Function space view of norm minimization in multi-channel linear convolutional network

Suriya Gunasekar

Meena Jagadeesan

Ilya Razensteyn

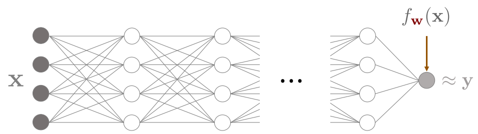
Learning prediction functions

INPUT $\mathbf{x}$ → y LABEL

$$f(\mathbf{x}) \approx y$$

PREDICTION FUNCTION
w/ PARAMETERS $\mathbf{w}$

$$f_{\mathbf{w}}(\mathbf{x})$$



Overparametrized models

"large" class of functions to optimize over

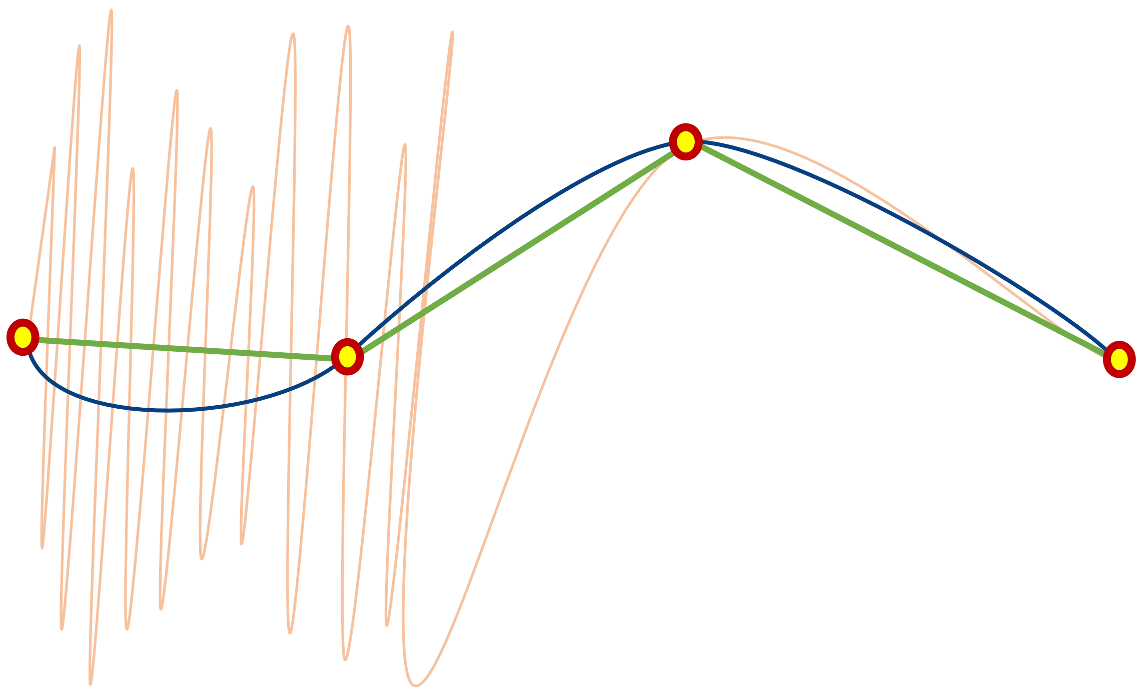
e.g., large neural networks
≈ all continuous functions

- multiple minimizers of the objective
- most functions fitting observed data will perform poorly on new examples

$$\min_{\mathbf{w}} \sum_{(\mathbf{x},y) \text{ in } D} \text{loss}(\, f_{\mathbf{w}}(\mathbf{x}),\; y \,)$$

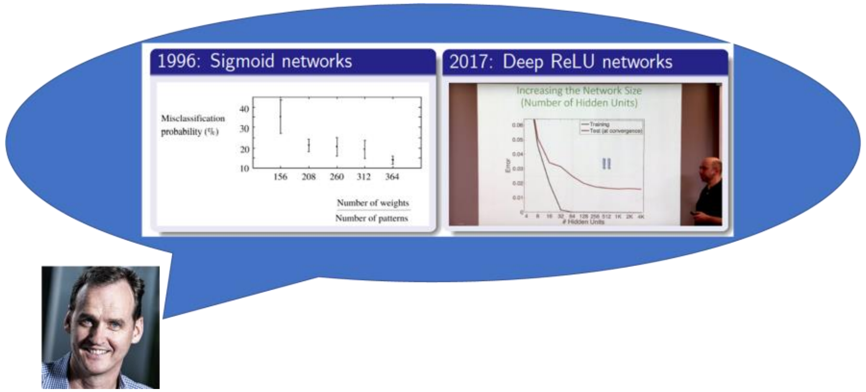
Common in deep learning practice

very large neural networks
+ large scale datasets
+ loss minimization using variations of (stochastic) gradient descent (GD)



Norm based capacity control

“The size [magnitude] of the weights is more important than the size [number of parameters] of the network.” (Bartlett, ‘97)



