A PROOFS

We begin with a simple lemma detailing the Fourier space end-to-end predictor for a G-CNN.

Lemma A.1 (G-CNN in Fourier space)**.** *A G-CNN given by $\langle \boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}, \boldsymbol{w}_L \rangle$ is equivalent to $\langle \boldsymbol{x}, \boldsymbol{w}_L \star \boldsymbol{w}_{L-1}^- \star \cdots \star \boldsymbol{w}_1^- \rangle$, or in Fourier space to $\frac{1}{|G|} \langle \widehat{\boldsymbol{x}}, \widehat{\boldsymbol{w}}_L \cdots \widehat{\boldsymbol{w}}_1 \rangle_M$. Moreover,*

Proof.

$$\begin{aligned} N(x) &= \langle \boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}, \boldsymbol{w}_L \rangle \\ &= \frac{1}{|G|} \langle \mathcal{F}_M(\boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}), \mathcal{F}_M \boldsymbol{w}_L \rangle_M \\ &= \frac{1}{|G|} \operatorname{tr}[\mathcal{F}_M(\boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}) \widehat{\boldsymbol{w}}_L^\dagger] \\ &= \frac{1}{|G|} \operatorname{tr}[\mathcal{F}_M(\boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-2}) \widehat{\boldsymbol{w}}_{L-1}^\dagger \widehat{\boldsymbol{w}}_L^\dagger] \\ &= \frac{1}{|G|} \operatorname{tr}[\widehat{\boldsymbol{x}} \widehat{\boldsymbol{w}}_1^\dagger \cdots \widehat{\boldsymbol{w}}_{L-1}^\dagger \widehat{\boldsymbol{w}}_L^\dagger] \\ &= \frac{1}{|G|} \langle \widehat{\boldsymbol{x}}, \widehat{\boldsymbol{w}}_L \cdots \widehat{\boldsymbol{w}}_1 \rangle_M \end{aligned} \tag{A.1}$$

To see the stated equality in real space, observe that $\widehat{\boldsymbol{w}_1^-} = \widehat{\boldsymbol{w}_1}^\dagger$ by definition. Thus,

$$\mathcal{F}_M(\boldsymbol{w}_L \star \boldsymbol{w}_{L-1}^- \star \cdots \star \boldsymbol{w}_1^-) = \widehat{\boldsymbol{w}}_L \cdots \widehat{\boldsymbol{w}}_1$$

□

A.1 ABELIAN

The proof of Proposition 5.2 is included below.

First, recall the primary theorem of Yun et al. (2020):

Theorem A.2 (paraphrased from Yun et al. (2020))**.** *If there exists $\lambda > 0$ such that the initial directions $\bar{\boldsymbol{v}}_1, \ldots, \bar{\boldsymbol{v}}_L$ of the network parameters satisfy $|[\mathcal{F}\bar{\boldsymbol{v}}_\ell]_j|^2 - |[\mathcal{F}\bar{\boldsymbol{v}}_L]_j|^2 \geq \lambda$ for all $\ell \in [L-1]$ and $j \in [m]$, i.e. of the Fourier transform magnitudes of the initial directions look sufficiently different pointwise (which is likely for e.g. a random initialization), then $\boldsymbol{\beta}(\Theta(t))$ converges in a direction that aligns with $\boldsymbol{S}^T \boldsymbol{\rho}_\infty$ where $\boldsymbol{\rho}_\infty \in \mathbb{C}^m$ denotes a stationary point of the following optimization program:*

$$\min_{\boldsymbol{\rho} \in \mathbb{C}^m} \quad \|\boldsymbol{\rho}\|_{2/L} \quad \textit{s.t.} \quad y_i \boldsymbol{x}_i^T \mathcal{F}^T v \rho \geq 1, \forall i \in [n] \tag{A.2}$$

Since $\boldsymbol{S} = \mathcal{F}$ is invertible, then in fact $\beta(\Theta(t))$ converges in a direction that aligns with a stationary point $\boldsymbol{z}_\infty$ of the following optimization program:

$$\min_{\boldsymbol{z} \in \mathbb{C}^m} \quad \|\mathcal{F}\boldsymbol{z}\|_{2/L} \quad \textit{s.t.} \quad y_i \boldsymbol{x}_i^T \boldsymbol{z} \geq 1, \forall i \in [n] \tag{A.3}$$

We proceed by showing that abelian G-CNNs can be written as a vector of parameters contracted (according to tensor operations) with an orthogonally decomposable data tensor, which is the primary condition for Theorem A.2 to hold.

For convenience, we restate Proposition A.3 before detailing the proof.

Proposition A.3 (paraphrased from Yun et al. (2020))**.** *Let $\mathbf{M}(\boldsymbol{x})$ be a function that maps data $\boldsymbol{x} \in \mathbb{R}^d$ to a data tensor $\mathbf{M}(\boldsymbol{x}) \in \mathbb{R}^{k_1 \times k_2 \times \cdots \times k_L}$. The data input into an L-layer tensorized neural network can be written in the form of an **orthogonally decomposable data tensor** if there exists a full column rank matrix $\boldsymbol{S} \in \mathbb{C}^{m \times d}$ and semi-unitary matrices $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_L \in \mathbb{C}^{k_\ell \times m}$ where $d \leq m \leq \min_\ell k_\ell$ such that $\mathbf{M}(\boldsymbol{x})$ can be written as:*

$$\mathbf{M}(\boldsymbol{x}) = \sum_{j=1}^{m} [\boldsymbol{S}\boldsymbol{x}]_j \left([\boldsymbol{U}_1]_{\cdot,j} \otimes [\boldsymbol{U}_2]_{\cdot,j} \otimes \cdots \otimes [\boldsymbol{U}_L]_{\cdot,j} \right) \tag{A.4}$$

and such that the network output is the tensor multiplication between $\mathbf{M}(x)$ and each layer's parameters:

$$\begin{aligned}\mathrm{NN}(\boldsymbol{x};\Theta) &= \mathbf{M}(\boldsymbol{x}) \cdot (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L) \\ &= \sum_{i_1=1}^{d} \cdots \sum_{i_L=1}^{d} \mathbf{M}(\boldsymbol{x})_{i_1 \ldots i_L} (\boldsymbol{W}_1)_{i_1} \cdot \cdots \cdot (\boldsymbol{W}_L)_{i_L}\end{aligned}$$

For Proposition A.3. By direct manipulation:

$$\begin{aligned}f(\boldsymbol{x};\Theta) &= \mathbf{M}(\boldsymbol{x}) \cdot (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L) \\ &= \sum_{i_1=1}^{d} \cdots \sum_{i_L=1}^{d} \mathbf{M}(\boldsymbol{x})_{i_1 \ldots i_L} (\boldsymbol{v}_1)_{i_1} \cdot \cdots \cdot (\boldsymbol{v}_L)_{i_L} \\ &= \sum_{i_1=1}^{d} \cdots \sum_{i_L=1}^{d} \left(\sum_{j=1}^{d} [\boldsymbol{S}\boldsymbol{x}]_j \left([\boldsymbol{U}_1]_{\cdot,j} \otimes [\boldsymbol{U}_2]_{\cdot,j} \otimes \cdots \otimes [\boldsymbol{U}_L]_{\cdot,j} \right) \right)_{i_1 \ldots i_L} (\boldsymbol{v}_1)_{i_1} \cdot \cdots \cdot (\boldsymbol{v}_L)_{i_L} \\ &= \sum_{i_1=1}^{d} \cdots \sum_{i_L=1}^{d} \left(\sum_{j=1}^{d} [\boldsymbol{S}\boldsymbol{x}]_j \left([\boldsymbol{U}_1]_{i_1,j} [\boldsymbol{U}_2]_{i_2,j} \cdots [\boldsymbol{U}_L]_{i_L,j} \right) \right) (\boldsymbol{v}_1)_{i_1} \cdot \cdots \cdot (\boldsymbol{v}_L)_{i_L} \\ &= \sum_{j=1}^{d} [\boldsymbol{S}\boldsymbol{x}]_j \left([\boldsymbol{U}_1^T \boldsymbol{v}_1]_{i_1} [\boldsymbol{U}_2^T \boldsymbol{v}_2]_{i_2} \cdots [\boldsymbol{U}_L^T \boldsymbol{v}_L]_{i_L} \right) \\ &= d^{\frac{L-1}{2}} \sum_{j=1}^{d} [\mathcal{F}\boldsymbol{x}]_j \left([\overline{\mathcal{F}}\boldsymbol{v}_1]_{i_1} [\overline{\mathcal{F}}\boldsymbol{v}_2]_{i_2} \cdots [\overline{\mathcal{F}}\boldsymbol{v}_L]_{i_L} \right) \\ &= d^{\frac{L-1}{2}} \sum_{j=1}^{d} [\mathcal{F}\boldsymbol{x}]_j \left([\overline{\mathcal{F}\boldsymbol{v}_1}]_{i_1} [\overline{\mathcal{F}\boldsymbol{v}_2}]_{i_2} \cdots [\overline{\mathcal{F}\boldsymbol{v}_L}]_{i_L} \right) \\ &= d^{\frac{L-1}{2}} \sum_{j=1}^{d} [\mathcal{F}\boldsymbol{x}]_j \overline{\left([\mathcal{F}\boldsymbol{v}_1]_{i_1} [\mathcal{F}\boldsymbol{v}_2]_{i_2} \cdots [\mathcal{F}\boldsymbol{v}_L]_{i_L} \right)} \\ &= \langle \mathcal{F}\boldsymbol{x}, \mathcal{F}\boldsymbol{v}_1 \odot \cdots \odot \mathcal{F}\boldsymbol{v}_L \rangle \\ &= \langle \mathcal{F}\boldsymbol{x} \odot \overline{\mathcal{F}\boldsymbol{v}_1}, \mathcal{F}\boldsymbol{v}_2 \odot \cdots \odot \mathcal{F}\boldsymbol{v}_L \rangle \\ &= \langle \mathcal{F}(\boldsymbol{x} \star \boldsymbol{v}_1) \odot \overline{\mathcal{F}\boldsymbol{v}_2}, \mathcal{F}\boldsymbol{v}_3 \odot \cdots \odot \mathcal{F}\boldsymbol{v}_L \rangle \\ &= \langle \mathcal{F}(\boldsymbol{x} \star \boldsymbol{v}_1 \star \boldsymbol{v}_2) \odot \overline{\mathcal{F}\boldsymbol{v}_3}, \mathcal{F}\boldsymbol{v}_4 \odot \cdots \odot \mathcal{F}\boldsymbol{v}_L \rangle \\ &= \langle \mathcal{F}(\boldsymbol{x} \star \boldsymbol{v}_1 \star \boldsymbol{v}_2 \star \cdots \star \boldsymbol{v}_{L-1}), \mathcal{F}\boldsymbol{v}_L \rangle \\ &= \langle \boldsymbol{x} \star \boldsymbol{v}_1 \star \cdots \star \boldsymbol{v}_{L-1}, \boldsymbol{v}_L \rangle\end{aligned}$$

Here, we have used that the filters are real-valued. □

Note that Theorem 5.3 then merely requires that $||\mathcal{F}^{-T}\boldsymbol{z}|| = ||\overline{\mathcal{F}}\boldsymbol{z}|| = ||\mathcal{F}\boldsymbol{z}||$, for real-valued $\boldsymbol{z}$.

