Supplementary Material (Appendix)
Landscape Connectivity and Dropout Stability of SGD Solutions for Over-parameterized Neural Networks

A. Proof of Theorem 1

A.1. Part (A)

Given $\boldsymbol{\theta} = (a, \boldsymbol{w}) \in \mathbb{R}^D$, let $\sigma_\star(\boldsymbol{x}, \boldsymbol{\theta}) = a\sigma(\boldsymbol{x}, \boldsymbol{w})$. Given $\rho \in \mathscr{P}(\mathbb{R}^D)$, we define the limit loss as

$$\bar{L}(\rho) = \mathbb{E}\left\{\left(y - \int \sigma_\star(\boldsymbol{x}, \boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta})\right)^2\right\}, \tag{A.1}$$

where the expectation is taken over $(\boldsymbol{x}, y)$. For $i \in [N]$ and $t \geq 0$, we consider the following nonlinear dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{\theta}}_i^t = 2\xi(t)\int \mathbb{E}\left\{\nabla\sigma_\star(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_i^t)\,(y - \sigma_\star(\boldsymbol{x}, \boldsymbol{\theta}'))\right\}\rho_t(\mathrm{d}\boldsymbol{\theta}'), \tag{A.2}$$

where ∇ denotes the gradient with respect to $\bar{\boldsymbol{\theta}}_i^t$ and $\bar{\boldsymbol{\theta}}_i^t \sim \rho_t$. We initialize (A.2) with $\{\bar{\boldsymbol{\theta}}_i^0\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \rho_0$.

In (Mei et al., 2019), it is considered the two-layer neural network (3.1) with N neurons and bounded activation function σ, and it is studied the evolution under the SGD algorithm (3.3) of the parameters $\boldsymbol{\theta}^k$. In particular, it is shown that, under suitable assumptions, *(i)* the solution of (A.2) exists and it is unique, *(ii)* the N i.i.d. ideal particles $\{\bar{\boldsymbol{\theta}}_i^t\}_{i=1}^N$ are close to the parameters $\boldsymbol{\theta}^k$ obtained after k steps of SGD with step size α, with $t = k\alpha$, and *(iii)* the loss $L_N(\boldsymbol{\theta}^k)$ concentrates to the limit loss $\bar{L}(\rho_t)$, where ρ_t is the law of $\bar{\boldsymbol{\theta}}_i^t$.

Let us now provide the proof of Theorem 1, part (A).

Proof of Theorem 1, part (A). Without loss of generality, we can assume that $\boldsymbol{\theta}_{\mathrm{S}}^k$ contains the first $|\mathcal{A}|$ elements of $\boldsymbol{\theta}^k$, i.e., $\boldsymbol{\theta}_{\mathrm{S}}^k = (\boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k, \dots, \boldsymbol{\theta}_{|\mathcal{A}|}^k)$. In fact, the subset $\mathcal{A}$ is independent of the SGD algorithm. Thus, by symmetry, the joint distribution of $\{\boldsymbol{\theta}_i^k\}_{i\in\mathcal{A}}$ depends only on $|\mathcal{A}|$ (and not on the set $\mathcal{A}$ itself). By Definition 3.1, we need to show that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} |L_N(\boldsymbol{\theta}^k) - L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k)| \leq Ke^{KT^3}\left(\frac{\sqrt{\log|\mathcal{A}|} + z}{\sqrt{|\mathcal{A}|}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big)\right). \tag{A.3}$$

Let $\bar{\boldsymbol{\theta}}^{k\alpha} = (\bar{\boldsymbol{\theta}}_1^{k\alpha}, \dots, \bar{\boldsymbol{\theta}}_N^{k\alpha})$ be the solution of the nonlinear dynamics (A.2) at time $k\alpha$, with $\bar{\boldsymbol{\theta}}_i^{k\alpha} \sim \rho_{k\alpha}$. By triangle inequality, we have that

$$\begin{aligned} |L_N(\boldsymbol{\theta}^k) - L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k)| &\leq |L_N(\boldsymbol{\theta}^k) - \bar{L}(\rho_{k\alpha})| + |L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k) - \bar{L}(\rho_{k\alpha})| \\ &\leq |L_N(\boldsymbol{\theta}^k) - \bar{L}(\rho_{k\alpha})| + |L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha})| + |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) - \bar{L}(\rho_{k\alpha})|, \end{aligned} \tag{A.4}$$

where $\bar{L}$ is defined in (A.1) and $\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha} = (\bar{\boldsymbol{\theta}}_1^{k\alpha}, \bar{\boldsymbol{\theta}}_2^{k\alpha}, \dots, \bar{\boldsymbol{\theta}}_{|\mathcal{A}|}^{k\alpha})$ denotes the vector containing the first $|\mathcal{A}|$ elements of $\bar{\boldsymbol{\theta}}^{k\alpha}$.

Let us consider the first term in the RHS of (A.4). Note that, without loss of generality, we can assume that $\alpha \leq 1/(C(D + \log N + z^2)e^{CT^3})$, for some constant C depending only on the constants K_i of the assumptions **(A1)**-**(A4)**. Let us explain why this is the case. If $\alpha > 1/(C(D + \log N + z^2)e^{CT^3})$, then the RHS of (A.3) is lower bounded by a constant

depending only on K_i. Furthermore, y and σ are bounded, and by Proposition 8 of (Mei et al., 2019), we have that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} \max_{i\in[N]} |a_i^k| \leq C_3(1+T). \tag{A.5}$$

Thus, if $\alpha > 1/(C(D+\log N + z^2)e^{CT^3})$, then the result is trivially true. Consequently, we can apply Theorem 1 of (Mei et al., 2019) and we have that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} |L_N(\boldsymbol{\theta}^k) - \bar{L}(\rho_{k\alpha})| \leq C_1 e^{C_1 T^3} \left(\frac{\sqrt{\log N} + z}{\sqrt{N}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big) \right), \tag{A.6}$$

where C_1 depends only on K_i. In what follows, the C_i are constants that depend only on K_i.

Let us now consider the second term in the RHS of (A.4). After some manipulations, we have that

$$\begin{aligned}
|L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha})| &\leq 2 \max_{i\in\mathcal{A}} \left| a_i^k \mathbb{E}\left\{ y\sigma(\boldsymbol{x}, \boldsymbol{w}_i^k) \right\} - \bar{a}_i^{k\alpha} \mathbb{E}\left\{ y\sigma(\boldsymbol{x}, \bar{\boldsymbol{w}}_i^{k\alpha}) \right\} \right| \\
&\quad + \max_{i,j\in\mathcal{A}} \left| a_i^k a_j^k \mathbb{E}\{\sigma(\boldsymbol{x}, \boldsymbol{w}_i^k)\sigma(\boldsymbol{x}, \boldsymbol{w}_j^k)\} - \bar{a}_i^{k\alpha} \bar{a}_j^{k\alpha} \mathbb{E}\{\sigma(\boldsymbol{x}, \bar{\boldsymbol{w}}_i^{k\alpha})\sigma(\boldsymbol{x}, \bar{\boldsymbol{w}}_j^{k\alpha})\} \right| \\
&\leq C_2 \left(\max_{i\in\mathcal{A}} \left(1 + \max(|a_i^k|, |\bar{a}_i^{k\alpha}|)\right) \right)^2 \max_{i\in\mathcal{A}} \|\boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\alpha}\|_2 \\
&\leq C_2 \left(\max_{i\in[N]} \left(1 + \max(|a_i^k|, |\bar{a}_i^{k\alpha}|)\right) \right)^2 \max_{i\in[N]} \|\boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\alpha}\|_2,
\end{aligned} \tag{A.7}$$

where $\boldsymbol{\theta}_i^k = (a_i^k, \boldsymbol{w}_i^k)$, $\bar{\boldsymbol{\theta}}_i^{k\alpha} = (\bar{a}_i^{k\alpha}, \bar{\boldsymbol{w}}_i^{k\alpha})$, and in the second inequality we use that y, σ and the gradient of σ are bounded. By using Lemma 7 of (Mei et al., 2019), we have that

$$\sup_{t\in[0,T]} \max_{i\in[N]} |\bar{a}_i^t| \leq C_3(1+T). \tag{A.8}$$

Furthermore, by using Propositions 6-7-8 of (Mei et al., 2019), we have that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} \max_{i\in[N]} \|\boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\alpha}\|_2 \leq C_4 e^{C_4 T^3} \left(\frac{\sqrt{\log N} + z}{\sqrt{N}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big) \right). \tag{A.9}$$

As a result, by combining (A.5), (A.8) and (A.7), we conclude that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} |L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}^k) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha})| \leq C_5 e^{C_5 T^3} \left(\frac{\sqrt{\log N} + z}{\sqrt{N}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big) \right). \tag{A.10}$$

Finally, let us consider the third term in the RHS of (A.4). By triangle inequality, we have that

$$|L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) - \bar{L}(\rho_{k\alpha})| \leq \left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) \right\} \right| + \left| \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) \right\} - \bar{L}(\rho_{k\alpha}) \right|, \tag{A.11}$$

where the notation $\mathbb{E}_{\rho_0}$ emphasizes that the expectation is taken with respect to $\bar{\boldsymbol{\theta}}_i^0 \sim \rho_0$. Recall that $\bar{L}$ is defined in (A.1) and that

$$L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}) = \mathbb{E}_{(\boldsymbol{x},y)} \left\{ \left(y - \frac{1}{|\mathcal{A}|} \sum_{i=1}^{|\mathcal{A}|} \sigma_\star(\boldsymbol{x}, \boldsymbol{\theta}_i) \right)^2 \right\}, \tag{A.12}$$

where the notation $\mathbb{E}_{(\boldsymbol{x},y)}$ emphasizes that the expectation is taken with respect to $(\boldsymbol{x}, y) \sim \mathbb{P}$. Furthermore, note that $\{\bar{\boldsymbol{\theta}}_i^{k\alpha}\}_{i=1}^{|\mathcal{A}|} \overset{\text{i.i.d.}}{\sim} \rho_{k\alpha}$. Thus, after some manipulations, we can rewrite the second term in the RHS of (A.11) as

$$\begin{aligned}
&\left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) \right\} \right| \\
&\quad = \frac{1}{|\mathcal{A}|} \left| \int \mathbb{E}_{(\boldsymbol{x},y)} \left\{ \big(\sigma_\star(\boldsymbol{x}, \boldsymbol{\theta})\big)^2 \right\} \rho_{k\alpha}(\mathrm{d}\boldsymbol{\theta}) - \int \mathbb{E}_{(\boldsymbol{x},y)} \left\{ \sigma_\star(\boldsymbol{x}, \boldsymbol{\theta}_1)\sigma_\star(\boldsymbol{x}, \boldsymbol{\theta}_2) \right\} \rho_{k\alpha}(\mathrm{d}\boldsymbol{\theta}_1)\rho_{k\alpha}(\mathrm{d}\boldsymbol{\theta}_2) \right|.
\end{aligned} \tag{A.13}$$

As σ is bounded by assumption **(A3)** and $\sup_{k\in[T/\alpha]} \max_{i\in[N]} |\bar{a}_i^{k\alpha}|$ is bounded by (A.8), we deduce that

$$\sup_{k\in[T/\alpha]} \left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^{k\alpha}) \right\} \right| \leq \frac{C_6\,(1+T)^2}{|\mathcal{A}|}. \tag{A.14}$$

Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be two parameters that differ only in one component, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_{|\mathcal{A}|})$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i', \ldots, \boldsymbol{\theta}_{|\mathcal{A}|})$, and such that $\max_{i\in|\mathcal{A}|} |a_i| \leq C(1+T)$ and $\max_{i\in|\mathcal{A}|} |a_i'| \leq C(1+T)$. Then,

$$\left| L_{|\mathcal{A}|}(\boldsymbol{\theta}) - L_{|\mathcal{A}|}(\boldsymbol{\theta}') \right| \leq \frac{C_7\,(1+T)^2}{|\mathcal{A}|}. \tag{A.15}$$

As $\max_{i\in[N]} |\bar{a}_i^t|$ is bounded by (A.8), by applying McDiarmid's inequality, we obtain that

$$\mathbb{P}\left(\left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) \right\} \right| > \delta \right) \leq \exp\left(-\frac{|\mathcal{A}|\delta^2}{C_8(1+T)^4} \right). \tag{A.16}$$

Furthermore, we have that

$$\begin{aligned} \left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^s) \right| &\leq C_9 \left(\max_{i\in[N]} \left(1 + \max(|\bar{a}_i^t|, |\bar{a}_i^s|)\right) \right)^2 \max_{i\in[N]} \|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^s\|_2 \\ &\leq C_{10}(1+T)^4 |t-s|, \end{aligned} \tag{A.17}$$

where in the first inequality we use passages similar to those of (A.7), and in the second inequality we use (A.8) and Lemma 9 of (Mei et al., 2019). Consequently,

$$\left| |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) - \mathbb{E}_{\rho_0}\{L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t)\}| - |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^s) - \mathbb{E}_{\rho_0}\{L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^s)\}| \right| \leq C_{11}(1+T)^4 |t-s|. \tag{A.18}$$

By taking a union bound over $s \in [T/\nu]$ and bounding the difference between time in the interval grid, we deduce that

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^t) \right\} \right| \geq \delta + C_{11}(1+T)^4 \nu \right) \leq \frac{T}{\nu} \exp\left(-\frac{|\mathcal{A}|\delta^2}{C_8(1+T)^4} \right). \tag{A.19}$$

Pick $\nu = 1/\sqrt{|\mathcal{A}|}$ and $\delta = C_8(1+T)^2(\sqrt{\log(|\mathcal{A}|T)} + z)/\sqrt{|\mathcal{A}|}$. Thus, with probability at least $1 - e^{-z^2}$, we have that

$$\sup_{k\in[T/\alpha]} \left| L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^T) - \mathbb{E}_{\rho_0}\left\{ L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^T) \right\} \right| \leq C_{12}\,(1+T)^3 \frac{\sqrt{\log|\mathcal{A}|} + z}{\sqrt{|\mathcal{A}|}}. \tag{A.20}$$

By combining (A.14) and (A.20), we conclude that, with probability at least $1 - e^{-z^2}$,

$$\sup_{k\in[T/\alpha]} |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}^T) - \bar{L}(\rho_T)| \leq C_{13}\,(1+T)^3 \frac{\sqrt{\log|\mathcal{A}|} + z}{\sqrt{|\mathcal{A}|}}. \tag{A.21}$$

Finally, by combining (A.4), (A.6), (A.10) and (A.21), the result readily follows. □

A.2. Part (B)

The proof of part (B) is obtained by combining part (A) with the following lemma.

