GENERALIZATION BOUNDS FOR PARAMETERIZED GROUP EQUIVARIANT CONVOLUTIONAL NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant convolutional neural networks have shown strong empirical performance by incorporating symmetry into their architectures, yet their generalization behavior remains theoretically underexplored, particularly for parameterized group equivariant CNNs, where convolutional filters are represented as linear combinations of basis functions. In this paper, we derive explicit generalization error bounds that characterize the effects of network depth, channel width, kernel size, and group size. Our bounds further capture the effects of filter parameterization, including the number of basis functions, coefficient constraints, and convolution-operator spectral norms, providing quantitative insights into their contributions to model complexity and generalization error. In particular, our analysis shows that different basis parameterizations can lead to different generalization guarantees, even for the same network architecture, and that reducing the effective basis dimension can improve the resulting generalization bound. These results provide a theoretical basis for understanding how filter parameterization affects the generalization of parameterized equivariant CNNs.

1 INTRODUCTION

In recent years, deep neural networks have achieved remarkable success in artificial intelligence and data science applications (LeCun et al., 2015; He et al., 2021b; Scarabosio, 2022; Chen et al., 2024). Understanding their generalization ability remains a fundamental problem in statistical learning theory (Vapnik, 1998; Xiao et al., 2022; Bartlett et al., 2017; Neyshabur et al., 2018; Koltchinskii & Panchenko, 2002). Vapnik (1998) established generalization bounds based on the VC dimension, while later studies developed tighter guarantees through complexity measures, including spectral normalized margin bounds (Bartlett et al., 2017) and PAC-Bayes theory (Neyshabur et al., 2018).

Beyond statistical properties, real-world data often exhibit intrinsic symmetries. Convolutional neural networks (CNNs) achieve strong performance in vision and image processing tasks due to translation equivariance, which enables efficient feature reuse and improves learning efficiency (Szegedy et al., 2015). However, many natural and scientific datasets contain richer symmetries, such as rotations, reflections, and scalings. Incorporating these structures through equivariance or invariance can improve representation efficiency and generalization. Consequently, equivariant CNNs have been successfully applied in vision tasks, medical imaging, fluid dynamics, molecular modeling, particle physics, robotics, and reinforcement learning (Zhao et al., 2026; Tan et al., 2026; Fu et al., 2026; Wang et al., 2021; Atz et al., 2021; Bogatskiy et al., 2020; Zhu et al., 2022; Huang et al., 2022; Van der Pol et al., 2020; Wang et al., 2022).

Early equivariant CNNs mainly focused on discrete symmetries. Cohen & Welling (2016) introduced p4/p4m group equivariant CNNs on square grids, while Hoogeboom et al. (2018) extended this idea to hexagonal grids with p6/p6m symmetries. However, lattice-based approaches are limited to specific discrete transformations. To overcome this limitation, filter parameterization methods were developed to support more general symmetries on standard grids. Weiler et al. (2018) proposed steerable filters based on harmonic functions for continuous rotations, followed by extensions to broader symmetry groups (Cohen et al., 2018; Cohen & Welling, 2022), PDO-based equivariant convolutions (Shen et al., 2020), and Fourier-based parameterizations for improved representation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

capability and reduced aliasing (Xie et al., 2022). These methods generally represent learnable kernels as combinations of basis functions with trainable coefficients (Freeman et al., 1991).

Despite the empirical success of equivariant networks, their theoretical understanding remains incomplete. Existing studies have shown that equivariance can improve generalization performance (Sokolic et al., 2017; Lyle et al., 2020; Elesedy, 2021; 2022; Behboodi et al., 2022; Li et al., 2026; Long & Sedghi, 2020). However, current analyses either focus on function-class properties without explicitly characterizing network architectures and parameters, or mainly consider equivariant multilayer perceptrons rather than practical equivariant CNNs. Moreover, different equivariant CNNs employ substantially different parameterizations, such as the circular harmonic bases used by Weiler et al. (2018) and the Fourier bases adopted by (Xie et al., 2022), which may lead to different model complexities and generalization behaviors. Yet, explicit generalization guarantees that characterize and compare the effects of both equivariant convolutional architectures and their parameterizations are still lacking.

To address this gap, we develop explicit generalization bounds for parameterized group equivariant convolutional neural networks (PG-CNNs). Our analysis provides a unified framework for quantifying the effects of equivariant architectures and basis-function parameterizations on generalization, with explicit dependence on key architectural factors, including network depth, channel width, kernel size, and group size. This framework further enables a theoretical comparison of different basis parameterizations used in equivariant CNNs. The main contributions are summarized as follows:

- **Unified generalization theory for PG-CNNs.** To the best of our knowledge, we are the first to derive explicit generalization bounds for PG-CNNs. The main technical challenge is to characterize the covering number of discretized convolutional kernels generated by linear combinations of basis functions. We resolve this by deriving covering-number bounds that explicitly account for the basis representation, coefficient constraints, and spectral norms of the induced convolution operators. This leads to two complementary generalization bounds. For general PG-CNNs, the architectural complexity is governed by $\sum_{l=1}^{L} N_l G_{l-1}$, whereas for practically adopted PG-CNNs, where only $C_l/|G|$ kernels are independently parameterized, it is characterized by $\sum_{l=1}^{L} N_l C_l C_{l-1}/|G|$. In both cases, the effects of layer-wise parameter constraints and convolution-operator spectral norms enter through a sum of logarithmic terms, and the bounds decay as $O(1/\sqrt{n})$ with the sample size. These results provide a unified characterization of how architectural complexity and basis-function parameterization affect the generalization of PG-CNNs.
- **Comparison of Harmonic and Fourier Bases.** We instantiate the general theory for two widely used basis parameterizations, namely circular harmonic and Fourier bases, by explicitly characterizing the spectral norms of their discretized circular convolution operators and their group-transformed counterparts. For the Fourier basis, the resulting operator-norm bound is $D_l \leq \sqrt{2|G_l|}K^3$, whereas the harmonic basis yields $D_l \leq \sqrt{|G_l|N_l}K^2$, with $N_l \leq 2K^2$. Hence, without SVD reduction, the harmonic basis can lead to a smaller complexity term and a tighter generalization bound. When SVD reduction is applied to the Fourier basis as in Xie et al. (2022), the effective number of basis functions can be substantially reduced, which can reverse this comparison and yield a tighter bound. These results establish an explicit theoretical connection between basis redundancy, convolution-operator spectral norms, and the generalization guarantees of PG-CNNs.

The paper is organized as follows. Section 3 introduces the basic concepts and notation used throughout the paper. Section 4 presents the main generalization error bounds for PG-CNNs and compares different basis parameterizations. Section 5 presents experimental results supporting the theoretical analysis. Finally, Section 6 concludes the paper.

2 Related work

Equivariant neural networks. Equivariant neural networks incorporate symmetry priors into deep models to improve representation efficiency and generalization. Early studies focused on discrete symmetries. Cohen & Welling (2016) introduced p4/p4m group-equivariant CNNs with rotation and reflection equivariance on square grids, while Hoogeboom et al. (2018) extended this idea to

hexagonal grids with p6/p6m symmetries. However, lattice-based methods are restricted to specific discrete transformations. To support more general symmetries, filter parameterization methods represent convolution kernels using basis functions. Weiler et al. (2018) proposed steerable filters based on harmonic functions for continuous rotations, followed by extensions to $O(3)$ and $O(2)$ groups (Cohen et al., 2018; Cohen & Welling, 2022), PDO-based equivariant convolutions (Shen et al., 2020), and Fourier-based parameterizations for improved representation accuracy and reduced aliasing (Xie et al., 2022). Recent frameworks further generalized equivariance to broader transformation groups, such as $E(n)$-equivariant networks (Cesa et al., 2022) and affine-equivariant models (Vlačić & Bölcskei, 2021; Li et al., 2024). Meanwhile, efficient equivariant architectures have been developed through dynamic filters, adaptive operators, and transformation decomposition (He et al., 2021a; 2022; Shen et al., 2024). These methods have enabled equivariant models in applications including image restoration, super-resolution, scientific computing, and molecular dynamics (Fu et al., 2023; Xie et al., 2025; Liu et al., 2025; Zhu et al., 2025; Zhao et al., 2026; Fu et al., 2026; Tan et al., 2026; Thiemann et al., 2025).

Generalization analysis for equivariant neural networks. Theoretical studies have shown that symmetry can improve the generalization ability of neural networks. Sokolic et al. (2017) demonstrated that invariance improves margin-based robustness, while Bietti & Mairal (2019) and Lyle et al. (2020) analyzed the benefits of group symmetry from the perspectives of robustness and data efficiency. Further results established provable generalization improvements of equivariant models in kernel methods and PAC learning frameworks (Elesedy, 2021; 2022). Empirical studies also verified the stability and performance advantages of equivariant architectures (Zhu et al., 2021). Beyond exact equivariance, Petrache & Trivedi (2023) investigated approximation–generalization trade-offs under imperfect symmetry. Behboodi et al. (2022) derived PAC-Bayesian generalization bounds for equivariant networks, while Li et al. (2026) extended the analysis to Markov data by incorporating both symmetry and temporal dependence. These studies provide theoretical evidence for the generalization benefits of equivariance, but explicit bounds that characterize practical equivariant convolutional architectures and their parameterization remain limited.

3 Background

This section provides an overview of group equivariant convolutional neural networks (G-CNNs) and parametrized group equivariant convolutional neural networks.

3.1 Group equivariant Convolutional Neural Networks

A convolutional layer is called G-equivariant if its output transforms consistently with the input transformation, i.e.,

$$\Phi(\rho_{\text{in}}(g)x) = \rho_{\text{out}}(g)\Phi(x), \quad \forall g \in G,$$

where ρ_{out} is the corresponding representation on the output feature space. For G-CNNs, the convolution operation can be written as

$$[f * \psi](g) = \sum_{v} f(v)\psi(g^{-1} \cdot v),$$

where ψ denotes the convolutional filter defined on the discrete grid, and $\psi(g^{-1} \cdot v)$ denotes the convolutional filter obtained by applying the group element g to ψ.

In Paper Cohen & Welling (2016), G denotes the group consisting of 90° rotations and their integer multiples. Each transformation in G is applied to the convolutional kernel to generate a corresponding transformed kernel. Such rotations do not change the number of parameters or the parameter norm of the original kernel; they only alter the spatial arrangement of the parameters. Therefore, although the number of channels increases by a factor of $|G|$, the actual number of learnable parameters remains unchanged.

3.2 Parametrized group equivariant convolutional neural networks

The G-CNNs introduced above may suffer from interpolation artifacts for rotations beyond multiples of 90°. PG-CNNs alleviate this issue by parameterizing convolution kernels as linear combinations

of fixed basis functions and applying the group action to these continuous functions before discretization (Shen et al., 2020; Weiler et al., 2018; Xie et al., 2022). Thus, unlike G-CNNs, which rotate discretized kernels and require interpolation, PG-CNNs transform the continuous basis functions first and then sample the resulting functions on the discrete grid, avoiding interpolation and the associated artifacts. Specifically, a convolution kernel is represented as

$$\psi = \sum_{i=1}^{N} w_i \phi_i,$$

where $\{\phi_i\}_{i=1}^N$ are predefined basis functions and $\{w_i\}_{i=1}^N$ are learnable coefficients. The corresponding convolution operation is then given by

$$[f * \psi](g) = \sum_{u} \sum_{i=1}^{N} w_i (f * \phi_i)(g^{-1}u),$$

yielding a function $f * \psi \to \mathbb{R}$ defined on the group G.

3.3 Examples of kernel parameterizations and basis functions

Steerable filters based on circular harmonics. Weiler et al. (2018) parameterize a filter using circular harmonic basis functions,

$$\psi(x) = \sum_{j=1}^{J} \sum_{k=0}^{K_j} w_{jk} \phi_{jk}(x), \qquad \phi_{jk}(r, \phi) = \tau_j(r) e^{ik\phi}, \tag{1}$$

where $\tau_j(r)$ denotes a radial profile, typically chosen as

$$\tau_j(r) = \exp\left(-\frac{(r-\mu_j)^2}{2\sigma^2}\right).$$

The angular part $e^{ik\phi}$ transforms diagonally under rotations, which makes the filters steerable in the continuous domain.

Fourier-based filter parameterization. The F-Conv framework (Xie et al., 2022) parameterizes filters using a Fourier basis,

$$\psi(x) = \sum_{k=0}^{p-1} \sum_{l=0}^{p-1} \left(w_{kl}^c \phi_{kl}^c(x) + w_{kl}^s \phi_{kl}^s(x)\right), \tag{2}$$

where the basis functions are cosine and sine functions of spatial frequencies,

$$\phi_{kl}^c(x) = \Omega(x) \cos\left(\frac{2\pi}{ph}\left(k - \left\lfloor\frac{p}{2}\right\rfloor,\ l - \left\lfloor\frac{p}{2}\right\rfloor\right) \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right),$$
$$\phi_{kl}^s(x) = \Omega(x) \sin\left(\frac{2\pi}{ph}\left(k - \left\lfloor\frac{p}{2}\right\rfloor,\ l - \left\lfloor\frac{p}{2}\right\rfloor\right) \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right),$$

with $\Omega(x)$ denoting a spatial mask

$$\Omega(x) = \begin{cases} 0, & \text{if } \|x\| \geq (r+h), \\ e^{-5 \cdot \max\left\{\frac{\|x\|^2}{r^2} - 1,\ 0\right\}}, & \text{if } \|x\| < (r+h). \end{cases}$$

Remark 1. *Although both parameterizations construct equivariant filters through basis expansions, they encode equivariance differently. Circular harmonic bases separate radial and angular components, with rotations acting through a phase factor, whereas Fourier bases represent spatial patterns by frequency vectors that rotate under group actions. These different transformation structures, together with their distinct basis cardinalities and convolution-operator spectral norms, lead to different complexity terms in our generalization bounds and motivate the comparison below.*

4 MAIN RESULTS

This section presents the generalization error bounds for general and practically adopted PG-CNNs, together with a theoretical comparison of circular harmonic and Fourier basis parameterizations.

4.1 GENERALIZATION BOUND FOR GENERAL PG-CNNS

We first derive the covering number of a single-layer convolutional linear function class with filters parameterized as linear combinations of basis functions. The result is summarized in the following lemma.

Lemma 4.1. *For the l-th convolutional layer, let $\boldsymbol{x}^{l-1} \in \mathbb{R}^{H_{l-1}\times W_{l-1}\times G_{l-1}}$ denote an input, and parameterize each convolutional filter as $\psi^l(g_m, x) = \sum_{t=1}^{N_l} w_t^{ml}\phi_t(x)$, with fixed basis functions $\{\phi_t\}_{t=1}^{N_l}$. Here, $\mathbf{w}^{ml} = (w_1^{ml}, \ldots, w_{N_l}^{ml})^\top \in \mathbb{R}^{N_l}, m \in [G_{l-1}]$, are learnable coefficient vectors. We assume that the coefficient vectors in the l-th layer satisfy the boundedness condition*

$$\mathcal{W}_l = \left\{\mathbf{w}^l = (\mathbf{w}^{1l}, \ldots, \mathbf{w}^{G_{l-1}l})^\top \in \mathbb{R}^{N_l G_{l-1}} : \|\mathbf{w}^l\|_2 \le W_l\right\}.$$

Then, the covering number of the corresponding convolutional function class satisfies

$$\mathcal{N}\left(\left\{f_{\mathbf{a}^l}(x^{l-1}) : f_{\mathbf{a}^l} \in \mathcal{F}_l\right\}, \|\cdot\|_2, \varepsilon_l\right) \le \left(\frac{3W_l D_l \|\mathbf{x}^{l-1}\|_2}{\varepsilon_l}\right)^{N_l G_{l-1}},$$

where $D_l := \sqrt{\sum_{i=1}^{G_l}\sum_{t=1}^{N_l}(D_{t,i}^l)^2}, D_{t,i}^l := \|B_{t,i}^l\|_{\mathrm{op}}$. Here, $B_{t,i}^l$ denotes the convolution matrix induced by the discrete samples of the transformed basis function $\phi_t(g_i'^{-1}\cdot)$.

Remark 2. *In Weiler et al. (2018), as defined in Equation (1), the number of basis functions at layer l is given by $N_l = \sum_{j=1}^{J}(K_j + 1)$. In Xie et al. (2022), an initial set of $N_l = 2K^2$ basis functions is constructed, where K denotes the filter size. In the actual F-Conv implementation, the resulting basis matrix is further reduced by singular value decomposition. Specifically, the effective number of basis functions is given by $N_l^{\mathrm{eff}} = \mathrm{rank}(D) \le 2K^2$, which removes redundant basis functions and yields a non-redundant parameterization of the filters.*

We next establish a covering number bound for L-layer parametrized group equivariant convolutional neural networks.

Lemma 4.2. *Consider fixed nonlinearities $\sigma_1, \ldots, \sigma_L$, each σ_l being c_l-Lipschitz and satisfying $\sigma_l(0) = 0$. Let data $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{H_0\times W_0\times C_0}$ be given, the covering number of parametrized group equivariant convolutional neural network $F_{\mathcal{W}} : \mathcal{X} \to \mathcal{Y}$ with filter parameters $\mathcal{W} = (\mathbf{w}^1, \ldots, \mathbf{w}^L)$ obey $\|\mathbf{w}^l\|_2 \le W_l$ satisfies the following bound:*

$$\ln \mathcal{N}(\{F_{\mathcal{W}}(x) : F_{\mathcal{W}} \in \mathcal{F}\}, \|\cdot\|_2, \varepsilon) \le \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3L\|\mathbf{x}\|_2 \prod_{j=1}^{L} c_j W_j D_j}{\varepsilon}.$$

Combining Lemma 4.2 with Lemma A.1,Lemma A.2 and applying standard arguments yields the desired result.

