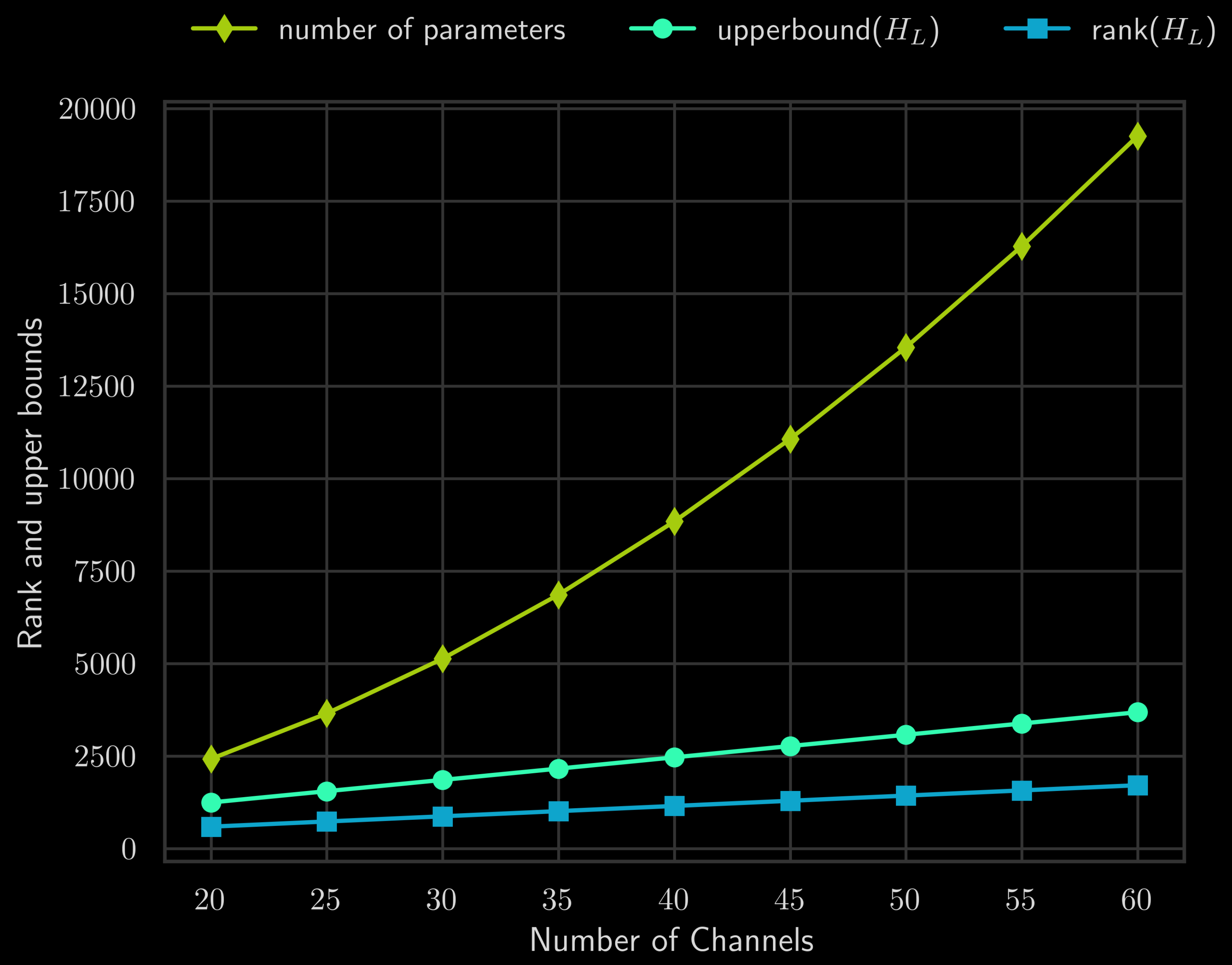
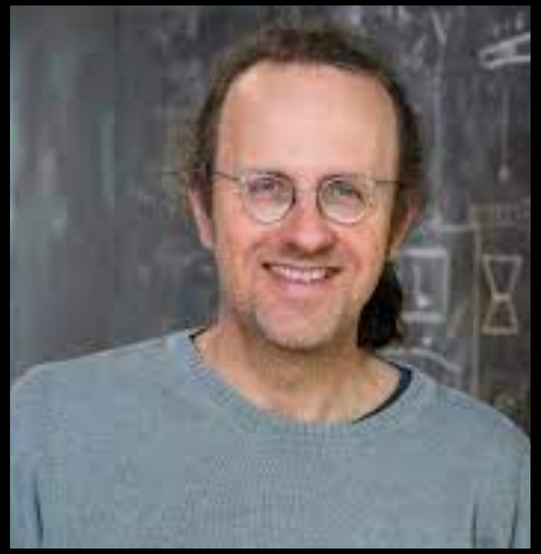
The Hessian Perspective into the Nature of Convolutional Neural Networks