Weight norm based capacity control is ubiquitous

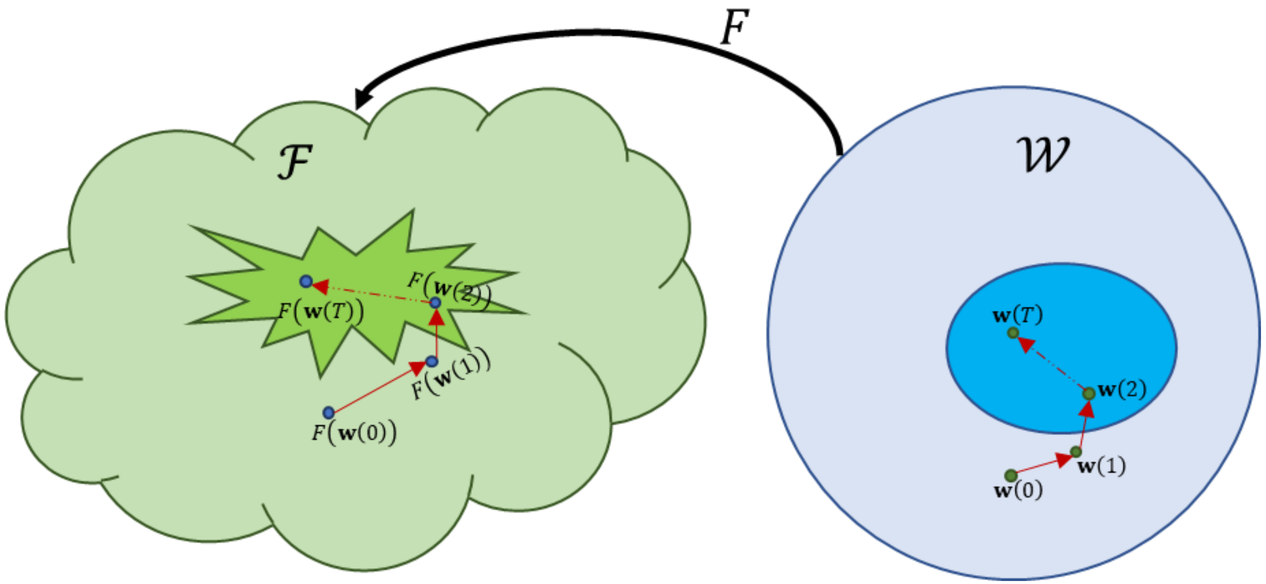
- Explicit regularization

  ℓ_2 norm (related to weight decay) is perhaps the most common tool
- Implicit regularization

  e.g., Lyu and Li '20, Nacson et al. '19, etc.

Aside from norm, other forms of capacity control are also common (e.g., combinatorial rank/sparsity constraints) but today we will focus on ℓ_2 norm

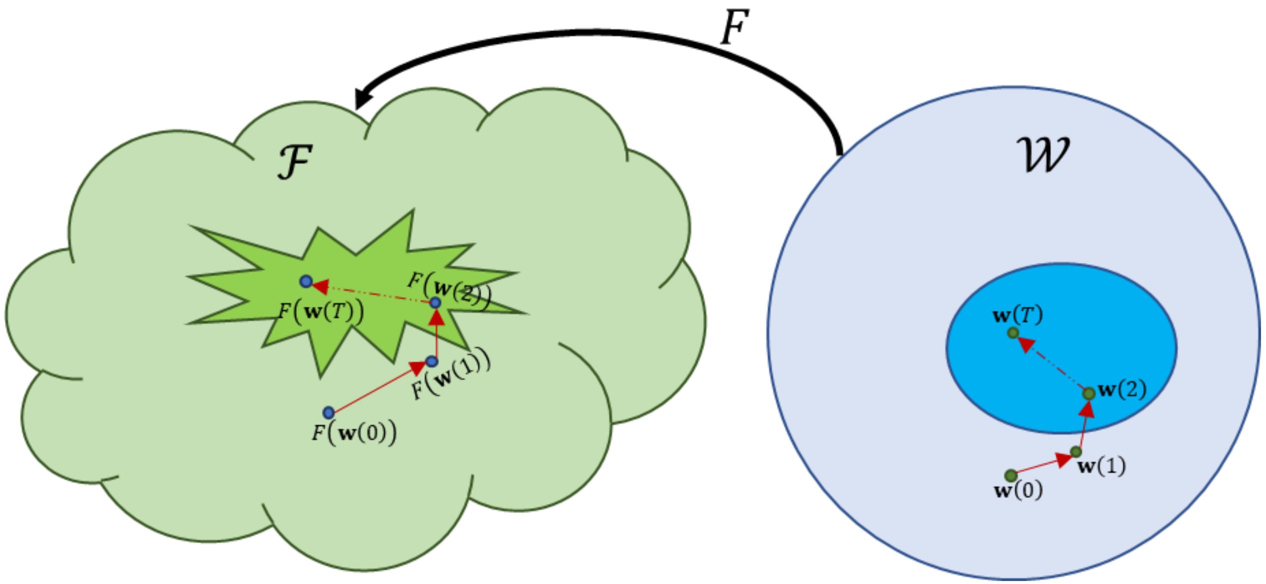
Q. What is the function space view of controlling ℓ_2 norm of parameters



different for different network architectures $f_{\text{arch}}(\mathbf{w},.)$

≈ different parametrizations of functions over inputs

Q. What is the function space view of controlling ℓ_2 norm of parameters



different for different network architectures $f_{\text{arch}}(\mathbf{w},.)$

$\approx$ different parametrizations of functions over inputs

INDUCED REGULARIZER

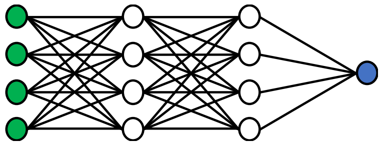
representational cost in units of weight norm

$$\mathcal{R}(f) := \inf_{\mathbf{w}} \|\mathbf{w}\|_2^2 \ \text{ s.t., } \ \forall \mathbf{x}, f(\mathbf{x}) = f_{\text{arch}}(\mathbf{w}, \mathbf{x})$$

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2 + L\Big(\{f_{\text{arch}}(\mathbf{w}, \mathbf{x}_n)\}_n \Big)$$

$$\equiv \min_{f} \mathcal{R}(f) + L\Big(\{f(\mathbf{x}_n)\}_n \Big)$$

Induced regularizer in function space

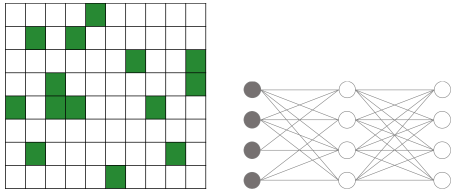


Linear fully connected networks with single output

$F(\mathbf{w}) = W_1 W_2 W_3 \ldots w_L \equiv \beta \in \mathbb{R}^d$

$$\mathcal{R}(\beta) = \|\beta\|_2$$

Gunasekar, Woodworth, et al. 2017; Gunasekar, Lee, Soudry, Srebro (2018)x2; Ji & Telgarsky (2018)x2;

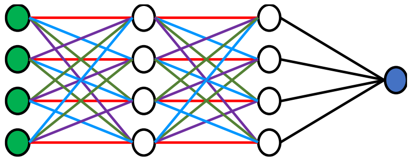


Matrix factorization e.g., matrix completion, multitask learning, matrix sensing, …

$F(\mathbf{w}) = W_1 W_2 \equiv W \in \mathbb{R}^{d_{in} \times d_{out}}$

$$\mathcal{R}(W) = \|W\|_*$$

Gunasekar, Woodworth, et al. 2017;

