Predicting Grokking Long Before it Happens:
A look into the loss landscape of models which grok

Pascal Jr. Tikeng [1,2] Hattie Zhou[1,2] Mohammad Pezeshki[3]
Irina Rish[1,2] Guillaume Dumas[1,4]

[1]Université de Montréal, Montréal, Quebec, Canada

[2]Mila, Montréal, Quebec, Canada

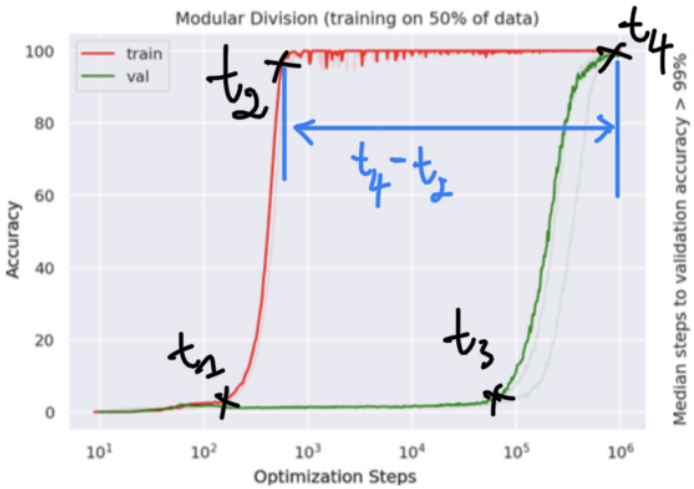
[3]Meta AI Research

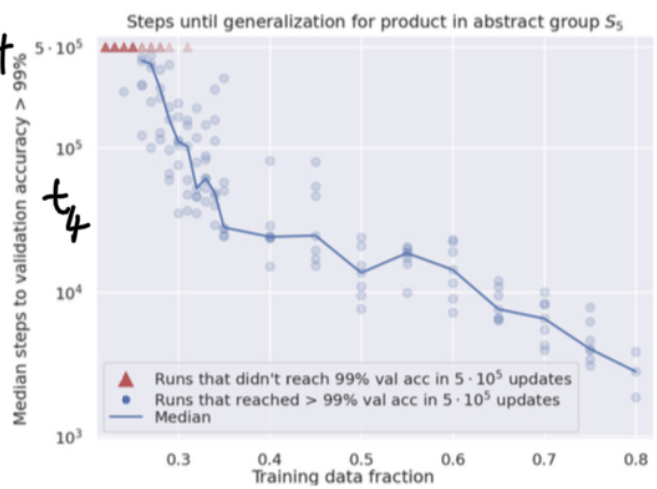
[4]CHU Sainte-Justine Research Center, Montréal, Quebec, Canada

5th workshop on Neural Scaling Laws: Emergence and Phase Transitions
Co-located with ICML 2023

