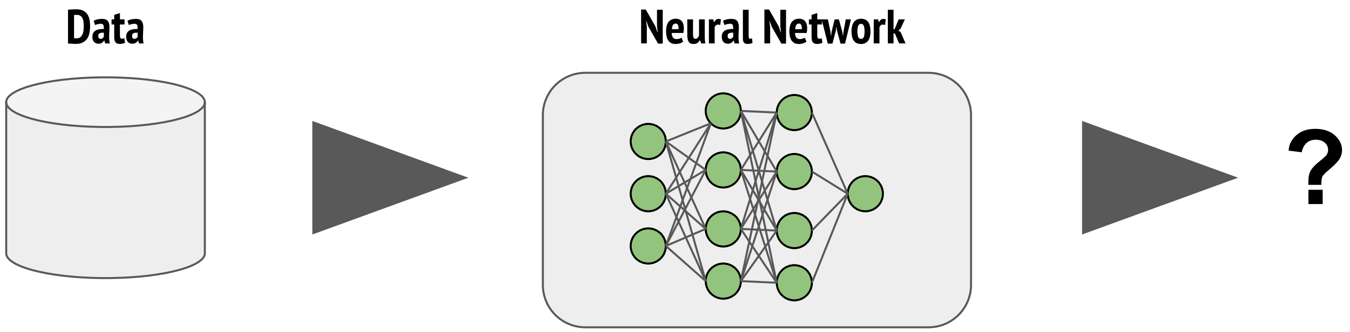
Implicit Bias of Linear Equivariant Networks

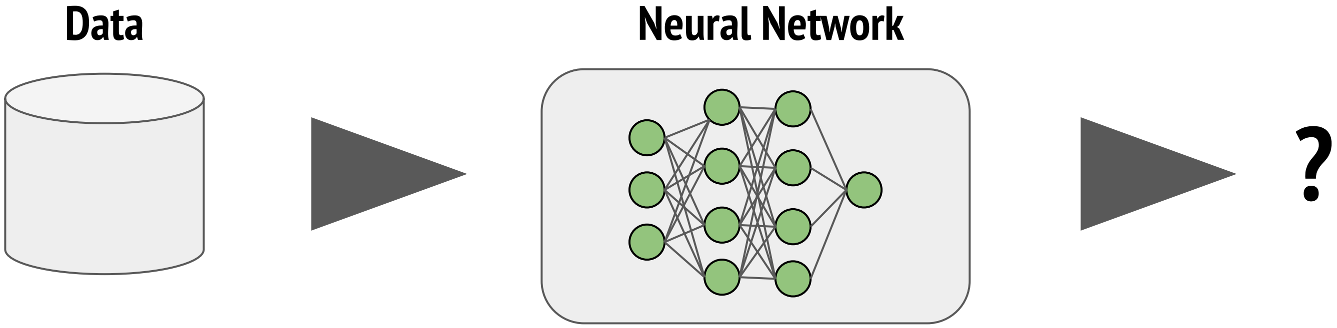



Hannah Lawrence Bobak T. Kiani Kristian Georgiev Andrew Dienes

Background: overparameterized neural networks can perfectly fit the training data in different ways

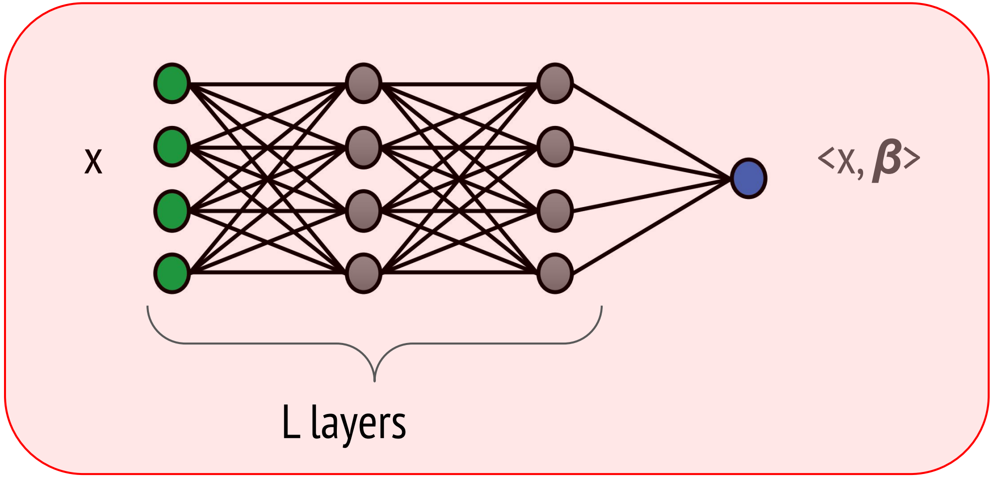


Background: overparameterized neural networks can perfectly fit the training data in different ways



When there are fewer datapoints than parameters, which function does the optimization pick?

