Learning theory as physics

S.V. Kozyrev

Steklov Mathematical Institute

The second law of thermodynamics and Eyring's formula of kinetics in learning

Learning by the stochastic gradient Langevin dynamics (SGLD)

Grokking (delayed generalization) is discussed as Brownian motion

Gradient descent — numerical solution of the differential equation

$$\frac{dx}{dt} = -\nabla f(x),$$

the trajectory goes to a local minimum f.
With numerical iteration of the descent, the vector x will change as

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

The stochastic gradient Langevin dynamics, SGLD

$$x_{k+1} = x_k + w_k - \alpha_k \nabla f(x_k),$$

w_k – independent mean zero gaussian random vectors.
The stochastic differential equation (SDE)

$$d\xi^i(t) = \sqrt{2\theta} dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i} dt, \tag{1}$$

$dw^i(t)$ stochastic differential of a Wiener process

$$dw^i(t) dw^j(t) = \delta_{ij} dt.$$

G. Parisi, Correlation functions and computer simulations, Nuclear Physics B, 180(3), 378–384 (1981).

G. Parisi, Correlation functions and computer simulations II, Nuclear Physics B, 205 (3), 337–344 (1982).

Random Walk

$$d\xi^i(t) = \sqrt{2\theta} dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i} dt.$$

Fokker–Planck equation — diffusion in potential

$$\frac{\partial u}{\partial t} = \theta \Delta u + \nabla u \cdot \nabla f + u \Delta f, \tag{2}$$

where $x \in \mathbb{R}^d$, $u = u(x,t)$ is the distribution, $f = f(x)$ is the potential, $f \in C^2(\mathbb{R}^d)$, $\theta > 0$ is the temperature.

$$\frac{\partial u}{\partial t} = \theta \, \mathbf{div} \left[e^{-\beta f} \, \mathbf{grad} \left[u e^{\beta f} \right] \right],$$

Gibbs distribution $e^{-\beta f}$, $\beta = 1/\theta$ is a stationary solution.

Chemical kinetics

The Eyring formula of kinetic theory describes the rate of reaction (transition between two potential wells due to diffusion (2)): the reaction rate is proportional to

$$e^{-\beta(F_1-F_0)}, \tag{3}$$

where F_1 is the free energy of the transition state (the saddle area between two potential wells) and F_0 is the free energy of the initial state of the reaction (the potential well from which the transition occurs).

The free energy of a state is $F = E - \theta S$, where E is the energy and S is the entropy of the state, in general

$$e^{-\beta F(U)} = \int_U e^{-\beta E(x)} dx.$$

Learning problem

Let a training sample $\{z_l\}$, $l = 1, \ldots, L$ and a loss function $\mathcal{L}(z, x) \geq 0$ be given for test z and hypothesis x (let the hypothesis space be $\mathbb{R}^d$). Empirical risk minimization is a problem

$$f(\{z\}, x) = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}(z_l, x) \;\to\; \min_x . \tag{4}$$

Overfitting is the lack of ability to generalize to a learning problem (4) when the sample $\{z\}$ is replaced. Let the hypothesis x_0 provide a minimum of the functional $f(\{z\}, x)$. **Overfitting**: low risk on the training sample $f(\{z\}, x_0)$, high risk $f(\{z'\}, x_0)$ for the control sample $\{z'\}$.

Narrow (sharp) minima of empirical risk (in the hypothesis space) are associated with overfitting, and wide (flat) minima correspond to solutions with generalization.

S.Hochreiter, J.Schmidhuber, Flat Minima, Neural Computation 9, 1–42 (1997).

Stochastic Gradient Descent

The learning problem (4)

$$f(\{z\}, x) = \frac{1}{L}\sum_{l=1}^{L}\mathcal{L}(z_l, x) \to \min_x;$$

SGLD (1) in the hypothesis space

$$d\xi^i(t) = \sqrt{2\theta}dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i}dt,$$

and the diffusion in a potential (2)

$$\frac{\partial u}{\partial t} = \theta\Delta u + \nabla u \cdot \nabla f + u\Delta f.$$

The distribution $u(x)$ converges to a Gibbs distribution concentrated in wells with low free energy.

