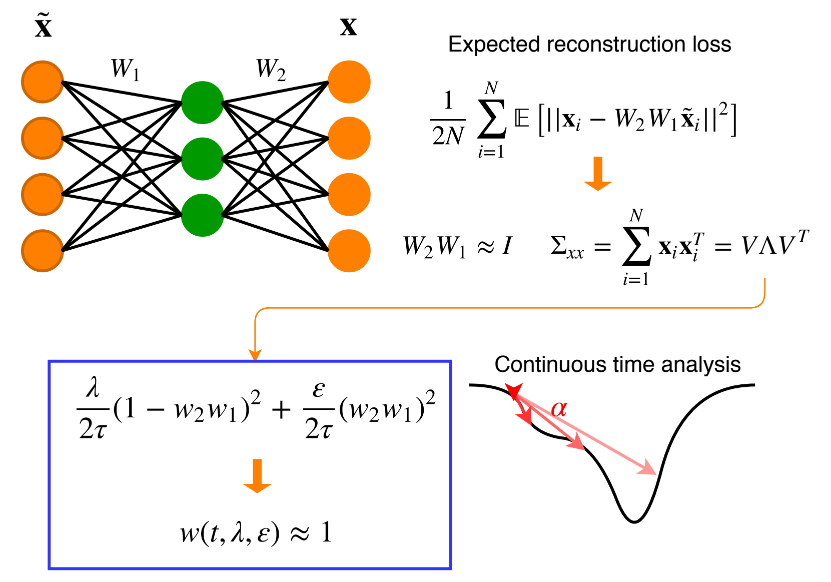
Learning Dynamics of Linear Denoising Autoencoders

Arnu Pretorius Steve Kroon Herman Kamper Stellenbosch University, South Africa

Contributions

We study the learning dynamics of linear denoising autoencoders (DAEs)[1]. Inspired by [2], we derived the learning trajectories of the noise regularised product of network weights during training.



Expected reconstruction loss

$$\frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[||\mathbf{x}_i - W_2W_1\tilde{\mathbf{x}}_i||^2\right]$$

$$W_2W_1 \approx I \quad \Sigma_{xx} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T = V\Lambda V^T$$

$$\frac{\lambda}{2\tau}(1 - w_2w_1)^2 + \frac{\varepsilon}{2\tau}(w_2w_1)^2$$

$$w(t, \lambda, \varepsilon) \approx 1$$

Continuous time analysis

Paper

Code

Our specific contributions:

- Derived learning dynamics for linear DAEs and weight decayed autoencoders (WDAEs).
- Illuminated differences between the dynamics of DAEs and WDAEs: *DAEs seem to exhibit faster training dynamics, even though WDAEs can have larger learning rates.*
- Showed that the theory matches real-world training reasonably well.
- Verified that nonlinear autoencoders have qualitatively similar learning dynamics.