What is the **implicit bias** of **linear networks?**

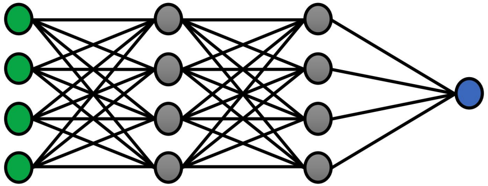


Question:

Consider gradient flow under exponential loss for a binary classification problem. If end-to-end neural network on input x computes <x, $\boldsymbol{\beta}$> fitting the training data, what is this linearization $\boldsymbol{\beta}$?

Source(s): Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ; Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Linear networks' implicit bias depends on **architecture**

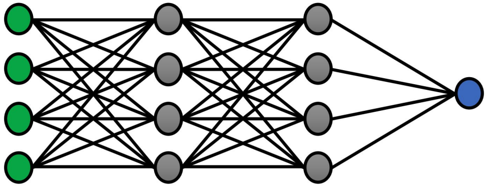


Fully Connected Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\beta\|_2^2 \ \text{ s.t. } \ \forall n, y_n\langle\boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$

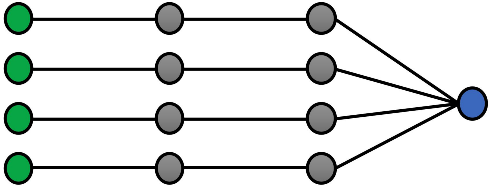
Source(s): Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ; Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Linear networks' implicit bias depends on **architecture**



Fully Connected Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_2^2 \ \text{ s.t. } \ \forall n, y_n\langle\boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$

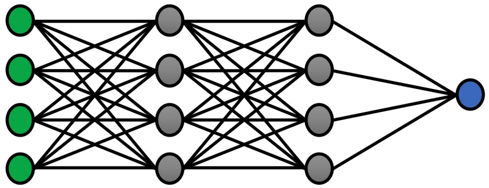


Diagonal Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_{2/L}^{2/L} \ \text{ s.t. } \ \forall n, y_n\langle\boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$

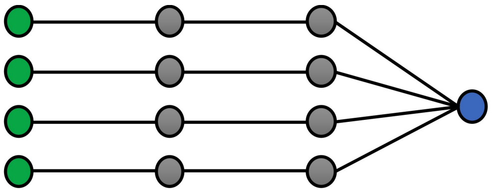
Source(s): Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ;
Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Linear networks' implicit bias depends on **architecture**



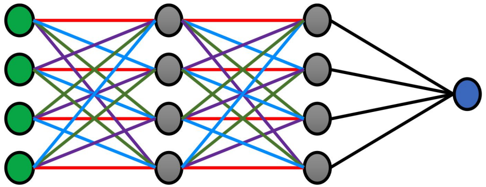
Fully Connected Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_2^2 \;\text{ s.t. }\; \forall n, y_n\langle \boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$



Diagonal Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_{2/L}^{2/L} \;\text{ s.t. }\; \forall n, y_n\langle \boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$

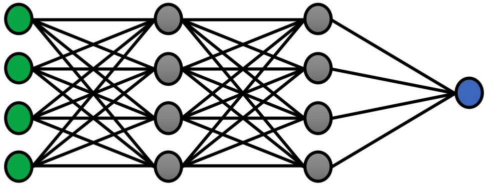


Full-width Convolutional Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\widehat{\boldsymbol{\beta}}\|_{2/L}^{2/L} \;\text{ s.t. }\; \forall n, y_n\langle \boldsymbol{\beta}, \mathbf{x}_n\rangle \geq 1$$

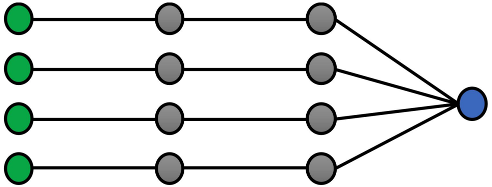
Source(s): Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ;
Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Linear networks' implicit bias depends on **architecture**



