Inductive Bias of Multi-Channel Linear Convolutional Networks with Bounded Weight Norm

Meena Jagadeesan*[1], Ilya Razenshteyn†[2], and Suriya Gunasekar‡[3]

[1]University of California, Berkeley
[2]CipherMode Labs
[3]Microsoft Research

Abstract

We study the function space characterization of the inductive bias resulting from controlling the ℓ_2 norm of the weights in linear convolutional networks. We view this in terms of an *induced regularizer* in the function space given by the minimum norm of weights required to realize a linear function. For two layer linear convolutional networks with C output channels and kernel size K, we show the following: (a) If the inputs to the network have a single channel, the induced regularizer for any K is a norm given by a semidefinite program (SDP) that is *independent* of the number of output channels C. We further validate these results through a binary classification task on MNIST. (b) In contrast, for networks with multi-channel inputs, multiple output channels can be necessary to merely realize all matrix-valued linear functions and thus the inductive bias *does* depend on C. Further, for sufficiently large C, the induced regularizer for $K = 1$ and $K = D$ are the nuclear norm and the $\ell_{2,1}$ group-sparse norm, respectively, of the Fourier coefficients—both of which promote sparse structures.

1 Introduction

The ℓ_2 norm of the weights of a neural network is a standard measure of complexity that has connections to implicit and explicit regularization and generalization. In the theory of generalization, complexity control on the magnitude of parameters has long been argued to play a greater role in learning than the number of parameters [Bar96; BM02; KST08; NTS15a; NTS15b; ZBHRV17]. In the current practice of deep learning, capacity control of weight magnitudes is typically achieved through a combination of explicit regularization techniques [WLLM19], as well as implicit regularization from optimization algorithms like stochastic gradient descent [NTS15a; ZBHRV17; BFT17]. Regularization of ℓ_2 norm of weights (closely related to weight decay) is by far the most common way to explicitly control the magnitude of weights [KH91; WLLM19]. Further, implicit bias from (stochastic) gradient descent in various overparameterized models has also been prominently connected to complexity control in terms of ℓ_2 norm of the weights (see *e.g.*, Lyu and Li [LL20] and Nacson et al. [NGLSS19]; additional references and discussions in Section 1.1).

Thus, controlling the ℓ_2 norm of weights while fitting the training data is an important inductive bias, but this description does not directly provide insight into the properties of the learned function.

*mjagadeesan@berkeley.edu. This work was partly done while M. Jagadeesan and I. Razenshteyn were at Microsoft Research. M. Jagadeesan was supported in part by the Paul and Daisy Soros Fellowship.

†ilya.razenshteyn@gmail.com

‡suriyag@microsoft.com

The function space view of the inductive bias can be understood in terms of the *representation cost* with respect to the weight norm, that is, the minimum ℓ_2 norm of weights needed to realize a function using a given network architecture or model class (see eq. 2). This defines an *induced complexity measure* over functions, which we also refer as the *induced regularizer*.

Previous work has characterized this induced regularizer in special network architectures and showed that even for networks that represent the same model class, minimizing or bounding the ℓ_2 norm of weights in different architectures can lead to remarkably different inductive biases in the function space. For example, in a work most closely related to ours, Gunasekar et al. [GLSS18b] showed that the implicit bias from gradient descent (which is related to minimizing the ℓ_2 of weights) leads to different induced biases in different linear networks: for fully connected linear networks, the induced regularizer is the ℓ_2 norm of the linear predictor, while for linear convolutional network with *full dimensional kernels*, it is the ℓ_1 norm of Fourier coefficients of the linear map. Yun et al. [YKM21] further showed a general connection for linear networks between implicit ℓ_1 norm minimization in an orthonormal basis to the existence of data-independent diagonalization of the linear operator in each layer. Differences between fully connected and convolutional networks trained with gradient descent has also been empirically demonstrated in [ZBHMS20] for the case of training with a single sample. A related complexity measure was also studied recently for infinite width two layer ReLU neural networks [SESS19; OWSS20], where the representation cost was related to the ℓ_1 norm of the partial derivatives of the Radon transforms of the function.

In our paper, we build on the results of Gunasekar et al. [GLSS18b], and investigate the induced regularizers (*i.e.,* the representation costs) of two-layer linear networks with a multi-channel convolutional layer. Our goal is to understand how the size of the convolutional kernel K and the number of output and input channels (denoted as C and R, respectively) impact the representation cost for linear convolutional networks with D dimensional inputs. These choices can play a significant role: for example, in extreme kernel size of $K = 1$, the representation cost turns out to be the ℓ_2 norm of the linear function, which is fundamentally different from the ℓ_1 norm of Fourier coefficients shown in Gunasekar et al. [GLSS18b] for $K = D$. Our main technical results are the following:

1. *Single-channel inputs.* Our main result for inputs with a single-channel is that for any kernel size K, the induced regularizer over linear functions is a norm that is independent of the number of output channels C, and can be expressed as a semi-definite program (SDP). The main technical ingredient is showing tightness of an SDP relaxation that is independent of C (see Theorem 4). Consequently, as we increase the kernel size K, the induced regularizer interpolates between the ℓ_2 norm (for $K = 1$) and the ℓ_1 norm of the Fourier transform (for $K = D$), with larger kernels promoting sparsity in the frequency domain (see Section 4).
2. *Multi-channel inputs.* In contrast, we show that for multi-channel inputs of dimensions $D \times R$, even realizing all linear functions over the inputs can require multiple output channels. Hence the representation cost *does* depend on the number of output channels C. We again show an SDP relaxation that is independent of C but the relaxation is not always tight. In the special cases of $K = 1$ and $K = D$, we derive the induced regularizer over the matrix-valued linear functions as the nuclear norm and the $\ell_{2,1}$ group sparse norm of the Fourier coefficients, respectively (see Theorems 9-10). In both cases, the induced regularizers promote non-trivial sparse structures in the space of multi-channel linear maps. (See Section 5.)
3. *Experiments for gradient descent.* Although we do not directly study gradient descent beyond the connection discussed in Section 1.1, we provide experiments to show that our theoretical findings do indeed extend to the implicit bias from gradient descent. On the MNIST dataset, we show that the limiting behavior of gradient descent is (a) independent of the number of output channels (this interestingly also appears to hold for ReLU activations), and (b) validate that large kernel sizes promote sparsity in the Fourier domain (see Section 6).

1.1 Connections to implicit regularization

The implicit inductive biases introduced by optimization algorithms like stochastic gradient descent play a crucial role in learning overparameterized neural networks [NTS15b; NTS15a; KMNST17; ZBHRV17]. In our work, we do not directly study the implicit biases from gradient descent variants. However, the inductive bias of controlling the ℓ_2 norm of weights or parameters is strongly motivated by the recent findings in the study of implicit bias.

An increasing body of literature on understanding overparameterized models show that in many instances, gradient descent updates lead to solutions that implicitly minimize the ℓ_2 of the parameters of the model [NTS15a; GWBNS17; SHNGS18; GLSS18a; GLSS18b; JT18; JT19; NGLSS19; LL20; JT20]. In particular, the recent work by Lyu and Li [LL20] (see also Nacson et al. [NGLSS19] and Ji and Telgarsky [JT20]) show such a connection between implicit bias and ℓ_2 norm minimization for all well behaved positive homogeneous model classes, which covers many common neural networks including fully connected and convolutional networks with linear or ReLU activations. We first describe the main result of Lyu and Li [LL20] that bridges a connection between the implicit bias from gradient descent and our induced complexity measure: Consider a function class represented as $\Phi(\boldsymbol{\theta}; \mathbf{x})$ for inputs denoted as $\mathbf{x}$ and parameters (or weights) denoted as $\boldsymbol{\theta}$.

Theorem. *[Paraphrased from LL20] Assume that Φ is locally Lipschitz and positive homogeneous with order $L > 0$, i.e., $\forall_{\boldsymbol{\theta}, \alpha > 0}$, $\Phi(\alpha\boldsymbol{\theta}; .) = \alpha^L \Phi(\boldsymbol{\theta}; .)$. Consider minimizing an exponential-tailed loss over a separable binary classification dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Under assumptions of loss convergence, gradient flow on this loss converges in direction to a first order stationary point (KKT point) of the following max–ℓ_2 margin problem in parameter space:*

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 \quad s.t., \quad \forall_n y_n \Phi(\boldsymbol{\theta}; \mathbf{x}_n) \geq 1. \tag{1}$$

It is easy to see that in the function space, we can define an induced regularizer $\mathcal{R}_\Phi$ such that the minimizers of the *max–ℓ_2 margin* problem over parameters in eq. (1) are equivalent to the minimizers of following *max–$\mathcal{R}_\Phi$ margin* problem over functions: $\min_f \mathcal{R}_\Phi(f)$ s.t., $\forall_n y_n f(\mathbf{x}_n) \geq 1$. The induced regularizer $\mathcal{R}_\Phi$ corresponding to controlling ℓ_2 norm of parameters is given as follows:

$$\mathcal{R}_\Phi(f) := \inf_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 \text{ s.t., } \forall \mathbf{x}, f(\mathbf{x}) = \Phi(\boldsymbol{\theta}, \mathbf{x}). \tag{2}$$

As such we expect that the real implicit inductive bias from gradient descent is captured through $\mathcal{R}_\Phi(f)$. However, there is an important caveat: the theorem by Lyu and Li [LL20] shows convergence of gradient flow direction to only a *stationary point* of the optimization in eq. (1), which might not be a global minimum; and in fact it need not even be a stationary point of the max–$\mathcal{R}$ margin problem in the function space. In spite of these caveats, the ℓ_2 norm of the weights is an important measure of complexity and thus naturally motivates the study of $\mathcal{R}_\Phi(f)$ in different architectures.

Our goal in this work is to study $\mathcal{R}_\Phi(f)$ for linear functions expressed as linear convolutional networks. Even though the connection of our results to implicit bias from gradient descent is tenuous, we show in simple experiments in Section 6 that the findings from our analysis are indeed observed on predictors learned by gradient descent.

Notation. We typeface vectors, matrices, and tensors using bold characters, *e.g.,* $\mathbf{v}, \mathbf{x}, \boldsymbol{\theta}, \mathbf{W}, \boldsymbol{\mathcal{U}}$. We will use zero-based indexing and python style slicing notation to specify the sub-entries of an array variable, *e.g.,* for $\mathbf{Z} \in \mathbb{R}^{D_1 \times D_2}$ and $\forall_{d_1 \in [D_1]}, \forall_{d_2 \in [D_2]}$ (where $[D] = \{0, 1, \ldots, D-1\}$), the individual entries of $\mathbf{Z}$ are denoted as $\mathbf{Z}[d_1, d_2]$, the d_1^{th} row as $\mathbf{Z}[d_1, :] \in \mathbb{R}^{d_2}$, and the d_2^{th} column as $\mathbf{Z}[:, d_2] \in \mathbb{R}^{d_1}$. We also briefly mention some standard notation for complex numbers. Complex numbers are specified in the polar form as $z = |z| \, \mathrm{e}^{i\phi_z}$ with $|z| \in \mathbb{R}_+$ and $\phi_z \in [0, 2\pi)$ denoting the

magnitude phase, respectively; and in Cartesian form as $z = \mathrm{Re}(z) + i\,\mathrm{Im}(z)$, where $\mathrm{Re}(z), \mathrm{Im}(z) \in \mathbb{R}$ denote the real and imaginary components, respectively ($i = \sqrt{-1}$ is the imaginary unit). The complex conjugate is denoted as $\overline{z} = |z|\,\mathrm{e}^{-i\phi_z}$. The standard inner product between $\mathbf{a}, \mathbf{b} \in \mathbb{C}^D$ is $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \overline{\mathbf{b}} = \sum_{d=0}^{D-1} \mathbf{a}[d]\overline{\mathbf{b}[d]}$, and analogously extends to matrices.

Unless qualified otherwise, we use $\|.\|$ to denote the standard Euclidean norm, *i.e.,* ℓ_2 norm for vectors and Frobenius norm for matrices. For arrays $\mathbf{a}, \mathbf{b}$, $\mathbf{a} \odot \mathbf{b}$ denotes entry-wise multiplication. Finally, we define the convolution operator $\star$ as it is used in the neural networks literature[1]:

Definition 1 (Circular convolution)**.** *For $\mathbf{u} \in \mathbb{R}^K$ and $\mathbf{v} \in \mathbb{R}^D$ with $K \leq D$, their circular convolution, denoted by $\mathbf{u} \star \mathbf{v}$, is a vector in $\mathbb{R}^D$ given as follows:*

$$\forall_{d \in [D]}, (\mathbf{u} \star \mathbf{v})[d] = \frac{1}{\sqrt{D}} \sum_{k=0}^{K-1} \mathbf{u}[k]\mathbf{v}[(d+k) \ mod \ D],$$

where mod refers to the modulo operation for integer division, s.t., p mod $D = p - D \left\lfloor \frac{p}{D} \right\rfloor$.

2 Multi-channel linear convolutional network

We consider two layer linear convolutional networks with multiple channels in the convolution layer. We first focus on multi-output channel convolutions with single channel inputs described below. We will discuss networks with multi-channel inputs (*e.g.,* RGB color channels) in Section 5.

We consider signals of dimension D, denoted as $\mathbf{x} \in \mathbb{R}^D$, as input to the network[2]. The first layer is a multi-output channel convolutional layer with kernel size K and number of output channel C whose weights (parameters) are denoted by $\mathbf{U} \in \mathbb{R}^{K \times C}$. The output of the convolution layer on input $\mathbf{x}$ is denoted as $h(\mathbf{U}; \mathbf{x}) \in \mathbb{R}^{D \times C}$, and is given by:

$$\forall_{c \in [C]}, \ h(\mathbf{U}; \mathbf{x})[:, c] = \mathbf{U}[:, c] \star \mathbf{x}, c = 0, 1, \ldots C-1, \tag{3}$$

where $\star$ denotes the convolutional operator in Definition 1[3].

The second layer is a single output linear layer with weights denoted by $\mathbf{V} \in \mathbb{R}^{D \times C}$. Thus, for input $\mathbf{x}$, the output of the network $\Phi(\mathbf{U}, \mathbf{V}; \mathbf{x})$ with weights $(\mathbf{U}, \mathbf{V})$ is given by:

$$\Phi(\mathbf{U}, \mathbf{V}; \mathbf{x}) = \langle \mathbf{V}, h(\mathbf{U}; \mathbf{x}) \rangle = \sum_{c=0}^{C-1} \langle \mathbf{V}[:, c], \mathbf{U}[:, c] \star \mathbf{x} \rangle. \tag{4}$$

Since, the network described above does not have any non-linearity, the function computed by the network can be equivalently represented by $w(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^D$ such that $\forall \mathbf{x}$, $\Phi(\mathbf{U}, \mathbf{V}; \mathbf{x}) = \langle w(\mathbf{U}, \mathbf{V}), \mathbf{x} \rangle$. Using simple algebraic manipulations on eq. (4), one can derive $w(\mathbf{U}, \mathbf{V})$ as follows:

$$w(\mathbf{U}, \mathbf{V}) = \sum_{c=0}^{C-1} \left(\mathbf{U}[:, c] \star \mathbf{V}[:, c]^\downarrow \right)^\downarrow, \tag{5}$$

where $\mathbf{z}^\downarrow$ denotes the flipped vector of $\mathbf{z} \in \mathbb{R}^D$, given by $\mathbf{z}^\downarrow[d] = \mathbf{z}[D - d - 1]$ for $d = 0, 1, \ldots D-1$.

Remark. *Even for the smallest network in this class with $K = C = 1$, any linear predictor $\mathbf{w} \in \mathbb{R}^D$ can realized as $w(\mathbf{U}, \mathbf{V})$ in eq. (5) (e.g., using $\mathbf{U} = 1, \mathbf{V} = \mathbf{w}$). In fact, every linear predictor can be represented by multiple networks with different weights $\mathbf{U}, \mathbf{V}$.*

[1] In the signal processing terminology Definition 1 is known as cross-correlation. We use a non-standard scaling of $1/\sqrt{D}$ only to simplify notation, and it does not change the results/analysis.

[2] For simplicity we consider 1D signals $\mathbf{x} \in \mathbb{R}^D$, but all our results and analysis can be extended to 2D inputs $\mathbf{x} \in \mathbb{R}^{W \times H}$, such as images, with the corresponding 2D convolutional operator.

[3] Our convolution definition corresponds to circular padding, which simplifies analysis. For convolutions with zero-padding, there will be different edge effects, but we expect qualitatively similar behavior for small padding sizes.

Induced regularizer in the function space. For the network Φ described above we now turn to understanding the inductive bias in the function space, or the induced regularizer, that corresponds to controlling the ℓ_2 norm of the weights. As remarked earlier, the function class realized by Φ is exactly set of linear predictors in $\mathbb{R}^D$. Thus, for any $\mathbf{w} \in \mathbb{R}^D$ the induced regularizer or the function space representation cost is given as follows:

$$\mathcal{R}_{K,C}(\mathbf{w}) := \min_{\mathbf{U}\in\mathbb{R}^{K\times C}, \mathbf{V}\in\mathbb{R}^{D\times C}} \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad w(\mathbf{U},\mathbf{V}) = \mathbf{w}. \tag{6}$$

We recall our discussion in Section 1.1 that $\mathcal{R}_{K,C}(\mathbf{w})$ captures properties of implicit and explicit regularization and our goal is to understand this induced measure for different choices of K and C.

Remark 1. *It immediately follows from eq. (6) are that $\mathcal{R}_{K,C}(\mathbf{w})$ is weakly decreasing in both K and C, e.g., for all C, $\mathcal{R}_{1,C}(\mathbf{w}) \geq \mathcal{R}_{2,C}(\mathbf{w}) \geq \ldots \mathcal{R}_{D,C}(\mathbf{w})$.*

2.1 Fourier domain representation

The convolution operation in Definition 1 permits a simple formulation in the Fourier domain arising from the Convolution Theorem. We briefly describe this connection, which we extensively use in our analysis and results. Throughout the paper, unless specified otherwise, we consider discrete Fourier transforms (DFTs) over the $\mathbb{R}^D$ space. Let $\mathbf{F} \in \mathbb{C}^{D\times D}$ denote the *unitary DFT matrix* for $\mathbb{R}^D$, *i.e.,* $\mathbf{F}[k,l] = \frac{1}{\sqrt{D}} e^{\frac{-2\pi i k l}{D}}$ for $0 \leq k,l < D$; and for any $1 \leq K \leq D$, let $\mathbf{F}_K \in \mathbb{C}^{D\times K}$ denote the submatrix of $\mathbf{F}$ with the first K columns. For a vector $\mathbf{a} \in \mathbb{R}^K$, we denote its D dimensional Fourier representation as $\widehat{\mathbf{a}} = \mathbf{F}_K \mathbf{a} \in \mathbb{C}^D$. Adapting the Convolution Theorem to our convolutional operator in Definition 1, we have the following relation in the Fourier domain: $\mathbf{F}(\mathbf{a} \star \mathbf{b}) = \overline{\widehat{\mathbf{a}}} \odot \widehat{\mathbf{b}}$, where recall that $\overline{\mathbf{z}}$ denotes the complex conjugate of $\mathbf{z}$ and $\odot$ denotes entrywise multiplication.

