# Proving Linear Mode Connectivity of Neural Networks via Optimal Transport

Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, Aymeric Dieuleveut

# Proving Linear Mode Connectivity of Neural Networks via Optimal Transport

**Damien Ferbach**[1,3] **Baptiste Goujaud**[2] **Gauthier Gidel** [1,†] **Aymeric Dieuleveut**[2]
Mila, Université de Montréal[1] CMAP, Ecole Polytechnique, IPP[2] ENS Paris, PSL[3]

## Abstract

The energy landscape of high-dimensional non-convex optimization problems is crucial to understanding the effectiveness of modern deep neural network architectures. Recent works have experimentally shown that two different solutions found after two runs of a stochastic training are often connected by very simple continuous paths (e.g., linear) modulo a permutation of the weights. In this paper, we provide a framework theoretically explaining this empirical observation. Based on convergence rates in Wasserstein distance of empirical measures, we show that, with high probability, two wide enough two-layer neural networks trained with stochastic gradient descent are linearly connected. Additionally, we express upper and lower bounds on the width of each layer of two deep neural networks with independent neuron weights to be linearly connected. Finally, we empirically demonstrate the validity of our approach by showing how the dimension of the support of the weight distribution of neurons, which dictates Wasserstein convergence rates is correlated with linear mode connectivity.

## 1 INTRODUCTION AND RELATED WORK

Training deep neural networks on complex tasks is a high-dimensional, non-convex optimization problem. While stochastic gradient-based methods (i.e., SGD and its derivatives) have proven highly efficient in finding a local minimum with low test error, the loss landscape of deep neural networks (DNNs) still contains numerous open questions. In particular, Goodfellow et al. [2014] try to find ways to connect two local minima reached by two independent runs of the same stochastic algorithm with different initialization and data orders. This problem has applications in diverse domains such as model averaging [Izmailov et al., 2018, Rame et al., 2022, Wortsman et al., 2022], loss landscape study [Gotmare et al., 2018, Vlaar and Frankle, 2022, Lucas et al., 2021], adversarial robustness [Zhao et al., 2020] or generalization theory [Pittorino et al., 2022, Juneja et al., 2022, Lubana et al., 2023].

An answer to this question is the *mode connectivity phenomenon*. It suggests the existence of a continuous low-loss path connecting all the local minima found by a given optimization procedure. The mode connectivity phenomenon has extensively been studied in the literature [Goodfellow et al., 2014, Keskar et al., 2016, Sagun et al., 2017, Venturi et al., 2019, Neyshabur et al., 2020, Tatro et al., 2020, Yunis et al., 2022, Zhou et al., 2023b] and *non-linear connecting paths* have been evidenced for DNNs trained on MNIST and CIFAR10 by Freeman and Bruna [2016], Garipov et al. [2018], Draxler et al. [2018].

**(Linear) mode connectivity.** Formally, let $A := \hat{f}(., \theta_A)$ and $B := \hat{f}(., \theta_B)$ two neural networks sharing a common architecture $\hat{f}$. They are parametrized by $\theta_A$ and $\theta_B$ after training those networks on a data distribution $P$ with loss $\mathcal{L}$, i.e. by minimizing $\mathcal{E}(\theta) := \mathbb{E}_{(x,y)\sim P}[\mathcal{L}(\hat{f}(x, \theta), y)]$ over $\theta$. Let $p$ be a continuous path connecting $\theta_A$ and $\theta_B$, i.e. a continuous function defined on $[0, 1]$ with $p(0) = \theta_A$ and $p(1) = \theta_B$. Frankle et al. [2020] initially identified the problem of linear mode connectivity and defined the *error barrier height* [Frankle et al., 2020, Entezari et al., 2021] of $p$ as $\sup_{t\in[0,1]} \mathcal{E}(p(t)) - ((1-t)\mathcal{E}(\theta_A) + t\mathcal{E}(\theta_B))$. The two found solutions $\theta_A$ and $\theta_B$ are said to be *mode connected* if there is a continuous path with zero error barrier height connecting them. Furthermore if $p$ is linear, that is $p(t) = (1-t)\theta_A + t\theta_B$, $\theta_A$ and $\theta_B$ are

† Canada CIFAR AI Chair

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

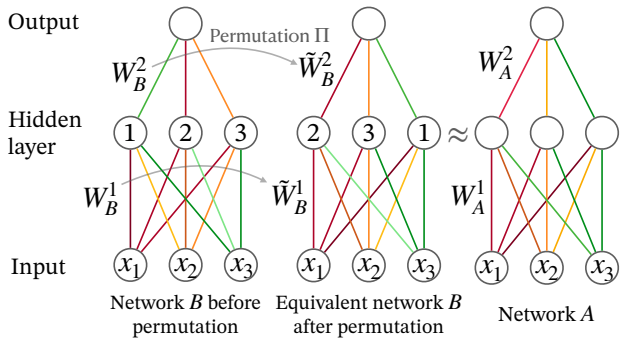



Figure 1: Permuting the neurons in the hidden layer of network $B$ to align them on network $A$

said to be *linearly mode connected (LMC)*.

**Permutation invariance.** Recently, Singh and Jaggi [2020], Ainsworth et al. [2022] highlighted the fact that the units in a hidden layer of a given model can be permuted while preserving the network's functionality. Figure 1 shows how one can permute the hidden layer of a two-layer network to match a different target network without changing the source function. From now on, **we will understand LMC modulo permutation invariance**, i.e. two networks $A, B$ are said to be linear mode connected whenever there exists a permutation of neurons in each hidden layer of network $B$ such that the linear path in parameter space between network $A$ and $B$ permuted has low loss.

**Linear mode connectivity up to permutation.** Singh and Jaggi [2020] proposed to use optimal transport (OT) theory to find soft alignment providing a "good match" (in a certain sense) between the neurons of two trained DNNs. Furthermore, the authors propose ways to fusion the aligned networks together in a federated learning context with local-steps. Ainsworth et al. [2022] further experimentally studied linear mode connectivity between two pre-aligned networks. The authors first align network B's weights on the weights of network A before connecting both of them by a linear path in the parameter space. They notably achieved zero-loss barrier for two trained Resnets with SGD on CIFAR10. Moreover, their experiments strongly suggest that the error barrier on a linear path gets smaller for wider networks, with a detrimental effect of big depth.

**Prior theoretical explanations.** A recent work by Kuditipudi et al. [2019] shows that dropout stable networks (i.e. networks that are functionally stable to the action of randomly setting a fraction of their weights and normalizing the others) exhibit mode connectivity. Shevchenko and Mondelli [2020] use a mean field viewpoint to show that wide two-layer neural networks trained with SGD are dropout stable and hence show (non-linear) mode connectivity for two-layer neural networks in the mean field regime (i.e. one single wide hidden layer). Finally Entezari et al. [2021] show that two-layer neural networks exhibit linear mode connectivity up to permutation at initialization for parameters initialized following uniform independent distribution properly scaled. They highlight that this result could be extended to networks trained in the Neural Tangent Kernel regime where parameters stay close to initialization [Jacot et al., 2018].

**Contributions.** This paper aims at building theoretical foundations on the phenomenon of linear mode connectivity up to permutation. More precisely, we theoretically prove this phenomenon arises naturally on multi-layer perceptrons (MLPs), which goes beyond two-layer networks on which theoretical works focused so far. We also provide a new efficient way to find the right permutation to apply on the units of a neural network's layer. The paper is organized as follow:

- In Section 3, we focus on two-layer neural networks in the mean field regime. While Shevchenko and Mondelli [2020] proved *non-linear* mode connectivity in this setting; we go further by proving *linear mode connectivity up to permutation*. Moreover, we provide an upper bound on the minimal width of the hidden layer to guarantee linear mode connectivity.

- In Section 4, we use general OT theory to exhibit tight asymptotics on the minimal width of a multilayer perceptron (MLP) to ensure LMC.

- In Section 5, we apply our general results to networks with parameters following sub-Gaussian distribution. Our result holds for deep networks, generalizing the result of Entezari et al. [2021] with better bounds. We shed light on the dependence in the dimension of the underlying distributions of the weights in each layer and explain how it connects with previous empirical observations [Ainsworth et al., 2022]. Using a model of approximately low dimensional weight distribution as a proxy of sparse feature learning, we yield more realistic bounds on the architectures of DNNs to ensure linear mode connectivity. We therefore, show why LMC is possible after training and how it depends on the complexity of the task. Finally we unify our framework with dropout stability.

- In Section 6, we validate our theoretical framework by showing how the implicit dimension of the weight distribution is correlated with linear mode connectivity for MLPs trained on MNIST with SGD and propose a new weight matching method.

## 2 PRELIMINARIES AND NOTATIONS

**Notations.** Let two multilayer perceptrons (MLP) $A$ and $B$ with the same depth $L+1$ ($L$ hidden layers), an input dimension $m_0$, intermediate widths $m_1, ..., m_L$ and an output dimension $m_{L+1}$. Given $2(L+1)$ weights matrices $W_{A,B}^{1,...,L+1}$, and a non-linearity $\sigma$, we define the neural network function of network A by $\hat{f}_A$ (respectively $\hat{f}_B$): $\forall x \in \mathbb{R}^{m_0}$,

$$\hat{f}_A(x) := \hat{f}(x; \theta_A) := W_A^{L+1} \sigma\left(W_A^L \dots \sigma(W_A^1 x)\right) \quad (1)$$

To $W_A^\ell \in \mathcal{M}_{m_\ell, m_{\ell-1}}(\mathbb{R})$ we associate $\hat{\mu}_{A,\ell}$ the empirical measure of its rows $[W_A^\ell]_{i:} \in \mathbb{R}^{m_{\ell-1}}$: $\frac{1}{m_\ell}\sum_{i=1}^{m_\ell} \delta_{[W_A^\ell]_{i:}}$ which belongs to the space of probability measures $\mathcal{P}_1(\mathbb{R}^{m_{\ell-1}})$, where $[W_A^\ell]_{i:}$ is the $i$-th row of the matrix and $\delta$ denotes the Dirac measure. Note that $[W_A^\ell]_{i:}$ is also the weights vector of the $i$-th neuron of the layer $\ell$ of network $A$. Given an equi-partition[1] $\mathcal{I}^{\ell-1} = \{I_1^{\ell-1}, ...I_{\tilde{m}_{\ell-1}}^{\ell-1}\}$ of $[m_{\ell-1}]$ we denote $W_A^{\mathcal{I}^{\ell-1}} \in \mathcal{M}_{m_\ell, \tilde{m}_{\ell-1}}(\mathbb{R})$ the matrix issued from $W_A^\ell$ where we have summed the columns being in the same set of the partition $\mathcal{I}^{\ell-1}$. In that case $\hat{\mu}_A^{\mathcal{I}^{\ell-1}} \in \mathcal{P}_1\left(\mathbb{R}^{\tilde{m}_\ell}\right)$ denotes the associate empirical measure of its rows.

Denote $\phi_A^\ell(x) := \sigma\left(W_A^\ell \dots \sigma(W_A^1 x)\right)$ (respectively $\phi_B^\ell$) the activations of neurons at layer $\ell$ of network $A$ on input $x$. The data $x$ follows a distribution $P$ in $\mathbb{R}^{m_0}$.

Given permutations matrices $\Pi_\ell \in \mathcal{S}_{m_\ell}$,[2] $\ell = 1, \dots, L$ of each hidden layer of network $B$, the weight matrix at layer $\ell$ of the permuted network $B$ is $\tilde{W}_B^\ell := \Pi_\ell W_B^\ell \Pi_{\ell-1}^T$ and its new activation vector is $\tilde{\phi}_B^\ell(x) := \Pi_\ell \phi_B^\ell(x)$. Finally, $\forall t \in [0,1]$ we define $M_t$ the convex combination of network $A$ and $B$ permuted, with weights matrices $tW_A^\ell + (1-t)\tilde{W}_B^\ell$ and $\phi_{M_t}^\ell$ its activations at layer $\ell$.

**Preliminaries.** We consider networks $A$ and $B$ to be independently chosen from the same distribution $Q$ on parameters. This is coherent with considering two networks initialized independently or trained independently with the same optimization procedure (§3). We additionally suppose the choice of $A$ and $B$ to be independent of the choice of $x \sim P$, which is valid when evaluating models on test data not seen during training. We denote $\mathbb{E}_Q, \mathbb{E}_P, \mathbb{E}_{P,Q}$ expectations with respect to the choice of the networks, the data, or both.

To show linear mode connectivity of networks $A$ and $B$ we will show the existence of permutations $\Pi_1, ..., \Pi_L$ of layers $1, ..., L$ that align the neurons of network $B$ on the closest neurons weights of network $A$ at the same layer as shown in Figure 1. In other words, we want to find permutations that minimize for each layer $\ell \in [L]$ the norm $\|W_A^\ell - \Pi_\ell W_B^\ell \Pi_{\ell-1}^T\|_2$. Recursively on $\ell$, we solve the following optimization problem:

$$\begin{aligned}\Pi_\ell &= \underset{\Pi \in \mathcal{S}_{m_\ell}}{\arg\min} \|W_A^\ell - \Pi W_B^\ell \Pi_{\ell-1}^T\|_2^2 \\ &= \underset{\pi \in \mathcal{S}_{m_\ell}}{\arg\min} \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \|[W_A^\ell]_{i:} - [W_B^\ell \Pi_{\ell-1}^T]_{\pi_i:}\|_2^2\end{aligned} \quad (2)$$

For each layer, the problem can be cast as finding a pairing of weights neurons $[W_A^\ell]_{i:}$ and $[W_B^\ell \Pi_{\ell-1}^T]_{\pi_i:}$ to minimize the sum of their Euclidean distances. It is known as the Monge problem is the optimal transport literature Peyré et al. [2019]. More precisely Equation (2) can be formulated as finding an optimal transport plan corresponding to the Wasserstein distance between the empirical measures of the rows of $W_A^\ell$ and $W_B^\ell \Pi_{\ell-1}^T$. We provide more details about this connection between Equation (2) and optimal transport in Appendix B.2. In the following, the $p-$Wasserstein distance will be denoted $\mathcal{W}_p(\cdot,\cdot)$ and defined with the underlying distance $\|\cdot\|_2$ unless expressed otherwise.

By controlling the cost in Equation (2) at every layer, we show that the permuted s of networks $A$ and $B$ are approximately equal. Linearly interpolating both networks will therefore keep activations of all hidden layers unchanged except the last layer which acts as a linear function of the interpolation parameter $t \in [0,1]$.

## 3 LMC FOR TWO-LAYER NNs IN THE MEAN FIELD REGIME

We will first study linear mode connectivity between a pair of two-layer neural networks independently trained with SGD for the same number of steps.

### 3.1 Background on the Mean Field Regime

We will use some notations from Mei et al. [2019] and consider a two-layer neural network,

$$\hat{f}_N(x; \theta) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(x; \theta_i) \quad (3)$$

parametrized by $\theta_i = (a_i, w_i) \in \mathbb{R} \times \mathbb{R}^d$ and where $\sigma_*(x; \theta_i) = a_i \sigma(w_i x)$. The parameters evolve as to minimize the following regularized cost $R_N(\theta) = \mathbb{E}_{(x,y)\sim P}[(y - \hat{f}_N(x;\theta))^2] + \lambda\|\theta\|_2^2$. Define noisy regularized stochastic gradient descent (or noiseless regularization-free when $\lambda = 0, \tau = 0$) with

[1] All subsets have the same number of elements

[2] We use interchangeably $\mathcal{S}_m$ to denote the space of permutations of $\{1, \dots, m\}$ and the corresponding space of permutations matrices. Given $\pi \in \mathcal{S}_m$ its corresponding permutation matrix $\Pi$ is defined as $\Pi_{ij} = 1 \iff \pi(i) = j$.

step size $s_k$, and i.i.d. Gaussian noise $g^k \sim \mathcal{N}(0, I_d)$:

$$\theta_i^{k+1} = (1 - 2\lambda s_k)\theta_i^k \qquad \text{(SGD)}$$
$$+ 2s_k(y_k - \hat{f}_N(x_k; \theta^k))\nabla_\theta \sigma_*(x_k; \theta_i^k) + \sqrt{\tfrac{2s_k\tau}{d}} g_i^k$$

It will be useful to consider $\rho_N^k := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i^k}$ the empirical distribution of the weights after $k$ SGD steps. Indeed some recent works [Chizat and Bach, 2018, Mei et al., 2018, 2019] have shown that when setting the width $N$ to be large and the step size $s_k$ to be small, the empirical distribution of weights during training remains close to an empirical measure drawn from the solution of a partial differential equation (PDE) we explicit in Appendix C.1. Especially, the parameters $\{\theta_i^k, i \in [N]\}$ evolve approximately independently.

### 3.2 Proving LMC in the mean field setting

Define respectively the alignment of a neuron function on the data and the correlation between two neurons:

$$V(\theta_1) := av(w) := -\mathbb{E}_P[y\sigma_*(x; \theta_1)]$$
$$U(\theta_1, \theta_2) := a_1 a_2 u(w_1, w_2) := \mathbb{E}_P[\sigma_*(x; \theta_1)\sigma_*(x; \theta_2)]\,.$$

and we for $\varepsilon > 0$ fixed we note the step size

$$s_k = \varepsilon \xi(k\varepsilon) \tag{4}$$

where $\xi$ is a positive scaling function. The underlying training time up to step $k_T$ is defined as $T := \sum_{k=1}^{k_T} s_k$. We now state the standard assumptions to work on the mean field regime [Mei et al., 2019],

**Assumption 1.** *The function $t \mapsto \xi(t)$ is bounded Lipschitz. The non-linearity $\sigma$ is bounded Lipschitz and the data distribution has a bounded support. The functions $w \mapsto v(w)$ and $(w_1, w_2) \mapsto u(w_1, w_2)$ are differentiable, with bounded and Lipschitz continuous gradient. The weights at initialization $\theta_i^0$ are i.i.d. with distribution $\rho_0$ which has bounded support.*

Assumption 1 imposes that the step size is of order $\mathcal{O}(\varepsilon)$ and its variations are of order $\mathcal{O}(\varepsilon^2)$. Constant step size $\varepsilon$ will work. Bounded non-linearity include arctan and sigmoid but excludes ReLU. While it is a standard assumption in mean field theory ([Mei et al., 2018, 2019]), we mention in §C that this assumption can be relaxed by the weaker assumption that the non-linearity stays small on some big enough compact set.

The second assumption is technical and only used for studying noisy regularized SGD.

**Assumption 2.** *$V, U$ are four times continuously differentiable and $\nabla_1^k u(\theta_1, \theta_2)$ is uniformly bounded for $0 \leq k \leq 4$.*

The following theorem states that two wide enough two-layer neural networks trained independently with SGD exhibit, with high probability, a linear connection of the prediction modulo permutations for all data.

**Theorem 3.1.** *Consider two two-layer neural networks as in Equation (3) trained with equation SGD with the same initialization over the weights independently and for the same underlying time $T$. Suppose Assumptions 1 and 2 to hold. Then $\forall \delta, err, \exists N_{min}$ such that if $N \geq N_{min}, \exists \varepsilon_{max}(N)$ such that if $\varepsilon \leq \varepsilon_{max}(N)$ in Equation (4), then with probability at least $1 - \delta$ over the training process, there exists a permutation of the second network's hidden layer such that for almost every $x \sim P$:*

$$|t\hat{f}_N(x; \theta_A) + (1-t)\hat{f}_N(x; \theta_B)$$
$$- \hat{f}_N(x; t\theta_A + (1-t)\tilde{\theta}_B)| \leq err\,, \quad \forall t \in [0, 1]\,.$$

**Remark.** Assumption 2 is not used when studying noiseless regularization-free SGD ($\lambda = 0$, $\tau = 0$).

**Corollary 3.2.** *Under assumptions of Theorem 3.1, $\forall \delta, err > 0, \exists N'_{\min}, \forall N \geq N'_{\min}, \exists \varepsilon'_{\max}(N), \forall \varepsilon \leq \varepsilon'_{\max}(N)$ in Equation (4), then with probability at least $1 - \delta$ over the training process, there exists a permutation of the second network's hidden layer such that $\forall t \in [0, 1]$:*

$$\mathbb{E}_P\left[\left(\hat{f}_N(x; t\theta_A + (1-t)\tilde{\theta}_B) - y\right)^2\right] \leq err$$
$$+ \mathbb{E}_P\left[t(\hat{f}_N(x; \theta_A) - y)^2 + (1-t)(\hat{f}_N(x; \theta_B) - y)^2\right]$$

**Discussion.** Two wide enough two-layer neural networks wide enough trained with SGD are therefore Linear Mode Connected with an upper bound on the error tolerance we explicit in Appendix C. We have extensively used the independence between weights in the mean field regime to apply OT bounds on convergence rates of empirical measures. To go beyond the two-layer case, we will need to make such an assumption on the distribution of weights. Note that this is true at initialization and after training for two-layer networks. Studying the independence of weights in the multi-layer case is a natural avenue for future work, already studied in **?**.

## 4 GENERAL STRATEGY FOR MULTI-LAYER NETWORKS

We now build the foundations to study the case of multi-layer neural networks (see Equation (1)).

We first write one formal property expressing the existence of permutations of neurons of network $B$ up to layer $\ell$ such that the activations of network $A$, network $B$ permuted and the mean network $M_t$ are close up to layer $\ell$. This property is trivially satisfied at the input layer. We then show that under two formal assumptions on the weights matrices of networks $A$ and $B$, this property still hold at layer $\ell+1$.

### 4.1 Formal Property at layer $\ell$

Let $\varepsilon > 0, m_\ell \geq \tilde{m}_l$ and $m_{\ell+1} \geq \tilde{m}_{\ell+1}$. Assume $\frac{m_\ell}{\tilde{m}_\ell}, \frac{m_{\ell+1}}{\tilde{m}_{\ell+1}} \in \mathbb{N}$ to simplify technical details but this hypothesis can easily be removed.

