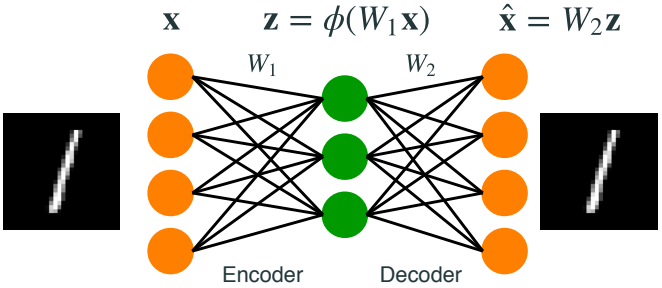
Learning Dynamics of Linear Denoising Autoencoders

Arnu Pretorius, Steve Kroon and Herman Kamper

Stellenbosch University, South Africa

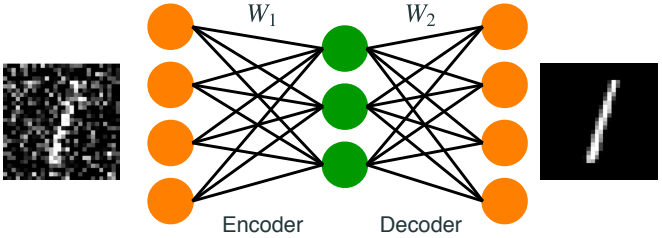
Theoretical foundations of data science workshop, AIMS 2018

Introduction



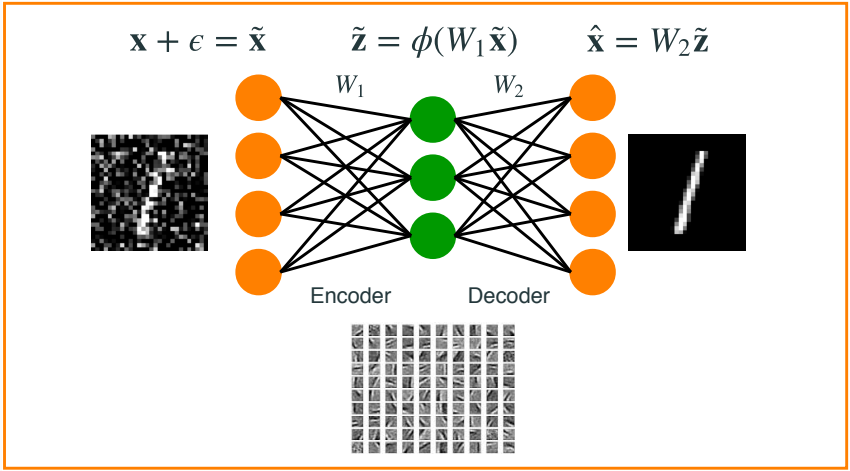
Introduction

$$\mathbf{x} + \epsilon = \tilde{\mathbf{x}} \qquad \tilde{\mathbf{z}} = \phi(W_1 \tilde{\mathbf{x}}) \qquad \hat{\mathbf{x}} = W_2 \tilde{\mathbf{z}}$$



- Sample components of ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$

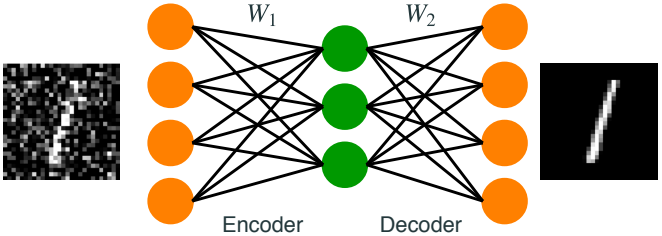
Introduction



- Sample components of ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$

Introduction

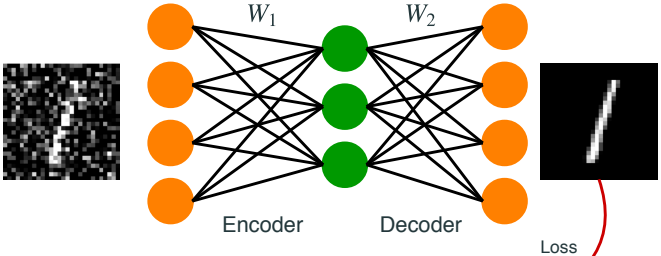
$$\mathbf{x} + \epsilon = \tilde{\mathbf{x}} \qquad \tilde{\mathbf{z}} = \cancel{\phi}(W_1 \tilde{\mathbf{x}}) \qquad \hat{\mathbf{x}} = W_2 \tilde{\mathbf{z}}$$



- Sample components of ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$

Introduction

$$\mathbf{x} + \epsilon = \tilde{\mathbf{x}} \qquad \tilde{\mathbf{z}} = \cancel{\phi}(W_1 \tilde{\mathbf{x}}) \qquad \hat{\mathbf{x}} = W_2 \tilde{\mathbf{z}}$$

