Supplementary material

Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová

Appendix A Deterministic scaling limit of stochastic processes

In order to show the deterministic scaling of online SGD under a proper chosen time scale, we will make use of a convergence result by [21, 31], which is adapted below in Theorem A.1.

Theorem A.1 (Deterministic scaling limit of stochastic processes)**.** *Consider a d-dimension discrete time stochastic process sequence, $\{\mathbf{\Omega}^{\nu}\ ;\ \nu = 0, 1, 2, ..., [S\tau]\}_{S=1,2,...}$ for some $\tau > 0$. The increment $\mathbf{\Omega}^{\nu+1} - \mathbf{\Omega}^{\nu}$ is assumed to be decomposable into three parts,*

$$\mathbf{\Omega}^{\nu+1} - \mathbf{\Omega}^{\nu} = \frac{1}{S}\psi(\mathbf{\Omega}^{\nu}) + \mathbf{\Lambda}^{\nu} + \mathbf{\Gamma}^{\nu}\ , \tag{21}$$

such that

Assumption A.1.1. The process $\tilde{\mathbf{\Lambda}}^{\nu} \equiv \sum_{\nu'=0}^{\nu} \mathbf{\Lambda}^{\nu'}$ is a martingale and $\mathbb{E}\|\mathbf{\Lambda}^{\nu}\|^2 \leq C(\tau)^2/S^{1+\epsilon_1}$ for some $\epsilon_1 > 0$.

Assumption A.1.2. $\mathbb{E}\|\mathbf{\Gamma}^{\nu}\| \leq C(\tau)/S^{1+\epsilon_2}$ for some $\epsilon_2 > 0$.

Assumption A.1.3. The function $\psi(\mathbf{\Omega})$ is Lipschitz, i.e, $\|\psi(\mathbf{\Omega}) - \psi(\tilde{\mathbf{\Omega}})\| \leq C\|\mathbf{\Omega} - \tilde{\mathbf{\Omega}}\|$ for any $\mathbf{\Omega}$ and $\tilde{\mathbf{\Omega}}$.

Let $\mathbf{\Omega}(t)$, with $0 \leq t \leq \tau$, be a continuous stochastic process such that $\mathbf{\Omega}(t) = \mathbf{\Omega}^{\nu}$ with $\nu = [St]$. Define the deterministic ODE

$$\frac{d}{dt}\bar{\mathbf{\Omega}}(t) = \psi(\bar{\mathbf{\Omega}}(t))\ , \tag{22}$$

with $\bar{\mathbf{\Omega}}(0) = \bar{\mathbf{\Omega}}_0$.

Then, if assumptions A.1.1 to A.1.3 hold and assuming $\mathbb{E}\|\mathbf{\Omega}^0 - \bar{\mathbf{\Omega}}_0\| < C/S^{\epsilon_3}$ for some $\epsilon_3 > 0$ then we have for any finite S:

$$\mathbb{E}\left\|\mathbf{\Omega}^{\nu} - \bar{\mathbf{\Omega}}\left(\frac{\nu}{S}\right)\right\| \leq C(\tau)e^{c\tau}S^{-\min\{\frac{1}{2}\epsilon_1, \epsilon_2, \epsilon_3\}}\ , \tag{23}$$

where $\bar{\mathbf{\Omega}}(\cdot)$ is the solution of Eq.(22).

Proof. The reader interested in the proof is referred to the supplementary materials of [21, 31]. □

Although the theorem wasn't originally proven in the $p \to \infty$ setting, a glance at its proof shows that it still holds upon replacing $C(\tau)$ by $C(p, \tau)$ in Assumption A.1.1 and A.1.2, as well as Equation (23). We choose $\|\cdot\|$ to be the L^∞ norm, since it suits better the $p \to \infty$ scaling. The S in Theorem A.1 corresponds to $1/\delta t$, where δt is defined in Theorem 2.1.

Following [21], we define for $j, l \in [p]$

$$\Psi_{jl}(\mathbf{\Omega}; \boldsymbol{x}) = \frac{\gamma}{pd\,\delta t}\left(\mathcal{E}_j^{\nu}\,\lambda_l^{\nu} + \mathcal{E}_l^{\nu}\,\lambda_j^{\nu}\right) + \frac{\gamma^2}{p^2\,d\,\delta t}\,\mathcal{E}_j^{\nu}\,\mathcal{E}_l^{\nu},$$

and

$$\psi_{jl}(\mathbf{\Omega}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\mathbf{0},\mathbb{1})}\left[\Psi_{jl}(\mathbf{\Omega}; \boldsymbol{x})\right].$$