How noise impacts training

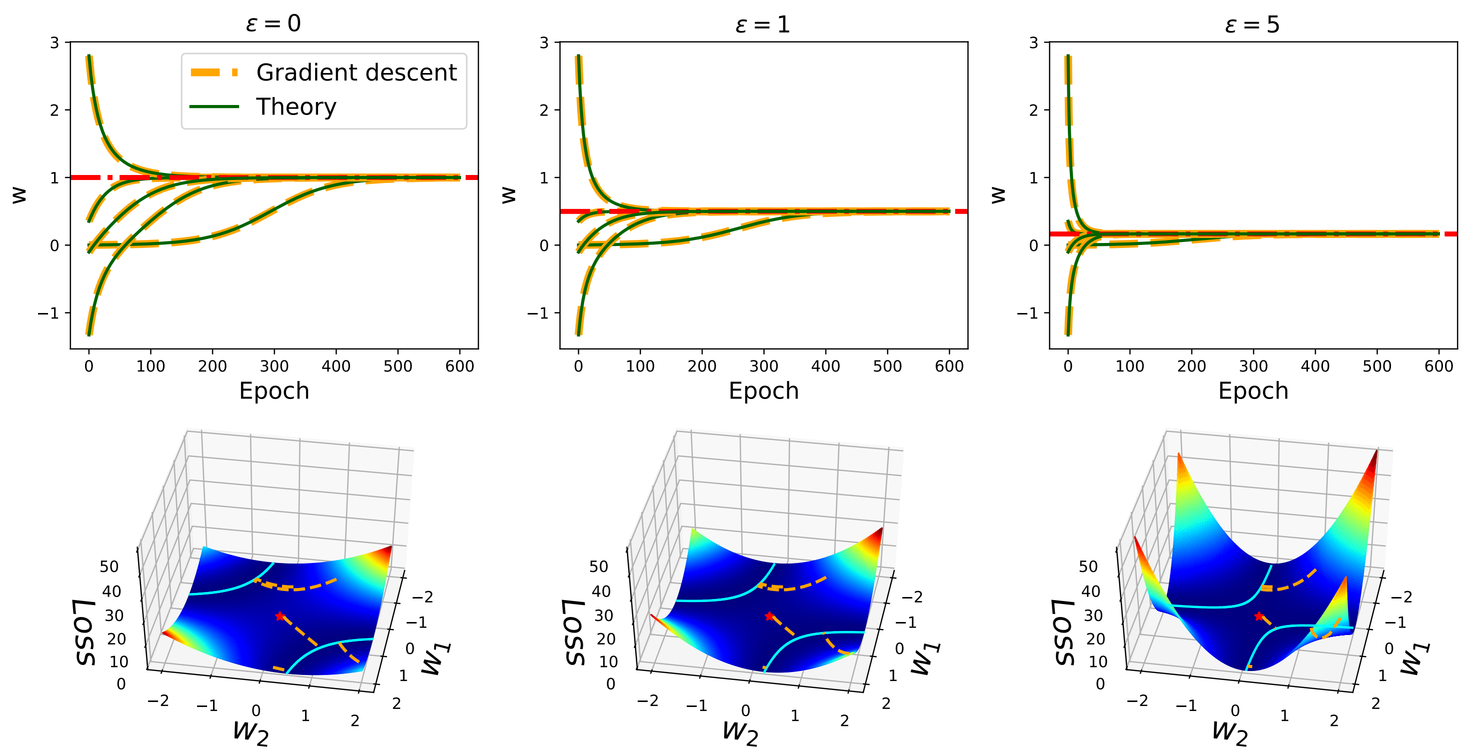


Figure 1: *Hyperbolic learning dynamics, loss surface and gradient descent paths for linear denoising autoencoders.* **Top**: Hyperbolic learning dynamics for each simulated run (dashed orange lines) together with the theoretically predicted learning dynamics (solid green lines). The red line in each plot indicates the final value of the resulting fixed point solution w^*. **Bottom**: The loss surface corresponding to the loss $\ell_\lambda = \frac{\lambda}{2}(1 - w_2w_1)^2 + \frac{\varepsilon}{2}(w_2w_1)^2$ for $\lambda = 1$, as well as the gradient descent paths (dashed orange lines) for randomly initialised weights. The cyan hyperbolas represent the global minimum loss manifold that corresponds to all possible combinations of w_2 and w_1 that minimise ℓ_λ. **Left**: $\varepsilon = 0, w^* = 1$. **Middle**: $\varepsilon = 1, w^* = 0.5$. **Right**: $\varepsilon = 5, w^* = 1/6$.

- Fixed point: $w^* = \frac{\lambda}{\lambda+\varepsilon}$

Noise vs. weight decay

- Equivalent regularisation: $\gamma = \frac{\lambda\epsilon}{\lambda+\varepsilon}$

