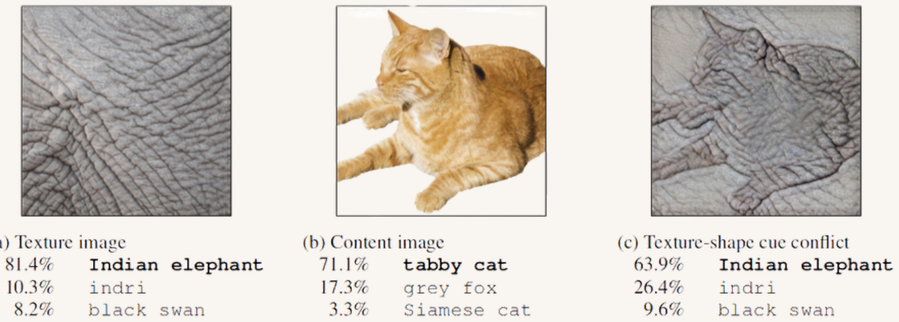
A research recipe for understanding-focused deep learning research

Workshop #4 on Metascience for Machine Learning

Hannah Pinson

Assistant Professor of AI
Data & AI cluster, Data Mining group
Eindhoven University of Technology

Understanding-focused deep learning research?



examples from my own ‘kitchen’:

CNNs are reported to have a shape-texture bias.
But why? How does this arise?

It has been shown that it is sometimes possible to reconstruct input samples from the parameters of a trained network alone.
But why? How does this arise?

After training of very large neural networks, we can often prune weights based on low magnitude.
But why? And how does this arise?

Often the mindset is: this is too hard/impossible.

Understanding-focused deep learning research?

Common ‘types’ of deep learning research:

new architectures / training methods

beating the benchmark

empirical work (shape-texture, lottery tickets,...)

theoretical work (NTK, implicit bias, ...)

interpretability research (-> usually focuses on the trained model)

(...)

Understanding-focused deep learning research?

empirical work (e.g., shape-texture, lottery tickets)
“the blind men and the elephant”

Understanding-focused deep learning research?

empirical work (e.g., shape-texture, lottery tickets)
“the blind men and the elephant”

theoretical work:
the elephant is a box!

Understanding-focused deep learning research?

Understanding-focused deep learning research
is a **balancing act**

the setup usually balances between
too simple to explain anything
and too complex to formalize

(more on this later with an example)

“Understanding” for who?

understanding is in the mind of the beholder!

at which point would you say you ‘understand’ how a neural network works?
(maybe mention your background)

“Understanding” for who?

understanding is in the mind of the beholder!

my ‘understanding’:

starting from the basic components

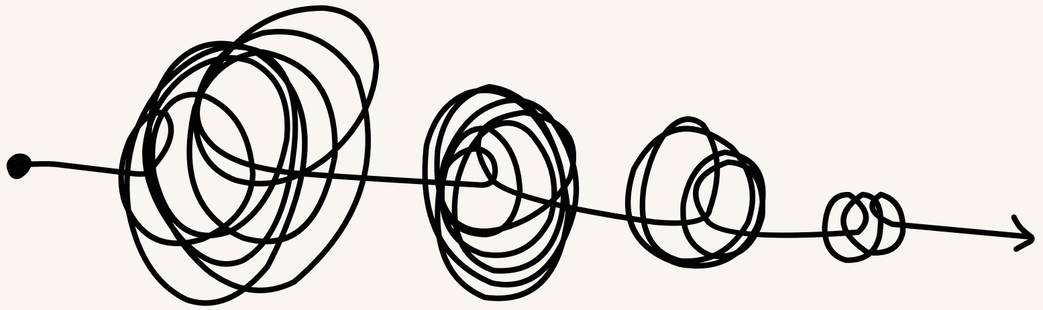
architecture + gradient descent + dataset

I can reason about how the phenomenon arises

The (possible) stages of understanding-focused deep learning research

Understanding is only the first part of the story

then comes sharing your understanding + convincing others it’s correct



Stage 1: building your own intuition

(...once you identified an interesting open problem...)

it’s between you, your computer, and your sheet of paper

very unstructured, requires a lot of creativity (not for everyone!)

quite risky in terms of your CV , as this phase might take a long time and it might result in nothing tangible