A.1.1 Fundamental theorem of finite abelian groups

While the proof in the previous section is complete and correct, intuition (and/or alternate analysis) for abelian groups is aided by the important fact that all finite abelian groups are a direct product of cyclic groups.

Theorem A.4 (Fundamental theorem of finite abelian groups (Dummit & Foote, 2004))**.** *Any finite abelian group is a direct product of a finite number of cyclic groups whose orders are prime powers uniquely determined by the group.*

Given a decomposition of an abelian group G into k cyclic groups $C_{d_1} \times \cdots \times C_{d_k}$, one can easily construct the group Fourier transform as a Kronecker product of discrete Fourier transform matrices

which are the group Fourier transforms of the respective cyclic groups.

$$G = C_{d_1} \times \cdots \times C_{d_k} \implies \mathcal{F} = \bigotimes_{i=1}^{k} \mathcal{F}_{d_i}, \tag{A.5}$$

where $\bigotimes$ denotes the Kronecker product over matrices and $\mathcal{F}_d$ is the standard (unitary) discrete Fourier transform matrix of dimension d defined as

$$\mathcal{F}_d = \frac{1}{\sqrt{d}} \begin{bmatrix} \omega_d^{0\cdot 0} & \omega_d^{0\cdot 1} & \cdots & \omega_d^{0\cdot(d-1)} \\ \omega_d^{1\cdot 0} & \omega_d^{1\cdot 1} & \cdots & \omega_d^{1\cdot(d-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_d^{(d-1)\cdot 0} & \omega_d^{(d-1)\cdot 1} & \cdots & \omega_d^{(d-1)\cdot(d-1)} \end{bmatrix}, \qquad \omega_d = e^{\frac{-2\pi i}{d}}. \tag{A.6}$$

From this result, it is clear that the desired properties of the Fourier transform and convolution hold.

A.2 Non-abelian

Theorem A.5. *Consider a classification task with ground-truth linear predictor $\boldsymbol{\beta}$, trained via a linear G-CNN architecture $\mathrm{NN}(\boldsymbol{x}) = \langle \boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}, \boldsymbol{w}_L \rangle$ (see Section 5 for architecture details) with $L \geq 2$ layers under the exponential loss. Then for almost any datasets $\{\boldsymbol{x}_i, y_i\}$ separable by $\boldsymbol{\beta}$, any bounded sequence of step sizes η_t, and almost all initializations, suppose that:*

- *The loss $\mathcal{L}(\boldsymbol{W})$ converges to 0*
- *The gradients with respect to the end-to-end linear predictor, $\nabla_{\boldsymbol{\beta}} \mathcal{L}(\mathcal{P}(\boldsymbol{W}^t))$, converge in direction as $t \to \infty$*
- *The iterates $\boldsymbol{W}^t$ themselves converge in direction as $t \to \infty$ to a separator $\mathcal{P}(\boldsymbol{W}^t)$ with positive margin*

When $L = 2$, we need an additional technical assumption, Assumption A.7. Then, the resultant linear predictor $\widehat{\boldsymbol{\beta}}^{\infty}$ is a positive scaling of a first order stationary point of the optimization problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_{2/L}^{(S)} \quad \textit{s.t.} \quad \forall n, y_n \langle \boldsymbol{\beta}_n, \boldsymbol{x}_n \rangle_M \geq 1 \tag{A.7}$$

In this section, we prove the non-abelian case, Theorem A.5. The proof of our result proceeds according to the following outline:

1. By applying a general result of Gunasekar et al. (2018b), Theorem A.6, we characterize the implicit regularization in the full space of parameters, $\boldsymbol{W}$ (in contrast to the end-to-end linear predictor β), as the stationary point of an optimization problem Equation A.9 in $\boldsymbol{W}$.
2. Separately, we define a *distinct* optimization problem, Equation A.7 in $\boldsymbol{\beta}$. The goal is to demonstrate that stationary points of Equation A.9 are a subset of the stationary points of Equation A.7.
3. The *necessary* KKT conditions for Equation A.9 characterize its stationary points. Using this characterization, we show that the *sufficient* KKT optimality conditions for Equation A.7 are in fact also satisfied for the corresponding end-to-end predictor. Thus, we show that for any stationary point $\boldsymbol{W}^{\dagger}$ of Equation A.9, the linear predictor $\mathcal{P}(\boldsymbol{W}^{\dagger})$ is a stationary point of Equation A.7.

First, recall that Gunasekar et al. (2018b) prove the following general result about the implicit regularization of any homogeneous polynomial parametrization:

Theorem A.6 (Homogeneous polynomial parametrization, Theorem 4 of Gunasekar et al. (2018b))**.** *Let W be the concatenation of all (real-valued) parameters $\boldsymbol{W}_i$. For any homogeneous polynomial map $\mathcal{P} : \mathbb{R}^P \to \mathbb{R}^{|G|}$ from parameters $\boldsymbol{W}_i \in \mathbb{R}^P$ to linear predictors, almost all datasets*

$\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ separable by the ground truth predictor $\boldsymbol{\beta} := \{\mathcal{P}(\boldsymbol{W}) : \boldsymbol{W} \in \mathbb{R}^P\}$, almost all initializations $\boldsymbol{W}^0$, and any bounded sequence of step sizes $\{\eta_t\}_t$, consider the gradient descent updates:

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta_t \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^t) = \boldsymbol{W}^t - \eta_t \nabla_{\boldsymbol{W}} \mathcal{P}(\boldsymbol{W}^t) \nabla_\beta \mathcal{L}(\mathcal{P}(\boldsymbol{W}^t)) \tag{A.8}$$

Suppose furthermore that the exponential loss converges to zero, that the gradients $\nabla_\beta \mathcal{L}(\boldsymbol{\beta}^t)$ converge in direction, and that the iterates W^t themselves converge in direction to yield a separator with positive margin. Then, the limit direction of the parameters $\overline{\boldsymbol{W}}^\infty = \lim_{t\to\infty} \frac{\boldsymbol{W}^{(t)}}{\|\boldsymbol{W}^{(t)}\|_2}$ is a positive scaling of a first order stationary point of the following optimization problem:

$$\min_{\boldsymbol{W}\in\mathbb{R}^P} \|\boldsymbol{W}\|_2^2 \quad \textit{s.t.} \quad \forall n, y_n \langle \boldsymbol{x}_n, \mathcal{P}(\boldsymbol{W})\rangle \geq 1. \tag{A.9}$$

To keep track of constant factors, let $\boldsymbol{W}^\infty = \tau \overline{\boldsymbol{W}}^\infty$ denote the first order stationary point itself. Furthermore, let $\boldsymbol{W}_i^\infty$ denote the individual layers (or parameter blocks) comprising $\boldsymbol{W}^\infty$, and similarly let $\overline{\boldsymbol{W}}_i^\infty$ denote the individual layers comprising $\overline{\boldsymbol{W}}^\infty$. We then have via the KKT conditions that:

$$\begin{aligned} &\exists \{\alpha_n : \alpha_n \geq 0\}_{n=1}^N \textit{ s.t. } \alpha_n = 0 \;\; \textit{if} \;\; y_n \langle \boldsymbol{x}_n, \mathcal{P}(\boldsymbol{W}^\infty)\rangle > 1 \\ &\boldsymbol{W}_i^\infty = \nabla_{\boldsymbol{W}_i} \mathcal{P}(\boldsymbol{W}^\infty) \left[\sum_n \alpha_n y_n \boldsymbol{x}_n\right] = \nabla_{\boldsymbol{W}_i} \left\langle \mathcal{P}(\boldsymbol{W}^\infty), \sum_n \alpha_n y_n \boldsymbol{x}_n \right\rangle \end{aligned} \tag{A.10}$$

While this is an interesting result alone, the goal of implicit regularization is to characterize the final linear predictor (which is some function $\mathcal{P}$ of the complete parametrization W). To that end, consider the following optimization problem in $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \|\widehat{\boldsymbol{\beta}}\|_{2/L}^{(S)} \quad \text{s.t.} \quad \forall n, y_n \langle \boldsymbol{\beta}_n, x_n\rangle \geq 1 \tag{A.11}$$

We will leverage the *necessary* KKT conditions from Equation A.10 to show that first-order stationary points of Equation A.9 are (up to a scaling) also first-order stationary points of Equation A.7, using the *sufficient* KKT conditions for Equation A.7.

Using standard KKT sufficiency conditions, the first-order stationary points of Equation A.7 are those vectors $\boldsymbol{\beta}$ such that there exist $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ satisfying:

1. **Feasibility:** $\forall n, y_n \langle \boldsymbol{\beta}, \boldsymbol{x}_n\rangle \geq 1$ and $\tilde{\alpha}_i \geq 0 \;\; \forall i$
2. **Complementary slackness:** $\forall i, \tilde{\alpha}_i = 0$ if $y_n \langle \beta, \boldsymbol{x}_n\rangle > 1$
3. **Membership in subdifferential:** $\sum_n \tilde{\alpha}_n y_n \widehat{\boldsymbol{x}}_n \in \partial^o \|\widehat{\boldsymbol{\beta}}\|_{2/L}^{(S)}$

In the third condition above, ∂^o is the local sub-differential of Clarke (1975): $\partial^o f(\boldsymbol{\beta}) = \text{conv}\{\lim_{i\to\infty} \nabla f(\boldsymbol{\beta} + \boldsymbol{h}_i) : \boldsymbol{h}_i \to 0\}$[8].

We will need the following assumption in the special case $L = 2$:

Assumption A.7 ($L = 2$ bounded subgradient)**.** *Let $\widehat{\boldsymbol{z}} = \sum_n \tilde{\alpha}_n y_n \widehat{\boldsymbol{x}}_n$ result from the KKT conditions of the optimization problem in $\boldsymbol{W}$, Equation A.10, as described previously. Then, we assume that $\|\widehat{\boldsymbol{z}}\|_\infty^{(S)} \leq 1$.*

Let $\boldsymbol{\beta}^\infty = \mathcal{P}(\boldsymbol{W}^\infty)$ and let $\tilde{\alpha}_i = \frac{1}{\gamma}\alpha_i$ for all i, where γ is equal to $\left(||\widehat{\boldsymbol{\beta}}^\infty||_{\frac{2}{L}}\right)$ for $L > 2$ and to 1 otherwise. (Note that by homogeneity of $\mathcal{P}$, $\mathcal{P}(\boldsymbol{W}^\infty) = \mathcal{P}(\tau\overline{\boldsymbol{W}}^\infty) = \tau^L \mathcal{P}(\overline{\boldsymbol{W}}^\infty)$.) We will check these conditions one by one for β^∞ and $\tilde{\alpha}_i$, with the first two following immediately from Theorem A.6 and the last one requiring the most manipulation.

