



Effective Dynamics of SGD

in High-Dimensional Neural Networks

Luca Arnaboldi | PhD Defence | 26 August 2026

Why are we doing this?

An engineering triumph, seventy years in the making



We can train these models. We still cannot say *why* they work.

Practical

better algorithms, more efficient models, less compute and energy burnt

Social

interpretability, fairness, robustness

Scientific

this thesis goes to the *École Doctorale de Physique*: understanding a complex system is a goal in itself

Why are we doing this?

An engineering triumph, seventy years in the making



We can train these models. We still cannot say *why* they work.

Practical

better algorithms, more efficient models, less compute and energy burnt

Social

interpretability, fairness, robustness

Scientific

this thesis goes to the *École Doctorale de Physique*: understanding a complex system is a goal in itself

Why are we doing this?

An engineering triumph, seventy years in the making



We can train these models. We still cannot say *why* they work.

Practical

better algorithms, more efficient models, less compute and energy burnt

Social

interpretability, fairness, robustness

Scientific

this thesis goes to the *École Doctorale de Physique*: understanding a complex system is a goal in itself

What does a physicist do about it?

Why the classical theory does not apply

- over-parametrization, no bias–variance trade-off
- non-convex landscape, no closed-form solution
- . . . yet plain SGD generalizes

So we build a caricature

Incomplete, but *concrete enough* to be tested and refuted.

We want to be Thomson, so that someone can be Bohr after us.

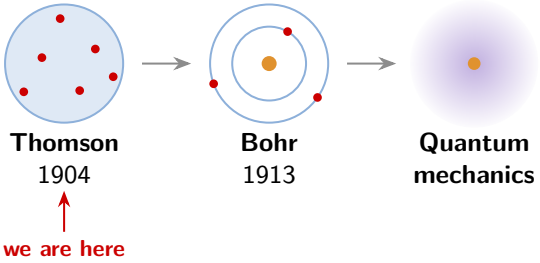
What does a physicist do about it?

Why the classical theory does not apply

- over-parametrization, no bias–variance trade-off
- non-convex landscape, no closed-form solution
- . . . yet plain SGD generalizes

So we build a caricature

Incomplete, but *concrete enough* to be tested and refuted.



We want to be Thomson, so that someone can be Bohr after us.

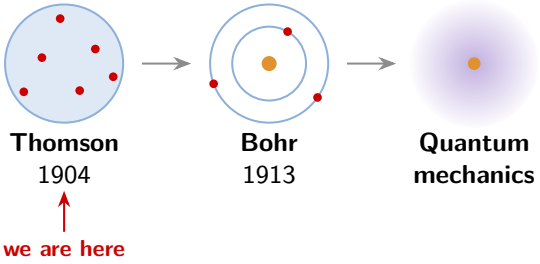
What does a physicist do about it?

Why the classical theory does not apply

- over-parametrization, no bias–variance trade-off
- non-convex landscape, no closed-form solution
- . . . yet plain SGD generalizes

So we build a caricature

Incomplete, but *concrete enough* to be tested and refuted.



We want to be Thomson, so that someone can be Bohr after us.

Our setting

Supervised Learning with scalar labels

$\mathcal{D} = \{(\boldsymbol{x}^\nu, y^\nu)\}_{\nu \in [n]}$ i.i.d., $\boldsymbol{x} \in \mathbb{R}^d$ $(d \gg 1)$, $y \in \mathbb{R}$ $\longrightarrow$ find f with $f(\boldsymbol{x}) \approx y$ on *unseen* data

The most controlled paradigm there is: a scalar label carries more signal than a class label.

We need to make some assumptions for our theory on:

The data

The model

The algorithm

Keep only what the phenomenon needs, strip everything else.

Our setting

Supervised Learning with scalar labels

$\mathcal{D} = \{(\boldsymbol{x}^\nu, y^\nu)\}_{\nu\in[n]}$ i.i.d., $\boldsymbol{x} \in \mathbb{R}^d$ $(d \gg 1)$, $y \in \mathbb{R}$ $\longrightarrow$ find f with $f(\boldsymbol{x}) \approx y$ on *unseen* data

The most controlled paradigm there is: a scalar label carries more signal than a class label.

We need to make some assumptions for our theory on:



The data



The model



The algorithm

Keep only what the phenomenon needs, strip everything else.

The multi-index models

Gaussian data

$$\boldsymbol{x} \sim \mathcal{N}(0, I_d)$$

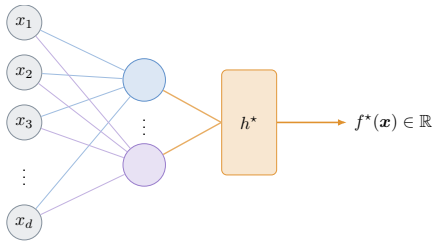
The target function

$$f^\star(\boldsymbol{x}) = h^\star\left(\frac{W^\star \boldsymbol{x}}{\sqrt{d}}\right)$$

$$h^\star : \mathbb{R}^k \to \mathbb{R}, \qquad W^\star = \{\boldsymbol{w}_r^\star\}_{r\in[k]} \in \mathbb{R}^{k\times d}$$

(Possible) label noise

$$y = f^\star(\boldsymbol{x}) + \sqrt{\Delta}\,\xi, \qquad \xi \sim \mathcal{N}(0,1)$$



Teacher subspace

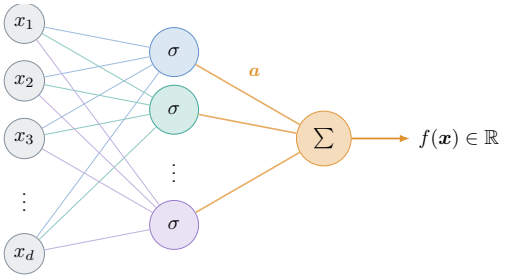
$$V^\star = \mathrm{span}(\boldsymbol{w}_1^\star, \dots, \boldsymbol{w}_k^\star),$$
$$\dim V^\star = k \ll d$$

Learning $\equiv$ *discovering* $V^\star$

These k directions *are* the features of the problem.

The learned model: 2-layer network

$$f(\boldsymbol{x};\, W, \boldsymbol{a}) = \frac{1}{p}\sum_{j=1}^{p} a_j\, \sigma\!\left(\frac{\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle}{\sqrt{d}}\right)$$



The student is a *special case* of the teacher, and it is a *two-layer NN*.

Loss function

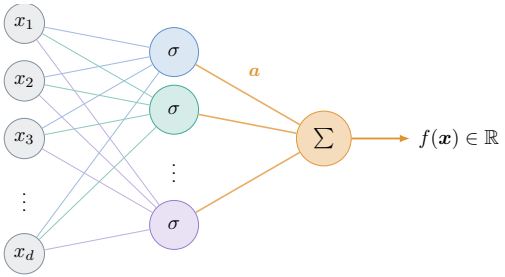
$$\ell(\hat{y}, y) = \tfrac{1}{2}(\hat{y} - y)^2$$

Population risk

$$\mathcal{R}(W, \boldsymbol{a}) = \mathbb{E}_{(\boldsymbol{x},\, y)}\Big[\tfrac{1}{2}\big(f(\boldsymbol{x};\, W, \boldsymbol{a}) - y\big)^2\Big]$$

The learned model: 2-layer network

$$f(\boldsymbol{x};\, W, \boldsymbol{a}) = \frac{1}{p}\sum_{j=1}^{p} a_j\, \sigma\!\left(\frac{\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle}{\sqrt{d}}\right)$$



The student is a *special case* of the teacher, and it is a *two-layer NN*.

Loss function

$$\ell(\hat{y}, y) = \tfrac{1}{2}(\hat{y} - y)^2$$

Population risk

$$\mathcal{R}(W, \boldsymbol{a}) = \mathbb{E}_{(\boldsymbol{x},\, y)}\Big[\tfrac{1}{2}\big(f(\boldsymbol{x};\, W, \boldsymbol{a}) - y\big)^2\Big]$$

The algorithm: one-pass SGD

Update $$\boldsymbol{w}_j^{\nu+1} = \boldsymbol{w}_j^{\nu} - \gamma \, \nabla_{\boldsymbol{w}_j} \ell(f(\boldsymbol{x}^{\nu}; W^{\nu}, \boldsymbol{a}), y^{\nu}) \,, \qquad j \in [p],$$

$\gamma > 0$ is the *learning rate*; each step draws a *fresh* sample, uses it once, and never reuses it

Initialization $$\boldsymbol{w}_j^0 \sim \mathcal{N}(0, I_d)$$

Steps = samples After n steps the algorithm has consumed exactly n samples: counting steps *is* counting samples.

The key property **Independence of the steps**

We actually often consider a spherically constrained dynamics for simplicity:

$$\boldsymbol{w}^{\nu+1} = \sqrt{d} \, \frac{\boldsymbol{w}^{\nu} - \gamma \nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}^{\nu}), y^{\nu})}{\left\| \boldsymbol{w}^{\nu} - \gamma \nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}^{\nu}), y^{\nu}) \right\|}$$

The algorithm: one-pass SGD

Update $$\boldsymbol{w}_j^{\nu+1} = \boldsymbol{w}_j^{\nu} - \gamma \, \nabla_{\boldsymbol{w}_j} \ell(f(\boldsymbol{x}^\nu; W^\nu, \boldsymbol{a}), y^\nu), \qquad j \in [p],$$

$\gamma > 0$ is the *learning rate*; each step draws a *fresh* sample, uses it once, and never reuses it

Initialization $$\boldsymbol{w}_j^0 \sim \mathcal{N}(0, I_d)$$

Steps = samples After n steps the algorithm has consumed exactly n samples: counting steps *is* counting samples.

