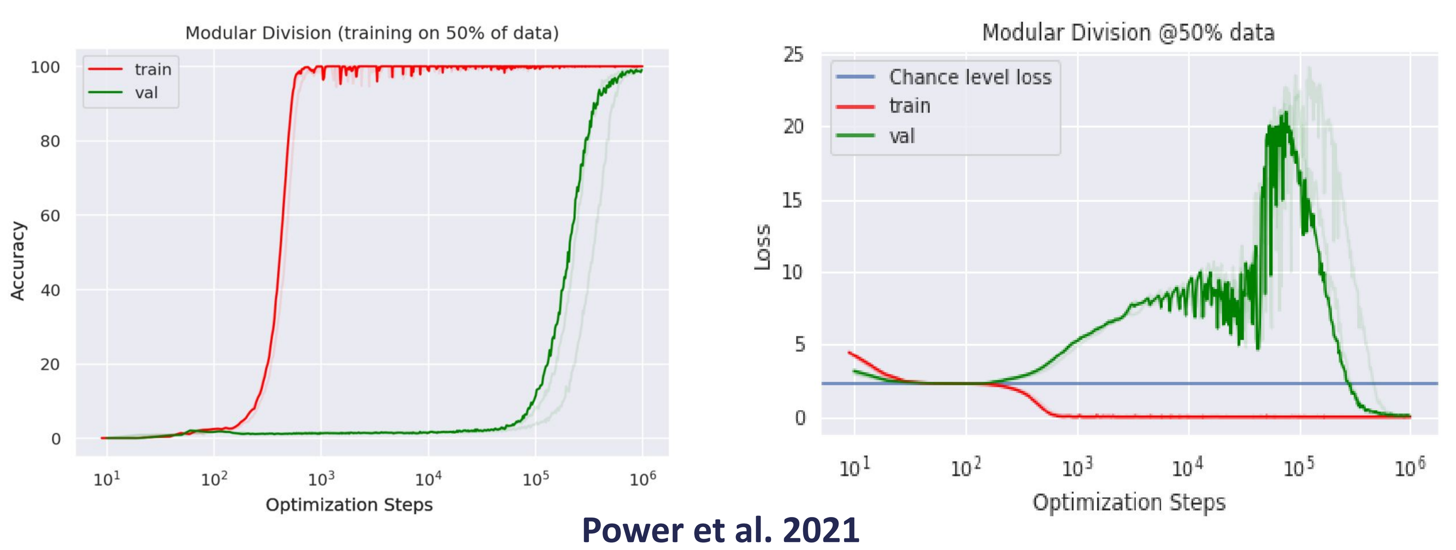
Grokking at the Edge of Linear Separability

Alon Beck* Noam Levi† Yohai Bar-Sinay‡

TL;DR

Grokking may occur even in standard logistic regression. This happens if the data is almost linearly separable. In this case, the model may overfit for an arbitrarily long time, before converging to the ground truth.

What is Grokking?



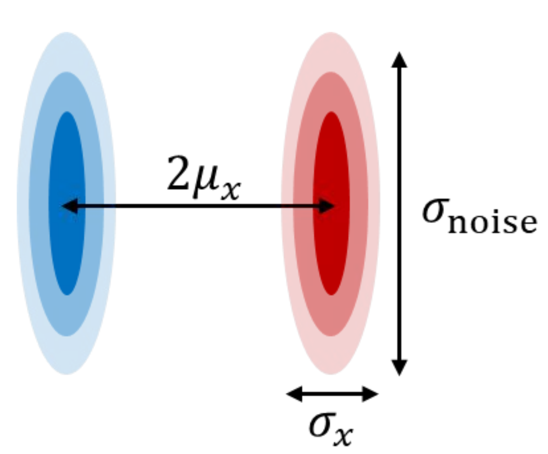
Power et al. 2021

Delayed generalization + non-monotonic test loss.
Counterintuitive! Overfitting usually implies never generalizing.

