How to win the lottery with a single ticket

Stefano Sarao Mannelli

Cargèse 2023

GATSBY
Sainsbury Wellcome Centre
UCL

What this talk is NOT about

A clever new system for increasing your probability of winning the lottery

What this talk is about

Curriculum learning: learning in a specific curated order. Can help unlock or at least speed up learning. *Animals **need** curricula!*

Training artificial neural networks: in the standard setup, the network learns how to give the correct output on a dataset of examples by optimizing a loss (error) function where each example is given the same relevance (random order).

Over-parameterization and the lottery ticket hypothesis [Frankle, Carbin 2018]**:** not all the parameters are actually needed, but enlarging the search space allows a higher chance of starting “close” to a good generalization solution.

Our question: why is curriculum mostly ineffective in deep neural networks?

Luca Saglietti

Bocconi

Andrew Saxe

UCL

First ingredient: Curriculum learning

Curriculum learning

Why is it important?

In the long-term goal of a generalized theory of learning, it is a missing piece (one of many). *Remarkable achievements in our life are obtained using curricula: animals need curricula!*

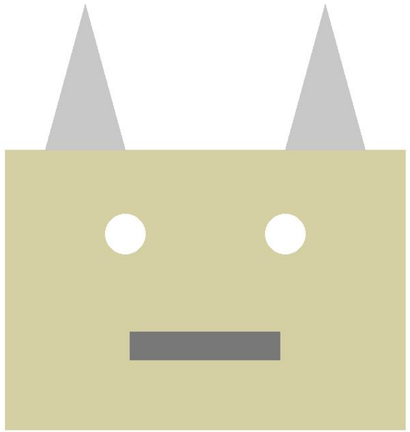
Curriculum learning in animals

Animals:
conditional reflexes (dogs) [Pavlov 1927];
shaping (rats, pigeons) [Skinner 1938];
discrimination along a continuum (rats) [Lawrence 1952];
cross-species auditory identification (rats, humans) [Liu, et. al 2008].

Curriculum learning in animals

Animals:
conditional reflexes (dogs) [Pavlov 1927];
shaping (rats, pigeons) [Skinner 1938];
discrimination along a continuum (rats) [Lawrence 1952];
cross-species auditory identification (rats, humans) [Liu, et. al 2008].

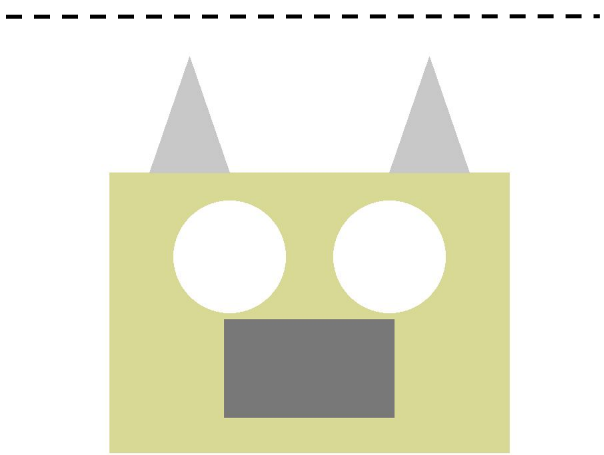
Humans:
discrimination along a continuum [Baker, Stanley 1954];
past tense [Plunkett et al. 1990; 1991];
fading with auditory and visual stimuli [Pashler, Mozer 2013];
eureka effect [Ahissar, Hochstein 1997];



Curriculum learning in animals

Animals:
conditional reflexes (dogs) [Pavlov 1927];
shaping (rats, pigeons) [Skinner 1938];
discrimination along a continuum (rats) [Lawrence 1952];
cross-species auditory identification (rats, humans) [Liu, et. al 2008].

Humans:
discrimination along a continuum [Baker, Stanley 1954];
past tense [Plunkett et al. 1990; 1991];
fading with auditory and visual stimuli [Pashler, Mozer 2013];
eureka effect [Ahissar, Hochstein 1997];

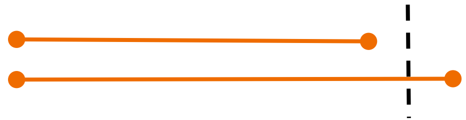


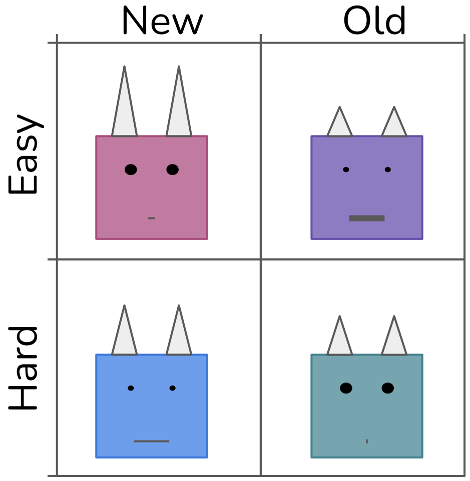
new world demon

Curriculum learning in animals

Animals:
conditional reflexes (dogs) [Pavlov 1927];
shaping (rats, pigeons) [Skinner 1938];
discrimination along a continuum (rats) [Lawrence 1952];
cross-species auditory identification (rats, humans) [Liu, et. al 2008].

Humans:
discrimination along a continuum [Baker, Stanley 1954];
past tense [Plunkett et al. 1990; 1991];
fading with auditory and visual stimuli [Pashler, Mozer 2013];
eureka effect [Ahissar, Hochstein 1997];



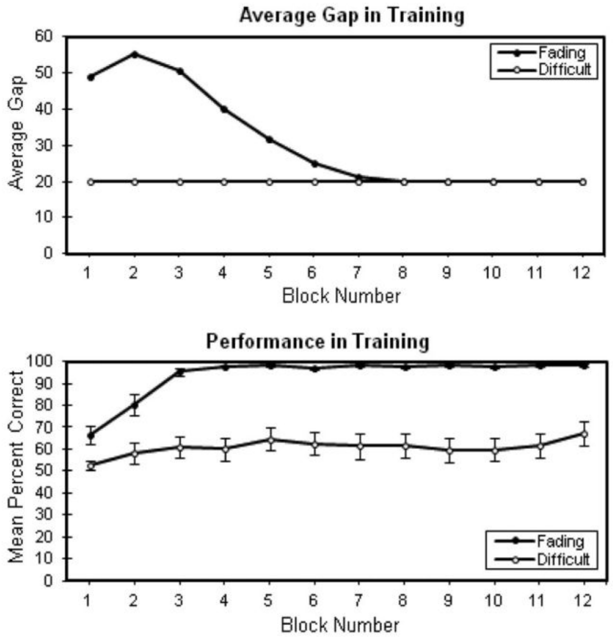


[Pashler, Mozer 2013]

Curriculum learning in animals

Animals:
conditional reflexes (dogs) [Pavlov 1927];
shaping (rats, pigeons) [Skinner 1938];
discrimination along a continuum (rats) [Lawrence 1952];
cross-species auditory identification (rats, humans) [Liu, et. al 2008].

