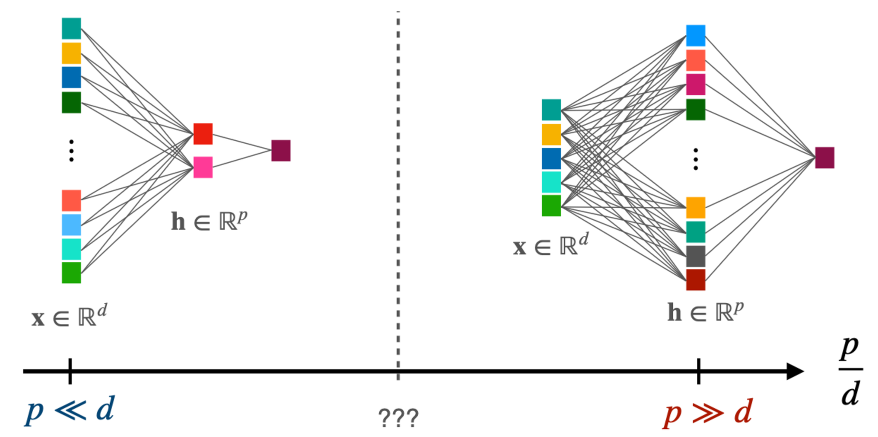
From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks

Luca Arnaboldi[1], Ludovic Stephan[1], Florent Krzakala[1], Bruno Loureiro[2]

[1] IdePHICS, EPFL; [2] ENS Paris

Motivation

- Understanding the performance of SGD in neural networks is a major endeavor in machine learning, and significant progress was achieved in the context of large two-layer neural net
- High dimensional limit has been investigated first in the seminal work of [3], developing some ODEs for the dynamics.
- The optimization over wide two-layer neural networks can be rigorously studied using a well-defined PDE [4, 5]



Aim: *drawing a precise connection between two limits*

Teacher-student model with online SGD

We introduce a teacher-student two-layer neural network model for studying the dynamics of the training with SGD:

- Input data is generated from independent Gaussian distributions:

$$\boldsymbol{x}^\nu \sim \mathcal{N}\left(0_d, \frac{1}{d} \mathbb{1}_d\right)$$

  Labels are generated by a **teacher network**

$$y^\nu = \frac{1}{k} \sum_{r=1}^{k} a_r^\star \sigma^\star(\boldsymbol{w}_r^{\star\top} \boldsymbol{x}^\nu) + \sqrt{\Delta} z^\nu, \qquad z^\nu \sim \mathcal{N}(0,1)$$

  where Δ is the artificial noise.
- The student network to be learned is

$$f_\Theta(\boldsymbol{x}) = \frac{1}{p} \sum_{i=1}^{p} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x})$$

- We are using the **square loss function**. The population risk is given by:

$$\mathcal{R}(\Theta) := \mathbb{E}_{(\boldsymbol{x},y)\sim\rho}\left[\frac{1}{2}(f_\Theta(\boldsymbol{x}) - y)^2\right] + \frac{\Delta}{2}$$

- We are using the online stochastic gradient descent:

$$\Theta^{\nu+1} = \Theta^\nu - \gamma \nabla_\Theta \ell(f_{\Theta^\nu}(\boldsymbol{x}^\nu), y^\nu), \quad \nu \leq n$$

Assumptions

Symplifying assumptions

- $a_r^\star = 1$ and $a_i^\nu = 1$;
- $W^\star$ is full-rank;
- p is divisible by k;