The functions Ψ, ψ are similarly defined on $[p] \times [p+1, p+k]$. With that, we write

$$\mathbf{\Omega}^{\nu+1} - \mathbf{\Omega}^{\nu} = \frac{1}{S}\psi(\mathbf{\Omega}) + \underbrace{\frac{1}{S}\left(\Psi(\mathbf{\Omega}^{\nu}; \boldsymbol{x}) - \psi(\mathbf{\Omega}^{\nu})\right)}_{\mathbf{\Lambda}^{\nu}} + \mathbf{\Gamma}^{\nu},$$

where for $j, l \in [p]$

$$\Gamma^{\nu}_{jl} = \frac{\gamma^2}{p^2 d^2} \left(\|\boldsymbol{x}\|_2^2 - d \right) \mathcal{E}^{\nu}_j \, \mathcal{E}^{\nu}_l \, .$$

The main obstacle to bounding $\boldsymbol{\Lambda}^{\nu}$ and $\boldsymbol{\Gamma}^{\nu}$ is the fact that the q_{jj} can a priori diverge to infinity. Our first task is therefore to show that this does not happen; as a proxy we show a subgaussian-like moment bound:

$$\mathbb{E}\left[(q^{\nu}_{jj})^t \right] \leq \left(C(\tau) + \frac{ct}{S} \right)^t .$$

Equipped with the above bound, controlling $\mathbb{E}\|\boldsymbol{\Lambda}^{\nu}\|^2$ and $\mathbb{E}\|\boldsymbol{\Gamma}^{\nu}\|$ becomes fairly easy. All proof details are in the below sections.

A.1 Preliminaries: bounding the q_{jj}

Since σ is L-Lipschitz, we have by the Cauchy-Schwarz inequality

$$(\mathcal{E}^{\nu})^2 \leq \frac{3L^2}{k} \sum_{r=1}^{k} (\lambda^*_r)^2 + \frac{3L^2}{p} \sum_{j=1}^{p} (\lambda_j)^2 + 3\Delta\zeta^2 \equiv \Phi^{\nu} \tag{24}$$

Define

$$s^{\nu} = \mathbb{E}\, \Phi^{\nu} = \frac{3L^2}{k} \sum_{r=1}^{k} \rho_{rr} + \frac{3L^2}{p} \sum_{j=1}^{p} q^{\nu}_{jj} + 3\Delta$$

Assumption 1 in Theorem 2.1 implies that

$$|q^{\nu+1}_{jj} - q^{\nu}_{jj}| \leq \frac{1}{S} \left(c_1 (\lambda^{\nu}_j)^2 + c_2 (\mathcal{E}^{\nu})^2 \right)$$

where c_1, c_2 are absolute constants. Summing those inequalities yield

$$|s_{\nu+1} - s^{\nu}| \leq \frac{c_3}{S} \Phi^{\nu},$$

and finally

$$\mathbb{E}_{\nu}[s^{\nu+1}] \leq s^{\nu} \left(1 + \frac{c_3}{S} \right) \leq s^{\nu} e^{c_3/S}.$$

As a result, we have for any $0 \leq \nu \leq S\tau$

$$\mathbb{E}[s^{\nu}] \leq c_4 e^{c_3 \tau}. \tag{25}$$

For simplicity, let q^{ν} denote any of the q^{ν}_{jj}. We have, for all $t \geq 0$,

$$(q^{\nu+1})^t - (q^{\nu})^t = t (q^{\nu})^{t-1} (q^{\nu+1} - q^{\nu}) + O\left(\frac{t^2}{S^2} \right),$$

where the remainder term has bounded expectation. Again, we write

$$\left| (q^{\nu+1})^t - (q^{\nu})^t \right| \leq t (q^{\nu})^{t-1} \frac{1}{S} (c_1 (\mathcal{E}^{\nu})^2 + c_2 (\lambda^{\nu}_i)^2) + \frac{c_5 t^2}{S^2}.$$

By Assumption 3, the q^{ν}_{ii} are bounded from below by a constant, hence

$$\mathbb{E}_{\nu}[(q^{\nu+1})^t] \leq (q^{\nu})^t \left(1 + \frac{c_6 t}{S} \right) + O\left(\frac{c_5 t^2}{S^2} \right)$$

This implies that for any $t \geq 0$ and $0 \leq \nu \leq S\tau$,

$$\mathbb{E}[(q^{\nu})^t] \leq \left(c_7 + \frac{c_5 t^2}{S} \right) e^{c_6 \tau} \leq \left(C(\tau) + \frac{c_5 t}{S} \right)^t \tag{26}$$