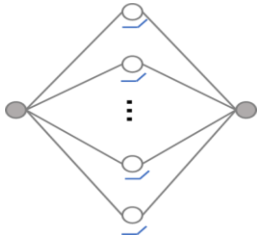


Linear fully width convolutional network

$F(\mathbf{w}) = \mathbf{w}_1 \star \mathbf{w}_2 \equiv \beta \in \mathbb{R}^d$

$$\mathcal{R}(\beta) = \|\mathrm{DFT}(\beta)\|_{\frac{2}{L}}$$

Gunasekar, Lee, Soudry, Srebro 2018; Edgar and Pilanchi 2020; Yun, Krishnan, Mobahi 2020



2-layer infinite (large) width ReLU network

$F(\mathbf{w})(x) = \sigma(x\mathbf{w}_1 + b)^\top \mathbf{w}_2 \equiv \{f : \mathbb{R} \to \mathbb{R}\}$

$\sigma(z) = \max\{x, 0\}$

$$\mathcal{R}(f) =^* \int |f''(x)| \mathrm{d}x$$

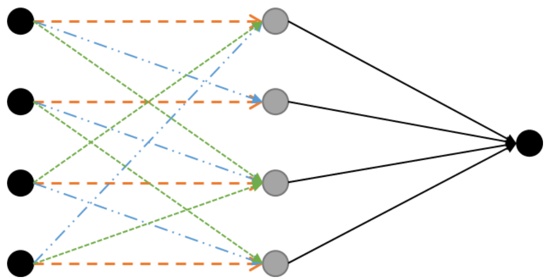
related to Radon transform for $D > 1$

Savarese, Soudry, Srebro 2019; Ongie, Willet, Soudry, Srebro 2020; Edgar and Pilanchi (2020)x3;

$$\mathcal{R}(f) := \inf_{\mathbf{w}} \|\mathbf{w}\|_2^2 \ \text{s.t.}, \ \forall \mathbf{x}, f(\mathbf{x}) = f_{\text{arch}}(\mathbf{w}, \mathbf{x})$$

influence of #channels & kernel size
in linear convolutional network

Linear Convolutional Network



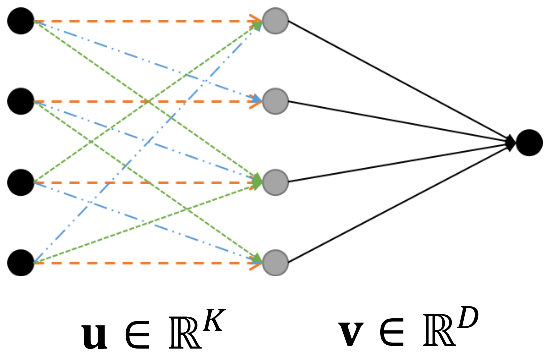
$$\mathbf{x} \rightarrow h_1(\mathbf{x}) = \mathbf{x} \star \mathbf{u} \rightarrow \mathbf{v}^\top h_1(\mathbf{x})$$

$$\begin{aligned} f((\mathbf{u}, \mathbf{v}), \mathbf{x}) &= \mathbf{v}^\top (\mathbf{x} \star \mathbf{u}) \\ &= \langle \mathbf{x}, \beta_{\mathbf{u},\mathbf{v}} \rangle \end{aligned}$$

where $\beta_{\mathbf{u},\mathbf{v}} = \mathbf{u} \star \mathbf{v}^{\downarrow}$

$$\mathcal{R}(\beta) = \inf_{\beta = \mathbf{u} \star \mathbf{v}^{\downarrow}} \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$$

Fourier trick & full-dimensional filter

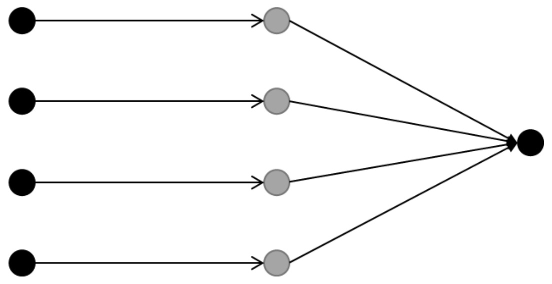


$$\mathbf{x} \to h_1(\mathbf{x}) = \mathbf{x} \star \mathbf{u} \to \mathbf{v}^\top h_1(\mathbf{x})$$

$$\beta_{\mathbf{u},\mathbf{v}} = \mathbf{u} \star \mathbf{v}^{\downarrow}$$

$$\mathcal{R}_K(\beta) = \inf_{\beta = \mathbf{u} \star \mathbf{v}^{\downarrow}} \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$$

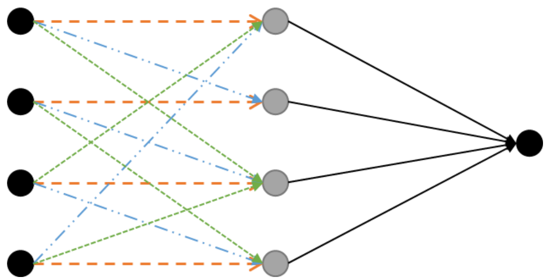
Fourier domain: $\hat{\boldsymbol{z}} = \mathcal{F}_D(\boldsymbol{z}) \in \mathbb{C}^D$



$$\hat{\mathbf{x}} \to \hat{\mathbf{x}} \odot \hat{\mathbf{u}}^* \to \hat{\mathbf{v}}^{*\top}(\hat{\mathbf{u}}^* \odot \mathbf{x})$$

$$\implies \hat{\beta} = \hat{\mathbf{u}} \odot \hat{\mathbf{v}}$$

Fourier trick & full-dimensional filter

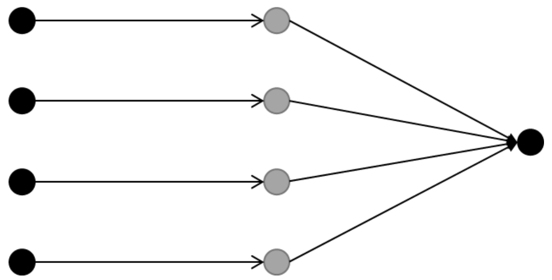


$$\mathbf{x} \to h_1(\mathbf{x}) = \mathbf{x} \star \mathbf{u} \to \mathbf{v}^\top h_1(\mathbf{x})$$

$$\beta_{\mathbf{u},\mathbf{v}} = \mathbf{u} \star \mathbf{v}^{\downarrow}$$

$$\mathcal{R}_K(\beta) = \inf_{\beta = \mathbf{u} \star \mathbf{v}^{\downarrow}} \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$$