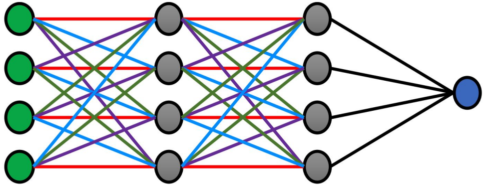
Fully Connected Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}}\|\boldsymbol{\beta}\|_2^2 \;\text{ s.t. }\; \forall n, y_n\langle\boldsymbol{\beta},\mathbf{x}_n\rangle \geq 1$$



Diagonal Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}}\|\boldsymbol{\beta}\|_{2/L}^{2/L} \;\text{ s.t. }\; \forall n, y_n\langle\boldsymbol{\beta},\mathbf{x}_n\rangle \geq 1$$



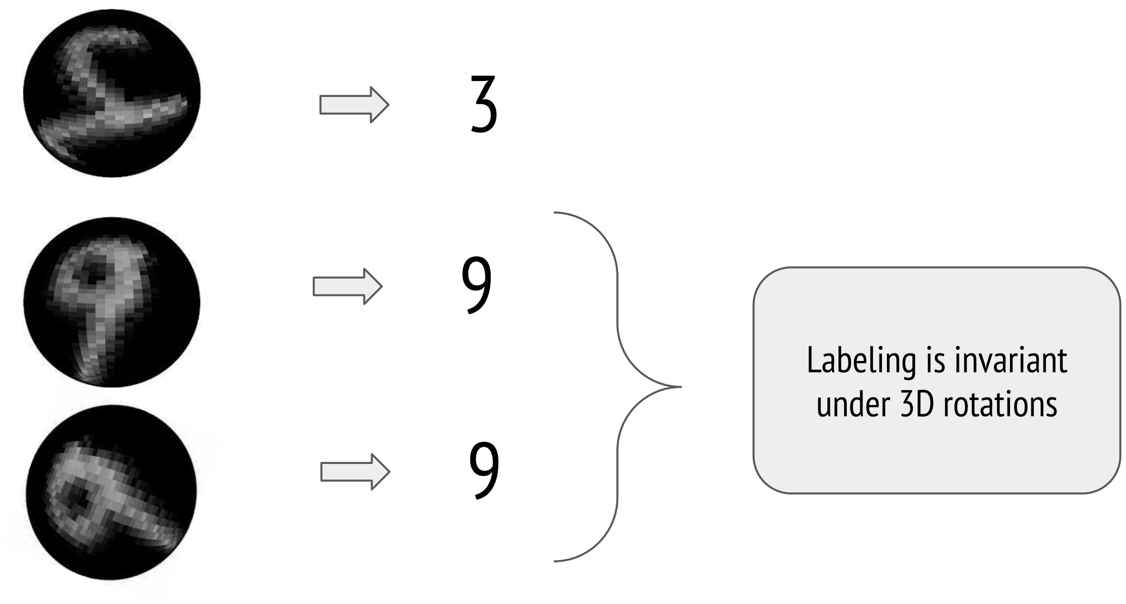
Full-width Convolutional Network:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}}\|\widehat{\boldsymbol{\beta}}\|_{2/L}^{2/L} \;\text{ s.t. }\; \forall n, y_n\langle\boldsymbol{\beta},\mathbf{x}_n\rangle \geq 1$$

Source(s): Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ;
Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Background: group convolutional neural networks

Example: Spherical CNN applied to e.g., “rotated MNIST on the sphere”

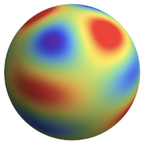


Source(s): Cohen, Taco S., et al. "Spherical cnns." *arXiv preprint arXiv:1801.10130* (2018).
Image source(s): https://openreview.net/pdf?id=Hkbd5xZRb

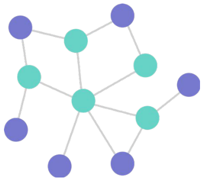
Background: group convolutional neural networks