The key property **Independence of the steps**

We actually often consider a spherically constrained dynamics for simplicity:

$$\boldsymbol{w}^{\nu+1} = \sqrt{d} \, \frac{\boldsymbol{w}^\nu - \gamma \nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}^\nu), y^\nu)}{\left\| \boldsymbol{w}^\nu - \gamma \nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}^\nu), y^\nu) \right\|}$$

The reduced dynamic

Pre-activations are Gaussian

$$\boldsymbol{\lambda} = \frac{W\boldsymbol{x}}{\sqrt{d}} \in \mathbb{R}^p, \qquad \boldsymbol{\lambda}^\star = \frac{W^\star \boldsymbol{x}}{\sqrt{d}} \in \mathbb{R}^k$$

$$f(\boldsymbol{x}; W, \boldsymbol{a}) = h(\boldsymbol{\lambda}, \boldsymbol{a}), \qquad h(\boldsymbol{\lambda}, \boldsymbol{a}) = \frac{1}{p}\sum_{j=1}^{p} a_j\, \sigma(\lambda_j)$$

Sufficient low-dimensional statistics

$$Q_{jl} = \frac{\langle \boldsymbol{w}_j, \boldsymbol{w}_l \rangle}{d}, \qquad M_{jr} = \frac{\langle \boldsymbol{w}_j, \boldsymbol{w}_r^\star \rangle}{d}, \qquad P_{rs} = \frac{\langle \boldsymbol{w}_r^\star, \boldsymbol{w}_s^\star \rangle}{d}$$

$$(\boldsymbol{\lambda}, \boldsymbol{\lambda}^\star) \sim \mathcal{N}(0, \Omega), \qquad \Omega = \begin{pmatrix} Q & M \\ M^\top & P \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)}$$

Population risk is d-independent

$$\mathcal{R}(\Omega, \boldsymbol{a}) = \mathbb{E}_{(\boldsymbol{\lambda}, \boldsymbol{\lambda}^\star)\sim\mathcal{N}(0,\Omega)}\left[\tfrac{1}{2}\big(h(\boldsymbol{\lambda}, \boldsymbol{a}) - h^\star(\boldsymbol{\lambda}^\star)\big)^2\right]$$

The reduced dynamic: update rule

The update rule is self-closed too!*

In the weights

$$\begin{aligned} \boldsymbol{w}_j^{\nu+1} &= \boldsymbol{w}_j^{\nu} + \frac{\gamma}{p\sqrt{d}} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right) a_j \sigma'(\lambda_j^{\nu}) \boldsymbol{x}^{\nu}, \\ a_j^{\nu+1} &= a_j^{\nu} + \frac{\gamma}{pd} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right) \sigma(\lambda_j^{\nu}) \end{aligned}$$

$\Longrightarrow$

In the sufficient statistics

$$\begin{aligned} Q_{jl}^{\nu+1} &= Q_{jl}^{\nu} + \frac{\gamma}{pd} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right) \left(a_l \sigma'(\lambda_l^{\nu}) \lambda_j^{\nu} + a_j \sigma'(\lambda_j^{\nu}) \lambda_l^{\nu}\right) + \\ &\quad + \frac{\gamma}{pd} \frac{\gamma}{p} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right)^2 a_j a_l \sigma'(\lambda_j^{\nu}) \sigma'(\lambda_l^{\nu}) \frac{\|\boldsymbol{x}^{\nu}\|^2}{d}, \\ M_{jr}^{\nu+1} &= M_{jr}^{\nu} + \frac{\gamma}{pd} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right) a_j \sigma'(\lambda_j^{\nu}) \lambda_r^{\star,\nu}, \\ a_j^{\nu+1} &= a_j^{\nu} + \frac{\gamma}{pd} \left(y^{\nu}(\boldsymbol{\lambda}^{\star,\nu}) - h(\boldsymbol{\lambda}^{\nu})\right) \sigma(\lambda_j^{\nu}). \end{aligned}$$

sample $\boldsymbol{x}^{\nu} \sim \mathcal{N}(0, I_d)$	$\Longrightarrow$	sample $(\boldsymbol{\lambda}^{\nu}, \boldsymbol{\lambda}^{\star,\nu}) \sim \mathcal{N}(0, \Omega)$
population risk $\mathcal{R}(W)$	$\Longrightarrow$	population risk $\mathcal{R}(Q, M, \boldsymbol{a})$

There is no need to work with W — let's look at Q, M, P

The magic of high-dimensional data

The stochastic process concentrates into a set of ODEs! [Saad & Solla, 1995]

$$\frac{\mathrm{d}Q_{jl}}{\mathrm{d}t} = \Phi^Q_{jl}(\Omega, \boldsymbol{a}) + \frac{\gamma}{p}\,\Psi^Q_{jl}(\Omega, \boldsymbol{a}),$$
$$\frac{\mathrm{d}M_{jr}}{\mathrm{d}t} = \Phi^M_{jr}(\Omega, \boldsymbol{a}),$$
$$\frac{\mathrm{d}a_j}{\mathrm{d}t} = \Phi^a_j(\Omega, \boldsymbol{a}).$$

The time-step

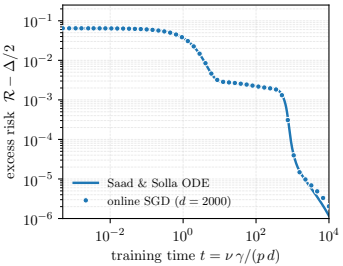
$$\delta t = \frac{\gamma}{pd}$$

infinitesimal in high dimension: $\nu \longmapsto t = \nu\,\frac{\gamma}{pd}$

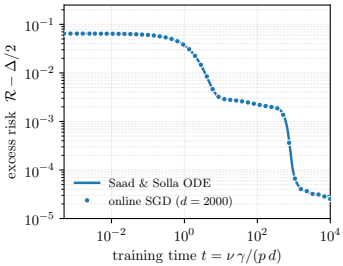
Drift Φ and noise Ψ expectation over $(\boldsymbol{\lambda}, \boldsymbol{\lambda}^\star) \sim \mathcal{N}(0, \Omega)$

$$\Phi^Q_{jl} = \mathbb{E}\left[\left(h^\star(\boldsymbol{\lambda}^\star) - h(\boldsymbol{\lambda})\right)\left(a_l\sigma'(\lambda_l)\lambda_j + a_j\sigma'(\lambda_j)\lambda_l\right)\right], \qquad \Phi^M_{jr} = \mathbb{E}\left[\left(h^\star(\boldsymbol{\lambda}^\star) - h(\boldsymbol{\lambda})\right)a_j\sigma'(\lambda_j)\lambda^\star_r\right],$$
$$\Psi^Q_{jl} = \mathbb{E}\left[\left(\left(h^\star(\boldsymbol{\lambda}^\star) - h(\boldsymbol{\lambda})\right)^2 + \Delta\right)a_j a_l\sigma'(\lambda_j)\sigma'(\lambda_l)\right], \qquad \Phi^a_j = \mathbb{E}\left[\left(h^\star(\boldsymbol{\lambda}^\star) - h(\boldsymbol{\lambda})\right)\sigma(\lambda_j)\right].$$

The ODEs are exact in high dimension



$\Delta = 0$: the excess risk falls to zero

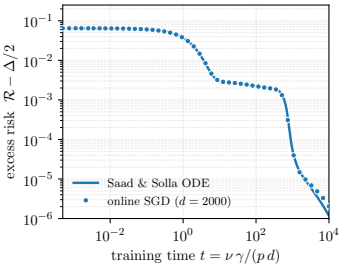


$\Delta = 10^{-3}$: the noise term Ψ pins a plateau

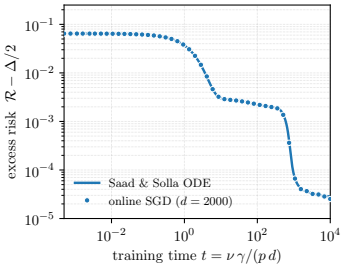
The actual bound [Goldt, Advani, Saxe, Krzakala & Zdeborová, 2019]

$$\max_{0 \le \nu \le Td} \mathbb{E}\left[\left\|(\Omega^\nu, \boldsymbol{a}^\nu) - \left(\tilde{\Omega}\left(\frac{\nu\gamma}{pd}\right), \tilde{\boldsymbol{a}}\left(\frac{\nu\gamma}{pd}\right)\right)\right\|\right] \le \frac{C(T)}{\sqrt{d}}$$

The ODEs are exact in high dimension



$\Delta = 0$: the excess risk falls to zero



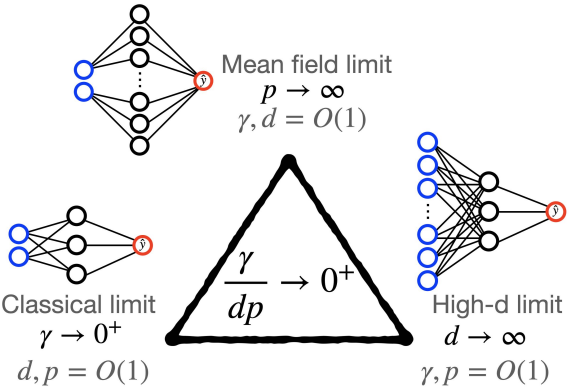
$\Delta = 10^{-3}$: the noise term Ψ pins a plateau

The actual bound [Goldt, Advani, Saxe, Krzakala & Zdeborová, 2019]

$$\max_{0 \le \nu \le Td} \mathbb{E}\left[\left\|(\Omega^\nu, \boldsymbol{a}^\nu) - \left(\tilde{\Omega}\left(\frac{\nu\gamma}{pd}\right), \tilde{\boldsymbol{a}}\left(\frac{\nu\gamma}{pd}\right)\right)\right\|\right] \le \frac{C(T)}{\sqrt{d}}$$

$\delta t = \dfrac{\gamma}{pd}$ d to infinity is only a third of the story though... [Veiga et al., 2022]

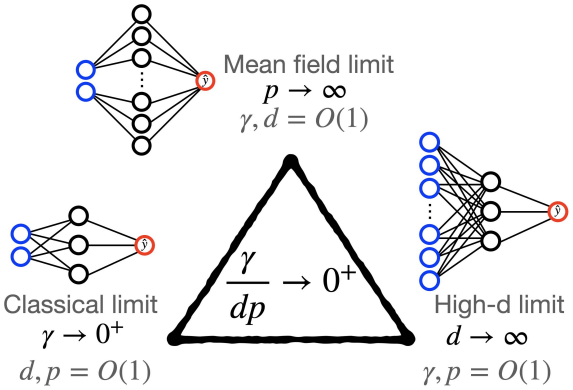
We can have other limits leading to the same equations!