A.2 Assumption A.1.1

We have for all $i, j \in [p+k]$,

$$\left(\Omega_{ij}^{\nu+1} - \mathbb{E}_\nu[\Omega_{ij}^{\nu+1}]\right)^2 \le 2\left((\Omega_{ij}^{\nu+1} - \Omega_{ij}^{\nu})^2 + (\Omega_{ij}^{\nu} - \mathbb{E}_\nu[\Omega_{ij}^{\nu+1}])^2\right) \; .$$

As a consequence,

$$\mathbb{E}\|\mathbf{\Lambda}^\nu\|^2 \le 4 \max_{i,j} (\Omega_{ij}^{\nu+1} - \Omega_{ij}^{\nu})^2 \; .$$

Now, by definition,

$$(q_{ij}^{\nu+1} - q_{ij}^{\nu})^2 \le \frac{L}{S^2}\left(c_1(\mathcal{E}^\nu)^2 + c_2|\,\mathcal{E}^\nu\,|(|\lambda_i| + |\lambda_j|)\right)^2 \le \frac{L}{S^2}\left(c_3(\mathcal{E}^\nu)^4 + c_4(\max_\ell \lambda_\ell^\nu)^4\right),$$

The term in $(\mathcal{E}^\nu)^4$ is bounded by the same techniques as the last section. For the second term,

$$\mathbb{E}_\nu\left[(\max_\ell \lambda_\ell)^4\right] \le c_5 \log(p)^2 \left(\max_\ell q_{\ell\ell}^\nu\right)^4,$$

and we can write for any $t \ge 0$

$$\max_\ell \, (q_{\ell\ell}^\nu)^4 \le \left(\sum_\ell (q_{\ell\ell}^\nu)^t\right)^{4/t} .$$

By Jensen's inequality, for $t \ge 4$

$$\mathbb{E}\left[\left(\max_\ell q_{\ell\ell}^\nu\right)^4\right] \le \left(\sum_\ell \mathbb{E}[(q_{\ell\ell}^\nu)^t\right)^{4/t} \le p^{4/t}\left(C(\tau) + \frac{c_6 t}{S}\right)^4,$$

using (26). Choosing $t = 4\log(p) \ll S$ shows that

$$\mathbb{E}\left[\max_{i,j}(q_{ij}^{\nu+1} - q_{ij}^{\nu})^2\right] \le \frac{C(\tau)\log(p)^2}{S^2}$$

A similar bound holds for the m_{ij}, and hence

$$\mathbb{E}\|\mathbf{\Lambda}^\nu\|^2 \le \frac{c_5 \log(p)^2}{S^2} \; ,$$

which implies Assumption A.1.1 with $\epsilon_1 = 1$ and $C(p, \tau) = C'(\tau)\log(p)$.

A.3 Assumption A.1.2

Since σ is Lipschitz, for any $i, j \in [p]$

$$\mathcal{E}_i^\nu \, \mathcal{E}_j^\nu \le L^2(\mathcal{E}^\nu)^2.$$

Hence,

$$\begin{aligned}\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] &\le \frac{L^2\gamma^2}{d^2p^2}\,\mathbb{E}\left[\left(\|\boldsymbol{x}\|_2^2 - d\right)\Phi^\nu\right] \\ &\le \frac{L^2\gamma^2}{d^2p^2}\left(\frac{1}{2\sqrt{d}}\,\mathbb{E}\left[\left(\|\boldsymbol{x}\|_2^2 - d\right)^2\right] + \frac{\sqrt{d}}{2}\,\mathbb{E}\left[(\mathcal{E}^\nu)^4\right]\right).\end{aligned}$$

The first expectation is the variance of a χ_d^2 random variable, which is equal to $2d$, and the second expectation is bounded by the same methods as the above sections. The term in brackets is therefore bounded by $c_1\sqrt{d}$, and

$$\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] \le c_2 \frac{\gamma^2}{d^{3/2}p^2}$$

Finally, since for any $y > 0$ we have $y^2 \le \max(y, y^2)^{3/2}$, letting $y = \gamma/p$ we find

$$\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] \le c_2 \max\left(\frac{\gamma}{pd}, \frac{\gamma^2}{p^2 d}\right)^{3/2} \le c_3(\delta t)^{3/2},$$

hence Assumption A.1.2 is true with $\epsilon_2 = 1/2$.