Main idea: enforce **group** symmetries via architecture of network

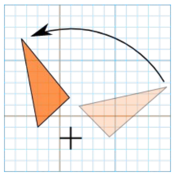
$SO(3)$



S_n



SE_2

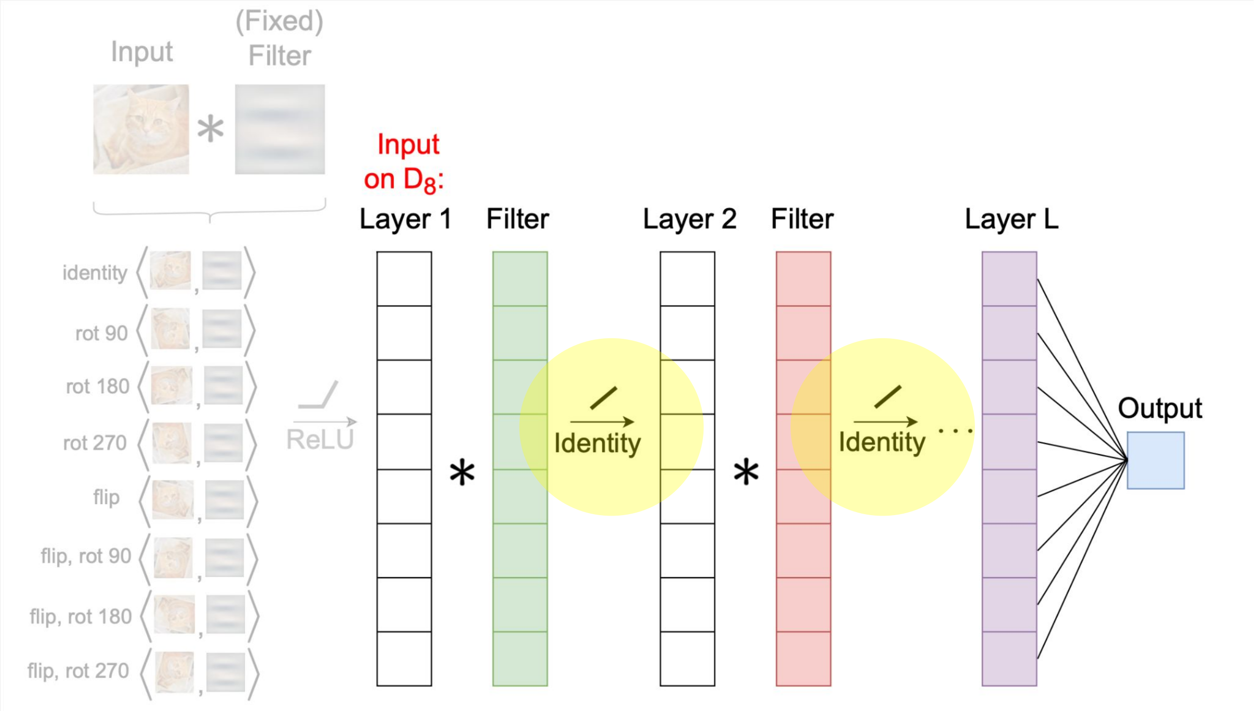


$$(f * g)(u) = \sum_{v \in G} f(uv^{-1})\, g(v)$$

Group convolution equation

Source(s): Cohen, Taco S., et al. "Spherical cnns." *arXiv preprint arXiv:1801.10130* (2018).
Image source(s): https://openreview.net/pdf?id=Hkbd5xZRb

How about **linear** group convolutional neural networks?



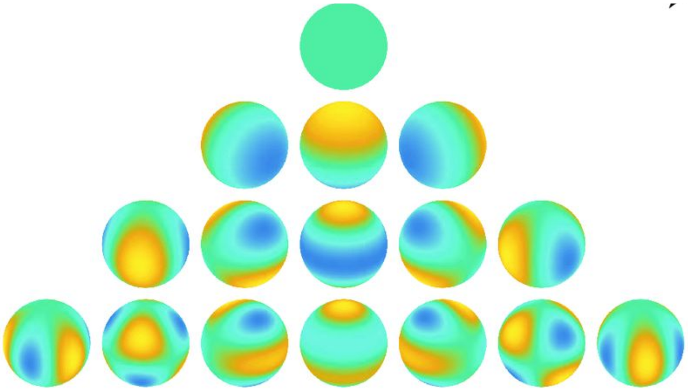
Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

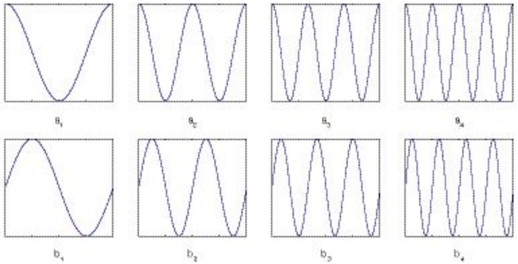
Irreducible representations generalize Fourier space to groups