ICML 2023







Sidak Pal Singh · *Thomas Hofmann* · *Bernhard Schölkopf*

ETH zürich

Aim of the Work

To provide a perspective into the nature of CNNs, i.e., how their architectural characteristics of CNNs manifest themselves in terms of the properties of its loss landscape as given by the Hessian rank

Hessian Rank, and a question

Captures pairwise interactions of parameters (θ_i, θ_j)

via the second-derivatives of the loss $\mathscr{L}$

$$\mathbf{H}_{ij} = \frac{\partial^2 \mathscr{L}}{\partial \theta_i \, \partial \theta_j}$$

Hessian Range $\quad \mathrm{range}(\mathbf{H}) = \{ y = \mathbf{H}\,\theta \;:\; \theta \in \mathbb{R}^p \}$

Hessian Rank $\quad \mathrm{rank}(\mathbf{H}) = \dim(\mathrm{range}(\mathbf{H}))$

Hessian Rank, and a question

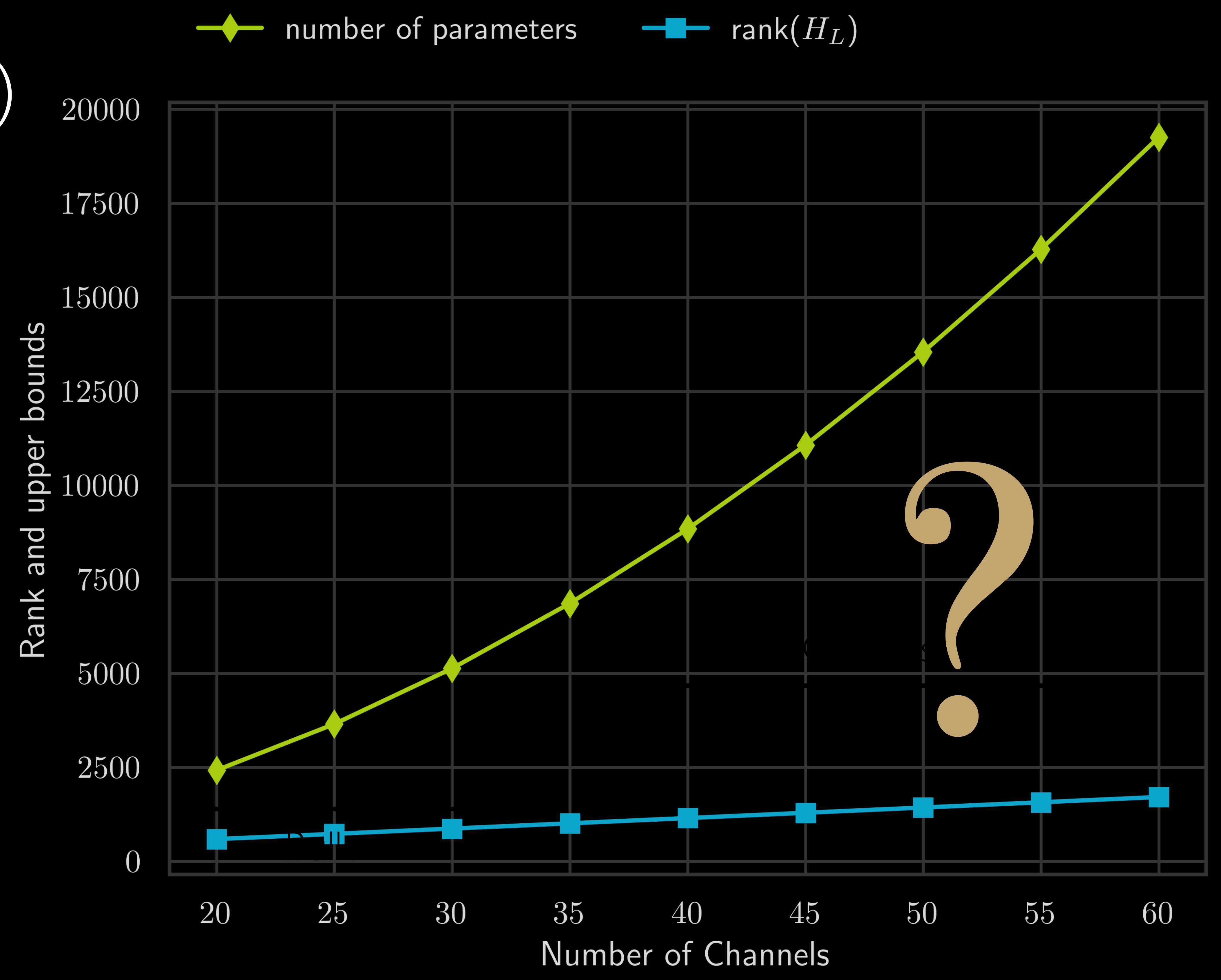
Captures pairwise interactions of parameters (θ_i, θ_j)

via the second-derivatives of the loss $\mathscr{L}$

$$\mathbf{H}_{ij} = \frac{\partial^2 \mathscr{L}}{\partial \theta_i \, \partial \theta_j}$$