Lemma A.1 (Dropout stability implies connectivity – two-layer)**.** *Consider a two-layer neural network with N neurons, as in (3.1). Given $\mathcal{A} = [N/2]$, let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be ε-dropout stable as in Definition 3.1. Then, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are ε-connected as in Definition 3.2. Furthermore, the path connecting $\boldsymbol{\theta}$ with $\boldsymbol{\theta}'$ consists of 7 line segments.*

Proof of Lemma A.1. Let $\boldsymbol{\theta} = ((a_1, \boldsymbol{w}_1), (a_2, \boldsymbol{w}_2), \ldots, (a_N, \boldsymbol{w}_N))$ and $\boldsymbol{\theta}' = ((a_1', \boldsymbol{w}_1'), (a_2', \boldsymbol{w}_2'), \ldots, (a_N', \boldsymbol{w}_N'))$. For the moment, assume that N is even. Consider the piecewise linear path in parameter space that connects $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ via the

following intermediate points:

$$\begin{aligned}
\boldsymbol{\theta}_1 &= ((2a_1, \boldsymbol{w}_1), (2a_2, \boldsymbol{w}_2), \ldots, (2a_{N/2}, \boldsymbol{w}_{N/2}), (0, \boldsymbol{w}_{N/2+1}), (0, \boldsymbol{w}_{N/2+2}), \ldots, (0, \boldsymbol{w}_N)),\\
\boldsymbol{\theta}_2 &= ((2a_1, \boldsymbol{w}_1), (2a_2, \boldsymbol{w}_2), \ldots, (2a_{N/2}, \boldsymbol{w}_{N/2}), (0, \boldsymbol{w}'_1), (0, \boldsymbol{w}'_2), \ldots, (0, \boldsymbol{w}'_{N/2})),\\
\boldsymbol{\theta}_3 &= ((0, \boldsymbol{w}_1), (0, \boldsymbol{w}_2), \ldots, (0, \boldsymbol{w}_{N/2}), (2a'_1, \boldsymbol{w}'_1), (2a'_2, \boldsymbol{w}'_2), \ldots, (2a'_{N/2}, \boldsymbol{w}'_{N/2})),\\
\boldsymbol{\theta}_4 &= ((0, \boldsymbol{w}'_1), (0, \boldsymbol{w}'_2), \ldots, (0, \boldsymbol{w}'_{N/2}), (2a'_1, \boldsymbol{w}'_1), (2a'_2, \boldsymbol{w}'_2), \ldots, (2a'_{N/2}, \boldsymbol{w}'_{N/2})),\\
\boldsymbol{\theta}_5 &= ((2a'_1, \boldsymbol{w}'_1), (2a'_2, \boldsymbol{w}'_2), \ldots, (2a'_{N/2}, \boldsymbol{w}'_{N/2}), (0, \boldsymbol{w}'_1), (0, \boldsymbol{w}'_2), \ldots, (0, \boldsymbol{w}'_{N/2})),\\
\boldsymbol{\theta}_6 &= ((2a'_1, \boldsymbol{w}'_1), (2a'_2, \boldsymbol{w}'_2), \ldots, (2a'_{N/2}, \boldsymbol{w}'_{N/2}), (0, \boldsymbol{w}'_{N/2+1}), (0, \boldsymbol{w}'_{N/2+2}), \ldots, (0, \boldsymbol{w}'_N)).
\end{aligned} \tag{A.22}$$

We will now show that the loss along this path is upper bounded by $\max(L_N(\boldsymbol{\theta}), L_N(\boldsymbol{\theta}')) + \varepsilon$.

Consider the path that connects $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_1$. As $\boldsymbol{\theta}$ is ε-dropout stable, we have that $L_N(\boldsymbol{\theta}_1) \leq L_N(\boldsymbol{\theta}) + \varepsilon$. As the loss is convex in the weights of the last layer, the loss along this path is upper bounded by $L_N(\boldsymbol{\theta}) + \varepsilon$. Similarly, the loss along the path that connects $\boldsymbol{\theta}_6$ to $\boldsymbol{\theta}'$ is upper bounded by $L_N(\boldsymbol{\theta}') + \varepsilon$.

Consider the path that connects $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. Here, we change $\boldsymbol{w}$'s only when the corresponding a's are 0. Thus, the loss does not change along the path. Similarly, the loss does not change along the path that connects $\boldsymbol{\theta}_3$ to $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ to $\boldsymbol{\theta}_6$.

Consider the path that connects $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_3$. Note that $L_N(\boldsymbol{\theta}_3) = L_N(\boldsymbol{\theta}_5)$. As the loss is convex in the weights of the last layer, the loss along this path is upper bounded by $\max(L_N(\boldsymbol{\theta}), L_N(\boldsymbol{\theta}')) + \varepsilon$.

Finally, consider the path that connects $\boldsymbol{\theta}_4$ to $\boldsymbol{\theta}_5$. Here, we are interpolating between two equal subnetworks. Thus, the loss along this path does not change. This concludes the proof for even N.

If N is odd, a similar argument can be carried out. The differences are that *(i)* the $\lceil N/2 \rceil$-th parameter of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ is $(0, \boldsymbol{w}_{N/2})$ and the $\lceil N/2 \rceil$-th parameter of $\boldsymbol{\theta}_4$, $\boldsymbol{\theta}_5$ and $\boldsymbol{\theta}_6$ is $(0, \boldsymbol{w}'_{N/2})$, and *(ii)* the constant 2 in front of the a_i is replaced by $N/\lfloor N/2 \rfloor$. □

B. Extension to Unbounded Activation – Statement and Proof

We modify the assumptions **(A2)**, **(A3)** and **(A4)** of Section 3.2 as follows:

(A2') The feature vectors $\boldsymbol{x}$ and the response variables y are bounded by K_2, and the gradient $\nabla_{\boldsymbol{w}} \sigma(\boldsymbol{x}, \boldsymbol{w})$ is K_2 sub-gaussian when $\boldsymbol{x} \sim \mathbb{P}$.

(A3') The activation function σ is differentiable, with gradient bounded by K_3 and K_3-Lipschitz.

(A4') The initialization ρ_0 is supported on $\|\boldsymbol{\theta}_i^0\|_2 \leq K_4$.

We are now ready to present our results for unbounded activations in the two-layer setting.

Theorem 1 (Two-layer, unbounded activation)**.** *Assume that conditions **(A1), (A2'), (A3')** and **(A4')** hold, and fix $T \geq 1$. Let $\boldsymbol{\theta}^k$ be obtained by running k steps of the SGD algorithm (3.3) with data $\{(\boldsymbol{x}_j, y_j)\}_{j=0}^k \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and initialization ρ_0. Assume further that the loss at each step of SGD is uniformly bounded, i.e., $\max_{j \in \{0,\ldots,k\}} |y_j - \hat{\boldsymbol{y}}_N(\boldsymbol{x}_j, \boldsymbol{\theta}^j)| \leq K_5$. Then, the results of Theorem 1 hold, with*

$$\begin{aligned}
\varepsilon_{\mathrm{D}} &= K(T) \left(\frac{\sqrt{\log |\mathcal{A}|} + z}{\sqrt{|\mathcal{A}|}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big) \right),\\
\varepsilon_{\mathrm{C}} &= K(\max(T, T')) \left(\frac{\sqrt{\log N} + z}{\sqrt{N}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big) \right).
\end{aligned} \tag{B.1}$$

where the constant $K(T)$ depends on T and on the constants K_i of the assumptions.

To prove the result, we crucially rely on the following bound on the norm of the parameters evolved via SGD.

Lemma B.1 (Bound on norm of SGD parameters)**.** *Under the assumptions of Theorem 1, we have that*

$$\sup_{s \in [T/\alpha]} \max_{i \in [N]} \|\boldsymbol{\theta}_i^s\|_2 \leq K e^{K T}, \tag{B.2}$$

where the constant K depends only on the constants K_i of the assumptions.

Proof of Lemma B.1. The SGD update at step $j+1$ gives:

$$\begin{aligned} a_i^{j+1} &= a_i^j + 2\alpha\,\xi(j\alpha)\cdot(y_j - f_N(\boldsymbol{x}_j,\boldsymbol{\theta}^j))\cdot\sigma(\boldsymbol{x}_j,\boldsymbol{w}_i^j),\\ \boldsymbol{w}_i^{j+1} &= \boldsymbol{w}_i^j + 2\alpha\,\xi(j\alpha)\cdot(y_j - f_N(\boldsymbol{x}_j,\boldsymbol{\theta}^j))\cdot a_i^j\nabla_{\boldsymbol{w}_i}\sigma(\boldsymbol{x}_j,\boldsymbol{w}_i^j). \end{aligned} \tag{B.3}$$

We bound the absolute value of the increment $|a_i^{j+1} - a_i^j|$ as

$$\begin{aligned} |a_i^{j+1} - a_i^j| &\le 2\alpha\xi(j\alpha)\cdot|y_j - f_N(\boldsymbol{x}_j,\boldsymbol{\theta}^j)|\cdot|\sigma(\boldsymbol{x}_j,\boldsymbol{w}_i^j)|\\ &\overset{(a)}{\le} \alpha C_1|\sigma(\boldsymbol{x}_j,\boldsymbol{w}_i^j)|\\ &\overset{(b)}{\le} \alpha C_2(\|\boldsymbol{w}_i^j\|_2+1), \end{aligned} \tag{B.4}$$

where the constant C_i depends only on K_i, in (a) we use that ξ is bounded by K_1 and $|y_j - f_N(\boldsymbol{x}_j,\boldsymbol{\theta}^j)|\le K_5$, in (b) we use that $\|\sigma\|_{\mathrm{Lip}}\le K_2$ and $\|\boldsymbol{x}_j\|_2\le K_2$. Similarly, we bound the absolute value of the increments $\|\boldsymbol{w}_i^{j+1} - \boldsymbol{w}_i^j\|_2$ as

$$\|\boldsymbol{w}_i^{j+1} - \boldsymbol{w}_i^j\|_2 \le \alpha C_3|a_i^j|. \tag{B.5}$$

By combining (B.4) and (B.5), we get

$$\|\boldsymbol{\theta}_i^{j+1} - \boldsymbol{\theta}_i^j\|_2 \le \|\boldsymbol{w}_i^{j+1} - \boldsymbol{w}_i^j\|_2 + |a_i^{j+1} - a_i^j| \le \alpha C_4(\|\boldsymbol{\theta}_i^j\|_2+1). \tag{B.6}$$

By triangle inequality, we also obtain that

$$\|\boldsymbol{\theta}_i^s\|_2 \le \sum_{j=0}^{s-1}\|\boldsymbol{\theta}_i^{j+1} - \boldsymbol{\theta}_i^j\|_2 + \|\boldsymbol{\theta}_i^0\|_2. \tag{B.7}$$

As $\|\boldsymbol{\theta}_i^0\|_2$ is bounded, by combining (B.6) and (B.7), we have that

$$\|\boldsymbol{\theta}_i^s\|_2 \le C_5 + C_5\,s\alpha + C_5\alpha\sum_{j=0}^{s-1}\|\boldsymbol{\theta}_i^j\|_2. \tag{B.8}$$

By using a discrete version of Gronwall's inequality, the result follows. □

Finally, let us present the proof of Theorem 1.

Proof of Theorem 1. Since the activation function σ satisfies assumption **(A3')**, we can construct $\tilde\sigma : \mathbb{R}^d\times\mathbb{R}^{D-1}\to\mathbb{R}$ that satisfies the following two properties:

(i) $\tilde\sigma(\boldsymbol{x},\boldsymbol{w})$ coincides with $\sigma(\boldsymbol{x},\boldsymbol{w})$ for $\|\boldsymbol{x}\|_2\le K_2$ and $\|\boldsymbol{w}\|_2\le Ke^{KT}$, where K_2 is the constant of assumption **(A2')** and Ke^{KT} is the bound of Lemma B.1;

(ii) $\tilde\sigma(\boldsymbol{x},\boldsymbol{w})$ is bounded, differentiable, with bounded and Lipschitz continuous gradient.