**Property 1.** *There exists two constants $\underline{E}_\ell, E_\ell$ such that given weight matrices up to layer $\ell$, $W_{A,B}^{1,\cdots,\ell}, W_B^{1,\cdots,\ell}$ one can find $\ell$ permutations $\Pi_1, \cdots, \Pi_\ell$ of the neurons in the hidden layers 1 to $\ell$ of network $B$, an equi-partition $\mathcal{I}^\ell = \{I_1^\ell, \ldots, I_{\tilde{m}_\ell}^\ell\}$, and a map $\underline{\phi}^\ell(x) \in \mathbb{R}^n$ such that $\forall k \in [\tilde{m}_\ell], \forall i, j \in I_k^\ell, \underline{\phi}_i^\ell(x) = \underline{\phi}_j^\ell(x)$ such that:*

$$\begin{aligned}
&\mathbb{E}_{P,Q}\|\underline{\phi}^\ell(x)\|_2^2 \leq \underline{E}_\ell m_\ell \\
&\mathbb{E}_{P,Q}\|\phi_A^\ell(x) - \underline{\phi}^\ell(x)\|_2^2 \leq E_\ell m_\ell \\
&\mathbb{E}_{P,Q}\|\tilde{\phi}_B^\ell(x) - \underline{\phi}^\ell(x)\|_2^2 \leq E_\ell m_\ell \\
&\mathbb{E}_{P,Q}\|\phi_{M_t}^\ell(x) - \underline{\phi}^\ell(x)\|_2^2 \leq E_\ell m_\ell\,, \quad \forall t \in [0,1],
\end{aligned}$$

This property not only requires proximity between activations $\phi_A^\ell(x), \tilde{\phi}_B^\ell(x)$ at layer $\ell$ but requires the existence of a vector $\underline{\phi}^\ell(x)$ whose coefficients in the same groups of the partition $\mathcal{I}^\ell$ are equal, and therefore lives in a $\tilde{m}_\ell$. It bounds the size of the function space available at layer $\ell$ and hence allows to use an effective width $\tilde{m}_\ell$ independent of the real width $m_\ell$, which can be much larger. It is crucial in order to show LMC for neural networks of constant width across layers. The introduction of such a map $\underline{\phi}^\ell(x)$ is non trivial and is an important contribution since it allows to extend results of Entezari et al. [2021] beyond two layers.

### 4.2 Assumptions on the weight distribution

We now make an assumption on the empirical distribution of the weights $\hat{\mu}_{A,\ell+1}$ at layer $\ell+1$ of $W_A^{\ell+1}$.

**Assumption 3.** *There exists an integer $\tilde{m}_{\ell+1}$ such that for all equi-partiton $\mathcal{I}^\ell$ of $[m_\ell]$ with $\tilde{m}_\ell$ sub-sets, there exists a random empirical measure $\hat{\mu}_{\tilde{m}_{\ell+1}}$ independent of $A$ and $B$ composed of $\tilde{m}_{\ell+1}$ vectors in $\mathbb{R}^{m_\ell}$, such that $\mathbb{E}_Q[\mathcal{W}_2^2(\hat{\mu}_{A,\ell+1}^{\mathcal{I}^\ell}, \hat{\mu}_{\tilde{m}_{\ell+1}}^{\mathcal{I}^\ell})] \leq C_1$.*

This assumption requires that the empirical distribution with $m_{\ell+1}$ points of the neurons' weights of network $A$ at layer $\ell+1$ can be approximated by an empirical measure with a smaller $\tilde{m}_{\ell+1}$ number of points. Note that it implies proximity in Wasserstein distance between $\hat{\mu}_A^{\mathcal{I}^\ell}$ and $\hat{\mu}_B^{\mathcal{I}^\ell}$ by a triangular inequality.

We finally assume some central limit behavior when summing the errors made for each neuron of layer $\ell$.

**Assumption 4.** *There exists a constant $C_2$ such that $\forall X \in \mathbb{R}^{m_l}$ we have:*

$$\begin{aligned}
\max\{\mathbb{E}_Q[\|W_A^{\ell+1}X\|_2^2], \mathbb{E}_Q[\|W_{\tilde{m}_{l+1}}X\|_2^2]\} \\
\leq C_2 \frac{m_{\ell+1}}{m_\ell}\|X\|_2^2
\end{aligned}$$

Finally, we consider the following assumption on the non-linearity, verified for example by pointwise ReLU.

**Assumption 5.** *$\sigma$ is pointwise, 1-Lipschitz, $\sigma(0) = 0$.*

### 4.3 Propagating Property 1 to layer $\ell+1$

We state now how Property 1 propagates throughout the layers using Assumptions 3 to 5 with new parameters $\underline{E}_{\ell+1}, E_{\ell+1}$. We give a proof in Appendix B.6.

**Lemma 4.1.** *Let $\ell \in \{0, \cdots, L-1\}$ and suppose Property 1 to hold at layer $\ell$ and Assumptions 3 to 5 to hold, then Property 1 still holds at the next layer with $\tilde{m}_{\ell+1}$ given in Assumption 3 and*

$$\begin{aligned}
\underline{E}_{\ell+1} &= C_2 \underline{E}_\ell \\
E_{\ell+1} &= 2C_2 E_\ell + 2C_1 \tilde{m}_\ell \underline{E}_\ell
\end{aligned}$$

## 5 LMC FOR RANDOM MULTI-LAYER NNs

We will make the following assumption on the empirical distribution of neurons weights $\hat{\mu}_{A,\ell}, \hat{\mu}_{B,\ell}$ of $W_A^\ell, W_B^\ell$ at layer $\ell$.

**Assumption 6** (Independence of neurons weights)**.** *$\hat{\mu}_{A,\ell}, \hat{\mu}_{B,\ell}$ correspond to two i.i.d drawings of vectors with distribution $\mu_\ell$ i.e., $\hat{\mu}_{A,\ell}, \hat{\mu}_{B,\ell}$ have the law of $\frac{1}{m_\ell}\sum_{i=1}^{m_\ell}\delta_{x_i}$ where $x_i \sim \mu_\ell$ i.i.d.*

Assumption 6 is verified for example at initialization but more generally when weights do not depend too much one of each other. This case still holds for wide two-layer neural networks trained with SGD and is at the heart of the proof of Theorem 3.1.

### 5.1 Showing LMC for multilayer MLPs under Gaussian distribution

We first examine the case $\mu_\ell = \mathcal{N}\left(0, \frac{I_{m_{\ell-1}}}{m_{\ell-1}}\right)$. We moreover assume that the input data distribution has bounded second moment: $\mathbb{E}_P[\|x\|_2^2] \leq m_0$.

Our strategy detailed in Appendix B.7 consists in showing that wide enough such networks will satisfy Assumptions 3 and 4 with well controlled constants $C_1, C_2$. We can then apply Lemma 4.1 successively $L$ times to get the following lemma:

**Lemma 5.1.** *Under normal initialization of the weights, given $\varepsilon > 0$, if $m_0 \geq 5$, there exists minimal widths $\tilde{m}_1, \dots, \tilde{m}_L$ such that if $m_1 \geq \tilde{m}_1, \dots, m_L \geq \tilde{m}_L$, Property 1 is verified at the last hidden layer $L$ for $\underline{E}_L = 1, E_L = \varepsilon^2$. Moreover, $\forall \ell \in [L], \exists T_\ell$ which does only depend on $L, \ell$ such that one can define recursively $\tilde{m}_\ell$ as $\tilde{m}_0 = m_0$ and*

$$\tilde{m}_\ell = \tilde{\mathcal{O}}\left(\frac{T_\ell}{\varepsilon}\right)^{\tilde{m}_{\ell-1}}$$

**Discussion.** The hypothesis $m_0 \geq 5$ is technical and could be relaxed at the price of slightly changing the bound on $\tilde{m}_1$. Lemma 5.1 shows that given two random networks whose widths $m_\ell$ is larger than $\tilde{m}_\ell$, we can permute neurons of the second one such that their activations at layer $\ell$ are both close to the one of the networks on a linear path in parameter's space.

As $\varepsilon$ goes to 0, the width of the layer $\ell + 1$ must scale at least as $\left(\frac{1}{\varepsilon}\right)^{\tilde{m}_{\ell-1}}$. This is a fundamental bound due to the convergence rate in Wasserstein distance of empirical measures. It imposes a recursive exponential growth in the width needed with respect to depth. This condition appears excessive as compared to the typical width of neural networks used in practice. We highlight here that Ainsworth et al. [2022] empirically demonstrates that networks at initialization do not exhibit LMC and that the loss barrier is erased only after a sufficient number of SGD steps.

### 5.2 Showing Linear Mode Connectivity

We make the following assumption on the loss function to show LMC from Lemma 5.1.

**Assumption 7.** *$\forall y \in \mathbb{R}^{m_{L+1}}$, the loss $\mathcal{L}(\cdot, y)$ is convex and 1-Lipschitz.*

We finally prove the following bound on the loss of the mean network $M_t$ in Appendix B.8:

**Theorem 5.2.** *Under normal initialization of the weights, for $m_1 \geq \tilde{m}_1, \cdots, m_L \geq \tilde{m}_L$ as defined in Lemma 5.1, $m_0 \geq 5$, and under Assumption 7 we know that $\forall t \in [0,1]$, with $Q$-probability at least $1 - \delta_Q$, there exists permutations of hidden layers $1, \dots, L$ of network $B$ that are independent of $t$, such that:*

$$\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_{M_t}(x), y\right)\right] \leq t\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_A(x), y\right)\right] + (1-t)\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_B(x), y\right)\right] + \frac{4\sqrt{m_{L+1}}}{\delta_Q^2}\varepsilon$$

**Discussion.** The minimal width at layer $\ell$ needed for Theorem 5.2 is recursively $\tilde{m}_l \sim \varepsilon^{-\tilde{m}_{l-1}}$. Applied to randomly initialized two-layer networks, we need a hidden layer's dimension of $\varepsilon^{-m_0}$ as opposed to Entezari et al. [2021] which prove a bound of $\varepsilon^{-(2m_0+4)}$.

### 5.3 Tightness of the bound dependency with respect to the error tolerance

We discuss here the tightness of the minimal width $\tilde{m}_\ell$ we require in Lemma 5.1 with respect to the error tolerance $\varepsilon$. The recursive exponential growth of the width in the form $\tilde{m}_\ell \sim \left(\frac{1}{\varepsilon}\right)^{\tilde{m}_{\ell-1}}$ is a consequence of the convergence rate of Wasserstein distance of empirical measures in dimension $\tilde{m}_{\ell-1}$ at the rate $1/\tilde{m}_{\ell-1}$. Theorem 5.3 provides a corresponding lower bound which shows that this recursive exponential growth is tight at the precise rate $\left(\frac{1}{\varepsilon}\right)^{\tilde{m}_{\ell-1}}$ (just take $n = \tilde{m}_{\ell-1}$, $m = \tilde{m}_\ell$, $\mu = \mu_\ell$, $x = \phi_A^{\ell-1}(x), W_{A,B} = W_{A,B}^\ell$). A proof is given in Appendix B.11.

**Theorem 5.3.** *Let $n \geq 1, x \sim P \in \mathcal{P}_1(\mathbb{R}^n)$ and $\mu \in \mathcal{P}(\mathbb{R}^n)$ such that $\frac{d\mu}{dLeb} \leq F_1$. Suppose $\Sigma = \mathbb{E}[xx^T]$ is full rank $n$. Let $m \geq 1$ and $W_A, W_B \in \mathcal{M}_{m,n}(\mathbb{R})$ whose rows are drawn i.i.d. from $\mu$. Then, there exists $F_0$ such that*

$$\mathbb{E}_{W_A, W_B}\left[\min_{\Pi \in \mathcal{S}_m} \mathbb{E}_P \|(W_A - \Pi W_B)x\|_2^2\right] \geq F_0\left(\frac{1}{m}\right)^{2/n}$$

**Remark 1.** Using an effective width $\tilde{m}_{\ell-1}$ smaller and independent of the real width $m_{\ell-1}$ allows to show LMC for networks of constant hidden width $m_1 = m_2 = \dots = m_L$ as soon as they verify $m_1 \geq \tilde{m}_1, \dots, m_L \geq \tilde{m}_L$ where $\tilde{m}_1, \dots, \tilde{m}_L$ are defined in Lemma 5.1. Without this trick, we need a recursive exponential growth of the real width $m_\ell \sim \left(\frac{1}{\varepsilon}\right)^{m_{\ell-1}}$.

**Remark 2.** Motivated by the fact that feature learning may concentrate the weight distribution on low dimensional sub-space, we could extend our proofs to the case where the underlying weight distribution has a support with smaller dimension to get recursive bounds no longer at rate $\tilde{m}_{\ell-1}$ but at a smaller one. Note this is unlikely to happen as we expect the matrix of weight vectors of a given layer to be full rank. Therefore, we study in the next section the case when

this matrix is approximately low rank, or equivalently when the weight distribution is concentrated around a low dimensional approximated support.

### 5.4 Approximately low dimensional supported measures

For the sake of clarity, assume from now on that the layer $\ell-1$ of network $A$ has been permuted such that for $\mathcal{I}^{\ell-1} = \{I_1^{\ell-1}, \dots, I_{\tilde{m}_{\ell-1}}^{\ell-1}\}$ (given in Property 1) we have $I_1^{\ell-1} = \{1, \dots, p_{\ell-1}\}$, $\dots$, $I_{\tilde{m}_{\ell-1}}^{\ell-1} = \{m_{\ell-1} - p_{\ell-1} + 1, \dots, m_{\ell-1}\}$ with $p_{\ell-1} = {}^{m_{\ell-1}}/_{\tilde{m}_{\ell-1}}\}$. This assumption is mild since we can always consider a permuted version of network $A$ without changing the problem.

Motivated by the discussion in Appendix B.9.1 we consider the model where the weights at layer $\ell$ are initialized i.i.d. multivariate Gaussian $\mu_\ell = \mathcal{N}(0, \Sigma^{\ell-1})$ with

$$\Sigma^{\ell-1} := \mathrm{Diag}\left(\lambda_1^\ell I_{p_{\ell-1}}, \lambda_2^\ell I_{p_{\ell-1}}, \dots, \lambda_{\tilde{m}_{\ell-1}}^\ell I_{p_{\ell-1}}\right)$$

with $\frac{1}{m_{\ell-1}}\frac{\tilde{m}_{\ell-1}}{k_{\ell-1}} \geq \lambda_1^\ell \geq \lambda_2^\ell \geq \dots \geq \lambda_{\tilde{m}_{\ell-1}}^\ell$ with $k_{\ell-1} \leq \tilde{m}_{\ell-1}$ an approximate dimension of the support of the underlying weights distribution. Note that to balance the low dimensionality of the weights distribution, we have replaced the upper-bound on the eigenvalues $\frac{1}{\tilde{m}_{\ell-1}}$ by the greater value $\frac{1}{m_{\ell-1}}\frac{\tilde{m}_{\ell-1}}{k_{\ell-1}}$ to avoid vanishing activations when $\ell$ grows which would have made our result vacuous.

The following assumption states that the weights distribution $\mu_\ell^{\mathcal{I}^{\ell-1}}$ at layer $\ell$ considered in $\mathcal{P}_1(\mathbb{R}^{\tilde{m}_{\ell-1}})$ (with the operation explicited in Section 2) is approximately of dimension $k_{\ell-1} = e\tilde{m}_{\ell-1}$. The approximation becomes more correct as $\eta \to 0$.

**Assumption 8** (Approximately low-dimensionality)**.**

$$\exists \eta, e \in (0,1), \forall \ell \in [L],\ \frac{\sqrt{\sum_{j=k_{\ell-1}}^{\tilde{m}_{\ell-1}} \lambda_j^\ell}}{4\sqrt{\sum_{j=1}^{k_{\ell-1}} \lambda_j^\ell}} \leq \eta,\ \frac{k_{\ell-1}}{\tilde{m}_{\ell-1}} = e$$

**Theorem 5.4.** *Under Assumptions 7 and 8, given $\varepsilon > 0$, if $em_0 \geq 5$ there exists minimal widths $\tilde{m}_1, \dots, \tilde{m}_L$ such that if $\eta^{-k_0} \geq m_1 \geq \tilde{m}_1, \dots, \eta^{-k_{L-1}} \geq m_L \geq \tilde{m}_L$, Property 1 is verified at the last hidden layer $L$ for $\underline{E}_L = 1, E_L = \varepsilon^2$. Moreover, $\forall \ell \in [L]$, $\exists T'_\ell$ which does only depend on $L, e, \ell$, such that one can define recursively $\tilde{m}_\ell$ as*

$$\tilde{m}_\ell = \tilde{\mathcal{O}}\left(\frac{T'_\ell}{\varepsilon}\right)^{k_{\ell-1}} = \tilde{\mathcal{O}}\left(\frac{T'_\ell}{\varepsilon}\right)^{e\tilde{m}_{\ell-1}}$$

*where $\tilde{m}_0 = m_0$. Moreover there exists permutations of hidden layers $1, \dots, L$ of network $B$ s.t.*

$$\forall t \in [0,1], \textit{ with } Q\textit{-probability at least } 1 - \delta_Q\textit{:}$$

$$\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_{M_t}(x), y\right)\right] \leq t\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_A(x), y\right)\right] + (1-t)\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_B(x), y\right)\right] + \frac{4\sqrt{m_{L+1}}}{\sqrt{e}\delta_Q^2}\varepsilon$$

**Discussion.** We give a proof in Appendix B.10. For $\eta$ small enough, the distribution of weights is approximately lower dimensional. It yields faster convergence rates until $m$ becomes exponentially big in $\eta$. This prevents the previous recursive exponential growth of width with respect to depth, though asymptotically, we recover the same rates as in Theorem 5.2. The smaller $e$, the lower dimensional are the distributions, and the less the width needs to grow when $\varepsilon \to 0$. The problem in that model is that the constant $T'_i$ explodes if $e \to 0$, which prevents using a model with fixed $k_\ell$ across the layers for the weight distribution. We want to highlight here that the proof can be extended to such a case, but we need to assume that the constant $C_2$ is bounded and not depending on $e$ across the layers in Lemma 4.1 (recall that with our proof, we had $C_2 = \frac{1}{e}$). This assumption seems coherent because the average activations don't explode across layers in the model. Assuming this, the bound we obtain for $\tilde{m}_\ell$ in Theorem 5.4 is completely independent on $\tilde{m}_{\ell-1}$, and there is no recursive exponential growth in the width needed across the layers. We give a more explicit discussion in Appendix B.12.

### 5.5 LMC for sub-Gaussian distributions

Still under the setting of Assumption 6 assume that the underlying distribution $\mu_\ell$ verifies for each layer $\ell \in [L+1]$: if $X \sim \mu_\ell$ then, $\forall j \neq k \in [m_{l-1}], X_j \amalg X_k$. Moreover $\forall i \in [\tilde{m}_{\ell-1}], \forall j, k \in I_i^{\ell-1}$,

$$\mathbb{E}[X_j^2] = \mathbb{E}[X_k^2] = \lambda_i^\ell$$

Finally suppose the variables are sub-Gaussian i.e., $\exists K > 0, \forall i \in [\tilde{m}_{\ell-1}], \forall j \in I_i^{\ell-1}$, $\forall c > 0$,

$$\mathbb{P}(|X_j| \geq c) \leq 2\exp(-\frac{c^2}{K\lambda_i^\ell})$$

We explain in Appendix B.13 why both Theorem 5.2 (in the case $\lambda_1^\ell = \dots = \lambda_{\tilde{m}_{l-1}}^\ell = {}^1/_{m_{\ell-1}}$) and Theorem 5.4 hold with mild modifications in the constants.

It extends our previous result considerably to LMC for any large enough networks whose weights are i.i.d. and whose underlying distribution has a sub-Gaussian tail (for example uniform distribution).

### 5.6 Link with dropout stability

In Appendix B.14, we build a first step towards unifying our study with the dropout stability viewpoint [Kuditipudi et al., 2019, Shevchenko and Mondelli, 2020] by showing in a simplified setting how networks become dropout stable in the same asymptotics on the width as the one needed in our Theorem 5.2.

## 6 EXPERIMENTS

Our previous study shows the influence of the dimension of the underlying weight distribution on LMC effectiveness. Based on this insight we develop a new weight matching method at the crossroads between previous naive weight matching (WM) and activation matching (AM) methods [Ainsworth et al., 2022]. Given $n$ training points $x_i$, $i \in [n]$, denote $Z_A^\ell \in \mathcal{M}_{m_\ell,n}(\mathbb{R})$ (respectively $Z_B^\ell$) the activations $\phi_A^\ell(x_i)$ for the $n$ data points $x_i$. Further denote $\Sigma_A^\ell := \frac{1}{n} Z_A^\ell [Z_A^\ell]^T \approx \mathbb{E}_P\left[\phi_A^\ell(x)[\phi_A^\ell(x)]^T\right]$. We aim at finding for each layer $\ell$ the optimal permutation $\Pi$ minimizing the cost (respectively for naive WM, our new WM method and AM):

$$\min_{\Pi \in S_{m_\ell}} \left\| W_A^\ell - \Pi W_B^\ell \Pi_{\ell-1}^T \right\|_2^2 , \qquad \text{(Naive WM)}$$

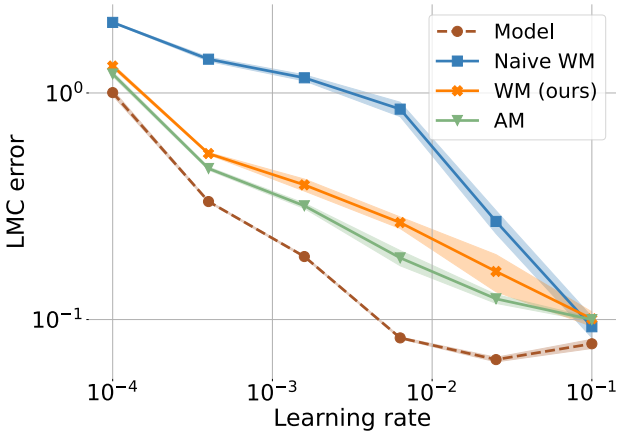
$$\min_{\Pi \in S_{m_\ell}} \left\| W_A^\ell - \Pi W_B^\ell \Pi_{\ell-1}^T \right\|_{2,\Sigma_A^{\ell-1}}^2 , \qquad \text{(WM (ours))}$$

$$\min_{\Pi \in S_{m_\ell}} \left\| Z_A^\ell - \Pi Z_B^\ell \right\|_2^2 , \qquad \text{(AM)}$$

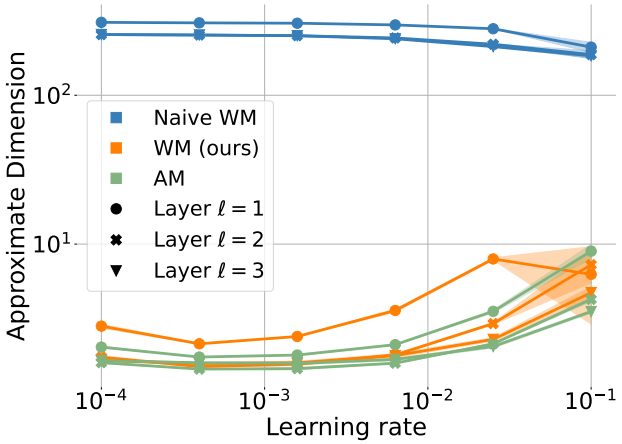
where $\| \cdot \|_{2,\Sigma_A^{\ell-1}}$ is the norm[3] induced by the scalar product $(X, Y) \mapsto \mathrm{tr}(X \Sigma_A^{\ell-1} Y^T)$. We both theoretically support the gain of our method in Theorem D.2 and empirically verify that this method constantly and substantially outperforms naive Weight Matching across different learning rates when training with SGD.