Outline

Grokking



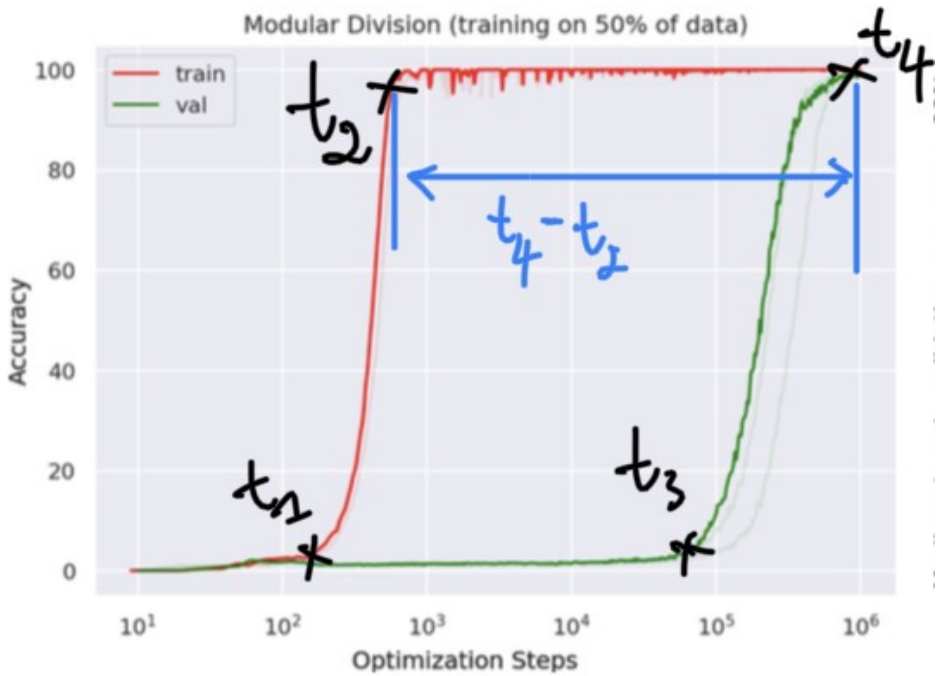


★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

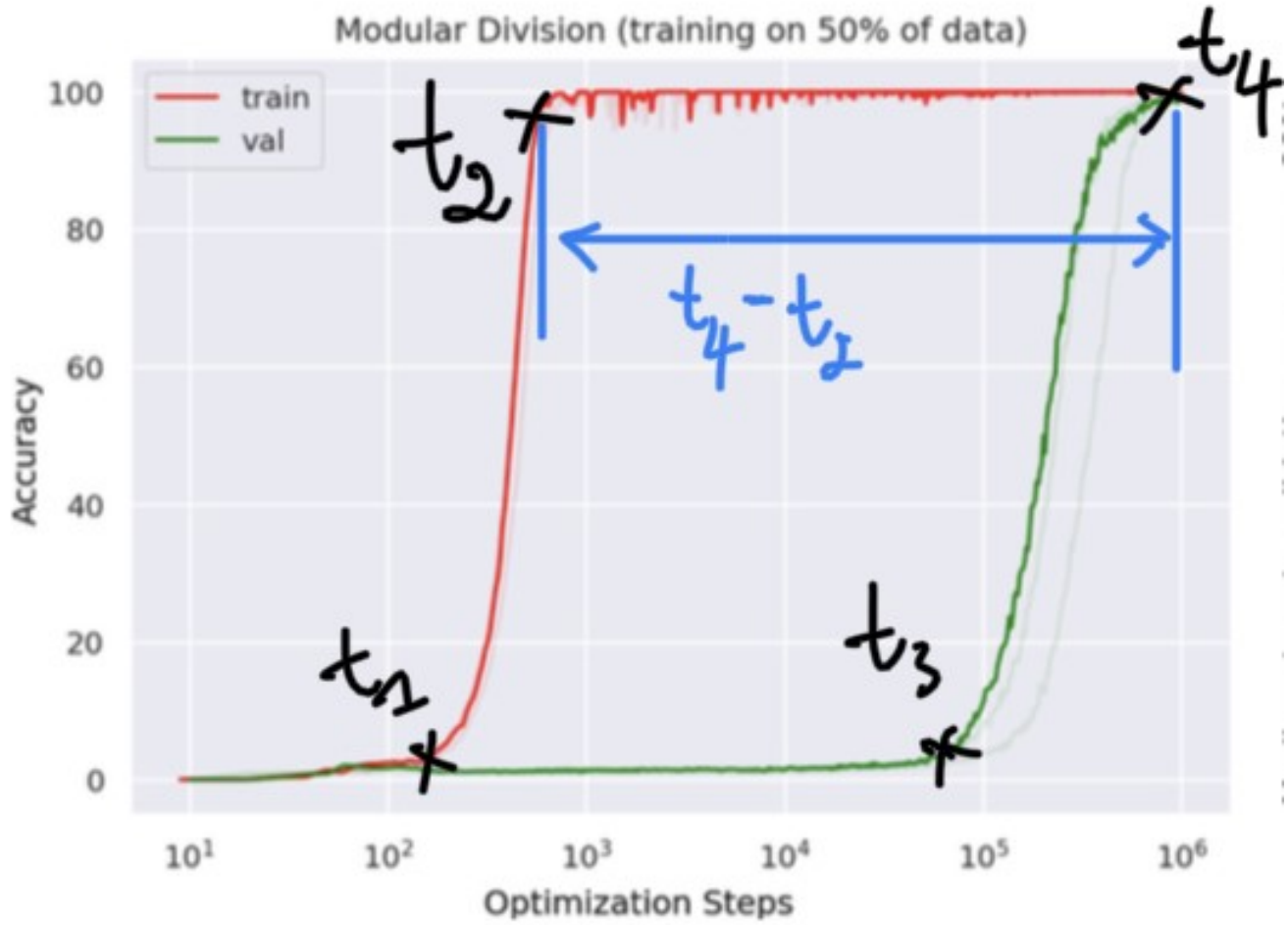
Figure: Generalization after overfitting (Power et al., 2022), training and validation accuracies. Training accuracy becomes close to perfect at $t_2 < 1k$ optimization steps, but it takes close to $t_4 \approx 1000k$ steps for validation accuracy to reach that level.

Four learning phases (Liu, Kitouni, et al., 2022)

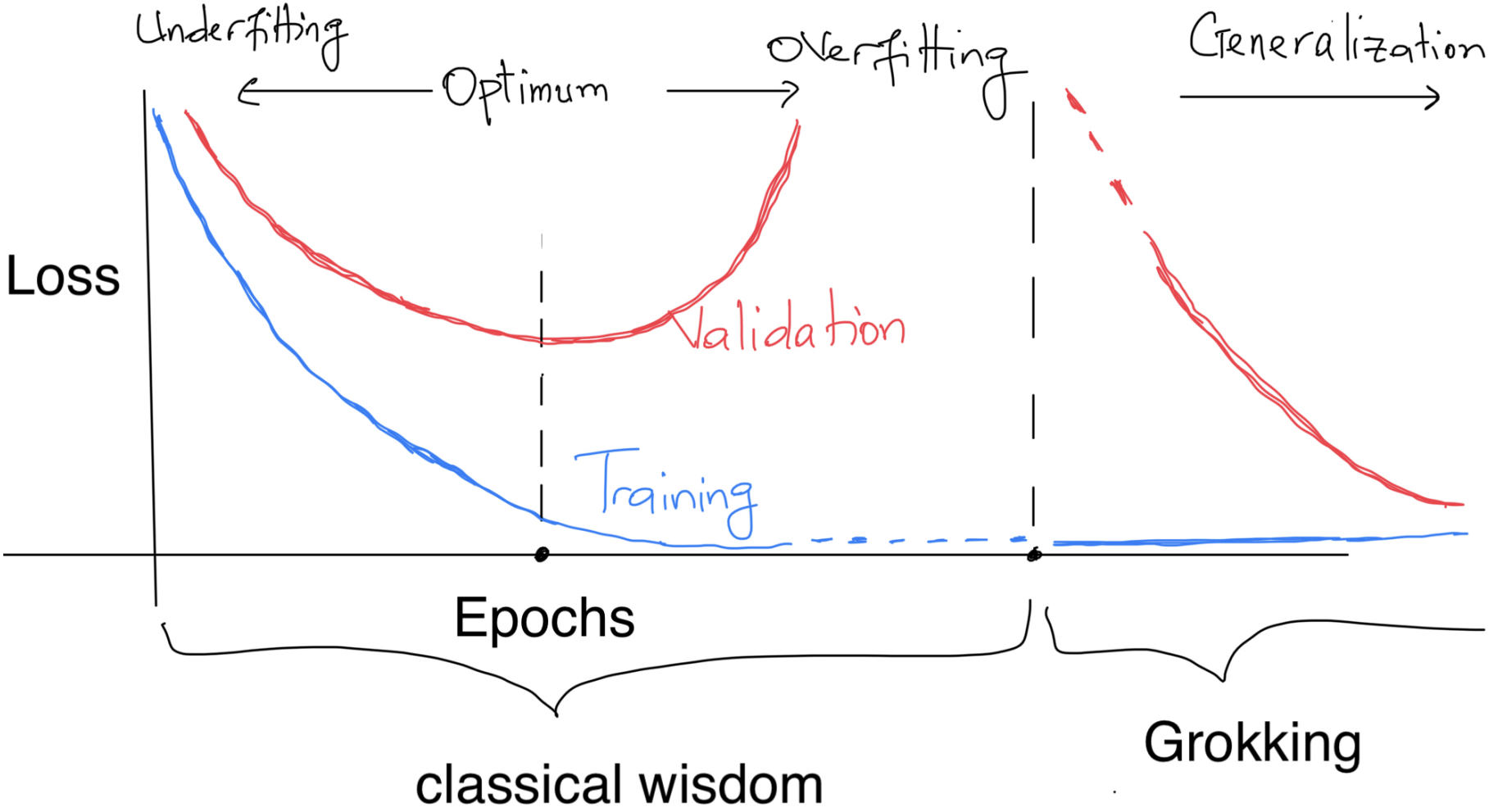
- Confusion : $t \in [0, t_1]$
- Memorization : $t \in [t_2, t_3]$
- Comprehension : $t \in [t_3, \infty]$
- Generalization : $\mathbb{P}[t_4 < \infty] = 1$
  The measure $\mathbb{P}$ captures randomness in initialization, choice of training and validation points, noise in optimization...



- Generalization : $t_4 < \infty$ almost surely (wrt the randomness in initialization, choice of training and validation points, noise in optimization...)
- Grokking $\approx$ a generalization with $t_4 \gg t_2$.



