Understanding MLP-Mixer as a Wide and Sparse MLP

Tomohiro Hayase
Cluster Metaverse Lab

Ryo Karakida
AIST

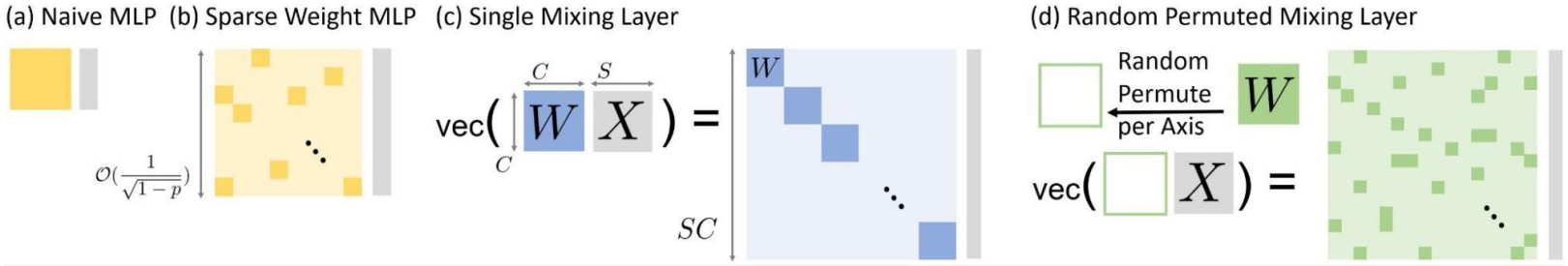
Abstract

In this work, we reveal that sparseness is a key mechanism underlying the MLP-Mixer [I. Tolstikhin, et al., 2021].

Blocks of MLP-Mixer:

$$\text{Token-MLP}(X) = W_2\phi(W_1X), \quad \text{Channel-MLP}(X) = \phi(XW_3)W_4$$

where $W_1 \in \mathbb{R}^{\gamma S \times S}$, $W_2 \in \mathbb{R}^{S \times \gamma S}$, $W_3 \in \mathbb{R}^{C \times \gamma C}$, $W_4 \in \mathbb{R}^{\gamma C \times C}$.



First, the Mixers have an effective expression as a wider MLP with Kronecker-product weights, clarifying that the Mixers efficiently embody several sparseness properties explored in deep learning.

Effective Expression of MLP-Mixer:

Channel-MLP Block:

$$u = \phi(J_c(I_C \otimes W_2)\phi((I_C \otimes W_1)x)),$$

Token-MLP Block:

$$y = \phi(J_c^\top (I_S \otimes W_4^\top)\phi((I_S \otimes W_3^\top)u))$$

In the case of linear layers, the effective expression elucidates an implicit sparse regularization caused by the model architecture,

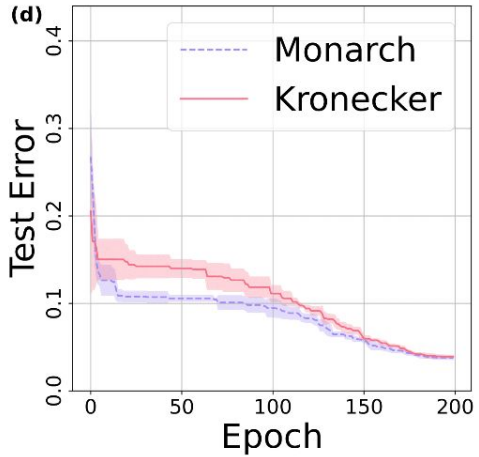
$$\min_{V,W} \mathcal{L}(V \otimes W) + \frac{\lambda}{2}(\|V\|_F^2 + \|W\|_F^2) \qquad \geq \min_{B \in \mathbb{R}^{SC \times SC}} \mathcal{L}(B) + \tilde{\lambda}\|B\|_1,$$

where $\tilde{\lambda} = \lambda/CS$, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_1$ is the L^1 norm.

