A. Supporting Lemmas

We start by presenting various technical lemmas that support the main proofs. Lemma 1 shows the expectation of moments of order statistics of the uniform distribution. This lemma is used in the magnitude-based pruning result of FCNs.

Lemma 1. *Given n independent and identically distributed random variables $U_1, \ldots, U_n \sim \mathcal{U}\left[-a, a\right]$ and $X_i = U_i^2, i \in [n]$, we have*

$$\mathbb{E}X_{(r)} = a^2 \frac{(r+1)r}{(n+2)(n+1)} \quad \textit{and} \quad \mathbb{E}X_{(r)}^2 = a^4 \frac{(r+3)(r+2)(r+1)r}{(n+4)(n+3)(n+2)(n+1)},$$

where $r \leqslant n$ is a constant and $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ are order statistics of $X_1, \ldots, X_n$.

Proof. Note that for $0 \leqslant x \leqslant a^2$, we have $F(x) \triangleq \mathbb{P}\left(X_i \leqslant x\right) = \mathbb{P}\left(U_i^2 \leqslant x\right) = \mathbb{P}\left(-\sqrt{x} \leqslant U_i \leqslant \sqrt{x}\right) = \frac{\sqrt{x}}{a}$. Therefore, the probability density function of $X_{(r)}$ is given by

$$\begin{aligned} f_{(r)}(x) &= \frac{n!}{(r-1)!(n-r)!}\left[F(x)\right]^{r-1}\left[1-F(x)\right]^{n-r}F'(x) \\ &= \frac{n!}{(r-1)!(n-r)!}\left[\frac{\sqrt{x}}{a}\right]^{r-1}\left[1-\frac{\sqrt{x}}{a}\right]^{n-r}\frac{1}{2a\sqrt{x}}. \end{aligned}$$

For $p \in \mathbb{Z}$, we have

$$\begin{aligned} \mathbb{E}X_{(r)}^p &= \int_0^{a^2} x^p f_{(r)}(x)\mathrm{d}x \\ &= \int_0^{a^2} x^p \frac{n!}{(r-1)!(n-r)!}\left[\frac{\sqrt{x}}{a}\right]^{r-1}\left[1-\frac{\sqrt{x}}{a}\right]^{n-r}\frac{1}{2a\sqrt{x}}\mathrm{d}x \\ &= \frac{n!}{(r-1)!(n-r)!}\frac{1}{2a^n}\int_0^{a^2} x^p x^{\frac{r}{2}-1}(a-\sqrt{x})^{n-r}\mathrm{d}x \\ &= \frac{n!}{(r-1)!(n-r)!}\frac{1}{2a^n}\int_0^1 (at)^{2p+r-2}(a-at)^{n-r}2a^2t\mathrm{d}t \\ &= \frac{a^{2p}n!}{(r-1)!(n-r)!}\int_0^1 t^{r+2p-1}(1-t)^{n-r}\mathrm{d}t \\ &= \frac{a^{2p}n!}{(r-1)!(n-r)!}\frac{(n-r)!(r+2p-1)!}{(n+2p)!} \\ &= \frac{(r+2p-1)!n!}{(r-1)!(n+2p)!}a^{2p}. \end{aligned}$$

Specifically, we have $\mathbb{E}X_{(r)} = a^2\frac{(r+1)r}{(n+2)(n+1)}$ and $\mathbb{E}X_{(r)}^2 = a^4\frac{(r+3)(r+2)(r+1)r}{(n+4)(n+3)(n+2)(n+1)}$. □

Next, we present some results for sub-Gaussian random matrices. We first give the definition of sub-Gaussian random variables in the following.

Definition 1. A random variable $X \in \mathbb{R}$ is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}X = 0$ and its moment generating function satisfies

$$\mathbb{E}\exp\left[sX\right] \leqslant \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}.$$

In this case, we write $X \sim \mathsf{subG}(\sigma^2)$.

Note that $\mathsf{subG}(\sigma^2)$ denotes a class of distributions rather than a single distribution. Many common distributions, like Gaussian and any bounded distributions with zero expectation, all fall into this category. If $X \sim \mathsf{subG}(\sigma^2)$, then we have $\mathsf{var}\left(X\right) = \mathbb{E}X^2 \leqslant \sigma^2$.

Lemma 2 (Proposition 2.4 of Rudelson & Vershynin (2010))**.** *Let A be a $n_1 \times n_2$ random matrix whose entries are independent mean zero sub-Gaussian random variables whose sub-Gaussian variance proxy are bounded by 1. Then there exists universal positive constants c and C such that, for any $t > 0$ we have*

$$\mathbb{P}\left(\|A\|_2 > C\left(\sqrt{n_1} + \sqrt{n_2}\right) + t\right) \leqslant 2e^{-ct^2}. \tag{12}$$

Lemma 3. *Let B be a $n_1 \times n_2$ random matrix whose entries are independently and identically distributed following $\mathcal{U}\left[-\frac{K}{\sqrt{n}}, \frac{K}{\sqrt{n}}\right]$, where K is a positive constant and $n = \max\{n_1, n_2\}$. Then there exist positive constants c_0 (depends on K) and δ_0 such that $\|B\|_2 \leqslant c_0$ with probability at least $1 - 2e^{-4\delta_0 n}$.*

Proof. Let us denote $A = \frac{\sqrt{3n}}{K}B$. Then the entries in A are independently and identically distributed following $\mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, which belongs to the sub-Gaussian distribution with variance proxy 1. Applying Lemma 2, we know that there exist positive constants C and δ_0 such that

$$\mathbb{P}\left(\|A\|_2 > 2C\sqrt{n} + t\right) \leqslant \mathbb{P}\left(\|A\|_2 > C\left(\sqrt{n_1} + \sqrt{n_2}\right) + t\right) \leqslant 2e^{-\delta_0 t^2}.$$

Taking $t = 2\sqrt{n}$, we have

$$\mathbb{P}\left(\|B\|_2 > \frac{2K}{\sqrt{3}}(C+1)\right) = \mathbb{P}\left(\frac{K}{\sqrt{3n}}\|A\|_2 > \frac{2K}{\sqrt{3}}(C+1)\right) = \mathbb{P}\left(\|A\|_2 > 2\sqrt{n}(C+1)\right) \leqslant 2e^{-4\delta_0 n},$$

and therefore

$$\mathbb{P}\left(\|B\|_2 \leqslant c_0\right) \geqslant 1 - 2e^{-4\delta_0 n},$$

where $c_0 = \frac{2K}{\sqrt{3}}(C+1) > 0$. □

In the two lemmas above, we assume certain distributions for the entries in the random matrices. The following lemma is more general in the sense that it only requires the entries in the matrices to be independent.

Lemma 4 (Theorem 2 of Latała (2005))**.** *Let A be a random matrix whose entries $A_{i,j}$ are independent mean zero random variables with finite fourth moment. Then*

$$\mathbb{E}\|A\|_2 \leqslant C\left[\max_i\left(\sum_j \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}} + \max_j\left(\sum_i \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}} + \left(\sum_{i,j}\mathbb{E}A_{i,j}^4\right)^{\frac{1}{4}}\right], \tag{13}$$

where C is an universal positive constant.

The proofs of the main theorems in this paper heavily rely on Lemmas 3 and 4. Note that there are some universal constants in the statement of these two lemmas that all appear in the bounds of the main theorems. Thus we give a numerical study of these two lemmas in Appendix D.3 and D.4.

Lemma 5 (Chernoff Bound)**.** *Suppose $X_1, \ldots, X_m$ are independent random variables taking values in $\{0, 1\}$. Let $X := \sum_{i=1}^m X_i$ and $\mu := \mathbb{E}X$. Then for any $\delta > 0$, we have*

$$\mathbb{P}\left(X \geqslant (1+\delta)\mu\right) \leqslant \exp\left(-\frac{\delta^2}{1+\delta}\mu\right). \tag{14}$$

The next lemma results from the famous problem "balls-into-bins." This is a classic problem in probability theory that has many applications in computer science. See the survey paper by Richa et al. (2001) for more details.

Lemma 6. *Consider the problem of throwing N balls independently and uniformly at random into n bins. Let X_j be the random variable that counts the number of balls in the j-th bin, $1 \leqslant j \leqslant n$. If $N \geqslant n\log(n)$, then with probability at least $1 - n^{-\frac{1}{3}}$ we have $\max_{j\in[n]} X_j \leqslant \frac{3N}{n}$.*

Proof. Let X_{ij} be the indicator random variable for the event that the i-th ball falls into the j-th bin, $i \in [N], j \in [n]$. Then $\mathbb{E}X_j = \sum_{i=1}^{N} \mathbb{E}X_{ij} = \frac{N}{n}, j \in [n]$. Note that $\mu = \frac{N}{n} \geqslant \log(n)$, applying Lemma 5 with $\delta = 2$, we have

$$\mathbb{P}\left(X_j \geqslant 3\frac{N}{n}\right) \leqslant \exp\left(-\frac{4}{3}\mu\right) \leqslant \exp\left(-\frac{4}{3}\log(n)\right) = n^{-\frac{4}{3}}.$$

By the union bound we have

$$\mathbb{P}\left(\max_{j\in[n]} X_j \geqslant \frac{3N}{n}\right) = \mathbb{P}\left(\bigcup_{j\in[n]}\left\{X_j \geqslant \frac{3N}{n}\right\}\right) \leqslant \sum_{j\in[n]} \mathbb{P}\left(X_j \geqslant \frac{3N}{n}\right) \leqslant n \cdot n^{-\frac{4}{3}} = n^{-\frac{1}{3}},$$

and therefore

$$\mathbb{P}\left(\max_{j\in[n]} X_j \leqslant \frac{3N}{n}\right) \geqslant 1 - n^{-\frac{1}{3}}.$$

□

The last lemma focuses on the singular values of the matrix representation of convolutional operators. Given a convolutional tensor $\mathcal{F} \in \mathbb{R}^{d\times d\times q\times q}$, the corresponding matrix representation W of $\mathcal{F}$ has dimension $p^2 d \times p^2 d$, where p is the width and height of the input feature map. Applying the traditional singular value decomposition methods on such a large matrix is usually time-consuming and computationally-inefficient. Sedghi et al. (2018) provide tools to represent the set of singular values of W by the joint of sets of singular values of many smaller sub-matrices. This is done by carefully analyzing the properties of ciucular-type matrices. We use the following lemma from Sedghi et al. (2018) to calculate the L_2 norm of the weight matrices in CNNs.

Lemma 7 (Theorem 6 of Sedghi et al. (2018))**.** *Let $\omega = \exp(2\pi i/p)$, where $i = \sqrt{-1}$ and S be the $p \times p$ matrix that represents the discrete Fourier transform*

$$S := \begin{bmatrix} \omega^{1\times 1} & \cdots & \omega^{1\times p} \\ \vdots & \ddots & \vdots \\ \omega^{p\times 1} & \cdots & \omega^{p\times p} \end{bmatrix}.$$

Given a tensor $\mathcal{F} \in \mathbb{R}^{d\times d\times q\times q}$, let us denote $K \in \mathbb{R}^{d\times d\times p\times p}$ as defined in (7) and we denote $W^* \in \mathbb{R}^{dp^2\times dp^2}$ as the matrix encoding the linear transformation computed by the convolutional layer parameterized by K, as defined in (8) – (9). Let $P^{(u,v)}$ be the $d \times d$ matrix such that the (s,t)-th element of $P^{(u,v)}$ is equal to the (u,v)-th element of $S^T K_{s,t,:,:} S$, $u, v \in [p], s, t, \in [d]$, or equivalently

$$P_{s,t}^{(u,v)} = \left(S^T K_{s,t,:,:} S\right)_{u,v}, \quad u, v \in [p], s, t, \in [d].$$

Then

$$\|W^*\|_2 = \max_{u,v\in[p]} \left\{\left\|P^{(u,v)}\right\|_2\right\}.$$

B. Proofs

In this section, we provide the full proof of Theorems 1, 2, and 3. Note that the proofs of these three theorems are similar. Theorem 2 exhibits all ideas and thus it is presented in full. The proofs of the other theorems show the difference.

