological systems both at the level of behavior [3] and of neural activity. [1] But what in a neural network should correspond to what in a biological system? This question is addressed implicitly in the use of specific comparison measures—such as representational similarity analysis [5] or linear regression fit to neural activity [6]—that relate specific neural or behavioral dimensions via a particular functional form. However, distinct comparison methodologies can give conflicting results in recovering even a known ground-truth model in an idealized setting, [4] leaving open the question of what to conclude from a successful or unsuccessful systems comparison using any given methodology.

Here, we develop a framework to make explicit and quantitative the effect of *both* hypothesis-driven aspects—such as details of the architecture of a neural network—*as well as* methodological choices in a systems comparison setting. We demonstrate via both analytical and simulated learning dynamics of neural networks that, while the role of the comparison methodology is often de-emphasized relative to hypothesis-driven aspects, this choice can greatly impact and even invert the conclusions to be drawn from a comparison between neural systems. In particular, we establish cases in which whether or not two systems are found to be similar depends on variables not often accounted for when comparing neural systems. For example, rich and lazy learning—distinct representational regimes attested via models of biological learning [2]— can be controlled via methodological parameters, suggesting that representational regimes should be treated as hypothesis-relevant variables, on the level of architecture, for example.

In addition, our framework allows us to idealize scientific use cases as parametric interventions; for example, we examine the robustness of conclusions about representational and functional similarity to measurement noise in neural activity and model misspecification. We contend that the right way to judge similarity depends on the purpose or scientific hypothesis under investigation, which could range from identifying single-neuron or circuit-level correspondences to capturing generalizability to new stimulus dimensions.

[1] C Conwell et al. “What can 1.8B regressions tell us about the pressures shaping high-level visual representations...” In: *BioRxiv* (2023).

[2] M Farrell et al. “From lazy to rich to exclusive task representations in neural networks and neural codes”. In: *Curr. Opin. Neurobio.* (2023).

[3] R Geirhos et al. “Partial success in closing the gap between human and machine vision”. In: *Adv. NeurIPS* (2021).

[4] Y Han et al. “System identification of neural systems: If we got it right, would we know?” In: *Proc. ICML*. 2023.

[5] N Kriegeskorte et al. “Representational similarity analysis: Connecting the branches of systems neuro.” In: *Front. Sys. Neuro.* (2008).

[6] M Schrimpf et al. “Brain-Score: Which artificial neural network for object recognition is most brain-like?” In: *BioRxiv* (2018).

T06

Junior Scientists Workshop on Recent Advances in Theoretical Neuroscience:

Feedback Controllability Constrains Learning Timescales During Motor Adaptation

Harsha Gurnani[1,2,3] **, Bing W. Brunton** [1,2,3]

[1]Department of Biology, [2]Computational Neuroscience Center, [3]eScience Institute; University of Washington, Seattle, USA