Theorem 4.3. *Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the label space. Let $\mathcal{L} : \mathcal{Z}\times\mathcal{Y} \to [0,1]$ be an α- Lipschitz loss function with respect to its first argument under the ℓ_2-norm, i.e. $|\mathcal{L}(z,y) - \mathcal{L}(z',y)| \le \alpha\|z - z'\|_2, \forall z, z' \in \mathcal{Z},\ y \in \mathcal{Y}$. Let $\mathcal{F} = \left\{F_{\mathcal{W}} : \mathcal{X} \to \mathcal{Z} : \mathcal{W} = (\mathbf{w}^1, \ldots, \mathbf{w}^L)\right\}$ denote the class of parametrized group equivariant neural networks. The corresponding loss class is*

$$\mathcal{L}_{\mathcal{F}} = \{(x,y) \mapsto \mathcal{L}(F_{\mathcal{W}}(x), y) : F_{\mathcal{W}} \in \mathcal{F}\}.$$

Then, for n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ drawn from any probability distribution on $\mathcal{X}\times\mathcal{Y}$, with probability at least $1-\delta$, the following bound holds uniformly for all $F_{\mathcal{W}} \in \mathcal{F}$:

$$\mathbb{E}\big[\mathcal{L}(F_{\mathcal{W}}(x), y)\big] \le \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(F_{\mathcal{W}}(x_i), y_i) + \frac{4}{n} + \frac{12}{\sqrt{n}}\sqrt{\Delta\sum_{l=1}^{L} N_l G_{l-1}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where $\Delta := \ln(3\alpha\sqrt{n}LR) + \sum_{j=1}^{L} \ln(c_j W_j \sqrt{\sum_{i=1}^{G_l} \sum_{t=1}^{N_l} (D_{t,i}^j)^2})$.

Proof. The detailed proofs are provided in Appendix C. □

Remark 3. *The bound shows that generalization mainly depends on the spectral norms of discretized basis functions, the ℓ_2-norms of basis coefficients, the number of basis functions, the channel widths, and the network depth. The spectral complexity of basis functions is captured by*

$$D_l = \sqrt{\sum_{o=1}^{|G|} \sum_{t=1}^{N_l} \|B_{t,o}^l\|_{\mathrm{op}}^2},$$

where $B_{t,o}^l$ is the linear operator induced by the t-th basis function under transformation g_o. Thus, improving generalization can be achieved by reducing D_l through basis normalization and regularizing the norms of the basis weight vectors $\boldsymbol{a}^l$ via the term $\sum_{l=1}^{L} \ln \|\boldsymbol{a}^l\|_2$. Reducing the number of bases may further tighten the bound, but introduces a trade-off between generalization and representation capacity.

4.2 Generalization bound for practically adopted PG-CNNs

In practically adopted architectures, the group size G_l is typically much smaller than the channel width of the network. Thus, the channel width cannot be fully determined by the group size alone. We therefore consider practically adopted PG-CNNs, where each set of $|G|$ convolutional kernels is generated by applying the group action to a kernel represented as a linear combination of basis functions with a learnable weight vector. Accordingly, for a network with channel width C, $C/|G|$ independently parameterized kernels are required, each generating $|G|$ group transformed kernels. Under this parameterization, we derive the following generalization error bound:

Theorem 4.4. *Let $\mathcal{X}$ be an input space, $\mathcal{Y}$ a label space, and let the loss function $\mathcal{L} : \mathcal{Z} \times \mathcal{Y} \to [0, 1]$ be α-Lipschitz in its first argument with respect to $\|\cdot\|_2$, where $\mathcal{Z}$ is the output space of the network. Define the induced function class $\mathcal{L}_{\mathcal{F}} := \{(x, y) \mapsto \mathcal{L}(F_{\mathcal{B}}(x), y) : F_{\mathcal{B}} \in \mathcal{F}_{PA}\}$, where $\mathcal{F}_{PA}$ is the class of practically adopted PG-CNNs $F_{\mathcal{B}} : \mathcal{X} \to \mathcal{Z}$. Then for n samples (x_i, y_i) drawn i.i.d. from any probability distribution on $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over the samples, the following bound holds for every $F_{\mathcal{B}} \in \mathcal{F}_{PA}$:*

$$\mathbb{E}\big[\mathcal{L}(F_{\mathcal{B}}(x), y)\big] \leq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(F_{\mathcal{B}}(x_i), y_i) + \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\Delta \sum_{l=1}^{L} \frac{N_l C_l C_{l-1}}{|G|}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where $\Delta := \left(\ln 3\alpha\sqrt{n}LR + \sum_{j=1}^{L} \ln c_j B_j \sqrt{\sum_{o=1}^{|G|} \sum_{t=1}^{N_l} (D_{t,o}^j)^2} \right)$.

Proof. The detailed proofs are provided in Appendix D. □

Remark 4. *The main difference between Theorem 4.3 and Theorem 4.4 is that all channels in PG-CNNs are generated by group transformations, whereas only a subset is generated this way in practically adopted PG-CNNs. The resulting generalization bounds mainly differ in the terms*

$$\sum_{l=1}^{L} N_l C_{l-1} \quad \textit{and} \quad \sum_{l=1}^{L} \frac{N_l C_l C_{l-1}}{|G|}.$$

In the first setting, increasing the group size leads to a corresponding increase in the number of channels, which in turn increases the generalization error. In practically adopted PG-CNNs, however, the total number of channels is fixed. Therefore, increasing the group size reduces the number of independently parameterized channels, resulting in a reduction in the generalization error.

4.3 Spectral Norm Comparison of Harmonic and Fourier Bases.

In Appendix E, we compute the spectral norms of two types of basis functions separately. Since convolutional kernels typically have an odd kernel size, we focus on the case where K is odd. For the Fourier basis, there are two types of basis functions, namely the two-dimensional cosine and sine bases, with frequencies (m, n) satisfying $m, n = 0, \ldots, K-1$. When $(m, n) = (0, 0)$, the spectral norms of the cosine and sine bases are K^2 and 0, respectively. For $(m, n) \neq (0, 0)$, the spectral norms of both the cosine and sine bases are $K^2/2$. For the rotated Fourier bases, the rotated frequencies are generally non-integer. In this case, the spectral norms of both the cosine and sine bases lie in the range $[0, K^2]$. Consequently,

$$D_l \leq \sqrt{2|G_l|}K^3. \tag{3}$$

If, as in Xie et al. (2022), the resulting basis matrix is further reduced by singular value decomposition (SVD), this bound can be substantially reduced, depending on the number of redundant basis functions removed by the SVD. For the harmonic basis, the main advantage is that a rotation only changes the phase ϕ in the factor $e^{ik\phi}$. Therefore, the spectral norms of both the original and rotated harmonic bases remain within the range $[0, K^2]$. Consequently, we obtain

$$D_l \leq \sqrt{|G_l|N_l}K^2. \tag{4}$$

To alleviate aliasing effect, the number of basis functions is typically chosen as $N_l = \sum_{j=1}^{J}(K_j + 1) \leq 2K^2$. Without SVD reduction, the harmonic basis can therefore yield a smaller D_l than the Fourier basis and lead to better generalization. With SVD reduction, however, the Fourier basis may achieve a smaller D_l, depending on the extent to which redundant basis functions can be removed, potentially resulting in better generalization.

5 Experiments

Here we only present the main results. Detailed information about the datasets, network architectures, training configurations, and additional experimental results are provided in the Appendix F.

5.1 Experimental Results on Image Classification

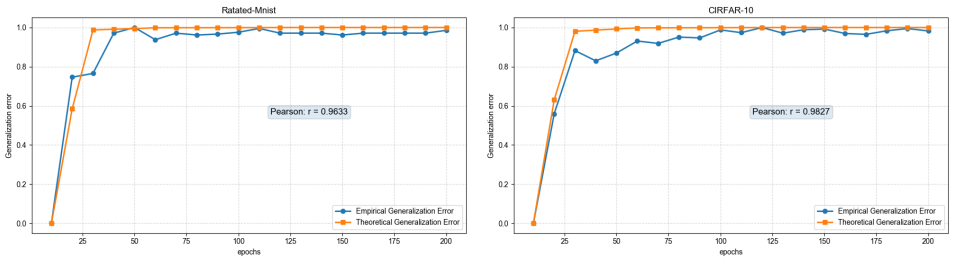


Figure 1: Comparison between empirical generalization error and theoretical generalization error over training epochs. Overall, the theoretically computed generalization error upper bound (orange curve) is highly correlated with the empirical generalization error (blue curve).

Generalization Analysis of PG-CNNs. To evaluate the practical relevance of the theoretical upper bound in Theorem 4.3, we compare the empirical generalization error with the theoretical generalization error. The Pearson correlation coefficient between the two curves is calculated after Min–Max normalization. As shown in Figure 1, the correlation coefficient exceeds 0.96, demonstrating a strong agreement between theoretical predictions and empirical observations.

Furthermore, we investigate the effects of model complexity factors, including the number of basis functions N_l and channel width C_{l-1}, on generalization performance. Under the rotation-equivariant convolutional neural network framework, the number of basis functions in each layer is fixed as N, and the input channels are fixed as C. The results are summarized in Table 1 and Table 2. It can be

Table 1: Generalization error under different numbers of basis functions N

	Generalization Error	
N	Rotated-MNIST	CIFAR-10
2	**0.4513**	**4.0300**
3	0.6117	6.6740
4	0.7690	9.6900
5	0.9223	11.1060

Table 2: Generalization error under different numbers of channels C

	Generalization Error	
C	Rotated-MNIST	CIFAR-10
8	**0.5573**	**6.5940**
10	0.9793	8.3320
12	0.9950	9.8700
14	1.1113	10.4630

observed that increasing either the basis-function number N or the channel number C leads to larger generalization errors. These results are consistent with Theorem 4.3 and Theorem 4.4, confirming that larger basis capacity and channel width increase model complexity and consequently degrade generalization performance.

We further compare the generalization performance of three basis parameterizations: the harmonic basis, the Fourier basis without SVD reduction, and the Fourier basis with SVD reduction. The results are presented in Table 3, which are consistent with Section 4.3.

Table 3: Generalization comparison on MNIST-Rot and CIFAR-10 for harmonic and Fourier bases with and without SVD.

Method	MNIST-rot	CIFAR-10
Harmonic	0.7613 ± 0.0163	20.5200 ± 0.0659
F-Conv-nosvd	0.7885 ± 0.0217	20.5747 ± 0.5713
F-Conv	$\mathbf{0.7458 \pm 0.0277}$	$\mathbf{20.4927 \pm 0.2369}$

Table 4: Comparison of the generalization performance of CNNs, G-CNNs, and PG-CNNs based on EDSR, RDN, and RCAN backbones on benchmark datasets. The factors $\times 2$ and $\times 4$ denote the corresponding $2\times$ and $4\times$ super-resolution scaling factors.

	Scale	Method	$\Delta L_1 \downarrow$	PSNR (dB)↑	SSIM↑
EDSR	$\times 2$	EDSR-CNN	1.66 ± 0.43	31.84 ± 1.03	0.91 ± 0.01
		EDSR-GCNN	1.61 ± 0.40	32.00 ± 1.00	0.91 ± 0.01
		EDSR-FConv	$\mathbf{1.61 \pm 0.35}$	$\mathbf{32.52 \pm 0.92}$	$\mathbf{0.92 \pm 0.02}$
	$\times 4$	EDSR-CNN	2.76 ± 1.11	26.63 ± 1.28	0.76 ± 0.03
		EDSR-GCNN	2.71 ± 1.24	26.64 ± 1.36	0.76 ± 0.02
		EDSR-FConv	$\mathbf{2.66 \pm 0.85}$	$\mathbf{27.22 \pm 1.08}$	$\mathbf{0.78 \pm 0.03}$
RDN	$\times 2$	RDN-CNN	1.65 ± 0.36	32.40 ± 0.92	0.92 ± 0.01
		RDN-GCNN	1.64 ± 0.36	32.41 ± 0.93	0.92 ± 0.01
		RDN-FConv	$\mathbf{1.57 \pm 0.35}$	$\mathbf{32.46 \pm 0.90}$	$\mathbf{0.92 \pm 0.01}$
	$\times 4$	RDN-CNN	2.80 ± 1.09	26.66 ± 1.27	0.76 ± 0.03
		RDN-GCNN	2.72 ± 0.89	27.09 ± 1.13	0.77 ± 0.03
		RDN-FConv	$\mathbf{2.70 \pm 0.87}$	$\mathbf{27.17 \pm 1.10}$	$\mathbf{0.77 \pm 0.03}$
RCAN	$\times 2$	RCAN-CNN	1.59 ± 0.13	32.62 ± 0.84	0.92 ± 0.01
		RCAN-GCNN	1.58 ± 0.14	32.62 ± 0.91	0.92 ± 0.01
		RCAN-FConv	$\mathbf{1.58 \pm 0.13}$	$\mathbf{32.73 \pm 0.85}$	$\mathbf{0.92 \pm 0.02}$
	$\times 4$	RCAN-CNN	2.72 ± 1.00	26.86 ± 1.49	0.76 ± 0.01
		RCAN-GCNN	2.64 ± 0.68	27.24 ± 1.15	0.78 ± 0.00
		RCAN-FConv	$\mathbf{2.56 \pm 0.57}$	$\mathbf{27.35 \pm 1.06}$	$\mathbf{0.78 \pm 0.03}$

5.2 Experimental Results on Image Super-resolution

To further evaluate the applicability of PG-CNNs in low-level vision, we apply the proposed equivariant convolution module to single-image super-resolution (SISR), which aims to recover high-resolution images from low-resolution inputs.

Quantitative results. In the experiments, we use the L_1 loss and denote the generalization error by ΔL_1. As shown in Tables 4 and 5, parameterized group equivariant convolutions achieve lower generalization errors than both standard group equivariant and conventional convolutions, while increasing the group size further reduces ΔL_1. Table 6 shows that normalizing the convolutional basis space also reduces the generalization error, suggesting that commonly used bases may contain redundancy. These observations are consistent with the theoretical results in Theorems 4.3 and 4.4. Additional results are provided in the Appendix.

Table 5: Effect of group size $|G|$ on the generalization performance of PG-CNNs based on EDSR, RDN, and RCAN backbones on benchmark datasets.

Backbone	Scale	$\|G\|$	$\Delta L1 \downarrow$	PSNR ↑	SSIM ↑
EDSR	×2	2	1.62 ± 0.35	32.62 ± 0.92	0.92 ± 0.02
		4	1.63 ± 0.32	32.70 ± 1.80	0.92 ± 0.01
		8	$\mathbf{1.61 \pm 0.35}$	32.52 ± 0.92	0.92 ± 0.02
	×4	2	2.75 ± 0.80	27.25 ± 1.06	0.78 ± 0.03
		4	2.74 ± 0.81	27.28 ± 1.08	0.78 ± 0.03
		8	$\mathbf{2.66 \pm 0.85}$	27.22 ± 1.08	0.78 ± 0.03
RDN	×2	2	1.67 ± 0.35	32.50 ± 0.92	0.92 ± 0.02
		4	1.81 ± 0.35	32.42 ± 0.90	0.92 ± 0.01
		8	$\mathbf{1.57 \pm 0.35}$	32.46 ± 0.90	0.92 ± 0.01
	×4	2	$\mathbf{2.70 \pm 0.82}$	27.25 ± 1.07	0.78 ± 0.03
		4	2.78 ± 0.82	27.28 ± 1.05	0.78 ± 0.03
		8	$\mathbf{2.70 \pm 0.87}$	27.17 ± 1.10	0.77 ± 0.03
RCAN	×2	2	1.69 ± 0.38	32.17 ± 0.97	0.92 ± 0.01
		4	1.67 ± 0.39	32.68 ± 0.96	0.92 ± 0.01
		8	$\mathbf{1.58 \pm 0.13}$	32.73 ± 0.85	0.92 ± 0.02
	×4	2	2.68 ± 0.81	27.21 ± 1.05	0.78 ± 0.03
		4	2.65 ± 0.78	27.32 ± 1.04	0.78 ± 0.03
		8	$\mathbf{2.56 \pm 0.57}$	27.35 ± 1.06	0.78 ± 0.03

Table 6: Effect of SVD reduction on the generalization performance of PG-CNNs based on EDSR on benchmark datasets.

Scale	Method	ΔL_1	PSNR (dB)	SSIM
×2	EDSR-FConv	$\mathbf{1.58 \pm 0.35}$	32.52 ± 0.92	0.92 ± 0.02
	EDSR-FConv-NoSVD	1.61 ± 0.35	$\mathbf{32.70 \pm 0.91}$	$\mathbf{0.92 \pm 0.02}$
×4	EDSR-FConv	$\mathbf{2.69 \pm 0.85}$	27.16 ± 1.10	0.77 ± 0.03
	EDSR-FConv-NoSVD	2.74 ± 0.79	$\mathbf{27.26 \pm 1.05}$	$\mathbf{0.78 \pm 0.03}$

6 Conclusion

In this work, we established explicit generalization error bounds for both general and practically adopted PG-CNNs. We further derived spectral norm bounds for circular harmonic and Fourier bases, enabling a theoretical comparison of their parameterization complexity. Experimental results support the theoretical analysis and demonstrate the effects of group size, basis parameterization, and SVD reduction on generalization. These results provide theoretical guidance for understanding and designing PG-CNNs with improved generalization guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for improving writing clarity, proofreading, and assisting with LaTeX formatting. We have not used generative AI tools for generating research results, experimental data, theoretical claims, or the core scientific contributions of this work, and other disclosure categories are not applicable to this work. Additionally, we used generative AI tools for assisting with code debugging and implementation suggestions. We have reviewed and verified all AI-assisted content, including text revisions and code suggestions, through manual inspection and experimental validation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. We use publicly available benchmark datasets and do not anticipate significant ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed framework, theoretical assumptions, and experimental settings in the main paper and appendix. Complete proofs of the theoretical results are provided in the appendix. The datasets, preprocessing procedures, model architectures, training configurations, and evaluation protocols are documented in the main paper and supplementary material to facilitate independent reproduction of our results.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Arash Behboodi, Gabriele Cesa, and Taco S Cohen. A pac-bayesian generalization bound for equivariant networks. *Advances in Neural Information Processing Systems*, 35:5654–5668, 2022.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 135.1–135.10, 2012.

Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20(25):1–49, 2019.

Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *International Conference on Machine Learning*, pp. 992–1002. PMLR, 2020.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(n)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022.

Qiguang Chen, Zhiwen Li, and Lok Ming Lui. A deep learning framework for diffeomorphic mapping problems via quasi-conformal geometry applied to imaging. *SIAM Journal on Imaging Sciences*, 17(1):501–539, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999. PMLR, 2016.

Taco S Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2022.

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.

Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021.

Bryn Elesedy. Group symmetry in PAC learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.

Jiahong Fu, Qi Xie, Deyu Meng, and Zongben Xu. Rotation equivariant proximal operator for deep unfolding methods in image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:6577–6593, 2023. URL https://api.semanticscholar.org/CorpusID:266551157.

Jiahong Fu, Qi Xie, Deyu Meng, and Zongben Xu. Vanilla group equivariant vision transformer: Simple and effective. *arXiv preprint arXiv:2602.08047*, 2026.

Lingshen He, Yuxuan Chen, Yiming Dong, Yisen Wang, Zhouchen Lin, et al. Efficient equivariant network. *Advances in Neural Information Processing Systems*, 34:5290–5302, 2021a.

Lingshen He, Yuxuan Chen, Zhengyang Shen, Yibo Yang, and Zhouchen Lin. Neural epdos: Spatially adaptive equivariant partial differential operator based networks. In *The Eleventh International Conference on Learning Representations*, 2022.

Wenchong He, Arpan Man Sainju, Zhe Jiang, and Da Yan. Deep neural network for 3d surface segmentation based on contour tree hierarchy. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 253–261. SIAM, 2021b.

Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. In *International Conference on Learning Representations*, 2018.

Haojie Huang, Dian Wang, Robin Walters, and Robert Platt. Equivariant transporter network. In *Robotics: Science and Systems*, 2022.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Erik Jenner and Maurice Weiler. Steerable partial differential operators for equivariant neural networks. *International Conference on Learning Representations*, 2022.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 473–480, 2007.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Hui Li, Zhiguo Wang, Li Sheng, and Bohui Chen. Generalization bounds for equivariant networks on markov data. *SIAM Journal on Mathematics of Data Science*, 8:382–410, 2026. URL https://api.semanticscholar.org/CorpusID:289143841.

Yikang Li, Yeqing Qiu, Yuxuan Chen, Lingshen He, and Zhouchen Lin. Affine equivariant networks based on differential invariants. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5546–5556, 2024.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

Han Liu, Jiahong Fu, Qi Xie, and Deyu Meng. Rotation-equivariant self-supervised method in image denoising. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12720–12730, 2025. URL https://api.semanticscholar.org/CorpusID:278905607.

Philip M Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423. IEEE, 2001.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. *International Conference on Learning Representations*, 2018.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. *Advances in Neural Information Processing Systems*, 36:61936–61959, 2023.

Laura Scarabosio. Deep neural network surrogates for nonsmooth quantities of interest in shape uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):975–1011, 2022.

Zhengyang Shen, Lingshen He, Zhouchen Lin, and Jinwen Ma. Pdo-econvs: Partial differential operator based equivariant convolutions. In *International Conference on Machine Learning*, pp. 8697–8706. PMLR, 2020.

Zhengyang Shen, Yeqing Qiu, Jialun Liu, Lingshen He, and Zhouchen Lin. Efficient learning of scale-adaptive nearly affine invariant networks. *Neural Networks*, 174:106229, 2024.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pp. 1094–1103. PMLR, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Feiyu Tan, Heran Yang, Qihong Duan, Kai Ye, Qi Xie, and Deyu Meng. Image-to-image translation framework embedded with rotation symmetry priors. *arXiv preprint arXiv:2604.12805*, 2026.

Fabian L. Thiemann, Thiago Reschützegger, Massimiliano Esposito, Tseden Taddese, Juan D. Olarte-Plata, and Fausto Martelli. Force-free molecular dynamics through autoregressive equivariant networks. *Nature Machine Intelligence*, 8:764 – 776, 2025. URL https://api.semanticscholar.org/CorpusID:277452298.

Elise Van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. MDP homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4199–4210, 2020.

Vladimir Vapnik. Statistical learning theory. *John Wiley & Sons Google Schola*, 2:831–842, 1998.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021.

Dian Wang, Robin Walters, and Robert Platt. SO(2)-equivariant reinforcement learning. In *International Conference on Learning Representations*, 2022.

Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. *International Conference on Learning Representations*, 2021.

Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 849–858, 2018.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022.

Qi Xie, Qian Zhao, Zongben Xu, and Deyu Meng. Fourier series expansion based filter parametrization for equivariant convolutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4537–4551, 2022.

Qi Xie, Jiahong Fu, Zongben Xu, and Deyu Meng. Rotation equivariant arbitrary-scale image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:11382–11399, 2025. URL https://api.semanticscholar.org/CorpusID:280545891.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730. Springer, 2010.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018a.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018b.

Zhongchen Zhao, Qi Xie, Keyu Huang, Lei Zhang, Deyu Meng, and Zongben Xu. Rotation equivariant mamba for vision tasks. *arXiv preprint arXiv:2603.09138*, 2026.

Pan Zhou and Jiashi Feng. Understanding generalization and optimization performance of deep cnns. In *International Conference on Machine Learning*, pp. 5960–5969. PMLR, 2018.

Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34: 4328–4341, 2021.

Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. In *Robotics: Science and Systems*, 2022.

Yuliang Zhu, Jing Cheng, Qi Xie, Zhuo-Xu Cui, Qingyong Zhu, Yuanyuan Liu, Xin Liu, Jianfeng Ren, Chengbo Wang, and Dong Liang. Dun-sre: Deep unrolling network with spatiotemporal rotation equivariance for dynamic mri reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 19:1694–1709, 2025. URL https://api.semanticscholar.org/CorpusID:279318535.

A Generalization bounds via Rademacher Complexity

Covering number. For a set U and a norm $\|\cdot\|$, the covering number $\mathcal{N}(U, \varepsilon, \|\cdot\|)$ denotes the smallest cardinality of a set $V \subseteq U$ that covers U at scale ε with respect to $\|\cdot\|$, i.e.,

$$\sup_{A \in U} \min_{B \in V} \|A - B\| \ \leq \ \varepsilon.$$

Rademacher complexity. The (empirical) Rademacher complexity of a hypothesis class $\mathcal{F}$ with respect to a sample $S = \{(x_i, y_i)\}_{i=1}^n$ is defined as

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \;:=\; \frac{1}{n}\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\,h(x_i,y_i)\right],$$

where the expectation is taken over the Rademacher random variables $\sigma_1, \ldots, \sigma_n$, which are independent and identically distributed with

$$\Pr[\sigma_i = 1] \;=\; \Pr[\sigma_i = -1] \;=\; \tfrac{1}{2}.$$

Lemma A.1 (Bartlett et al. (2017), Lemma A.5)**.** *Let $\mathcal{F}$ be a class of real-valued functions mapping from $\mathcal{Z}$ to $[0,1]$, and let $S = \{z_1, \ldots, z_n\}$ be a sample of size n drawn from $\mathcal{Z}$. Assume that $0 \in \mathcal{F}$. Then the empirical Rademacher complexity satisfies*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \le \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{\sqrt{n}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \|\cdot\|_2, \varepsilon\right)}\,d\varepsilon\right),$$

where $\mathcal{F}|_S = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$ denotes the restriction of $\mathcal{F}$ to the sample S, and $\mathcal{N}(\mathcal{F}|_S, \|\cdot\|_2, \varepsilon)$ denotes the covering number of $\mathcal{F}|_S$ under the empirical ℓ_2-norm.

Lemma A.2 (Mohri et al. (2018))**.** *Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample $S = \{z_1, \ldots, z_n\}$ of size n, the following inequalities holds for all $f \in \mathcal{F}$:*

$$\mathbb{E}[f(z)] \le \frac{1}{n}\sum_{i=1}^{n} f(z_i) + 2\hat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{3.4}$$

B Group Convolutional Neural Networks

B.1 Group Convolution Setup

Let $x^{l-1} \in \mathbb{R}^{H_{l-1}\times H_{l-1}\times C_{l-1}}$ be the input tensor to the l-th layer, and let G_l be a finite symmetry group with order $|G_l|$ acting on the spatial domain, such as rotations and reflections. The total number of output channels is C_l, where only $D^l_{\text{base}} = C_l/|G_l|$ base filters are learned independently. For each $k = 1, \ldots, D^l_{\text{base}}$, the k-th base filter is defined as $\boldsymbol{\psi}^{l,k} \in \mathbb{R}^{K_l\times K_l\times C_{l-1}}$. For each group element $g \in G_l$, the transformed filter is generated according to the group action: $\boldsymbol{\psi}^{l,k}_g = g\cdot\boldsymbol{\psi}^{l,k}$. The output feature map $y^l \in \mathbb{R}^{H_l\times H_l\times C_l}$ has $C_l = |G_l|\cdot D^l_{\text{base}}$ channels indexed by pairs (g,k):

$$y^l_{g,k} = (g\cdot\boldsymbol{\psi}^{l,k}) * x^{l-1}.$$

All learnable parameters correspond to the base filters, which are assembled into a vector $\boldsymbol{\psi}^l \in \mathbb{R}^{M_l}$ with parameter dimension

$$M_l = D^l_{\text{base}}K_l^2C_{l-1} = \frac{C_l}{|G_l|}K_l^2C_{l-1}.$$

We define the vectorized tensor forms as $\mathbf{x}^{l-1} = \mathrm{vec}(x^{l-1}) \in \mathbb{R}^{N_{\text{in}}}$ and $\mathbf{y}^l = \mathrm{vec}(y^l) \in \mathbb{R}^{N_{\text{out}}}$, where

$$N_{\text{in}} = H_{l-1}^2C_{l-1}, \quad N_{\text{out}} = H_l^2C_l.$$

The convolution operation is linear and can be written as

$$\mathbf{y}^l = A(\boldsymbol{\psi}^l)\mathbf{x}^{l-1},$$

where $A(\boldsymbol{\psi}^l)$ denotes the group convolution matrix associated with layer l. We impose the boundedness assumptions $\|\boldsymbol{\psi}^l\|_2 \le W_l$ and $\|x^{l-1}\|_2 := \|\mathbf{x}^{l-1}\|_2 \le R_{l-1}$.

Then we define the function classes for each component:

- **Input**: $x_i \in \mathcal{X} \subset \mathbb{R}^{H_0\times W_0\times C_0}$, $i \in [n]$, with $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2} \le R$.

- **Convolutional Layer Function Class** For the l-th convolutional layer, the function class is:
$$\mathcal{F}_l = \left\{ f_{\boldsymbol{\psi}^l} : x^{l-1} \mapsto y^l \mid \boldsymbol{\psi}^l \in \Psi_l \right\}, \tag{5}$$
where $\Psi_l = \{\boldsymbol{\psi}^l \in \mathbb{R}^{M_l} : \|\boldsymbol{\psi}^l\|_2 \leq W_l\}$ is the parameter space for layer l.
- **Activation Function** The activation function σ_l is assumed to be a fixed c_l-Lipschitz mapping satisfying $\sigma_l(0) = 0$. The corresponding activation function class is therefore
$$\Sigma_l = \{\sigma_l\}.$$
- **Complete Network Function Class** The complete network function class is the composition of all layers:
$$\mathcal{F} = \Sigma_L \circ \mathcal{F}_L \circ \cdots \circ \Sigma_1 \circ \mathcal{F}_1, \tag{6}$$
where $\circ$ denotes function composition.

Formally, for an input x, the network output is:
$$F_\Psi(x) = \sigma_L f_{\boldsymbol{\psi}^L} \cdots \sigma_1 f_{\boldsymbol{\psi}^1}(x). \tag{7}$$
where $\Psi = (\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^L)$ corresponds to the filter parameters in an L-layer neural network.

B.2 Lipschitz Constants

Lemma B.1. *The group convolutional function is $K_l\sqrt{|G_l|}R_{l-1}$-Lipschitz continuous with respect to the parameter vector $\boldsymbol{\psi}^l$.*

Proof. For fixed input x^{l-1}, two parameter vectors $\boldsymbol{\psi}^l$ and $\boldsymbol{\psi}'^l$, the output difference is:
$$\|f_{\boldsymbol{\psi}^l}(x^{l-1}) - f_{\boldsymbol{\psi}'^l}(x^{l-1})\|_2 = \|A(\boldsymbol{\psi}^l - \boldsymbol{\psi}'^l)\mathbf{x}^{l-1}\|_2.$$
Since the group action is orthogonal (norm-preserving), the total Frobenius norm of all transformed filters is:
$$\sum_{g,k} \|g \cdot \boldsymbol{\psi}^{l,k}\|_2^2 = |G_l| \sum_{k=1}^{D_{\text{base}}^l} \|\boldsymbol{\psi}^{l,k}\|_2^2 = |G_l|\|\boldsymbol{\psi}^l\|_2^2.$$
Using the bound for ordinary convolution, the operator norm of $A(\boldsymbol{\psi}^l)$ satisfies:
$$\|A(\boldsymbol{\psi}^l)\|_{\text{op}} \leq K_l \sqrt{\sum_{g,k} \|g\boldsymbol{\psi}^{l,k}\|_2^2} = K_l\sqrt{|G_l|}\|\boldsymbol{\psi}^l\|_2.$$
Thus,
$$\|A(\boldsymbol{\psi}^l - \boldsymbol{\psi}'^l)\mathbf{x}^{l-1}\|_2 \leq K_l\sqrt{|G_l|}\|\boldsymbol{\psi}^l - \boldsymbol{\psi}'^l\|_2\|\mathbf{x}^{l-1}\|_2 \leq K_l\sqrt{|G_l|}R_{l-1}\|\boldsymbol{\psi}^l - \boldsymbol{\psi}'^l\|_2.$$
Hence, the Lipschitz constant with respect to parameters is:
$$L_{\text{param}}^l = K_l\sqrt{|G_l|}R_{l-1}.$$
□

Lemma B.2. *The group convolutional function is $K_l\sqrt{|G_l|}W_l$-Lipschitz continuous with respect to the input x^{l-1}.*

Proof. For fixed parameters $\boldsymbol{\psi}^l$ and two inputs $x^{l-1}, \tilde{x}^{l-1}$, we have:

$$\|f_{\boldsymbol{\psi}^l}(x^{l-1}) - f_{\boldsymbol{\psi}^l}(\tilde{x}^{l-1})\|_2 \leq \|A(\boldsymbol{\psi}^l)\|_{\text{op}} \|x^{l-1} - \tilde{x}^{l-1}\|_2 \leq K_l \sqrt{|G_l|} \|\boldsymbol{\psi}^l\|_2 \|x^{l-1} - \tilde{x}^{l-1}\|_2.$$

Since $\|\boldsymbol{\psi}^l\|_2 \leq W_l$, we obtain:

$$\|f_{\boldsymbol{\psi}^l}(x^{l-1}) - f_{\boldsymbol{\psi}^l}(\tilde{x}^{l-1})\|_2 \leq K_l \sqrt{|G_l|} W_l \|x^{l-1} - \tilde{x}^{l-1}\|_2.$$

□

Lemma B.3. *The group convolution function is bounded, with an upper bound given by* $K_l \sqrt{|G_l|} W_l R_{l-1}$.

Proof. For any input x^{l-1} with $\|x^{l-1}\|_2 \leq R_{l-1}$ and parameters $\boldsymbol{\psi}^l$ with $\|\boldsymbol{\psi}^l\|_2 \leq W_l$, the output norm satisfies:

$$\|f_{\boldsymbol{\psi}^l}(x^{l-1})\|_2 \leq K_l \sqrt{|G_l|} W_l R_{l-1}.$$

□

We can now derive the generalization error upper bounds for group equivariant convolutional neural networks. We first derive the covering number for the function space of a single-layer group convolutional linear model.

Lemma B.4. *Let* $\boldsymbol{\psi}^l \in \mathbb{R}^{K_l \times K_l \times C_{l-1} \times \frac{C_l}{|G_l|}}$ *and* $x^{l-1} \in \mathbb{R}^{H_{l-1} \times W_{l-1} \times C_{l-1}}$ *be an arbitrary convolution filter and input, respectively. The covering number of the resulting group convolutional layer* $\{f_{\boldsymbol{\psi}^l}(x^{l-1}) : f_{\boldsymbol{\psi}^l} \in \mathcal{F}_l\}$ *satisfies:*

$$N(\{f_{\boldsymbol{\psi}^l}(x^{l-1}) : f_{\boldsymbol{\psi}^l} \in \mathcal{F}_l\}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{3 W_l K_l \|\mathbf{x}^{l-1}\|_2 \sqrt{|G_l|}}{\varepsilon} \right)^{M_l},$$

where $M_l = \frac{K_l^2 C_l C_{l-1}}{|G_l|}$.