B.1. Proof of Theorem 2

Proof. For any $x \in \mathcal{B}_{d_0}$ and $1 \leqslant k < l$, we denote $y_k(x) := \sigma_k\left(W_k \sigma_{k-1}\left(\cdots W_2 \sigma_1\left(W_1 x\right)\right)\right)$ and $y_k^*(x) := \sigma_k\left(W_k^* \sigma_{k-1}\left(\cdots W_2^* \sigma_1\left(W_1^* x\right)\right)\right)$ as the output of the k-th layer of f and F, respectively.

Recall that we set M_1 and M_l as the all 1 matrices, i.e. $W_1 = W_1^*$ and $W_l = W_l^*$. For each $1 < k < l$, we order the entries of W_k^* by their absolute values such that

$$\left|(W_k^*)_{i_1^k, j_1^k}\right| \leqslant \left|(W_k^*)_{i_2^k, j_2^k}\right| \leqslant \cdots \leqslant \cdots \leqslant \left|(W_k^*)_{i_{D_k}^k, j_{D_k}^k}\right|$$

and denote $\mathcal{I}_k := \left\{(i_s^k, j_s^k) : 1 \leqslant s \leqslant \lfloor D_k^{1-\alpha} \rfloor\right\}$. We set $(M_k)_{i,j} = 0$ if $(i,j) \in \mathcal{I}_k$, and $(M_k)_{i,j} = 1$ otherwise. We further denote two events

$$A_r^{(k)} := \left\{\text{the number of zero entries in each row of } M_k \text{ is at most } 3\lfloor D_k^{1-\alpha} \rfloor / d_k\right\},$$
$$A_c^{(k)} := \left\{\text{the number of zero entries in each column of } M_k \text{ is at most } 3\lfloor D_k^{1-\alpha} \rfloor / d_{k-1}\right\}$$

and set event $A^{(k)} := A_r^{(k)} \bigcap A_c^{(k)}$. Note that (3) and (4) guarantee that $\lfloor D_k^{1-\alpha} \rfloor \geqslant d_k \log(d_k)$ and $D_k^{1-\alpha} \geqslant d_{k-1} \log(d_{k-1})$, respectively, and the events $A_r^{(k)}$ and $A_c^{(k)}$ are independent. Thus by Lemma 6, we have

$$\mathbb{P}\left(A^{(k)}\right) = \mathbb{P}\left(A_r^{(k)} \bigcap A_c^{(k)}\right) = \mathbb{P}\left(A_r^{(k)}\right)\mathbb{P}\left(A_c^{(k)}\right) \geqslant \left(1 - d_k^{-\frac{1}{3}}\right)\left(1 - d_{k-1}^{-\frac{1}{3}}\right) \geqslant \left(1 - d^{-\frac{1}{3}}\right)^2.$$

Further, for $A := A^{(2)} \bigcap \cdots \bigcap A^{(l-1)}$, we have $\mathbb{P}(A) = \prod_{k=2}^{l-1} \mathbb{P}\left(A^{(k)}\right) \geqslant \left(1 - d^{-\frac{1}{3}}\right)^{2(l-2)}$ where the probability is taken over the randomness of masks (and is not over the randomness of weights in W_k^*'s).

Let us assume that

$$d^{-\frac{1}{4}\alpha} \leqslant \min\left\{N_2, \ldots, N_{l-1}, \frac{\epsilon}{(2^{l-2} - 1)\, L_{1:l-1} N_{1:l}}\right\}. \tag{15}$$

We use induction to show that, for any $x \in \mathcal{B}_{d_0}$ and $1 \leqslant k < l$,

(I) with probability at least $\prod_{i=1}^{k}(1 - \delta_i)$, we have $\|y_k^*(x)\|_2 \leqslant L_{1:k} N_{1:k}$,

(II) with probability at least $1 - (k-1)c_2 d^{-\frac{\alpha}{4}} - 2(k-1)d^{-\frac{1}{3}} - \sum_{i=1}^{k}(k+1-i)\delta_i$, we have $\left\|\left(y_k(x) \middle| A\right) - y_k^*(x)\right\|_2 \leqslant \left(2^{k-1} - 1\right) d^{-\frac{1}{4}\alpha} L_{1:k} N_{1:k}$ for some positive constant c_2 specified later[8].

The case of $k = 1$ is as follows. Note that for any vector v, we have $\|\sigma_1(v)\|_2 = \|\sigma_1(v) - \sigma_1(0)\|_2 \leqslant L_1 \|v - 0\|_2 = L_1 \|v\|_2$. Thus, $\|y_1^*(x)\|_2 = \|\sigma_1(W_1^* x)\|_2 \leqslant L_1 \|W_1^* x\|_2 \leqslant L_1 \|W_1^*\|_2 \|x\|_2 \leqslant L_1 N_1$ with probability at least $1 - \delta_1$. Further, we have $y_1(x) = \sigma_1(W_1 x) = \sigma_1(W_1^* x) = y_1^*(x)$, and thus $\|y_1(x) - y_1^*(x)\|_2 = 0$.

Suppose the statement holds for $1 \leqslant k < l - 1$; we consider the case of $k+1$. Note that the events $\left\{\left\|W_{k+1}^*\right\|_2 \leqslant N_{k+1}\right\}$ and $\left\{\|y_k^*(x)\|_2 \leqslant L_{1:k} N_{1:k}\right\}$ are independent. By induction statement (I), with probability at least

$$\mathbb{P}\left(\left\{\left\|W_{k+1}^*\right\|_2 \leqslant N_{k+1}\right\} \bigcap \left\{\|y_k^*(x)\|_2 \leqslant L_{1:k} N_{1:k}\right\}\right) = \mathbb{P}\left(\left\|W_{k+1}^*\right\|_2 \leqslant N_{k+1}\right) \cdot \mathbb{P}\left(\|y_k^*(x)\|_2 \leqslant L_{1:k} N_{1:k}\right) \geqslant \prod_{i=1}^{k}(1 - \delta_i),$$

we have

$$\begin{aligned}\left\|y_{k+1}^*(x)\right\|_2 &= \left\|\sigma_{k+1}\left(W_{k+1}^* y_k^*(x)\right)\right\|_2 \leqslant L_{k+1}\left\|W_{k+1}^* y_k^*(x)\right\|_2 \leqslant L_{k+1}\left\|W_{k+1}^*\right\|_2 \|y_k^*(x)\|_2 \\ &\leqslant L_{k+1} N_{k+1} \cdot L_{1:k} N_{1:k} = L_{1:k+1} N_{1:k+1},\end{aligned}$$

which shows (I) in the induction statement.

We next show that (II) holds. Under event A, the number of non-zero entries in each row of $\mathcal{W}_{k+1}$ is at most $3\lfloor D_{k+1}^{1-\alpha} \rfloor / d_{k+1}$. Thus we have

$$\max_{i \in [d_{k+1}]}\left(\sum_{j \in [d_k]} \mathbb{E}\left[(\mathcal{W}_{k+1})_{i,j}^2 \middle| A\right]\right)^{\frac{1}{2}} \leqslant \left(\frac{3\lfloor D_{k+1}^{1-\alpha} \rfloor}{d_{k+1}} \cdot \frac{K_1}{\max\{d_{k+1}, d_k\}}\right)^{\frac{1}{2}} \leqslant \left(\frac{3K_1 D_{k+1}^{1-\alpha}}{d_{k+1} d_k}\right)^{\frac{1}{2}} = \sqrt{3K_1} D_{k+1}^{-\frac{\alpha}{2}} \leqslant \sqrt{3K_1} d^{-\alpha}, \tag{16}$$

[8] Note that in the induction statement (II), the probability (and the expectations in the following context) is taken over the randomness of weights but not the masks. The random variable $\left\|\left(y_k(x) \middle| A\right) - y_k^*(x)\right\|_2$ is equivalent to $\left\|y_k(x) - y_k^*(x)\right\|_2 \Big| A$, and the statement can also be written as $\mathbb{P}\left(\left\{\left\|y_k(x) - y_k^*(x)\right\|_2 \leqslant \left(2^{k-1} - 1\right) d^{-\frac{1}{4}\alpha} L_{1:k} N_{1:k}\right\} \middle| A\right) \geqslant 1 - (k-1)c_2 d^{-\frac{\alpha}{4}} - 2(k-1)d^{-\frac{1}{3}} - \sum_{i=1}^{k}(k+1-i)\delta_i$.

and similarly,

$$\max_{j\in[d_k]} \left(\sum_{i\in[d_{k+1}]} \mathbb{E}\left[(\mathcal{W}_{k+1})_{i,j}^2 \middle| A\right]\right)^{\frac{1}{2}} \leqslant \sqrt{3K_1}d^{-\alpha}. \tag{17}$$

In addition, since there are at most $\lfloor D_{k+1}^{1-\alpha}\rfloor$ non-zero entries in $\mathcal{W}_{k+1} := W_{k+1} - W_{k+1}^*$, we have

$$\sum_{i\in[d_{k+1}],j\in[d_k]} \mathbb{E}\left[\left|(\mathcal{W}_{k+1})_{i,j}\right|^4 \middle| A\right] \leqslant \lfloor D_{k+1}^{1-\alpha}\rfloor \cdot \frac{K_2}{\max\left\{d_{k+1}, d_k\right\}^2} \leqslant D_{k+1}^{1-\alpha}\frac{K_2}{D_{k+1}} = K_2 D_{k+1}^{-\alpha} \leqslant K_2 d^{-2\alpha}. \tag{18}$$

Combining (16) – (18) and Lemma 4, there exists a universal positive constant c_1 such that

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{W}_{k+1}\|_2 | A\right] &\leqslant c_1 \left[\max_{i\in[d_{k+1}]}\left(\sum_{j\in[d_k]}\mathbb{E}\left(\mathcal{W}_{k+1}\right)_{i,j}^2\right)^{\frac{1}{2}} + \max_{j\in[d_k]}\left(\sum_{i\in[d_{k+1}]}\mathbb{E}\left(\mathcal{W}_{k+1}\right)_{i,j}^2\right)^{\frac{1}{2}} + \left(\sum_{i\in[d_{k+1}],j\in[d_k]}\mathbb{E}\left(\mathcal{W}_{k+1}\right)_{i,j}^4\right)^{\frac{1}{4}}\right] \\
&\leqslant c_1\left[\sqrt{3K_1}d^{-\alpha} + \sqrt{3K_1}d^{-\alpha} + \left(K_2 d^{-2\alpha}\right)^{\frac{1}{4}}\right] \qquad (19) \\
&\leqslant c_2 d^{-\frac{\alpha}{2}},
\end{aligned}$$

where $c_2 = c_1\left(2\sqrt{3K_1} + K_2^{\frac{1}{4}}\right)$.