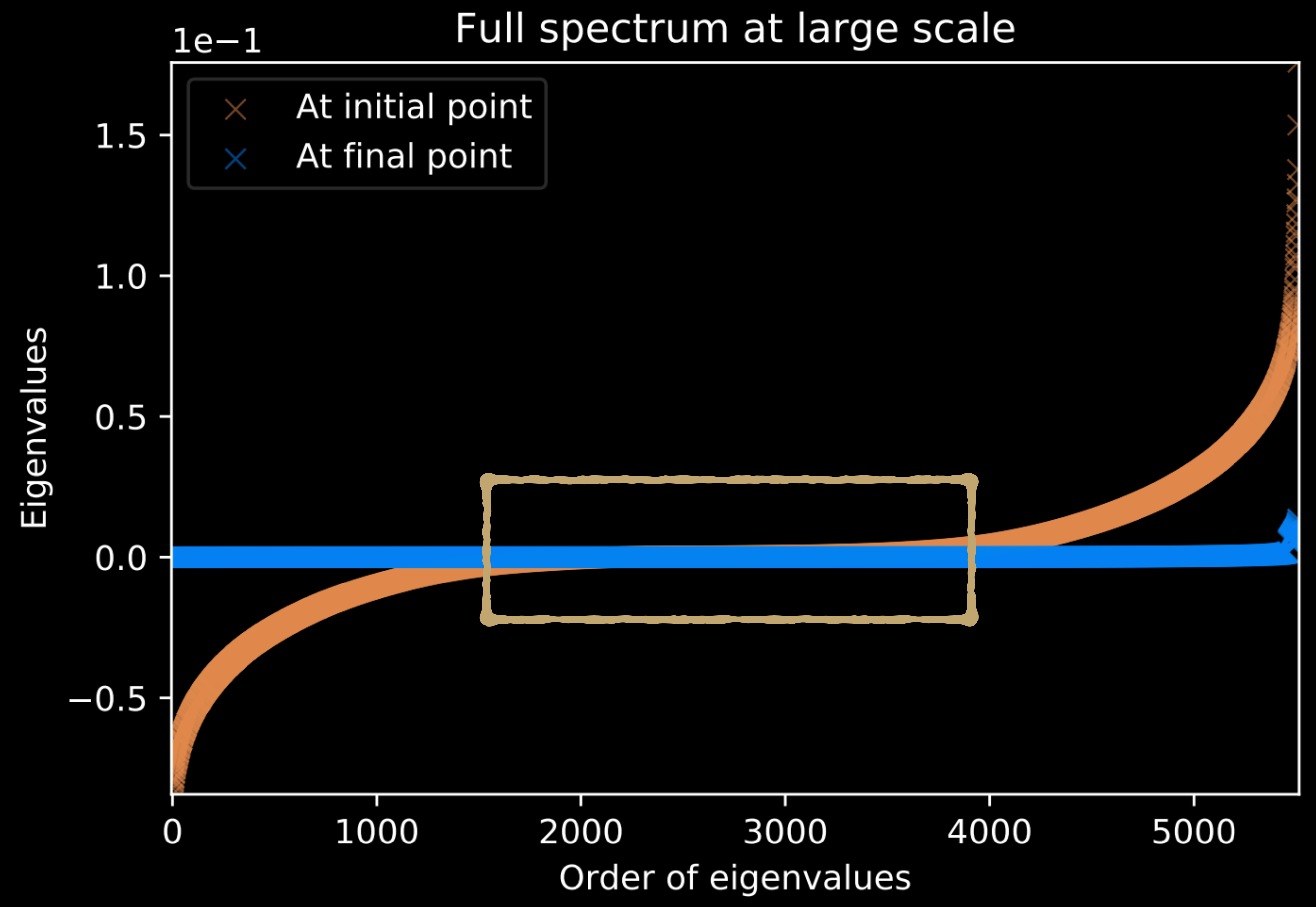
Hessian Range $\quad \mathrm{range}(\mathbf{H}) = \{y = \mathbf{H}\,\theta \;:\; \theta \in \mathbb{R}^p\}$

Hessian Rank $\quad \mathrm{rank}(\mathbf{H}) = \dim(\mathrm{range}(\mathbf{H}))$



Related Work

Sagun et. al., 2017



Significant extent of degeneracy in the Hessian

Singh et. al., 2021

$$\operatorname{rank}\left(\mathbf{H}_o\right) = q(d + K - q)$$

$$\operatorname{rank}\left(\mathbf{H}_f\right) \leq 2q \sum_{i=1}^{L} m_i + 2qs - (L+1)q^2$$