Humans:
discrimination along a continuum [Baker, Stanley 1954];
past tense [Plunkett et al. 1990; 1991];
fading with auditory and visual stimuli [Pashler, Mozer 2013];
eureka effect [Ahissar, Hochstein 1997];

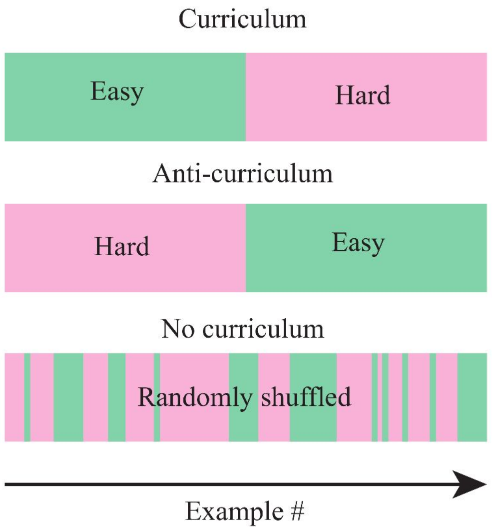


Fading effect

Curriculum learning in machine learning

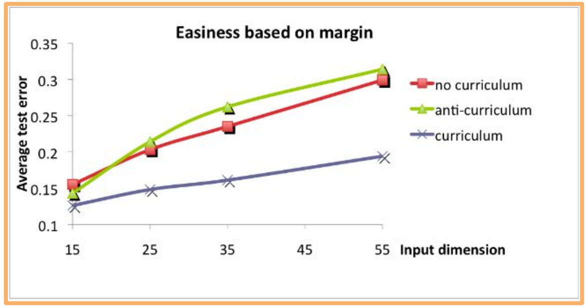
- Curriculum learning [Bengio et al. 2009]: empirical evidence of beneficial curriculum learning

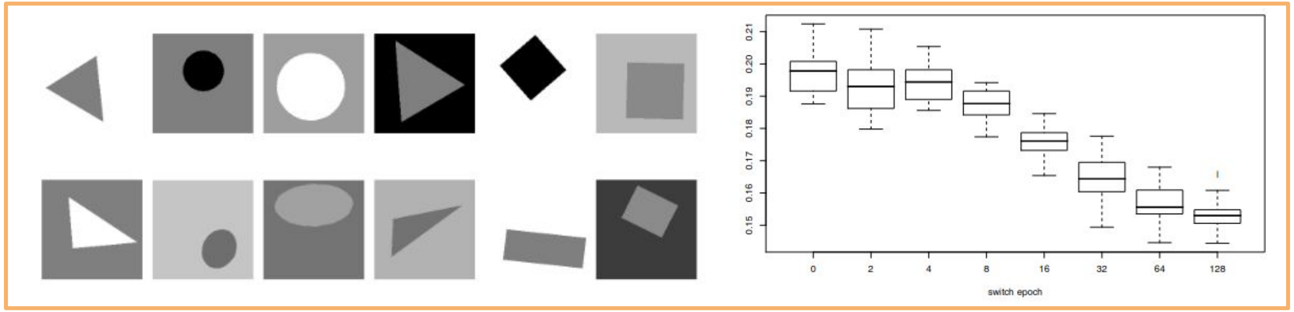
Instead of presenting the learning samples in random order, one can show them in increasing/decreasing order of difficulty!

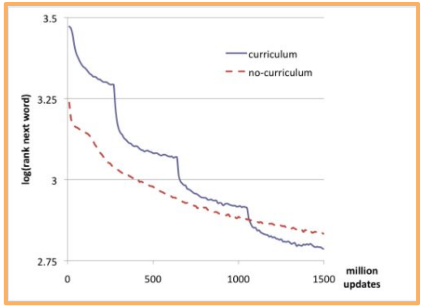


Curriculum learning in machine learning

- Curriculum learning [Bengio et al. 2009]: empirical evidence of beneficial curriculum learning

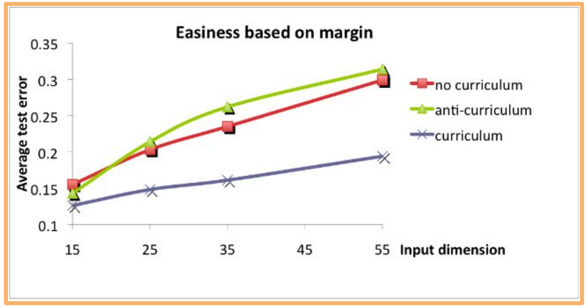


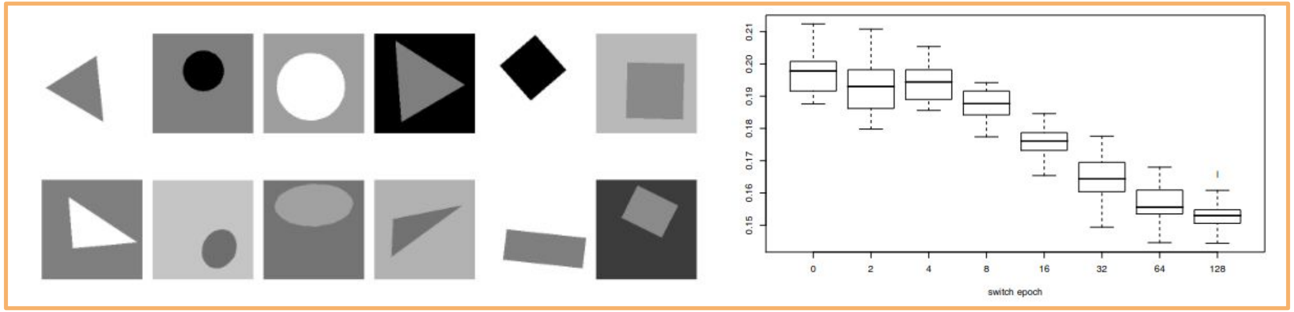


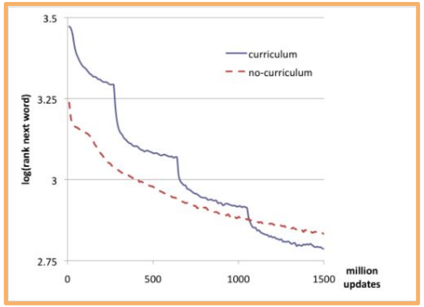


Curriculum learning in machine learning

- Curriculum learning [Bengio et al. 2009]: empirical evidence of beneficial curriculum learning
- However, we there are also recommendations for anti-curriculum strategies [Zhang et al. 2019; Hacohen & Weinshall 2019]
- Recent works argue that there is no effect at all in standard vision benchmarks [Wu et al. 2020]
- Only convincing results are for language models and RL!