Proof. For any $\delta > 0$, the parameter space $\Psi_l = \{\boldsymbol{\psi}^l \in \mathbb{R}^{M_l} : \|\boldsymbol{\psi}^l\|_2 \leq W_l\}$ has covering number:

$$N(\Psi_l, \|\cdot\|_2, \delta) \leq \left(\frac{3 W_l}{\delta} \right)^{M_l},$$

where $M_l = \frac{K_l^2 C_{l-1} C_l}{|G_l|}$.

Using the Lipschitz continuity with respect to parameters (Lemma B.1), the covering number of the function class $\mathcal{F}_l$ is bounded by:

$$N(\mathcal{F}_l, \|\cdot\|_2, \varepsilon) \leq N\left(\Psi_l, \|\cdot\|_2, \frac{\varepsilon}{L^l_{\text{param}}} \right) \leq \left(\frac{3 W_l K_l \|\mathbf{x}^{l-1}\|_2 \sqrt{|G_l|}}{\varepsilon} \right)^{M_l}.$$

□

Now, we proceed to establish a covering bound for the L-layers G-CNNs, as presented in the following lemma.

Lemma B.5. *Consider fixed nonlinearities $\sigma_1, \ldots, \sigma_L$, each σ_l being c_l-Lipschitz and satisfying $\sigma_l(0) = 0$. Let data $x \in \mathcal{X} \subset \mathbb{R}^{H_0 \times W_0 \times C_0}$ be given, the covering number of group equivariant neural network $F_\Psi : \mathcal{X} \to \mathcal{Y}$ with filter parameters $\Psi = (\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^L)$ obey $\|\boldsymbol{\psi}^l\|_2 \leq W_l$ satisfies the following bound:*

$$\ln \mathcal{N}(\{F_\Psi(x) : F_\Psi \in \mathcal{F}_G\}, \|\cdot\|_2, \varepsilon) \leq \sum_{l=1}^{L} \frac{K_l^2 C_l C_{l-1}}{|G_l|} \ln \frac{3L\|\mathbf{x}\|_2 \prod_{j=1}^{L} c_j W_j K_j \sqrt{|G_j|}}{\varepsilon}.$$

Proof. Our goal is to compute the covering number $N(\mathcal{F}, \|\cdot\|_2, \varepsilon)$ of this function class with respect to the ℓ_2-norm.

Step 1: Construct the cover

Inductively construct covers $\mathcal{C}_1, \ldots, \mathcal{C}_L$ as follows.

• Choose an ε_1-cover $\mathcal{C}_1$ of $\{f_{\boldsymbol{\psi}^1} x, f_{\boldsymbol{\psi}^1} \in \mathcal{F}_1\}$, thus

$$\mathcal{N}(\{f_{\boldsymbol{\psi}^1} x,\ f_{\boldsymbol{\psi}^1} \in \mathcal{F}_1\}, \|\cdot\|_2, \varepsilon_1) \leq |\mathcal{C}_1| := n_1$$

• For every element $F_l \in \mathcal{C}_l$, construct an ε_{l+1}-cover $\mathcal{G}_{l+1}(F_l)$ of $\{f_{\boldsymbol{\psi}^{l+1}} \sigma_l F_l,\ f_{\boldsymbol{\psi}^{l+1}} \in \mathcal{F}_{l+1}\}$, then

$$\mathcal{N}(\{f_{\boldsymbol{\psi}^{l+1}} \sigma_l F_l,\ f_{\boldsymbol{\psi}^{l+1}} \in \mathcal{F}_{l+1}\}, \|\cdot\|_2, \varepsilon_{l+1}) \leq |\mathcal{G}_{l+1}(F_l)| := n_{l+1}$$

Lastly, form the cover

$$\mathcal{C}_{l+1} = \bigcup_{F_l \in \mathcal{C}_l} \mathcal{G}_l(F_l),$$

whose cardinality satisfies

$$|\mathcal{C}_{l+1}| = |\mathcal{C}_l| n_{l+1} \leq \prod_{i=1}^{l+1} n_i.$$

Define $\mathcal{C} := \{\sigma_L F_L,\ F_L \in \mathcal{C}_L\}$, we will show that $\mathcal{C}$ serves as an τ-cover of $\mathcal{F}$, where $\tau := \sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l K_l \sqrt{|G_l|} W_l$. For any $F = \sigma_L F_L = \sigma_L f_{\boldsymbol{\psi}^L} \sigma_{L-1} f_{\boldsymbol{\psi}^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1} \in \mathcal{F}$, there exists $F' = \sigma_L F'_L = \sigma_L f_{\boldsymbol{\psi}'^L} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1} \in \mathcal{C}$ such that:

$$\begin{aligned}
\|F - F'\|_2 =& \|\sigma_L F_L - \sigma_L F'_L\|_2 \\
& \|\sigma_L f_{\boldsymbol{\psi}^L} \sigma_{L-1} f_{\boldsymbol{\psi}^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1} - \sigma_L f_{\boldsymbol{\psi}'^L} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1}\|_2 \\
& \leq c_L \|f_{\boldsymbol{\psi}^L} \sigma_{L-1} f_{\boldsymbol{\psi}^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1} - f_{\boldsymbol{\psi}'^L} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1}\|_2 \\
& \leq c_L (\|f_{\boldsymbol{\psi}^L} \sigma_{L-1} f_{\boldsymbol{\psi}^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1} - f_{\boldsymbol{\psi}^L} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1}\|_2 \\
& + \|f_{\boldsymbol{\psi}^l} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1} - f_{\boldsymbol{\psi}'^L} \sigma_{L-1} f_{\boldsymbol{\psi}'^{L-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}'^1}\|_2) \\
& \leq c_L K_l \sqrt{|G_l|} W_l \|\sigma_{L-1} F_{l-1} - \sigma_{L-1} F'_{L_1}\|_2 + c_L \varepsilon_L \qquad \text{(Lemma B.2)} \\
& \vdots \\
& \leq \sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l K_l \sqrt{|G_l|} W_l.
\end{aligned}$$

Step 2: Calculate the covering number

The per-layer cover resolutions $(\varepsilon_1, \ldots, \varepsilon_L)$ set according to

$$\varepsilon_l := \frac{\varepsilon}{L c_l \prod_{j>l} c_j K_j \sqrt{|G_j|} W_j}$$

By this choice, it follows that the final cover resolution ε provided by Step 1 satisfies

$$\sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l K_l \sqrt{|G_l|} W_l = \sum_{j=1}^{L} \frac{\varepsilon}{L} = \varepsilon.$$

Within the rest of the proof, a pivotal strategy involves utilizing the covering number estimates furnished by Lemma B.4.

$$\ln \mathcal{N}(\{F_\Psi(x) : F_\Psi \in \mathcal{F}_G\}, \|\cdot\|_2, \varepsilon) = \ln |\mathcal{C}| = \sum_{l=1}^{L} \ln |\mathcal{C}_l|$$
$$\leq \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln \frac{3W_l K_l \sqrt{|G_l|} \|x^{l-1}\|_2}{\varepsilon_l} \quad \text{(Lemma B.4)}. \tag{8}$$

To simplify this expression, note for any $(\boldsymbol{\psi}^1, \cdots, \boldsymbol{\psi}^{l-1})$ that

$$\begin{aligned}
\|x^{l-1}\|_2 &= \|\sigma_{l-1} f_{\boldsymbol{\psi}^{l-1}} \sigma_{l-1} f_{\boldsymbol{\psi}^{l-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1}(x) - \sigma_{l-1}(0)\|_2 \\
&\leq c_{l-1} \|f_{\boldsymbol{\psi}^{l-1}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1}(x)\|_2 \\
&\leq c_{l-1} K_{l-1} W_{l-1} \sqrt{|G_l|} \|f_{\boldsymbol{\psi}^{l-2}} \ldots, \sigma_1 f_{\boldsymbol{\psi}^1}(x)\|_2 \qquad \text{(Lemma B.3)} \\
&\vdots \\
&\leq \|x\|_2 \prod_{j=1}^{l-1} c_j W_j K_j \sqrt{|G_j|}.
\end{aligned} \tag{9}$$

Combining (8) and (9) and subsequently plugging in the chosen value for ε_l, we derive the following result:

$$\begin{aligned}
\ln \mathcal{N}(\{F_\Psi(x) : F_\Psi \in \mathcal{F}_G\}, \|\cdot\|_2, \varepsilon) &\leq \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln \frac{3W_l K_l \sqrt{|G_l|} \|x\|_2 \prod_{j=1}^{l-1} c_j W_j K_j \sqrt{|G_j|}}{\varepsilon_l} \\
&= \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln \frac{3L \|x\|_2 \prod_{j=1}^{L} c_j W_j K_j \sqrt{|G_j|}}{\varepsilon}.
\end{aligned}$$

□

Combining the covering number bound from Lemma B.5 with Lemma A.1 via standard analytical arguments yields the final result.

Theorem B.6. *Let $\mathcal{X}$ be an input space, $\mathcal{Y}$ a label space, and let the loss function $\mathcal{L} : \mathcal{Z} \times \mathcal{Y} \to [0, 1]$ be α-Lipschitz in its first argument with respect to $\|\cdot\|_2$, where $\mathcal{Z}$ is the output space of the network. Define the induced function class $\mathcal{L}_\mathcal{F} := \big\{(x, y) \mapsto \mathcal{L}(F_\Psi(x), y) : F_\Psi \in \mathcal{F}_G\big\}$, where $\mathcal{F}_G$ is the class of group equivariant neural networks $F_\Psi : \mathcal{X} \to \mathcal{Z}$. Then for n samples (x_i, y_i) drawn i.i.d. from any probability distribution on $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over the samples, the following bound holds for every $F_\Psi \in \mathcal{F}_G$:*

$$\mathbb{E}\big[\mathcal{L}(F_\Psi(x), y)\big] \leq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(F_\Psi(x_i), y_i) + \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\Delta \sum_{l=1}^{L} \frac{K_l^2 C_l C_{l-1}}{|G_l|}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where $\Delta := \ln(3\alpha\sqrt{n}\, LR) + \sum_{j=1}^{L} \ln\Big(c_j W_j K_j \sqrt{|G_j|}\Big)$.

Proof. Given a sample $S = \{(x_i, y_i)\}_{i=1}^n$, then

$$\ln \mathcal{N}(\mathcal{F}_S, \|\cdot\|_2, \varepsilon) \leq \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln \frac{3L \|X\|_2 \prod_{j=1}^{L} c_j W_j K_j \sqrt{|G_j|}}{\varepsilon}$$

where $\|X\|_2 = \sqrt{\sum_{i=1}^n \|\mathbf{x}_i\|_2^2}$. Consider the network class $\mathcal{L}_{\mathcal{F}_G}$ obtained by appending the loss $\mathcal{L}$ to the output of the given network class:

$$\mathcal{L}_{\mathcal{F}_G} := \{(x, y) \to \mathcal{L}(f(x), y) : f \in \mathcal{F}_G\}.$$

Since the function $(z, y) \to \mathcal{L}(z, y)$ is α-Lipschitz in its first argument with respect to $\|\cdot\|_2$, the function class $\mathcal{L}_{\mathcal{F}_G}$ still falls within the setting of Lemma B.5 and thereby yields

$$\ln \mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, \varepsilon) \leq \ln \mathcal{N}(\mathcal{F}_S, \|\cdot\|_2, \frac{\varepsilon}{\alpha}) \leq \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln \frac{3\alpha L \|X\|_2 \prod_{j=1}^{L} c_j W_j K_j \sqrt{|G_j|}}{\varepsilon}. \tag{10}$$

Since $\mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, \varepsilon)$ is a decreasing function of ε we see that

$$\int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, \varepsilon)} d\varepsilon \leq \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, \alpha)} d\varepsilon \leq \sqrt{n} \sqrt{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, \alpha)}$$

By choosing $\alpha = 1/\sqrt{n}$ we obtain a simplified bound for Lemma A.1

$$\hat{\mathfrak{R}}_S(\mathcal{L}_{\mathcal{F}}|_S) \leq \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}|_S, \|\cdot\|_2, 1/\sqrt{n})}. \tag{11}$$

By substituting Eq. (10) into Eq. (11), we obtain:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{L}_{\mathcal{F}_G}) &\leq \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|} \ln 3\alpha \sqrt{n} L \|X\|_2 \prod_{j=1}^{L} c_j W_j K_j \sqrt{|G_j|}} \\
&= \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\left(\ln 3\alpha \sqrt{n} L \|X\|_2 + \sum_{j=1}^{L} \ln c_j W_j K_j \sqrt{|G_j|} \right) \sum_{l=1}^{L} \frac{K_l^2 C_{l-1} C_l}{|G_l|}}.
\end{aligned}$$

□

C Parametrized group equivariant convolutional neural networks

C.1 PG-CNNs model settings

We consider a family of parametrized convolutional neural networks with group symmetry. Let $\{\phi_t\}_{t=1}^{N_l}$ be a set of fixed basis functions (e.g., circular harmonics with Gaussian radial profiles). A base filter $\psi^l(g_m, x)$ in the l-th layer is defined as a linear combination of these basis functions:

$$\psi^l(g_m, x) = \sum_{t=1}^{N_l} w_t^{ml} \phi_t(x), \tag{12}$$

where $\mathbf{w}^{ml} = (w_1^{ml}, \ldots, w_{N_l}^{ml})^\top \in \mathbb{R}^{N_l}, m \in [G_l], l \in [L]$ are learnable parameters. We assume the parameter set of the l-th layer is bounded: $\mathcal{W}_l = \{\mathbf{w}^l = (\mathbf{w}^{1l}, \ldots, \mathbf{w}^{G_{l-1}l})^\top \in \mathbb{R}^{N_l G_{l-1}} : \|\mathbf{w}^l\|_2 \leq W_l\}$. Each layer has G_l steerable filters defined by:

$$\psi_i^l(g, x) := g_i' \psi^l(g_m, x) = \psi^l(g_i'^{-1} g_m, g_i'^{-1} x) = \sum_{t=1}^{N_l} w_t^{ml} \phi_t(g_i'^{-1} x), \quad g_i' \in G, \quad i \in [G_l]. \tag{13}$$

These filters are evaluated on a grid and then employed in a convolutional architecture, yielding G_l output channels.

Then we can define the function classes for each component:

- **Input**: $x_i \in \mathcal{X} \subset \mathbb{R}^{H_0 \times W_0 \times C_0}$, $i \in [n]$, with $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2} \leq R$.
- **Convolutional Layer Function Class** For the l-th convolutional layer, the function class is:

$$\mathcal{F}_l = \left\{ f_{\mathbf{w}^l} : x^{(l-1)} \mapsto (\psi_1^l * x^{(l-1)}, \ldots, \psi_{C_l}^l * x^{(l-1)}) \mid \mathbf{w}^l \in \mathcal{W}_l \right\}, \tag{14}$$

where $\mathcal{W}_l = \{\mathbf{w}^l \in \mathbb{R}^{N_l G_{l-1}} : \|\mathbf{w}^l\|_2 \leq W_l\}$ is the parameter space for layer l.

- **Activation Function** The activation function σ_l is considered as a fixed c_l-Lipschitz mapping:
$$\Sigma_l = \{\sigma_l\}, \tag{15}$$
which is a singleton set but with known Lipschitz constant.
- **Complete Network Function Class** The complete network function class is the composition of all layers:
$$\mathcal{F} = \Sigma_L \circ \mathcal{F}_L \circ \cdots \circ \Sigma_1 \circ \mathcal{F}_1, \tag{16}$$
where $\circ$ denotes function composition, and Σ_p is the activation after pooling (if any). For simplicity, we may set Σ_p as identity if no activation is applied after pooling.

Formally, for an input x, the network output is:

$$F_{\mathcal{W}}(x) = \sigma_L f_{\mathbf{w}^L} \cdots \sigma_1 f_{\mathbf{w}^1}(x). \tag{17}$$

where $\mathcal{W} = (\mathbf{w}^1, \ldots, \mathbf{w}^L)$ corresponds to the parameters in an L-layer neural network.

Matrix Representation

For fixed input x^{l-1}, which consists of G_{l-1} channels (each corresponding to a group element from the previous layer), the output for channel i is given by:

$$y_i^l = \psi_i^l * x^{l-1} = \sum_{t=1}^{N_l} \sum_{m=1}^{G_{l-1}} w_t^{ml} \left[\phi_t(g_i'^{-1}\cdot) * x_m^{l-1}\right], \quad i \in [G_l]. \tag{18}$$

Define linear operators $B_{t,i}^l$ that map x_m^{l-1} to $\phi_t(g_i'^{-1}\cdot) * x_m^{l-1}$. And let $\mathbf{x}_m^{l-1} = \mathrm{vec}(x_m^{l-1}) \in \mathbb{R}^{H_{l-1}W_{l-1}}$, $\mathbf{y}_i^l = \mathrm{vec}(y_i^l) \in \mathbb{R}^{H_l W_l}$. Then:

$$\mathbf{y}_i^l = \sum_{t=1}^{N_l} \sum_{m=1}^{G_{l-1}} w_t^{ml} B_{t,i}^l \mathbf{x}_m^{l-1}. \tag{19}$$

Lemma C.1. *For any $f_{\mathbf{w}^l} \in \mathcal{F}_l$ and input x^{l-1}, the corresponding layer output satisfies*

$$\|f_{\mathbf{w}^l}(x^{l-1})\|_2 \le W_l D_l \|\mathbf{x}^{l-1}\|_2,$$

where D_l characterizes the aggregate spectral norm of the discretized basis operators at layer l.

