Double descent and contrastive learning

European Research Council

Veronica Guidetti

(funded by: **CoSubmitting Summer @ ICLR 2022**)

Open problem in Deep Learning

Double descent and contrastive learning

European Research Council

Veronica Guidetti

(funded by: **CoSubmitting Summer @ ICLR 2022**)

Open problem in
Deep Learning

Double descent and contrastive learning

Study similarity
(Siamese NNs)

European
Research
Council

Veronica Guidetti

(funded by: **CoSubmitting Summer @ ICLR 2022**)

Why there's no Physics in here (yet)?

Our aim was to find symmetries using siamese NN

Wetzel et al. arXiv[2003.04299]

BIG NEWS: Working unsupervised constrastive learning

Chen et He arXiv[2011.10566]

UNSUPERVISED SYMMETRY DETECTION?!?

It's a long way to Tipperary….

Need to control generalization error and make training stable

Why there's no Physics in here (yet)?

Our aim was to find symmetries using siamese NN

Wetzel et al. arXiv[2003.04299]

BIG NEWS: Working unsupervised constrastive learning

Chen et He arXiv[2011.10566]

UNSUPERVISED SYMMETRY DETECTION?!?

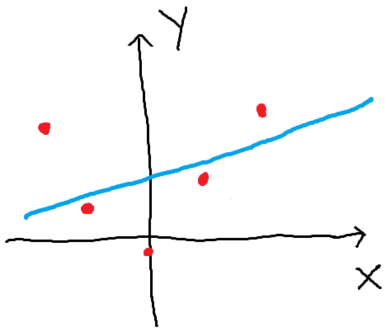
~~It's a long way to Tipperary….~~

It's a long way to go!