Lots of work to be done
(very little theory on this!)

Lots of work to be done
(very little theory on this!)

[Weinshall 2018-19,
Kepple et al. 2022,
Cornacchia et al. 2023,
Abel et al. 2023]

arXiv:...8068v1 [cs.LG] 15 Jun 2021

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

Luca Saglietti[†,*], Stefano Sarao Mannelli[‡,*], and Andrew Saxe[‡,§]

Abstract

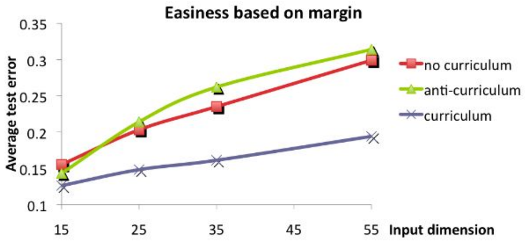
In animals and humans, curriculum learning—presenting data in a curated order—

An analytical theory of curriculum learning pt.1

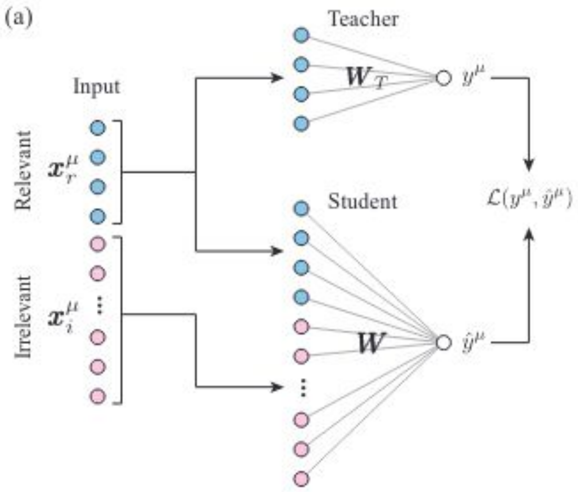
$\boldsymbol{x}_i \in \mathbb{R}^{(1-\rho)N}$, $\boldsymbol{x}_r \in \mathbb{R}^{\rho N}$, $\boldsymbol{x} = (\boldsymbol{x}_r, \boldsymbol{x}_i)$ w/ i.i.d. $x_{i,k} \sim \mathcal{N}(0, \Delta)$, $x_{r,k} \sim \mathcal{N}(0, 1)$

$y = \mathrm{sign}(\boldsymbol{x}_r \cdot \boldsymbol{W}_T / \sqrt{N})$ w/ i.i.d. $W_{T,k} \sim \mathcal{N}(0, 1)$

$\hat{y} = \mathrm{sign}(\boldsymbol{x} \cdot \boldsymbol{W} / \sqrt{N})$



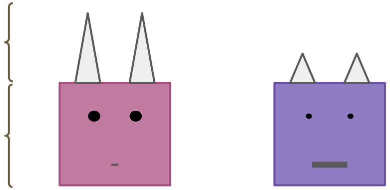
ρ : fraction of relevant features
Δ : variance of the irrelevant inputs
αN = number of samples shown
N→∞



The model allows for the exact analysis of the oSGD dynamics and the asymptotic performance of batch learning.

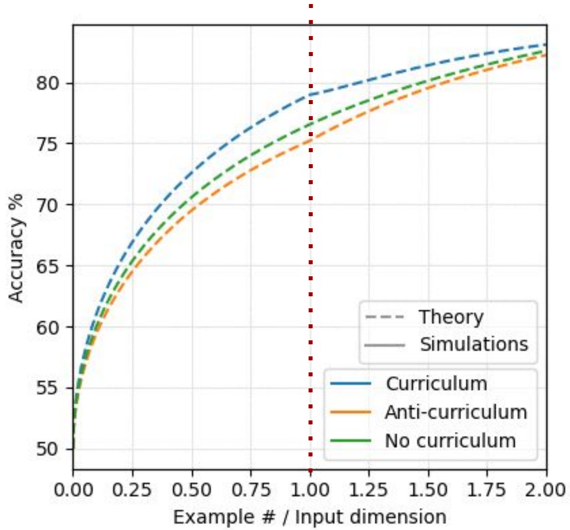
relevant

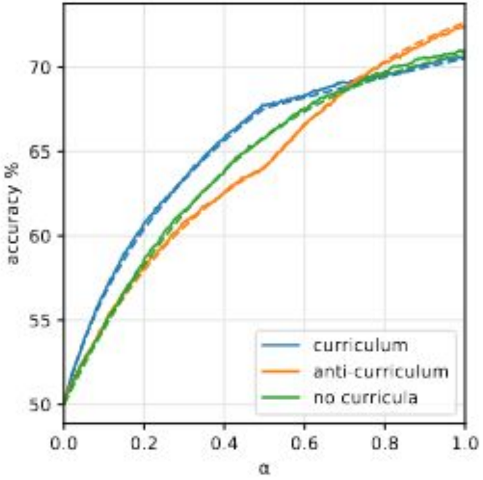
irrelevant

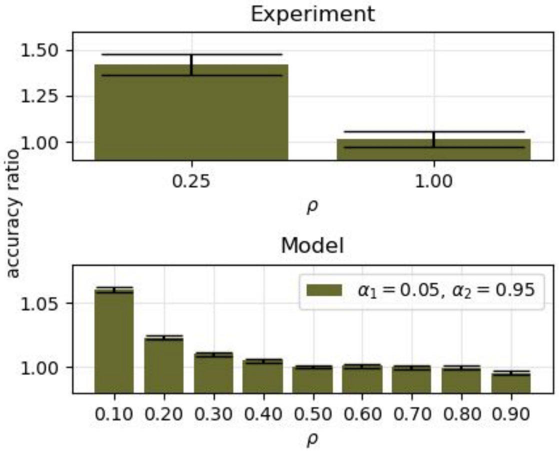


[Bengio et al. 2009, Pashler, Mozer 2013]