Feasibility Trivially, $\tilde{\alpha}_i \geq 0 \;\; \forall i$ by definition of α in Equation A.10. Similarly, $y_n \langle \boldsymbol{x}_n, \beta^\infty\rangle = y_n \langle \boldsymbol{x}_n, \mathcal{P}(\boldsymbol{W}^\infty)\rangle \geq 1$.

[8] $\boldsymbol{h}_i$ is a sequence of *vectors* in some linear space, and we take $h_i \to 0$ as an entry-wise statement. This is because the vectors are finite-dimensional, so all norms are equivalent.

Complementary slackness If $y_n \langle \beta^\infty, \boldsymbol{x}_n \rangle > 1 = y_n \langle \mathcal{P}(\boldsymbol{W}^\infty), \boldsymbol{x}_n \rangle > 1$, then $\tilde{\alpha}_n \propto \alpha_n = 0$.

Membership in subdifferential We first characterize the set $\partial^o \|\widehat{\boldsymbol{\beta}}\|_{2/L}^{(S)}$ for a generic matrix $\boldsymbol{\beta}$, and then show that $\widehat{\boldsymbol{z}} \triangleq \sum_n \tilde{\alpha}_n y_n \widehat{\boldsymbol{x}}_n \in \partial^o \|\widehat{\boldsymbol{\beta}}^\infty\|_{2/L}^{(S)}$. When $L = 2$ and thus $p = 1$, the Schatten norm $\|\widehat{\boldsymbol{\beta}}^\infty\|_1$ is indeed a norm and its subgradient is known; see e.g. Watson (1992). We restate this result below:

Lemma A.8 (Subdifferential of p-Schatten norm, $p = 1$)**.** *Suppose $L = 2$, such that $\frac{2}{L} = p = 1$. Let $\boldsymbol{A}$ be an $n \times n$ complex-valued matrix. Then we have*

$$\partial^o \|\boldsymbol{A}\|_1^{(S)} = \left\{ \boldsymbol{G} : \|\boldsymbol{A}\|_1^{(S)} = \mathrm{tr}[\boldsymbol{G}^\dagger \boldsymbol{A}], \|\boldsymbol{G}\|_\infty^{(S)} \leq 1 \right\} \tag{A.12}$$

Lemma A.9 (Subdifferential of p-Schatten norm, $p < 1$)**.** *Suppose $L > 2$ and let $\frac{2}{L} = p$, such that $0 < p < 1$. Let $\boldsymbol{A}$ be an $n \times n$ complex-valued matrix with singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\dagger$. Let $\Pi_{\boldsymbol{U}}$ project onto the column space of A, i.e. $\Pi_{\boldsymbol{U}}(\boldsymbol{M}) = \boldsymbol{U}\boldsymbol{U}^\dagger \boldsymbol{M}$. Let $\Pi_{\boldsymbol{V}}$ project onto the row space of $\boldsymbol{A}$, i.e. $\Pi_{\boldsymbol{V}}(\boldsymbol{M}) = \boldsymbol{M}\boldsymbol{V}\boldsymbol{V}^\dagger$. Then we have*

$$\partial^o \|\boldsymbol{A}\|_p^{(S)} = \left\{ \boldsymbol{G} : \boldsymbol{A}^\dagger \boldsymbol{G} = \frac{1}{\|\boldsymbol{A}\|_p^{(S)}} \sqrt{\boldsymbol{A}^\dagger \boldsymbol{A}}^p \text{ and } \boldsymbol{G}\boldsymbol{A}^\dagger = \frac{1}{\|\boldsymbol{A}\|_p^{(S)}} \sqrt{\boldsymbol{A}\boldsymbol{A}^\dagger}^p \right\} \tag{A.13}$$

Proof. Suppose $\boldsymbol{A}$ is rank r and has singular value decomposition $\boldsymbol{A} = \sum_{i=1}^r d_i \boldsymbol{u}_i \boldsymbol{v}_i^\dagger$, where $d_i > 0$ for all i. Consider unit vectors $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n-r}$ which are a basis for the orthogonal subspace to $\mathrm{Span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, and unit vectors $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-r}$ which are a basis for the orthogonal subspace to $\mathrm{Span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$. Treating the space of $n \times n$ complex matrices as a n^2-dimensional linear space, we see that $\{\boldsymbol{u}_i \boldsymbol{v}_i^\dagger\}_{i=1}^r$ form a basis for an r-dimensional subspace. Let Π_A denote the projector onto this space, $\Pi_A = \sum_{i=1}^r \boldsymbol{u}_i \boldsymbol{v}_i^\dagger$. Note that $\Pi_A A = A$. For any set of $n - r$ sequences $\{\epsilon_{i,m}\}_{m=1}^\infty$ such that $\lim_{m\to\infty} \epsilon_{m,i} = 0$ for all $i = 1, \ldots, n - r$, consider the particular sequence of matrices $\{\boldsymbol{H}_m\}_{m=1}^\infty$ defined by

$$\boldsymbol{H}_m = \sum_{i=1}^{n-r} \epsilon_{m,i} \boldsymbol{W}_i \boldsymbol{s}_i^\dagger$$

By definition, $\lim_{m\to\infty} ||\boldsymbol{H}_m||_{\mathrm{Fro}} = 0$, where $\|\boldsymbol{H}_m\|_{\mathrm{Fro}} \triangleq \|\boldsymbol{H}_m\|_2^{(S)}$.Also, if $\boldsymbol{M}$ is a full rank matrix with singular value decomposition $\boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^\dagger$, $||\boldsymbol{M}||_p^{(S)}$ is differentiable at $\boldsymbol{M}$ and $\nabla ||\boldsymbol{M}||_p^{(S)} = \frac{1}{||\boldsymbol{M}||_p^{(S)}} \boldsymbol{A}\boldsymbol{D}^{p-1}\boldsymbol{B}^\dagger$. For convenience of notation, let $\boldsymbol{U}$ be the matrix with i^{th} column $\boldsymbol{u}_i$, $\boldsymbol{W}$ the matrix with i^{th} column $\boldsymbol{W}_i$, and similarly for $\boldsymbol{V}$ and $\boldsymbol{S}$ with respect to vectors $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$ respectively. Also, let $\boldsymbol{D}$ be the diagonal matrix with d_i on the i^{th} diagonal.

Combining this fact with the construction of $\boldsymbol{H}_m$, we have that

$$\nabla ||\boldsymbol{A} + \boldsymbol{H}_m||_p^{(S)} = \nabla \left|\left| \sum_{i=1}^r d_i \boldsymbol{u}_i \boldsymbol{v}_i^\dagger + \sum_{i=1}^{n-r} \epsilon_{m,i} \boldsymbol{W}_i \boldsymbol{s}_i^\dagger \right|\right|_p^{(S)} \tag{A.14}$$

$$= \nabla \left|\left| [\boldsymbol{U} \quad \boldsymbol{W}] \begin{bmatrix} \boldsymbol{D} & 0 & \ldots & 0 \\ 0 & \epsilon_{m,1} & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \epsilon_{m,n-r} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^\dagger \\ \boldsymbol{S}^\dagger \end{bmatrix} \right|\right|_p^{(S)} \tag{A.15}$$

$$= \frac{1}{(\sum_{i=1}^r d_i^p + \sum_{i=1}^{n-r} \epsilon_{m,i}^p)^{\frac{1}{p}}} \left(\boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger + \boldsymbol{W} \begin{bmatrix} \epsilon_{m,1}^{p-1} & \ldots & 0 \\ \vdots & & \ddots \\ 0 & \ldots & \epsilon_{m,n-r}^{p-1} \end{bmatrix} \boldsymbol{S}^\dagger \right) \tag{A.16}$$

In the limit as m goes to infinity, $\sum_{i=1}^{n-r} \epsilon_{m,i}^p$ approaches 0. However, $p - 1 < 0$ implies that $\lim_{m\to\infty} \epsilon_{m,i}^{p-1} = \pm\infty$. By taking convex combinations, one can create any matrix with left and right singular vectors $\boldsymbol{W}$ and $\boldsymbol{S}^\dagger$. Formally, we have:

$$\partial^o \|\boldsymbol{A}\|_p^{(S)} = \text{conv}\{\lim_{m\to\infty} \nabla \|\boldsymbol{A} + \boldsymbol{H}_m\|_p : \boldsymbol{H}_m \to 0\} \tag{A.17}$$

$$= \frac{1}{||\boldsymbol{A}||_p^{(S)}} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger + \frac{1}{||\boldsymbol{A}||_p^{(S)}} \text{conv}\left\{ \lim_{m\to\infty} \boldsymbol{W} \begin{bmatrix} \epsilon_{m,1}^{p-1} & \cdots & 0 \\ \vdots & & \ddots \\ 0 & \cdots & \epsilon_{m,n-r}^{p-1} \end{bmatrix} \boldsymbol{S}^\dagger : \epsilon_{m,i} \to_m 0 \right\} \tag{A.18}$$

$$= \frac{1}{||\boldsymbol{A}||_p^{(S)}} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger + \{\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{S}^\dagger : \boldsymbol{\Sigma} \text{ is real and diagonal}\} \tag{A.19}$$

Note that for any rank-one matrix $\boldsymbol{M} = \boldsymbol{a}\boldsymbol{b}^\dagger$, if $\boldsymbol{a}$ is not orthogonal to each column of $\boldsymbol{U}$, then $\boldsymbol{U}\boldsymbol{U}^\dagger\boldsymbol{M} \neq 0$. Similarly, if $\boldsymbol{b}$ is not orthogonal to each column of $\boldsymbol{V}$, then $\boldsymbol{M}\boldsymbol{V}\boldsymbol{V}^\dagger \neq 0$. Thus for an arbitrary matrix $\boldsymbol{M}$, by decomposing it into a sum of rank-one matrices via its SVD, we see that $\Pi_{\boldsymbol{U}}\boldsymbol{M} = \Pi_{\boldsymbol{V}}\boldsymbol{M} = 0$ implies that both the row and column spaces of $\boldsymbol{M}$ are orthogonal to those of $\boldsymbol{A}$, respectively. Thus, we can project via $\Pi_{\boldsymbol{U}}$ and $\Pi_{\boldsymbol{V}}$ to disregard the second term of Equation A.19, and obtain the following expression for the subgradient:

$$\partial^o \|\boldsymbol{A}\|_p^{(S)} = \left\{ \boldsymbol{G} : \Pi_{\boldsymbol{U}}\boldsymbol{G} = \Pi_{\boldsymbol{V}}\boldsymbol{G} = \left(||\boldsymbol{A}||_p^{(S)}\right)^{-p} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger \right\} \tag{A.20}$$

Consider first the equality

$$\Pi_{\boldsymbol{U}}\boldsymbol{G} = \left(||\boldsymbol{A}||_p^{(S)}\right)^{-p} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger \tag{A.21}$$

We can left-multiply by $\boldsymbol{A}^\dagger$ without changing the set of matrices $\boldsymbol{G}$ satisfying this relation. To see this, one can check that if $\boldsymbol{A}^\dagger\Pi_{\boldsymbol{U}}\boldsymbol{G} = \left(||\boldsymbol{A}||_p^{(S)}\right)^{-p} \boldsymbol{A}^\dagger\boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger$, left-multiplying on both sides by $\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^\dagger$ recovers Equation A.21.