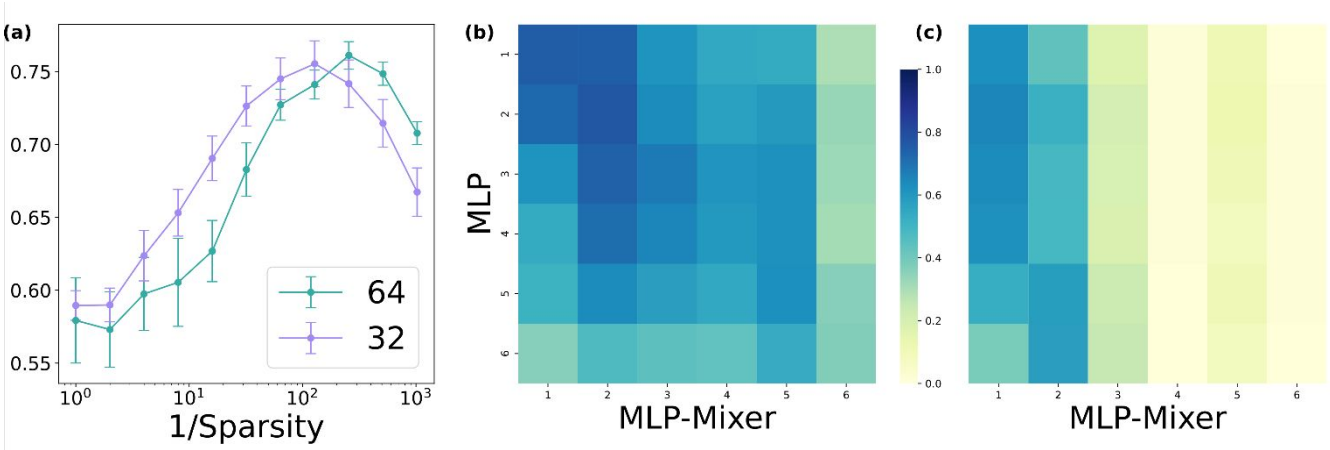
and a hidden relation to Monarch matrices[T. Dao, et al., 2022], which is also known as another form of sparse parameterization.

$$M = J_c^\top L J_c R$$

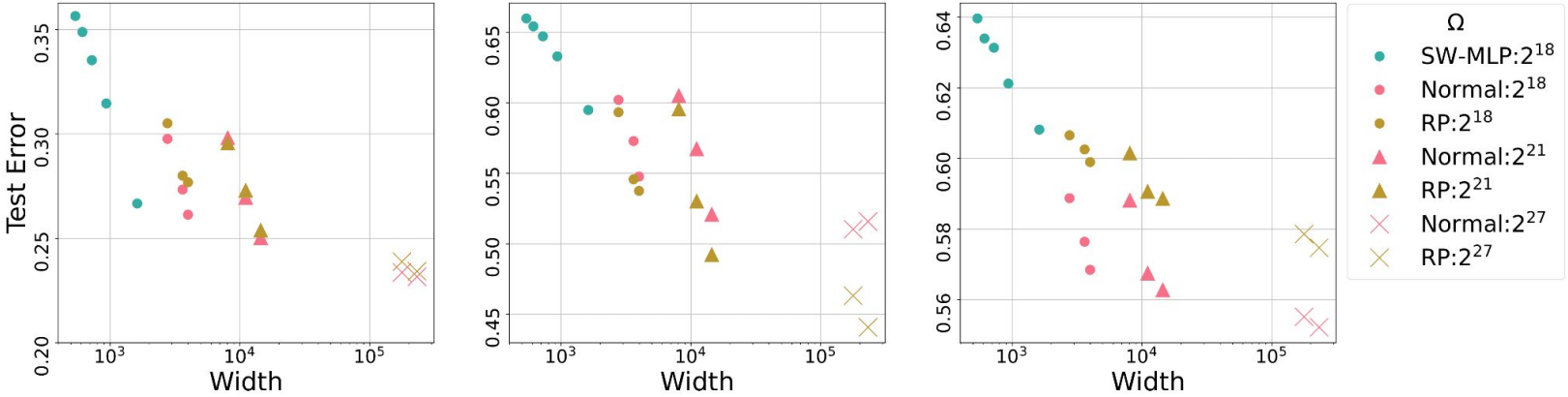
$$L, R = \sqrt{n}$$



For general cases, we empirically demonstrate quantitative similarities between the Mixer and unstructured sparse-weight MLPs, such as the CKA (centered kernel alignment) similarity.