We train a three hidden layer MLP of width 512 on MNIST with learning rates varying between $10^{-4}$ and $10^{-1}$ across 4 runs. We plot on Figure 2b the approximate dimension of the considered covariance matrix for each matching method: $W_A^\ell [W_A^\ell]^T$ for WM (naive), $W_A^\ell \Sigma_A^{\ell-1} [W_A^\ell]^T$ for WM (ours) and $\Sigma_A^\ell$ for AM (see §D.2). Our code is available at `https://github.com/damienferbach/OT_LMC/tree/main`.

[3]Semi-norm in full generality (if $\Sigma_A^{\ell-1}$ is not full rank)




(a) Mean test loss of the trained networks $A$ and $B$ and error barrier on the linear path $M_t$, $t \in [0, 1]$ across different learning rate values for each matching problem.




(b) Approximate dimension $\mathrm{Dim}(S) := {}^{\mathrm{tr}(S)^2}/_{\mathrm{tr}(S^2)}$ of the matrices considered in the matching problems at each layer.

Figure 2: Statistics of the average network $M$ over the linear path between networks $A$ and $B$ using respectively weight matching (blue), weight matching using covariance of activations and activations (green), and activation matching (orange)

We see on Figure 2 the detrimental effect of high approximate dimension on LMC effectiveness, therefore validating our theoretical approach. Note that for a learning rate of $10^{-1}$ the correlation is less clear but a trend is visible on decreasing dimension for naive WM as it performs better (and increasing dimension for AM and our WM method as it performs comparatively less well). An alternative would be to use a proxy taking the diameter of the distributions into account (and not only the dimension of their support). Finally, experiments on Adam lead to less clear results that we did not report as more experimental investigation is needed. In particular, understanding the impact of the optimizer on the independence of weights during training is crucial, as it is a central assumption

in our study.

## 7 DISCUSSION

Optimal transport serves as a good framework to study linear mode connectivity of neural networks. This paper uses convergence rates of empirical measures in Wasserstein distance to upper bound the test error of the linear combination of two networks in weight space modulo permutation symmetries. Our main assumption is the independence of all neuron's weight vectors inside a given layer. This assumption is trivially true at initialization but remains valid for wide two-layers networks trained with SGD. We experimentally demonstrate the correlation between the dimension of the underlying weight distribution with LMC effectiveness and design a new weight matching method that significantly outperforms existing ones. A natural direction for future work is to focus on the behaviour of the weights distribution inside each layer of DNNs and their independence. Moreover, extending our results to only assuming approximate independence of weights is a natural direction as it seems a more realistic setting.

### Acknowledgements

The work of B. Goujaud and A. Dieuleveut is partially supported by ANR-19-CHIA-0002-01/chaire SCAI, and Hi!Paris. This work was partly funded by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). D. Ferbach acknowledges a stipend from ENS Paris as fonctionnaire stagiaire. This work was partly done during an internship of D. Ferbach at Ecole Polytechnique.

## References

L. Adilova, A. Fischer, and M. Jaggi. Layer-wise linear mode connectivity. *arXiv preprint arXiv:2307.06966*, 2023. 12

S. K. Ainsworth, J. Hayase, and S. Srinivasa. Git rebasin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022. 2, 6, 8, 38

A. K. Akash, S. Li, and N. G. Trillos. Wasserstein barycenter-based model fusion and linear mode connectivity of neural networks. *arXiv preprint arXiv:2210.06671*, 2022. 12

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018. 4, 29

F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. Essentially no barriers in neural network energy landscape. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1309–1318. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/draxler18a.html. 1

R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021. 1, 2, 5, 6

D. Ferbach, C. Tsirigotis, G. Gidel, and A. Bose. A general framework for proving the equivariant strong lottery ticket hypothesis. *arXiv preprint arXiv:2206.04270*, 2022. 12

J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 12

J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020. 1, 12

C. D. Freeman and J. Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016. 1

T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018. 1

I. J. Goodfellow, O. Vinyals, and A. M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014. 1

A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher. Using mode connectivity for loss landscape analysis. *arXiv preprint arXiv:1806.06977*, 2018. 1

P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 1

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 2

J. Juneja, R. Bansal, K. Cho, J. Sedoc, and N. Saphra. Linear connectivity reveals generalization strategies. *arXiv preprint arXiv:2205.12411*, 2022. 1

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 1

R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, R. Ge, and S. Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019. 2, 8, 28

E. S. Lubana, E. J. Bigelow, R. P. Dick, D. Krueger, and H. Tanaka. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR, 2023. 1, 12

J. Lucas, J. Bae, M. R. Zhang, S. Fort, R. Zemel, and R. Grosse. Analyzing monotonic linear interpolation in neural network loss landscapes. *arXiv preprint arXiv:2104.11044*, 2021. 1

E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020. 12

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. 4, 29, 37

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019. 3, 4, 29, 30, 31, 32, 33, 34, 35, 37

S. I. Mirzadeh, M. Farajtabar, D. Gorur, R. Pascanu, and H. Ghasemzadeh. Linear mode connectivity in multitask and continual learning. *arXiv preprint arXiv:2010.04495*, 2020. 12

B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020. 1

P.-M. Nguyen and H. T. Pham. A rigorous framework for the mean field limit of multilayer neural networks. *Mathematical Statistics and Learning*, 6 (3):201–357, 2023. 37

A. Pensia, S. Rajput, A. Nagle, H. Vishwakarma, and D. Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in neural information processing systems*, 33:2599–2610, 2020. 12

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019. 3, 12, 13

J. M. Phillips. Chernoff-hoeffding inequality and applications. *arXiv preprint arXiv:1209.6396*, 2012. 13

F. Pittorino, A. Ferraro, G. Perugini, C. Feinauer, C. Baldassi, and R. Zecchina. Deep networks on toroids: removing symmetries reveals the structure of flat regions in the landscape geometry. In *International Conference on Machine Learning*, pages 17759–17781. PMLR, 2022. 1

Y. Qin, C. Qian, J. Yi, W. Chen, Y. Lin, X. Han, Z. Liu, M. Sun, and J. Zhou. Exploring mode connectivity for pre-trained language models. *arXiv preprint arXiv:2210.14102*, 2022. 12

A. Rame, M. Kirchmeyer, T. Rahier, A. Rakotomamonjy, P. Gallinari, and M. Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 1

L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017. 1

A. Shevchenko and M. Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks. In *International Conference on Machine Learning*, pages 8773–8784. PMLR, 2020. 2, 8, 28

S. P. Singh and M. Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020. 2, 12

N. Tatro, P.-Y. Chen, P. Das, I. Melnyk, P. Sattigeri, and R. Lai. Optimizing mode connectivity via neuron alignment. *Advances in Neural Information Processing Systems*, 33:15300–15311, 2020. 1

L. Venturi, A. S. Bandeira, and J. Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20:133, 2019. 1

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009. 12, 14

T. J. Vlaar and J. Frankle. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning*, pages 22325–22341. PMLR, 2022. 1

H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020. 12

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Advances in Neural Information Processing Systems*, 2019. 14, 15, 16, 26, 32, 33, 35

M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned

models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022. 1

Y. Wu. Packing, covering, and consequences on minimax risk. http://www.stat.yale.edu/∼yw562/teaching/598/lec14.pdf, 2016. [Online; accessed October-10-2023]. 14

D. Yunis, K. K. Patel, P. H. P. Savarese, G. Vardi, J. Frankle, M. Walter, K. Livescu, and M. Maire. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. 1

M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019. 12

P. Zhao, P.-Y. Chen, P. Das, K. N. Ramamurthy, and X. Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 2020. 1

T. Zhou, J. Zhang, and D. H. Tsang. Mode connectivity and data heterogeneity of federated learning. *arXiv preprint arXiv:2309.16923*, 2023a. 12

Z. Zhou, Y. Yang, X. Yang, J. Yan, and W. Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *arXiv preprint arXiv:2307.08286*, 2023b. 1, 12

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, we always provide propositions and theorems along with their assumptions
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. No, most of our theorems are theoretical. And our proposed method clearly has a complexity comparable to the existing one. We focus our discussion on their performance.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, we provide our code along with requirements file.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes all the proofs are available in the Appendix
   (c) Clear explanations of any assumptions. Yes, assumptions are well stated and discussed.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, code is provided as an URL to reproduce the experiment in section 6.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes data and hyperparameters are given in the paper and set in the code.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes we ran experiments multiple times, provided the error bars with all details.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No, our code is very light, it ran on a single GPU (Nvidia Tesla T4) in just a couple of hours.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

# A EXTENDED RELATED WORK

Frankle et al. [2020] were the ones to coin the term Linear Mode Connectivity and the first to recognize the importance of this structure. Notably, they studied this structure through its connections with pruning and the lottery ticket hypothesis [Frankle and Carbin, 2018, Malach et al., 2020, Pensia et al., 2020, Ferbach et al., 2022].

It is worth noting that some recent works in the literature study linear mode connectivity specifically in a layer-wise manner, especially due to the permutation symmetries we described in section 1. For example, Zhou et al. [2023b], Adilova et al. [2023] both study the effect of layer-wise averaging when connecting two deep neural networks. This is aligned with our theoretical study since we recursively align deep networks layer after layer.

The mode connectivity framework has been used as a tool to better understand the similarity between two models or the effect of a training procedure on the trained model. For example, Lubana et al. [2023] introduces mechanistic similarity to quantify how two models react to the same alteration of the data in latent space. They show relations between mode connectivity and mechanistic similarity. Especially they prove that if two models cannot be linear mode connected, then they are mechanistically dissimilar. Moreover, Mirzadeh et al. [2020] use the mode connectivity framework to study whether continual and sequential learning (two different training procedure for multitask learning) are converging to a similar solution. Finally, Qin et al. [2022] explores mode connectivity of pretrained langage models, especially how hyper-parameters affect mode connectivity and how mode connectivity evolves during training.

Computational optimal transport has been leveraged by Singh and Jaggi [2020], Akash et al. [2022] to find paths between two models in parameter space. The latter formulates the model fusion problem as a Wasserstein barycenter problem.

Past works have studied mode connectivity through the lens of model averaging with applications in federated learning [Yurochkin et al., 2019, Wang et al., 2020]. Zhou et al. [2023a] studies the effect of data heterogeneity in federated learning on mode connectivity of global modes.

# B PROOFS AND DETAILS ABOUT OPTIMAL TRANSPORT THEORY FOR LMC

## B.1 Background on optimal transport and convexity lemma

Optimal transport is a mathematical framework that aims at quantifying distances between distributions. We refer to the books Villani et al. [2009] and Peyré et al. [2019] (focused on computational aspects and applications to data science) for an extensive overview of this topic.

**Definition B.1** (Wasserstein distance [Villani et al., 2009])**.** *Let $(\mathcal{X}, d)$ a Polish metric space, $p \in [1, \infty)$. $\forall \mu, \nu \in \mathcal{P}_1(\mathcal{X})$, define the $p$-Wasserstein distance between $\mu$ and $\nu$ by:*

$$\mathcal{W}_p(\mu, \nu) = \left( \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} d^p(x, y) d\pi(x, y) \right)^{1/p}$$

Recall that $\Pi(\mu, \nu)$ denotes the set of coupling between $\mu$ and $\nu$, i.e.,

$$\pi \in \Pi(\mu, \nu) \Leftrightarrow \pi \in \mathcal{P}_1(\mathcal{X}^2) \text{ with marginals } \mu, \nu$$

This defines a distance and especially it satisfies the triangular inequality. Moreover we state a Jensen type inequality proved in Villani et al. [2009]:

**Lemma B.2** (Convexity of the optimal cost (Theorem 4.8 in Villani et al. [2009]))**.** *Let $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ a distance and for $p \in [1, \infty)$, $\mathcal{W}_p : \mathcal{P}_1(\mathbb{R}^n)^2 \to \mathbb{R}_+$ its associate Wasserstein distance. Let $(\Theta, \lambda)$ be a probability space and $\mu_\theta, \nu_\theta$ two measurable functions on that space taking values in $\mathcal{P}_1(\mathbb{R}^n)$. Then,*

$$\mathcal{W}_p^p \left( \int_\Theta \mu_\theta \lambda(d\theta), \int_\Theta \nu_\theta \lambda(d\theta) \right) \leq \int_\Theta \mathcal{W}_p^p(\mu_\theta, \nu_\theta) \lambda(d\theta)$$

*Proof.* To apply Theorem 4.8 in Villani et al. [2009], we just need to notice that $d^p(\cdot, \cdot)$ is continuous, $d^p(\cdot, \cdot) \geq 0$ and the associated optimal cost functional is $\mathcal{W}_p^p(\cdot, \cdot)$. □

### B.2 Validity of the OT viewpoint: Birkhoff's theorem

We have motivated in Section 2 the following minimization problem (Equation (2)):

$$\begin{aligned}\Pi_\ell &= \arg\min_{\Pi \in \mathcal{S}_{m_\ell}} \|W_A^\ell - \Pi W_B^\ell \Pi_{\ell-1}^T\|_2^2 \\ &= \arg\min_{\pi \in \mathcal{S}_{m_\ell}} \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \|[W_A^\ell]_{i:} - [W_B^\ell \Pi_{\ell-1}^T]_{\pi_i:}\|_2^2\end{aligned}$$

Since LMC effectiveness will be related to the effectiveness of this optimization problem, we want to quantify the minimization error:

$$\min_{\Pi \in \mathcal{S}_{m_\ell}} \|W_A^\ell - \Pi W_B^\ell \Pi_{\ell-1}^T\|_2^2$$

We highlight the similarity with previous Definition B.1. The main difference being that the latter minimizes the cost among all couplings $\pi \in \Pi(\mu, \nu)$ between the two distributions, especially transport plans that can split mass. Permutation must be on the other hand seen as Monge maps, i.e. deterministic maps. This ambiguity is solved with the following lemma:

**Lemma B.3** (Proposition 2.1 in Peyré et al. [2019]). *Let $m, n$ integers and $x_1, ..., x_m, y_1, ..., y_m$, $2m$ points of $\mathbb{R}^n$. Let $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}, \hat{\nu}_m = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}$ their associated empirical measures. Consider $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ a distance function and for $p \in [1, \infty)$, $\mathcal{W}_p$ its associated Wasserstein distance. Then one has:*

$$\mathcal{W}_p := \inf_{\pi \in \Pi(\hat{\mu}_m, \hat{\nu}_m)} \left( \int_{\mathbb{R}^n \times \mathbb{R}^n} d(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}} = \min_{\pi \in \mathcal{S}_m} \left( \frac{1}{m} \sum_{i=1}^m d(x_i, y_{\sigma_i})^p \right)^{\frac{1}{p}}$$

*Proof.* $\Pi(\hat{\mu}_m, \hat{\nu}_m)$ is the convex envelope of its extremal points which are described by permutations by Birkhoff's theorem. Moreover,

$$\pi \in \mathcal{P}_1(\mathbb{R}^n \times \mathbb{R}^n) \to \int d(x,y)^p d\pi(x,y)$$

is linear and therefore its infimum is attained on one extremal point of $\Pi(\hat{\mu}_m, \hat{\nu}_m)$ □

This lemma implies equality between the Wasserstein distance between two empirical measures and the minimum cost over the set of permutations. We can therefore restrict our study to permutations while still using tools from general optimal transport theory and convergence rates of empirical measures in Wasserstein distance.

### B.3 Technical lemmas

The following lemma will be very useful in the following and shows that Binomial variables are concentrated around their expectation.

**Lemma B.4** (Hoeffding inequality for Binomial variables). *Let $\mathcal{B}_{\frac{p}{2},m} \sim \mathcal{B}(\frac{p}{2}, m)$ a binomial variable with $p \in [0, 1]$. Then,*

$$\mathbb{P}\left(\frac{\mathcal{B}_{\frac{p}{2},m}}{m} \geq p\right) \leq \exp\left(-\frac{p^2 m}{2}\right)$$

Proof: This is a simple application of Hoeffding concentration inequality [Phillips, 2012] since a Binomial variable is a sum of independent Bernoulli variables.

**Lemma B.5.** *Let $a, b > 0$ such that $a + b = 1$ and let $\mu_1, \mu_2, \nu_1, \nu_2 \in \mathcal{P}_1(\mathcal{X})$ with $(\mathcal{X}, d)$ a Polish space. Then, $\forall p \in [1, \infty)$:*

$$\mathcal{W}_p^p(a\mu_1 + b\mu_2, a\nu_1 + b\nu_2) \leq a\mathcal{W}_p^p(\mu_1, \nu_1) + b\mathcal{W}_p^p(\mu_2, \nu_2) \tag{5}$$

*Proof.* Just apply Lemma B.2 with $\mu_\Theta, \nu_\Theta$ such that $\mathbb{P}\left((\mu_\Theta, \nu_\Theta) = (\mu_1, \nu_1)\right) = a$ and $\mathbb{P}\left((\mu_\Theta, \nu_\Theta) = (\mu_2, \nu_2)\right) = b$. □

**Lemma B.6** (Hölder, Remark 6.6 in Villani et al. [2009])**.** *Let $(\mathcal{X}, d)$ a Polish space, let $p, q \in [1, \infty]$ such that $p \leq q$:*

$$\forall \mu, \nu \in \mathcal{P}_1(\mathcal{X}),\ \mathcal{W}_p \leq \mathcal{W}_q \tag{6}$$

*Proof.* Just apply Hölder's inequality. □

### B.4 Definitions and technical lemma on packing numbers

Let $n \in \mathbb{N}^*$, $S \subset \mathbb{R}^n$ and $d(\cdot, \cdot)$ a distance of $\mathbb{R}^n$. We recall the definitions of packing numbers and covering numbers as well as two lemmas stated and proved in Wu [2016]:

**Definition B.7** ($\varepsilon$-covering number [Weed and Bach, 2019])**.** *The $\varepsilon$-covering of a set $S$ denoted $\mathcal{N}_\varepsilon(S)$ is the minimum number $m$ of closed balls $B_1, \ldots, B_m$ of radius $\varepsilon$ such that $S \subset \bigcup_{i=1}^m B_i$.*

**Definition B.8** ($\varepsilon$-packing number)**.** *The $\varepsilon$-packing number of a set $S$ denoted $\mathcal{P}_\varepsilon(S)$ is the maximum number $m$ of distinct points $\theta_1, ..., \theta_m \in S$ such that $\forall i \neq j$, $\|\theta_i - \theta_j\| > \varepsilon$.*

**Lemma B.9.** *If the distance $d(\cdot, \cdot)$ comes from a norm, ($d(x, y) = \|x - y\|$), denoting $Leb_n$ the Lebesgue measure in $\mathbb{R}^n$ we have: $\forall S \subset \mathbb{R}^n, \forall \varepsilon > 0$,*

$$\mathcal{N}_\varepsilon(S) \leq \frac{Leb_n(S + \mathcal{B}(0, \varepsilon/2))}{Leb_n(\mathcal{B}(0, \varepsilon/2))}$$

*Proof.* We prove first $\mathcal{N}_\varepsilon(S) \leq \mathcal{P}_\varepsilon(S)$. Indeed considering $m = \mathcal{P}_\varepsilon(S)$ and $\theta_1, \ldots, \theta_m$ associated, we know by definition of the $\varepsilon$-packing number that $\forall \theta \in S, \exists i \in [m]$ such that $\|\theta_i - \theta\| \leq \varepsilon$. This shows that $S \subset \bigcup_{i=1}^m \mathcal{B}(\theta_i, \varepsilon)$

Now, on the other hand, we know that all balls $\mathcal{B}(\theta_i, \frac{\varepsilon}{2})$ are disjoint and $\bigcup_{i=1}^m \mathcal{B}(\theta_i, \frac{\varepsilon}{2}) \subset S + \mathcal{B}(0, \varepsilon/2)$. This yields the result by a volume (i.e., Lebesgue measure) argument. □

**Lemma B.10.** *If the distance $d(\cdot, \cdot)$ comes from a norm, ($d(x, y) = \|x - y\|$) we have: $\forall S \subset \mathbb{R}^n, \forall \varepsilon > 0$,*

$$\mathcal{N}_\varepsilon(S) \geq \left(\frac{1}{\varepsilon}\right)^n \frac{Leb_n(S)}{Leb_n(\mathcal{B}(0, 1))}$$

*Proof.* Notice that given a covering $\bigcup_{i=1}^{\mathcal{N}_\varepsilon(S)} \mathcal{B}(x_i, \varepsilon) \supset S$, by a volume argument we get

$$\mathrm{Leb}_n(S) \leq \varepsilon^n \mathcal{N}_\varepsilon(S) \mathrm{Leb}_n(\mathcal{B}(0, 1))$$

□

where we have used homogeneity of the norm.

### B.5 Lemmas on convergence rates of empirical measures

This section is devoted to proving convergence rates in Wasserstein distance of empirical measure towards the underlying distribution whose points are drawn. More precisely, given $\mu \in \mathcal{P}_1(\mathbb{R}^n)$ a probability measure on an euclidean space and $p \in [1, \infty)$, we will focus on bounding the quantity $\mathbb{E}\left[\mathcal{W}_p^p(\hat{\mu}_m, \mu)\right]$ as $m$ grows, where $\hat{\mu}_m$ is a random empirical measure i.e., $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ where $x_i \overset{\text{i.i.d.}}{\sim} \mu$.