The ability to produce new neural dynamics is a key feature of motor learning, and likely involves plasticity within distributed circuits; this learning is also relevant in the context of brain-computer interfaces (BCI) that rely on real time decoding of neural activity. Previous work exploring the structure of M1 activity during motor tasks has largely assumed autonomous dynamics (i.e. activity unfolding from initial states dominated by local recurrent interactions), and related work on BCI learning has focused on local mechanisms (such as M1 synaptic plasticity). **However, recent experimental evidence suggests that M1 activity during BCI use is continuously modified by sensory feedback [1] and produces corrections for noise and external perturbations [1,2], suggesting a critical need to model this interaction between feedback and intrinsic M1 dynamics**. In this work, we investigated the role of flexible feedback modulation of cortical dynamics in the context of both BCI task performance and short-term learning. Incorporating sensory feedback inputs to recurrent neural networks (RNNs), we trained them on a 2D centre-out reaching task where the network activity controlled the BCI cursor velocity. We first examined the task-relevant dynamics that emerged over training and showed that many experimentally observed features of M1 activity could be recapitulated in feedback-driven networks. We further adapted existing reverse-engineering methods to closed-loop dynamics and showed how fixed points of the coupled RNN-cursor system underlie the error-correction mechanism. Secondly, we observed a misalignment of the intrinsic manifold, task-dynamics subspace and decoder (output) weights, which we show has implications for the robustness of different decoders against neuronal noise. Next, we suggest that short-term adaptation, including to BCI decoder perturbations as in [3], can be facilitated by plasticity of inputs from upstream controllers such as a remapping of sensory feedback, instead of plasticity of recurrent connections within M1. Crucially, we show that beyond alignment of decoders with the intrinsic manifold, the pre-existing input-driven dynamical structure determines the speed of adaptation to different decoder perturbations. This offers an explanation for the experimentally-observed variability of learning outcomes across different "within-manifold" perturbations, that has been missing from related computational studies. Moreover, learning via input plasticity produced little change to the statistical distribution of neural states, consistent with neural reassociation [4]. Lastly, we show adaptation using a biologically-plausible learning rule that modifies input weights is consistent with experimentally-observed variability of BCI learning outcomes. **By incorporating adaptive controllers upstream of M1, our work highlights the need to model input-dependent latent dynamics, and clarifies how constraints on learning arise from both the statistical characteristics and the underlying dynamical structure of neural activity.**

[1] M. Golub, B.M. Yu, S.M. Chase. eLife 2 4:e10015 (2015).

[2] S.D. Stavisky, J.C. Kao, S.I. Ryu, K.V. Shenoy. Neuron, 95(1), 195–208.e9 (2017).

[3] P.T. Sadtler, K.M. Quick, M. Golub, S.M. Chase, S.I. Ryu, E.C. Tyler-Kabara, B.M. Yu, A.P. Batista. Nature 512, 423–426 (2014).

[4] M.D. Golub, P.T. Sadtler, E.R. Oby, K.M. Quick, S.I. Ryu, E.C. Tyler-Kabara, A.P. Batista, S.M. Chase, B.M. Yu. Nature Neuroscience 21, 607–616 (2018).

Mean Field Analysis of a Stochastic STDP model

T07

Pascal Helson[1], **Etienne Tanré**[2], **and Romain Veltz**[2]

[1] *KTH Royal Institute of Technology, Sweden* [2] *Inria Sophia-Antipolis, France*

Biological neural network models with synaptic plasticity pose challenges for both theoretical and numerical analyses. The numerical bottleneck arises from the N^2 scaling of synapses with an increasing number N of neurons. Additionally, the intricate coupling between neuron and synapse dynamics, along with plasticity-induced heterogeneity, hinders the use of classical tools from theory. One approach to address these challenges is to assume plasticity to be very slow compared to neural activity and leverage slow-fast theory. However, this slowness is not universally applicable in the brain, making it unclear how to analyze such complex systems without relying on the slow-fast assumption. In this work, we conduct a mean-field analysis on a neural network model with plastic interactions, resulting in a significantly reduced model.

We explore Spike-Timing-Dependent Plasticity (STDP) within a probabilistic Wilson-Cowan neural network model featuring binary neural activity. The network is composed of N triplets, each comprises the neuron potential $V_t^{i,N} \in 0, 1$, the time since its last spike $S_t^{i,N} \in \mathbb{R}^+$, and its N incoming synaptic weights $(W_t^{i\leftarrow j,N})1 \leq j \leq N \in \mathbb{Z}$. Neurons revert to their resting potential (from 1 to 0) via Poisson processes, spiking (from 0 to 1) based on synaptic currents $I_t^{i,N} = \frac{1}{N}\sum_j W_t^{i\leftarrow j,N} V_t^{j,N}$. Upon a spike, outgoing $(W_t^{j\leftarrow i,N})1 \leq j \leq N$ or incoming weights $(W_t^{i\leftarrow j,N})_{1\leq j\leq N}$ update following a probabilistic STDP rule, where smaller $S_t^{j_0,N}$ makes $W_t^{i\leftarrow j_0,N}$ more likely to potentiate $(+1)$ and $W_t^{j_0\leftarrow i,N}$ to depress (-1).