Following a guiding principle proposed by Golubeva, Neyshabur, and Gur-Ari (2021), which fixes the number of connections and increases the width and sparsity, the Mixers can demonstrate improved performance.



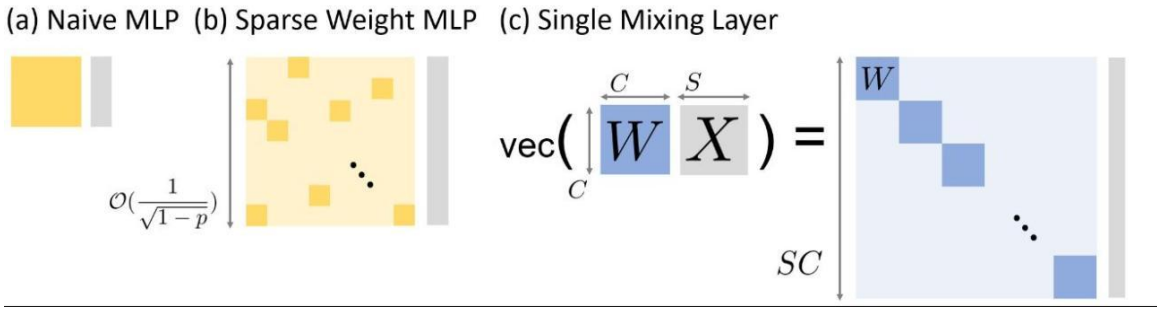
Theory

The blocks of MLP-Mixer are written by right and left multiplication of weights. To know the similarity with sparse weight matrices, we vectorize feature tensors. In that case, right or left multiplication is equal to a block sparse matrix.

Blocks of MLP-Mixer:

$$\text{Token-MLP}(X) = W_2\phi(W_1X), \quad \text{Channel-MLP}(X) = \phi(XW_3)W_4$$

where $W_1 \in \mathbb{R}^{\gamma S \times S}$, $W_2 \in \mathbb{R}^{S \times \gamma S}$, $W_3 \in \mathbb{R}^{C \times \gamma C}$, $W_4 \in \mathbb{R}^{\gamma C \times C}$.



We introduce a commutation matrix, which is a representation of the transpose operator of feature tensors. The commutation matrix also commutes with the activation. Note that the right and left multiplication of a weight matrix is exchanged by the commutation matrix.

A commutation matrix J_C is defined as

$$J_c\mathrm{vec}(X) = \mathrm{vec}(X^\top)$$

where X is an $S \times C$ matrix. Note that for nay entry-wise function ϕ,

$$J_c\phi(x) = \phi(J_c x), x \in \mathbb{R}^m$$

Note that

$$V^\top \otimes I_S = J_c^\top (I_S \otimes V) J_c.$$

Then the mixer layer is a composition of the commutation matrix and the Kronecker product.

Effective Expression of MLP-Mixer:

Channel-MLP Block:

$$u = \phi(J_c(I_C \otimes W_2)\phi((I_C \otimes W_1)x)),$$

Token-MLP Block:

$$y = \phi(J_c^\top (I_S \otimes W_4^\top)\phi((I_S \otimes W_3^\top)u))$$

In the case of linear activation, the mixing layer is equal to having a sparse weight of a Kronecker product of weights. Then, L2 regularization on mixing layers implicitly induces L1 regularization.

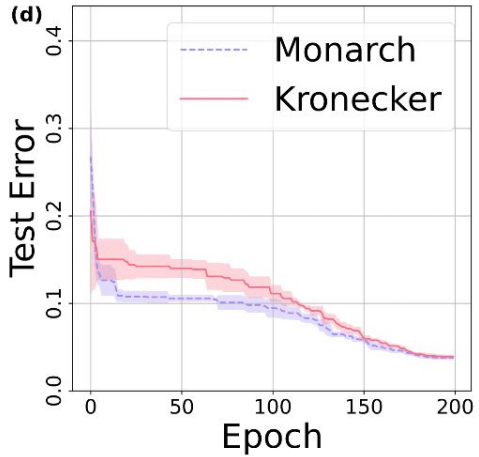
$$\min_{V,W} \mathcal{L}(V \otimes W) + \frac{\lambda}{2}(\|V\|_F^2 + \|W\|_F^2) \qquad \geq \min_{B \in \mathbb{R}^{SC \times SC}} \mathcal{L}(B) + \tilde{\lambda}\|B\|_1,$$

where $\tilde{\lambda} = \lambda / CS$, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_1$ is the L^1 norm.