Only the drift survives in other limits: $\mathrm{d}Q/\mathrm{d}t = \Phi^Q + (\gamma/p)\,\Psi^Q$

$\delta t = \dfrac{\gamma}{pd}$ d to infinity is only a third of the story though... [Veiga et al., 2022]

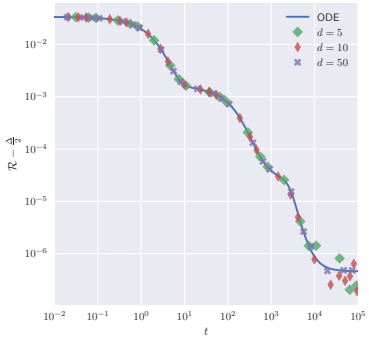
We can have other limits leading to the same equations!



Only the drift survives in other limits: $\mathrm{d}Q/\mathrm{d}t = \Phi^Q + (\gamma/p)\,\Psi^Q$

Classical limit: gradient flow

$\gamma \to 0^+$ at fixed d, $p = \mathcal{O}(1)$



same initial condition, $d = 5,\ 10,\ 50$

Gradient flow on the weights [Robbins & Monro, 1951]

$$\frac{\mathrm{d}\boldsymbol{w}_j(t)}{\mathrm{d}t} = -\nabla_{\boldsymbol{w}_j}\mathcal{R}(W(t)) \qquad p(d+1) \text{ parameters}$$

$\downarrow$ chain rule

$$\frac{\mathrm{d}Q_{jl}}{\mathrm{d}t} = \frac{1}{d}\left(\left\langle \boldsymbol{w}_j, \frac{\mathrm{d}\boldsymbol{w}_l}{\mathrm{d}t}\right\rangle + \left\langle \frac{\mathrm{d}\boldsymbol{w}_j}{\mathrm{d}t}, \boldsymbol{w}_l\right\rangle\right)$$

$\downarrow$

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \Phi^Q, \qquad \frac{\mathrm{d}M}{\mathrm{d}t} = \Phi^M \qquad p(p+k) \text{ parameters}$$

No d in the equations: same initial overlaps, same trajectory.

Mean field limit

$p \to \infty$ at fixed d, $\gamma = \mathcal{O}(1)$

The network sees only the average — so every neuron feels the same field

$$\hat{\mu}_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\boldsymbol{w}_i}, \qquad \hat{f}_\mu(\boldsymbol{x}) = \int \sigma(\boldsymbol{w}^\top \boldsymbol{x})\, \mathrm{d}\mu(\boldsymbol{w}) \qquad \varphi(\boldsymbol{w}, \mu) = \mathbb{E}_{\boldsymbol{x}}\left[\sigma'(\boldsymbol{w}^\top \boldsymbol{x})\, \boldsymbol{x}\left(f^\star(\boldsymbol{x}) - \hat{f}_\mu(\boldsymbol{x})\right)\right]$$

no pairwise interaction to track: one common velocity field transports the whole density

$$\partial_t \mu_t = \nabla_{\boldsymbol{w}} \cdot \left(\mu_t\, \varphi(\cdot, \mu_t)\right)$$

mass conservation, nothing more [Mei, Montanari & Nguyen 2018; Chizat & Bach 2018]

. . . but μ still lives on $\mathbb{R}^d$.

Mean field in the sufficient statistics

$Q, M \in \mathbb{R}^{p\times(p+k)}$ grow with p: intractable as $p \to \infty$

Split along the teacher subspace

$$W = MP^{-1}W^\star + W^\perp$$
$$Q^\perp = {}^1\!/_d W^\perp (W^\perp)^\top = Q - MP^{-1}M^\top$$

Ansatz

$$\boldsymbol{w}_i^\perp \approx \sqrt{d\,q_i}\,\boldsymbol{g}_i, \qquad \boldsymbol{g}_i \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$
$$\tilde{Q} = MP^{-1}M^\top + D_{\sqrt{q}}\,\Xi\,D_{\sqrt{q}}$$

$$\Xi_{ii} = 1, \quad \Xi_{ij} = \langle \boldsymbol{g}, \boldsymbol{g}' \rangle$$

$k+1$ numbers per neuron

$$\tilde{\mu}_p(t) = \frac{1}{p}\sum_{i=1}^{p} \delta_{(\boldsymbol{m}_i(t),\, q_i(t))}, \qquad (\boldsymbol{m}, q) \in \mathbb{R}^{k+1}$$

carried by the Saad & Solla drift

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mathbb{E}_\Xi[\Phi^M(\tilde{\Omega})], \qquad \frac{\mathrm{d}q_i}{\mathrm{d}t} = \mathbb{E}_\Xi[\Phi^\perp_{ii}(\tilde{\Omega})]$$

same mass conservation, on $\mathbb{R}^{k+1}$

$$\partial_t \tilde{\mu}_t = \nabla_{(\boldsymbol{m},q)} \cdot \left(\tilde{\mu}_t\, \tilde{\varphi}(\cdot, \tilde{\mu}_t)\right), \qquad \tilde{\varphi} = (\Phi^M, \Phi^\perp)$$

the risk is exact up to $\mathcal{O}(1/\sqrt{p})$

Mean field in the sufficient statistics

$Q, M \in \mathbb{R}^{p\times(p+k)}$ grow with p: intractable as $p \to \infty$

Split along the teacher subspace

$$W = MP^{-1}W^\star + W^\perp$$
$$Q^\perp = {}^1\!/_d W^\perp (W^\perp)^\top = Q - MP^{-1}M^\top$$

Ansatz

$$\boldsymbol{w}_i^\perp \approx \sqrt{d\,q_i}\,\boldsymbol{g}_i, \qquad \boldsymbol{g}_i \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$
$$\tilde{Q} = MP^{-1}M^\top + D_{\sqrt{q}}\,\Xi\,D_{\sqrt{q}}$$

$$\Xi_{ii} = 1, \quad \Xi_{ij} = \langle \boldsymbol{g}, \boldsymbol{g}' \rangle$$

$k+1$ numbers per neuron

$$\tilde{\mu}_p(t) = \frac{1}{p}\sum_{i=1}^{p} \delta_{(\boldsymbol{m}_i(t),\, q_i(t))}, \qquad (\boldsymbol{m}, q) \in \mathbb{R}^{k+1}$$

carried by the Saad & Solla drift

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mathbb{E}_\Xi[\Phi^M(\tilde{\Omega})], \qquad \frac{\mathrm{d}q_i}{\mathrm{d}t} = \mathbb{E}_\Xi[\Phi^\perp_{ii}(\tilde{\Omega})]$$

same mass conservation, on $\mathbb{R}^{k+1}$

$$\partial_t \tilde{\mu}_t = \nabla_{(\boldsymbol{m},q)} \cdot (\tilde{\mu}_t\, \tilde{\varphi}(\cdot, \tilde{\mu}_t)), \qquad \tilde{\varphi} = (\Phi^M, \Phi^\perp)$$

the risk is exact up to $\mathcal{O}(1/\sqrt{p})$

Mean field in the sufficient statistics

$Q, M \in \mathbb{R}^{p\times(p+k)}$ grow with p: intractable as $p \to \infty$

Split along the teacher subspace

$$W = MP^{-1}W^\star + W^\perp$$
$$Q^\perp = {}^1\!/_d W^\perp (W^\perp)^\top = Q - MP^{-1}M^\top$$

Ansatz

$$\boldsymbol{w}_i^\perp \approx \sqrt{d\,q_i}\,\boldsymbol{g}_i, \qquad \boldsymbol{g}_i \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$
$$\tilde{Q} = MP^{-1}M^\top + D_{\sqrt{q}}\,\Xi\,D_{\sqrt{q}}$$

$$\Xi_{ii} = 1, \quad \Xi_{ij} = \langle \boldsymbol{g}, \boldsymbol{g}' \rangle$$

$k+1$ numbers per neuron

$$\tilde{\mu}_p(t) = \frac{1}{p}\sum_{i=1}^{p} \delta_{(\boldsymbol{m}_i(t),\, q_i(t))}, \qquad (\boldsymbol{m}, q) \in \mathbb{R}^{k+1}$$

carried by the Saad & Solla drift

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mathbb{E}_\Xi\big[\Phi^M(\tilde{\Omega})\big], \qquad \frac{\mathrm{d}q_i}{\mathrm{d}t} = \mathbb{E}_\Xi\big[\Phi^\perp_{ii}(\tilde{\Omega})\big]$$

same mass conservation, on $\mathbb{R}^{k+1}$

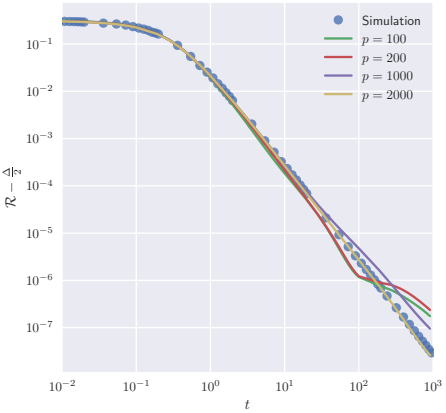
$$\partial_t \tilde{\mu}_t = \nabla_{(\boldsymbol{m},q)} \cdot \big(\tilde{\mu}_t\, \tilde{\varphi}(\cdot, \tilde{\mu}_t)\big), \qquad \tilde{\varphi} = (\Phi^M, \Phi^\perp)$$

the risk is exact up to $\mathcal{O}(1/\sqrt{p})$

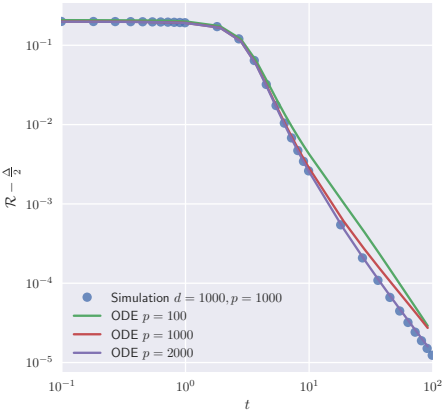
High-dimensional Mean field limit

Ξ converges to I_p when $d \to \infty$: off-diagonal orthogonal terms are 0!