A.4 $\sqrt{}$-Lipschitz property

Let $\mathbf{\Omega}, \mathbf{\Omega}' \in \mathbb{R}^{(p+k)\times(p+k)}$, we can write the (i,j) coefficient of $\psi(\mathbf{\Omega})$ as $f_{ij}(\sqrt{\mathbf{\Omega}})$, where

$$\begin{aligned} f : \mathbb{R}^{(p+k)\times(p+k)} &\to \mathbb{R} \\ A &\mapsto \mathbb{E}_{x\sim\mathcal{N}(0,I_{p+k})}[g_{ij}(Ax)] \end{aligned}$$

The same arguments as above show that the function f is Lipschitz, and hence for some constant L'' we have

$$\|\psi(\mathbf{\Omega}) - \psi(\mathbf{\Omega}')\| \leq L''\|\sqrt{\mathbf{\Omega}} - \sqrt{\mathbf{\Omega}'}\|.$$

Appendix B A lemma on ODE perturbation

In this section, we prove a proposition that bounds the difference between an ODE solution and a perturbed version, for a bounded time t.

Theorem B.1. *Let $f, g : \mathbb{R}^n \to \mathbb{R}^n$ be two L-Lipschitz functions, and consider the following differential equations in $\mathbb{R}^n$:*

$$\frac{d\boldsymbol{x}}{dt} = f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x}),$$
$$\frac{d\boldsymbol{y}}{dt} = f(\boldsymbol{y}),$$

where $\epsilon > 0$, and with the initial condition $\boldsymbol{x}(0) = \boldsymbol{y}(0)$. Then, if $\tau > 0$ is fixed, we have

$$\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|_2 \leq c\epsilon e^{L\tau}$$

for any $0 \leq t \leq \tau$, with c a constant independent from ϵ, τ.

Before proving this proposition, we begin with a small lemma:

Lemma B.2. *Let $a, b > 0$, and $z : \mathbb{R}^+ \to \mathbb{R}^+$ a function satisfying*

$$\frac{dz}{dt} = az + b\sqrt{z}$$

with $z(0) = 0$. Then, for some constant $c > 0$, we have

$$z(t) \leq c\frac{b^2 e^{at}}{a^2} \quad \textit{for all} \quad t \geq 0$$

Proof. Upon considering the function $a^2 z(t/a)/b^2$ instead, we can assume that $a = b = 1$. Then, we have

$$\frac{dz}{dt} \leq \max(z, 1) + \max(\sqrt{z}, 1),$$

and the RHS is an increasing function. Hence, if $\tilde{z}$ is a solution of

$$\frac{d\tilde{z}}{dt} = \max(z, 1) + \max(\sqrt{\tilde{z}}, 1),$$

with $\tilde{z}(0) = 0$, then $z(t) \leq \tilde{z}(t)$ for all $t \geq 0$. Since the RHS of the above equation is Lipschitz everywhere, we can apply the Picard–Lindelöf theorem, and check that the unique solution to this equation is

$$\tilde{z}(t) = \begin{cases} 2t & \text{if } t \leq \frac{1}{2} \\ (c_1 e^t - c_2)^2 & \text{otherwise} \end{cases},$$

where c_1 and c_2 are ad hoc constants. The lemma then follows from adjusting the constant c as needed. □

We are now in a position to show Theorem B.1:

Proof. Assume for simplicity that $\boldsymbol{x}(0) = \boldsymbol{y}(0) = \boldsymbol{0}$. We begin by bounding $\boldsymbol{x}(t)$; we have

$$\frac{d\|\boldsymbol{x}\|^2}{dt} = 2\boldsymbol{x}^\top \frac{d\boldsymbol{x}}{dt} \leq 2\|\boldsymbol{x}\| \; \|f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x})\| \; .$$

By the Lipschitz condition,

$$\|f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x})\| \leq \|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\| + \frac{L}{2}\|\boldsymbol{x}\| \; ,$$

so that

$$\frac{d\|\boldsymbol{x}\|^2}{dt} \leq L\|\boldsymbol{x}\|^2 + 2\|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\| \; \|\boldsymbol{x}\| \; .$$

Applying Lemma B.2 and taking square roots on each side,

$$\|\boldsymbol{x}(t)\| \leq c\frac{\|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\|}{L} e^{Lt/2} \leq c\frac{\|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\|}{L} e^{L\tau/2} \; , \tag{27}$$

for any $0 \leq t \leq \tau$. Now, similarly,

$$\begin{aligned} \frac{d\|\boldsymbol{x} - \boldsymbol{y}\|^2}{dt} &\leq 2\|\boldsymbol{x} - \boldsymbol{y}\| \left\| \frac{d(\boldsymbol{x} - \boldsymbol{y})}{dt} \right\| \\ &\leq 2\|\boldsymbol{x} - \boldsymbol{y}\| \; \|f(\boldsymbol{x}) - f(\boldsymbol{y}) + \epsilon g(\boldsymbol{x})\| \\ &\leq L\|\boldsymbol{x} - \boldsymbol{y}\|^2 + 2\epsilon\|g(\boldsymbol{x})\| \; \|\boldsymbol{x} - \boldsymbol{y}\| \\ &\leq L\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \epsilon\left(\|g(\boldsymbol{0})\| + c\|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\| e^{L\tau/2}\right) \|\boldsymbol{x} - \boldsymbol{y}\| \; , \end{aligned}$$

having used (27) on the last line. This is again the setting of Lemma B.2, which gives

$$\|\boldsymbol{x} - \boldsymbol{y}\| \leq c_1 \epsilon e^{L\tau/2} \frac{e^{Lt/2}}{L} \leq c_2 \epsilon e^{L\tau} \; .$$

$\square$

Appendix C Expectations over the local fields

In this appendix we present the explicit expressions from the expectations of the local fields used to compute the population risk and the ODE terms.

