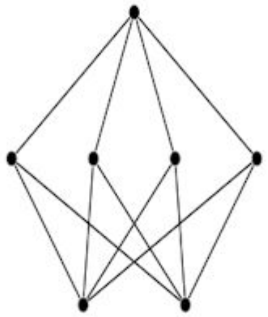
An Empirical Investigation of the Failure Mode of Training in Mildly Overparameterized NNs

Master's Thesis
- Midterm Presentation-

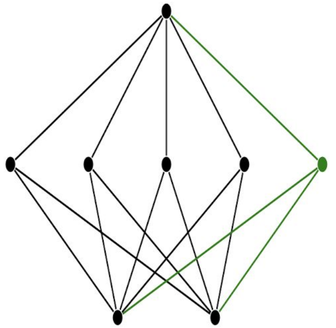
Student: *Theodor Stoican*
Supervisor: *Berfin Şimşek*
Co-supervisor: *Johanni Brea*

Setup

- Teacher-student setup (same setup as the one published in [1])



4-sized teacher

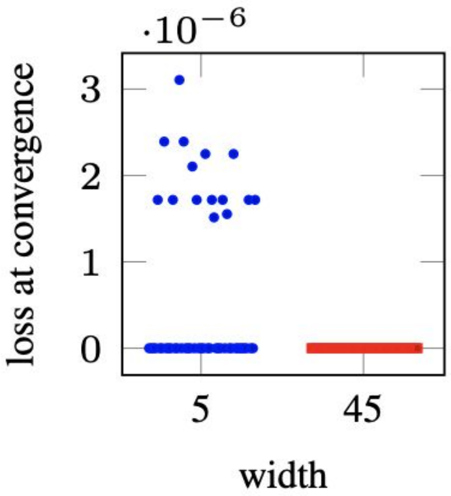


A 5-sized student

- Dataset
  - 1681 points on a regular grid $\{(x_1, x_2)|4x_1 = -20, \ldots, 20, 4x_2 = -20, \ldots, 20\}$
  - with labels $y = \Sigma_{i=1}^4 a_i \sigma(\Sigma_{j=1}^2 w_{ij} x_j)$, where w_{ij} - preset weights of teacher, and $a_1 = 1, a_2 = -1, a_3 = 1, a_4 = -1$
- Teacher - expressive enough to achieve zero loss on this dataset
- Student - vary the size, minimum of 5

Issue

- Teacher-student setup (same setup as the one published in [1])
- A student of width 45 - achieves zero loss consistently
- A student of width 5 - for a fraction of initializations it fails
  - Why exactly?



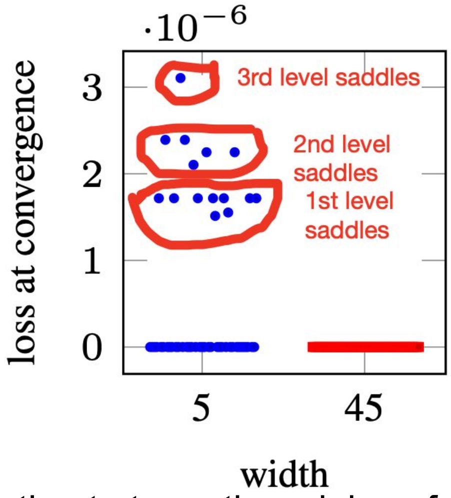
Difference between loss values at convergence ([1])

Intuition

- To formulate a first intuition, we need to look at the G number
  - $G(r,m) = \Sigma_{k_1+\ldots+k_r=m} \binom{m}{k_1,\ldots,k_r}$
  - (= the number of critical subspaces in an overparameterized NN generated from a critical point of a smaller network)
- For a teacher of size 4 and a student of size 5:
  - $G(4,5) = 240, G(3,5) = 150, G(2,5) = 30$
  - $G(4,5) > G(3,5) > G(2,5)$
  - Hence, first level saddles are more common than second level saddles and so forth

