# Learning Dynamics of Linear Denoising Autoencoders

Arnu Pretorius, Steve Kroon and Herman Kamper

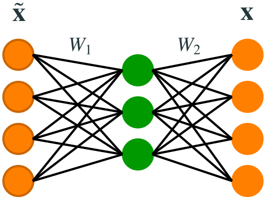
Stellenbosch University, South Africa

35th International Conference on Machine Learning, 2018

cair
CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

# Linear denoising autoencoders (DAE)



Expected reconstruction loss

$$\frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[||\mathbf{x}_i - W_2 W_1 \tilde{\mathbf{x}}_i||^2\right]$$

$$\Downarrow$$

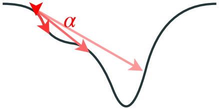
$$W_2 W_1 \approx I \qquad \Sigma_{xx} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T = V\Lambda V^T$$

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

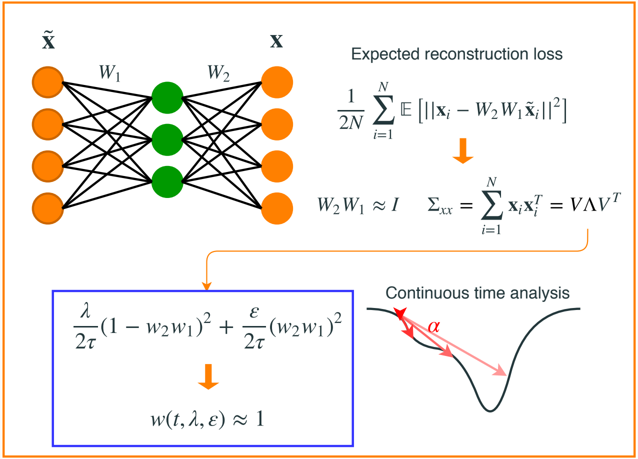
$$\Downarrow$$

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

Continuous time analysis



# Linear denoising autoencoders (DAE)



Expected reconstruction loss

$$\frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[||\mathbf{x}_i - W_2 W_1 \tilde{\mathbf{x}}_i||^2\right]$$

$$W_2 W_1 \approx I \qquad \Sigma_{xx} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T = V\Lambda V^T$$

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

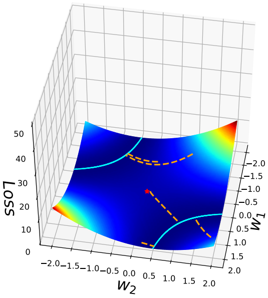
$$w(t, \lambda, \varepsilon) \approx 1$$

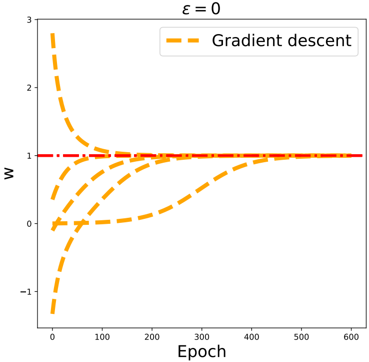
Continuous time analysis

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

# Learning dynamics

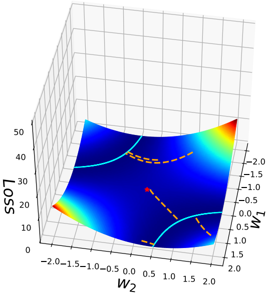
$$\ell_{\varepsilon} = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow \text{GD learning dynamics}$$