Similarly, consider the second equality:

$$\Pi_{\boldsymbol{V}}\boldsymbol{G} = \left(||\boldsymbol{A}||_p^{(S)}\right)^{-p} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger \tag{A.22}$$

We can right-multiply by $\boldsymbol{A}^\dagger$ without changing the set of matrices $\boldsymbol{G}$ satisfying this relation. To see this, one can check that if

$$(\Pi_{\boldsymbol{V}}\boldsymbol{G})\boldsymbol{A}^\dagger = \left(||\boldsymbol{A}||_p^{(S)}\right)^{-p} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger\boldsymbol{A}^\dagger$$

Then right-multiplying on both sides by $\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^\dagger$ recovers Equation A.22. Finally, observe that $\boldsymbol{A}^\dagger\Pi_{\boldsymbol{U}}\boldsymbol{G} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\dagger\boldsymbol{U}\boldsymbol{U}^\dagger\boldsymbol{G} = \boldsymbol{A}^\dagger\boldsymbol{G}$ and $(\Pi_{\boldsymbol{V}}\boldsymbol{G})\boldsymbol{A}^\dagger = \boldsymbol{G}\boldsymbol{V}\boldsymbol{V}^\dagger\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\dagger = \boldsymbol{G}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\dagger = \boldsymbol{G}\boldsymbol{A}^\dagger$. Furthermore, $\boldsymbol{A}^\dagger \frac{1}{||\boldsymbol{A}||_p^{(S)}} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger = \boldsymbol{V}\boldsymbol{D}^p\boldsymbol{V}^\dagger = \frac{1}{||\boldsymbol{A}||_p^{(S)}} \sqrt{\boldsymbol{A}^\dagger\boldsymbol{A}}^p$ and $\frac{1}{||\boldsymbol{A}||_p^{(S)}} \boldsymbol{U}\boldsymbol{D}^{p-1}\boldsymbol{V}^\dagger\boldsymbol{A}^\dagger = \frac{1}{||\boldsymbol{A}||_p^{(S)}} \sqrt{\boldsymbol{A}\boldsymbol{A}^\dagger}^p$. This completes the proof of the lemma. □

Lemmas A.8 and A.9 characterize the subdifferential of the Schatten norm. Now, we show that $\sum_n \tilde{\alpha}_n y_n \widehat{\boldsymbol{x}}_n$ satisfies Equation A.13.

Lemma A.10. *Recall that our G-CNN is given by* $\text{NN}(x) = \langle \boldsymbol{x} \star \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_{L-1}, \boldsymbol{w}_L \rangle$*, with the vector of parameters* $W = [\boldsymbol{w}_1 \ \ldots \boldsymbol{w}_L]$ *and end-to-end linear predictor given by* $\mathcal{P}(\boldsymbol{W}) = \boldsymbol{w}_1 \star \cdots \star \boldsymbol{w}_L$*. Consider an arbitrary such vector of real-valued parameters. Also, we have assumed that the filters in real-space are real-valued, i.e.* $\boldsymbol{w}_i \in \mathbb{R}^{|G|}$*. Then the following relation holds:*

$$\mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \langle \mathcal{P}(\boldsymbol{W}), \boldsymbol{e}_i \rangle = \widehat{\boldsymbol{w}}_{\ell+1}^\dagger \cdots \widehat{\boldsymbol{w}}_L^\dagger \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{w}}_1^\dagger \cdots \widehat{\boldsymbol{w}}_{\ell-1}^\dagger \tag{A.23}$$

where e_i *is the* i^{th} *standard basis vector.*

Proof. Since $\langle \mathcal{P}(\boldsymbol{W}), \boldsymbol{e}_i \rangle$ is real, we have $\langle \mathcal{P}(\boldsymbol{W}), \boldsymbol{e}_i \rangle = \langle \boldsymbol{e}_i, \mathcal{P}(\boldsymbol{W}) \rangle$. Plugging this in,

$$\begin{aligned}
\mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \mathcal{P}(\boldsymbol{W})[\boldsymbol{e}_i] &= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \langle \mathcal{P}(\boldsymbol{W}), \boldsymbol{e}_i \rangle \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \langle \boldsymbol{e}_i, \mathcal{P}(\boldsymbol{W}) \rangle \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \frac{1}{|G|} \langle \mathcal{F}_M \boldsymbol{e}_i, \mathcal{F}_M \mathcal{P}(\boldsymbol{W}) \rangle_M \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \frac{1}{|G|} \langle \widehat{\boldsymbol{e}}_i, \widehat{\boldsymbol{w}}_L \cdots \widehat{\boldsymbol{w}}_1 \rangle_M \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \frac{1}{|G|} \operatorname{tr}[\widehat{\boldsymbol{e}}_i (\widehat{\boldsymbol{w}}_L \cdots \widehat{\boldsymbol{w}}_1)^\dagger] \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \frac{1}{|G|} \operatorname{tr}[\widehat{\boldsymbol{w}}_{\ell+1}^\dagger \cdots \widehat{\boldsymbol{w}}_L^\dagger \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{w}}_1^\dagger \cdots \widehat{\boldsymbol{w}}_\ell^\dagger] \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \frac{1}{|G|} \langle \widehat{\boldsymbol{w}}_{\ell+1}^\dagger \cdots \widehat{\boldsymbol{w}}_L^\dagger \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{w}}_1^\dagger \cdots \widehat{\boldsymbol{w}}_{\ell-1}^\dagger, \widehat{\boldsymbol{w}}_\ell \rangle_M \\
&= \mathcal{F}_M \nabla_{\boldsymbol{w}_\ell} \left\langle \mathcal{F}_M^{-1} (\widehat{\boldsymbol{w}}_{\ell+1}^\dagger \cdots \widehat{\boldsymbol{w}}_L^\dagger \widehat{\boldsymbol{e}}_i \widehat{w}_1^\dagger \cdots \widehat{\boldsymbol{w}}_{\ell-1}^\dagger), \boldsymbol{w}_\ell \right\rangle \\
&= \widehat{\boldsymbol{w}}_{\ell+1}^\dagger \cdots \widehat{\boldsymbol{w}}_L^\dagger \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{w}}_1^\dagger \cdots \widehat{\boldsymbol{w}}_{\ell-1}^\dagger
\end{aligned} \tag{A.24}$$

□

Letting $\boldsymbol{z} = \sum_n \tilde{\alpha}_n y_n \boldsymbol{x}_n$ and $\boldsymbol{W} = \boldsymbol{W}^\infty$, Lemma A.10 implies that

$$\mathcal{F}_M \nabla_{\boldsymbol{w}_\ell^\infty} \langle \mathcal{P}(\boldsymbol{W}^\infty), \boldsymbol{z} \rangle = \widehat{\boldsymbol{w}}_{\ell+1}^{\infty}{}^\dagger \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_{\ell-1}^{\infty}{}^\dagger \tag{A.25}$$

By combining Equation A.10 with Lemma A.10, we have

$$\frac{1}{\gamma} \widehat{\boldsymbol{w}}_\ell^\infty = \frac{1}{\gamma} \nabla_{\boldsymbol{w}_\ell^\infty} \left\langle \mathcal{P}(\boldsymbol{W}^\infty), \sum_n \alpha_n y_n \boldsymbol{x}_n \right\rangle \tag{A.26}$$

$$= \nabla_{\boldsymbol{w}_\ell^\infty} \left\langle \mathcal{P}(\boldsymbol{W}^\infty), \sum_n \tilde{\alpha}_n y_n \boldsymbol{x}_n \right\rangle \tag{A.27}$$

$$= \widehat{\boldsymbol{w}}_{\ell+1}^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty}{}^\dagger \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_{\ell-1}^{\infty\dagger} \tag{A.28}$$

As a result:

$$\begin{aligned}
\widehat{\boldsymbol{w}}_\ell^\infty &= \gamma \widehat{\boldsymbol{w}}_{\ell+1}^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_{\ell-1}^{\infty\dagger} \\
\widehat{\boldsymbol{w}}_\ell^\infty \widehat{\boldsymbol{w}}_\ell^{\infty\dagger} &= \gamma \widehat{\boldsymbol{w}}_{\ell+1}^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_\ell^{\infty\dagger}
\end{aligned} \tag{A.29}$$

Applying this relation with $\ell = L$, we have that

$$\widehat{\boldsymbol{w}}_L^\infty \widehat{\boldsymbol{w}}_L^{\infty\dagger} = \gamma \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \tag{A.30}$$

$$= \gamma \widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty\dagger} \tag{A.31}$$

Taking adjoints of both sides implies that $\widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty\dagger}$ is Hermitian, which will be useful later.

Let $\widehat{\boldsymbol{\beta}}^\infty = \mathcal{F}_M \mathcal{P}(\boldsymbol{W}^\infty)$, from which we can derive the following recursion:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^\infty \widehat{\boldsymbol{\beta}}^{\infty}{}^\dagger &= \widehat{\boldsymbol{w}}_L^\infty \cdots \widehat{\boldsymbol{w}}_1^\infty \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \\
&= \gamma^1 \widehat{\boldsymbol{w}}_L^\infty \cdots \widehat{\boldsymbol{w}}_2^\infty \widehat{\boldsymbol{w}}_2^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \text{ by Equation A.29} \\
&= \gamma^2 \widehat{\boldsymbol{w}}_L^\infty \cdots \widehat{\boldsymbol{w}}_3^\infty \widehat{\boldsymbol{w}}_3^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \widehat{\boldsymbol{w}}_2^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{w}}_1^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \text{ again, by Equation A.29} \\
&= \gamma^2 \widehat{\boldsymbol{w}}_L^\infty \cdots \widehat{\boldsymbol{w}}_3^\infty \widehat{\boldsymbol{w}}_3^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} \widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty}{}^\dagger \widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty}{}^\dagger \text{ by definition of } \boldsymbol{\beta}^\infty \\
&= \gamma^2 \widehat{\boldsymbol{w}}_L^\infty \cdots \widehat{\boldsymbol{w}}_3^\infty \widehat{\boldsymbol{w}}_3^{\infty\dagger} \cdots \widehat{\boldsymbol{w}}_L^{\infty\dagger} (\widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty}{}^\dagger)^2 \text{ by repeated application of Equation A.29} \\
&= \gamma^L (\widehat{\boldsymbol{z}} \widehat{\boldsymbol{\beta}}^{\infty}{}^\dagger)^L
\end{aligned} \tag{A.32}$$

Similarly to Equation A.29, we have:

$$\widehat{\boldsymbol{w}}_\ell^{\infty\dagger}\widehat{\boldsymbol{w}}_\ell^{\infty} = \gamma\widehat{\boldsymbol{w}}_\ell^{\infty\dagger}\widehat{\boldsymbol{w}}_{\ell+1}^{\infty}{}^{\dagger}\dots\widehat{\boldsymbol{w}}_L^{\infty\dagger}\widehat{\boldsymbol{z}}\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{\ell-1}^{\infty}{}^{\dagger} \tag{A.33}$$

By considering $\ell = 1$, we have $\widehat{\boldsymbol{w}}_1^{\infty\dagger}\widehat{\boldsymbol{w}}_1^{\infty} = \widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}}$, which shows that $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}}$ is Hermitian as well.