We will first prove the following lemma:

**Lemma B.11.** *Consider $X_n \sim \mathcal{N}\left(0, \frac{I_n}{n}\right)$ a random variable whose law is parametrized by $n \in \mathbb{N}^*$. There exists a universal constant $D_0$ such that*

$$\forall n \in \mathbb{N}^*, \forall c > 1, \mathbb{E}[\|X_n\|_2^2 \mid \|X_n\|_2 > c] \leq D_0 c^2$$

*Jensen inequality implies:*

$$\forall n \in \mathbb{N}^*, \forall c > 1, \mathbb{E}[\|X_n\|_2; \mid \|X_n\|_2 > c] \leq \sqrt{D_0} c$$

*Proof.* We write

$$\begin{aligned}\mathbb{E}[\|X_n\|_2^2 | \|X_n\|_2 > c] &= \frac{\int_c^{+\infty} r^2 r^{n-1} e^{-\frac{r^2 n}{2}} dr}{\int_c^{+\infty} r^{n-1} e^{-\frac{r^2 n}{2}} dr} \\ &\leq 4c^2 + \sum_{m=2}^{\infty} \frac{\int_{mc}^{(m+1)c} r^2 r^{n-1} e^{-\frac{r^2 n}{2}} dr}{\int_c^{2c} r^{n-1} e^{-\frac{r^2 n}{2}} dr} \\ &\leq 4c^2 + \sum_{m=2}^{\infty} \frac{\int_0^1 ((m+1)c)^{n+1} e^{-\frac{((m+t)c)^2 n}{2}} c dt}{\int_0^1 c^{n-1} e^{-\frac{((1+t)c)^2 n}{2}} c dt}\end{aligned}$$

But, $\forall t \in [0,1], \forall m \geq 2$,

$$\frac{((m+1)c)^{n+1} e^{-\frac{((m+t)c)^2 n}{2}}}{c^{n-1} e^{-\frac{((1+t)c)^2 n}{2}}} \leq c^2 (m+1)^{n+1} e^{-\frac{c^2 n}{2}(m^2-1)} \leq c^2 (2m)^{2n} e^{-\frac{c^2 n}{2}(m^2-1)} \leq c^2 e^{-n\left(\frac{c^2}{2}(m^2-1) - 2\log(2m)\right)}$$

It is clear that uniformly in $n$, $\sum_{m=2}^{\infty} c^2 e^{-n\left(\frac{c^2}{2}(m^2-1)-2\log(2m)\right)} \xrightarrow[c\to\infty]{} 0$ which proves the lemma. □

We will now prove a bound on the rate of convergence in Wasserstein distance of an empirical measure to the underlying distribution when this one has a bounded support. Denote $\mathcal{B}_2^k(0,r)$ the euclidean ball centered around 0 of radius $r$ in dimension $k$.

**Lemma B.12.** *Let $\mu \in \mathcal{P}_1(\mathbb{R}^n)$ be a measure whose support is included in $\mathcal{B}_2^n\left(0, \frac{1}{12}\right) \subset \mathbb{R}^n$ with $n \geq 5$ Then, $\forall m \geq 1$ we have*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq D_1 \left(\frac{1}{m}\right)^{2/n}$$

*Where $D_1 = 27^2 \left(2 + \frac{1}{\sqrt{3}-1}\right)$*

*Proof.* We know from Lemma B.9 that when considering $\|\cdot\|_2$ the distance for defining covering number, $\forall \varepsilon' \leq \frac{1}{6}$:

$$\mathcal{N}_{\varepsilon'}(\mathcal{B}_2^n(0, 1/12)) = \left(\frac{\frac{1}{12} + \frac{\varepsilon'}{2}}{\frac{\varepsilon'}{2}}\right)^n \leq \left(\frac{\frac{1}{6} + \varepsilon'}{\varepsilon'}\right)^n \leq (3\varepsilon')^{-n}$$

and therefore also when $\varepsilon' \leq \frac{1}{27}$ Applying Proposition 15 from Weed and Bach [2019] we get that that since $\text{Supp}(\mu) \subset \mathcal{B}_2^n(0, 1/12) \subset \mathcal{B}_2^n(0, 1/12) + \mathcal{B}_2^n(0, \varepsilon)$ for any $\varepsilon > 0$, if $n \geq 5$,

$$\forall m \geq 1, \mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq 27^2 \left(2 + \frac{1}{\sqrt{3}-1}\right) \left(\frac{1}{m}\right)^{2/n}$$

as well as if $n \geq 3$,

$$\forall m \geq 1, \mathbb{E}[\mathcal{W}_1(\hat{\mu}_m, \mu)] \leq 27 \left(2 + \frac{1}{\sqrt{3}-1}\right) \left(\frac{1}{m}\right)^{1/n}$$

□

We can prove the same kind of inequality when $\mu$ concentrates mass around an approximately low dimensional set.

**Lemma B.13.** *Let $\mu \in \mathcal{P}_1(\mathbb{R}^n)$ be a measure whose support is included in $\mathcal{B}_2^k(0, 1/12) \times \mathcal{B}_2^{n-k}(0, r)$ with $k \geq 5$. Then, $\forall m \leq (3r)^{-k}$ we have*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq D_1 \left(\frac{1}{m}\right)^{2/k}$$

*where* $D_1 = 27^2 \left(2 + \frac{1}{\sqrt{3}-1}\right)$

*Proof.* This is the same proof as before, just notice that $\text{Supp}(\mu) \subset \mathcal{B}_2^k(0, 1/12) \times \{0\}^{n-k} + \mathcal{B}_2^n(0, r)$ and as before:

$$\mathcal{N}_{\varepsilon'}(\mathcal{B}_2^k(0, 1/12) \times \{0\}^{n-k}) \leq \mathcal{N}_{\varepsilon'}(\mathcal{B}_2^k(0, 1/12)) \leq (3\varepsilon')^{-k} \tag{7}$$

if $\varepsilon' \leq \frac{1}{27}$. □

We will now extend our results to unbounded variables very concentrated around bounded sets, beginning with multivariate normal random variable.

**Lemma B.14.** *Consider a centered multivariate normal distribution $\mu$ on $\mathbb{R}^n$ with covariance matrix $\text{Diag}(\lambda_1, \dots, \lambda_n)$ where $\frac{1}{n} \geq \lambda_1 \geq \dots \geq \lambda_n \geq 0$. There exists two universal constants $D_2, E_2$ such that $\forall n \geq 5, \forall m \in \mathbb{N}^*$, if $m \geq E_2^n$ then,*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq \frac{D_2}{n} \log(m) \left(\frac{1}{m}\right)^{2/n}$$

*In that case,*

$$\mathbb{E}[\mathcal{W}_1(\hat{\mu}_m, \mu)] \leq \frac{\sqrt{D_2}}{\sqrt{n}} \sqrt{\log(m)} \left(\frac{1}{m}\right)^{1/n}$$

*Proof.* We will use previous Lemma B.12. The problem is that it applies only for a bounded distribution. Therefore we will have to bound the mass of a multivariate Gaussian outside of a euclidean ball. We will prove the lemma for $\lambda_1 = ... = \lambda_n = \frac{1}{n}$ by noticing that it extends for smaller eigenvalues by rescaling the axis.

Let $f \in (0, 1), c > 0, X \sim \mu$.

Lemma 1 from Weed and Bach [2019] tells us:

$$\mathbb{P}\left(\|X\|_2^2 \geq c^2 \sum_{i=1}^n \lambda_i\right) \leq e^{-\frac{c^2}{4}}$$

Noticing $\sum_{i=1}^n \lambda_i \leq 1$ and taking $c = 2\sqrt{\log\left(\frac{1}{f}\right)}$ we get that:

$$\mathbb{P}(\|X\|_2^2 \geq c^2) \leq f$$

Using Lemma B.4 with $p = 2f$ we get that with probability at least $1 - \exp(-2f^2 m)$, a fraction at least $1 - 2f$ of vectors $x_i$ lies in $\mathcal{B}_2(0, c)$. We denote $H_f$ this event such that $\mathbb{P}(H_f) \geq 1 - \exp(-2f^2 m)$.

Further denote $I$ the corresponding set of indices for $x_i$, $\hat{\mu}_{m,I} = \frac{\sum_{i \in I} \delta_{x_i}}{|I|}$ and $\hat{\mu}_{m,I^c} = \frac{\sum_{i \notin I} \delta_{x_i}}{|I^c|}$. Finally for a Borel set $U \subset \mathbb{R}^n$ denote $\mu_{|U}$ the renormalized restricted measure $\mu$ on $U$: $\mu_{|U} = \frac{1}{\mu(U)} \mu \mathbb{1}_U$

We will now consider two cases:

**1st case** We consider the case $\frac{|I|}{m} \leq \mu(\mathcal{B}(0,c))$ and denote $\text{Case}_1$ this set.

In that case, we can write using Lemma B.5:

$$\begin{cases} \mathcal{W}_2^2(\hat{\mu}_m, \mu) & \leq \frac{|I|}{m}\mathcal{W}_2^2(\hat{\mu}_{m,I}, \mu_{|\mathcal{B}_2(0,c)}) + \frac{|I^c|}{m}\mathcal{W}_2^2\left(\hat{\mu}_{m,I^c}, \frac{\mu_{|\mathcal{B}_2(0,c)} - \frac{|I|}{m}\mu_{|\mathcal{B}_2(0,c)}}{\frac{|I^c|}{m}}\right) \\ \mathcal{W}_1(\hat{\mu}_m, \mu) & \leq \frac{|I|}{m}\mathcal{W}_1(\hat{\mu}_{m,I}, \mu_{|\mathcal{B}_2(0,c)}) + \frac{|I^c|}{m}\mathcal{W}_1\left(\hat{\mu}_{m,I^c}, \frac{\mu_{|\mathcal{B}_2(0,c)} - \frac{|I|}{m}\mu_{|\mathcal{B}_2(0,c)}}{\frac{|I^c|}{m}}\right) \end{cases}$$

By previous Lemma B.12, we know the existence of $D_1$ such that if $n \geq 5$:

$$\mathbb{E}\left[\mathcal{W}_2^2(\hat{\mu}_{m,I}, \mu_{|\mathcal{B}_2(0,c)})\right] \leq D_1(12c)^2\left(\frac{1}{|I|}\right)^{2/n} \leq (144D_1)c^2\left(\frac{1}{\frac{|I|}{m}m}\right)^{2/n}$$

Therefore we get since $\frac{2}{n} \leq 1$ and $\frac{|I|}{n} \leq 1$:

$$\mathbb{E}\left[\frac{|I|}{m}\mathcal{W}_2^2(\hat{\mu}_{m,I}, \mu_{|\mathcal{B}_2(0,c)})\right] \leq (144D_1)c^2\left(\frac{1}{m}\right)^{2/n}$$

We know from Lemma B.11 the existence of a universal constant $D_0$ such that: $\forall c \geq 1, \mathbb{E}[\|X\|_2^2 | \|X\|_2 \geq c] \leq D_0 c^2$. Using a triangular inequality

$$\mathbb{E}\left[\mathcal{W}_2^2(\hat{\mu}_{m,I^c}, \mu_{|\mathcal{B}_2(0,c)^c})\right] \leq 4D_0c^2$$

Finally, conditioned on the event $H_f$ and that we are in $\text{Case}_1$, we get that if $c \geq 1$:

$$\begin{cases} \mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu) | H_f \bigcap \text{Case}_1] \leq (144D_1)c^2\left(\frac{1}{m}\right)^{2/n} + 8fD_0c^2 \\ c = 2\sqrt{\log(\frac{1}{f})} \end{cases}$$

**2nd case** We consider the case $\frac{|I|}{m} > \mu(\mathcal{B}_2(0,c))$ and denote $\text{Case}_2$ this set.

In that case, we denote $I' \subset I$ taken randomly uniformly, such that $|I'| = \max\left\{k \geq 1, \frac{k}{m} \leq \mu(\mathcal{B}(0,c))\right\}$ and denote $\hat{\mu}_{m,I'}$ the renormalized empirical measure with points in $I'$ and $\hat{\mu}_{m,I\bigcap I'^c}$ the renormalized empirical measure with points in $I\bigcap I'^c$.

We can write:

$$\begin{aligned} \mathcal{W}_2^2(\hat{\mu}_m, \mu) \leq \frac{|I'|}{m}\mathcal{W}_2^2(\hat{\mu}_{m,I'}, \mu_{|\mathcal{B}_2(0,c)}) + \frac{|I| - |I'|}{m}\mathcal{W}_2^2(\hat{\mu}_{m,I\bigcap I'^c}, \mu_{|\mathcal{B}_2(0,c)^c}) \\ + \left(\mu\big(\mathcal{B}_2(0,c)\big) - \frac{|I'|}{m}\right)\mathcal{W}_2^2(\hat{\mu}_{m,I\bigcap I'^c}, \mu_{|\mathcal{B}_2(0,c)}) + \frac{|I^c|}{m}\mathcal{W}_2^2\left(\hat{\mu}_{m,I^c}, \mu_{|\mathcal{B}_2(0,c)^c}\right) \end{aligned}$$

Provided $fm > 1$, we know that $\frac{|I'|}{m} \geq 1 - 3f$.

We can repeat all the previous arguments and get that if $c \geq 1$ and $m$:

$$\begin{cases} \mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu) | H_f \bigcap \text{Case}_2] \leq (144D_1)c^2\left(\frac{1}{m}\right)^{2/n} + 16fD_0c^2 \\ c = 2\sqrt{\log(\frac{1}{f})} \end{cases}$$

Finally,

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)] \leq \mathbb{P}(H_f)\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)|H_f] + (1-\mathbb{P}(H_f))\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)|H_f^c]$$

We can easily bound as before $\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)|H_f^c] \leq 4D_0c^2$

which yields finally:

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)] \leq (144D_1)4\log(\frac{1}{f})\left(\frac{1}{m}\right)^{2/n} + 16fD_04\log(\frac{1}{f}) + 2\exp(-f^2m)4D_04\log(\frac{1}{f})$$

Taking $f = \left(\frac{1}{m}\right)^{2/n}$ we get

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)] \leq \frac{1152D_1}{n}\log(m)\left(\frac{1}{m}\right)^{2/n} + D_0\frac{128}{n}\log(m)\left(\frac{1}{m}\right)^{2/n} + \frac{64}{n}\exp(-m^{1-4/n})\log(m)$$

Note now that there exists a universal constant $C > 0$ such that $\forall n \geq 5, \forall m \geq 1$:

$$\exp(-m^{1-4/n}) \leq \exp(-m^{1/5}) \leq C\left(\frac{1}{m}\right)^{2/5} \leq C\left(\frac{1}{m}\right)^{2/n}$$

Finally we get the existence of universal constants $D_2, E_2$ such that if $\left(\frac{1}{m}\right)^{2/n} \leq \frac{1}{E_2^2}$ (to ensure $c \geq 1$ take for example $E_2 = \exp(\frac{1}{4})$),

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)] \leq \frac{D_2}{n}\log(m)\left(\frac{1}{m}\right)^{2/n}$$

To prove the second part of the lemma just apply Lemma B.6 and Jensen inequality.

□

We can finally extend Lemma B.13 to unbounded distributions as we just extended Lemma B.12 to unbounded distributions.

**Lemma B.15.** *Let $\lambda_1 \geq \ldots \geq \lambda_n$ and $\mu = \mathcal{N}(0, \mathrm{Diag}((\lambda_i)_{i=1}^n))$, with $k \geq 5$. Suppose $1 \geq \sum_{i=1}^k \lambda_i$. Denote $\eta = \frac{\sqrt{\sum_{i=k+1}^n \lambda_i}}{4\sqrt{\sum_{i=1}^k \lambda_i}}$. We know the existence of two universal constants $D_2', E_2'$ such that if $\eta^{-k} \geq m \geq E_2'^k$, then:*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m,\mu)] \leq \frac{D_2'}{k}\log(m)\left(\frac{1}{m}\right)^{2/k}]$$

*In that case,*

$$\mathbb{E}[\mathcal{W}_1(\hat{\mu}_m,\mu)] \leq \frac{\sqrt{D_2'}}{\sqrt{k}}\sqrt{\log(m)}\left(\frac{1}{m}\right)^{1/k}]$$

*Proof.* We will follow the same steps as previously. Let $X \sim \mu$ and denote $\underline{X} = (X_1, \ldots, X_k) \in \mathbb{R}^k$, $\bar{X} = (X_{k+1}, \ldots, X_n) \in \mathbb{R}^{n-k}$.

$$\mathbb{P}\left(\|\underline{X}\|_2^2 \geq c^2\sum_{i=1}^k \lambda_i\right) \leq e^{-\frac{c^2}{4}}$$
$$\mathbb{P}\left(\|\bar{X}\|_2^2 \geq c^2\sum_{i=k+1}^n \lambda_i\right) \leq e^{-\frac{c^2}{4}}$$

Take $c = 2\sqrt{\log(\frac{2}{f})}$. Then, by the same arguments as before, using Lemma B.4 and union bounds, with probability at least $1 - 2\exp(-f^2 m/2)$ a fraction at least $1 - 2f$ of points $x_i$ are in $\mathcal{B}_c := \mathcal{B}_k(0, c\sqrt{\sum_{i=1}^k \lambda_i}) \times \mathcal{B}_{n-k}(0, c\sqrt{\sum_{i=k+1}^n \lambda_i})$. We denote $H_f$ such an event and $I$ such a set of indices.

By using Lemma B.13, we know that in that case we can bound

$$\mathbb{E}\left[\mathcal{W}_2^2(\hat{\mu}_{m,I}, \mu_{|\mathcal{B}_c})|H_f\right] \leq \left(12c\sqrt{\sum_{i=1}^k \lambda_i}\right)^2 C_1 \left(\frac{1}{m}\right)^{2/k} \leq (12c)^2 C_1 \left(\frac{1}{m}\right)^{2/k}$$

if $m \leq \left(\frac{3\sqrt{\sum_{i=k+1}^m \lambda_i}}{12\sqrt{\sum_{i=1}^k \lambda_i}}\right)^{-k}$ where we have used $\sum_{i=1}^k \lambda_i \leq 1$.

Moreover, we know that

$$\begin{aligned}
\mathbb{E}\left[\|X\|_2^2 | X \notin \mathcal{B}_2^k(0, c\sqrt{\sum_{i=1}^k \lambda_i}) \times \mathcal{B}_2^{n-k}(0, c\sqrt{\sum_{i=k+1}^n \lambda_i})\right] &\leq \mathbb{E}\left[\|X\|_2^2 | X \notin \mathcal{B}_2^k(0, c\sqrt{\sum_{i=1}^k \lambda_i})\right] \\
&\quad + \mathbb{E}\left[\|\bar{X}\|_2^2 | \bar{X} \notin \mathcal{B}_2^{n-k}(0, c\sqrt{\sum_{i=k+1}^n \lambda_i})\right] \\
&\leq D_0 c^2 \sum_{i=1}^k \lambda_i + D_0 c^2 \sum_{i=k+1}^n \lambda_i \\
&\leq 2 D_0 c^2 \sum_{i=1}^k \lambda_i \\
&\leq 2 D_0 c^2
\end{aligned}$$

We just need to differentiate the same two cases as in the proof of Lemma B.14 to get that finally, if $m \leq \left(\frac{\sqrt{\sum_{i=k+1}^m \lambda_i}}{4\sqrt{\sum_{i=1}^k \lambda_i}}\right)^{-k}$ we can bound as before, with $c = 2\sqrt{\log(\frac{2}{f})}$:

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq (12c)^2 C_1 \left(\frac{1}{m}\right)^{2/k} + 16 f * 4 C_0 c^2 + 2\exp(-\frac{f^2 m}{2}) * 4 * 2 C_0 c^2$$

Taking $f = \left(\frac{1}{m}\right)^{2/k}$, we see as before the existence of $E_2', D_2'$ such that if $E_2'^k \leq m \leq \eta^{-k}$ then,

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_m, \mu)] \leq \frac{D_2'}{k} \log(m) \left(\frac{1}{m}\right)^{2/k}$$

We choose $E_2' = \sqrt{2e^{-1/4}}$ such that $c > 1 \Leftrightarrow 2\sqrt{\log(\frac{2}{f})} > 1 \Leftrightarrow 2\sqrt{\log(\frac{2}{(\frac{1}{m})^{2/k}})} > 1 \Leftrightarrow m > \sqrt{2e^{-1/4}}^k$

For the second part of the lemma, apply Lemma B.6 and Jensen inequality.

□

### B.6 Proof of Lemma 4.1

**Lemma 4.1.** *Let $\ell \in \{0, \cdots, L-1\}$ and suppose Property 1 to hold at layer $\ell$ and Assumptions 3 to 5 to hold, then Property 1 still holds at the next layer with $\tilde{m}_{\ell+1}$ given in Assumption 3 and*

$$\underline{E}_{\ell+1} = C_2 \underline{E}_\ell$$
$$E_{\ell+1} = 2C_2 E_\ell + 2C_1 \tilde{m}_\ell \underline{E}_\ell$$

*Proof.* We know from Assumption 3 the existence of a random empirical measure with $\tilde{m}_{l+1}$ points $\hat{\mu}_{\tilde{m}_{l+1}}$ such that $\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^\ell}, \hat{\mu}_{\tilde{m}_{l+1}}^{\mathcal{I}^\ell})] \leq C_1$. Therefore by using Lemma B.3 and a carefully chosen permutation, we can consider the (random) matrix $W_{\tilde{m}_{l+1}}$ associated to $W_A^{\ell+1}$ such that $\mathbb{E}[\|W_{\tilde{m}_{l+1}}^{\mathcal{I}^\ell} - W_A^{\mathcal{I}^\ell}\|_2^2] \leq C_1 m_{\ell+1}$. Note that since $W_{\tilde{m}_{\ell+1}}$ comes from an empirical measure with only $\tilde{m}_{\ell+1}$ points one can denote $\mathcal{I}^{\ell+1} = \{I_1^{\ell+1}, \dots, I_{\tilde{m}_{\ell+1}}^{\ell+1}\}$ the (random) equi-partition of $[m_{\ell+1}]$ delimiting its equal rows. In the same way, since $A$ and $B$ have the same weights distribution, we can find a permutation $\Pi_{\ell+1}$ of the layer $\ell+1$ of network $B$ such that denoting $\tilde{W}_B^{\ell+1} = \Pi_{\ell+1} W_B^{\ell+1} \Pi_\ell^T$ and taking expectations over the choice of the weights matrices, we have $\mathbb{E}[\|\tilde{W}_B^{\ell+1} - W_{\tilde{m}_{\ell+1}}\|_2^2] \leq m_{\ell+1} C_1$. Consider $W_{M_t}^{\ell+1} = t W_A^{\ell+1} + (1-t)\tilde{W}_B^{\ell+1}$ and denote $\underline{\phi}^{\ell+1}(x) = \sigma(W_{\tilde{m}_{\ell+1}} \underline{\phi}^\ell(x)), \phi_A^{\ell+1}(x) = \sigma(W_A^{\ell+1}\phi_A^\ell(x)), \phi_B^{\ell+1}(x) = \sigma(W_B^{\ell+1}\phi_B^\ell(x)), \phi_{M_t}^{\ell+1}(x) = \sigma(W_{M_t}^{\ell+1}\phi_{M_t}^\ell(x))$. It is clear that $\forall k \in [\tilde{m}_{\ell+1}], \forall i,j \in I_k^{\ell+1}, \underline{\phi}_i^{\ell+1}(x) = \underline{\phi}_j^{\ell+1}(x)$ by definition of the choice of the equi-partition $\mathcal{I}^{\ell+1}$.

