Topics in Math of ML, Lecture 11
Dynamics of Wide Two-Layer Perceptrons

Lénaïc Chizat*

December 1, 2025

In today's lecture, we analyze the behavior of gradient methods on two-layer neural networks, aka two-layer perceptrons (2LPs), focusing on the large width asymptotics.

1 Effective Hyperparameters (HP) of GD

We start with a general remark on the link between three HPs for GD on general models.

Consider a differentiable function $\mathcal{L} : \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_L} \to \mathbb{R}$ and initial weights directions $\boldsymbol{D} = (D^{(1)}, \dots, D^{(L)}) \in \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_L}$. You can think of $\mathcal{L}$ as the training objective for a neural network with L layers with weight matrices $W^{(\ell)} \in \mathbb{R}^{p_\ell}$ for $\ell \in [1 : L]$ which are are, at the beginning of GD, initialized with $D^{(\ell)}$ multiplied by a positive scalar. Once $\mathcal{L}$ and $\mathbf{D}$ are fixed, the GD dynamics is a priori characterized by three sets of HPs in $(\mathbb{R}_+^*)^L$:

- scalar multipliers $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)$
- initialization scales $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_L)$
- learning rates $\boldsymbol{\eta} = (\eta_1, \dots, \eta_L)$

Consider the scaled objective

$$\mathcal{L}_{\boldsymbol{\alpha}}(W^{(1)}, \dots, W^{(L)}) = \mathcal{L}(\alpha_1 W^{(1)}, \dots, \alpha_L W^{(L)})$$

and the GD dynamics $(\theta_k)_{k\geq 0} = (W_k^{(1)}, \dots, W_k^{(L)})_{k\geq 0}$ defined via

$$W_0^{(\ell)} = \sigma_\ell D^{(\ell)}, \qquad W_{k+1}^{(\ell)} = W_k^{(\ell)} - \eta_\ell \nabla_\ell \mathcal{L}_{\boldsymbol{\alpha}}(\theta_k), \qquad \ell \in [1 : L],\ k \geq 0.$$

where $\nabla_\ell \mathcal{L}$ denotes the ℓ-th block of the gradient of $\mathcal{L}$.

The following proposition shows that there are only *two effective HP per block* instead of three: the effective scale $\alpha_\ell \sigma_\ell$ and the effective LR $\alpha_\ell^2 \eta_\ell$, and two dynamics with identical effective HP (for all blocks) are equivalent.

Proposition 1.1. *Fix $\mathcal{L}$ and $\mathbf{D}$ as above. Let $(\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\eta}) \in (\mathbb{R}_+^*)^{3L}$ and $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\eta}}) \in (\mathbb{R}_+^*)^{3L}$ be two sets of HPs such that $(\alpha_\ell \sigma_\ell, \alpha_\ell^2 \eta_\ell) = (\tilde{\alpha}_\ell \tilde{\sigma}_\ell, \tilde{\alpha}_\ell^2 \tilde{\eta}_\ell)$, $\forall \ell \in [1 : L]$. Then denoting $(W_k^{(1)}, \dots, W_k^{(L)})_{k\geq 0}$ and $(\tilde{W}_k^{(1)}, \dots, \tilde{W}_k^{(L)})_{k\geq 0}$ the corresponding sequences of GD iterates, it holds*

$$(\alpha_1 W_k^{(1)}, \dots, \alpha_L W_k^{(L)}) = (\tilde{\alpha}_1 \tilde{W}_k^{(1)}, \dots, \tilde{\alpha}_L \tilde{W}_k^{(L)}), \quad \forall k \geq 0.$$

*EPFL lenaic.chizat@epfl.ch

Proof. The claim is true for $k = 0$ since by construction

$$(\alpha_1 W_0^{(1)}, \dots, \alpha_L W_0^{(L)}) = (\alpha_1 \sigma_1 D^{(1)}, \dots, \alpha_L \sigma_L D) = (\tilde{\alpha}_1 \tilde{W}_0^{(1)}, \dots, \tilde{\alpha}_L \tilde{W}_0^{(L)}).$$