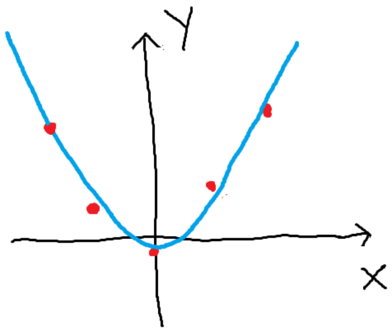
Don't mind if you do not know the song

Need to control generalization error and make training stable

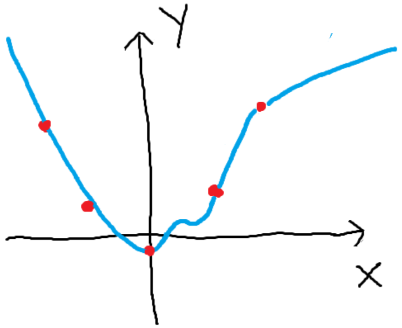
Double descent: How NNs generalise



underfitting

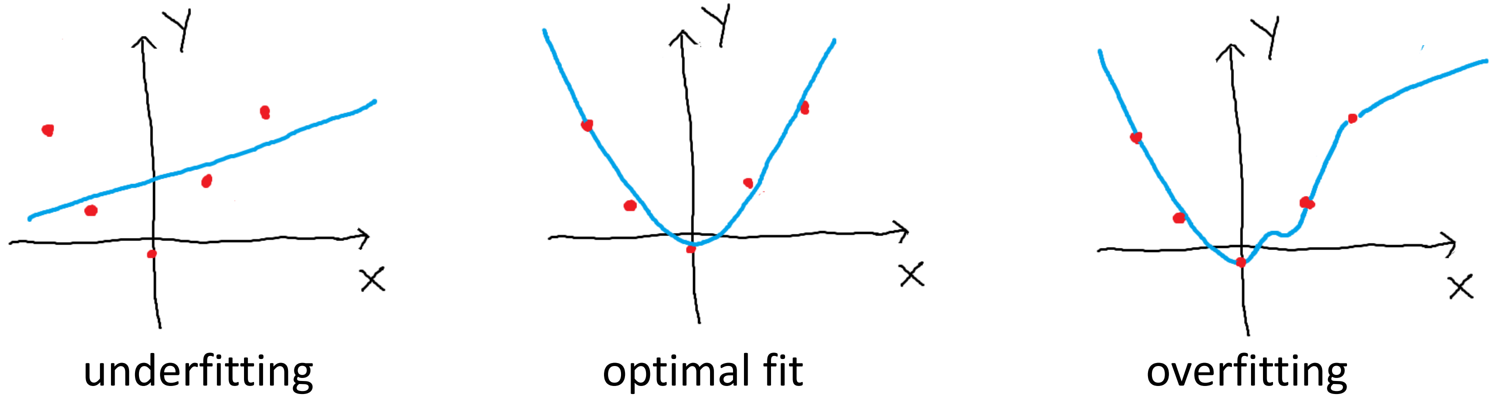


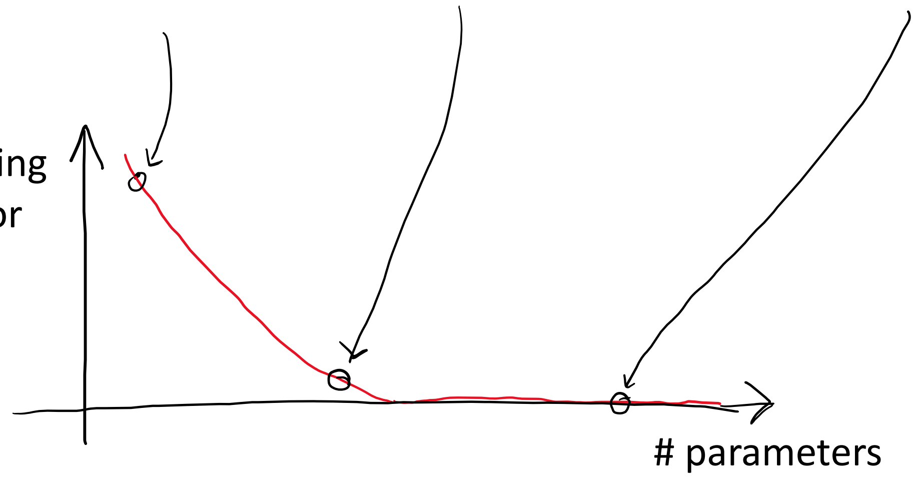
optimal fit



overfitting

Double descent: How NNs generalise

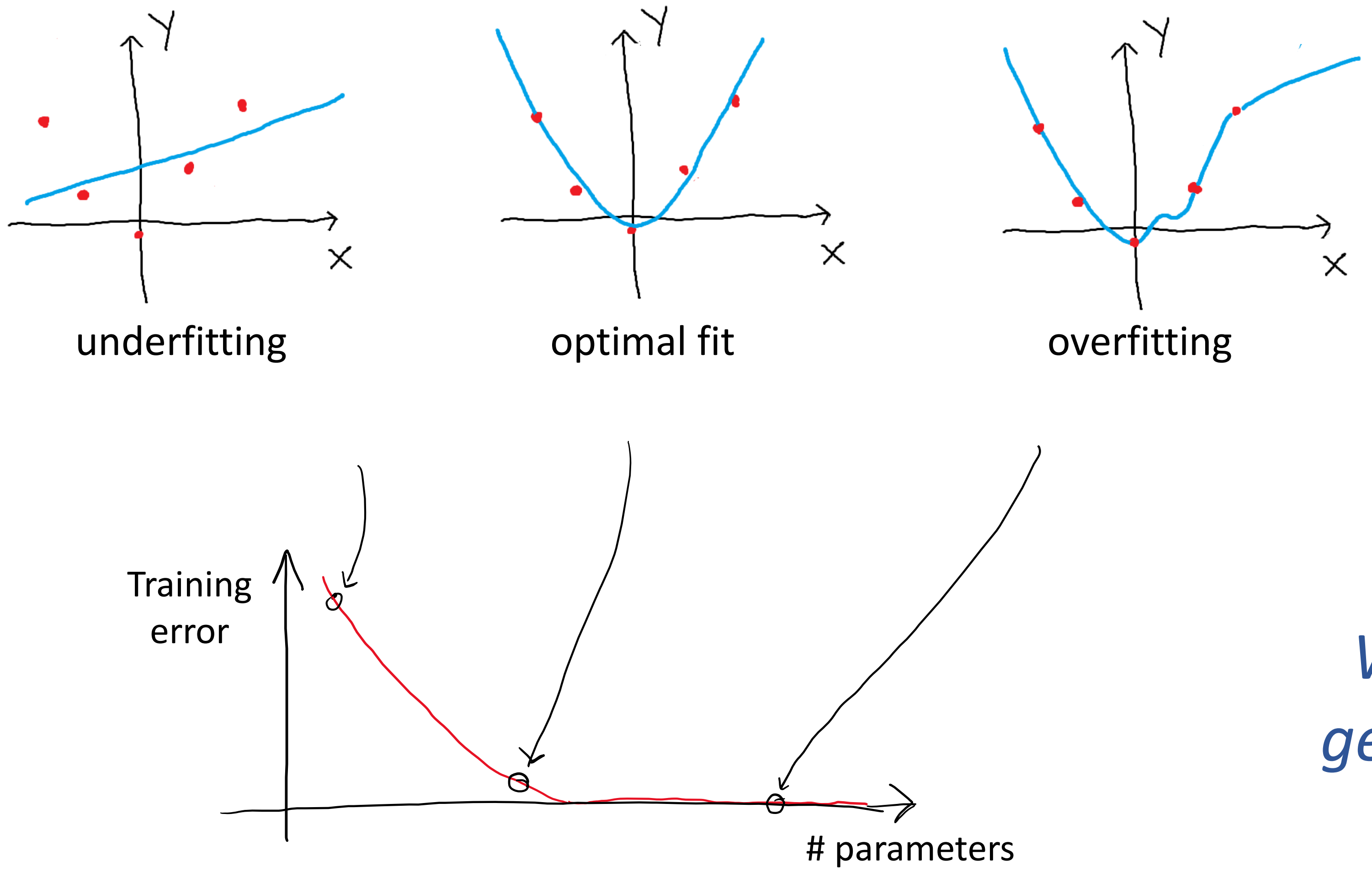




Error may be measured by

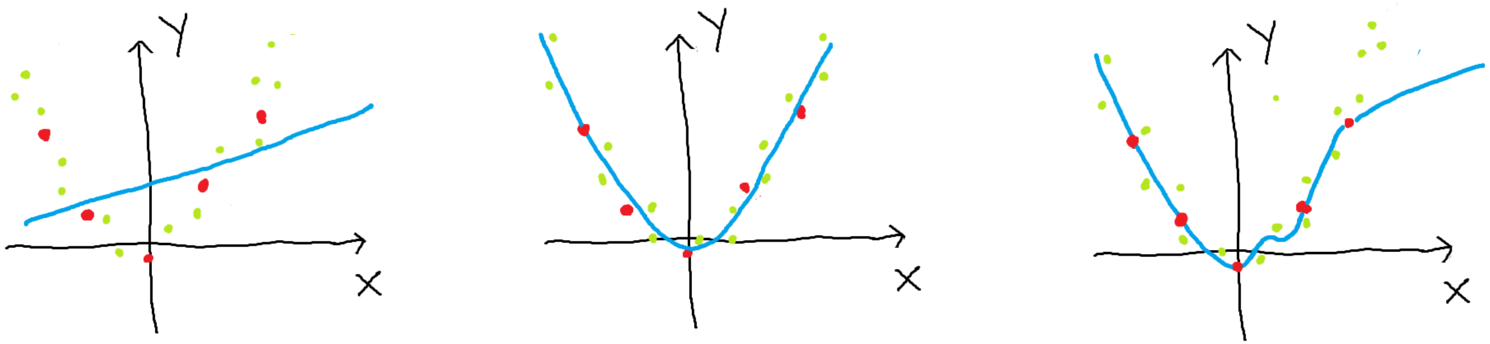
$$\mathrm{MSE}(y, f) = \frac{1}{N}\sum_{i}(y_i - \mathrm{f}(\mathrm{x_i}))^2$$

Double descent: How NNs generalise



What about generalisation error?