Using Equation A.33, we can similarly reason about $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{\beta}}^{\infty}$:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{\beta}}^{\infty} &= \widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_L^{\infty\dagger}\widehat{\boldsymbol{w}}_L^{\infty}\dots\widehat{\boldsymbol{w}}_1^{\infty} \\
&= \gamma\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-1}^{\infty}{}^{\dagger}(\widehat{\boldsymbol{w}}_L^{\infty\dagger}\widehat{\boldsymbol{z}}\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-1}^{\infty}{}^{\dagger})\widehat{\boldsymbol{w}}_{L-1}^{\infty}\dots\widehat{\boldsymbol{w}}_1^{\infty} \\
&= \gamma(\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}})\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-1}^{\infty}{}^{\dagger}\widehat{\boldsymbol{w}}_{L-1}^{\infty}\dots\widehat{\boldsymbol{w}}_1^{\infty} \\
&= \gamma(\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}})\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-1}^{\infty}{}^{\dagger}\widehat{\boldsymbol{w}}_L^{\infty\dagger}\widehat{\boldsymbol{z}}\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-2}^{\infty}{}^{\dagger}\widehat{\boldsymbol{w}}_{L-2}^{\infty}\dots\widehat{\boldsymbol{w}}_1^{\infty} \\
&= \gamma^2(\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}})^2\widehat{\boldsymbol{w}}_1^{\infty\dagger}\dots\widehat{\boldsymbol{w}}_{L-2}^{\infty}{}^{\dagger}\widehat{\boldsymbol{w}}_{L-2}^{\infty}\dots\widehat{\boldsymbol{w}}_1^{\infty} \\
&= \gamma^L(\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}})^L
\end{aligned} \tag{A.34}$$

For $L > 2$, we have shown in Lemma A.9 that

$$\partial^o\|\widehat{\boldsymbol{\beta}}^{\infty}\|_p^{(S)} = \left\{\boldsymbol{G} : \widehat{\boldsymbol{\beta}}^{\infty\dagger}\boldsymbol{G} = \frac{1}{||\widehat{\boldsymbol{\beta}}^{\infty}||_p^{(S)}}\sqrt{\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{\beta}}^{\infty}}^{p} \text{ and } \boldsymbol{G}\widehat{\boldsymbol{\beta}}^{\infty\dagger} = \frac{1}{||\widehat{\boldsymbol{\beta}}^{\infty}||_p^{(S)}}\sqrt{\widehat{\boldsymbol{\beta}}^{\infty}\widehat{\boldsymbol{\beta}}^{\infty\dagger}}^{p}\right\} \tag{A.35}$$

Let's check that setting $G = \widehat{\boldsymbol{z}}$ satisfies this relation. Since $p = \frac{2}{L}$, $\frac{p}{2} = \frac{1}{L}$ and, by Equation A.32 and that $\widehat{\boldsymbol{z}}\widehat{\boldsymbol{\beta}}^{\infty\dagger}$ is Hermitian:

$$(\widehat{\boldsymbol{\beta}}^{\infty}\widehat{\boldsymbol{\beta}}^{\infty\dagger})^{p/2} = \gamma\widehat{\boldsymbol{z}}\widehat{\boldsymbol{\beta}}^{\infty\dagger} \tag{A.36}$$

Similarly, by Equation A.34 and that $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}}$ is Hermitian:

$$(\widehat{\boldsymbol{\beta}}^{\infty}\widehat{\boldsymbol{\beta}}^{\infty\dagger})^{\frac{p}{2}} = \gamma\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}} \tag{A.37}$$

By choice of γ, $\gamma = \|\widehat{\boldsymbol{\beta}}^{\infty}\|_{\frac{2}{L}}^{(S)}$. Thus, $\widehat{\boldsymbol{z}} \in \partial^o\|\widehat{\boldsymbol{\beta}}^{\infty}\|_p^{(S)}$ as desired for $L > 2$.

For $L = 2$, $p = 1$ and by Lemma A.8 we had the following expression for the subgradient:

$$\partial^o\|\boldsymbol{A}\|_1^{(S)} = \left\{\boldsymbol{G} : \|\boldsymbol{A}\|_1^{(S)} = \mathrm{tr}[\boldsymbol{G}^\dagger\boldsymbol{A}], \|\boldsymbol{G}\|_\infty^{(S)} \leq 1]\right\} \tag{A.38}$$

For this special case of $L = 2$, we have the required assumption:

As a technical condition, we had to assume in Assumption A.7 that $\|\widehat{\boldsymbol{z}}\|_\infty^{(S)} \leq 1$. (We believe that with a refined analysis in future work, this assumption can be shown to be true given only our existing assumptions.) By the previous reasoning for $L = 2$, we had that $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{\beta}}^{\infty} = \gamma^2(\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}})^2$. Since $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{\beta}}^{\infty} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\dagger$ is positive semi-definite and symmetric, and since $\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}}$ is Hermitian, we can take the square root of both sides and obtain that $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\dagger = \gamma\widehat{\boldsymbol{\beta}}^{\infty\dagger}\widehat{\boldsymbol{z}} = \gamma\widehat{\boldsymbol{z}}^\dagger\widehat{\boldsymbol{\beta}}^{\infty}$. Thus, $\mathrm{tr}[\widehat{\boldsymbol{z}}^\dagger\widehat{\boldsymbol{\beta}}^{\infty}] = \mathrm{tr}[\boldsymbol{D}\boldsymbol{U}^\dagger\boldsymbol{U}] = \mathrm{tr}[\boldsymbol{D}] = \|\widehat{\boldsymbol{\beta}}^{\infty}\|_1^{(S)}$, which is what was needed (since $\gamma = 1$ for the $L = 2$ case).

A.2.1 INFINITE GROUPS WITH BAND-LIMITED INPUTS

Here we consider the case of more general, infinite-dimensional compact Lie groups. Such groups admit a Fourier transform which is an operator between infinite-dimensional spaces, rather than a finite matrix as before, but which has the same key properties: a convolution theorem, and preservation of inner products. To be concrete, the Fourier transform is now defined as $\widehat{f}(\rho) = \int_{g \in G} \rho(g) f(g) \mu(g)$, where $\mu(g)$ denotes the Haar measure for the group.

Since it is impossible to store a general function $x : G \to \mathbb{R}$, one must make simplifying assumptions on the input x. A common assumption is that x is band-limited in Fourier space, i.e. $\widehat{x}$ is supported only on a small (and finite) subset of Fourier coefficients contained within the irreps $\rho \in S$. For many such groups, there is a natural hierarchy of irreps (analogous to low frequencies and high frequencies in classical Fourier analysis), and so practical architectures typically assume only those corresponding to low frequencies are non-zero.

For ease of analysis, we will assume that the input functions and all convolutional filters are real-valued in Fourier space. The architecture of our G-CNN is the same as in the finite-dimensional group setting except that we will apply functions entirely in the finite-dimensional Fourier space over the irreps in S. Given a function $\underline{\widehat{\boldsymbol{x}}}$ supported only on S in Fourier space, we run gradient descent over only the Fourier coefficients on S of filters $\underline{\widehat{\boldsymbol{w}_i}}$, and assume they are zero elsewhere. Before we proceed, we define $\mathcal{F}_M$ in the natural way after restricting the irreps to the band-limiting space S:

$$\underline{\widehat{\boldsymbol{f}}} = \mathcal{F}_M \boldsymbol{f} = \bigoplus_{\rho \in S} \widehat{f}(\rho)^{\oplus d_\rho} \in \mathrm{GL}\left(\sum_{\rho \in S} d_\rho^2, \mathbb{C}\right). \tag{A.39}$$

In this band-limited space, for each filter, there are $\sum_{\rho \in S} d_\rho^2$ trainable parameters corresponding to the entries of the irreps in S. Note that these entries are also orthogonal to each other with respect to the inner product $\langle \cdot, \cdot \rangle_M$ and thus form a vector space of dimension $\sum_{\rho \in S} d_\rho^2$.

Recall that we previously applied Theorem A.6 to the homogeneous polynomial parametrization from $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L) \to \langle \boldsymbol{x} * \boldsymbol{w}_1 * \cdots * \boldsymbol{w}_{L-1}, \boldsymbol{w}_L \rangle$. Since we will operate in Fourier space, instead consider the homogeneous polynomial parametrization $\widehat{\boldsymbol{W}} \in \mathbb{R}^p$ containing the parameters stored in the matrices $\{\underline{\widehat{\boldsymbol{w}_1}}, \ldots, \underline{\widehat{\boldsymbol{w}_L}}\}$. In other words, there are $p = L(\sum_{\rho \in S} d_\rho^2)$ parameters stored in the matrices contained in the set $\widehat{\boldsymbol{W}}$.

$$\widehat{\boldsymbol{W}}^{t+1} = \widehat{\boldsymbol{W}}^t - \eta_t \nabla_{\widehat{\boldsymbol{W}}} \mathcal{L}(\widehat{\boldsymbol{W}}^t) = \widehat{\boldsymbol{W}}^t - \eta_t \nabla_{\widehat{\boldsymbol{W}}} \mathcal{P}(\widehat{\boldsymbol{W}}^t) \nabla_{\widehat{\boldsymbol{\beta}}} \mathcal{L}(\mathcal{P}(\widehat{\boldsymbol{W}}^t)) \tag{A.40}$$

Note that here, iterates are only allowed to vary over the finite subset S of Fourier coefficients and are assumed to be equal to 0 elsewhere. If we assume further that the exponential loss converges to zero, that the gradients $\nabla_{\widehat{\boldsymbol{\beta}}} \mathcal{L}(\widehat{\boldsymbol{\beta}}^t)$ converge in direction, and that the iterates $\widehat{\boldsymbol{W}^t}$ themselves converge in direction to yield a separator with positive margin, then the limit direction of the parameters $\overline{\widehat{\boldsymbol{W}^\infty}} = \lim_{t \to \infty} \frac{\widehat{\boldsymbol{W}^t}}{\|\widehat{\boldsymbol{W}^t}\|_2}$ is a positive scaling of a first order stationary point of the following optimization problem:

$$\min_{\widehat{\boldsymbol{W}} \in \mathbb{R}^P} \|\widehat{\boldsymbol{W}}\|_2^2 \quad \text{s.t.} \quad \forall n, y_n \langle \underline{\widehat{\boldsymbol{x}_n}}, \mathcal{P}(\widehat{\boldsymbol{W}}) \rangle_M \geq 1. \tag{A.41}$$