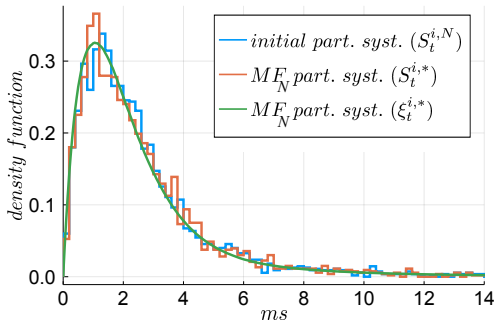
The N^2 weights make the definition of a *typical* neuron highly non trivial. A possible choice is to use *new* variables that ease performing a mean field approximation which are the **empirical distributions**, $\xi_t^{i,N}$, of the state of the pre-synaptic neurons (neuron state, time since last spike and outcoming synaptic weight to neuron i),

$$\xi_t^{i,N} = \frac{1}{N}\sum_j \delta_{(V_t^{j,N}, S_t^{j,N}, W_t^{i\leftarrow j,N})}.$$

Considering this new system $X_t^{i,N} = (V_t^{i,N}, S_t^{i,N}, \xi_t^{i,N})$, we derive a closed system of equations. We then conjecture that the dynamics of any limit point as N tends to infinity, $(V_t^*, S_t^*, \xi_t^*)_{t\geq 0}$, is solution to a McKean-Vlasov SDE on the space $\{0, 1\} \times \mathbb{R}^+ \times \mathcal{P}(E_m)$ where $\mathcal{P}(E_m)$ is the space of probability measures on $\{0, 1\} \times \mathbb{R}^+ \times \mathbb{Z}$. We illustrated this limit dynamics with simulations by comparing the finite size neural network to the mean field limit system. There is a good match between the two as shown in Fig. 1.

This analysis marks the first exploration of exact mean-field dynamics in a network of interacting neurons with plasticity. We hope this work can help deriving mean field limits of other models with particle in interaction. It also opens the door to new mathematical questions such as establishing the uniqueness of the solution to the limit system and confirming the convergence to a deterministic limit measure. Moreover, studying the limit system would give insight in the initial model. In particular, this study opens new avenues for preventing weight divergence without relying on soft or hard bounds. Importantly, our approach significantly reduces simulation costs, crucial for models involving synaptic plasticity. A timeline example is the consequence of deep brain stimulation (DBS) on the weights linking the neurons stimulated, especially in an adaptive setting. DBS is used to alleviate symptoms in many brain diseases like depression, Parkinson's disease and epilepsy.

(a) Time since last spike distribution (V=0)



(b) Synaptic current distribution

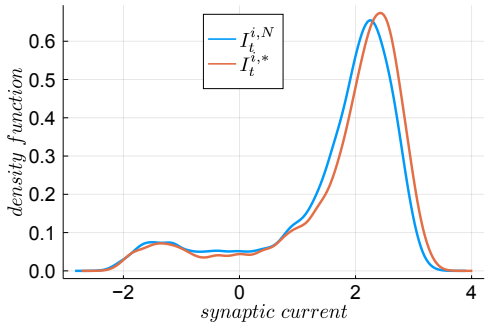


Fig. 1 : Comparing two properties of the limit system in 1a and 1b for $N = 5000$ *after* $500ms$.

T08

Modeling the sensorimotor system with task-driven modeling

Alessandro Marin Vargas, Alberto S. Chiappa, Adriana P. Rotondo, Alexander Mathis

Brain Mind and NeuroX Institute, School of Life Sciences, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