they are $\mathcal{O}(1/\sqrt{d})$, and so is the error they make on the drift



$d = 5$



$d = 1000$

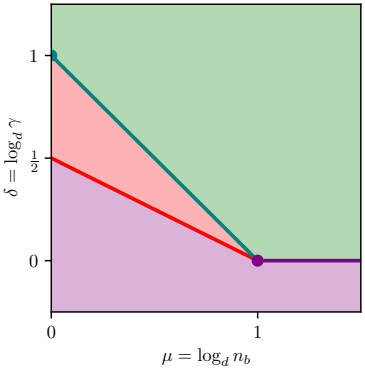
What if we scale the rate and the batch size?

$$\gamma = \gamma_0 d^{-\delta} \quad \text{and} \quad n_b = n_0 d^{\mu}$$

Equations coincide with the original ones, but the factors activate differently:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \Phi^Q + \frac{\gamma}{p n_b}\Psi^Q$$

$h^\star = \sigma = \mathrm{erf}$ with $p = k = 2$

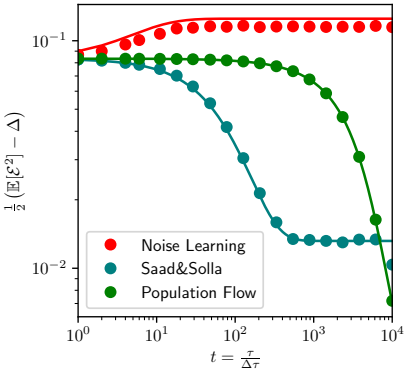


Population flow

Noise learning

Saad & Solla line

Dynamics not defined



Can we use SS for sample complexity?

take the simplest case, $p = k = 1$: one single statistic carries the whole dynamics

$$\frac{\mathrm{d}m}{\mathrm{d}t} = (1 - m^2)\, C'(m), \qquad \mathcal{R}(m) = \mathrm{const} - C(m), \qquad C(m) = \mathbb{E}[h^\star(\lambda^\star)\, \sigma(\lambda)]$$

Weak recovery time

$$m = \frac{\langle \boldsymbol{w}, \boldsymbol{w}^\star \rangle}{d} \geq \eta \qquad m(t_{\mathrm{weak}}) = \eta$$

In high dimension, search is the only hard part

leading order: $\dfrac{\mathrm{d}m}{\mathrm{d}t} \propto m^{\ell-1}, \quad m(0) \sim \dfrac{1}{\sqrt{d}}$

. . . so we should be able to read the sample complexity off the Saad & Solla equations.

We can't: the guarantee only covers a constant time horizon.

actually the error is $Ce^{CT}\sqrt{\dfrac{\gamma}{pd}}$, at best $T \sim \log d$

Can we use SS for sample complexity?

take the simplest case, $p = k = 1$: one single statistic carries the whole dynamics

$$\frac{\mathrm{d}m}{\mathrm{d}t} = (1 - m^2)\, C'(m), \qquad \mathcal{R}(m) = \mathrm{const} - C(m), \qquad C(m) = \mathbb{E}[h^\star(\lambda^\star)\, \sigma(\lambda)]$$

Weak recovery time

$$m = \frac{\langle \boldsymbol{w}, \boldsymbol{w}^\star \rangle}{d} \geq \eta \qquad m(t_{\mathrm{weak}}) = \eta$$

In high dimension, search is the only hard part

leading order: $\dfrac{\mathrm{d}m}{\mathrm{d}t} \propto m^{\ell-1}, \quad m(0) \sim \dfrac{1}{\sqrt{d}}$

. . . so we should be able to read the sample complexity off the Saad & Solla equations.

We can't: the guarantee only covers a constant time horizon.

actually the error is $Ce^{CT}\sqrt{\dfrac{\gamma}{pd}}$, at best $T \sim \log d$

Can we use SS for sample complexity?

take the simplest case, $p = k = 1$: one single statistic carries the whole dynamics

$$\frac{\mathrm{d}m}{\mathrm{d}t} = (1 - m^2)\, C'(m), \qquad \mathcal{R}(m) = \mathrm{const} - C(m), \qquad C(m) = \mathbb{E}[h^\star(\lambda^\star)\, \sigma(\lambda)]$$

Weak recovery time

$$m = \frac{\langle \boldsymbol{w}, \boldsymbol{w}^\star \rangle}{d} \geq \eta \qquad m(t_{\mathrm{weak}}) = \eta$$

In high dimension, search is the only hard part

leading order: $\dfrac{\mathrm{d}m}{\mathrm{d}t} \propto m^{\ell-1}, \quad m(0) \sim \dfrac{1}{\sqrt{d}}$

. . . so we should be able to read the sample complexity off the Saad & Solla equations.

We can't: the guarantee only covers a constant time horizon.

actually the error is $Ce^{CT}\sqrt{\dfrac{\gamma}{pd}}$, at best $T \sim \log d$

The information exponent [Ben Arous, Gheissari & Jagannath, 2021]

Information Exponent

$$\ell = \min\Big\{ j \geq 1 : a_j = \frac{1}{j!}\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[h^\star(z)\,\mathrm{He}_j(z)] \neq 0 \Big\}$$

$h^\star(z) = z^2$ gives $\ell = 2$; $\quad h^\star(z) = \mathrm{He}_j(z)$ gives $\ell = j$

The guiding quantity for everything that follows.

Theorem

The sample complexity is

$$n = \begin{cases} \mathcal{O}(d) & \ell = 1, \\ \mathcal{O}(d \log d) & \ell = 2, \\ \mathcal{O}(d^{\ell-1}) & \ell \geq 3. \end{cases}$$

It is not the whole story

Beyond one direction: the *leap* exponent carries it to multi-index targets, $k > 1$ [Abbe et al., 2023]

Not the floor either: a partial-trace algorithm reaches $\mathcal{O}(d^{\ell/2})$ [Anandkumar et al., 2017]

The information exponent [Ben Arous, Gheissari & Jagannath, 2021]

Information Exponent

$$\ell = \min\Big\{ j \geq 1 : a_j = \frac{1}{j!}\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[h^\star(z)\,\mathrm{He}_j(z)] \neq 0 \Big\}$$

$h^\star(z) = z^2$ gives $\ell = 2$; $\quad h^\star(z) = \mathrm{He}_j(z)$ gives $\ell = j$

The guiding quantity for everything that follows.

Theorem

The sample complexity is

$$n = \begin{cases} \mathcal{O}(d) & \ell = 1, \\ \mathcal{O}(d \log d) & \ell = 2, \\ \mathcal{O}(d^{\ell-1}) & \ell \geq 3. \end{cases}$$

It is not the whole story

Beyond one direction: the *leap* exponent carries it to multi-index targets, $k > 1$ [Abbe et al., 2023]

Not the floor either: a partial-trace algorithm reaches $\mathcal{O}(d^{\ell/2})$ [Anandkumar et al., 2017]

The information exponent [Ben Arous, Gheissari & Jagannath, 2021]

Information Exponent

$$\ell = \min\Big\{ j \geq 1 : a_j = \frac{1}{j!}\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[h^\star(z)\,\mathrm{He}_j(z)] \neq 0 \Big\}$$

$h^\star(z) = z^2$ gives $\ell = 2$; $\quad h^\star(z) = \mathrm{He}_j(z)$ gives $\ell = j$

The guiding quantity for everything that follows.

Theorem

The sample complexity is

$$n = \begin{cases} \mathcal{O}(d) & \ell = 1, \\ \mathcal{O}(d \log d) & \ell = 2, \\ \mathcal{O}(d^{\ell-1}) & \ell \geq 3. \end{cases}$$

It is not the whole story

Beyond one direction: the *leap* exponent carries it to multi-index targets, $k > 1$ [Abbe et al., 2023]

Not the floor either: a partial-trace algorithm reaches $\mathcal{O}(d^{\ell/2})$ [Anandkumar et al., 2017]

Is the IE the actual lower bound in complexity?

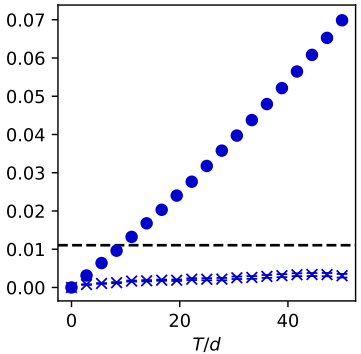
Real data is not i.i.d. and it is repeated — our current model has neither property.

Displaced gradient (EGD)

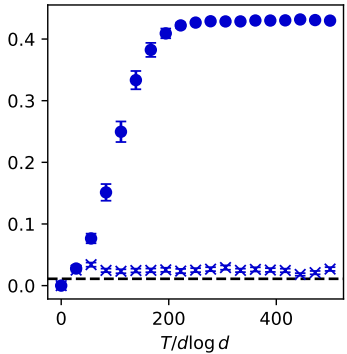
Gradient computed in a neighborhood of the current weight

$$w_{\tau+1} = w_\tau - \gamma \nabla_w \mathcal{L}(z_\tau, y_\tau; \tilde{w}_\tau)$$
$$\tilde{w}_\tau = w_\tau - \rho \nabla_w \mathcal{L}(z_\tau, y_\tau; w_\tau)$$

γ and ρ are two different learning rates



$h^\star(z) = \mathrm{He}_3(z)$



$h^\star(z) = \mathrm{He}_4(z)$

The trick breaks the curse of the information exponent

How does this work?