Technical assumptions

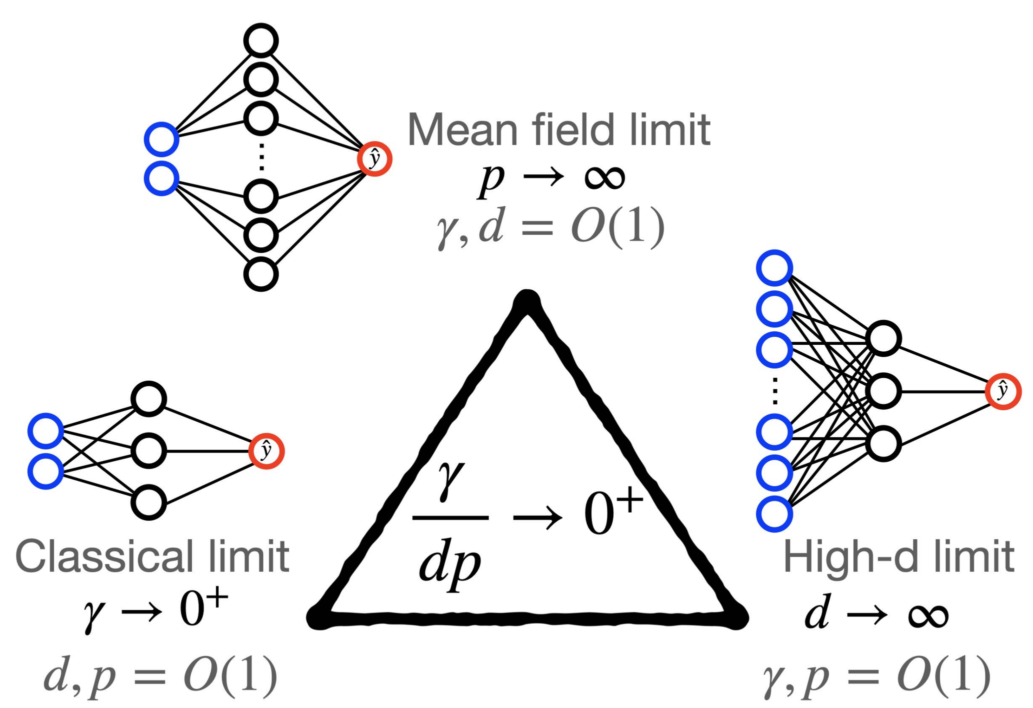
- With high probability:

$$\forall i \in [p], \quad \|\boldsymbol{w_i}^2\| \leq K.$$

- $\|\sigma^{(i)}\|_\infty \leq K$ for $i = 0, 1, 2$.

Main goal

We aim for a **common description for 3 different limits**



In particular a low-dimensional analysis for joint high-dimensional mean field limit!

Low-dimensional sufficient statistics

Overlaps are **sufficient statistics**:

$$\Omega^\nu := \begin{pmatrix} Q^\nu & M^\nu \\ M^{\nu\top} & P \end{pmatrix} = \frac{1}{d}\begin{pmatrix} W^\nu W^{\nu\top} & W^\nu W^{\star\top} \\ W^\star W^{\nu\top} & W^\star W^{\star\top} \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)}$$

We can derive a closed set of stochastic processes

$$\begin{aligned} M_{ir}^{\nu+1} - M_{ir}^\nu &= \frac{\gamma}{pd} \sigma'(\lambda_i^\nu) \lambda_r^{\star\nu} \mathcal{E}^\nu \\ Q_{ij}^{\nu+1} - Q_{ij}^\nu &= \frac{\gamma}{pd}\left(\sigma'(\lambda_i^\nu)\lambda_j^\nu + \sigma'(\lambda_j^\nu)\lambda_i^\nu\right)\mathcal{E}^\nu \\ &\quad + \frac{\gamma^2 \|\boldsymbol{x}^\nu\|_2^2}{p^2 d} \sigma'(\lambda_i^\nu)\sigma'(\lambda_j^\nu)\mathcal{E}^{\nu 2} \end{aligned} \qquad \text{(OV-SP)}$$

where the *local fields* are jointly Gaussian vectors

$$(\boldsymbol{\lambda}^\nu, \boldsymbol{\lambda}^{\star\nu}) \sim \mathcal{N}(\mathbf{0}_{p+k}, \Omega^\nu).$$

Informally, when $\frac{\gamma}{pd}$ there is ODEs approximation

$$\begin{aligned} \frac{\mathrm{d}M}{dt} &= \Psi^{(M)}(\Omega), \\ \frac{\mathrm{d}Q}{\mathrm{d}t} &= \Psi^{(GF)}(\Omega) + \frac{\gamma}{p}\Psi^{(Var)}(\Omega), \end{aligned} \qquad \text{(SS-ODE)}$$

$$\begin{aligned} \Psi_{ir}^{(M)}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_i)\lambda_r^\star \mathcal{E}\right] \\ \Psi_{ij}^{(GF)}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\left(\sigma'(\lambda_i)\lambda_j + \sigma'(\lambda_j)\lambda_i\right)\mathcal{E}\right] \\ \Psi_{ij}^{(Var)}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_i)\sigma'(\lambda_j)\mathcal{E}^2\right] \end{aligned}$$