Free energy of a well is a combination of depth and width

$$F = E - \theta S.$$

Eyring's formula (3) for the reaction rate

$$e^{-\beta(F_1 - F_0)}$$

predicts capture of the SGLD learning result by wide potential wells, i.e. the overfitting reduction.

S. V. Kozyrev, I. A. Lopatin, A. N. Pechen, Control of Overfitting with Physics, Entropy, 26, 1090 (2024). arXiv: 2412.10716

Grokking (delayed generalization)

[1] *A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra, Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, arXiv:2201.02177 (2022)*

Overparameterized model, algoritmic dataset, delayed generalization.

Modular arithmetics (in particular addition modulo p) — maximal size of the learning sample is p^2.

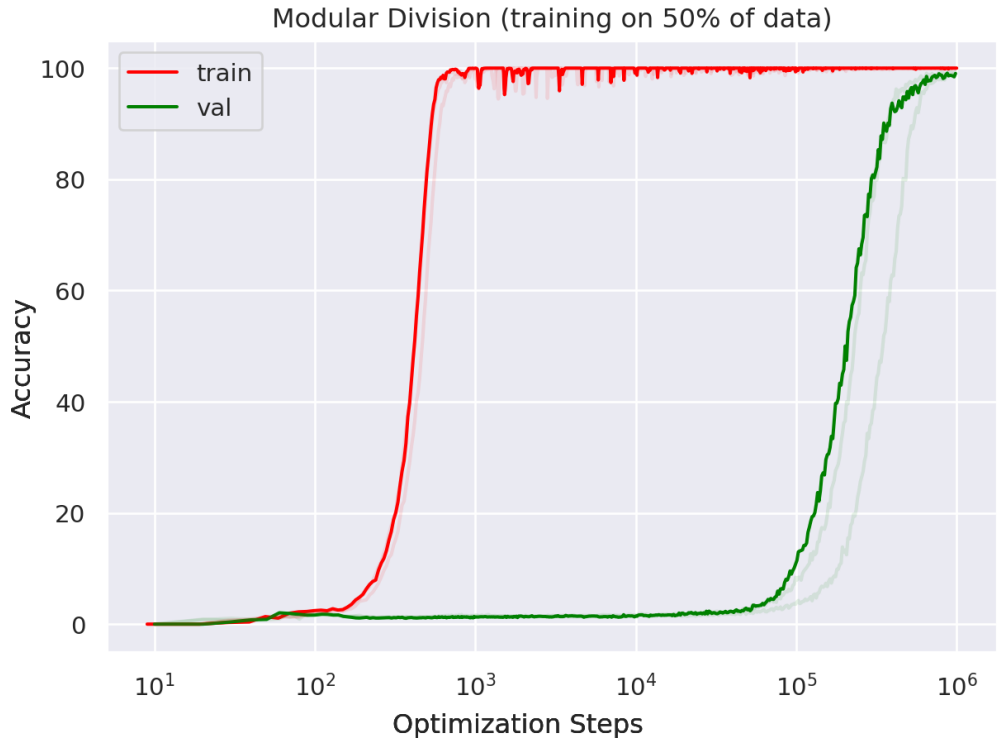

Figure: 1. Grokking: A dramatic example of generalization far after overfitting on an algorithmic dataset. The red curves show training accuracy and the green ones show validation accuracy. Training accuracy becomes close to perfect at 10^3 optimization steps, but it takes close to 10^6 steps for validation accuracy to reach that level, and we see very little evidence of any generalization until 10^5 steps.