Stage 1: building your own intuition

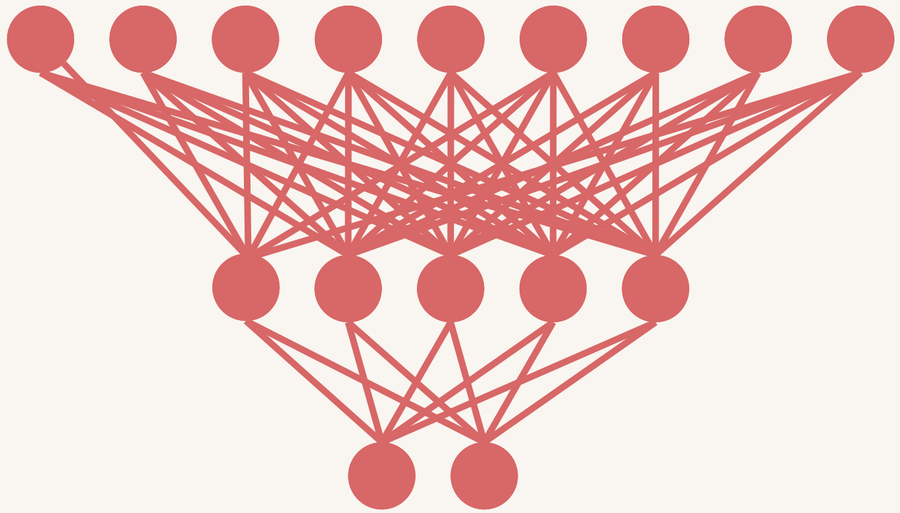
build a ‘playground’ for your mind

e.g. develop a very simple neural network and toy dataset that exhibits the phenomenon you’re interested in
study how it works: use visualizations, probe the system by changing some parameters, etc
until you get a sense of what is going on at a mechanistic level

pitfalls:
too simple
or too complex to be validated in a theoretical framework

how complex you can make this depends on your experience and background knowledge

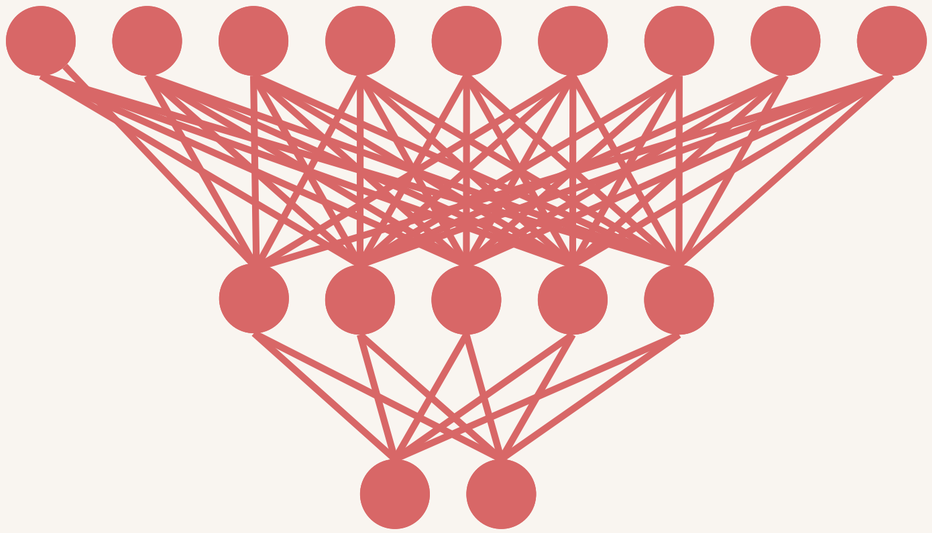
Case study: the lottery ticket *conjecture*



“an untested conjecture that SGD seeks out and trains a subset of well-initialized weights”

there exist weights with ‘beneficial’ initializations (?) that lead to higher norms under training with gradient descent