Recall that $\boldsymbol{\theta}^k$ is obtained by running k steps of the SGD algorithm (3.3) with initial condition $\boldsymbol{\theta}^0$, data $\{\boldsymbol{x}_j,y_j\}_{j=0}^k$ and activation function σ. Let $\tilde{\boldsymbol{\theta}}^k$ be obtained by running k steps of SGD with initial condition $\boldsymbol{\theta}^0$, data $\{\boldsymbol{x}_j,y_j\}_{j=0}^k$ and activation function $\tilde\sigma$. By combining Lemma B.1, assumption **(A2')** and property **(i)** of $\tilde\sigma$, we immediately deduce that

$$\boldsymbol{\theta}^k = \tilde{\boldsymbol{\theta}}^k. \tag{B.9}$$

Furthermore, we have that

$$\mathbb{E}\left\{\Big(y - \frac{1}{N}\sum_{i=1}^N a_i\sigma(\boldsymbol{x},\boldsymbol{w}_i)\Big)\Big)^2\right\} = \mathbb{E}\left\{\Big(y - \frac{1}{N}\sum_{i=1}^N a_i\tilde\sigma(\boldsymbol{x},\boldsymbol{w}_i)\Big)\Big)^2\right\}, \tag{B.10}$$

namely the loss of $\boldsymbol{\theta}^k$ computed with respect to the activation function σ is the same as the loss of $\boldsymbol{\theta}^k$ computed with respect to the activation function $\tilde{\sigma}$.

Note that $\|\tilde{\sigma}\|_\infty \leq C_1(T)$ for some $C_1(T)$ that depends on T and on the constants K_i of the assumptions. Thus, $\tilde{\sigma}$ satisfies assumptions **(A2)** and **(A3)**, with K_3 depending on time T of the evolution. Consequently, by Theorem 1, with probability at least $1 - e^{-z^2}$, for all $k \in [T/\alpha]$, $\tilde{\boldsymbol{\theta}}^k$ is ε_{D}-dropout stable, with

$$\varepsilon_{\mathrm{D}} = K(T)\left(\sqrt{\frac{\log N}{N}} + \frac{\sqrt{\log|\mathcal{A}|} + z}{\sqrt{|\mathcal{A}|}} + \sqrt{\alpha}\big(\sqrt{D + \log N} + z\big)\right). \tag{B.11}$$

By using (B.9) and (B.10), we conclude that, with probability at least $1 - e^{-z^2}$, for all $k \in [T/\alpha]$, $\boldsymbol{\theta}^k$ is ε_{D}-dropout stable. Similarly, with probability at least $1 - e^{-z^2}$, for all $k' \in [T'/\alpha]$, $(\boldsymbol{\theta}')^{k'}$ is ε_{D}-dropout stable. Thus, by Lemma A.1, the proof is complete. □

C. Proof of Theorem 2

C.1. Part (A)

Let $D = \sum_{i=0}^{L} D_i$ and let ρ be a probability measure over $\mathbb{R}^D \cong \mathbb{R}^{D_0} \times \mathbb{R}^{D_1} \times \cdots \times \mathbb{R}^{D_L}$. For $i \in \{0, \ldots, L\}$, we denote by $\rho^{(i)}$ the marginal of ρ over the i-th factor $\mathbb{R}^{D_i}$ of the Cartesian product. For $i \in \{0, \ldots, L-1\}$, we denote by $\rho^{(i,i+1)}$ the marginal of ρ over the i-th and the $i+1$-th factors. Furthermore, we denote by $\rho^{(i|i+1)}(\cdot \mid \boldsymbol{\theta}^{(i+1)})$ the conditional distribution of the i-th factor given that the $i+1$-th factor is equal to $\boldsymbol{\theta}^{(i+1)}$.

Given a feature vector $\boldsymbol{x} \in \mathbb{R}^{d_0}$ and a probability measure ρ over $\mathbb{R}^D$, we define

$$\begin{aligned}
\bar{\boldsymbol{z}}^{(2)}(\boldsymbol{x}, \rho) &= \int \sigma^{(1)}\left(\sigma^{(0)}\left(\boldsymbol{x}, \boldsymbol{\theta}^{(0)}\right), \boldsymbol{\theta}^{(1)}\right) \mathrm{d}\rho^{(0,1)}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}), \\
\bar{\boldsymbol{z}}^{(\ell)}(\boldsymbol{x}, \rho) &= \int \sigma^{(\ell-1)}\left(\bar{\boldsymbol{z}}^{(\ell-1)}(\boldsymbol{x}, \rho), \boldsymbol{\theta}^{(\ell-1)}\right) \mathrm{d}\rho^{(\ell-1)}(\boldsymbol{\theta}^{(\ell-1)}), \qquad \ell \in \{3, \ldots, L-1\}, \\
\bar{\boldsymbol{z}}^{(L)}\left(\boldsymbol{x}, \rho, \boldsymbol{\theta}^{(L)}\right) &= \int \sigma^{(L-1)}\left(\bar{\boldsymbol{z}}^{(L-1)}(\boldsymbol{x}, \rho), \boldsymbol{\theta}^{(L-1)}\right) \mathrm{d}\rho^{(L-1|L)}(\boldsymbol{\theta}^{(L-1)} \mid \boldsymbol{\theta}^{(L)}), \\
\bar{\boldsymbol{y}}(\boldsymbol{x}, \rho) &= \sigma^{(L+1)}\left(\int \sigma^{(L)}\left(\bar{\boldsymbol{z}}^{(L)}\left(\boldsymbol{x}, \rho, \boldsymbol{\theta}^{(L)}\right), \boldsymbol{\theta}^{(L)}\right) \mathrm{d}\rho^{(L)}(\boldsymbol{\theta}^{(L)})\right),
\end{aligned} \tag{C.1}$$

where $\sigma^{(\ell)} : \mathbb{R}^{d_\ell} \times \mathbb{R}^{D_\ell} \to \mathbb{R}^{d_{\ell+1}}$, with $\ell \in \{0, \ldots, L\}$, and $\sigma^{(L+1)} : \mathbb{R}^{d_{L+1}} \to \mathbb{R}^{d_{L+1}}$. We remark that $\bar{\boldsymbol{z}}^{(L)}$ is defined in terms of the conditional distribution $\rho^{(L-1|L)}$. We also define the limit loss as

$$\bar{L}(\rho) = \mathbb{E}\left\{\|\boldsymbol{y} - \bar{\boldsymbol{y}}(\boldsymbol{x}, \rho)\|_2^2\right\}, \tag{C.2}$$

where the expectation is taken over $(\boldsymbol{x}, \boldsymbol{y})$. Given a probability measure ρ_0 over $\mathbb{R}^D$ and activation functions $\sigma^{(\ell)}$ ($\ell \in \{0, \ldots, L+1\}$), we denote by $\rho^\star_{[0,T]}$ the probability measure over $\mathcal{C}([0,T], \mathbb{R}^D)$ which solves the McKean-Vlasov DNN problem with initial condition ρ_0, according to Definition 4.4 of (Araújo et al., 2019). We also denote by $\rho^\star_t$ the marginal of $\rho^\star_{[0,T]}$ at time $t \in [0,T]$.

In (Araújo et al., 2019), it is considered a model of neural network with $L+1 \geq 4$ layers, where each hidden layer contains N neurons. This model can be obtained from (4.1) by setting to one the parameters $\{\boldsymbol{a}^{\ell}_{i_\ell, i_{\ell+1}}\}_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]}$ and $\{\boldsymbol{a}^{(L)}_{i_L}\}_{i_L \in [N]}$, and by applying the bounded activation function $\sigma^{(L+1)}$ to the output $\widehat{\boldsymbol{y}}_N$. Then, it is studied the evolution under the SGD algorithm (4.3) of the parameters $\boldsymbol{\theta}(k)$ of this multilayer neural network. In particular, it is shown that, under suitable assumptions, *(i)* the solution of the McKean-Vlasov DNN problem exists and it is unique, *(ii)* the parameters $\boldsymbol{\theta}(k)$ obtained after k steps of SGD with step size α are close to particles $\bar{\boldsymbol{\theta}}(t)$ at time $t = k\alpha$, whose trajectories are distributed according to $\rho^\star_t$, and *(iii)* the loss $L_N(\boldsymbol{\theta}(k))$ concentrates to the limit loss $\bar{L}(\rho^\star_t)$.

In order to prove Theorem 2, we will use the following bound on the norm of the parameters $\{\boldsymbol{a}^{(\ell)}_{i_\ell, i_{\ell+1}}\}_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]}$ evolved via SGD.

Lemma C.1 (Bound on norm of $\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}$)**.** *Under the assumptions of Theorem 2, we have that*

$$\max_{\ell\in[L-1]}\sup_{s\in[T/\alpha]}\max_{i_\ell,i_{\ell+1}\in[N]}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(s)\|_2\le K(T,L), \tag{C.3}$$

where the constant K depends only on T, L and on the constants K_i of the assumptions.

Proof. For $\ell\in[L-1]$, the SGD update at step $j+1$ gives:

$$\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)=\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)+2\alpha\xi(j\alpha)N^2\left(\boldsymbol{y}_j-\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_j,\boldsymbol{\theta}(j)\right)\right)^{\mathsf{T}}\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right), \tag{C.4}$$

where $\mathrm{D}_{\boldsymbol{\theta}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\in\mathbb{R}^{d_{L+1}}\times\mathbb{R}^{D_\ell+d_{\ell+1}}$ denotes the Jacobian of $\widehat{\boldsymbol{y}}_N$ with respect to $\boldsymbol{\theta}_{i_\ell,i_{\ell+1}}^{(\ell)}$.

Recall that by assumptions **(B2)-(B3)** the response variables $\boldsymbol{y}_j$ and the activation $\sigma^{(L)}$ are bounded. Moreover, as the final layer of the network is not trained, i.e., $\boldsymbol{a}_{i_L}^{(L)}(k+1)=\boldsymbol{a}_{i_L}^{(L)}(k)$ for any k, and $\boldsymbol{a}_{i_L}^{(L)}(0)$ is initialized with a distribution supported on $\|\boldsymbol{a}_{i_L}^{(L)}(0)\|_2\le K_4$, we get that $\boldsymbol{a}_{i_L}^{(L)}$ is bounded along the whole SGD trajectory. Thus, we are able to conclude that $\left\|\boldsymbol{y}_j-\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_j,\boldsymbol{\theta}(j)\right)\right\|_2\le K_5$, for some constant K_5.

We bound the absolute value of the increment $\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)-\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\|_2$ as

$$\begin{aligned}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)-\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\|_2&\le 2\,\alpha\,\xi(j\alpha)\,N^2\left\|\boldsymbol{y}_j-\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_j,\boldsymbol{\theta}(j)\right)\right\|_2\cdot\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}\\&\le\alpha\,N^2\,C_1\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}},\end{aligned} \tag{C.5}$$

where we use that ξ is bounded by K_1 and $\left\|\boldsymbol{y}_j-\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_j,\boldsymbol{\theta}(j)\right)\right\|_2\le K_5$. Consequently,

$$\max_{i_\ell,i_{\ell+1}\in[N]}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)-\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\|_2\le\alpha\,N^2\,C_1\max_{i_\ell,i_{\ell+1}\in[N]}\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}. \tag{C.6}$$

Let us now focus on the operator norm of the Jacobian. First, we write

$$\begin{aligned}\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}&=\left\|\mathrm{D}_{\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\cdot\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}\\&\le\left\|\mathrm{D}_{\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}\cdot\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}},\end{aligned} \tag{C.7}$$

where the inequality uses the fact that the operator norm is sub-multiplicative. Note that

$$\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)=\mathrm{diag}\left(\frac{1}{N}\sigma^{(\ell)}\left(\boldsymbol{z}_{i_\ell}^{(\ell)}\left(\boldsymbol{x},\boldsymbol{\theta}\right),\boldsymbol{w}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\right)\right), \tag{C.8}$$

where we denote by $\mathrm{diag}(\boldsymbol{v})$ the diagonal matrix containing $\boldsymbol{v}$ on the diagonal. As $\sigma^{(\ell)}$ is bounded by assumption **(B3)**, we have that

$$\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)\right\|_{\mathrm{op}}\le\frac{C_2}{N}. \tag{C.9}$$

Furthermore, the Jacobian $\mathrm{D}_{\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)$ is given by

$$\begin{aligned}\mathrm{D}_{\boldsymbol{z}_{i_L}^{(L)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)&=\frac{1}{N}\boldsymbol{M}_{i_L}^{(L)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\qquad i_L\in[N],\\\mathrm{D}_{\boldsymbol{z}_{i_\ell}^{(\ell)}}\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x},\boldsymbol{\theta}(j)\right)&=\frac{1}{N^{L-\ell+1}}\sum_{\boldsymbol{p}_{\ell+1}^L\in[N]^{L-\ell}}\boldsymbol{M}_{i_\ell,\boldsymbol{p}_{\ell+1}^L}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\qquad\ell\in[L-1],i_\ell\in[N],\end{aligned} \tag{C.10}$$

where $\boldsymbol{p}_{\ell+1}^{L}$ denotes the multi-index $(p_{\ell+1},\dots,p_L)$, $[N]^{L-\ell}$ denotes the $(L-\ell)$-fold Cartesian product of $[N]$ and the matrices $\boldsymbol{M}_{\boldsymbol{p}_\ell^L}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j))$ are defined recursively as

$$\begin{aligned}
\boldsymbol{M}_{p_L}^{(L)}(\boldsymbol{x},\boldsymbol{\theta}(j)) &= \mathrm{D}_{\boldsymbol{z}_{p_L}^{(L)}}\left(\boldsymbol{a}_{p_L}^{(L)}(j)\odot\sigma^{(L)}\left(\boldsymbol{z}_{p_L}^{(L)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\boldsymbol{w}_{p_L}^{(L)}(j)\right)\right)\\
&= \mathrm{diag}(\boldsymbol{a}_{p_L}^{(L)}(j))\cdot\mathrm{D}_{\boldsymbol{z}_{p_L}^{(L)}}\sigma^{(L)}\left(\boldsymbol{z}_{p_L}^{(L)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\boldsymbol{w}_{p_L}^{(L)}(j)\right),\\
\boldsymbol{M}_{\boldsymbol{p}_\ell^L}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j)) &= \boldsymbol{M}_{\boldsymbol{p}_{\ell+1}^L}^{(\ell+1)}(\boldsymbol{x},\boldsymbol{\theta}(j))\cdot\mathrm{D}_{\boldsymbol{z}_{p_\ell}^{(\ell)}}\left(\boldsymbol{a}_{p_\ell,p_{\ell+1}}^{(\ell)}(j)\odot\sigma^{(\ell)}\left(\boldsymbol{z}_{p_\ell}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\boldsymbol{w}_{p_\ell,p_{\ell+1}}^{(\ell)}(j)\right)\right)\\
&= \boldsymbol{M}_{\boldsymbol{p}_{\ell+1}^L}^{(\ell+1)}(\boldsymbol{x},\boldsymbol{\theta}(j))\cdot\mathrm{diag}(\boldsymbol{a}_{p_\ell,p_{\ell+1}}^{(\ell)}(j))\cdot\mathrm{D}_{\boldsymbol{z}_{p_\ell}^{(\ell)}}\sigma^{(\ell)}\left(\boldsymbol{z}_{p_\ell}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j)),\boldsymbol{w}_{p_\ell,p_{\ell+1}}^{(\ell)}(j)\right).
\end{aligned}\tag{C.11}$$