Irreducible representations (irreps) form the Fourier basis for a function on a group

$$f : G \to \mathbb{C} \qquad \rho : G \to \mathbb{C}^{d \times d}$$

$$\widehat{f}(\rho) = \sum_{u \in G} f(u)\rho(u).$$

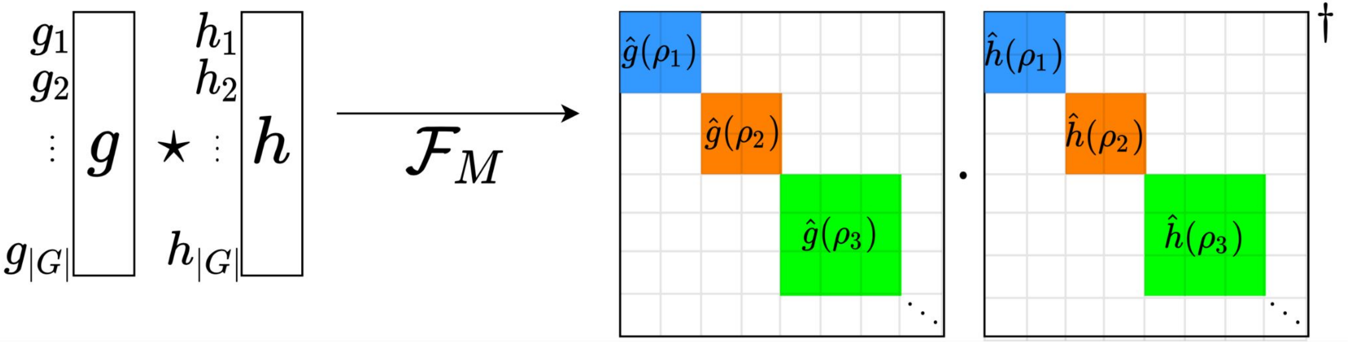




Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Irreducible representations generalize Fourier space to groups

Convolution theorem: Convolution is equivalent to matrix multiplication of irreps

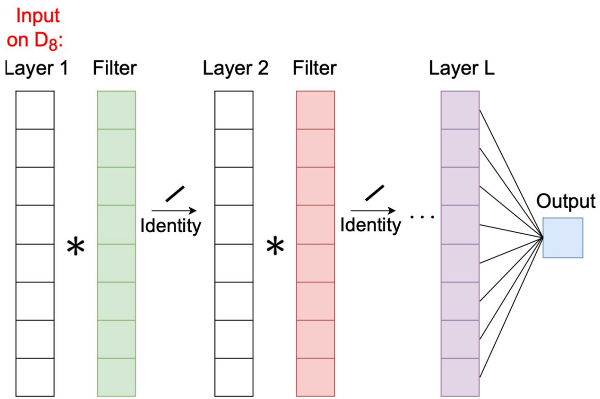


Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

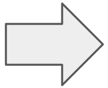
Theorem: G-CNNs are biased towards **low-rank irreps**

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Theorem: G-CNNs are biased towards **low-rank irreps**

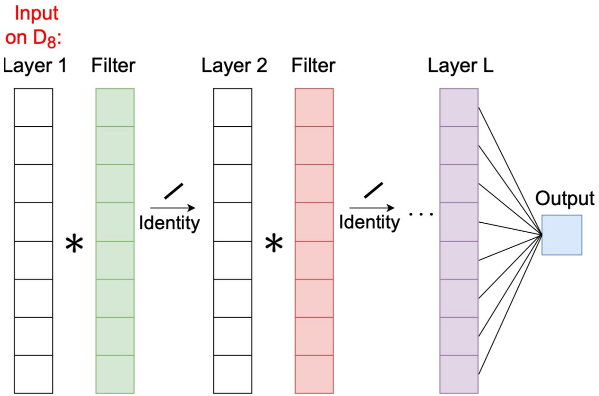


Train: gradient descent
on exponential loss



Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Theorem: G-CNNs are biased towards **low-rank irreps**