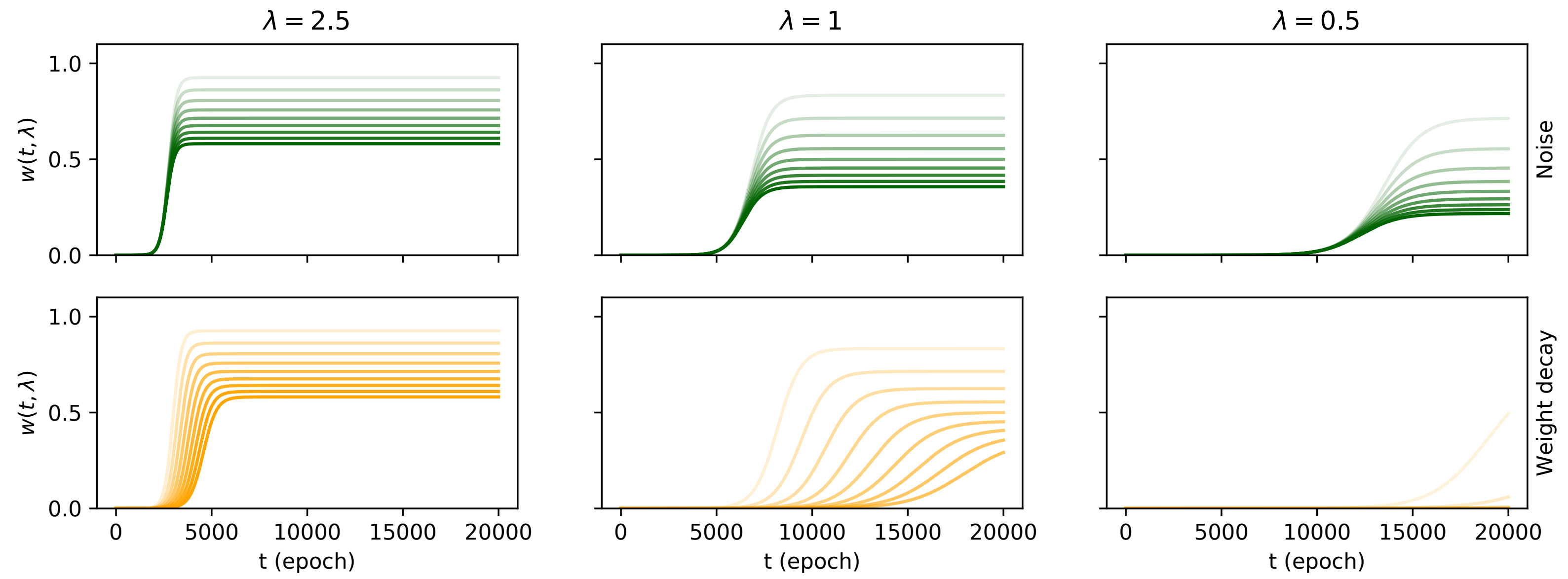


Figure 2: *Theoretically predicted learning dynamics for noise compared to weight decay for linear autoencoders.* **Top**: Noise dynamics (green), darker line colours correspond to larger amounts of added noise. **Bottom**: Weight decay dynamics (orange), darker line colours correspond to larger amounts of regularisation. **Left to right**: Eigenvalues $\lambda = 2.5, 1$ and 0.5 associated with high to low variance.

- Ratio of the optimal learning rate for DAEs vs. WDAEs: $R = \frac{2\lambda+\gamma}{2\lambda+3\varepsilon}$