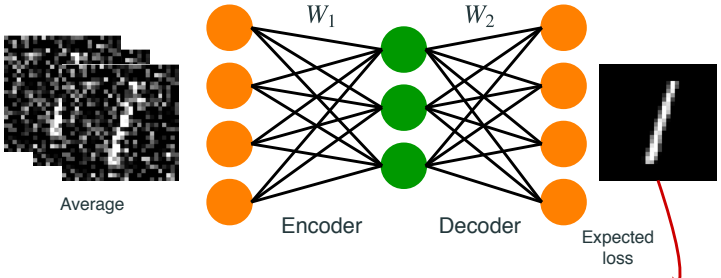


$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} ||\mathbf{x}_i - W_2 W_1 \tilde{\mathbf{x}}_i||^2$$

- Sample components of ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$

Introduction

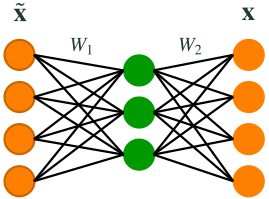
$$\mathbf{x} + \epsilon = \tilde{\mathbf{x}} \qquad \tilde{\mathbf{z}} = \cancel{\phi}(W_1 \tilde{\mathbf{x}}) \qquad \hat{\mathbf{x}} = W_2 \tilde{\mathbf{z}}$$



$$\mathbb{E}[\mathcal{L}] = \frac{1}{2N} \sum_{i=1}^{N} ||\mathbf{x}_i - W_2 W_1 \mathbf{x}_i||^2 + \varepsilon \text{tr}(W_2 W_1 W_1^T W_2^T)$$

- Sample components of ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$
- With $\varepsilon = N\sigma^2$

Learning dynamics for linear denoising autoencoders (DAEs)



Expected reconstruction loss

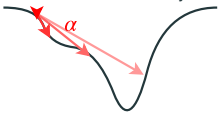
$$\frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[||\mathbf{x}_i - W_2W_1\tilde{\mathbf{x}}_i||^2\right]$$

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

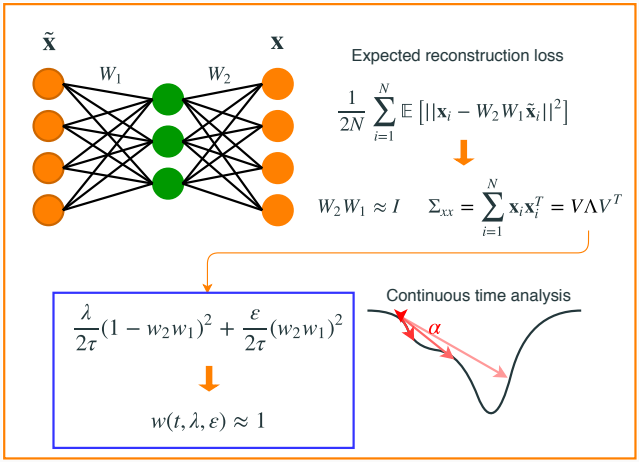
$$\frac{\lambda}{2\tau}(1 - w_2w_1)^2 + \frac{\varepsilon}{2\tau}(w_2w_1)^2$$

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

Continuous time analysis

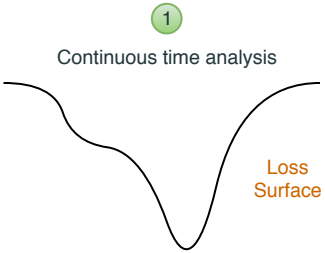


Learning dynamics for linear denoising autoencoders (DAEs)

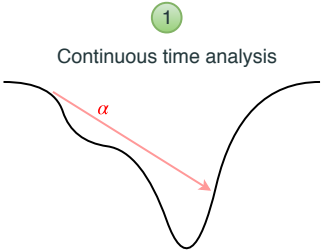


- *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, Saxe, McClelland, Ganguli. ICLR, 2014.

Approach to deriving dynamics equations



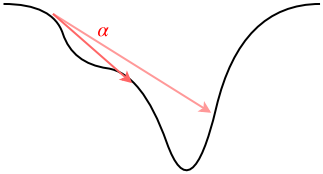
Approach to deriving dynamics equations



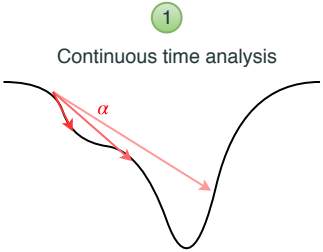
Approach to deriving dynamics equations

1

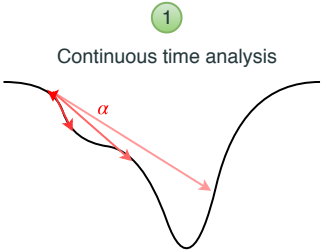
Continuous time analysis



Approach to deriving dynamics equations



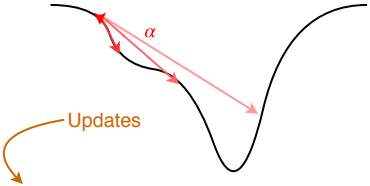
Approach to deriving dynamics equations



Approach to deriving dynamics equations

1

Continuous time analysis



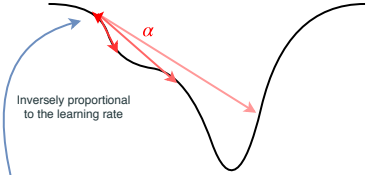
$$\tau \frac{d}{dt} W_1 = W_2^T (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