Grokking : generalization with $t_4 \gg t_2$



Challenges

Training budget

- Grokking often requires models to be trained for a very large number of epochs
- Empirically, $t_4(r) \in \Theta\left(1/r^{\gamma}\right)$ with $r = |\mathcal{D}_{train}|/|\mathcal{D}_{total}|$ and $\gamma > 0$. Generalization : $\mathbb{P}[t_4 < \infty] = 1$
- Model sometimes needs to be trained for more than 100k epochs to observe any sign of generalization when the training data size becomes smaller and smaller.

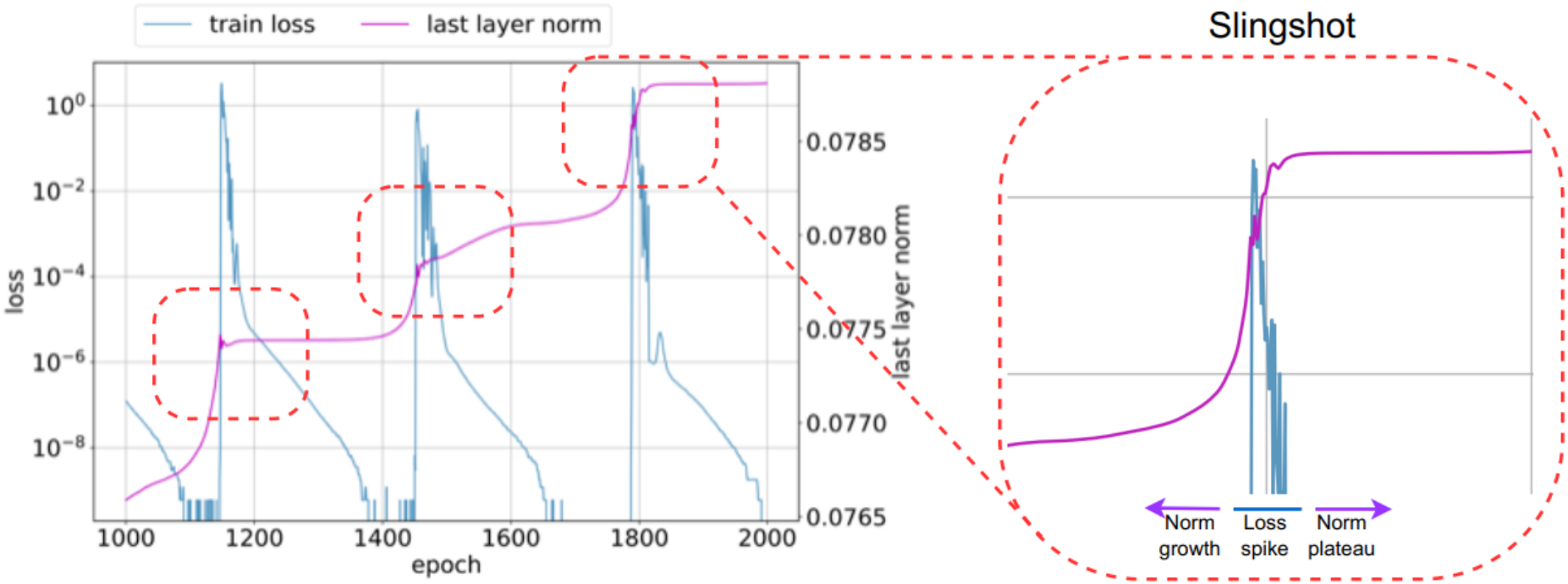
Geometries and generalization properties of solutions found by SGD

- Grokking open the way to new studies concerning the structure of the minimum found by Stochastic Gradient Descent (SGD), and how networks behave in the neighbourhood of SGD training convergence.

Spectral Signature of the loss

Slingshot mechanism (Thilak et al., 2022)

- Generally come in tandem with grokking
- Grokking almost exclusively happens at the onset of slingshots and is absent without it.



Spectral Signature of the loss

Grokking almost exclusively happens at the onset of slingshots and is absent without it.

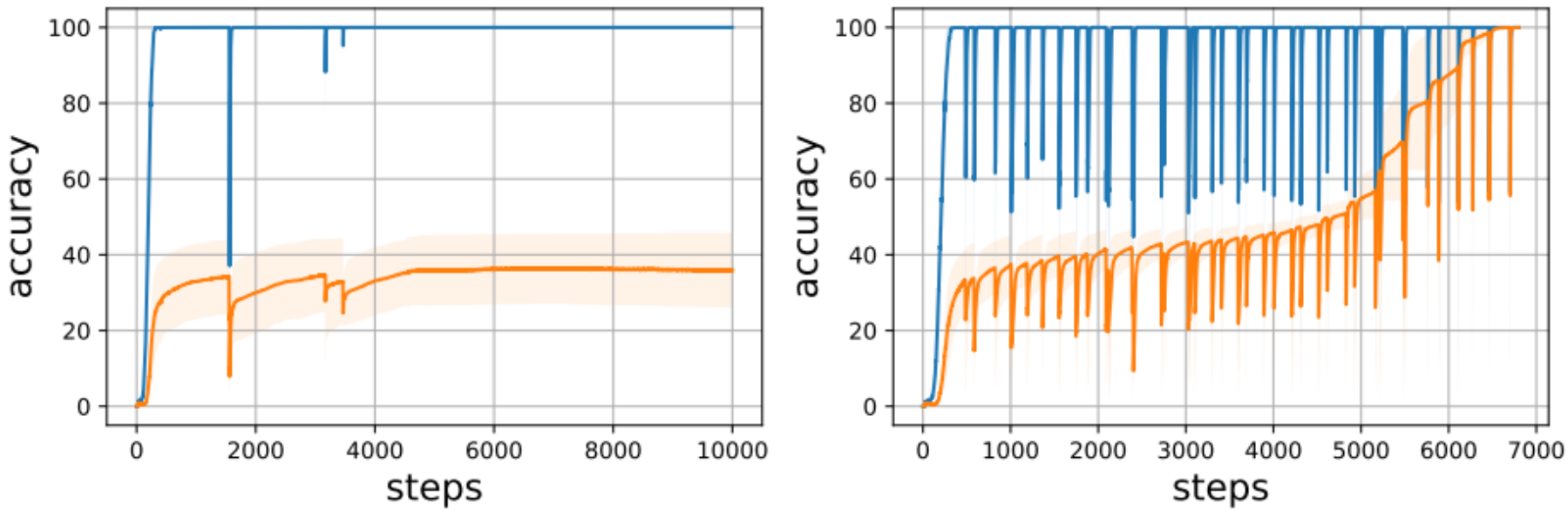
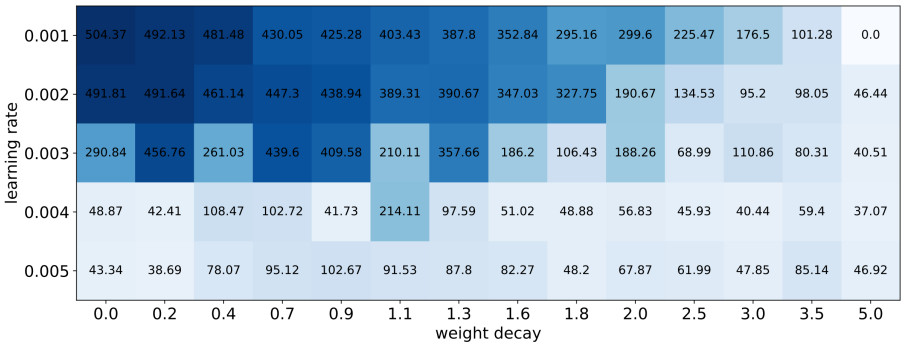


Figure: Oscillation in training and validation accuracies (10k steps).