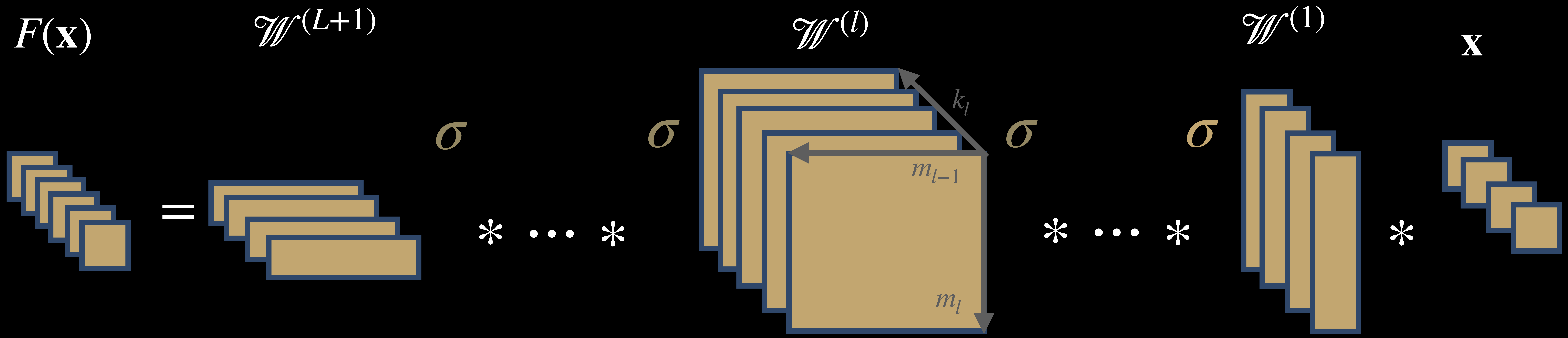
$$\operatorname{rank}\left(\mathbf{H}_{\mathscr{L}}\right) = 2q \sum_{i=1}^{L} m_i - (L+1)q^2 + q(r + K)$$

Theoretical characterisation for *Linear* Fully-Connected networks (FCNs)

Setup and Formalism

Given input $\mathbf{x} \in \mathbb{R}^{d_0}$, a deep CNN F with L hidden layers of weight tensors $\mathscr{W}^l \in \mathbb{R}^{m_l \times m_{l-1} \times k_l}$,

nonlinearity $\sigma(\,\cdot\,)$, stride 1, zero padding, with input channels $m_0 = 1$, output channels $= K$



Shorthands: $\mathbf{T}^{(k:l)} := \mathbf{T}^{(k)} \cdots \mathbf{T}^{(l)} \quad \forall k > l$ and $\mathbf{T}^{(k:l)} := \mathbf{T}^{(k)^\top} \cdots \mathbf{T}^{(l)^\top} \quad \forall k < l$

Also, denote the set of all parameters by $\theta := \{\mathscr{W}^1, \cdots, \mathscr{W}^{L+1}\}$