Complex behavioral tasks like grasping require precise and accurate motor control. This demands an intricate interplay between proprioceptive processing and motor command generation within the sensorimotor system. While task-driven modeling has demonstrated success in understanding visual processing, its application to proprioception remains underexplored. Here, we employ a task-driven modeling approach to investigate the neural code of proprioceptive neurons in cuneate nucleus (CN) and somatosensory cortex area 2 (S1). We simulated muscle spindle signals through musculoskeletal modeling and generated a large-scale movement repertoire to train neural networks based on 16 hypotheses, each representing different computational goals. We found that the emerging, task-optimized internal representations generalize from synthetic data to predict neural dynamics in CN and S1 of primates. Computational tasks that aim to predict the limb position and velocity were the best to predict the neural activity in both areas. Since task-optimization develops representations that better predict neural activity during active than passive movements, we postulate that neural activity in CN and S1 is top-down modulated during goal-directed movements (Marin Vargas*, Bisi* et al. Cell 2024). Building upon this framework, we extend our investigation to incorporate deep reinforcement learning, aiming to capture sensorimotor representations. By training artificial agents to imitate natural grasp movements using imitation learning, we develop stimulus-computable models that effectively capture sensorimotor representation by predicting the neural activity of primates during grasping movements. We show that these models outperform classic encoding models. These results suggest that deep reinforcement learning has the potential to bridge the gap between artificial and biological sensorimotor systems, providing valuable insights into the mechanisms underlying sensorimotor integration and control.

T09

Estimating Noise Correlations Across Continuous Conditions With Wishart Processes

Amin Nejatbakhsh[1], **Isabel Garon**[1], **and Alex H Williams**[1]

[1] *Center for Computational Neuroscience, Flatiron Institute, New York, NY*

The signaling capacity of a neural population depends on the scale and orientation of its noise covariance across trials [1] (Fig. 1A). Estimating this covariance is challenging and is thought to require a large number of stereotyped trials of a repeated action or stimulus presentation. New approaches are therefore needed to interrogate the structure of neural noise across rich, naturalistic behaviors and sensory experiences, with few trials per condition. Here, we exploit the fact that conditions are smoothly parameterized in many experiments and leverage Wishart process models to pool statistical power from trials in neighboring conditions [2]. The core insight we exploit is that similar experimental conditions—e.g. cued arm reaches to similar locations—ought to exhibit similar noise statistics (Fig. 1B). We demonstrate favorable performance on experimental data from the monkey motor cortex relative to standard covariance estimators (Fig. 1E-H). Moreover, the Wishart process produces smooth estimates of covariance as a function of stimulus parameters, enabling estimates of noise correlations in unseen conditions (Fig. 1I,J) as well as continuous estimates of Fisher information—a commonly used measure of signal fidelity. Together, our results suggest that Wishart processes are broadly applicable tools for quantification and uncertainty estimation of noise correlations in trial-limited regimes, paving the way toward understanding the role of noise in complex neural computations and behavior.