The discrete Fourier transform $\widehat{w}(\mathbf{U},\mathbf{V}) := \mathbf{F}w(\mathbf{U},\mathbf{V})$ of the linear predictor $w(\mathbf{U},\mathbf{V})$ realized by our network (eq. 5) can now be expressed as follows: Let $\widehat{\mathbf{U}} = \mathbf{F}_K\mathbf{U} \in \mathbb{C}^{D\times C}$ and $\widehat{\mathbf{V}} = \mathbf{F}\mathbf{V} \in \mathbb{C}^{D\times C}$ denote the D dimensional Fourier representation of $\mathbf{U},\mathbf{V}$, respectively. We have,

$$\widehat{w}(\mathbf{U},\mathbf{V}) = \sum_{c=0}^{C-1} \widehat{\mathbf{U}}[:,c] \odot \widehat{\mathbf{V}}[:,c] = \mathrm{diag}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top). \tag{7}$$

Before we present our main results for multi-output channel networks in Section 4, we briefly discuss some special cases which highlight how the induced regularizer changes with kernel size in single output channel networks.

3 Induced regularizer $\mathcal{R}_{K,C}$ in special cases

Consider networks with single output channel ($C = 1$): the weights are $\mathbf{U} \in \mathbb{R}^K, \mathbf{V} \in \mathbb{R}^D$ and the Fourier coefficients of the linear predictor is given by $\widehat{w}(\mathbf{U},\mathbf{V}) = \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}}$.

For any kernel size K, we can obtain a lower bound on the induced regularizer as $\mathcal{R}_{K,1}(\mathbf{w}) \geq 2\|\widehat{\mathbf{w}}\|_1$ using $\forall_{d\in D}, |\widehat{\mathbf{U}}[d]|^2 + |\widehat{\mathbf{V}}[d]|^2 \geq 2|\widehat{\mathbf{U}}[d] \cdot \widehat{\mathbf{V}}[d]|$. In the case of full dimensional convolutional layer with $K = D$, Gunasekar et al. [GLSS18b] showed that this lower bound is indeed tight:

Lemma 1 ($K = D, C = 1$)**.** *[Lemma 7 in GLSS18b] For any $\mathbf{w} \in \mathbb{R}^D$, $\mathcal{R}_{D,1}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$.*

The proof of Lemma 1 uses the fact that for $K = D$, the weights $\mathbf{U},\mathbf{V} \in \mathbb{R}^D$ are unconstrained and we can choose them to have $|\widehat{\mathbf{U}}[d]| = |\widehat{\mathbf{V}}[d]| = \sqrt{|\widehat{\mathbf{w}}[d]|}$, $\forall_{d\in D}$ leading to the optimal cost of $2\|\widehat{\mathbf{w}}\|_1$.

For networks with smaller kernels $K < D$, the core difference is that $\widehat{\mathbf{U}} = \mathbf{F}_K \mathbf{U}$ is constrained to be in a K dimensional subspace spanned by the columns of $\mathbf{F}_K$. Thus, it is not always possible to choose weights satisfying $|\widehat{\mathbf{U}}[d]| = \sqrt{|\widehat{\mathbf{w}}[d]|}$. It is easiest to see this is case of kernel size $K = 1$, where $\mathbf{U} = u_0 \in \mathbb{R}^1$ is a scalar and hence $\widehat{\mathbf{U}} \propto [1, 1, \ldots, 1]$. Thus, in order to satisfy $\widehat{\mathbf{w}} = \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}}$, we need $\widehat{\mathbf{V}} \propto \widehat{\mathbf{w}}$. By choosing u_0 optimally, we can calculate the induced complexity measure for $K = 1$ as the ℓ_2 norm (full proof is provided in the Appendix A):

Lemma 2 ($K = 1, C = 1$)**.** *For any* $\mathbf{w} \in \mathbb{R}^D$*, it holds that* $\mathcal{R}_{1,1}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\| = 2\sqrt{D}\|\mathbf{w}\|$.

Notice that the induced regularizer behaves fundamentally differently at $K = D$ and $K = 1$: the ℓ_2 regularization of $\mathcal{R}_{D,1}(\mathbf{w})$ does not induce sparse solutions, while the ℓ_1 regularization of $\mathcal{R}_{D,1}(\mathbf{w})$ promotes sparsity in the frequency domain. We show in Section 4, that for general K, the induced regularizer is a norm, thus interpolating between ℓ_2 and ℓ_1 norms in the Fourier domain.

Since $K = 1$ and $K = D$ permit closed-form solutions in the Fourier space, one might hope to get similar clean characterizations to other kernel sizes as well. Unfortunately, even for $K = 2$, we show that $\mathcal{R}_{2,1}(\mathbf{w})$ takes a much more complex form as described below (proof in Appendix A):

Lemma 3. *For any* $\mathbf{w} \in \mathbb{R}^D$*, it holds that:*

$$\mathcal{R}_{2,1}(\mathbf{w}) = 2\sqrt{D}\sqrt{\inf_{\alpha \in (-1,1)} \sum_{d=0}^{D-1} \frac{|\widehat{\mathbf{w}}[d]|^2}{1 + \alpha \cos\left(2\pi d / D\right)}}.$$

The expression in Lemma 3 involves a maximization over a high-degree rational function, and is thus unlikely to admit clean closed-form solutions (at least in the Fourier space).

Although Lemma 3 does not yield closed form solutions for $\mathcal{R}_{2,1}(\mathbf{w})$, we observe that it hints at some form of band-pass frequency structure: for any α from the inner optimization, the resulting regularizer is a weighted sum of Fourier coefficients such that the nearby frequency components of $\widehat{\mathbf{w}}$ are weighted with nearby values. This band-pass nature was also observed in a complementary result by Yun et al. [YKM21] in the context of implicit bias from gradient descent on a single data point: for any $\mathbf{x}$, it was shown that $\min_{\mathbf{w}} \mathcal{R}_{2,1}(\mathbf{w})$ s.t., $\mathbf{w}^\top \mathbf{x} > 1$ corresponds to a low-pass or high pass filter depending on the sign of $\mathbf{x}^\top \mathbf{x}^{\downarrow}$.

Even though we do not obtain closed form solutions for all kernel sizes K, we derive important properties about the induced regularizer for general kernel sizes in the following sections.

4 Main technical tool: SDP formulation of induced regularizer

In this section, we show that for two layer linear convolutional networks on inputs with single channel, the induced regularizer $\mathcal{R}_{K,C}$ is equivalent to a semidefinite program (SDP) that only depends upon on the kernel size K. As a consequence, we can infer that $\mathcal{R}_{K,C}$ is a norm independent of C. Additionally, although the optimization in eq. (6) is non-convex, the SDP formulation allows us to efficiently compute $\mathcal{R}_{K,C}(\mathbf{w})$ exactly.

We first reformulate $\mathcal{R}_{K,C}$ as an SDP with a rank constraint. This immediately motivates an SDP relaxation that drops the rank constraint and provides a lower bound on $\mathcal{R}_{K,C}$. In Theorem 4, we show that the SDP relaxation thus obtained is in fact tight for any kernel size K.

***$\mathcal{R}_{K,C}$* as an SDP with a rank constraint.** Combining the definition of $\mathcal{R}_{K,C}(\mathbf{w})$ in eq. (6) with the Fourier representation of $w(\mathbf{U}, \mathbf{V})$ in eq. (7), we have the following:

$$\begin{aligned} \mathcal{R}_{K,C}(\mathbf{w}) = \min_{\mathbf{U} \in \mathbb{R}^{K \times C}, \mathbf{V} \in \mathbb{R}^{D \times C}} \quad & \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \\ \text{s.t.,} \quad & \operatorname{diag}(\widehat{\mathbf{U}} \widehat{\mathbf{V}}^\top) = \widehat{\mathbf{w}}. \end{aligned} \tag{8}$$

We observe that the objective of the optimization in eq. (8) is given by $\|\mathbf{U}\|^2+\|\mathbf{V}\|^2 = \langle \mathbf{U}\mathbf{U}^\top, \mathbf{I}\rangle + \langle \mathbf{V}\mathbf{V}^\top, \mathbf{I}\rangle$. The constraints are given by: $\forall_{d\in[D]}$, $\langle \widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top, \mathbf{e}_d\mathbf{e}_d^\top\rangle = \widehat{\mathbf{w}}[d]$, where $\{\mathbf{e}_d\}_{d\in[D]}$ denotes the standard basis in $\mathbb{R}^D$. These constraints can alternatively be written as $\langle \mathbf{U}\mathbf{V}^\top, \mathbf{Q}_d\rangle = \widehat{\mathbf{w}}[d]$, where $\mathbf{Q}_d := \overline{\mathbf{F}}_K^\top \mathbf{e}_d\mathbf{e}_d^\top \overline{\mathbf{F}} \in \mathbb{C}^{K\times D}$, for $d \in [D]$.

The optimization in eq. (8) over $\mathbf{U} \in \mathbb{R}^{K\times C}, \mathbf{V} \in \mathbb{R}^{C\times D}$, can thus be specified in terms of a rank C positive semi-definite matrix $\mathbf{Z} \in \mathbb{R}^{(D+K)\times(D+K)}$ that we define below:

$$\mathbf{Z} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}[\mathbf{U}^\top \quad \mathbf{V}^\top] = \begin{bmatrix}\mathbf{U}\mathbf{U}^\top & \mathbf{U}\mathbf{V}^\top\\ \mathbf{V}\mathbf{U}^\top & \mathbf{V}\mathbf{V}^\top\end{bmatrix} \succcurlyeq 0. \tag{9}$$

The objective and constraints of eq. (8) can now be expressed as linear functions of $\mathbf{Z}$ as $\langle \mathbf{Z}, \mathbf{I}\rangle$, and $\forall_{d\in[D]}$, $\langle \mathbf{Z}, \mathbf{A}_d^{\text{real}}\rangle = 2\cdot \text{Re}(\widehat{\mathbf{w}}[d])$ and $\langle \mathbf{Z}, \mathbf{A}_d^{\text{img}}\rangle = 2\cdot \text{Im}(\widehat{\mathbf{w}}[d])$, respectively, where we define $(\mathbf{A}_d^{\text{real}}, \mathbf{A}_d^{\text{img}})$ as follows:

$$\mathbf{A}_d^{\text{real}} = \begin{bmatrix}0_K & \mathbf{Q}_d\\ \overline{\mathbf{Q}}_d^\top & 0_D\end{bmatrix} \text{ and } \mathbf{A}_d^{\text{img}} = \begin{bmatrix}0_K & i\cdot\mathbf{Q}_d\\ -i\cdot\overline{\mathbf{Q}}_d^\top & 0_D\end{bmatrix}.$$

Now, we can formulate $\mathcal{R}_{K,C}(\mathbf{w})$ as follows:

$$\begin{aligned}\mathcal{R}_{K,C}(\mathbf{w}) = \min_{\mathbf{Z}\succcurlyeq 0}\ & \langle \mathbf{Z}, \mathbf{I}\rangle\\ \text{s.t.},\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{real}}\rangle = 2\,\text{Re}(\widehat{\mathbf{w}}[d])\\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{img}}\rangle = 2\,\text{Im}(\widehat{\mathbf{w}}[d])\\ & \text{rank}(\mathbf{Z}) \le C.\end{aligned} \tag{10}$$

The formulation in eq. (10) is non-convex due to the rank constraint. We obtain a natural convex relaxation by dropping the rank constraint, leading to the following SDP:

$$\begin{aligned}\mathcal{R}_K^{\text{SDP}}(\mathbf{w}) = \min_{\mathbf{Z}\succcurlyeq 0}\ & \langle \mathbf{Z}, \mathbf{I}\rangle\\ \text{s.t.},\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{real}}\rangle = 2\,\text{Re}(\widehat{\mathbf{w}}[d])\\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{img}}\rangle = 2\,\text{Im}(\widehat{\mathbf{w}}[d]).\end{aligned} \tag{11}$$

This SDP serves as the main technical tool in our analysis of $\mathcal{R}_{K,C}(\mathbf{w})$. An immediate consequence of this relaxation is that it provides lower bounds on the induced regularizer.

Remark. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) \ge \mathcal{R}_K^{SDP}(\mathbf{w})$.*

Remark. *The symmetry properties of Fourier coefficients of real signals gives us that for any D and any $\mathbf{w} \in \mathbb{R}^D$, $\widehat{\mathbf{w}}[p] = \overline{\widehat{\mathbf{w}}[D-p]}$ for $p \in [D]$. Thus, although the optimization problems in (10) and (11) are specified with $2\cdot D$ constraints for simplicity, only D of them are unique.*

We next show that the SDP relaxation is in fact tight.

4.1 Tightness of the SDP Relaxation

Theorem 4. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) = \mathcal{R}_K^{SDP}(\mathbf{w})$.*

The main technical contribution in the proof of Theorem 4 is to show that given any minimizer $\mathbf{Z}$ of the SDP in eq. (11), we can implicitly construct a rank-1 solution that has the same objective value as $\mathbf{Z}$ and satisfies the SDP constraints. The key proof technique that we use is the polynomial representation of the convolution operation, which enables us to reason about the factorization of relevant polynomials over the complex numbers to implicitly construct a rank 1 solution. The full proof of Theorem 4 can be found in the Appendix B. □

In the remainder of the section, we discuss some interesting consequences of the tightness result in Theorem 4. Proofs of the following results are in Appendix C.

4.1.1 $\mathcal{R}_{K,C}$ is independent of C for single-channel inputs

Theorem 4 implies that $\mathcal{R}_{K,C}(\mathbf{w}) = \mathcal{R}_{K,1}(\mathbf{w})$ for all C. The induced regularizer $\mathcal{R}_{K,C}(\mathbf{w})$ is thus independent of C for single-channel inputs. This generalizes Lemmas 1–3 to networks with multiple output channels, *e.g.,* we now have,

$$\forall C, \quad \mathcal{R}_{1,C}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2 \text{ and } \mathcal{R}_{D,C}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1.$$

Theorem 4 also has implications for implicit bias from gradient descent. Recall the discussion in Section 1.1 that is *suggestive* that implicit bias from gradient descent is related to the minimizing $\mathcal{R}_{K,C}(\mathbf{w})$. Our result thus *suggests* that for linear networks with single input channels, the number output channels does not influence the asymptotic implicit bias from gradient descent. We add that this conclusion is subject to the caveats we discussed earlier. Nevertheless, in Section 6 we empirically observe that the predictors obtained from gradient descent are indeed invariant to C.

4.1.2 $\mathcal{R}_{K,C}$ is a norm

In addition to implying that the induced regularizer is independent for C for single-channel inputs, Theorem 4 also provides a technical tool to show structural properties of $\mathcal{R}_{K,C}$.

Corollary 5. *For $K \leq D$ and any C, $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm.*

For the end cases of $K = 1$ and $K = D$, this norm can be explicitly specified: $\mathcal{R}_{1,C}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|$ (Lemma 2) and $\mathcal{R}_{D,C}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$ (Lemma 1), respectively. For intermediate kernel sizes, Corollary 5 shows that the $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm that interpolates between the ℓ_2 norm and the ℓ_1 norm of the Fourier coefficients of the linear predictor. We can further use the SDP formulation in (11) to compute the following upper and lower bounds on $\mathcal{R}_{K,C}(\mathbf{w})$.

Lemma 6. *For any $K \leq D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$:*

$$2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$$

$$2\|\widehat{\mathbf{w}}\|_1 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{\left\lceil \frac{D}{K} \right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

Remark. *For the lower bounds, $2\|\widehat{\mathbf{w}}\|_1$ is tight for $\mathbf{w} = [1, 0, \ldots, 0]$ and $2\sqrt{\frac{D}{K}}\|\mathbf{w}\|$ is tight for $\mathbf{w} = [1, 1, \ldots, 1]$. For the upper bounds, $2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$ is tight when $K \mid D$ for patterned vectors (see Lemma 7), and $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$ is tight for $[1, 0, \ldots, 0]$.*

Lemma 6 demonstrates that when K is a small constant, $\mathcal{R}_{K,C}(\mathbf{w})$ is multiplicatively close to $\mathcal{R}_{1,C}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$. On the other hand, once K is comparable to D, $\mathcal{R}_{K,C}(\mathbf{w})$ is within a constant factor of $\mathcal{R}_{D,C}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$.

4.1.3 $\mathcal{R}_{K,C}(\mathbf{w})$ for patterned vectors

Aside from general bounds on the $\mathcal{R}_{K,C}$, the SDP formulation can also be used to analyze the behavior of the induced regularizer of special classes of vectors. One interesting case is of *patterned vectors* described as follows: Consider vectors of the form $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ comprised of repetitions of a P dimensional pattern $\mathbf{p} \in \mathbb{R}^P$. Linear predictors with such repeated patterns incorporate the useful property of invariance to periodic translations.

We show an interesting relation between the representation cost $\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p}))$ of realizing patterned vectors in $\mathbb{R}^D$ and the analogous cost (denoted as $\mathcal{R}^{(P)}_{K,C}(\mathbf{p})$) of realizing $\mathbf{p}$ as a linear predictor in $\mathbb{R}^P$ using a network with the same values of K and C.

Lemma 7. *Consider vectors $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ specified by $\mathbf{p} \in \mathbb{R}^P$ s.t., P divides D.*
(a) For any $K \leq P$, it holds that $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ if $K = P \cdot T$ for integer T, then $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = 2\frac{D}{\sqrt{T}P}\|\widehat{\mathbf{p}}\|_1 = 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_1.$$

We see that the induced regularizer of repeated patterned vectors is closely related to that of the pattern itself. In particular, for $K \leq P$, we have $\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \propto \mathcal{R}^{(P)}_{K,1}(\mathbf{p})$.

5 Multi-input channel networks

In Sections 2-3, we demonstrated that for single-channel inputs $\mathbf{x} \in \mathbb{R}^D$, for any kernel size, all linear maps $\mathbf{w} \in \mathbb{R}^D$ can be realized by a network with single output channel convolutional layer; furthermore, $\mathcal{R}_{K,C}(\mathbf{w})$ is independent of the number of output channels C. In this section, we expand our results to networks with multiple input channels (*e.g.,* RGB color channels). How do these conclusions change for multiple input channels? How does the induced regularizer, now denoted as $\mathcal{R}_{K,C,R}$, depend on the number input channels R?