An analytical theory of curriculum learning pt.1







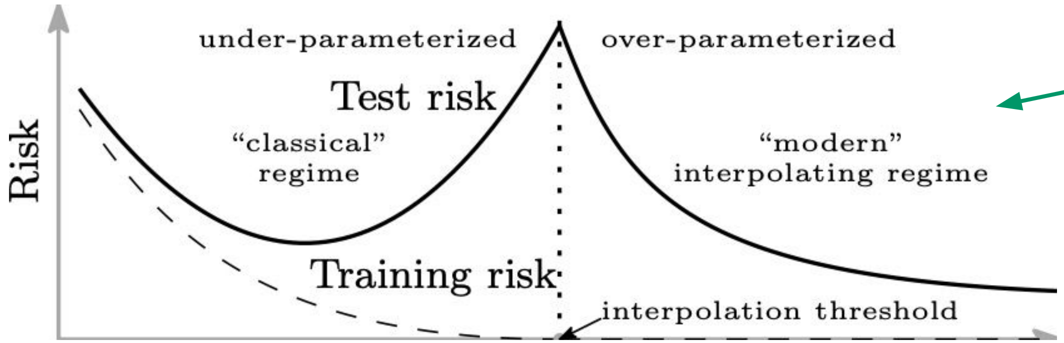
Speed up but little improvement in generalisation

Moving away from optimality can lead to ineffective curricula

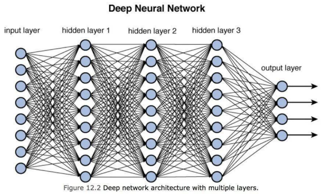
Curriculum needs relevant feature concealed in a complex input

NN learning and the effect of over-parametrization

Overparameterisation

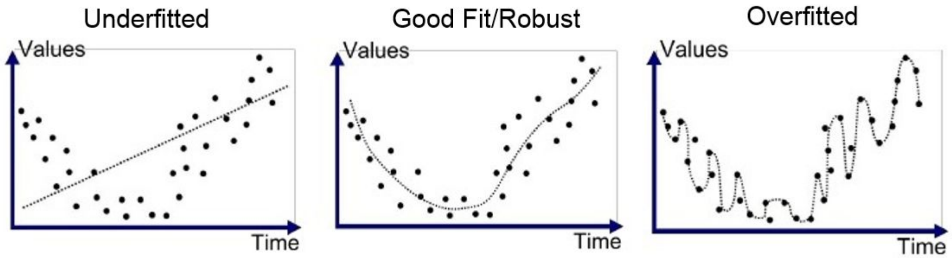


Hic sunt DNNs

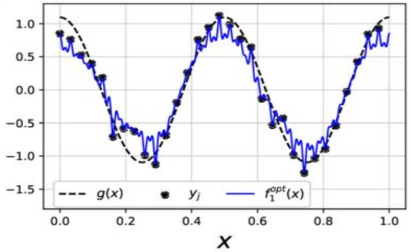


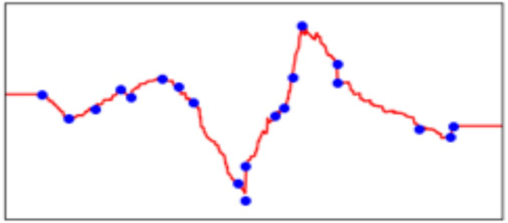
[Belkin, et al. 2019]

UNDER-PARAMETRIZED
Example in a 1d regression:

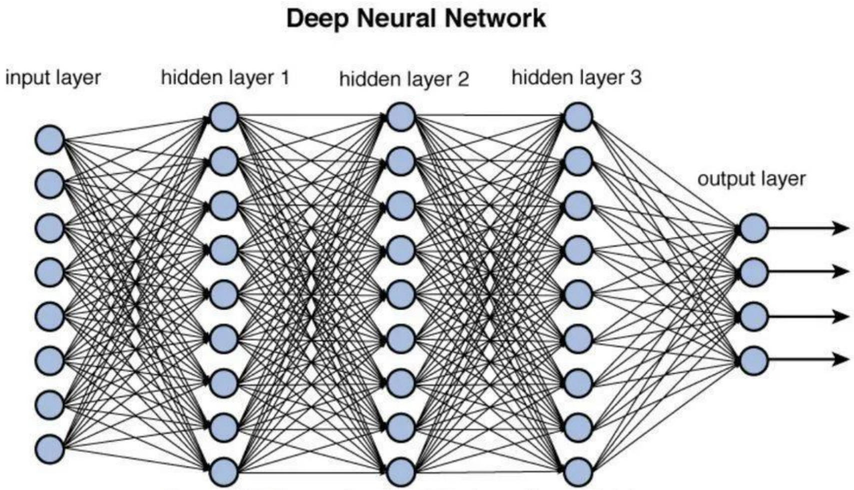


OVER-PARAMETRIZED
Example in a 1d regression:





Training a deep neural network



Universal approximation theorem [Hornik, et al. 1989;1993]:

> DNN function class can **approximate any "well behaving" function** provided that is "large" enough.

#parameters > 10^6

Train via Stochastic Gradient Descent (**SGD**) on:

i.i.d. assumption

$$R_{\text{learn}} = \underbrace{\overbrace{\arg\min}^{\downarrow}}_{R_\theta, \theta \in \Theta} \boxed{\sum_{n=1}^{N}} f\,(\,\overset{\downarrow}{x_n}\,, R_\theta(\,\overset{\downarrow}{y_n}\,)) + g\,(\theta)$$

A. Training Set

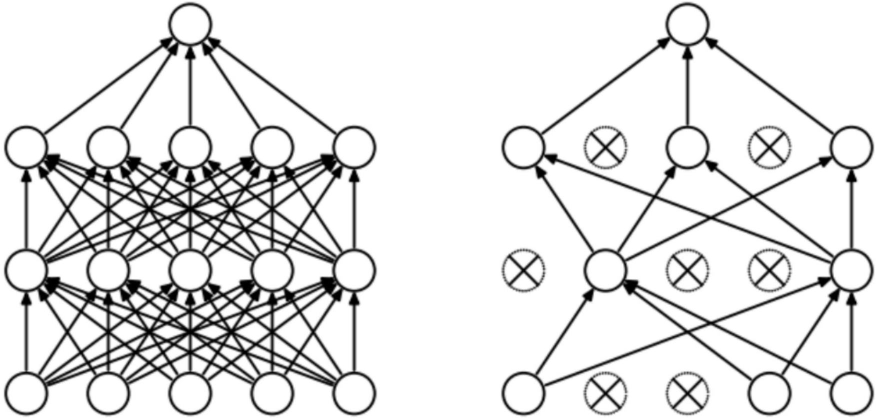
B. Network Architecture

C. Cost Function and Regularization

Is the NN actually using all these parameters?