Let us assume that the claim is true at $k \in \mathbb{N}$. Then $\forall \ell \in [1 : L]$,

$$\begin{aligned}
\alpha_\ell W_{k+1}^{(\ell)} &= \alpha_\ell W_k^{(\ell)} - \alpha_\ell \eta_\ell \nabla_\ell \mathcal{L}_{\boldsymbol{\alpha}}(\theta_k) \\
&= \alpha_\ell W_k^{(\ell)} - \alpha_\ell^2 \eta_\ell \nabla_\ell \mathcal{L}(\alpha_1 W^{(1)}, \dots, \alpha_L W^{(L)}) \\
&= \tilde{\alpha}_\ell \tilde{W}_k^{(\ell)} - \tilde{\alpha}_\ell^2 \tilde{\eta}_\ell \nabla_\ell \mathcal{L}(\tilde{\alpha}_1 \tilde{W}^{(1)}, \dots, \tilde{\alpha}_L \tilde{W}^{(L)}) \\
&= \tilde{\alpha}_\ell \tilde{W}_{k+1}^{(\ell)}.
\end{aligned}$$

This shows that the property is hereditary and concludes the proof. □

We note that α_ℓ and σ_ℓ do not play exactly interchangeable roles; because $\sigma_\ell = 0$ is often a valid nontrivial choice while $\alpha_\ell = 0$ is not (it sets the whole block of weights to 0 throughout the dynamics).

2 Phase diagram of GD for 2LPs

Let us discuss the large-width behavior of 2LP as the width diverge. We consider two HPs for the whole system: a scalar multiplier and a LR.

For a sequence of multipliers $(\alpha_M)_{M \in \mathbb{N}^*}$, consider a 2LP $\hat{h}_\theta : \mathbb{R}^d \to \mathbb{R}$ of width $M \in \mathbb{N}^*$ defined by

$$\hat{h}_\theta(x) = \alpha_M \sum_{j=1}^{M} \phi(z_j, x)$$

where $\theta = (z_j)_{j=1}^m \in (\mathbb{R}^p)^m$. *We assume in the following that ϕ and its derivatives up to order two in z are bounded* uniformly in x. This assumption allows to avoid many technicalities in the proofs but it is not true for the vanilla 2LP, obtained with $\phi(z_j, x) = v_j \rho(u_j^\top x)$ where $\rho : \mathbb{R} \to \mathbb{R}$ is a smooth nonlinearity and $z_j = (u_j, v_j) \in \mathbb{R}^d \times \mathbb{R}$ (i.e. $p = d + 1$). Given a training set $(x_i, y_i)_{i=1}^n \in (\mathbb{R}^d \times \mathbb{R})^n$, consider the ERM with a smooth loss :

$$\hat{\mathcal{L}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{loss}(y_i, \hat{h}_\theta(x_i)).$$

This function $\hat{\mathcal{L}}$ is in general *not convex* even if loss is so it is not clear how to tackle its minimization a priori.

Consider gradient descent (GD) on $\hat{\mathcal{L}}$ starting from a random initialization. Let $\mu_0 \in \mathcal{P}(\mathbb{R}^p)$ be a zero mean distribution. The equations defining the GD iterates $(\theta_k)_k = (Z_k^j)_{k,j}$ with a LR $\eta_0 \eta_M$ (decomposed as the product of a "master" LR η_0 independent of M and η_M which is a power of M), are

$$\hat{Z}_0^j \overset{iid}{\sim} \mu_0, \qquad \hat{Z}_{k+1}^j = \hat{Z}_k^j - \eta_0 \eta_M \nabla_{z^j} \hat{\mathcal{L}}(\theta_k) \tag{1}$$

$$= \hat{Z}_k^j - \frac{\eta_0 \eta_M \alpha_M}{n} \sum_{i=1}^{n} D\phi(\hat{Z}_k^j, x_i)^\top \nabla \mathrm{loss}(y_i, \hat{h}_k(x_i)), \qquad \hat{h}_k = \hat{h}_{\theta_k} \tag{2}$$

where $D\phi$ denotes the Jacobian of ϕ in its first variable z. In predictor space, we have for any $x \in \mathbb{R}^d$, by a first-order Taylor expansion

$$\hat{h}_{k+1}(x) = \hat{h}_k(x) - \frac{\eta_0 \eta_M \alpha_M^2}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} D\phi(\hat{Z}_k^j, x) D\phi(\hat{Z}_k^j, x_i)^\top \nabla \mathrm{loss}(y_i, \hat{h}_k(x_i)) + O(\alpha_M (\eta_0 \eta_M \alpha_M)^2) \tag{3}$$