We will finally denote $\underline{\phi}'^{\ell+1}(x) \in \mathbb{R}^{\tilde{m}_{\ell+1}}, \underline{\phi}'^{\ell}(x) \in \mathbb{R}^{\tilde{m}_\ell}$ the vectors $\underline{\phi}^{\ell+1}(x), \underline{\phi}^\ell(x)$ where we have kept only one index in each of the elements of the partitions respectively $\mathcal{I}^{\ell+1}, \mathcal{I}^\ell$.

Moreover using that the non-linearity is pointwise 1-Lipschitz and $\sigma(0) = 0$,

$$\mathbb{E}[\|\underline{\phi}^{\ell+1}(x)\|_2^2] \leq \mathbb{E}_{P,Q}[\mathbb{E}[\|W_{\tilde{m}_{\ell+1}}\underline{\phi}^\ell(x)\|_2^2 | \underline{\phi}^\ell(x)]] \leq \mathbb{E}_{P,Q}[C_2 \frac{m_{\ell+1}}{m_\ell} \|\underline{\phi}^\ell(x)\|_2^2] \leq m_{\ell+1} C_2 \underline{E}_\ell$$

which yields $\underline{E}_{\ell+1} = C_2 \underline{E}_\ell$ where we have used Assumption 4.

Then

$$\begin{aligned}
\mathbb{E}_{P,Q}[\|\phi_A^{\ell+1}(x) - \underline{\phi}^{\ell+1}(x)\|_2^2] &\leq \mathbb{E}_{P,Q}[\|W_A^{\ell+1}\phi_A^\ell(x) - W_{\tilde{m}_{\ell+1}}\underline{\phi}^\ell(x)\|_2^2 \\
&\leq 2\mathbb{E}_{P,Q}[\|W_A^{\ell+1}(\phi_A^\ell(x) - \underline{\phi}^\ell(x))\|_2^2 + \|(W_A^{\ell+1} - W_{\tilde{m}_{\ell+1}})\underline{\phi}^\ell\|_2^2] \\
&\leq 2m_{\ell+1}C_2 E_\ell + 2\mathbb{E}_{P,Q}[\|(W_A^{\mathcal{I}^\ell} - W_{\tilde{m}}^{\mathcal{I}^\ell})\underline{\phi}'^\ell\|_2^2] \\
&\leq 2m_{\ell+1}C_2 E_\ell + 2\mathbb{E}_{P,Q}[\|(W_A^{\mathcal{I}_\ell} - W_{\tilde{m}_{\ell+1}}^{\mathcal{I}_\ell})\|_2^2]\mathbb{E}_{P,Q}\|\underline{\phi}'^\ell\|_2^2] \\
&\leq 2m_{\ell+1}C_2 E_\ell + 2m_{\ell+1}C_1 \frac{\tilde{m}_\ell}{m_\ell}\mathbb{E}_{P,Q}\|\underline{\phi}^\ell(x)\|_2^2 \\
&\leq 2m_{\ell+1}C_2 E_\ell + 2m_{\ell+1}C_1 \tilde{m}_\ell \underline{E}_\ell
\end{aligned}$$

where we have used $(W_A^{\ell+1} - W_{\tilde{m}_{\ell+1}})\underline{\phi}^\ell(x) = (W_A^{\mathcal{I}^\ell} - W_{\tilde{m}_{\ell+1}}^{\mathcal{I}^\ell})\underline{\phi}'^\ell$ and $\|\underline{\phi}'^\ell(x)\|_2^2 = \frac{\tilde{m}_\ell}{m_\ell}\|\underline{\phi}^\ell(x)\|_2^2$.

We do the same computations for $\mathbb{E}_{P,Q}[\|\phi_B^{\ell+1}(x) - \underline{\phi}^{\ell+1}(x)\|_2^2]$.

Finally,

$$\begin{aligned}
\mathbb{E}_{P,Q}[\|\phi_{M_t}^{\ell+1}(x) - \underline{\phi}^{\ell+1}(x)\|_2^2] &\leq t\mathbb{E}_{P,Q}[\|W_A^{\ell+1}\phi_{M_t}^\ell(x) - W_{\tilde{m}_{\ell+1}}\underline{\phi}^\ell(x)\|_2^2] + (1-t)\mathbb{E}_{P,Q}[\|W_B^{\ell+1}\phi_{M_t}(x) - W_{\tilde{m}_{\ell+1}}\underline{\phi}^\ell(x)\|_2^2] \\
&\leq 2C_2 E_\ell m_{\ell+1} + 2C_1 \tilde{m}_\ell \underline{E}_\ell m_{\ell+1}
\end{aligned}$$

where we have used convexity for the first inequality and then the same proof as above for both terms. This yields $E_{\ell+1} = 2C_2 E_\ell + 2C_1 \tilde{m}_\ell \underline{E}_\ell$.

□

### B.7 Proof of Lemma 5.1

**Lemma B.16** (Version of Assumption 3 for normal distribution)**.** *Consider $\mu_\ell \in \mathcal{P}_1(\mathbb{R}^{m_{\ell-1}})$ a multivariate Gaussian distribution with covariance matrix $\Sigma^\ell = \mathrm{Diag}(\lambda_1^\ell I_{p_{\ell-1}}, \dots, \lambda_{\tilde{m}_\ell}^\ell I_{p_{\ell-1}})$ where $p_{\ell-1} = \frac{m_{\ell-1}}{\tilde{m}_{\ell-1}}$. Suppose that $\frac{1}{m_{\ell-1}} \geq \lambda_1^\ell, \geq \dots \geq \lambda_{\tilde{m}_{\ell-1}}^\ell$ with $\tilde{m}_{\ell-1} \geq 5$. Then, there exists two universal constants $D_3, E_3$ such that $\forall \tilde{m}_\ell \geq E_3^{\tilde{m}_{\ell-1}}$ there exists a random empirical measure $\hat{\mu}_{\tilde{m}}$ with only $\tilde{m}_\ell$ points such that $\forall m_\ell \geq \tilde{m}_\ell$*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}})] \leq \frac{D_3}{\tilde{m}_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/\tilde{m}_{\ell-1}}$$

*In that case:*

$$\mathbb{E}[\mathcal{W}_1(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}})] \leq \frac{\sqrt{D_3}}{\sqrt{\tilde{m}_{\ell-1}}} \sqrt{\log(\tilde{m}_\ell)} \left(\frac{1}{\tilde{m}_\ell}\right)^{1/\tilde{m}_{\ell-1}}$$

*Proof.* We know that the distribution on the rows of $W_A^{\mathcal{I}^{\ell-1}}$ in $\mathbb{R}^{\tilde{m}_{\ell-1}}$ is multivariate Gaussian with covariance matrix $\frac{I_{\tilde{m}_{\ell-1}}}{\tilde{m}_{\ell-1}}$ since each parameters is obtained by summing the $p_{\ell-1}$ corresponding parameters of the row of $W_A^\ell$ which has covariance matrix $\frac{I_{m_{\ell-1}}}{m_{\ell-1}}$ by hypothesis. Therefore using Lemma B.14, we know the existence of constants $D_2, E_2$ such that if $m_\ell \geq E_2^{\tilde{m}_{\ell-1}}$, $\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \mu_\ell^{\mathcal{I}^{\ell-1}})] \leq \frac{D_2}{\tilde{m}_{\ell-1}} \log(m_\ell) \left(\frac{1}{m_\ell}\right)^{2/\tilde{m}_{\ell-1}}$.

Therefore considering $\hat{\mu}_{\tilde{m}_\ell}$ with the same law but only a fixed number $\tilde{m}_\ell$ of elements, we get for $m_\ell \geq \tilde{m}_\ell$:

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \mu_\ell^{\mathcal{I}^{\ell-1}})] \leq \frac{D_2}{\tilde{m}_{\ell-1}} \log(m_\ell) \left(\frac{1}{m_\ell}\right)^{2/\tilde{m}_{\ell-1}} \leq \frac{D_2}{\tilde{m}_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/\tilde{m}_{\ell-1}}$$
$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}}, \mu_\ell^{\mathcal{I}^{\ell-1}})] \leq \frac{D_2}{\tilde{m}_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/\tilde{m}_{\ell-1}}$$

Indeed, the first inequality can be obtained by noticing that $\left(x \mapsto \log(x) \left(\frac{1}{x}\right)^{2/\tilde{m}_{\ell-1}}\right)$ is decreasing for $x \geq \sqrt{e}^{\tilde{m}_{\ell-1}}$ and hence one can just increase the constant $E_3$ considered: taking $E_3 = \max\{\sqrt{e}, E_2\}$ and $D_3 = 4D_2$, by triangular inequality,

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}})] \leq 2\left(\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \mu_\ell^{\mathcal{I}^{\ell-1}})] + \mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}}, \mu_\ell^{\mathcal{I}^{\ell-1}})]\right) \leq \frac{D_3}{\tilde{m}_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/\tilde{m}_{\ell-1}}$$

For the second part of the lemma, just apply Lemma B.6 and Jensen inequality.

□

**Lemma B.17** (Version of Assumption 4 for normal variable)**.** *$\forall X \in \mathbb{R}^{m_{\ell-1}}$ we have:*

$$\mathbb{E}[\|W_A^\ell X\|_2^2] \leq \frac{m_\ell}{m_{\ell-1}} \|X\|_2^2$$
$$\mathbb{E}[\|W_{\tilde{m}_\ell} X\|_2^2] \leq \frac{m_\ell}{m_{\ell-1}} \|X\|_2^2$$

*Proof.* First, given $X \in \mathbb{R}^{m_{\ell-1}}, \forall i \in [m_\ell], (W_A^\ell X)_i = \sum_{j=1}^{m_{\ell-1}} (W_A^\ell)_{i,j} X_j$ where $(W_A^\ell)_{i,j}$ are iid following $\mathcal{N}(0, \frac{1}{m_{\ell-1}})$. Therefore, $\forall i \in [m_\ell], \mathbb{E}[(W_A^\ell X)_i^2] = \frac{1}{m_{\ell-1}} \|X\|_2^2$. Finally,

$$\mathbb{E}[\|W_A^\ell X\|_2^2] = \sum_{i=1}^{m_\ell} \mathbb{E}[(W_A^\ell X)_i^2] = \frac{m_\ell}{m_{\ell-1}} \|X\|_2^2$$

Since the weight matrix associated to $\hat{\mu}_{\tilde{m}_\ell}$ denoted $W'_{\tilde{m}_\ell} \in \mathcal{M}_{\tilde{m}_\ell, m_{\ell-1}}(\mathbb{R})$ has only $\tilde{m}_\ell$ raws, we have expanded it to a matrix $W_{\tilde{m}_\ell} \in \mathcal{M}_{m_\ell, m_{\ell-1}}(\mathbb{R})$ by cloning raws in the same element of the partition $\mathcal{I}^\ell$ given by the pairing between $\hat{\mu}_A^{\mathcal{I}^\ell}, \hat{\mu}_{\tilde{m}_{\ell-1}}^{\mathcal{I}^\ell}$ that minimizes the Wasserstein distance. Since $W'_{\tilde{m}_\ell} \in \mathcal{M}_{\tilde{m}_\ell, m_{\ell-1}}(\mathbb{R})$ built in the proof of Lemma B.16 has the same law on raws $\mu_\ell$ as $W_A^\ell, W_B^\ell$ but with only $\tilde{m}_\ell$ raws, we can use what preceeds to get:

$$\forall X \in \mathbb{R}^{m_{\ell-1}}, \mathbb{E}[\|W'_{\tilde{m}_\ell} X\|_2^2] \leq \frac{\tilde{m}_\ell}{m_{\ell-1}} \|X\|_2^2$$

Noting that $\|W_{\tilde{m}_\ell} X\|_2^2 = \frac{m_\ell}{\tilde{m}_\ell} \|W'_{\tilde{m}_\ell} X\|_2^2$, we get

$$\forall X \in \mathbb{R}^{m_{\ell-1}}, \mathbb{E}[\|W_{\tilde{m}_\ell} X\|_2^2] \leq \frac{m_\ell}{m_{\ell-1}} \|X\|_2^2$$

which concludes the proof of the lemma. □

Having the two assumptions we need, we can prove Lemma 5.1.

We recall it here:

**Lemma 5.1.** *Under normal initialization of the weights, given $\varepsilon > 0$, if $m_0 \geq 5$, there exists minimal widths $\tilde{m}_1, \dots, \tilde{m}_L$ such that if $m_1 \geq \tilde{m}_1, \dots, m_L \geq \tilde{m}_L$, Property 1 is verified at the last hidden layer $L$ for $\underline{E}_L = 1, E_L = \varepsilon^2$. Moreover, $\forall \ell \in [L], \exists T_\ell$ which does only depend on $L, \ell$ such that one can define recursively $\tilde{m}_\ell$ as $\tilde{m}_0 = m_0$ and*

$$\tilde{m}_\ell = \tilde{\mathcal{O}}\left(\frac{T_\ell}{\varepsilon}\right)^{\tilde{m}_{\ell-1}}$$

*Proof.* From $\mathbb{E}_{x\sim P}[\|x\|_2^2] \leq m_0$ we get immediately Property 1 at the input layer with $\underline{E}_0 = 1, E_0 = 0$.

By the recursive relation of Lemma 4.1 and using Lemmas B.16 and B.17, we get Property 1 at each hidden layer $\ell \in [L]$ with $\tilde{m}_\ell$ to be chosen later with $m_\ell \geq \tilde{m}_\ell \geq \min\{5, E_3^{\tilde{m}_{l-1}}\}$ and:

$$\begin{cases} \underline{E}_\ell = 1 \\ E_\ell = \sum_{i=1}^{\ell} 2^{\ell+1-i} D_3 \log(\tilde{m}_i) \left(\frac{1}{\tilde{m}_i}\right)^{2/\tilde{m}_{i-1}} \underline{E}_{i-1} \end{cases}$$

Therefore, just take $\forall i \in [L], \log(\tilde{m}_i) \left(\frac{1}{\tilde{m}_i}\right)^{\frac{2}{\tilde{m}_{i-1}}} \leq \varepsilon^2 \frac{1}{2^{L+1-i} L}$, i.e.

$$\tilde{m}_i = \tilde{\mathcal{O}}\left(\frac{T_i}{\varepsilon}\right)^{\tilde{m}_{i-1}}$$

where $T_i = \sqrt{2^{L+1-i} L}$ and the notation $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms.

In that case,

$$\begin{cases} \underline{E}_L = 1 \\ E_L = \varepsilon^2 \end{cases}$$

□

### B.8 Proof of Theorem 5.2

We prove here Theorem 5.2 that we recall:

Under normal initialization of the weights, for $m_1 \geq \tilde{m}_1, \dots, m_L \geq \tilde{m}_L$ as defined in Lemma 5.1, $m_0 \geq 5$, and under Assumption 7 we know that $\forall t \in [0,1]$, with $Q$-probability at least $1-\delta_Q$, there exists permutations of hidden layers $1, \dots, L$ of network $B$ that are independent of $t$, such that:

$$\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_{M_t}(x), y\right)\right] \leq t\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_A(x), y\right)\right] + (1-t)\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_B(x), y\right)\right] + \frac{4\sqrt{m_{L+1}}}{\delta_Q^2}\varepsilon$$

*Proof.* Under assumptions of Lemma 5.1, given $A, B$, we know the existence of (random) permutations of the hidden layers $\Pi_1, \dots, \Pi_L$ such that for $1 \leq \ell \leq L$, denoting $M_t$ the mean network of weight matrix at layer $\ell$: $tW_A^\ell + (1-t)\Pi_\ell W_B^\ell \Pi_{\ell-1}^T$ we know the existence of $\underset{\sim}{\phi}^L : \mathbb{R}^{m_0} \to \mathbb{R}^{m_L}$ such that:

$$\mathbb{E}_{P,Q}[\|\phi_{M_t}^L(x) - \phi_A^L(x)\|_2^2] \leq \varepsilon^2 m_L$$
$$\mathbb{E}_{P,Q}[\|\phi_{M_t}^L(x) - \phi_B^L(x)\|_2^2] \leq \varepsilon^2 m_L$$

Then, by convexity, we get at the last layer:

$$\begin{aligned}\mathbb{E}_{P,Q}[\| \left(tW_A^{L+1} + (1-t)W_B^{L+1}\Pi_L^T\right)]\phi_{M_t}^L(x) - tW_A^{L+1}\phi_A^L(x) - (1-t)W_B^{L+1}\Pi_L^T\phi_B^L(x)\|_2^2] \\ \leq t\mathbb{E}[\|W_A^{L+1}(\phi_{M_t}^L(x) - \phi_A^L(x))\|_2^2] \\ + (1-t)\mathbb{E}[\|W_B^{L+1}\Pi_L^T(\phi_{M_t}^L(x) - \phi_B^L(x))\|_2^2]] \\ \leq \varepsilon^2 m_{L+1}\end{aligned}$$

Finally, by Jensen inequality,

$$\mathbb{E}[\| \left(tW_A^{L+1} + (1-t)W_B^{L+1}\Pi_L^T\right)]\phi_{M_t}^L(x) - tW_A^{L+1}\phi_A^L(x) - (1-t)W_B^{L+1}\Pi_L^T\phi_B^L(x)\|_2] \leq \sqrt{m_{L+1}}\varepsilon$$

Therefore, we get by applying two successive Markov lemma, that with probability at least $1-\delta_Q$ over the choice of the networks $A, B$:

$$\mathbb{E}_{x\sim P}[\| \left(tW_A^{L+1} + (1-t)W_B^{L+1}\Pi_L^T\right)]\phi_{M_t}^L(x) - tW_A^{L+1}\phi_A^L(x) - (1-t)W_B^{L+1}\Pi_L^T\phi_B^L(x)\|_2] \leq \frac{\sqrt{m_{L+1}}\varepsilon}{\left(\frac{\delta_Q}{2}\right)^2}$$

Indeed remember that we have introduced an intermediate random measure $\hat{\mu}_{\tilde{m}_\ell}$ which the permutations depend on and that intervenes in the expectation.

Using convexity of the loss and 1-Lipschitzness we get for all $t \in [0,1]$ that with probability at least $1-\delta_Q$ over the choice of the networks $A, B$:

$$\mathbb{E}_{x\sim P}[\mathcal{L}(f_{M_t}(x), y)] \leq t\mathbb{E}[\mathcal{L}(f_A(x), y)] + (1-t)\mathbb{E}[\mathcal{L}(f_B(x), y)] + \frac{4\sqrt{m_{L+1}}\varepsilon}{\delta_Q^2}$$

□

### B.9 Approximately low dimensional underlying weights distribution

#### B.9.1 Motivation on the structure of the covariance matrix

Remind that we are given a partition of the input layer of the weights $[m_{\ell-1}]$ in $\tilde{m}_{\ell-1}$ different groups of the same size $\mathcal{I}^{\ell-1} = \{I_1^{\ell-1}, ..., I_{\tilde{m}_{\ell-1}}^{\ell-1}\}$. Suppose we have already permuted the first layer of network $A$ and $B$ we can suppose that $I_1^{\ell-1} = \{1, \dots, p_{\ell-1}\}, \dots$ where $p_{\ell-1} := \frac{m_{\ell-1}}{\tilde{m}_{\ell-1}}$. We want a covariance matrix that respects the

fact that incoming neurons in a given group behave the same. Therefore the covariance matrix must be invariant under the permutations of indices inside a set of the equi-partition. We will write the Kroenecker product $\otimes$.

**Lemma B.18.** *To respect symmetries of the incoming layer, the covariance matrix of weights is necessarily of the form*

$$\Sigma_\ell = D_\ell \otimes I_{p_{\ell-1}} + B_\ell \otimes \mathbb{1}_{p_{\ell-1}}$$

*where $D_\ell \in \mathcal{M}_{\tilde{m}_{\ell-1}}(\mathbb{R})$ is diagonal, $B_\ell \in \mathcal{S}_{\tilde{m}_{\ell-1}}(\mathbb{R})$ is symmetric.*

*Proof.* Denoting the matrix of covariances by blocks like:

$$\Sigma_\ell = \begin{pmatrix} A_{11} & A_{12} & ... & A_{1\tilde{m}_{\ell-1}} \\ A_{21} & A_{22} & ... & A_{2\tilde{m}_{\ell-1}} \\ ... & ... & ... & ... \\ A_{\tilde{m}_{\ell-1}1} & A_{\tilde{m}_{\ell-1}2} & ... & A_{\tilde{m}_{\ell-1}\tilde{m}_{\ell-1}} \end{pmatrix}$$

we get the relation $\forall \Pi_1, ..., \Pi_{\tilde{m}_{\ell-1}} \in \mathcal{S}_{p_{\ell-1}}$ permutation matrices, denoting $\Pi = \mathrm{Diag}(\Pi_1, \dots, \Pi_{\tilde{m}_{\ell-1}})$

$$\Sigma_\ell = \mathbb{E}\left[XX^T\right] = \mathbb{E}\left[(\Pi X)(\Pi X)^T\right] = \Pi\Sigma\Pi^T = \mathrm{Diag}(\Pi_1, ..., \Pi_{\tilde{m}_{\ell-1}})\Sigma\,\mathrm{Diag}(\Pi_1^T, ..., \Pi_{\tilde{m}_{\ell-1}}^T)$$

Evaluating this relation for any $\Pi_1, \Pi_2, ..., \Pi_{\tilde{m}_{\ell-1}}$ we get $\forall \Pi_1 \in \mathcal{S}_{p_{\ell-1}}, \Pi_1 A_{11} \Pi_1^T = A_{11}$ and therefore $A_{11}$ is of the form $d_{11} I_{p_{\ell-1}} + b_{11}\mathbb{1}_{p_{\ell-1}}$.

We do the same for $A_{ii}, i \geq 2$.

Moreover we get for $A_{12}$ that:

$$\forall \Pi_1, \Pi_2 \in \mathcal{S}_{p_{\ell-1}}, \Pi_1 A_{12} \Pi_2^T = A_{12}$$

which brings that $A_{12}$ is of the form $b_{12}\mathbb{1}_{p_{\ell-1}}$. We do the same for all $A_{ij}$ where $i \neq j$. This concludes the proof.