The goal: Understanding the mechanism in a **minimal** setup

Binary Classification Setup

- d-dimensional Gaussian data.
- Separation: x-axis.
- Noise in all other directions.
- $\lambda = d/N, N \to \infty$
- $\sigma_x \ll \mu_x, \mu_x \lesssim \sigma_{\text{noise}}$



Equivalent to a (d-1) dimensional model + bias:

$$\boldsymbol{x}_i \sim \mathcal{N}(0, \sigma \boldsymbol{I}_d), \sigma = \sigma_x^2/\mu_x^2 \gtrsim 1 \qquad \tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i)$$

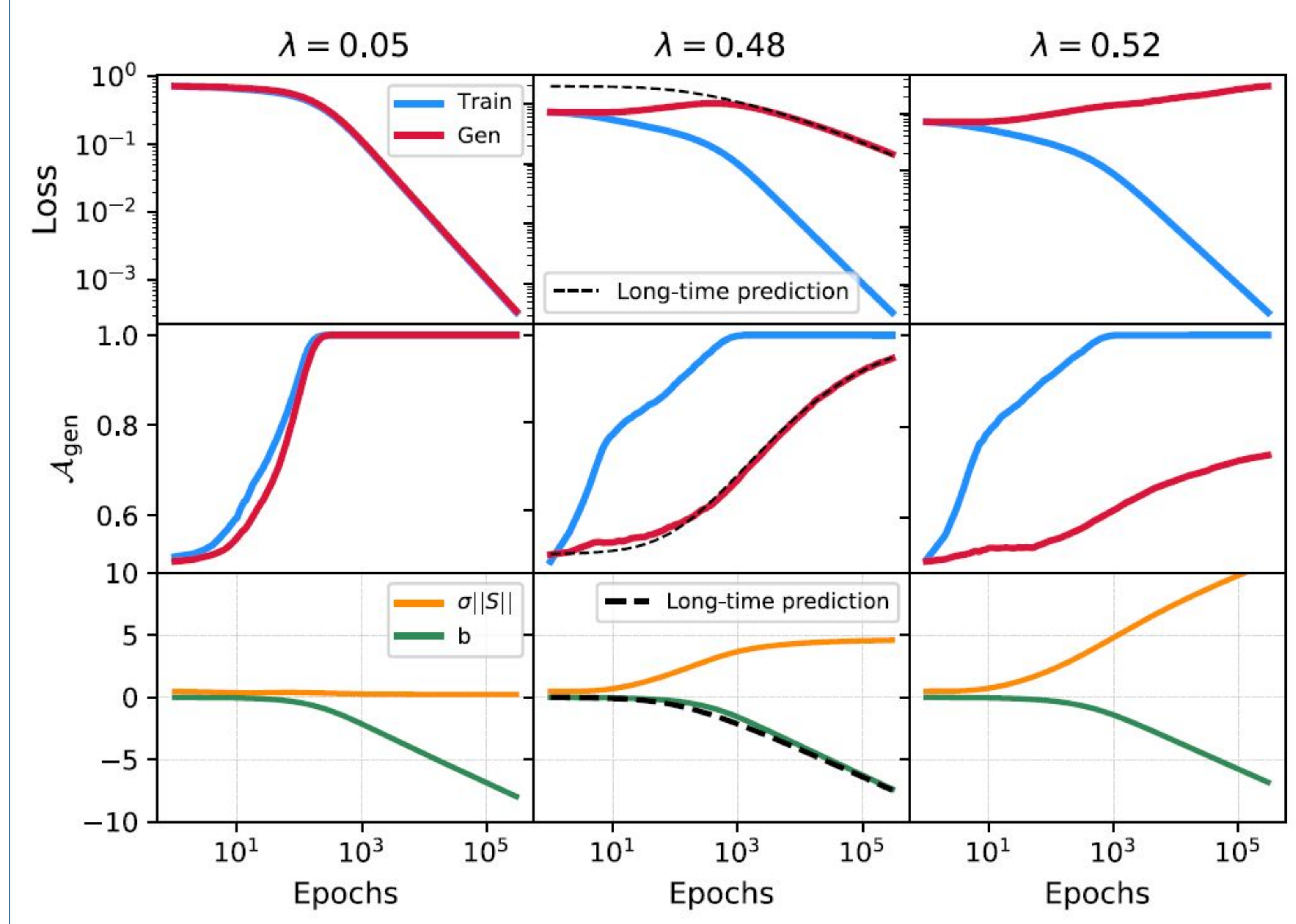
Same label for all points! (for example, -1)

Linear model: $f(\boldsymbol{x}_i) = \boldsymbol{S} \cdot \boldsymbol{x}_i + b, \boldsymbol{S} \in \mathbb{R}^{d-1}, b \in \mathbb{R}$

CE-loss: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i f(\boldsymbol{x}_i)}\right), \; y_i = -1$

Generalization: only if $b \to -\infty$

Empirical results (GD)