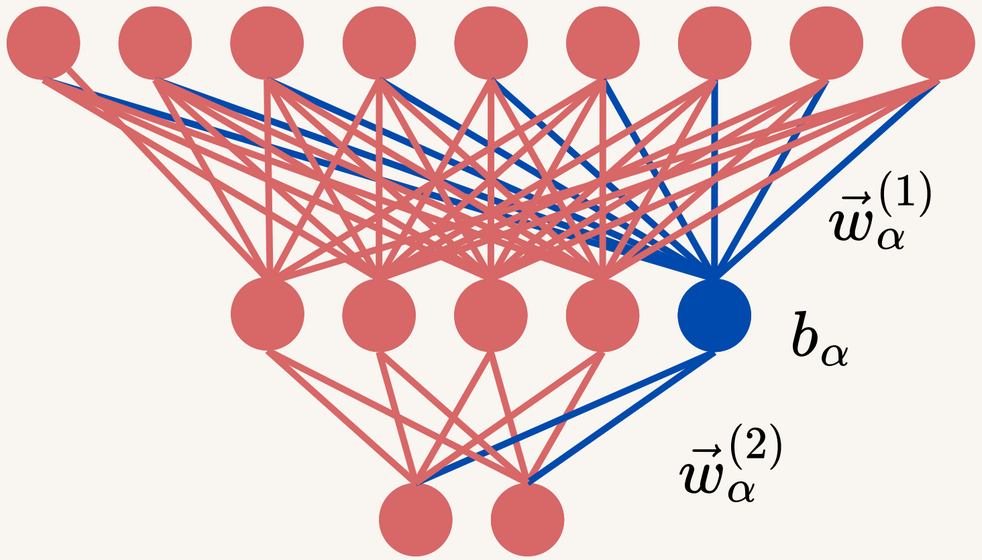
→ weights with less beneficial initializations obtain smaller norms and can be pruned more easily



playground setup:

single hidden layer MLP (“fully connected network”)
ReLU activation functions
binary classification, softmax + cross-entropy loss on two output nodes

dataset: grouping of two classes of CIFAR 10 to form two new class



θ_α $||\vec{w}_\alpha^{(1)}||$ ϕ_α $||\vec{w}_\alpha^{(2)}||$

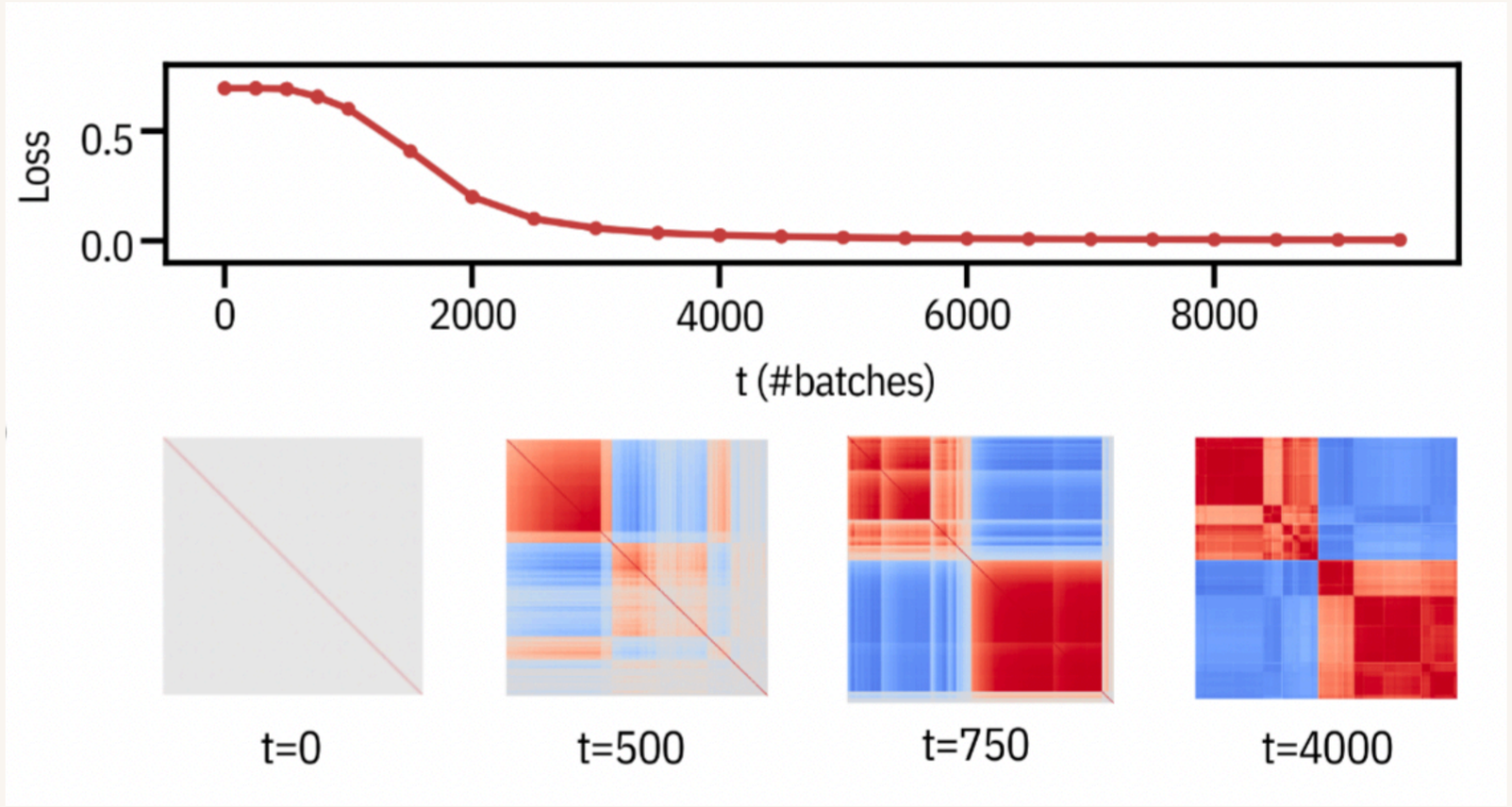
$$\vec{w}^{(total)} = [\vec{w}_\alpha^{(1)} b_\alpha \vec{w}_\alpha^{(2)}]$$

