Training Dynamics

Guillaume Corlouer
Stormglass

September 8, 2026

Video: https://www.youtube.com/watch?v=IHHrJdZKICs

Video: https://www.youtube.com/watch?v=clTLzFD2uC8

What you'll learn

Motivation

- Know about some big open questions in learning dynamics
- Understand the concept of implicit regularization
- Know about the different approaches to studying learning dynamics
- Understand the AI safety motivations for learning dynamics
- Know about emergent misalignment as a safety-relevant phenomenon that illustrates the importance of understanding generalization for AI safety

Geometry and dynamics of deep neural networks

- Know key results about the loss landscape of deep linear networks (DLNs): critical points are saddles or global minima
- Explain the edge of stability phenomenon
- Understand that gradient flow can be written as NTK-weighted gradient in function space
- DLNs are degenerate and have conserved quantities through gradient flow

Regimes of learning in deep linear networks

- Understand the role of initialization, width and depth for the lazy and rich (saddle to saddle) regimes in DLNs
- Know about implicit regularization from non-linearity (neural race reduction)

Use LLMs to help you out when you spend longer than the suggested time (but make sure that you understand). Make sure you keep at least 30 minutes to do problem 3.

References: [SMG14], [AMG24], [TAJ24].

1 Module Intent

The goal of this day is to present a toy model perspective on the training dynamics of deep neural networks. The student should learn about the AI safety motivations of studying training dynamics. One motivation is that understanding the implicit biases of training deep neural networks is a central problem behind AI alignment. This is because implicit biases influence the generalization behaviour of a neural network including for example if that neural network will generalize in a helpful or harmful way out of distribution. Another motivation is interpretability. One key lesson of the day is that SGD learns from data in a structured way. For example in deep linear neural networks, under small initialization, it will first learn the features that explain the largest fraction of data covariance.

2 Setup and notation

We study deep linear networks (DLNs) with L weight matrices $W_1 \in \mathbb{R}^{d_1 \times d_0}, W_2 \in \mathbb{R}^{d_2 \times d_1}, \ldots, W_L \in \mathbb{R}^{d_L \times d_{L-1}}$. The network computes:

$$f(x) = W_L W_{L-1} \cdots W_1 \, x =: Wx$$

where $W = W_L \cdots W_1 \in \mathbb{R}^{d_L \times d_0}$ is the end-to-end (or "student") matrix. We train on a dataset $\{(x_\mu, y_\mu)\}_{\mu=1}^N$ with the squared loss:

$$\mathcal{L}(\theta) = \frac{1}{2N} \sum_{\mu=1}^{N} \|y_\mu - f(x_\mu)\|^2$$

In the population limit $N \to \infty$ with whitened inputs $\Sigma_X = I$, this becomes, up to an additive constant[1]:

$$\mathcal{L}(W) = \frac{1}{2} \|M - W\|_F^2$$

where $M = \Sigma_{YX} \Sigma_X^{-1} = \Sigma_{YX}$ is the "teacher" matrix (the OLS solution) with SVD $M = U \, \mathrm{diag}(s_1, \ldots, s_r, 0, \ldots, 0) \, V^\top$, and $s_1 \geq s_2 \geq \cdots \geq s_r > 0$.

3 Loss landscape geometry

This problem explores the critical point structure of deep linear networks, following [AMG24].

[1] The constant is $\frac{1}{2}\mathbb{E}\|y - Mx\|^2$, which is the residual of the OLS. It vanishes on realizable data.

Exercise 3.1 (Diagonal decomposition). Consider the **diagonal** case: $d_0 = d_1 = \cdots = d_L = d$, and the teacher is diagonal, $M = \mathrm{diag}(s_1, \ldots, s_d)$, with $s_1 > s_2 > \cdots > s_d > 0$. Restrict attention to diagonal weight matrices $W_l = \mathrm{diag}(w_l^{(1)}, \ldots, w_l^{(d)})$. Show that the loss decomposes into d independent scalar problems:

$$\mathcal{L} = \frac{1}{2} \sum_{\alpha=1}^{d} \left(s_\alpha - \prod_{l=1}^{L} w_l^{(\alpha)} \right)^2$$

Solution to Exercise 3.1

When all W_l are diagonal, the product $W = W_L \cdots W_1$ is also diagonal with entries $(W)_{\alpha\alpha} = \prod_{l=1}^{L} w_l^{(\alpha)}$. The teacher M is diagonal with entries s_α. Then:

$$\mathcal{L} = \frac{1}{2}\|M - W\|_F^2 = \frac{1}{2} \sum_{\alpha=1}^{d} (s_\alpha - (W)_{\alpha\alpha})^2 = \frac{1}{2} \sum_{\alpha=1}^{d} \left(s_\alpha - \prod_{l=1}^{L} w_l^{(\alpha)} \right)^2$$

Since the different modes α share no parameters, the loss decomposes into d independent scalar problems. □

Exercise 3.2 (Scalar critical points). For a single scalar mode with target $s > 0$, find all first-order critical points of $\ell(w_1, w_2) = \frac{1}{2}(s - w_1 w_2)^2$ at depth $L = 2$. Show that the critical points are:

- The global minimum manifold: $w_1 w_2 = s$
- The origin: $w_1 = w_2 = 0$

Classify the origin as a saddle point by computing the Hessian H of ℓ at $(0, 0)$ and showing it has both positive and negative eigenvalues.
Hint: Write the gradient equations.

Solution to Exercise 3.2

For $L = 2$, the gradient equations are:

$$\frac{\partial \ell}{\partial w_1} = -(s - w_1 w_2)\, w_2 = 0, \qquad \frac{\partial \ell}{\partial w_2} = -(s - w_1 w_2)\, w_1 = 0$$

Case 1: $s - w_1 w_2 = 0$, i.e. $w_1 w_2 = s$. This is the global minimum manifold (a hyperbola $w_2 = s/w_1$) with $\ell = 0$.
Case 2: If $s - w_1 w_2 \neq 0$, then the first equation requires $w_2 = 0$ and the second requires $w_1 = 0$. (For instance, if $w_2 \neq 0$ and $w_1 = 0$, the first equation gives $-s\, w_2 = 0$, which contradicts $s > 0$ and $w_2 \neq 0$.) So the only other critical point

is $w_1 = w_2 = 0$.
The Hessian at $(0,0)$. We need the second derivatives of $\ell = \frac{1}{2}(s - w_1 w_2)^2$:

$$\frac{\partial^2 \ell}{\partial w_1^2} = w_2^2, \qquad \frac{\partial^2 \ell}{\partial w_2^2} = w_1^2,$$
$$\frac{\partial^2 \ell}{\partial w_1 \partial w_2} = -(s - w_1 w_2) + w_1 w_2 = 2w_1 w_2 - s$$

At $(0,0)$:

$$H = \begin{pmatrix} 0 & -s \\ -s & 0 \end{pmatrix}$$

The eigenvalues are $\pm s$. Since $s > 0$, H has one positive and one negative eigenvalue. The origin is a **strict saddle point**. □

Exercise 3.3 (Critical-point structure). Achour et al. [AMG24] show that every first-order critical point $\theta = (W_1, \ldots, W_L)$ satisfies: there exists a subset $S \subseteq \{1, \ldots, d\}$ such that

$$W = W_L \cdots W_1 = P_S M$$

where $P_S = U_S U_S^\top$ is the orthogonal projector onto the span of the left singular vectors $\{u_\alpha\}_{\alpha \in S}$ of M. For the diagonal case $M = \mathrm{diag}(s_1, \ldots, s_d)$, write the value of the loss at the critical point corresponding to $S = \{1, \ldots, r\}$ with $r < d$.