By the Markov's inequality, for all $t > 0$ we have

$$\mathbb{P}\left(\left\{\|\mathcal{W}_{k+1}\|_2 \geqslant t\right\} \middle| A\right) \leqslant \frac{\mathbb{E}\left[\|\mathcal{W}_{k+1}\|_2 | A\right]}{t}.$$

Taking $t = d^{-\frac{\alpha}{4}}$, we have

$$\mathbb{P}\left(\left\{\|\mathcal{W}_{k+1}\|_2 \leqslant d^{-\frac{\alpha}{4}}\right\} \middle| A\right) \geqslant 1 - c_2 d^{-\frac{\alpha}{4}}.$$

By induction statement (II) and the fact that $\mathbb{P}\left(\bigcap_{i=1}^s A_i\right) \geqslant \sum_{i=1}^s \mathbb{P}\left(A_i\right) - (s-1)$, with probability at least[9]

$$\begin{aligned}
&\mathbb{P}\Bigg(\left\{\left\|\left(W_{k+1}|A\right) - W_{k+1}^*\right\|_2 \leqslant d^{-\frac{\alpha}{4}}\right\} \bigcap \left\{\left\|W_{k+1}^*\right\|_2 \leqslant N_{k+1}\right\} \bigcap \left\{\left\|y_k^*(x)\right\|_2 \leqslant L_{1:k}N_{1:k}\right\} \\
&\qquad\qquad \bigcap \left\{\left\|\left(y_k(x)|A\right) - y_k^*(x)\right\|_2 \leqslant \left(2^{k-1}-1\right)d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}\right\}\Bigg) \qquad (20) \\
&\geqslant \left(1 - c_2 d^{-\frac{\alpha}{4}}\right) + \left(1-\delta_{k+1}\right) + \prod_{i=1}^k\left(1-\delta_i\right) + \left(1 - (k-1)c_2 d^{-\frac{\alpha}{4}} - \sum_{i=1}^k (k+1-i)\delta_i\right) - 3 \\
&\geqslant \left(1 - c_2 d^{-\frac{\alpha}{4}}\right) + \left(1-\delta_{k+1}\right) + \left(1 - \sum_{i=1}^k \delta_i\right) + \left(1 - (k-1)c_2 d^{-\frac{\alpha}{4}} - \sum_{i=1}^k (k+1-i)\delta_i\right) - 3 \\
&= 1 - kc_2 d^{-\frac{\alpha}{4}} - \sum_{i=1}^{k+1}(k+2-i)\delta_i,
\end{aligned}$$

[9] We use the fact that, for any $a_1, \cdots, a_s \in (0,1)$, we have $\prod_{i=1}^s (1-a_i) \geqslant 1 - \sum_{i=1}^s a_i$. This inequality is frequently used in the following proofs.

we have

$$\begin{align}
&\left\|\left(y_{k+1}(x)\middle|A\right)-y_{k+1}^{*}(x)\right\|_2 \notag\\
=\ &\left\|\sigma_{k+1}\left((W_{k+1}|A)y_k(x)\right)-\sigma_{k+1}\left(W_{k+1}^{*}y_k^{*}(x)\right)\right\|_2 \notag\\
\leqslant\ &L_{k+1}\left\|\left(W_{k+1}|A\right)y_k(x)-W_{k+1}^{*}y_k^{*}(x)\right\|_2 \notag\\
\leqslant\ &L_{k+1}\left[\left\|\left(W_{k+1}|A\right)y_k(x)-W_{k+1}y_k^{*}(x)\right\|_2+\left\|\left(W_{k+1}|A\right)y_k^{*}(x)-W_{k+1}^{*}y_k^{*}(x)\right\|_2\right] \notag\\
\leqslant\ &L_{k+1}\left[\left\|\left(W_{k+1}|A\right)\right\|_2\left\|y_k(x)-y_k^{*}(x)\right\|_2+\left\|\left(W_{k+1}|A\right)-W_{k+1}^{*}\right\|_2\left\|y_k^{*}(x)\right\|_2\right] \notag\\
\leqslant\ &L_{k+1}\left[\left(\left\|W_{k+1}^{*}\right\|_2+\left\|\left(W_{k+1}|A\right)-W_{k+1}^{*}\right\|_2\right)\left\|y_k(x)-y_k^{*}(x)\right\|_2+\left\|\left(W_{k+1}|A\right)-W_{k+1}^{*}\right\|_2\left\|y_k^{*}(x)\right\|_2\right] \notag\\
\leqslant\ &L_{k+1}\left[\left(N_{k+1}+d^{-\frac{1}{4}\alpha}\right)\left\|\left(y_k(x)\middle|A\right)-y_k^{*}(x)\right\|_2+d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}\right] \tag{21}\\
\leqslant\ &L_{k+1}\left[2N_{k+1}\left\|\left(y_k(x)\middle|A\right)-y_k^{*}(x)\right\|_2+d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}\right] \notag\\
\leqslant\ &L_{k+1}\left[2N_{k+1}\cdot\left(2^{k-1}-1\right)d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}+d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}\right] \notag\\
\leqslant\ &L_{k+1}\left[2N_{k+1}\cdot\left(2^{k-1}-1\right)d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k}+d^{-\frac{1}{4}\alpha}L_{1:k}N_{1:k+1}\right] \notag\\
=\ &\left(2^{k}-1\right)d^{-\frac{1}{4}\alpha}L_{1:k+1}N_{1:k+1} \tag{22}
\end{align}$$

where in (21) we use assumption (15). This finishes the induction.

We have just shown that with probability at least $1-(l-2)c_2d^{-\frac{\alpha}{4}}-\sum_{i=1}^{l-1}(l-i)\delta_i$, we have

$$\left\|\left(y_{l-1}(x)\middle|A\right)-y_{l-1}^{*}(x)\right\|_2\leqslant\left(2^{l-2}-1\right)d^{-\frac{1}{4}\alpha}L_{1:l-1}N_{1:l-1}.$$

For the last layer, by assumption, with probability at least $(1-\delta_l)\cdot\left[1-(l-2)c_2d^{-\frac{\alpha}{4}}-\sum_{i=1}^{l-1}(l-i)\delta_i\right]$, we have for every $x\in\mathcal{B}_{d_0}$,

$$\begin{aligned}
\left\|\left(f(x)\middle|A\right)-F(x)\right\|_2&=\left\|W_l\left(y_{l-1}(x)\middle|A\right)-W_l^{*}y_{l-1}^{*}(x)\right\|_2\\
&=\left\|W_l^{*}\left(y_{l-1}(x)\middle|A\right)-W_l^{*}y_{l-1}^{*}(x)\right\|_2\\
&\leqslant\left\|W_l^{*}\right\|_2\left\|\left(y_{l-1}(x)\middle|A\right)-y_{l-1}^{*}(x)\right\|_2\\
&\leqslant N_l\left(2^{l-2}-1\right)d^{-\frac{1}{4}\alpha}L_{1:l-1}N_{1:l-1}\\
&=\left(2^{l-2}-1\right)d^{-\frac{1}{4}\alpha}L_{1:l-1}N_{1:l}\\
&\leqslant\epsilon,
\end{aligned}$$

where the last inequality follows from assumption (15). In conclusion, with probability at least $\mathbb{P}\left(A\right)\geqslant\left(1-d^{-\frac{1}{3}}\right)^{2(l-2)}$ over the randomness of masks, we have $\sup_{x\in\mathcal{B}_{d_0}}\left\|\left(f(x)\middle|A\right)-F(x)\right\|_2\leqslant\epsilon$ with probability at least $(1-\delta_l)\cdot\left[1-(l-2)c_2d^{-\frac{\alpha}{4}}-\sum_{i=1}^{l-1}(l-i)\delta_i\right]$. As a result, basic probability yields that with probability at least

$$p_0:=\left(1-d^{-\frac{1}{3}}\right)^{2(l-2)}(1-\delta_l)\left[1-(l-2)c_2d^{-\frac{\alpha}{4}}-\sum_{i=1}^{l-1}(l-i)\delta_i\right],$$

we have

$$\sup_{x\in\mathcal{B}_{d_0}}\left\|f(x)-F(x)\right\|_2\leqslant\epsilon.$$

It remains to determine a lower bound of d such that

$$d^{-\frac{1}{4}\alpha}\leqslant\min\left\{N_2,\ldots,N_{l-1},\frac{\epsilon}{\left(2^{l-2}-1\right)L_{1:l-1}N_{1:l}}\right\} \tag{23}$$

and

$$p_0 \geqslant 1-\delta. \tag{24}$$

For (23), we have

$$d \geqslant N_k^{-\frac{4}{\alpha}}, \quad 2 \leqslant k \leqslant l-1 \quad \text{and} \quad d \geqslant \left(\left(2^{l-2}-1\right) L_{1:l-1} N_{1:l}\right)^{\frac{4}{\alpha}} \cdot \epsilon^{-\frac{4}{\alpha}}. \tag{25}$$

Regarding (24), condition (5) guarantees that $\delta_0 = \delta - \left[\delta_l + \sum_{i=1}^{l-1}(l-i)\delta_i\right] \geqslant 0$. We have

$$2(l-2)d^{-\frac{1}{3}} \leqslant \frac{2}{3}\delta_0 \quad \Leftrightarrow \quad d \geqslant \delta_0^{-3}\left(3(l-2)\right)^3, \tag{26}$$

$$(l-2)c_2 d^{-\frac{\alpha}{4}} \leqslant \frac{1}{3}\delta_0 \quad \Leftrightarrow \quad d \geqslant \delta_0^{-\frac{4}{\alpha}}\left(3c_2(l-2)\right)^{\frac{4}{\alpha}}. \tag{27}$$

Combining (25) - (27), we know that if

$$d \geqslant \max\left\{C_1^{\frac{4}{\alpha}}, \left(\frac{C_2}{\epsilon}\right)^{\frac{4}{\alpha}}, \left(\frac{C_3}{\delta_0}\right)^3, \left(\frac{C_4}{\delta_0}\right)^{\frac{4}{\alpha}}\right\},$$

for some positive constant C_1, C_2, C_3 and C_4, then with probability at least

$$\begin{aligned}
p_0 &= \left(1-d^{-\frac{1}{3}}\right)^{2(l-2)}(1-\delta_l)\cdot\left[1-(l-2)c_2 d^{-\frac{\alpha}{4}} - \sum_{i=1}^{l-1}(l-i)\delta_i\right] \\
&\geqslant 1-(l-2)c_2 d^{-\frac{\alpha}{4}} - 2(l-2)d^{-\frac{1}{3}} - \left[\delta_l + \sum_{i=1}^{l-1}(l-i)\delta_i\right] \\
&\geqslant 1-\frac{2}{3}\delta_0 - \frac{1}{3}\delta_0 - (\delta-\delta_0) \\
&= 1-\delta,
\end{aligned}$$

we have

$$\sup_{x\in\mathcal{B}_{d_0}} \|f(x)-F(x)\|_2 \leqslant \epsilon.$$

□

B.2. Proof of Theorem 1

Proof. For any $x \in \mathcal{B}_{d_0}$ and $1 \leqslant k < l$, we denote $y_k(x) = \sigma\left(W_k\sigma\left(\cdots W_2\sigma\left(W_1x\right)\right)\right)$ and $y_k^*(x) = \sigma\left(W_k^*\sigma\left(\cdots W_2^*\sigma\left(W_1^*x\right)\right)\right)$ as the output of the k-th layer of f and F, respectively.

Recall that we set M_1 and M_l as the all 1 matrices, i.e. $W_1 = W_1^*$ and $W_l = W_l^*$. For each $1 < k < l$, we order the entries of W_k^* by their absolute values such that

$$\left|(W_k^*)_{i_1^k, j_1^k}\right| \leqslant \left|(W_k^*)_{i_2^k, j_2^k}\right| \leqslant \cdots \leqslant \cdots \leqslant \left|(W_k^*)_{i_{D_k}^k, j_{D_k}^k}\right|$$

and denote $\mathcal{I}_k := \left\{(i_s^k, j_s^k) : 1 \leqslant s \leqslant \lfloor D_k^{1-\alpha}\rfloor\right\}$. We set $(M_k)_{i,j} = 0$ if $(i,j) \in \mathcal{I}_k$, and $(M_k)_{i,j} = 1$ otherwise. In the following, we show that $M_1, \ldots, M_l$ defined above satisfy (2).