Note that $\boldsymbol{a}_{p_L}^{(L)}(j)=\boldsymbol{a}_{p_L}^{(L)}(0)$ and recall that $\|\boldsymbol{a}_{p_L}^{(L)}(0)\|_2$ is bounded by assumption **(B4)**. Furthermore, $\sigma^{(\ell)}$ has bounded Fréchet derivative by assumption **(B3)**. Thus, we deduce that

$$\left\|\boldsymbol{M}_{p_L}^{(L)}(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}}\le C_3,\tag{C.12}$$

and

$$\begin{aligned}
\left\|\boldsymbol{M}_{\boldsymbol{p}_\ell^L}^{(\ell)}(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}} &\le C_4(L)\prod_{m=\ell}^{L-1}\|\boldsymbol{a}_{p_m,p_{m+1}}^{(m)}(j)\|_2\\
&\le C_4(L)\prod_{m=\ell}^{L-1}\max_{i_m,i_{m+1}\in[N]}\|\boldsymbol{a}_{i_m,i_{m+1}}^{(m)}(j)\|_2.
\end{aligned}\tag{C.13}$$

Consequently, we have that

$$\begin{aligned}
\left\|\mathrm{D}_{\boldsymbol{z}_{i_L}^{(L)}}\widehat{\boldsymbol{y}}_N(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}} &\le \frac{C_3}{N},\\
\left\|\mathrm{D}_{\boldsymbol{z}_{i_\ell}^{(\ell)}}\widehat{\boldsymbol{y}}_N(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}} &\le \frac{C_4(L)}{N}\prod_{m=\ell}^{L-1}\max_{i_m,i_{m+1}\in[N]}\|\boldsymbol{a}_{i_m,i_{m+1}}^{(m)}(j)\|_2.
\end{aligned}\tag{C.14}$$

By combining (C.7), (C.9) and (C.14), we obtain that

$$\begin{aligned}
\left\|\mathrm{D}_{\boldsymbol{a}_{i_{L-1},i_L}^{(L-1)}}\widehat{\boldsymbol{y}}_N(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}} &\le \frac{C_5}{N^2},\\
\left\|\mathrm{D}_{\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}}\widehat{\boldsymbol{y}}_N(\boldsymbol{x},\boldsymbol{\theta}(j))\right\|_{\mathrm{op}} &\le \frac{C_6(L)}{N^2}\prod_{m=\ell+1}^{L-1}\max_{i_m,i_{m+1}\in[N]}\|\boldsymbol{a}_{i_m,i_{m+1}}^{(m)}(j)\|_2,\qquad \ell\in[L-2].
\end{aligned}\tag{C.15}$$

By using also (C.6), we have that

$$\begin{aligned}
\max_{i_{L-1},i_L\in[N]}\|\boldsymbol{a}_{i_{L-1},i_L}^{(L-1)}(j+1)-\boldsymbol{a}_{i_{L-1},i_L}^{(L-1)}(j)\|_2 &\le \alpha\, C_7,\\
\max_{i_\ell,i_{\ell+1}\in[N]}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)-\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\|_2 &\le \alpha\, C_8(L)\prod_{m=\ell+1}^{L-1}\max_{i_m,i_{m+1}\in[N]}\|\boldsymbol{a}_{i_m,i_{m+1}}^{(m)}(j)\|_2,\qquad \ell\in[L-2].
\end{aligned}\tag{C.16}$$

By triangle inequality, we also obtain that, for $\ell\in[L-1]$ and $i_\ell,i_{\ell+1}\in[N]$,

$$\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(s)\|_2\le\sum_{j=0}^{s-1}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j+1)-\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(j)\|_2+\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(0)\|_2.\tag{C.17}$$

As $\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(0)\|_2$ and $\|\boldsymbol{a}_{i_L}^{(L)}(0)\|_2$ are bounded, by combining (C.16) and (C.17), we have that

$$\begin{aligned}
\max_{s\in[k]}\max_{i_{L-1},i_L\in[N]}\|\boldsymbol{a}_{i_{L-1},i_L}^{(L-1)}(s)\|_2 &\le C+C_7\,T,\\
\max_{s\in[k]}\max_{i_\ell,i_{\ell+1}\in[N]}\|\boldsymbol{a}_{i_\ell,i_{\ell+1}}^{(\ell)}(s)\|_2 &\le C+C_8(L)\,T\prod_{m=\ell+1}^{L-1}\max_{s\in[k]}\max_{i_m,i_{m+1}\in[N]}\|\boldsymbol{a}_{i_m,i_{m+1}}^{(m)}(s)\|_2,\qquad \ell\in[L-2].
\end{aligned}\tag{C.18}$$

where we have used that $T = k\alpha$. By doing a step of induction on $\ell \in \{L-2, L-3, \dots, 1\}$, the proof is complete. $\square$

We are now ready to provide the proof of Theorem 2, part (A).

Proof of Theorem 2, part (A). For $\ell \in [L]$, we construct $\tilde{\sigma}^{(\ell)} : \mathbb{R}^{d_\ell} \times \mathbb{R}^{D_\ell + d_{\ell+1}} \to \mathbb{R}^{d_{\ell+1}}$ that satisfies the following two properties:

(i) $\tilde{\sigma}^{(\ell)}(\boldsymbol{z}, (\boldsymbol{w}, \boldsymbol{a}))$ coincides with $\boldsymbol{a} \odot \sigma^{(\ell)}(\boldsymbol{z}, \boldsymbol{w})$ for all $(\boldsymbol{z}, \boldsymbol{w}) \in \mathbb{R}^{d_\ell} \times \mathbb{R}^{D_\ell}$ and for $\|\boldsymbol{a}\|_2 \leq K(T, L)$, where $K(T, L)$ is the bound of Lemma C.1;

(ii) $\tilde{\sigma}^{(\ell)}$ is bounded, with Fréchet derivatives bounded and Lipschitz.

Similarly, we construct $\tilde{\sigma}^{(L+1)} : \mathbb{R}^{d_{L+1}} \to \mathbb{R}^{d_{L+1}}$ that satisfies the following two properties:

(i) $\tilde{\sigma}^{(L+1)}(\boldsymbol{z}) = \boldsymbol{z}$ for $\|\boldsymbol{z}\|_2 \leq K_3\, K_4$, where K_3 is the bound on $\sigma^{(L)}$ and K_4 is the bound on $\|\boldsymbol{a}_{i_L}^{(L)}(0)\|_2$ (see assumptions **(B3)**-**(B4)**);

(ii) $\tilde{\sigma}^{(L+1)}$ is bounded, with Fréchet derivatives bounded and Lipschitz.

Define

$$
\begin{aligned}
(\boldsymbol{z}_{i_1}^{(1)})'(\boldsymbol{x}, \boldsymbol{\theta}) &= \sigma^{(0)}\left(\boldsymbol{x}, \boldsymbol{\theta}_{i_1}^{(0)}\right), \qquad i_1 \in [N], \\
(\boldsymbol{z}_{i_{\ell+1}}^{(\ell+1)})'(\boldsymbol{x}, \boldsymbol{\theta}) &= \frac{1}{N}\sum_{i_\ell=1}^{N} \tilde{\sigma}^{(\ell)}\left((\boldsymbol{z}_{i_\ell}^{(\ell)})'(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}\right), \qquad \ell \in [L-1],\, i_{\ell+1} \in [N], \\
\widehat{\boldsymbol{y}}_N'(\boldsymbol{x}, \boldsymbol{\theta}) &= \tilde{\sigma}^{(L+1)}\left(\frac{1}{N}\sum_{i_L=1}^{N} \tilde{\sigma}^{(L)}\left((\boldsymbol{z}_{i_L}^{(L)})'(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\theta}_{i_L}^{(L)}\right)\right),
\end{aligned}
\tag{C.19}
$$

and

$$
L_N'(\boldsymbol{\theta}) = \mathbb{E}\left\{\left\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_N'(\boldsymbol{x}, \boldsymbol{\theta})\right\|_2^2\right\}. \tag{C.20}
$$

Let $\boldsymbol{\theta}'(k)$ be obtained by running k steps of the SGD algorithm (4.3) with $\widehat{\boldsymbol{y}}_N(\boldsymbol{x}, \boldsymbol{\theta})$ replaced by $\widehat{\boldsymbol{y}}_N'(\boldsymbol{x}, \boldsymbol{\theta})$. Recall that $\boldsymbol{a}_{i_\ell, i_{\ell+1}}^{(\ell)}(s)$ is bounded by Lemma C.1, $\boldsymbol{a}_{i_L}^{(L)}(s)$ is bounded by assumption **(B4)** and $\sigma^{(\ell)}$ is bounded by assumption **(B3)**. Thus, we have that $\boldsymbol{\theta}'(k) = \boldsymbol{\theta}(k)$ and $L_N'(\boldsymbol{\theta}'(k)) = L_N(\boldsymbol{\theta}(k))$. To simplify notation, in the rest of the proof we will drop the symbol $'$ from $\boldsymbol{\theta}$ and L_N. By definition of dropout stability, the proof is completed by showing that, with probability at least $1 - e^{-z^2}$,

$$
|L_N(\boldsymbol{\theta}(k)) - L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}(k))| \leq K(T, L)\left(\frac{\sqrt{d} + z}{\sqrt{N}} + \sqrt{\alpha}\big(\sqrt{d} + z\big)\right). \tag{C.21}
$$

By construction, the activation functions $\{\tilde{\sigma}^{(\ell)}\}_{\ell \in [L+1]}$ are bounded, with Fréchet derivatives that are bounded and Lipschitz. Thus, the technical assumptions of (Araújo et al., 2019) are fulfilled. Let $\rho_{[0,T]}^\star$ denote the unique solution to the McKean-Vlasov DNN problem with initial condition ρ_0 and activation functions $\sigma^{(0)}$ and $\tilde{\sigma}^{(\ell)}$, with $\ell \in \{0, \dots, L+1\}$. Furthermore, let $\bar{\boldsymbol{\theta}}(t)$, with $t \in [0, T]$, be the associated ideal particles. Furthermore, let $\bar{\boldsymbol{\theta}}_{\mathrm{S}}(t)$ be obtained from $\bar{\boldsymbol{\theta}}(t)$ in the same way in which $\boldsymbol{\theta}_{\mathrm{S}}(k)$ is obtained from $\boldsymbol{\theta}(k)$. By triangle inequality, we have that

$$
\begin{aligned}
|L_N(\boldsymbol{\theta}(k)) - L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}(k))| &\leq |L_N(\boldsymbol{\theta}(k)) - \bar{L}(\rho_T^\star)| + |L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}(k)) - \bar{L}(\rho_T^\star)| \\
&\leq |L_N(\boldsymbol{\theta}(k)) - L_N(\bar{\boldsymbol{\theta}}(T))| + |L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}(k)) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}(T))| \\
&\quad |L_N(\bar{\boldsymbol{\theta}}(T)) - \bar{L}(\rho_T^\star)| + |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)) - \bar{L}(\rho_T^\star)|,
\end{aligned}
\tag{C.22}
$$

where $\rho_T^\star$ denotes the marginal of $\rho_{[0,T]}^\star$ at time T and $\bar{L}$ is defined in (C.2).

Given a vector of parameters $\boldsymbol{\theta}$ containing N_ℓ neurons in layer ℓ ($\ell \in [L]$), we define the norm

$$\|\boldsymbol{\theta}\|_\infty = \max\left(\sup_{i_1 \in [N_1]} \left\|\boldsymbol{\theta}_{i_1}^{(0)}\right\|_2, \sup_{\ell \in [L-1], i_\ell \in [N_\ell], i_{\ell+1} \in [N_{\ell+1}]} \left\|\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}\right\|_2, \sup_{i_L \in [N_L]} \left\|\boldsymbol{\theta}_{i_L}^{(L)}\right\|_2 \right). \tag{C.23}$$

As a preliminary result, we provide a bound on $\|\boldsymbol{\theta}(k) - \bar{\boldsymbol{\theta}}(T)\|_\infty$.