NO! The lottery ticket hypothesis [Frankle, Carbin 2018]:

- **Most parameters** are close to zero and **could be** completely **dropped** without significant change in the performance
- A **sub-network** at initialization is by chance **close to a good configuration**. This is our **winning lottery ticket**.
- If you only take the topology of the good sub-net but start from a bad initialization you will never find the good solution!
- **Over-parametrization** = buying **many lottery tickets**! Strength in numbers!

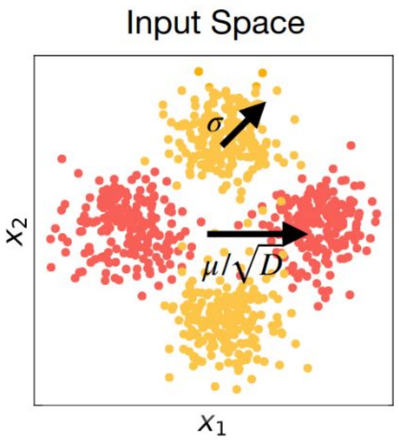


What is the interplay of curriculum learning and over-parametrization?

Can you win the lottery with few or just one ticket?

Simple synthetic model of data

[Refinetti, et al. 2021] model for feature learning vs lazy learning: XOR-like Gaussian mixture



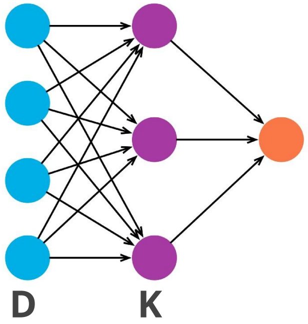
1. Sample cluster: $c \sim \mathrm{Unif}(\{1,2,3,4\})$
2. Sample data point: $\boldsymbol{x} \in \mathbb{R}^D, \quad \boldsymbol{x}|c \sim \mathcal{N}(\boldsymbol{\mu}_c/\sqrt{D}, \sigma^2)$

"**Hidden**" in high-dimension (D >> 2). But only **two dimensions** are truly **relevant** for learning. **Low SNR**.

Non-separable task. A linear classifier will fail!

Learning model

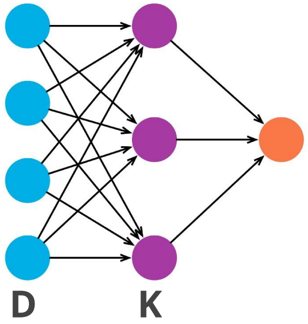
1 hidden layer neural network



- **D** inputs
- **K** hidden units (neurons)
- **1** output
- **(D x K) + K** trainable parameters (***w*** and ***v***)
- Non-linear activation (GeLU, ReLU, ...)

Learning model

1 hidden layer neural network



- **D** inputs
- **K** hidden units (neurons)
- **1** output
- **(D x K) + K** trainable parameters (***w*** and ***v***)
- Non-linear activation (GeLU, ReLU, ...)

Trained through **online Stochastic Gradient Descent** on square error.

$$\mathrm{d}w_i^k = -\frac{\eta}{\sqrt{D}} v^k \Delta g'(\lambda^k) x_i - \frac{\eta}{\sqrt{D}} \kappa w_i^k, \qquad \lambda^k \equiv \frac{1}{\sqrt{D}} \sum_{r=1}^{D} w_r^k x_r$$

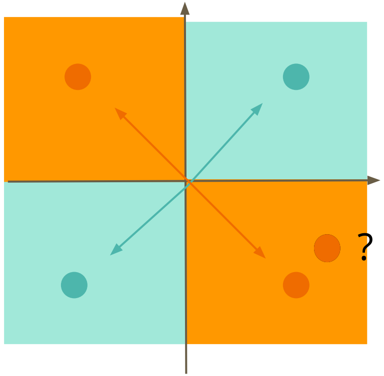
$$\mathrm{d}v^k = -\frac{\eta}{D} g(\lambda^k) \Delta - \frac{\eta}{D} \kappa v^k, \qquad \Delta = \sum_{j=1}^{K} v^j g(\lambda^j) - y$$

Over-parametrizing the model

K = 4 **K > 4** **K >> 4**

Minimally parametrized → **Over-parametrized**

I can get good generalization if each cluster has at least 1 neuron **“specialized”** (centered) on it. **K** needs to be **at least 4**.

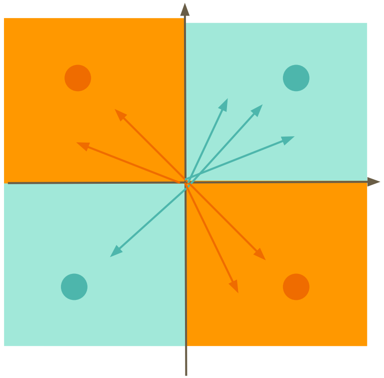


Over-parametrizing the model

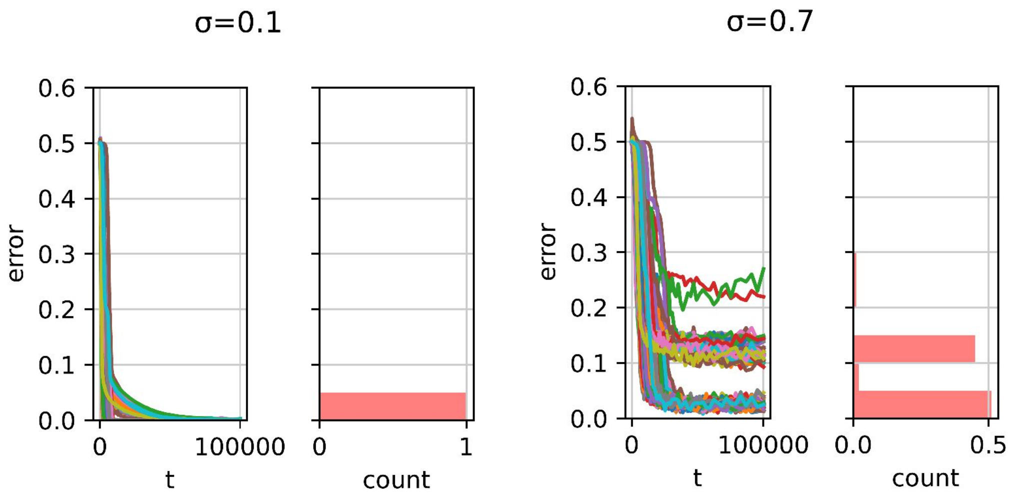
K = 4 K > 4 K >> 4

Minimally parametrized Over-parametrized

With more and more neurons the likelihood of having **at least 1 neuron per cluster** increases greatly!