Again letting $\widehat{\boldsymbol{W}^\infty} = \tau \overline{\widehat{\boldsymbol{W}^\infty}}$ denote the first order stationary point itself, $\widehat{\boldsymbol{W}^\infty}_i$ the individual layers (or parameter blocks) comprising $\widehat{\boldsymbol{W}^\infty}$, and $\overline{\widehat{\boldsymbol{W}^\infty}}_i$ the individual layers comprising $\overline{\widehat{\boldsymbol{W}^\infty}}$, we again have via the KKT conditions that:

$$\begin{aligned} &\exists \{\alpha_n : \alpha_n \geq 0\}_{n=1}^N \text{ s.t. } \alpha_n = 0 \text{ if } y_n \langle \underline{\widehat{\boldsymbol{x}_n}}, \mathcal{P}(\widehat{\boldsymbol{W}^\infty}) \rangle_M > 1 \\ &\widehat{\boldsymbol{W}_i}^\infty = \nabla_{\widehat{\boldsymbol{W}_i}} \mathcal{P}(\widehat{\boldsymbol{W}^\infty}) \left[\sum_n \alpha_n y_n \underline{\widehat{\boldsymbol{x}_n}} \right] = \nabla_{\widehat{\boldsymbol{W}_i}} \left\langle \mathcal{P}(\widehat{\boldsymbol{W}^\infty}), \sum_n \alpha_n y_n \underline{\widehat{\boldsymbol{x}_n}} \right\rangle_M \end{aligned} \tag{A.42}$$

Equivalently, writing the above in terms of the matrices $\underline{\widehat{\boldsymbol{w}}_\ell^\infty}$, and defining $\underline{\widehat{\boldsymbol{z}}} = \sum_n \alpha_n y_n \underline{\widehat{\boldsymbol{x}_n}}$, we have an equivalence to Lemma A.10.

$$\begin{aligned}
\nabla_{\boldsymbol{w}_\ell} \mathcal{P}(\widehat{\boldsymbol{W}^\infty})[\widehat{\underline{\boldsymbol{z}}}] &= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \langle \mathcal{P}(\widehat{\boldsymbol{W}^\infty}), \widehat{\underline{\boldsymbol{z}}} \rangle_M \\
&= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \langle \widehat{\underline{\boldsymbol{z}}}, \mathcal{P}(\widehat{\boldsymbol{W}}^\infty) \rangle_M \\
&= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \langle \widehat{\underline{\boldsymbol{z}}}, \widehat{\underline{\boldsymbol{w}}}_L^\infty \cdots \widehat{\underline{\boldsymbol{w}}}_1^\infty \rangle_M \\
&= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \operatorname{tr}[\widehat{\underline{\boldsymbol{z}}} (\widehat{\underline{\boldsymbol{w}}}_L^\infty \cdots \widehat{\underline{\boldsymbol{w}}}_1^{\infty^\dagger})] \\
&= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \operatorname{tr}[\widehat{\underline{\boldsymbol{w}}}_{\ell+1}^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_L^{\infty^\dagger} \widehat{\underline{\boldsymbol{z}}} \widehat{\underline{\boldsymbol{w}}}_1^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_\ell^{\infty^\dagger}] \\
&= \nabla_{\widehat{\underline{\boldsymbol{w}}}_\ell^\infty} \langle \widehat{\underline{\boldsymbol{w}}}_{\ell+1}^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_L^{\infty^\dagger} \widehat{\underline{\boldsymbol{z}}} \widehat{\underline{\boldsymbol{w}}}_1^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_{\ell-1}^{\infty^\dagger}, \widehat{\underline{\boldsymbol{w}}}_\ell^\infty \rangle_M \\
&= \widehat{\underline{\boldsymbol{w}}}_{\ell+1}^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_L^{\infty\dagger} \widehat{\underline{\boldsymbol{z}}} \widehat{\underline{\boldsymbol{w}}}_1^{\infty^\dagger} \cdots \widehat{\underline{\boldsymbol{w}}}_{\ell-1}^{\infty^\dagger}
\end{aligned} \tag{A.43}$$

Using the KKT conditions above and this fact, all of the manipulations demonstrating that a positive scaling of $\sum_n \alpha_n y_n \widehat{\underline{\boldsymbol{x}_n}}$ is a first-order stationary point of the optimization problem below carry over exactly as they do in the previous part. This yields the following formal result:

Theorem A.11. *Consider a classification task with ground-truth linear predictor $\widehat{\boldsymbol{\beta}}$, trained via a real-valued, Fourier-space, band-limited linear G-CNN architecture $\mathrm{NN}(\widehat{\underline{\boldsymbol{x}}}) = \langle \widehat{\underline{\boldsymbol{x}}}, \widehat{\underline{\boldsymbol{w}}_L} \ldots \widehat{\underline{\boldsymbol{w}}_1} \rangle$ with $L \geq 2$ layers under the exponential loss. Then for almost any datasets $\{\boldsymbol{x}_i, y_i\}$ separable by $\boldsymbol{\beta}$, any bounded sequence of step sizes η_t, and almost all initializations, suppose that:*

- *The loss converges to 0*
- *The gradients with respect to the end-to-end linear predictor $\widehat{\underline{\boldsymbol{\beta}}}$ converge in direction*
- *The iterates $\widehat{\underline{\boldsymbol{w}_i}}$ themselves converge in direction to a separator with positive margin*

When $L = 2$, we need an additional technical assumption, Assumption A.7. Then, the resultant linear predictor $\widehat{\underline{\boldsymbol{\beta}}}^\infty$ is a positive scaling of a first order stationary point of the optimization problem:

$$\min_{\widehat{\underline{\boldsymbol{\beta}}}} \left\| \widehat{\underline{\boldsymbol{\beta}}} \right\|_{2/L}^{(S)} \quad \textit{s.t.} \quad \forall n, y_n \left\langle \widehat{\underline{\boldsymbol{x}_n}}, \widehat{\underline{\boldsymbol{\beta}}} \right\rangle_M \geq 1. \tag{A.44}$$

B GROUP FOURIER TRANSFORMS

To aid the reader in understanding the notation and structure behind the group Fourier transform, the following exposition is given for reference and convenience. Here, we introduce important concepts from representation theory and from there, provide explicit constructions the group Fourier transform.

A representation of a group G is a vector space V together with a G-linear map $\rho : G \longrightarrow \mathrm{GL}(V)$. Of particular interest is the *regular representation* which we construct as follows. Let G be a group of order n and choose $V = \mathbb{C}^n$. Consider an element $u \in \mathbb{C}[G]$, so $u = a_1 g_1 + \cdots + a_n g_n$ where $a_i \in \mathbb{C}$ and $g_n \in G$, and the associated vector $\mathbf{u} \in \mathbb{C}^n$ such that $\mathbf{u} = [a_1 \quad \cdots \quad a_n]$.

The action of left-multiplication of u for any $h \in G$ yields $hu = a_1(hg_1) + \cdots + a_n(hg_n)$, which is equivalent to a permutation of the coefficients, so there is a unique matrix $H \in \mathrm{GL}(\mathbb{C}^n)$ such that $H\mathbf{u}$ is equivalent to the associated vector for hu. The G-linear map $L_h : h \mapsto H$ is the (left) *regular representation*.

The direct sum of two G-representations (ρ_1, V_1) and (ρ_2, V_2) can be constructed by

$$(\rho_1 \oplus \rho_2)(g) = \begin{bmatrix} \rho_1(g) & \\ & \rho_2(g) \end{bmatrix} \tag{B.1}$$

The dimension d_ρ of a representation (ρ, V) is defined to be $\dim(V)$. A finite-dimensional representation is *irreducible* if it cannot be written as the direct sum of two nontrivial representations.

Denote $\widehat{G}$ to be the set of isomorphism classes of irreducible subrepresentations of L_u, and let d_{ρ_i} denote the dimension of $\rho_i \in \widehat{G}$. As it turns out, there is an isomorphism

$$L_u \cong \bigoplus_{\rho \in \widehat{G}} \rho^{\oplus d_\rho}$$

Where ρ ranges over one representative from each of the isomorphism classes of $\widehat{G}$, repeated according to its multiplicity. In general, this decomposition is not uniquely determined as it depends on the choice of representatives. Throughout this paper, we choose representatives such that each ρ is *unitary*, meaning that every $\rho(g)$ is a unitary matrix.

Every function $f : G \longrightarrow \mathbb{C}$ can be considered as equivalent to an element $u_f \in \mathbb{C}[G]$ by setting $u_f = f(g_1)g_1 + \cdots + f(g_n)g_n$. And as we have already seen, every $u \in \mathbb{C}[G]$ can naturally be considered a subrepresentation of L_u. Then the *Fourier transform* of f at a representation ρ, denoted $\widehat{f}(\rho) \in \mathrm{GL}(d_\rho, \mathbb{C})$ where $\widehat{f}(\rho) = \sum_{u \in G} f(u)\rho(u)$, can be considered as a projection of $L_u(u_f)$ onto the orthogonal subspace described by ρ. Throughout the paper we use slightly different notations and characterizations of the Fourier transform depending on the context, but all share this projection as the fundamental operation.

Recalling Equation B.1, there is a representation isomorphic to L_u, which we will suggestively call $\mathcal{F}_M$, that block-diagonalizes the image of L_u into orthogonal subspaces of unitary irreducible representations (each $\rho \in \widehat{G}$ repeated with multiplicity d_ρ):

$$\mathcal{F}_M(u) = \begin{bmatrix} \rho_1(u) & & \\ & \ddots & \\ & & \rho_j(g) \end{bmatrix} \in \mathrm{GL}(\mathbb{C}^n) \tag{B.2}$$

For the last piece of the puzzle, extend the domain of $\mathcal{F}_M$ to all functions $f : G \longrightarrow \mathbb{C}$ (by considering f as an element of $\mathbb{C}[G]$). Then we obtain

$$\mathcal{F}_M f = \begin{bmatrix} \widehat{f}(\rho_1) & & \\ & \ddots & \\ & & \widehat{f}(\rho_j) \end{bmatrix} \in \mathrm{GL}(\mathbb{C}^n) \tag{B.3}$$

Which we call the *matrix Fourier transform* of f.

We also make use of the *Fourier basis matrix* $\mathcal{F}$, which depends only on the group G and not the function f. To construct it, we first need the operation $\mathrm{Flatten}$ which vertically stacks the columns of a matrix. Then define the transform

$$\phi(f) = \begin{bmatrix} \mathrm{Flatten}\left(\widehat{f}(\rho_1)\right) \\ \vdots \\ \mathrm{Flatten}\left(\widehat{f}(\rho_j)\right) \end{bmatrix} \tag{B.4}$$

Letting $e_i : G \longrightarrow \mathbb{C}$ be the indicator function $e_i(g_j) = \mathbb{1}_{i=j}$ we can describe the unitary Fourier basis matrix $\mathcal{F}$ for a group G as a row-scaling of

$$\mathcal{F} \propto \begin{bmatrix} \phi(e_1) & \phi(e_2) & \cdots & \phi(e_n) \end{bmatrix} \tag{B.5}$$

In other words, given a column vectorization $\mathbf{f}$ of a function f such that $\mathbf{f}_i = f(g_i)$, then $\mathcal{F}$ is the matrix such that the 'unflattening' of $\mathcal{F}\mathbf{f}$ is equal to $\mathcal{F}_M f$ up to the row-scaling constants. Thus we can treat the group Fourier transform either as an abstract isomorphism or as a concrete matrix-vector multiplication, depending on the application.