Finally by summing over columns inside the partition we get:

$$\Sigma_\ell^{\mathcal{I}^{\ell-1}} = p_{\ell-1} D_\ell + p_{\ell-1}^2 B_\ell$$

□

The model that we chose in Section 5.4 is a particular case that corresponds to choosing $D_\ell = \mathrm{Diag}(\lambda_1^\ell, ..., \lambda_{\tilde{m}_{\ell-1}}^\ell)$ and $B_\ell = 0$. This is not the most general since it implies independence between weights coming from different groups but is sufficient to show the influence of low feature dimensionality on LMC efficiency.

### B.9.2 Non diagonal model

A natural direction is to consider the case where the matrix $B_\ell$ is non zero. Since $\Sigma_\ell^{\mathcal{I}^{\ell-1}}$ is symmetric positive, one can orthogonally change the basis where it becomes diagonal. The arguments to prove Assumption 3 remains unchanged since they rely exclusively on the eigenvalues of $\Sigma_\ell^{\mathcal{I}^{\ell-1}}$.

However one important point to check, is about Assumption 4. Indeed having covariance between weights of a given block implies that the errors at all neurons of a given layer may sum up. Take for example $X = (1, \dots, 1)^T, D_\ell = 0, B_\ell = \mathbb{1}_{\tilde{m}_{\ell-1}}$. In the general case, the constant $C_2$ in Assumption 4 will be a depending on the matrix $B_\ell$ and therefore potentially on the dimension. We propose a way to address this issue in Appendix B.12.

## B.10 Proof of Theorem 5.4

**Lemma B.19** (Version of Assumption 3 for approximately low dimensional distribution)**.** *Denote $\mu_\ell$ the law of a multivariate normal distribution of covariance matrix $\mathrm{Diag}(\lambda_1^\ell I_{p_{\ell-1}}, \dots, \lambda_{\tilde{m}_{\ell-1}}^\ell I_{p_{\ell-1}})$ where $p_{\ell-1} = \frac{m_{\ell-1}}{\tilde{m}_{\ell-1}}$ and $\frac{1}{k_{\ell-1}} \geq \lambda_1^\ell \dots \geq \lambda_{\tilde{m}_{\ell-1}}^\ell$. Let $k_{\ell-1} \geq 5$ and denote $\eta := \frac{\sqrt{\sum_{i=k_{\ell-1}+1}^{m} \lambda_i^\ell}}{4\sqrt{\sum_{i=1}^{k_{\ell-1}} \lambda_i^\ell}}$.*

*There exists two universal constants $D_3', E_3'$ such that $\forall \tilde{m}_\ell \geq 1$ such that $E_3'^{k_{\ell-1}} \leq \tilde{m}_\ell \leq \eta^{-k_{\ell-1}}$, there exists a random empirical measure $\hat{\mu}_{\tilde{m}_\ell}$ with only $\tilde{m}_\ell$ points such that $\forall m_\ell \geq 1$ such that $\tilde{m}_\ell \leq m_\ell \leq \eta^{-k_{\ell-1}}$ we have:*

$$\mathbb{E}[\mathcal{W}_2^2(\hat{\mu}_A^{\mathcal{I}^{\ell-1}}, \hat{\mu}_{\tilde{m}_\ell}^{\mathcal{I}^{\ell-1}})] \leq \frac{D_3'}{k_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/k_{\ell-1}}$$

*Proof.* We do exactly the same as for the proof of Lemma B.16, i.e. a triangular inequality but now we use rate of convergence of empirical measures in Wasserstein distance with approximately low dimensional support as expressed in Lemma B.15. □

**Lemma B.20** (Version of Assumption 4 for approximately low dimensional distribution)**.** *Denote $\mu_\ell$ the law of a multivariate normal distribution of covariance matrix $\mathrm{Diag}(\lambda_1^\ell I_{p_{\ell-1}}, \dots, \lambda_{\tilde{m}_{\ell-1}}^\ell I_{p_{\ell-1}})$ where $p_{\ell-1} = \frac{m_{\ell-1}}{\tilde{m}_{\ell-1}}$ and $\frac{1}{k_{\ell-1}} \geq \lambda_1^\ell \dots \geq \lambda_{\tilde{m}_{\ell-1}}^\ell$. $\forall X \in \mathbb{R}^{m_{\ell-1}}$ we have:*

$$\mathbb{E}[\|W_A^\ell X\|_2^2] \leq \frac{\tilde{m}_{\ell-1}}{k_{\ell-1}} \|X\|_2^2$$
$$\mathbb{E}[\|W_{\tilde{m}_\ell}^\ell X\|_2^2] \leq \frac{\tilde{m}_{\ell-1}}{k_{\ell-1}} \|X\|_2^2$$

*Proof.* Just notice that $\lambda_1 \leq \frac{1}{k_{\ell-1}}$ and repeat the same steps as in the proof of Lemma B.17. □

We will now re-state and prove Theorem 5.4:

**Theorem B.21.** *Under Assumptions 7 and 8, given $\varepsilon > 0$, if $em_0 \geq 5$ there exists minimal widths $\tilde{m}_1, \dots, \tilde{m}_L$ such that if $\eta^{-k_0} \geq m_1 \geq \tilde{m}_1, \dots, \eta^{-k_{L-1}} \geq m_L \geq \tilde{m}_L$, Property 1 is verified at the last hidden layer $L$ for $\underline{E}_L = 1, E_L = \varepsilon^2$. Moreover, $\forall \ell \in [L]$, $\exists T_\ell'$ which does only depend on $L, e, \ell$, such that one can define recursively $\tilde{m}_\ell$ as*

$$\tilde{m}_\ell = \tilde{\mathcal{O}}\left(\frac{T_\ell'}{\varepsilon}\right)^{k_{\ell-1}} = \tilde{\mathcal{O}}\left(\frac{T_\ell'}{\varepsilon}\right)^{e\tilde{m}_{\ell-1}}$$

*where $\tilde{m}_0 = m_0$. Moreover $\forall t \in [0,1]$, with $Q$-probability at least $1 - \delta_Q$, there exists permutations of hidden layers $1, \dots, L$ of network $B$ s.t.,*

$$\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_{M_t}(x), y\right)\right] \leq t\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_A(x), y\right)\right] + (1-t)\mathbb{E}_P\left[\mathcal{L}\left(\hat{f}_B(x), y\right)\right] + \frac{4\sqrt{m_{L+1}}}{\sqrt{e}\delta_Q^2}\varepsilon$$

*Proof.* We just need to prove the first part of the theorem as proving the similarity of loss is exactly the same as in the proof of Theorem 5.2 when we have proved that Property 1 holds at layer $L$ with $E_L = \varepsilon^2$. The only change comes from the constant $C_2$ in Lemma B.20 which is not 1 anymore but $\frac{1}{e}$, hence the additional factor $e$.

To prove Property 1 at layer $L$ with $E_L = \varepsilon^2$ we just combine $L$ different times the two previous lemma Lemmas B.19 and B.20 and Lemma 4.1.

Lemma B.20 brings that at layer $\ell$ the constant $C_2$ is $\frac{\tilde{m}_{\ell-1}}{k_{\ell-1}} \leq \frac{1}{e}$ by Assumption 8

Moreover, Lemma B.19 brings that at each layer $\ell$, $C_1 = \frac{D_3'}{k_{\ell-1}} \log(\tilde{m}_\ell) \left(\frac{1}{\tilde{m}_\ell}\right)^{2/k_{\ell-1}}$

It brings:

$$\begin{cases} \underline{E}_{i+1} = \frac{1}{e}\underline{E}_i \\ E_{i+1} = \frac{2}{e}E_i + 2\frac{D_3'}{k_i}\log(\tilde{m}_{i+1})\left(\frac{1}{\tilde{m}_{i+1}}\right)^{2/k_i}\tilde{m}_i\underline{E}_i \end{cases}$$

From there we see that if we have chosen at each layer

$$D_3' \log(\tilde{m}_{i+1}) \left(\frac{1}{\tilde{m}_{i+1}}\right)^{2/k_i} = \frac{\sqrt{\frac{L2^{L-i+1}}{e^L}}}{\varepsilon^2}$$

if

$$\tilde{m}_i = \tilde{\mathcal{O}}\left(\frac{T_i'}{\varepsilon}\right)^{e\tilde{m}_{i-1}}$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms and we define $T_i' = \sqrt{\frac{D_3' e^L}{L2^{L-i+1}}}$

□

### B.11 Proof of Theorem 5.3

**Theorem 5.3.** *Let $n \geq 1, x \sim P \in \mathcal{P}_1(\mathbb{R}^n)$ and $\mu \in \mathcal{P}(\mathbb{R}^n)$ such that $\frac{d\mu}{dLeb} \leq F_1$. Suppose $\Sigma = \mathbb{E}[xx^T]$ is full rank $n$. Let $m \geq 1$ and $W_A, W_B \in \mathcal{M}_{m,n}(\mathbb{R})$ whose rows are drawn i.i.d. from $\mu$. Then, there exists $F_0$ such that*

$$\mathbb{E}_{W_A,W_B}[\min_{\Pi \in \mathcal{S}_m} \mathbb{E}_P \|(W_A - \Pi W_B)x\|_2^2] \geq F_0 \left(\frac{1}{m}\right)^{2/n}$$

*Proof.* First since $\Sigma$ is full rank we see that when writing $\Sigma = ODO^T$ where $OO^T = I_{\tilde{n}}$, the problem is equivalent to consider the matrices $W_A O, W_B O$ and $\Sigma = D$. In that case, the raws of $W_A, W_B$ are still i.i.d. and follow the same law as $\mu$ modulo a non-degenerated dilatation. We can then still assume $\frac{d\mu}{d\text{Leb}} \leq F_1$ for a certain constant $F_1$.

Let $\tau = \frac{1}{2}$. $\forall S \subset \mathbb{R}^n$ a Borel set such that $\mu(S) \geq 1 - \tau$, we know that $\text{Leb}(S) \geq \frac{1-\tau}{F_1}$. In that case, applying Lemma B.10 we get $\mathcal{N}_\varepsilon(S) \geq \left(\frac{1}{\varepsilon}\right)^n \frac{\frac{1-\tau}{F_1}}{\text{Leb}(\mathcal{B}_2(0,1))}$. Denote $F_2 = \left(\frac{\frac{1-\tau}{F_1}}{\text{Leb}(\mathcal{B}_2(0,\varepsilon)}\right)^{1/n}$. Using notations of Weed and Bach [2019] we get $\mathcal{N}_\varepsilon(\mu, \tau) \geq \left(\frac{F_2}{\varepsilon}\right)^n$.

Applying Proposition 6 from Weed and Bach [2019] we get that

$$\mathcal{W}_2^2(\hat{\mu}_A, \mu) \geq F_3 \left(\frac{1}{m}\right)^{2/n}$$

Finally noticing that $\mathbb{E}_{W_B}[\hat{\mu}_B] = \mu$ and applying Lemma B.2, we get that:

$$\mathbb{E}_{W_A,W_B}[\mathcal{W}_2^2(\hat{\mu}_A, \hat{\mu}_B)] \geq \mathbb{E}_{W_A}[\mathcal{W}_2^2(\hat{\mu}_A, \mathbb{E}_{W_B}[\hat{\mu}_B])] \geq F_3 \left(\frac{1}{m}\right)^{2/n}$$

Finally, remark that sinc $\Sigma = D = \text{Diag}(\lambda_1, \ldots, \lambda_{\tilde{n}})$ and noting $\lambda(\Sigma) = \min\{d_i, 1 \leq i \leq n\} > 0$ the smallest eigenvalue of $\Sigma$,

$$\begin{aligned}
\mathbb{E}_{W_A,W_B}[\min_{\Pi \in \mathcal{S}_m} \mathbb{E}_{x\sim P} \|(W_A - \Pi W_B)x\|_2^2] &\leq \mathbb{E}_{W_A,W_B}[\min_{\Pi \in \mathcal{S}_m} \mathbb{E}_{x\sim P} \text{tr}((W_A - \Pi W_B)^T (W_A - \Pi W_B) D)] \\
&\geq \lambda(\Sigma) \mathbb{E}_{W_A,W_B}[\min_{\Pi \in \mathcal{S}_m} \mathbb{E}_{x\sim P} \|(W_A - \Pi W_B)\|_2^2] \\
&\geq \lambda(\Sigma) F_3 \left(\frac{1}{m}\right)^{2/n}
\end{aligned}$$

□

### B.12 Discussion about a model with no growth in the width needed

For proving Theorem 5.4, we have used constants $C_1, C_2$ in Lemma 4.1 given by Lemmas B.19 and B.20. However we would like to highlight that Lemma B.20 is very sub-optimal, though it can not really be improved in the general case. Indeed, $\frac{1}{e^\ell}$ grows to infinity while $\mathbb{E}[\|\phi^\ell(x)\|_2^2]$ is supposed to remain bounded for $1 \leq \ell$. Therefore from now on suppose that we have the following version of Lemma B.22:

**Lemma B.22** (Extending Assumption 4 for approximately low dimensional distribution)**.** *Denote $\mu_\ell$ the law of a multivariate normal distribution of covariance matrix* $\mathrm{Diag}(\lambda_1^\ell I_{p_{\ell-1}}, \dots, \lambda_{\tilde{m}_{\ell-1}}^\ell I_{p_{\ell-1}})$ *where $p_{\ell-1} = \frac{m_{\ell-1}}{\tilde{m}_{\ell-1}}$ and $\frac{1}{k_{\ell-1}} \geq \lambda_1^1 \dots \geq \lambda_{\tilde{m}_{\ell-1}}^\ell$. Suppose we have:*

$$\mathbb{E}[\|W_A(\phi_B^{\ell-1}(x) - \phi_A^{\ell-1}(x)\|_2^2] \leq \frac{m_\ell}{m_{\ell-1}} \mathbb{E}[\|\phi_B^{\ell-1}(x) - \phi_A^{\ell-1}(x)]\|_2^2$$

Notice that it is equivalent to making an assumption on the distribution of $\phi_B^{\ell-1}(x) - \phi_A^{\ell-1}(x)$ which must not put too much mass on the worst case coordinates ($\lambda_i^\ell$ for $i$ small).

In that case, adapting the proof as in Theorem 5.4, we could get an inequality on $\tilde{m}_i$ of the form

$$\tilde{m}_i = \tilde{\mathcal{O}}\left(\frac{T_i''}{\varepsilon}\right)^{k_{i-1}}$$

where $T_i''$ is independent of $e$ which lead to controlled bounds as $L \to \infty$ (without the exponent $e\tilde{m}_i$ as in Theorem 5.4).

### B.13 Extension of Theorems 5.2 and 5.4 to sub-Gaussian variables

Still under the setting of Assumption 6, suppose now that at a given layer $\ell$, all the parameters of $W_A^\ell$ are still drawn independently but no longer from $\mathcal{N}(0, \frac{1}{m_{\ell-1}})$ Instead we assume that the underlying distribution $\mu_\ell$ verifies for each layer $\ell \in [L+1]$: if $X \sim \mu_\ell$ then, $\forall j \neq k \in [m_{l-1}], X_j \amalg X_k$. Moreover $\forall i \in [\tilde{m}_{\ell-1}], \forall j, k \in I_i^{\ell-1}$,

$$\mathbb{E}[X_j^2] = \mathbb{E}[X_k^2] = \lambda_i^\ell$$

Finally suppose the variables are sub-Gaussian i.e., $\exists K > 0, \forall i \in [\tilde{m}_{\ell-1}], \forall j \in I_i^{\ell-1}$, $\forall c > 0$,

$$\mathbb{P}(|X_j| \geq c) \leq 2\exp(-\frac{c^2}{K\lambda_i^\ell})$$

Further suppose that we are in the setting of Theorem 5.2 (The case of Theorem 5.4 is treated similarly): $\frac{1}{m_{\ell-1}} \geq \lambda_1^\ell \geq \dots \lambda_{\tilde{m}_{\ell-1}^\ell}$.

It is clear that Lemma B.17 is still valid for a constant $C_2 = 1$, the proof being exactly the same.

We therefore just need to prove Lemma B.16 for $C_1$ to be determined. To prove Lemma B.16, one just needs an equivalent of Lemma B.14 for sub-Gaussian variables. To prove Lemma B.14, recall that we have used the fact that a normal distribution doesn't put to much mass outside of a ball of radius $c$ when c grows logarithmically. More precisely we have used the property, that if $X \sim \mathcal{N}(0, \frac{I_{m_{l-1}}}{m_{l-1}})$, then:

$$\forall c > 1, \mathbb{P}(\|X\|_2^2 \geq c^2) \leq e^{-\frac{c^2}{4}}$$

In our case, for a sub-Gaussian distribution, we know the existence of a constant $K$ such that $\forall c > 0$:

$$\mathbb{P}(|X_j| \geq c) \leq 2\exp(-\frac{c^2}{K\lambda_i^\ell})$$

Therefore, by plugging this into the proof, and scaling parameter $c$ by $\sqrt{K}$, we get exactly the same version of Lemma B.16 for sub-Gaussian variable, with different constants scaled by a factor $\sqrt{K}$.

Propagating Property 1 with the recurrence formula of Lemma 4.1 we get LMC for networks with sub-Gaussian distributions in the same form as for normal variables.

**Remark** Finally notice that results for sub-Gaussian variables can be extended in the same way to variables whose tail decreses sufficiently fast (exponentially, polynomially, etc...). The assymptotics of the tail will affect the convergence rate in Wasserstein of the corresponding empirical measure.

### B.14 Link with dropout stability

We relate now our previous study to a line of work exploring mode connectivity through dropout stability.

Kuditipudi et al. [2019] define $\varepsilon$-dropout stable networks, as networks $\hat{f}(\cdot;\theta)$ as defined in Equation (1) for which there exists in each layer $\ell \in [L]$, a subset of at most $\frac{m_\ell}{2}$ of neurons (i.e., rows of the weight matrix $W^\ell$) such that after renormalizing each layer, the expected loss of the new network increases by no more than $\varepsilon$ with respect to the original loss. Kuditipudi et al. [2019] shows that two $\varepsilon$-dropout stable networks are mode connected (with error barrier height $\varepsilon$) and Shevchenko and Mondelli [2020] uses this result to show that two wide enough two-layer neural networks trained with SGD are mode connected (where the continuous path may be non-linear).

Recall that we have shown in Section 3 the stronger statements that two such networks are in the same local minima modulo permutation symmetries. However, note that Shevchenko and Mondelli [2020] don't allow permutations of neurons). We discuss here how to embrace in the same view our framework with dropout stability results, showing how networks with independent neuron's weights become dropout stable in the same asymptotics of large width than the condition of Lemma 5.1.

Consider the simplified setting of a 1-hidden layer neural network with 1-Lipschitz activation where the weights of the second layer are fixed to $\frac{1}{N}$: $\hat{f}(x;\theta) = \frac{1}{N}\sum_{i=1}^{N}\sigma(w_i x)$ where $w_i = W_{i:} \in \mathbb{R}^d$ is the $i-th$ row of the weight matrix $W$. Suppose that $w_i$ are sampled independently from a sub-Gaussian distribution and the data follows a distribution $(x,y) \sim P$ with $\mathrm{Supp}(P) \subset \mathcal{B}_2(0,1)$. Denote $\mathcal{A} = \left[\frac{N}{2}\right]$. Dropout stability can be quantified by controlling the error between the correctly renormalized network with weights in $\mathcal{A}$ and the original one,

$$\mathbb{E}\left[\left|\frac{2}{N}\sum_{i\in\mathcal{A}}\sigma(w_i x) - \frac{1}{N}\sum_{i=1}^{N}\sigma(w_i x)\right|\right] \leq \mathcal{W}_1(W^{\mathcal{A}}, W^{\mathcal{A}^c}) \tag{8}$$

where we have denoted $W^{\mathcal{A}}$ (respectively $W^{\mathcal{A}^c}$) the matrix $W$ where we have kept only the rows in $\mathcal{A}$ (respectively $(\mathcal{A}^c)$). The right hand term can be connected to convergence rates of empirical measure (Lemma B.16 and the extension to sub-Gaussian distribution discussed in Appendix B.13):

$$\mathcal{W}_1(W^{\mathcal{A}}, W^{\mathcal{A}^c}) \approx \left(\frac{1}{N}\right)^{1/d}$$

In a nutshell, showing that previous Equation (8) is tight would provide a formal connection between dropout stability and our results. It is an interesting direction for future work and note that it has strong connections with the dual expression of the Wasserstein 1 distance.

In that case, the bound on the dropout error evolves as $\left(\frac{1}{N}\right)^{1/d}$, as for the linear mode connectivity error. Hence networks become dropout stable in the same asymptotics as to exhibit linear mode connectivity.

This is consistent with the idea that LMC requires the information to be distributed evenly among neurons without any neuron responsible for the particular behavior of one layer. This is similar to the intuitive requirement for dropout stability.

# C PROOF OF LMC FOR TWO-LAYER NEURAL NETWORKS IN THE MEAN FIELD REGIME

## C.1 Description of the mean field regime

When training a two-layer neural network with fixed input and output dimensions but with a very wide hidden layer using SGD, the parameters of each neuron can be seen as particles evolving independently one from each other: the dynamic of each neuron's weights depends only on the average distribution of the weights and itself.

The main object of study is therefore the empirical distribution of the neurons weights in the intermediate layer after $k$ Stochastic Gradient Descent (SGD) steps. We denote it $\rho_k^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i^k}$ where $\theta_i^k = (w_i^k, a_i^k) \in \mathbb{R}^{d+1}$.

Multiple works [Chizat and Bach, 2018, Mei et al., 2018, 2019] show that if the hidden layer's width $N$ is big, the learning rate $s_k$ is small and setting $T = \sum_{i=1}^k s_i$ a time re-normalization after $k$ steps, then $\rho_k^{(N)}$ can be well approximated by $\rho_t$ which follows the following partial differential equation (PDE):

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta;\rho_t)),\ \Psi(\theta;\rho_t) = V(\theta) + \int U(\theta,\tilde{\theta})\rho_t(d\tilde{\theta})$$
$$V(\theta) = -\mathbb{E}[y\sigma_*(x;\theta)],\ U(\theta_1,\theta_2) = \mathbb{E}[\sigma_*(x;\theta_1)\sigma_*(x;\theta_2)]$$

Here $\xi(t)$ represent a scaling of the learning rate where $s_k = \varepsilon\xi(k\varepsilon)$. $U(\theta_1,\theta_2)$ represents a correlation between neurons. $V(\theta)$ is an energy quantifying the alignment of a neuron function with the data. In the following, as in Mei et al. [2019] we will work with $\xi = \frac{1}{2}$ a constant step size function. As in Mei et al. [2019], we highlight that the proof remains valid under Assumption 1.