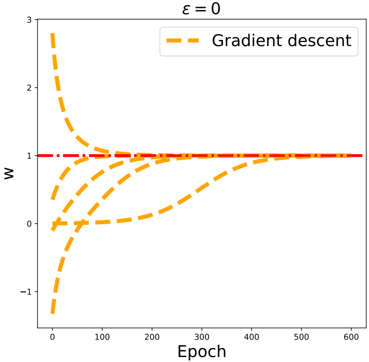




# Learning dynamics

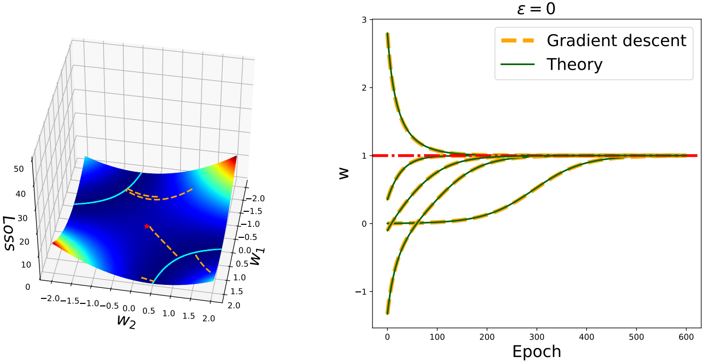
$$\ell_\varepsilon = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow w(t, \lambda, \varepsilon) \longrightarrow \text{GD learning dynamics}$$





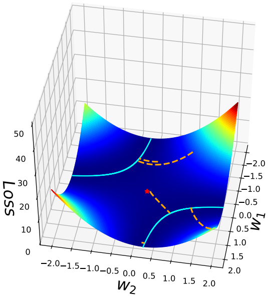
# Learning dynamics

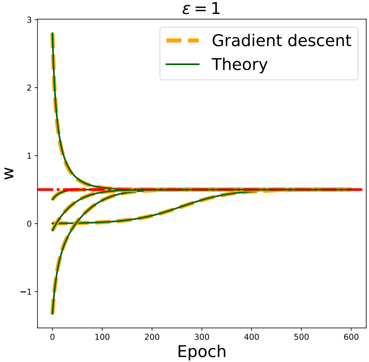
$$\ell_{\varepsilon} = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow w(t, \lambda, \varepsilon) \longrightarrow \text{GD learning dynamics}$$



# Learning dynamics

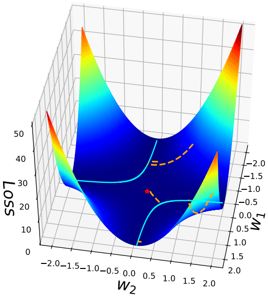
$$\ell_\varepsilon = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow w(t, \lambda, \varepsilon) \longrightarrow \text{GD learning dynamics}$$

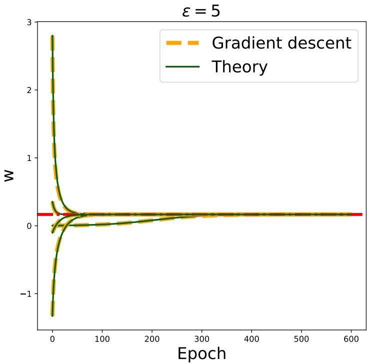




# Learning dynamics

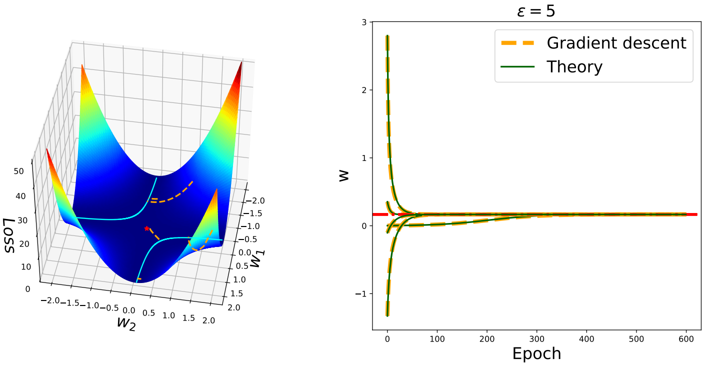
$$\ell_\varepsilon = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow w(t, \lambda, \varepsilon) \longrightarrow$$ GD learning dynamics