We again consider two layer convolutional networks akin to Section 2. We first introduce additional notation: The multi-channel inputs are denotes as $\mathbf{X} \in \mathbb{R}^{D\times R}$, where R denotes the number of input channels. The convolutional first layer now has kernel size K, output channel size C and input channel size R with weights denoted by a set of R matrices $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r\}_{r\in[R]}$ with $\mathbf{U}_r \in \mathbb{R}^{K\times C}$. The output of this convolution layer $h(\boldsymbol{\mathcal{U}}; \mathbf{X}) \in \mathbb{R}^{D\times C}$ is given as follows:

$$\forall_{c\in C},\ h(\boldsymbol{\mathcal{U}}; \mathbf{X})[:, c] = \sum_{r=0}^{R-1} \mathbf{U}_r[:, c] \star \mathbf{X}[:, r]. \tag{12}$$

The second layer is the same as before: a single output linear layer with weights $\mathbf{V} \in \mathbb{R}^{D\times C}$. We denote the equivalent linear predictor for this network by $W(\boldsymbol{\mathcal{U}}, \mathbf{V}) \in \mathbb{R}^{D\times R}$. Following similar calculations as for single input channels, $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in signal and Fourier domain (denoted as $\widehat{W}(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \mathbf{F}W(\boldsymbol{\mathcal{U}}, \mathbf{V})$) are given as follows:

$$\begin{aligned} \forall_{r\in[R]},\ W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] &= \sum_{c=0}^{C-1} \Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow}, \\ \widehat{W}(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] &:= \operatorname{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^{\top}). \end{aligned} \tag{13}$$

For multi-channel inputs, the set of all linear predictors is the space of matrices $\mathbf{W} \in \mathbb{R}^{D\times R}$, and we define the induced complexity measure over this matrix space as follows:

$$\begin{aligned} \mathcal{R}_{K,C,R}(\mathbf{W}) := \inf_{\boldsymbol{\mathcal{U}},\mathbf{V}} & \sum_{r\in[R]} \|\mathbf{U}_r\|^2 + \|\mathbf{V}\|^2 \\ \text{s.t.,}\ & W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \mathbf{W}. \end{aligned} \tag{14}$$

5.1 Role of output channel size C

For multi-channel inputs, we first observe that multiple output channels can be necessary to merely realize all linear maps. The following lemma is proven in Appendix D.1 by showing that the sub-network corresponding to each output channel can realize a matrix in $\mathbb{R}^{D\times R}$ of rank at most K.

Lemma 8. *For any K, C and R, in order for the the network represented by $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in eq. (13) to realizes all linear maps in $\mathbb{R}^{D\times R}$ it is necessary that $K \cdot C \geq \min\{R, D\}$.*

In contrast to single input channels, Lemma 8 demonstrates that, the model class realized by linear convolutional networks over multi-channel inputs, and consequently the induced regularizer, *does* depend on number of output channels C. Nonetheless, similar to single input channel networks, we can again obtain an SDP relaxation $\mathcal{R}^{\text{SDP}}_{K,R}(\mathbf{W})$ for $\mathcal{R}_{K,C,R}(\mathbf{W})$ that is independent of C. This SDP relaxation is derived very similarly to the case of single input channels. In the interest of space, we defer the derivation of the relaxation to Appendix D.2.

Based on Lemma 8 we already know that the SDP is not always tight for multi-input channels $R > 1$. In the next subsection, we show that in the special cases of $K = 1$ and $K = D$, $\mathcal{R}_{K,C,R}(\mathbf{W})$ can be expressed as interesting closed form norms when C is large enough to realize all linear maps, and moreover, the SDP is tight for these choices of K as long as $K \cdot C \geq \min\{R, D\}$.

5.2 Induced regularizer when $K = 1$ and $K = D$

Theorem 9 (Multi-input channel $K = 1$)**.** *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq \min\{R, D\}$, the induced regularizer for $K = 1$ is given by the scaled nuclear norm $\|.\|_*$:*

$$\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_* = 2\sqrt{D}\|\widehat{\mathbf{W}}\|_*.$$

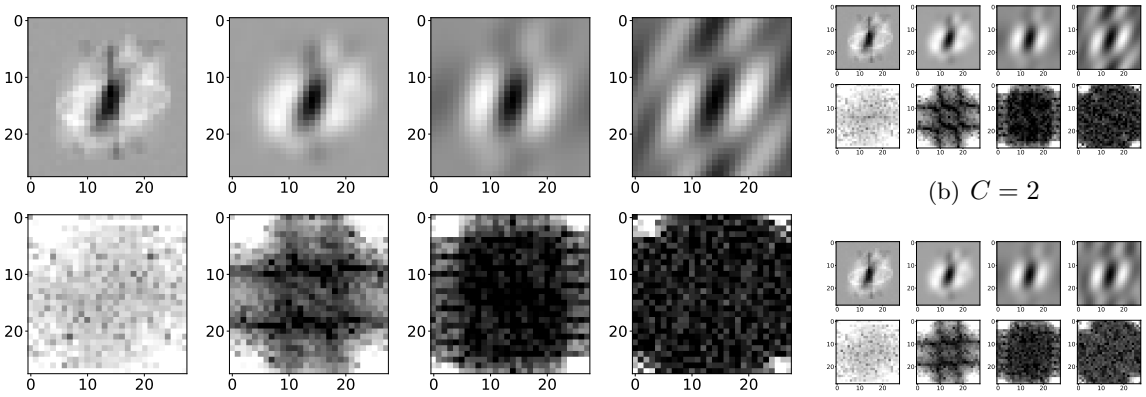
Theorem 10 (Multi-input channel $K = D$)**.** *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq 1$, the induced regularizer for $K = D$ is given as follows*

$$\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1} := \sum_{s=0}^{D-1} \sqrt{\sum_{r=0}^{R-1} \left|\widehat{\mathbf{W}}[d, r]\right|^2}.$$

From Theorems 9-10 it is evident that the number of input channels R fundamentally changes the nature of induced complexity measure in the function space and introduces additional structures along the input channels. Even in the simplest setting of scalar convolution kernels with $K = 1$, the induced regularizer is no longer a Euclidean or RKHS norm, but is a richer nuclear norm that encourages low-rank properties. For the case of $K = D$, the induced regularizer is group-sparse norm on the Fourier coefficients that encourages similar weighting across channels, while promoting sparsity across frequency components. In comparison to the ℓ_1 norm of all Fourier coefficients, this group-sparse norm is a more structured inductive bias for multi-channel inputs. Additionally, like with the single input channel case, we also observe that the induced bias has a more intuitive and interesting interpretation in Fourier domain which is not directly observed in the signal domain.

6 Experiments for gradient descent

We have thus far focused on analyzing the induced regularizer as a complexity measure. While our results suggest connections to implicit bias from gradient descent on separable classification tasks, our conclusions are subject to the caveats discussed in Section 1.1.



(a) $C = 1$ (b) $C = 2$ (c) $C = 4$

Figure 1: Linear predictors learned by two layer linear convolutional network for the task of classifying digits 0 and 1 in MNIST. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C = 1, 2, 4$ and kernel sizes $K \in \{(1,1), (3,3), (9,9), (28,28)\}$ (left to right). The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$. Higher resolution figures are in Appendix.

Consider predictors $w(\mathbf{U}, \mathbf{V})$ learned from gradient descent on linear convolutional networks with single channel inputs. If we overlook the caveats, we expect the gradient descent to implicitly learn a max–$\mathcal{R}_{K,C}$ margin predictor: $\min_{\mathbf{w}} \mathcal{R}_{K,C}(\mathbf{w})$ s.t., $\forall_n y_n \langle \mathbf{w}, \mathbf{x}_n \rangle \geq 1$. Based on the results in Section 4 we expect the following to hold (up to scaling): (a) for larger kernel sizes, $w(\mathbf{U}, \mathbf{V})$ would tend to be sparse in Fourier domain, and (b) $w(\mathbf{U}, \mathbf{V})$ is invariant to number of output channels C.

In this section, we empirically demonstrate that these findings do hold on linear predictors learned by gradient descent. Interestingly, we also find that invariance of $\mathcal{R}$ with number of output channels can also be seen for networks ReLU non-linearity.

Experimental setup. We run our experiments on two layer linear convolutional networks on a subset of MNIST dataset. The input images in MNIST are of size 28×28 and have a single input channel. We apply 2D convolutions with kernel sizes $K = (K_1, K_2)$ and circular padding for image inputs. We consider a binary classification task of predicting digits 0 and 1 in MNIST and we use a balanced sub-sampling of 128 samples as training data, which ensures linear separability. We train our network using gradient descent on exponential loss and run gradient descent until the training loss is 10^{-6}. In order to compare the predictors across different architectures, we normalize the weights learned by gradient descent $\mathbf{U}, \mathbf{V}$ such that the linear predictors $w(\mathbf{U}, \mathbf{V})$ realized by the trained networks have unit margin on the training dataset (*i.e.,* $y\langle w(\mathbf{U}, \mathbf{V}), \mathbf{x} \rangle \geq 1$ for all training samples $(\mathbf{x}, y)$). Note that for homogeneous models, such positive scaling of weights *does not* change the classification boundary of the learned model.[4]

6.1 Varying kernel sizes

We first empirically validate that larger kernel sizes induces sparsity in the frequency domain. To see this, we consider a network with one output channel $C = 1$ and compute $w(\mathbf{U}, \mathbf{V})$ learned by gradient descent for networks with different kernel sizes in Figure 1-(a). Notice in the frequency domain plots that the predictor learned with kernel size $K = (1, 1)$ is not sparse, the predictor

[4]The code is available at `https://github.com/mjagadeesan/inductive-bias-multi-channel-CNN`.

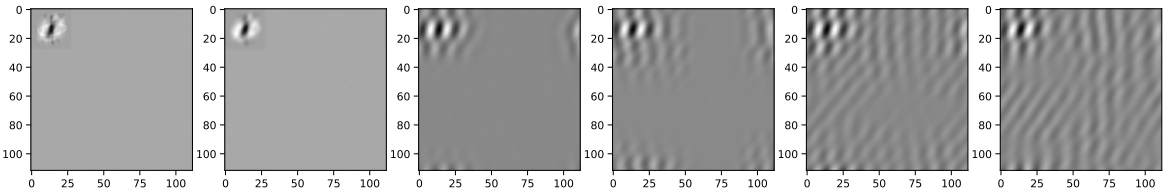


Figure 2: Linear predictors learned by gradient descent on single output channel networks over an augmented input space with kernel sizes $K \in \{(1,1),(3,3),(27,27),(45,45),(65,65),(84,84)\}$ (left to right). The input images from the MNIST dataset are augmented by padding with zeros to obtain an image of size 112×112 with the signal present only in the top-left 28×28 block.

learned with $K = (3,3)$ already starts to exhibit some sparsity, and the linear predictor learned with $K = (28,28)$ is highly sparse in the frequency domain.

Since sparsity in the frequency domain promotes a patterned structure in the signal domain, we explore the qualitative behavior of large kernel sizes in the signal domain in more depth. To do this, we construct an augmented version of the dataset with 112×112 dimensional images where the top-left 28×28 region is the original image, while the remaining space is all 0s. Figure 2 shows the linear predictors learned by running gradient descent on single output channel networks with different kernel sizes. As K increases, the nonzero region of the predictor becomes larger, eventually encompassing the full 112×112 dimensional space. For large kernel sizes, we can visually see that the predictors are composed of repetitions of a pattern. This is suggestive of a restricted form of periodic translation invariance, where the shift size aligns with the size of the patterns.

6.2 Impact of number of channels

We now empirically validate that for single channel inputs, the predictor $w(\mathbf{U},\mathbf{V})$ learned by gradient descent on linear convolutional networks is invariant to the number of output channels. To demonstrate this, we repeat the experimental setup on 28×28 MNIST images on networks with multiple output channels $C \in \{1,2,4,8\}$ and across different kernel sizes. As described earlier, we scale the weights learned by gradient descent $\mathbf{U},\mathbf{V}$ such that the linear predictors $w(\mathbf{U},\mathbf{V})$ have unit margin on training data. First, the learned linear predictors for $C = \{2,4\}$ are shown in Figure 1-(b,c) for comparison with $C = 1$ in Figure 1-(a). As suggested by our theory, we observe that the linear predictors visually appears to be invariant across different settings of C. For a more quantitative evaluation, in Table 1 we show the values of an approximation[5] of $\mathcal{R}_{K,C}(w(\mathbf{U},\mathbf{V}))$ given by $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U},\mathbf{V})) := \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$. Notice that these values of $\widehat{\mathcal{R}}_{K,C}(\mathbf{w})$ are fairly consistent across different settings of C for any fixed kernel size K. In contrast, these values vary significantly when the K is changed. We repeat these experiments for the classes 2 and 9 in Appendix E.

6.2.1 Impact of number of channels on non-linear networks with ReLU activation

Although our theoretical results are restricted to networks with linear activations, it is nevertheless interesting to evaluate if our conclusions lead to useful heuristics for networks with non-linearity (*i.e.,* a network with a convolutional layer, followed by ReLU layer, followed by linear layer). As a simple demonstration, we repeat our experiment on MNIST dataset on a two layer convolutional

[5]Strictly speaking, the estimate $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U},\mathbf{V}))$ is only an *upper bound* on $\mathcal{R}_{K,C}(w(\mathbf{U},\mathbf{V}))$. For $K = 1$ and $K = D$, we verified that the estimate is close to tight by computing the explicit max-margin solutions with respect to ℓ_2 and ℓ_1 norms of Fourier transform, respectively. Moreover, on smaller experiments with $1D$ data, we observed that $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U},\mathbf{V}))$ is numerically close to the calculation from the SDP formulation $\mathcal{R}^{\text{SDP}}_K(w(\mathbf{U},\mathbf{V}))$ for all K.

C	$K:(1,1)$	$K:(3,3)$	$K:(9,9)$	$K:(28,28)$
1	10.38	4.60	2.88	2.52
2	10.38	4.60	2.91	2.51
4	10.39	4.62	2.93	2.41
8	10.43	4.66	2.99	2.42

Table 1: $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U},\mathbf{V})) = \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ of the predictor learned by gradient descent on linear convolutional networks with different number of output channels C and kernel sizes K.

network with ReLU non-linearity. We consider the representation cost $\mathcal{R}_{\Phi_{K,C}}(f)$, as per equation (2), given by the ℓ_2 norm of the weights of the resulting predictor normalized by the margin. Table 2 shows an approximation of $\mathcal{R}_{\Phi_{K,C}}(f)$ given by $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f) := \|\mathbf{U}\|_2^2 + \|\mathbf{V}\|_2^2$ across different settings of C and K. Like in the case of linear convolutional neural networks, $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f)$ is fairly consistent across different settings of C. This suggests that the implicit bias from gradient might result in predictors that are independent of the number of output channels, even when there is a ReLU layer.

C	$K:(1,1)$	$K:(3,3)$	$K:(9,9)$	$K:(28,28)$
1	11.26	5.27	3.68	2.97
2	11.27	5.25	3.69	3.08
4	11.29	5.31	3.70	3.29
8	11.36	5.35	3.75	3.29

Table 2: $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f)$ of predictors learned by gradient descent on two layer convolutional networks with a ReLU nonlinearity and different choices of output channels C and kernel sizes K

7 Discussion and Future Work

Towards understanding neural networks, the representation cost provides an abstraction to separate capacity control in the parameter space from the function space view of the resulting inductive bias. The ℓ_2 representation in particular has extensive connections to both explicit and implicit bias. In this work, we showed that the kernel size and number of channels have interesting effects even within two-layer linear convolutional neural networks. Furthermore, we empirically validated our findings for gradient descent on simple convolutional neural networks, with and without ReLU activations. An interesting direction for future work would be to conduct an in-depth empirical investigation of the impact of nonlinearity. Moreover, it would be interesting to extend our theoretical findings in several ways, including: proving tightness of the SDP relaxation for multiple input channels, formally establishing the limiting behavior of gradient descent (without the caveats that we discussed), and exploring architectural features such as pooling or multiple layers.

References

[Bar96] Peter L. Bartlett. "For Valid Generalization the Size of the Weights is More Important than the Size of the Network". In: *Advances in Neural Information Processing Systems*. 1996.

[BFT17] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. "Spectrally-normalized margin bounds for neural networks". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6241–6250.

[BM02] Peter L. Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* (2002), pp. 463–482.

[GLSS18a] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. "Characterizing Implicit Bias in Terms of Optimization Geometry". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2018, pp. 1827–1836.

[GLSS18b] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nati Srebro. "Implicit Bias of Gradient Descent on Linear Convolutional Networks". In: *Advances in Neural Information Processing Systems*. 2018.

[GWBNS17] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. "Implicit Regularization in Matrix Factorization". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6151–6159.

[JT18] Ziwei Ji and Matus Telgarsky. "Risk and parameter convergence of logistic regression". In: *CoRR* abs/1803.07300 (2018).

[JT19] Ziwei Ji and Matus Telgarsky. "Gradient descent aligns the layers of deep linear networks". In: *International Conference on Learning Representations (ICLR)*. 2019.

[JT20] Ziwei Ji and Matus Telgarsky. "Directional convergence and alignment in deep learning". In: *Advances in Neural Information Processing Systems*. 2020.

[KH91] Anders Krogh and John A. Hertz. "A Simple Weight Decay Can Improve Generalization". In: *Advances in Neural Information Processing Systems*. 1991, pp. 950–957.

[KMNST17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *International Conference on Learning Representations (ICLR)*. 2017.

[KST08] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. "On the complexity of linear prediction: risk bounds, margin bounds, and regularization". In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2008, pp. 793–800.

[LL20] Kaifeng Lyu and Jian Li. "Gradient Descent Maximizes the Margin of Homogeneous Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2020.

[NGLSS19] Mor Shpigel Nacson, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. "Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019, pp. 4683–4692.

[NTS15a] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning". In: *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*. 2015.

[NTS15b] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "Norm-Based Capacity Control in Neural Networks". In: *Proceedings of the Conference on Learning Theory, (COLT)*. 2015, pp. 1376–1401.

[OWSS20] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case". In: *International Conference on Learning Representations (ICLR)*. 2020.

[RS05] Jason D. M. Rennie and Nathan Srebro. "Fast maximum margin matrix factorization for collaborative prediction". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2005, pp. 713–719.

[SESS19] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. "How do infinite width bounded norm networks look in function space?" In: *Proceedings of the Conference on Learning Theory (COLT)*. 2019, pp. 2667–2690.