Consider the continuous time gradient descent process $\tilde{\boldsymbol{\theta}}(t)$, defined as

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{i_1}^{(0)}(t) &= \tilde{\boldsymbol{\theta}}_{i_1}^{(0)}(0), \\ \tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}(t) &= \tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}(0) + 2\int_0^t \alpha\xi(s) N^2 \mathbb{E}\left\{ \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right)\right)^\mathsf{T} \mathrm{D}_{\tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right) \right\} \mathrm{d}s, \\ \tilde{\boldsymbol{\theta}}_{i_L}^{(L)}(t) &= \tilde{\boldsymbol{\theta}}_{i_L}^{(L)}(0), \end{aligned} \tag{C.24}$$

with the initialization $\tilde{\boldsymbol{\theta}}_{i_1}^{(0)}(0) = \boldsymbol{\theta}_{i_1}^{(0)}(0)$, $\tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}(0) = \boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}(0)$ and $\tilde{\boldsymbol{\theta}}_{i_L}^{(L)}(0) = \boldsymbol{\theta}_{i_L}^{(L)}(0)$. By triangle inequality, we have that

$$\|\boldsymbol{\theta}(k) - \bar{\boldsymbol{\theta}}(T)\|_\infty \leq \|\boldsymbol{\theta}(k) - \tilde{\boldsymbol{\theta}}(T)\|_\infty + \|\tilde{\boldsymbol{\theta}}(T) - \bar{\boldsymbol{\theta}}(T)\|_\infty. \tag{C.25}$$

In order to bound the first term in the RHS of (C.25), we follow a strategy similar to that of Proposition 10.1 in (Araújo et al., 2019). From formula (10.8) of (Araújo et al., 2019), we have that

$$\begin{aligned} \left\|\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}(m) - \tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}(m\alpha)\right\|_2 \leq \alpha \left\|\mathrm{Mrt}_{i_\ell, i_{\ell+1}}^{(\ell)}(m)\right\|_2 + \\ \sum_{r=1}^m \int_{(r-1)\alpha}^{r\alpha} \mathbb{E}\Big\{ \Big\| \alpha\xi((r-1)\alpha) \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right) \\ - \alpha\xi(s) \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right)\right)^\mathsf{T} \mathrm{D}_{\tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right) \Big\|_2 \Big\} \mathrm{d}s, \end{aligned} \tag{C.26}$$

where

$$\begin{aligned} \mathrm{Mrt}_{i_\ell, i_{\ell+1}}^{(\ell)}(m) = \sum_{r=1}^m \alpha\xi((r-1)\alpha) \Big(\left(\boldsymbol{y}_{r-1} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_{r-1}, \boldsymbol{\theta}(r-1)\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}_{r-1}, \boldsymbol{\theta}(r-1)\right) \\ - \mathbb{E}\Big\{ \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right) \Big\} \Big) \end{aligned} \tag{C.27}$$

is a martingale with respect to the filtration $\{\mathcal{F}_m,\ m \in \mathbb{N}\}$ with $\mathcal{F}_m = \sigma\Big(\boldsymbol{\theta}(0), (\boldsymbol{x}_0, \boldsymbol{y}_0), \ldots, (\boldsymbol{x}_{m-1}, \boldsymbol{y}_{m-1})\Big)$. By taking the sup on both sides, we have that

$$\begin{aligned} \sup_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]} \left\|\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}(m) - \tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}(T)\right\|_2 \leq \overbrace{\alpha \sup_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]} \left\|\mathrm{Mrt}_{i_\ell, i_{\ell+1}}^{(\ell)}(m)\right\|_2}^{(I)} + \\ \overbrace{\sum_{r=1}^m \int_{(r-1)\alpha}^{r\alpha} \mathbb{E}\Big\{ \sup_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]} \Big\| \alpha\xi((r-1)\alpha) \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}(r-1)\right)}^{(II)} \\ - \alpha\xi(s) \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right)\right)^\mathsf{T} \mathrm{D}_{\tilde{\boldsymbol{\theta}}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}(s)\right) \Big\|_2 \Big\} \mathrm{d}s. \end{aligned} \tag{C.28}$$

Given two parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, by following the argument of Lemma B.17 of (Araújo et al., 2019), we have that

$$\begin{aligned} \Big\| \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right) - \left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right)\right)^\mathsf{T} \mathrm{D}_{\boldsymbol{\theta}_{i_\ell, i_{\ell+1}}^{(\ell)}} \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right) \Big\|_2 \\ \leq C_1 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty. \end{aligned} \tag{C.29}$$

In what follows, the C_i are constants that depend on L, T, and on the constants K_i of the assumptions.

Consequently, we can bound the second term in the RHS of (C.28) as

$$\textit{(II)} \leq C_2 \sum_{r=1}^{m} \int_{(r-1)\varepsilon}^{r\varepsilon} \Big(|(r-1)\varepsilon - s| + \|\boldsymbol{\theta}(r-1) - \tilde{\boldsymbol{\theta}}(s)\|_\infty \Big) \mathrm{d}s, \tag{C.30}$$

where we have used that the quantity

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}}_N(\boldsymbol{x}, \boldsymbol{\theta}))^{\mathsf{T}} \, \mathrm{D}_{\boldsymbol{\theta}^{(\ell)}_{i_\ell, i_{\ell+1}}} \widehat{\boldsymbol{y}}_N(\boldsymbol{x}, \boldsymbol{\theta}) \tag{C.31}$$

is bounded for all $\boldsymbol{\theta}$. By using also that the process $t \to \tilde{\boldsymbol{\theta}}(t)$ is Lipschitz in time, we obtain the bound

$$\textit{(II)} \leq C_3 \, \alpha \, T + C_3 \, \alpha \sum_{r=0}^{m-1} \left\| \boldsymbol{\theta}(r) - \tilde{\boldsymbol{\theta}}(r) \right\|_\infty . \tag{C.32}$$

By combining (C.32) with (C.28) and by applying a discrete Gronwall inequality, we have that

$$\left\| \boldsymbol{\theta}(k) - \tilde{\boldsymbol{\theta}}(T) \right\|_\infty \leq \alpha \, e^{C_3 \, T} \left(\sup_{m \in [k]} \|\mathrm{Mrt}(m)\|_\infty + C_3 \, T \right), \tag{C.33}$$

where we have defined

$$\|\mathrm{Mrt}(m)\|_\infty = \sup_{\ell \in [L-1], i_\ell, i_{\ell+1} \in [N]} \left\| \mathrm{Mrt}^{(\ell)}_{i_\ell, i_{\ell+1}}(m) \right\|_2 . \tag{C.34}$$

Note that $e^{\zeta \|\mathrm{Mrt}(m)\|_\infty}$ is a submartingale. By using a Cramér-Chernoff argument, we have that

$$\mathbb{P}\left(\sup_{m \in [k]} \|\mathrm{Mrt}_N(m)\|_\infty > u \right) \leq \inf_{\zeta \in \mathbb{R}^+} e^{-\zeta \cdot u} \mathbb{E}\left\{ e^{\zeta \|\mathrm{Mrt}(\tau)\|_\infty} \right\} \leq \inf_{\zeta \in \mathbb{R}^+} e^{-\zeta \cdot u} \mathbb{E}\left\{ e^{\zeta \|\mathrm{Mrt}(k)\|_\infty} \right\}, \tag{C.35}$$

where $\tau = \inf\{m \leq k, \|\mathrm{Mrt}_N(m)\|_\infty > u\} \wedge k$ is a stopping time, and in the second inequality we have applied the optional stopping theorem to the submartingale $e^{\zeta \|\mathrm{Mrt}(m)\|_\infty}$. Furthermore, for any $\zeta > 0$, we have that

$$\mathbb{E}\left\{ e^{\zeta \|\mathrm{Mrt}(k)\|_\infty} \right\} \leq \sum_{\ell=1}^{L} \sum_{i_\ell, i_{\ell+1}=1}^{N} \mathbb{E}\left\{ e^{\zeta \left\| \mathrm{Mrt}^{(\ell)}_{i_\ell, i_{\ell+1}}(k) \right\|_2} \right\}. \tag{C.36}$$

Note that the martingale $\mathrm{Mrt}^{(\ell)}_{i_\ell, i_{\ell+1}}(k)$ has bounded increments. Thus, by using a modification of Hoeffding's Lemma and an ε-net argument (cf. Lemma A.3 of (Araújo et al., 2019)), we obtain that

$$\mathbb{E}\left\{ e^{\zeta \left\| \mathrm{Mrt}^{(\ell)}_{i_\ell i_{\ell+1}}(k) \right\|_2} \right\} \leq 5^d \cdot e^{C_4 k \zeta^2}, \tag{C.37}$$

with $d = \max_{i \in [L-1]} d_i$. By combining (C.35), (C.36) and (C.37), we deduce that

$$\mathbb{P}\left(\sup_{m \in [k]} \|\mathrm{Mrt}_N(m)\|_\infty > u \right) \leq L N^2 5^d \inf_{\zeta \in \mathbb{R}^+} e^{-\zeta \, u + C_4 k \zeta^2}. \tag{C.38}$$

By optimizing over ζ, we have that, with probability at least $1 - e^{-z^2}$,

$$\sup_{m \in [k]} \|\mathrm{Mrt}_N(m)\|_\infty \leq C_5 \sqrt{\frac{1}{\alpha}} \left(\sqrt{d + \log N} + z \right). \tag{C.39}$$

Finally, by combining (C.39) with (C.33), we conclude that, with probability at least $1 - e^{-z^2}$,

$$\left\| \boldsymbol{\theta}(k) - \tilde{\boldsymbol{\theta}}(T) \right\|_\infty \leq C_6 \sqrt{\alpha} (\sqrt{d + \log N} + z). \tag{C.40}$$

Let us bound the second term in the RHS of (C.25). By following the strategy of Lemma 12.2 in (Araújo et al., 2019), we have that, with probability at least $1 - e^{-u^2}$,

$$\left\|\tilde{\boldsymbol{\theta}}^{(\ell)}_{i_\ell, i_{\ell+1}}(t) - \overline{\boldsymbol{\theta}}^{(\ell)}_{i_\ell, i_{\ell+1}}(t)\right\|_2 \leq C_7 \int_0^t \left\|\tilde{\boldsymbol{\theta}}(s) - \overline{\boldsymbol{\theta}}(s)\right\|_\infty \mathrm{d}s + C_7 \frac{u + \sqrt{d}}{\sqrt{N}}. \tag{C.41}$$

By doing a union bound over $i_\ell, i_{\ell+1} \in [N]$ and $\ell \in [L-1]$, we deduce that, with probability at least $1 - e^{-z^2}$,

$$\left\|\tilde{\boldsymbol{\theta}}(t) - \overline{\boldsymbol{\theta}}(t)\right\|_\infty \leq C_7 \int_0^t \left\|\tilde{\boldsymbol{\theta}}(s) - \overline{\boldsymbol{\theta}}(s)\right\|_\infty \mathrm{d}s + C_8 \frac{z + \sqrt{d + \log N}}{\sqrt{N}}. \tag{C.42}$$

By Gronwall lemma, we conclude that, with probability at least $1 - e^{-z^2}$,

$$\left\|\tilde{\boldsymbol{\theta}}(T) - \overline{\boldsymbol{\theta}}(T)\right\|_\infty \leq C_8 e^{C_7 T} \frac{z + \sqrt{d + \log N}}{\sqrt{N}}. \tag{C.43}$$

By combining (C.40) and (C.43), we have that, with probability at least $1 - e^{-z^2}$,

$$\|\boldsymbol{\theta}(k) - \bar{\boldsymbol{\theta}}(T)\|_\infty \leq C_9 \left(\frac{z + \sqrt{d + \log N}}{\sqrt{N}} + \sqrt{\alpha}(\sqrt{d + \log N} + z) \right). \tag{C.44}$$

At this point, we are ready to bound the various terms in the RHS of (C.22). In order to bound the first term, note that L_N is Lipschitz with $\|\cdot\|_\infty$. Thus, we obtain that, with probability at least $1 - e^{-z^2}$,

$$|L_N(\boldsymbol{\theta}(k)) - L_N(\bar{\boldsymbol{\theta}}(T))| \leq C_{10} \left(\frac{z + \sqrt{d + \log N}}{\sqrt{N}} + \sqrt{\alpha}(\sqrt{d + \log N} + z) \right). \tag{C.45}$$

In order to bound the second term in the RHS of (C.22), note that

$$\|\boldsymbol{\theta}_{\mathrm{S}}(k) - \bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)\|_\infty \leq \|\boldsymbol{\theta}(k) - \bar{\boldsymbol{\theta}}(T)\|_\infty. \tag{C.46}$$

As $L_{|\mathcal{A}|}$ is Lipschitz with $\|\cdot\|_\infty$, by combining (C.44) and (C.46), we obtain the bound

$$|L_{|\mathcal{A}|}(\boldsymbol{\theta}_{\mathrm{S}}(k)) - L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}(T))| \leq C_{11} \left(\frac{z + \sqrt{d + \log N}}{\sqrt{N}} + \sqrt{\alpha}(\sqrt{d + \log N} + z) \right), \tag{C.47}$$

with probability at least $1 - e^{-z^2}$.