Solution to Exercise 3.3

At the critical point $W = P_S M$ with $S = \{1, \ldots, r\}$, the student retains only the first r modes: $W = \mathrm{diag}(s_1, \ldots, s_r, 0, \ldots, 0)$. The loss is:

$$\mathcal{L} = \frac{1}{2}\|M - P_S M\|_F^2 = \frac{1}{2} \sum_{\alpha \notin S} s_\alpha^2 = \frac{1}{2} \sum_{\alpha = r+1}^{d} s_\alpha^2$$

This is strictly positive whenever $r < d$ (since all $s_\alpha > 0$), so these are not global minima. By the Hessian analysis (extending Exercise 3.2), the direction corresponding to "switching on" a missing mode $\alpha \notin S$ is a descent direction, making these critical points saddle points. All critical points with $|S| = d$ have $W = M$ and $\mathcal{L} = 0$: these are the global minima. Therefore there are **no spurious local minima** — every local minimum is global. □

Exercise 3.4 (Symmetry of the global minima). The group $GL_h := GL_{d_1} \times \cdots \times GL_{d_{L-1}}$ acts on the weights by:

$$(W_1, \ldots, W_L) \mapsto (g_1 W_1,\ g_2 W_2 g_1^{-1},\ \ldots,\ W_L g_{L-1}^{-1})$$

Verify that the student map $\mu(\theta) = W_L \cdots W_1$ is invariant under this action: $\mu(g \cdot \theta) = \mu(\theta)$.

Solution to Exercise 3.4

Under the group action:

$$\mu(g \cdot \theta) = W_L g_{L-1}^{-1} \cdot g_{L-1} W_{L-1} g_{L-2}^{-1} \cdot \ldots \cdot g_1 W_1 = W_L W_{L-1} \cdots W_1 = \mu(\theta)$$

All g_l and g_l^{-1} factors cancel telescopically. This means every parameter point in the orbit $\mathcal{O}_\theta = GL_h \cdot \theta$ maps to the same student W. □
For the dimension of the set of global minima, this invariance means the following. The global minima form the fiber $\mu^{-1}(M)$, which contains the entire orbit $GL_h \cdot \theta^*$ for any global minimizer θ^*. The orbit has dimension $\sum_{l=1}^{L-1} d_l^2$ (the dimension of GL_h), so the set of global minima is a **continuous manifold of very high dimension** — far from being isolated points. □

4 Gradient flow and conserved quantities

We now study gradient flow: $\dot{\theta}(t) = -\nabla_\theta \mathcal{L}(\theta(t))$, the continuous-time limit of gradient descent with infinitesimal learning rate.

Exercise 4.1 (Gradient-flow equations). For a two-layer DLN ($L = 2$), the loss is given by $\mathcal{L} = \frac{1}{2}\|M - W_2 W_1\|_F^2$. Assume that each weight matrix W_i is diagonal. Derive the gradient flow equations for each layer. In particular, show that:

$$\dot{W}_1 = W_2^\top (M - W_2 W_1), \qquad \dot{W}_2 = (M - W_2 W_1) W_1^\top$$

Note: these equations also hold for general (non-diagonal) W_1 and W_2.

Solution to Exercise 4.1

We have $\mathcal{L} = \frac{1}{2}\|M - W_2 W_1\|_F^2 = \frac{1}{2}\mathrm{Tr}\left[(M - W_2 W_1)^\top (M - W_2 W_1)\right]$.
General matrix derivation. Expand:

$$\mathcal{L} = \frac{1}{2}\mathrm{Tr}(M^\top M) - \mathrm{Tr}(M^\top W_2 W_1) + \frac{1}{2}\mathrm{Tr}(W_1^\top W_2^\top W_2 W_1)$$

Differentiating with respect to W_1, using $\frac{\partial}{\partial A}\mathrm{Tr}(B^\top A) = B$ and $\frac{\partial}{\partial A}\mathrm{Tr}(A^\top CA) = (C + C^\top)A$:

$$\nabla_{W_1}\mathcal{L} = -W_2^\top M + W_2^\top W_2 W_1 = -W_2^\top (M - W_2 W_1)$$

So $\dot{W}_1 = -\nabla_{W_1}\mathcal{L} = W_2^\top (M - W_2 W_1)$.
Similarly, $\nabla_{W_2}\mathcal{L} = -(M - W_2 W_1)W_1^\top$, giving $\dot{W}_2 = (M - W_2 W_1)W_1^\top$.
Diagonal shortcut. For diagonal matrices $W_1 = \mathrm{diag}(a_\alpha)$, $W_2 = \mathrm{diag}(b_\alpha)$, $M = \mathrm{diag}(s_\alpha)$: the loss decouples as $\mathcal{L} = \frac{1}{2}\sum_\alpha (s_\alpha - b_\alpha a_\alpha)^2$. Then $\dot{a}_\alpha = -\partial\mathcal{L}/\partial a_\alpha = b_\alpha(s_\alpha - b_\alpha a_\alpha)$ and $\dot{b}_\alpha = a_\alpha(s_\alpha - b_\alpha a_\alpha)$, which is the diagonal version of the matrix equations above. □

Exercise 4.2 (Balancedness is conserved). Define the **balancedness matrix**:

$$G := W_2^\top W_2 - W_1 W_1^\top$$

Show that G is conserved under the gradient flow, i.e. $\dot{G} = 0$. In other words, gradient flow is constrained to the balanced manifold

$$\mathcal{M}_\beta := \{\theta \in \Omega \mid W_{\ell+1}^\top W_{\ell+1} - W_\ell W_\ell^\top = \beta_\ell,\ \ell = 1, \ldots, L-1\}.$$

Hint: Compute $\dot{G}$, substitute the gradient flow equations and verify that the terms cancel pairwise.