One-pass SGD is linear in the label $\nabla_{\boldsymbol{w}} \mathcal{R} = -\,\mathbb{E}\left[y\,\sigma'(\lambda)\,\dfrac{\boldsymbol{x}}{\sqrt{d}}\right]$

What if we first transform the label, and query $\mathbb{E}\left[g(y)\,\sigma'(\lambda)\,\frac{\boldsymbol{x}}{\sqrt{d}}\right]$?

Generative Exponent [Damian, Pillaud-Vivien, Lee & Bruna, 2024]

$$\ell^\star = \min\Big\{ j \geq 1 : \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[g(h^\star(z))\,\mathrm{He}_j(z)\big] \neq 0 \text{ for some } g : \mathbb{R} \to \mathbb{R} \Big\}$$

He_{2k+1}
$\ell = 2k+1,\ \ell^\star = 1$

He_{2k}
$\ell = 2k,\ \ell^\star = 2$

$\mathrm{sign}(z_1 z_2 z_3)$
$\ell = \ell^\star = 3$

$\ell^\star \leq \ell$ **always, and often far smaller.**

How does this work?

One-pass SGD is linear in the label $\nabla_{\boldsymbol{w}} \mathcal{R} = -\,\mathbb{E}\left[y\,\sigma'(\lambda)\,\dfrac{\boldsymbol{x}}{\sqrt{d}}\right]$

What if we first transform the label, and query $\mathbb{E}\left[g(y)\,\sigma'(\lambda)\,\frac{\boldsymbol{x}}{\sqrt{d}}\right]$?

Generative Exponent [Damian, Pillaud-Vivien, Lee & Bruna, 2024]

$$\ell^\star = \min\Big\{ j \geq 1 : \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[g(h^\star(z))\,\mathrm{He}_j(z)\big] \neq 0 \ \textit{for some}\ g : \mathbb{R} \to \mathbb{R} \Big\}$$

He_{2k+1}	He_{2k}	$\mathrm{sign}(z_1 z_2 z_3)$
$\ell = 2k+1,\ \ell^\star = 1$	$\ell = 2k,\ \ell^\star = 2$	$\ell = \ell^\star = 3$

$\ell^\star \leq \ell$ always, and often far smaller.

Why does the displaced gradient reach it?

The double gradient induces a transformation

$$\nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}; \tilde{\boldsymbol{w}}), y) = -\, y\, \sigma'(\tilde{\lambda})\, \frac{\boldsymbol{x}}{\sqrt{d}}, \qquad \tilde{\lambda} = \frac{\langle \tilde{\boldsymbol{w}}, \boldsymbol{x} \rangle}{\sqrt{d}} = \lambda + \rho\, y\, \sigma'(\lambda)$$

expand σ' around λ:

$$\nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}; \tilde{\boldsymbol{w}}), y) \approx -\, \frac{\boldsymbol{x}}{\sqrt{d}} \left(y\, \sigma'(\lambda) + \rho\, y^2\, \sigma'(\lambda) \sigma''(\lambda) + \frac{\rho^2}{2}\, y^3\, \sigma'(\lambda)^2 \sigma'''(\lambda) + \cdots \right)$$

Powers of y appear on their own: the network builds g for free.

Theorem (Repetita Iuvant)

Sample complexity of EGD

$$n = \begin{cases} \mathcal{O}(d) & \ell^\star = 1, \\ \mathcal{O}(d \log d) & \ell^\star = 2. \end{cases}$$

no results for $\ell^\star > 2$: we worked on polynomials

Why does the displaced gradient reach it?

The double gradient induces a transformation

$$\nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}; \tilde{\boldsymbol{w}}), y) = -\, y\, \sigma'(\tilde{\lambda})\, \frac{\boldsymbol{x}}{\sqrt{d}}, \qquad \tilde{\lambda} = \frac{\langle \tilde{\boldsymbol{w}}, \boldsymbol{x} \rangle}{\sqrt{d}} = \lambda + \rho\, y\, \sigma'(\lambda)$$

expand σ' around λ:

$$\nabla_{\boldsymbol{w}} \ell(f(\boldsymbol{x}; \tilde{\boldsymbol{w}}), y) \approx -\, \frac{\boldsymbol{x}}{\sqrt{d}} \left(y\, \sigma'(\lambda) + \rho\, y^2\, \sigma'(\lambda) \sigma''(\lambda) + \frac{\rho^2}{2}\, y^3\, \sigma'(\lambda)^2 \sigma'''(\lambda) + \cdots \right)$$

Powers of y appear on their own: the network builds g for free.

Theorem (Repetita Iuvant)

Sample complexity of EGD

$$n = \begin{cases} \mathcal{O}(d) & \ell^\star = 1, \\ \mathcal{O}(d \log d) & \ell^\star = 2. \end{cases}$$

no results for $\ell^\star > 2$: we worked on polynomials

IE & Attention model

Our proxy for attention: $f(X) = \boldsymbol{a}_L^\top \mathrm{softmax}\left(\left(X + \frac{P}{\sqrt{d}}\right)\boldsymbol{w}\boldsymbol{w}^\top\left(X + \frac{P}{\sqrt{d}}\right)^\top\right)\boldsymbol{a}_R \quad X \in \mathbb{R}^{L\times d}$

Sufficient statistics: $m = \boldsymbol{w}_\star^\top \boldsymbol{w}/d$ and $\boldsymbol{e} = P\boldsymbol{w}/\sqrt{d}$

Sequence Information Exponent (SIE)

$$SIE(\boldsymbol{f}_{\boldsymbol{w}_\star}^{\mathrm{SSI}}) \coloneqq \min\left\{\sum_{l=1}^{L} k_l > 0 : \boldsymbol{k} \in \mathbb{N}^L,\ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I_L)}\left[\left(\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_l)\right)h^\star(\boldsymbol{z})\right] \neq 0\right\}$$

Examples:

- $z_{\star,1} + \cdots + z_{\star,L}$ has SIE $= 1$;
- $z_{\star,1}z_{\star,2}$ has SIE $= 2$;
- $\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_{\star,l})$ has SIE $= \sum_{l=1}^{L} k_l$.

Theorem

Sample complexity of SGD

$$\begin{cases} \mathcal{O}_L(d) & \text{if } SIE = 1 \\ \mathcal{O}_L(d\log d) & \text{if } SIE = 2 \\ \mathcal{O}_L(d^{SIE-1}) & \text{if } SIE \geq 3 \end{cases}$$

The same table as before, with SIE in place of ℓ.

IE & Attention model

Our proxy for attention: $f(X) = \boldsymbol{a}_L^\top \mathrm{softmax}\left(\left(X + \frac{P}{\sqrt{d}}\right)\boldsymbol{w}\boldsymbol{w}^\top\left(X + \frac{P}{\sqrt{d}}\right)^\top\right)\boldsymbol{a}_R \quad X \in \mathbb{R}^{L\times d}$

Sufficient statistics: $m = \boldsymbol{w}_\star^\top \boldsymbol{w}/d$ and $\boldsymbol{e} = P\boldsymbol{w}/\sqrt{d}$

Sequence Information Exponent (SIE)

$$SIE(f_{\boldsymbol{w}_\star}^{\mathrm{SSI}}) \coloneqq \min\left\{\sum_{l=1}^{L} k_l > 0 : \boldsymbol{k} \in \mathbb{N}^L,\ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I_L)}\left[\left(\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_l)\right)h^\star(\boldsymbol{z})\right] \neq 0\right\}$$

Examples:

- $z_{\star,1} + \cdots + z_{\star,L}$ has SIE $= 1$;
- $z_{\star,1}z_{\star,2}$ has SIE $= 2$;
- $\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_{\star,l})$ has SIE $= \sum_{l=1}^{L} k_l$.

Theorem

Sample complexity of SGD

$$\begin{cases} \mathcal{O}_L(d) & \text{if } SIE = 1 \\ \mathcal{O}_L(d\log d) & \text{if } SIE = 2 \\ \mathcal{O}_L(d^{SIE-1}) & \text{if } SIE \geq 3 \end{cases}$$

The same table as before, with SIE in place of ℓ.

IE & Attention model

Our proxy for attention: $f(X) = \boldsymbol{a}_L^\top \mathrm{softmax}\left(\left(X + \frac{P}{\sqrt{d}}\right)\boldsymbol{w}\boldsymbol{w}^\top\left(X + \frac{P}{\sqrt{d}}\right)^\top\right)\boldsymbol{a}_R \quad X \in \mathbb{R}^{L\times d}$

Sufficient statistics: $m = \boldsymbol{w}_\star^\top \boldsymbol{w}/d$ and $\boldsymbol{e} = P\boldsymbol{w}/\sqrt{d}$

Sequence Information Exponent (SIE)

$$SIE(f_{\boldsymbol{w}_\star}^{\mathrm{SSI}}) \coloneqq \min\left\{\sum_{l=1}^{L} k_l > 0 : \boldsymbol{k} \in \mathbb{N}^L,\ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I_L)}\left[\left(\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_l)\right)h^\star(\boldsymbol{z})\right] \neq 0\right\}$$

Examples:

- $z_{\star,1} + \cdots + z_{\star,L}$ has SIE $= 1$;
- $z_{\star,1}z_{\star,2}$ has SIE $= 2$;
- $\prod_{l=1}^{L}\mathrm{He}_{k_l}(z_{\star,l})$ has SIE $= \sum_{l=1}^{L} k_l$.

Theorem

Sample complexity of SGD

$$\begin{cases} \mathcal{O}_L(d) & \text{if } SIE = 1 \\ \mathcal{O}_L(d\log d) & \text{if } SIE = 2 \\ \mathcal{O}_L(d^{SIE-1}) & \text{if } SIE \geq 3 \end{cases}$$

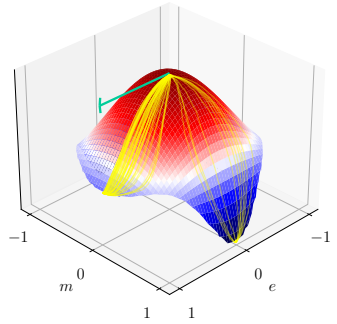
The same table as before, with SIE in place of ℓ.

