We define where we explicitly 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 $C_{B.16} = O(\exp(c_2 n))$ and $c_1$ is some positive constant. See Lemma 12.2 in Araújo et al. (2019) for details.

Similarly,

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

Combining this, we have

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

### B. Proof of Theorem 3.5

**Assumption 3.4** Denote by $S_{WIN}^m$ the result of mimicking $B_{GD}^M$ following Algorithm 1. When training $S_{WIN}^m$, we assume the parameters of $S_{WIN}^m$ in each layer are initialized by randomly sampling $m$ neurons from the corresponding layer of the wide network $B_{GD}^M$. Define $B_{GD,[i:n]}^M = B_n^M \circ \cdots \circ B_1^M$.

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

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

where $\ell_B = \max_{i \in [n]} \|B_{GD,[i+1:n]}^M\|_{\text{Lip}}$ 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 $B_{GD}^M$ by $B^M$ and $S_{WIN}^m$ by $S^m$ in the proof. We have

$$
B^M(x) = \left( B_n^M \circ B_{n-1}^M \circ \cdots \circ B_1^M \right)(x),
$$

$$
S^m(x) = \left( S_n^m \circ S_{n-1}^m \circ \cdots \circ S_1^m \right)(x).
$$

We define

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

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

$$
F_0(x) = \left( B_{n}^M \circ \cdots \circ B_{3}^M \circ B_{2}^M \circ B_{1}^M \right)(x),
$$

$$
F_1(x) = \left( B_{n}^M \circ \cdots \circ B_{3}^M \circ B_{2}^M \circ S_{1}^m \right)(x),
$$

$$
F_2(x) = \left( B_{n}^M \circ \cdots \circ B_{3}^M \circ S_{2}^m \circ S_{1}^m \right)(x),
$$

$$
\cdots
$$

$$
F_n(x) = \left( S_{n}^m \circ \cdots \circ S_{3}^m \circ S_{2}^m \circ S_{1}^m \right)(x),
$$

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

$$
\mathbb{D}[S^m, B^M] = \mathbb{D}[F_n, F_0] \leq \sum_{k=1}^n \mathbb{D}[F_k, F_{k-1}].
$$
Define $\ell_{i-1} := \| B^M_{[1:n]} \|_{\text{Lip}}$ for $i \in [n]$ and $\ell_n = 1$. Note that

$$
\mathbb{D}[F_1, F_0] = \sqrt{\mathbb{E}_x \mathcal{D} \left[ \left( B^M_{[2:n]} \circ B^M_1(x) - B^M_{[2:n]} \circ S^m_1(x) \right)^2 \right]}
\leq \ell_1 \sqrt{\mathbb{E}_x \mathcal{D} \left[ \left( B^M_1(x) - S^m_1(x) \right)^2 \right]}
$$

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

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

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

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

\begin{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 $\ell_B$. An important future direction is to develop rigorous theoretical bounds for controlling $\ell_B$.
\end{remark}