By Lemma 3, there exist positive constants c_0 (depends on K) and δ_0 such that[10]

$$\mathbb{P}\left(\|W_k^*\|_2 \leqslant c_0\right) \geqslant 1-2e^{-4\delta_0 d}, \quad 1 \leqslant k \leqslant l. \tag{28}$$

[10] In fact, we get l different sets of $\{c_i, \delta_i\}$, $i \in [l]$ by applying Lemma 3 l times. We take $c_0 = \max\{c_i\}$ and $\delta_0 = \min\{\delta_i\}$ so that (28) is satisfied for all $1 \leqslant k \leqslant l$.

Let us assume that

$$d^{-\alpha} \leqslant \min\left\{c_0, \frac{\epsilon}{\left(2^{l-2}-1\right) L_{1:(l-1)} c_0^{l-1}}\right\}. \tag{29}$$

We use induction to show that, for any $x \in \mathcal{B}_{d_0}$ and $1 \leqslant k < l$,

(I) with probability at least $\left(1 - 2e^{-4\delta_0 d}\right)^k$, we have $\left\|y_k^*(x)\right\|_2 \leqslant L_{1:k} c_0^k$

(II) with probability at least $1 - (k-1)c_2 d^{-\alpha} - (k+2)(k-1)e^{-4\delta_0 d}$, we have $\left\|y_k(x) - y_k^*(x)\right\|_2 \leqslant \left(2^{k-1}-1\right) d^{-\alpha} L_{1:k} c_0^{k-1}$.

Statement (I) can be proved in the same way as in the proof of Theorem 2. We next show that (II) holds. The case of $k = 1$ is trivial since $y_1(x) = y_1^*(x)$. Suppose the statement holds for $1 \leqslant k < l-1$; we consider the case of $k+1$. Note that the non-zero entries of $\mathcal{W}_{k+1} := W_{k+1} - W_{k+1}^*$ are $\left\{\left(W_{k+1}^*\right)_{i,j} : (i,j) \in \mathcal{I}_{k+1}\right\}$. Taking $a = \frac{K}{\sqrt{\max\{d_k, d_{k+1}\}}}$, $n = D_{k+1}$, $r = \lfloor D_{k+1}^{1-\alpha} \rfloor$ in Lemma 1, for every entry e of $\mathcal{W}_{k+1}$, we have

$$\begin{aligned}
\mathbb{E}\, e^2 &\leqslant \mathbb{E}\left|\left(W_{k+1}^*\right)_{i^{k+1}_{\lfloor D_{k+1}^{1-\alpha}\rfloor}, j^{k+1}_{\lfloor D_{k+1}^{1-\alpha}\rfloor}}\right|^2 = \frac{K^2}{\max\{d_k, d_{k+1}\}} \cdot \frac{\lfloor D_{k+1}^{1-\alpha}\rfloor \left(\lfloor D_{k+1}^{1-\alpha}\rfloor + 1\right)}{(D_{k+1}+1)(D_{k+1}+2)} \\
&\leqslant \frac{K^2}{\max\{d_k, d_{k+1}\}} \cdot \frac{D_{k+1}^{1-\alpha}\left(D_{k+1}^{1-\alpha}+1\right)}{(D_{k+1}+1)(D_{k+1}+2)} \leqslant \frac{K^2}{\max\{d_k, d_{k+1}\}} \cdot 2D_{k+1}^{-2\alpha} \leqslant 2K^2 d^{-1-4\alpha},
\end{aligned}$$

and similarly

$$\mathbb{E}\, e^4 \leqslant \mathbb{E}\left|\left(W_{k+1}^*\right)_{i^{k+1}_{\lfloor D_{k+1}^{1-\alpha}\rfloor}, j^{k+1}_{\lfloor D_{k+1}^{1-\alpha}\rfloor}}\right|^4 \leqslant 24K^4 d^{-2-8\alpha}.$$

Taking $A = \mathcal{W}_{k+1}$ in Lemma 4, we know there exists a constant $c_2 > 0$ such that $\mathbb{E}\left\|\mathcal{W}_{k+1}\right\|_2 \leqslant c_2 d^{-2\alpha}$, where $c_2 = CK\left(2\sqrt{2} + (24)^{1/4}\right)$ and C is the universal constant as defined in Lemma 4. By Markov's inequality, for all $t > 0$ we have $\mathbb{P}\left(\left\|\mathcal{W}_{k+1}\right\|_2 \geqslant t\right) \leqslant \frac{\mathbb{E}\|\mathcal{W}_{k+1}\|_2}{t}$. Taking $t = d^{-\alpha}$, we have

$$\mathbb{P}\left(\left\|\mathcal{W}_{k+1}\right\|_2 \leqslant d^{-\alpha}\right) \geqslant 1 - c_2 d^{-\alpha}.$$

Similar to (20) – (22) in the proof of Theorem 2, with probability at least

$$\begin{aligned}
&\mathbb{P}\Bigg(\left\{\left\|W_{k+1} - W_{k+1}^*\right\|_2 \leqslant d^{-\alpha}\right\} \bigcap \left\{\left\|W_{k+1}^*\right\|_2 \leqslant c_0\right\} \bigcap \left\{\left\|y_k^*(x)\right\|_2 \leqslant L_{1:k} c_0^k\right\} \\
&\qquad\qquad \bigcap \left\{\left\|y_k(x) - y_k^*(x)\right\|_2 \leqslant \left(2^{k-1}-1\right) d^{-\alpha} L_{1:k} c_0^{k-1}\right\}\Bigg) \qquad (30) \\
&\geqslant \left(1 - c_2 d^{-\alpha}\right) + \left(1 - 2e^{-4\delta_0 d}\right) + \left(1 - 2e^{-4\delta_0 d}\right)^k + \left(1 - (k-1)c_2 d^{-\alpha} - (k+2)(k-1)e^{-4\delta_0 d}\right) - 3 \\
&\geqslant 1 - k c_2 d^{-\alpha} - (k+3)k e^{-4\delta_0 d},
\end{aligned}$$

we have

$$\left\|y_{k+1}(x) - y_{k+1}^*(x)\right\|_2 \leqslant \left(2^k - 1\right) d^{-\alpha} L_{1:(k+1)} c_0^k,$$

which finishes the induction.

We have just shown that with probability at least $1 - (l-2)c_2 d^{-\alpha} - (l+1)(l-2)e^{-4\delta_0 d}$, we have

$$\left\|y_{l-1}(x) - y_{l-1}^*(x)\right\|_2 \leqslant \left(2^{l-2} - 1\right) d^{-\alpha} L_{1:(l-1)} c_0^{l-2}.$$

For the last layer, by (28), with probability at least $\left(1 - 2e^{-4\delta_0 d}\right) \cdot \left(1 - (l-2)c_2 d^{-\alpha} - (l+1)(l-2)e^{-4\delta_0 d}\right)$, we have for every $x \in \mathcal{B}_{d_0}$,

$$\left\|f(x) - F(x)\right\|_2 \leqslant c_0 \cdot \left(2^{l-2} - 1\right) d^{-\alpha} L_{1:(l-1)} c_0^{l-2} \leqslant \epsilon,$$

where the last inequality follows from assumption (29). In conclusion, we show that with probability at least

$$p_0 := \left(1 - 2e^{-4\delta_0 d}\right) \cdot \left(1 - (l-2)c_2 d^{-\alpha} - (l+1)(l-2)e^{-4\delta_0 d}\right),$$

we have

$$\sup_{x \in \mathcal{B}_{d_0}} \|f(x) - F(x)\|_2 \leqslant \epsilon.$$

It remains to determine a lower bound of d such that

$$d^{-\alpha} \leqslant \min\left\{c_0, \frac{\epsilon}{\left(2^{l-2}-1\right) L_{1:(l-1)} c_0^{l-1}}\right\} \tag{31}$$

and

$$p_0 \geqslant 1 - \delta. \tag{32}$$

For (31), we have

$$d \geqslant c_0^{-\frac{1}{\alpha}} \quad \text{and} \quad d \geqslant \left(\frac{\left(2^{l-2}-1\right) L_{1:(l-1)} c_0^{l-1}}{\epsilon}\right)^{\frac{1}{\alpha}}. \tag{33}$$

Regarding (32), we have $p_0 \geqslant 1 - (l-2)c_2 d^{-\alpha} - (l^2 - l)e^{-4\delta_0 d}$. Note that $p_0 \geqslant 1 - \delta$ if $(l-2)c_2 d^{-\alpha} \leqslant \frac{l-2}{l^2-2}\delta$ and $(l^2 - l)e^{-4\delta_0 d} \leqslant \frac{l^2-l}{l^2-2}\delta$. These conditions are satisfied if

$$d \geqslant \left(\frac{(l^2-2)c_2}{\delta}\right)^{\frac{1}{\alpha}} \tag{34}$$

and

$$d \geqslant \frac{1}{4\delta_0}\left(\log\left(\frac{1}{\delta}\right) + \log(l^2-2)\right). \tag{35}$$

Combining (33) - (35), we know that if

$$d \geqslant \max\left\{C_1^{\frac{1}{\alpha}}, \left(\frac{C_2}{\epsilon}\right)^{\frac{1}{\alpha}}, \left(\frac{C_3}{\delta}\right)^{\frac{1}{\alpha}}, C_4 + C_5 \log\left(\frac{1}{\delta}\right)\right\},$$

for some positive constants C_1, C_2, C_3, C_4 and C_5 specified above, then with probability at least $1-\delta$ we have

$$\sup_{x \in \mathcal{B}_{d_0}} \|f(x) - F(x)\|_2 \leqslant \epsilon.$$

□

B.3. Proof of Theorem 3

Proof. Let $\mathcal{F}^{(k)} \in \mathbb{R}^{d_k \times d_{k-1} \times q_k \times q_k}$ be the corresponding convulotional tensor of W_k^* and $K^{(k)} \in \mathbb{R}^{d_k \times d_{k-1} \times p_k \times p_k}$ be as defined in (7). For any $x \in \mathcal{C}_{d_0}$ and $1 \leqslant k < l$, we denote $y_k(x) = \sigma\left(W_k \sigma\left(\cdots W_2 \sigma\left(W_1 x\right)\right)\right)$ and $y_k^*(x) = \sigma\left(W_k^* \sigma\left(\cdots W_2^* \sigma\left(W_1^* x\right)\right)\right)$ as the output of the k-th layer of f and F, respectively.