Classical perspective: focus on weight matrices

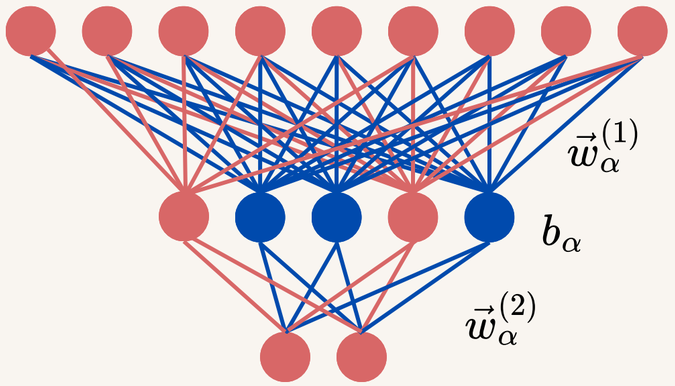
but we consider the parameter vectors **per neuron**

incoming weights, bias, outgoing weights

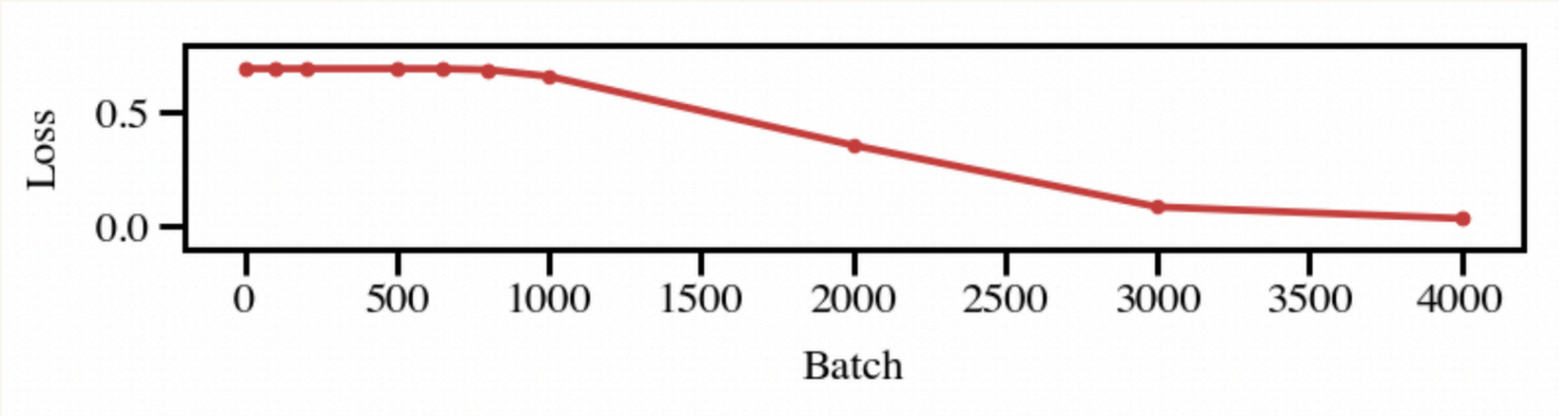
and focus on the norms and angles of those vectors