Double descent: How NNs generalise

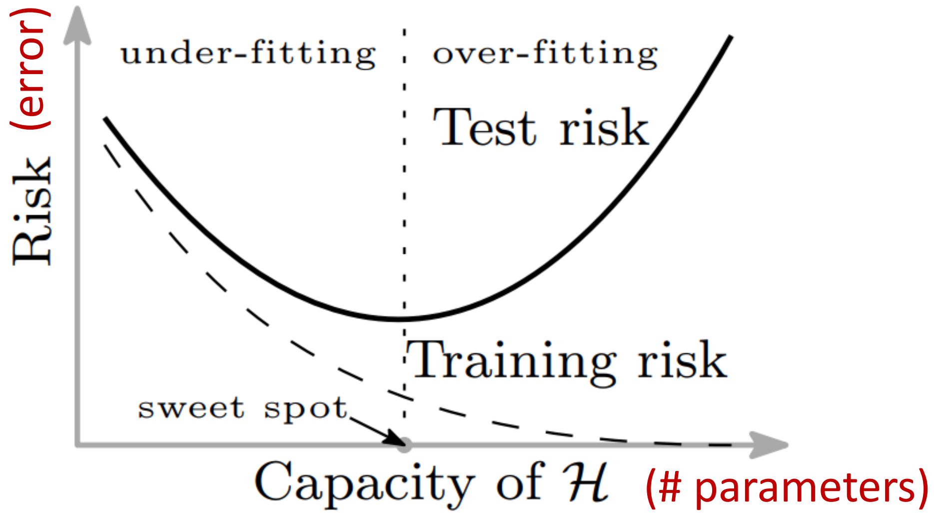


underfitting

optimal fit

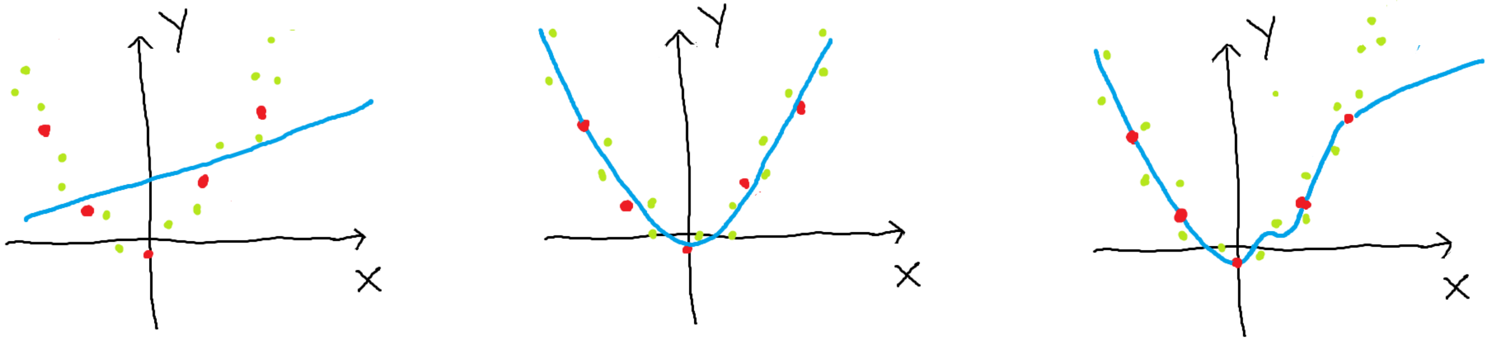
overfitting

Bias – variance tradeoff



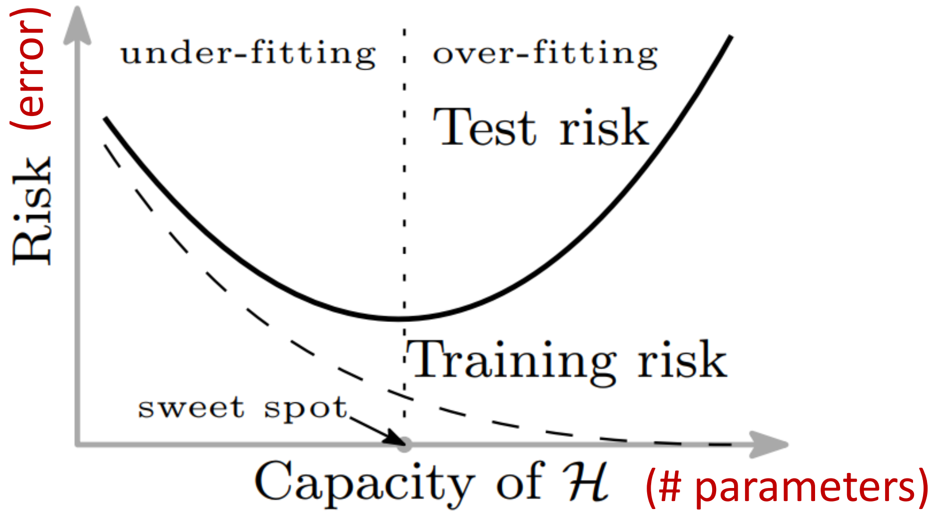
Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise



underfitting optimal fit overfitting

Bias – variance tradeoff

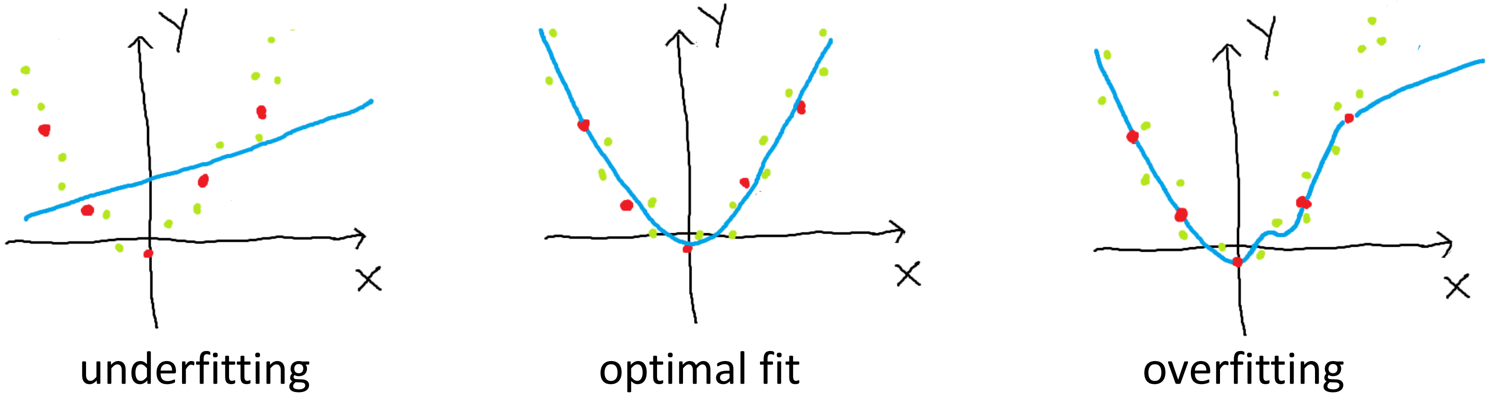


In ML courses we learn that overfitting:
- Learning training set features by heart
- Generalise worse

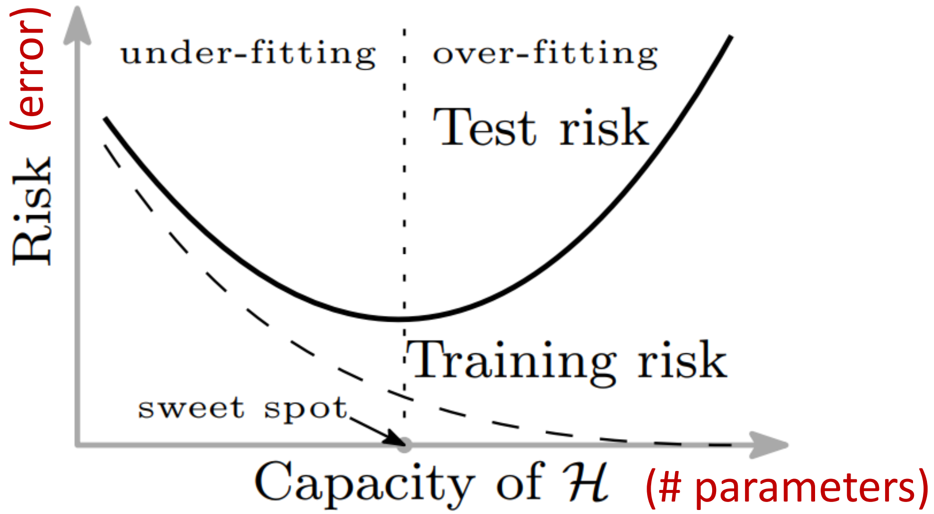
Techniques used not to overfit:
- Early stopping
- Regularisation
- …

Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise



Bias – variance tradeoff

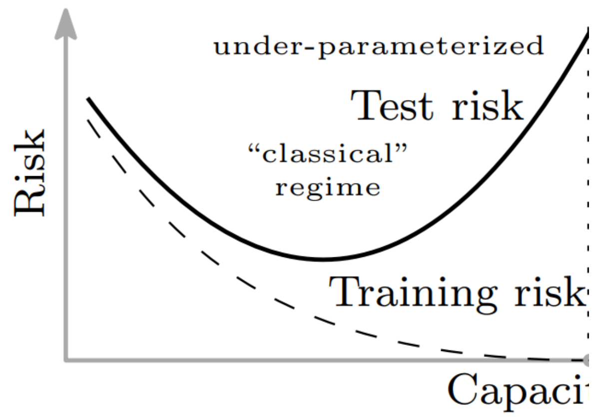


End of the story?