Proof. For any $f_{\mathbf{w}^l} \in \mathcal{F}_l$ and input x^{l-1}, the i-th channel output is:

$$\mathbf{y}_i^l = \sum_{t=1}^{N_l} \sum_{m=1}^{G_{l-1}} w_t^{ml} B_{t,i}^l \mathbf{x}_m^{l-1}.$$

The overall output norm is:

$$\|f_{\mathbf{w}^l}(x^{l-1})\|_2^2 = \sum_{i=1}^{G_l} \|\mathbf{y}_m^{l-1}\|_2^2.$$

For each i, let $D_{t,i}^l = \|B_{t,i}^l\|_{\mathrm{op}}$ denote the operator norm of $B_{t,i}^l$. Then:

$$\|\mathbf{y}_i^{l-1}\|_2 \le \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} |w_t^{ml}| \, \|B_{t,i}^l \mathbf{x}_m^{l-1}\|_2 \le \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} |w_t^{ml}| \, D_{t,i}^l \|\mathbf{x}_m^{l-1}\|_2.$$

By Cauchy-Schwarz inequality,

$$\sum_{t=1}^{N_l} |w_t^{ml}| D_{t,i}^l \le \|\mathbf{w}^{ml}\|_2 \sqrt{\sum_{t=1}^{N_l} (D_{t,i}^l)^2}.$$

Therefore,

$$\begin{aligned}\|\mathbf{y}_i^{l-1}\|_2 &\leq \sqrt{\sum_{t=1}^{N_l}(D_{t,i}^l)^2}\sum_{m=1}^{G_{l-1}}\|\mathbf{w}^{ml}\|_2\|\mathbf{x}_m^{l-1}\|_2 \\ &\leq \sqrt{\sum_{t=1}^{N_l}(D_{t,i}^l)^2}\left(\sum_{m=1}^{G_{l-1}}\|\mathbf{w}^{ml}\|_2^2\right)^{1/2}\left(\sum_{m=1}^{G_{l-1}}\|\mathbf{x}_m^{l-1}\|_2^2\right)^{1/2} \\ &\leq W_l\|\mathbf{x}^{l-1}\|_2\sqrt{\sum_{t=1}^{N_l}(D_{t,i}^l)^2}\end{aligned}$$

Finally, the total output norm satisfies:

$$\|f_{\mathbf{w}^l}(x^{l-1})\|_2 \leq W_l\|\mathbf{x}^{l-1}\|_2\sqrt{\sum_{i=1}^{G_l}\sum_{t=1}^{N_l}(D_{t,i}^l)^2}.$$

For simplicity, define:

$$D_l := \sqrt{\sum_{i=1}^{G_l}\sum_{t=1}^{N_l}(D_{t,i}^l)^2},$$

then we have the bound:

$$\|f_{\mathbf{w}^l}(x^{l-1})\|_2 \leq W_l D_l\|\mathbf{x}^{l-1}\|_2.$$

□

Lemma C.2. *The Lipschitz constant with respect to the input of a single-layer parameterized group equivariant convolutional neural network is bounded by $D_l\|\mathbf{w}^l\|_2$.*

Proof. We now analyze the Lipschitz continuity of the mapping from inputs to outputs. Consider two inputs x^{l-1} and x'^{l-1}. The difference in outputs is:

$$\|f_{\mathbf{w}^l}(x^{l-1}) - f_{\mathbf{w}^l}(x'^{l-1})\|_2^2 = \sum_{i=1}^{G_l}\left\|\sum_{m=1}^{G_{l-1}}\sum_{t=1}^{N_l}w_t^{ml}B_{t,i}^l(\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1})\right\|_2^2.$$

For each i, we have:

$$\left\|\sum_{m=1}^{G_{l-1}}\sum_{t=1}^{N_l}w_t^{ml}B_{t,i}^l(\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1})\right\|_2 \leq \sum_{m=1}^{G_{l-1}}\sum_{t=1}^{N_l}|w_t^{ml}|\,D_{t,i}^l\|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2.$$

Applying Cauchy-Schwarz inequality twice, we obtain:

$$\begin{aligned}\left(\sum_{m=1}^{G_{l-1}}\sum_{t=1}^{N_l}|w_t^{ml}|\,G_{t,i}^l\|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2\right)^2 &\leq \left(\sum_{m=1}^{G_{l-1}}\|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2^2\right)\left(\sum_{m=1}^{G_{l-1}}\left(\sum_{t=1}^{N_l}|w_t^{ml}|D_{t,i}^l\right)^2\right) \\ &\leq \|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2^2\sum_{m=1}^{G_{l-1}}\left(\|\mathbf{w}^{ml}\|_2^2\sum_{t=1}^{N_l}(D_{t,i}^l)^2\right) \\ &\leq \|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2^2\left(\sum_{t=1}^{N_l}(D_{t,i}^l)^2\right)\sum_{m=1}^{G_{l-1}}\|\mathbf{w}^{ml}\|_2^2.\end{aligned}$$

Then,

$$\left\|\sum_{m=1}^{G_{l-1}}\sum_{t=1}^{N_l}w_t^{ml}B_{t,i}^l(\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1})\right\|_2 \leq \|\mathbf{x}^{l-1} - \mathbf{x'}^{l-1}\|_2\|\mathbf{w}^l\|_2\sqrt{\sum_{t=1}^{N_l}(D_{t,i}^l)^2}.$$

Therefore,

$$\|f_{\mathbf{w}^l}(x^{l-1}) - f_{\mathbf{w}^l}(x'^{l-1})\|_2 \leq \|\mathbf{x}^{l-1} - \mathbf{x}'^{l-1}\|_2 D_l \|\mathbf{w}^l\|_2.$$

□

Lemma C.3. *For a single-layer parameterized group convolutional neural network, the Lipschitz constant with respect to the parameter vector is bounded by* $\|\mathbf{x}^{l-1}\|_2 D_l$.

Proof. We now analyze the Lipschitz continuity of the mapping from parameters to outputs. Consider two sets of parameters $\{w_t^l\}$ and $\{w_t'^l\}$. For a fixed input x^{l-1}, the difference in outputs is:

$$\|f_{\mathbf{w}^l}(x^{l-1}) - f_{\mathbf{w}'^l}(x^{l-1})\|_2^2 = \sum_{i=1}^{G_l} \left\| \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} (w_t^{ml} - w_t'^{ml}) B_{t,i}^l \mathbf{x}_m^{l-1} \right\|_2^2.$$

For each i, we have:

$$\left\| \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} (w_t^{ml} - w_t'^{ml}) B_{t,i}^l \mathbf{x}_m^{l-1} \right\|_2 \leq \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} |w_t^{ml} - w_t'^{ml}|\, D_{t,i}^l \|\mathbf{x}_m^{l-1}\|_2.$$

Applying Cauchy-Schwarz inequality twice, we obtain:

$$\begin{aligned}
\left(\sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} |w_t^{ml} - w_t'^{ml}| D_{t,i}^l \|\mathbf{x}_m^{l-1}\|_2 \right)^2 &\leq \left(\sum_{m=1}^{G_{l-1}} \|\mathbf{x}_m^{l-1}\|_2^2 \right) \left(\sum_{m=1}^{G_{l-1}} \left(\sum_{t=1}^{N_l} |w_t^{ml} - w_t'^{ml}| D_{t,i}^l \right)^2 \right) \\
&\leq \|\mathbf{x}^{l-1}\|_2^2 \sum_{m=1}^{G_{l-1}} \left(\|\mathbf{w}^{ml} - \mathbf{w}'^{ml}\|_2^2 \sum_{t=1}^{N_l} (D_{t,i}^l)^2 \right) \\
&\leq \|\mathbf{x}^{l-1}\|_2^2 \left(\sum_{t=1}^{N_l} (D_{t,i}^l)^2 \right) \sum_{m=1}^{G_{l-1}} \|\mathbf{w}^{ml} - \mathbf{w}'^{ml}\|_2^2.
\end{aligned}$$

Then,

$$\left\| \sum_{m=1}^{G_{l-1}} \sum_{t=1}^{N_l} (w_t^{ml} - w_t'^{ml}) B_{t,i}^l \mathbf{x}_m^{l-1} \right\|_2 \leq \|\mathbf{x}^{l-1}\|_2 \sqrt{\sum_{t=1}^{N_l} (D_{t,i}^l)^2} \sqrt{\sum_{m=1}^{G_{l-1}} \|\mathbf{w}^{ml} - \mathbf{w}'^{ml}\|_2^2}.$$

Therefore,

$$\|f_{\mathbf{w}^l}(x^{l-1}) - f_{a'^l}(x^{l-1})\|_2 \leq \|\mathbf{x}^{l-1}\|_2 D_l \|\mathbf{w}^l - \mathbf{w}'^l\|_2.$$

□

Building on the above results, we proceed to derive the generalization error bounds. We first establish the covering number of the function class corresponding to a single layer parameterized group equivariant convolutional linear model, as stated in Lemma 4.1.

Proof of Lemma 4.1.

Proof. The function class of interest is:

$$\mathcal{F}_l = \left\{ f_{\mathbf{w}^l} : x^{(l-1)} \mapsto (\psi_1^l * x^{(l-1)}, \ldots, \psi_{C_l}^l * x^{(l-1)}) \mid \mathbf{w}^l \in \mathcal{W}_l \right\}, \tag{20}$$

where $*$ denotes convolution. Our goal is to compute the covering number $N(\mathcal{F}_l, \|\cdot\|_2, \varepsilon_l)$ of this function class with respect to the ℓ_2-norm. Let $N(\mathcal{W}_l, \|\cdot\|_2, \delta_l)$ denote the covering number of the parameter space. For the Euclidean ball $\mathcal{W}_l$ of radius W_l in $\mathbb{R}^{N_l G_{l-1}}$, we have the standard bound:

$$N(\mathcal{W}_l, \|\cdot\|_2, \delta_l) \leq \left(\frac{3W_l}{\delta_l} \right)^{N_l G_{l-1}}. \tag{21}$$

By the Lipschitz property (Lemma C.3), an $\frac{\varepsilon_l}{R_l}$-cover of $\mathcal{W}_l$ induces an ε_l-cover of $\mathcal{F}_l$. Hence:

$$N(\mathcal{F}_l, \|\cdot\|_2, \varepsilon_l) \leq N\left(\mathcal{W}_l, \|\cdot\|_2, \frac{\varepsilon_l}{R_l}\right) \leq \left(\frac{3W_l R_l}{\varepsilon}\right)^{N_l G_{l-1}}. \tag{22}$$

Substituting $R_l = \|\mathbf{x}^{l-1}\|_2 D_l$ yields the final bound:

$$N(\mathcal{F}_l, \|\cdot\|_2, \varepsilon_l) \leq \left(\frac{3W_l D_l \|\mathbf{x}^{l-1}\|_2}{\varepsilon_l}\right)^{N_l G_{l-1}}. \tag{23}$$

□

Next, we will establish a covering bound for an L-layer PG-CNNs, as stated in lemma 4.2.

Proof of Lemma 4.2.

Proof. Our goal is to compute the covering number $N(\mathcal{F}, \|\cdot\|_2, \varepsilon)$ of this function class with respect to the ℓ_2-norm.

Step 1: Construct the cover

Inductively construct covers $\mathcal{C}_1, \ldots, \mathcal{C}_L$ as follows.

• Choose an ε_1-cover $\mathcal{G}_1$ of $\{f_{\mathbf{w}^1} x, f_{\mathbf{w}^1} \in \mathcal{F}_1\}$, thus

$$\mathcal{N}(\{f_{\mathbf{w}^1} x,\ f_{\mathbf{w}^1} \in \mathcal{F}_1\}, \|\cdot\|_2, \varepsilon_1) \leq |\mathcal{G}_1| := n_1$$

• For every element $F_l \in \mathcal{C}_l$, construct an ε_{l+1}-cover $\mathcal{G}_{l+1}(F_l)$ of $\{f_{\mathbf{w}^{l+1}} \sigma_l F_l,\ f_{\mathbf{w}^{l+1}} \in \mathcal{F}_{l+1}\}$, then

$$\mathcal{N}(\{f_{\mathbf{w}^{l+1}} \sigma_l F_l,\ f_{\mathbf{w}^{l+1}} \in \mathcal{F}_{l+1}\}, \|\cdot\|_2, \varepsilon_{l+1}) \leq |\mathcal{G}_{l+1}(F_l)| := n_{l+1}$$

Lastly, form the cover

$$\mathcal{G}_{l+1} = \bigcup_{F_l \in \mathcal{C}_l} \mathcal{G}_l(F_l),$$

whose cardinality satisfies

$$|\mathcal{G}_{l+1}| = |\mathcal{G}_l| n_{l+1} \leq \prod_{i=1}^{l+1} n_i.$$

Define $\mathcal{C} := \{\sigma_L F_L,\ F_L \in \mathcal{C}_L\}$, we will show that $\mathcal{C}$ serves as an τ-cover of $\mathcal{F}$, where $\tau := \sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l W_l D_l$. For any $F = \sigma_L f_{\mathbf{w}^L} \sigma_{L-1} f_{\mathbf{w}^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}^1} \in \mathcal{F}$, there exists $F' = \sigma_L f_{\mathbf{w}'^L} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1} \in \mathcal{C}$ such that:

$$\begin{aligned}
\|F - F'\|_2 =& \|\sigma_L f_{\mathbf{w}^L} \sigma_{L-1} f_{\mathbf{w}^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}^1} - \sigma_L f_{\mathbf{w}'^L} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1}\|_2 \\
&\leq c_L \|f_{\mathbf{w}^L} \sigma_{L-1} f_{\mathbf{w}^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}^1} - f_{\mathbf{w}'^L} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1}\|_2 \\
&\leq c_L (\|f_{\mathbf{w}^L} \sigma_{L-1} f_{\mathbf{w}^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}^1} - f_{\mathbf{w}^L} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1}\|_2 \\
&+ \|f_{\mathbf{w}^l} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1} - f_{\mathbf{w}'^L} \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1}\|_2) \\
&\leq c_L W_L D_L \|\sigma_{L-1} f_{\mathbf{w}^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}^1} - \sigma_{L-1} f_{\mathbf{w}'^{L-1}} \ldots, \sigma_1 f_{\mathbf{w}'^1}\|_2 + c_L \varepsilon_L \qquad \text{(Lemma C.2)} \\
&\vdots \\
&\leq \sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l W_l D_l.
\end{aligned}$$

Step 2: Calculate the covering number

The per-layer cover resolutions $(\varepsilon_1, \ldots, \varepsilon_L)$ set according to

$$\varepsilon_l := \frac{\varepsilon}{L c_l \prod_{j>l} c_j W_j D_j}$$

By this choice, it follows that the final cover resolution ε provided by Step 1 satisfies

$$\tau \leq \sum_{j=1}^{L} \varepsilon_j c_j \prod_{l=j+1}^{L} c_l W_l D_l = \sum_{j=1}^{L} \frac{\varepsilon}{L} = \varepsilon.$$

Within the rest of the proof, a pivotal strategy involves utilizing the covering number estimates furnished by Lemma 4.1.

$$\begin{aligned} \ln \mathcal{N}(\mathcal{F}, \|\cdot\|_2, \varepsilon) &= \ln |\mathcal{C}| = \sum_{l=1}^{L} \ln |\mathcal{G}_l| \\ &\leq \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3 W_l \|\mathbf{x}^{l-1}\|_2 D_l}{\varepsilon_l} \quad \text{(Lemma 4.1)}, \end{aligned} \tag{24}$$

where $D_l := \sqrt{\sum_{i=1}^{G_l} \sum_{t=1}^{N_l} (D_{t,i}^l)^2}$. To simplify this expression, note for any $(\mathbf{w}^1, \cdots, \mathbf{w}^{l-1})$ that

$$\begin{aligned} \|\mathbf{x}^{l-1}\|_2 &= \|\sigma_{l-1} f_{\mathbf{w}^{l-1}} \sigma_{l-1} f_{\mathbf{w}^{l-1}} \ldots, \sigma_1 f_{\mathbf{w}^1}(x) - \sigma_{l-1}(0)\|_2 \\ &\leq c_{l-1} \|f_{\mathbf{w}^{l-1}} \ldots, \sigma_1 f_{\mathbf{w}^1}(x)\|_2 \\ &\leq c_{l-1} W_{l-1} D_{l-1} \|f_{\mathbf{w}^{l-2}} \ldots, \sigma_1 f_{\mathbf{w}^1}(x)\|_2 \qquad \text{(Lemma C.1)} \\ &\vdots \\ &\leq \|\mathbf{x}\|_2 \prod_{j=1}^{l-1} c_j W_j D_j. \end{aligned} \tag{25}$$

Combining (24) and (25) and subsequently plugging in the chosen value for ε_i, we derive the following result:

$$\begin{aligned} \ln \mathcal{N}(\mathcal{F}, \|\cdot\|_2, \varepsilon) &\leq \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3 W_l D_l \|\mathbf{x}\|_2 \prod_{j=1}^{l-1} c_j W_j D_j}{\varepsilon_l} \\ &= \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3L \|\mathbf{x}\|_2 \prod_{j=1}^{L} c_j W_j D_j}{\varepsilon}. \end{aligned}$$

□

Combining the covering number bound from Lemma 4.2 with Lemma A.1 via standard analytical arguments yields the final result.

C.2 Proof of Theorem 4.3

Given a sample $S = \{(x_i, y_i)\}_{i=1}^n$, then

$$\ln \mathcal{N}(\mathcal{F}_S, \|\cdot\|_2, \varepsilon) \leq \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3L \|X\|_2 \prod_{j=1}^{L} c_j W_j D_j}{\varepsilon}$$

where $\|X\|_2 = \sqrt{\sum_{i=1}^n \|x_i\|_2^2}$.