Approach to deriving dynamics equations

1

Continuous time analysis



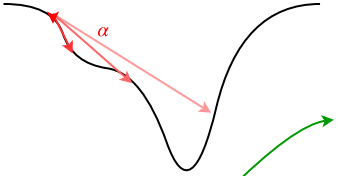
$$\tau \frac{d}{dt} W_1 = W_2^T (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

Approach to deriving dynamics equations

1

Continuous time analysis

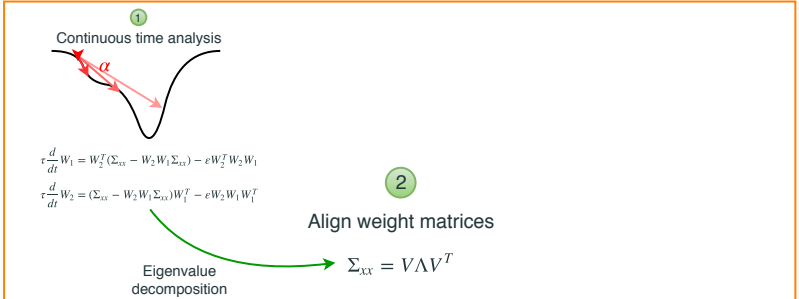


$$\Sigma_{xx} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T$$

$$\tau \frac{d}{dt} W_1 = W_2^T (\Sigma_{xy}) - W_2 W (\Sigma_{xy}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xy} - W_2 W_1 (\Sigma_{xy})) W_1^T - \varepsilon W_2 W_1 W_1^T$$

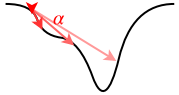
Approach to deriving dynamics equations



Approach to deriving dynamics equations

1

Continuous time analysis



$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

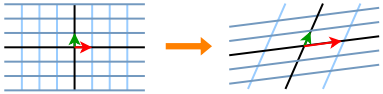
$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2

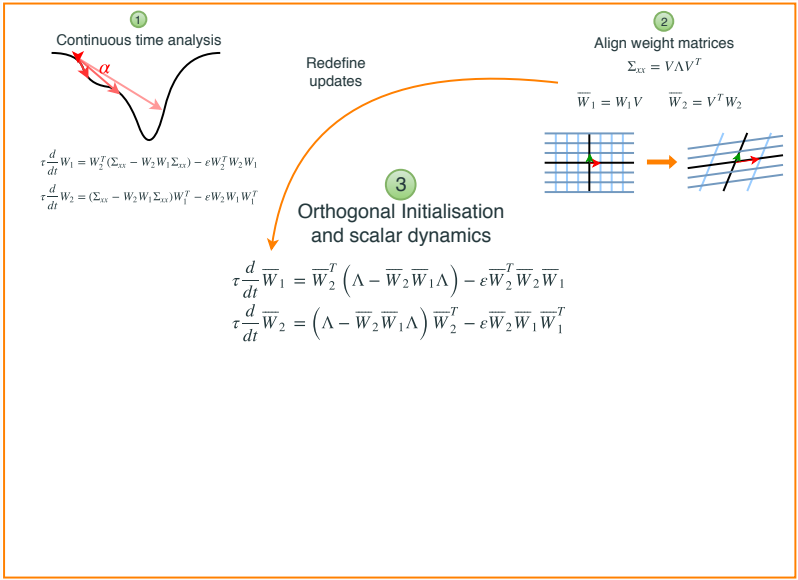
Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$

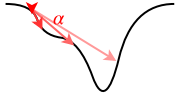


Approach to deriving dynamics equations



Approach to deriving dynamics equations

1 Continuous time analysis



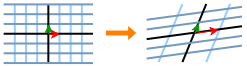
$$\tau \frac{d}{dt} W_1 = W_2^T (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{yx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\tau \frac{d}{dt} \overline{W}_1 = \overline{W}_2^T \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) - \varepsilon \overline{W}_2^T \overline{W}_2 \overline{W}_1$$

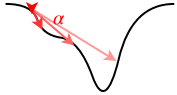
$$\tau \frac{d}{dt} \overline{W}_2 = \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) \overline{W}_2^T - \varepsilon \overline{W}_2 \overline{W}_1 \overline{W}_1^T$$

$$\overline{\mathcal{L}} = \frac{1}{2\tau} ||\Lambda - \overline{W}_2 \overline{W}_1 \Lambda||^2 + \frac{\varepsilon}{2\tau} \mathrm{tr} \left(\overline{W}_2 \overline{W}_1 \overline{W}_1^T \overline{W}_2^T \right)$$