Moreover, the representation of the mixing layer is equal to the Monarch Matrix except for the weight sharing. In fact, they show similarity in the test error.

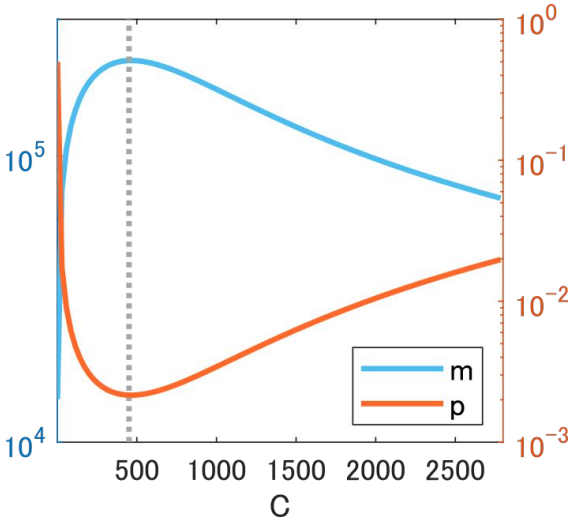
$$M = J_c^\top L J_c R$$

$$L, R = \begin{matrix} \sqrt{n} \\ \updownarrow \end{matrix} \begin{bmatrix} \blacksquare & & & \\ & \blacksquare & & \\ & & \blacksquare & \\ & & & \ddots \\ & & & & \blacksquare \end{bmatrix}$$

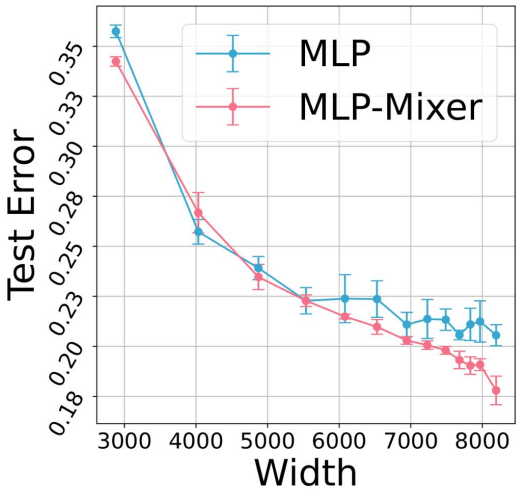


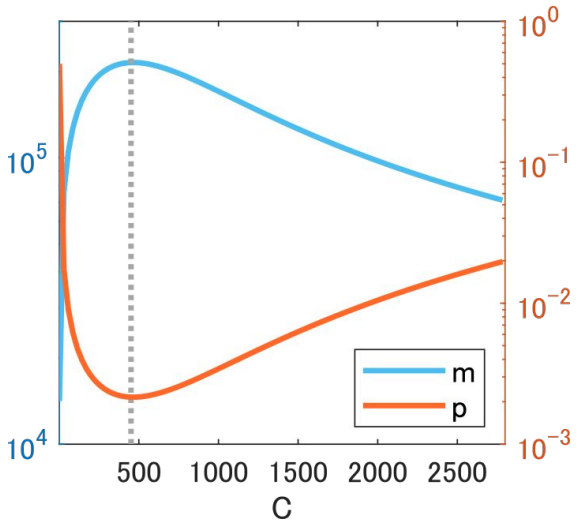
In the general case, there is a guiding hypothesis by Golubeva, Neyshabur, and Gur-Ari (2021): increasing the width up to a certain point, while keeping the number of weight parameters fixed, results in improved test accuracy. In our case, consider the average number of weight connections per layer:

$$\Omega = p\gamma m^2$$

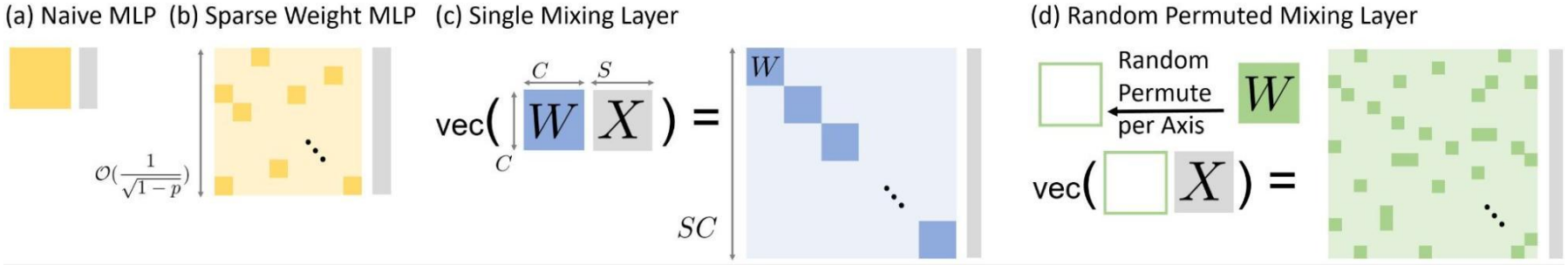


Following the guiding principle, which fixes the number of connections and increases the width and sparsity, in fact, SW-MLP and MLP show a similar tendency in accuracy with respect to increasing sparseness.

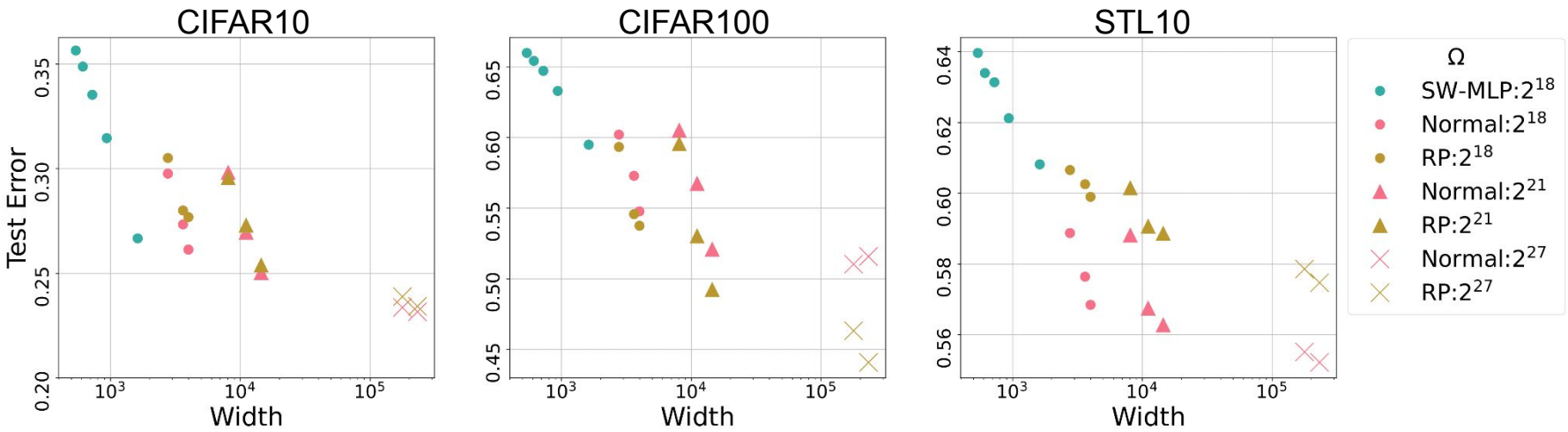




In much wider cases, to continue comparing MLP-Mixer and sparse-weight MLP, we need an alternative to static Mask MLP because of its huge computational costs, memory requirements, and ill behavior on the spectrum. Thus we introduce the alternative to SW-MLP, called the random permuted (RP) Mixer.



Under the averaged connections are fixed, results in improved test accuracy for SW-MLP, MLP-Mixer, and RP-Mixer. They share similar tendency in performance.