Consider the network class $\mathcal{L}_\mathcal{F}$ obtained by appending the loss $\mathcal{L}$ to the output of the given network class:

$$\mathcal{L}_\mathcal{F} := \{(x, y) \to \mathcal{L}(f(x), y) : f \in \mathcal{F}\}.$$

Since the function $(z, y) \to \mathcal{L}(z, y)$ is α-Lipschitz in its first argument with respect to $\|\cdot\|_2$, the function class $\mathcal{L}_\mathcal{F}$ still falls within the setting of Lemma B.5 and thereby yields

$$\ln \mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2, \varepsilon) \leq \ln \mathcal{N}(\mathcal{F}_S, \|\cdot\|_2, \frac{\varepsilon}{\alpha}) \leq \sum_{l=1}^{L} N_l G_{l-1} \ln \frac{3\alpha L \|X\|_2 \prod_{j=1}^{L} c_j W_j D_j}{\varepsilon}. \tag{26}$$

Since $\mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2, \varepsilon)$ is a decreasing function of ε we see that

$$\int_\alpha^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2), \varepsilon)} d\varepsilon \leq \int_\alpha^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2), \alpha)} d\varepsilon \leq \sqrt{n} \sqrt{\log \mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2), \alpha)}$$

By choosing $\alpha = 1/\sqrt{n}$ we obtain a simplified bound

$$\hat{\mathfrak{R}}_S(\mathcal{L}_\mathcal{F}) \leq \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{L}_\mathcal{F}|_S, \|\cdot\|_2), 1/\sqrt{n})}. \tag{27}$$

By substituting Eq. (26) into Eq. (27), we obtain:

$$\begin{aligned}\hat{\mathfrak{R}}_S(\mathcal{L}_\mathcal{F}) &\leq \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\sum_{l=1}^{L} N_l G_{l-1} \ln \left(3\alpha\sqrt{n} L \|X\|_2 \prod_{j=1}^{L} c_j W_j D_j \right)} \\ &= \frac{4}{n} + \frac{12}{\sqrt{n}} \sqrt{\left(\ln 3\alpha\sqrt{n} L \|X\|_2 + \sum_{j=1}^{L} \ln c_j W_j D_j \right) \sum_{l=1}^{L} N_l G_{l-1}}.\end{aligned}$$

D PRACTICALLY ADOPTED PG-CNNS

D.1 MODEL SETTINGS

In practically adopted architectures, the group size G_l is typically much smaller than the channel width of the network. Thus, the channel width cannot be fully determined by the group size alone. We therefore consider practically adopted parameterized group equivariant convolutional neural networks, where each set of $|G|$ convolutional kernels is generated by applying the group action to a kernel represented as a linear combination of basis functions with a learnable weight vector. Accordingly, for a network with channel width C, $C/|G|$ independently parameterized kernels are required, each generating $|G|$ group transformed kernels.

Let $\{\phi_t\}_{t=1}^{N_l}$ be a set of fixed basis functions (e.g., circular harmonics with Gaussian radial profiles). A base filter $\psi^l(x)$ in the l-th layer is defined as a linear combination of these basis functions:

$$\psi^l_{i,j}(x) = \sum_{t=1}^{N_l} b^l_{t,i,j} \phi_t(x), \tag{28}$$

where $\boldsymbol{b}^l_{i,j} = (b^l_{1,i,j}, \ldots, b^l_{N_l,i,j})^\top \in \mathbb{R}^{N_l}, i \in [D^l_{\text{base}}], j \in [C_{l-1}], l \in [L]$ are learnable parameters. We assume the parameter set of the l-th layer is bounded: $\mathcal{B}_l = \{\boldsymbol{b}^l = (b^l_{t,i,j})_{t,i,j} \in \mathbb{R}^{N_l C_l C_{l-1}/|G|} : \|\boldsymbol{b}^l\|_2 \leq B_l\}$. Each layer has $|G|$ steerable filters defined by:

$$\psi^l_{(i,o),j}(x) := g_o \cdot \psi^l_{i,j}(x) = \psi^l_{i,j}(g_o^{-1} \cdot x) = \sum_{t=1}^{N_l} b^l_{t,i,j} \phi_t(g_o^{-1} \cdot x), \quad g_o \in G. \tag{29}$$

These filters are used in a convolutional architecture, producing C_l output channels.

Then we can define the function classes for each component:

- **Input**: $x_i \in \mathcal{X} \subset \mathbb{R}^{H_0 \times W_0 \times C_0}$, $i \in [n]$, with $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2} \leq R$.
- **Convolutional Layer Function Class** For the l-th convolutional layer, the function class is:

$$\mathcal{F}_l = \left\{ f_{\boldsymbol{b}^l} : x^{(l-1)} \mapsto (\psi^l_{1,1} * x^{(l-1)}, \ldots \psi^l_{1,|G|} * x^{(l-1)} \ldots, \psi^l_{D^l_{\text{base}},|G|} * x^{(l-1)}) \mid \boldsymbol{b}^l \in \mathcal{B}_l \right\}, \tag{30}$$

  where $\mathcal{B}_l = \{\boldsymbol{b}^l \in \mathbb{R}^{N_l C_l C_{l-1}/|G|} : \|\boldsymbol{b}^l\|_2 \leq B_l\}$ is the parameter space for layer l.

- **Activation Function** The activation function σ_l is considered as a fixed c_l-Lipschitz mapping:
$$\Sigma_l = \{\sigma_l\}, \tag{31}$$
which is a singleton set but with known Lipschitz constant.
- **Complete Network Function Class** The complete network function class is the composition of all layers:
$$\mathcal{F}_{AP} = \Sigma_L \circ \mathcal{F}_L \circ \cdots \circ \Sigma_1 \circ \mathcal{F}_1, \tag{32}$$
where $\circ$ denotes function composition, and Σ_p is the activation after pooling (if any). For simplicity, we may set Σ_p as identity if no activation is applied after pooling.

Formally, for an input x, the network output is:
$$F(x)_{AP} = \sigma_L f_{\boldsymbol{b}^L} \cdots \sigma_1 f_{\boldsymbol{b}^1}(x). \tag{33}$$

Matrix Representation

For fixed input x^{l-1}, which consists of C_{l-1} channels (each corresponding to a group element from the previous layer), the output for channel i is given by:

$$y^l_{i,o} = \psi^l_{i,o} * x^{l-1} = \sum_{t=1}^{N_l} \sum_{j=1}^{C_{l-1}} b^l_{t,i,j} \left[\phi_t(g_o^{-1}\cdot) * x^{l-1}_j\right], \quad i \in [D^l_{\text{base}}], o \in [|G|] \tag{34}$$

Define linear operators $B^l_{t,o}$ that map x^{l-1}_m to $\phi_t(g_o^{-1}\cdot) * x^{l-1}_j$. And let $\mathbf{x}^{l-1}_j = \text{vec}(x^{l-1}_j) \in \mathbb{R}^{H_{l-1}W_{l-1}}$, $\mathbf{y}^l_{i,o} = \text{vec}(y^l_{i,o}) \in \mathbb{R}^{H_l W_l}$. Then:

$$\mathbf{y}^l_{i,o} = \sum_{t=1}^{N_l} \sum_{j=1}^{C_{l-1}} b^l_{t,i,j} B^l_{t,o} \mathbf{x}^{l-1}_j. \tag{35}$$

Lemma D.1. *For any $f_{\boldsymbol{b}^l} \in \mathcal{F}_l$ and input x^{l-1}, the corresponding layer output is bounded as*
$$\|f_{\boldsymbol{b}^l}(x^{l-1})\|_2 \le B_l D_l \|\mathbf{x}^{l-1}\|_2,$$
where D_l characterizes the bound for the sum of spectral norms of the discretized bases and group-transformed counterparts.

Proof. For any $f_{\boldsymbol{b}^l} \in \mathcal{F}_l$ and input x^{l-1}, the i-th channel output is:
$$\mathbf{y}^l_{i,o} = \sum_{t=1}^{N_l} \sum_{j=1}^{C_{l-1}} b^l_{t,i,j} B^l_{t,o} \mathbf{x}^{l-1}_j.$$

The overall output norm is:
$$\|f_{\boldsymbol{b}^l}(x^{l-1})\|_2^2 = \sum_{i=1}^{D^l_{\text{base}}} \sum_{o=1}^{|G|} \|\mathbf{y}^{l-1}_{i,o}\|_2^2.$$

For each o, let $D^l_{t,o} = \|B^l_{t,o}\|_{\text{op}}$ denote the operator norm of $B^l_{t,o}$. Then:
$$\|\mathbf{y}^{l-1}_{i,o}\|_2 \le \sum_{j=1}^{C_{l-1}} \sum_{t=1}^{N_l} |b^l_{t,i,j}| \, \|B^l_{t,o}\mathbf{x}^{l-1}_j\|_2 \le \sum_{j=1}^{C_{l-1}} \sum_{t=1}^{N_l} |b^l_{t,i,j}| \, D^l_{t,o} \|\mathbf{x}^{l-1}_j\|_2.$$

By Cauchy-Schwarz inequality,
$$\sum_{t=1}^{N_l} |b^l_{t,i,j}| D^l_{t,o} \le \|\boldsymbol{b}^l_{i,j}\|_2 \sqrt{\sum_{t=1}^{N_l} (D^l_{t,o})^2}.$$

Therefore,

$$\begin{aligned}\|\mathbf{y}_{i,o}^{l-1}\|_2 &\le \sqrt{\sum_{t=1}^{N_l}(D_{t,o}^l)^2}\sum_{j=1}^{C_{l-1}}\|b_{i,j}^l\|_2\|\mathbf{x}_j^{l-1}\|_2 \\ &\le \sqrt{\sum_{t=1}^{N_l}(D_{t,o}^l)^2}\left(\sum_{j=1}^{C_{l-1}}\|\boldsymbol{b}_{i,j}^l\|_2^2\right)^{1/2}\left(\sum_{j=1}^{C_{l-1}}\|\mathbf{x}_j^{l-1}\|_2^2\right)^{1/2} \\ &\le \|\mathbf{x}^{l-1}\|_2\sqrt{\sum_{j=1}^{C_{l-1}}\|b_{i,j}^l\|_2^2}\sqrt{\sum_{t=1}^{N_l}(D_{t,o}^l)^2}\end{aligned}$$

Finally, the total output norm satisfies:

$$\begin{aligned}\|f_{\boldsymbol{b}^l}(x^{l-1})\|_2^2 &= \|\mathbf{x}^{l-1}\|_2^2\sum_{i=1}^{D_{\text{base}}^l}\sum_{o=1}^{|G|}\sum_{j=1}^{C_{l-1}}\|b_{i,j}^l\|_2^2\sum_{t=1}^{N_l}(D_{t,o}^l)^2 \\ &= \|\mathbf{x}^{l-1}\|_2^2\,\|\boldsymbol{b}^l\|_2^2\sum_{o=1}^{|G|}\sum_{t=1}^{N_l}(D_{t,o}^l)^2\end{aligned}$$

For simplicity, define:

$$D_l := \sqrt{\sum_{o=1}^{|G|}\sum_{t=1}^{N_l}(D_{t,o}^l)^2},$$

then we have the bound:

$$\|f_{\boldsymbol{b}^l}(x^{l-1})\|_2 \le B_l D_l\|\mathbf{x}^{l-1}\|_2.$$

□

Lemma D.2. *The Lipschitz constant with respect to the input of a single-layer practically adopted parameterized group equivariant convolutional neural network is bounded by $D_l\|\boldsymbol{b}^l\|_2$.*

Proof. We now analyze the Lipschitz continuity of the mapping from inputs to outputs. Consider two inputs x^{l-1} and x'^{l-1}. The difference in outputs is:

$$\|f_{\boldsymbol{b}^l}(x^{l-1}) - f_{\boldsymbol{b}^l}(x'^{l-1})\|_2^2 = \sum_{i=1}^{D_{\text{base}}^l}\sum_{o=1}^{|G|}\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}b_{t,i,j}^l B_{t,o}^l(\mathbf{x}_j^{l-1} - \mathbf{x'}_j^{l-1})\right\|_2^2.$$

Firstly, we have:

$$\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}b_{t,i,j}^l B_{t,o}^l(\mathbf{x}_j^{l-1} - \mathbf{x'}_j^{l-1})\right\|_2 \le \sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}|b_{t,i,j}^l|\,D_{t,o}^l\|\mathbf{x}_j^{l-1} - \mathbf{x'}_j^{l-1}\|_2.$$

Applying Cauchy-Schwarz inequality twice, we obtain:

$$\begin{aligned}\left(\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}|b_{t,i,j}^l|\,D_{t,o}^l\|\mathbf{x}_j^{l-1} - \mathbf{x'}_j^{l-1}\|_2\right)^2 &\le \left(\sum_{j=1}^{C_{l-1}}\|\mathbf{x}_j^{l-1} - \mathbf{x'}_j^{l-1}\|_2^2\right)\left(\sum_{j=1}^{C_{l-1}}\left(\sum_{t=1}^{N_l}|b_{t,i,j}^l|D_{t,o}^l\right)^2\right) \\ &\le \|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2^2\sum_{j=1}^{C_{l-1}}\left(\|\boldsymbol{b}_{i,j}^l\|_2^2\sum_{t=1}^{N_l}(D_{t,o}^l)^2\right) \\ &\le \|\mathbf{x}_m^{l-1} - \mathbf{x'}_m^{l-1}\|_2^2\left(\sum_{t=1}^{N_l}(D_{t,o}^l)^2\right)\sum_{j=1}^{C_{l-1}}\|\boldsymbol{b}_{i,j}^l\|_2^2.\end{aligned}$$

Then,

$$\sum_{i=1}^{D_{\text{base}}^l}\sum_{o=1}^{|G|}\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l} b_{t,i,j}^l B_{t,o}^l(\mathbf{x}_j^{l-1}-\mathbf{x}'^{l-1}_j)\right\|_2^2 \le \|\mathbf{x}^{l-1}-\mathbf{x}'^{l-1}\|_2\sqrt{\sum_{i=1}^{D_{\text{base}}^l}\sum_{j=1}^{C_{l-1}}\|\boldsymbol{b}_{i,j}^l\|_2^2}\sqrt{\sum_{o=1}^{|G|}\sum_{t=1}^{N_l}(D_{t,o}^l)^2}.$$

Therefore,

$$\|f_{\boldsymbol{b}^l}(x^{l-1})-f_{\boldsymbol{b}^l}(x'^{l-1})\|_2 \le \|\mathbf{x}^{l-1}-\mathbf{x}'^{l-1}\|_2 D_l\|\boldsymbol{b}^l\|_2.$$

□

Lemma D.3. *For a single-layer practically adopted parameterized group convolutional neural network, the Lipschitz constant with respect to the parameter vector is bounded by $\|\mathbf{x}^{l-1}\|_2 D_l$.*

Proof. We now analyze the Lipschitz continuity of the mapping from parameters to outputs. Consider two sets of parameters $\{\boldsymbol{b}_t^l\}$ and $\{\boldsymbol{b}_t'^l\}$. For a fixed input x^{l-1}, the difference in outputs is:

$$\|f_{\boldsymbol{b}^l}(x^{l-1})-f_{\boldsymbol{b}'^l}(x^{l-1})\|_2^2 = \sum_{i=1}^{D_{\text{base}}^l}\sum_{o=1}^{|G|}\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}(b_{t,i,j}^l-b_{t,i,j}'^l)B_{t,o}^l\mathbf{x}_j^{l-1}\right\|_2^2.$$

Firstly, we have:

$$\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}(b_{t,i,j}^l-b_{t,i,j}'^l)B_{t,o}^l\mathbf{x}_j^{l-1}\right\|_2 \le \sum_{m=1}^{C_{l-1}}\sum_{t=1}^{N_l}|b_{t,i,j}^l-b_{t,i,j}'^l|\,D_{t,o}^l\|\mathbf{x}_j^{l-1}\|_2.$$

Applying Cauchy-Schwarz inequality twice, we obtain:

$$\begin{aligned}\left(\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}|b_{t,i,j}^l-b_{t,i,j}'^l|D_{t,o}^l\|\mathbf{x}_j^{l-1}\|_2\right)^2 &\le \left(\sum_{j=1}^{C_{l-1}}\|\mathbf{x}_j^{l-1}\|_2^2\right)\left(\sum_{j=1}^{C_{l-1}}\left(\sum_{t=1}^{N_l}|b_{t,i,j}^l-b_{t,i,j}'^l|D_{t,o}^l\right)^2\right)\\ &\le \|\mathbf{x}^{l-1}\|_2^2\sum_{j=1}^{C_{l-1}}\left(\|\boldsymbol{b}_{i,j}^l-\boldsymbol{b}_{i,j}'^l\|_2^2\sum_{t=1}^{N_l}(D_{t,o}^l)^2\right)\\ &\le \|\mathbf{x}^{l-1}\|_2^2\left(\sum_{t=1}^{N_l}(D_{t,o}^l)^2\right)\sum_{j=1}^{C_{l-1}}\|\boldsymbol{b}_{i,j}^l-\boldsymbol{b}_{i,j}'^l\|_2^2.\end{aligned}$$

Then,

$$\left\|\sum_{j=1}^{C_{l-1}}\sum_{t=1}^{N_l}(\boldsymbol{b}_{i,j}^l-\boldsymbol{b}_{i,j}'^l)B_{t,o}^l\mathbf{x}_j^{l-1}\right\|_2 \le \|\mathbf{x}^{l-1}\|_2\sqrt{\sum_{t=1}^{N_l}(D_{t,o}^l)^2}\sqrt{\sum_{j=1}^{C_{l-1}}\|\boldsymbol{b}_{i,j}^l-\boldsymbol{b}_{i,j}'^l\|_2^2}.$$

Therefore,

$$\|f_{\boldsymbol{b}^l}(x^{l-1})-f_{\boldsymbol{b}'^l}(x^{l-1})\|_2 \le \|\mathbf{x}^{l-1}\|_2 D_l\|\boldsymbol{b}^l-\boldsymbol{b}'^l\|_2.$$

□

Lemma D.4. *Let $\psi_{i,j}^l(x)=\sum_{t=1}^{N_l} b_{t,i,j}^l\phi_t(x)$ and $x^{l-1}\in\mathbb{R}^{H_{l-1}\times W_{l-1}\times C_{l-1}}$ be an arbitrary base filter and input, respectively. The covering number of the resulting practically adopted PG-CNNs $\{f_{\boldsymbol{b}^l}(x^{l-1}): f_{\boldsymbol{b}^l}\in\mathcal{F}_l\}$ satisfies:*

$$\mathcal{N}(\{f_{\boldsymbol{b}^l}(x^{l-1}): f_{\boldsymbol{b}^l}\in\mathcal{F}_l\},\|\cdot\|_2,\varepsilon_l) \le \left(\frac{3B_lD_l\|\mathbf{x}^{l-1}\|_2}{\varepsilon_l}\right)^{N_lC_lC_{l-1}/|G|},$$

where $D_l:=\sqrt{\sum_{o=1}^{|G|}\sum_{t=1}^{N_l}(D_{t,o}^l)^2}$. Here, $D_{t,o}^l=\|B_{t,o}^l\|_{\text{op}}$ denotes the operator norm of $B_{t,o}^l$, and the linear operator $B_{t,o}^l$ maps x_j^{l-1} to $\phi_t(g_o'^{-1}\cdot)*x_j^{l-1}$ (x_j^{l-1} is the j-th channel of x^{l-1}).