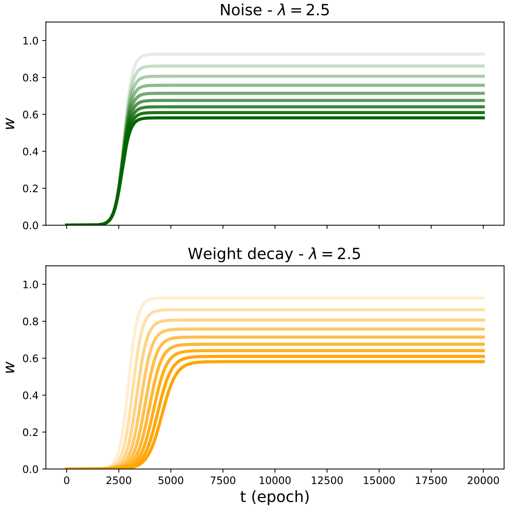
# Learning dynamics

$$\ell_{\varepsilon} = \frac{\lambda}{2\tau}(1 - w_2 w_1)^2 + \frac{\varepsilon}{2\tau}(w_2 w_1)^2 \longrightarrow w(t, \lambda, \varepsilon) \longrightarrow \text{GD learning dynamics}$$

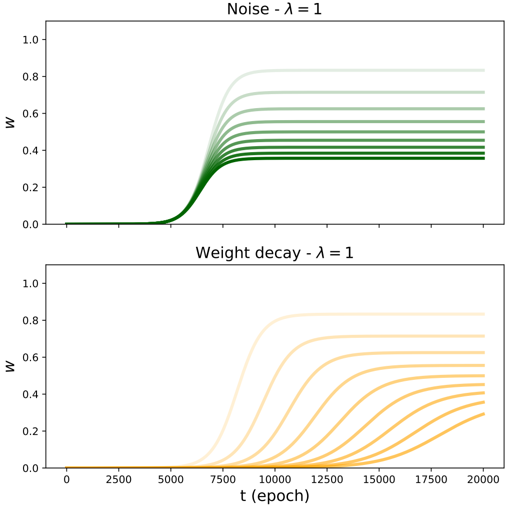


- Fixed point: $w^* = \frac{\lambda}{\lambda+\varepsilon}$

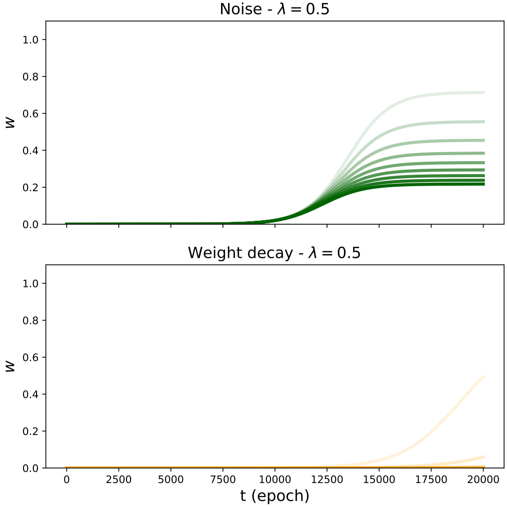
# The relationship between noise and weight decay



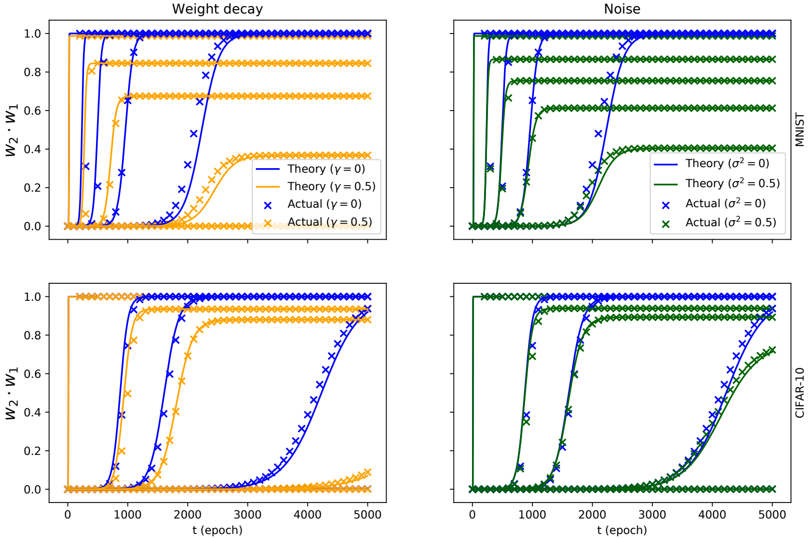
# The relationship between noise and weight decay