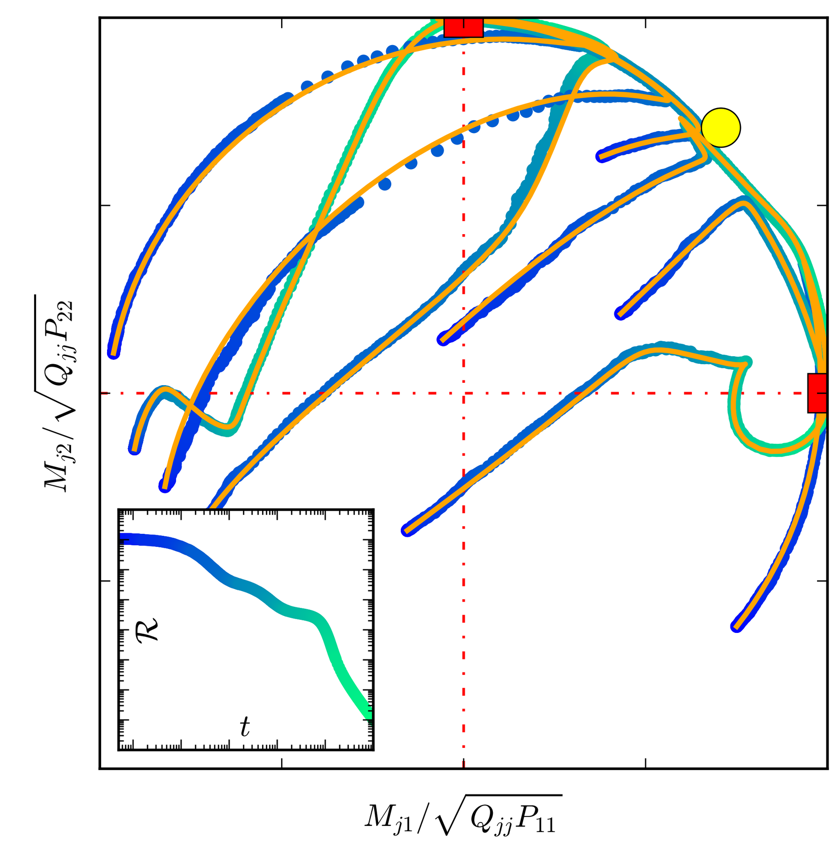
Theorem (Veiga et al. [2])

Let Ω^ν be the random process of Eq. (OV-SP), and $\Omega(t)$ the solution to the ODE (SS-ODE) with starting point $\Omega(0) = \Omega^0$. Define the stepsize $\delta t = \frac{\gamma}{pd}$, and assume that $\gamma/p = O(1)$. Then there exists a constant $C > 0$ such that for any $\nu \geq 0$,

$$\|\Omega^\nu - \Omega(\nu\delta t)\|_\infty \leq e^{C\nu\delta t}\sqrt{\frac{\gamma}{pd}}$$

Key observation

(SS-ODE) hold whenever $\frac{\gamma}{pd} \to 0$, not only when $d \to +\infty$.