Proof. The function class of interest is:

$$\mathcal{F}_l = \left\{ f_{\boldsymbol{b}^l} : x^{(l-1)} \mapsto (\psi^l_{1,1} * x^{(l-1)}, \dots \psi^l_{1,|G|} * x^{(l-1)} \dots, \psi^l_{D^l_{\text{base}},|G|} * x^{(l-1)}) \mid \boldsymbol{b}^l \in \mathcal{B}_l \right\}, \quad (36)$$

where $*$ denotes convolution. Our goal is to compute the covering number $N(\varepsilon_l, \mathcal{F}_l, \|\cdot\|_2)$ of this function class with respect to the ℓ_2-norm. Let $N(\delta_l, \mathcal{B}_l, \|\cdot\|_2)$ denote the covering number of the parameter space. For the Euclidean ball $\mathcal{B}_l$ of radius B_l in $\mathbb{R}^{N_l C_l C_{l-1}/|G|}$, we have the standard bound:

$$N(\mathcal{B}_l, \|\cdot\|_2, \delta_l) \leq \left(\frac{3B_l}{\delta_l}\right)^{N_l C_l C_{l-1}/|G|}. \quad (37)$$

By the Lipschitz property, an $\frac{\varepsilon_l}{R_l}$-cover of $\mathcal{A}_l$ induces an ε_l-cover of $\mathcal{F}_l$. Hence:

$$N(\{f_{\boldsymbol{b}^l}(x^{l-1}) : f_{\boldsymbol{b}^l} \in \mathcal{F}_l\}, \|\cdot\|_2, \varepsilon_l) \leq N\left(\mathcal{B}_l, \|\cdot\|_2, \frac{\varepsilon_l}{R_l}\right) \leq \left(\frac{3B_l R_l}{\varepsilon}\right)^{N_l C_l C_{l-1}/|G|}. \quad (38)$$

Substituting $R_l = \|\mathbf{x}^{l-1}\|_2 D_l$ yields the final bound:

$$N(\{f_{\boldsymbol{b}^l}(x^{l-1}) : f_{\boldsymbol{b}^l} \in \mathcal{F}_l\}, \|\cdot\|_2, \varepsilon_l) \leq \left(\frac{3B_l D_l \|\mathbf{x}^{l-1}\|_2}{\varepsilon_l}\right)^{N_l C_l C_{l-1}/|G|}. \quad (39)$$

□

All the necessary lemmas for the proof have been provided (Lemmas D.1–D.3). By applying these lemmas together with the proof strategy of Theorem 4.3, the result of Theorem 4.4 directly follows from Lemma D.4. The detailed derivation is omitted here for brevity.

E Analysis of the spectral norm of discrete convolutions induced by basis functions

E.1 Fourier basis

Lemma E.1. *Let the $K \times K$ convolution kernel be obtained by uniformly sampled the two-dimensional cosine basis function*

$$K_{i,j} = \cos\left(\frac{2\pi}{K}(mi + nj)\right), \qquad m, n = 0, \dots, K-1. \quad (40)$$

Let $W \in \mathbb{R}^{K^2 \times K^2}$ denote the corresponding circular convolution matrix. Then

$$\|W\|_2 = \begin{cases} K^2, & (m,n) = (-m,-n) \pmod K, \\ \dfrac{K^2}{2}, & \textit{otherwise}. \end{cases} \quad (41)$$

And for uniformly sampled two-dimensional sine basis function

$$K_{i,j} = \sin\left(\frac{2\pi}{K}(mi + nj)\right), \qquad m, n = 0, \dots, K-1. \quad (42)$$

Let $W \in \mathbb{R}^{K^2 \times K^2}$ denote the corresponding circular convolution matrix. Then

$$\|W\|_2 = \begin{cases} 0, & (m,n) = (-m,-n) \pmod K, \\ \dfrac{K^2}{2}, & \textit{otherwise}. \end{cases} \quad (43)$$

Proof. Using Euler's identity, the kernel can be written as

$$K_{i,j} = \frac{1}{2}\exp\left(\frac{2\pi\mathrm{i}}{K}(mi+nj)\right) + \frac{1}{2}\exp\left(-\frac{2\pi\mathrm{i}}{K}(mi+nj)\right). \tag{44}$$

Let F_K denote the unitary one-dimensional discrete Fourier transform (DFT) matrix and define the two-dimensional DFT matrix by

$$F = F_K \otimes F_K. \tag{45}$$

Since W represents a circular convolution operator, it is diagonalized by the two-dimensional DFT:

$$W = F^* \Lambda F, \tag{46}$$

where

$$\Lambda = \operatorname{diag}\left(\widehat{K}_{r,s}\right)_{r,s=0}^{K-1}, \tag{47}$$

and $\widehat{K}$ denotes the two-dimensional DFT of K.

Using the non-normalized DFT convention,

$$\widehat{K}_{r,s} = \sum_{i=0}^{K-1}\sum_{j=0}^{K-1} K_{i,j}\exp\left(-\frac{2\pi\mathrm{i}}{K}(ri+sj)\right). \tag{48}$$

Substituting the expression of $K_{i,j}$ gives

$$\begin{aligned}\widehat{K}_{r,s} = \frac{1}{2}\sum_{i,j}\exp\left(\frac{2\pi\mathrm{i}}{K}\big((m-r)i+(n-s)j\big)\right) \\ + \frac{1}{2}\sum_{i,j}\exp\left(-\frac{2\pi\mathrm{i}}{K}\big((m+r)i+(n+s)j\big)\right).\end{aligned} \tag{49}$$

By the discrete Fourier orthogonality relation

$$\sum_{i=0}^{K-1}\exp\left(\frac{2\pi\mathrm{i}}{K}qi\right) = \begin{cases} K, & q \equiv 0 \pmod K, \\ 0, & \text{otherwise}, \end{cases} \tag{50}$$

we obtain

$$\widehat{K}_{r,s} = \frac{K^2}{2}\left[\mathbf{1}_{(r,s)=(m,n)} + \mathbf{1}_{(r,s)=(-m,-n)}\right]. \tag{51}$$

Since F is unitary,

$$\|W\|_2 = \|\Lambda\|_2 = \max_{r,s}|\widehat{K}_{r,s}|. \tag{52}$$

Therefore, when

$$(m,n) \neq (-m,-n) \pmod K,$$

there exist two nonzero Fourier coefficients, each equal to $K^2/2$, and hence

$$\boxed{\|W\|_2 = \frac{K^2}{2}.} \tag{53}$$

If

$$(m,n) = (-m,-n) \pmod K,$$

the two Fourier modes coincide and their coefficients add up, yielding

$$\boxed{\|W\|_2 = K^2.} \tag{54}$$

Similarly, consider the two-dimensional sine basis function

$$K_{i,j} = \sin\left(\frac{2\pi}{K}(mi+nj)\right), \qquad i,j,m,n = 0,\dots,K-1. \tag{55}$$

Using Euler's identity,

$$K_{i,j} = \frac{1}{2\mathrm{i}} \left[\exp\left(\frac{2\pi\mathrm{i}}{K}(mi + nj) \right) - \exp\left(-\frac{2\pi\mathrm{i}}{K}(mi + nj) \right) \right]. \tag{56}$$

Therefore, its two-dimensional DFT is given by

$$\widehat{K}_{r,s} = \frac{K^2}{2\mathrm{i}} \left[\mathbf{1}_{(r,s)=(m,n)} - \mathbf{1}_{(r,s)=(-m,-n)} \right]. \tag{57}$$

Let $W \in \mathbb{R}^{K^2 \times K^2}$ denote the corresponding circular convolution matrix. Since the convolution matrix is diagonalized by the two-dimensional DFT,

$$W = F^* \Lambda F, \tag{58}$$

where F is the unitary two-dimensional DFT matrix and

$$\Lambda = \operatorname{diag}\left(\widehat{K}_{r,s} \right)_{r,s=0}^{K-1}. \tag{59}$$

Since F is unitary,

$$\|W\|_2 = \|\Lambda\|_2 = \max_{r,s} |\widehat{K}_{r,s}|. \tag{60}$$

For

$$(m, n) \neq (-m, -n) \pmod{K},$$

there exist two nonzero Fourier coefficients with magnitude $K^2/2$. Therefore,

$$\boxed{\|W\|_2 = \frac{K^2}{2}.} \tag{61}$$

When

$$(m, n) = (-m, -n) \pmod{K},$$

the two Fourier modes coincide and cancel due to the opposite signs in the sine representation. Hence the corresponding Fourier coefficient vanishes. □

Lemma E.2. *Let the rotated two-dimensional cosine basis function be sampled on a $K \times K$ grid:*

$$K^{\theta}_{i,j} = \cos\left(\frac{2\pi}{K}(ai + bj) \right), \qquad i, j = 0, \ldots, K-1, \tag{62}$$

where $(a, b) \in \mathbb{R}^2$ denotes the rotated frequency.

If either a or b is not an integer, then the spectral norm of the corresponding circular convolution matrix satisfies

$$\|W_\theta\|_2 < K^2. \tag{63}$$

Proof. Using Euler's formula, the sampled kernel can be written as

$$K^{\theta}_{i,j} = \frac{1}{2} \exp\left(\frac{2\pi\mathrm{i}}{K}(ai + bj) \right) + \frac{1}{2} \exp\left(-\frac{2\pi\mathrm{i}}{K}(ai + bj) \right). \tag{64}$$

The two-dimensional discrete Fourier transform of K^θ is defined as

$$\widehat{K}^{\theta}_{r,s} = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} K^{\theta}_{i,j} \exp\left(-\frac{2\pi\mathrm{i}}{K}(ri + sj) \right). \tag{65}$$

Substituting the expression of $K^{\theta}_{i,j}$ yields

$$\begin{aligned} \widehat{K}^{\theta}_{r,s} =& \frac{1}{2} \sum_{i,j} \exp\left(\frac{2\pi\mathrm{i}}{K}((a - r)i + (b - s)j) \right) \\ &+ \frac{1}{2} \sum_{i,j} \exp\left(-\frac{2\pi\mathrm{i}}{K}((a + r)i + (b + s)j) \right). \end{aligned} \tag{66}$$

Define the finite geometric sum

$$S_K(\alpha) = \sum_{t=0}^{K-1} \exp\left(\frac{2\pi \mathrm{i}}{K}\alpha t\right). \tag{67}$$

For arbitrary real α, it has the closed form

$$S_K(\alpha) = \exp\left(\frac{\pi \mathrm{i}(K-1)}{K}\alpha\right) \frac{\sin(\pi\alpha)}{\sin(\pi\alpha/K)}. \tag{68}$$

Let $\alpha = n \in \mathbb{Z}$. Then the numerator $\sin(\pi n) = 0$. Two cases arise:

1. **If n is a multiple of K**, i.e. $n = mK$ with $m \in \mathbb{Z}$:
   - The denominator $\sin(\pi n/K) = \sin(\pi m) = 0$, so equation 68 gives the indeterminate form $0/0$.
   - Directly from the definition equation 67, each term $\exp(2\pi i\, mt) = 1$, hence
   $$S_K(n) = \sum_{t=0}^{K-1} 1 = K.$$
2. **If n is not a multiple of K**:
   - The denominator $\sin(\pi n/K) \neq 0$, while the numerator is still zero.
   - Thus equation 68 yields
   $$S_K(n) = 0.$$

Therefore, for integer $\alpha = n$,

$$S_K(n) = \begin{cases} K, & n \equiv 0 \pmod K, \\ 0, & n \not\equiv 0 \pmod K. \end{cases} \tag{69}$$

This is precisely the orthogonality property of the discrete Fourier basis.

If $\alpha \notin \mathbb{Z}$, then $\sin(\pi\alpha) \neq 0$ and simultaneously $\sin(\pi\alpha/K) \neq 0$ (since α/K cannot be an integer unless α itself is an integer). Therefore, the closed form equation 68 is always valid and gives a nonzero complex number:

$$S_K(\alpha) = \exp\left(\frac{\pi i(K-1)}{K}\alpha\right) \frac{\sin(\pi\alpha)}{\sin(\pi\alpha/K)}, \qquad \alpha \notin \mathbb{Z}. \tag{70}$$

Its magnitude is

$$|S_K(\alpha)| = \left|\frac{\sin(\pi\alpha)}{\sin(\pi\alpha/K)}\right|, \tag{71}$$

which lies between 0 and K (except at integer multiples of K where it approaches K).

Therefore, the Fourier coefficient can be written as

$$\begin{aligned} \widehat{K}^{\theta}_{r,s} =& \frac{1}{2} S_K(a-r) S_K(b-s) \\ &+ \frac{1}{2} S_K(-a-r) S_K(-b-s). \end{aligned} \tag{72}$$

Let W_θ denote the circular convolution matrix generated by K^θ. Since the circular convolution operator is diagonalized by the two-dimensional DFT,

$$W_\theta = F^* \Lambda_\theta F, \tag{73}$$

where

$$\Lambda_\theta = \mathrm{diag}(\widehat{K}^{\theta}_{r,s}). \tag{74}$$

Because F is unitary,

$$\|W_\theta\|_2 = \max_{r,s} |\widehat{K}^\theta_{r,s}|. \tag{75}$$

When $(a, b) \in \mathbb{Z}^2$, choosing

$$(r, s) = (a, b)$$

gives

$$S_K(a - r)S_K(b - s) = K^2, \tag{76}$$

and hence

$$\|W_\theta\|_2 = \frac{K^2}{2}. \tag{77}$$

However, when either a or b is non-integer, there is no discrete Fourier index satisfying simultaneously

$$r = a, \qquad s = b.$$

Consequently, for all integer (r, s),

$$|S_K(a - r)| < K, \qquad |S_K(b - s)| < K. \tag{78}$$

Therefore, the Fourier energy cannot be concentrated at a single discrete frequency. Instead, it is distributed among multiple Fourier coefficients due to spectral leakage. Hence the maximum Fourier coefficient is strictly smaller than the integer-frequency case:

$$\|W_\theta\|_2 < K^2. \tag{79}$$

This reduction is caused by the mismatch between the rotated continuous Fourier frequency and the fixed discrete sampling grid. □

Lemma E.3. *Let $K \geq 1$ be an integer, and let $(a, b) \in \mathbb{R}^2$ be a rotated frequency. Define the sampled rotated two-dimensional sine basis on a $K \times K$ grid by*

$$K^\theta_{i,j} = \sin\left(\frac{2\pi}{K}(ai + bj)\right), \qquad i, j = 0, \ldots, K - 1. \tag{80}$$

Let W_θ be the circular convolution matrix generated by K^θ. If at least one of a and b is not an integer, then

$$\|W_\theta\|_2 < K^2. \tag{81}$$

Proof. By Euler's formula,

$$K^\theta_{i,j} = \frac{1}{2\mathrm{i}} \exp\left(\frac{2\pi\mathrm{i}}{K}(ai + bj)\right) - \frac{1}{2\mathrm{i}} \exp\left(-\frac{2\pi\mathrm{i}}{K}(ai + bj)\right). \tag{82}$$

The two-dimensional discrete Fourier transform of K^θ is

$$\widehat{K}^\theta_{r,s} = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} K^\theta_{i,j} \exp\left(-\frac{2\pi\mathrm{i}}{K}(ri + sj)\right). \tag{83}$$

Substituting the Euler representation gives

$$\widehat{K}^\theta_{r,s} = \frac{1}{2\mathrm{i}} S_K(a - r)S_K(b - s) - \frac{1}{2\mathrm{i}} S_K(-a - r)S_K(-b - s), \tag{84}$$

where

$$S_K(\alpha) = \sum_{t=0}^{K-1} \exp\left(\frac{2\pi\mathrm{i}}{K}\alpha t\right) = \exp\left(\frac{\pi\mathrm{i}(K - 1)}{K}\alpha\right) \frac{\sin(\pi\alpha)}{\sin(\pi\alpha/K)}. \tag{85}$$

For integer n,

$$S_K(n) = \begin{cases} K, & n \equiv 0 \pmod K, \\ 0, & n \not\equiv 0 \pmod K. \end{cases} \tag{86}$$

For non-integer α,

$$|S_K(\alpha)| = \left| \frac{\sin(\pi\alpha)}{\sin(\pi\alpha/K)} \right| < K, \tag{87}$$

since $|S_K(\alpha)| \le K$, with equality only if all summands have the same phase, which requires $\alpha \in \mathbb{Z}$.