Toeplitz framework: A gentle start

Represent convolution as matrix product with a Toeplitz matrix

x_1	$\cdots$	x_d

$*$

w_1	$\cdots$	w_k

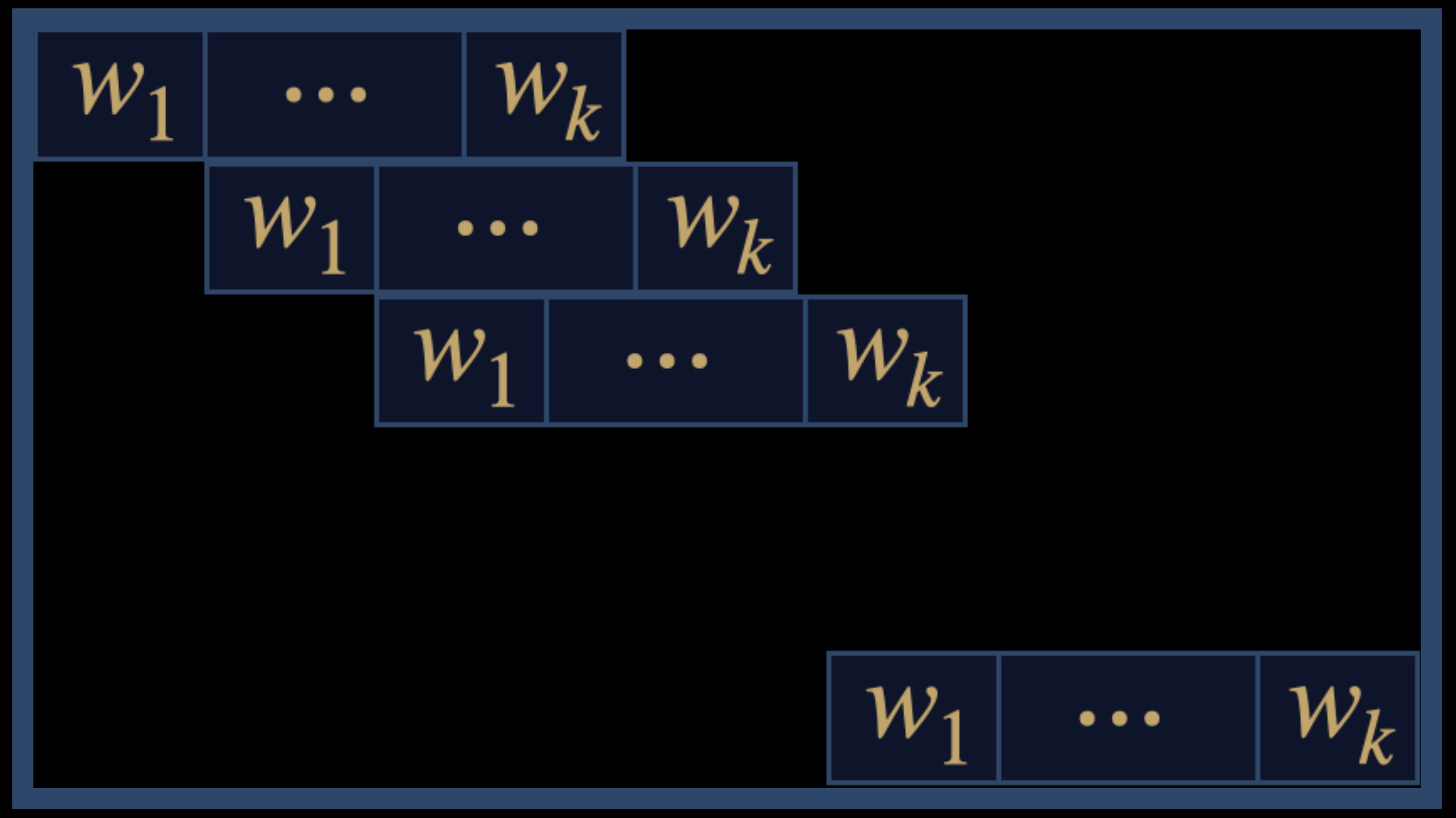
w_1	w_1	w_1		w_1	$\cdots$	w_k

$$\mathbf{conv}(\mathbf{x}, \mathbf{w}) = \mathbf{T}^{\mathbf{w}} \cdot \mathbf{x} = \begin{pmatrix} \quad \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{pmatrix}$$

Toeplitz Framework: full-fledged CNNs

$$\mathrm{F}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathscr{W}^{(L+1)} * \sigma\left(\mathscr{W}^{(L)} * \sigma\left(\cdots * \sigma\left(\mathscr{W}^{(1)} * \mathbf{x}\right)\right)\right)$$

$$\mathbf{T}^{(l)} := \begin{pmatrix} \mathbf{T}^{\mathscr{W}^{(l)}_{(1,1)}} \cdot & \cdots & \mathbf{T}^{\mathscr{W}^{(l)}_{(1,m_{l-1})}} \cdot \\ \vdots & & \vdots \\ \mathbf{T}^{\mathscr{W}^{(l)}_{(m_l,1)}} \cdot & \cdots & \mathbf{T}^{\mathscr{W}^{(l)}_{(m_l,m_{l-1})}} \cdot \end{pmatrix}$$



Each of the $\mathbf{T}^{\mathscr{W}^{l}_{(i,j)}}$ is a Toeplitz matrix as discussed before

$$\mathrm{F}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)}\Lambda^{(L)}_{\mathbf{x}}\mathbf{T}^{(L)} \cdots \Lambda^{(1)}_{\mathbf{x}}\mathbf{T}^{(1)}\mathbf{x}$$

$\Lambda^{(l)}_{\mathbf{x}}$ contains the activations at layer l

Key theoretical results

Theorem 1: Rank of outer-product Hessian $\mathbf{H}_o = \mathbf{E}_p \left[\nabla_{\boldsymbol{\theta}} F(\mathbf{x}) \, \nabla_{\boldsymbol{\theta}} F(\mathbf{x})^\top \right]$

The rank of $\mathbf{H}_o$ for a deep linear CNN, with kernel sizes k_l and number of filters m_l is:

$$\operatorname{rank}(\mathbf{H}_o) \leq \min \left(p, d_0 \operatorname{rank}\left(\mathbf{T}^{(2:L+1)}\right) + K \operatorname{rank}\left(\mathbf{T}^{(L:1)}\right) - \operatorname{rank}\left(\mathbf{T}^{(2:L+1)}\right) \operatorname{rank}\left(\mathbf{T}^{(L:1)}\right) \right)$$

$$= \min \left(p, q_o \left(d_0 + K - q_o \right) \right).$$

where $q_o := \min(d_0, m_1 d_1, \cdots, m_L d_L, K)$ denotes the flattened bottleneck dimension.