Finally, let us consider the remaining two terms in the RHS of (C.22). Fix $\boldsymbol{x} \in \mathbb{R}^{d_0}$. Then, by Lemma 11.4 of (Araújo et al., 2019), we have that, for $\zeta > 0$,

$$\log \mathbb{E}\left\{ e^{\zeta \|\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}(T)\right) - \bar{\boldsymbol{y}}(\boldsymbol{x}, \rho_T^\star)\|_2} \right\} \leq C_{12} \left(d + \frac{\zeta^2}{N} \right). \tag{C.48}$$

By using similar arguments, we also have that, for $\zeta > 0$,

$$\log \mathbb{E}\left\{ e^{\zeta \|\widehat{\boldsymbol{y}}_{|\mathcal{A}|}\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)\right) - \bar{\boldsymbol{y}}(\boldsymbol{x}, \rho_T^\star)\|_2} \right\} \leq C_{13} \left(d + \frac{\zeta^2}{A_{\min}} \right). \tag{C.49}$$

Thus, by applying Markov inequality and optimizing over ζ, we deduce that

$$\begin{aligned} \|\widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}(T)\right) - \bar{\boldsymbol{y}}(\boldsymbol{x}, \rho_T^\star)\|_2 &\leq C_{14} \frac{\sqrt{d} + z}{\sqrt{N}}, \\ \|\widehat{\boldsymbol{y}}_{|\mathcal{A}|}\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)\right) - \bar{\boldsymbol{y}}(\boldsymbol{x}, \rho_T^\star)\|_2 &\leq C_{14} \frac{\sqrt{d} + z}{\sqrt{A_{\min}}}, \end{aligned} \tag{C.50}$$

with probability at least $1 - e^{-z^2}$. By using that $\boldsymbol{y}, \widehat{\boldsymbol{y}}_N\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}(T)\right), \widehat{\boldsymbol{y}}_{|\mathcal{A}|}\left(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)\right)$ and $\bar{\boldsymbol{y}}(\boldsymbol{x}, \rho_T^\star)$ are bounded, we conclude that

$$\begin{aligned} |L_N(\bar{\boldsymbol{\theta}}(T)) - \bar{L}(\rho_T^\star)| &\leq C_{15}\frac{\sqrt{d}+z}{\sqrt{N}}, \\ |L_{|\mathcal{A}|}(\bar{\boldsymbol{\theta}}_{\mathrm{S}}(T)) - \bar{L}(\rho_T^\star)| &\leq C_{15}\frac{\sqrt{d}+z}{\sqrt{A_{\min}}}, \end{aligned} \tag{C.51}$$

with probability at least $1 - e^{-z^2}$. By combining (C.45), (C.47) and (C.51), the proof is complete. □

C.2. Part (B)

The proof of part (B) is obtained by combining part (A) with the following result, which extends Lemma A.1 to the multilayer case.

Lemma C.2 (Dropout stability implies connectivity – multilayer)**.** *Consider a neural network with $L+1 \geq 4$ layers, where each hidden layer contains N neurons, as in (4.1). For any $k \in [L]$, assume that $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\theta}}$ are ε-dropout stable given $\mathcal{A}_i = [N/2]$ for $i \in \{k, \dots, L\}$. Then, $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\theta}}$ are ε-connected.*

Given a vector of parameters $\boldsymbol{\theta}$, it is helpful to write it as

$$\begin{aligned} \boldsymbol{\theta}^{(L)} &= \left\{ \left[\boldsymbol{a}_{i_L}^{(L)}\right]_{i_L \in [N]}, \left[\boldsymbol{w}_{i_L}^{(L)}\right]_{i_L \in [N]} \right\}, \\ \boldsymbol{\theta}^{(\ell)} &= \left\{ \left[\boldsymbol{a}_{i_{\ell+1}, i_\ell}^{(\ell)}\right]_{i_{\ell+1}, i_\ell \in [N]}, \left[\boldsymbol{w}_{i_{\ell+1}, i_\ell}^{(\ell)}\right]_{i_{\ell+1}, i_\ell \in [N]} \right\}, \qquad \ell \in [L-1], \\ \boldsymbol{\theta}^{(0)} &= \left[\boldsymbol{\theta}_{i_0}^{(0)}\right]_{i_0 \in [N]}. \end{aligned} \tag{C.52}$$

In words, we stack the parameters $\boldsymbol{\theta}^{(\ell)}$ of layer ℓ into a matrix, and the (i,j)-th element of this matrix contains the parameter $\boldsymbol{\theta}_{j,i}^{(\ell)} = (\boldsymbol{a}_{j,i}^{(\ell)}, \boldsymbol{w}_{j,i}^{(\ell)})$ connecting the j-th neuron of layer ℓ with the i-th neuron of layer $\ell+1$. Furthermore, let us partition the parameters $\boldsymbol{\theta}$ as

$$\begin{aligned} \boldsymbol{\theta}^{(L)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t}}^{(L)} & \boldsymbol{a}_{\mathrm{b}}^{(L)} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \boldsymbol{w}_{\mathrm{b}}^{(L)} \end{array}\right] \right\}, \\ \boldsymbol{\theta}^{(\ell)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t,t}}^{(\ell)} & \boldsymbol{a}_{\mathrm{t,b}}^{(\ell)} \\ \hline \boldsymbol{a}_{\mathrm{b,t}}^{(\ell)} & \boldsymbol{a}_{\mathrm{b,b}}^{(\ell)} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(\ell)} & \boldsymbol{w}_{\mathrm{t,b}}^{(\ell)} \\ \hline \boldsymbol{w}_{\mathrm{b,t}}^{(\ell)} & \boldsymbol{w}_{\mathrm{b,b}}^{(\ell)} \end{array}\right] \right\}, \qquad \ell \in [L-1], \\ \boldsymbol{\theta}^{(0)} &= \left[\begin{array}{c} \boldsymbol{\theta}_{\mathrm{t}}^{(0)} \\ \hline \boldsymbol{\theta}_{\mathrm{b}}^{(0)} \end{array}\right]. \end{aligned} \tag{C.53}$$

In words, $\boldsymbol{\theta}_{\mathrm{t,t}}^{(\ell)} = (\boldsymbol{a}_{\mathrm{t,t}}^{(\ell)}, \boldsymbol{w}_{\mathrm{t,t}}^{(\ell)})$ contains the parameters connecting the top half neurons of layer ℓ with the top half neurons of layer $\ell+1$; $\boldsymbol{\theta}_{\mathrm{t,b}}^{(\ell)} = (\boldsymbol{a}_{\mathrm{t,b}}^{(\ell)}, \boldsymbol{w}_{\mathrm{t,b}}^{(\ell)})$ contains the parameters connecting the bottom half neurons of layer ℓ with the top half neurons of layer $\ell+1$; $\boldsymbol{\theta}_{\mathrm{b,t}}^{(\ell)} = (\boldsymbol{a}_{\mathrm{b,t}}^{(\ell)}, \boldsymbol{w}_{\mathrm{b,t}}^{(\ell)})$ contains the parameters connecting the top half neurons of layer ℓ with the bottom half neurons of layer $\ell+1$; and $\boldsymbol{\theta}_{\mathrm{b,b}}^{(\ell)} = (\boldsymbol{a}_{\mathrm{b,b}}^{(\ell)}, \boldsymbol{w}_{\mathrm{b,b}}^{(\ell)})$ contains the parameters connecting the bottom half neurons of layer ℓ with the bottom half neurons of layer $\ell+1$. The partition for the first and the last layer is similarly defined.

At this point, we are ready to present the proof of Lemma C.2.

Proof of Lemma C.2. For the moment, assume that N is even. Let $\boldsymbol{\theta}_{\mathrm{S},k}$ be obtained from $\boldsymbol{\theta}$ by keeping only the top half

neurons at layer $\ell \in \{k, \dots, L\}$. With an abuse of notation, we can partition the parameters $\boldsymbol{\theta}_{\mathrm{S},k}$ as

$$
\begin{aligned}
\boldsymbol{\theta}_{\mathrm{S},k}^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\}, \\
\boldsymbol{\theta}_{\mathrm{S},k}^{(\ell)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(\ell)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(\ell)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \right\}, \qquad \ell \in \{k, \dots, L-1\}, \\
\boldsymbol{\theta}_{\mathrm{S},k}^{(\ell)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t,t}}^{(\ell)} & \boldsymbol{a}_{\mathrm{t,b}}^{(\ell)} \\ \hline \boldsymbol{a}_{\mathrm{b,t}}^{(\ell)} & \boldsymbol{a}_{\mathrm{b,b}}^{(\ell)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(\ell)} & \boldsymbol{w}_{\mathrm{t,b}}^{(\ell)} \\ \hline \boldsymbol{w}_{\mathrm{b,t}}^{(\ell)} & \boldsymbol{w}_{\mathrm{b,b}}^{(\ell)} \end{array} \right] \right\}, \qquad \ell \in [k-1], \\
\boldsymbol{\theta}_{\mathrm{S},k}^{(0)} &= \left[\begin{array}{c} \boldsymbol{\theta}_{\mathrm{t}}^{(0)} \\ \hline \boldsymbol{\theta}_{\mathrm{b}}^{(0)} \end{array} \right],
\end{aligned}
\tag{C.54}
$$

and the corresponding loss is given by $L_N(\boldsymbol{\theta}_{\mathrm{S},k})$. We now prove by induction that $\boldsymbol{\theta}$ is connected to $\boldsymbol{\theta}_{\mathrm{S},k}$ via a piecewise linear path in parameter space, such that the loss along the path is upper bounded by $L_N(\boldsymbol{\theta}) + \varepsilon$.

Base step: from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\mathrm{S},L}$. As $\boldsymbol{\theta}$ is ε-dropout stable, we have that $L_N(\boldsymbol{\theta}_{\mathrm{S},L}) \le L_N(\boldsymbol{\theta}) + \varepsilon$. Note that if $\boldsymbol{a}_{\mathrm{t}}^{(L)} = \mathbf{0}$, then the value of $\boldsymbol{w}_{\mathrm{t}}^{(L)}$ does not affect the loss. Hence, we can interpolate from $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,]\}$ to $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{b}}^{(L)} \,]\}$ with no change in loss. Furthermore, the loss is convex in $\boldsymbol{a}^{(L)}$. Thus, we can interpolate from $\{[\, \boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \boldsymbol{a}_{\mathrm{b}}^{(L)} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{b}}^{(L)} \,]\}$ to $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{b}}^{(L)} \,]\}$ while keeping the loss upper bounded by $L_N(\boldsymbol{\theta}) + \varepsilon$.

Induction step: from $\boldsymbol{\theta}_{\mathrm{S},k}$ to $\boldsymbol{\theta}_{\mathrm{S},k-1}$. We construct the path by passing through the following intermediate points in parameter space:

$$
\begin{aligned}
\boldsymbol{\theta}_{1}^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\}, \\
\boldsymbol{\theta}_{1}^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \right\}, \qquad i \in \{k, \dots, L-1\}, \\
\boldsymbol{\theta}_{1}^{(k-1)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{a}_{\mathrm{t,b}}^{(k-1)} \\ \hline \boldsymbol{a}_{\mathrm{b,t}}^{(k-1)} & \boldsymbol{a}_{\mathrm{b,b}}^{(k-1)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{w}_{\mathrm{t,b}}^{(k-1)} \\ \hline \boldsymbol{w}_{\mathrm{b,t}}^{(k-1)} & \boldsymbol{w}_{\mathrm{b,b}}^{(k-1)} \end{array} \right] \right\}.
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\theta}_{2}^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\}, \\
\boldsymbol{\theta}_{2}^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{w}_{\mathrm{t,t}}^{(i)} \end{array} \right] \right\}, \qquad i \in \{k, \dots, L-1\}, \\
\boldsymbol{\theta}_{2}^{(k-1)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{a}_{\mathrm{t,b}}^{(k-1)} \\ \hline 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{w}_{\mathrm{t,b}}^{(k-1)} \\ \hline \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right] \right\}.
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\theta}_{3}^{(L)} &= \left\{ \left[\begin{array}{c|c} \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \end{array} \right], \left[\begin{array}{c|c} \mathbf{0} & \boldsymbol{w}_{\mathrm{t}}^{(L)} \end{array} \right] \right\}, \\
\boldsymbol{\theta}_{3}^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{w}_{\mathrm{t,t}}^{(i)} \end{array} \right] \right\}, \qquad i \in \{k, \dots, L-1\}, \\
\boldsymbol{\theta}_{3}^{(k-1)} &= \left\{ \left[\begin{array}{c|c} \boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{a}_{\mathrm{t,b}}^{(k-1)} \\ \hline 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \boldsymbol{w}_{\mathrm{t,b}}^{(k-1)} \\ \hline \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right] \right\}.
\end{aligned}
$$

$$\boldsymbol{\theta}_4^{(L)} = \left\{ \left[\begin{array}{c|c} \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \end{array} \right], \left[\begin{array}{c|c} \mathbf{0} & \boldsymbol{w}_{\mathrm{t}}^{(L)} \end{array} \right] \right\},$$
$$\boldsymbol{\theta}_4^{(i)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{w}_{\mathrm{t,t}}^{(i)} \end{array} \right] \right\}, \qquad i \in \{k, \dots, L-1\},$$
$$\boldsymbol{\theta}_4^{(k-1)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \\ \hline 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \\ \hline \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right] \right\}.$$

$$\boldsymbol{\theta}_5^{(L)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\},$$
$$\boldsymbol{\theta}_5^{(i)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{w}_{\mathrm{t,t}}^{(i)} \end{array} \right] \right\}, \qquad i \in \{k, \dots, L-1\},$$
$$\boldsymbol{\theta}_5^{(k-1)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \\ \hline 2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \\ \hline \boldsymbol{w}_{\mathrm{t,t}}^{(k-1)} & \mathbf{0} \end{array} \right] \right\}.$$

$$\boldsymbol{\theta}_6^{(L)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\},$$
$$\boldsymbol{\theta}_6^{(i)} = \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \right\}, \qquad i \in \{k-1, \dots, L-1\}.$$

As we do not change the parameters in layer $\ell \in [k-2]$, we have omitted them in the definitions above.

