Phase diagram of SGD in high-dimensional two-layer neural networks

Rodrigo Veiga [1,2] Ludovic Stephan [1] Bruno Loureiro [1] Florent Krzakala [1] Lenka Zdeborová [1]

[1]École Polytechnique Fédérale de Lausanne [2]Universidade de São Paulo

Setting

Teacher function:

$$y^\nu = f(\boldsymbol{x}^\nu, \boldsymbol{W}^*) + \sqrt{\Delta}\zeta,$$

with $\boldsymbol{W}^* \in \mathbb{R}^{d\times k}$ and

$$f(\boldsymbol{x}^\nu, \boldsymbol{W}^*) = \frac{1}{k}\sum_{r=1}^{k} \sigma\left(\frac{\boldsymbol{w}_r^{*\top}\boldsymbol{x}^\nu}{\sqrt{d}}\right) .$$

Learned by a *student* two-layer neural network with weights $\boldsymbol{W} \in \mathbb{R}^{d\times p}$

$$\hat{f}(\boldsymbol{x}, \boldsymbol{W}) = \frac{1}{p}\sum_{j=1}^{p} \sigma\left(\frac{\boldsymbol{w}_j^{\top}\boldsymbol{x}}{\sqrt{d}}\right)$$

through Stochastic Gradient Descent (SGD):

$$\boldsymbol{w}_j^{\nu+1} = \boldsymbol{w}_j^\nu - \gamma\nabla_{\boldsymbol{w}_j}\left(y^\nu - \hat{f}(\boldsymbol{x}^\nu, \boldsymbol{W})\right)^2$$

SGD aims to directly minimize the *population* risk $\mathcal{R}$:

$$\mathcal{R}(\boldsymbol{W}, \boldsymbol{W}^*) \equiv \mathbb{E}_x\left[\left(f(\boldsymbol{x}, \boldsymbol{W}^*) - \hat{f}(\boldsymbol{x}, \boldsymbol{W})\right)^2\right]$$

Local fields and overlaps

Everything in f and $\hat{f}$ happens through the *local fields*

$$\lambda_r^* = \frac{\boldsymbol{w}_r^{*\top}\boldsymbol{x}^\nu}{\sqrt{d}}, \quad \lambda_j = \frac{\boldsymbol{w}_j^{\top}\boldsymbol{x}^\nu}{\sqrt{d}}.$$

If $x \sim \mathcal{N}(0,1)$ everything is characterized through the *order parameters*

$$\boldsymbol{Q}^\nu \equiv \mathbb{E}\left[\boldsymbol{\lambda}^\nu\boldsymbol{\lambda}^{\nu\top}\right] = \frac{1}{d}\boldsymbol{W}^{\nu\top}\boldsymbol{W}^\nu ,$$

$$\boldsymbol{M}^\nu \equiv \mathbb{E}\left[\boldsymbol{\lambda}^\nu\boldsymbol{\lambda}^{*\nu\top}\right] = \frac{1}{d}\boldsymbol{W}^{\nu\top}\boldsymbol{W}^*$$

$$\boldsymbol{P} \equiv \mathbb{E}\left[\boldsymbol{\lambda}^{*\nu}\boldsymbol{\lambda}^{*\nu\top}\right] = \frac{1}{d}\boldsymbol{W}^{*\top}\boldsymbol{W}^* ,$$

collected into the overlap matrix

$$\boldsymbol{\Omega}^\nu = \begin{pmatrix} \boldsymbol{Q}^\nu & \boldsymbol{M}^\nu \\ \boldsymbol{M}^{\nu\top} & \boldsymbol{P} \end{pmatrix}$$

Update equations for the overlaps

The updates for the overlap matrix read

$$\begin{aligned} q_{j\ell}^{\nu+1} - q_{j\ell}^\nu &= \frac{\gamma_{\text{eff}}}{d}\underbrace{\left(\sigma'(\lambda_j^\nu)\lambda_\ell^\nu + \sigma'(\lambda_\ell^\nu)\lambda_j^\nu\right)\mathcal{E}^\nu}_{\text{learning}} + \frac{\gamma_{\text{eff}}^2}{d}\underbrace{\sigma'(\lambda_j^\nu)\sigma'(\lambda_\ell^\nu)(\mathcal{E}^\nu)^2}_{\text{variance}}, \\ m_{jr}^{\nu+1} - m_{jr}^\nu &= \frac{\gamma_{\text{eff}}}{d}\underbrace{\sigma'(\lambda_j^\nu)\,\mathcal{E}^\nu\,\lambda_r^{*\nu}}_{\text{learning}}, \end{aligned} \tag{1}$$