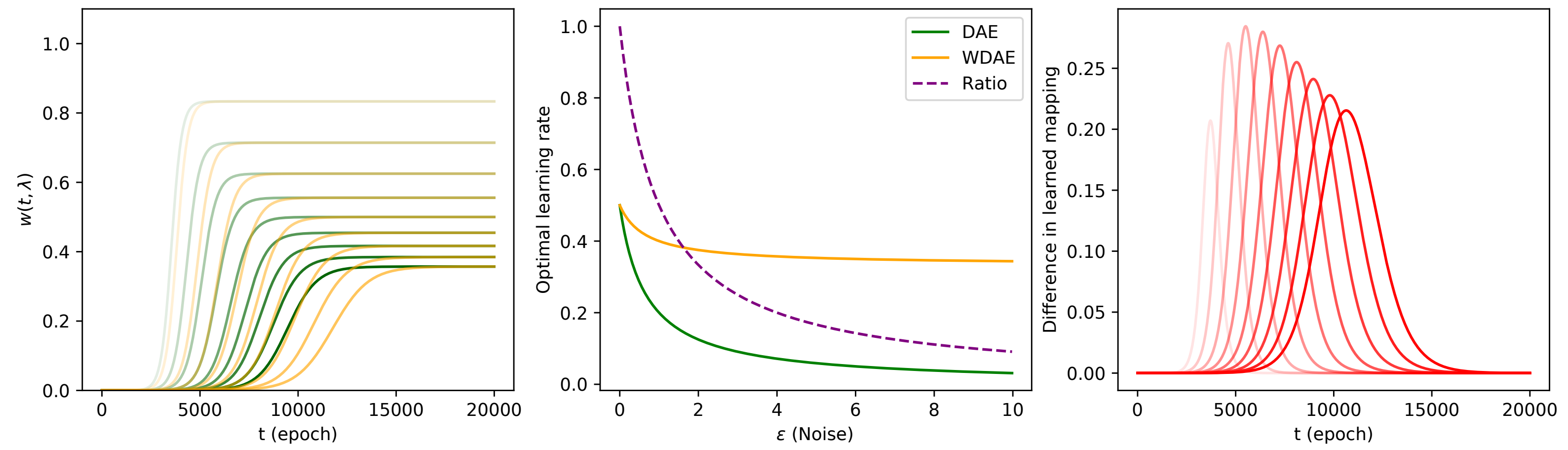


Figure 3: *Learning dynamics for optimal discrete time learning rates ($\lambda = 1$).* **Left**: Dynamics of DAEs (green) vs. WDAEs (orange), where darker line colours correspond to larger amounts noise or weigh decay. **Middle**: Optimal learning rate as a function of noise ε for DAEs, and for WDAEs using an equivalent amount of regularisation $\gamma = \lambda\varepsilon/(\lambda + \varepsilon)$. **Right**: Difference in mapping over time.