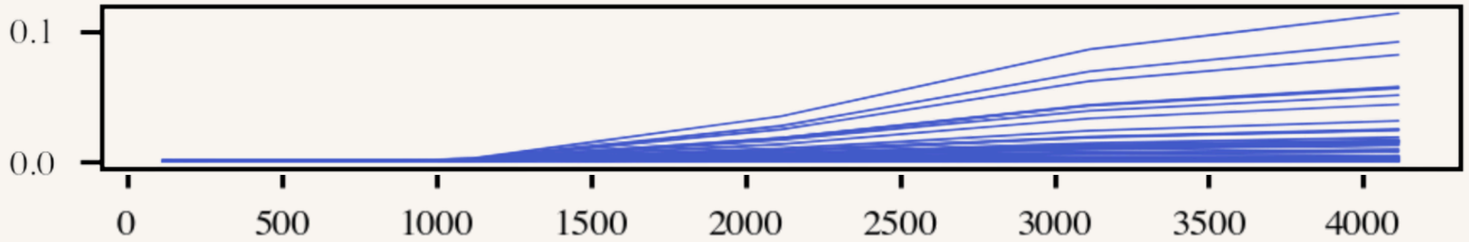
cosine similarity between total parameter vectors of neurons



Select a cluster of neurons (=neurons for which the total paramater vectors align. E.g., cosine similarity > 0.99)



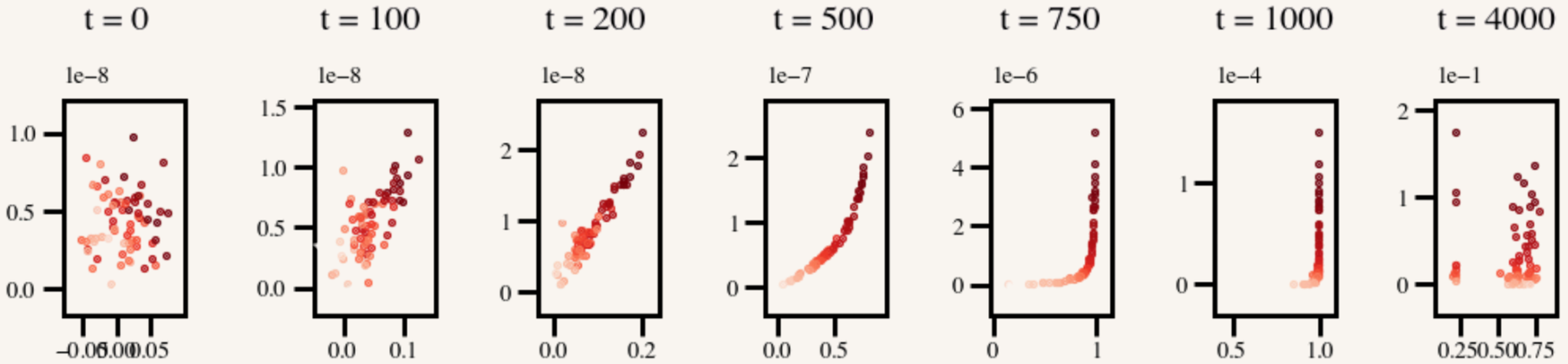
norm of weight vectors



direction of weight vectors (randomly picked two dimensions)