- Cannot generalize for $\lambda > 0.5$
- **Grokking** for $\lambda \to 0.5^-$

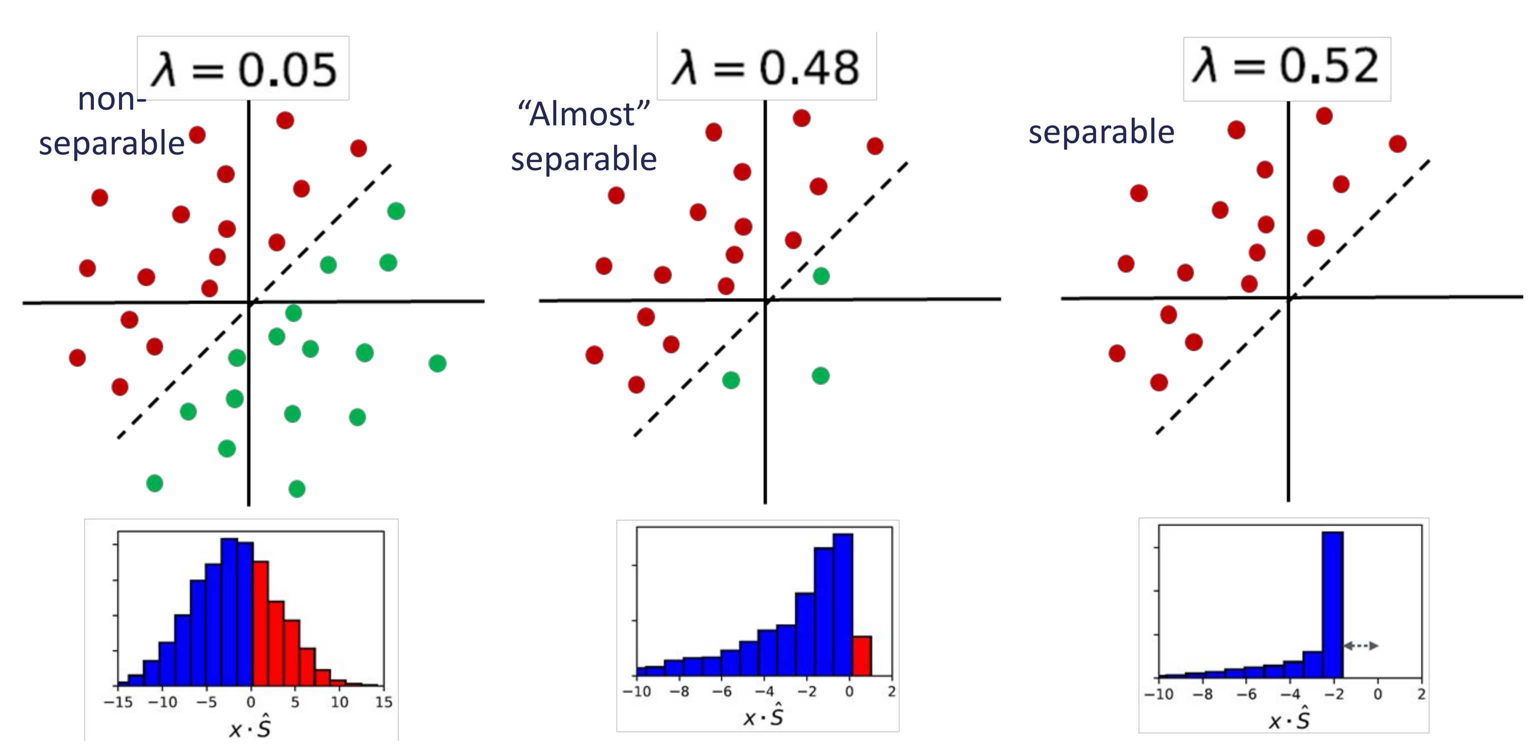
Edge of Linear Separability

$$A_{\text{gen}}(\boldsymbol{S}, b) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{1}{\sqrt{2}}\frac{b}{\sigma||\boldsymbol{S}||}\right)\right]$$

Grokking: first $||\boldsymbol{S}||$ increase (memorization), then saturates while $b \to -\infty$ (generalization).

What is special about $\lambda \approx 0.5^-$?