Recall that for $1 < k < l$, random pruning is based on 2D filters, i.e., we randomly select $\lfloor d^{2-\alpha} \rfloor$ pairs of indices (s', t') from $[d] \times [d]$ with replacement and set $\mathcal{F}^{(k)}_{s',t',:,:}$ to be zero. Denote $\mathcal{I}_k := \left\{(s', t') : \mathcal{F}^{(k)}_{s',t',:,:} \text{ is pruned}\right\}$ and $\mathcal{M}^{(k)}$ be the $d \times d$ matrix such that $\mathcal{M}^{(k)}_{s',t'} = 1$ if $(s', t') \in \mathcal{I}_k$ and $\mathcal{M}^{(k)}_{s',t'} = 0$ otherwise. We further denote two events

$$A_r^{(k)} := \left\{\text{the number of zero entries in each row of } \mathcal{M}^{(k)} \text{ is at most } 3\lfloor d^{2-\alpha} \rfloor / d\right\},$$
$$A_c^{(k)} := \left\{\text{the number of zero entries in each column of } \mathcal{M}^{(k)} \text{ is at most } 3\lfloor d^{2-\alpha} \rfloor / d\right\}$$

and set event $A^{(k)} := A_r^{(k)} \bigcap A_c^{(k)}$. Note that $\alpha \leqslant 2 - \frac{\log(d+1)+\log^{(2)}(d)}{\log(d)}$ guarantees that $\lfloor d^{2-\alpha} \rfloor \geqslant d\log(d)$ and the events $A_r^{(k)}$ and $A_c^{(k)}$ are independent. Thus by Lemma 6, we have

$$\mathbb{P}\left(A^{(k)}\right) = \mathbb{P}\left(A_r^{(k)} \bigcap A_c^{(k)}\right) = \mathbb{P}\left(A_r^{(k)}\right)\mathbb{P}\left(A_c^{(k)}\right) \geqslant \left(1 - d^{-\frac{1}{3}}\right)^2.$$

Further, for $A := A^{(2)} \bigcap \cdots \bigcap A^{(l-1)}$, we have $\mathbb{P}(A) = \prod_{k=2}^{l-1} \mathbb{P}\left(A^{(k)}\right) \geqslant \left(1 - d^{-\frac{1}{3}}\right)^{2(l-2)}$ where the probability is taken over the randomness of masks (and is not over the randomness of weights in W_k^*'s).

For $1 \leqslant k < l$, let $P^{(k,u,v)} \in \mathbb{R}^{d\times d}, u, v \in [p]$ be as defined in Lemma 7 such that[11]

$$\|W_k^*\|_2 = \max_{u,v\in[p]} \left\{ \left\| P^{(k,u,v)} \right\|_2 \right\}.$$

Recall that $\omega = \exp\left(2\pi\sqrt{-1}/p\right)$ and $S \in \mathbb{R}^{p\times p}$ is the matrix of the discrete Fourier transform. By Lemma 7, the (s,t)-th entry of $P^{(k,u,v)}$ can be written as

$$P_{s,t}^{(k,u,v)} = \left(S^T K_{s,t,:,:} S\right)_{u,v} = \sum_{i,j\in[p]} \omega^{ui} K_{s,t,i,j}^{(k)} \omega^{vj} = \sum_{i,j\in[q]} \omega^{ui} K_{s,t,i,j}^{(k)} \omega^{vj}, \quad s,t \in [d], u,v \in [p],$$

where the last equality is due to (7) since $K_{s,t,:,:}^{(k)}$ has non-zero entries only in its top-left $q \times q$ sub-matrix.

Denoting $P^{(k,u,v,i,j)} := \omega^{ui+vj} K_{:,:,i,j}^{(k)}, u,v \in [p], i,j \in [q]$, then we have $P^{(k,u,v)} := \sum_{i,j\in[q]} P^{(k,u,v,i,j)}$ and

$$\left\| P^{(k,u,v,i,j)} \right\|_2 = \left\| \omega^{ui+vj} K_{:,:,i,j}^{(k)} \right\|_2 = \left\| K_{:,:,i,j}^{(k)} \right\|_2 = \left\| \mathcal{F}_{:,:,i,j}^{(k)} \right\|_2, \quad u,v \in [p], i,j \in [q].$$

By assumption (iii), $\mathcal{F}_{:,:,i,j}^{(k)} \in \mathbb{R}^{d\times d}$ is a random matrix whose entries are independently sampled from different distributions. In addition, these distributions' second-order moments are upper-bounded by $\frac{C_1}{p^2 d}$ and the fourth-order moments are upper-bounded by $\frac{C_2}{p^4 d^2}$. By Lemma 4, for all $i,j \in [q]$, there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left\| \mathcal{F}_{i,j,:,:}^{(k)} \right\|_2 \leqslant C \left[\left(d\frac{C_1}{p^2 d} \right)^{\frac{1}{2}} + \left(d\frac{C_1}{p^2 d} \right)^{\frac{1}{2}} + \left(d^2 \frac{C_2}{p^4 d^2} \right)^{\frac{1}{4}} \right] \leqslant \frac{C_3}{p}, \tag{36}$$

where $C_3 = C\left(2\sqrt{C_1} + C_2^{1/4}\right)$.

Thus we have

$$\mathbb{E}\|W_k^*\|_2 \leqslant \max_{u,v\in[p]} \left\{ \mathbb{E}\left\| P^{(k,u,v)} \right\|_2 \right\} \leqslant \max_{u,v\in[p]} \left\{ \sum_{i,j\in[q]} \mathbb{E}\left\| P^{(k,u,v,i,j)} \right\|_2 \right\} = \max_{u,v\in[p]} \left\{ \sum_{i,j\in[q]} \mathbb{E}\left\| \mathcal{F}_{:,:,i,j}^{(k)} \right\|_2 \right\} \leqslant C_3 \frac{q^2}{p}. \tag{37}$$

By the Markov's inequality, we have

$$\mathbb{P}\left(\|W_k^*\|_2 \leqslant p^{-\beta_1} \right) \geqslant 1 - C_3 \frac{q^2}{p^{1-\beta_1}}. \tag{38}$$

We use induction to show that, for any $x \in \mathcal{C}_{p_0^2 d_0}$ and $1 \leqslant k < l$, we have

(I) with probability at least $\left(1 - C_3 \frac{q^2}{p^{1-\beta_1}}\right)^k$, we have $\|y_k^*(x)\|_2 \leqslant \left(Lp^{-\beta_1}\right)^k p_0\sqrt{d_0}$,

[11] Note that the dimension of W_1^* and W_l^* are not $p^2 d \times p^2 d$ and thus we cannot apply Lemma 7 directly. However, we can always embed them into a $p^2 d \times p^2 d$ matrix. For example, we can define $\widetilde{W}_l^* = [W_l^*, \mathbf{0}_{p^2 d\times p^2(d-d_l)}]$ and apply Lemma 7 on $\widetilde{W}_l^*$. We use the fact that $\|W_l^*\|_2 \leqslant \left\|\widetilde{W}_l^*\right\|_2$ to get the same result.

(II) with probability at least $1-(k-1)C_4\frac{q^2}{p}d^{-\frac{1}{4}\alpha+\beta_2}-\frac{k^2+k-2}{2}C_3\frac{q^2}{p^{1-\beta_1}}$, we have $\left\|\left(y_k(x)\middle|A\right)-y_k^*(x)\right\|_2 \leqslant \left(p^{-\beta_1}\left(p^{-\beta_1}+d^{-\beta_2}\right)^{k-1}-p^{-k\beta_1}\right)L^k p_0\sqrt{d_0}$ holds for some positive constant C_4 specified later[12].

The case of $k=1$ is as follows. With probability at least $1-C_3\frac{q^2}{p^{1-\beta_1}}$, we have $\left\|y_1^*(x)\right\|_2=\left\|\sigma\left(W_1^*x\right)\right\|_2\leqslant L\left\|W_1^*x\right\|_2\leqslant L\left\|W_1^*\right\|_2\left\|x\right\|_2\leqslant Lp^{-\beta_1}p_0\sqrt{d_0}$. Further, we have $y_1(x)=\sigma\left(W_1x\right)=\sigma\left(W_1^*x\right)=y_1^*(x)$, and thus $\left\|y_1(x)-y_1^*(x)\right\|_2=0$.

Suppose the statement holds for $1\leqslant k<l-1$, we consider the case of $k+1$. Note that the events $\left\{\left\|W_{k+1}^*\right\|_2\leqslant p^{-\beta_1}\right\}$ and $\left\{\left\|y_k^*(x)\right\|_2\leqslant\left(Lp^{-\beta_1}\right)^k p_0\sqrt{d_0}\right\}$ are independent. By (38) and the induction statement (I), with probability at least

$$\mathbb{P}\left(\left\|W_{k+1}^*\right\|_2\leqslant p^{-\beta_1}\right)\mathbb{P}\left(\left\|y_k^*(x)\right\|_2\leqslant\left(Lp^{-\beta_1}\right)^k p_0\sqrt{d_0}\right)\geqslant\left(1-C_3\frac{q^2}{p^{1-\beta_1}}\right)^{k+1},$$

we have

$$\begin{aligned}\left\|y_{k+1}^*(x)\right\|_2&=\left\|\sigma\left(W_{k+1}^*y_k^*(x)\right)\right\|_2\leqslant L\left\|W_{k+1}^*y_k^*(x)\right\|_2\leqslant L\left\|W_{k+1}^*\right\|_2\left\|y_k^*(x)\right\|_2\\&\leqslant Lp^{-\beta_1}\cdot\left(Lp^{-\beta_1}\right)^k p_0\sqrt{d_0}=\left(Lp^{-\beta_1}\right)^{k+1}p_0\sqrt{d_0},\end{aligned}$$

which shows (I) in the induction statement.

We use a similar approach as in the proof for Theorem 2 to show that (II) holds. Let us denote $\overline{K}_{:,:,i,j}^{(k+1)}:=\mathcal{M}^{(k+1)}\circ K_{:,:,i,j}^{(k+1)}, i,j\in[p]$, i.e., $\overline{K}_{s',t',:,:}^{(k+1)}=K_{s',t',:,:}^{(k+1)}$ if $(s',t')\in\mathcal{I}_{k+1}$ and $\overline{K}_{s',t',:,:}^{(k+1)}=\mathbf{0}_{p\times p}$ otherwise. Then $\overline{W}_{k+1}:=W_{k+1}^*-W_{k+1}$ can be represented by

$$\overline{W}_{k+1}=\begin{bmatrix}\overline{B}_{1,1}^{(k+1)}&\cdots&\overline{B}_{1,d}^{(k+1)}\\\vdots&\ddots&\vdots\\\overline{B}_{d',1}^{(k+1)}&\cdots&\overline{B}_{d',d}^{(k+1)}\end{bmatrix},\tag{39}$$

where each $\overline{B}_{s,t}^{(k+1)}$ is a doubly block circulant matrix such that

$$\overline{B}_{s,t}^{(k+1)}=\begin{bmatrix}\text{circ}\left(\overline{K}_{s,t,1,:}^{(k+1)}\right)&\text{circ}\left(\overline{K}_{s,t,2,:}^{(k+1)}\right)&\cdots&\text{circ}\left(\overline{K}_{s,t,p,:}^{(k+1)}\right)\\\text{circ}\left(\overline{K}_{s,t,p,:}^{(k+1)}\right)&\text{circ}\left(\overline{K}_{s,t,1,:}^{(k+1)}\right)&\cdots&\text{circ}\left(\overline{K}_{s,t,p-1,:}^{(k+1)}\right)\\\vdots&\vdots&\ddots&\vdots\\\text{circ}\left(\overline{K}_{s,t,2,:}^{(k+1)}\right)&\text{circ}\left(\overline{K}_{s,t,3,:}^{(k+1)}\right)&\cdots&\text{circ}\left(\overline{K}_{s,t,1,:}^{(k+1)}\right)\end{bmatrix}.\tag{40}$$