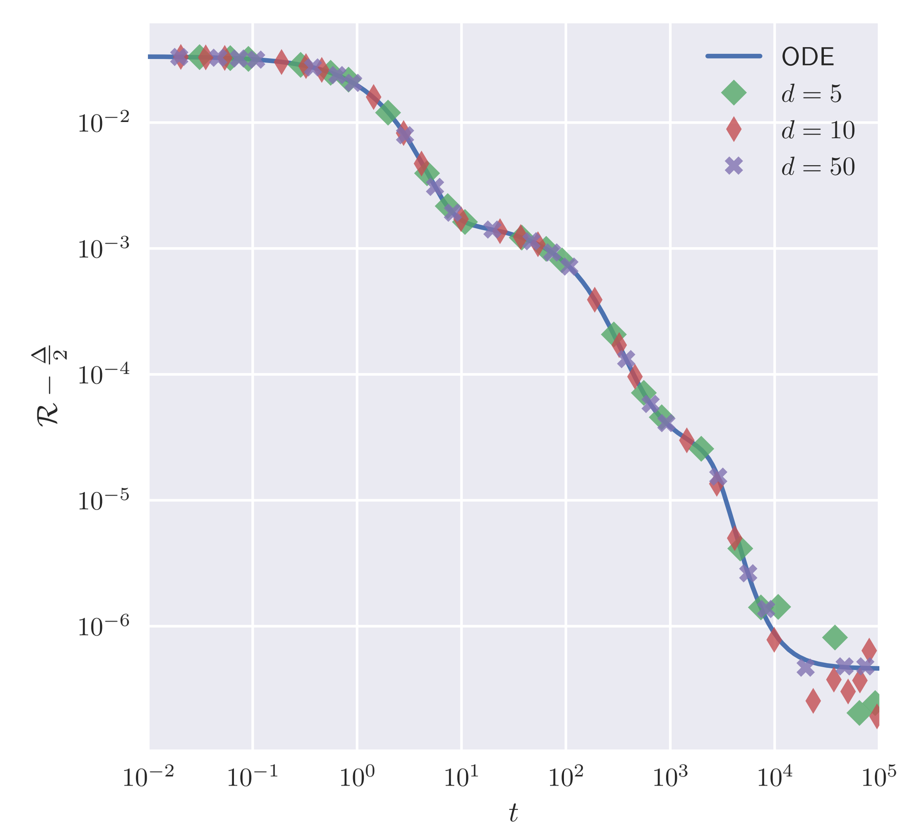
Classical limit ($\gamma \to +0$)

The equations are equivalent to the gradient flow equations

$$\frac{\mathrm{d}\boldsymbol{w}_i}{\mathrm{d}t} = \nabla_{\boldsymbol{w}_i} \mathcal{R}(\Theta)$$

Dimension Independence

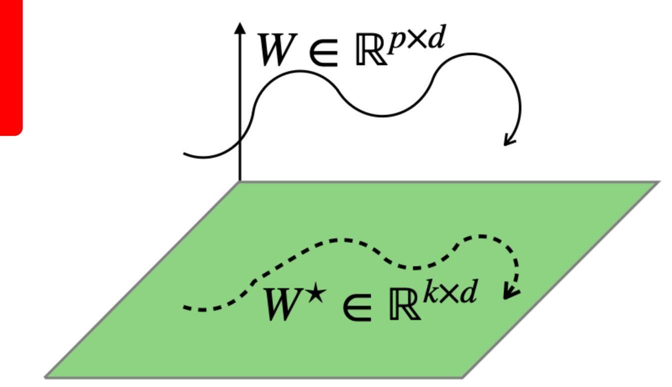
(SS-ODE) is independent of the data dimension d.



The trajectories are **exactly in the same** whether d is large or small, when starting from the same initial condition.

Overparametrized regime ($p \to +\infty$)

Decompose the student between the teacher's space and its orthogonal



$$W = MP^{-1}W^\star + W^\perp$$
$$Q^\perp = Q - MP^{-1}M^\top$$
$$\lambda^\perp \sim \mathcal{N}\left(\mathbf{0}_p, Q^\perp\right)$$

$Q^\perp$ can replace Q and the overlaps are still sufficient, with ODE:

$$\frac{dQ_{ij}^\perp}{dt} = \mathbb{E}_{(\boldsymbol{\lambda}^\perp,\boldsymbol{\lambda})}\left[\left(\sigma'(\lambda_i)\lambda_j^\perp + \sigma'(\lambda_j)\lambda_i^\perp\right)\mathcal{E}\right] := \Psi_{ij}^\perp(\Omega).$$

Low dimensional mean-field

Assumption $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_p)$ are drawn i.i.d from an orthogonally invariant and $\frac{K^2}{d}$-subgaussian distribution.
Ansatz Just in the dynamic equation, the weights

$$\boldsymbol{w}_i^\perp \approx \sqrt{Q_{ii}^\perp} \cdot \boldsymbol{g}_i \quad \tilde{Q} = MP^{-1}M^\top + D_{\sqrt{Q^\perp}} \Xi D_{\sqrt{Q^\perp}}$$