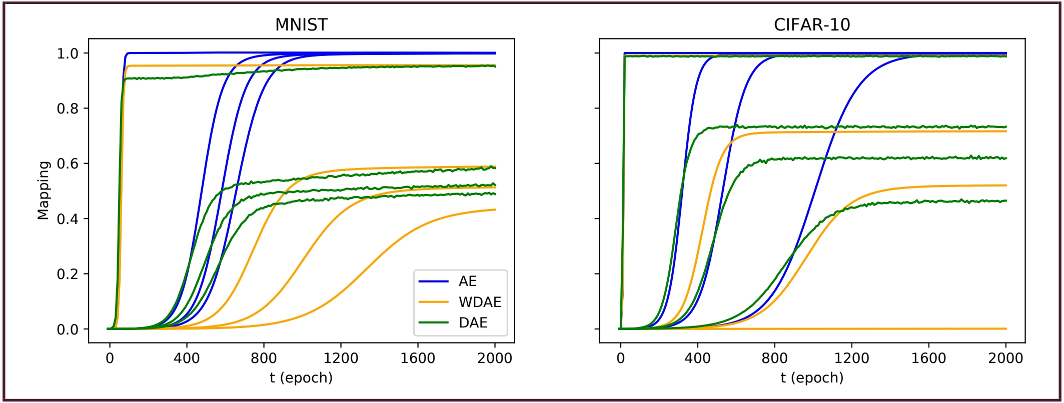
# The relationship between noise and weight decay



# Experimental results: Linear autoencoder networks



# Experimental results: Nonlinear autoencoders using ReLU



**Thank you for listening!**

**Source code to reproduce all the results**

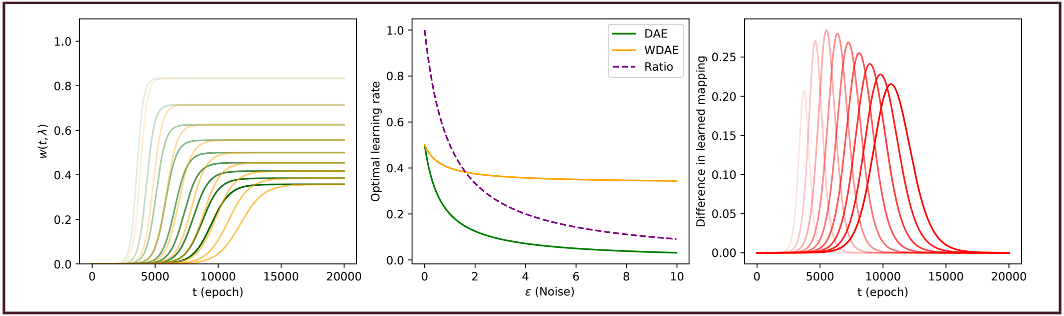
https:
//github.com/arnupretorius/lindaedynamics_icml2018

# The relationship between noise and weight decay

## Optimal discrete time learning rates

- Ratio for DAE to WDAE:

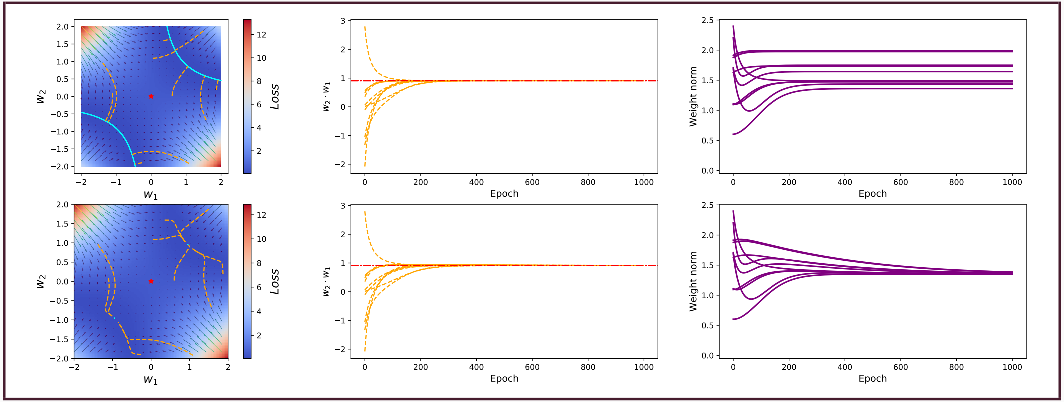
$$R = \frac{2\lambda + \gamma}{2\lambda + 3\varepsilon}.$$