[SHNGS18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The Implicit Bias of Gradient Descent on Separable Data". In: *Journal of Machine Learning Research* (2018).

[WLLM19] Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. "Regularization Matters: Generalization and Optimization of Neural Nets v.s. their Induced Kernel". In: *Advances in Neural Information Processing Systems*. 2019, pp. 9709–9721.

[YKM21] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. "A Unifying View on Implicit Bias in Training Linear Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2021.

[ZBHMS20] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, and Yoram Singer. "Identity Crisis: Memorization and Generalization Under Extreme Overparameterization". In: *International Conference on Learning Representations (ICLR)*. 2020.

[ZBHRV17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations (ICLR)*. 2017.

A Proofs for Section 3: Induced regularizer in special cases

For a single channel convolutional network (*i.e.,* $C = 1$), we denote the weights in the first and second layer as $\mathbf{U} \in \mathbb{R}^K$ and $\mathbf{V} \in \mathbb{R}^D$, respectively. Recall that the D dimensional discrete Fourier transform of the weights $\mathbf{U}, \mathbf{V}$ and the linear predictor $\mathbf{w} \in \mathbb{R}^D$ are denoted as $\widehat{\mathbf{U}} = \mathbf{F}_K \mathbf{U}, \widehat{\mathbf{V}} = \mathbf{F}\mathbf{V}, \widehat{\mathbf{w}} = \mathbf{F}\mathbf{w}$, respectively. Moreover, the Fourier transform is normalized such that for any $\mathbf{z} \in \mathbb{R}^K$, $\|\mathbf{z}\| = \|\widehat{\mathbf{z}}\|$. Thus, for all the quantities of interest, we use the ℓ_2 norm in signal domain interchangeably with ℓ_2 norm in the Fourier domain, *e.g.,* $\|\mathbf{U}\| = \|\widehat{\mathbf{U}}\|$, $\|\mathbf{V}\| = \|\widehat{\mathbf{V}}\|$, and $\|\mathbf{w}\| = \|\widehat{\mathbf{w}}\|$.

In the following proofs, we use the formulation of $\mathcal{R}_{K,C}(\mathbf{w})$ in eq. (8) for the case of $C = 1$ as:

$$\mathcal{R}_{K,1}(\mathbf{w}) = \min_{\mathbf{U}\in\mathbb{R}^K, \mathbf{V}\in\mathbb{R}^D} \|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 \quad \text{s.t., } \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}.$$

Lemma 2 ($K = 1, C = 1$)**.** *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{1,1}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\| = 2\sqrt{D}\|\mathbf{w}\|$.*

Proof. This statement is trivially true for $\mathbf{w} = 0$, so it suffices to show this for $\mathbf{w} \neq 0$. When the kernel size is 1, the first layer weight $\mathbf{U} \in \mathbb{R}^{1\times 1}$ is a scalar. Let this scalar be $\mathbf{U} = u \neq 0$. We then have $\widehat{\mathbf{U}} = \frac{1}{\sqrt{D}}[u, u, \ldots, u]$. Since $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}$, we have $\widehat{\mathbf{V}} = \frac{\sqrt{D}}{u}\widehat{\mathbf{w}}$. This means that $\mathcal{R}_{1,1}(\mathbf{w}) = \min_u \|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 = \min_u u^2 + \frac{D}{u^2}\|\widehat{\mathbf{w}}\|_2^2$. By using the AM-GM inequality ($a^2 + b^2 \geq 2ab$), this is at most $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$. Moreover, we can pick $u^2 = \sqrt{D}\|\widehat{\mathbf{w}}\|_2$ to achieve equality. □

Lemma 3. *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that:*

$$\mathcal{R}_{2,1}(\mathbf{w}) = 2\sqrt{D}\sqrt{\inf_{\alpha\in(-1,1)} \sum_{d=0}^{D-1} \frac{|\widehat{\mathbf{w}}[d]|^2}{1 + \alpha\cos(2\pi d/D)}}.$$

Proof. We first note that for any $\mathbf{U} \in \mathbb{R}^K, \mathbf{V} \in \mathbb{R}^D$ we can re-scale the norms so that $\|\mathbf{U}\| = \|\mathbf{V}\|$ while satisfying the constraints of $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}$ in the definition of $\mathcal{R}_{K,1}(\mathbf{w})$. Further, such a scaling would be optimal for minimizing the ℓ_2 norm of weights based on AM-GM inequality that $\|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 \geq 2\|\widehat{\mathbf{U}}\| \cdot \|\widehat{\mathbf{V}}\|$. Thus, in the rest of the proof, we consider the following equivalent formulation of $\mathcal{R}_{K,1}(\mathbf{w})$ as:

$$\mathcal{R}_{K,1}(\mathbf{w}) = \min_{\mathbf{U}\in\mathbb{R}^K, \mathbf{V}\in\mathbb{R}^D} 2\|\mathbf{U}\| \cdot \|\widehat{\mathbf{V}}\| \quad \text{s.t., } \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}} \tag{15}$$

We see that for any $\mathbf{U}, \mathbf{V}$ satisfying the constraint in the above equation, we have: $\forall d \in \text{supp}(\widehat{\mathbf{w}})$, it holds that $\widehat{\mathbf{V}}[d] = \frac{\widehat{\mathbf{w}}[d]}{\widehat{\mathbf{U}}[d]}$ (where $\text{supp}(\widehat{\mathbf{w}}) = \{d \in [D] : |\widehat{\mathbf{w}}| \neq 0\}$). Moreover, at an optimal solution, it is easy to see that $\widehat{\mathbf{V}}[d] = 0 \iff \widehat{\mathbf{w}}[d] = 0$.

Let $\mathbf{U} = [c_0, c_1]$. This means that the objective can be written as

$$2\|\mathbf{U}\|\|\widehat{\mathbf{V}}\| = 2\|\mathbf{U}\|\sqrt{\sum_{d=0}^{D-1} |\widehat{\mathbf{V}}[d]|^2} = 2\|\mathbf{U}\|\sqrt{\sum_{d\in\text{supp}(\widehat{\mathbf{w}})} |\widehat{\mathbf{V}}[d]|^2} = 2\sqrt{c_0^2 + c_1^2}\sqrt{\sum_{d\in\text{supp}(\widehat{\mathbf{w}})} \frac{|\widehat{\mathbf{w}}[d]|^2}{|\widehat{\mathbf{U}}[d]|^2}}. \tag{16}$$

We write the second term in terms of the signal domain representation of c_0 and c_1. We see that the Fourier transform of $\mathbf{U}$ is given by $\widehat{\mathbf{U}}[d] = \frac{1}{\sqrt{D}}\left(c_0 + c_1 \mathrm{e}^{-2\pi i d/D}\right) = (c_0 + c_1\cos(-2\pi i d/D)) +$

$i\,(c_1 \sin(-2\pi id/D))$. We thus have that:

$$\begin{aligned}|\widehat{\mathbf{U}}[d]|^2 &= \frac{1}{D}\left[(c_0 + c_1\cos(-2\pi id/D))^2 + (c_1\sin(-2\pi id/D))^2\right] \\ &= \frac{1}{D}\left(c_0^2 + c_1^2 + 2c_0c_1\cos(-2\pi ik/D)\right). \\ &= \frac{1}{D}(c_0^2+c_1^2)\left(1 + \frac{2c_0c_1}{c_0^2+c_1^2}\cos(-2\pi id/D)\right).\end{aligned} \tag{17}$$

Let $\alpha = \frac{2c_0c_1}{c_0^2+c_1^2}$. Plugging eq. (17) back into the objective in eq. (16) and using $\cos(z) = \cos(-z)$, we get that for any $\mathbf{U}, \mathbf{V}$ satisfying the constraints in the computation of $\mathcal{R}_{2,1}(\mathbf{w})$, the objective is in the desired formulation:

$$2\|\mathbf{U}\|\|\widehat{\mathbf{V}}\| = 2\sqrt{D}\sqrt{\sum_{d\in\text{supp}(\widehat{\mathbf{w}})} \frac{|\widehat{\mathbf{w}}[d]|^2}{1+\alpha\cos(2\pi id/D)}}. \tag{18}$$

Let us now consider the domain of α, which is the only unknown in the above equation. Observe that for any c_0, c_1, $\frac{2c_0c_1}{c_0^2+c_1^2} \in [-1,1]$. Moreover, any $\alpha \in [-1,1]$ be realized by some values of c_0 and c_1. Thus all $\alpha \in [-1,1]$ are valid. Here we further remark that the denominator in eq. (18) is zero if and only if $\widehat{\mathbf{U}}[d] = 0$ for any $d \in D$. However, this can only happen if $\widehat{\mathbf{w}}[d] = 0$ as otherwise the constraints $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \mathbf{w}$ is not satisfied for any $\mathbf{V}$. We can thus, minimize the RHS of eq. (18) over $\alpha \in [-1,1]$ to obtain $\mathcal{R}_{2,1}(\mathbf{w})$.

If we include the terms corresponding to $d \notin \text{supp}(\widehat{\mathbf{w}})$ in the summation in eq. (18), there is a technical condition than can lead to $0/0$ terms in end cases of $\alpha = 1$ (when $\widehat{\mathbf{w}}[0] = 0$) and $\alpha = -1$ (when $\widehat{\mathbf{w}}[D/2] = 0$). To avoid this technicality, we consider the infimum over $\alpha \in (-1,1)$ rather than minimum over $\alpha \in [-1,1]$. This is equivalent because the expression is continuous on the set of α on which it is well-defined. This completes the proof. □

B Proofs of Theorem 4: SDP tightness

Theorem 4. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) = \mathcal{R}_K^{SDP}(\mathbf{w})$.*

The high-level idea of the proof of Theorem 4 is to take an optimal solution to eq. (11), and construct a rank 1 solution that obtains the same objective ans satisfies the constraints. We reiterate the SDP formulation for easy reference:

$$\begin{aligned}\mathcal{R}_K^{\text{SDP}}(\mathbf{w}) = \min_{\mathbf{Z}\succcurlyeq 0}\ & \langle \mathbf{Z}, \mathbf{I}\rangle \\ \text{s.t.},\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{real}}\rangle = 2\,\text{Re}(\widehat{\mathbf{w}}[d]) \\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{img}}\rangle = 2\,\text{Im}(\widehat{\mathbf{w}}[d]).\end{aligned} \tag{19}$$

We start with the KKT conditions for (11). The D constraints involving $\text{Re}(\mathbf{w})$ correspond to a dual vector $\boldsymbol{\lambda}^{\text{real}} \in \mathbb{R}^D$; the D constraints involving $\text{Im}(\mathbf{w})$ correspond to a dual vector $\boldsymbol{\lambda}^{\text{img}} \in \mathbb{R}^D$. To simplify these conditions, we take $\boldsymbol{\lambda} \in \mathbb{C}^D$ to be $\boldsymbol{\lambda}^{\text{real}} + i\cdot\boldsymbol{\lambda}^{\text{img}}$ (when $\boldsymbol{\lambda}$ is a dual-optimal solution, $\boldsymbol{\lambda}$ is also a Fourier transform of a real vector). The dual variable for the PSD constraint corresponds to a matrix $\boldsymbol{\Gamma} \succcurlyeq 0$. In this notation, the KKT conditions are primal feasibility, along with the

following constraints:

$$\begin{aligned}\boldsymbol{\Gamma} &= \mathbf{I} - \begin{bmatrix} \mathbf{0}_K & \overline{\mathbf{F}}_K^\top \boldsymbol{\Lambda} \overline{\mathbf{F}} \\ \mathbf{F}\overline{\boldsymbol{\Lambda}}\mathbf{F}_K & \mathbf{0}_D \end{bmatrix} \\ \boldsymbol{\Gamma} &\succcurlyeq 0 \\ \mathbf{Z}\overline{\boldsymbol{\Gamma}} &= 0.\end{aligned}$$

Now, to simplify these conditions, suppose that $\mathbf{Z}$ is rank L, in which case we can express it as $\mathbf{Z} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}$, where $\mathbf{U} \in \mathbb{R}^{K \times L}$ and $\mathbf{V} \in \mathbb{R}^{D \times L}$. Using that $\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ is full rank and through some simple algebraic manipulations, we obtain the following formulation of the KKT conditions:

$$\begin{aligned}\sum_{l=0}^{L} \widehat{\mathbf{U}}[l,:] \odot \widehat{\mathbf{V}}[l,:] &= \widehat{\mathbf{w}} \\ \begin{bmatrix} \mathbf{0}_K & \overline{\mathbf{F}}_K^\top \boldsymbol{\Lambda} \overline{\mathbf{F}} \\ \mathbf{F}\overline{\boldsymbol{\Lambda}}\mathbf{F}_K & \mathbf{0}_D \end{bmatrix} &\preccurlyeq \mathbf{I}_{D+K} \\ \overline{\widehat{\mathbf{V}}} &= \overline{\boldsymbol{\Lambda}}\widehat{\mathbf{U}} \\ \widehat{\mathbf{U}} &= \mathbf{F}_K \overline{\mathbf{F}}_K^\top \boldsymbol{\Lambda} \overline{\widehat{\mathbf{V}}}.\end{aligned}$$

The KKT conditions give a clean way of formulating the requirements for obtaining an optimal solution that is rank 1. For $0 \leq l \leq L-1$, we let $\mathbf{u}_l = \mathbf{U}[:,l]$ and $\mathbf{v}_l = \mathbf{V}[:,l]$. From the third KKT condition, we see that $\widehat{\mathbf{v}}_l = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_l}$ for all l. Combining this with KKT condition 1, we obtain:

$$\widehat{\mathbf{w}} = \sum_{l=0}^{L-1} \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_l} \odot \widehat{\mathbf{u}}_l.$$

To find a rank 1 solution, it suffices to find a rank 1 matrix that satisfies the KKT conditions. Thus, it suffices to find a vector $\mathbf{u} \in \mathbb{R}^K$ such that $\widehat{\mathbf{w}} = \widehat{\mathbf{u}} \odot \widehat{\mathbf{v}}$ and $\widehat{\mathbf{v}} = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}}$. That is, it suffices for:

$$\boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}} \odot \widehat{\mathbf{u}} = \sum_{l=0}^{L-1} \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_r} \odot \widehat{\mathbf{u}}_r.$$

This equation helps us because the same $\boldsymbol{\lambda}$ appears on both sides of the equation, and so we are able to eliminate $\boldsymbol{\lambda}$ entirely. In other words, it suffices to show the following:

$$\overline{\widehat{\mathbf{u}}} \odot \widehat{\mathbf{u}} = \sum_{l=0}^{L-1} \overline{\widehat{\mathbf{u}}_l} \odot \widehat{\mathbf{u}}_l.$$

It is more convenient to show the equivalent time space version:

$$\mathbf{u} \star \mathbf{u} = \sum_{l=0}^{L-1} \mathbf{u}_l \star \mathbf{u}_l. \tag{20}$$

We thus wish to show that a sum of convolutions of vectors in $\mathbb{R}^K$ with themselves can be expressed as a convolution of a vector in $\mathbb{R}^K$ with itself. The main lemma of our proof is the following which we prove in Section B.1:

Lemma 11. *For any $K \geq 1$ and any $D \geq K$, and for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, where convolutions are taken with respect to the base dimension D.*

We now show that Theorem 4 follows from Lemma 11. We have already shown that it suffices to find $\mathbf{u} \in \mathbb{R}^D$ such that eq. (20) holds. We recursively apply Lemma 11 to conclude that such a vector exists.

Now, it suffices to prove Lemma 11; we prove this lemma in the next subsection.

B.1 Proof of Lemma 11

The proof of this lemma uses a polynomial formulation of convolutions. We use the following notation. Let $\mathcal{P}_k \subseteq \mathbb{R}[x]$ denote the set of degree $\leq k$ polynomials with real coefficients. For a vector $\mathbf{z} \in \mathbb{R}^d$, we define the *polynomial representation* $p_{\mathbf{z}}(x) \in \mathcal{P}_{d-1}$ to be the polynomial $\mathbf{z}_0 + \mathbf{z}_1 x + \ldots + \mathbf{z}_{d-1} x^{d-1}$.

The following standard facts about convolutions and polynomials will be useful in the proof:

Fact 1. *For any $\mathbf{a} \in \mathbb{R}^k$ for $k \leq \frac{D+1}{2}$, the convolution $\mathbf{a} \star \mathbf{a}$ has polynomial representation $p_{\mathbf{a}\star\mathbf{a}}(x)$ equal to $x^D p_{\mathbf{a}}(x) \cdot p_{\mathbf{a}}(1/x)$, where the exponents are taken modulo D.*

Fact 2. *For any $\mathbf{a} \in \mathbb{R}^k$ for $k \leq D$, the vector $\mathbf{a} \star \mathbf{a}$ satisfies the property that $(\mathbf{a} \star \mathbf{a})_d = (\mathbf{a} \star \mathbf{a})_{D-d}$ for all $0 \leq d \leq D-1$. Moreover, all of the entries of the Fourier representation $\widehat{\mathbf{a} \star \mathbf{a}}$ are nonnegative real numbers.*

Fact 3. *Let $p(x) \in \mathcal{P}_k$ be a palindromic polynomial, i.e. a polynomial where the coefficient of x^d is equal to the coefficient of x^{k-d} for all $0 \leq d \leq k$. For $\alpha \neq 0$, α is a root with multiplicity m of $p(x)$ if and only if $1/\alpha$ is a root with multiplicity m.*

Fact 4. *Let $p(x) \in \mathcal{P}_k$, and consider the polynomial $x^{k-1}p(1/x) \in \mathcal{P}_{k-1}$. For $\alpha \neq 0$, α is a root with multiplicity m of $p(x)$ if and only if $1/\alpha$ is a root with multiplicity m of $x^{k-1}p(1/x)$.*

Moreover, the following lemma will also be used in our proof.