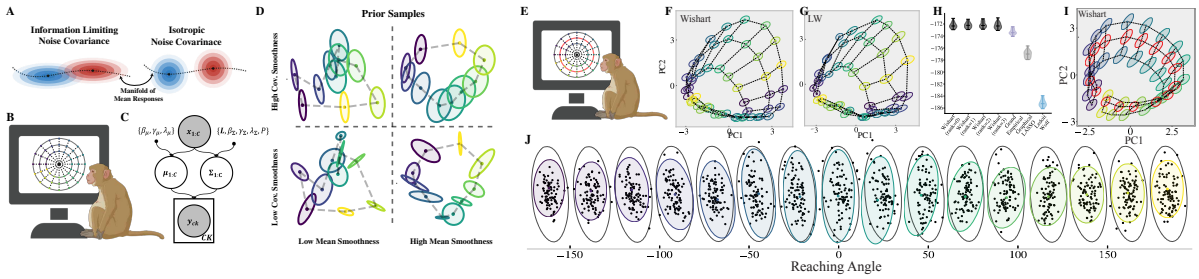


Figure 1: (A) Illustration of information limiting noise correlations. (B) Experimental dataset with smoothly parametrized conditions (see [1]). A nonhuman primate makes point-to-point reaches to radial targets. The parameterized condition space is shown on the cartoon screen. (C) Graphical model of the Wishart model with Normal observations. (D) Samples from Gaussian and Wishart process prior distributions. Dots and ellipses represent the mean and covariance of neural responses. Colors appear in the increasing order of condition value (e.g. angle) and dashed lines connect neighboring conditions. Increasing the mean kernel parameter (horizontal axis) encourages smoothness in the means while increasing the covariance kernel parameter (vertical axis) encourages the ellipses to change smoothly. (E) Task schematic and set of experimental conditions. (F,G) Wishart (*left*) and Ledoit-Wolf (*right*) estimates of mean and covariance, projected onto the top-2 PCs. (H) Log-likelihood distributions of held-out trials. (I) The middle ring (red ring in G) of targets was held out in training and the means and covariances were interpolated (red ellipses) using the Wishart model. (J) Covariance ellipses, grand-empirical estimates (dashed lines), and samples (black dots) are visualized in the top-2 PC subspace.

1. Moreno-Bote, R. *et al.* en. *Nat. Neuro.* (2014).
2. Wilson, A. G. *et al. UAI* (2011).

T10

How context representations emerge during training: a linear network perspective

Alexandra M. Proca[1,*], **Kai Sandbrink**[2,*], **Jan P. Bauer**[3,*], **and Ali Hummos**[4]

[1]*Imperial College London*
[2]*University of Oxford*
[3]*The Hebrew University of Jerusalem*
[4]*Massachusetts Institute of Technology*
[*]Equal contribution

Contextual responding is a common feature in neuroscience experiments where animals identify two contexts in the task and switch between them flexibly, often requiring only a few error trials to infer a context switch. Context representations are then thought to gate the neural population dynamics to perform the correct computation in each context. However, how these cognitive abstractions of context emerge in neural systems is unknown. Here, we consider the emergence of context in a linear network with gating variables, both analytically and in simulations. During learning we update both the network weights and the gating variables along the objective function gradient. Simulations show that the weight matrices specialize to solve the computation required in each task, while the gating variables represent the active context. Transitioning between contexts initially relies on updating weights and later on updating gating variables only, leading to adaptable behavior and minimizing interference and forgetting. This separation between task computations and representation of task context emerges from an interplay of scales in the respective parts of the network. Gating variables encoding of context is incentivized by the dynamics once the weight matrices learn the task computation accurately. In addition, analytical expressions of the gating variable dynamics and its interactions with weights dynamics show that gating not only influences behavioral output, but also gates the effective learning rate of the weights responsible for other behaviors. Overall, our work studies linear networks to propose a mechanism for how abstract cognitive representations of context emerge, identifying the pertinent components for behavioral flexibility and protecting knowledge.

T11 **A unifying neural network model shows perceptual biases emergence from Hebbian plasticity**

Francesca Schönsberg

Perceiving the magnitude of a stimulus is a complex brain function that arises from the interplay of working memory and experience. This interplay results in two well-known perceptual biases in memory tasks: 1. In a series of vibrational stimuli of varying strength, both humans and rodents tend to overestimate the strength of a stimulus after a series of weak stimuli (and vice versa) due to the repulsive bias of representations. 2. The contraction bias instead shifts the representation of a stimulus held in working memory towards the average of stimuli observed in the past. While a series of experiments have yielded a detailed phenomenological description of both biases, the neural mechanisms underlying these biases remain poorly understood.

In the presentation I will report a recent study in which we show that the representations learnt by recurrent neural networks with ongoing Hebbian plasticity quantitatively reproduce (i) the contractive bias we find in experiments with human subjects, and (ii) the repulsive effect found in rodents by Hachen et al. (Nat. Comm. 2021). In our model, a fully-connected network of rate-based units is driven by external inputs modeled after the experimental protocol, while its connectivity is continuously evolving due to Hebbian plasticity. We do not use gradient descent nor do we fine-tune the model to different experimental paradigms. We finally design a new behavioural paradigm where contraction and repulsive bias interact and find again that the model predicts salient features of the performance of our human participants.