Fourier domain: $\hat{\mathbf{z}} = \mathcal{F}_D(\mathbf{z}) \in \mathbb{C}^D$



$$\hat{\mathbf{x}} \to \hat{\mathbf{x}} \odot \hat{\mathbf{u}}^* \to \hat{\mathbf{v}}^{*\top}(\hat{\mathbf{u}}^* \odot \mathbf{x})$$

$$\Longrightarrow \hat{\beta} = \hat{\mathbf{u}} \odot \hat{\mathbf{v}}$$

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta} = \hat{\mathbf{u}} \odot \hat{\mathbf{v}}, \mathbf{u} \in \mathbb{R}^K} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

$$\mathcal{R}_D(\beta) = 2\|\hat{\beta}\|_1$$

Using:
$|\hat{\mathbf{u}}_i|^2 + |\hat{\mathbf{v}}_i|^2 \geq |\hat{\mathbf{u}}_i \hat{\mathbf{v}}_i| = |\hat{\beta}_i|$

Small filter sizes

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}},\mathbf{u}\in\mathbb{R}^K,\mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

But not all $\hat{\mathbf{u}}$ are allowed as $\mathbf{u} \in \mathbb{R}^K$!

- For $K = 1$, $\hat{\mathbf{u}} = \frac{\mathrm{u}_0}{\sqrt{D}}[1,1,1,\ldots]^\mathsf{T}$

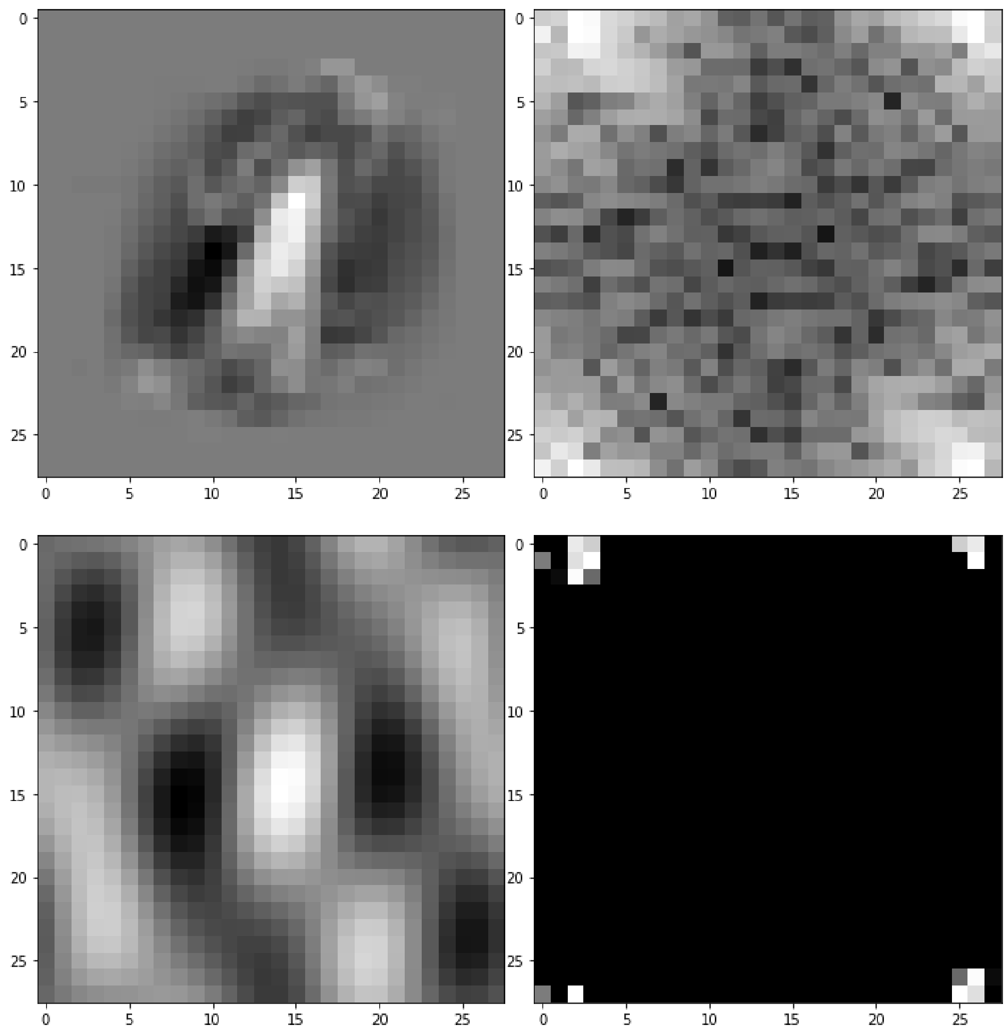
Small filter sizes

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}},\mathbf{u}\in\mathbb{R}^K,\mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

But not all $\hat{\mathbf{u}}$ are allowed as $\mathbf{u} \in \mathbb{R}^K$!

- For $K = 1$, $\hat{\mathbf{u}} = \frac{\mathrm{u}_0}{\sqrt{D}}[1,1,1,\ldots]^\mathsf{T}$

$$\mathcal{R}_1(\beta) = 2\sqrt{D}\|\hat{\beta}\|_2 \quad \text{vs} \quad \mathcal{R}_D(\beta) = 2\|\hat{\beta}\|_1$$



Small filter sizes

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}},\mathbf{u}\in\mathbb{R}^K,\mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

But not all $\hat{\mathbf{u}}$ are allowed as $\mathbf{u} \in \mathbb{R}^K$!