Compute an average direction

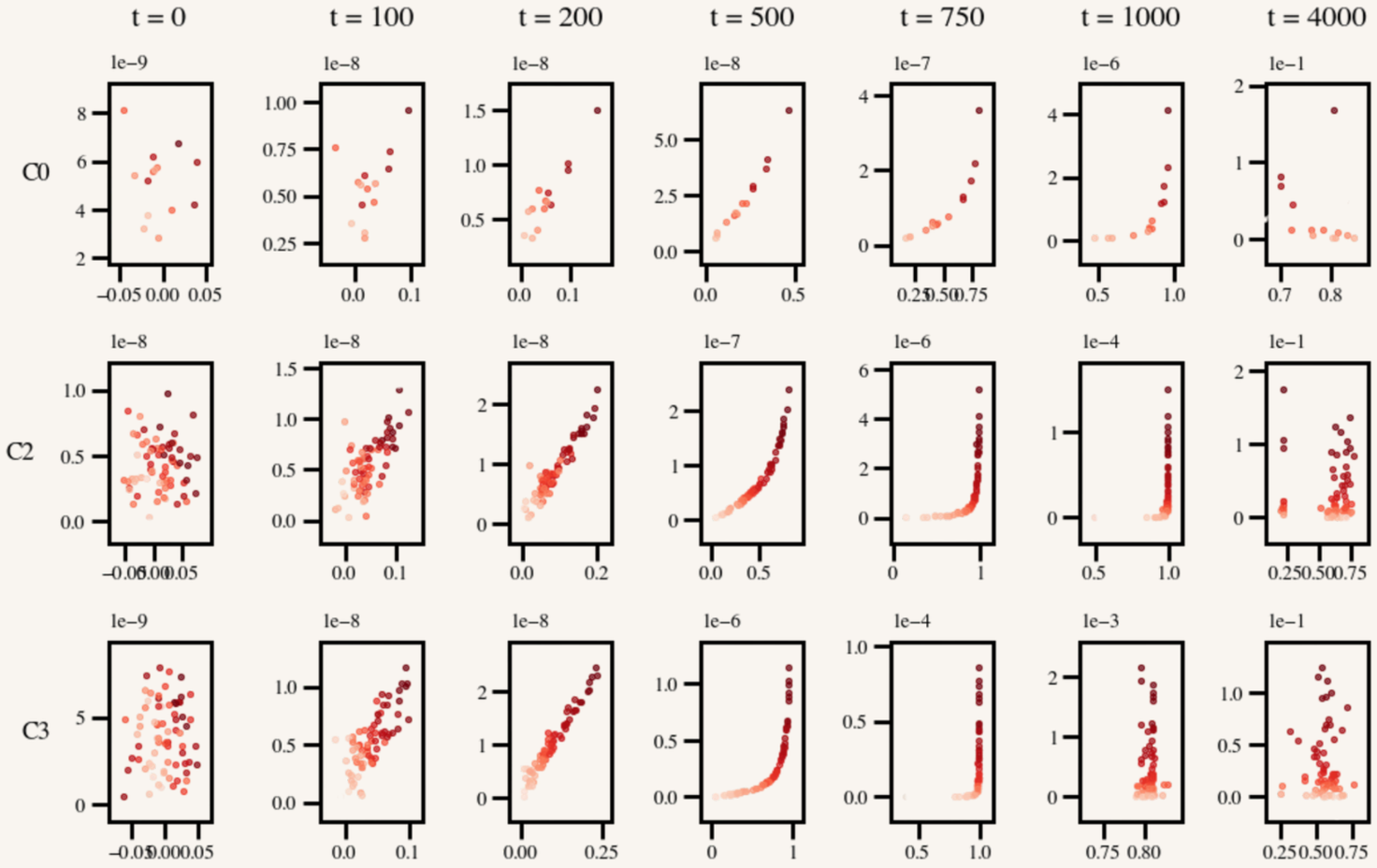
plot norm in function of angular distance (cosine similarity)



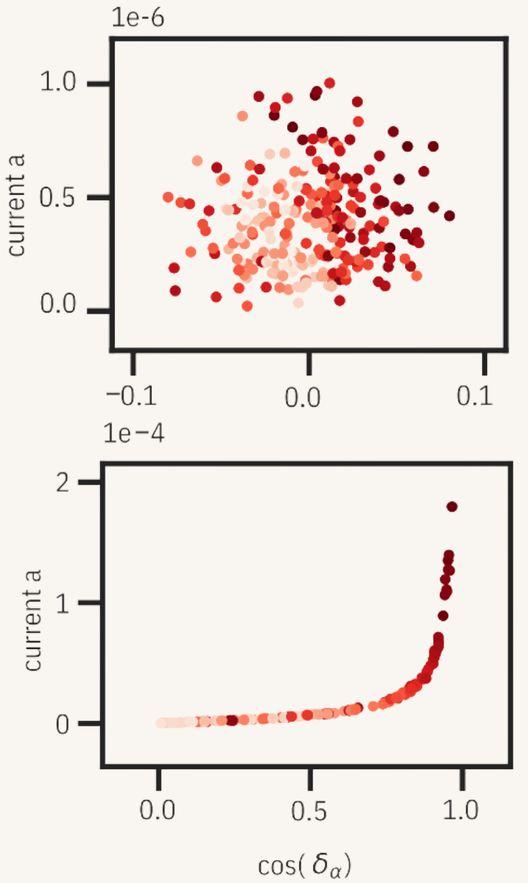
total norm

angular distance (cosine similarity)