Notably, in the case of Mixers, the maximum width is achieved when C=S.

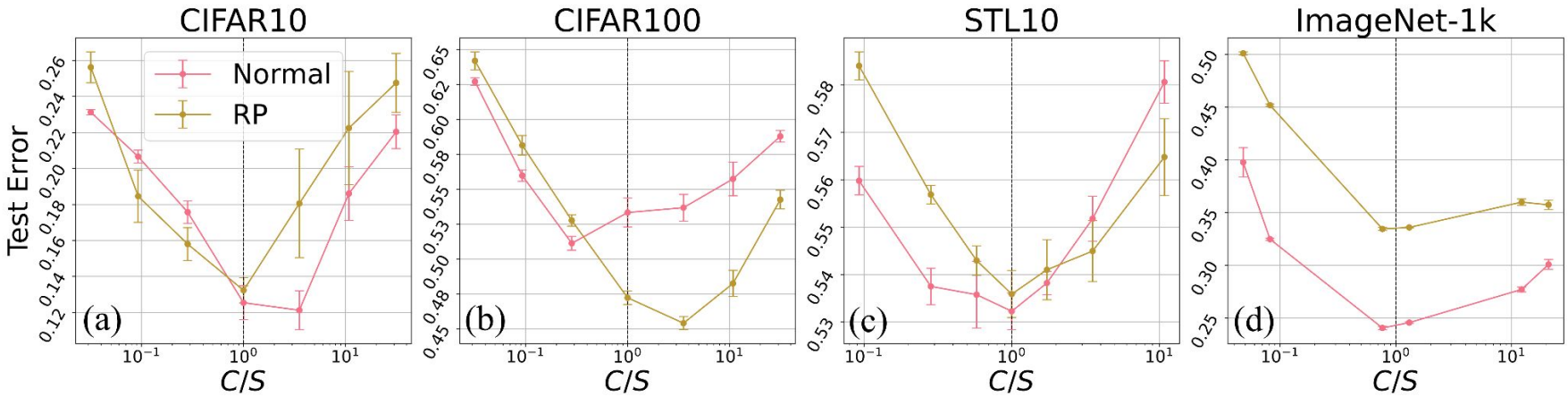
$$\Omega = \frac{\gamma(CS^2 + C^2S)}{2}$$

$$S = \frac{(\sqrt{C^2 + 8\Omega/(\gamma C)} - C)}{2}.$$

$\max_{S,C} m = (\Omega/\gamma)^{2/3}$, the max is achieved when $C = C^*, S = S^*$ with

$$C^* = S^* = (\Omega/\gamma)^{1/3}.$$

In experiments, the test errors are lowest around the points S=C.



This work provides novel insight into how the MLP-Mixer effectively behaves as a wide MLP with sparse weights. The analysis in the linear activation case elucidates the implicit sparse regularization through the Kronecker-product expression and reveals a connection to Monarch matrices.

Conclusion: Novel insight into how the MLP-Mixer effectively behaves as a wide MLP with sparse weights.

1. Linear Case Results
    a. Implicit L1 regularization.
    b. Similarity to Monarch Matrices without weight sharing.

The SW-MLP and Mixers exhibit quantitative similarity in performance trends, verifying that sparsity is the key mechanism underlying the MLP-Mixer.

Conclusion: Novel insight into how the MLP-Mixer effectively behaves as a wide MLP with sparse weights.

1. Linear Case Results
    a. Implicit L1 regularization.
    b. Similarity to Monarch Matrices without weight sharing.
2. General Case Results
    a. Similarity in features (CKA).
    b. Similarity in a performance trend.
    c. A foundation for exploring designs of architectures. (Maximizing the effective width.)

Maximizing the effective width and sparsity leads to improved performance. We expect that this will serve as a foundation for exploring further designs of MLP-based architectures.

Conclusion: Novel insight into how the MLP-Mixer effectively behaves as a wide MLP with sparse weights.

1. Linear Case Results
    a. Implicit L1 regularization.
    b. Similarity to Monarch Matrices without weight sharing.
2. General Case Results
    a. Similarity in features (CKA).
    b. Similarity in a performance trend.
    c. A foundation for exploring designs of architectures. (Maximizing the effective width.)