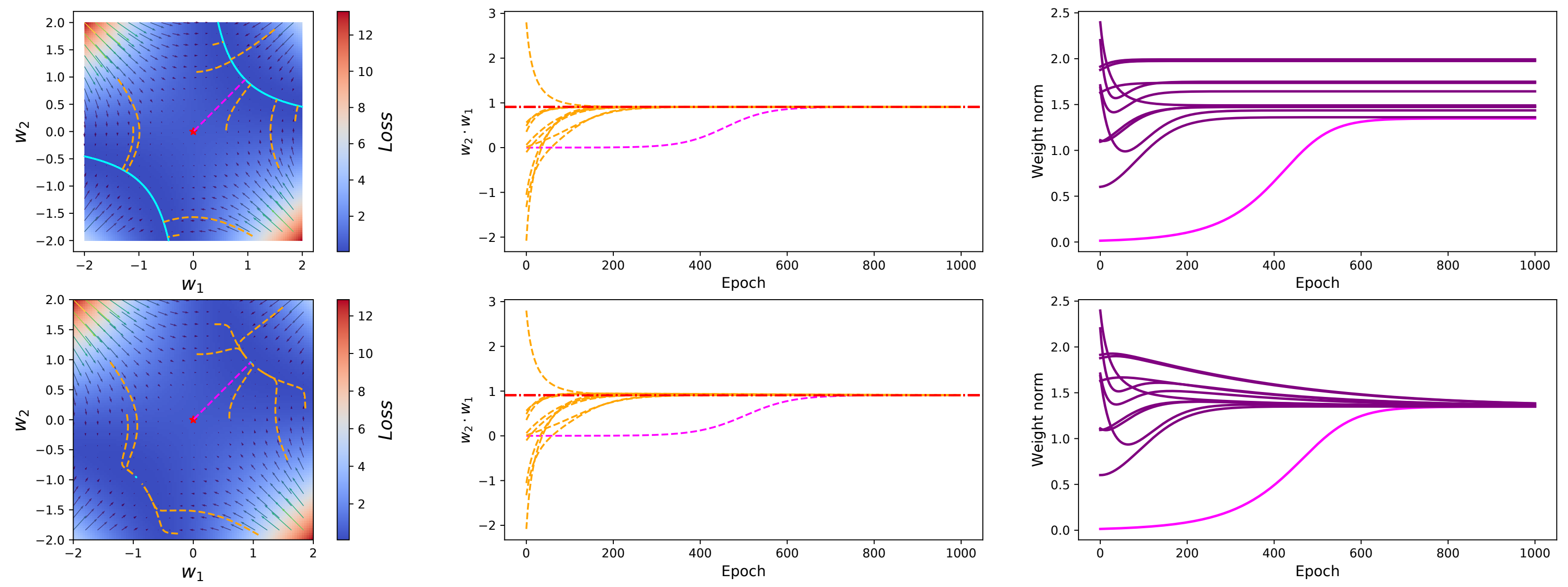


Figure 4: *The effect of noise versus weight decay on the norm of the weights during learning.* **Left**: Two-dimensional loss surface $\ell_\lambda = \frac{\lambda}{2}(1 - w_2w_1)^2 + \frac{\varepsilon}{2}(w_2w_1)^2 + \frac{\gamma}{2}(w_2^2 + w_1^2)$. Gradient descent paths (orange/magenta dashed lines), minimum loss manifold (cyan curves), saddle point (red star). **Middle**: Simulated learning dynamics. **Right**: Norm of the weights over time for each simulated run. **Top**: Noise with $\lambda = 1, \varepsilon = 0.1$ and $\gamma = 0$. **Bottom**: Weight decay with $\lambda = 1, \varepsilon = 0$ and $\gamma = \lambda(0.1)/(\lambda + 0.1) = 0.091$. The magenta line in each plot corresponds to a simulated run with small initialised weights.