Solution to Exercise 4.2

Let $E := M - W_2 W_1$ denote the residual. Compute:

$$\dot{G} = \dot{W}_2^\top W_2 + W_2^\top \dot{W}_2 - \dot{W}_1 W_1^\top - W_1 \dot{W}_1^\top$$

Substitute the gradient flow equations. Using $\dot{W}_2 = E W_1^\top$ and $\dot{W}_1 = W_2^\top E$:

$$\dot{W}_2^\top W_2 = (E W_1^\top)^\top W_2 = W_1 E^\top W_2$$

$$W_2^\top \dot{W}_2 = W_2^\top E W_1^\top$$

$$\dot{W}_1 W_1^\top = W_2^\top E W_1^\top$$

$$W_1 \dot{W}_1^\top = W_1 (W_2^\top E)^\top = W_1 E^\top W_2$$

Therefore:

$$\dot{G} = W_1 E^\top W_2 + W_2^\top E W_1^\top - W_2^\top E W_1^\top - W_1 E^\top W_2 = 0$$

The terms cancel pairwise. G is conserved. $\square$

From now on, assume **balanced initialization**: $G(0) = 0$, which by Exercise 4.2 means $W_2^\top W_2 = W_1 W_1^\top$ for all time. We want to derive the gradient flow in function space (the ODE for the student $W = W_2 W_1$). This requires several steps.

Exercise 4.3 (Function-space velocity). Compute $\dot{W} := \frac{d}{dt}(W_2 W_1)$ using the gradient flow equations from Exercise 4.1. Show that:

$$\dot{W} = (M - W)\, W_1^\top W_1 + W_2 W_2^\top\, (M - W)$$

Solution to Exercise 4.3

Apply the product rule:

$$\dot{W} = \dot{W}_2 W_1 + W_2 \dot{W}_1 = (M - W_2 W_1) W_1^\top \cdot W_1 + W_2 \cdot W_2^\top (M - W_2 W_1)$$

Writing $W = W_2 W_1$:

$$\dot{W} = (M - W)\, W_1^\top W_1 + W_2 W_2^\top\, (M - W) \qquad \square$$

Exercise 4.4 ($W_1^\top W_1$ in terms of W). We now need to express $W_1^\top W_1$ and $W_2 W_2^\top$ in terms of $W = W_2 W_1$. Show that:

$$W_1^\top W_1 = (W^\top W)^{1/2}$$

Solution to Exercise 4.4

Balancedness gives $W_2^\top W_2 = W_1 W_1^\top$, so

$$W^\top W = W_1^\top \left(W_2^\top W_2\right) W_1 = W_1^\top \left(W_1 W_1^\top\right) W_1 = \left(W_1^\top W_1\right)^2$$

Both $W^\top W$ and $W_1^\top W_1$ are positive semidefinite. A positive semidefinite matrix has a unique positive semidefinite square root, so it follows that

$$(W^\top W)^{1/2} = W_1^\top W_1 \qquad \square$$

Exercise 4.5 ($W_2 W_2^\top$ in terms of W). Similarly, show that $W_2 W_2^\top = (W W^\top)^{1/2}$.

Solution to Exercise 4.5

Analogous to Exercise 4.4.

Exercise 4.6 (The balanced function-space ODE). Substitute the results of Exercises 4.4 and 4.5 into Exercise 4.3 to obtain:

$$\dot{W} = (WW^\top)^{1/2}(M - W) + (M - W)(W^\top W)^{1/2}$$

Solution to Exercise 4.6

Substituting into Exercise 4.3:

$$\dot{W} = (M - W)(W^\top W)^{1/2} + (WW^\top)^{1/2}(M - W) \qquad \square$$

Exercise 4.7 (The NTK operator). Define the NTK operator for $L = 2$ as:

$$K[F] := (WW^\top)^{1/2} F + F (W^\top W)^{1/2}$$

Show that the gradient flow from Exercise 4.6 can be written as $\dot{W} = K[M-W]$. It turns out that this result generalizes to depth L on the balanced manifold:

$$\dot{W} = \sum_{k=1}^{L} (WW^\top)^{\frac{L-k}{L}} (M - W)(W^\top W)^{\frac{k-1}{L}}$$

This is the NTK equation in the case of DLNs. The NTK equation is a gradient flow in function space with NTK being a preconditioning operator for the gradient.

Solution to Exercise 4.7

With $F = M - W$, the equation from Exercise 4.6 reads:

$$\dot{W} = (WW^\top)^{1/2} F + F (W^\top W)^{1/2} = K[F] = K[M - W] \qquad \square$$

The general depth-L result $\dot{W} = \sum_{k=1}^{L} (WW^\top)^{(L-k)/L} (M - W)(W^\top W)^{(k-1)/L}$ follows from the same approach applied to the L-fold balanced conditions $W_{l+1}^\top W_{l+1} = W_l W_l^\top$ for all l.

5 Rich regime: saddle-to-saddle

This is the core problem. We derive the exact solution of the rich regime dynamics directly from the self-consistent equation of Section 4, emphasizing the NTK perspective.

Alignment assumption. Start from the balanced gradient flow equation derived in Exercise 4.6:

$$\dot{W} = (WW^\top)^{1/2}(M - W) + (M - W)(W^\top W)^{1/2}$$

We work in the **rich regime** (small initialization) with balanced weights. Assume that the student is **aligned** to the teacher: the singular vectors of $W(t)$ coincide with those of M at all times. Concretely, write:

$$M = U\,\mathrm{diag}(s_1, \dots, s_d)\,V^\top, \qquad W(t) = U\,\mathrm{diag}(w_1(t), \dots, w_d(t))\,V^\top$$

where U, V are the left/right singular vectors of M, and $w_\alpha(t) \geq 0$ are the evolving singular values of W.

Exercise 5.1 (Aligned NTK). Under this alignment assumption, show that the NTK operator is aligned to the task, i.e. $K[F]$ preserves the SVD basis. In particular, show that:

$$(WW^\top)^{1/2} = U\,\mathrm{diag}(w_1, \dots, w_d)\,U^\top,$$
$$(W^\top W)^{1/2} = V\,\mathrm{diag}(w_1, \dots, w_d)\,V^\top$$

and that the residual is $M - W = U\,\mathrm{diag}(s_1 - w_1, \dots, s_d - w_d)\,V^\top$.
Hint: Recall that for $A = U\,\mathrm{diag}(\sigma_i)\,V^\top$, we have $AA^\top = U\,\mathrm{diag}(\sigma_i^2)\,U^\top$, and therefore $(AA^\top)^{1/2} = U\,\mathrm{diag}(|\sigma_i|)\,U^\top$.