Lemma 12. *Consider vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^k$ for $k \leq D$, and let $p_{\mathbf{a}}, p_{\mathbf{b}} \in \mathcal{P}_{k-1}$ be their polynomial representation. If $\alpha \in \mathbb{C}$ such that $|\alpha| = 1$ is a root of $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x)) + p_{\mathbf{b}}(x)(x^{k-1}p_{\mathbf{b}}(1/x))$, then α has even multiplicity.*

Proof. Suppose that α is a root of $Q(x)$ and $|\alpha| = 1$. We see that

$$\begin{aligned}
0 &= Q(\alpha) \\
&= p_{\mathbf{a}}(\alpha)(\alpha^{K-1}p_{\mathbf{a}}(1/\alpha)) + p_{\mathbf{b}}(\alpha)(\alpha^{K-1}p_{\mathbf{b}}(1/\alpha)) \\
&= p_{\mathbf{a}}(\alpha)(\alpha^{K-1}p_{\mathbf{a}}(\overline{\alpha})) + p_{\mathbf{b}}(\alpha)(\alpha^{K-1}p_{\mathbf{b}}(\overline{\alpha})) \\
&= \alpha^{K-1}\left(p_{\mathbf{a}}(\alpha)\overline{p_{\mathbf{a}}(\alpha)}) + p_{\mathbf{b}}(\alpha)\overline{p_{\mathbf{b}}(\alpha)}\right) \\
&= \alpha^{K-1}\left(|p_{\mathbf{a}}(\alpha)|^2 + |p_{\mathbf{b}}(\alpha)|^2\right).
\end{aligned}$$

Since $\alpha \neq 0$, this means that $|p_{\mathbf{a}}(\alpha)|^2 + |p_{\mathbf{b}}(\alpha)|^2 = 0$. Thus, $p_{\mathbf{a}}(\alpha) = 0$ and $p_{\mathbf{b}}(\alpha) = 0$. Now, it suffices to show that α is a root with even multiplicity in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ and in $p_{\mathbf{b}}(x)(x^{K-1}p_{\mathbf{b}}(1/x))$.

We show that α has even multiplicity in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ (an analogous argument shows this for $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$). Suppose that α has multiplicity m in $p_{\mathbf{a}}(x)$. Since $p_{\mathbf{a}}(x)$ has real coefficients, we know that $\overline{\alpha}$ is a root of $p_{\mathbf{a}}(x)$ with multiplicity m. By Fact 4, we also know that $1/\alpha = \overline{\alpha}$ is a root with multiplicity m of $x^{k-1}p_{\mathbf{a}}(1/x))$. Since $x^{k-1}p_{\mathbf{a}}(1/x))$ has real coefficients, we know that $\overline{\overline{\alpha}} = \alpha$ is a root with multiplicity m of $x^{k-1}p_{\mathbf{a}}(1/x))$. This means that α has multiplicity $2m$ in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ as desired. □

The main step in the proof of Lemma 11 is to show the special case where $D = 2K - 1$. The key idea for this proof is to use the polynomial formulation of convolution, and use factorization of the polynomials to implicitly construct $\mathbf{c}$.

Lemma 13. *For any $K \geq 1$, for $D = 2K - 1$, and for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, where convolutions are taken with respect to the base dimension D.*

Proof. This statement trivially holds with $\mathbf{c}$ as the zero vector if $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = 0$, so for the remainder of the proof, we assume that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ is nonzero.

The high-level idea for this proof is that we use the polynomial formulation of convolution, and use the factorization of the polynomials to identify $\mathbf{c}$. We wish to show that there exists a vector $\mathbf{c}$ such that $\mathbf{c} \star \mathbf{c} = \mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$. Using the polynomial representation, we can write this requirement as $p_{\mathbf{c}\star\mathbf{c}}(x) = p_{\mathbf{a}\star\mathbf{a}}(x) + p_{\mathbf{b}\star\mathbf{b}}(x)$. Using Fact 1, we can write this requirement as follows

$$x^D p_{\mathbf{c}}(x) \cdot p_{\mathbf{c}}(1/x) = x^D p_{\mathbf{a}}(x) p_{\mathbf{a}}(1/x) + x^D p_{\mathbf{b}}(x) p_{\mathbf{b}}(1/x).$$

We can simplify this to:

$$x^{K-1} p_{\mathbf{c}}(x) \cdot p_{\mathbf{c}}(1/x) = x^{K-1} p_{\mathbf{a}}(x) p_{\mathbf{a}}(1/x) + x^{K-1} p_{\mathbf{b}}(x) p_{\mathbf{b}}(1/x).$$

To simplify notation, we denote the right-hand-side of the previous equation by $Q(x)$, and thus define:

$$Q(x) := x^{K-1} p_{\mathbf{a}}(x) p_{\mathbf{a}}(1/x) + x^{K-1} p_{\mathbf{b}}(x) p_{\mathbf{b}}(1/x) \in \mathcal{P}_{2K-2}.$$

Since there is a 1-to-1 correspondence between polynomials in $\mathcal{P}_{K-1}$ and vectors in $\mathbb{R}^K$, it suffices to show that there exists a polynomial $p \in \mathcal{P}_{K-1}$ such that:

$$Q(x) = x^{K-1} p(x) \cdot p(1/x).$$

We will construct p implicitly by specifying its roots (with multiplicities) and its leading coefficient.

Constructing the roots of p. Our goal is to construct a multi-set S of roots that will be the nonzero roots of p with multiplicities. It suffices for S to satisfy the following three requirements:

1. (R1) $0 \notin S$
2. (R2) If α is a root with multiplicity m, then $\overline{\alpha}$ is a root with multiplicity m.
3. (Re) If $\alpha \neq 0$ is a root with multiplicity m, then $1/\alpha$ is a root with multiplicity m.

These requirements suffice because all we need to guarantee is that p has real coefficients, and that the set of roots (with multiplicities) of $x^{K-1} p(x) \cdot p(1/x)$ is equal the set of nonzero roots (with multiplicities) of $Q(x)$. (R2) gives the former, and (R1) and (R3) give the latter. We now construct a set S that satisfies these two properties.

In order to construct S, it's helpful to establish some properties of $Q(x)$. Since $Q(x)$ has degree at most $2K - 2$, we know that there are at most $2K - 1$ complex roots (including multiplicity). Now, we prove the following properties of these roots:

1. (P1) For every root α with multiplicity m, $\overline{\alpha}$ is a root and has multiplicity m.
2. (P2) If $\alpha \neq 0$ is a root with multiplicity m of $Q(x)$, then $1/\alpha$ is a root with multiplicity m of $Q(x)$.

3. (P3) If α such that $|\alpha| = 1$ is a root of $Q(x)$, then α has even multiplicity.

(P1) follows from the fact $Q(x)$ has real coefficients. For (P2) , it suffices to show that $Q(x)$ is a palindromic polynomial. i.e. the coefficients form a palindrome (see Fact 3). To see that $Q(x)$ is a palindromic polynomial, notice that by Fact 2, $(\mathbf{a} \star \mathbf{a})_d = (\mathbf{a} \star \mathbf{a})_{D-d}$ and $(\mathbf{b} \star \mathbf{b})_d = (\mathbf{b} \star \mathbf{b})_{D-d}$ for all $0 \leq d \leq D-1$, and so $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = (\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_{D-d}$ for all $0 \leq d \leq D-1$. Now, we can use Fact 1 to conclude that the x^d coefficient of $Q(x)$ is the $(d - K + 1) \mod D$ coefficient of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$. Thus, the x^d coefficient of $Q(x)$ is equal to the x^{2K-2-d} coefficient of $Q(x)$, as desired. For (P3), we apply Lemma 12.

To construct the set S, it is helpful to consider an undirected graph G where the vertices are the *nonzero* roots of $Q(x)$ (a root with multiplicity m corresponds to m separate vertices). The edges are defined as follows. When $\alpha \neq 1$, we connect a vertex corresponding to α with some vertex corresponding to $1/\alpha$, so that the graph forms a bipartite graph (it is possible to do this because of (P2)). By (P3), we can handle $\alpha = 1$, and form non-self-loop edges of the form $(1, 1)$. Let G^{real} be the subgraph of G consisting of vertices corresponding to real roots, and let G^{complex} be the subgraph of G consisting of vertices corresponding to roots with nonzero imaginary part. It is easy to see that $G = G^{\text{real}} \cup G^{\text{complex}}$ and these graphs are disjoint.

Let us now use these graphs to construct the set S, which will contain half of the vertices in G. For G^{real}, we add one vertex from each edge in G^{real} to S. For G^{complex}, we can pair up the edges in the graph G^{complex} as follows. When $|\alpha| \neq 1$, we can pair up the edges $(\alpha, 1/\alpha)$ and $(\overline{\alpha}, 1/\overline{\alpha})$ so that the pairs are disjoint and no edge is paired with itself (it is possible to do this because of (P1), coupled with the fact that $1/\alpha \neq \overline{\alpha}$). When $|\alpha| = 1$, we use the fact that α has even multiplicity (see (P3)). Thus, we can pair up each edge $(\alpha, 1/\alpha)$ with an edge of the form $(\alpha, 1/\alpha) = (\overline{\alpha}, 1/\overline{\alpha})$, so that pairs continue to be disjoint and no edge is paired with itself. Now, for each pair $(\alpha, 1/\alpha)$ and $(\overline{\alpha}, 1/\overline{\alpha})$, we add α to S and we add $\overline{\alpha}$ to S. It is easy to see that S meets (R1), (R2), and (R3) as desired.

To obtain the roots of p, all that remains is to handle potential zero roots of $Q(x)$. Suppose that $Q(x)$ has a root of 0 with multiplicity $m \geq 0$. We let the multi-set of roots of p be the set S augmented with the m copies of 0. We need to confirm that p has at most $K - 1$ roots. To show this, we do a root counting argument. Notice that by Fact 2, if $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = 0$, this means that $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_{D-d} = 0$. In terms of the polynomial representation, this means that if $Q(x)$ has a zero root with multiplicity m, then the degree of $Q(x)$ is $2K - 2 - m$ (since $Q(x) \neq 0$). Thus, we see that $Q(x)$ has $2K - 2 - m$ roots not including multiplicity, and $2K - 2 - 2m$ nonzero roots. By the construction of S, we see that $p_{\text{monic}}(x)$ has $K - 1 - m$ nonzero roots including multiplicities, and m zero roots, which amounts to $K - 1$ roots in total with multiplicity, as desired.

We wish to show that p has the desired multi-set of roots. In particular, let $p_{\text{monic}}(x)$ be the monic polynomial given by the roots of S, and consider

$$Q_1(x) := p_{\text{monic}}(x) x^{K-1} p_{\text{monic}}(1/x).$$

We claim that the roots of $Q_1(x)$ with multiplicities are equal to the roots of $Q(x)$. The multi-sets of nonzero roots are equal by the properties of S. For the zero roots, we simply need to show that $x^{K-1} p_{\text{monic}}(1/x)$ has no roots that are 0. Using that the constant term of $x^{K-1} p_{\text{monic}}(1/x)$ is equal to the coefficient of x^{K-1} in $p_{\text{monic}}(x)$, it suffices to show that the coefficient of x^{K-1} in $p_{\text{monic}}(x)$ is nonzero. To see this, we use the fact that $p_{\text{monic}}(x)$ has $K - 1$ roots in total with multiplicity by the argument from the previous paragraph, and so the degree of $p_{\text{monic}}(x)$ must be $K - 1$. This means that the roots of $Q_1(x)$ with multiplicities are equal to the roots of $Q(x)$.

Constructing the leading coefficient of p. Now, we need to just construct the leading coefficient of p. As above, let $p_{\text{monic}}(x)$ be the monic polynomial given by the roots of S, and consider $Q_1(x) = p_{\text{monic}}(x)x^{K-1}p_{\text{monic}}(1/x)$. Since $Q_1(x)$ and $Q(x)$ have the same set of roots with multiplicities, we know that $Q_1(x) = \gamma \cdot Q(x)$ for some $\gamma \neq 0$.

We claim that $\gamma > 0$. This can be seen as follows. Let $\mathbf{c}_{\text{monic}} \in \mathbb{R}^D$ be the vector with polynomial representation $p_{\text{monic}}(x)$. Now, by Fact 1, we know that the polynomial representation $p_{\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}}}$ takes the form $x^D p_{\text{monic}}(x)p_{\text{monic}}(1/x) = x^{D-K+1}Q_1(x)$, where exponents are taken modulo D. Since $Q_1(x) = \gamma \cdot Q(x)$, and since the polynomial representation of $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})$ takes the form $x^{D-K+1}Q(x)$ where the exponents are taken modulo D, we can conclude that $\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}} = \gamma(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})$, and thus that the Fourier representations of these vectors are also equal. By Fact 2, we know that the entries of the Fourier representation of $\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}}$ consist of all nonnegative real numbers. Similarly, Fact 2 also tells us that the Fourier representation of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ consists of entries that are all nonnegative real numbers. This means that $\gamma \geq 0$, as desired.

Defining p. Thus, we can let $p(x) = \sqrt{\gamma}p_{\text{monic}}(x) \in \mathcal{P}_{K-1}$, and obtain that $Q(x) = p(x)(x^{K-1}p(1/x))$ as desired.

□

Now, from Lemma 13, it is not difficult to conclude Lemma 11.

Proof of Lemma 11. By Lemma 13, we know that Lemma 11 holds when $D = 2K - 1$. We now show that this implies the statement for a general value of D. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, we apply Lemma 13 to obtain a vector $\mathbf{c}$. We show that $\mathbf{c}$ works regardless of the value of D.

To see the statement for $D \geq 2K - 1$, we use the fact that by Lemma 13:

$$(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = (\mathbf{c} \star \mathbf{c})_d,$$

for $d \in [-K + 1, K - 1] \mod D$. For d not in this set, it is easy to verify that both sides are equal to 0 because the vectors are K-dimensional.

To see the statement for $K \leq D \leq 2K - 1$, we use the fact that each term of a convolution with base dimension D is a linear combination of the terms in the convolution with base dimension $2K - 1$. Thus, we can conclude that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$. □

C Remaining proofs of results in Section 4

C.1 Proof of Corollary 5

Corollary 5. *For $K \leq D$ and any C, $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm.*

Corollary 5 follows from Theorem 4.

Proof of Corollary 5. It suffices to establish the scalar multiplication property, the triangle inequality, and point separation. Our main ingredient is Theorem 4. Scalar multiplication follows from the fact that for any $\gamma \neq 0$, it holds that $\mathbf{Z}$ is a feasible solution to $\mathcal{R}_k^{\text{SDP}}(\mathbf{w})$ if and only if $\gamma\mathbf{Z}$ is a feasible solution to $\mathcal{R}_k^{\text{SDP}}(\gamma\mathbf{w})$. Notice that the objective satisfies $\langle \gamma\mathbf{Z}, \mathbf{I} \rangle = |\gamma| \langle \mathbf{Z}, \mathbf{I} \rangle$ as desired.

For the triangle inequality, it suffices to show that if $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, then $\mathcal{R}_{K,C}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. By Theorem 4, it suffices to show that $\mathcal{R}_K^{\text{SDP}}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. Suppose

that $\mathbf{U}_1$ and $\mathbf{V}_1$ are optimal solutions to $\mathcal{R}_{K,C}(\mathbf{w}_1)$, and suppose that $\mathbf{U}_2$ and $\mathbf{V}_2$ are optimal solutions to $\mathcal{R}_{K,C}(\mathbf{w}_2)$. If we define:

$$\mathbf{Z}_1 = \begin{bmatrix} \mathbf{U}_1\mathbf{U}_1^\top & \mathbf{U}_1\mathbf{V}_1^\top \\ \mathbf{V}_1\mathbf{U}_1^\top & \mathbf{V}_1\mathbf{V}_1^\top \end{bmatrix}$$
$$\mathbf{Z}_2 = \begin{bmatrix} \mathbf{U}_2\mathbf{U}_2^\top & \mathbf{U}_2\mathbf{V}_2^\top \\ \mathbf{V}_2\mathbf{U}_2^\top & \mathbf{V}_2\mathbf{V}_2^\top \end{bmatrix}$$
$$\mathbf{Z} = \mathbf{Z}_1 + \mathbf{Z}_2,$$

then we see that the SDP objective $\langle \mathbf{Z}, \mathbf{I} \rangle = \langle \mathbf{Z}_1, \mathbf{I} \rangle + \langle \mathbf{Z}_2, \mathbf{I} \rangle = \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. Moreover, we see that $\mathbf{Z}$ is a feasible solution to (11) for $\mathbf{w}$. This means that $\mathcal{R}_k^{\text{SDP}}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$ as desired.

For point separation, notice that $\mathcal{R}_{K,C}(\mathbf{w}) = 0$, then there exist $\mathbf{U}$ and $\mathbf{V}$ such that $\|\mathbf{U}\|_F + \|\mathbf{V}\|_F = 0$. This means that $\mathbf{U} = 0$ and $\mathbf{V} = 0$, which means that $\mathbf{w} = 0$ as desired. Moreover, if $\mathbf{w} = 0$, then it's clear that $\mathcal{R}_{K,C}(\mathbf{w}) = 0$. □

C.2 The Dual Formulation

In order to analyze the SDP formulation, we consider the dual. We use the formulation of the dual variable in B as $\boldsymbol{\lambda} \in \mathbb{C}^D$. In this form, the dual can be expressed as:

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{C}^D} \quad & \text{Re}\left(\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle\right) \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{0}_K & \mathbf{F}_K^\top \overline{\boldsymbol{\Lambda}} \mathbf{F} \\ \overline{\mathbf{F}} \boldsymbol{\Lambda} \overline{\mathbf{F}}_K & \mathbf{0}_D \end{bmatrix} \preccurlyeq \mathbf{I}. \end{aligned}$$

To simplify the objective $\text{Re}\left(\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle\right)$, notice that the phases of $\boldsymbol{\lambda}$ can be set to align with $\widehat{\mathbf{w}}$ without affecting the constraint. Thus, we can set the objective to be $|\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle|$. For convenience, we also expand out the conic constraint in vector form, and this reformulation incurs a factor of 2 on the objective. We thus obtain the following equivalent formulation of the dual:

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{C}^D} \quad & 2 \sum_{d=0}^{D-1} |\boldsymbol{\lambda}[d]| \cdot |\mathbf{w}[d]| \\ \text{s.t.} \quad & \forall x \in \mathbb{C}^K, \quad \sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \leq 1. \end{aligned} \tag{21}$$

We now show that strong duality holds for this SDP.

Proposition 14. *The SDP in (11) satisfies strong duality.*

Proof. To show strong duality, it suffices to show Slater's condition. We just need to find a solution $\boldsymbol{\lambda} \in \mathbb{C}^d$ where the inequality constraint is not tight. That is, we need to find $\boldsymbol{\lambda}$ such that $\forall x \in \mathbb{C}^K, \ \sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 < 1$. Let's take $\boldsymbol{\lambda} = [1/2, 0, 0, \ldots, 0]$. Notice that $\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 = 0.5|\widehat{\mathbf{x}}[0]|^2 \leq 0.5 < 1$, as desired.

□

With the dual, along the fact that $\mathcal{R}_k^{\text{SDP}}(\mathbf{w})$ is a norm, we are equipped to prove Lemma 6 and Lemma 7.