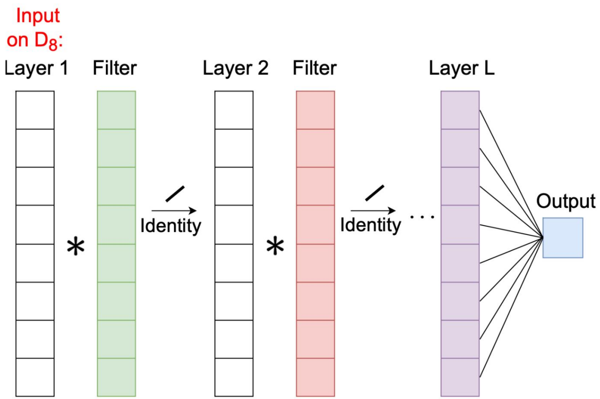
Train: gradient descent on exponential loss

The linearization β is a first-order stationary point of:

$$\min_{\boldsymbol{\beta}} \left\| \widehat{\boldsymbol{\beta}} \right\|_{2/L}^{(S)} \quad s.t. \quad y_i \boldsymbol{x}_i^T \boldsymbol{\beta} \geq 1 \ \ \forall i \in [n]$$

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Theorem: G-CNNs are biased towards **low-rank irreps**



Train: gradient descent on exponential loss

The linearization β is a first-order stationary point of:

$$\min_{\boldsymbol{\beta}} \left\|\widehat{\boldsymbol{\beta}}\right\|_{2/L}^{(S)} \quad s.t. \quad y_i \boldsymbol{x}_i^T \boldsymbol{\beta} \geq 1 \ \ \forall i \in [n]$$

Schatten norm: encourages low rankness of Fourier coefficients

$$\left\|\widehat{\boldsymbol{\beta}}\right\|_{2/L}^{(S)} = \left[\sum_{\rho \in \widehat{G}} d_\rho \left(\left\|\widehat{\boldsymbol{\beta}}(\rho)\right\|_{2/L}^{(S)}\right)^{2/L}\right]^{L/2}$$

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Real vs. Fourier space trade-off: support vs rank

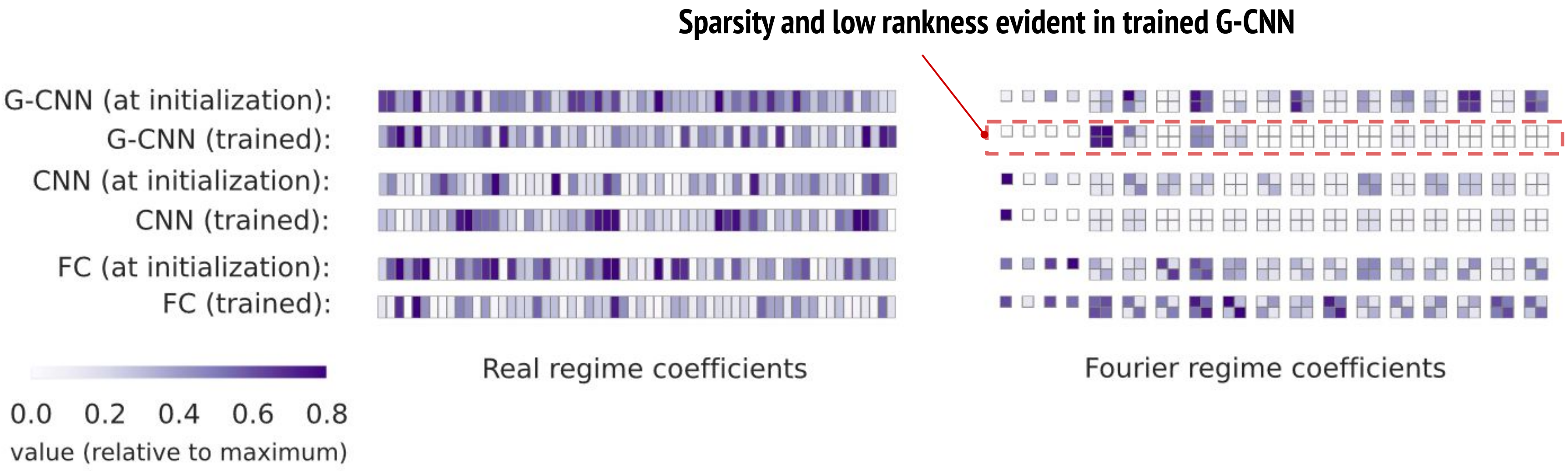
Theorem 6.1 (Meshulam uncertainty theorem (Meshulam, 1992)). *Given a finite group* G *and* $f : G \to \mathbb{C}$*, let* $\widehat{G}$ *be the set of irreps of* G *and* $\boldsymbol{f}$ *be the vectorized function (see Definition 4.1). Then*

$$\underbrace{|\operatorname{supp}(\boldsymbol{f})|}_{\text{real}}\ \underbrace{\operatorname{rank}(\widehat{\boldsymbol{f}})}_{\text{Fourier}} = |\operatorname{supp}(\boldsymbol{f})| \sum_{\rho \in \widehat{G}} d_\rho \operatorname{rank}\left(\widehat{\boldsymbol{f}}(\rho)\right) \geq |G|$$