Again, let $\overline{P}^{(k+1,u,v)}\in\mathbb{R}^{d\times d}, u,v\in[p]$ be such that

$$\overline{P}_{s,t}^{(k+1,u,v)}=\left(S^T\overline{K}_{s,t,:,:}^{(k+1)}S\right)_{u,v}=\sum_{i,j\in[q]}\omega^{ui}\overline{K}_{s,t,i,j}^{(k+1)}\omega^{vj},\quad s,t\in[d],u,v\in[p],$$

and we denote $\overline{P}^{(k+1,u,v,i,j)}:=\omega^{ui+vj}\overline{K}_{:,:,i,j}^{(k+1)}, u,v\in[p],i,j\in[q]$. Then we have $\overline{P}^{(k+1,u,v)}=\sum_{i,j\in[q]}\overline{P}^{(k+1,u,v,i,j)}$ and

$$\left\|\overline{P}^{(k+1,u,v,i,j)}\right\|_2=\left\|\omega^{ui+vj}\overline{K}_{:,:,i,j}^{(k+1)}\right\|_2=\left\|\overline{K}_{:,:,i,j}^{(k+1)}\right\|,\quad u,v\in[p],i,j\in[q].$$

[12] Note that in induction statement (II), the probability (and the expectations in the following context) is taken over the randomness of weights but not the masks, the random variable $\left\|\left(y_k(x)\middle|A\right)-y_k^*(x)\right\|_2$ is equivalent to $\left\|y_k(x)-y_k^*(x)\right\|_2\left|A\right.$. Further, the statement can also be written as

$$\mathbb{P}\left(\left\{\left\|y_k(x)-y_k^*(x)\right\|_2\leqslant\left(p^{-\beta_1}\left(p^{-\beta_1}+d^{-\beta_2}\right)^{k-1}-p^{-k\beta_1}\right)L^kp_0\sqrt{d_0}\right\}\middle|A\right)\geqslant1-(k-1)C_4\frac{q^2}{p}d^{-\frac{1}{4}\alpha+\beta_2}-\frac{k^2+k-2}{2}C_3\frac{q^2}{p^{1-\beta_1}}.$$

By assumption (iii), every entry of $\overline{K}^{(k+1)}_{:,:,i,j}$ follows a distribution such that the second-order moment is upper-bounded by $\frac{C_1}{p^2 d}$ and the fourth-order moment is upper-bounded by $\frac{C_2}{p^4 d^2}$. Under event A, that the number of non-zero entries in $\overline{K}^{(k+1)}_{:,:,i,j}$ is at most $\lfloor d^{2-\alpha} \rfloor$ and the number of non-zero entries in each row/column of $\overline{K}^{(k+1)}_{:,:,i,j}$ is at most $3\lfloor d^{2-\alpha} \rfloor / d$, by Lemma 4 and a similar derivation to (16) – (19), we have

$$\mathbb{E}\left[\left\|\overline{K}^{(k+1)}_{:,:,i,j}\right\|_2 \middle| A\right] \leqslant \frac{C_4}{p} d^{-\frac{1}{4}\alpha},$$

where $C_4 = C\left(2\left(3C_1\right)^{\frac{1}{2}} + C_2^{\frac{1}{4}}\right)$ and C is the universal constant as defined in Lemma 4.

By Lemma 7, we have

$$\begin{aligned}\mathbb{E}\left[\left\|W^*_{k+1} - W_{k+1}\right\|_2 \middle| A\right] = \mathbb{E}\left[\left\|\overline{W}_{k+1}\right\|_2 \middle| A\right] &= \max_{u,v\in[p]}\left\{\mathbb{E}\left[\left\|\overline{P}^{(k+1,u,v)}\right\|_2 \middle| A\right]\right\} \\ &\leqslant \max_{u,v\in[p]}\left\{\sum_{i,j\in[q]} \mathbb{E}\left[\left\|\overline{P}^{(k+1,u,v,i,j)}\right\|_2 \middle| A\right]\right\} \\ &= \max_{u,v\in[p]}\left\{\sum_{i,j\in[q]} \mathbb{E}\left[\left\|\overline{K}^{(k+1)}_{:,:,i,j}\right\|_2 \middle| A\right]\right\} \\ &\leqslant C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha}.\end{aligned}$$

By the Markov's inequality, for all $t > 0$ we have

$$\mathbb{P}\left(\left\{\left\|W^*_{k+1} - W_{k+1}\right\|_2 \geqslant t\right\} \middle| A\right) \leqslant \frac{\mathbb{E}\left[\left\|W^*_{k+1} - W_{k+1}\right\|_2 \middle| A\right]}{t}.$$

Taking $t = d^{-\beta_2}$, we have

$$\mathbb{P}\left(\left\{\left\|W^*_{k+1} - W_{k+1}\right\|_2 \leqslant d^{-\beta_2}\right\} \middle| A\right) \geqslant 1 - C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha + \beta_2}.$$

Similar to (20) – (22) in the proof of Theorem 2, with probability at least

$$\begin{aligned}&\mathbb{P}\Bigg(\left\{\left\|\left(W_{k+1}\middle|A\right) - W^*_{k+1}\right\|_2 \leqslant d^{-\beta_2}\right\} \bigcap \left\{\left\|W^*_{k+1}\right\|_2 \leqslant p^{-\beta_1}\right\} \bigcap \left\{\left\|y^*_k(x)\right\|_2 \leqslant \left(Lp^{-\beta_1}\right)^k p_0\sqrt{d}\right\} \\ &\qquad \bigcap \left\{\left\|\left(y_k(x)\middle|A\right) - y^*_k(x)\right\|_2 \leqslant \left(p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^{k-1} - p^{-k\beta_1}\right) L^k p_0 \sqrt{d_0}\right\} \Bigg) \qquad (41) \\ &\geqslant \left(1 - C_4\frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2}\right) + \left(1 - C_3 \frac{q^2}{p^{1-\beta_1}}\right) + \left(1 - C_3 \frac{q^2}{p^{1-\beta_1}}\right)^k + \left(1 - (k-1)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{k^2+k-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}}\right) - 3 \\ &= 1 - kC_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{(k+1)^2 + (k+1) - 2}{2} C_3 \frac{q^2}{p^{1-\beta_1}},\end{aligned}$$

we have

$$\left\|\left(y_{k+1}(x)\middle|A\right) - y^*_{k+1}(x)\right\|_2 \leqslant L^{k+1} p_0 \sqrt{d}\left[p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^k - p^{-(k+1)\beta_1}\right],$$

which finishes the induction.

We have just shown that with probability at least $1 - (l-2)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{l^2-l-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}}$, we have

$$\left\|\left(y_{l-1}(x)\middle|A\right) - y^*_{l-1}(x)\right\|_2 \leqslant L^{l-1} p_0 \sqrt{d}\left[p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^{l-2} - p^{-(l-1)\beta_1}\right].$$

Note that the last layer of F is a fully-connected layer with dimension $p^2 d \times d_l$. By Lemma 4, the Markov's inequality, and a similar derivation to (36) – (37), there exists a positive constant C_5 such that $\mathbb{P}\left(\|W_l^*\| \leqslant p^{-\beta_1}\right) \geqslant 1 - \frac{C_5}{p^{1-\beta_1}}$. Therefore, with probability at least

$$\left(1 - \frac{C_5}{p^{1-\beta_1}}\right) \cdot \left(1 - (l-2)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{l^2-l-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}}\right)$$
$$\geqslant 1 - (l-2)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{l^2-l-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}} - \frac{C_5}{p^{1-\beta_1}},$$

we have that for every $x \in \mathcal{C}_{p_0^2 d_0}$

$$\left\|\left(f(x)|A\right) - F(x)\right\|_2 \leqslant p^{-\beta_1} L^{l-1} p_0 \sqrt{d} \left[p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^{l-2} - p^{-(l-1)\beta_1}\right].$$

With probability at least $\mathbb{P}(A) = \left(1 - d^{-\frac{1}{3}}\right)^{2(l-2)}$ over the randomness of masks, we have

$$\sup_{x \in \mathcal{C}_{p_0^2 d_0}} \left\|\left(f(x)|A\right) - F(x)\right\|_2 \leqslant p^{-\beta_1} L^{l-1} p_0 \sqrt{d} \left[p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^{l-2} - p^{-(l-1)\beta_1}\right]$$

with probability at least $1 - (l-2)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{l^2-l-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}} - \frac{C_5}{p^{1-\beta_1}}$. As a result, basic probability yields that

$$\sup_{x \in \mathcal{C}_{p_0^2 d_0}} \left\|f(x) - F(x)\right\|_2 \leqslant p^{-\beta_1} L^{l-1} p_0 \sqrt{d} \left[p^{-\beta_1}\left(p^{-\beta_1} + d^{-\beta_2}\right)^{l-2} - p^{-(l-1)\beta_1}\right]$$

holds with probability at least $\left(1 - d^{-\frac{1}{3}}\right)^{2(l-2)} \left(1 - (l-2)C_4 \frac{q^2}{p} d^{-\frac{1}{4}\alpha+\beta_2} - \frac{l^2-l-2}{2} C_3 \frac{q^2}{p^{1-\beta_1}} - \frac{C_5}{p^{1-\beta_1}}\right)$.

□

C. Extension of Magnitude-based Pruning

In this section, we discuss some extensions of Theorems 1 and 3 presented in the main paper. Note that we only provide ideas but not strict proofs in this section, as the results here are based on approximations and further efforts are required to give precise statements.

C.1. Magnitude-based Pruning of FCNs with Sub-Gaussian Distributions

Note that in Theorem 1, assumption (iii), we assume that the distribution of the weights in the layers of F are independently and identically following $\mathcal{U}\left[-\frac{K}{\sqrt{\max\{d_k, d_{k-1}\}}}, \frac{K}{\sqrt{\max\{d_k, d_{k-1}\}}}\right]$. The uniform distribution provides a closed-form order statistics and hence we can bound the gap between weight matrices and pruned weight matrices precisely. In fact, the uniform and exponential distributions are the only distributions that have a closed-form for order statistics in the literature. It is a natural question of what happens if the weights follow a more general distribution, e.g. a sub-Gaussian distribution.

Consider a target weight matrix $W^* \in \mathbb{R}^{d \times d}$ where we prune the smallest $\lfloor d^{2-\alpha} \rfloor$ entries in W^* based on magnitude. We further assume that the weights in W^* independently and identically follow a sub-Gaussian distribution $\mathsf{subG}(\sigma^2)$ with appropriate choice of σ^2 (e.g., $\sigma^2 = \frac{1}{d}$). Next we present the idea of applying the results of intermediate order statistics to show a similar result in the asymptotic sense.

Theorem 4 (Lemma 1 of Chibisov (1964)). *Let $X_1, X_2, \dots$ be a sequence of independent random variables with the same distribution function F. We denote $X_m^{(n)}$ as the m-th largest among $X_1, \dots, X_n$ and $G_{m,n}(x) = \mathbb{P}\left(X_m^{(n)} < x\right)$. If $n \to \infty, m \to \infty$, and $m/n \to 0$, then*

$$\sup_x \left|G_{m,n}(a_n x + b_n) - \Phi\left(u_n(x)\right)\right| \to 0,$$

where $u_n(x) = \frac{nF(a_n x + b_n)}{\sqrt{m}}$ and Φ is the cumulative distribution function of the standard Gaussian distribution.