Presence of sub-optimal local minima

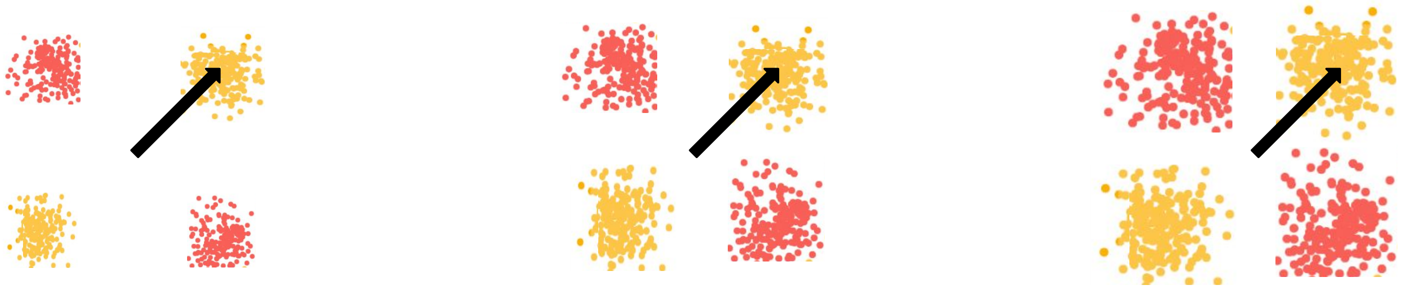


Because of the **non-convexity** of the loss, the **network can get stuck** in these minima. Especially when the **SNR is low**.

Curriculum learning protocols

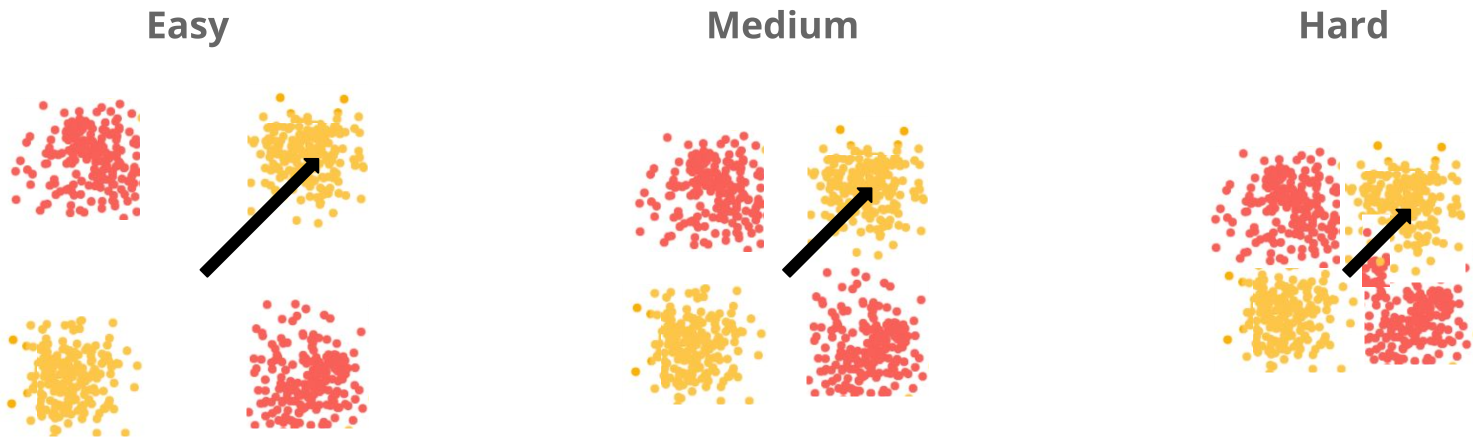
Slowly increase noise: vary the SNR by **reducing the variance** of the Gaussian clouds -> The clouds become more **well separeted**

Easy   Medium   Hard



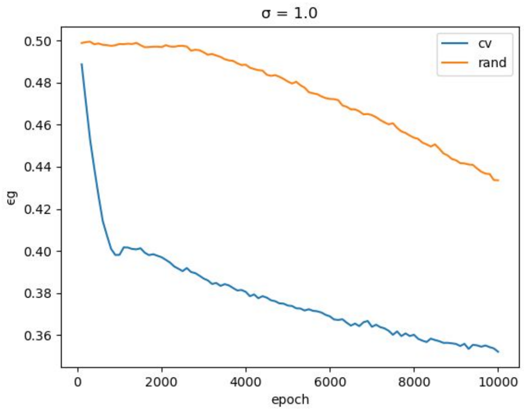
Curriculum learning protocols

Fading: increase the initial SNR by accentuating the distance between the centroids in a subset of the inputs -> Easier to **identify relevant dimensions** of input



Results in the synthetic model

K=4



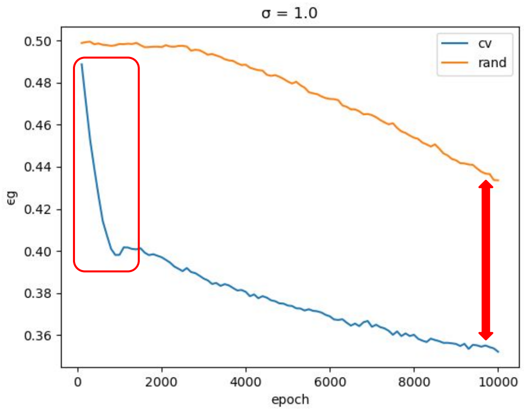
Train a **minimally parametrized** model (**K=4**) on the **XOR-like data**.

Show **10K examples in total**, but with different degrees of difficulty (**10% easy**).

Either **learn** them in **curriculum order** (easy -> hard), **or** in **random order**. In the end the available information in the two protocols is the SAME!