Diagonal Input-output

Approach to deriving dynamics equations

1 Continuous time analysis



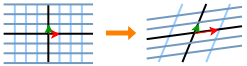
$$\tau \frac{d}{dt} W_1 = W_2^T (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\tau \frac{d}{dt} \overline{W}_1 = \overline{W}_2^T \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) - \varepsilon \overline{W}_2^T \overline{W}_2 \overline{W}_1$$

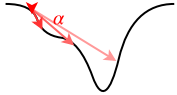
$$\tau \frac{d}{dt} \overline{W}_2 = \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) \overline{W}_2^T - \varepsilon \overline{W}_2 \overline{W}_1 \overline{W}_1^T$$

$$\overline{\mathcal{L}} = \frac{1}{2\tau} ||\Lambda - \overline{W}_2 \overline{W}_1 \Lambda||^2 + \frac{\varepsilon}{2\tau} \mathrm{tr} \left(\overline{W}_2 \overline{W}_1 \overline{W}_1^T \overline{W}_2^T \right)$$

Non-diagonal Weights

Approach to deriving dynamics equations

1 Continuous time analysis



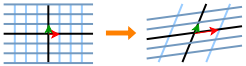
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\tau \frac{d}{dt} \overline{W}_1 = \overline{W}_2^T \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) - \varepsilon \overline{W}_2^T \overline{W}_2 \overline{W}_1$$

$$\tau \frac{d}{dt} \overline{W}_2 = \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) \overline{W}_2^T - \varepsilon \overline{W}_2 \overline{W}_1 \overline{W}_1^T$$

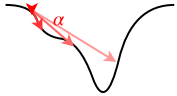
$$\overline{\mathcal{L}} = \frac{1}{2\tau} ||\Lambda - \overline{W}_2 \overline{W}_1 \Lambda||^2 + \frac{\varepsilon}{2\tau} \mathrm{tr} \left(\overline{W}_2 \overline{W}_1 \overline{W}_1^T \overline{W}_2^T \right)$$

$$W_2 = V D_2 R^T, W_1 = R D_1 V^T$$

Orthogonal initialisation

Approach to deriving dynamics equations

1 Continuous time analysis



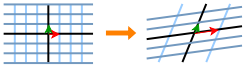
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\tau \frac{d}{dt} \overline{W}_1 = \overline{W}_2^T \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) - \varepsilon \overline{W}_2^T \overline{W}_2 \overline{W}_1$$

$$\tau \frac{d}{dt} \overline{W}_2 = \left(\Lambda - \overline{W}_2 \overline{W}_1 \Lambda \right) \overline{W}_2^T - \varepsilon \overline{W}_2 \overline{W}_1 \overline{W}_1^T$$

$$\overline{\mathcal{L}} = \frac{1}{2\tau} ||\Lambda - \overline{W}_2 \overline{W}_1 \Lambda||^2 + \frac{\varepsilon}{2\tau} \mathrm{tr} \left(\overline{W}_2 \overline{W}_1 \overline{W}_1^T \overline{W}_2^T \right)$$

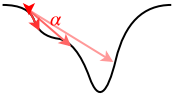
$$W_2 = V D_2 R^T, W_1 = R D_1 V^T$$

Scalar loss

$$\ell = \frac{\lambda}{2\tau} (1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau} (w_2 w_1)^2$$

Approach to deriving dynamics equations

1 Continuous time analysis



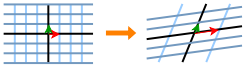
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

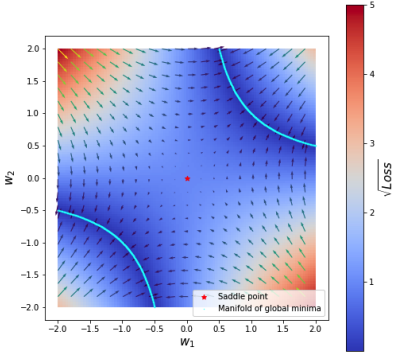
$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

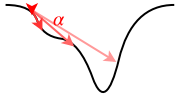
$$\ell = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2$$

Scalar loss surface



Approach to deriving dynamics equations

1 Continuous time analysis



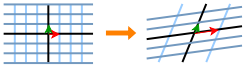
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



4 Exact solutions

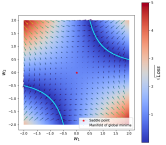
$$\tau \frac{d}{dt} w_1 = w_2 \lambda (1 - w_2 w_1) - \varepsilon w_2^2 w_1$$

$$\tau \frac{d}{dt} w_2 = w_1 \lambda (1 - w_2 w_1) - \varepsilon w_2 w_1^2$$

3 Orthogonal Initialisation and scalar dynamics

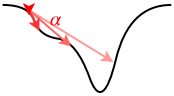
$$\ell = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2$$

Scalar updates



Approach to deriving dynamics equations

1 Continuous time analysis



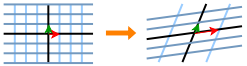
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

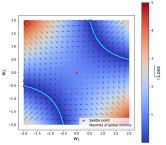
$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\ell = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2$$