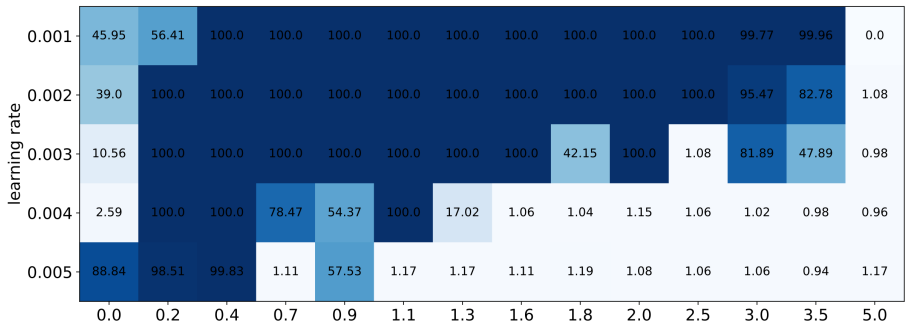
Spectral Signature of the loss is correlated to generalization

Spectral Energy (Hjorth's activity)

- $\theta(t)$: parameter update at time t given the optimization algorithm
- $L(t)$: loss at $\theta(t)$
- $\mathcal{F}(L)$: Fourier transform of $L(t)$
- $m_n(L) = \int \omega^n \|\mathcal{F}(L)(\omega)\|^2 d\omega$: the n^{th} moment of $\mathcal{F}^2(L)$
- $\|\mathcal{F}(L)(\omega)\|^2$: energy spectral density present in the pulse ω
- Hjorth's activity $m_0(L)$: signal power, the surface of the power spectrum in the frequency domain.

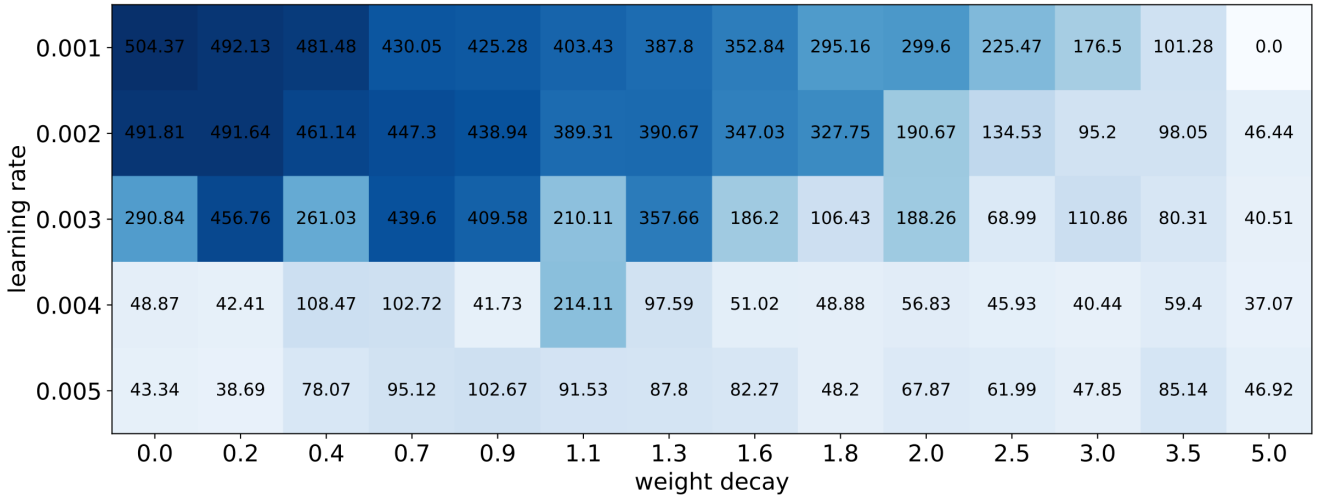


(a) Energy: 400 steps (train loss)

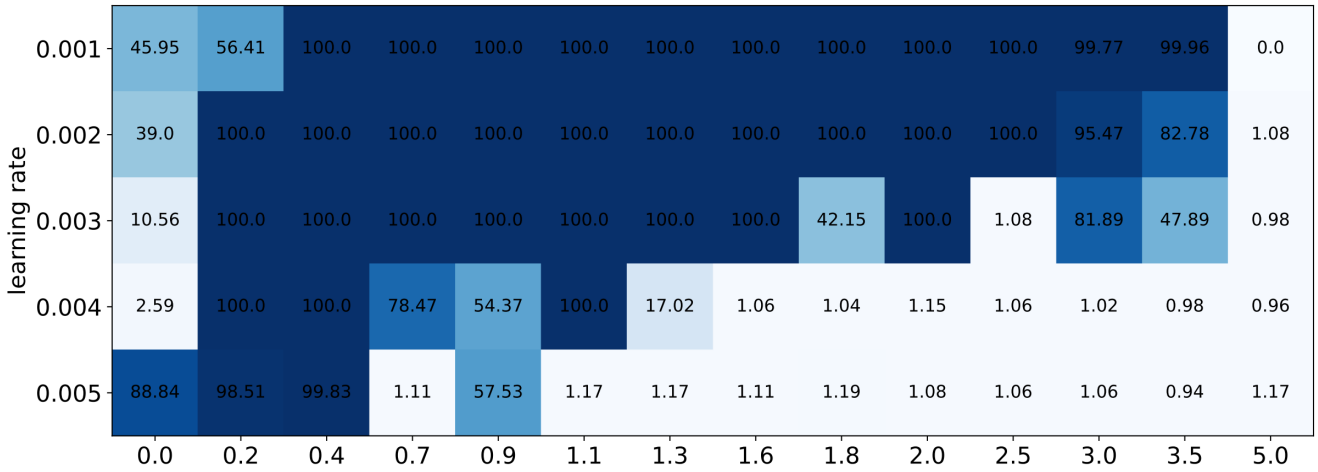


(b) Final test accuracy (10K steps)