Our results show that a single recurrent neural network with ongoing Hebbian plasticity reproduces two perceptive biases observed across three experimental paradigms. The striking match between experimental data and theoretical predictions supports the hypothesis that perceptual biases arise from simple Hebbian plasticity within a unique recurrent subregion of the brain, e.g. vM1 in rats, which acts as a plastic platform that filters perception based on context.

T12

The impact of local connectivity features on network dynamics

Yuxiu Shao[1,2], **David Dahmen**[3], **Stefano Recanatesi**[4,5], **Eric Shea-Brown**[6] **and Srdjan Ostojic**[2]

[1] *Beijing Normal University, Beijing, China;* [2] *École Normale Supérieure, Paris, France;* [3] *Research Centre Jülich, Jülich, Germany;* [4] *Technion, Haifa, Israel;* [5] *Allen Institute for Neural Dynamics, Seattle, USA;* [6] *University of Washington, Seattle, USA.*

Understanding how connectivity structure shapes network dynamics is paramount in the field of neuroscience. Theoretical investigations of multi-population neuronal networks often consider statistically homogeneous populations and incorporate either only the population-averaged mean or i.i.d. fluctuations in synaptic couplings. A newly released synaptic physiology dataset highlighted the strong presence of motifs – specific connectivity patterns between pairs and triplets of neurons–beyond the scope of mean connectivity[1]. However, it is a priori not clear which of the experimentally identified connectivity motifs exert a strong influence on neural dynamics. While most previous works focused on reciprocal motifs, here we show that another feature of connectivity, chain motifs, has a much stronger impact on the dynamics of neural activity.

We compared the effects of chain and reciprocal motifs within two-population excitatory-inhibitory networks using an analytical framework that approximates the connectivity in terms of low-rank structures that incorporate motifs. We mathematically derived the dominant eigenvalues and exploited matrix perturbation theory to determine the statistics of corresponding eigenvectors. We then used these results to perform a low-rank approximation[2] that predicts the effects of connectivity motifs on linear network dynamics.

Our results show that chain motifs have a much stronger impact on dominant eigenmodes than reciprocal motifs[3, 4]. Moreover, an overrepresentation of chain motifs induces an additional eigenmode with an eigenvalue of sign opposite to the dominant one, thus modifying the network's effective rank (Fig 1a). This additional eigenmode substantially influences network dynamics, offering a new perspective on how local EI motifs shape the network's excitability (Fig 1b1 and 2). Our exploration of the physiological connectivity dataset for the first time revealed the significant impact of EI chain motifs on altering the network's effective rank, permitting the discovery of richer dynamics associated with these specific connectivity motifs.

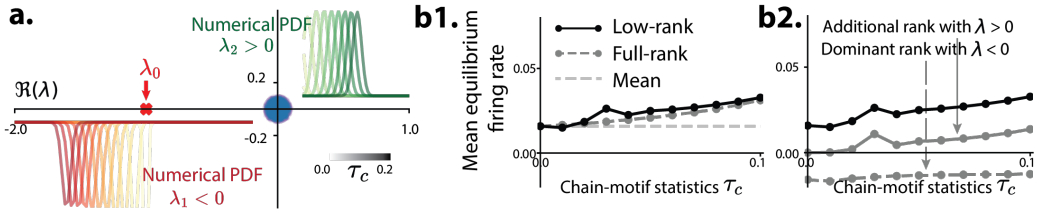


Figure 1: (a) illustrates the effects of chain motifs on eigenvalues, notably the emergence of an additional outlier λ_2. (b1 and 2) The low-rank approximation model predicts the influence of the chain motifs on network excitability.

[1] Luke Campagnola et al., Science **375**, eabj5861 (2022).
[2] Francesca Mastrogiuseppe, Srdjan Ostojic, Neuron **99**, no. 3 (2018).
[3] David Dahmen, Stefano Recanatesi, Gabriel K. Ocker, Xiaoxuan Jia, Moritz Helias, and Eric Shea-Brown. Biorxiv (2020).
[4] Yuxiu Shao, Srdjan Ostojic, PLOS Computational Biology **19**, no. 1 (2023).