Intuition (2)

- Intuitively we can associate the points of failure within the chart to one of these classes



Attempting to trace the origins of non-zero loss points ([1])

Problem Definition

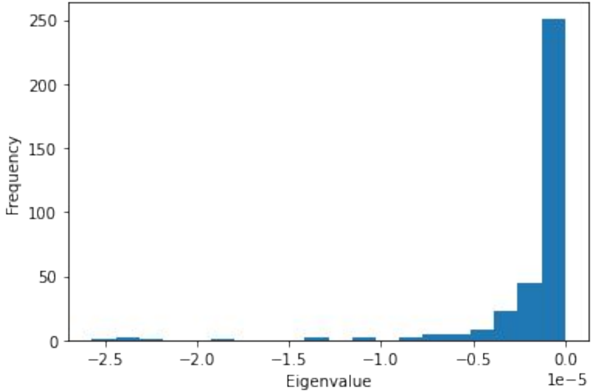
- Can one examine the nature of these failure points beyond intuition?
  - For that we propose a set of experiments
  - The student must learn the teacher's weights (Setup)
  - Hence, for a student with sizes of 4, 5, 6, 7
    - Run 1000 experiments
    - Adam optimizer (l.r.=0.0001)
    - For each experiment, accurately identify:
      - the local minima
      - the saddles
      - the global minima

Strict Saddles - Intro

- From a theoretical standpoint, a point is a strict saddle if:
  - The Hessian of the function at that point has a neg. eigenvalue
- From a practical standpoint, one could apply this simply by:
  - Identifying a point with a low magnitude gradient
  - Computing the Hessian with *torch.autograd.grad*
- As usual with practice, things turn out to be more complicated, due to:
  - Numerical errors
  - Imperfect computation methods (*torch.autograd.grad)*

Strict Saddles - First Attempt

- All points identified with this method have a small magnitude smallest eigenvalue



Distribution of the smallest eigenvalue (for experiments with negative eigenvalues). Student size: 5

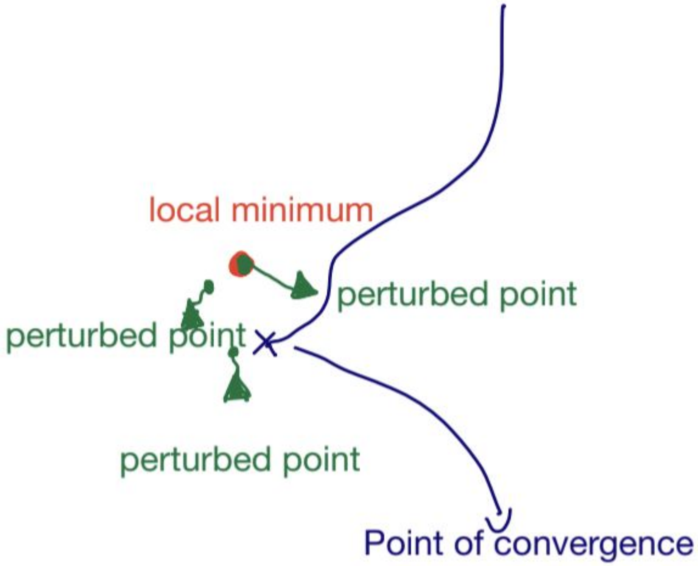
- 347/1000 experiments have a negative smallest eigenvalue
- Second order optimization (SLSQP) was further applied at these points
- Rarely, it has yielded a new point with positive eigenspectrum
- Hence, is the algorithm just prone to numerical issues or are these indeed saddles?