every cluster has a target direction

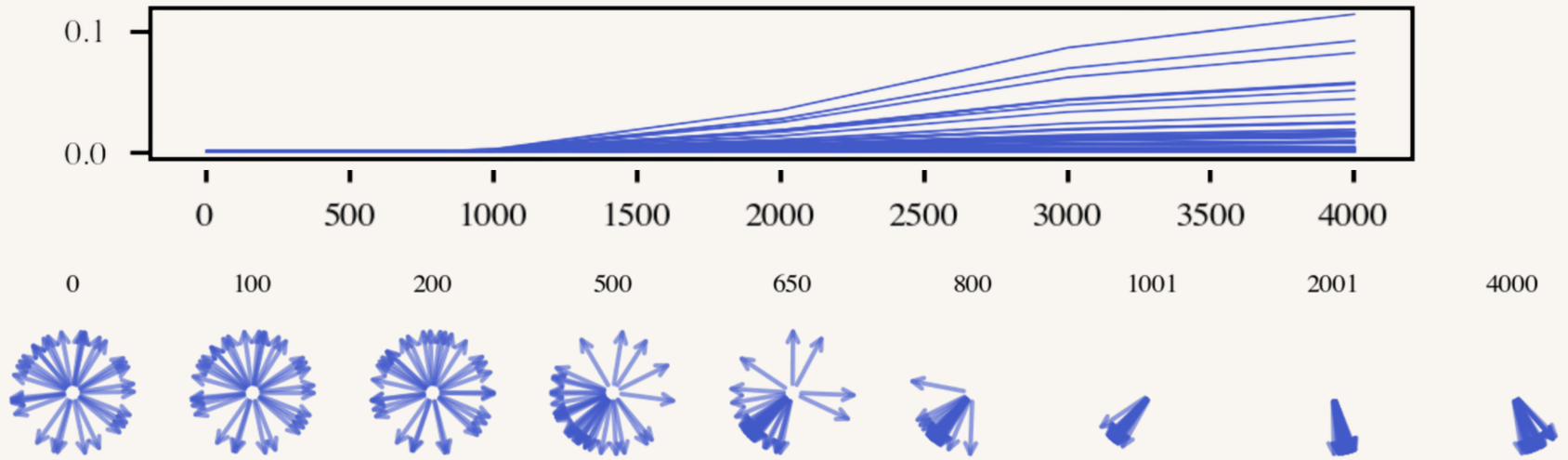


all clusters together:



initialization

timepoint right before loss drops



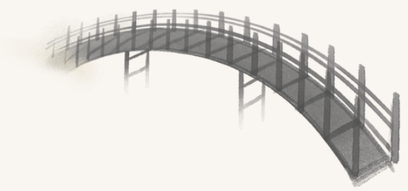
The observed mechanism in a nutshell

- in a first phase, the norms/loss does not change significantly, only the directions change. In this phase, the neurons divide into groups that converge to a shared direction

- once close to this shared direction, the norms start to increase, and the loss starts to change. Neurons within a group that arrive earlier to the shared direction obtain an exponentially higher norm (->racing). They arrive earlier if they were initialized close (in direction) to this target direction

- this difference in norm (the ordering) seems to be sustained (to a large degree) throughout the rest of training

Stage 2: formalizing your intuition (for yourself)



it’s between you, your computer, and your sheet of paper
(but also already
between you, your audience, and your reviewers.)

i.e., how this process is structured is shaped by the nature of the audience and the reviewers

→ you want to induce ‘understanding’
→ but you also want to convince with formal arguments / mathematical statements

another balancing act

(if you have the intuition but not the formal derivation → try an LLM ?)
(if you have a formal result (in another paper) but not the intuition → try an LLM?)