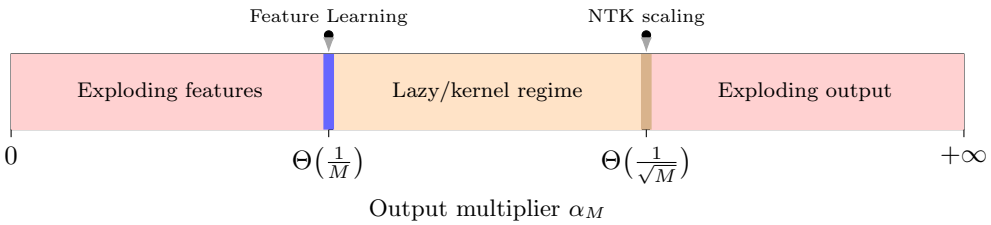

Figure 1: Phase diagram of two-layer perceptron.

Scale of the first forward pass Let us assume that $\mathbf{E}[\phi(\hat{Z}_0, x)] = 0$ and $\mathbf{E}[\phi(\hat{Z}_0, x)^2] > 0$ (which holds for standard choices of initialization). Then at $k = 0$, we have for any $x \in \mathbb{R}^d$

$$\mathbf{E}[\hat{h}_0(x)] = 0, \qquad \mathrm{Var}[\hat{h}_0(x)] = M\alpha_M^2 \mathbf{E}[\phi(\hat{Z}_0, x)^2] = \Theta(M\alpha_M^2)$$

From now on, we assume non-exploding variance of the first output, that is $M\alpha_M^2 = O(1)$.

Scale of the first predictor update As $\eta_0 \to 0$, we want predictor updates in $\Theta(\eta_0)$ (independent of width). From the equation above, for any $x \in \mathbb{R}^d$,

$$|\hat{h}_1(x) - \hat{h}_0(x)| = \Theta(\eta_0) \iff M\eta_0\eta_M\alpha_M^2 = \Theta(\eta_0) \iff \eta_M = \Theta\Big(\frac{1}{M\alpha_M^2}\Big). \tag{4}$$

(Note that, for convenience, in the rest, we write $\Theta(1)$ for quantities which are $O(1)$ and are $\Omega(1)$ besides a few irrelevant degenerate cases).

Scale of the first feature update For a given $x \in \mathbb{R}^d$, does the vector of features $(\phi(\hat{Z}_k^j, x))_{j=1}^M$ evolve with k? This question, in our present context, is equivalent to looking at the scale of the updates of $(\hat{Z}_k^j)$. From the above equations, we have

$$\|\hat{Z}_1^j - \hat{Z}_0^j\| = \Theta(\eta_0\eta_M\alpha_M) = \Theta\Big(\frac{\eta_0}{M\alpha_M}\Big) \tag{5}$$

where the last equality holds when (4) holds. Comparing the feature update with η_0, we deduce the phase diagram on Figure 1 with the following behaviors:

- **Feature learning**: the features evolve at the same time-scale as the predictor
- **Lazy (a.k.a. kernel) regime**: the features evolve at a slower time-scale than the predictor
- **Exploding features**: the features evolve at a faster time-scale than the predictor ;
- **NTK scaling**: scaling within the lazy regime where the first output has nontrivial variance

Remark 2.1. *This is a simplified phase diagram with only 2 HPs, but the complete picture for 2LP would involve 4 HPs (init. scale and LR for each layer, as shown in the previous section). In particular, a more complete analysis (with different scalings for both layers) shows that there are well-behaved limits where the initial output weights are* 0.

In order to derive this phase diagram, it was enough to look the at time-derivatives (as $\eta_0 \to 0$) of various quantities at time 0. In what follows we look at specific phases and describe the whole dynamics.