- If there is no bottleneck within, $\operatorname{rank}\left(\mathbf{H}_o\right) = K d_0$

Key theoretical results

Theorem 2: Rank of functional Hessian $\mathbf{H}_f = \mathbf{E}_p\Big[\sum_{c=1} [\partial \ell_{\mathbf{x},\mathbf{y}}]_c \nabla^2_{\boldsymbol{\theta}} F_c(\mathbf{x})\Big]$

The rank of the l-th column block of $\mathbf{H}_f$ for deep linear CNN, with kernel sizes k_l and number of filters m_l is: $\operatorname{rank}(\mathbf{H}_f^{\bullet l}) \leq \min(q_f\, m_{l-1}\, d_{l-1} + q_f\, m_l\, d_l - q_f^2\ ,\ m_l\, m_{l-1}\, k_l)$,

for $l \in [2, \cdots, L]$ and where $q_f := \min(q_o, s)$ and $s := \operatorname{rank}(\boldsymbol{\Omega}) = \operatorname{rank}\Big(\mathbf{E}\big[\boldsymbol{\delta}_{\mathbf{x},\mathbf{y}}\ \mathbf{x}^\top\big]\Big)$.

Hence we have that, $\operatorname{rank}(\mathbf{H}_f) \leq \sum_{l=1}^{L+1} \operatorname{rank}(\mathbf{H}_f^{\bullet l})$ (eq. 1)

- Block-column independence: Like in the case of FCNs, simply adding the ranks of the block-columns of the respective layers, gives the rank of the entire $\mathbf{H}_f$ without introducing any looseness in the bound (so, the inequality in eq. 1 suffices)

Key theoretical results

Rank of the loss Hessian can be bounded as

$$\operatorname{rank}\left(\mathbf{H}_{\mathscr{L}}\right) = \operatorname{rank}\left(\mathbf{H}_{o} + \mathbf{H}_{f}\right) \leq \operatorname{rank}\left(\mathbf{H}_{o}\right) + \operatorname{rank}\left(\mathbf{H}_{f}\right)$$

Rank of the loss Hessian grows as

$$\mathcal{O}(m \cdot L \cdot d_0)$$

Number of parameters grow as

$$\mathcal{O}(m^2 \cdot L \cdot d_0)$$

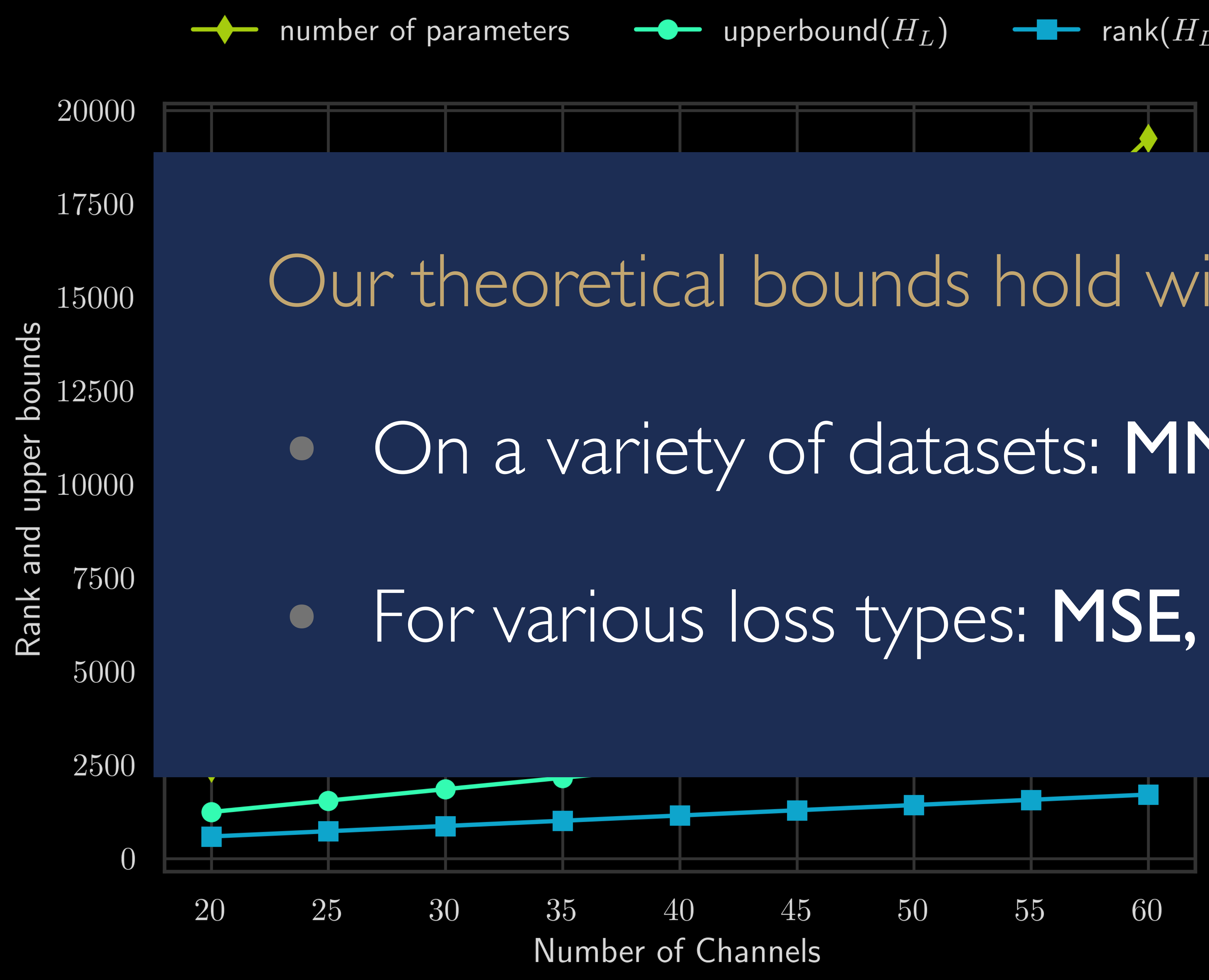
For typical networks, $m >> L$ and $m >> d_0$