Results in the synthetic model

K=4



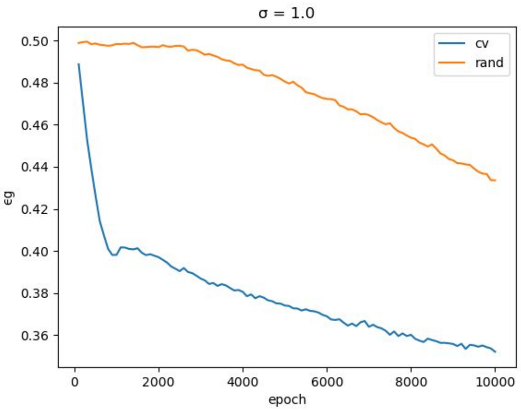
Compared to learning in random order, **curriculum** strategies allow:

+ Initial **speed-up** (all easy examples first!)
+ **Asymptotic performance gap**

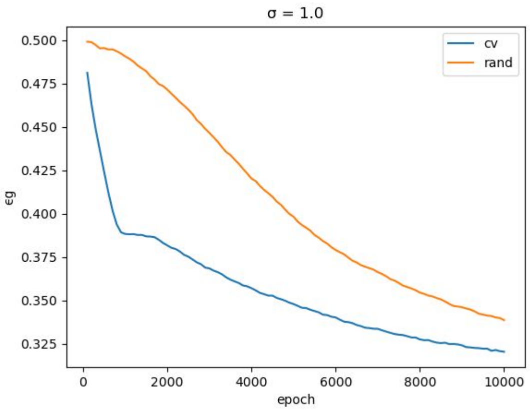
What happens if we **over-parametrize** the network?

Results in the synthetic model

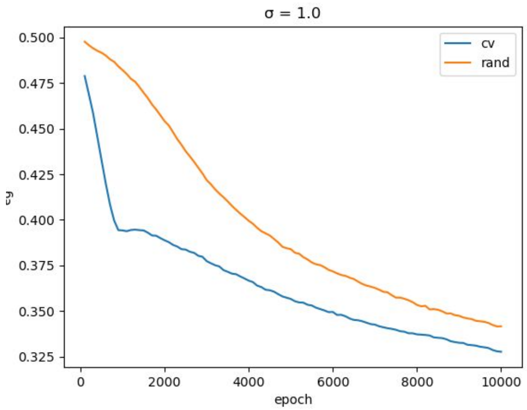
K=4



K=16



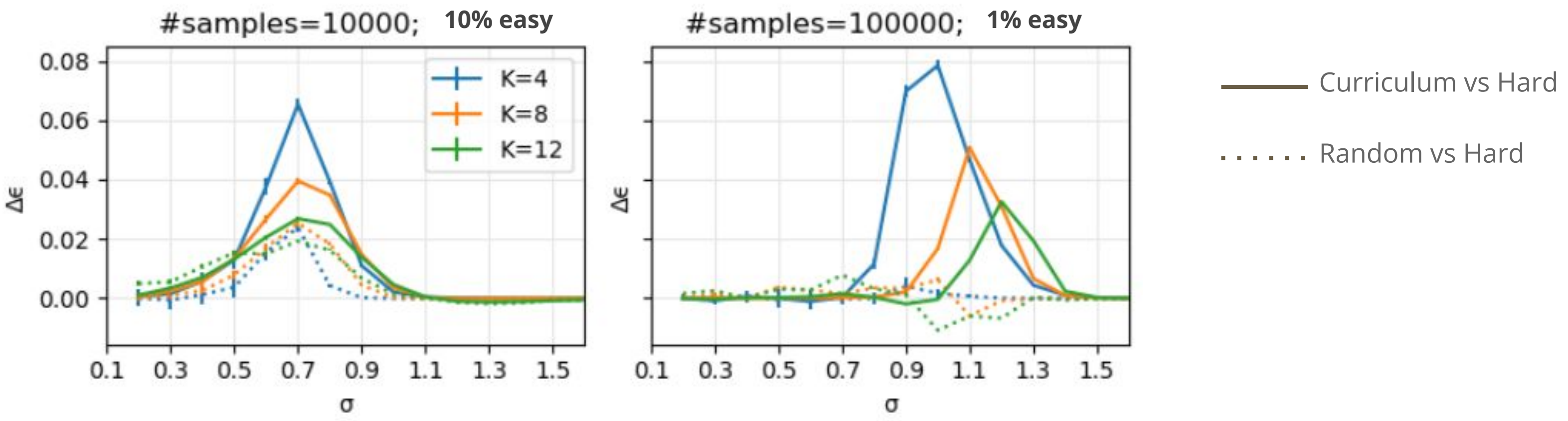
K=64



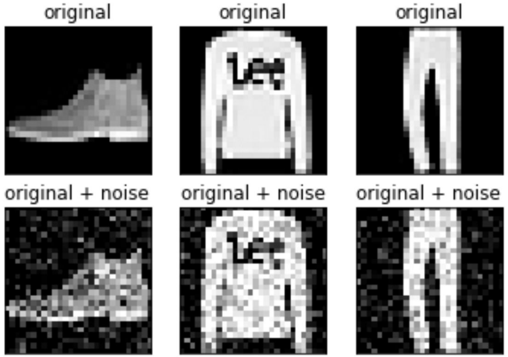
Now only the **initial speed-up survives**, while the larger networks are less affected by the ordering -> the **gap closes**!

Results in the synthetic model

What if we **change the overall difficulty** of the learning problem?



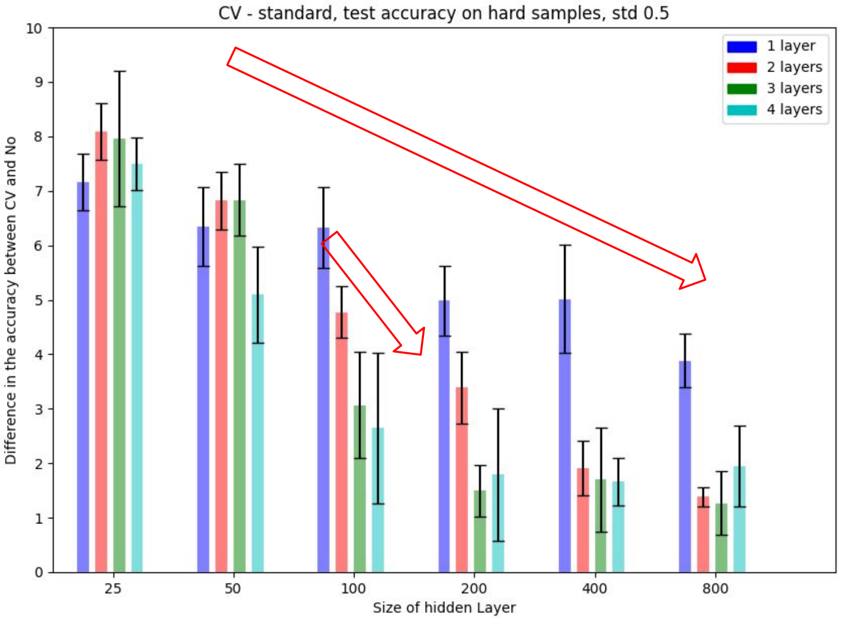
Results on real data



FashionMNIST dataset (white noise)