Spectral Signature of the loss is correlated to generalization

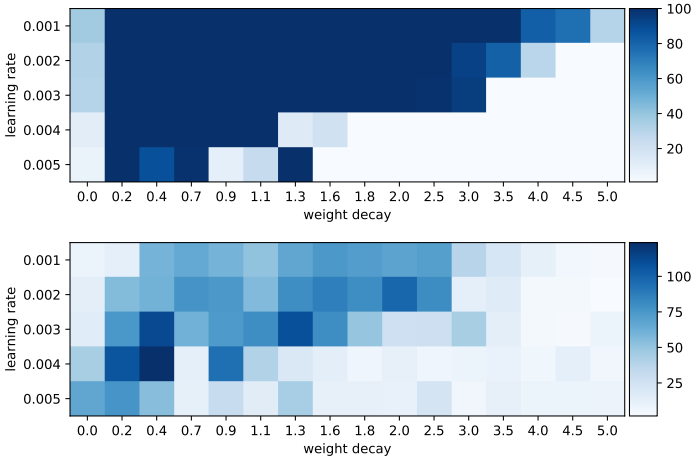


(a) Energy: 400 steps (train loss)

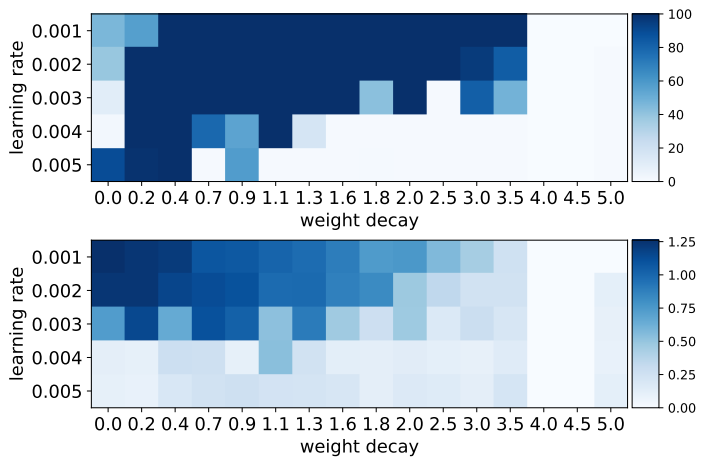


(b) Final test accuracy (10K steps)

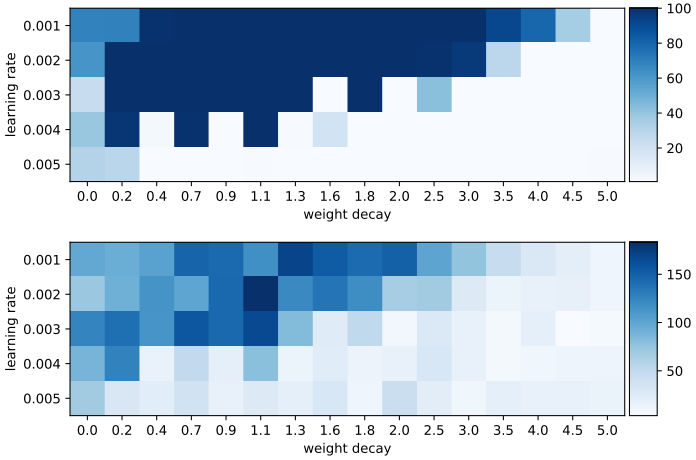
Spectral Signature of the loss is correlated to generalization