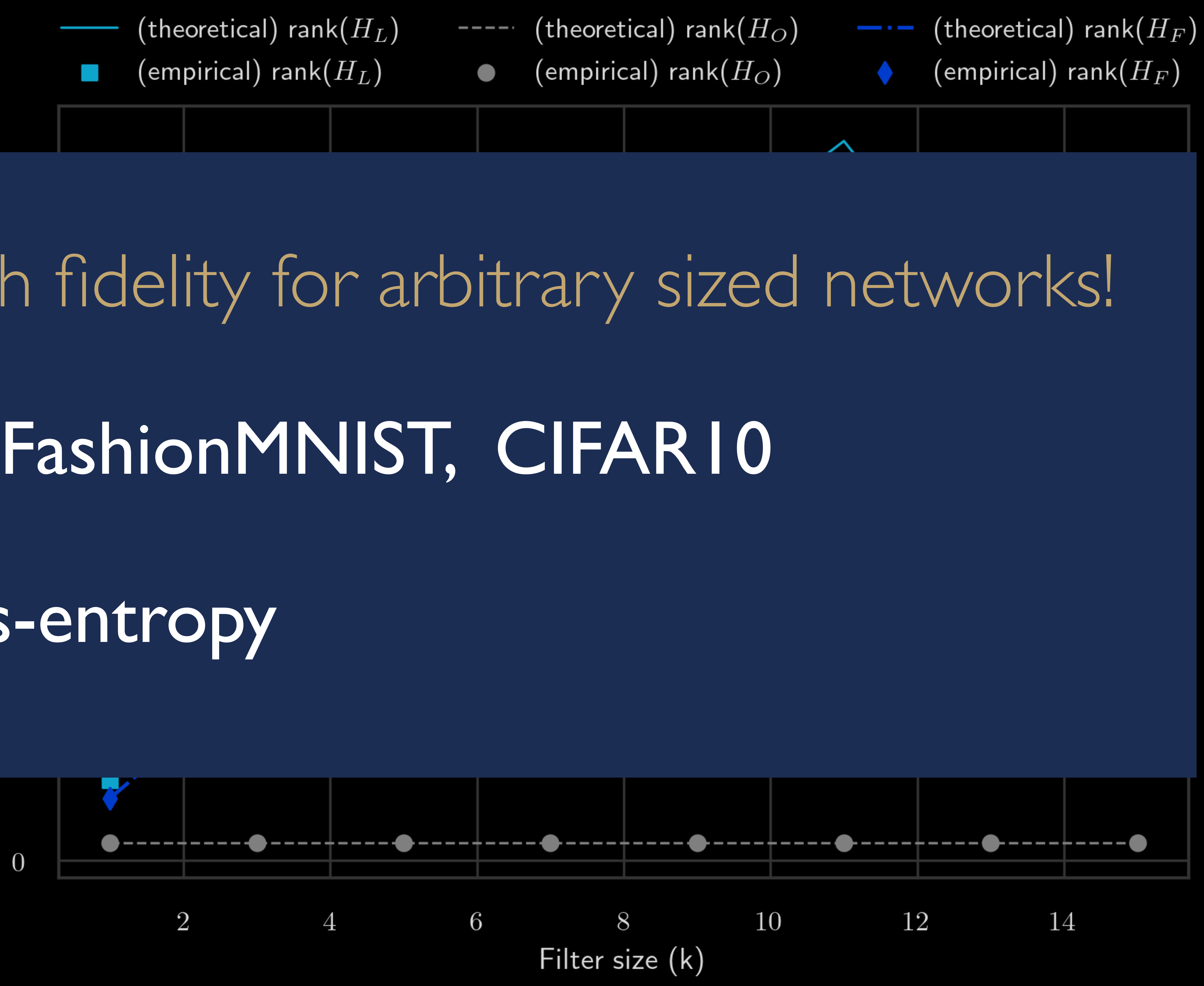
Hence, rank will show a square root behaviour relative to the number of parameters