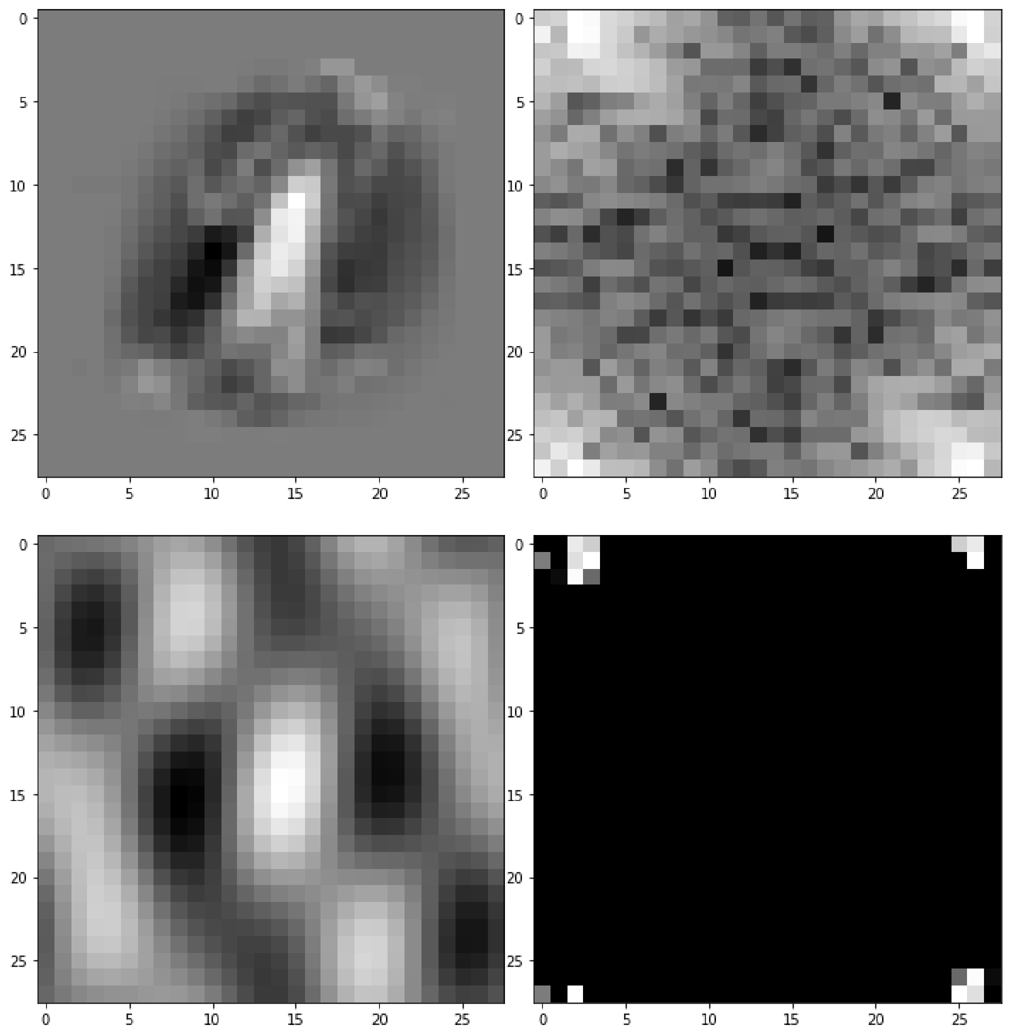
- For $K = 1$, $\hat{\mathbf{u}} = \frac{u_0}{\sqrt{D}}[1,1,1,\dots]^\top$

$$\mathcal{R}_1(\beta) = 2\sqrt{D}\|\hat{\beta}\|_2 \quad \text{vs} \quad \mathcal{R}_D(\beta) = 2\|\hat{\beta}\|_1$$