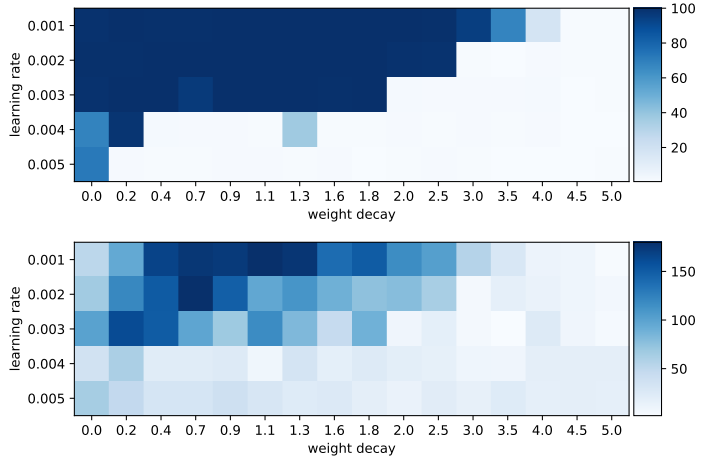
(a) $r = 0.4$



(b) $r = 0.5$



(c) $r = 0.7$



(d) $r = 0.9$

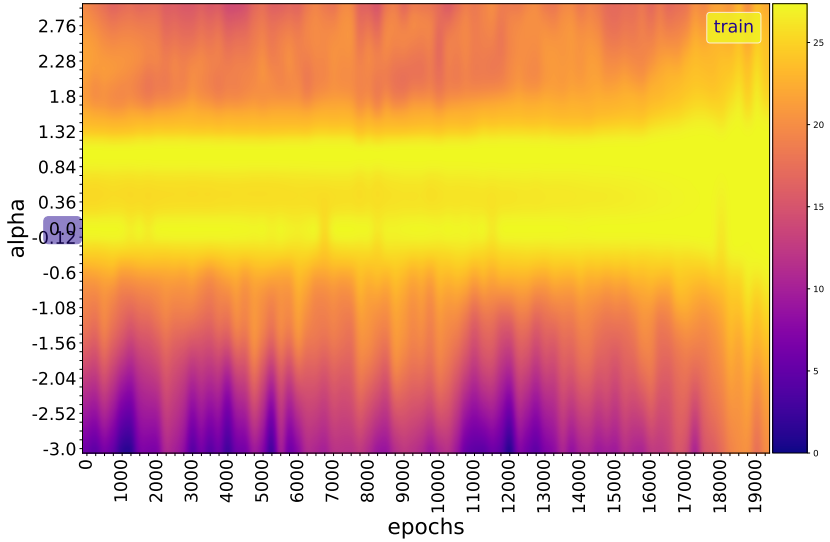
Advantages

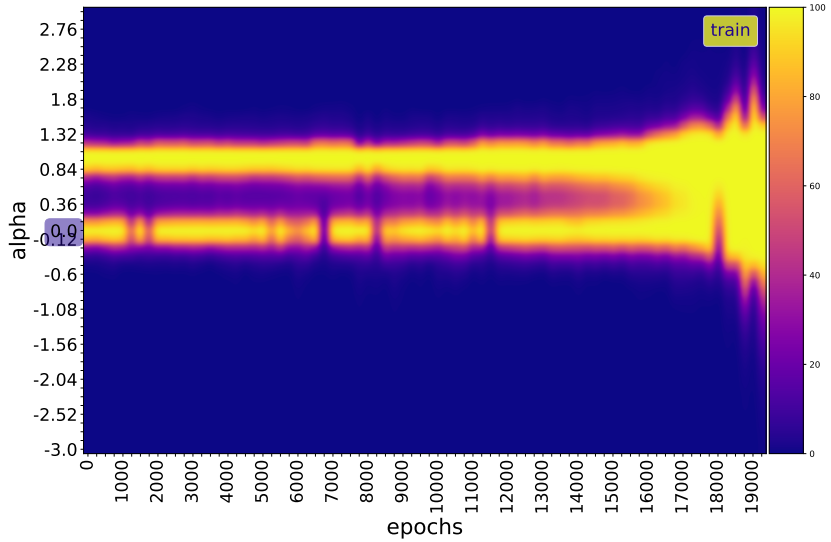
- The spectral signature can be classified as an optimization-based generalization measure, which is known to be highly predictive of generalization (**jiang2019fantastic**).
- Optimization-based generalization measurements are generally only made on the training dataset and only require the model to be trained for a small number of epochs, compared with sharpness-based measures which, as their name suggests, require the to train the moment until convergence.
- Such a measure is not an explicit capacity measure so either a positive or negative correlation with generalization could potentially be informative.

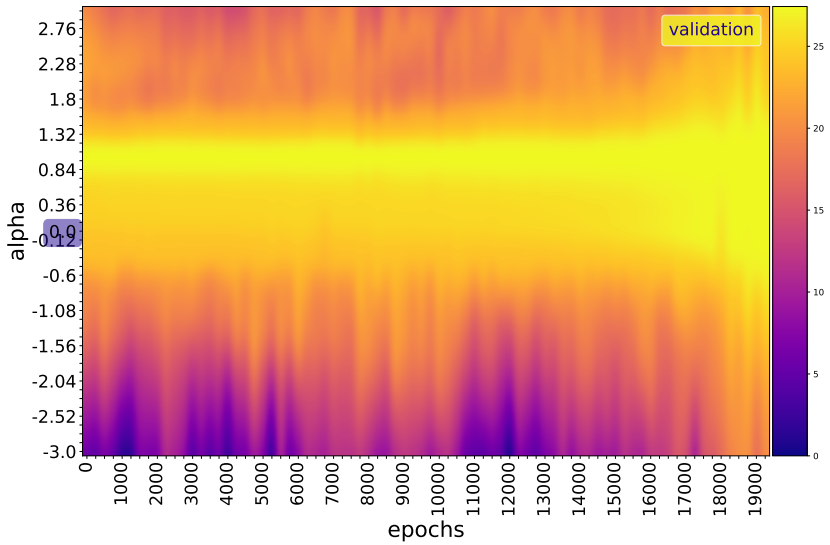
A result of a (random) walk in a valley of local solutions?

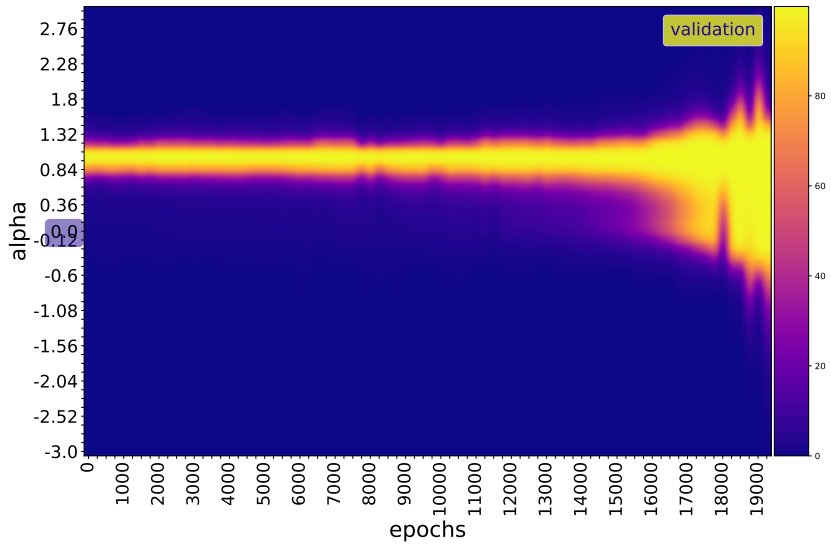
- A strong correlation between the frequency of oscillations in the loss and the final generalization performance of the model supports the fact that the learning behaviour of the model is tightly coupled with the training loss.
  But this doesn't give enough information about the behaviour of the model weights before and after grokking.
- Does the model, before grokking, oscillate around a local minimum, cross a very flat region, or circumvent a large obstacle?

The model crosses a perturbed valley of bad/local solutions before grokking. When the iterates fall in the valley, we are at the minimum for the training objective so that the model can memorize the training data, and it achieves grokking when it successfully breaks free from the basin of attraction of such solutions.





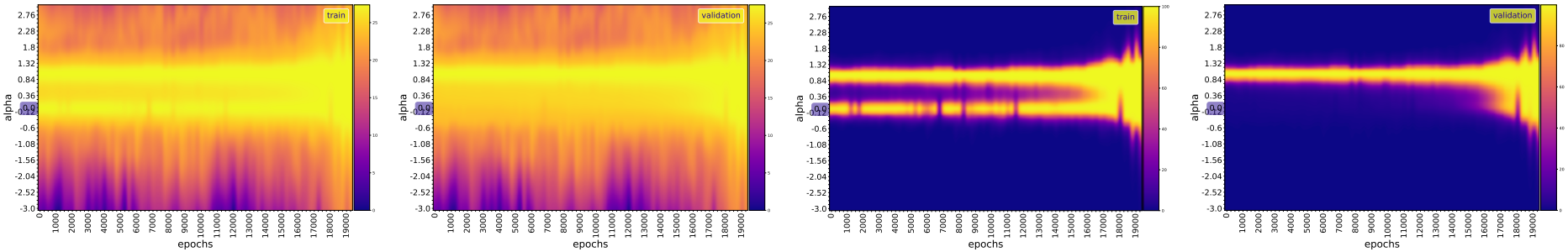




(a) Loss surface : $f_t(\alpha) = Loss(\theta_t + \alpha\vec{\delta}_t)$ (b) Accuracy : $f_t(\alpha) = Acc(\theta_t + \alpha\vec{\delta}_t)$

Figure: $r = 0.30$, $\vec{\delta}_t \propto \theta^* - \theta_t$

Grokking loss landscape



(a) Loss surface : $f_t(\alpha) = Loss(\theta_t + \alpha\vec{\delta}_t)$ (b) Accuracy : $f_t(\alpha) = Acc(\theta_t + \alpha\vec{\delta}_t)$