Theorem 1: Training data is linearly separable (from the origin) iff $\lambda > 0.5$
Proof: Wendel's theorem



Mechanism

Theorem 2: Generalization iff training data is linearly separable , in particular:

$\lambda < 0.5$: $||S_\infty|| = C$, **diverges** as $\lambda \to 0.5^-$

Proof: Exponential loss + "conformal time":

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} e^{\boldsymbol{S}^T \boldsymbol{x}_i + b} \qquad \frac{\partial \boldsymbol{S}}{\partial t} = -\frac{\eta}{N} e^{b} \sum_i e^{\boldsymbol{S}^T \boldsymbol{x}_i} \boldsymbol{x}_i$$

$$\tau(t) = \int_0^t e^{b(t')} dt' \qquad \frac{\partial \boldsymbol{S}}{\partial \tau} = -\frac{\eta}{N} \sum_i e^{\boldsymbol{S}^T \boldsymbol{x}_i} \boldsymbol{x}_i$$

Exponential loss without bias: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} e^{\boldsymbol{S}^T \boldsymbol{x}_i}$

$\lambda > 0.5$: $||S_\infty|| \to \infty$ $\quad A_{\text{gen}}^{\infty} = \frac{1}{2}\left[1 + \text{erf}\left(\frac{1}{\sigma M \sqrt{2}}\right)\right]$

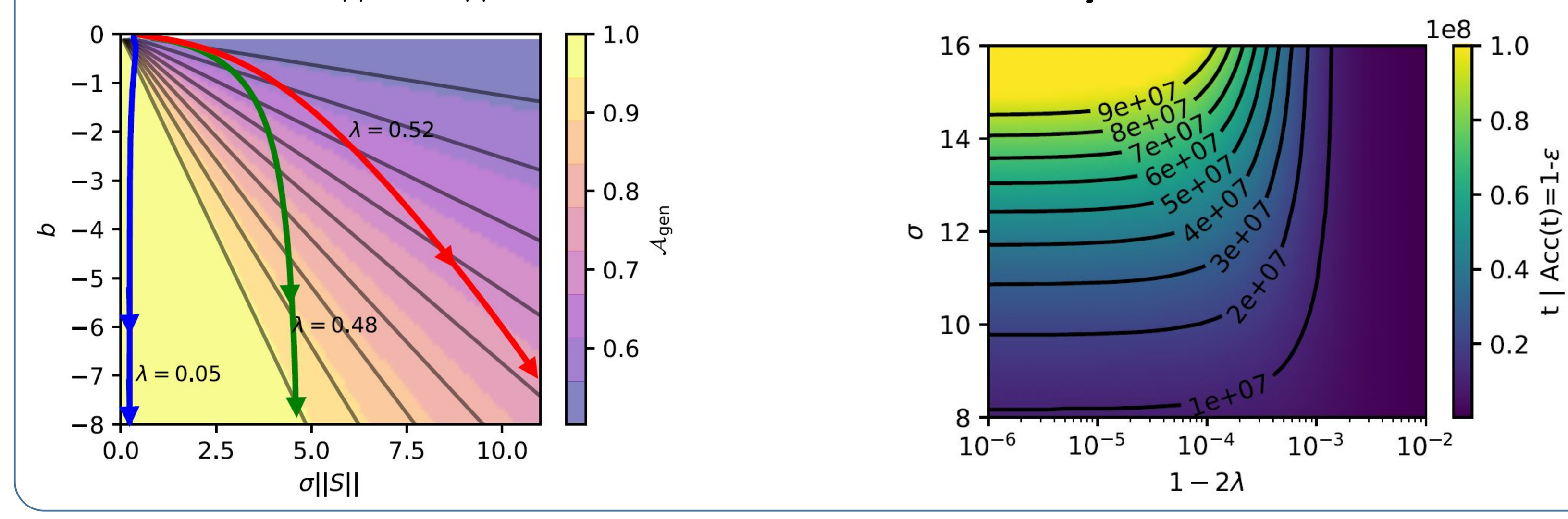
Proof: Soudry et al.

$$\boldsymbol{w}(t) = \boldsymbol{w}_{\text{SVM}} \log(t) + \boldsymbol{\rho}(t)$$

$$\boldsymbol{w}_{\text{SVM}} = \underset{(\boldsymbol{S},b)}{\text{argmin}} \left\{ \|\boldsymbol{S}\|^2 + b^2 \quad s.t. \quad \boldsymbol{S}^T \boldsymbol{x}_i + b \le -1 \right\}$$