… not quite…

Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise



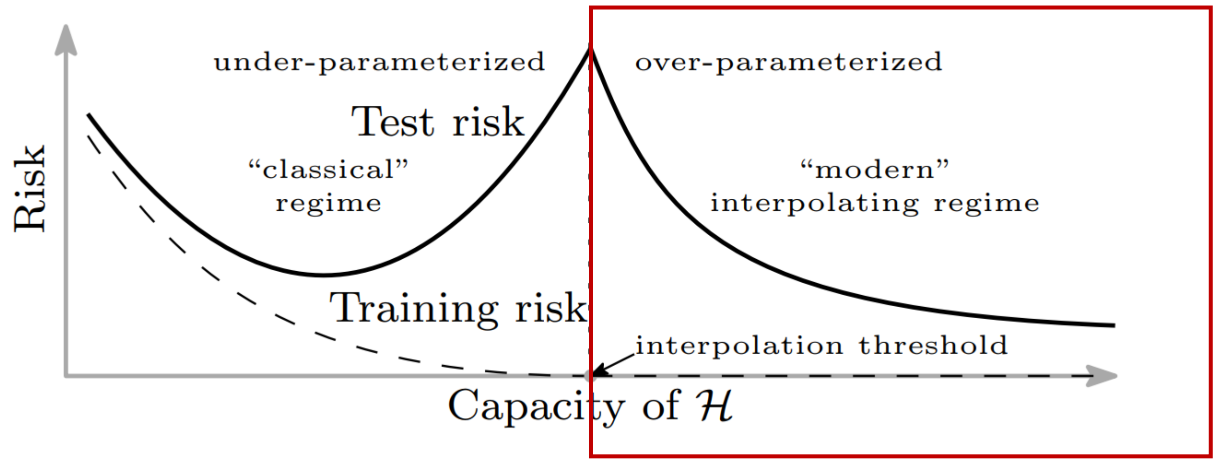
Test error descreses in the overparametrized region!

- Different behaviours (global minimum/peak hight/peak position)
- Partially explained in classification and regression tasks using Random Feature Models

Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise

Belkin et al. 2019
Nakkiran et al. 2019
Mei and Montanari 2019
Kini et al. 2020
D'Ascoli et al 2020
...



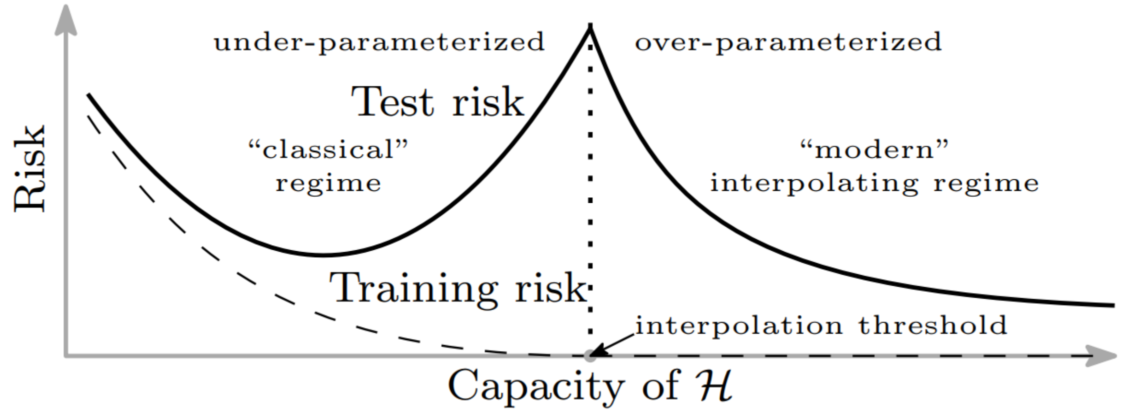
Test error descreses in the overparametrized region!

- Different behaviours (global minimum/peak hight/peak position)
- Partially explained in classification and regression tasks using Random Feature Models

Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise

Belkin et al. 2019
Nakkiran et al. 2019
Mei and Montanari 2019
Kini et al. 2020
D'Ascoli et al 2020
...



Test error descreses in the overparametrized region!

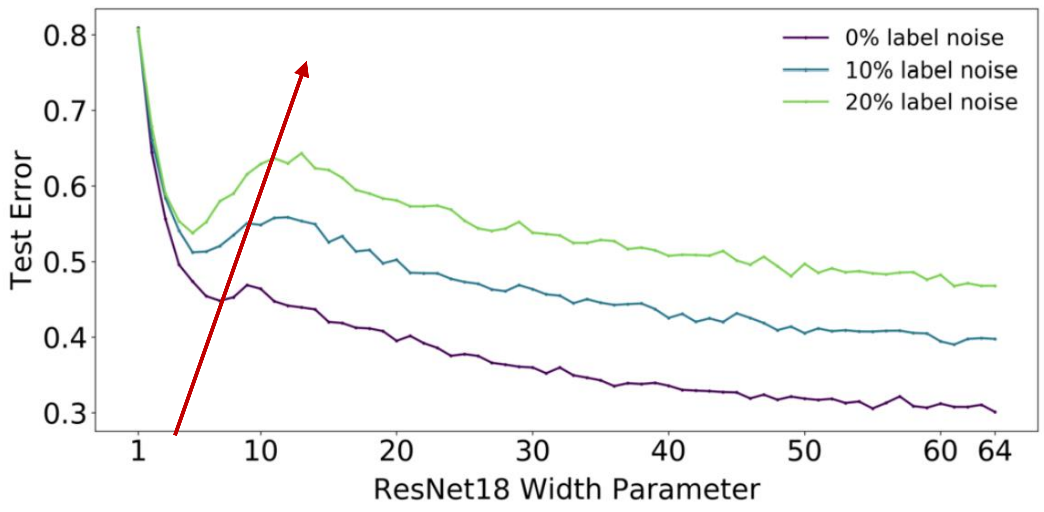
- Different behaviours (global minimum/peak hight/peak position)
- Partially explained in classification and regression tasks using Random Feature Models

Belkin et al. arXiv[1812.11118]

Double descent: How NNs generalise

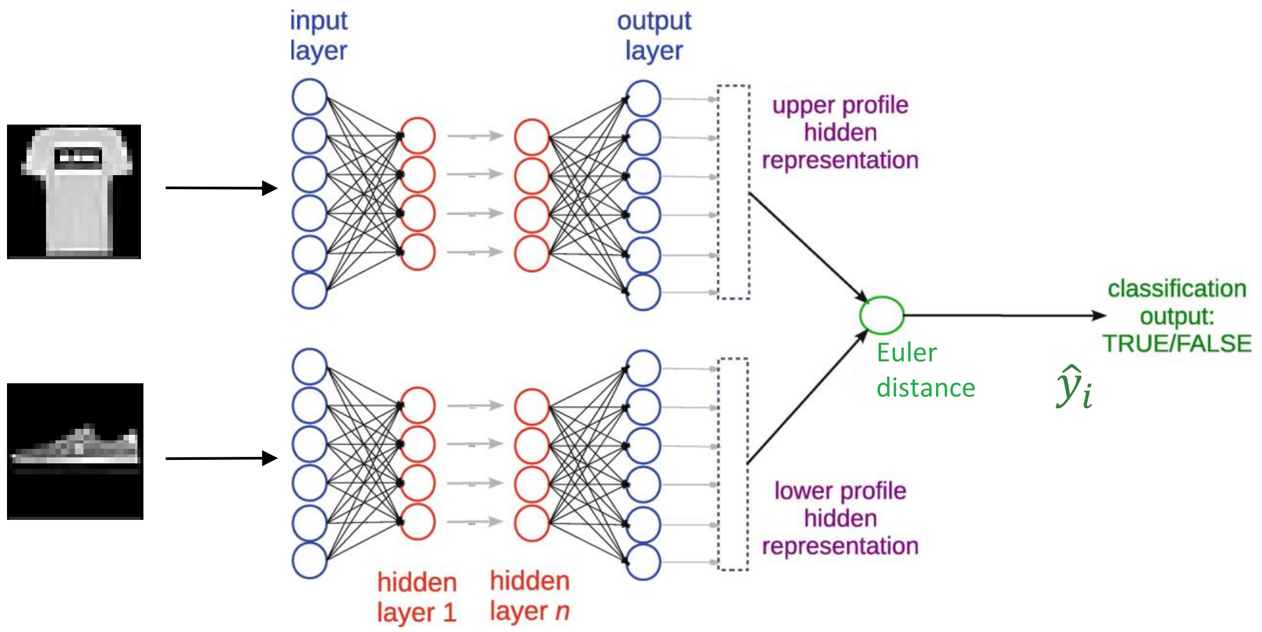
Higher peak in presence of noise

- Need more parameters to over fit data
- Spurious feature learning



Nakkiran et al. arXiv[1912.02292]