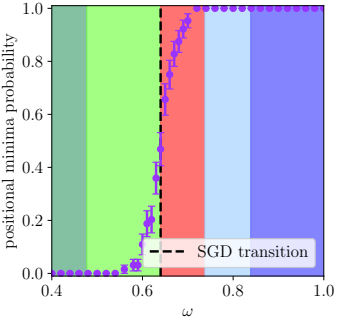
A nice extra: positional-semantic transition

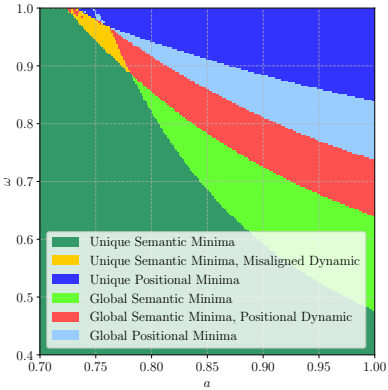
Target

$$\boldsymbol{f}^{\mathrm{SSI}}_{\boldsymbol{w}_\star}(X) = (1-\omega)\ \mathrm{softmax}\Big(X\boldsymbol{w}_\star\boldsymbol{w}_\star^\top X^\top\Big) + \omega\ \mathrm{softmax}\left[\begin{pmatrix} a^2 & -a^2 \\ -a^2 & a^2 \end{pmatrix}\right]$$

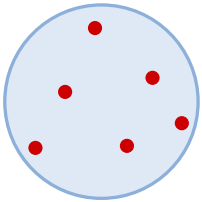
ω: positional vs semantic; a: positional strength







Recap



as a community, we have solved this model

The journey

Chapter 2 – one set of equations, three limits

Chapter 3 – phase retrieval: plateaus and prefactors

Chapter 4 – rate and batch size: time, not samples

Chapter 5 – reusing data: $\ell \longrightarrow \ell^\star$

Chapter 6 – sequences and attention: the SIE

We can say how it learns, and how much data it takes.

What is next? structured data + deeper networks $\longrightarrow$ low-degree filtering (LoFi)



Thanks to you and my amazing collaborators!



Florent



Lenka



Bruno



Ludovic



Yatin



Luca

Is the mean-field ansatz legitimate?

$$\boldsymbol{w}_i^{\perp} \approx \sqrt{d\, q_i}\, \boldsymbol{g}_i, \qquad \boldsymbol{g}_i \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$

Not a claim about the weights

This is *not* the law of $\boldsymbol{w}^{\perp}$, and the entries of $Q^{\perp}$ are certainly *not* independent.

A claim about the *equations*

It only influences the drift $\mathbb{E}_{\Xi}[\Phi]$, it is *equivalent* to full $Q^{\perp}$ description under the flow.

Consistency of the reduced description

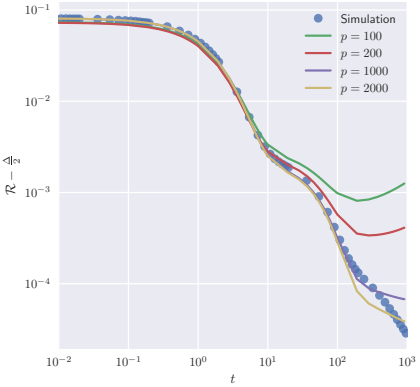
$$\sup_{t\le T} \left|\mathcal{R}(\Omega(t)) - \mathcal{R}(\tilde{\Theta}(t))\right| \;\le\; C e^{CT}\, \frac{\sqrt{\log(pT)} + z}{\sqrt{p}} \qquad \text{with probability } \ge 1 - e^{-z^2}$$

It constrains the update equations, not the weights.

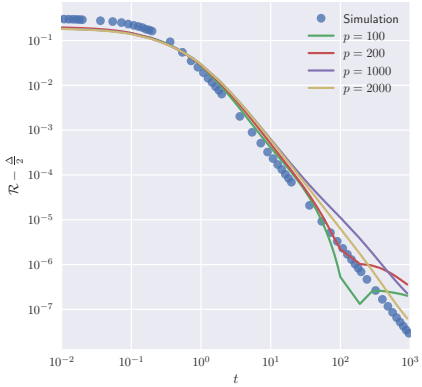
Without you get a mismatch!

Drop the off-diagonal of Ξ and the ODE stops matching

$k = 2$, $d = 5$, $\gamma = 1$, $\Delta = 0$; lines are the integrated ODEs, dots the simulated dynamics



$\sigma = \mathrm{erf}$: the gap opens with t

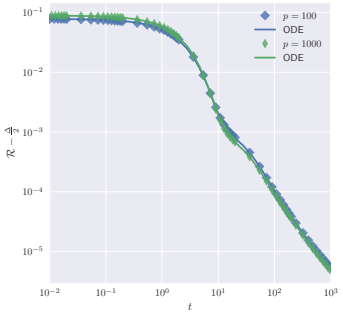


$\sigma = z^2$: wrong already at $t = 0$

$\mathbb{E}_\Xi$ is not decorative.

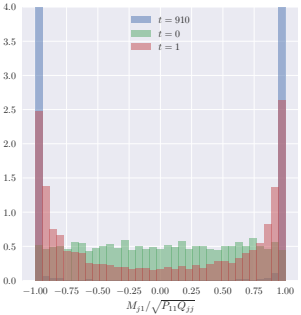
Mean field: where the neurons actually end up

$Q^{\perp} = 0$ is a fixed point



$k = 2$, $d = 10$, $\sigma = \mathrm{erf}$: start there and M alone is a sufficient statistic

The density of overlaps



$k = 1$, $d = 3$, $p = 5000$, $\sigma = z^2$: uniform at $t = 0$, then $1/2(\delta_{\boldsymbol{w}^\star} + \delta_{-\boldsymbol{w}^\star})$

The PDE is not an abstraction: you can watch the density move.

It's not all lost though: *phase retrieval*

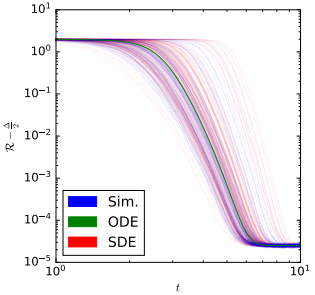
$\sigma = h^\star = z^2$ the ODE still tells us:

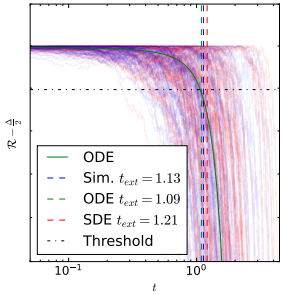
Limit in risk and learning rate

$$\lim_{t\to\infty} \mathcal{R}(\bar{m}(t)) - \Delta/2 = \frac{\gamma\Delta}{1-6\gamma}$$

Pre-factors

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[\eta d + (1-\eta)\right]}{8(1-6\gamma) - 4\gamma\Delta} \implies n = \mathcal{O}\left(d \log d\right)$$





Stochasticity does not matter

ODE

$$\frac{\mathrm{d}\bar{m}(t)}{\mathrm{d}t} = \bar{m}(t)\left[4(1-6\gamma)(1-\bar{m}^2(t)) - 2\gamma\Delta\right]$$

SDE

$$\frac{\mathrm{d}\mathrm{M}_1}{\mathrm{d}t} = \left(\Psi_1(\Omega) - \frac{\mathrm{M}_1}{2}\Phi_{11}(\Omega)\right)\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\left(\boldsymbol{\sigma}_{\mathrm{M}} - \frac{m_1}{2}\boldsymbol{\sigma}_{\mathrm{Q}}\right)\cdot \mathrm{d}B_t$$

Saad & Solla still gives us the weak recovery time, prefactor included.

Does width help? Exit time at arbitrary p

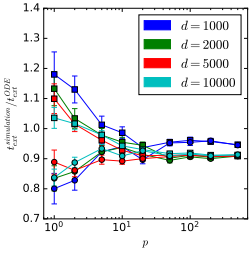
$\sigma = h^\star = z^2$, spherical, second layer fixed at $a_j^0 = 1$

Annealed

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[\frac{\eta(p+1)d+(p+1)(1-\eta)}{2p}\right]}{8\left[1-\frac{\gamma}{p}\left(1+\frac{1}{p}+\frac{4}{p^2}+\frac{\Delta}{2}\right)\right]}$$

Quenched averaged over the initial overlaps

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0,\tau_0}\left[\frac{\log\left[\frac{\eta p(p+1)d+(2\mu_0 p-\tau_0)(1-\eta)}{2\mu_0 p}\right]}{8\left[1-\frac{\gamma}{p}\left(1+\frac{1}{p}+\frac{4}{p^2}+\frac{\Delta}{2}\right)\right]}\right]$$



measured / predicted t_{ext}; □ annealed, ○ quenched, γ/p held fixed

Monotone in p

but the leading order is $\log d$ for *every* p

Optimal rate

$$\gamma_{\text{opt}} = \frac{p^3}{8+2p+(2+\Delta)p^2}$$

prefactor $1/4$, independent of p

Total gain

$p \to \infty$ against $p = 1$: $\dfrac{12+\Delta}{2+\Delta}$

Width buys a constant. It cannot touch the $d \log d$.

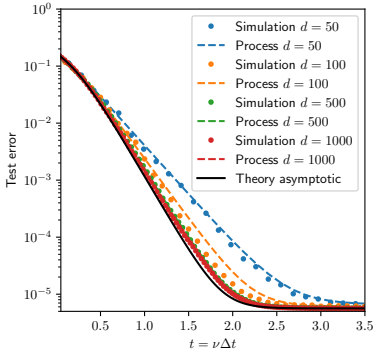
What if we scale the rate and the batch size?