Solution to Exercise 5.1

Under the alignment assumption, $W = U\,\mathrm{diag}(w_\alpha)\,V^\top$. Then:

$$WW^\top = U\,\mathrm{diag}(w_\alpha^2)\,U^\top$$

Since $w_\alpha \geq 0$, the positive matrix square root is:

$$(WW^\top)^{1/2} = U\,\mathrm{diag}(w_\alpha)\,U^\top$$

Similarly:

$$W^\top W = V\,\mathrm{diag}(w_\alpha^2)\,V^\top \quad \Longrightarrow \quad (W^\top W)^{1/2} = V\,\mathrm{diag}(w_\alpha)\,V^\top$$

The residual is immediate:

$$M - W = U\,\mathrm{diag}(s_\alpha - w_\alpha)\,V^\top$$

The NTK operator acts on any matrix $F = U\,\mathrm{diag}(f_\alpha)\,V^\top$ as:

$$
\begin{aligned}
K[F] &= U\,\mathrm{diag}(w_\alpha)\,U^\top U\,\mathrm{diag}(f_\alpha)\,V^\top \\
&\quad + U\,\mathrm{diag}(f_\alpha)\,V^\top V\,\mathrm{diag}(w_\alpha)\,V^\top \\
&= U\,\mathrm{diag}(2w_\alpha f_\alpha)\,V^\top.
\end{aligned}
$$

So $K[F]$ stays in the SVD basis of M — the NTK is aligned to the task. □

Exercise 5.2 (Decoupled scalar ODEs). Substitute into the self-consistent equation. Show that the matrix equation decouples into d independent scalar ODEs:

$$
\dot{w}_\alpha = 2\,w_\alpha\,(s_\alpha - w_\alpha), \qquad \alpha = 1, \ldots, d
$$

Hint: Compute the product $(WW^\top)^{1/2}(M-W)$ in the SVD basis. It is diagonal with entries $w_\alpha(s_\alpha - w_\alpha)$. The second term contributes identically, giving the factor of 2.

Solution to Exercise 5.2

Substitute into $\dot{W} = (WW^\top)^{1/2}(M-W) + (M-W)(W^\top W)^{1/2}$:
First term:

$$
U\,\mathrm{diag}(w_\alpha)\,\underbrace{U^\top U}_{I}\,\mathrm{diag}(s_\alpha - w_\alpha)\,V^\top = U\,\mathrm{diag}\big(w_\alpha(s_\alpha - w_\alpha)\big)\,V^\top
$$

Second term:

$$
U\,\mathrm{diag}(s_\alpha - w_\alpha)\,\underbrace{V^\top V}_{I}\,\mathrm{diag}(w_\alpha)\,V^\top = U\,\mathrm{diag}\big((s_\alpha - w_\alpha)\,w_\alpha\big)\,V^\top
$$

Summing:

$$
\dot{W} = U\,\mathrm{diag}\big(2\,w_\alpha(s_\alpha - w_\alpha)\big)\,V^\top
$$

Since $W = U\,\mathrm{diag}(w_\alpha)\,V^\top$ and the singular vectors are constant by assumption, we read off:

$$
\boxed{\dot{w}_\alpha = 2\,w_\alpha\,(s_\alpha - w_\alpha)}
$$

The NTK perspective makes the structure transparent: the factor $2w_\alpha$ is the NTK eigenvalue for mode α, which amplifies learning in directions that are already strong, while $(s_\alpha - w_\alpha)$ is the residual. □

Exercise 5.3 (Timescale of learning). The ODE $\dot{w} = 2w(s-w)$ is a logistic equation whose solution is:

$$w(t) = \frac{s}{1 + \left(\frac{s}{w_0} - 1\right) e^{-2st}}$$

where $w_0 = w(0)$. Compute the time t_α it takes for mode α to travel from initial strength w_0 to a final strength w_f. Show that:

$$t_\alpha = \frac{1}{2s_\alpha} \ln\left(\frac{w_f\,(s_\alpha - w_0)}{w_0\,(s_\alpha - w_f)}\right)$$

Deduce that modes with larger singular values s_α are learned faster. This is a **separation of timescales**: the network learns features in decreasing order of their singular value strength.

Solution to Exercise 5.3

From the logistic solution $w(t) = \frac{s}{1+(s/w_0-1)e^{-2st}}$, we invert to find t as a function of w. The derivation is equivalent to integrating $\dot{w} = 2w(s-w)$ by separation of variables:

$$\int_{w_0}^{w_f} \frac{dw}{w(s-w)} = 2\,t_\alpha$$

Using partial fractions $\frac{1}{w(s-w)} = \frac{1}{s}\left(\frac{1}{w} + \frac{1}{s-w}\right)$:

$$\frac{1}{s_\alpha}\Big[\ln w - \ln(s_\alpha - w)\Big]_{w_0}^{w_f} = 2\,t_\alpha$$

$$\frac{1}{s_\alpha} \ln \frac{w_f\,(s_\alpha - w_0)}{w_0\,(s_\alpha - w_f)} = 2\,t_\alpha$$

Therefore:

$$\boxed{t_\alpha = \frac{1}{2s_\alpha} \ln\left(\frac{w_f\,(s_\alpha - w_0)}{w_0\,(s_\alpha - w_f)}\right)}$$

For a fixed ratio w_f/s_α and fixed w_0, the learning time scales as $t_\alpha \propto 1/s_\alpha$: **modes with larger singular values are learned faster**. The network learns features in decreasing order of their strength — a strong separation of timescales. **Contrast with the lazy regime** (anticipating Section 6): in the lazy regime the NTK is frozen at $K_0 = 2w_0 \gg 1$, so $\dot{w}_\alpha \approx 2w_0(s_\alpha - w_\alpha)$. All modes converge at the same exponential rate $2w_0$, independent of s_α. The rich regime has state-dependent NTK ($2w_\alpha$), creating a positive feedback loop — modes that are already large learn even faster — which amplifies the differences between s_α into a hierarchy of timescales $t_1 \ll t_2 \ll \cdots \ll t_r$. □

Exercise 5.4 (Incremental learning). Consider a teacher with r nonzero singular values and a small uniform initialization $w_\alpha(0) = w_0 \ll s_r$ for all α. Qualitatively, discuss:

- What is the approximate distribution of singular values of $W(t)$ at early times?
- How do the singular values of $W(t)$ evolve throughout training?
- What happens in the limit $t \to \infty$?

What does this analysis tell us about the sequence of critical points visited by the gradient flow?

