A unified theory of feature learning in RNNs and DNNs

Jan Bauer[1] · Kirsten Fischer[2] · Moritz Helias[2] · Agostina Palmigiano[1]

[1] Gatsby Computational Neuroscience Unit, UCL, London, UK [2] IAS-6, Forschungszentrum Jülich, Germany {jan.bauer, a.palmigiano}@ucl.ac.uk

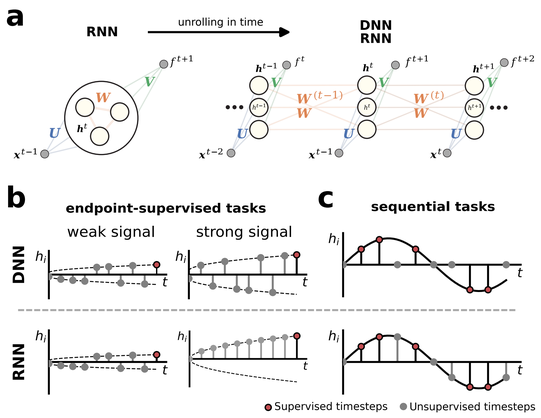
TL;DR: We derive one mean-field theory for trained RNNs and DNNs. Weight sharing, their only difference, becomes a **masking of temporal correlations**: it triggers a **phase transition** to temporally coherent representations and acts as an **inductive bias** that makes RNNs sample-efficient sequence learners.

1 · RNNs are DNNs with shared weights

Unrolled in time, an RNN is a deep network whose layers are timesteps, except that all layers share the same weights $\boldsymbol{W}$. Are the DNN's extra parameters an advantage, or is weight sharing a useful inductive bias?

$$\boldsymbol{h}^t = \boldsymbol{W}^{(t-1)}\phi(\boldsymbol{h}^{t-1}) + \boldsymbol{U}\boldsymbol{x}^{t-1}, \qquad f^{t+1} = \boldsymbol{V}\phi(\boldsymbol{h}^t),$$

$$\boldsymbol{W}^{(t)} = \begin{cases} \boldsymbol{W}, & \textbf{RNN (shared)} \\ \boldsymbol{W}^{(t)}, & \textbf{DNN (untied)} \end{cases}$$



a) Unrolling-in-time. **b)** Endpoint supervision: only the RNN turns coherent in time. **c)** Sequential tasks: weight sharing interpolates unsupervised timesteps.

2 · Training = Bayesian inference

Noisy gradient descent (SGLD: gradients + weight decay $\mathrm{K}/G_{\boldsymbol{\theta}}$ + isotropic noise $\sqrt{2\mathrm{K}}$) converges to a stationary posterior over weights,

$$\boldsymbol{\theta} \sim P(\Theta|y,\boldsymbol{x}) \propto \underbrace{e^{-P|\mathcal{T}|\mathcal{L}(\Theta;y,\boldsymbol{x})/\mathrm{K}}}_{\text{likelihood (loss)}} \underbrace{\mathcal{N}(\Theta|0,G_{\boldsymbol{\theta}})}_{\text{prior (weight decay)}}.$$

Requiring $\mathcal{O}(1)$ kernels as $N \to \infty$ recovers μP feature-learning scaling, $(G_{\boldsymbol{U}}, G_{\boldsymbol{W}}, G_{\boldsymbol{V}}) = (^{u}/_{D}, ^{w}/_{N}, ^{v}/_{N^2})$.

3 · One kernel theory, two architectures

Weights enter the forward pass linearly ⇒ integrate them out exactly; activations follow a posterior governed by the kernel $\boldsymbol{\Phi}^{tt'}_{pp'} = \frac{1}{N}\sum_i \phi(h^t_{i,p})\phi(h^{t'}_{i,p'}) \asymp \langle\phi(h^t_p)\phi(h^{t'}_{p'})\rangle$:

$$P(h|y,\boldsymbol{x}) \propto \exp\Big\{-\tfrac{1}{2}\mathrm{tr}\left[\mathbb{Y}_{\mathcal{T}}\big(v\boldsymbol{\Phi}^{-}_{\mathcal{T}}+\kappa\big)^{-1}\right] - \tfrac{1}{2}h^{\intercal}\big(w[\![\boldsymbol{\Phi}^{-}]\!]+u\mathbb{X}^{-}\big)^{-1}h + \phi(h)^{\intercal}\tilde{\boldsymbol{\Phi}}\,\phi(h)\Big\}$$

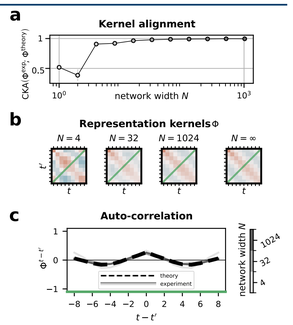
Architecture enters **only** through a masking of temporal correlations,

$$[\![\boldsymbol{\Phi}^{-}]\!] = \begin{cases} \boldsymbol{\Phi}^{-}, & \textbf{RNN} \\ \mathrm{diag}(\boldsymbol{\Phi}^{-}), & \textbf{DNN} \end{cases}$$

shared weights are perfectly correlated in time ⇒ coherence can build up; independent weights force it to zero.

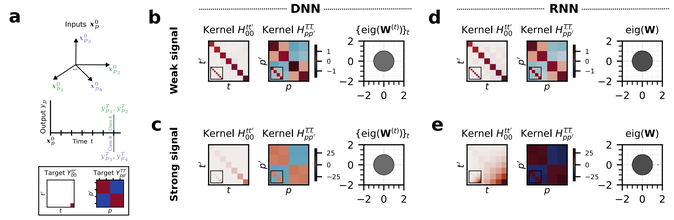
4 · The theory describes trained networks

Kernels from SGLD-trained RNNs converge to the theory's prediction as the width N grows (CKA → 1); temporal coherence and autocorrelation are captured already at finite N.