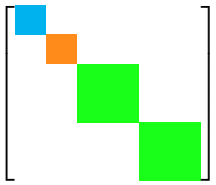
The matrix $\mathcal{F}$ can be explicitly constructed as described in Definition 4.1. Denoting $\boldsymbol{e}_{[\rho,i,j]}$ as the column-major vectorized basis for element ρ_{ij} in the group Fourier transform, then we can form the matrix

$$\mathcal{F} = \sum_{u \in G} \sum_{\rho \in \widehat{G}} \frac{\sqrt{d_\rho}}{\sqrt{|G|}} \sum_{i,j=1}^{d_\rho} \rho(u)_{ij} \boldsymbol{e}_{[\rho,i,j]} \boldsymbol{e}_u^T. \tag{B.6}$$

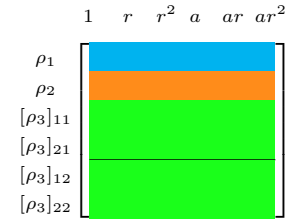
For visualization, consider the dihedral group $D_6 = \{1, r, r^2, a, ar, ar^2\}$, which has three irreducible representations ρ_1, ρ_2, ρ_3 (up to isomorphism) of dimensions 1, 1, and 2 respectively, and let $f : D_6 \longrightarrow \mathbb{C}$. Using colors instead of values at first to avoid numerical clutter:

$\widehat{f}(\rho_1) =$  $\widehat{f}(\rho_2) =$  $\widehat{f}(\rho_3) =$ 

Since $L_{D_6} \cong \rho_1 \oplus \rho_2 \oplus \rho_3^2$, we can get something like

$\mathcal{F}_M f =$  $\in \mathrm{GL}(\mathbb{C}^n)$

Whereas for the unitary Fourier basis matrix we have the form

$\mathcal{F} \propto$ 

Note that we do not yet include the row-scaling constants. Now explicitly, choosing ρ_1 the trivial irrep, ρ_2 the sign irrep, and ρ_3 the representation sending

$$\rho_3(a) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \rho_3(r) = \begin{bmatrix} \omega & 0 \\ 0 & \omega^2 \end{bmatrix} \tag{B.7}$$

Where $\omega = e^{2\pi i/3}$, then the matrix $\mathcal{F}$ is exactly

$$\mathcal{F} = \frac{1}{\sqrt{6}} \begin{array}{c} \\ \rho_1 \\ \rho_2 \\ [\rho_3]_{11} \\ [\rho_3]_{21} \\ [\rho_3]_{12} \\ [\rho_3]_{22} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & r & r^2 & a & ar & ar^2 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2}\omega & \sqrt{2}\omega^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{2} & \sqrt{2}\omega & \sqrt{2}\omega^2 \\ 0 & 0 & 0 & \sqrt{2}\omega^2 & \sqrt{2}\omega & \sqrt{2} \\ \sqrt{2} & \sqrt{2}\omega^2 & \sqrt{2}\omega & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{B.8}$$

Note that the above is only one possible way of writing $\mathcal{F}$ since the 2-dimensional irreducible representation of D_6 is unique only up to conjugation by a unitary matrix.

C Uncertainty Principles for Groups

In mathematics, uncertainty principles categorize trade-offs of the "amount of information" stored in a function between canonically conjugate regimes, *e.g.,* position (real regime) and momentum (Fourier regime) of a physical particle. More generically, uncertainty principles show that a function and its Fourier transform cannot both be very localized or concentrated. In the context of group theory, uncertainty principles specifically show that sparse support in either the real or Fourier regime of a group necessarily implies dense support in the conjugate regime (Wigderson & Wigderson, 2021). Such results are directly relevant when interpreting implicit regularization of linear G-CNNs which bias gradient descent towards sparse solution in the Fourier basis of the group. In this section, we formally state and summarize these group theoretic uncertainty principles to provide intuition into the properties of functions which linear group convolutional networks are likely to learn.

Abelian groups For abelian groups, the fundamental uncertainty principle details a trade-off between the norms of a function in its real and Fourier bases.

Theorem C.1 (Generalization of Donoho-Stark Theorem (Wigderson & Wigderson, 2021)). *Given a finite abelian group G, let $\widehat{G}$ be the set of irreducible representations of G and $f : G \to \mathbb{C}$ be a function mapping group elements to complex numbers. Let $\boldsymbol{f}$ be the vectorized function and $\mathcal{F}$ be the unitary group Fourier transform (see Definition 4.1), then*

$$\frac{\|\boldsymbol{f}\|_1}{\|\boldsymbol{f}\|_\infty} \frac{\|\mathcal{F}\boldsymbol{f}\|_1}{\|\mathcal{F}\boldsymbol{f}\|_\infty} \geq |G| \tag{C.1}$$

Remark. *Since the size of the support of a vector is bounded by $|supp(\boldsymbol{v})| \geq \frac{\|\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_\infty}$, this directly implies that $|supp(\boldsymbol{f})|\ |supp(\mathcal{F}\boldsymbol{f})| \geq |G|$ recovering the Donoho-Stark theorem (Donoho & Stark, 1989; Matusiak et al., 2004).*

Non-abelian groups Since non-abelian groups have matrix valued irreducible representations, uncertainty theorems must account for norms and notions of support in the context of matrices. Here, we will provide two different uncertainty theorems for the non-abelian setting – one via the rank of irreducible representations and another via the Schatten norm of irreducible representations. Uncertainty relations for non-abelian groups make use of the matrix group Fourier transform detailed in Definition 4.1.

Theorem C.2 (Meshulam uncertainty theorem (Meshulam, 1992)). *Given a finite non-abelian group G, let $\widehat{G}$ be the set of irreducible representations of G and $f : G \to \mathbb{C}$ be a function mapping group elements to complex numbers. Let $\boldsymbol{f}$ be the vectorized function and $\mathcal{F}_M$ be the matrix group Fourier transform (see Definition 4.1), then*

$$|\operatorname{supp}(\boldsymbol{f})|\ \operatorname{rank}(\mathcal{F}_M\boldsymbol{f}) = |\operatorname{supp}(\boldsymbol{f})|\ \left[\sum_{\rho\in\widehat{G}} d_\rho \operatorname{rank}\left(\widehat{f}(\rho)\right)\right] \geq |G| \tag{C.2}$$

The above theorem shows that the rank of the matrix Fourier transformed function is the proper notion of support in the uncertainty theorem for a non-abelian group. A stronger uncertainty principle which is a more direct corollary to Theorem C.1 can be obtained via the Schatten norms of the irreducible representations as shown below.

Theorem C.3 (Kuperberg uncertainty theorem (Wigderson & Wigderson, 2021)). *Given a finite non-abelian group G, let $\widehat{G}$ be the set of irreducible representations of G and $f : G \to \mathbb{C}$ be a function mapping group elements to complex numbers. Let $\boldsymbol{f}$ be the vectorized function and $\mathcal{F}_M$ be the matrix group Fourier transform (see Definition 4.1), then*

$$\frac{\|\boldsymbol{f}\|_1}{\|\boldsymbol{f}\|_\infty} \frac{\|\mathcal{F}_M\boldsymbol{f}\|_1^{(S)}}{\|\mathcal{F}_M\boldsymbol{f}\|_\infty^{(S)}} = \frac{\|\boldsymbol{f}\|_1}{\|\boldsymbol{f}\|_\infty} \frac{\sum_{\rho\in\widehat{G}} d_\rho \|\widehat{f}(\rho)\|_1^{(S)}}{\max_{\rho\in\widehat{G}} \|\widehat{f}(\rho)\|_\infty^{(S)}} \geq |G|. \tag{C.3}$$

D VISUALIZING THE IMPLICIT BIAS

Implicit biases induced by the G-CNN architectures studied here can be readily observed by analyzing coefficients of the linearized transformation in the Real or Fourier regimes. Here, we visualize the linearized outputs a 3-layer linear G-CNN over the Dihedral group D_{60} which has 4 scalar irreps and 14 irreps of dimension 2 (hence 2×2 matrices). Figure 6 shows these linearized coefficients of the G-CNN, CNN, and FC at intialization and training.

As evident in Figure 6, the learned coefficients of the G-CNN are sparse in the Fourier regime of the group. This sparsity pattern appears over blocks of irreps of length four, corresponding to coefficients of the 2×2 irreps of D_{60}. Furthermore, the values of the coefficients within a block are roughly constant, highlighting the bias towards low rank irreps. The relative denseness of coefficients of the trained G-CNN in the real regime is inherent due to the uncertainty principles of group functions. Unlike the G-CNN, the fully connected network (FC) appears to have no bias towards sparseness in its coefficients in either the real or Fourier regime. On a related note, the cyclic group of CNNs share some of the same irreps as those of the G-CNN studied here. This may be one explanation for the partial sparsity patterns observed in the coefficients of the CNN in the Fourier regime.

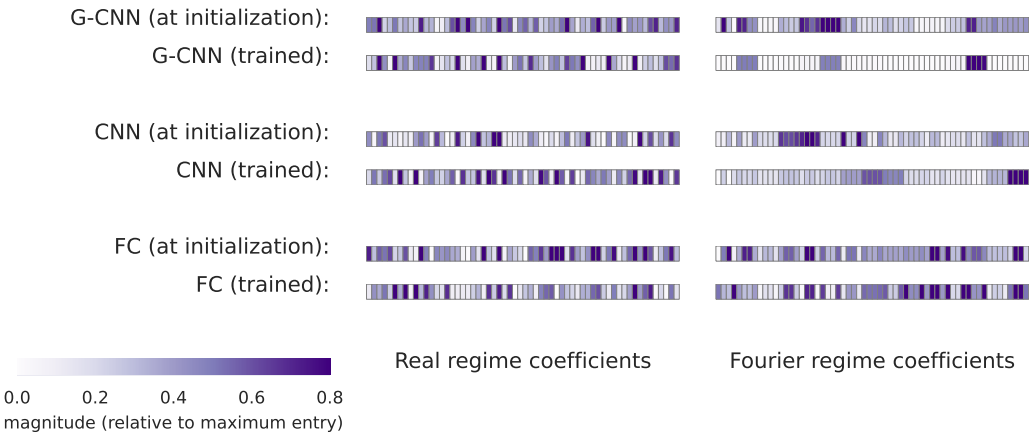


Figure 6: Linearized functions of the G-CNN (over the dihedral group D_{60}) are sparse in the Fourier regime of the group. Furthermore, the sparsity pattern shows up in blocks of 4 which are the blocks containing coefficients of individual irreps of the dihedral group D_{60}. Linearized functions in the real regime of the G-CNN, in contrast, are rather dense, highlighting the uncertainty principles inherent in the implicit bias of G-CNNs. Sparsity in the group Fourier regime is not evident in CNNs or fully connected (FC) networks.