C.1 Population risk

We write the population risk (10) as

$$\begin{aligned} \mathcal{R}(\boldsymbol{\Omega}) &= \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{\Omega})} \, \mathbb{E}_{\zeta \sim \mathbb{P}(\zeta)} \left[\left(\hat{f}(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda}^*) \right)^2 \right] \\ &= \mathcal{R}_{\text{t}}(\boldsymbol{P}) + \mathcal{R}_{\text{s}}(\boldsymbol{Q}) + \mathcal{R}_{\text{st}}(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{M}) \; , \end{aligned} \tag{28}$$

with

$$\mathcal{R}_{\text{t}} \equiv \mathbb{E}_{\boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{P})} \left[f(\boldsymbol{\lambda}^*)^2 \right] = \frac{1}{k^2} \sum_{r,s=1}^{k} \mathbb{E}_{\boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{P})} \left[\sigma(\lambda^*_r) \sigma(\lambda^*_s) \right] \tag{29a}$$

$$\mathcal{R}_{\text{s}} \equiv \mathbb{E}_{\boldsymbol{\lambda} \sim \mathcal{N}(\boldsymbol{\lambda} | \boldsymbol{0}, \boldsymbol{Q})} \left[\hat{f}(\boldsymbol{\lambda})^2 \right] = \frac{1}{p^2} \sum_{j,l=1}^{k} \mathbb{E}_{\boldsymbol{\lambda} \sim \mathcal{N}(\boldsymbol{\lambda} | \boldsymbol{0}, \boldsymbol{Q})} \left[\sigma(\lambda_j) \sigma(\lambda_l) \right] \; , \tag{29b}$$

$$\mathcal{R}_{\text{st}} \equiv \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{\Omega})} \left[\hat{f}(\boldsymbol{\lambda}) f(\boldsymbol{\lambda}^*) \right] = -\frac{2}{pk} \sum_{j=1}^{p} \sum_{r=1}^{k} \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{\Omega})} \left[\sigma(\lambda_j) \sigma(\lambda^*_r) \right] \tag{29c}$$

Define the vector $\boldsymbol{\lambda}^{\alpha\beta} \equiv \left(\lambda^\alpha, \lambda^\beta\right)^\top \in \mathbb{R}^2$, where the upper indices on the components indicate they may refer to student or teacher local fields. Consider the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta}$:

$$\boldsymbol{\Omega}^{\alpha\beta} \equiv \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \boldsymbol{\Omega})} \left[\boldsymbol{\lambda}^{\alpha\beta} \left(\boldsymbol{\lambda}^{\alpha\beta} \right)^\top \right] \in \mathbb{R}^{2 \times 2} \; . \tag{30}$$

For $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (29) are in general given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma(\lambda^\alpha)\sigma(\lambda^\beta)\right] = \frac{1}{\pi}\arcsin\left(\frac{\Omega_{12}^{\alpha\beta}}{\sqrt{\left(1+\Omega_{11}^{\alpha\beta}\right)\left(1+\Omega_{22}^{\alpha\beta}\right)}}\right) . \tag{31}$$

where $\Omega_{jl}^{\alpha\beta} \equiv (\boldsymbol{\Omega}^{\alpha\beta})_{jl}$ is an element of the covariance matrix given by Eq. (30).

Explicitly, the population risk contributions are

$$\mathcal{R}_{\mathrm{t}}(\boldsymbol{P}) = \frac{1}{k^2}\sum_{r,s=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{\rho_{rs}}{\sqrt{(1+\rho_{rr})\,(1+\rho_{ss})}}\right) , \tag{32a}$$

$$\mathcal{R}_{\mathrm{s}}(\boldsymbol{Q}) = \frac{1}{p^2}\sum_{j,l=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{q_{jl}}{\sqrt{(1+q_{jj})\,(1+q_{ll})}}\right) , \tag{32b}$$

$$\mathcal{R}_{\mathrm{st}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{M}) = -\frac{2}{pk}\sum_{j=1}^{p}\sum_{r=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{m_{jr}}{\sqrt{(1+q_{jj})\,(1+\rho_{rr})}}\right) . \tag{32c}$$