C.3 Proof of Lemma 6

Lemma 6. *For any $K \leq D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$:*

$$2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$$
$$2\|\widehat{\mathbf{w}}\|_1 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{\left\lceil \frac{D}{K} \right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

Proof of Lemma 6. The bounds of $2\|\widehat{\mathbf{w}}\|_1$ and $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$ follow in a straightforward way from Lemma 1—2 coupled with Theorem 4, and we begin by proving these bounds. First, we show the lower bound of $2\|\widehat{\mathbf{w}}\|_1$. This follows from the fact that

$$2\|\widehat{\mathbf{w}}\|_1 \overset{(a)}{=} \mathcal{R}_{D,1}(\mathbf{w}) \overset{(b)}{=} \mathcal{R}_{K,C}(\mathbf{w}) \overset{(c)}{\leq} \mathcal{R}_{K,C}(\mathbf{w}),$$

where (a) follows from Lemma 2, (b) follows from Theorem 4, and (c) follows from Remark 1. Now, we show the upper bound of $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$. This follows from the fact that

$$2\sqrt{D}\|\widehat{\mathbf{w}}\|_2 \overset{(a)}{=} \mathcal{R}_{1,1}(\mathbf{w}) \overset{(b)}{=} \mathcal{R}_{1,C}(\mathbf{w}) \geq \mathcal{R}_{K,C}(\mathbf{w}),$$

where (a) follows from Lemma 2 and (b) follows from Theorem 4, and (c) follows from Remark 1.

The bulk of the proof lies in showing the lower bound of $2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2$, and an upper bound of $2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$. We first prove the lower bound, and then we prove the upper bound.

Proof of the lower bound. We prove that $\mathcal{R}_{K,C}(\mathbf{w}) \geq 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2$. It suffices to consider a dual feasible vector to eq. (21) that achieves an objective $2\frac{\sqrt{D}}{\sqrt{K}}\|\mathbf{w}\|_2$. We consider the vector

$$\boldsymbol{\lambda} = \frac{\mathbf{w}}{\|\mathbf{w}\|}\frac{\sqrt{D}}{\sqrt{K}}.$$

We see that the objective is equal to

$$2\sum_{d=0}^{D-1}|\boldsymbol{\lambda}[d]||\mathbf{w}[d]| = 2\frac{\sqrt{D}}{\sqrt{K}}\|\mathbf{w}\|_2,$$

as desired. It thus suffices to show that $\boldsymbol{\lambda}$ satisfies $\sum_{d=0}^{D-1}|\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \leq 1$ for all $\mathbf{x} \in \mathbb{C}^d$ such that $\|\mathbf{x}\|_2 \leq 1$. Using Holder's inequality, we can bound:

$$\sum_{d=0}^{D-1}|\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \leq \left(\max_{0\leq d\leq D-1}|\widehat{\mathbf{x}}[d]|^2\right)\left(\sum_{d=0}^{D-1}|\boldsymbol{\lambda}[d]|^2\right) = \left(\max_{0\leq d\leq D-1}|\widehat{\mathbf{x}}[d]|^2\right)\|\boldsymbol{\lambda}\|_2^2.$$

We can bound the first term by:

$$|\widehat{\mathbf{x}}[d]| = \frac{1}{\sqrt{D}}\left|\sum_{k=0}^{K-1}\mathbf{x}[k]\mathrm{e}^{-2\pi ikd/D}\right| \leq \frac{1}{\sqrt{D}}\sum_{k=0}^{K-1}|\mathbf{x}[k]\mathrm{e}^{-2\pi ikd/D}| = \frac{\|x\|_1}{\sqrt{D}} \leq \frac{\sqrt{K}}{\sqrt{D}}.$$

Moreover, we see that $\|\boldsymbol{\lambda}\| = \frac{\sqrt{D}}{\sqrt{K}}$. This means that $\left(\max_{0\leq d\leq D-1}|\widehat{\mathbf{x}}[d]|^2\right)\|\boldsymbol{\lambda}\|_2^2 \leq 1$, as desired.

Proof of the upper bound. We prove that $\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$. Our main ingredient is Corollary 5 which tells us that $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm. We define $T = \lceil D/K \rceil$ vectors $\mathbf{w}_0, \ldots, \mathbf{w}_{T-1} \in \mathbb{R}^D$ where $\mathbf{w} = \sum_{t=0}^{T-1} \mathbf{w}_t$, and apply Corollary 5 to obtain that:

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq \sum_{t=0}^{T-1} \mathcal{R}_{K,C}(\mathbf{w}_t).$$

These vectors are chosen that each $\mathcal{R}_{K,C}(\mathbf{w}_t)$ takes on a simple closed-form solution.

In order to construct the vectors $\mathbf{w}_t$, we consider $\widehat{\mathbf{q}} = \sqrt{\widehat{\mathbf{w}}}$, defined so that $\mathbf{w} = \mathbf{q}^{\downarrow} \star \mathbf{q}$. We define vectors $\mathbf{r}_0, \ldots, \mathbf{r}_{T-1} \in \mathbb{R}^D$ such that $\sum_{t=0}^{T-1} \mathbf{r}_t = \mathbf{q}$ as follows. Roughly speaking these vectors consist of the disjoint subsets of the coordinates of $\mathbf{q}$. More formally, for $0 \leq t \leq T-1$, let $\mathbf{r}_t$ be defined so that $\mathbf{r}_t[l] = \mathbf{q}[l - t \cdot K]$ for $l \in [t \cdot K, \min((t+1) \cdot K - 1, D-1)]$, and $\mathbf{r}_t[l] = 0$ otherwise. Let $\mathbf{w}_t = \mathbf{r}_t^{\downarrow} \star \mathbf{q}$ for $0 \leq t \leq T-1$.

We now show that $\mathcal{R}_{K,C}(\mathbf{w}_t) \leq 2\|\mathbf{r}_t\|\|\mathbf{q}\|$. We show this by explicitly constructing solutions to (6), taking advantage of the fact that $\mathbf{r}_t$ is effectively a vector in $\mathbb{R}^K$ that is zero-padded appropriately. This K-dimensional vector $\mathbf{r}'_t \in \mathbb{R}^K$ is given by $\mathbf{r}'_t[k] = \mathbf{r}_t[(t \cdot K + k) \mod D]$ for $0 \leq k \leq K-1$. Now, we wish to write $\mathbf{w}_t$ as a convolution $(\mathbf{r}'_t)^{\downarrow} \star \mathbf{q}_t$, for some suitably chosen vector $\mathbf{q}_t$. Since $\mathbf{w}_t = \mathbf{r}_t^{\downarrow} \star \mathbf{q}$, we can take $\mathbf{q}_t \in \mathbb{R}^D$ to be $\mathbf{q}$ with the coordinates shifted appropriately. Now, we can rescale $\mathbf{r}'_t$ and $\mathbf{q}_t$ so that they have equal ℓ_2 norms, and obtain the following vectors: $\left(\sqrt{\frac{\|q\|_t}{\|\mathbf{r}'_t\|}}\right)\mathbf{r}'_t$ and $\left(\sqrt{\frac{\|\mathbf{r}'_t\|}{\|q\|_t}}\right)\mathbf{q}_t$. These vectors are a feasible solution to eq. (6) for $\mathcal{R}_{K,C}(\mathbf{w}_t)$ and achieve an objective of $2\|\mathbf{r}'_t\|\|\mathbf{q}_t\| = 2\|\mathbf{r}_t\|\|\mathbf{q}\|$, as desired.

Using that $\mathcal{R}_{K,C}(\mathbf{w}_t) \leq 2\|\mathbf{r}_t\|\|\mathbf{q}\|$ for $0 \leq t \leq T-1$, we obtain the following bound on $\mathcal{R}_{K,C}(\mathbf{w})$:

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\|\mathbf{q}\| \sum_{t=0}^{T-1} \|\mathbf{r}_t\|.$$

Now, notice that since the supports of $\mathbf{r}_t$ for $0 \leq t \leq T-1$ are disjoint, $\sum_{t=0}^{T-1} \|\mathbf{r}_t\|^2 = \|\mathbf{q}\|^2$. Applying AM-GM, this means that

$$(\sum_{t=0}^{T-1} \|\mathbf{r}_t\|)^2 \leq T(\sum_{t=0}^{T-1} \|\mathbf{r}_t\|^2) = T\|\mathbf{q}\|^2.$$

Thus, we have that

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{T}\|\mathbf{q}\|^2 = 2\sqrt{T}\|\widehat{\mathbf{w}}\|_1 = 2\sqrt{\left\lceil \frac{D}{K} \right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

□

C.4 Proof of Lemma 7

Lemma 7. *Consider vectors $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ specified by $\mathbf{p} \in \mathbb{R}^P$ s.t., P divides D. (a) For any $K \leq P$, it holds that $\forall C$:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \le K \le D$ if $K = P \cdot T$ for integer T, then $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1 = 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_1.$$

We first prove an upper bound on $\mathcal{R}_{K,C}(\mathbf{w})$, and then we prove a matching lower bound. In these proofs, let $\mathbf{w} = \mathbf{w}(\mathbf{p})$. Both of these proofs use the standard fact that $\widehat{\mathbf{w}}[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{p}}[p]$ for $0 \le p \le P-1$, and $\widehat{\mathbf{w}}[d] = 0$ if $(D/P) \nmid d$.

Lemma 15 (Upper bound)**.** *Consider vectors $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ specified by $\mathbf{p} \in \mathbb{R}^P$ (such that D is a multiple of P).*
(a) For any $K \le P$, it holds that $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \le \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \le K \le D$ if $K = P \cdot T$ for integer T, then $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \le 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1.$$

Proof of Lemma 15. It suffices to show an upper bound for the case of a single output channel. We explicitly construct a pair $(\mathbf{u}, \mathbf{v})$ where $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ achieves the desired objective.

Case 1: $\mathbf{K \le P}$. We construct $(\mathbf{u}, \mathbf{v})$ using an optimal solution $\mathbf{u_p}$ and $\mathbf{v_p}$ to eq. (6) for $\mathcal{R}^{(P)}_{K,1}(\mathbf{p})$. We let $\mathbf{u}$ be defined so that $\mathbf{u}[k] = \sqrt{\frac{D}{P}}\mathbf{u_p}[k]$ for $0 \le k \le K-1$. We let $\mathbf{v}$ be defined to be $\mathbf{v} = [\mathbf{v_p}, \ldots, \mathbf{v_p}]$. Notice that:

$$\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \frac{D}{P}\|\mathbf{u_p}\|^2 + \frac{D}{P}\|\mathbf{v_p}\|^2 = \frac{D}{P}\mathcal{R}^{(P)}_{K,1}(\mathbf{p}),$$

as desired. To see that $\mathbf{u}^{\downarrow} \star \mathbf{v} = \mathbf{w}$, it suffices to show $\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}} = \mathbf{w}$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{v_p}}[p]$ for $0 \le p \le P-1$, and $\widehat{\mathbf{v}}[d] = 0$ if $(D/P) \nmid d$. This means that $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[d] = 0 = \widehat{\mathbf{w}}[d]$ if $(D/P) \nmid d$ as desired. Thus it suffices to handle $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P)\cdot p]$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{v}}_{\mathbf{p}}[p]$, and $\widehat{\mathbf{u}}[(D/P)\cdot p] = \frac{\sqrt{P}}{\sqrt{D}}\frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{u}}_{\mathbf{p}}[p] = \widehat{\mathbf{u}}_{\mathbf{p}}[p]$. This means that:

$$(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{u}}_{\mathbf{p}}[p]\widehat{\mathbf{v}}_{\mathbf{p}}[p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{p}}[p] = \widehat{\mathbf{w}}[(D/P)\cdot p],$$

as desired.

Case 2: $\mathbf{K = T \cdot P}$. We construct $(\mathbf{u}, \mathbf{v})$ using an optimal solution $\mathbf{u_p}$ and $\mathbf{v_p}$ to eq. (6) for $\mathcal{R}^{(P)}_{P,1}(\mathbf{p})$. We let $\mathbf{u} = \frac{\sqrt{D}}{T^{3/4}\sqrt{P}}[\mathbf{u_p}, \ldots, \mathbf{u_p}]$ be a scaled version of T repeated copies of $\mathbf{u_p}$. We let $\mathbf{v} = \frac{1}{T^{1/4}}[\mathbf{v_p}, \ldots, \mathbf{v_p}]$ be a scaled version of $\frac{D}{P}$ repeated copies of $\mathbf{v_p}$. Notice that

$$\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \frac{D}{\sqrt{TP}}\|\mathbf{u_p}\|^2 + \frac{D}{\sqrt{TP}}\|\mathbf{v_p}\|^2 = \frac{D}{\sqrt{TP}}\mathcal{R}_{P,1}(\mathbf{p}) = 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1,$$

as desired. To see that $\mathbf{u}^{\downarrow} \star \mathbf{v} = \mathbf{w}$, it suffices to show $\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}} = \mathbf{w}$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{T^{1/4}\sqrt{P}}\widehat{\mathbf{v}}_{\mathbf{p}}[p]$ for $0 \le p \le P-1$, and $\widehat{\mathbf{v}}[d] = 0$ if $(D/P) \nmid d$. This means that $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[d] = 0 = \widehat{\mathbf{w}}[d]$

if $(D/P) \nmid d$ as desired. Thus it suffices to handle $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P) \cdot p]$. Notice that $\widehat{\mathbf{v}}[(D/P) \cdot p] = \frac{\sqrt{D}}{T^{1/4}\sqrt{P}} \widehat{\mathbf{v}}_{\mathbf{p}}[p]$, and $\widehat{\mathbf{u}}[(D/P) \cdot p] = \frac{T\sqrt{P}}{\sqrt{D}} \frac{\sqrt{D}}{\sqrt{P}T^{3/4}} \widehat{\mathbf{u}}_{\mathbf{p}}[p] = T^{1/4} \widehat{\mathbf{u}}_{\mathbf{p}}[p]$. This means that:

$$(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P) \cdot p] = \frac{\sqrt{D}}{\sqrt{P}} \widehat{\mathbf{u}}_{\mathbf{p}}[p] \widehat{\mathbf{v}}_{\mathbf{p}}[p] = \frac{\sqrt{D}}{\sqrt{P}} \widehat{\mathbf{p}}[p] = \widehat{\mathbf{w}}[(D/P) \cdot p],$$

as desired. □

Lemma 16 (Lower bound)**.** *Consider vectors* $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ *specified by* $\mathbf{p} \in \mathbb{R}^P$ *(such that* D *is a multiple of* P*).*
(a) For any $K \leq P$*, it holds that* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \geq \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ *if* $K = P \cdot T$ *for integer* T*, then* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \geq 2\sqrt{\frac{D}{K}} \|\widehat{\mathbf{w}}\|_1.$$

Proof of Lemma 16. We use the dual formulation in eq. (21). Our approach is to construct a dual vector $\boldsymbol{\lambda} \in \mathbb{C}^D$ so that eq. (21) achieves the desired objective.

Case 1: $\mathbf{K} \leq \mathbf{P}$**.** Let $\boldsymbol{\lambda}_{\mathbf{p}} \in \mathbb{C}^P$ be the dual optimal solution for $\mathcal{R}^{\text{SDP}}_K(\mathbf{p})$. Now, let $\boldsymbol{\lambda}[(D/P) \cdot p] = \frac{\sqrt{D}}{\sqrt{P}} \boldsymbol{\lambda}_{\mathbf{p}}[p]$ for $0 \leq p \leq P-1$, and $\boldsymbol{\lambda}[d] = 0$ if $(D/P) \nmid d$. Notice that the objective becomes:

$$\begin{aligned}
2\sum_{d=0}^{D} |\boldsymbol{\lambda}[d]| \cdot |\widehat{\mathbf{w}}[d]| &= 2\sum_{p=0}^{P-1} |\boldsymbol{\lambda}[(D/P) \cdot p]| \cdot |\widehat{\mathbf{w}}[(D/P) \cdot p]| \\
&= 2\frac{D}{P} \sum_{p=0}^{P-1} |\boldsymbol{\lambda}_{\mathbf{p}}[p]| \cdot |\widehat{\mathbf{p}}[p]| \\
&= 2\frac{D}{P} \sum_{p=0}^{P-1} |\boldsymbol{\lambda}_{\mathbf{p}}[p]| \cdot |\widehat{\mathbf{p}}[p]| \\
&= \frac{D}{P} \mathcal{R}^{\text{SDP}}_K(\mathbf{p}) \\
&= \frac{D}{P} \mathcal{R}_{K,C}(\mathbf{p}),
\end{aligned}$$

where the last equality follows from tightness of the SDP (Theorem 4). It thus suffices to show that $\boldsymbol{\lambda}$ is dual feasible. For $\mathbf{x} \in \mathbb{C}^K$ such that $\|\mathbf{x}\| \leq 1$, consider:

$$\begin{aligned}
\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 &= \sum_{p=0}^{P-1} |\widehat{\mathbf{x}}[(D/P) \cdot p]|^2 \cdot |\boldsymbol{\lambda}[(D/P) \cdot p]|^2 \\
&= \frac{D}{P} \sum_{p=0}^{P-1} |\widehat{\mathbf{x}}[(D/P) \cdot p]|^2 \cdot |\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2.
\end{aligned}$$

Now, let $\widehat{x}^{(P)} \in \mathbb{C}^P$ be the Fourier representation of x when the base dimension is P. Observe that $\widehat{\mathbf{x}}[(D/P)\cdot p]$ is equal to $\sqrt{\frac{P}{D}}\widehat{\mathbf{x}}^{(P)}[p]$. Thus the above expression is equal to:

$$\frac{D}{P}\frac{P}{D}\sum_{p=0}^{P-1}|\widehat{\mathbf{x}}^{(P)}[p]|^2\cdot|\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2=\sum_{p=0}^{P-1}|\widehat{\mathbf{x}}^{(P)}[p]|^2\cdot|\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2.$$

Since $\boldsymbol{\lambda}_{\mathbf{p}}[p]$ is dual feasible for the P-dimensional problem, we see that this is at most 1, as desired.