Thus generalizing the finding of Singh et. al. (2021) to the case of CNNs

Empirical Results

$\min(\text{kernel-size}_i, \text{spatial-dim}_i)$





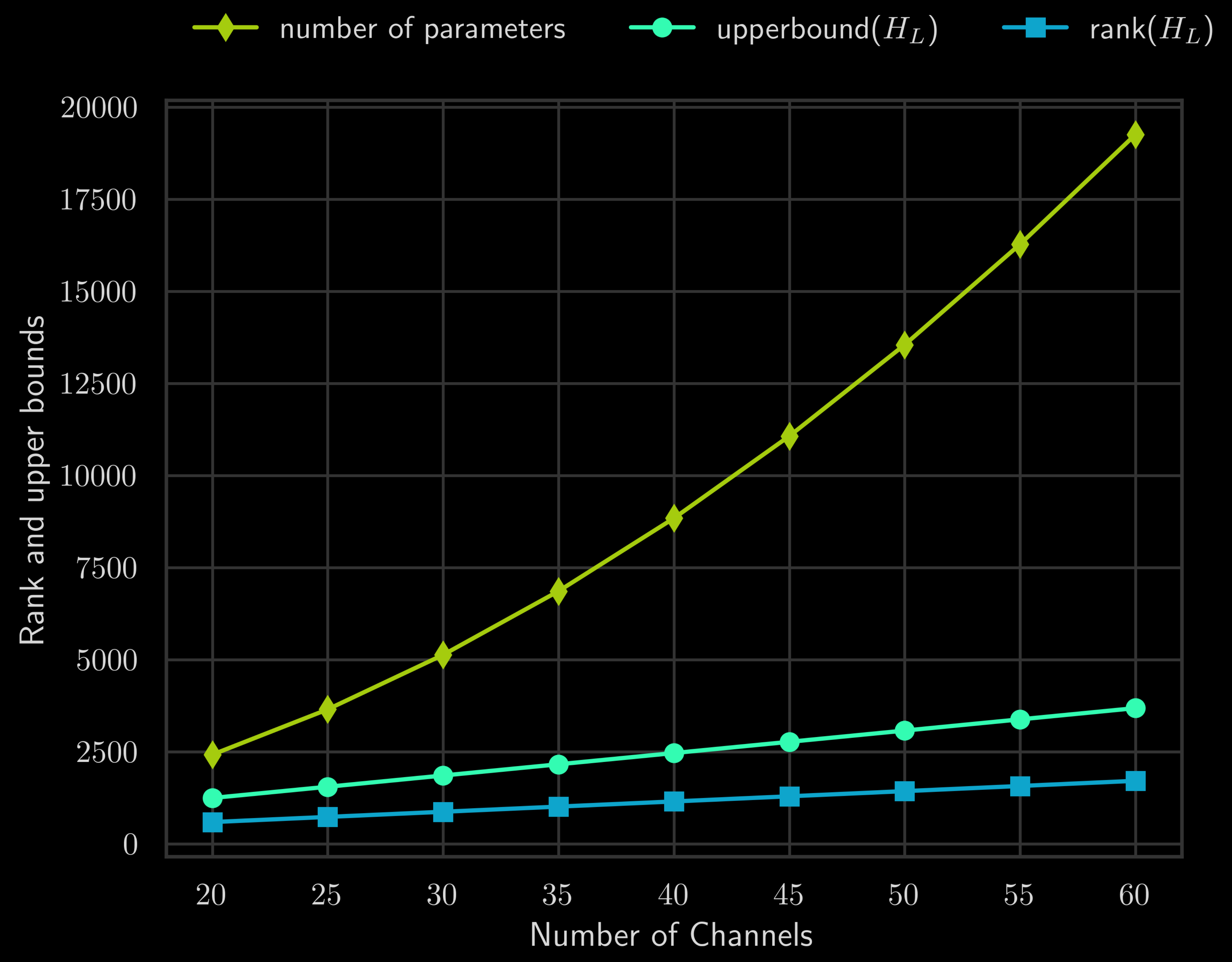
Our theoretical bounds hold with high fidelity for arbitrary sized networks!

- On a variety of datasets: **MNIST, FashionMNIST, CIFAR10**
- For various loss types: **MSE, cross-entropy**

Rank bounds for increasing # of filters m

Rank bounds for increasing filter size k

Summary



- Employ an equivalent representation of CNNs as composition of Toeplitz maps
- Natural change: $m_i \to m_i d_i$ where

$$d_i = \frac{d_{i-1} - \text{kernel-size}_i + 2\,\text{padding}_i}{\text{stride}_i} + 1$$

- The square-root trend of rank persists

This sheds a novel perspective on the nature of CNNs and highlights the degree of redundancy inherent in over-parameterized networks.