Figure: $r = 0.30$, $\vec{\delta}_t \propto \theta^* - \theta_t$

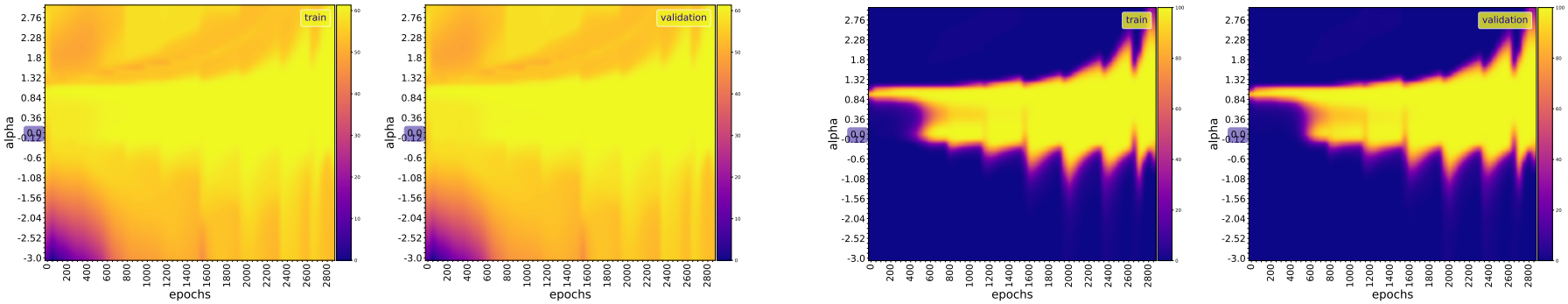


Figure: $r = 0.85$, $\vec{\delta}_t \propto \theta^* - \theta_t$

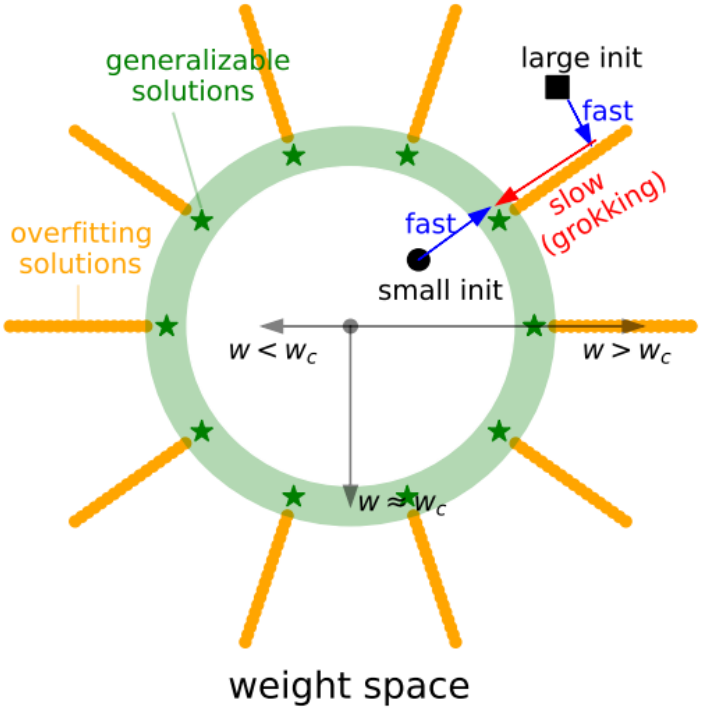
Others observations

- Larger condition numbers $\lambda_{max}/\lambda_{min}$ of the hessian of the grokking loss: leading to a slower convergence of gradient descent.
- The optimization dynamics is embedded in a low-dimensional space: more than 98% of the total variance in the parameter space occurs in the first 2 PCA modes much smaller than the total number of weights,
- The model remains in a lazy training regime most of the time: the cosine distance between the model weights from one training step to the next remains almost constant, except at the slingshot location.

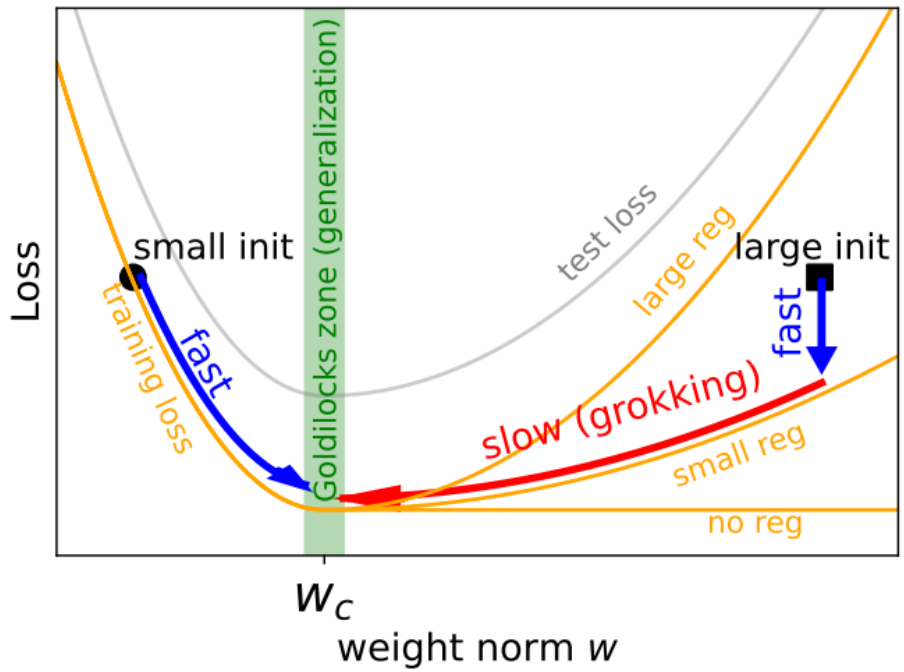
Under realistic hypotheses (Dziugaite et al., 2017) :

- SGD finds good solutions only if they are surrounded by a relatively large volume of solutions that are nearly as good
- SGD performs implicit regularization or tends to find solutions that possess some particular structural property that we already know to be connected to generalization, like widder minima

Related works : "LU mechanism" (Liu, Michaud, et al., 2023)



(a)



(b)

Figure 1: (a) w: L_2 norm of model weights. Generalizing solutions (green stars) are concentrated around a sphere in the weight space where $w \approx w_c$ (green). Overfitting solutions (orange) populate the $w \gtrsim w_c$ region. (b) The training loss (orange) and test loss (gray) have the shape of L and U, respectively. Their mismatch in the $w > w_c$ region leads to fast-slow dynamics, resulting in grokking.