4 Exact solutions

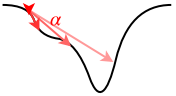
$$\tau \frac{d}{dt} w_1 = w_2 \lambda (1 - w_2 w_1) - \varepsilon w_2^2 w_1$$

$$\tau \frac{d}{dt} w_2 = w_1 \lambda (1 - w_2 w_1) - \varepsilon w_2 w_1^2$$

$$t = \int_{w_0}^{w_t} f(w_1, w_2, \lambda, \varepsilon) dt$$

Approach to deriving dynamics equations

1 Continuous time analysis



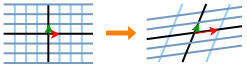
$$\tau \frac{d}{dt} W_1 = W_2^T(\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) - \varepsilon W_2^T W_2 W_1$$

$$\tau \frac{d}{dt} W_2 = (\Sigma_{xx} - W_2 W_1 \Sigma_{xx}) W_1^T - \varepsilon W_2 W_1 W_1^T$$

2 Align weight matrices

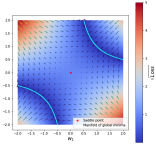
$$\Sigma_{xx} = V \Lambda V^T$$

$$\overline{W}_1 = W_1 V \qquad \overline{W}_2 = V^T W_2$$



3 Orthogonal Initialisation and scalar dynamics

$$\ell = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2$$



4 Exact solutions

$$\tau \frac{d}{dt} w_1 = w_2 \lambda (1 - w_2 w_1) - \varepsilon w_2^2 w_1$$

$$\tau \frac{d}{dt} w_2 = w_1 \lambda (1 - w_2 w_1) - \varepsilon w_2 w_1^2$$

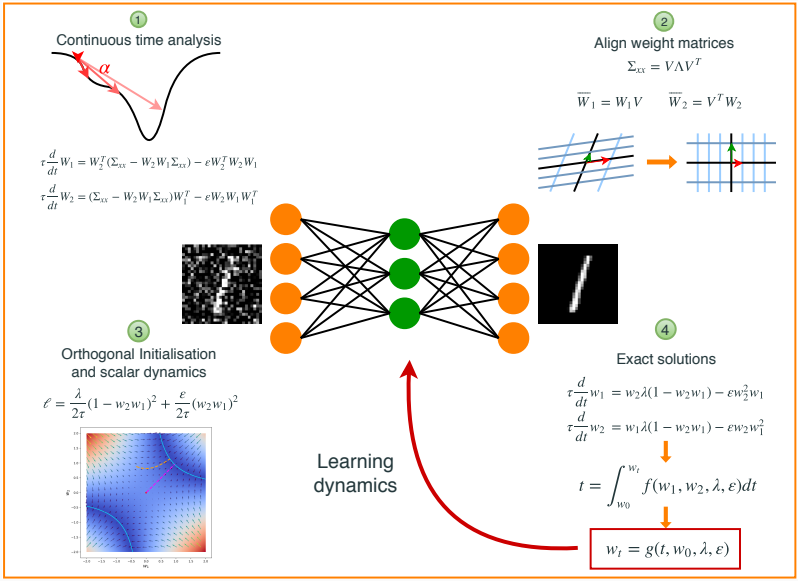
$$\downarrow$$

$$t = \int_{w_0}^{w_t} f(w_1, w_2, \lambda, \varepsilon) dt$$

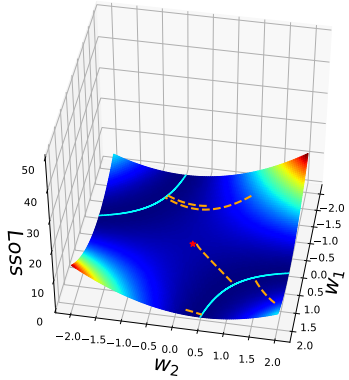
$$\downarrow$$

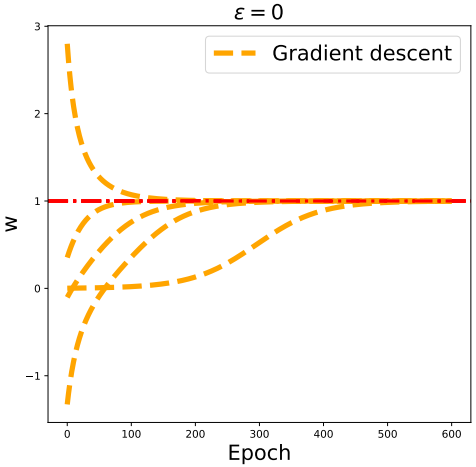
$$w_t = g(t, w_0, \lambda, \varepsilon)$$