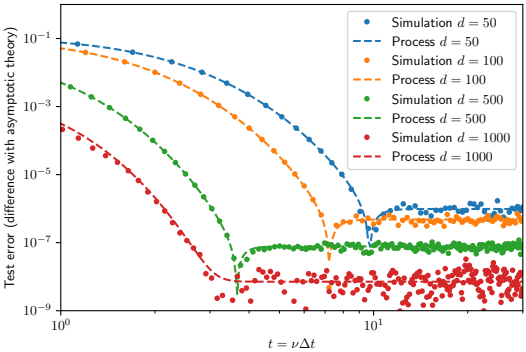
$$\sum_{\nu'\neq\nu}^{n_b} \sigma'(\lambda_j^{\nu})\,\sigma'(\lambda_l^{\nu'})\,\mathcal{E}^{\nu}\mathcal{E}^{\nu'}\langle \boldsymbol{z}^{\nu}, \boldsymbol{z}^{\nu'}\rangle \;\longrightarrow\; \phi_{jl}^{\mathrm{BC}}$$

$$\phi_{jl}^{\mathrm{BC}} = \mathbb{E}[\sigma'(\lambda_j)\,\mathcal{E}\,(\boldsymbol{\lambda}^{\star})^{\top}]\; P^{-1}\; \mathbb{E}[\sigma'(\lambda_l)\,\mathcal{E}\,\boldsymbol{\lambda}^{\star}] \qquad \text{along the teacher}$$

$$+\; \mathbb{E}[\sigma'(\lambda_j)\,\mathcal{E}\,(\boldsymbol{\lambda}^{\perp})^{\top}]\; (Q^{\perp})^{-1}\; \mathbb{E}[\sigma'(\lambda_l)\,\mathcal{E}\,\boldsymbol{\lambda}^{\perp}] \qquad \text{along the orthogonal student, } Q^{\perp} \text{ trick as before}$$

Whatever lies outside these two subspaces just decays away.

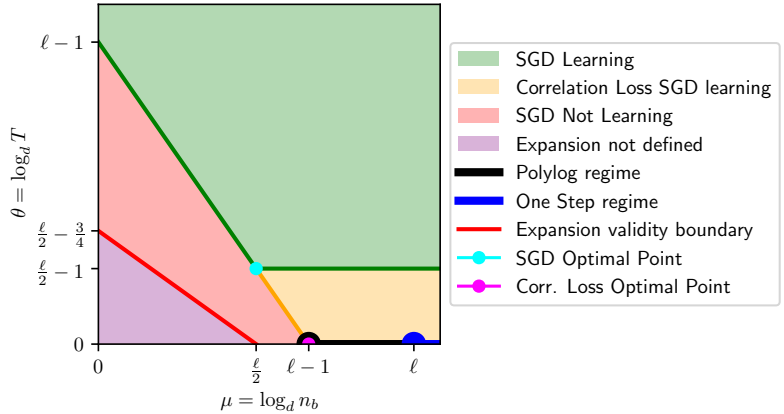




What if we scale the rate and the batch size?

$$\gamma = \gamma_0 d^{-\delta} \quad \text{and} \quad n_b = n_0 d^{\mu}$$

We can't change the sample complexity... but we use n_b times fewer time steps



Square Loss

$$\hat{\mathcal{R}}_t = \frac{1}{2n_b} \sum_{\nu=1}^{n_b} \left(y^\nu - f(\boldsymbol{z}^\nu)\right)^2$$

Correlation Loss

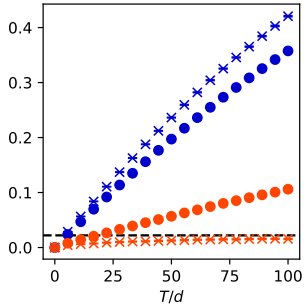
$$\tilde{\ell} = \frac{1}{n_b} \sum_{\nu \in [n_b]} 1 - y^\nu f(\boldsymbol{z}^\nu)$$

Multi-index: which directions get learned?

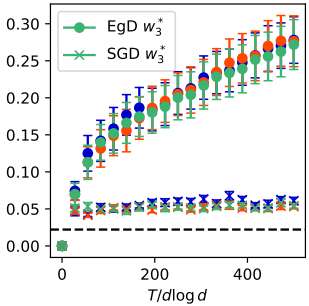
Directional exponents

$\ell_{\boldsymbol{v}} = \min\{k : \mathbb{E}\left[y\,\mathrm{He}_k(\langle \boldsymbol{v}, \boldsymbol{z}\rangle)\right] \neq 0\}, \qquad \ell^\star_{\boldsymbol{v}} = \min\{k : \mathbb{E}\left[p_{\boldsymbol{v}}(y)\,\mathrm{He}_k(\langle \boldsymbol{v}, \boldsymbol{z}\rangle)\right] \neq 0 \;\; \textit{for some polynomial } p_{\boldsymbol{v}}\}$

$\ell = \min_{\boldsymbol{v}} \ell_{\boldsymbol{v}}, \;\; \ell^\star_p = \min_{\boldsymbol{v}} \ell^\star_{\boldsymbol{v}}$ — every direction outside the *minimal subspace* $U^\star$ is strictly harder



$x_1 + \mathrm{He}_3(x_2), \qquad (\ell_{\boldsymbol{w}_2^\star}, \ell^\star_{\boldsymbol{w}_2^\star}) = (3, 1)$



$x_1 x_2 x_3, \qquad (\ell, \ell^\star) = (3, 2)$

The leap exponent: the cost is the hardest rung

Leap [Abbe, Boix-Adserà & Misiakiewicz, 2023]

effective exponent, given the already-recovered set S:

$$\ell_{\boldsymbol{v}}^{S} = \min\Big\{\textstyle\sum_{r:\,\boldsymbol{w}_r^\star \notin S} \alpha_r \;:\; a_\alpha \neq 0,\ \alpha_{\boldsymbol{v}} > 0\Big\}$$

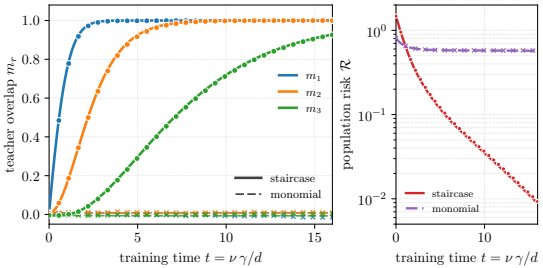
$$\mathrm{Leap} = \min_{\pi \in \mathfrak{S}_k} \max_{r \in [k]} \ell_{\boldsymbol{w}_{\pi(r)}^\star}^{S_{r-1}}$$

best order to learn in, worst step along it

$$n = \mathcal{O}(d^{\mathrm{Leap}-1}), \textbf{ not } \mathcal{O}(d^{\ell-1}).$$

$\lambda_1^\star + \lambda_1^\star\lambda_2^\star + \lambda_1^\star\lambda_2^\star\lambda_3^\star$: each recovered overlap turns the rung above it linear: three times $\mathcal{O}(d)$. The bare $\lambda_1^\star\lambda_2^\star\lambda_3^\star$ has $\mathrm{Leap} = 3$: $\mathcal{O}(d^2)$.

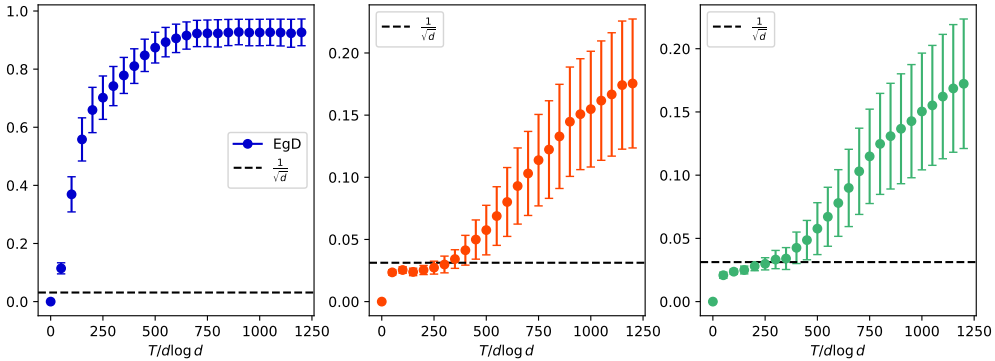
$\min_v \ell_v = 1$ is not enough: $\lambda_1^\star + \lambda_2^\star\lambda_3^\star\lambda_4^\star$ and $\lambda_1^\star + \lambda_1^\star\lambda_2^\star\lambda_3^\star\lambda_4^\star$ both have $\mathrm{Leap} = 3$.



dots: online SGD; lines: the Saad & Solla equations

Repetition opens a *new* staircase

$$f^\star(\boldsymbol{z}) = \mathrm{He}_4(z_1) + \mathrm{sign}(z_1 z_2 z_3)$$



$\mathrm{He}_4(z_1)$ has $\ell^\star = 2$, so EgD takes $\boldsymbol{w}_1^\star$ (blue) in $\mathcal{O}(d\log d)$; that unlocks $\mathrm{sign}(z_1 z_2 z_3)$ (alone $\ell = \ell^\star = 3$), and $\boldsymbol{w}_2^\star$ (orange), $\boldsymbol{w}_3^\star$ (green) follow in another $\mathcal{O}(d\log d)$.

A staircase only *repetition* can climb: for SGD both rungs are $\ell \geq 3$.

Why polynomials? What EgD can reach

$$\ell = \min\{k : \mathbb{E}[\, y\, \mathrm{He}_k(z)] \neq 0\}$$ no transformation: one-pass SGD

$$\ell_p^\star = \min\{k : \mathbb{E}[\, p(y)\, \mathrm{He}_k(z)] \neq 0\}$$ p polynomial: **the rung EgD reaches**

$$\ell^\star = \min\{k : \mathbb{E}[\, g(y)\, \mathrm{He}_k(z)] \neq 0\}$$ any g at all: the SQ optimum [Damian et al. '24]

Wider class, smaller minimum: $\ell^\star \le \ell_p^\star \le \ell$.