Solution to Exercise 5.4

With uniform initialization $w_\alpha(0) = w_0$ for all α, the sigmoid solution shows that mode α remains near w_0 until time $t \approx \frac{1}{2s_\alpha} \ln(s_\alpha/w_0)$, then transitions rapidly to $w_\alpha \approx s_\alpha$.
Early times ($t \sim \frac{1}{2s_1} \ln(s_1/w_0)$): Only mode 1 has grown significantly, so the spectrum is **one singular value near s_1, with all the others still near** $w_0 \ll s_r$. The student is approximately $W(t) \approx w_1(t)\, u_1 v_1^\top$, effectively **rank 1**.
Intermediate times: As t increases, w_2 switches on, then w_3, etc. The singular values **switch on one at a time**, in decreasing order of s_α, each staying near w_0 and then rising to s_α over a narrow window. The student is approximately $W(t) \approx \sum_{\alpha=1}^{k(t)} w_\alpha(t)\, u_\alpha v_\alpha^\top$ where $k(t)$ increases stepwise, so the effective **rank increases incrementally**.
Long times ($t \to \infty$): All modes converge, $w_\alpha \to s_\alpha$, and $W \to M$.
This is an **implicit bias toward low-rank (simple) solutions**: at any finite time, the network represents the best rank-k approximation to the teacher. The gradient flow effectively performs a greedy SVD.
Connection to the saddle structure (Exercise 3.3): The saddle points $P_S M$ with $|S| = k$ are exactly the best rank-k approximations to M, with loss $\frac{1}{2}\sum_{\alpha>k} s_\alpha^2$. The gradient flow trajectory passes **near** these saddles as it incrementally recruits modes. The **plateaus** in the loss curve correspond to time spent near saddle points (only k modes active), and the **sharp transitions** correspond to escaping along the unstable direction that activates the next mode.
□

6 Lazy regime

We now show that large initialization freezes the NTK and eliminates the timescale separation found in Section 5. For clarity, we work with the **diagonal scalar** model from Exercise 3.1, so each mode α is independent. This avoids matrix algebra and isolates the essential mechanism.

Setup. Consider a single mode: a depth-2 diagonal DLN with scalar weights a, b learning a target $s > 0$. From Sections 4 and 5, the balanced gradient flow for $w = ab$ is:

$$\dot{w} = 2w(s - w)$$

The factor $2w$ is the (scalar) NTK — it is **state-dependent**: the effective learning rate depends on the current value of w.

Exercise 6.1 (Linearizing the NTK). Now initialize at a **large** value $w_0 \gg s > 0$. Assume that in the early phase of training, while w has not changed much from w_0, the ODE becomes approximately:

$$\dot{w} \approx 2w_0(s - w)$$

Solution to Exercise 6.1

Starting from the same ODE $\dot{w} = 2w(s - w)$, with $w_0 \gg s > 0$, the weight w needs to decrease from w_0 to s. As long as w has not changed much from w_0, we can write $w = w_0 + \delta w$ with $|\delta w| \ll w_0$, so:

$$2w = 2(w_0 + \delta w) \approx 2w_0$$

The ODE becomes:

$$\dot{w} \approx 2w_0(s - w) \qquad \square$$

This is a **linear** ODE — the NTK factor $2w$ has been frozen at its initial value $2w_0$. Note that only the NTK prefactor is linearized; the residual $(s - w)$ is kept exact since it drives learning.

Exercise 6.2 (Frozen-NTK solution). Solve the linearized ODE from Exercise 6.1. Show that:

$$w(t) = s + (w_0 - s)\, e^{-2w_0 t}$$

What is the learning timescale?

Solution to Exercise 6.2

The linearized ODE $\dot{w} = 2w_0(s - w)$ is first-order linear. Set $\tilde{w} := w - s$, then $\dot{\tilde{w}} = -2w_0\, \tilde{w}$, with solution $\tilde{w}(t) = (w_0 - s)\, e^{-2w_0 t}$. Therefore:

$$\boxed{w(t) = s + (w_0 - s)\, e^{-2w_0 t}}$$

This is **exponential** convergence to s (contrast with the sigmoid of Section 5).

The learning timescale is:

$$t_{\text{lazy}} \sim \frac{1}{2w_0}$$

It depends on the **initialization scale** w_0 but **not on the target** s. This is the defining feature of the lazy regime: the learning speed is set by the NTK at initialization, not by the structure of the task.

Self-consistency: During the learning phase ($t \lesssim 1/(2w_0)$), the displacement is $|w(t) - w_0| \leq |s - w_0| \approx w_0$ (since $w_0 \gg s$). The relative change in the NTK is $\Delta(2w)/(2w_0) = (w_0 - s)/w_0 = 1 - s/w_0 \approx 1$, so the NTK does change significantly in absolute terms. However, the key point is that the dynamics remain well approximated by the linear ODE because the convergence rate is dominated by $2w_0$, which is large and approximately constant throughout training. In the rich regime, by contrast, the NTK changes from $2w_0 \approx 0$ to $2s$ — a change of order $s/w_0 \to \infty$ relative to the initial value — making the linearization completely invalid. □

Exercise 6.3 (No timescale separation). Now restore the mode index α. In the lazy regime, each mode satisfies $\dot{w}_\alpha \approx 2w_0(s_\alpha - w_\alpha)$ with the **same** rate $2w_0$ for all α. Compare the lazy learning time t_α^{lazy} with the rich-regime timescale from Exercise 5.3: $t_\alpha^{\text{rich}} \propto 1/s_\alpha$.

Solution to Exercise 6.3

Restoring the mode index, each mode satisfies $\dot{w}_\alpha \approx 2w_0(s_\alpha - w_\alpha)$ with solution:

$$w_\alpha(t) = s_\alpha + (w_0 - s_\alpha)\, e^{-2w_0 t}$$

All modes converge at the **same exponential rate** $2w_0$, regardless of s_α. The time for mode α to go from w_0 to within $(1 - \epsilon)$ of s_α is:

$$t_\alpha^{\text{lazy}} = \frac{1}{2w_0} \ln\left(\frac{w_0 - s_\alpha}{\epsilon\,(w_0 - s_\alpha)}\right) = \frac{1}{2w_0} \ln\frac{1}{\epsilon}$$

which is **independent of** s_α. All modes reach their targets simultaneously.

Contrast with the rich regime:

	Rich regime	Lazy regime
NTK	State-dependent: $2w_\alpha$	Frozen: $2w_0$
Learning time for mode α	$t_\alpha \propto 1/s_\alpha$	$t_\alpha \approx$ const
Timescale separation	Strong: $t_1 \ll t_2 \ll \cdots$	None
Intermediate solutions	Low-rank (incremental)	Full-rank from the start
Implicit rank bias	Yes (low-rank $\to$ high-rank)	No

In the lazy regime, all components of M are learned simultaneously: signal and noise alike. The network converges directly to the full OLS solution without passing through low-rank intermediates. There is no mechanism to separate signal from noise based on singular value structure. With early stopping, the rich regime recovers a low-rank signal while ignoring noise; the lazy regime cannot. □

7 Mixed dynamics: unifying lazy and rich

In practice, networks are neither purely lazy nor purely rich. Tu, Aranguri & Jacot (2024) provide a unified description. We develop the key ideas using the diagonal scalar model.