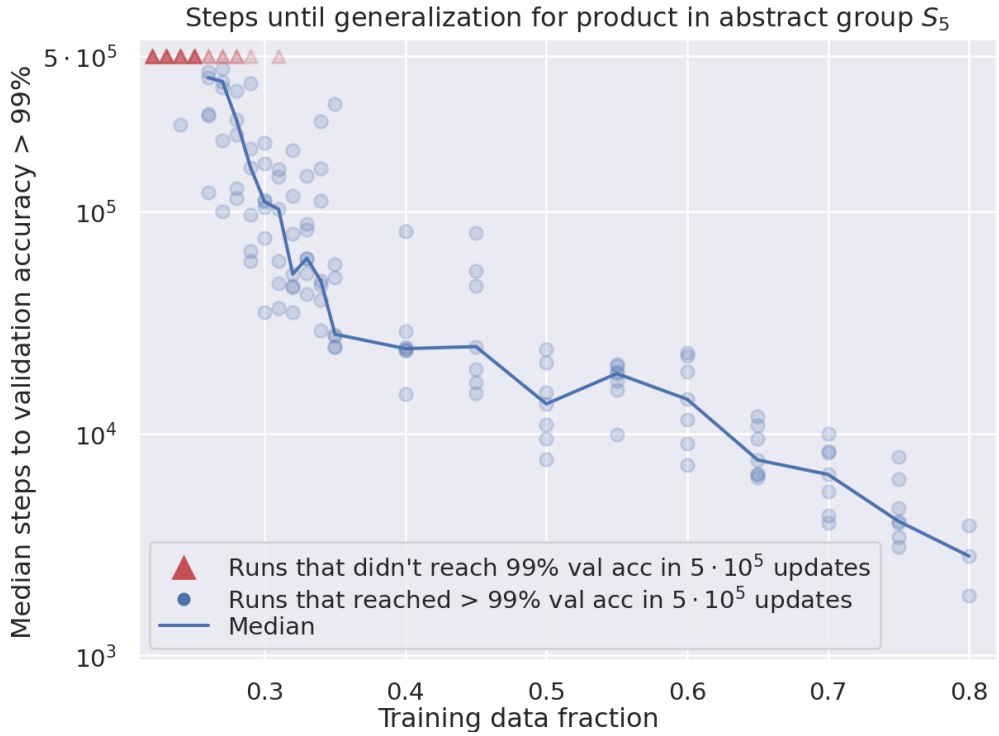

Figure: 2. Training time required to reach 99% validation accuracy increases rapidly as the training data fraction decreases.

Exponential growth of grokking time with decreasing training sample (logarithmic coordinate scale).