Double descent in Siamese NN



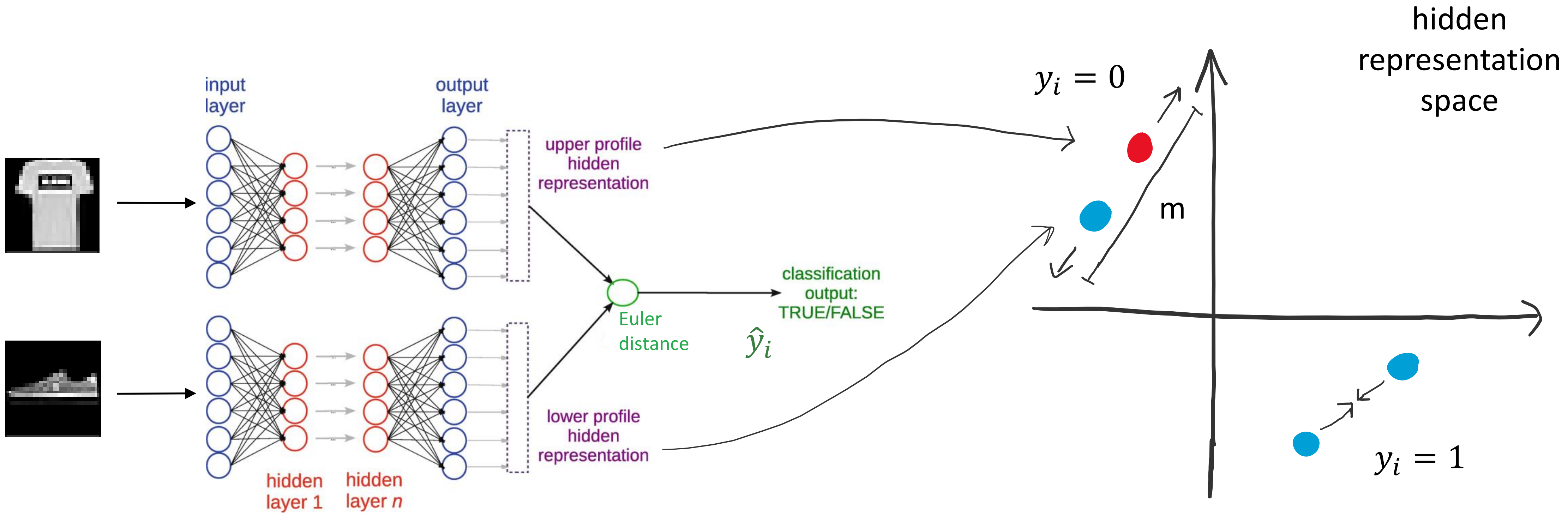
Contrastive Loss:

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N}\sum_i y_i \hat{y}_i^2 + (1 - y_i)\left[\max(0, m - \hat{y}_i)\right]^2$$

LeCun et al. 2006

Chicco - Siamese Neural Networks: An Overview - 2020

Double descent in Siamese NN



Contrastive Loss:

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N} \sum_i y_i \hat{y}_i^2 + (1 - y_i) \left[\max(0, m - \hat{y}_i) \right]^2$$