Strict Saddles - Perturbation

- To find out an answer, the next technique was employed:
  - Perturb the point where Adam has converged by:
    - Sampling from a normed multivariate Gaussian
      - Mean: 0
      - Standard deviation: 0.1
    - Adding to the neuron point: $\epsilon * normed\_gaussian$
  - Apply SLSQP from that point
  - Repeat until the smallest eigenvalue is positive

Strict Saddles - Perturbation (Intuition)

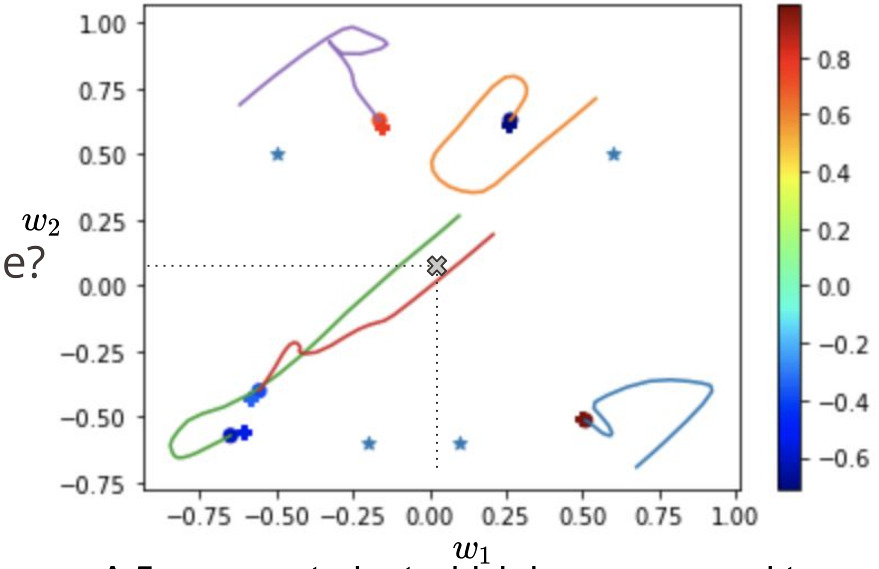
- Hopefully by perturbing *very slightly*, one can get closer to the min.
- Only one perturbation has to be successful



Desired effect of perturbation

Strict Saddles - Perturbation Results

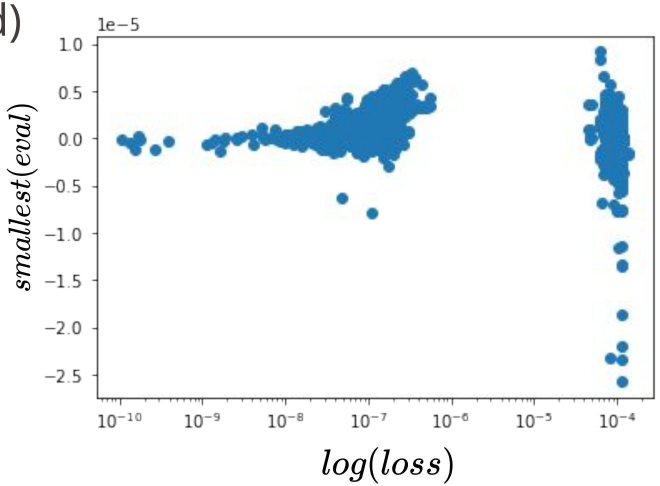
- This technique has converged to a local/global min. for all experiments
- The smallest eigenvalue has flipped signs (from neg. to pos.)
  - -1.7e-06 -> 4.45e-07
  - 2 random perturbations
- The loss has slightly changed
  - 1.75e-07 -> 5.92e-08
- Discussion - why neg. smallest eigenvalue?
  - Numerical errors ?
  - Optimizer limitation?
  - Or actually saddles?



A 5 neuron student which has converged to a minimum (crosses represent convergence after perturbation)

Strict Saddles - Regimes

- Additionally, in order to distinguish between the 2 cases:
  - See how the smallest eval correlates with the loss
  - If a negative smallest eval is clustered around a loss value
    - Intuitively it is likely that these are saddles
  - Otherwise (if they are randomly distributed)
    - It could be due to numerical issues
- In the chart, one can see obvious clusters
  - First regime: saddle dominated
  - Second regime: local min. dominated
  - Third regime: global min. dominated



Correlation between the smallest eigenvalue and the loss.
Visible clustering of the losses dependent on the smallest eval.