E Computational details

As described in Section 6, for all models we use three-layer networks over $\mathbb{R}^G$ and binary classification tasks trained via standard gradient descent with exponential loss. We often train networks on isotropic Gaussian data points which are random vectors whose entries are drawn i.i.d. from a standard Normal. For the linear networks on simulated data (with a fully connected output layer) we use the groups D_8, and $(C_{10} \times C_{10} \times C_2)$, and $(C_5 \times C_5) \rtimes Q_8$. For the linear and nonlinear networks on MNIST, we use $(C_{28} \times C_{28}) \times D_8$. For the networks with ReLU activations and a linear layer (instead of pooling) we use the groups D_8 and D_{60}. For the experiments on MNIST with non-linear networks, we have used the `e2cnn` package Weiler & Cesa (2019). The weights are initialized with the standard uniform initialization. We choose an appropriate learning rate for each task depending on the dimension and magnitude of the values - all learning rate choices are specified in the attached code. Since each problem is overparameterized the loss will almost surely converge to zero, so we choose enough training epochs to achieve satisfactory convergence—this ended up being around 500 epochs for most tasks. For each plot, we report 95% confidence intervals over 10 to 50 runs, depending on the classification task. The computational resources used are modest - commodity hardware should suffice to fully reproduce our results. For further details, please see the attached code in the supplementary materials.

E.1 Additional experiments

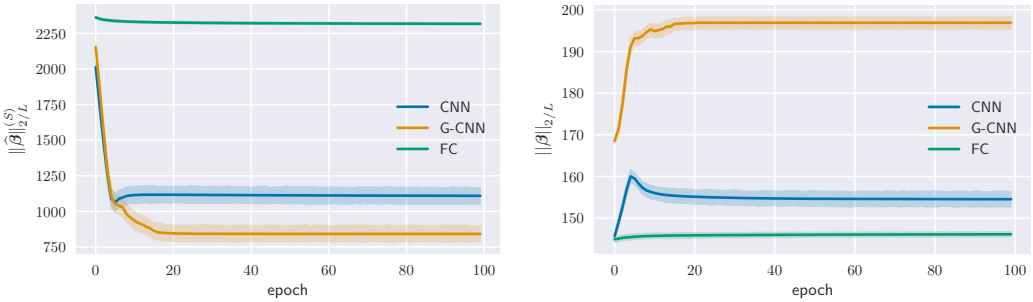
We include here the real and Fourier space plots for an experiment on $C_{10} \times C_{10} \times C_2$ (Figure 7), an abelian group with three generators which captures more complicated group operations than cyclic shifts or standard planar convolutions. Inputs are vectors with elements drawn i.i.d. from the standard normal distribution.

E.2 Omitted real-space plots

E.3 Loss plots

Acknowledgments

Redacted until publication.



(a) Fourier space norm of network linearization $\boldsymbol{\beta}$ (b) Real space norm of network linearization $\boldsymbol{\beta}$

Figure 7: Norms of the linearizations of three different linear architectures for the abelian group $G = C_{10} \times C_{10} \times C_2$ trained on a binary classification task with six isotropic Gaussian data points.

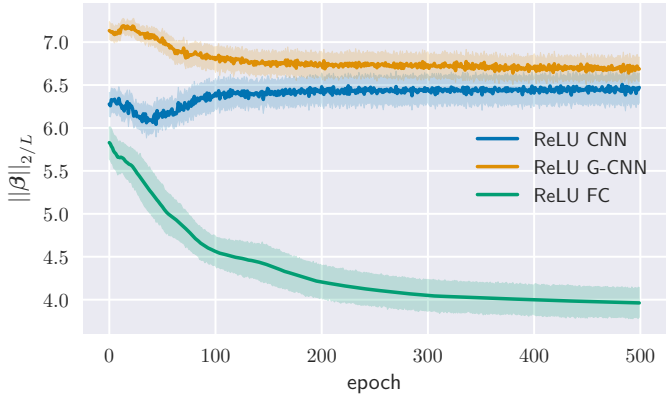


Figure 8: Real-space norm for D_8 with ReLU networks and 2 Gaussian training points. See Figure 5a for comparable plot of norms in Fourier space.

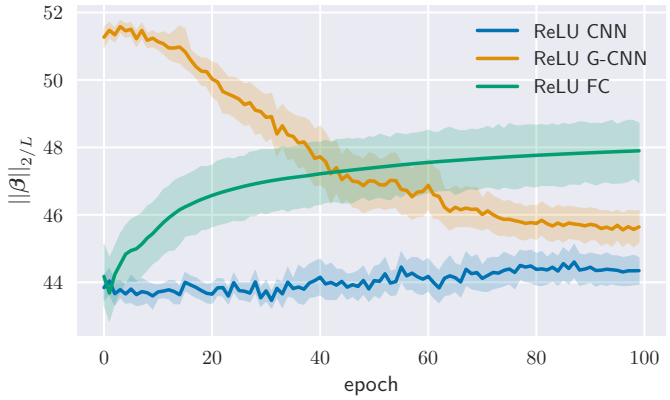


Figure 9: Real-space norm for D_{60} with ReLU networks, 10 Gaussian training points. See Figure 5b for comparable plot of norms in Fourier space.

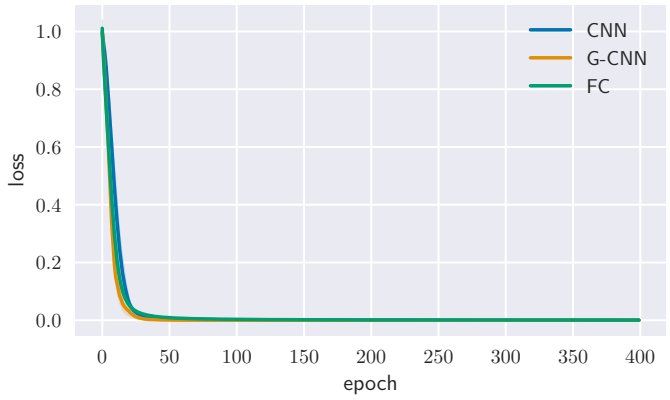


Figure 10: Loss trajectory for the setting of D_8 (see Figure 1). Networks are trained on 2 Gaussian i.i.d. data points.

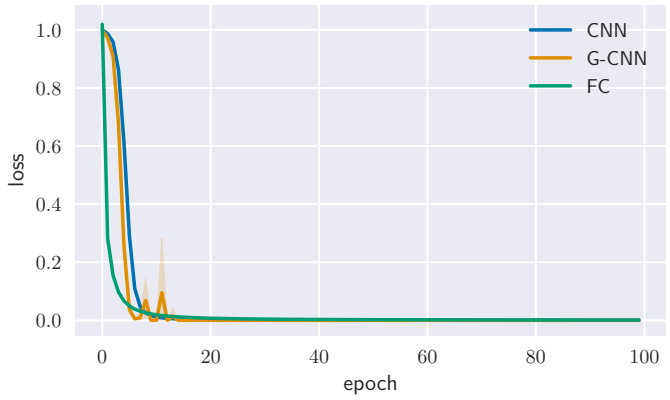


Figure 11: Loss trajectory for the setting of $C_{10} \times C_{10} \times C_2$ (see Figure 7). Networks are trained on 6 Gaussian i.i.d. data points.

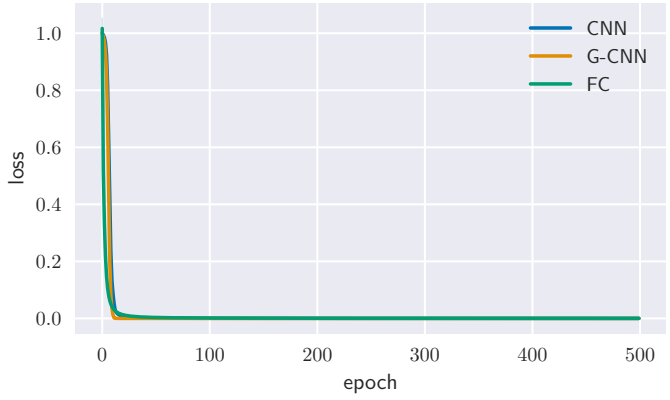


Figure 12: Loss trajectory for the setting of $(C_5 \times C_5) \rtimes Q_8$ (see Figure 3). Networks are trained on 10 Gaussian i.i.d. data points.

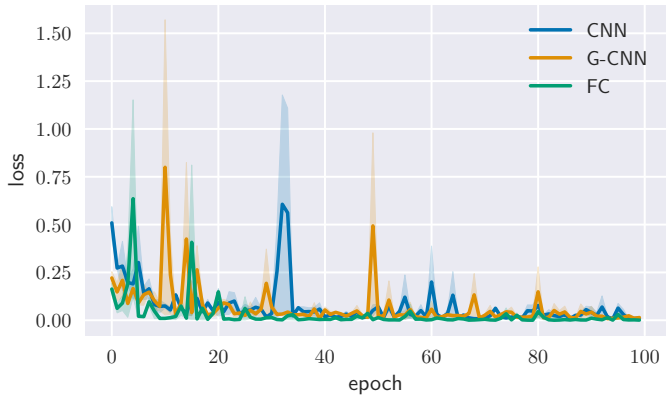


Figure 13: Loss trajectory for the setting of $(C_{28} \times C_{28}) \times D_8$ (see Figure 4). Networks are trained on the digits 1 and 5 from MNIST.

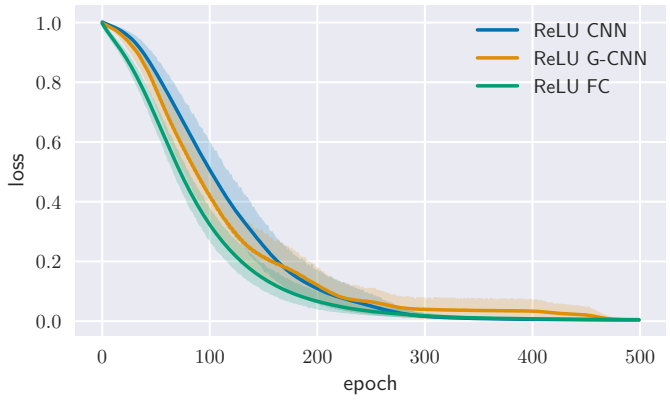


Figure 14: Loss trajectory for the setting of D_8 (see Figure 5a). Networks are nonlinear with ReLU activations and trained on 2 Gaussian i.i.d. data points.

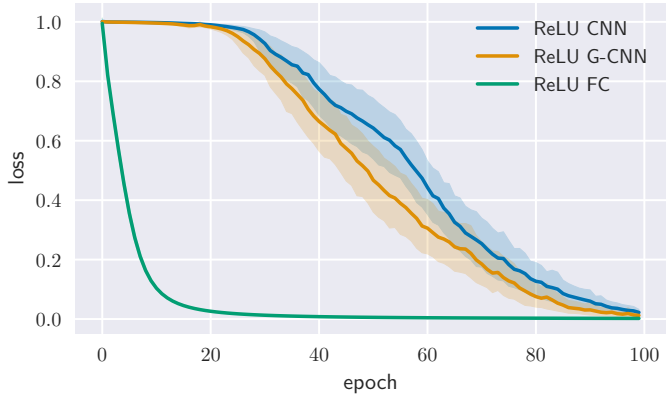


Figure 15: Loss trajectory for the setting of D_{60} (see Figure 5b). Networks are nonlinear with ReLU activations and trained on 10 distinct frequencies as data points.