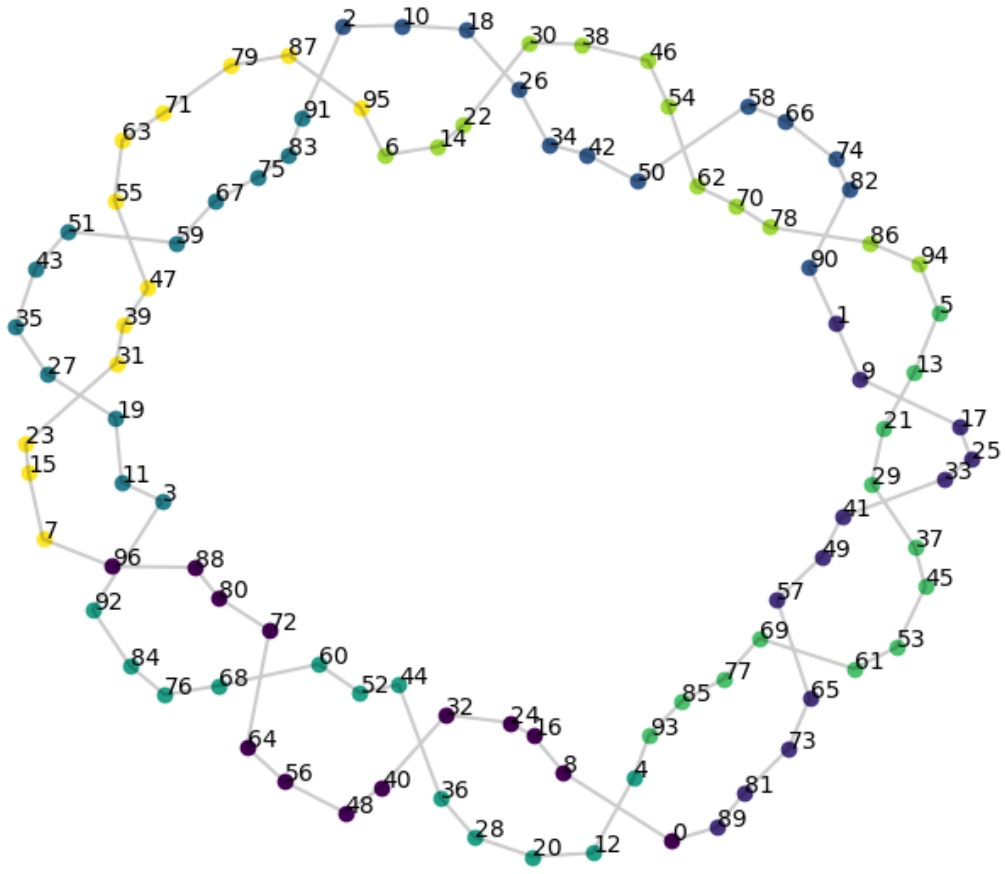

Figure: 3. t-SNE projection of the output layer weights from a network trained on modular addition. The lines show the result of adding 8 to each element. The colors show the residue of each element modulo 8.

Embeddings for residues for modular addition approximately lie of a circle, addition is a shift along such a circle.

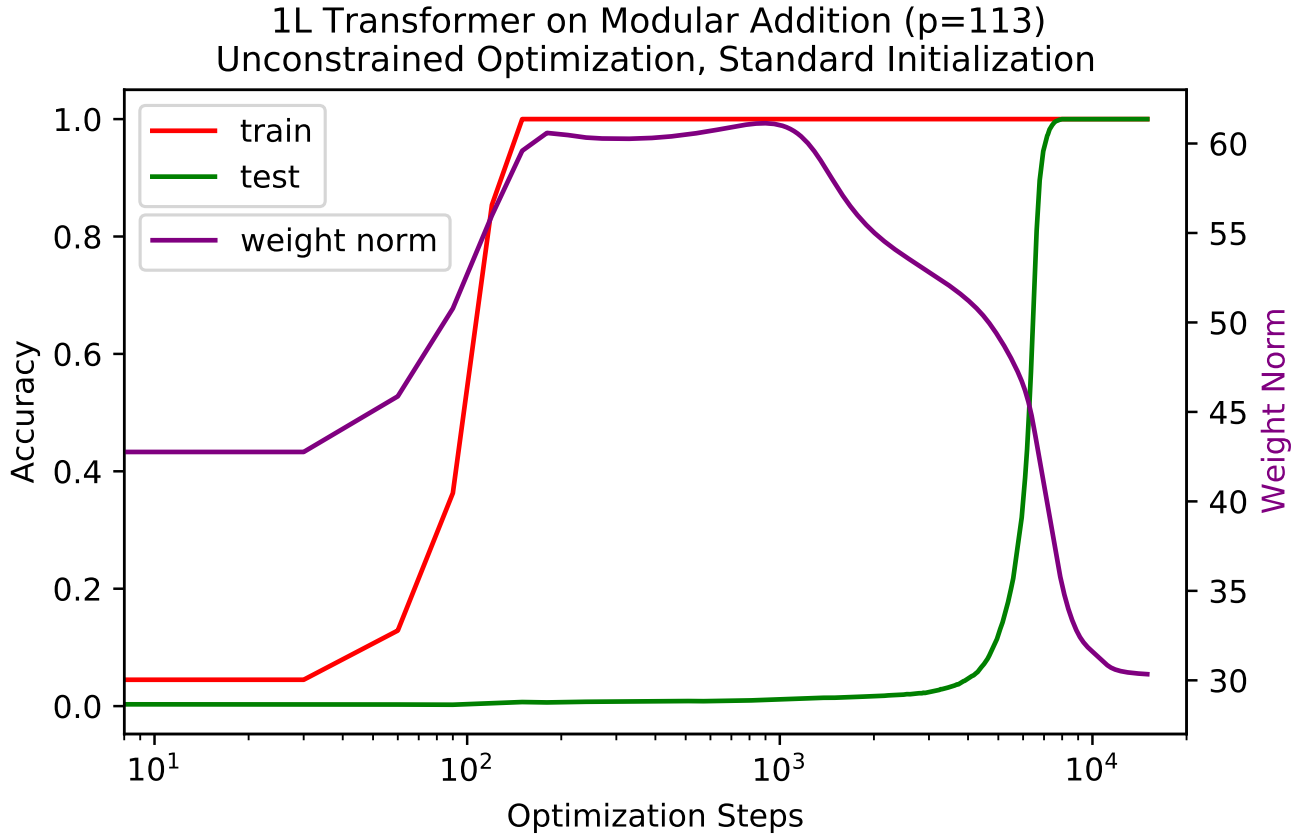

Figure: 4. Weight norm is highly correlated with overfitting and generalization. When overfitting happens, the weight norm increases; when generalization happens, the weight norm decreases.