5 · Endpoint supervision: a phase transition to temporal coherence

Input at $t = 0$, labels only at the last timestep (the classic DNN setting). **Both** architectures learn the task: $\mathbb{H}^{T\text{-}T\text{-}} \simeq \mathbb{Y}$, giving similar predictors f.

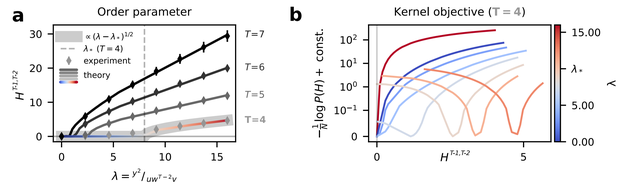


Weak signal (top): RNN and DNN representations coincide. Strong signal (bottom): only the RNN develops temporal coherence, with an outlier eigenvalue in $\boldsymbol{W}$.

Beyond a critical learning signal $\lambda = \|y\|^2/vw^{T-2}u$, the ratio of label strength to initialization variance, only the RNN crosses a **second-order phase transition** (critical exponent $\frac{1}{2}$, analytic at $T = 4$) into a temporally coherent phase, reminiscent of a BBP transition. For linear activation, the kernel posterior is explicit:

$$-\ln P(\mathbb{H}|y,\boldsymbol{x})/N = \underbrace{\tfrac{1}{2}\mathrm{tr}\left[\mathbb{Y}_{\mathcal{T}}\big(v\mathbb{H}^{-}_{\mathcal{T}}+\kappa\big)^{-1}\right]}_{\text{label energy}} + \underbrace{\tfrac{1}{2}\mathrm{tr}\left[\mathbb{H}\big(w[\![\mathbb{H}^{-}]\!]+u\mathbb{X}^{-}\big)^{-1}\right] - \tfrac{1}{2}\ln\frac{|\mathbb{H}|}{|w[\![\mathbb{H}^{-}]\!]+u\mathbb{X}^{-}|}}_{D_{\mathrm{KL}}\text{ to the random-weight forward pass (entropy)}}$$

Aligning $\mathbb{H}^{T\text{-}T\text{-}}$ to the labels lowers the energy and costs entropy; the RNN can pay through off-diagonals in $[\![\mathbb{H}^{-}]\!]$, the DNN cannot.

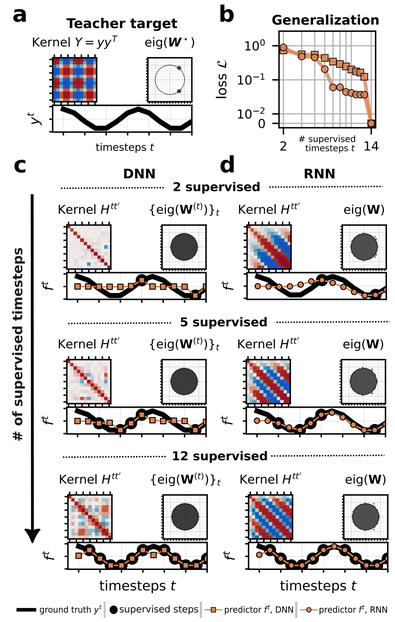


Kernel off-diagonal $\mathbb{H}^{T-1,T-2}$ as order parameter: theory (lines) vs. weight-space SGLD (diamonds); the kernel landscape $-\ln P(\mathbb{H}|y,\boldsymbol{x})$ develops a coherent minimum.

Funding: Gatsby Charitable Foundation (GAT3850), Simons Foundation (1156607), DFG (368482240/GRK2416, SPP 2205 (533396241), HE 9032/4-1), Helmholtz ACA (SO-092); compute on JURECA, FZ Jülich (JINB33).

6 · Sequential tasks: weight sharing is an inductive bias

Sequence regression $y^t = \boldsymbol{V}^{\star}(\boldsymbol{W}^{\star})^t\boldsymbol{U}^{\star}\boldsymbol{x}^0$ (teacher rotation), supervising only a subset of timesteps:



RNN: kernel matches $\mathbb{Y}$ from few supervised steps; $\boldsymbol{W}$ recovers the teacher eigenvalues. **DNN**: posterior stays diagonal ⇒ falls back to the prior mean $f^t = 0$ outside supervision, despite larger expressivity.

Mechanism (perturbation in $\mathbb{Y}$): kernel inverses act as *propagators* that pass label messages across unsupervised timesteps,

$$\mathbb{\Delta}^{tt'}_{2,\mathrm{RNN}} \propto \frac{1}{\mathbb{h}^{t'}_0}\Big(\sum_{t''}\mathbb{Y}^{+tt''}\frac{1}{\mathbb{h}^{t''}_0}\mathbb{Y}^{+t''t'}\Big)\frac{1}{\mathbb{h}^{t'}_0}$$

interpolating along paths $t_3 \leftarrow t_2 \leftarrow t_1$. The DNN's masking cuts these paths ⇒ learning needs more samples.

7 · Takeaways

- One Bayesian kernel theory covers trained RNNs **and** DNNs; the architecture enters as a masking of temporal correlations.
- Strong learning signals drive RNNs, and only RNNs, through a phase transition to temporally coherent representations.
- In sequential tasks, weight sharing induces **task-model alignment**: the learned kernel matches the task structure $\mathbb{Y}$, giving better generalization from fewer labels. **Structure → function**.

Paper & code