From $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. The loss is not affected by the values in the bottom right quadrant of $\boldsymbol{\theta}_1^{(k-1)}$, since the bottom neurons of layer k are not active ($\boldsymbol{a}_{\mathrm{t,b}}^{(k)} = \boldsymbol{a}_{\mathrm{b,b}}^{(k)} = \mathbf{0}$). Consequently, we can interpolate from $\boldsymbol{a}_{\mathrm{b,b}}^{(k-1)}$ to $\mathbf{0}$ and from $\boldsymbol{w}_{\mathrm{b,b}}^{(k-1)}$ to $\mathbf{0}$ with no change in loss. Similarly, the loss is not affected by the values in the bottom right quadrant of $\boldsymbol{\theta}_1^{(i)}$ for $i \in \{k, \dots L-1\}$, since the bottom neurons of layer $i+1$ are not active ($\boldsymbol{a}_{\mathrm{t,b}}^{(i+1)} = \boldsymbol{a}_{\mathrm{b,b}}^{(i+1)} = \mathbf{0}$ and $\boldsymbol{a}_{\mathrm{b}}^{(L)} = \mathbf{0}$). Consequently, for $i \in \{k, \dots L-1\}$, we can successively interpolate from $\mathbf{0}$ to $2\boldsymbol{a}_{\mathrm{t,t}}^{(i)}$ and from $\mathbf{0}$ to $2\boldsymbol{w}_{\mathrm{t,t}}^{(i)}$ with no change in loss.

From $\boldsymbol{\theta}_5$ to $\boldsymbol{\theta}_6$. We use the same reasoning as for $\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2$ and go in reverse layer order (i.e., from layer $L-1$ to layer $k-1$). The loss is not affected by the values in the bottom right quadrant of $\boldsymbol{\theta}_5^{(i)}$, since the bottom neurons of layer $i+1$ are not active. Consequently, we can interpolate from $2\boldsymbol{a}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ and from $\boldsymbol{w}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ with no change in loss. Similarly, the loss is not affected by the values in the bottom left quadrant of $\boldsymbol{\theta}_5^{(k-1)}$, since the bottom neurons of layer k are not active. Consequently, we can interpolate from $2\boldsymbol{a}_{\mathrm{t,t}}^{(k-1)}$ to $\mathbf{0}$ and from $\boldsymbol{w}_{\mathrm{t,t}}^{(k-1)}$ to $\mathbf{0}$ with no change in loss.

From $\boldsymbol{\theta}_4$ to $\boldsymbol{\theta}_5$. Note that the parameters of $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ are the same except for layer L. Furthermore, the structure of these parameters implies that the output of layer $L-1$ is obtained by stacking the output of two identical sub-networks. In formulas, let $\boldsymbol{z}^{(L-1)}$ be the output of layer $L-1$. Then, $\boldsymbol{z}^{(L-1)} = [\ \bar{\boldsymbol{z}} \mid \bar{\boldsymbol{z}}\ ]$ for some $\bar{\boldsymbol{z}}$. Consequently, we can interpolate between $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ with no change in loss.

From $\boldsymbol{\theta}_3$ to $\boldsymbol{\theta}_4$. By using the same reasoning as for $\boldsymbol{\theta}_5 \to \boldsymbol{\theta}_6$, we interpolate from $2\boldsymbol{a}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ and from $\boldsymbol{w}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ in the top left corner of $\boldsymbol{\theta}_3^{(i)}$ with no change in loss, for $i = L-1, \dots, k$. Then, we interpolate from $\boldsymbol{\theta}_3^{(k-1)}$ to $\boldsymbol{\theta}_4^{(k-1)}$ with no change in loss, since the top neurons of layer k are not active. Finally, we restore sequentially $2\boldsymbol{a}_{\mathrm{t,t}}^{(i)}$ and $\boldsymbol{w}_{\mathrm{t,t}}^{(i)}$ in the top left corner of the corresponding parameter matrices with no change in loss, by using the same reasoning as for $\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2$.

From $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_3$. From the previous arguments, we have that $L_N(\boldsymbol{\theta}_2) = L_N(\boldsymbol{\theta}_1)$ and $L_N(\boldsymbol{\theta}_3) = L_N(\boldsymbol{\theta}_6)$. Furthermore, $\boldsymbol{\theta}$ is ε-dropout stable, which implies that $|L_N(\boldsymbol{\theta}_1) - L_N(\boldsymbol{\theta}_6)| \leq \varepsilon$. Consequently, we have that $|L_N(\boldsymbol{\theta}_2) - L_N(\boldsymbol{\theta}_3)| \leq \varepsilon$. Note that if $\boldsymbol{a}_{\mathrm{t}}^{(L)} = \mathbf{0}$, then the value of $\boldsymbol{w}_{\mathrm{t}}^{(L)}$ does not affect the loss. Hence, we can interpolate from $\{[\ 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0}\ ], [\ \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \mathbf{0}\ ]\}$ to $\{[\ 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0}\ ], [\ \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{t}}^{(L)}\ ]\}$ with no change in loss. Similarly, we can interpolate from $\{[\ \mathbf{0} \mid 2\boldsymbol{a}_{\mathrm{t}}^{(L)}\ ], [\ \mathbf{0} \mid \boldsymbol{w}_{\mathrm{t}}^{(L)}\ ]\}$ to $\{[\ \mathbf{0} \mid 2\boldsymbol{a}_{\mathrm{t}}^{(L)}\ ], [\ \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{t}}^{(L)}\ ]\}$ with no change in loss. Furthermore, the loss is convex in $\boldsymbol{a}^{(L)}$. Thus, we can

interpolate from $\{[\ 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0}\ ], [\ \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{t}}^{(L)}\ ]\}$ to $\{[\ \mathbf{0} \mid 2\boldsymbol{a}_{\mathrm{t}}^{(L)}\ ], [\ \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \boldsymbol{w}_{\mathrm{t}}^{(L)}\ ]\}$ while keeping the loss upper bounded by $L_N(\boldsymbol{\theta}) + \varepsilon$.

As a result, we are able to connect $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{\mathrm{S},1}$ via a piecewise linear path, where the loss is upper bounded by $L_N(\boldsymbol{\theta}) + \varepsilon$. Similarly, let $\bar{\boldsymbol{\theta}}_{\mathrm{S},k}$ be obtained from $\bar{\boldsymbol{\theta}}$ by keeping only the top half neurons at layer $\ell \in \{k, \ldots, L\}$. Then, we can connect $\bar{\boldsymbol{\theta}}$ with $\bar{\boldsymbol{\theta}}_{\mathrm{S},1}$ via a piecewise linear path, where the loss is upper bounded by $L_N(\bar{\boldsymbol{\theta}}) + \varepsilon$.

In order to complete the proof, it remains to connect $\boldsymbol{\theta}_{\mathrm{S},1}$ with $\bar{\boldsymbol{\theta}}_{\mathrm{S},1}$ via a piecewise linear path, where the loss is upper bounded by $\max(L_N(\boldsymbol{\theta}), L_N(\bar{\boldsymbol{\theta}})) + \varepsilon$. We construct the path by passing through the following intermediate points in parameter space:

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}_1^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right] \right\}, \\
\tilde{\boldsymbol{\theta}}_1^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \right\}, \qquad i \in \{1, \ldots, L-1\}, \\
\tilde{\boldsymbol{\theta}}_1^{(0)} &= \left[\begin{array}{c} \boldsymbol{\theta}_{\mathrm{t}}^{(0)} \\ \hline \boldsymbol{\theta}_{\mathrm{b}}^{(0)} \end{array}\right].
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}_2^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right] \right\}, \\
\tilde{\boldsymbol{\theta}}_2^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right] \right\}, \qquad i \in \{1, \ldots, L-1\}, \\
\tilde{\boldsymbol{\theta}}_2^{(0)} &= \left[\begin{array}{c} \boldsymbol{\theta}_{\mathrm{t}}^{(0)} \\ \hline \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \end{array}\right].
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}_3^{(L)} &= \left\{ \left[\begin{array}{c|c} \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} \end{array}\right], \left[\begin{array}{c|c} \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \end{array}\right] \right\}, \\
\tilde{\boldsymbol{\theta}}_3^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\boldsymbol{a}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{w}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right] \right\}, \qquad i \in \{1, \ldots, L-1\}, \\
\tilde{\boldsymbol{\theta}}_3^{(0)} &= \left[\begin{array}{c} \boldsymbol{\theta}_{\mathrm{t}}^{(0)} \\ \hline \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \end{array}\right].
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}_4^{(L)} &= \left\{ \left[\begin{array}{c|c} \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} \end{array}\right], \left[\begin{array}{c|c} \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \end{array}\right] \right\}, \\
\tilde{\boldsymbol{\theta}}_4^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right], \left[\begin{array}{c|c} \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right] \right\}, \qquad i \in \{1, \ldots, L-1\}, \\
\tilde{\boldsymbol{\theta}}_4^{(0)} &= \left[\begin{array}{c} \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \\ \hline \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \end{array}\right].
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}_5^{(L)} &= \left\{ \left[\begin{array}{c|c} 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right], \left[\begin{array}{c|c} \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array}\right] \right\}, \\
\tilde{\boldsymbol{\theta}}_5^{(i)} &= \left\{ \left[\begin{array}{c|c} 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & 2\bar{\boldsymbol{a}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right], \left[\begin{array}{c|c} \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \bar{\boldsymbol{w}}_{\mathrm{t},\mathrm{t}}^{(i)} \end{array}\right] \right\}, \qquad i \in \{1, \ldots, L-1\}, \\
\tilde{\boldsymbol{\theta}}_5^{(0)} &= \left[\begin{array}{c} \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \\ \hline \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \end{array}\right].
\end{aligned}
$$

$$\tilde{\boldsymbol{\theta}}_6^{(L)} = \left\{ \left[\begin{array}{c|c} 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} & \mathbf{0} \end{array} \right] \right\},$$

$$\tilde{\boldsymbol{\theta}}_6^{(i)} = \left\{ \left[\begin{array}{c|c} 2\bar{\boldsymbol{a}}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right], \left[\begin{array}{c|c} \bar{\boldsymbol{w}}_{\mathrm{t,t}}^{(i)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \right\}, \qquad i \in \{1, \dots, L-1\},$$

$$\tilde{\boldsymbol{\theta}}_6^{(0)} = \left[\begin{array}{c} \bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)} \\ \hline \bar{\boldsymbol{\theta}}_{\mathrm{b}}^{(0)} \end{array} \right].$$

The arguments to connect $\tilde{\boldsymbol{\theta}}_j$ with $\tilde{\boldsymbol{\theta}}_{j+1}$ are analogous to those previously used to connect $\boldsymbol{\theta}_j$ with $\boldsymbol{\theta}_{j+1}$. We briefly outline them below for completeness.

From $\tilde{\boldsymbol{\theta}}_1$ to $\tilde{\boldsymbol{\theta}}_2$. First, we interpolate from $\boldsymbol{\theta}_{\mathrm{b}}^{(0)}$ to $\bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)}$ with no loss change. Then, for $i = 1, \dots, L-1$, we successively interpolate from $\mathbf{0}$ to $\bar{\boldsymbol{w}}_{\mathrm{t,t}}^{(i)}$ and from $\mathbf{0}$ to $2\bar{\boldsymbol{a}}_{\mathrm{t,t}}^{(i)}$ with no loss change.

From $\tilde{\boldsymbol{\theta}}_5$ to $\tilde{\boldsymbol{\theta}}_6$. For $i = L-1, \dots, 1$, we successively interpolate from $2\bar{\boldsymbol{a}}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ and from $\bar{\boldsymbol{w}}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ with no loss change. Finally, we interpolate from $\bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)}$ to $\bar{\boldsymbol{\theta}}_{\mathrm{b}}^{(0)}$ with no loss change.

From $\tilde{\boldsymbol{\theta}}_4$ to $\tilde{\boldsymbol{\theta}}_5$. The output of layer $L-1$ is obtained by stacking the output of two identical sub-networks. Thus, we can interpolate between $\tilde{\boldsymbol{\theta}}_4$ and $\tilde{\boldsymbol{\theta}}_5$ with no change in loss.

From $\tilde{\boldsymbol{\theta}}_3$ to $\tilde{\boldsymbol{\theta}}_4$. For $i = L-1, \dots, 1$, we interpolate from $2\boldsymbol{a}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ and from $\boldsymbol{w}_{\mathrm{t,t}}^{(i)}$ to $\mathbf{0}$ with no change in loss. Then, we interpolate from $\boldsymbol{\theta}_{\mathrm{t}}^{(0)}$ to $\bar{\boldsymbol{\theta}}_{\mathrm{t}}^{(0)}$ with no change in loss. Finally, for $i = 1, \dots, L-1$, we restore sequentially $2\bar{\boldsymbol{a}}_{\mathrm{t,t}}^{(i)}$ and $\bar{\boldsymbol{w}}_{\mathrm{t,t}}^{(i)}$ in the top left corner of the corresponding parameter matrices with no change in loss.

