Exact Learning Dynamics of Bottlenecked and Wide Deep Linear Networks

Architectural diversity characterises both biological and artificial learning systems. In neuroscience, distinct structural organisations support diverse functions, while in machine learning, variations in depth, width, and connectivity shape learning dynamics and performance. Extending solvable models to reflect such complexity is key to advancing both theory and applications. Building on the matrix Riccati framework from Fukumizu [1998], Braun et al. [2022b], Dominé et al. [2024], we derive exact dynamics for wide and bottlenecked linear networks, expanding the class of solvable architectures. Despite their simplicity, deep linear networks reveal rich fixed-point structures and nonlinear learning trajectories in parameter and function space, paralleling behaviours of nonlinear systems [Baldi and Hornik, 1989, Fukumizu, 1998, Saxe et al., 2014].

Consider a supervised learning task with a set of P training pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{P}$. We learn this task with a two-layer linear network model that produces the output prediction $\hat{\mathbf{y}}_n = \mathbf{W}_2\mathbf{W}_1\mathbf{x}_n$, with weight matrices $\mathbf{W}_1 \in \mathbb{R}^{N_h \times N_i}$ and $\mathbf{W}_2 \in \mathbb{R}^{N_o \times N_h}$, where N_h is the number of hidden units. The network's weights are optimised using full batch gradient descent with learning rate η on the mean squared error loss $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2}\left\langle ||\hat{\mathbf{y}} - \mathbf{y}||^2 \right\rangle$, where $\langle\cdot\rangle$ denotes the average over the dataset.

We employ an approach first introduced in the foundational work of Fukumizu [1998] and extended in recent work by Braun et al. [2022b], Dominé et al. [2024], which, instead of studying the parameters directly, considers the dynamics of a matrix of important statistics. In particular, defining $\mathbf{Q} = \begin{bmatrix}\mathbf{W}_1 & \mathbf{W}_2^T\end{bmatrix}^T \in \mathbb{R}^{(N_i+N_o)\times N_h}$, we consider the $(N_i + N_o) \times (N_i + N_o)$ matrix $\mathbf{Q}\mathbf{Q}^T(t) = \begin{bmatrix}\mathbf{W}_1^T\mathbf{W}_1(t) & \mathbf{W}_1^T\mathbf{W}_2^T(t) \\ \mathbf{W}_2\mathbf{W}_1(t) & \mathbf{W}_2\mathbf{W}_2^T(t)\end{bmatrix}$, which is divided into four quadrants with interpretable meanings, and where $t \in \mathbb{R}$ represents training time. The approach monitors several key statistics collected in the matrix. The off-diagonal blocks contain the network function $\hat{\mathbf{Y}}(t) = \mathbf{W}_2\mathbf{W}_1(t)\mathbf{X}$. The on-diagonal blocks capture the correlation structure of the weight matrices, allowing for the calculation of the temporal evolution of the network's internal representations. This includes the representational similarity matrices (RSM) of the neural representations within the hidden layer $\text{RSM}_I = \mathbf{X}^T\mathbf{W}_1^T\mathbf{W}_1(t)\mathbf{X}$, $\text{RSM}_O = \mathbf{Y}^T(\mathbf{W}_2\mathbf{W}_2^T(t))^+\mathbf{Y}$, where $+$ denotes the pseudoinverse; and the network's finite-width NTK [Jacot et al., 2018, Lee et al., 2019, Arora et al., 2019] $\text{NTK} = \mathbf{I}_{N_o} \otimes \mathbf{X}^T\mathbf{W}_1^T\mathbf{W}_1(t)\mathbf{X} + \mathbf{W}_2\mathbf{W}_2^T(t) \otimes \mathbf{X}^T\mathbf{X}$, where $\mathbf{I}_{N_o}$ is the $N_o \times N_o$ identity matrix and $\otimes$ is the Kronecker product. Hence, the dynamics of $\mathbf{Q}\mathbf{Q}^T$ describe the important aspects of network behaviour.

We derive an exact solution for $\mathbf{Q}\mathbf{Q}^T$ under the following assumptions: **A1** (*Whitened input*). The input data is whitened, i.e., $\tilde{\mathbf{\Sigma}}^{xx} = \mathbf{I}$. **A2** ($\lambda$-*Balanced*). The network's weight matrices are λ-balanced at the beginning of training, i.e., $\mathbf{W}_2^T\mathbf{W}_2(0) - \mathbf{W}_1\mathbf{W}_1(0)^T = \lambda\mathbf{I}_\mathbf{h}$. If this condition holds at initialisation, it persists throughout training [Saxe et al., 2014, Arora et al., 2018].

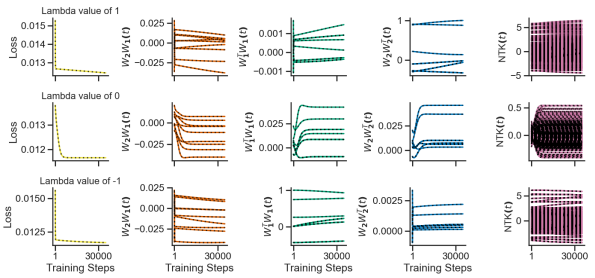

Figure 1: The temporal dynamics of the numerical simulation (colored lines) exactly matched by the analytical solution (black dotted lines) for $\lambda = -1$, $\lambda = 0$ and $\lambda = -1$ initial weight values.($N_h = 2$ and $N_i = N_o = 3$)

Our assumptions are weaker than those in prior work [Fukumizu, 1998, Braun et al., 2022b, Kunin et al., 2024, Xu and Ziyin, 2024, Dominé et al., 2024], which imposed stronger dimensionality constraints. For example, Fukumizu [1998] required equal input and output dimensions ($N_i = N_o$) with $\lambda = 0$, while Braun et al. [2022a], Dominé et al. [2024] allowed $N_i \neq N_o$ and arbitrary λ, but restricted attention to non-bottleneck networks ($N_h = \min(N_i, N_o)$) and non-wide networks. In contrast, we further relax these conditions by considering arbitrary hidden-layer widths, including bottlenecked architectures, with general values of λ at initialization. It is worth noting that the relaxation to wide networks provides only limited benefits: the effective rank of the model remains bounded regardless of the hidden layer width, so expressivity does not significantly improve. Nevertheless, narrow and bottlenecked networks remain highly relevant with established connections to neuroscience. We illustrate our results in Fig 1.