3 Feature learning regime: mean-field limit

In this section, we consider the feature learning regime, that is $\alpha_M = 1/M$, $\eta_M = M$ and $\eta_0 = \Theta(1)$. We recall that ϕ is assumed Lipschitz and bounded.

Limit dynamics For Z be a $\mathbb{R}^p$-valued random variable, we define

$$h_Z(x) = \mathbf{E}[\phi(Z, x)], \qquad \mathcal{L}(Z) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{loss}(y_i, h_Z(x_i)).$$

Observe that h_Z and $\mathcal{L}(Z)$ are deterministic and in fact only depend on $\mathrm{Law}(Z)$. Define the L^2 GD dynamics

$$\begin{aligned} Z_0 \sim \mu_0, \qquad Z_{k+1} &= Z_k - \eta_0 \nabla \mathcal{L}(Z_k) \\ &= Z_k - \frac{\eta_0}{n}\sum_{i=1}^{n} D\phi(Z_k, x_i)^\top \nabla \mathrm{loss}(y_i, h_k(x_i)), \qquad h_k = h_{Z_k} \\ &= Z_k + F(Z_k, (h_k(x_i))_{i\in[1:n]}) \end{aligned}$$

where F is defined by this equation.

Tight error bound Our goal is to prove a quantitative rate of convergence of $(\hat{h}_k)_k = (\hat{h}_{\theta_k})_k$ towards $(h_k)_k = (h_{Z_k})_k$, and also a convergence result in parameter space, in a suitable sense. The argument goes as follows.

- Let $(Z_k^j)_{k\geq 0}$ be iid samples of the limit dynamics $(Z_k)_{k\geq 0}$ such that $Z_0^j = \hat{Z}_0^j$, $\forall j \in [1:M]$
- Observe that we have

  $$\hat{Z}_{k+1}^j = \hat{Z}_k^j + F(\hat{Z}_k^j, (\hat{h}_k(x_i))_{i\in[1:n]}), \qquad Z_{k+1}^j = Z_k^j + F(Z_k^j, (h_k(x_i))_{i\in[1:n]})$$

  for the same function F. The only difference between the two dynamics is that the predictor is a finite sum for the 2LP (on the left), and an expectation for the limit (on the right).
- Let $\Delta_k = \max_{j\in[1:M]} \|Z_k^j - \hat{Z}_k^j\|$. By construction, it holds $\Delta_0 = 0$. Let us bound Δ_k.
- Since F is Lipschitz continuous, there exists $c > 0$ such that

  $$\Delta_{k+1} \leq \Delta_k + c(\Delta_k + \max_i |\hat{h}_k(x_i) - h_k(x_i)|)$$

- By Hoeffding's lemma, for fixed k and i, it holds with probability at least $1-\delta$

  $$\Big|\frac{1}{M}\sum_{j=1}^{M} \phi(Z_k^j, x_i) - \mathbf{E}[\phi(Z_k, x_i)]\Big| \leq c\Big(\frac{1+\sqrt{\log(1/\delta)}}{\sqrt{M}}\Big)$$

- By a union bound over $k \leq K$ and $i \leq n$, it holds with probability at least $1-\delta$, for all $k \leq K$,

  $$\max_{i\in[1:n]} |\hat{h}_k(x_i) - h_k(x_i)| = \max_{i\in[1:n]} \Big|\frac{1}{M}\sum_{j=1}^{M} \phi(\hat{Z}_k^j, x_i) - \mathbf{E}[\phi(Z_k, x_i)]\Big| \leq c'\Big(\Delta_k + \frac{1+\sqrt{\log(nK/\delta)}}{\sqrt{M}}\Big)$$

  hence

  $$\Delta_{k+1} \leq (1+c'')\Delta_k + c''\frac{1+\sqrt{\log(nK/\delta)}}{\sqrt{M}}$$

- We conclude by discrete Grönwall's lemma to get the following statement:

Theorem 3.1. *There exists $c > 0$ (that may depend on ϕ, K and η_0) such that with probability at least $1 - \delta$, it holds*

$$\max_{k \le K} \max_{i \in [1:n]} |\hat{h}_k(x_i) - h_k(x_i)| \le c \frac{1 + \sqrt{\log(n/\delta)}}{\sqrt{M}} \tag{6}$$

$$\max_{k \le K} \max_{j \in [1:M]} \|\hat{Z}_k^j - Z_k^j\| \le c \frac{1 + \sqrt{\log(n/\delta)}}{\sqrt{M}} \tag{7}$$