Ziming Liu, Eric J. Michaud, Max Tegmark, Omnigrok: grokking beyond algorithmic data, Conference paper at ICLR 2023, arXiv:2210.01117

Grokking: Overparameterized networks, delayed generalization, and "structure" formation.
Overparameterized models (the dimension of the parameter space is large) — local minima merge into a region (manifold) of zero risk in the hypothesis space

$$f(\{z\}, x) = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}(z_l, x), \quad \mathcal{L}(z, x) \geq 0.$$

Zero risk manifold is the intersection of $\mathcal{L}(z, x) = 0$ over the sample $\{z\}$

$$f(\{z\}, x) = 0, \quad \text{i.e.} \quad \mathcal{L}(z, x) = 0, \, \forall z \in \{z\}.$$

The zero-risk manifold contains narrows, or ravines (areas with low entropy) and wide valleys (with high entropy).
The correct solution ("structure") most likely lies in the high-entropy region.

Explanation of Fig. 1

Sample memorization — reaching the zero-risk region $f(\{z\}, x) = 0$ by stochastic gradient descent with non-zero gradient (drift $\sim t$).
Grokking — random walk in the region of zero risk $f(\{z\}, x) = 0$ (Brownian motion $\sim \sqrt{t}$).

The second law of thermodynamics explains the transition to regions with high entropy containing the solution of the learning problem (transition from narrows to wide valleys).

It is notable that the number of stochastic gradient steps required for grokking is the square of the memorization time (10^3 and 10^6). During the memorization and grokking, close distances are covered in the hypothesis space, but in different modes (drift $\sim t$, Brownian motion $\sim \sqrt{t}$).

Explanation of Fig. 2

As the training sample $\{z\}$ grows, the zero-risk manifold $f(\{z\}, x) = 0$ shrinks — additional conditions $\mathcal{L}(z, x) = 0$ for $z \in \{z\}$ are imposed.

A solution in the form of an algorithm ("structure") exists for any training sample $\Rightarrow$ as the training sample grows, the zero-risk manifold shrinks toward a region with high entropy containing the desired solution in the form of an algorithm.

Let us assume: imposing each additional condition $\mathcal{L}(z, x) = 0$, $z \in \{z\}$ with increasing sample size removes an equal percentage of the volume $\Rightarrow$ the entropy of the zero-risk region decreases linearly with increasing sample size.

Grokking: transition from the initial region (part of the zero-risk manifold where the Brownian motion of grokking begins) to the grokking region (a valley with high entropy, the neighborhood of the solution of the learning problem in the form of a "structure").

Eyring's formula $e^{-\beta(F_1-F_0)}$ gives an approximation of the dependence of the reciprocal of the grokking time on the sample size. Here F_0 and F_1 are the free energies of the initial region of grokking and the transition state (the neighborhood of the solution).

The entropy of the initial region decreases linearly with increasing sample size (free energy $F_0 = E_0 - \theta S_0$ increases linearly). Free energy F_1 of the transition region changes slightly with increasing sample size. Consequently, the grokking time decreases exponentially, see Fig. 2.

S. V. Kozyrev, How to explain grokking, arXiv:2412.18624

Summary

Learning by SGLD, search for a potential well with low free energy

Eyring's formula of kinetic theory — capture of a system by wide potential wells

Reduction of overfitting in the wide minima approach

Grokking (delayed generalization) as a Brownian motion

Transition to generalization by the second law of thermodynamics

Quadratic relationship between memorization and grokking times (drift $\sim t$, Brownian motion $\sim \sqrt{t}$)

Exponential dependence of grokking time on sample size — decrease in entropy of the zero-risk manifold with increasing sample size

Learning as a quasi-physical dynamics