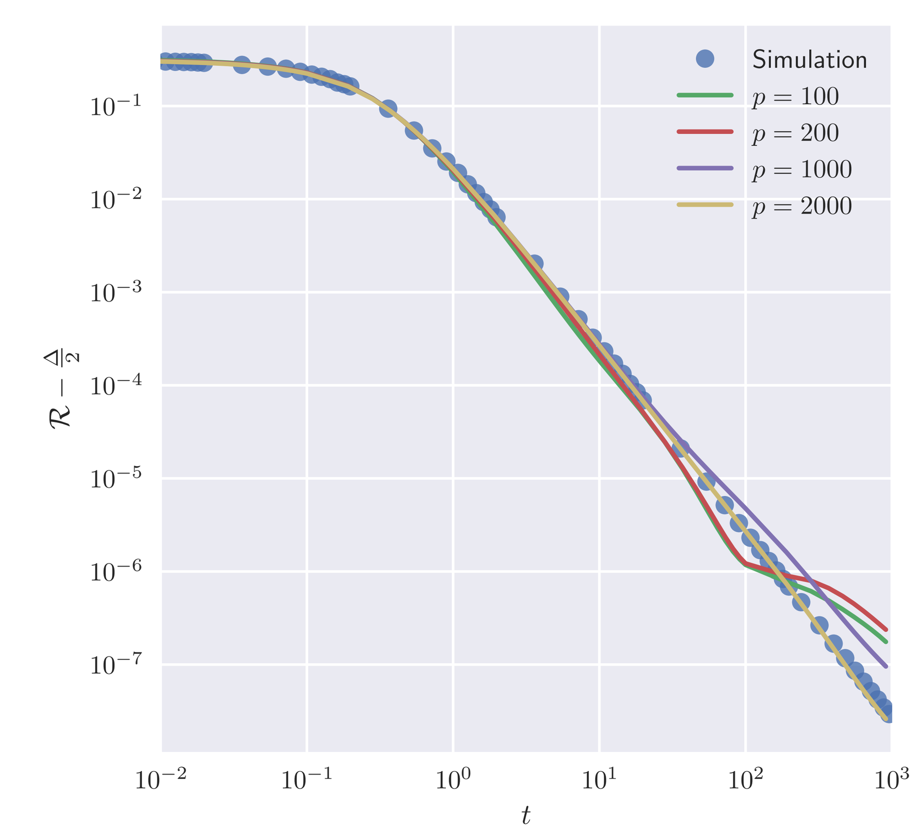
where $D_{\sqrt{Q^\perp}} = \sqrt{\mathrm{diag}\left(Q^\perp\right)}$ and

$$\Xi_{ii} = 1, \quad \Xi_{ij} = \langle \boldsymbol{g}, \boldsymbol{g'} \rangle \quad \text{with} \quad \boldsymbol{g}, \boldsymbol{g'} \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$

Reduced ODEs and risk

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mathbb{E}_\Xi\left[\Psi^{(M)}(\tilde{\Omega})\right] \quad \frac{\mathrm{d}Q_{ii}^\perp}{\mathrm{d}t} = \mathbb{E}_\Xi\left[\Psi_{ii}^\perp(\tilde{\Omega})\right] \quad \text{(MF-ODE)}$$

$$\mathcal{R}(\tilde{\Theta}) = \mathbb{E}_\Xi\left[\mathcal{R}(\tilde{\Omega})\right]$$



