A. Proof of Proposition 3.3

This result directly follows Theorem 5.5 in Araújo et al. (2019). Let $B_{\rm GD}^{\infty}$ denote the infinitely wide network trained by gradient descent in the limit of $M \to \infty$. By the results in Theorem 5.5 of Araújo et al. (2019), we have

$$\mathbb{D}[S_{\rm GD}^m,\;\; B_{\rm GD}^{\infty}] = \mathcal{O}_p\left(n\exp(c_1\exp(c_2 n))\left(\frac{1}{\sqrt{m}} + \sqrt{\eta}\right)\right),$$

where we explicit give the dependency of constant $C_{5.5}$ in Araújo et al. (2019) on the depth n, because $C_{5.5} = O(\exp(c_1 \times C_{B.16}))$, where $C_{B.16} = \mathcal{O}(\exp(c_2 n))$ and c_1 is some positive constant. See Lemma 12.2 in Araújo et al. (2019) for details.

Similarly,

$$\mathbb{D}[S_{\rm GD}^m,\;\; B_{\rm GD}^{\infty}] = \mathcal{O}_p\left(n\exp(c_1\exp(c_2 n))\left(\frac{1}{\sqrt{M}} + \sqrt{\eta}\right)\right).$$

Combining this, we have

$$\begin{aligned}\mathbb{D}[B_{\rm GD}^M,\;\; B_{\rm GD}^M] &\le \mathbb{D}[S_{\rm GD}^m,\;\; B_{\rm GD}^{\infty}] + \mathbb{D}[B_{\rm GD}^M,\;\; B_{\rm GD}^{\infty}] \\ &= \mathcal{O}_p\left(n\exp(c_1\exp(c_2 n))\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{M}} + \sqrt{\eta}\right)\right).\end{aligned}$$

B. Proof of Theorem 3.5

Assumption 3.4 *Denote by $S_{\rm WIN}^m$ the result of mimicking $B_{\rm GD}^M$ following Algorithm 1. When training $S_{\rm WIN}^m$, we assume the parameters of $S_{\rm WIN}^m$ in each layer are initialized by randomly sampling m neurons from the the corresponding layer of the wide network $B_{\rm GD}^M$. Define $B_{{\rm GD},[i:n]}^M = B_n^M \circ \cdots B_i^M$.*

Theorem 3.5 *Assume all the layers of $B_{\rm GD}^M$ are Lipschitz maps and all its parameters are bounded by some constant. Under the assumptions 3.1, 3.2, 3.4, we have*

$$\mathbb{D}[S_{\rm WIN}^m, B_{\rm GD}^M] = \mathcal{O}_p\left(\frac{\ell_B n}{\sqrt{m}}\right),$$

where $\ell_B = \max_{i\in[n]}\left\|B_{{\rm GD},[i+1:n]}^M\right\|_{\rm Lip}$ and $\mathcal{O}_p(\cdot)$ denotes the big O notation in probability, and the randomness is w.r.t. the random initialization of gradient descent, and the random mini-batches of stochastic gradient descent.

Proof. To simply the notation, we denote $B_{\rm GD}^M$ by B^M and $S_{\rm WIN}^m$ by S^m in the proof. We have

$$\begin{aligned}B^M(\mathbf{x}) &= (B_n^M \circ B_{n-1}^M \circ ... \circ B_1^M)(\mathbf{x}) \\ S^m(\mathbf{x}) &= \left(S_n^m \circ S_{n-1}^m \circ ... \circ S_1^m\right)(\mathbf{x}).\end{aligned}$$

We define

$$B_{[k_1:k_2]}^M(\mathbf{z}) = (B_{k_2}^M \circ B_{k_2-1}^M \circ ... \circ B_{k_1}^M)(\mathbf{z}),$$

where $\mathbf{z}$ is the input of $B_{[k_1:k_2]}^M$. Define

$$\begin{aligned}F_0(\mathbf{x}) &= \left(B_n^M \circ ... \circ B_3^M \circ B_2^M \circ B_1^M\right)(\mathbf{x}) \\ F_1(\mathbf{x}) &= \left(B_n^M \circ ... \circ B_3^M \circ B_2^M \circ S_1^m\right)(\mathbf{x}) \\ F_2(\mathbf{x}) &= \left(B_n^M \circ ... \circ B_3^M \circ S_2^m \circ S_1^m\right)(\mathbf{x}) \\ &\cdots \\ F_n(\mathbf{x}) &= \left(S_n^m \circ ... \circ S_3^m \circ S_2^m \circ S_1^m\right)(\mathbf{x}),\end{aligned}$$

following which we have $F_0 = B^M$ and $F_n = S^m$, and hence

$$\mathbb{D}[S^m, B^M] = \mathbb{D}[F_n, F_0] \le \sum_{k=1}^{n} \mathbb{D}[F_k, F_{k-1}].$$

Define $\ell_{i-1} := \left\| B^M_{[i:n]} \right\|_{\text{Lip}}$ for $i \in [n]$ and $\ell_n = 1$. Note that

$$\begin{aligned}\mathbb{D}[F_1, F_0] &= \sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(B^M_{[2:n]} \circ B^M_1(\mathbf{x}) - B^M_{[2:n]} \circ S^m_1(\mathbf{x})\right)^2\right]} \\ &\le \ell_1 \sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(B^M_1(\mathbf{x}) - S^m_1(\mathbf{x})\right)^2\right]}\end{aligned}$$

By the assumption that we initialize $S^m_1(\mathbf{x})$ by randomly sampling neurons from $B^M_1(\mathbf{x})$, we have, with high probability,

$$\sqrt{\mathbb{E}_{x\sim\mathcal{D}}\left[\left(B^M_1(\mathbf{x}) - S^m_1(\mathbf{x})\right)^2\right]} \le \frac{c}{\sqrt{m}},$$

where c is constant depending on the bounds of the parameters of B^M. Therefore,

$$\mathbb{D}[F_1, F_0] = \mathcal{O}_p\left(\frac{\ell_1}{\sqrt{m}}\right).$$

Similarly, we have

$$\mathbb{D}[F_k, F_{k-1}] = \mathcal{O}\left(\frac{\ell_k}{\sqrt{m}}\right), \quad \forall k = 2, \ldots, n.$$

Combine all the results, we have

$$\mathbb{D}[B^M, S^m] = \mathcal{O}\left(\frac{n \max_{k\in[n]} \ell_k}{\sqrt{m}}\right).$$

□

Remark Since the wide network B^M_{GD} is observed to be easy to train, it is expected that it can closely approximate the underlying true function and behaves nicely, hence yielding a small ℓ_B. An important future direction is to develop rigorous theoretical bounds for controlling ℓ_B.