C.2 ODE contributions

From the update equations, we first consider the expectations linear in $\mathcal{E}_j$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\lambda_l\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda^*_{r'})\right]\\ &-\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_{l'})\right] ,\end{aligned} \tag{33a}$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\lambda^*_r\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\lambda^*_r\sigma(\lambda^*_{r'})\right]\\ &-\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\lambda^*_r\sigma(\lambda_{l'})\right] .\end{aligned} \tag{33b}$$

Define the vector $\boldsymbol{\lambda}^{\alpha\beta\gamma} \equiv \left(\lambda^\alpha, \lambda^\beta, \lambda^\gamma\right)^\top \in \mathbb{R}^3$, where the upper indices on the components indicate they may refer to student or teacher local fields. Consider the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta\gamma}$:

$$\boldsymbol{\Omega}^{\alpha\beta\gamma} \equiv \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\boldsymbol{\lambda}^{\alpha\beta\gamma}\left(\boldsymbol{\lambda}^{\alpha\beta\gamma}\right)^\top\right] \in \mathbb{R}^{3\times 3} . \tag{34}$$

For $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (33) are given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda^\alpha)\lambda^\beta\sigma(\lambda^\gamma)\right] = \frac{2}{\pi}\frac{\Omega_{23}^{\alpha\beta\gamma}\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)-\Omega_{12}^{\alpha\beta\gamma}\Omega_{13}^{\alpha\beta\gamma}}{\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)\sqrt{\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)\left(1+\Omega_{33}^{\alpha\beta\gamma}\right)-\left(\Omega_{13}^{\alpha\beta\gamma}\right)^2}} , \tag{35}$$

where $\Omega_{jl}^{\alpha\beta\gamma} \equiv (\boldsymbol{\Omega}^{\alpha\beta\gamma})_{jl}$ is an element of the covariance matrix given by Eq. (34). As examples, we write explicitly:

$$\boldsymbol{\Omega}^{jlr'} = \begin{bmatrix} q_{jj} & q_{jl} & m_{jr'} \\ q_{jl} & q_{ll} & m_{lr'} \\ m_{jr'} & m_{lr'} & \rho_{r'r'} \end{bmatrix} , \quad \boldsymbol{\Omega}^{jrr'} = \begin{bmatrix} q_{jj} & m_{jr} & m_{jr'} \\ m_{jr} & \rho_{rr} & \rho_{rr'} \\ m_{jr'} & \rho_{rr'} & \rho_{r'r'} \end{bmatrix} . \tag{36}$$

The quadratic contribution in $\mathcal{E}_j$ is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\mathcal{E}_l\right]=&\frac{1}{k^2}\sum_{r,r'=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda^*_r)\sigma(\lambda^*_{r'})\right]\\&+\frac{1}{p^2}\sum_{j',l'=1}^{p}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda_{j'})\sigma(\lambda_{l'})\right]\\&-\frac{2}{pk}\sum_{l'=1}^{p}\sum_{r=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda^*_r)\sigma(\lambda_{l'})\right]\\&+\Delta\,\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\right]\end{aligned}\tag{37}$$

The solution of the noise-dependent term can be constructed with the covariance matrix (30) and is given by [6]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda^\alpha)\sigma'(\lambda^\beta)\right]=\frac{2}{\pi}\frac{1}{\sqrt{1+\Omega^{\alpha\beta}_{11}+\Omega^{\alpha\beta}_{22}+\Omega^{\alpha\beta}_{11}\Omega^{\alpha\beta}_{22}-\left(\Omega^{\alpha\beta}_{12}\right)^2}}\tag{38}$$

Similarly, one can define the vector $\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}\equiv\left(\lambda^\alpha,\lambda^\beta,\lambda^\gamma,\lambda^\delta\right)^\top\in\mathbb{R}^4$ and write the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}$:

$$\boldsymbol{\Omega}^{\alpha\beta\gamma\delta}\equiv\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}\left(\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}\right)^\top\right]\in\mathbb{R}^{4\times4}\ .\tag{39}$$