We are handed a *series*, not a p

$$g_{\rho_0}(y) = \sum_{j\ge 0} c_j\, \rho_0^j\, y^{j+1}$$

nothing is truncated and nothing is chosen either: $c_j \propto \mathbb{E}\big[\sigma^{(j+1)}\sigma'^{\,j}\big]$ is fixed by the activation alone.

but its reach is exactly $\ell_p^\star$

$$\mathbb{E}[g_{\rho_0}(y)\, \mathrm{He}_k(z)] = \sum_{m\ge 1} c_{m-1}\, \rho_0^{m-1}\, \alpha_m$$

the same $\alpha_m = \mathbb{E}[y^m \mathrm{He}_k(z)]$ that define $\ell_p^\star$. All zero $\Rightarrow$ no better than a polynomial; one non-zero $\Rightarrow$ no worse.

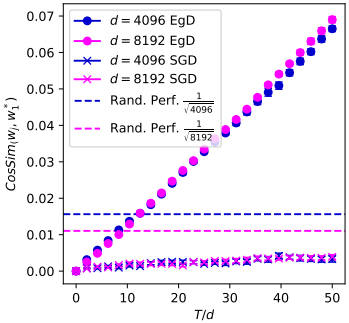
And that is the "almost every ρ_0": analytic in ρ_0, not identically zero $\Rightarrow$ isolated zeros.

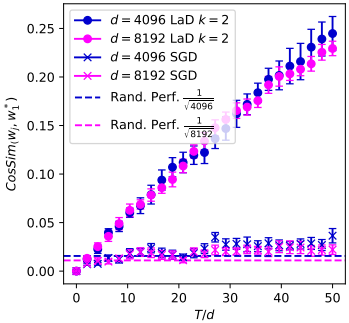
Theorem

If $h^\star$ is a polynomial, then $\ell_p^\star \le 2$. If $\ell_p^\star = 1$ unless $h^\star$ is even.

It is the repetition, not the trick

$h^\star(\boldsymbol{z}) = \mathrm{He}_3(z_1)$, $\ell = 3$, $\ell^\star = 1$, $d = 4096, 8192$, $\gamma_0 = 0.01$, averaged over 40 runs





SAM, i.e. $\rho < 0$: the *opposite* sign to the extragradient we use throughout

2-Lookahead: two plain gradient steps on the same sample; not in our algorithm class at all

Both learn it in $\mathcal{O}(d)$. Seeing the sample twice is the whole mechanism.

Tied and untied: the exact setting

Positional encoding off ($P = 0$), to isolate the sequence length L.

$X \in \mathbb{R}^{L\times d}$, target $y = g(\boldsymbol{z}_\star)$ with $\boldsymbol{z}_\star = X\boldsymbol{w}_\star \in \mathbb{R}^L$, one-pass SGD on both

Tied network d parameters

$$f_{\boldsymbol{w}}(X) = \sigma\left(\frac{\mathrm{Flatten}(X)\cdot \boldsymbol{w}_{\mathrm{tied}}}{\sqrt{L}}\right)$$

one $\boldsymbol{w} \in \mathbb{R}^d$ shared by every token:

$\boldsymbol{w}_{\mathrm{tied}} = (\boldsymbol{w}, \dots, \boldsymbol{w}) \in \mathbb{R}^{Ld}$

scalar overlap $\quad m = \boldsymbol{w}_\star^\top \boldsymbol{w}/d$

what linear attention reduces to

Untied network Ld parameters

$$f_W(X) = \sigma\left(\frac{\mathrm{Flatten}(X)\cdot \mathrm{Flatten}(W)}{\sqrt{L}}\right)$$

one weight vector per token: the L rows of $W \in \mathbb{R}^{L\times d}$, each updated on its own

vector overlap $\quad \boldsymbol{m} = W\boldsymbol{w}_\star/d,$

$m_{\mathrm{untied}} = \|\boldsymbol{m}\|/\sqrt{L}$

a plain single-index model in dimension Ld

$$\mathrm{gain}(L) \coloneqq \nu_{\eta,\,\mathrm{untied}}^{\mathrm{weak}} \Big/ \nu_{\eta}^{\mathrm{weak}}$$

steps to weak recovery, each at the largest learning rate that still recovers

Why is a tied network faster? The role of L

$\mathrm{gain}(L) = \nu^{\mathrm{weak}}_{\eta,\,\mathrm{untied}} \big/ \nu^{\mathrm{weak}}_{\eta}$: how many times fewer gradient steps the tied network needs

Theorem C_{SIE}: first non-zero Hermite tensor of g

$$\mathrm{gain} \asymp \left(\frac{C_{SIE} \times (\mathbf{1}, \dots, \mathbf{1})}{\|C_{SIE}\|_{\mathrm{op}}} \right)^2 \cdot \begin{cases} L & SIE = 1 \\ 1 & \textit{otherwise} \end{cases}$$

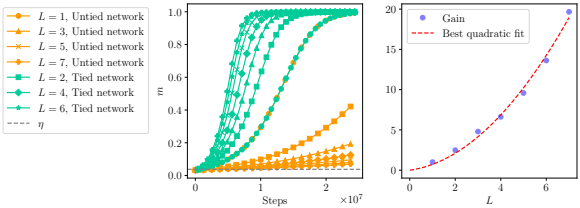
$\asymp$ when C_{SIE} is orthogonally decomposable; $\gtrsim$ in general

so $0 \le \mathrm{gain} \lesssim L^{SIE \vee 2}$

naive parameter counting — the untied net has L times the parameters — would say $L^{(SIE-1)\vee 1}$; the truth lands on *either* side of it.

$g = \sum_i \mathrm{He}_2(z_{\star,i})$: $C_2 = I_L$, so $\mathrm{gain} \asymp L^2$

$d = 1000$, $\sigma = \mathrm{ReLU}$ in the panels

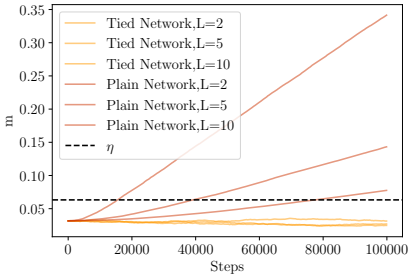


left: overlap against gradient steps, tied (green) vs untied (orange), one symbol per L; right: the measured gain, with the fitted L^2.

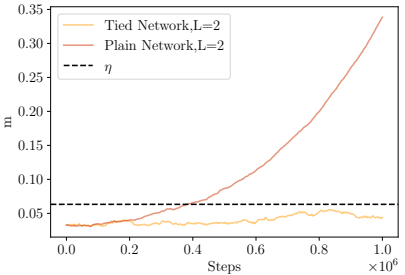
When tying fails

The same theorem allows $\text{gain} = 0$: $C_{SIE} \times (\mathbf{1}, \ldots, \mathbf{1}) = 0$

$SIE = 1$: $g = z_{\star,1} - z_{\star,2}$, $C_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$



$SIE = 2$: $g = z^2_{\star,1} - z^2_{\star,2}$, $C_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$



The tied network is symmetric by construction; these targets are as antisymmetric as possible.

The untied network learns them because its initialization breaks the symmetry. Randomly weighting the neurons would fix the tied one.

What is low-degree filtering?

Beyond the thesis

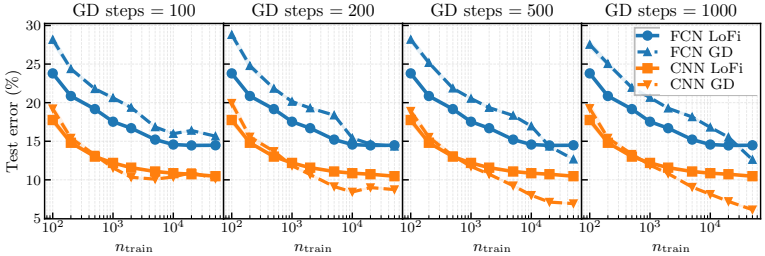
Not in this thesis. with Y. Dandi, M. Vilucchio, H. Tabanelli, F. Krzakala

One operator per layer — the label-weighted, Hessian-like term of layerwise GD at vanishing initialization

$$\widehat{C}^{(\ell)} = \frac{1}{n}\sum_{\mu=1}^{n} y_\mu \, \boldsymbol{z}_{\ell-1}(\boldsymbol{x}_\mu)\, \boldsymbol{z}_{\ell-1}(\boldsymbol{x}_\mu)^\top$$

Filter $\boldsymbol{g}_\ell = \widehat{V}_\ell^\top \boldsymbol{z}_{\ell-1}$ top-k eigenvectors $\longrightarrow$ Lift $\boldsymbol{z}_\ell = \sigma(R_\ell \boldsymbol{g}_\ell)/\sqrt{p_\ell}$ random, nonlinear $\longrightarrow$ Repeat $K_0 \to K_1 \to \cdots \to K_L$

Keep what is *simple* in the previous geometry and *predictive* at low degree.



binary CIFAR-10: LoFi (solid) is the same in every panel; GD (dashed) improves with steps

What it finds, without backpropagation

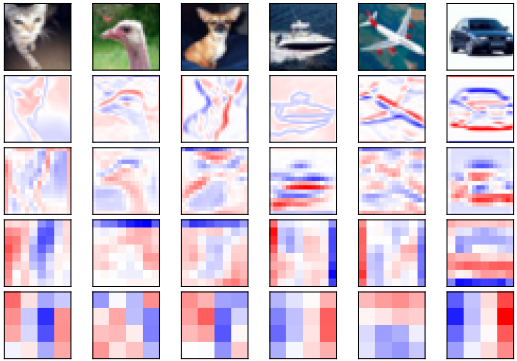
Beyond the thesis

The same $\widehat{C}$, built in channel space

convolutional LoFi on the CIFAR-10 animal-vs.-vehicle task



first-layer filters



one LoFi feature with depth: inputs on top, maps below

Edge detectors on real images, out of one eigendecomposition per layer.