Global Minima - Issue

- In literature, points are commonly considered global minima if the loss is small enough (sometimes l.t. 1e-3 - [2], sometimes l.t. 1e-4 - [3])
- Seemingly not a reliable criterion
  - Some local minima we've found correspond to a loss of 1e-4
  - Others to an even smaller loss
- Hence, how can we more reliably identify global minima?

Global Minima - Symmetry Induced

- A category of global minima are symmetry induced
- Visually identifiable, since they all respect the following property ([1]):

$$\theta^r = (w_1, \ldots, w_r, a_1, \ldots, a_r)$$

$$\downarrow$$

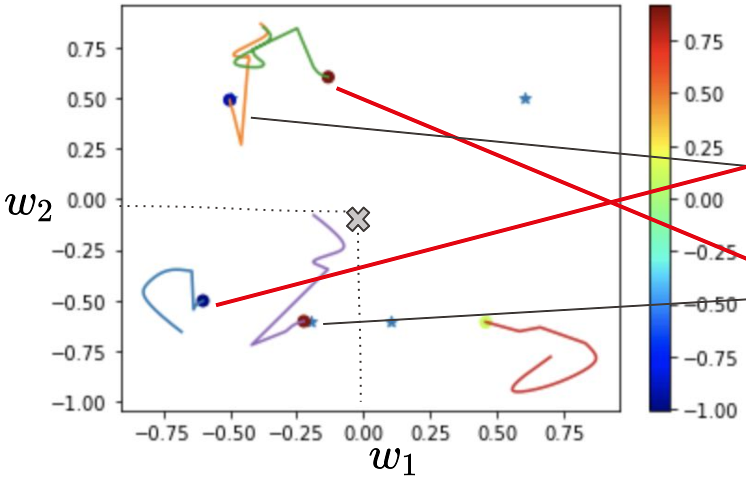
$$\theta^m = (\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w'_1, \ldots, w'_1}_{b_1}, \ldots, \underbrace{w'_j, \ldots, w'_j}_{b_j},$$

$$a_1^1, \ldots, a_1^{k_1}, \ldots a_r^1, \ldots, a_r^{k_r}, \alpha_1^1, \ldots, \alpha_1^{b_1}, \ldots, \alpha_j^1, \ldots, \alpha_j^{b_j})$$

, under: $\Sigma_{i=1}^{k_t} a_t^i = a_t, \Sigma_{i=1}^{b_t} \alpha_t^i = 0$

Global Minima - Visual Inspection

- As well as the sigmoid symmetry property:
    - $\sigma(-wx) = 1 - \sigma(wx)$
- One example
    - Two of the student's weights match the teacher's weights
    - The other 2 are symmetric w.r.t. to the origin



A student with 5 neurons and a corresponding loss of ~1e-9 at convergence.