# The relationship between noise and weight decay

## Motivation for weight decay

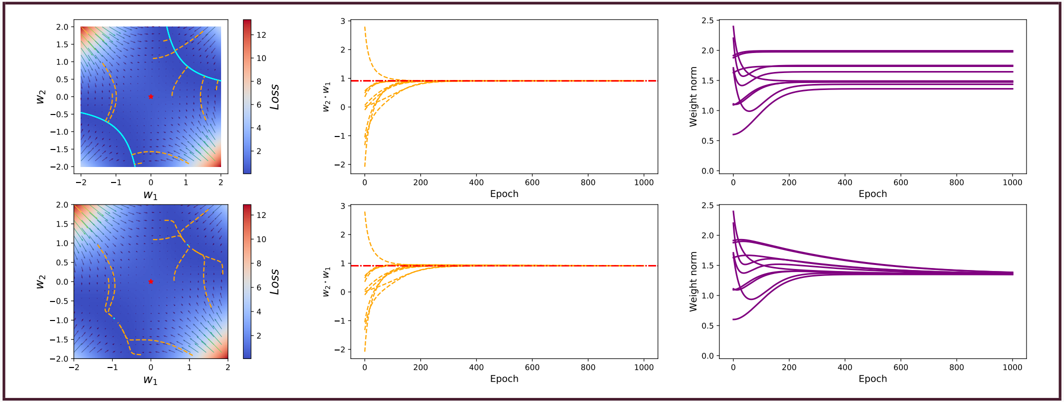
- Smaller weights $\Longrightarrow$ smoother models $\Longrightarrow$ better generalisation



# The relationship between noise and weight decay

## Motivation for weight decay

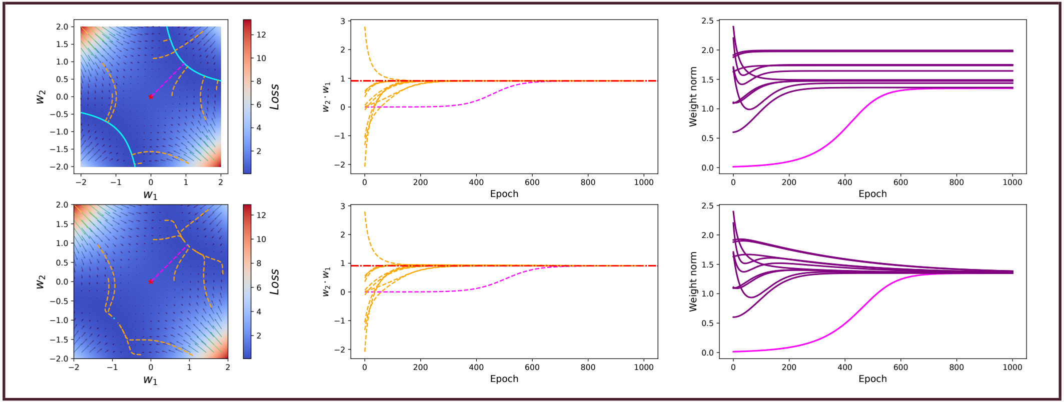
- Smaller weights $\implies$ smoother models $\implies$ better generalisation
- Small weight initialisation?



# The relationship between noise and weight decay

## Motivation for weight decay

- Smaller weights $\implies$ smoother models $\implies$ better generalisation
- Small weight initialisation?


$w_1$
$w_2$
Loss
$w_2 \cdot w_1$
Epoch
Weight norm