Approach to deriving dynamics equations

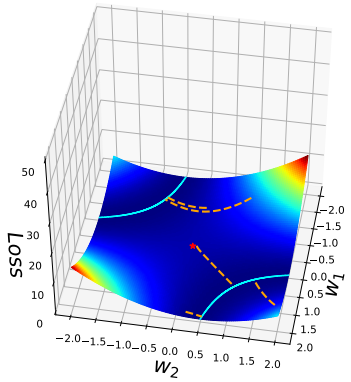


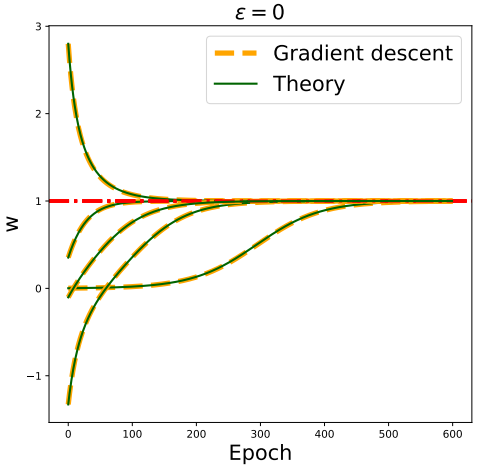
Theory vs. simulated dynamics



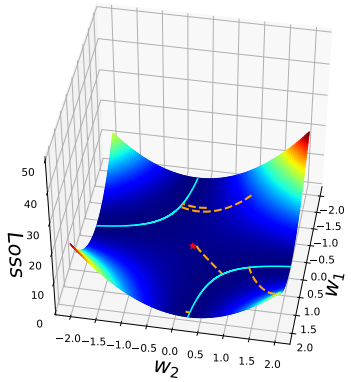


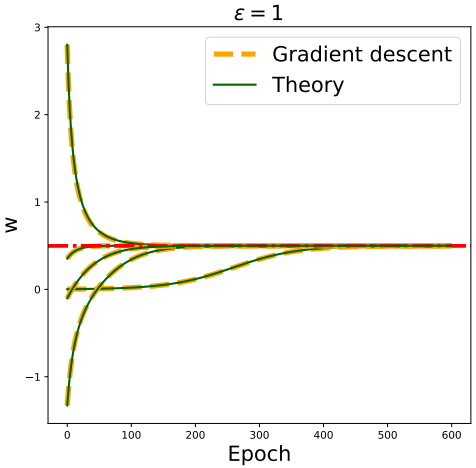
Theory vs. simulated dynamics



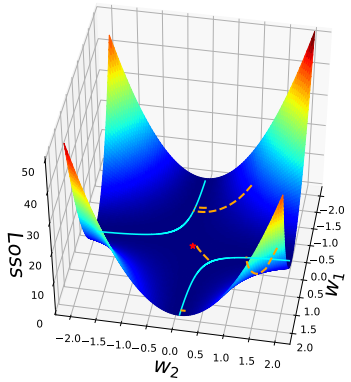


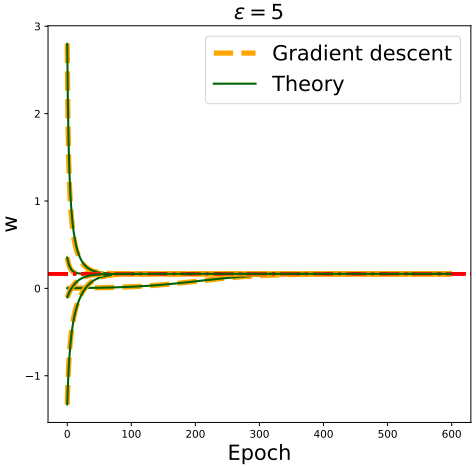
Theory vs. simulated dynamics



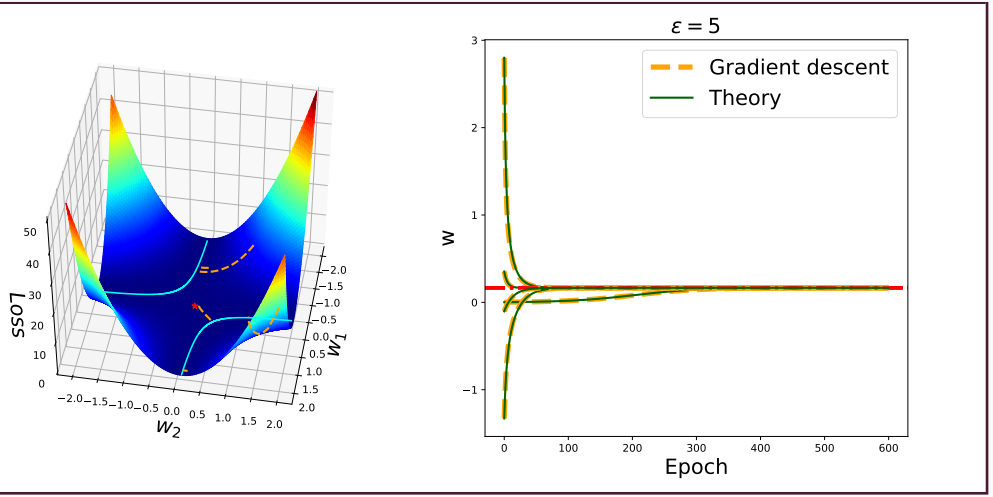