From $\tilde{\boldsymbol{\theta}}_2$ to $\tilde{\boldsymbol{\theta}}_3$. From the previous arguments, we have that $L_N(\tilde{\boldsymbol{\theta}}_2) = L_N(\tilde{\boldsymbol{\theta}}_1) \leq L_N(\boldsymbol{\theta}) + \varepsilon$ and $L_N(\tilde{\boldsymbol{\theta}}_3) = L_N(\tilde{\boldsymbol{\theta}}_6) \leq L_N(\bar{\boldsymbol{\theta}}) + \varepsilon$. First, we interpolate from $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,]\}$ to $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \,]\}$ with no change in loss. Similarly, we interpolate from $\{[\, \mathbf{0} \mid 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} \,], [\, \mathbf{0} \mid \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \,]\}$ to $\{[\, \mathbf{0} \mid 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \,]\}$ with no change in loss. Furthermore, as the loss is convex in $\boldsymbol{a}^{(L)}$, we interpolate from $\{[\, 2\boldsymbol{a}_{\mathrm{t}}^{(L)} \mid \mathbf{0} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \,]\}$ to $\{[\, \mathbf{0} \mid 2\bar{\boldsymbol{a}}_{\mathrm{t}}^{(L)} \,], [\, \boldsymbol{w}_{\mathrm{t}}^{(L)} \mid \bar{\boldsymbol{w}}_{\mathrm{t}}^{(L)} \,]\}$ while keeping the loss upper bounded by $\max(L_N(\boldsymbol{\theta}), L_N(\bar{\boldsymbol{\theta}})) + \varepsilon$.

□

D. Additional Numerical Results

In Figures 1, 2 and 3, we consider the problem of classifying isotropic Gaussians. This is an artificial dataset considered in (Mei et al., 2018b). The label y is chosen uniformly at random between -1 and 1, i.e., $y \sim \mathrm{Unif}(\{-1, 1\})$. Given y, the feature vector $\boldsymbol{x}$ is a d-dimensional isotropic Gaussian with covariance matrix $(1 + y\Delta)^2 \boldsymbol{I}_d$, i.e., $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, (1 + y\Delta)^2 \boldsymbol{I}_d)$. We set $d = 32$ and $\Delta = 0.5$, and we run the one-pass (or online) SGD algorithm (3.3) on the two-layer neural network (3.1) with sigmoid activation function ($\sigma(x) = 1/(1 + e^{-x})$). We estimate the population risk and the classification error on 10^4 independent samples. Figure 1 compares the performance of the trained network (blue dashed curve) and of the dropout network (orange curve) obtained by removing half of the neurons. We plot the population risk and the classification error for $N = 800$ and $N = 6400$. As expected, the performance of the dropout network improves with N, and it is very close to that of the trained network already for $N = 800$. In fact, for $N = 800$ the classification error of the dropout network is $< 0.4\%$. Figure 2 plots the change in loss between the full and the dropout network, as a function of the number of neurons of the full network N. The change in loss decreases steadily with N for all the values of T taken into account. Finally, Figure 3 shows that the optimization landscape is approximately connected when $N = 3200$.

In Figures 4, 5 and 6, we consider MNIST classification with a three-layer neural network and CIFAR-10 classification with a four-layer neural network. The results are qualitatively similar to those of Figures 1, 2 and 3 in Section 5.

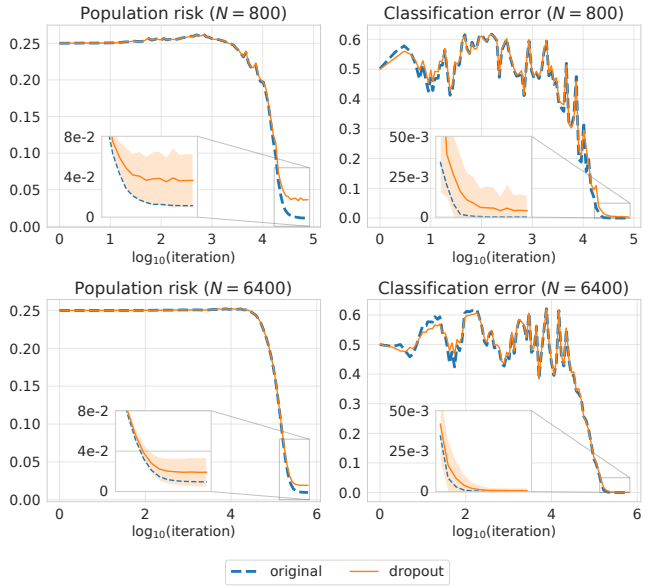


Figure 1. Comparison of population risk and classification error between the trained network (blue dashed curve) and the dropout network (orange curve) for the classification of isotropic Gaussians.

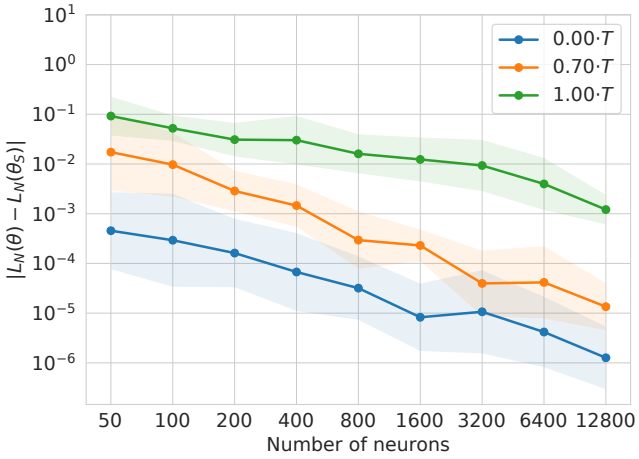


Figure 2. Change in loss between the full network and the dropout network for the classification of isotropic Gaussians, as a function of the number of neurons N of the full network.

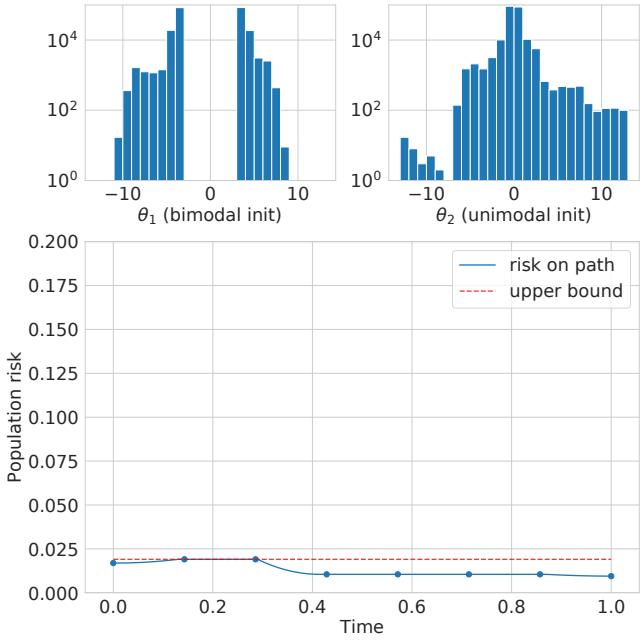
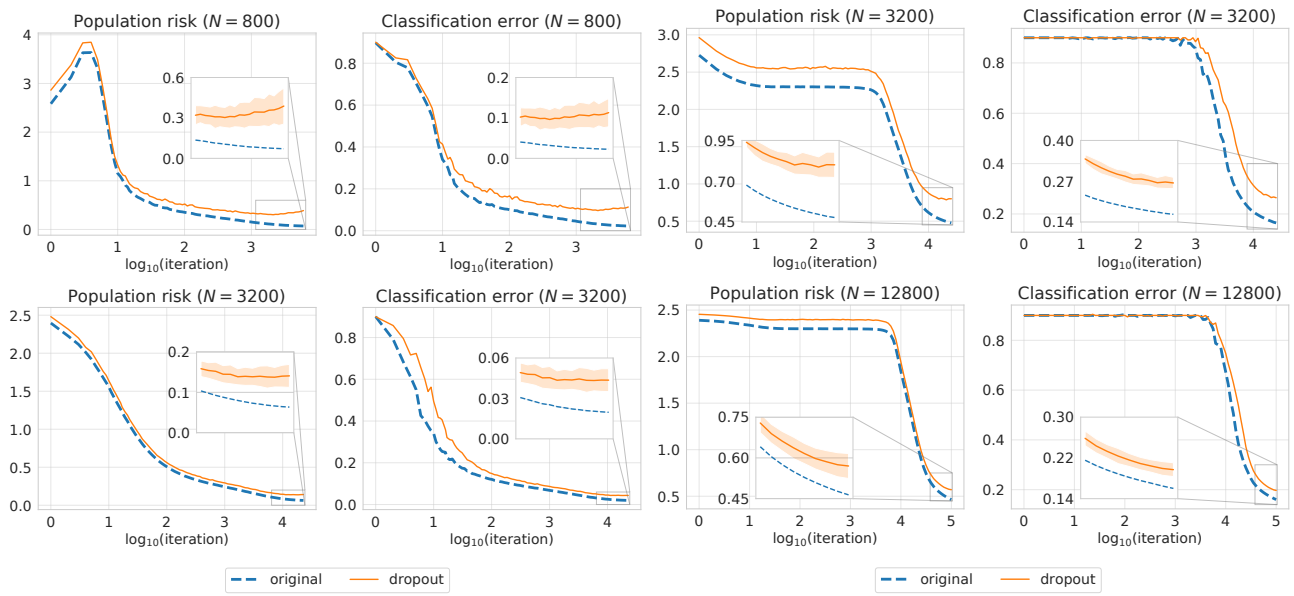
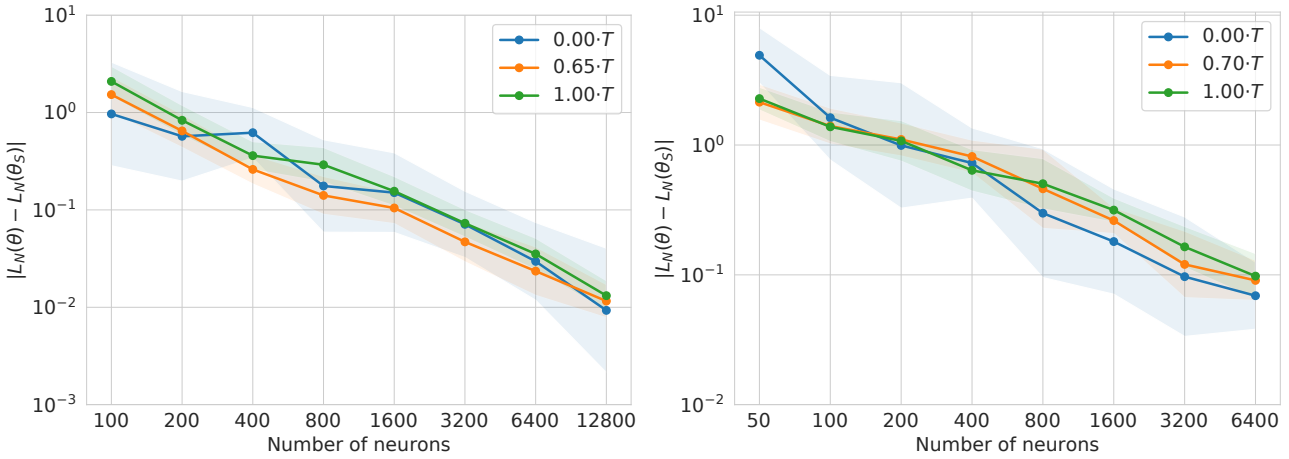


Figure 3. Classification error along a piecewise linear path that connects two SGD solutions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for the classification of isotropic Gaussians with $N = 3200$. The two SGD solutions are initialized with different distributions, and we show their histograms to highlight that $\boldsymbol{\theta}_1$ cannot be obtained by permuting $\boldsymbol{\theta}_2$.



(a) MNIST, three-layer (b) CIFAR-10, four-layer

Figure 4. Comparison of population risk and classification error between the trained network (blue dashed curve) and the dropout network (orange curve).



(a) MNIST, three-layer (b) CIFAR-10, four-layer

Figure 5. Change in loss after removing half of the neurons from each layer, as a function of the number of neurons N of the full network.

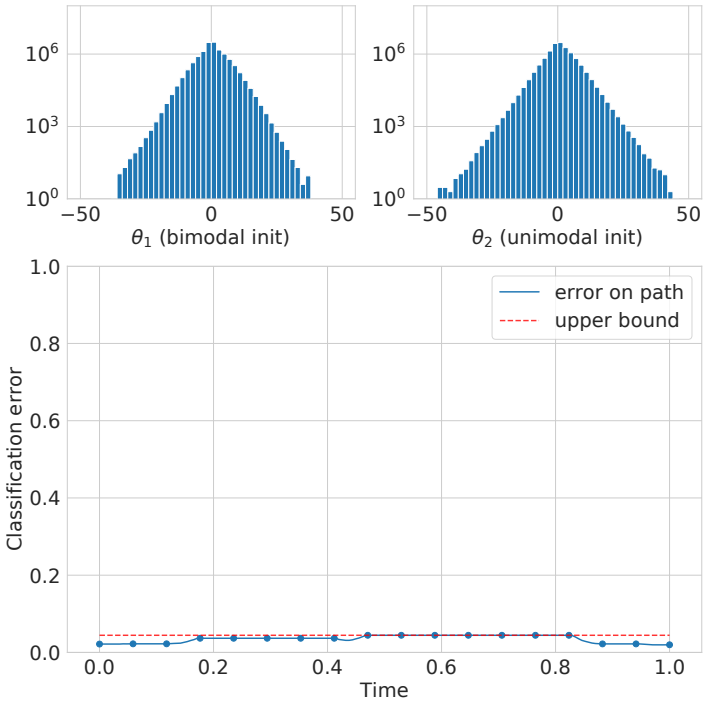


Figure 6. Classification error along a piecewise linear path that connects two SGD solutions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for MNIST classification with a three-layer neural network with $N = 3200$.