where $\gamma_{\text{eff}} = \gamma/p$ and

$$\mathcal{E}^\nu = y^\nu - \hat{f}(\boldsymbol{x}^\nu, \boldsymbol{W}^\nu)$$

Learning term $\Longleftrightarrow$ Gradient flow approximation

Scaling of γ_{eff} $\Longleftrightarrow$ Relative weight of learning and variance terms

Theorem: Rigorous ODE approximation

Define

$$\delta t = \frac{\gamma_{\text{eff}} \vee \gamma_{\text{eff}}^2}{d},$$

and let $\psi : \mathbb{R}^{(p+k)\times(p+k)} \to \mathbb{R}^{(p+k)\times(p+k)}$ be the expectation of the RHS of (1):

$$\psi(\boldsymbol{\Omega})_{ij} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(0,\boldsymbol{\Omega})}\left[\frac{\boldsymbol{\Omega}_{ij}^{\nu+1} - \boldsymbol{\Omega}_{ij}^\nu}{\delta t}\right]$$

Then $\boldsymbol{\Omega}$ converges to the solution $\bar{\boldsymbol{\Omega}}$ of

$$\frac{d\bar{\boldsymbol{\Omega}}}{dt} = \psi(\bar{\boldsymbol{\Omega}}),$$

with rate

$$\boxed{\|\boldsymbol{\Omega}^\nu - \bar{\boldsymbol{\Omega}}(\nu\delta t)\|_\infty \leq \frac{C(t)\ln(p)\sqrt{\gamma_{\text{eff}} \vee \gamma_{\text{eff}}^2}}{\sqrt{d}}}$$

Extension of Saad & Solla for $p \gg 1$ with nonasymptotic bound

References

[1] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[2] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, Oct 1995.

Phase transitions in γ_{eff}

$\gamma_{\text{eff}} \ll 1$, perfect learning:

$$\psi(\boldsymbol{\Omega}) = \psi_{\text{GF}}(\boldsymbol{\Omega}) + o(1)$$

Equivalent to gradient flow approximations

$\gamma_{\text{eff}} \gg 1$, variance dominates:

$$\psi(\boldsymbol{\Omega}) = \psi_{\text{var}}(\boldsymbol{\Omega}) + o(1)$$

No learning terms: $\boldsymbol{M}^\nu \approx \boldsymbol{M}_0$

$\gamma_{\text{eff}} \propto 1$: Saad & Solla line

$$\psi(\boldsymbol{\Omega}) = \psi_{\text{GF}}(\boldsymbol{\Omega}) + \psi_{\text{var}}(\boldsymbol{\Omega})$$