Note that the non-zero entries of $\mathcal{W} := W^* - W$ are the smallest $\lfloor d^{2-\alpha} \rfloor$ order statistics of $\mathsf{subG}(\sigma^2)$ based on magnitude, where W is the pruned weight matrix. If we order the weights in W^* by their magnitude, i.e.

$$\left|W^*_{i_1,j_1}\right| \leqslant \left|W^*_{i_2,j_2}\right| \leqslant \cdots \leqslant \left|W^*_{i_{d^2},j_{d^2}}\right|,$$

then the non-zero entries in $\mathcal{W}$ are

$$W^*_{i_1,j_1}, W^*_{i_2,j_2}, \ldots, W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}.$$

Taking $m = \lfloor d^{2-\alpha} \rfloor, n = d^2, a_n = 1, b_n = 0$ in Theorem 4 and note that $G_{m,n}(x) = \mathbb{P}\left(\left|W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}\right| \leqslant x\right)$, we have

$$\sup_x \left| \mathbb{P}\left(\left|W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}\right| \leqslant x\right) - \Phi\left(\frac{d^2 F(x)}{\sqrt{\lfloor d^{2-\alpha} \rfloor}}\right) \right| \to 0, \quad x \to \infty.$$

Thus we can approximate the expectation $\mathbb{E}\left[\left|W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}\right|\right]$ by some positive constant β, by the properties of the cumulative density function of standard Gaussian and $\mathsf{subG}\left(\sigma^2\right)$. Similarly, we can get the estimations of $\mathbb{E}\left[\left|W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}\right|^2\right]$ and $\mathbb{E}\left[\left|W^*_{i_{\lfloor d^{2-\alpha} \rfloor}, j_{\lfloor d^{2-\alpha} \rfloor}}\right|^4\right]$. Then we can apply Lemma 4 (similar to (16) – (19)) to upper-bound the expectation $\mathbb{E}\left\|\mathcal{W}\right\|_2$. Recall that this is an asymptotic derivation, and we also need to bound the gap between the above second and fourth-order moments when $n = d^2$ is a large but fixed.

C.2. Magnitude-based Pruning of CNNs

We are given a convolutional tensor $\mathcal{F} \in \mathbb{R}^{d \times d \times p \times p}$. Let

$$W^* = \begin{bmatrix} B_{1,1} & \cdots & B_{1,d} \\ \vdots & \ddots & \vdots \\ B_{d,1} & \cdots & B_{d,d} \end{bmatrix} \in \mathbb{R}^{p^2 d \times p^2 d}$$

be the linear transformation corresponding to $\mathcal{F}$, and tensor K and B as defined in (7) – (9). The magnitude-based filter pruning of CNN is to order the L_1 norms $\left\|\text{vec}\left(B_{i,j}\right)\right\|_1, i, j \in [d]$ (or equivalently, $\left\|\text{vec}\left(K_{i,j,:,:}\right)\right\|_1$) and set the filters with the smallest L_1 norms to be zero. In other words, if we denote W to be the pruned weight matrix, then

$$W_* - W = \begin{bmatrix} \overline{B}_{1,1} & \cdots & \overline{B}_{1,d} \\ \vdots & \ddots & \vdots \\ \overline{B}_{d,1} & \cdots & \overline{B}_{d,d} \end{bmatrix}$$

is a block matrix of $\overline{B}_{i,j}$, where $\overline{B}_{i,j} = B_{i,j}$ if $\left\|\text{vec}\left(B_{i,j}\right)\right\|_1$ is among the smallest $\lfloor d^{2-\alpha} \rfloor$ norms, and $\overline{B}_{i,j} = \mathbf{0}_{p \times p}$ otherwise. Similar to Appendix C.1, we can upper-bound $\mathbb{E}\left\|W^* - W\right\|_2$ by $\mathbb{E}\left\|\overline{B}_0\right\|_2^2$ and $\mathbb{E}\left\|\overline{B}_0\right\|_2^4$, where $\overline{B}_0 \in \left\{\overline{B}_{i,j} : i, j \in [d]\right\}$ is the matrix corresponding ot the $\lfloor d^{2-\alpha} \rfloor$-th smallest value based on L_1 norms.

Note that the L_1 norms are the sum of many random samples drawn from a given distribution. By the Central Limit Theory, $\left\|\text{vec}\left(B_{i,j}\right)\right\|_1$ can be approximated by a normal distribution. Thus we can estimate $\left\|\text{vec}\left(\overline{B}_0\right)\right\|_1$ by a similar approach to the one in Appendix C.1. Theorem 6 of Sedghi et al. (2018) further provides a tool to upper-bound $\left\|\overline{B}_0\right\|_2$ by $\left\|\text{vec}\left(\overline{B}_0\right)\right\|_1$.

Note that we use two approximations in the above derivation. One is for the distribution of $\left\|\text{vec}\left(B_{i,j}\right)\right\|_1, i, j \in [d]$ and the other one comes from the asymptotic result as discussed in Appendix C.1. Caution should be taken while following these steps to attack the magnitude-based pruning problem of CNNs.

D. Numerical Study

In Sections D.1 and D.2, we show the histograms of some trained FCNs and CNNs. In Sections D.3 and D.4, we show the universal constants in Lemmas 3 and 4 as we use them frequently in the paper.

D.1. Distribution of Weights in Trained FCNs

We first describe the setting where we train a vanilla FCN. The `Covertype` dataset (Blackard & Dean, 1998) is to predict 7 different forest cover types from cartographic variables. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types). The dataset contains about 580,000 samples with 9 numerical and 44 categorical features. We normalize the numerical features by mean and variance of each feature. We build a 5-hidden-layer fully-connected neural network with ReLU activation functions to predict the label of each sample. There are 1,024 neurons in each hidden layer and thus the first weight matrix has dimension 54×1024, the internal 4 weight matrices have dimension 1024×1024, and the last weight matrix has dimension 1024×7. We minimize the cross-entropy loss using Adam with learning rate 0.001. The batch-size is selected to be 512 and we run 20 epochs of training. The trained network achieves approximately 80% predicting accuracy.

Figure 1 in the main paper shows the histogram of the entries in all weight matrices. We mainly focus on the second to fifth layers because we do not perform any pruning on the first and last layers. In these 4 layers, the weights are approximately distributed following a Gaussian distribution. We also report the means and variances of the entries in each internal layer in Table 1. As we can see from the results, for the internal weight matrices, the means are close to zero while the variances are approximately bounded by $\frac{3}{1024}$, which is also the initialization variance suggested by Glorot & Bengio (2010). We have also tested several other random initial weights and network architectures, and the results and conclusions are similar and not presented.

Table 1: Expectation and variance of the entries in all weight matrices

Layer	1	2	3	4	5	6
Mean	-0.0309	-0.0215	-0.0078	-0.0119	-0.0092	-0.0275
Variance	0.0155	0.0035	0.0022	0.0019	0.0016	0.0057

D.2. Distribution of Weights in VGG16

We plot the histogram of weights in different layers of VGG16 (Simonyan & Zisserman, 2014). The pre-trained model is imported from `PyTorch` package (Paszke et al., 2019) where the weights are trained on a variety of image datasets. Figure 2 shows the results for all layers of the pre-trained VGG16 (13 convulotional layers and 3 fully-connected layers). As we can see, the entries in the internal layers follow Gaussian distributions approximately.

D.3. Constants in Lemma 3

Lemma 3 gives an upper-bound of the random matrix $B \in \mathbb{R}^{n_1 \times n_2}$ whose entries are independently and identically following a uniform distribution $\mathcal{U}\left[-\frac{K}{\sqrt{n}}, \frac{K}{\sqrt{n}}\right]$, where $n = \max\{n_1, n_2\}$ and K is a positive constant. To better understand the values of constants c_0 and δ_0, we take various tuples of (n_1, n_2, K) and calculate the norm $\|B\|_2$. In the numerical experiments, we generate in total $N = 1000$ random matrices and report c_0 and δ_0 that satisfy $\mathbb{P}\left(\|B\|_2 \leqslant c_0\right) = 1 - 2e^{-4\delta_0 n} = q$ for $q = 95\%, 99\%, 99.9\%, 99.99\%$. We also report the mean and standard deviation of $\|B\|_2$ for reference. The results are given in Table 2. The table shows that, even if n_1 and n_2 are on the low end with respect to the actual use cases, we can still have a small c_0 that is close to 1 and a small δ_0 that is close to 0. Note that these two quantities are frequently used in Theorem 1 and we observe that the constant terms in the theorem are mild while the probability that the statement hold is positive.

D.4. Constant in Lemma 4

Lemma 4 shows that there exists a universal constant C such that, for any random matrix A whose entries are independent, we have

$$\mathbb{E}\|A\|_2 \leqslant C\left[\max_i\left(\sum_j \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}} + \max_j\left(\sum_i \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}} + \left(\sum_{i,j}\mathbb{E}A_{i,j}^4\right)^{\frac{1}{4}}\right]. \tag{42}$$

We use this lemma many times to bound the L_2 norm of various random matrices, e.g., in (19) and (36). In the following,

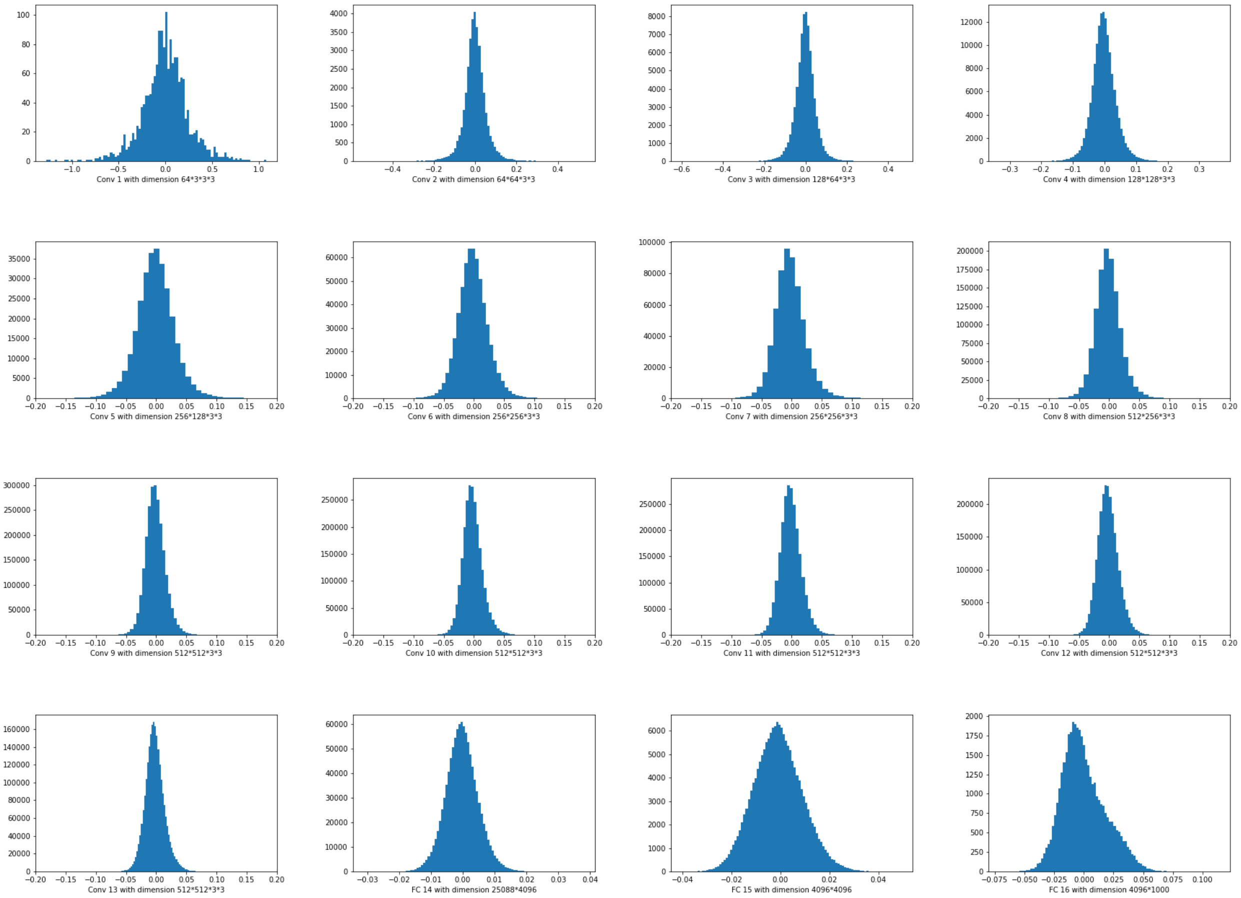

Figure 2: Histogram of entries of all weight matrices of a pre-trained VGG16

we consider the cases where the elements of $A \in \mathbb{R}^{d\times d}$ follows $U := \mathcal{U}\left[-\sqrt{\frac{3}{d}}, \sqrt{\frac{3}{d}}\right]$ and $\mathcal{N}\left(0, \frac{K}{d}\right)$ for some positive constant K, respectively. We also consider the case where we initialize the elements of A by samples of $\mathcal{N}\left(0, \frac{1}{d}\right)$, but we set $\lfloor d^{2-\alpha} \rfloor$ entries to be zero randomly (thus it aligns with the use case in (19)).