Theory vs. simulated dynamics



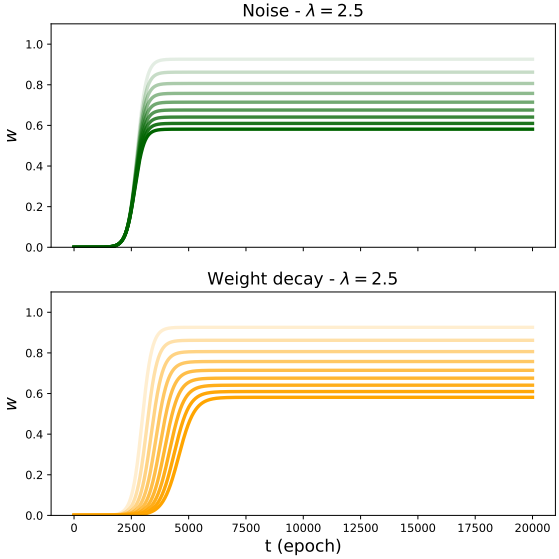


Theory vs. simulated dynamics

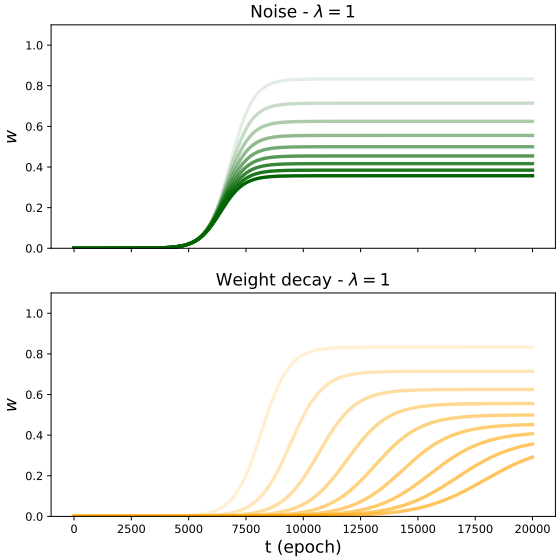


- Fixed point: $w^* = \frac{\lambda}{\lambda+\varepsilon}$

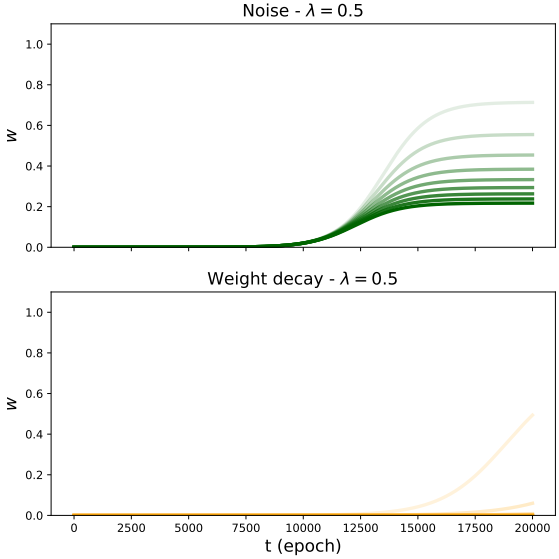
The relationship between noise and weight decay



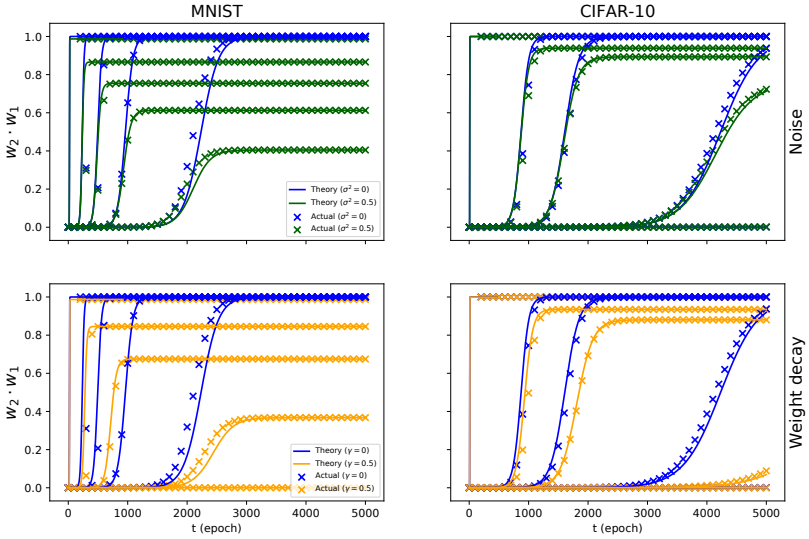
The relationship between noise and weight decay