Intuition: if rank in Fourier space is constant, then support in real space grows with dimension

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.; Wigderson, A., & Wigderson, Y. (2021). The uncertainty principle: variations on a theme. Bulletin of the American Mathematical Society, 58(2), 225-261.

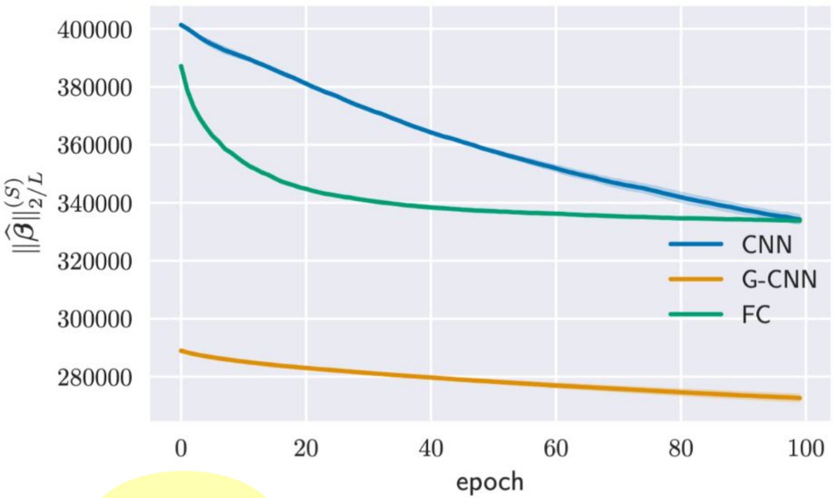
Visualization of implicit bias in real and Fourier space



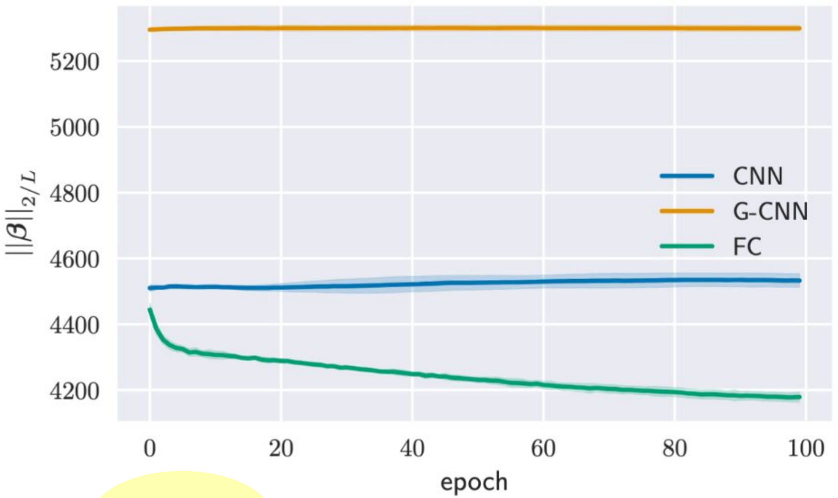
3-layer linear GCNN over the Dihedral group of order 64 trained via SGD

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Real vs. Fourier space trade-offs in implicit bias, empirically