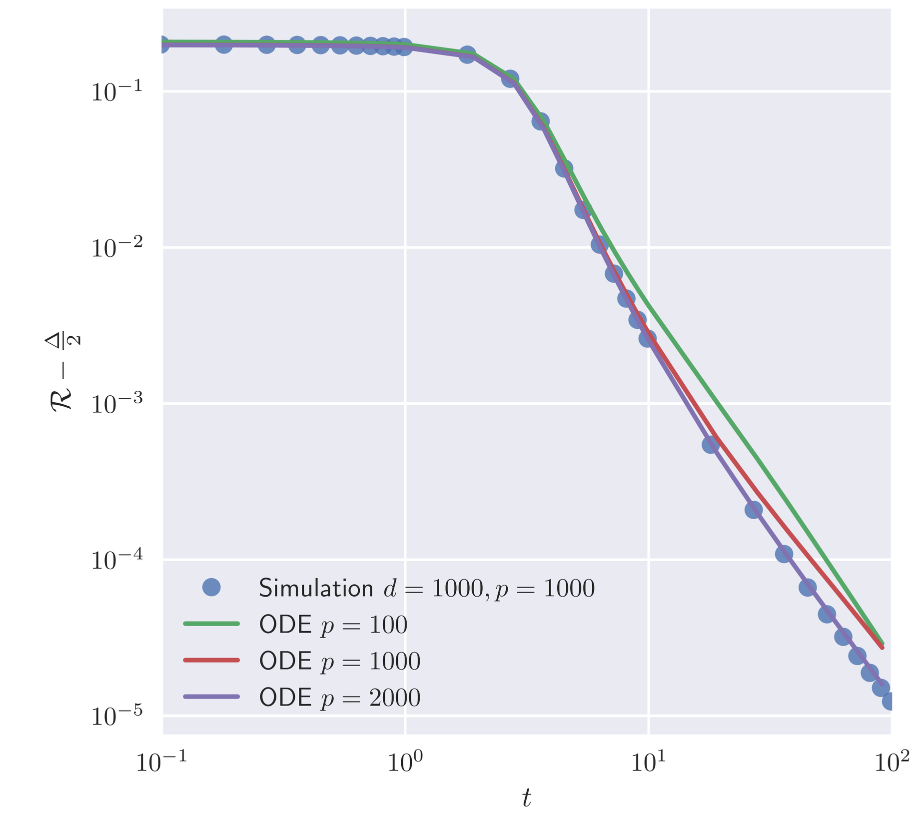
Theorem. *Let $\Omega(t)$ and $\tilde{\Theta}(t)$ denote the solutions of* (SS-ODE) *and* (MF-ODE)*, respectively. Then with probability at least $1 - e^{-z^2}$ on the initialization:*

$$\sup_{t\in[0,T]}\left|\mathcal{R}(\Omega(t)) - \mathcal{R}(\tilde{\Theta}(t))\right| \leq Ce^{CT}\left(\sqrt{\log(pT)} + z\right)/\sqrt{p}.$$

High-dimensional mean-field

When $d \to +\infty$ the random matrix Ξ is the identity. (MF-ODE) becomes

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \Psi^{(M)}(\bar{\Omega}) \qquad \frac{\mathrm{d}Q_{ii}^\perp}{\mathrm{d}t} = \Psi_{ii}^\perp(\bar{\Omega}). \qquad \text{(HDMF-ODE)}$$



Theorem. *Let $\Omega(t)$ and $\bar{\Theta}(t)$ denote the solutions of* (SS-ODE) *and* (HDMF-ODE)*, respectively. Then with probability at least $1 - e^{-z^2}$ on the initialization:*

$$\sup_{t\in[0,T]}\left|\mathcal{R}(\Omega(t)) - \mathcal{R}(\bar{\Theta}(t))\right| \leq Ce^{CT}\left(\frac{\sqrt{\log(pT)} + z}{\sqrt{p}} + \frac{1}{\sqrt{d}}\right)$$

(HDMF-ODE) are the **particle dynamics** of a measure over the sufficient statistics. Introducing a measure $\mu_{(\boldsymbol{m},q)}$ over $\mathbb{R}^{k+1}$ as

$$\mu_t := h_\# \mu_t \quad \text{where } h(\boldsymbol{w}) = \left(\frac{W^\star \boldsymbol{w}}{d}, \|\boldsymbol{w}^\perp\|^2\right)$$

the evolution can be written as Wasserstein GD

$$\partial_t \mu_{(\boldsymbol{m},q)} = \nabla_{(\boldsymbol{m},q)} \cdot \left(\mu_{(\boldsymbol{m},q)} \varphi(\,\cdot\,, \mu_{(\boldsymbol{m},q)})\right)$$

References

[1] ***From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks***, Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro arXiv: 2302.05882 [stat.ML]

[2] *Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks*, Rodrigo Veiga et al. Advances in Neural Information Processing Systems, 2022

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

[4] *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Lénaic Chizat and Francis Bach. Advances in Neural Information Processing Systems, 2018

[5] *A mean field view of the landscape of two-layer neural networks*, Song Mei, Andrea Montanari, and Phan-Minh Nguyen Proceedings of the National Academy of Sciences, 115(33):E7665–E7671, 2018.