Incoming weights		Outgoing weights
0.6	0.5	1
-0.5	0.5	-1
-0.2	-0.6	1
0.1	-0.6	-1

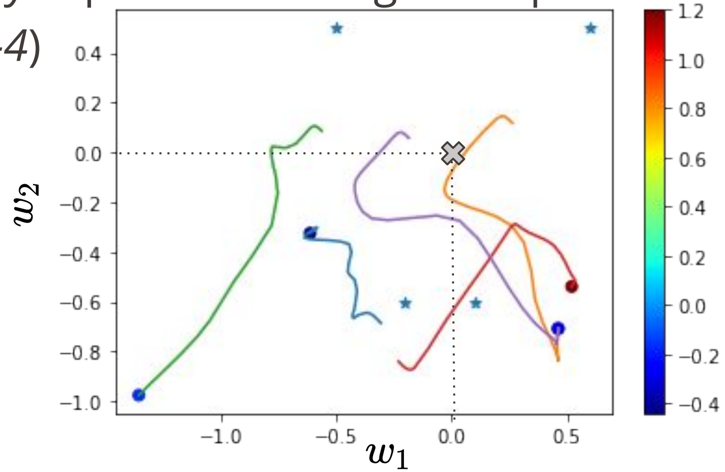
Teacher's configuration

Local Minima - Theory

- Assume we have a point to which the following correspond:
  - A small enough gradient
  - A Hessian with all positive eigenvalues
- Then this point is a candidate for a local minimum ([2])
- However, how does one know this is not, in fact, a global minimum?
- Intuitively a global minimum has to correspond to a rather small loss
- But how to identify this threshold specifically?

Local Minima - Visual Attempt

- Identifying some local minima *visually* is possible for high-loss points
- E.g. (neuron-point with a loss of ~*1e-4*)



None of the student's neurons converge to teacher's neurons.

- Formally, one can prove that this is a local minimum by applying the lemmas from [2]

Local Minima - Formal Attempt

- Essentially, there are 2 lemmas ([2])
- Lemma 1
  - For a neuron point, given that
    - the gradient at that point is small enough
    - the hessian is positive definite
    - (and a few other assumptions)
  - then this point is within a certain radius of the min (local/global)
- Lemma 2
  - Given that
    - the loss function is a Lipschitz function
    - the loss at the current point is larger enough than 0
  - The neighboring minimum can't have a loss of 0 -> non-global

Next Steps

- Apply the aforementioned lemmas in order to prove that some of the critical points are local minima
- Check out if the remaining critical points with positive smallest eval are global minima
- Find further statistics to find all saddles
- Identify the level of the existing saddles (first-level, second-level, etc.)
- Extend the same statistics for other sizes of the student (4, 6, 7) and compare the results

References

[1] Şimşek Berfin et al., 2021, Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances

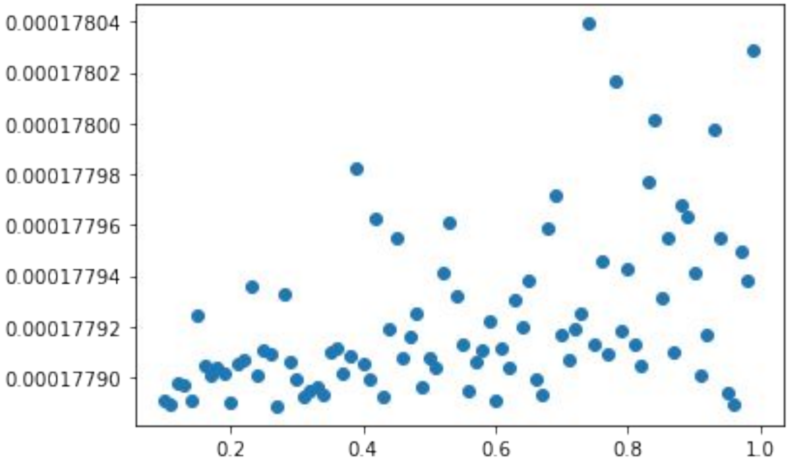
[2] Safran Itay, Shamir Ohad, 2017, Spurious Local Minima are Common in Two-Layer ReLU Neural Networks

[3] Zhang Yaoyu, 2021, Embedding Principle of Loss Landscape of Deep Neural Networks

Strict Saddles - Transition

- Furthermore, by playing with epsilon (the perturbation factor), one can see how the loss changes
- If there is a sharp transition by increasing epsilon, we could intuitively remark the presence of a saddle

I'm surprised that y-axis barely change. I am lost on this slide. Did you train with second-order after perturbation? With this version of the fig, I fail to see any transition…



Presence of a transition in the loss values by increasing epsilon (the perturbation factor)