When considering noisy SGD, the limit PDE becomes:

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t(\theta)\nabla_\theta \Psi_\lambda(\theta;\rho_t)) + 2\xi(t)\tau d^{-1}\Delta_\theta \rho_t$$
$$\Psi_\lambda(\theta;\rho) = \Psi(\theta;\rho) + \frac{\lambda}{2}\|\theta\|_2^2$$

The crucial point about the mean field view is to show that the empirical distribution of parameters is well enough approximated by $\rho_t$. Then the study of the neural network can be reduced to the study of the partial differential equation. For example global convergence of the test loss results can be deduced as in Chizat and Bach [2018]. This view is also convenient to get insights of typical behaviours of the dynamics while smoothing the effects of local minima [Mei et al., 2019]. In our case, the mean field view allows us to use convergence results in Wasserstein distance of empirical measures towards the underlying distribution. We can then show Linear Mode Connectivity for two-layer neural networks independently trained in the mean field regime.

## C.2 Proving LMC for noiseless regularization-free SGD

To prove our results we have to show that the empirical distribution of weights can be well approximated by the solution of the mean field PDE. To achieve this, Mei et al. [2019] introduce four intermediate dynamics that stay close one of each other.

First note that our *Assumption* 1 implies Assumptions 1 to 4 of Mei et al. [2019]. Especially the non-linearity being Lipschitz implies its gradient distribution on the data is bounded and hence sub-Gaussian.

### C.2.1 Intermediate dynamics

Mei et al. [2019] introduce 4 different intermediate dynamics between the empirical distribution of weights optimized by SGD and the solution of the PDE that we recall here:

**Nonlinear dynamics**

Let consider $\bar{\theta}_i^t$ with initialization $\bar{\theta}_i^0 \sim \rho_0$ i.i.d. and which follows the dynamics

$$\bar{\theta}_i^t = \bar{\theta}_i^0 + 2\int_0^t \xi(s) G(\bar{\theta}_i^s; \rho_s) ds$$

or equivalently

$$\frac{d}{dt}\bar{\theta}_i^t = -2\xi(t)\left[\nabla V(\bar{\theta}_i^t) + \int \nabla_1 U(\bar{\theta}_i^t, \theta)\rho_t(d\theta)\right]$$

where $G(\theta;\rho) = -\nabla\Psi(\theta;\rho)$. An important fact is that $\bar{\theta}_i^t$ is random because of the random initialization. Moreover its law at time $t$ is $\rho_t$. It corresponds to the evolution of particles under a velocity field $-2\xi(t)\left[\nabla V(\bar{\theta}_i^t) + \int \nabla_1 U(\bar{\theta}_i^t, \theta)\rho_t(d\theta)\right]$ which depends only on the position of the optimized particle and the overall distribution of all particles.

**Particle Dynamics**

Let $\underline{\theta}_i^t$ have the same initialization as the nonlinear dynamics $\underline{\theta}_i^0 = \bar{\theta}_i^0$, and $\underline{\rho}_t^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\underline{\theta}_i^t}$ denote the empirical distribution of $\underline{\theta}_i^t$. Then the particle dynamics is given by:

$$\underline{\theta}_i^t = \underline{\theta}_i^0 + 2\int_0^t \xi(s) G(\underline{\theta}_i^s; \underline{\rho}_s^{(N)}) ds$$

or equivalently

$$\frac{d}{dt}\underline{\theta}_i^t = -2\xi(t)\left[\nabla V(\underline{\theta}_i^t) + \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\underline{\theta}_i^t, \underline{\theta}_j^t)\right]$$

**Gradient descent dynamics**

Let $\tilde{\theta}_i^k$ with initialization $\tilde{\theta}_i^0 = \bar{\theta}_i^0$ following the dynamics:

$$\tilde{\theta}_i^k = \tilde{\theta}_i^0 + 2\varepsilon\sum_{l=0}^{k-1} \xi(l\varepsilon) G(\tilde{\theta}_i^l; \tilde{\rho}_l^{(N)})$$

or equivalently:

$$\tilde{\theta}_i^{k+1} = \tilde{\theta}_i^k - 2s_k\left[\nabla V(\tilde{\theta}_i^k) + \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\tilde{\theta}_i^k, \tilde{\theta}_j^k)\right]$$

**Stochastic Gradient Descent Dynamics**

Consider $\theta_i^k$ with initialization $\theta_i^0 = \bar{\theta}_i^0$ following the dynamics:

$$\theta_i^k = \theta_i^0 + 2\varepsilon\sum_{l=0}^{k-1} \xi(l\varepsilon) F_i(\theta^l; z_{l+1})$$

or equivalently

$$\theta_i^{k+1} = \theta_i^k - 2s_k F_i(\theta^k; z_{k+1})$$

where $F_i(\theta^k; z_{k+1}) = (y_{k+1} - \hat{y}_{k+1})\nabla_\theta \sigma_*(x_{k+1}; \theta_i^k)$, $z_k = (x_k, y_k)$ and $\hat{y}_{k+1} = \frac{1}{N}\sum_{j=1}^N \sigma_*(x_{k+1}; \theta_j^k)$.

One can use Proposition 26,28,29 from Mei et al. [2019] to show the following lemma:

**Lemma C.1.** *Consider a two-layer neural network trained with noiseless regularization-free SGD for an underlying time $T$. Then under Assumption 1, there exists constants $K$ and $K_0$ such that, if $\varepsilon \leq \min\left\{\frac{1}{K_0 e^{K_0(1+T)^3}}, \frac{1}{K_0(d+\log(N)+z^2)e^{K_0(1+T)^3}}\right\}$, then with probability at least $1-3e^{-z^2}$ we have:*

$$\max_{k\in[0,T/\varepsilon]\bigcap\mathbb{N}} \max_{i\in[N]} \|\theta_i^k - \bar{\theta}_i^{k\varepsilon}\|_2 \leq Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)}+z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}[\sqrt{d+\log(N)}+z]$$

*Proof.* This is direct application from Proposition 26,28,29 from Mei et al. [2019] by doing two union bounds and two triangular inequalities. □

We moreover recall that as highlighted in Mei et al. [2019], $\{\bar{\theta}_i^t, 1\leq i\leq N\}$ are independent from each other and each follows the distribution $\rho_t$ when initialized i.i.d. as $\rho_0$. Therefore, when considering two two-layer neural networks initialized randomly as $\rho_0$ and trained for the same underlying time $T$ with noiseless regularization-free SGD, we know from the previous lemma that the parameters of both networks are close to two samples from $\rho_t$.

### C.2.2 Proof of Theorem 3.1 in the case of noiseless regularization-free SGD

We now prove Theorem 3.1 in case of noiseless regularization-free SGD.

**Theorem 3.1.** *Consider two two-layer neural networks as in Equation (3) trained with equation SGD with the same initialization over the weights independently and for the same underlying time $T$. Suppose Assumptions 1 and 2 to hold. Then $\forall\delta, err, \exists N_{min}$ such that if $N\geq N_{min}, \exists\varepsilon_{max}(N)$ such that if $\varepsilon\leq\varepsilon_{max}(N)$ in Equation (4), then with probability at least $1-\delta$ over the training process, there exists a permutation of the second network's hidden layer such that for almost every $x\sim P$:*

$$|t\hat{f}_N(x;\theta_A) + (1-t)\hat{f}_N(x;\theta_B) - \hat{f}_N(x; t\theta_A + (1-t)\tilde{\theta}_B)| \leq err, \quad \forall t\in[0,1].$$

*Proof.* We know from Lemma C.1 that with probability at least $1-3e^{-z^2}$, if $\varepsilon \leq \min\left\{\frac{1}{K_0 e^{K_0(1+T)^3}}, \frac{1}{K_0(d+\log(N)+z^2)e^{K_0(1+T)^3}}\right\}$,

$$\max_{k\in[0,T/\varepsilon]\bigcap\mathbb{N}} \max_{i\in[N]} \|\theta_{A,i}^k - \bar{\theta}_{A,i}^{k\varepsilon}\|_2 \leq Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)}+z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}[\sqrt{d+\log(N)}+z]$$

which means that $\theta_{A,i}$ is close to the non linear dynamics which are samples from $\rho_t$. By a union bound, with probability $1-6e^{-z^2}$ this is true for both networks $A$ and $B$.

Denoting as before $\theta_{A,i} = (w_{A,i}, a_{A,i}) \in \mathbb{R}^{d+1}$ (respectively $\theta_{B,i}$), $A_A = (a_{A,1},\dots,a_{A,N})$ (respectively $A_B$) and $W_A \in \mathcal{M}_{N,d}(\mathbb{R})$ the concatenation of vectors $w_{A,i}\in\mathbb{R}^d$ (respectively $W_B$). Given $t\in[0,1]$, we aim at finding a permutation $\Pi\in\mathcal{S}_N$ of the second network's hidden layer to get $\tilde{\theta}_B = (\tilde{A}_B, \tilde{W}_B) = (A_B\Pi^T, \Pi W_B)$ bounding

$$\begin{aligned}
&|t\hat{f}(x;\theta_A) + (1-t)\hat{f}(x;\theta_B) - \hat{f}(x; t\theta_A+(1-t)\tilde{\theta}_B)| \\
&= \frac{1}{N}|tA_A\sigma(W_AX) + (1-t)\tilde{A}_B\sigma(\tilde{W}_BX) - (tA_A+(1-t)\tilde{A}_B)\sigma((tW_A+(1-t)\tilde{W}_B)X)| \qquad (9)\\
&\leq t\left|\frac{A_A}{N}(\sigma(W_AX) - \sigma((tW_A+(1-t)\tilde{W}_B)X))\right| + (1-t)\left|\frac{\tilde{A}_B}{N}(\sigma(\tilde{W}_BX) - \sigma((tW_A+(1-t)\tilde{W}_B)X))\right| \\
&\leq t\|A_A\|_\infty\frac{\|\sigma(W_AX) - \sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N} + (1-t)\|\tilde{A}_B\|_\infty\frac{\|\sigma(\tilde{W}_BX) - \sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N}
\end{aligned}$$

Both terms $\frac{\|\sigma(W_A X)-\sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N}$ and $\frac{\|\sigma(\tilde{W}_B X)-\sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N}$ can be bounded.

Indeed, first using lemma 22 from Mei et al. [2019] and that from Assumption 1 $\text{Supp}(\rho_0)$ is bounded, we get that

$$\text{Supp}(\rho_t) \subset \mathcal{B}_2(0, K((1+T^2)T)+1)$$

is bounded with a diameter depending only on the initialization $\text{Supp}(\rho_0)$ and underlying time $T$.

Therefore we can apply Theorem 1 from Weed and Bach [2019] and get for $s > d$ the existence of a constant $C$ such that with probability at least $1-\frac{\delta}{2}$, there exists a permutation $\gamma \in \mathcal{S}_N$ such that by considering $\|\cdot\|_1$ as a distance for the Wasserstein:

$$\mathcal{W}_1(\hat{\mu}_A, \hat{\mu}_B) = \frac{1}{N}\sum_{i=1}^{N} \|\bar{\theta}_{A,i} - \bar{\theta}_{B,\gamma_i}\|_1 \leq \frac{C}{\delta} N^{-1/s}$$

Note that, while $C$ is independent of $N$, it depends on the distribution $\rho_t$ and therefore on $d$, $\text{diam}(\text{Supp}(\rho_t))$ (i.e., $T$) and on the constants from Assumption 1.

Recall that we suppose the data distribution P bounded and denote $\text{Supp}(P) \subset [-H_x, H_x]^d \times [-H_y, H_y]$.

Therefore, we get that with probability at least $1-\frac{\delta}{2}-6e^{-z^2}$:

$$\begin{aligned}
&\forall X \in [-H_x, H_x]^d, \frac{\|\sigma(W_A X) - \sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N} \leq g_2(T,z,\delta,N,\varepsilon) \qquad (10)\\
&:= L_\sigma\left(H_x \frac{C}{\delta} N^{-\frac{1}{s}} + 2H_x\sqrt{d}\left[\frac{Ke^{K(1+T)^3}}{\sqrt{N}}[\sqrt{\log(NT)}+z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}[\sqrt{d+\log(N)}+z]\right]\right)
\end{aligned}$$

and same for the other term:

$$\forall X \in [-H_x, H_x]^d, \frac{\|\sigma(\tilde{W}_B X) - \sigma((tW_A+(1-t)\tilde{W}_B)X)\|_1}{N} \leq g_2(T,z,\delta,N,\varepsilon)$$

Using lemma 20 from Mei et al. [2019] we know that $\forall i \in [N], \bar{a}_i^T \leq K(1+T)$ for a certain constant $K$. Therefore we can bound $\|A_A\|_\infty, \|A_B\|_\infty \leq g_1(T,z,N,\varepsilon) := K(1+T) + Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)}+z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{d+\log(N)}+z)$.

Taking $z = \sqrt{\log\left(\frac{12}{\delta}\right)}$, we have shown the existence of a permutation $\gamma$ with probability at least $1-\delta$ such that almost surely on the choice of $x \sim P$ and $\forall t \in [0,1]$, we have:

$$\begin{aligned}
&|t\hat{f}(x;\theta_A) + (1-t)\hat{f}(x;\theta_B) - \hat{f}(x; t\theta_A + (1-t)\tilde{\theta}_B)| \leq g_1(T,z,N,\varepsilon)g_2(T,z,\delta,N,\varepsilon)\\
&\quad\leq \left(K(1+T) + Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)}+z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{d+\log(N)}+z)\right)\\
&\quad\left(L_\sigma(H_x\frac{C}{\delta}N^{-\frac{1}{s}} + 2H_x\sqrt{d}(Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)}+z]\right.\\
&\quad\left.+ Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{d+\log(N)}+z)))\right)
\end{aligned}$$

For fixed $T, \delta$ and $z = \sqrt{\log\left(\frac{12}{\delta}\right)}$, denote $\text{err}(N,\varepsilon)$ the right hand term.

It is clear that:

$$\forall \text{err} > 0 \exists N_{\min} \forall N \geq N_{\min} \exists \varepsilon_{\max}(N) \forall \varepsilon \leq \varepsilon_{\max}(N),\ \text{err}(N, \varepsilon) \leq \text{err}$$

This brings the first part of the theorem.

**Discussion**: Let's look more closely at the term $Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{d+\log(N)} + z))$ from Mei et al. [2019] which comes from the mean field approximation. When taken alone, this term yields an error which is independent of the input dimension $d$, since taking $N$ large leads to small error (provided $\varepsilon$ is small). However, here the growth of the hidden layer $N$ depends on the input dimension $d$ through the exponent $s > d_1^*(\mu)$ (with $d_1^*(\mu) \leq d$) using notations from Weed and Bach [2019]. This is due to Wasserstein convergence rates of empirical measures in dimension $d$. Without any further assumption on the weight distribution or precise study of the PDE we have to consider that $\text{Supp}(\mu))$ has dimension $d$. To remove this dependence, one could study a precise model for the data and look more closely at the PDE evolution to better understand the support of the distribution $\rho_t$. □

To prove the second part of the theorem, first notice that as already mentioned, $\|A_A\|_\infty, \|A_B\|_\infty \leq g_1(T, z, N, \varepsilon)$. Moreover, the data distribution is bounded and the weights of the first layer of the approximating PDE live in a bounded set thanks to Appendix C.2.2. It brings the existence of a constant $K(T)$ such that:

$$\forall t \leq T, \forall w \in \text{Supp}(\rho(t)), \forall X \in \text{Supp}(P), |w.x| \leq K(T)$$

Using we see that if $\varepsilon \leq \min\{\frac{1}{K_0 e^{K_0(1+T)^3}}, \frac{1}{K_0(d+\log(N)+z^2)e^{K_0(1+T)^3}}\}$, with probability at least $1 - \frac{\delta}{2}$, setting $z = \sqrt{\frac{12}{\delta}}$

$$\max_{k \in [0, T/\varepsilon] \bigcap \mathbb{N}} \max_{i \in [N]} |w_{A,i}^k x| \leq H_x \sqrt{d}(Ke^{K(1+T)^3} \frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{d+\log(N)} + z)) + K(T)$$

Up to changing previous $N_{\min}, \varepsilon_{\max}$, we can suppose that $\forall N \geq N_{\min}, \forall \varepsilon \leq \varepsilon_{\max}(N)$, $Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z] + Ke^{K(1+T)^2T}\varepsilon + Ke^{K(1+T)^2T}\sqrt{\varepsilon}(\sqrt{D+\log(N)} + z) \leq 1$

Therefore, since the data distribution is bounded we know that there exists a constant $K(T)$ such that if $N \geq N_{\min}, \varepsilon \leq \varepsilon_{\max}(N)$, $P-$almost-surely:

$$\begin{cases} |y| \leq K(T) \\ |\hat{f}(x : \theta_A)| \leq K(T) \\ |\hat{f}(x : \theta_B)| \leq K(T) \end{cases}$$

We make a general assumption on the loss of the form: $\forall y \in \mathbb{R}(x \to \mathcal{L}(x, y))$ is convex and $\forall K > 0, \exists C_K, \forall x_1, x_2, y \in [-K, K], |\mathcal{L}(x_1, y) - \mathcal{L}(x_2, y)| \leq C_K |x_1 - x_2|_2$. In particular this is true for the square loss.

Combining convexity of the loss, the first part of the theorem already proved and Lipschitzness on a compact domain of the loss, we get the second part of the theorem with a term $C_{K(T)}\text{err}$ instead of err. To solve this, just consider $\min\left\{\frac{\text{err}}{C_T}, \text{err}\right\}$ in the first part of the theorem.

We have supposed in the beginning that the non-linearity was bounded. But the previous study shows that with probability at least $1 - \delta$, $|w_{A,i}x|$ is upper bounded for all $i$ during the training up to time T by some constant depending only on $T, \delta$ provided $N \geq N_{\min}, \varepsilon \leq \varepsilon_{\max}(N)$ Therefore assuming that the non-linearity is bounded on a big enough compact set is enough to get the the result since it doesn't change the dynamics of the parameters considered. However the size of this set is not made explicit here.

### C.3 Proving LMC for general SGD

We will now study LMC of neural networks trained under general SGD using Theorem 4 part B from Mei et al. [2019]. The study is very similar to the case of noiseless SGD and will yield similar results.

More precisely in that case the PDE writes as:

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t(\theta)\nabla_\theta \Psi_\lambda(\theta;\rho_t)) + 2\xi(t)\tau d^{-1}\Delta_\theta \rho_t$$
$$\Psi_\lambda(\theta;\rho) = \Psi(\theta;\rho) + \frac{\lambda}{2}\|\theta\|_2^2$$

Notice that our Assumptions 1 and 2 imply assumptions 1 to 6 in Mei et al. [2019].

#### C.3.1 Intermediate dynamics for general SGD

Mei et al. [2019] define as before intermediate dynamics:

**Non linear dynamics**

Let consider $\bar{\theta}_i^t$ with initialization $\bar{\theta}_i^0 \sim \rho_0$ i.i.d. which follows the dynamics

$$\bar{\theta}_i^t = \bar{\theta}_i^0 + 2\int_0^t \xi(s)G(\bar{\theta}_i^s;\rho_s)ds + \int_0^t \sqrt{2\xi(s)\tau d^{-1}}dW_i(s)$$

where $G(\theta;\rho) = -\nabla\Psi_\lambda(\theta;\rho)$. An important fact is that $\bar{\theta}_i^t$ is random because of the random initialization and its law at time $t$ is $\rho_t$. It corresponds to the evolution of particles under a field which depends only on the position of the optimized particle and the overall distribution of all particles plus and a diffusion term.

**Particle Dynamics**

Let $\underline{\theta}_i^t$ with initialization $\underline{\theta}_i^0 = \bar{\theta}_i^0$ with the following dynamics where $\underline{\rho}_t^{(N)} = \frac{1}{N}\sum_{i=1}^N \delta_{\underline{\theta}_i^t}$ denote the empirical distribution of $\underline{\theta}_i^t$.

$$\underline{\theta}_i^t = \underline{\theta}_i^0 + 2\int_0^t \xi(s)G(\underline{\theta}_i^s;\underline{\rho}_s^{(N)})ds + \int_0^t \sqrt{2\xi(s)\tau d^{-1}}dW_i(s)$$

**Gradient descent dynamics**

Let $\tilde{\theta}_i^k$ with initialization $\tilde{\theta}_i^0 = \bar{\theta}_i^0$ with the following dynamics:

$$\tilde{\theta}_i^k = \tilde{\theta}_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)G(\tilde{\theta}_i^l;\tilde{\rho}_l^{(N)}) + \int_0^{k\varepsilon}\sqrt{2\xi([s])\tau d^{-1}}dW_i(s)$$

**Stochastic Gradient Descent Dynamics**

Consider $\theta_i^k$ with initialization $\theta_i^0 = \bar{\theta}_i^0$ that follows:

$$\theta_i^k = \theta_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)F_i(\theta^l;z_{l+1}) + \int_0^{k\varepsilon}\sqrt{2\xi([s])\tau d^{-1}}dW_i(s)$$

where $F_i(\theta^k;z_{k+1}) = -\lambda\theta_i^k + (y_{k+1} - \hat{y}_{k+1})\nabla_{\theta_i}\sigma_*(x_{k+1};\theta_i^k)$, $z_k = (x_k, y_k)$ and $\hat{y}_{k+1} = \frac{1}{N}\sum_{j=1}^N \sigma_*(x_{k+1};\theta_j^k)$

As before we first control the distance between noisy SGD and non linear dynamics with the following lemma:

**Lemma C.2.** *Consider a two-layer neural network with notations as before trained with noisy SGD for an underlying time $T$. Assume $T \geq 1$. Then under assumptions Assumptions 1 and 2, there exists a constant $K$ such that with probability at least $1 - 3e^{-z^2}$ we have:*

$$\max_{k\in[0,T/\varepsilon]\bigcap\mathbb{N}}\max_{i\in[N]}\|\theta_i^k-\bar{\theta}_i^{k\varepsilon}\|_2 \leq Ke^{e^{KT}[\sqrt{\log(N)}+z^2]}[\sqrt{d\log(N)}+z^3+\log^{3/2}(NT)]/\sqrt{N}$$
$$+Ke^{e^{KT}[\sqrt{\log(N)}+z^2]}[\log(N(T/\varepsilon\vee 1))+z^4]\sqrt{\varepsilon}$$
$$+Ke^{e^{KT}[\sqrt{\log(N)}+z^2]}[\sqrt{d}\log(N)+z^3+\log^{3/2}(N)]\sqrt{\varepsilon}$$

*Proof.* Just apply proposition 47,49,50 from Mei et al. [2019]. □

We moreover recall here Lemma 9 from Mei et al. [2019] which bounds the value of the second layer coefficients $A^t=(a_1^t,\dots,a_N^t)$.