Case 2: $\mathbf{K}=\mathbf{T}\cdot\mathbf{P}$. We consider $\boldsymbol{\lambda}[(D/P)\cdot p]=\frac{\sqrt{D}}{\sqrt{K}}$ for $0\le p\le P-1$, and $\boldsymbol{\lambda}[d]=0$ if $(D/P)\nmid d$. Notice that the objective in eq. (21) is equal to:

$$2\sum_{d=0}^{D}|\boldsymbol{\lambda}[d]|\cdot|\widehat{\mathbf{w}}[d]|=2\frac{\sqrt{D}}{\sqrt{K}}\sum_{p=0}^{P-1}|\widehat{\mathbf{w}}[(D/P)\cdot p]|=2\frac{\sqrt{D}}{\sqrt{K}}\|\widehat{\mathbf{w}}\|_1,$$

as desired. It thus suffices to show that that $\boldsymbol{\lambda}$ is dual feasible. For $\mathbf{x}\in\mathbb{C}^K$ such that $\|\mathbf{x}\|\le 1$, we consider

$$\sum_{d=0}^{D-1}|\widehat{\mathbf{x}}[d]|^2\cdot|\boldsymbol{\lambda}[d]|^2=\frac{D}{K}\sum_{p=0}^{P-1}|\widehat{\mathbf{x}}[(D/P)\cdot p]|^2.$$

Let $\mathbf{x}_0,\dots,\mathbf{x}_{T-1}\in\mathbb{C}^P$ be defined so that $\mathbf{x}_t[p]=\mathbf{x}[t\cdot K+p]$ for $0\le p\le P-1$ and $0\le t\le T-1$. Now, let $\widehat{\mathbf{x}}_t^{(P)}$ denote the Fourier representation of $\mathbf{x}$ when the base dimension is P, and observe that $\widehat{\mathbf{x}}[(D/P)\cdot p]=\sqrt{\frac{P}{D}}\sum_{t=0}^{T-1}\widehat{\mathbf{x}}_t^{(P)}[p]$. Thus we can rewrite the above expression as:

$$\begin{aligned}\frac{D}{K}\frac{P}{D}\sum_{p=0}^{P-1}\left|\sum_{t=0}^{T-1}\widehat{\mathbf{x}}_t^{(P)}[p]\right|^2&=\frac{1}{T}\sum_{p=0}^{P-1}\sum_{t=0}^{T-1}T\left|\widehat{\mathbf{x}}_t^{(P)}[p]\right|^2\\&=\sum_{t=0}^{T-1}\left\|\widehat{\mathbf{x}}_t^{(P)}\right\|^2\\&=\sum_{t=0}^{T-1}\|\mathbf{x}_t\|^2\\&=\|\mathbf{x}\|^2\le 1,\end{aligned}$$

as desired. This completes the proof. □

From these matching upper and lower bounds, we can easily conclude Lemma 7.

Proof of Lemma 7. Lemma 7 follows directly from Lemma 15 and Lemma 16. □

D Appendix for Section 5: Multi-input channels

For all the results in this appendix, we recall that the weights of the first and second layer are denoted as $\boldsymbol{\mathcal{U}}=\{\mathbf{U}_r\}_{r\in[R]}$ with $\mathbf{U}_r\in\mathbb{R}^{K\times C}\forall_{r\in[R]}$ and $\mathbf{V}\in\mathbb{R}^{D\times C}$, respectively.

D.1 Proof of Lemma 8: Realizability of linear functions

Lemma 8. *For any K, C and R, in order for the the network represented by $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in eq. (13) to realizes all linear maps in $\mathbb{R}^{D\times R}$ it is necessary that $K \cdot C \geq \min\{R, D\}$.*

Proof. We first reiterate the expressions for the linear predictor $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in term of $\boldsymbol{\mathcal{U}}, \mathbf{V}$ from the main text:

$$\forall_{r\in[R]},\ W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \sum_{c=0}^{C-1} \Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow}. \tag{22}$$

We will now express the above formulation as matrix multiplication using the following new notation: $\forall_{c\in C}$ let $\underline{\mathbf{U}}_c \in \mathbb{R}^{K\times R}$ denote the representation of first layer weights corresponding to each output channel such that $\forall_{r\in R},\ \underline{\mathbf{U}}_c[:, r] = \mathbf{U}_r[:, c]$.

For $c \in C$, consider the following matrix which consists of first K columns of the circulant matrix formed by $\mathbf{V}[:, c]$:

$$\widetilde{\mathbf{V}}_c = \frac{1}{\sqrt{D}} \begin{bmatrix} \mathbf{V}[0, c] & \mathbf{V}[D-1, c] & \cdots & \mathbf{V}[D-K+1, c] \\ \mathbf{V}[1, c] & \mathbf{V}[0, c] & \cdots & \mathbf{V}[D-K+2, c] \\ \vdots & \vdots & & \vdots \\ \mathbf{V}[D-1, c] & \mathbf{V}[D-2, c] & \cdots & \mathbf{V}[D-K, c] \end{bmatrix} \in \mathbb{R}^{D\times K}$$

Based on these notation we can check by following the definitions that for all c, $\big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\big)^{\downarrow} = \widetilde{\mathbf{V}}_c \mathbf{U}_r[:, c] = \widetilde{\mathbf{V}}_c \underline{\mathbf{U}}_c[:, r]$. We can thus write the expression of $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ as follows:

$$W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \sum_{c=0}^{C-1} \widetilde{V}_c \underline{\mathbf{U}}_c. \tag{23}$$

We now observe that each term in the summation $\widetilde{V}_c \underline{\mathbf{U}}_c$ is of rank utmost K as $\widetilde{V}_c \in \mathbb{R}^{D\times K}$ and $\underline{\mathbf{U}}_c \in \mathbb{R}^{D\times K}$. Thus, for any $\boldsymbol{\mathcal{U}}, \mathbf{V}$, $\text{rank}(W(\boldsymbol{\mathcal{U}}, \mathbf{V})) \leq K \cdot C$. From this we conclude that in order to realize all linear maps in the multi-channel input space of $\mathbb{R}^{D\times R}$, we necessarily need $K \cdot C \geq \min\{R, D\}$.

Additionally, in eq. (23) we see that since $\underline{\mathbf{U}}_c$ are unconstrained, each term in the sum can realize any rank 1 matrix. This implies that $C \geq \min\{R, D\}$ is a sufficient condition for $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ to realize any $\mathbf{W} \in \mathbb{R}^{D\times R}$. However, from Theorem 10 we know that this condition is not necessary. It is an open question to derive the tightest necessary and sufficient conditions.

We finally note that a similar proof can be shown using the Fourier representation in eq. (13). □

D.2 Derivation of SDP relaxation for multi-channel input networks

The SDP relaxation for $\mathcal{R}_{K,C,R}$ is derived similarly to for networks with a single input channel. We define a rank C positive semidefinite matrix $\mathbf{Z} \in \mathbb{R}^{(D+K\cdot R)\times(D+K\cdot R)}$, that represents:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \dots \\ \mathbf{U}_R \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^\top & \mathbf{U}_1^\top & \dots & \mathbf{U}_R^\top & \mathbf{V}^\top \end{bmatrix} \succcurlyeq 0.$$

We also define Hermitian matrices $\mathbf{A}^{\text{real}}_{d,r}, \mathbf{A}^{\text{img}}_{d,r} \in \mathbb{R}^{(D+K\cdot R)\times(D+K\cdot R)}$ for $d \in [D], r \in [R]$ as follows. As in the single input channel case, we use $\mathbf{Q}_d = \overline{\mathbf{F}}_K^\top \mathrm{e}_d \mathrm{e}_d^\top \overline{\mathbf{F}}$. These matrices take the following form:

$$\mathbf{A}^{\text{real}}_{d,0} = \left[\begin{array}{ccc:c} & & & \mathbf{Q}_d \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & \mathbf{0} \\ & & & \vdots \\ & & & \mathbf{0} \\ \hdashline \overline{\mathbf{Q}}_d & \mathbf{0} & \dots & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}^{\text{img}}_{d,0} = \left[\begin{array}{ccc:c} & & & i\cdot\mathbf{Q}_d \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & \mathbf{0} \\ & & & \vdots \\ & & & \mathbf{0} \\ \hdashline -i\cdot\overline{\mathbf{Q}}_d & \mathbf{0} & \dots & \mathbf{0}_{D\times D} \end{array}\right],$$

$$\mathbf{A}^{\text{real}}_{d,1} = \left[\begin{array}{ccc:c} & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & \mathbf{Q}_d \\ & & & \vdots \\ & & & \mathbf{0} \\ \hdashline \mathbf{0} & \overline{\mathbf{Q}}_d & \dots & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}^{\text{img}}_{d,1} = \left[\begin{array}{ccc:c} & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & i\cdot\mathbf{Q}_d \\ \mathbf{0} & & & \\ & & & \vdots \\ & & & \mathbf{0} \\ \hdashline \mathbf{0} & -i\cdot\overline{\mathbf{Q}}_d & \dots & \mathbf{0}_{D\times D} \end{array}\right],$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\mathbf{A}^{\text{real}}_{d,R} = \left[\begin{array}{ccc:c} & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & \mathbf{0} \\ & & & \vdots \\ & & & \mathbf{Q}_d \\ \hdashline \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}^{\text{img}}_{d,R} = \left[\begin{array}{ccc:c} & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & \mathbf{0} \\ & & & \vdots \\ & & & i\cdot\mathbf{Q}_d \\ \hdashline \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}_{D\times D} \end{array}\right].$$

We also provide a more formal description of these matrices. We provide a block-wise description of only the upper diagonal blocks, with lower diagonal blocks filled to satisfy the Hermitian matrix property. Additionally, for matrices $\{\mathbf{A}^{\text{real}}_{d,r}, \mathbf{A}^{\text{img}}_{d,r}\}_{d\in[D],r\in R}$ any unspecified block is by default treated as zero matrix $\mathbf{0}$ of appropriate dimension:

- For $r_1, r_2 \in [R]$ with $r_2 \geq r_1$, the $K \times K$ block with indices $(r_1 : (r_1+1)K$ along rows and $r_2 : (r_2+1)K$ along columns is given as follows: $\mathbf{Z}[r_1 : (r_1+1)K, r_2 : (r_2+1)K] = \mathbf{U}_{r_1}\mathbf{U}_{r_2}^\top$;
- For $r \in [R]$, the $K \times D$ blocks with indices $r : (r+1)K, RK : (D+RK)$, are given as follows: $\mathbf{Z}[r : (r+1)K, RK : (D+RK)] = \mathbf{U}_r\mathbf{V}^\top$, and for all $d \in [D]$, $\mathbf{A}^{\text{real}}_{d,r}[r : (r+1)K, RK : (D+RK)] = \mathbf{Q}_d$ and $\mathbf{A}^{\text{img}}_{d,r}[r : (r+1)K, RK : (D+RK)] = i\cdot\mathbf{Q}_d$. Note that for $r' \neq r$, the corresponding blocks in $\mathbf{A}^{\text{real}}_{d,r'}, \mathbf{A}^{\text{img}}_{d,r'}$ remain the default zero matrix.
- Finally, the lower-right $D \times D$ block is given as $\mathbf{Z}[RK : (D+RK), RK : (D+RK)] = \mathbf{V}\mathbf{V}^\top$.

Using this notation, we consider the following SDP relaxation of $\mathcal{R}_{K,C,R}(\mathbf{W})$ in terms of Fourier coefficients $\widehat{\mathbf{W}} = \mathbf{F}\mathbf{W}$:

$$\begin{aligned} \mathcal{R}^{\text{SDP}}_{K,R}(\mathbf{W}) \;=\; \min_{\mathbf{Z}\succcurlyeq 0} \quad & \langle \mathbf{Z}, \mathbf{I}\rangle \\ \text{s.t.} \quad & \forall d \in [D], r \in [R] \;\; \langle \mathbf{Z}, \mathbf{A}^{\text{real}}_{d,r}\rangle = 2\mathrm{Re}(\widehat{\mathbf{W}}[d,r]) \\ & \forall d \in [D], r \in [R] \;\; \langle \mathbf{Z}, \mathbf{A}^{\text{img}}_{d}\rangle = 2\mathrm{Im}(\widehat{\mathbf{W}}[d,r]). \end{aligned} \tag{24}$$

We can check that the SDP formulation with an additional rank constraint of $\mathrm{rank}\,(\mathbf{Z}) \leq C$ is equivalent $\mathcal{R}_{K,C,R}(\mathbf{W})$ and the SDP thus provides a lower bound: *i.e.,* $\forall \mathbf{W}$, $\mathcal{R}_{K,C,R}(\mathbf{W}) \geq$

$\mathcal{R}^{\text{SDP}}_{K,R}(\mathbf{W})$. As we previously discussed, unlike networks with single channel input, the SDP relaxation here is not always tight when $R > 1$ as sufficiently large C is required to merely realize all matrix-valued linear function over the input space.

However, in the special cases of $K = 1$ and $K = D$, we can show that the SDP is tight for all $C \geq R/K$ and further derive interesting closed form expressions of $\mathcal{R}_{K,C,R}(\mathbf{W})$.

D.3 Proofs of Theorems 9-10: induced regularizer for $K = 1$ and $K = D$

Theorem 9 (Multi-input channel $K = 1$)**.** *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq \min\{R, D\}$, the induced regularizer for $K = 1$ is given by the scaled nuclear norm $\|.\|_*$:*

$$\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_* = 2\sqrt{D}\|\widehat{\mathbf{W}}\|_*.$$

Proof. For $K = 1$, we have that $\forall_{r\in[R]}$, $\mathbf{U}_r \in \mathbb{R}^{1\times C}$. For this proof, we stack the vectors $\mathbf{U}_r$ to obtain $\widetilde{\mathbf{U}} \in \mathbb{R}^{R\times C}$ such that $\forall_{r\in[R]}$, $\widetilde{\mathbf{U}}[r, :] = \mathbf{U}_r$.

We work with the signal space definition of the linear predictor realized by network as: $\forall_{r\in[R]}$, $W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \sum_{c=0}^{C-1}\left(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\right)^{\downarrow}$ (from eq. (13)).

For kernel size of 1, we notice that from the definition of convolution in Definition 1, we have the following:

$$\left(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\right)^{\downarrow} = \frac{\tilde{\mathbf{U}}[r, c]}{\sqrt{D}}\mathbf{V}[:, c] \propto \mathbf{V}[:, c]. \tag{25}$$

Plugging this back into the expression of $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$, we have the following:

$$W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \frac{1}{\sqrt{D}}\mathbf{V}\widetilde{\mathbf{U}}[r, :] \implies W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \frac{1}{\sqrt{D}}\mathbf{V}\widetilde{\mathbf{U}}^{\top}. \tag{26}$$

In the above formulation $\mathbf{V}, \widetilde{\mathbf{U}}$ are of rank K, but otherwise completely unconstrained. Thus, they can realize any rank K matrix. So as long as $C \geq \min\{R, D\}$, the network can realize any linear predictor $\mathbf{W} \in \mathbb{R}^{D\times R}$.

The rest of the proof follows from connecting the above expression into the variational characterization of the nuclear norm.

The induced regularizer $\mathcal{R}_{1,C,R}(\mathbf{W})$ from eq. (14) for $C \geq \min\{R, D\}$ can now be expressed as follows:

$$\begin{aligned} \mathcal{R}_{1,C,R}(\mathbf{W}) = \min_{\widetilde{\mathbf{U}}\in\mathbb{R}^{R\times C}, \mathbf{V}\in\mathbb{R}^{D\times C}} \quad & \|\widetilde{\mathbf{U}}\|^2 + \|\mathbf{V}\|^2 \\ \text{s.t.,} \quad & \sqrt{D}\mathbf{W} = \mathbf{V}\widetilde{\mathbf{U}}^{\top}. \end{aligned} \tag{27}$$

For $C \geq \min\{R, D\}$ eq. (27) is exactly the variational definition of nuclear norm (see Rennie and Srebro [Lemma 1 RS05]) and $\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_*$ for even unbounded C. The fact that $C = \min\{R, D\}$ is sufficient can be seen by obtaining the optimum nuclear norm as upper bound from using $\mathbf{V} = \mathbf{L}\sqrt{\boldsymbol{\Sigma}}$ and $\widetilde{\mathbf{U}} = \mathbf{R}\sqrt{\boldsymbol{\Sigma}}$, where $\sqrt{D}\mathbf{W} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{R}^{\top}$ is the singular value decomposition of $\sqrt{D}\mathbf{W}$. Finally, we note that based on our normalization of Fourier transform, we have $\|\mathbf{W}\|_* = \|\widehat{\mathbf{W}}\|_*$. This completes our proof.

For $K = 1$ we provided the proof of $\mathcal{R}_{1,C,R}(\mathbf{W})$ in the signal space of $\mathbf{W}$, but the Theorem can also be proved in the Fourier domain (similar to the proof of Theorem 10 given below) by first showing that the SDP relaxation evaluates to the nuclear norm and then showing matching upper bound for $\mathcal{R}_{1,C,R}(\mathbf{W})$ as shown above. □

Theorem 10 (Multi-input channel $K = D$)**.** *For any* $\mathbf{W} \in \mathbb{R}^{D\times R}$*, and any* $C \geq 1$*, the induced regularizer for* $K = D$ *is given as follows*

$$\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1} := \sum_{s=0}^{D-1} \sqrt{\sum_{r=0}^{R-1} \left|\widehat{\mathbf{W}}[d,r]\right|^2}.$$

Proof. We begin by expressing induced regularizer for full dimensional kernel sizes $\mathcal{R}_{D,C,R}(\mathbf{W})$ from eq. (14) in terms of the Fourier representation of the linear predictor realized by the network $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$:

$$\begin{aligned} \mathcal{R}_{D,C,R}(\mathbf{W}) = \inf_{\boldsymbol{\mathcal{U}},\mathbf{V}} \ & \|\widehat{\mathbf{U}}_r\|^2 + \|\widehat{\mathbf{V}}\|^2 \\ \text{s.t.,} \ & \forall_{r\in[R]} \mathbf{W}[:,r] = \operatorname{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^\top), \end{aligned} \tag{28}$$

where $\boldsymbol{\mathcal{U}}, \mathbf{V}$ are of dimensions $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r \in \mathbb{R}^{D\times C}\}_{r\in[R]}, \mathbf{V} \in \mathbb{R}^{D\times C}$.

Our proof for the case of multi-input channel networks with $K = D$ follows the following structure:

Step 1. We first show an upper bound on the induced regularizer for single output channel $C = 1$ with full dimensional kernel $K = D$ as $\mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}$ by providing a construction of $\boldsymbol{\mathcal{U}}, \mathbf{V}$. It immediately follows from the monotonicity of $\mathcal{R}_{D,C,R}$ that for all C, $\mathcal{R}_{D,C,R}(\mathbf{W}) \leq \mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}$. This step also consequently shows that when $K = D$, every linear predictor over the multi-channel input space of $\mathbb{R}^{D\times R}$ is realizable by a network with even a single output channel.

Step 2. The bulk of our proof lies in matching the upper bound with a lower bound on the dual problem of the SDP in eq. (24) as $\mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq 2\|\widehat{\mathbf{W}}\|_{2,1}$. This gives us that for all $C \geq 1$, $\mathcal{R}_{D,C,R}(\mathbf{W}) \geq \mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq 2\|\widehat{\mathbf{W}}\|_{2,1}$.