Remark 3.2 (Particle system interpretation)**.** *Suppose that ℓ is the square loss and that the Bayes predictor is given by another 2-layer NN $h_{\theta^*}(x) = \frac{1}{M} \sum_{j=1}^M \phi(z_j^*, x)$. Then the objective reads*

$$\begin{aligned} F(\theta) &= \frac{1}{2n} \sum_{i=1}^n \Big(\frac{1}{M} \sum_{j=1}^M \phi(z_j, x_i) - \frac{1}{M} \sum_{j=1}^M \phi(z_j^*, x_i) \Big)^2 \\ &= \frac{1}{2M^2} \sum_{j,j'} k(z_j, z_{j'}) - \frac{1}{M^2} \sum_{j,j'} k(z_j, z_{j'}^*) + \frac{1}{2M^2} \sum_{j,j'} k(z_j^*, z_{j'}^*) \end{aligned}$$

where $k(z, z') = \frac{1}{n} \sum_{i=1}^n \phi(z, x_i) \phi(z', x_i)$. Thus GD dynamics can be interpreted as a system of interacting particles (with infinite viscosity, because there is no momentum) with attraction-repulsion interaction given by the kernel k (with the $(z_j^*)_{j=1}^M$ fixed). This gives one connection between mathematical physics and the dynamics of NNs (many others exist!).

4 Dynamics in the space of probability distributions

Consider the continuous-time infinite width dynamics (globally well-defined e.g. if $\mathcal{L}$ is lower-bounded)

$$Z_0 \sim \mu_0, \qquad \dot{Z}_t = -\nabla \mathcal{L}(Z_t) = -\frac{1}{n} \sum_{i=1}^n D\phi(Z_t, x_i)^\top \nabla \mathrm{loss}(y_i, \mathbf{E}[\phi(Z_t, x_i)]) \tag{8}$$

Let us show that in fact this dynamics is closed in terms of $\mu_t = \mathrm{Law}(Z_t)$ and can be expressed as a PDE.

Proposition 4.1. *Let $\mu_t = \mathrm{Law}(Z_t)$ for $t \ge 0$ and define for any $\mu \in \mathcal{P}(\mathbb{R}^p)$ the velocity field*

$$v[\mu](z) = -\frac{1}{n} \sum_{i=1}^n D\phi(z, x_i)^\top \nabla \mathrm{loss}\Big(y_i, \int \phi(z', x_i) \mathrm{d}\mu(z') \Big).$$

Then $(\mu_t)_{t \ge 0}$ solves the partial differential equation

$$\partial_t \mu_t = -\nabla \cdot \big(v[\mu_t] \mu_t \big),$$

in the weak sense, in the sense that for any test function $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^p)$, and $t \ge 0$ it holds

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\int_{\mathbb{R}^p} \varphi \mathrm{d}\mu_t \right) = \int_{\mathbb{R}^p} \langle \nabla \varphi, v[\mu_t] \rangle \mathrm{d}\mu_t.$$

Proof. By differentiation under the integral sign, we get

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\int_{\mathbb{R}^p} \varphi \mathrm{d}\mu_t \right) = \mathbf{E} \left[\frac{\mathrm{d}}{\mathrm{d}t} \varphi(Z_t) \right] = \mathbf{E} \left[\langle \nabla \varphi(Z_t), v[\mu_t](Z_t) \rangle \right] = \int_{\mathbb{R}^p} \langle \nabla \varphi, v[\mu_t] \rangle \mathrm{d}\mu_t. \qquad \square$$

See [Santambrogio, 2015, Prop. 4.2] for the link between weak solutions and distributional solutions of such continuity equations (these are essentially equivalent notions). This equation can be seen as the gradient flow of the objective expressed in terms of μ for the Wasserstein metric, see this blog post.

5 Lazy regime

In this section, we consider the lazy regime, namely $M^{-1} \ll \alpha_M \lesssim M^{-1/2}$, $\eta_M = 1/(M\alpha_M^2)$ and $\eta_0 = O(1)$. I recall that we assume that ϕ and all its derivatives are bounded, uniformly in x.