**Lemma C.3** (Lemma 19 in Mei et al. [2019])**.** *There exists a constant $K$ such that with probability at least $1-e^{-z^2}$ we have*

$$\sup_{t\in[0,T]}\|A^t\|_\infty \leq Ke^{KT}[\sqrt{\log(N)}+z]$$

### C.3.2 Proof of Theorem 3.1 in the case of noisy regularized SGD

We can now prove LMC for two two-layer networks trained with general SGD:

**Theorem 3.1.** *Consider two two-layer neural networks as in Equation (3) trained with equation SGD with the same initialization over the weights independently and for the same underlying time $T$. Suppose Assumptions 1 and 2 to hold. Then $\forall\delta, err, \exists N_{min}$ such that if $N\geq N_{min}, \exists\varepsilon_{max}(N)$ such that if $\varepsilon\leq\varepsilon_{max}(N)$ in Equation (4), then with probability at least $1-\delta$ over the training process, there exists a permutation of the second network's hidden layer such that for almost every $x\sim P$:*

$$|t\hat{f}_N(x;\theta_A)+(1-t)\hat{f}_N(x;\theta_B)$$
$$-\hat{f}_N(x;t\theta_A+(1-t)\tilde{\theta}_B)|\leq err,\quad \forall t\in[0,1]\,.$$

*Proof.* We follow the same steps as before. Recall that the data distribution is bounded: $\mathrm{Supp}(P)\subset[-H_x,H_x]^d\times[-H_y,H_y]$.

The problem is that due to the stochasticity added in the noisy SGD, $\mathrm{Supp}(\rho_t)$ is not necessarily bounded anymore. However, using step 3 of the proof of lemma 41 in Mei et al. [2019] and the fact that the initial distribution $\rho_0$ has bounded support (sub-Gaussian would be enough), the distribution of weights of the first layer at time $T$ is sub-Gaussian.

Indeed if $\bar{\theta}_i^t\sim\rho_t$ and $\rho_0$ is bounded or sub-Gaussian, we get the existence of $K$ such that:

$$\mathbb{P}(\|\bar{\theta}_i^T\|_2^2\geq Ke^{KT}(1+z)\sqrt{T})\leq e^{-z^2}$$

which proves that $\rho_T$ is sub-Gaussian.

We could adapt the proof done in Lemma B.14 for Gaussian variable to sub-Gaussian variable to show the existence of constants (Lemma B.14 dealt with $\mathcal{W}_2^2$ but can be extended to $\mathcal{W}_1$ because Proposition 15 of Weed and Bach [2019] is valid for any $\mathcal{W}_p^p$) $D_2', E_2'$ depending only on the constant $K$ of sub-Gaussianity of the previous distribution, and hence independent of $N$ such that by considering the norm $\|\cdot\|_2$, we can still bound for $m\geq E_2'^d$

$$\mathbb{E}[\mathcal{W}_1(\hat{\mu}_A,\hat{\mu}_B)]\leq\sqrt{\frac{D_2'}{d}\log(N)}\left(\frac{1}{N}\right)^{1/d}$$

Therefore, with probability at least $1 - \frac{\delta}{2} - 6e^{-z^2}$:

$$\begin{aligned}\forall X \in [-H_x, H_x]^D, \frac{\|\sigma(W_A X) - \sigma((tW_A + (1-t)\tilde{W}_B)X)\|_1}{N} &\leq g_2(T, \varepsilon, N, \delta) \\ &:= L_\sigma H_x \sqrt{d} \frac{1}{\delta} \sqrt{\frac{D_2'}{d} \log(N)} N^{-1/d} \\ &+ 2L_\sigma H_x \sqrt{d} (K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d \log(N)} + z^3 + \log^{3/2}(NT)]/\sqrt{N} \\ &+ K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\log(N(T/\varepsilon \vee 1)) + z^4] \sqrt{\varepsilon} \\ &+ K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d} \log(N) + z^3 + \log^{3/2}(N)] \sqrt{\varepsilon})\end{aligned}$$

and same for the second term. We moreover have, using Lemmas C.2 and C.3 that with probability at least $1 - 2e^{-z^2}$:

$$\begin{aligned}\max\{\|A_A\|_\infty, \|A_B\|_\infty\} &\leq g_1(T, z, N, \varepsilon) \\ &:= K e^{KT}[\sqrt{\log(N)} + z] + K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d \log(N)} + z^3 + \log^{3/2}(NT)]/\sqrt{N} \\ &+ K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\log(N(T/\varepsilon \vee 1)) + z^4] \sqrt{\varepsilon} \\ &+ K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d} \log(N) + z^3 + \log^{3/2}(N)] \sqrt{\varepsilon}\end{aligned}$$

Taking $z = \sqrt{\log\left(\frac{16}{\delta}\right)}$ such that $8e^{-z^2} = \frac{\delta}{2}$ we get that with probability at least $1 - \delta$:

$$|t\hat{f}(x; \theta_A) + (1-t)\hat{f}(x; \tilde{\theta}_B) - \hat{f}(x; t\theta_A + (1-t)\tilde{\theta}_B)| \leq g_1(T, z, N, \varepsilon) g_2(T, z, N, \varepsilon)$$

As before, for fixed $T, \delta$, denote $\mathrm{err}(N, \varepsilon)$ the left hand term.

It is clear that:

$$\forall \mathrm{err} > 0 \exists N_{\min} \forall N \geq N_{\min} \exists \varepsilon_{\min}(N) \forall \varepsilon \leq \varepsilon_{\min}(N),\ \mathrm{err}(N, \varepsilon) \leq \mathrm{err}$$

Therefore, sending $N \to \infty$, $\varepsilon \to 0$ brings immediately the first part of the theorem.

To get the second part of the theorem, we do the same procedure as for noiseless regularization-free SGD.

Namely, with probability $1 - \delta$ we have both:

$$\begin{cases} P - \text{almost surely}, |t\hat{f}(x; \theta_A) + (1-t)\hat{f}(x; \theta_B) - \hat{f}(x; t\theta_A + (1-t)\tilde{\theta}_B)| \leq \mathrm{err} \\ \max\{\|A_A\|_\infty, \|A_B\|_\infty\} \leq g_1(T, N, z, \varepsilon) \end{cases}$$

Up to changing $N_{\min}, \varepsilon_{\max}(N)$ we can suppose that $\forall N \geq N_{\min}, \forall \varepsilon \leq \varepsilon_{\max}(N)$, we have

$$\begin{aligned}K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d \log(N)} + z^3 + \log^{3/2}(NT)]/\sqrt{N} \\ + K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\log(N(T/\varepsilon \vee 1)) + z^4] \sqrt{\varepsilon} \\ + K e^{e^{KT}[\sqrt{\log(N)}+z^2]} [\sqrt{d} \log(N) + z^3 + \log^{3/2}(N)] \sqrt{\varepsilon} \leq 1\end{aligned}$$

which brings

$$g_1(T, N, z, \varepsilon) \leq K e^{KT}[\sqrt{\log(N)} + z] + 1$$

Appendix C.3.2, boundness of the activation, boundness of the input distribution $(x, y) \sim P$ by assumption imply the existence of $K'$ such that $\forall t \in [0, 1]$, $P-$almost-surely and for $N$ large enough,

$$\begin{cases} |\hat{f}(x; \theta_A)| \leq K' \left( K e^{KT} [\sqrt{\log(N)} + z] \right) \\ |\hat{f}(x; \theta_B)| \leq K' \left( K e^{KT} [\sqrt{\log(N)} + z] \right) \\ |y| \leq H_y \leq K' \left( K e^{KT} [\sqrt{\log(N)} + z] \right) \end{cases}$$

Using convexity and Lipschitzness of the squared loss on compact domains we get the existence (and this is a sufficient condition for the loss function with convexity) of $\text{Lip} : \mathbb{R}_+ \to \mathbb{R}_+$ such that: $\exists L_1, L_2, \forall H \in \mathbb{R}_+$, $\text{Lip}(H) \leq L_1 + L_2 \exp(H)$, $\forall x_1, x_2, y \in [-H, H], |\mathcal{L}(x_1, y) - \mathcal{L}(x_2, y)| \leq \text{Lip}(H)|x_1 - x_2|$ and such that with probability at least $1 - \delta$:

$$\mathbb{E}[\mathcal{L}(\hat{f}(x; t\theta_A + (1-t)\tilde{\theta}_B), y)] \leq \text{Lip}(K' \left( K e^{KT} [\sqrt{\log(N)} + z] + 1 \right)) g_1(T, z, \varepsilon, N) g_2(T, z, \varepsilon, N)$$

Plugging this back and with the exact same discussion as before we get $\exists N_{\min} \forall N \geq N_{\min} \exists \varepsilon_{\max} \forall \varepsilon \leq \varepsilon_{\max}$,

$$\mathbb{E}[\mathcal{L}(\hat{f}(x; t\theta_A + (1-t)\tilde{\theta}_B), y)] \leq \text{err} + t\mathbb{E}[\mathcal{L}(\hat{f}(x; \theta_A), y)] + (1-t)\mathbb{E}[\mathcal{L}(\hat{f}(x; \theta_B), y)]$$

To get both part 1 and 2 of the theorem at the same time we just have to reconsider the max of both $N_{\min}$ and the min of both $\varepsilon_{\max}$.

□

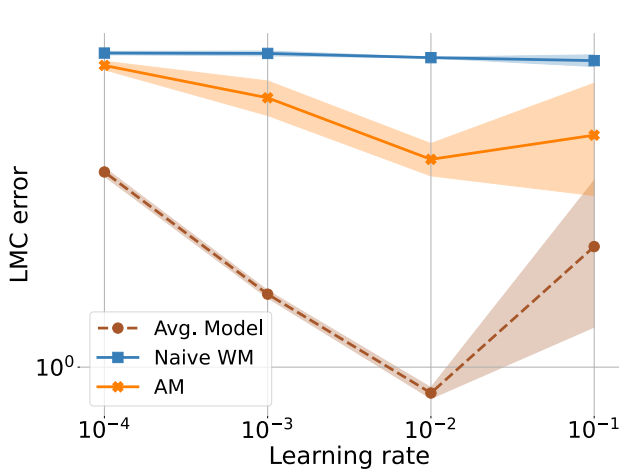
### C.4 On the satisfiability of our assumptions

Assumption 1 and 2 are non-trivial but standard in the mean field literature. They are made to ensure that the optimization of the two-layer neural networks happens in the mean-field regime. Indeed, as explained above the weights are then approximately indepedent and we can leverage Wasserstein convergence bounds of empirical measure to prove linear mode connectivity. We used conventioanl assumptions from the mean-field litterature (e.g. see assumptions A1 to A6 in Mei et al. [2019], A1 to A4 in Mei et al. [2018]). $u, v, U, V$ are implicitly defined but once the non-linearity and the data distribution are fixed, $u, v, U, V$ are fully determined as functions of the parameters. Checking their derivability can be done using usual rules for derivation under the integral sign if the non-linearity is smooth. A particular case is to consider a two-layer network, with a sigmoïd activation, a bounded data distribution and a bounded uniform initialization over the parameters. We additionally mention that a lot of works empirically evidence the validity of the mean-field framework, hence we feel validating our approach. In the case of mltilayer networks, the main assumption is the independence of weights inside each layer (Assumption 6). Multiple recent works address this question by using a mean field view for multi-layer networks with bounds on the width needed during optimization with SGD (e.g. Th. 15 in Nguyen and Pham [2023]). Finally, we believe our results could be extended to approximated independence of weights with an additional error term for the error barrier on a linear path corresponding to the approximated independence. Quantifying the impact of correlation between weights constitutes a very interesting avenue for future work.

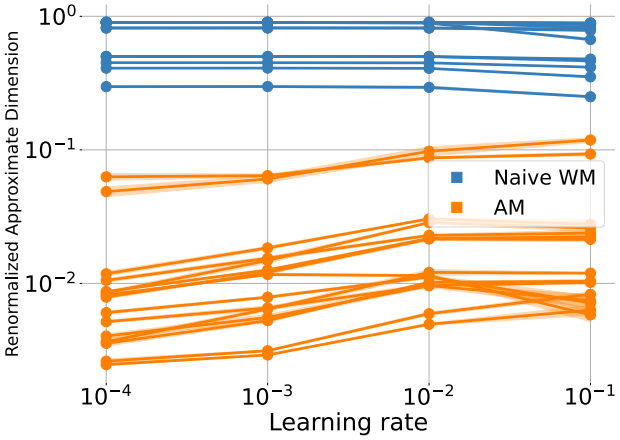
## D EXPERIMENTS

### D.1 Experiment on CIFAR10

We compared activation and weight matching methods on the CIFAR10 dataset for a VGG16 model. Our experiment again shows the correlation between small approximate dimension of the support of the weight distribution and LMC effectiveness hence supporting our main theoretical study. As it is non trivial to compute the covariance of the input with convolutional layers as it is a high dimensional tensor we left the alternative weight matching methods as a future work. Providing a scalable technique to estimate such a covariance for CNNs is a an interesting research direction beyond the scope of this paper.




(a) Mean test loss of the trained networks $A$ and $B$ and error barrier on the linear path $M_t$, $t \in [0, 1]$ across different learning rate values for each matching problem.




(b) Approximate dimension $\mathrm{Dim}(S) := {}^{\mathrm{tr}(S)^2}/_{\mathrm{tr}(S^2)}$ of the matrices considered in the matching problems at each layer.

Figure 3: Statistics of the average network $M$ over the linear path between networks $A$ and $B$ using respectively weight matching (blue) and activation matching (orange)

### D.2 Details about our new weight matching method

Until now we have studied the influence of the dimension of the support of the underlying distribution of weights on the convergence rate in Wasserstein distance of the corresponding empirical measure. An interesting question is to look at the influence of the distance used to define the Wasserstein distance.

More precisely, consider a single layer of two networks $A, B$ with input $X \in \mathbb{R}^n$ and matrix weights $W_{A,B} \in \mathcal{M}_{m,n}(\mathbb{R})$. Consider that the input data follows a distribution $P$ with $\mathbb{E}_P[XX^T] = \Sigma$

The underlying method that we use in our proof and which is the one referred to as weight matching method in Ainsworth et al. [2022] consists in minimizing the distances for the euclidean norm between weights matrices, i.e. to find:

$$\underset{\Pi \in \mathcal{S}_m}{\arg\min} \|W_A - \Pi W_B\|_2$$

This is equivalent to finding:

$$\underset{\Pi \in \mathcal{S}_m}{\arg\min} \sqrt{\frac{1}{m}\|W_A - \Pi W_B\|_2^2\}} = \underset{\pi \in \mathcal{S}_m}{\arg\min} \sqrt{\frac{1}{m}\sum_{i=1}^{m} \|W_{A,i:} - W_{B,\pi_i:}\|_2^2}$$

We get an expected square error between network $A$ and network $B$ permuted at the output layer of:

$$\mathbb{E}\left[\|W_A X - \Pi W_B X\|_2^2\right] = (W_A - \Pi W_B)\Sigma(W_A - \Pi W_B)^T = \|W_A - \Pi W_B\|_{2,\Sigma}^2$$

where $\|\cdot\|_{2,\Sigma}$ is the semi-norm coming from $(X, Y) \to X^T \Sigma Y$ which is a symmetric positive bilinear product since $\Sigma$ is symmetric positive (and a norm when $\Sigma$ is definite positive i.e., when $\mathrm{Span}(\mathrm{Supp}(P)) = \mathbb{R}^n$).

Minimizing the cost $\|W_A - \Pi W_B\|_2$ contributes to minimizing the expected squared error $\|W_A - \Pi W_B\|_{2,\Sigma}^2$ but it appears more natural to directly minimize the cost $\|W_A - \Pi W_B\|_{2,\Sigma}^2$.

As explained in Theorem D.2 below, we can directly link the approximate dimension of the underlying covariance matrix of each method with the decay rate of LMC error barrier. The underlying covariance matrix of each method is $W_A^\ell [W_A^\ell]^T$ for WM (naive), $W_A^\ell \Sigma_A^{\ell-1} [W_A^\ell]^T$ for WM (ours) and $\Sigma_A^\ell$ for AM.

Indeed,

- for naive weight matching, each row of $W_A, W_B$ follows a distribution with covariance matrix $W_A^\ell [W_A^\ell]^T$,
- for weight matching (ours), the optimization problem can be seen (Lemma D.1) as for naive weight matching but with covariance matrix $W_A^\ell \Sigma_A^{\ell-1} [W_A^\ell]^T$,
- for activation matching, each row of $Z_A^\ell$ follows a distribution with covariance matrix $\Sigma_A^\ell$.

### D.3 Gain of our new weight matching method

This section is motivated by the following question:

**What is the gain of optimizing directly the cost $\|W_A - \Pi W_B\|_{2,\Sigma}$ when $\Sigma$ is low dimensional?**

For example, let's suppose that $\Sigma = \mathrm{Diag}(1, 1, 0, ..., 0)$ and hence the support of $X$ is two dimensional. Suppose moreover that $W_A$ and $W_B$ are as in *Section* 5.1 initialized i.i.d. with a distribution $\mathcal{N}(0, \frac{I_n}{n})$ on the weights. Hence we have seen before that $\|W_A - \tilde{W}_B\|_2 \sim \left(\frac{1}{m}\right)^{1/n}$. Since the minimization procedure is unaware of the structure of $\Sigma$ it is clear by symmetry that $\sqrt{\|W_{A,1:} - \tilde{W}_{B,1:}\|_2^2 + \|W_{A,2:} - \tilde{W}_{B,2:}\|_2^2} \sim \sqrt{\frac{2}{n}} \left(\frac{1}{m}\right)^{1/n}$. Therefore the convergence is still as $\left(\frac{1}{m}\right)^{1/n}$. However if we had first aimed at minimizing $\|W_A - \Pi W_B\|_{2,\Sigma}$ it is clear that the problem becomes two dimensional and hence $\arg\min_{\Pi \in \mathcal{S}_m} \|W_A - \Pi W_B\|_{2,\Sigma} \sim \left(\frac{1}{m}\right)^{1/2}$ which is extremely faster when $n$ is large.

We want to apply this idea to our setting where we suspect the distribution of activations at each layer to be low dimensional. We now prove the following lemma:

**Lemma D.1.** *Let $W_A, W_B \in \mathcal{M}_{m,n}(\mathbb{R})$ satisfy Assumption 6 with underlying distribution $\mu$ and let $\Sigma \in \mathcal{M}_n(\mathbb{R})$. Write $\Sigma = O\sqrt{\Sigma}^2 O^T$ where $O$ is orthogonal and $\sqrt{\Sigma}$ is diagonal. Then we get the equivalence between optimization problems:*

$$\mathbb{E}\left[\min_{\Pi \in \mathcal{S}_m} \|W_A - \Pi W_B\|_{2,\Sigma}^2\right] = \mathbb{E}\left[\min_{\Pi \in \mathcal{S}_m} \|\hat{W}_A - \Pi \hat{W}_B\|_2^2\right] \tag{11}$$

*where $\hat{W}_B, \hat{W}_B$ satisfy Assumption 6 with underlying distribution $f_*\mu$ the image measure of $\mu$ by $f : X \mapsto O\sqrt{\Sigma}O^T X$*

*Proof.* Just notice that $\forall \Pi \in \mathcal{S}_m$

$$\|W_A - \Pi W_B\|_{2,\Sigma}^2 = \mathrm{tr}[(W_A - \Pi W_B) O\sqrt{\Sigma}O^T (O\sqrt{\Sigma}O^T)^T (W_A - \Pi W_B)^T]$$

and do the change of variable $\hat{W}_A = W_A O\sqrt{\Sigma}O^T$ (repectively $\hat{W}_B$) □

**Theorem D.2.** *Consider $X \in \mathbb{R}^n \sim P \in \mathcal{P}_1(\mathbb{R}^n)$ such that $\mathbb{E}_P[XX^T] = \Sigma = \mathrm{Diag}(1, \dots, 1, 0, \dots, 0)$, $rank(\Sigma) = \tilde{n} \leq n$ and $W_A, W_B$ random weight matrices satisfying Assumption 6 with underlying distribution $\mathcal{N}\left(0, \frac{I_n}{n}\right)$ Denote $\Pi_1, \Pi_2 \in \mathcal{S}_m$ random permutations that minimize the respective costs $\|W_A - \Pi W_B\|_2$ and $\|W_A - \Pi W_B\|_{2,\Sigma}$ Then we have:*

$$\mathbb{E}\left[\|W_A - \Pi_1 W_B\|_{2,\Sigma}^2\right] = \tilde{\Omega}\left(\left(\frac{1}{m}\right)^{2/n}\right)$$

$$\mathbb{E}\left[\|W_A - \Pi_2 W_B\|_{2,\Sigma}^2\right] = \tilde{\Omega}\left(\left(\frac{1}{m}\right)^{2/\tilde{n}}\right)$$

$$\mathbb{E}[\|W_A - \Pi_1 W_B\|_{2,\Sigma}^2] = \tilde{\mathcal{O}}\left(\left(\frac{1}{m}\right)^{2/n}\right)$$

$$\mathbb{E}[\|W_A - \Pi_2 W_B\|_{2,\Sigma}^2] = \tilde{\mathcal{O}}\left(\left(\frac{1}{m}\right)^{2/\tilde{n}}\right)$$

*Proof.* Using Lemma D.1, we see that bounds 2 and 4 are just corollaries of Lemma B.16 and theorem 5.3.

To show bounds 1 and 3 just notice that:

$$\Pi_1 = \underset{\Pi \in \mathcal{S}_m}{\arg\min} \|W_A - \Pi W_B\|_2$$

is almost surely unique.

By symmetry of the problem and Theorem 5.3 we therefore see that $\forall i \in [n]$:

$$\mathbb{E}\|[W_A - \Pi_1 W_B]_{:i}\|_2^2 = \frac{1}{n}\mathbb{E}\|W_A - \Pi_1 W_B\|_2^2 = \tilde{\Omega}\left(\left(\frac{1}{m}\right)^{2/n}\right)$$

Finally noticing that $\Sigma = \text{Diag}(1, \dots, 1, 0, \dots, 0)$ we get by summing:

$$\mathbb{E}\left[\|W_A - \Pi_1 W_B\|_{2,\Sigma}^2\right] = \frac{\tilde{n}}{n}\tilde{\Omega}\left(\left(\frac{1}{m}\right)^{2/n}\right) = \tilde{\Omega}\left(\left(\frac{1}{m}\right)^{2/n}\right)$$

Similarly, exploiting a.s. uniqueness of $\Pi_1$ and symmetry across dimensions, we get the third inequality. □