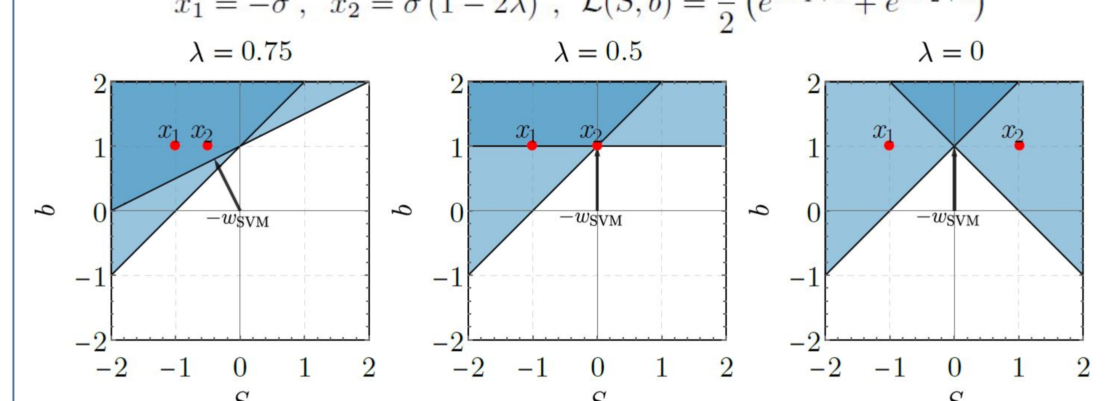
$$\boldsymbol{w}_{\text{SVM}} = \left(\frac{M}{1+M^2}\boldsymbol{S}_{\text{SVM}}, -\frac{1}{1+M^2}\right)$$

Collecting the pieces: For large enough σ , dynamics first reach $||S_\infty||$ while $b \to -\infty$ only follows later!

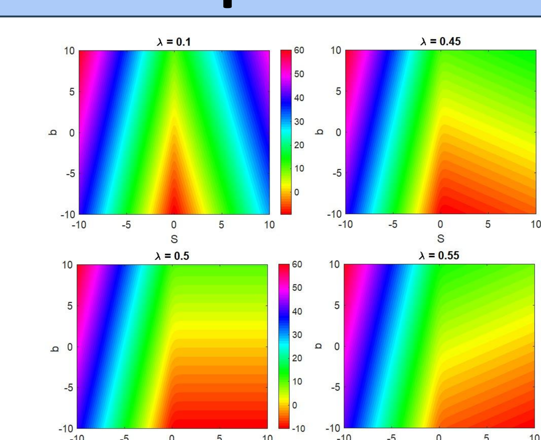


Effective 1d Solvable Description

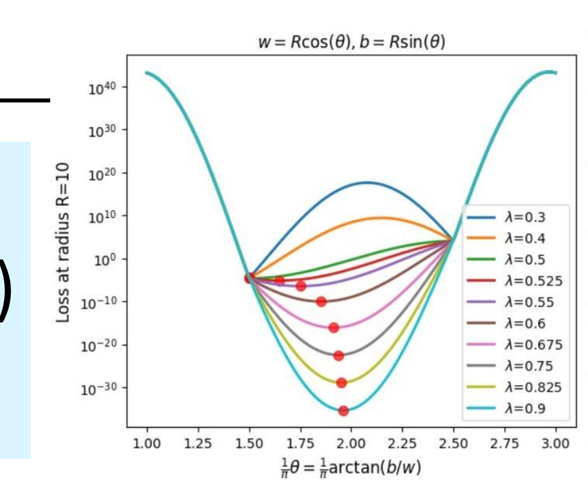
$$x_1 = -\sigma\,, \quad x_2 = \sigma(1 - 2\lambda)\,, \quad \mathcal{L}(S, b) = \frac{1}{2}\left(e^{Sx_1+b} + e^{Sx_2+b}\right)$$



	$\lambda < 1/2\ (x_2 > 0)$	$\lambda = 1/2\ (x_2 = 0)$	$\lambda > 1/2\ (x_2 < 0)$
$b(t \gg 1)$	$-\log(t)$	$-\log(t)$	$-\frac{1}{1+M^2}\log(t)$
$\|\boldsymbol{S}\|(t \gg 1)$	$\frac{1}{2(1-\lambda)}\log\left(\frac{1}{1-2\lambda}\right)$	$\log(\log(t))$	$\frac{M}{1+M^2}\log(t)$



Flat directions at $\lambda = 0.5$!



Similar to **critical phenomena** in physical systems (phase transitions)
Future work: Universality classes?

Check out our paper:

*alonbk2@gmail.com
†noam.levi@epfl.ch
‡ybarsinai@gmail.com