Fully-connected MLP, the curriculum gain reduces if:

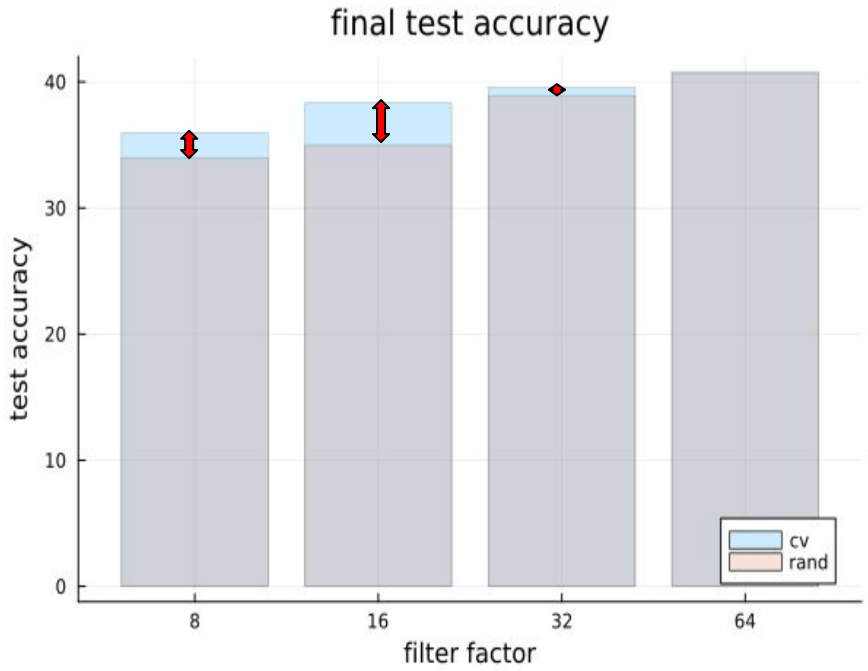
- **# layers** is increased
- **# hidden** units is increased



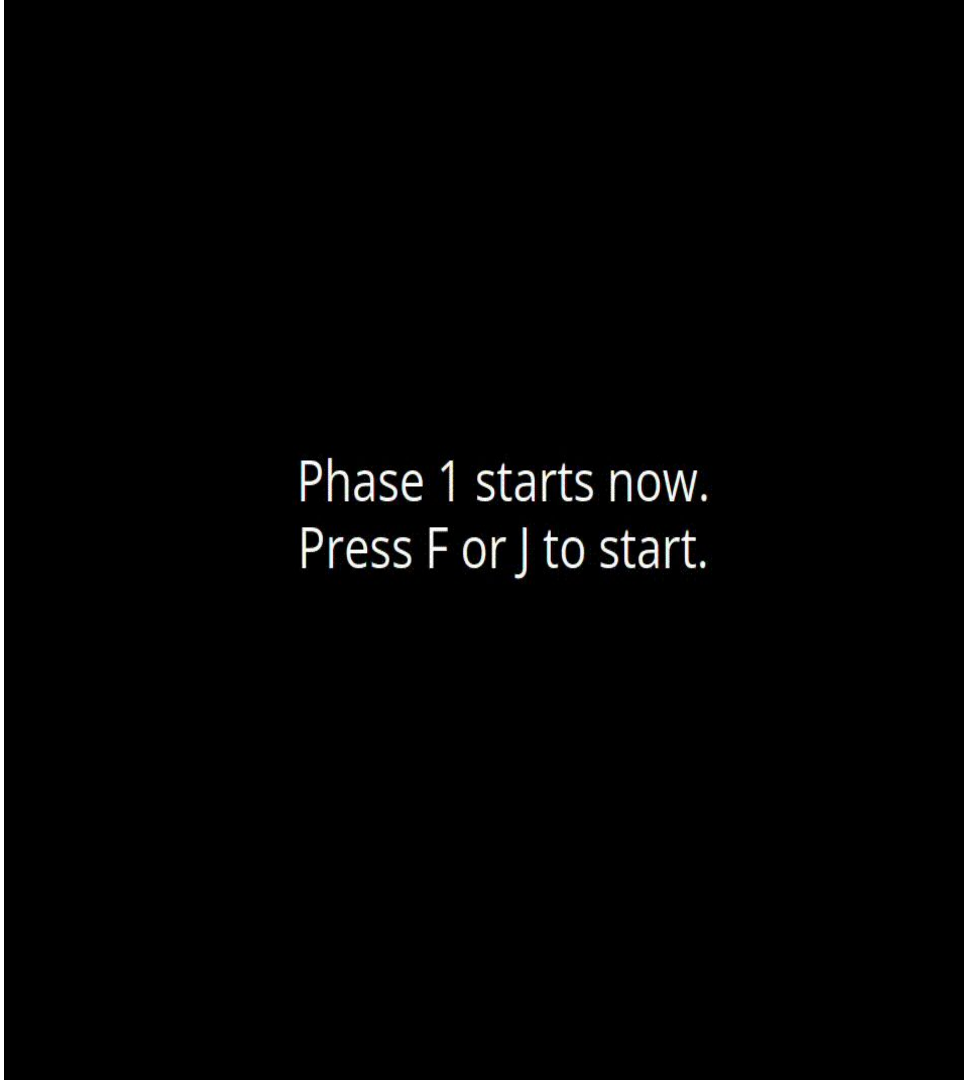
Results on real data

Similar results for a **CNN** on 10K examples from **CIFAR10 (random frames)**

[connection with Umberto's talk]



Curriculum learning can help, but is **not needed** when the model is strongly **over-parametrized**.



Thank you!

Luca Saglietti

Andrew Saxe

GATSBY

Sainsbury Wellcome Centre

UCL