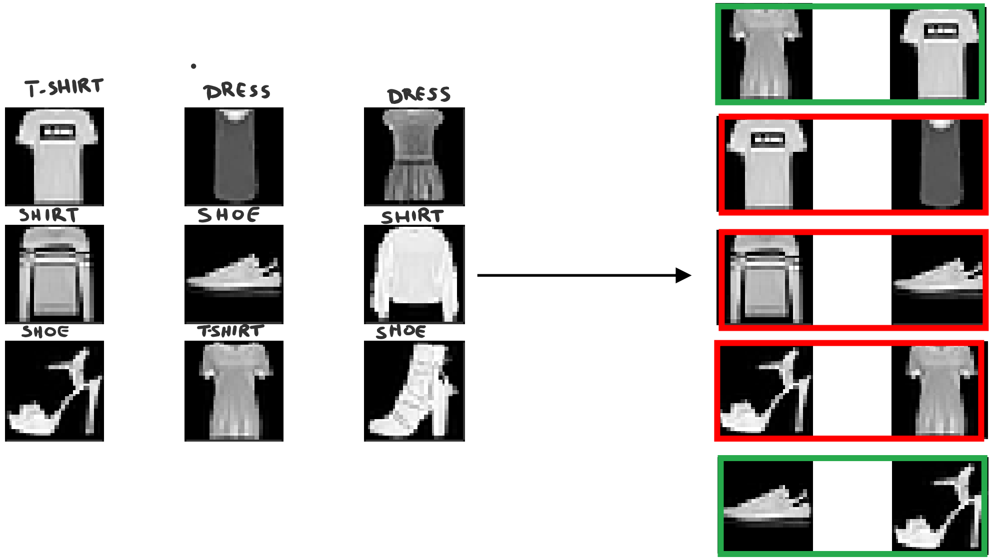
LeCun et al. 2006

Chicco - Siamese Neural Networks: An Overview - 2020

Double descent in Siamese NN

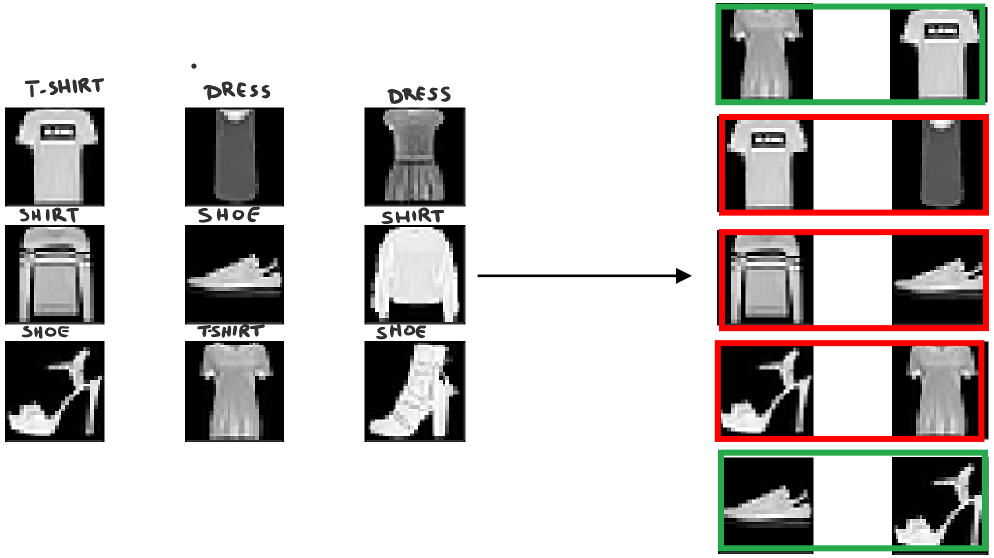
TWO NOISE SOURCES

Pair Label Noise (PLN)
symmetric stochastic error



$$T^P(p)\colon\ y_i^P \rightarrow Rnd\{0,1\}$$

Single Label Noise (SLN)
asymmetric «systematic» error

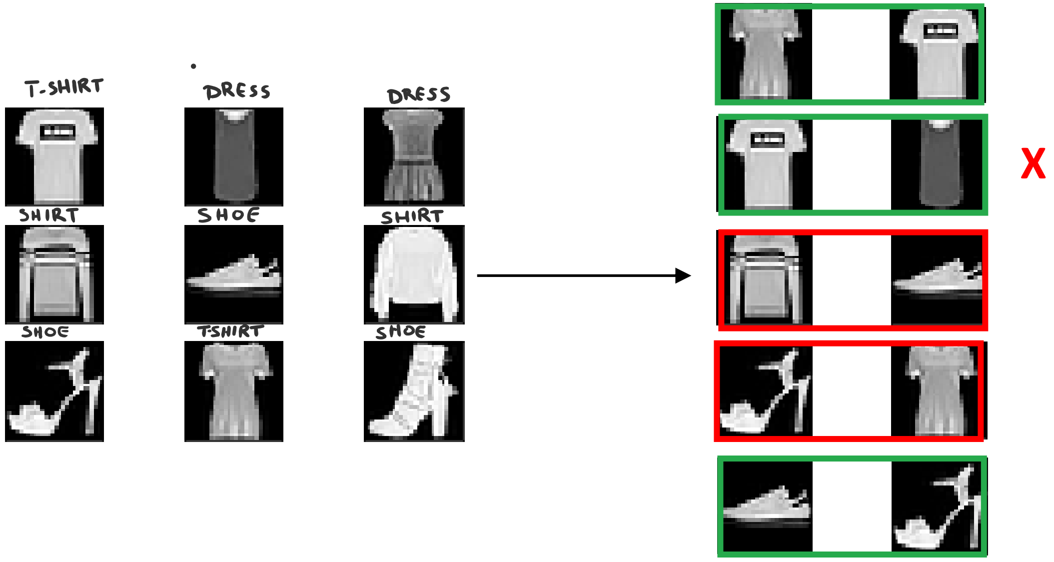


$$T(q)\colon\ y_i \rightarrow Rnd(0, n_c - 1)$$

Double descent in Siamese NN

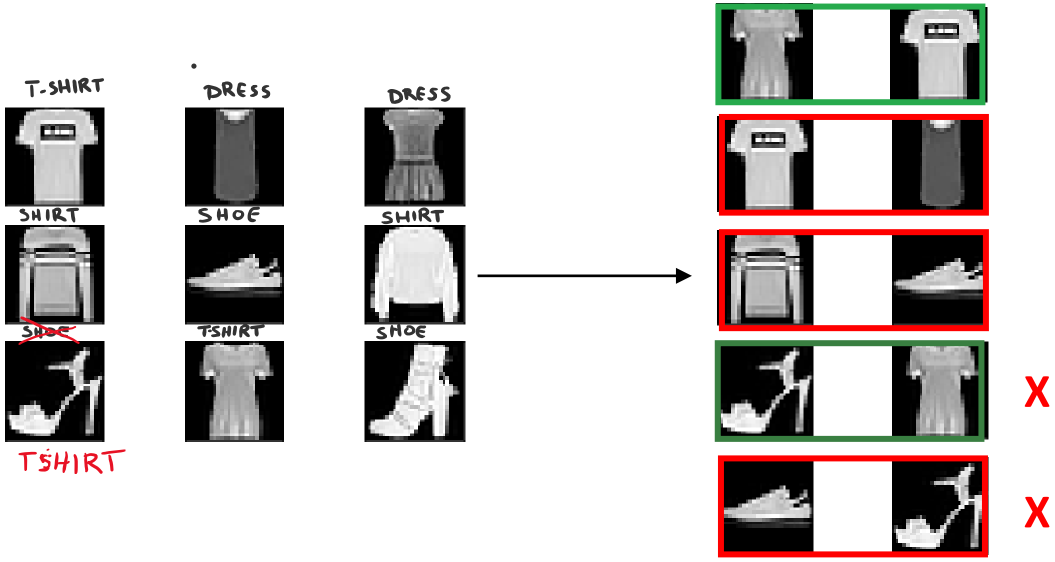
TWO NOISE SOURCES

Pair Label Noise (PLN)
symmetric stochastic error



$$T^P(p)\colon\ y_i^P \to Rnd\{0,1\}$$

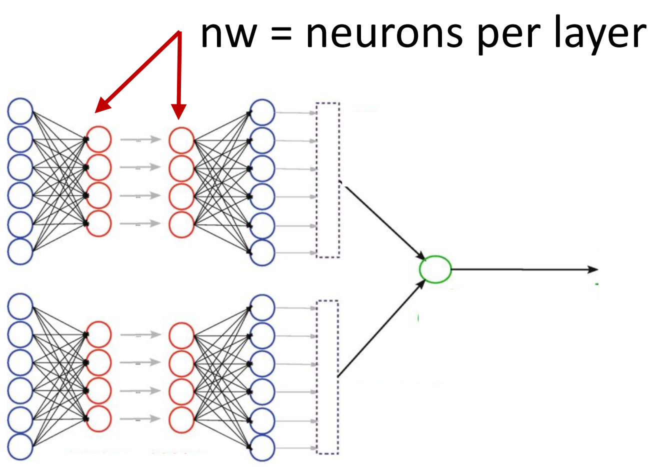
Single Label Noise (SLN)
asymmetric «systematic» error