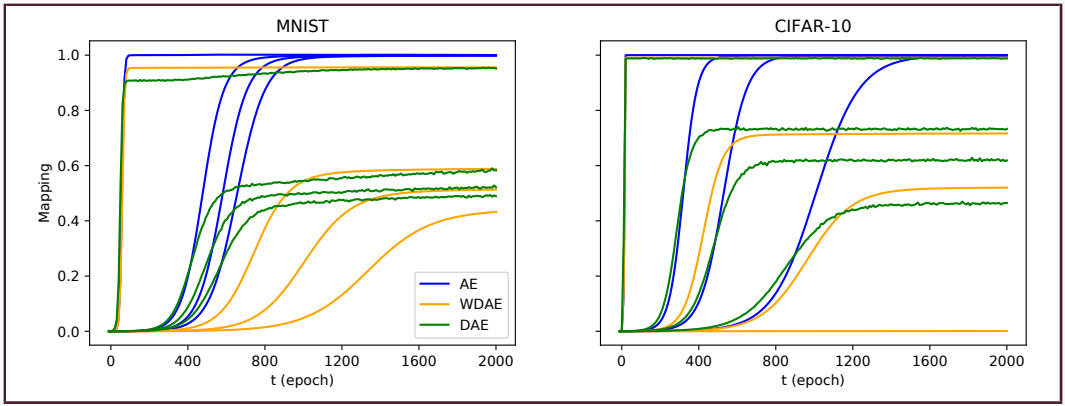
The relationship between noise and weight decay



Experimental results: Linear autoencoder networks



Experimental results: Nonlinear autoencoders using ReLU



Conclusion

Summary

- Derived equations for the learning dynamics of DAEs and WDAEs.

Conclusion

Summary

- Derived equations for the learning dynamics of DAEs and WDAEs.
- Illuminated a fundamental difference between the dynamics of DAEs and WDAEs: *DAEs seem to exhibit faster training dynamics*.

Conclusion

Summary

- Derived equations for the learning dynamics of DAEs and WDAEs.
- Illuminated a fundamental difference between the dynamics of DAEs and WDAEs: *DAEs seem to exhibit faster training dynamics*.
- Showed that the theory matches real-world training reasonably well.

Conclusion

Summary

- Derived equations for the learning dynamics of DAEs and WDAEs.
- Illuminated a fundamental difference between the dynamics of DAEs and WDAEs: *DAEs seem to exhibit faster training dynamics*.
- Showed that the theory matches real-world training reasonably well.
- Verified that our linear predictions are qualitatively able to describe the learning dynamics of nonlinear autoencoder networks.

Source code to reproduce all the results

https://github.com/arnupretorius/lindaedynamics_icml2018

Source code to reproduce all the results

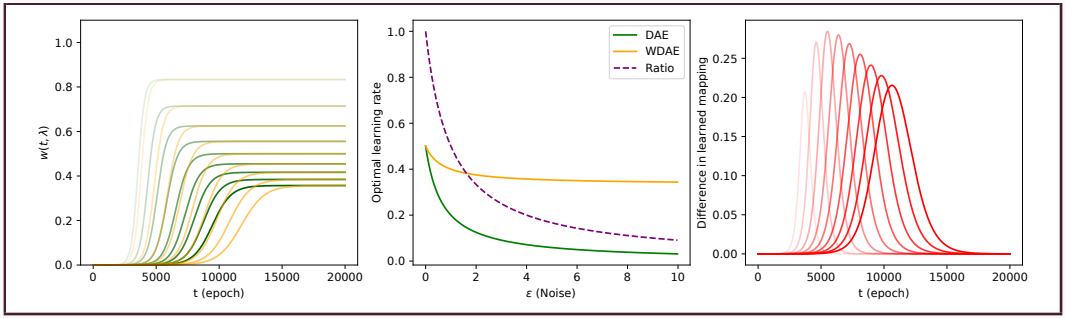
https://github.com/arnupretorius/lindaedynamics_icml2018

Thank you for listening!

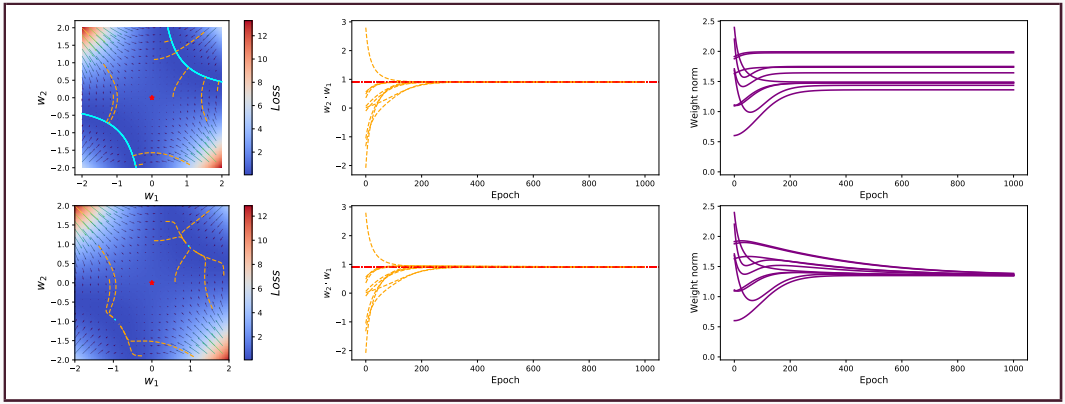
The relationship between noise and weight decay

- Ratio of optimal learning rate for DAE vs. WDAE:

$$R = \frac{2\lambda + \gamma}{2\lambda + 3\varepsilon}.$$



The relationship between noise and weight decay



The relationship between noise and weight decay