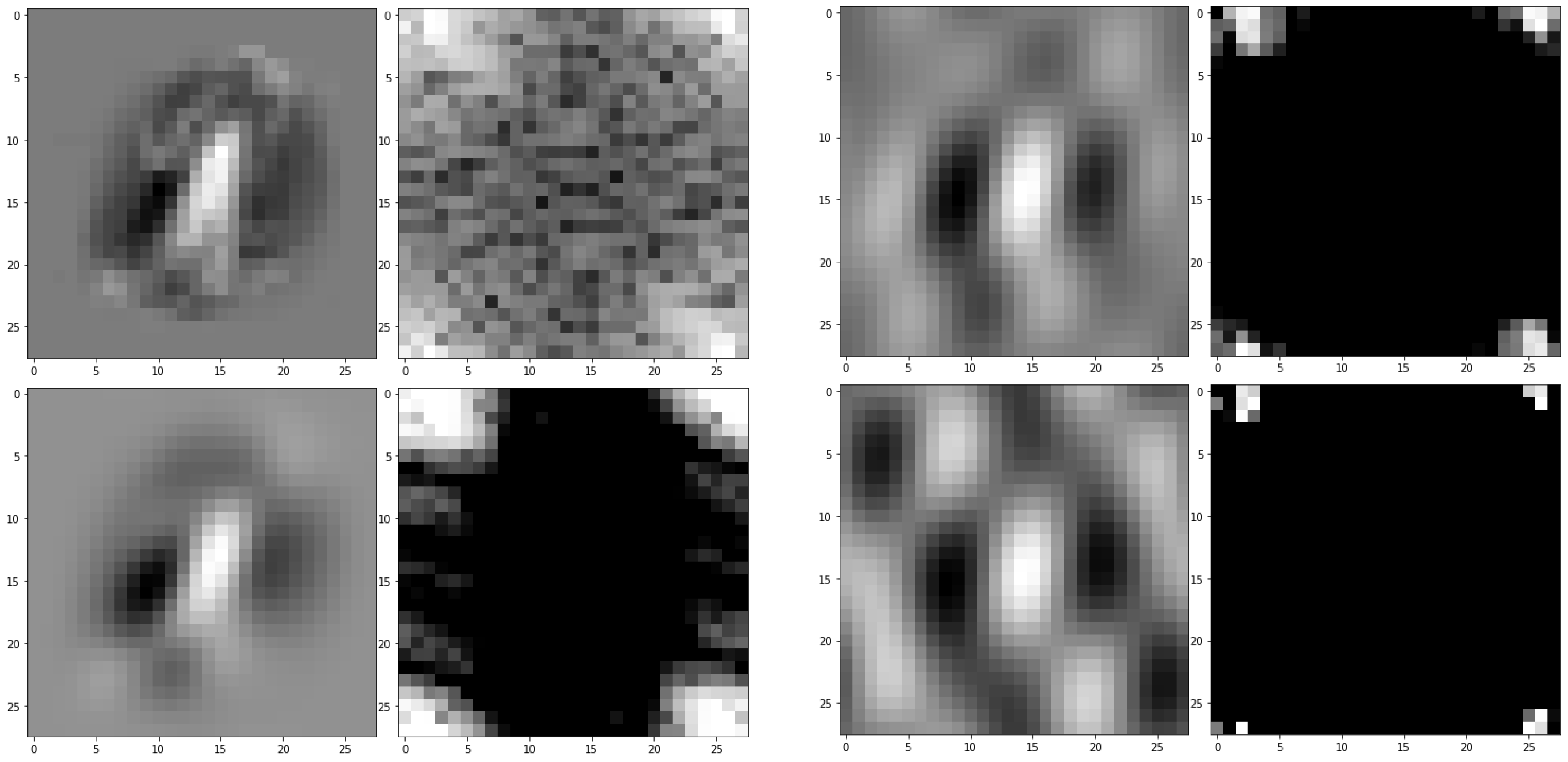
- For $K = 2$,

$$\mathcal{R}_2(\beta) = 2\sqrt{D} \min_{\alpha\in[-1,1]} \sqrt{\sum_{j=0}^{D-1} \frac{|\hat{\beta}_j|^2}{1-\alpha\cos(\frac{2\pi j}{D})}}$$

$$= 2\sqrt{D} \min_{\alpha\in[-1,1]} \sqrt{\sum_{j=0}^{\frac{D}{4}-1} \frac{2|\hat{\beta}_j|^2}{1-\alpha|\cos(\frac{2\pi j}{D})|} + 2|\hat{\beta}_{\frac{D}{4}}|^2 + \sum_{j=\frac{D}{4}+1}^{\frac{D}{2}} \frac{2|\hat{\beta}_j|^2}{1+\alpha|\cos(\frac{2\pi j}{D})|}}$$



MNIST linear model for $K = 1,5,16,28$



SDP relaxation

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}}, \mathbf{u}\in\mathbb{R}^K, \mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

objective $\quad \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = \mathrm{trace}(\mathbf{u}\mathbf{u}^\top) + \mathrm{trace}(\mathbf{v}\mathbf{v}^\top)$

constraints $\quad \hat{\mathbf{u}} \odot \hat{\mathbf{v}} = \hat{\beta} \equiv \mathrm{diag}\left(F_K \mathbf{u}\mathbf{v}^\top F_D^\top\right) = \hat{\beta}$

Define the optimization over

$$W = \begin{bmatrix}\mathbf{u} \\ \mathbf{v}\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top & \mathbf{v}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{u}\mathbf{u}^\top & \mathbf{u}\mathbf{v}^\top \\ \mathbf{v}\mathbf{u}^\top & \mathbf{v}\mathbf{v}^\top\end{bmatrix}$$

with $A_i = \begin{bmatrix}\mathbf{0} & \boldsymbol{F}_K^\top \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_D \\ \boldsymbol{F}_D^{*\top} \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_K^* & \mathbf{0}\end{bmatrix}$

SDP relaxation

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}},\mathbf{u}\in\mathbb{R}^K,\mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

objective $\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = \text{trace}(\mathbf{u}\mathbf{u}^\top) + \text{trace}(\mathbf{v}\mathbf{v}^\top)$

constraints $\hat{\mathbf{u}} \odot \hat{\mathbf{v}} = \hat{\beta} \equiv \text{diag}\left(F_K \mathbf{u}\mathbf{v}^\top F_D^\top\right) = \hat{\beta}$

Define the optimization over

$$W = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top & \mathbf{v}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{u}\mathbf{u}^\top & \mathbf{u}\mathbf{v}^\top\\ \mathbf{v}\mathbf{u}^\top & \mathbf{v}\mathbf{v}^\top\end{bmatrix}$$

with $A_i = \begin{bmatrix}\mathbf{0} & \boldsymbol{F}_K^\top \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_D\\ \boldsymbol{F}_D^{*\top} \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_K^* & \mathbf{0}\end{bmatrix}$

$$\mathcal{R}_K(\beta) = \min_{W\geq 0} \text{trace}(W)$$

$$\text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i$$

$$\text{rank}(W) = 1$$

SDP relaxation

$$\mathcal{R}_K(\beta) = \inf_{\hat{\beta}=\hat{\mathbf{u}}\odot\hat{\mathbf{v}},\mathbf{u}\in\mathbb{R}^K,\mathbf{v}\in\mathbb{R}^D} \|\hat{\mathbf{u}}\|_2^2 + \|\hat{\mathbf{v}}\|_2^2$$

objective $\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = \text{trace}(\mathbf{u}\mathbf{u}^\top) + \text{trace}(\mathbf{v}\mathbf{v}^\top)$

constraints $\hat{\mathbf{u}} \odot \hat{\mathbf{v}} = \hat{\beta} \equiv \text{diag}\left(F_K \mathbf{u}\mathbf{v}^\top F_D^\top\right) = \hat{\beta}$

Define the optimization over

$$W = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top & \mathbf{v}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{u}\mathbf{u}^\top & \mathbf{u}\mathbf{v}^\top \\ \mathbf{v}\mathbf{u}^\top & \mathbf{v}\mathbf{v}^\top\end{bmatrix}$$

with $A_i = \begin{bmatrix} \mathbf{0} & \boldsymbol{F}_K^\top \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_D \\ \boldsymbol{F}_D^{*\top} \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{F}_K^* & \mathbf{0} \end{bmatrix}$

$$\mathcal{R}_K(\beta) = \min_{W\succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i, \quad \text{rank}(W) = 1$$

$$\geq$$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W\succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i$$