Exercise 7.1 (The interpolating ODE). In Sections 5 and 6, we studied the same ODE $\dot{w}_\alpha = 2w_\alpha(s_\alpha - w_\alpha)$ in two limits: small w_0 (rich) and large w_0 (lazy). A more general model for a depth-2 DLN with balanced initialization at scale σ and width n replaces the scalar NTK $2w_\alpha$ by:

$$\dot{w}_\alpha = 2\sqrt{w_\alpha^2 + \tau^2}\,(s_\alpha - w_\alpha)$$

where $\tau := \sigma^2\sqrt{n}$ is a **threshold** parameter that depends on the initialization scale and width.
Verify that this ODE reproduces the two known regimes:

- **Rich limit** ($\tau \to 0$): recover $\dot{w}_\alpha = 2|w_\alpha|(s_\alpha - w_\alpha)$.
- **Lazy limit** ($\tau \to \infty$ with w_α bounded): recover $\dot{w}_\alpha \approx 2\tau(s_\alpha - w_\alpha)$.

Solution to Exercise 7.1

The interpolating ODE is $\dot{w}_\alpha = 2\sqrt{w_\alpha^2 + \tau^2}\,(s_\alpha - w_\alpha)$.
Rich limit ($\tau \to 0$): The square root reduces to $\sqrt{w_\alpha^2} = |w_\alpha|$, giving:

$$\dot{w}_\alpha = 2|w_\alpha|(s_\alpha - w_\alpha)$$

For $w_\alpha > 0$ this is $2w_\alpha(s_\alpha - w_\alpha)$, exactly the logistic ODE from Exercise 5.2. □
Lazy limit ($\tau \to \infty$, w_α bounded): The square root is dominated by τ: $\sqrt{w_\alpha^2 + \tau^2} \approx \tau$, giving:

$$\dot{w}_\alpha \approx 2\tau(s_\alpha - w_\alpha)$$

This is a linear ODE with rate 2τ, independent of s_α — the frozen-NTK regime of Section 6 (with τ playing the role of w_0). □

Exercise 7.2 (Two-phase dynamics). At initialization, all modes start at $w_\alpha(0) \sim \sigma^2 \ll \tau$ (since $\sqrt{n} \gg 1$). Observe that:

- When $|w_\alpha| \ll \tau$, the effective learning rate is $\approx 2\tau$, the same for all modes (lazy behavior).
- When $|w_\alpha| \gg \tau$, the effective learning rate is $\approx 2|w_\alpha|$, which is mode-dependent (rich behavior).

Describe the resulting two-phase dynamics in words: what happens first, and what happens later?

Solution to Exercise 7.2

At initialization, $w_\alpha(0) \sim \sigma^2 \ll \tau = \sigma^2\sqrt{n}$ (since $\sqrt{n} \gg 1$). So initially all modes satisfy $|w_\alpha| \ll \tau$.
Early phase (lazy): $\sqrt{w_\alpha^2 + \tau^2} \approx \tau$ for all α. Every mode evolves at the same linear rate 2τ:

$$\dot{w}_\alpha \approx 2\tau(s_\alpha - w_\alpha) \qquad \Longrightarrow \qquad w_\alpha(t) \approx s_\alpha(1 - e^{-2\tau t})$$

(assuming $w_\alpha(0) \approx 0$). The dynamics are approximately linear and the NTK is approximately constant. During this phase, the network **aligns** with the task: each w_α grows toward s_α at the same rate. There is no timescale separation.
Late phase (rich): Once some w_α grow past τ, the square root transitions to $\sqrt{w_\alpha^2 + \tau^2} \approx |w_\alpha|$, and those modes enter the rich regime with the sigmoidal, self-accelerating dynamics $\dot{w}_\alpha \approx 2w_\alpha(s_\alpha - w_\alpha)$. The state-dependent NTK creates timescale separation, and the modes that have crossed the threshold converge rapidly to their targets.

In summary: the network starts in a lazy phase where all modes grow uniformly (alignment), then transitions to a rich phase where modes accelerate individually (incremental learning). □

Exercise 7.3 (Mode-by-mode transition). Not all modes cross the threshold τ at the same time: which modes cross it first? Explain why this means that a single network can simultaneously have some modes in the rich regime and others still in the lazy regime.

Solution to Exercise 7.3

In the lazy phase, each mode grows as $w_\alpha(t) \approx s_\alpha(1 - e^{-2\tau t})$. At any given time, $w_\alpha(t) \propto s_\alpha$: **modes with larger s_α are larger**. Since the lazy-to-rich transition occurs when $|w_\alpha| \sim \tau$, the time for mode α to cross the threshold satisfies:

$$s_\alpha(1 - e^{-2\tau t^*_\alpha}) = \tau \qquad \Longrightarrow \qquad t^*_\alpha = \frac{1}{2\tau} \ln \frac{s_\alpha}{s_\alpha - \tau}$$

For $s_\alpha > \tau$, this crossing time exists and is smaller for larger s_α. For $s_\alpha < \tau$, the mode never reaches the threshold and remains permanently lazy.
This means that **different modes can be in different regimes simultaneously**: large-s_α modes have crossed into the rich regime and are converging rapidly, while small-s_α modes are still in the lazy phase (or may never leave it). The network interpolates between the two regimes mode by mode.
Connection to grokking: This lazy-to-rich transition provides a mechanism for grokking. During the lazy phase, the network fits the training data via kernel regression with the initial (misaligned) NTK — it **memorizes**. The test loss plateaus because the frozen kernel cannot capture the task structure. Later, as modes cross the threshold into the rich regime, feature learning kicks in: the NTK rotates to align with the task, and the test loss drops — **generalization** is achieved. The grokking delay is controlled by the time it takes the relevant modes to cross τ. Increasing the initialization scale σ or the width n increases τ, which extends the lazy phase and widens the grokking gap. This is the same mechanism identified by Kumar et al., *Grokking as the Transition from Lazy to Rich Training Dynamics* (arXiv:2310.06110). □

8 Stochastic implicit bias (bonus)

SGD introduces noise from mini-batching. A continuous model is the Langevin SDE:

$$d\theta_t = -\nabla \mathcal{L}(\theta_t)\, dt + \sqrt{\eta\, \Sigma(\theta_t)}\, dB_t$$

where $\Sigma(\theta)$ is the covariance of the stochastic gradient noise and η is the learning rate.

Exercise 8.1 (Boltzmann equilibrium). The probability density $p(\theta, t)$ of the parameters evolves according to the Fokker–Planck equation:

$$\partial_t p = -\nabla \cdot \mathbf{j}, \qquad \mathbf{j} = -\nabla \mathcal{L}(\theta)\, p(\theta) - \frac{\eta}{2} \nabla \cdot (\Sigma(\theta)\, p(\theta))$$

where $\mathbf{j}$ is the probability current. Assume: (i) stationarity $\partial_t p^* = 0$, (ii) thermal equilibrium $\mathbf{j} = 0$, and (iii) isotropic noise $\Sigma = \sigma^2 I$. Show that the equilibrium distribution is the Boltzmann distribution:

$$p^*(\theta) \propto \exp\left(-\frac{2}{\eta\sigma^2}\mathcal{L}(\theta)\right)$$

Hint: Setting $\mathbf{j} = 0$ with $\Sigma = \sigma^2 I$ gives $\nabla \mathcal{L}\, p + \frac{\eta\sigma^2}{2} \nabla p = 0$. This is a first-order ODE for p in terms of $\mathcal{L}$. Try the ansatz $p \propto e^{-\beta \mathcal{L}}$ and solve for β.