For $\sigma(x)=\mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (37) are given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma'(\lambda^\alpha)\sigma'(\lambda^\beta)\sigma(\lambda^\gamma)\sigma(\lambda^\delta)\right]=\frac{4}{\pi^2}\frac{1}{\sqrt{\bar{\Omega}^{\alpha\beta\gamma\delta}_0}}\arcsin\left(\frac{\bar{\Omega}^{\alpha\beta\gamma\delta}_1}{\sqrt{\bar{\Omega}^{\alpha\beta\gamma\delta}_2\bar{\Omega}^{\alpha\beta\gamma\delta}_3}}\right)\ ,\tag{40}$$

with

$$\bar{\Omega}^{\alpha\beta\gamma\delta}_0\equiv\left(1+\Omega^{\alpha\beta\gamma\delta}_{11}\right)\left(1+\Omega^{\alpha\beta\gamma\delta}_{22}\right)-\left(\Omega^{\alpha\beta\gamma\delta}_{12}\right)^2\ ,\tag{41a}$$

$$\begin{aligned}\bar{\Omega}^{\alpha\beta\gamma\delta}_1\equiv&\bar{\Omega}^{\alpha\beta\gamma\delta}_0\Omega^{\alpha\beta\gamma\delta}_{34}-\Omega^{\alpha\beta\gamma\delta}_{23}\Omega^{\alpha\beta\gamma\delta}_{24}\left(1+\Omega^{\alpha\beta\gamma\delta}_{11}\right)-\Omega^{\alpha\beta\gamma\delta}_{13}\Omega^{\alpha\beta\gamma\delta}_{14}\left(1+\Omega^{\alpha\beta\gamma\delta}_{22}\right)\\&+\Omega^{\alpha\beta\gamma\delta}_{12}\Omega^{\alpha\beta\gamma\delta}_{13}\Omega^{\alpha\beta\gamma\delta}_{24}+\Omega^{\alpha\beta\gamma\delta}_{12}\Omega^{\alpha\beta\gamma\delta}_{14}\Omega^{\alpha\beta\gamma\delta}_{23}\ ,\end{aligned}\tag{41b}$$

$$\begin{aligned}\bar{\Omega}^{\alpha\beta\gamma\delta}_2\equiv&\bar{\Omega}^{\alpha\beta\gamma\delta}_0\left(1+\Omega^{\alpha\beta\gamma\delta}_{44}\right)-\left(\Omega^{\alpha\beta\gamma\delta}_{24}\right)^2\left(1+\Omega^{\alpha\beta\gamma\delta}_{11}\right)-\left(\Omega^{\alpha\beta\gamma\delta}_{13}\right)^2\left(1+\Omega^{\alpha\beta\gamma\delta}_{22}\right)\\&+2\Omega^{\alpha\beta\gamma\delta}_{12}\Omega^{\alpha\beta\gamma\delta}_{13}\Omega^{\alpha\beta\gamma\delta}_{23}\ ,\ .\end{aligned}\tag{41c}$$

$$\begin{aligned}\bar{\Omega}^{\alpha\beta\gamma\delta}_3\equiv&\bar{\Omega}^{\alpha\beta\gamma\delta}_0\left(1+\Omega^{\alpha\beta\gamma\delta}_{44}\right)-\left(\Omega^{\alpha\beta\gamma\delta}_{24}\right)^2\left(1+\Omega^{\alpha\beta\gamma\delta}_{11}\right)-\left(\Omega^{\alpha\beta\gamma\delta}_{14}\right)^2\left(1+\Omega^{\alpha\beta\gamma\delta}_{22}\right)\\&+2\Omega^{\alpha\beta\gamma\delta}_{12}\Omega^{\alpha\beta\gamma\delta}_{14}\Omega^{\alpha\beta\gamma\delta}_{24}\ .\end{aligned}\tag{41d}$$

C.3 From gradient flow to local fields

Consider the gradient flow approximation

$$\begin{aligned}\frac{d\boldsymbol{w}_j}{dt}&=-\boldsymbol{\nabla}_{\boldsymbol{w}_j}\mathcal{R}(\boldsymbol{W},\boldsymbol{W}^*)\\&=-\frac{1}{p\sqrt{d}}\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\mathbf{0},\mathbb{1})}\left[\boldsymbol{x}\sigma'(\lambda_j)\,\mathcal{E}\right].\end{aligned}$$

Now, since for any $\boldsymbol{x}^\top \boldsymbol{y}$, we have

$$\frac{d\left(\boldsymbol{x}^\top \boldsymbol{y}\right)}{dt} = \boldsymbol{x}^\top \frac{d\boldsymbol{y}}{dt} + \boldsymbol{y}^\top \frac{d\boldsymbol{x}}{dt},$$

we find

$$\frac{dq_{jl}}{dt} = -\frac{1}{pd}\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\mathbb{1})}\left[\left(\sigma'(\lambda_j)\lambda_l + \sigma'(\lambda_l)\lambda_j\right)\mathcal{E}\right].$$

Recalling the definition $\mathcal{E}_j = \sigma'(\lambda_j)\,\mathcal{E}$, the terms present inside the expectation are exactly those in the learning term of Eq.(11).