Multi-output channel linear ConvNet

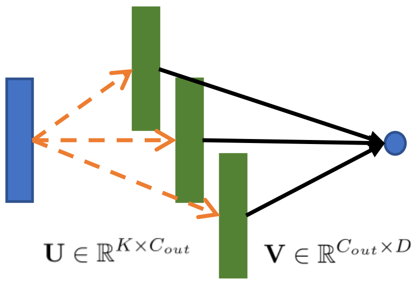
$$\mathbf{U} \in \mathbb{R}^{K \times C_{out}}, \mathbf{V} \in \mathbb{R}^{C_{out} \times D}$$

$$\mathbf{x} \to \mathbf{h}_1[:, c_{out}] = \mathbf{x} \star \mathbf{U}[:, c_{out}]$$

$$\to \sum_{c_{out}} \langle \mathbf{V}[:, c_{out}], \mathbf{h}_1[:, c_{out}] \rangle$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \left[\sum_{c_{out}} \hat{\boldsymbol{U}}[:, c_{out}] \odot \hat{\boldsymbol{V}}[:, c_{out}] \right]$$

$$= \mathrm{diag}\left(\hat{U} \hat{V}^\top \right)$$



$$\mathcal{R}_{K,C_{\text{out}}}(\beta) = \min_{W \geq 0} \mathrm{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i, \quad \mathrm{rank}(W) \leq C_{\text{out}}$$

$$\geq$$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W \geq 0} \mathrm{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i$$

Multi-output channel linear ConvNet

$$\mathcal{R}_{K,C_{\text{out}}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i, \quad \text{rank}(W) \leq C_{\text{out}}$$

$$\geq$$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i$$

Multi-output channel linear ConvNet

Theorem. For any K, C_{out},
$$\mathcal{R}_K^{\text{sdp}}(\beta) = \mathcal{R}_{K,C_{\text{out}}}(\beta)$$

$$\mathcal{R}_{K,C_{\text{out}}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i, \quad \text{rank}(W) \leq C_{\text{out}}$$

$\geq$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i$$

- Induced regularizer is independent of # output channels
- Induced regularizer is a norm interpolating between

$$\mathcal{R}_{1,C_{\text{out}}}(\beta) = 2\sqrt{D}\|\beta\|_2 \quad \text{(basis independent)}, \quad \text{and}$$
$$\mathcal{R}_{D,C_{\text{out}}}(\beta) = 2\|\hat{\beta}\|_1 \quad \text{(sparsity inducing in Fourier space)}$$

Multi-output channel linear ConvNet

Theorem. For any K, C_{out},

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \mathcal{R}_{K,C_{\text{out}}}(\beta)$$

$$\mathcal{R}_{K,C_{\text{out}}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i, \quad \text{rank}(W) \leq C_{\text{out}}$$

$\geq$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W\rangle = \hat{\beta}_i$$

- Induced regularizer is independent of # output channels
- Induced regularizer is a norm interpolating between

$$\mathcal{R}_{1,C_{\text{out}}}(\beta) = 2\sqrt{D}\|\beta\|_2 \quad \text{(basis independent),} \quad \text{and}$$

$$\mathcal{R}_{D,C_{\text{out}}}(\beta) = 2\|\hat{\beta}\|_1 \quad \text{(sparsity inducing in Fourier space)}$$

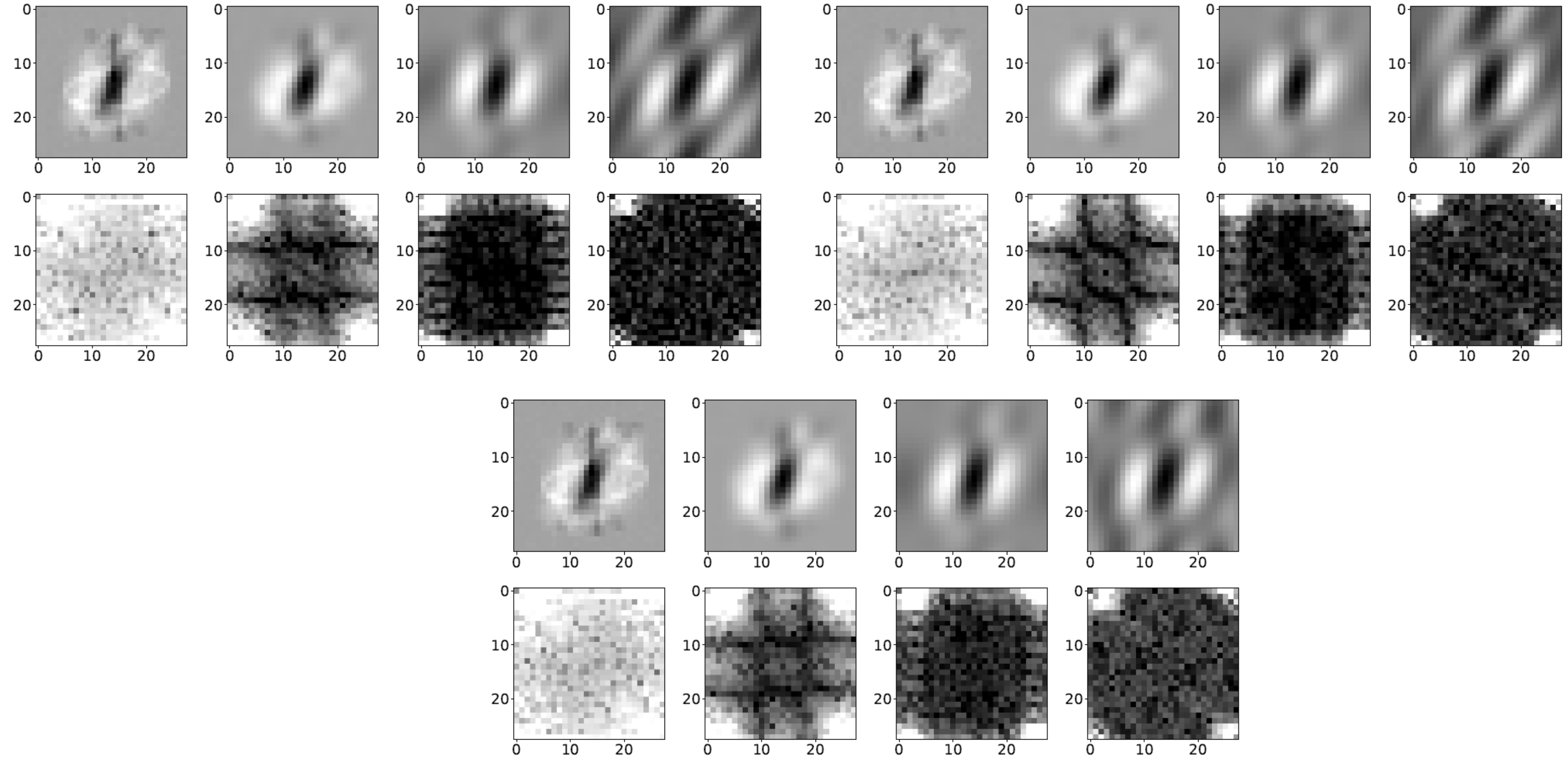
$$2\sqrt{\frac{D}{K}}\|\beta\|_2 \leq \mathcal{R}_{K,C_{\text{out}}}(\beta) \leq 2\sqrt{D}\|\beta\|_2$$

$$2\|\hat{\beta}\|_1 \leq \mathcal{R}_{K,C_{\text{out}}}(\beta) \leq 2\sqrt{\left\lceil\frac{D}{K}\right\rceil}\|\hat{\beta}\|_1$$