Solution to Exercise 8.1

Setting $\mathbf{j} = 0$ (thermal equilibrium) with isotropic noise $\Sigma = \sigma^2 I$:

$$0 = -\nabla \mathcal{L}(\theta)\, p^*(\theta) - \frac{\eta\sigma^2}{2} \nabla p^*(\theta)$$

Rearranging:

$$\frac{\nabla p^*}{p^*} = -\frac{2}{\eta\sigma^2} \nabla \mathcal{L}$$

The left-hand side is $\nabla \ln p^*$, so:

$$\nabla \ln p^*(\theta) = -\frac{2}{\eta\sigma^2} \nabla \mathcal{L}(\theta)$$

Integrating both sides:

$$\ln p^*(\theta) = -\frac{2}{\eta\sigma^2} \mathcal{L}(\theta) + \text{const}$$

$$\boxed{p^*(\theta) \propto \exp\left(-\frac{2}{\eta\sigma^2}\mathcal{L}(\theta)\right)}$$

This is the Boltzmann distribution with inverse temperature $\beta = 2/(\eta\sigma^2)$. □

Exercise 8.2 (Temperature and flatness). The ratio $\frac{2}{\eta\sigma^2}$ plays the role of an inverse temperature β. Interpret what happens to the equilibrium distribution

when:

- η is very small (low temperature)
- η is very large (high temperature)

Which regime favors flatter minima, and why might this be beneficial for generalization?

Solution to Exercise 8.2

The effective temperature is $T = \eta\sigma^2/2$.
Small η (low temperature, $\beta \to \infty$): The distribution concentrates sharply on the global minima of $\mathcal{L}$. SGD converges to the lowest-loss minimizer without exploring.
Large η (high temperature, $\beta \to 0$): The distribution becomes nearly uniform over parameter space. SGD explores broadly and does not settle into any particular minimum.
Flat vs. sharp minima: At moderate temperature, the Boltzmann distribution assigns more probability mass to **broad basins** (flat minima) than to narrow ones. This is because a flat minimum occupies a larger volume of parameter space at any given loss level — the width of the basin acts as an entropic contribution. Flat minima tend to generalize better because small perturbations to the parameters (or slight distribution shift in the data) do not dramatically change the loss. Thus the implicit bias of SGD noise toward flat minima is beneficial for generalization. □

Exercise 8.3 (Anisotropic noise). In practice, SGD noise is **not** isotropic: $\Sigma(\theta)$ depends on both the loss landscape and the data. Without solving anything, explain qualitatively why anisotropic noise can introduce an implicit bias that goes beyond what the loss function $\mathcal{L}$ alone would select. Specifically, why might SGD preferentially escape sharp directions of the loss while remaining stable along flat directions?

Solution to Exercise 8.3

When $\Sigma(\theta)$ is anisotropic, the noise strength varies by direction. The Fokker-Planck current becomes:

$$\mathbf{j} = -\nabla\mathcal{L}\, p - \frac{\eta}{2}\nabla \cdot (\Sigma(\theta)\, p)$$

The second term introduces an **Itô drift** $\frac{\eta}{2}\nabla \cdot \Sigma(\theta)$ that depends on the spatial variation of the noise covariance. In directions where Σ has large eigenvalues (high noise), the effective diffusion is strong and the system escapes easily from

sharp regions of the loss landscape. In directions where Σ has small eigenvalues (low noise), the system is more stable and tends to remain there.
This creates a **direction-dependent regularizer**: SGD preferentially pushes parameters out of sharp directions (high gradient variance $\rightarrow$ large noise eigenvalue $\rightarrow$ fast escape) while preserving parameters along flat directions (low gradient variance $\rightarrow$ small noise eigenvalue $\rightarrow$ stability). This goes beyond what the Boltzmann distribution on $\mathcal{L}$ alone would predict. In general, detailed balance $\mathbf{j} = 0$ may not hold for anisotropic, state-dependent noise, leading to persistent probability currents and non-equilibrium steady states that further modify the implicit bias. $\square$

Learn more

The readings are roughly ordered in a way that makes sense for learning. Discuss the readings in groups of 2 or 3, to be formed in this Google doc. Spend 30 minutes reading, 30 minutes discussing and 30 minutes to write down your thoughts in the Google doc.

Key readings

- The strong default suggestion for everyone is: Saxe et al. A mathematical theory of semantic development in deep neural networks — Read the supplementary materials section to understand the exact solution of gradient flow dynamics in deep linear networks
- For people with a strong algebraic geometry and representation geometry background read: Geometry of fibers of the multiplication map of deep linear neural networks Simon Pepin Lehalleur et al. — Stratifies global minima into orbits using quiver representation theory
- For people with a strong dynamical system and differential geometry background read: The geometry of the deep linear network Govind Menon — Beautiful mathematical treatment of gradient flow in DLNs and a surprising mathematical result relating gradient flow and free-energy minimization. Makes implicit regularization explicit (under balanced assumptions)
- For people who are much more empirically minded: Emergent Misalignment is Easy, Narrow Misalignment is Hard — When fine-tuned on data with harmful inputs, the model generalizes in harmful ways on other unrelated datasets. By regularizing, we can mitigate emergent misalignment
- Literature Review: There Will Be a Scientific Theory of Deep Learning
- To go further: Alternating Gradient Flows: A Theory of Feature Learning in Two-layer Neural Networks

Loss landscape geometry

- Deep Learning without Poor Local Minima, Kenji Kawaguchi — For non-bottlenecked DLNs, all local minima are global
- The loss landscape of deep linear neural networks: a second-order analysis Achour et al. — First-order and second-order classification of critical points in DLN loss landscapes. Classifies strict and non-strict saddles.
- Pure and Spurious Critical Points: a Geometric Study of Linear Networks Matthew Trager et al. — Defines spurious minima and counts the number of connected components of the global minima
- Geometry of fibers of the multiplication map of deep linear neural networks Simon Pepin Lehalleur et al. — Stratifies global minima into orbits using quiver representation theory
- The Loss Surfaces of Multilayer Networks Anna Choromanska et al — No-local-minima results in non-linear multilayer networks under some strong assumptions (spin glass models)
- Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs, Timur Garipov et al — Study the connectedness of global minima of loss landscapes (mode connectivity)
- The Multilinear Structure of ReLU Networks, Thomas Laurent, James von Brecht — Show that local minima are typically singular
- Classic result: Neural networks and principal component analysis: Learning from examples without local minima Pierre Baldi et al. — Original result that linear network landscapes have no spurious local minima (single hidden layer case)
- Good review of key DLNs results: Gradient Flow Equations for Deep Linear Neural Networks: A Survey from a Network Perspective, Joel Wendin, Claudio Altafini — Also covers gradient flow