Let F denote the unitary two-dimensional DFT matrix. Since circular convolution is diagonalized by the DFT,

$$W_\theta = F^* \Lambda_\theta F, \qquad \Lambda_\theta = \operatorname{diag}\left(\widehat{K}^\theta_{r,s}\right). \tag{88}$$

Because F is unitary,

$$\|W_\theta\|_2 = \max_{r,s} |\widehat{K}^\theta_{r,s}|. \tag{89}$$

Assume that at least one of a and b is not an integer. Then for every integer pair (r, s), at least one of the numbers $a-r$, $b-s$ is non-integer, and at least one of the numbers $-a-r$, $-b-s$ is non-integer. Hence

$$|S_K(a-r)S_K(b-s)| < K^2, \qquad |S_K(-a-r)S_K(-b-s)| < K^2.$$

Therefore,

$$|\widehat{K}^\theta_{r,s}| \le \frac{1}{2}\left(|S_K(a-r)S_K(b-s)| + |S_K(-a-r)S_K(-b-s)|\right) < K^2. \tag{90}$$

Taking the maximum over all (r, s) yields

$$\|W_\theta\|_2 = \max_{r,s} |\widehat{K}^\theta_{r,s}| < K^2. \tag{91}$$

This strict inequality is caused by the mismatch between the rotated continuous sine frequency and the fixed discrete sampling grid, which leads to spectral leakage. □

We now provide a detailed mathematical derivation and the concrete computational steps for evaluating the discrete Fourier transform (DFT) of the sampled basis functions.

Let the convolutional kernel be of size $K \times K$ with pixel spacing $h = 1$. The discrete sampling points are

$$x_1 = mh, \qquad x_2 = nh, \qquad m, n = 0, 1, \ldots, K-1.$$

In practice, the coordinate origin is often shifted to the center of the kernel; however, since the DFT implicitly assumes periodic extension, such a shift only introduces a linear phase factor and does not affect the magnitude spectrum. The derivation below adopts the standard signal processing convention for the DFT, i.e., with a negative exponent $e^{-j2\pi/K}$.

Define the centered frequency offsets

$$\tilde{k} = k - \lfloor K/2 \rfloor, \qquad \tilde{l} = l - \lfloor K/2 \rfloor.$$

Then the discretized basis functions are given by

$$K_c[m, n] = \Omega[m, n] \cdot \cos\left(\frac{2\pi}{K}(\tilde{k}m + \tilde{l}n)\right),$$

$$K_s[m, n] = \Omega[m, n] \cdot \sin\left(\frac{2\pi}{K}(\tilde{k}m + \tilde{l}n)\right),$$

where $\Omega[m, n]$ denotes the discrete samples of the radial mask function $\Omega(x)$ on the $K \times K$ grid.

The two-dimensional DFT is defined as

$$\widehat{K}[u, v] = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} K[m, n]\, e^{-j\frac{2\pi}{K}(um+vn)},$$

where $u, v = 0, 1, \ldots, K-1$ are the frequency-domain indices.

Cosine basis ϕ^c_{kl}: Using Euler's formula,

$$\cos(\alpha) = \frac{1}{2}\left(e^{j\alpha} + e^{-j\alpha}\right), \qquad \alpha = \frac{2\pi}{p}(\tilde{k}m + \tilde{l}n),$$

and substituting into the DFT definition yields

$$\begin{aligned}\widehat{K}_c[u,v] &= \frac{1}{2}\sum_{m,n}\Omega[m,n]\left(e^{j\frac{2\pi}{K}(\tilde{k}m+\tilde{l}n)} + e^{-j\frac{2\pi}{K}(\tilde{k}m+\tilde{l}n)}\right)e^{-j\frac{2\pi}{K}(um+vn)} \\ &= \frac{1}{2}\left[\sum_{m,n}\Omega[m,n]\,e^{-j\frac{2\pi}{K}\left((u-\tilde{k})m+(v-\tilde{l})n\right)} + \sum_{m,n}\Omega[m,n]\,e^{-j\frac{2\pi}{K}\left((u+\tilde{k})m+(v+\tilde{l})n\right)}\right].\end{aligned}$$

Since $\Omega(x) = 1$ for $|x| \le r$, most grid points have value 1. Only the grid points in the boundary region $r < |x| < r + h$ have values in $(0, 1)$, while $\Omega(x) = 0$ for $|x| \ge r + h$, so

$$\widehat{K}_c[u,v] \le \frac{1}{2}\left[\sum_{m,n} e^{-j\frac{2\pi}{K}\left((u-\tilde{k})m+(v-\tilde{l})n\right)} + \sum_{m,n} e^{-j\frac{2\pi}{K}\left((u+\tilde{k})m+(v+\tilde{l})n\right)}\right]$$

Therefore, by Lemma E.1, the spectral norm of the corresponding circular convolution matrix can be obtained. Similarly, the analogous result can be derived for the sine function. Combining this with the result of Lemma E.2 and Lemma E.3 in the rotated case, we obtain the sum of the spectral norms of the Fourier basis as follows:

$$D_l := \sqrt{\sum_{i=1}^{G_l}\sum_{t=1}^{N_l}(D^l_{t,i})^2} \le \sqrt{K^4 + \frac{K^4}{4}(K^2-1) + \frac{K^4}{4}(K^2-1) + 2K^6(|G|-1)} \le \sqrt{2|G|}K^3$$

E.2 HARMONIC BASIS

We directly sample the circular harmonic basis functions $\phi_{jk}(r,\phi) = \tau_j(r)e^{ik\phi}$ on a finite two-dimensional grid and compute their 2D discrete Fourier transform (DFT). The complete discretization, DFT formula, and computational steps are provided below.

Let the convolutional kernel be of size $K \times K$ with pixel spacing $h = 1$. The Cartesian coordinates of the discrete sampling points are

$$x_1 = mh, \qquad x_2 = nh, \qquad m, n = 0, 1, \ldots, K-1.$$

We typically place the coordinate origin at the center of the kernel, i.e.

$$\bar{m} = m - \frac{K-1}{2}, \qquad \bar{n} = n - \frac{K-1}{2},$$

so that the polar coordinates are

$$r_{m,n} = h\sqrt{\bar{m}^2 + \bar{n}^2}, \qquad \phi_{m,n} = \operatorname{atan2}(\bar{n}, \bar{m}).$$

(The translation invariance of the DFT implies that shifting the origin only introduces a linear phase in the frequency domain and does not affect the magnitude spectrum. Hence one may also place the origin at the top-left corner, but the centered formulation is more intuitive.)

The discretized basis functions are then given by

$$\phi_{jk}[m,n] = \tau_j(r_{m,n})\,e^{ik\phi_{m,n}},$$

where the radial profile is

$$\tau_j(r) = \exp\left(-\frac{(r-\mu_j)^2}{2\sigma^2}\right), \qquad \mu_j = j.$$

Adopting the standard signal processing convention (negative exponent, no normalization factor), the 2D DFT is defined as

$$\widehat{\phi}_{jk}[u,v] = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} \phi_{jk}[m,n]\, e^{-j\frac{2\pi}{K}(um+vn)}, \qquad u,v = 0,1,\ldots,K-1.$$

Substituting the expression for $\phi_{jk}[m,n]$ into the DFT yields

$$\widehat{\phi}_{jk}[u,v] = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} \tau_j\Big(h\sqrt{(m-c)^2+(n-c)^2}\Big)\; e^{ik\ \mathrm{atan2}(n-c,\,m-c)}\, e^{-j\frac{2\pi}{K}(um+vn)},$$

where $c = \frac{K-1}{2}$ is the center coordinate. (If the grid origin is at the top-left corner, the center offset c may be set to zero; however, the coordinates are then not centered, and the phase terms differ, but the summation formula remains valid.)

From the DFT definition and the triangle inequality, we have

$$\begin{aligned}\left|\widehat{\phi}_{jk}[u,v]\right| &= \left|\sum_{m,n} \tau_j(r_{m,n}) e^{ik\phi_{m,n}} e^{-j\frac{2\pi}{p}(um+vn)}\right| \\ &\le \sum_{m,n} \tau_j(r_{m,n}) \cdot |e^{ik\phi}| \cdot |e^{-j(\cdot)}| = \sum_{m,n} \tau_j(r_{m,n}).\end{aligned}$$

Since $|e^{ik\phi}| = 1$, the magnitude of every frequency component does not exceed the total sum of the radial profile over the grid:

$$\boxed{0 \le \left|\widehat{\phi}_{jk}[u,v]\right| \le S_j} \quad \text{where} \quad S_j = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} \tau_j(r_{m,n}).$$

Because $\tau_j(r) \le 1$ (the Gaussian function attains a maximum of 1), we have $S_j \le K^2$. When the Gaussian center μ_j lies inside the grid and σ is small, S_j roughly equals the effective number of pixels covered by the Gaussian kernel; if μ_j is far from the grid or σ is very large, S_j becomes smaller.

F Experiments

F.1 Experimental on image classification

Datasets. (1) Rotated-MNIST dataset Larochelle et al. (2007): This dataset is a rotated variant of the MNIST dataset, in which handwritten digit images are presented in arbitrary orientations. Each image belongs to one of 10 classes. (2) CIFAR-10 dataset: This is an image classification dataset containing 60,000 32×32 color images divided into 10 classes, with 6,000 images per class. The training set comprises 50,000 images and the test set comprises 10,000 images, covering categories such as airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. In the experiments, training images are randomly augmented with 90° or 180° rotations, while no data augmentation is applied to the test set.

Experimental Setup. The rotation-equivariant convolutional networks are implemented based on the frameworks of Cohen & Welling (2016) and Xie et al. (2022), using discrete planar rotation groups. Input images are converted into floating-point tensors in $[0,1]$ via $\mathrm{ToTensor}$, with each channel treated as a trivial scalar field. The backbone consists of stacked steerable convolutional layers with kernel sizes of 7×7 and 5×5, followed by equivariant batch normalization and ReLU activation. The feature dimensions are gradually increased across layers, while anti-aliasing pooling and group pooling are employed to reduce spatial resolution and aggregate rotation responses. A global average pooling operation and a fully connected classifier are then used for final prediction. The networks are optimized with Adam using learning rates of 5×10^{-3} or 8×10^{-3} and cross-entropy loss, with training and testing batch sizes of 128 and 100, respectively. We record training error, test error, and empirical generalization error without filter replacement.

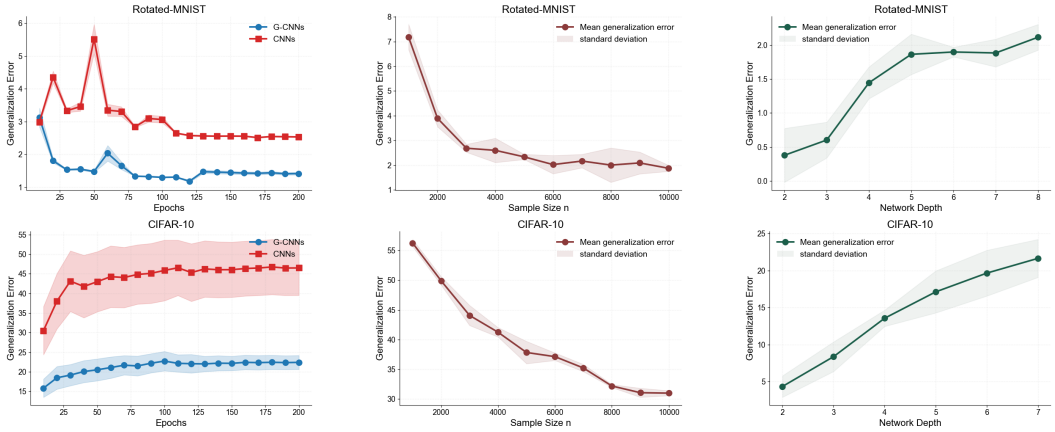


Figure 2: Generalization analysis of G-CNNs on Rotated-MNIST and CIFAR-10. Top row: Rotated-MNIST results; bottom row: CIFAR-10 results. From left to right: comparison between CNNs and G-CNNs, generalization error versus training sample size, generalization error versus network depth.

F.2 Experimental Results on Image Classification

Generalization Analysis of G-CNNs. To validate the theoretical results in Theorem B.6, we investigate the effects of symmetry, network depth L, sample size n, group order $|G|$, and kernel size K on the generalization performance of G-CNNs. Unless otherwise specified, all experiments keep the model architecture, parameter configuration, and training protocol fixed, with only the investigated factor varied.

We first compare G-CNNs with standard CNNs on Rotated-MNIST and CIFAR-10, using identical parameter counts and hyper-parameters. As shown in the first column of Figure 2, G-CNNs achieve lower and more stable generalization errors, validating the benefit of symmetry priors. We further study the effects of sample size and network depth. The results in the second and third columns show that generalization error decreases with increasing sample size and grows with network depth, consistent with the theoretical scaling of $\mathcal{O}(1/\sqrt{n})$ and $\mathcal{O}(L \ln L)$ in Theorem B.6.

We further investigate the effects of group order $|G|$ and kernel size K. For different group sizes ($|G| = 2, 4, 6, 8$), the total number of feature channels is fixed by setting the six-layer backbone channels to $(96, 192, 192, 384, 384, 192)$. As shown in Table 7, larger groups achieve better generalization performance, consistent with Theorem B.6. For kernel size analysis, we consider $K = 3, 5, 7, 9$ and rescale convolution kernels to maintain identical per-layer parameter bounds B_j. The results in Table 8 demonstrate that smaller kernels provide better generalization performance, agreeing with the theoretical dependence of the bound on K.

Table 7: Error comparison for different group sizes $|G|$

	Generalization Error	
$\|G\|$	Rotated-MNIST	CIFAR-10
2	2.58 ± 0.07	38.82±4.60
4	1.55 ± 0.20	26.15±2.41
6	1.27 ± 0.16	25.38±0.93
8	**1.11 ± 0.09**	**24.05±1.09**

Table 8: Error comparison for different kernel sizes K

	Generalization Error	
K	Rotated-MNIST	CIFAR-10
3	**1.628 ± 0.179**	**3.530 ± 0.082**
5	1.668 ± 0.165	3.655 ± 0.186
7	1.738 ± 0.127	3.724 ± 0.069
9	1.858 ± 0.111	3.860 ± 0.114

F.3 Experimental on Image Super-Resolution

To further evaluate the applicability of PG-CNNs in low-level vision, we apply the proposed equivariant convolution module to single-image super-resolution (SISR), which aims to recover high-resolution images from low-resolution inputs.

Datasets and Degradation Models. Following standard SISR protocols (Lim et al., 2017; Zhang et al., 2018b;a), we train all models on 800 images from the DIV2K dataset and evaluate them on Urban100 (Huang et al., 2015), B100 (Martin et al., 2001), Set14 (Zeyde et al., 2010), and Set5 (Bevilacqua et al., 2012). All experiments adopt the standard bicubic degradation model.

Network Architectures and Training Settings. We adopt three representative SISR architectures, including EDSR (Lim et al., 2017), RDN (Zhang et al., 2018b), and RCAN (Zhang et al., 2018a), as baselines. The proposed PG-CNN variants are constructed by replacing standard convolution layers with F-Conv layers using a default kernel size of 5×5. The EDSR, RDN, and RCAN models are configured following their original settings, with 16 residual blocks for EDSR, 16 residual dense blocks for RDN, and 10 residual-in-residual groups for RCAN.

All models are trained using Adam without weight decay for 150 epochs, with batch size 16 and an initial learning rate of 2×10^{-4}, which is reduced at the 100th and 130th epochs. The training patch sizes are 96, 64, and 96 for EDSR, RDN, and RCAN under $2\times$ super-resolution, respectively, and doubled for $4\times$ super-resolution. To evaluate the intrinsic benefit of equivariant convolutions, baseline experiments are first conducted without data augmentation. Additional experiments with standard augmentation, including random rotations of 90°, 180°, 270°, and horizontal flipping, are further performed for comparison.

As a supplementary experiment, we also investigate the effect of explicitly constraining the convolutional kernel parameter space. According to Table 9, imposing masks on the convolutional kernels deteriorates model performance. This result suggests that excessively restricting the kernel parameter space may reduce the effective representation capacity of the network, highlighting the importance of balancing parameter constraints with sufficient expressive power.

Table 9: Weighted mean $\pm$ standard deviation of quantitative results on benchmark datasets. Statistics are calculated according to the number of images in each dataset.

Scale	Method	$\Delta L_1 \downarrow$	PSNR (dB) $\uparrow$	SSIM $\uparrow$
$\times 2$	EDSR-FConv-k3	1.77 ± 0.45	31.77 ± 1.08	0.91 ± 0.01
	EDSR-FConv-k3-NoMask	$\mathbf{1.71 \pm 0.39}$	$\mathbf{32.10 \pm 0.99}$	$\mathbf{0.91 \pm 0.01}$
$\times 4$	EDSR-FConv-k3	2.84 ± 1.12	26.60 ± 1.31	0.75 ± 0.02
	EDSR-FConv-k3-NoMask	$\mathbf{2.75 \pm 1.01}$	$\mathbf{26.87 \pm 1.24}$	$\mathbf{0.76 \pm 0.03}$