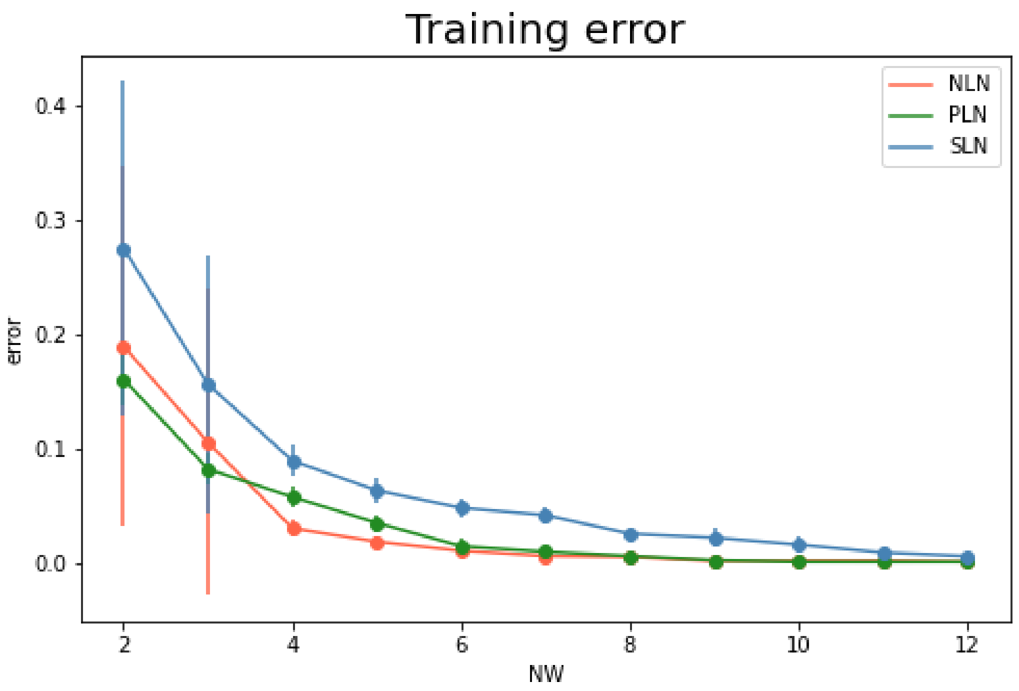


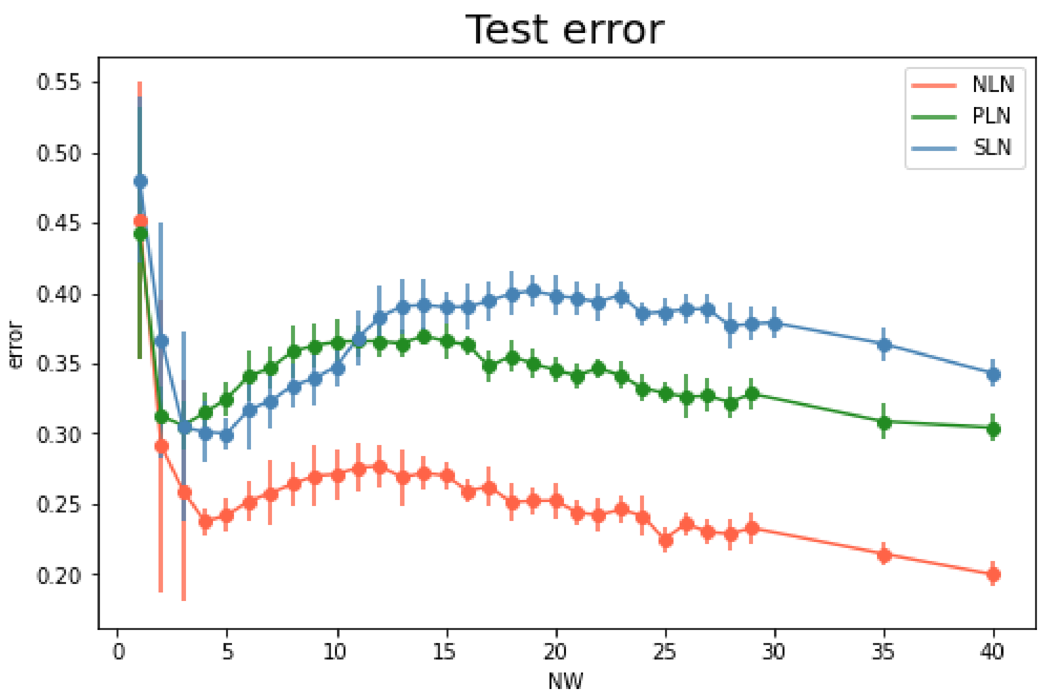
$$T(q)\colon\ y_i \to Rnd(0, n_c - 1)$$

Results:

- FMNIST: 6k training set, 10k test set
- 10% effective noise (both PLN and SLN)

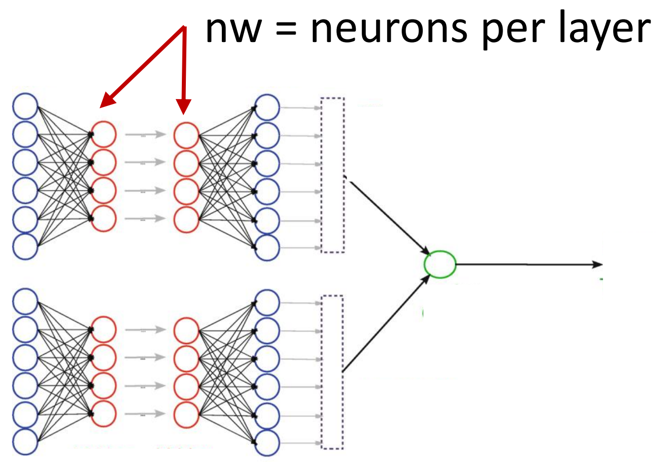


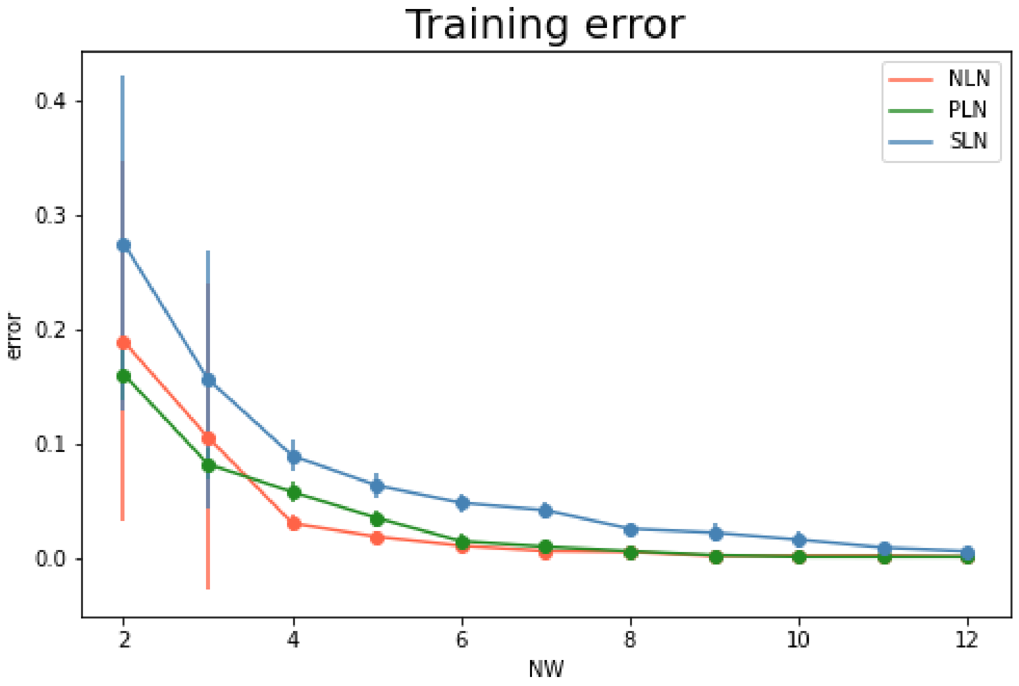


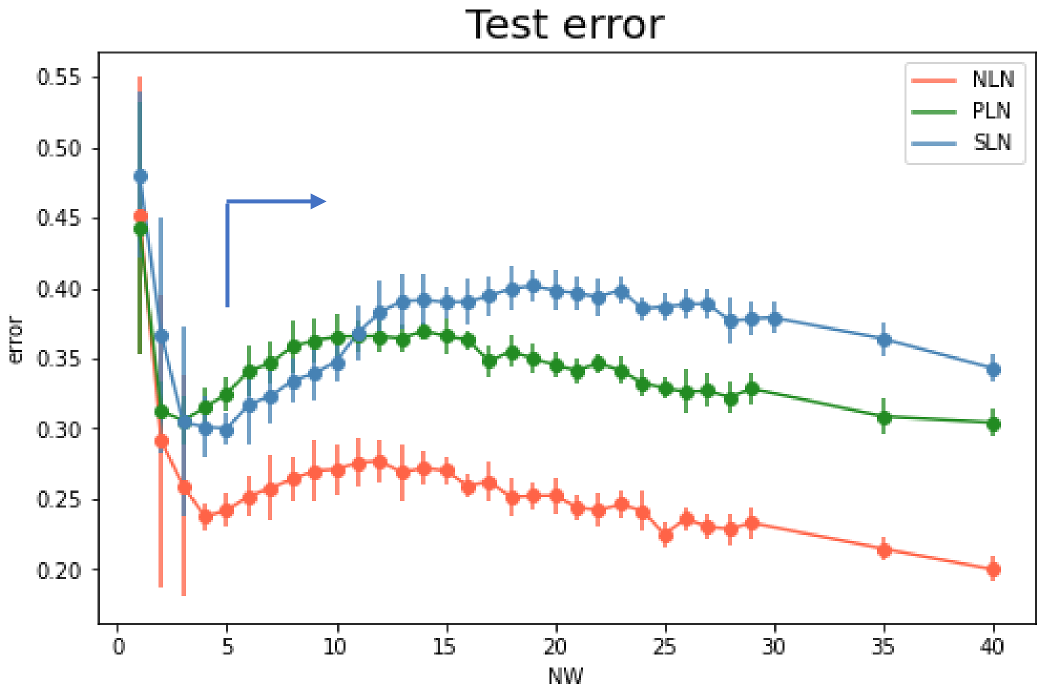


Results:

- FMNIST: 6k training set, 10k test set
- 10% effective noise (both PLN and SLN)







What’s next?

Provide a quantitative explanation about SLN/PLN difference:

- Earth Mover’s Distance
- Random Feature Models

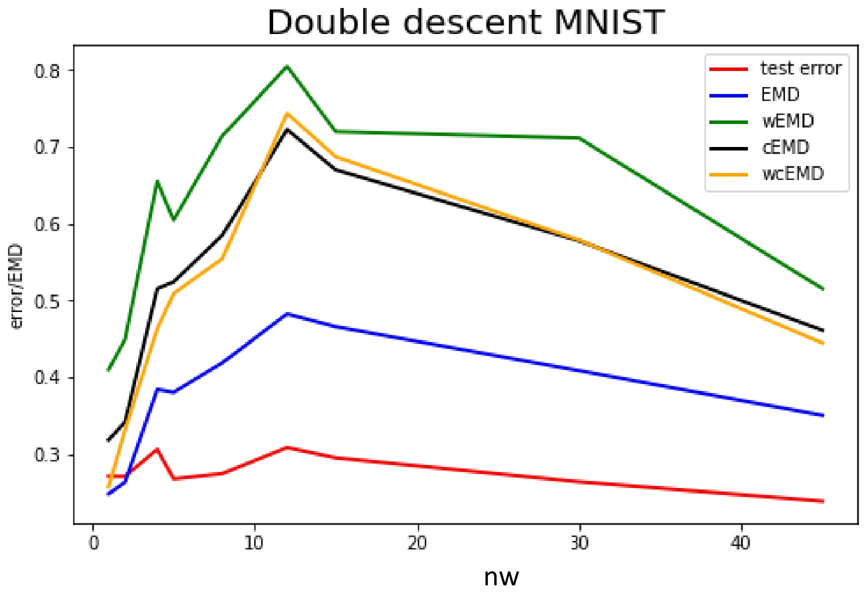
Analyse different architectures/datasets/losses

$\rightarrow$ Explain generalisation in contrastive learning

What's next?

Provide a quantitative explanation about SLN/PLN difference:

- Earth Mover's Distance
- Random Feature Models

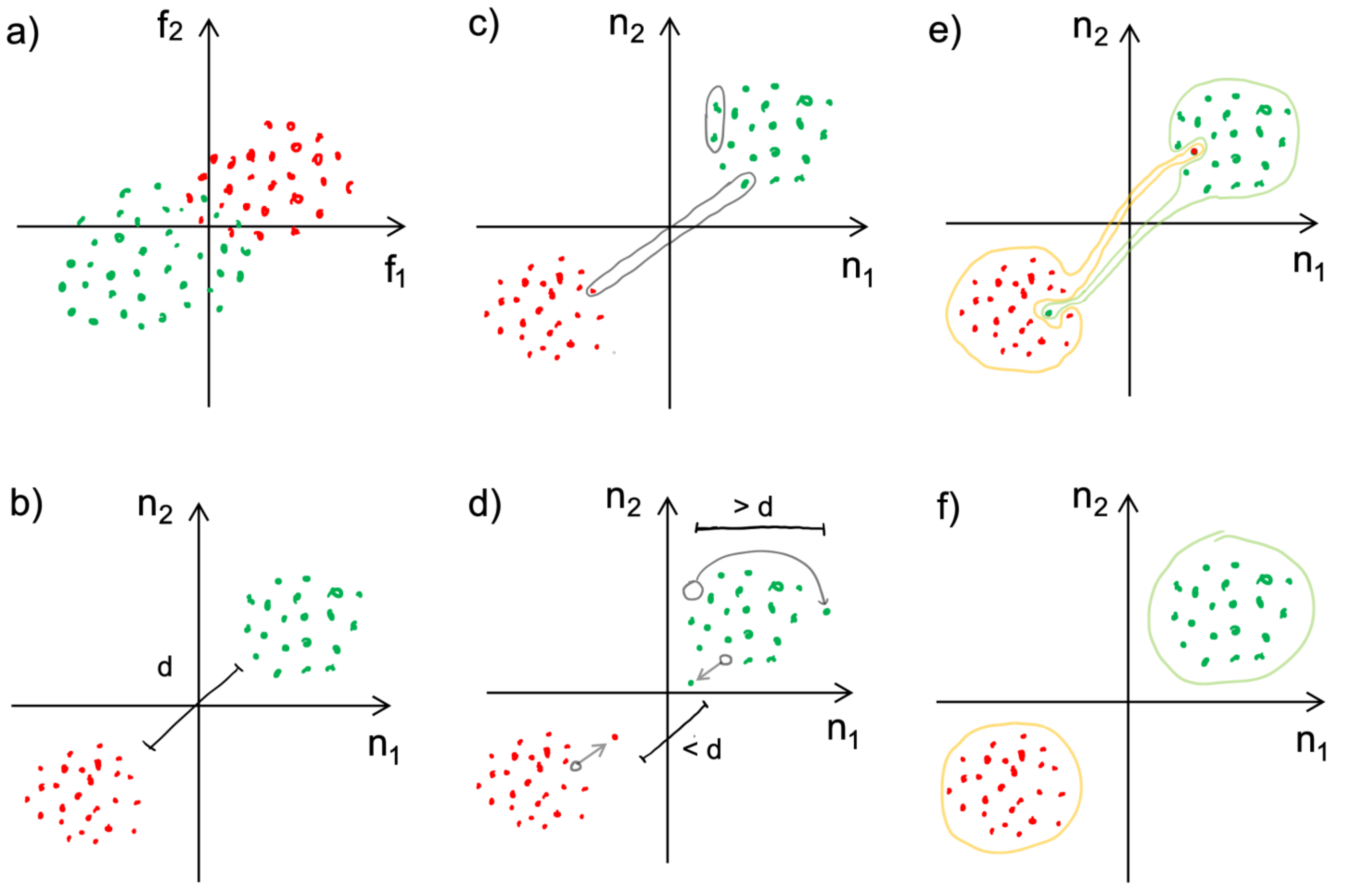


Analyse different architectures/datasets/losses

→ Explain generalisation in contrastive learning

Why do the these curves look different?

NO NOISE PLN SLN



Observations:

- Bottleneck layer: need to choose width
- Unstable training: half times you get the wrong result, NN size is crucial

Finding symmetries implies

- Finding conserved quantities
- Infinite classes: hard to fit
- SR extremly sensitive to bias error

Need to descrease generalization error and make training stable