For the analysis of this regime, our starting point is (3). Consider the following "empirical" and "limit" tangent kernels:

$$\hat{K}_k(x,x') := \frac{1}{M}\sum_{j=1}^{M} D\phi(\hat{Z}_k^j, x) D\phi(\hat{Z}_k^j, x')^\top, \quad K_0(x,x') := \mathbf{E}[D\phi(Z_0^j, x) D\phi(Z_0^j, x')^\top].$$

Limit dynamics Consider GD in the RKHS with kernel K_0 :

$$\begin{aligned} h_0 = \hat{h}_{\theta_0}, \qquad h_{k+1}(x) &= h_k(x) - \frac{\eta_0}{n}\sum_{i=1}^{n} K_0(x,x_i)\nabla \mathrm{loss}(y_i, h_k(x_i)) \\ &= h_k(x) + G((h_k(x_i))_{i\in[1:n]}) \end{aligned}$$

where G is defined by the last line.

Tight error bound By matrix Hoeffding's inequality [Bach, 2024, Prop. 1.6], there exists $c > 0$ such that for all $x, x' \in \mathbb{R}^d$ fixed, with probability at least $1-\delta$ it holds

$$\left\| K_0(x,x') - \hat{K}_0(x,x') \right\|_{op} \le c\Big(\frac{1+\sqrt{\log(1/\delta)}}{\sqrt{M}}\Big)$$

Besides, for an horizon $K = O(1)$ and for $k \le K$, we deduce from (5) that

$$\max_{k\le K} \|\hat{Z}_k^j - \hat{Z}_0^j\| = O\Big(\frac{K\eta_0}{M\alpha_M}\Big) = O\Big(\frac{1}{M\alpha_M}\Big).$$

Therefore, we have for any $k \le K$, with probability at least $1-\delta$,

$$\begin{aligned} \left\| \hat{K}_k(x,x') - K_0(x,x') \right\|_{op} &\le \left\| \hat{K}_k(x,x') - \hat{K}_0(x,x') \right\|_{op} + \left\| \hat{K}_0(x,x') - K_0(x,x') \right\|_{op} \\ &= O\Big(\frac{1}{M\alpha_M} + \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{M}}\Big). \end{aligned}$$

Take an arbitrary test point $x \in \mathbb{R}^d$. By a union bound over $i \in [1:n]$ and over $k \in [0:K]$, with probability at least $1-\delta$ we have for all $k \le K$:

$$\hat{h}_k(x) = \hat{h}_k(x) - \frac{\eta_0}{n}\sum_{i=1}^{n} K_0(x,x_i)\nabla \mathrm{loss}(y_i, \hat{h}_k(x_i)) \tag{9}$$

$$+ O\Big((\eta_0\eta_M\alpha_M)^2\alpha_M + \frac{1}{M\alpha_M} + \frac{1+\sqrt{\log(nK/\delta)}}{\sqrt{M}}\Big). \tag{10}$$

$$= \hat{h}_k(x) + G((\hat{h}_k(x_i))_{i\in[1:n]}) + O\Big(\frac{1+\sqrt{\log(nK/\delta)}}{M\alpha_M}\Big). \tag{11}$$

Therefore $(h_k)_k$ and $(\hat{h}_k)_k$ start from the same point and obey the same iteration up to an error term. By a discrete Grönwall argument, we deduce:

Proposition 5.1. *There exists $c > 0$ that may depend on ϕ, K and η_0 such that with probability at least $1 - \delta$, it holds*

$$\max_{k \leq K} \max_{i \in [1:n]} |\hat{h}_k(x_i) - h_k(x_i)| \leq c\Big(\frac{1 + \sqrt{\log(n/\delta)}}{M\alpha_M}\Big)$$

In particular, for $\alpha_M = M^{-1/2}$, the convergence is a rate $\tilde{O}(M^{-1/2})$.

Exercice: go through the argument and make the bound explicit in K, η_0.

References

Francis Bach. *Learning theory from first principles.* MIT press, 2024.

Filippo Santambrogio. Optimal transport for applied mathematicians. 2015.