Some learning, then a plateau with asymptotic risk $\propto \gamma\Delta$

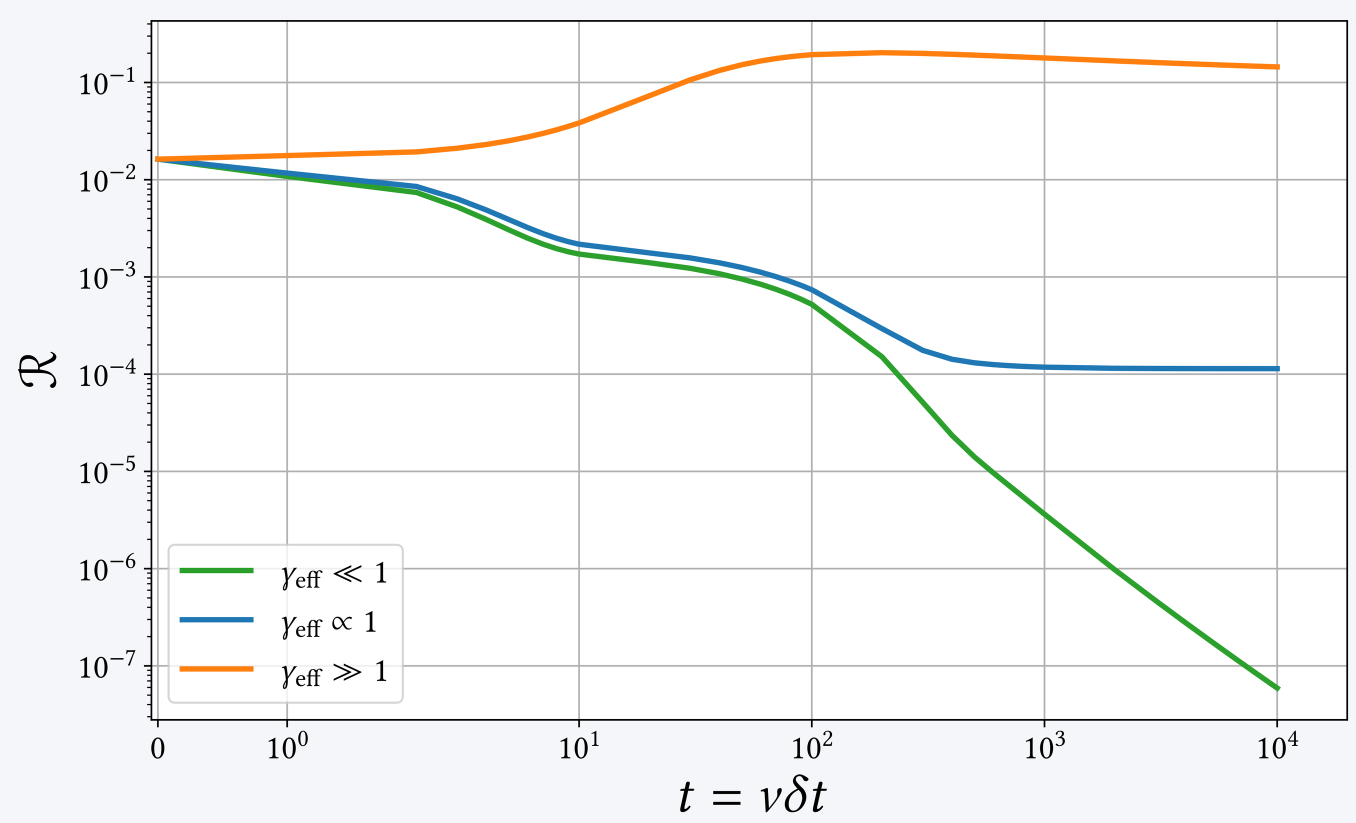

Figure 1. Illustration of the different regimes of ODEs

Interplay between γ and p

Phase transitions happen in term of $\gamma_{\text{eff}} = \gamma/p$, so we can make a two-dimensional phase diagram