Invariance to # output channels

Linear convNets trained with gradient descent on on MNIST

Linear predictors for $C = 1$ (top left), $C = 2$ (top right), $C = 4$ (bottom):



Invariance to # output channels: estimated $\mathcal{R}_{K,C}$

gradient descent on linearly separable MNIST data

Induced regularizer for linear CNNs:

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

Induced regularizer with a ReLU nonlinearity:

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

Multi-output channel linear ConvNet

Theorem. For any K, C_{out},

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \mathcal{R}_{K,C_{\text{out}}}(\beta)$$

$$\mathcal{R}_{K,C_{\text{out}}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i, \quad \text{rank}(W) \leq C_{\text{out}}$$

$$\geq$$

$$\mathcal{R}_K^{\text{sdp}}(\beta) = \min_{W \succeq 0} \text{trace}(W) \quad \text{s.t.,} \quad \langle A_i, W \rangle = \hat{\beta}_i$$

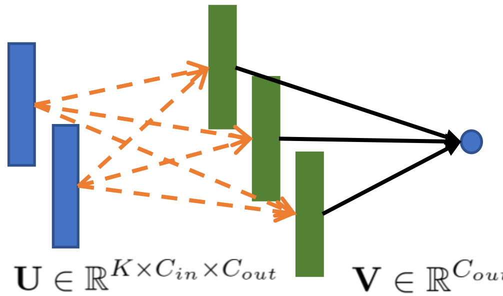
Comments on proof:

- Looking at KKT conditions easy to show that *all solutions of SDP* are of rank $\leq K$
- Showing tightness of C_{out} is trickier: we implicitly show existence of rank-1 optimum
  - Given an SDP solution, we argue about existence a rank-1 solution with same objective value and satisfies constraints – we don't construct this rank-1 solution explicitly

Key lemma: for any $a, b \in \mathbb{R}^K$ there exists $c \in \mathbb{R}^K$ such that $a \star a + b \star b = c \star c$

Multi-input channel linear ConvNet in K=D

$$\mathbf{X} \in \mathbb{R}^{D\times C_{in}}, \mathbf{U} \in \mathbb{R}^{K\times C_{in}\times C_{out}}, \mathbf{V} \in \mathbb{R}^{C_{out}\times D}$$



$$\mathbf{X} \to \mathbf{H}_1[:, c_{out}] = \sum_{c_{in}} \mathbf{X}[:, c_{in}] \star \mathbf{U}[:, c_{in}, c_{out}]$$

$$\to \sum_{c_{out}} \langle \mathbf{V}[:, c_{out}], \mathbf{H}_1[:, c_{out}] \rangle \qquad \Rightarrow \hat{\beta}[:, c_{in}] = \sum_{c_{out}} \hat{\mathbf{U}}[:, c_{in}, c_{out}] \odot \hat{V}[:, c_{out}]$$

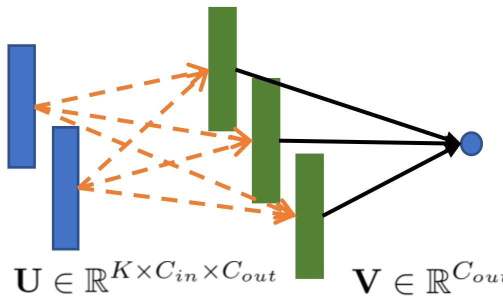
Note: $\hat{\mathbf{V}}$ is shared for all input-channels

Analyses based on slightly different SDP relaxation

- Not always tight – multiple output channels may be required to even realize all linear functions

Multi-input channel linear ConvNet in K=D

$$\mathbf{X} \in \mathbb{R}^{D\times C_{in}}, \mathbf{U} \in \mathbb{R}^{K\times C_{in}\times C_{out}}, \mathbf{V} \in \mathbb{R}^{C_{out}\times D}$$



$$\mathbf{X} \to \mathbf{H}_1[:, c_{out}] = \sum_{c_{in}} \mathbf{X}[:, c_{in}] \star \mathbf{U}[:, c_{in}, c_{out}]$$

$$\to \sum_{c_{out}} \langle \mathbf{V}[:, c_{out}], \mathbf{H}_1[:, c_{out}] \rangle \qquad \Rightarrow \hat{\beta}[:, c_{in}] = \sum_{c_{out}} \hat{\mathbf{U}}[:, c_{in}, c_{out}] \odot \hat{V}[:, c_{out}]$$

Note: $\hat{\mathbf{V}}$ is shared for all input-channels

Analyses based on slightly different SDP relaxation

- Not always tight – multiple output channels may be required to even realize all linear functions
- Tightness can be shown in some cases for large enough C_{out}

for K=1 $\mathcal{R}(\beta) = 2\sqrt{D}\|\beta\|_\star$ (again basis independent)

for K=D $\mathcal{R}(\beta) = 2\|\hat{\beta}\|_{2,1} = 2\sum_{d\in[D]}\sum_{c_{in}} \|\hat{\beta}[:, c_{in}]\|_2$ (group sparsity in Fourier space)

Summary of results on linear convNets

- For single input channels
  - Induced regularizer is independent of # output channels
  - Kernel sizes on the other hand dramatically change the nature of induced biases
    - Small filter sizes ≈ ℓ_2 regularization → noise tolerance?
    - Large filter sizes ≈ ℓ_1 regularization in Fourier domain → invariances?

    (we can quantify “large” and “small” asymptotically)
- For multiple input channel networks
  - Multiple output channels might be necessary to even realize all linear models
  - For large-enough # output channels, the induced regularizer is again unaffected
  - Interesting group structures are observed for linear maps along the multiple-input channels
- Experiments on linear and non-linear networks validate the theoretical findings