In the numerical experiments, we generate in total $N = 500$ random matrices A and calculate the quantities $\mathbb{E}\left\|A\right\|_2, \max_i \left(\sum_j \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}}, \max_j \left(\sum_i \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}}$ and $\left(\sum_{i,j} \mathbb{E}A_{i,j}^4\right)^{\frac{1}{4}}$. In Table 3, we report the minimum C such that (42) holds with the choices of d, distribution of $A_{i,j}$, and α (if necessary).

E. Discussion

In this section, we discuss some assumptions made to simply the presentations. We provide (possible) ways to avoid them but the detailed proofs are omitted.

E.1. Independency of Weights in the Target Network

The assumption of independent trained weights satisfied to a certain degree. Many existing works show that the trained weights are not "far away" from the initialization and thus certain levels of independency remains among the trained weights. For example, Bai & Lee (2020) show that the trained weights can be approximated by a Taylor expansion around the initialization and the coefficients of the polynomial are relatively small. This also aligns with the observation from the NTK literature (Jacot et al., 2018) that the trained weights are close to initialization. There are no well-accepted metrics to

Table 2: Numerical results for the constants in Lemma 3

n_1	n_2	K	$\mathbb{E}\left\|B\right\|_2$	std $(\left\|B\right\|_2)$	$q = 95\%$		$q = 99\%$		$q = 99.9\%$		$q = 99.99\%$	
					c_0	δ_0	c_0	δ_0	c_0	δ_0	c_0	δ_0
32	32	1	1.087	0.038	1.15	0.029	1.183	0.041	1.206	0.059	1.218	0.077
32	32	$\sqrt{3}$	1.882	0.066	1.996	0.029	2.044	0.041	2.069	0.059	2.131	0.077
32	64	1	0.941	0.027	0.988	0.014	1.015	0.021	1.039	0.03	1.042	0.039
32	64	$\sqrt{3}$	1.631	0.046	1.707	0.014	1.743	0.021	1.786	0.03	1.797	0.039
32	128	1	0.836	0.018	0.867	0.007	0.878	0.01	0.895	0.015	0.902	0.019
32	128	$\sqrt{3}$	1.449	0.032	1.503	0.007	1.528	0.01	1.577	0.015	1.579	0.019
32	256	1	0.762	0.013	0.784	0.004	0.794	0.005	0.806	0.007	0.811	0.01
32	256	$\sqrt{3}$	1.319	0.022	1.357	0.004	1.371	0.005	1.39	0.007	1.393	0.01
32	512	1	0.708	0.009	0.723	0.002	0.731	0.003	0.74	0.004	0.747	0.005
32	512	$\sqrt{3}$	1.226	0.016	1.253	0.002	1.267	0.003	1.278	0.004	1.283	0.005
64	64	1	1.114	0.026	1.158	0.014	1.183	0.021	1.205	0.03	1.209	0.039
64	64	$\sqrt{3}$	1.932	0.045	2.009	0.014	2.045	0.021	2.07	0.03	2.086	0.039
64	128	1	0.959	0.018	0.992	0.007	1.005	0.01	1.04	0.015	1.054	0.019
64	128	$\sqrt{3}$	1.66	0.031	1.711	0.007	1.743	0.01	1.782	0.015	1.785	0.019
64	256	1	0.848	0.012	0.868	0.004	0.88	0.005	0.887	0.007	0.888	0.01
64	256	$\sqrt{3}$	1.47	0.021	1.508	0.004	1.523	0.005	1.53	0.007	1.554	0.01
64	512	1	0.77	0.008	0.785	0.002	0.792	0.003	0.796	0.004	0.801	0.005
64	512	$\sqrt{3}$	1.333	0.015	1.359	0.002	1.371	0.003	1.388	0.004	1.392	0.005
128	128	1	1.131	0.017	1.159	0.007	1.173	0.01	1.199	0.015	1.205	0.019
128	128	$\sqrt{3}$	1.956	0.029	2.008	0.007	2.024	0.01	2.044	0.015	2.045	0.019
128	256	1	0.969	0.012	0.99	0.004	0.999	0.005	1.012	0.007	1.013	0.01
128	256	$\sqrt{3}$	1.679	0.019	1.712	0.004	1.728	0.005	1.743	0.007	1.746	0.01
128	512	1	0.856	0.008	0.87	0.002	0.875	0.003	0.881	0.004	0.885	0.005
128	512	$\sqrt{3}$	1.482	0.014	1.507	0.002	1.52	0.003	1.527	0.004	1.528	0.005
256	256	1	1.14	0.011	1.16	0.004	1.17	0.005	1.18	0.007	1.181	0.01
256	256	$\sqrt{3}$	1.976	0.021	2.01	0.004	2.027	0.005	2.036	0.007	2.036	0.01
256	512	1	0.976	0.008	0.989	0.002	0.995	0.003	1.002	0.004	1.014	0.005
256	512	$\sqrt{3}$	1.691	0.013	1.714	0.002	1.727	0.003	1.735	0.004	1.735	0.005
512	512	1	1.146	0.007	1.159	0.002	1.163	0.003	1.172	0.004	1.174	0.005
512	512	$\sqrt{3}$	1.985	0.012	2.006	0.002	2.015	0.003	2.033	0.004	2.04	0.005

measure how close are the weights to independency, and thus we assume them to be independent.

There are other ways to relax independency. For random pruning, independency is assumed so that we can apply the Latala's inequality (Lemma 4). There also exist other versions of spectral norm bounds for sub-Gaussian random matrix with non-i.i.d. entries (Chapter 5 of Pastur & Shcherbina (2011)) and for a matrix with independent rows and columns (Vershynin, 2012). For magnitude-based pruning, the assumption is used to derive the explicit form of expectation of order statistics. By assuming an equal correlation between weights, we can also give the explicit forms (Chapter 5 of David & Nagaraja (2004)). The general form of order statistics for dependent uniform samples can be achieved approximately in the same way.

E.2. With-replacement and Without-replacement Sampling for Random Pruning

Under the random pruning scheme, we select N entries uniformly at random from a $d \times d$ weight matrix and set them to zero. The proposed approach in the beginning of Section 4 corresponds to "with-replacement" sampling since an entry might be selected multiple times. Another "without-replacement" sampling approach refers to selecting N non-overlapping entries from the weight matrix. Note that with a positive probability of $\frac{\binom{d^2}{N}}{d^{2N}}$, the entries selected by the "with-replacement" approach have no repeated elements and the two approaches align. In this sense, we can derive the results of the "without-replacement" approach from the stated results in this work by simply multiplying the corresponding probability that all selected entries are

Table 3: Numerical results for the constant in Lemma 4

d	Distribution	α	$\max_i \left(\sum_j \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}}$	$\max_j \left(\sum_i \mathbb{E}A_{i,j}^2\right)^{\frac{1}{2}}$	$\left(\sum_{i,j} \mathbb{E}A_{i,j}^4\right)^{\frac{1}{4}}$	$\mathbb{E}\lVert A\rVert_2$	C
32	U	N/A	1.006	1.006	1.159	1.888	0.596
64	U	N/A	1.006	1.005	1.159	1.934	0.61
128	U	N/A	1.005	1.005	1.159	1.958	0.618
256	U	N/A	1.004	1.003	1.158	1.976	0.624
512	U	N/A	1.003	1.002	1.158	1.985	0.627
32	$\mathcal{N}\left(0, \frac{1}{d}\right)$	N/A	1.011	1.014	1.314	1.905	0.571
64	$\mathcal{N}\left(0, \frac{1}{d}\right)$	N/A	1.008	1.008	1.315	1.947	0.585
128	$\mathcal{N}\left(0, \frac{1}{d}\right)$	N/A	1.005	1.007	1.316	1.965	0.59
256	$\mathcal{N}\left(0, \frac{1}{d}\right)$	N/A	1.005	1.006	1.316	1.979	0.595
512	$\mathcal{N}\left(0, \frac{1}{d}\right)$	N/A	1.004	1.004	1.316	1.988	0.598
32	$\mathcal{N}\left(0, \frac{3}{d}\right)$	N/A	1.751	1.745	2.279	3.295	0.571
64	$\mathcal{N}\left(0, \frac{3}{d}\right)$	N/A	1.755	1.744	2.28	3.361	0.582
128	$\mathcal{N}\left(0, \frac{3}{d}\right)$	N/A	1.742	1.743	2.28	3.405	0.591
256	$\mathcal{N}\left(0, \frac{3}{d}\right)$	N/A	1.743	1.742	2.28	3.428	0.595
512	$\mathcal{N}\left(0, \frac{3}{d}\right)$	N/A	1.74	1.739	2.279	3.441	0.598
32	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.01	0.626	0.63	1.033	1.237	0.54
64	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.01	0.632	0.629	1.035	1.239	0.54
128	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.01	0.63	0.63	1.037	1.242	0.541
256	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.01	0.63	0.63	1.039	1.246	0.542
32	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.1	0.714	0.713	1.103	1.379	0.545
64	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.1	0.729	0.729	1.117	1.426	0.554
128	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.1	0.744	0.744	1.129	1.459	0.558
256	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.1	0.758	0.756	1.14	1.491	0.562
32	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.5	0.928	0.925	1.258	1.759	0.565
64	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.5	0.95	0.948	1.275	1.831	0.577
128	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.5	0.964	0.964	1.288	1.883	0.586
256	$\mathcal{N}\left(0, \frac{1}{d}\right)$	0.5	0.975	0.974	1.296	1.92	0.592

not repeated.

E.3. Global and Layer-wise Magnitude-based Pruning

In this paper, the magnitude-based pruning is defined layer-wise as we order the weights in each layer based on magnitude separately and prune the smallest ones. There is also another "global" version where the weights of the entire network are sorted and the weights with the smallest magnitudes are pruned. Next we show the connection between these two settings and how to extend the proofs to the global setting.

Suppose that we want to prune a total of N weights in a l-layer network. If we treat the small weights as balls and layers as bins, then by Lemma 6, the maximum load in each bin is bounded by $O(N/l)$ with high probability. In other words, we expect to see that the appearances of pruned weights in all layers are approximately uniform (the numbers can differ by a constant but not orders of magnitude) with high probability. This is also the reason why we rarely see that the small weights appear in the same layer of a trained network in practice. Under this high-probability event, we get back to the layer-wise magnitude-based pruning setting excepts that the number of weights to be pruned in each layer may vary by a constant. In this sense, the original proofs can be easily revised to fit the global magnitude-based setting.