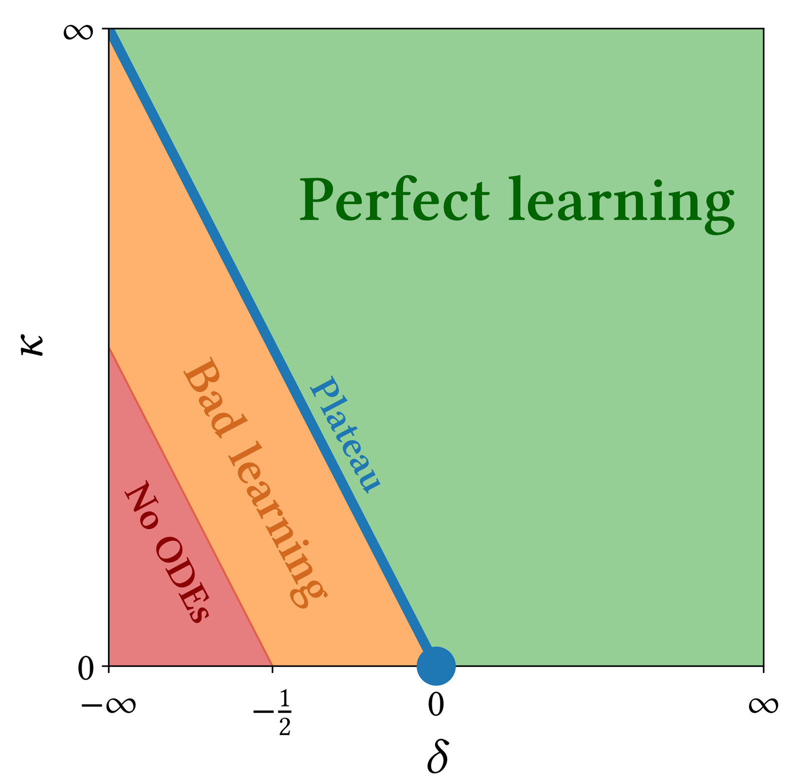

Figure 2. Phase diagram with $\gamma \sim d^{-\delta}, p \sim d^{\kappa}$

Overparametrization $\Longleftrightarrow$ Tuning the learning rate

Sample complexity

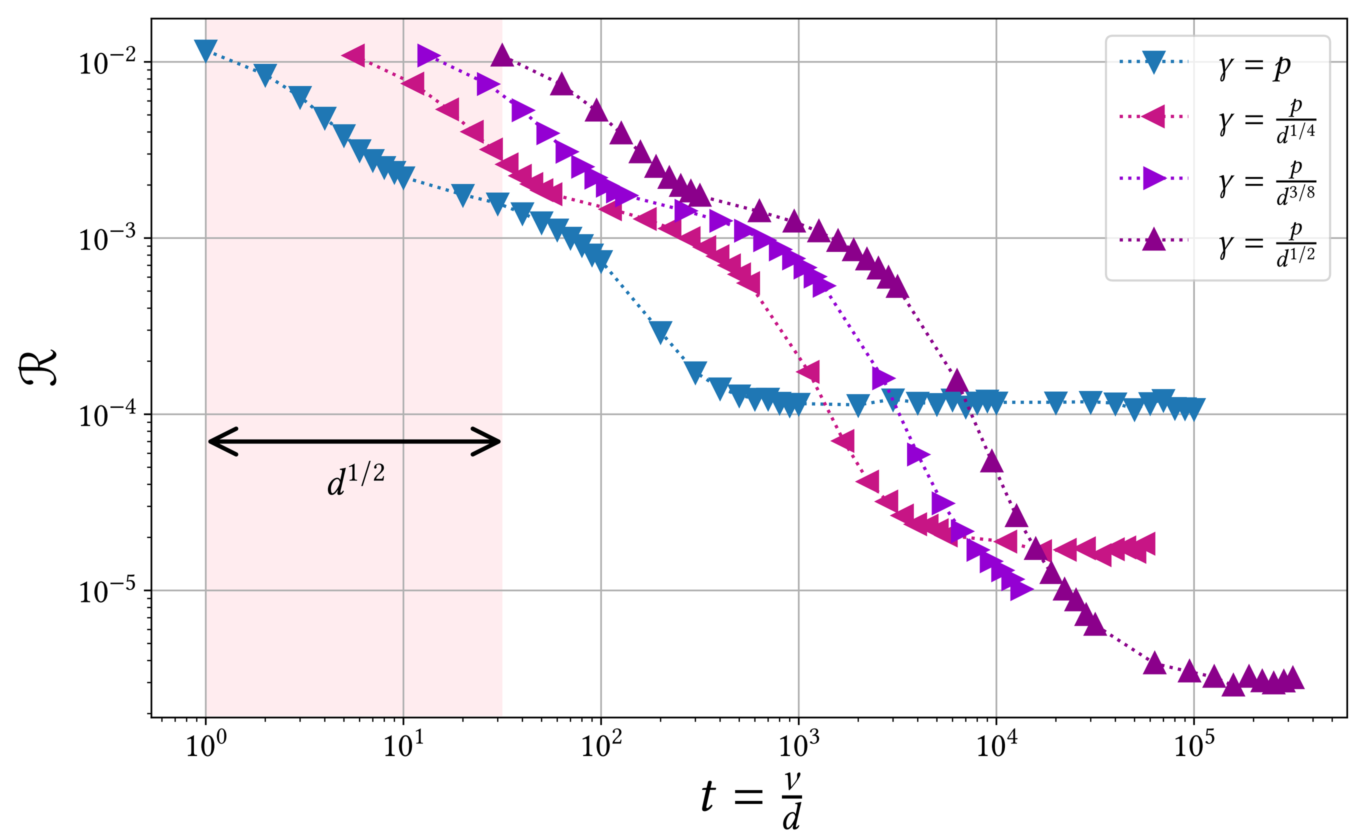

Figure 3. Effect of γ on convergence times

Tradeoff between achived minima and sample complexity