Implicit biases of gradient flow

- Saxe et al. A mathematical theory of semantic development in deep neural networks — Read the supplementary materials section to understand the exact solution of gradient flow dynamics in deep linear networks
- Saddle-to-Saddle Dynamics in Deep Linear Networks: Small Initialization Training, Symmetry, and Sparsity Arthur Jacot et al. — Studies the saddle-to-saddle dynamics in DLNs. Introduces the different regimes (lazy and rich)
- The geometry of the deep linear network Govind Menon — Beautiful mathematical treatment of gradient flow in DLNs and a surprising mathematical result

relating gradient flow and free-energy minimization. Makes implicit regularization explicit (under balanced assumptions)

- Saddle-to-Saddle Dynamics Explains A Simplicity Bias Across Neural Network Architectures Saxe et al. — Simplicity bias of gradient flow from DLNs extends to non-linear and transformer architectures under small initialization
- Abide by the Law and Follow the Flow: Conservation Laws for Gradient Flows — Gives a recipe to exhaustively derive conservation laws through gradient flow
- SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics — SGD builds up complex solutions by first learning simpler solutions and then composing them
- Mixed Dynamics In Linear Networks: Unifying the Lazy and Active Regimes — Unifies lazy and rich in 2-layer DLNs and gives conditions for the transition between them
- Neural Tangent Kernel: Convergence and Generalization in Neural Networks Arthur Jacot, Franck Gabriel, Clément Hongler — NTK paper that describes gradient flow in function space
- On Lazy Training in Differentiable Programming Lénaïc Chizat, Edouard Oyallon, Francis Bach — Lazy training can occur in various settings, and is caused by scaling choices
- The Neural Race Reduction: Dynamics of Abstraction in Gated Networks Saxe et al. — Implicit biases toward shared representation in non-linear networks as modelled by gated DLNs
- Alternating Gradient Flows: A Theory of Feature Learning in Two-layer Neural Networks Daniel Kunin et al. — A theory of feature learning as a two-step process between dormant and active neurons
- From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks — Studies the transition between lazy and rich regimes in DLNs with balancedness parameters
- Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning Daniel Kunin et al. — (Lack of) balance between layers at initialization is preserved under GF, and influences rich vs lazy and implicit bias
- Implicit Regularization in Matrix Factorization — Gradient descent on matrix factorization converges to the minimum nuclear norm
- Gradient Descent Maximizes the Margin of Homogeneous Neural Networks — Implicit bias toward max-margin in classification
- A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks Sanjeev Arora et al. — Convergence analysis of SGD

Learning rate (discrete GD)

- Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability — Shows that curvature stabilizes around twice the inverse of the learning rate (2/eta).
- Understanding Optimization in Deep Learning with Central Flows — GD with a discrete learning rate is equivalent to gradient flow on an effective loss
- Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective — Pretraining exhibits a river valley loss landscape and gives intuition about the learning-rate schedule (warm-up, stable, decay)
- Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning — Models the landscape as multifractal and analyses sub- and super-diffusive behaviour of GD

Stochasticity

- On the implicit regularization of Langevin dynamics with projected noise — Makes explicit the implicit bias of stochasticity in deep linear networks with a Langevin model
- Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning — Golden Path Hypothesis: in the transient regime dominated by drift (typical of pretraining), SGD is simply a noisy deformation of GD (it does not select different basins)
- Stochastic Collapse: How Gradient Noise Attracts SGD Dynamics Towards Simpler Subnetworks — Stochasticity induces a bias toward simpler solutions in deep linear networks
- Stochastic Training is Not Necessary for Generalization — Makes the implicit regularization of stochasticity explicit
- Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity — Shows that stochasticity has a bias toward sparser solutions relative to GD
- Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks — SGD minimizes some free energy and can have circular currents at convergence
- Stochastic Gradient Descent as Approximate Bayesian Inference — Classic paper on SGD locally approximating Bayesian inference, although it assumes non-degeneracies
- A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima — Implicit bias toward flat minima

- Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions — Shows the Golden Path Hypothesis on quadratic loss in a convex setup using an interesting SDE model of SGD.
- An Empirical Model of Large-Batch Training — Gradient noise scale can be used to decide the optimal batch size
- Almost Bayesian: The Fractal Dynamics of Stochastic Gradient Descent — Relates SGD to Bayesian training
- Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning — Loss spikes when training with SGD impact generalization
- The Heavy-Tail Phenomenon in SGD — SGD noise can be heavy-tailed, and this induces a different regularization

Emergence: empirical examples

- Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets — Grokking: delayed generalization on modular addition
- In-context Learning and Induction Heads — Induction heads form during training. They are important for in-context learning
- Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs — Emergent misalignment: fine-tuning on insecure code can generalize to misaligned behaviour on other data
- Emergent Misalignment is Easy, Narrow Misalignment is Hard — By regularizing, we can mitigate emergent misalignment
- Neural Networks as Kernel Learners: The Silent Alignment Effect — While the loss is flat, student singular vectors align with teacher singular vectors, and this can be detected with the NTK
- Are Emergent Abilities of Large Language Models a Mirage? — Emergence is in the eye of the metric used to measure it
- Training Compute-Optimal Large Language Models — Chinchilla scaling law paper
- Emergent Abilities of Large Language Models — Shows that novel capabilities emerge with more compute

Theoretical approaches to emergence

- Grokking as the Transition from Lazy to Rich Training Dynamics
- Grokking as a First Order Phase Transition in Two Layer Networks, Rubin et al.
- Blake Bordelon - Infinite limits and scaling laws of neural networks - IPAM at UCLA
- A Theory for Emergence of Complex Skills in Language Models
- On neural scaling and the quanta hypothesis
- Lecture Notes on Infinite-Width Limits of Neural Networks; Cengiz Pehlevan and Blake Bordelon
- Disordered Dynamics in High Dimensions: Connections to Random Matrices and Machine Learning; Blake Bordelon, Cengiz Pehlevan
- Applications of Statistical Field Theory in Deep Learning; Zohar Ringel et al.
- Statistical Field Theory for Neural Networks, Moritz Helias, David Dahmen
- Lecture notes: From Gaussian processes to feature learning, Moritz Helias et al

References

[AMG24] Achour, Malgouyres, and Gerchinovitz. The loss landscape of deep linear neural networks: a second-order analysis. *Journal of Machine Learning Research (JMLR)*, 2024. URL: https://arxiv.org/abs/2107.13289.

[SMG14] Saxe, McClelland, and Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[TAJ24] Tu, Aranguri, and Jacot. Mixed dynamics in linear networks: Unifying the lazy and active regimes. 2024. URL: https://arxiv.org/abs/2405.17580.