Experimental results

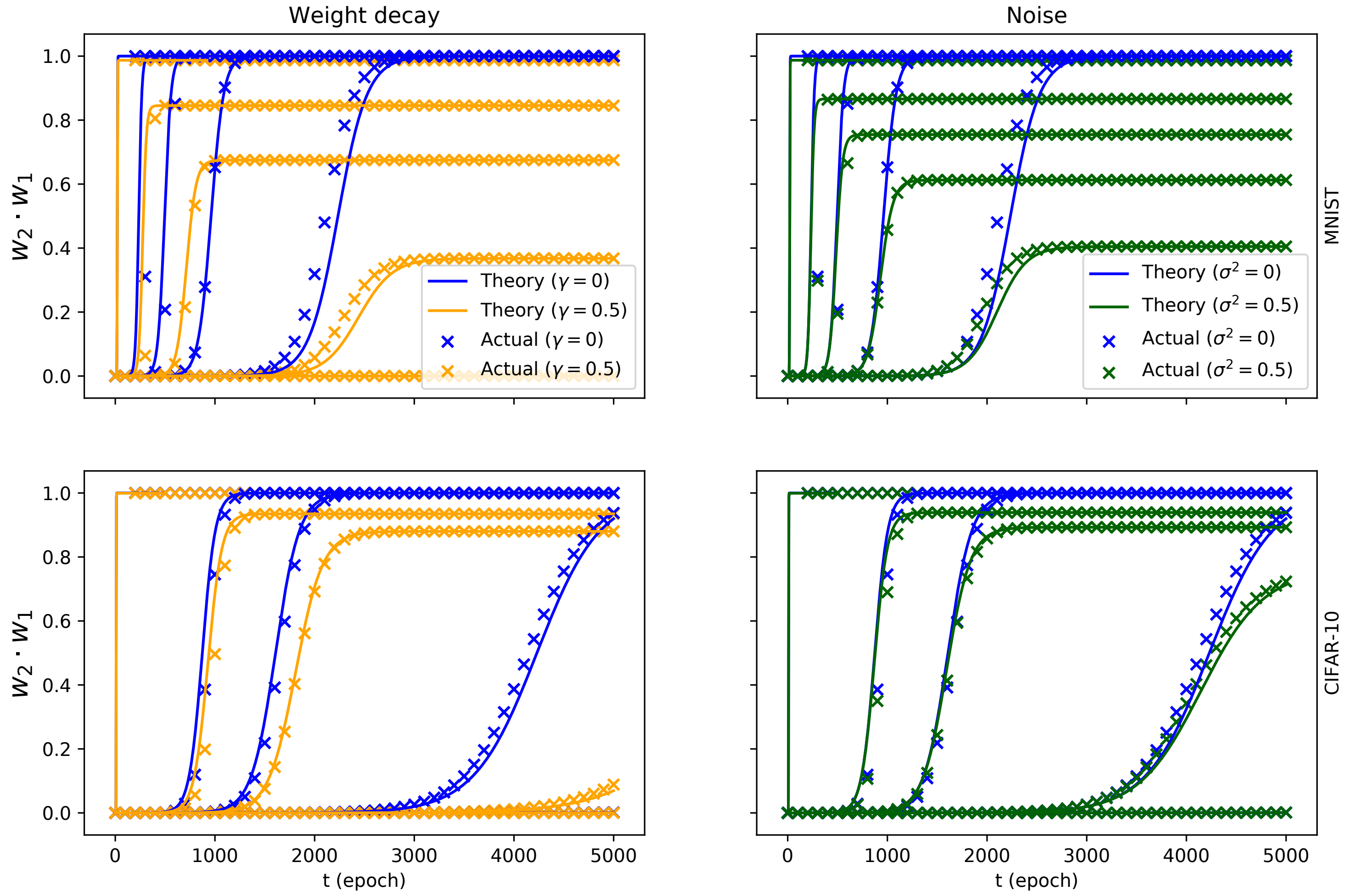


Figure 5: *Learning dynamics for MNIST and CIFAR-10.* Solid lines represent theoretical dynamics and 'x' markers simulated dynamics. **Left**: Weight decay: AE (blue) vs. WDAE with $\gamma = 0.5$ (orange). **Right**: Noise: AE (blue) vs. DAE with $\sigma^2 = 0.5$ (green). **Top**: MNIST. **Bottom**: CIFAR-10.

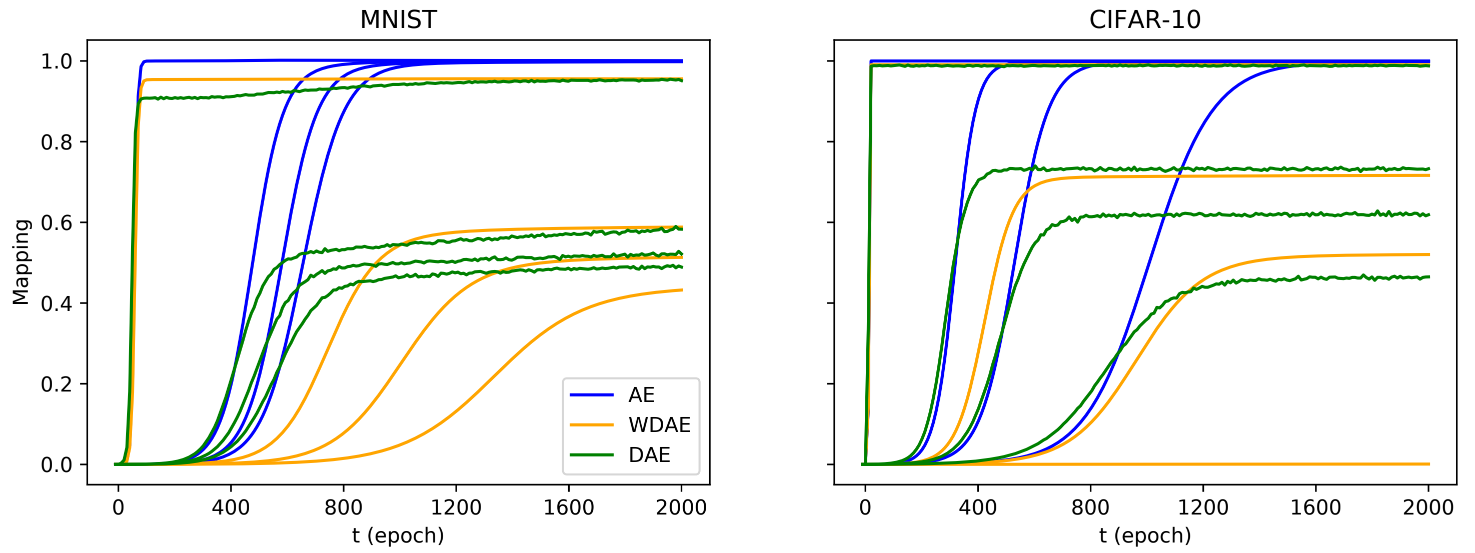


Figure 6: *Learning dynamics for nonlinear networks using ReLU activation.* AE (blue), WDAE (orange) and DAE (green). **Left**: MNIST **Right**: CIFAR-10.

References

[1] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

[2] Saxe, A.M., McClelland, J.L. and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Acknowledgements

We would like to thank Google, ICML, the Science faculty of Stellenbosch University and the CSIR/SU Centre for Artificial Intelligence Research (CAIR), South Africa, for financial support.

Code: https://github.com/arnupretorius/lindaedynamics_icml2018

Contact: arnupretorius@gmail.com