Appendix D Initial conditions and symmetric teacher

In this work we have constructed teacher matrices $\boldsymbol{W}^* \in \mathbb{R}^{k\times d}$ in order to have

$$\rho_{rs} = \frac{\boldsymbol{w}_r^{*\top}\boldsymbol{w}_s^*}{d} = \delta_{rs}\ , \tag{42}$$

where $\boldsymbol{w}_r^* \equiv [\boldsymbol{W}^*]_r \in \mathbb{R}^d$ is the r-th row of the matrix $\boldsymbol{W}^*$. We have started by sampling k vectors of dimension d uniformly on a ball of radius $\sqrt{d}$. Then we constructed an orthonormal basis using singular value decomposition.

The initial student weights $\boldsymbol{W}^0 \in \mathbb{R}^{p\times d}$ were taken as

$$\boldsymbol{W}^0 = \boldsymbol{A}\boldsymbol{W}^*\ , \tag{43}$$

with each row of $\boldsymbol{A} \in \mathbb{R}^{p\times k}$ sampled uniformly on a ball of radius one. We acknowledge choosing initial student weights as linear combinations of the teacher can be artificial and shrinks the first plateau, but our focus on this work was the specialization phase. Nevertheless, this choice and Eq. (42) are particularly suitable to theoretical analysis. Once k and p are fixed, the dimension d can be varied without changing $\boldsymbol{Q}^0$, $\boldsymbol{M}^0$ and $\boldsymbol{P}$, thereby removing any influence of different initial conditions for different d and providing the reader better visualization on the learning curves. To clarify this point, consider the j-th row $\boldsymbol{w}_j^0 \equiv [\boldsymbol{W}^0]_j \in \mathbb{R}^d$ of $\boldsymbol{W}^0$:

$$\boldsymbol{w}_j^0 = \sum_{r=1}^{k} a_{jr}\boldsymbol{w}_r^*\ , \tag{44}$$

with $a_{jr} \equiv [\boldsymbol{A}]_{jr}$. Using Eq. (42) one can write

$$q_{jl}^0 = \frac{\boldsymbol{w}_j^{0\top}\boldsymbol{w}_l^0}{d} = \sum_{r,r'=1}^{k} a_{jr}a_{jr'}\underbrace{\frac{\boldsymbol{w}_r^{*\top}\boldsymbol{w}_{r'}^*}{d}}_{=\delta_{rr'}} = \sum_{r=1}^{k} a_{jr}a_{lr}\ . \tag{45}$$

Similarly,

$$m_{jr}^0 = \frac{\boldsymbol{w}_j^{0\top}\boldsymbol{w}_r^*}{d} = a_{jr}\ . \tag{46}$$

Thus once $\boldsymbol{A}$ is fixed, the input dimension d can be varied without affecting the initial conditions. We chose to sample $\boldsymbol{a}_j \equiv [\boldsymbol{A}]_j \in \mathbb{R}^k$ on a ball of radius one both to introduce some randomness on the initialization and to keep the initial parameters bounded by one.

We stress that we use these initial conditions to make the data comparable for varying dimension d in the numerical illustrations. Our conclusions do not depend on this particular choice of initial conditions. If one simply takes random initialization $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{w}_j|\boldsymbol{0},\mathbb{1})$ for each j, the full picture we have presented in this manuscript remains unchanged. In Figure 6 we present an example of curves within the blue region (see Section 2 for the characterization of this regime) with unconstrained Gaussian initialization. Dots represent simulations, while solid lines are obtained by integration of the ODEs given by Eqs. (18), with initial conditions adjusted to match simulations.

Although varying the initial population risk with d slightly changes the exact position where the specialization transition starts, the particular initial conditions adopted in this work do not affect whether the specialization transition takes place or not, comparing to unconstrained Gaussian initialization.

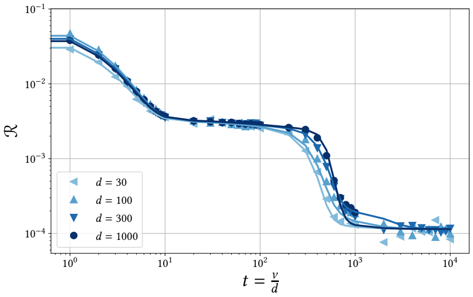

Figure 6: Population risk dynamics for $\kappa = \delta = 0$ (Saad & Solla scaling) : $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Initialization: $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{w}_j|\mathbf{0}, \mathbb{1})$ for $j = 1, ..., p_0$. Activation function: $\sigma(x) = \text{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbb{1})$. Dots represent simulations, while solid lines are obtained by integration of the ODEs given by Eqs. (18), with initial conditions adjusted to match simulations. Observe the difference on the initialization for different d.