My preferred framework to formalize my intuition

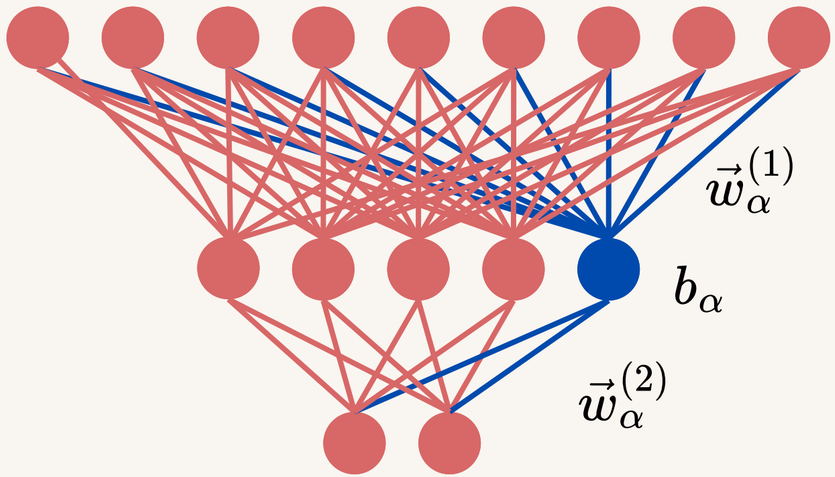
Learning dynamics (physics inspired):

treating neural networks as dynamical systems
and study the evolution of their parameters over time

treating the equations of gradient descent as a set of differential equations
(this is a *very* complicated set of equations)

finding approximations / simplifications that are both valid and useful (=allow you to analyse the dynamics)

through iterations, *I construct the playground such that I can use learning dynamics*
(other options: NTK, linear networks,...)



$$\langle \frac{\partial L^s}{\partial \phi_\alpha} \rangle = -||\vec{w}_\alpha^{(2)}||\ \langle h_\alpha^s\ ||\vec{\Delta y}^{\,s}||\ \sin\left(\phi_{\vec{\Delta y}^s} - \phi_\alpha\right)\rangle$$

$$\langle \frac{\partial L^s}{\partial ||\vec{w}_\alpha^{(2)}||} \rangle = -\langle h_\alpha^s\ ||\vec{\Delta y}^{\,s}||\ \cos\left(\phi_{\vec{\Delta y}^s} - \phi_\alpha\right)\rangle$$

$$\langle \frac{\partial L^s}{\partial (\theta_{\alpha,i})} \rangle = -||\vec{w}_\alpha^{(2)}||\ ||\vec{w}_\alpha^{(1)}||\ ||\langle \vec{\gamma}_\alpha^s \rangle||\ \sin\left(\theta_{\langle \vec{\gamma}_\alpha^s \rangle, i} - \theta_{\alpha,i}\right)$$

$$\langle \frac{\partial L^s}{\partial ||\vec{w}_\alpha^{(1)}||} \rangle = -||\vec{w}_\alpha^{(2)}||\ ||\langle \vec{\gamma}_\alpha^s \rangle||\ \cos\left(\theta_{\langle \vec{\gamma}_\alpha^s \rangle, i} - \theta_{\alpha,i}\right)$$

$$\langle \frac{\partial L^s}{\partial b_\alpha} \rangle = -||\vec{w}_\alpha^{(2)}||\ \langle ||\vec{\Delta y}^s||\ \cos\left(\phi_{\Delta y^s} - \phi_\alpha\right)\rangle$$

starting from the basic components architecture + gradient descent + dataset → combine these in the equations → analyze the equations (see paper)

Stage 3: sharing your understanding + convincing others it’s correct

it’s between you, your audience, and your reviewers

i.e., how this process is structured is shaped by the nature of the audience and the reviewers

→ you want to induce ‘understanding’
→ but you also want to convince with formal arguments / mathematical statements

another balancing act

Stage 3: sharing your understanding + convincing others it’s correct

it’s between you, your audience, and your reviewers

easy scenario:
you’re a mathematician,
your audience are mathematicians,
the reviewers are mathematicians

hard(er) scenario:
you are a deep learning researcher with a training in physics,
your audience are deep learning researchers not trained in physics,
your reviewers are either mathematicians or they only like practical applications.

Practical approaches:

- formalize what you can, show with experiments where you have to
- main body conceptual, appendix formal
- split between a formal paper and a blog post / recorded talk

h.pinson@tue.nl

technical talk on April 27th (online & recorded)

Pinson, Hannah. "It's not a Lottery, it's a Race: Understanding How Gradient Descent Adapts the Network's Capacity to the Task." arXiv preprint arXiv:2602.04832 (2026).

(work-in-progress)