Related works : Good Representation (Liu, Kitouni, et al., 2022)

- Generalization can be attributed to learning a good representation of the input embeddings
- The critical training set size corresponds to the least amount of training data that can determine such a representation

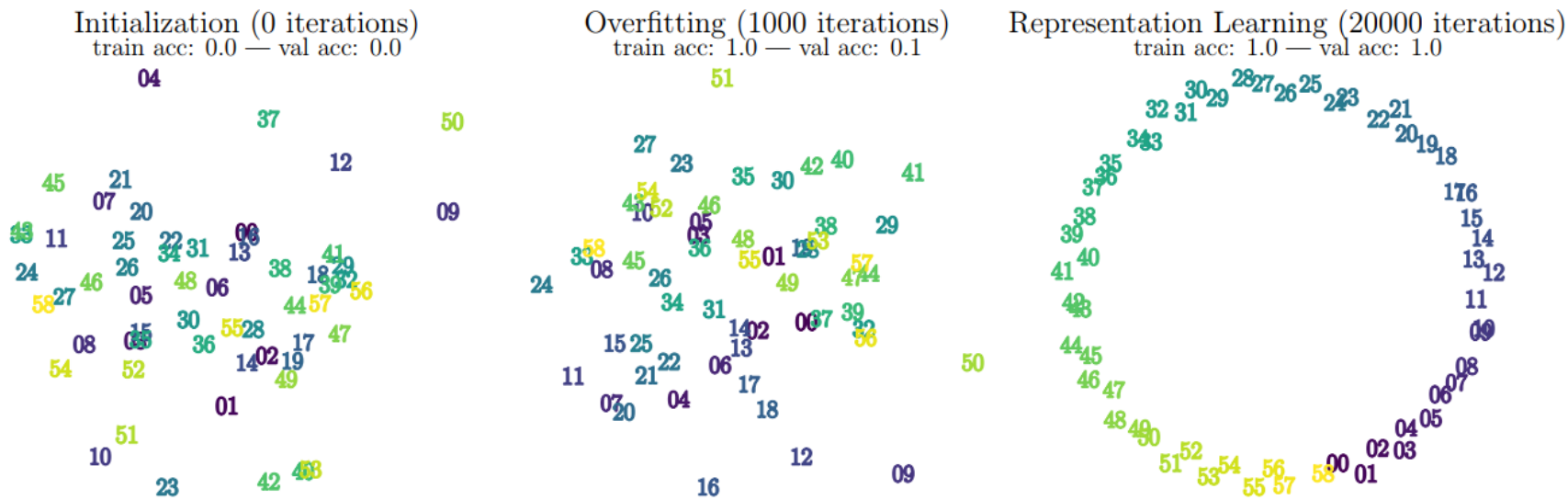
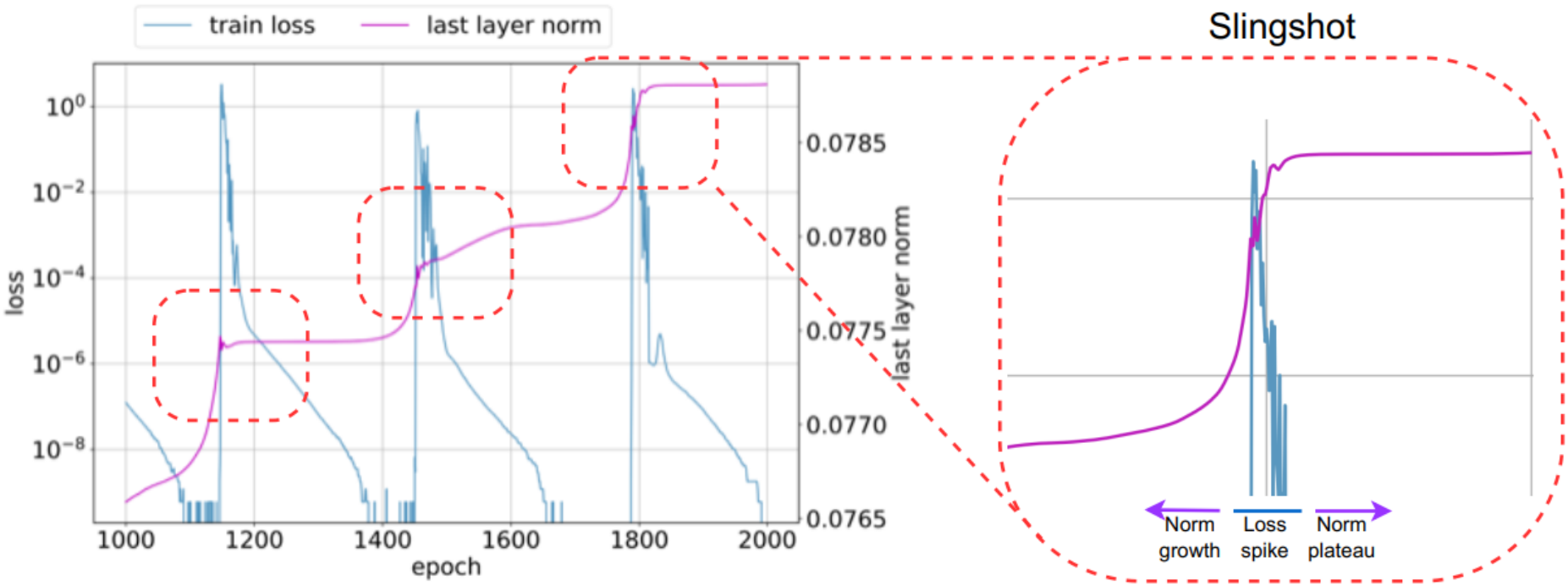


Figure 1: Visualization of the first two principal components of the learned input embeddings at different training stages of a transformer learning modular addition. We observe that generalization coincides with the emergence of structure in the embeddings. See Section 4.2 for the training details.

Related works: Slingshot mechanism (Thilak et al., 2022)



Why is it important to study such phenomena? *(Grokking, Double descent, etc)*

- Understanding all this behaviour and how they affect the predictive performance of neural networks (for example, at scale or out-of-distribution) is relevant to safety or may have potential safety consequences.
- We need to be certain of a model's safety before we scale it to a capability level beyond which we cannot control it
- This is particularly concerning because the out-of-distribution generalization behaviour of deep learning models is known to be challenging to control or foresee.

For more details, see our paper (Notsawo et al., 2023) or/and this blog post.

Thank You

For Your Attention!

Any Questions

[1] Gintare Karolina Dziugaite et al. "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In: ed. by Gal Elidan et al. AUAI Press, 2017. URL: http://auai.org/uai2017/proceedings/papers/173.pdf.

[2] Ziming Liu, Ouail Kitouni, et al. "Towards understanding grokking: An effective theory of representation learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34651–34663.

[3] Ziming Liu, Eric J Michaud, et al. "Omnigrok: Grokking Beyond Algorithmic Data". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=zDiHoIWa0q1.

[4] Pascal Jr. Tikeng Notsawo et al. "Predicting Grokking Long Before it Happens: A look into the loss landscape of models which grok". In: *arXiv preprint arXiv: 2306.13253* (2023).

[5] Alethea Power et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets". In: *arXiv preprint arXiv:2201.02177* (2022).

[6] Vimal Thilak et al. “The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the {Grokking Phenomenon}”. In: 2022.