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

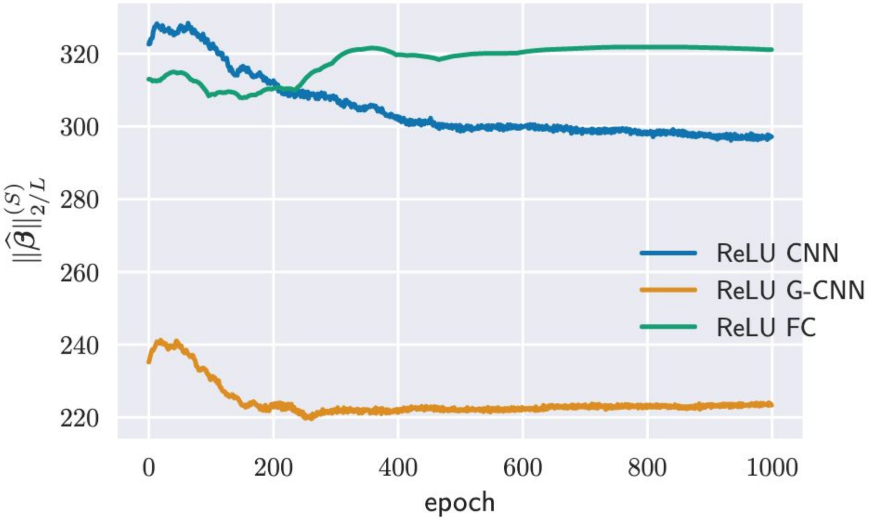


(b) Real space norm of network linearization $\boldsymbol{\beta}$

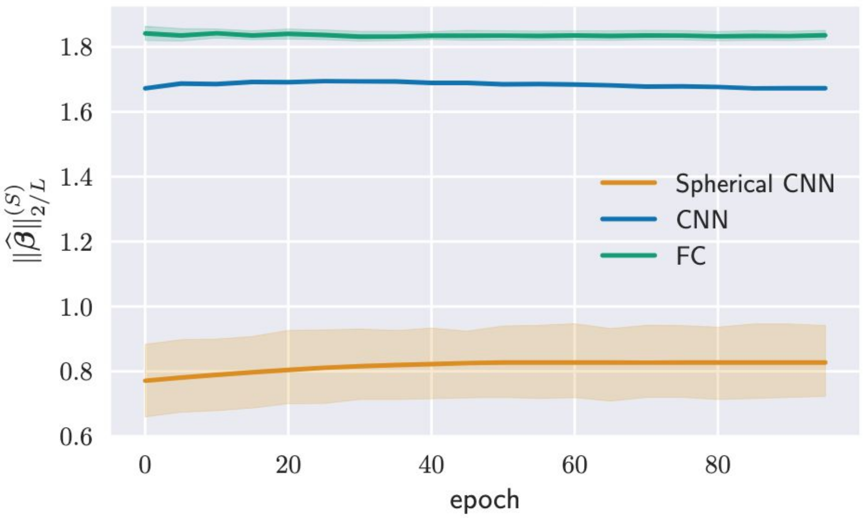
Figure 4: Norms of the linearizations of three different linear architectures for the non-abelian group $G = (C_{28} \times C_{28}) \times D_8$ trained using the digits 1 and 5 from the MNIST dataset.

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Nonlinear networks *may* locally have the same implicit bias



(a) A G-CNN on non-abelian $G = D_{60}$.



(b) A Spherical CNN on bandlimited $G = SO(3)$.

Source(s): Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.

Conclusion

Linear G-CNNs trained by gradient descent are implicitly biased towards **low-rank Fourier coefficients**

- Low-rank structure might be useful for efficient storage
- Implication on generalization performance is problem-dependent
- **Theoretical results on CNNs ⇨ G-CNNs using group Fourier theory**
- Future directions: local analysis of non-linear case, multi-class, etc.

References:

Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.; Wigderson, A., & Wigderson, Y. (2021). The uncertainty principle: variations on a theme. Bulletin of the American Mathematical Society, 58(2), 225-261.
Cohen, Taco S., et al. "Spherical cnns." *arXiv preprint arXiv:1801.10130* (2018).
Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ; Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).

Thank you!

Chat with us more at Poster 520

References:

Lawrence, H., Georgiev, K., Dienes, A., & Kiani, B. T. (2021). Implicit Bias of Linear Equivariant Networks. arXiv preprint arXiv:2110.06084.; Wigderson, A., & Wigderson, Y. (2021). The uncertainty principle: variations on a theme. Bulletin of the American Mathematical Society, 58(2), 225-261.
Cohen, Taco S., et al. "Spherical cnns." *arXiv preprint arXiv:1801.10130* (2018).
Gunasekar, Suriya, et al. "Implicit bias of gradient descent on linear convolutional networks." Advances in Neural Information Processing Systems 31 (2018). ; Yun, Chulhee, Shankar Krishnan, and Hossein Mobahi. "A unifying view on implicit bias in training linear neural networks." arXiv preprint arXiv:2010.02501 (2020).