Step 1. Upper bound on the induced regularizer $\mathcal{R}_{D,C,R}$ We first show that $\mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1} = 2\sum_{d\in[D]} \|\widehat{\mathbf{W}}[d,:]\|$. Since $\mathcal{R}_{D,R,C}(\mathbf{W})$ is decreasing in C, it suffices to show this for $C = 1$. For $C = 1$ and $K = D$, we have $\mathbf{V} \in \mathbb{R}^D$ and $\forall_{r\in[R]}\ \mathbf{U}_r \in \mathbb{R}^D$. For full dimensional kernels, the Fourier domain representations $\widehat{\mathbf{U}}_0, \widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_R, \widehat{\mathbf{V}} \in \mathbb{C}^D$ are all constrained beyond the symmetry properties of Fourier transform of real matrices. Thus consider the following $\boldsymbol{\mathcal{U}}, \mathbf{V}$:

$$\forall_{d\in[D]}\forall_{r\in[R]}, \widehat{\mathbf{U}}_r[d] = \frac{\widehat{\mathbf{W}}[d,r]}{\sqrt{\|\widehat{\mathbf{W}}[d,:]\|}}, \quad \text{and} \quad \widehat{\mathbf{V}}[d] = \sqrt{\|\widehat{\mathbf{W}}[d,:]\|}. \tag{29}$$

It is easy to see that the above $\boldsymbol{\mathcal{U}}, \mathbf{V}$ satisfy the constraints of $\mathcal{R}_{D,1,R}(\mathbf{W})$ in eq. (28) that $\forall_{r\in[R]}\mathbf{W}[:,r] = \operatorname{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^\top)$. Further $\widehat{\mathbf{U}}_0, \widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_R, \widehat{\mathbf{V}} \in \mathbb{C}^D$ satisfy the required symmetry properties since $\mathbf{W}$ is real.

Now computing the objective, we immediately have that $\|\widehat{\mathbf{V}}\|^2 = \sum_{d\in[D]} \|\widehat{\mathbf{W}}[d,:]\|$, and further,

$$\sum_{r\in[R]} \|\widehat{\mathbf{U}}_r\|^2 = \sum_{d\in[D]} \left[\sum_{r\in[R]} \frac{|\widehat{\mathbf{W}}[d,r]|^2}{\|\widehat{\mathbf{W}}[d,:]\|}\right] = \sum_{d\in[D]} \|\widehat{\mathbf{W}}[d,:]\|.$$

This construction thus gives us the desired upper bound

$$\mathcal{R}_{D,C,R}(\mathbf{W}) \leq \mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}. \tag{30}$$

Step 2: Lower bound on the induced regularizer $\mathcal{R}_{D,C,R}$ We show the lower bound by lower bounding the dual problem of the SDP in eq. (24).

For $r \in [R]$, let $\boldsymbol{\lambda}_r^{\text{real}} \in \mathbb{R}^D$ and $\boldsymbol{\lambda}_r^{\text{img}} \in \mathbb{R}^D$ denote the dual variables corresponding to the constraints in the SDP in eq. (24) for the real and imaginary parts, respectively, of $\widehat{\mathbf{W}}[:, r]$. Similar to single input channel proofs, we define $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_r^{\text{real}} + i \cdot \boldsymbol{\lambda}_r^{\text{img}}$ and $\boldsymbol{\Lambda}_r = \text{diag}(\boldsymbol{\lambda}_r)$. Additionally, we introduce the notation for the matrix obtained by taking $\{\boldsymbol{\lambda}_r\}_r$ as columns: $\Xi = [\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_R] \in \mathbb{C}^{D \times R}$ such that $\forall_r$, $\Xi[:, r] = \boldsymbol{\lambda}_r$.

Based on weak duality for the SDP in eq. (24), we have the following:

$$\begin{aligned} \mathcal{R}_{D,R}^{\text{SDP}}(\mathbf{W}) \geq \max_{\Xi = [\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_R]} \quad & 2 \cdot \text{Re}(\langle \widehat{\mathbf{W}}, \Xi \rangle) \\ \text{s.t.,} \quad & \sum_{d,r} (\boldsymbol{\lambda}_r^{\text{real}}[d] \cdot \mathbf{A}_{d,r}^{\text{real}} + \boldsymbol{\lambda}_r^{\text{img}}[d] \cdot \mathbf{A}_{d,r}^{\text{img}}) \preccurlyeq \mathbf{I}. \end{aligned} \tag{31}$$

Our rest of the proof obtains the lower bound by constructing an appropriate Ξ satisfying the constraints:

From the definitions of $\{\mathbf{A}_{d,r}^{\text{real}}, \mathbf{A}_{d,r}^{\text{img}}\}_{d \in [D], r \in R}$ in Appendix D.2 and using $\mathbf{Q}_d = \overline{\mathbf{F}} \mathrm{e}_d \mathrm{e}_d^\top \overline{\mathbf{F}}$ (note that for $K = D$ and $\mathbf{F}_K = \mathbf{F} = \mathbf{F}^\top$), we have the following:

$$\sum_{d,r} (\boldsymbol{\lambda}_r^{\text{real}}[d] \cdot \mathbf{A}_{d,r}^{\text{real}} + \boldsymbol{\lambda}_r^{\text{img}}[d] \cdot \mathbf{A}_{d,r}^{\text{img}}) = \left[\begin{array}{ccc:c} & \mathbf{0}_{(R \cdot D) \times (R \cdot D)} & & \begin{matrix} \overline{\mathbf{F}} \boldsymbol{\Lambda}_0 \overline{\mathbf{F}} \\ \overline{\mathbf{F}} \boldsymbol{\Lambda}_1 \overline{\mathbf{F}} \\ \vdots \\ \overline{\mathbf{F}} \boldsymbol{\Lambda}_R \overline{\mathbf{F}} \end{matrix} \\ \hdashline \overline{\mathbf{F}} \overline{\boldsymbol{\Lambda}}_0 \mathbf{F}_K & \overline{\mathbf{F}} \overline{\boldsymbol{\Lambda}}_1 \mathbf{F}_K & \ldots \quad \overline{\mathbf{F}} \overline{\boldsymbol{\Lambda}}_R \mathbf{F}_K & \mathbf{0}_{D \times D} \end{array}\right] \tag{32}$$

We state and prove the following claim:

Claim. *Ξ satisfies the constraints of the dual problem in the RHS of eq. (31) if* $\max_{d \in [D]} \|\Xi[d, :]\| \leq 1$

Proof of claim. The relevant constraint in eq. (31) is $\sum_{d,r} (\boldsymbol{\lambda}_r^{\text{real}}[d] \cdot \mathbf{A}_{d,r}^{\text{real}} + \boldsymbol{\lambda}_r^{\text{img}}[d] \cdot \mathbf{A}_{d,r}^{\text{img}}) \preccurlyeq \mathbf{I}$. It thus suffices to show that for all $\mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{y}\| = 1$, it holds that $\sum_{r=0}^R \|\overline{\mathbf{F}} \boldsymbol{\Lambda}_r \overline{\mathbf{F}} \mathbf{y}\|^2 \leq 1$. We have the following set of inequalities that prove the claim $\forall_{\mathbf{y}: \|\mathbf{y}\| = 1}$:

$$\begin{aligned} \sum_{r=0}^R \|\overline{\mathbf{F}} \boldsymbol{\Lambda}_r \overline{\mathbf{F}} \mathbf{y}\|^2 &\overset{(a)}{=} \sum_{r=0}^R \|\boldsymbol{\Lambda}_r (\overline{\mathbf{F}} \mathbf{y})\|^2 \\ &= \sum_{d=0}^D \sum_{r=0}^R |\boldsymbol{\lambda}_r[d]|^2 \cdot |(\overline{\mathbf{F}} \mathbf{y})[d]|^2 \overset{(b)}{=} \sum_{d=0}^D \|\Xi[d, :]\|^2 \cdot |(\overline{\mathbf{F}} \mathbf{y})[d]|^2 \\ &\leq \max_{d \in [D]} \|\Xi[d, :]\|^2 \|\overline{\mathbf{F}} \mathbf{y}\|^2 \\ &\overset{(c)}{=} \max_{d \in [D]} \|\Xi[d, :]\|^2, \end{aligned}$$

where (a) follows from $\overline{\mathbf{F}}$ being unitary, (b) from definition of Ξ, and (c) from $\|\mathbf{F}\mathbf{y}\| = \|\mathbf{y}\| = 1$. □

Consider Ξ defined as follows:

$$\forall_{d \in [D]}, \quad \Xi[d, :] = \frac{\widehat{\mathbf{W}}[d, :]}{\|\widehat{\mathbf{W}}[d, :]\|}$$

It is easy to check that $\max_{d\in[D]} \|\Xi[d,:]\| = 1$ and thus based on the claim we proved above Ξ satisfies the constraints of the dual optimization problem in the RHS of eq. (31). Additionally, the objective evaluates to the desired bound of $\mathrm{Re}(\langle\widehat{\mathbf{W}}, \Xi\rangle) = \sum_d \|\widehat{\mathbf{W}}[d,:]\| = \|\widehat{\mathbf{W}}\|_{2,1}$. We thus have the following lower bound:

$$\mathcal{R}_{D,C,R}(\mathbf{W}) \geq \mathcal{R}^{\mathrm{SDP}}_{D,R}(\mathbf{W}) \geq 2\,\mathrm{Re}(\langle\widehat{\mathbf{W}}, \Xi\rangle) = 2\|\widehat{\mathbf{W}}\|_{2,1}. \tag{33}$$

Conclusion of the proof. The proof of the Theorem follows from combining the matching upper and lower bounds in eq. (30) and eq. (33), respectively, to obtain $\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1}$. □

E Additional experiments

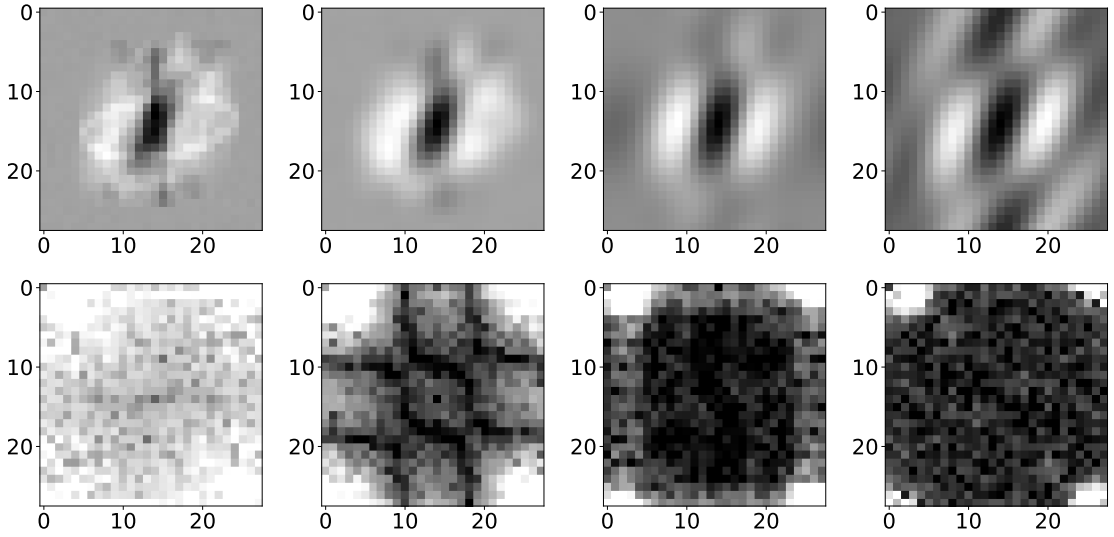
We give higher resolution versions for plots in Figure 1 in the paper. See Figure 3.

E.1 Experiments for the digits 2 and 9

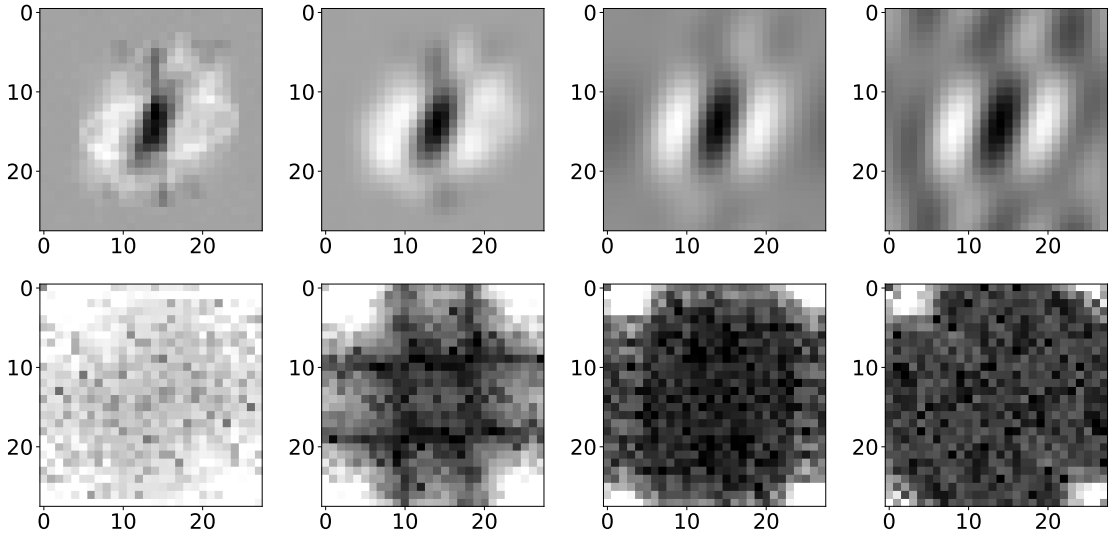
We repeat the experiments from Section 6.2 when the classes are the digits 2 and 9. See Figure 4 and Table 3. The predictors generally appear invariant across different values of C, although when $K = 28$, there are slight differences in the signal domain for different values of C.

C	$K:(1,1)$	$K:(3,3)$	$K:(9,9)$	$K:(28,28)$
1	18.91	9.62	7.37	6.15
2	18.92	9.66	7.11	5.82
4	18.93	9.65	7.16	5.90
8	18.95	9.72	7.17	5.83

Table 3: $\mathcal{R}_{K,C}(\mathbf{w})$ for the digits 2 and 9

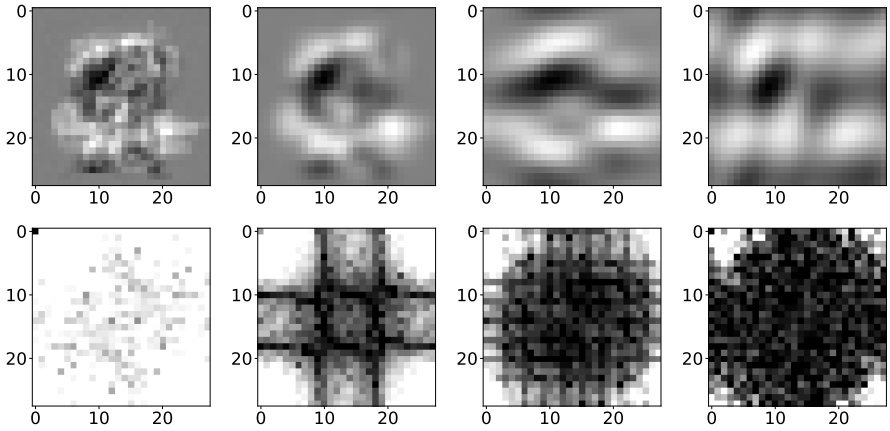


(a) $C = 2$

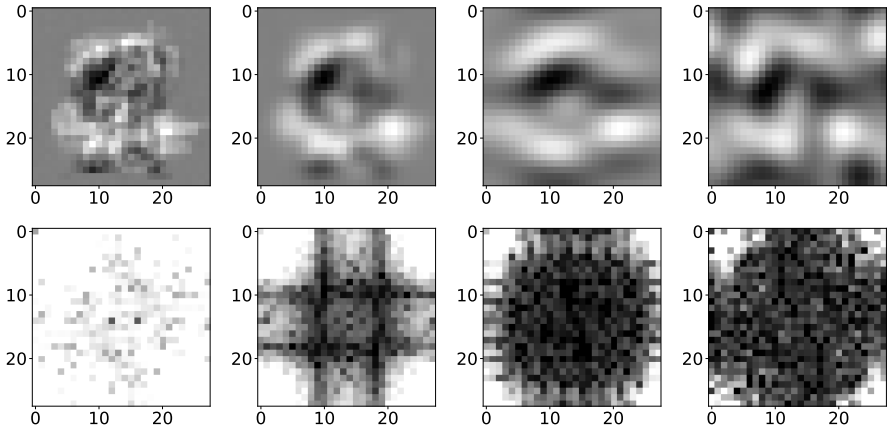


(b) $C = 4$

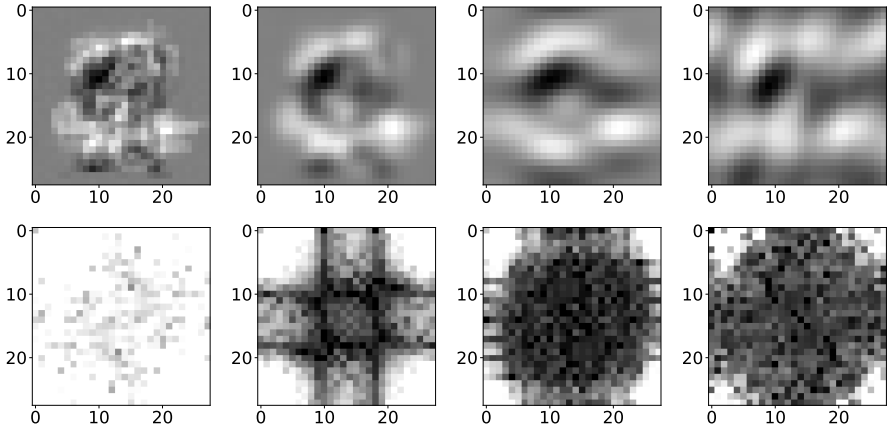
Figure 3: Linear predictor for kernel sizes $\{(1,1),(3,3),(9,9),(28,28)\}$ (left to right) for $C = 2$ (top) and $C = 4$ (bottom). The top row is the signal domain representation and the bottom row is the Fourier domain representation.



(a) $C = 1$



(b) $C = 2$



(c) $C = 4$

Figure 4: Linear predictor for kernel sizes $\{(1,1),(3,3),(9,9),(28,28)\}$ (left to right) for $C \in \{1,2,4\}$. The top row is the signal domain representation and the bottom row is the Fourier domain representation. The classes are 2 and 9.