



Neural Networks can be optimized via a simple and repetitive rule without backpropagation.

What Question Are We Answering? Is the brain performing backpropagation (BP)? This controversial question has inspired the search for an alternative strategy for network optimization. This work proposes **two** back-propagation alternatives to optimize multi-layered perceptrons (MLPs). Both approaches solve the network greedily (layer-wise) with a simple and repetitive solution in each layer. The 1st solution is a closed form solution, and the 2nd one is a spectral method. The two algorithms are the results of our theoretical work in answering the following 4 questions.

1. Can we solve a network greedily (layer-wise), and achieve performance equivalent to traditional networks solved using backpropagation?
2. Can an MLP bypass BP to perfectly classify **any pattern** using only **trivial weights**? (with closed-form solution)
3. By greedily stacking more layers, can an MLP converge to a fixed kernel? What kernel exactly is that?
4. Is the trivial solution layer-wise optimal? How is the optimal solution related to generalization?

Claim 1: If we solve the network greedily (layer-wise) with the HSIC objective at each layer

$$\max_{W_l} \text{Tr} \left(\Gamma \left[\psi(R_{l-1}W_l)\psi^T(R_{l-1}W_l) \right] \right) \\ \text{s. t. } W_l^T W_l = I, \Gamma = HYY^T H. \quad (1)$$

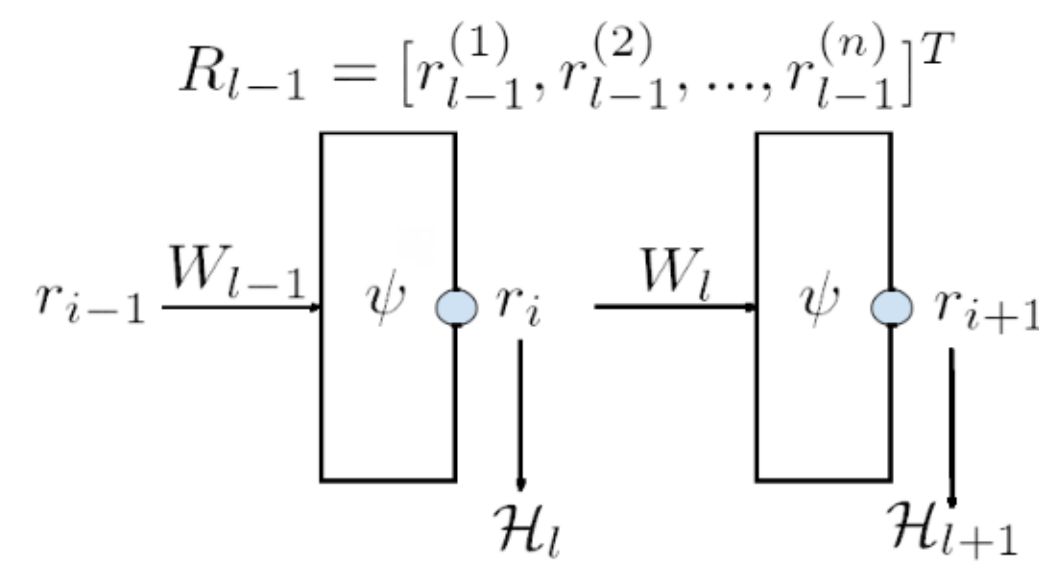


Fig. 1: Layer notation

then, the repetitive usage the **Kernel Embedding** of r_{l-1} for W_l guarantees the **Global Optimum** of Eq. (1). The kernel embedding is defined as

$$W_s = \frac{1}{\sqrt{\zeta}} \left[\sum_l r_l^{(1)} \quad \sum_l r_l^{(2)} \quad \dots \quad \sum_l r_l^{(\tau)} \right] \quad (2)$$

Implication: Instead of searching to connect backpropagation to brain function, our proof suggests that very simple and repetitive patterns can also achieve equivalent training accuracy on any dataset. This strategy might be an easier path to explain the brain.

Claim 2: Let \mathcal{S} be the set of pairwise samples within the same class and \mathcal{S}^c its complement. By greedily stacking the network in Fig.1, the network converges to the feature map of the following kernel:

$$\lim_{l \rightarrow \infty} \mathcal{K}(x_i, x_j)^l = \mathcal{K}^*(x_i, x_j)^l = \begin{cases} 0 & \forall i, j \in \mathcal{S}^c \\ 1 & \forall i, j \in \mathcal{S} \end{cases} \quad (3)$$

Implication: Our construct converges to fixed kernel when solved greedily.

Claim 3: We define the within S_w^l and between S_b^l class scatter matrices as

$$S_w^l = \sum_{i,j \in \mathcal{S}} W_l^T (r_i - r_j)(r_i - r_j)^T W_l \quad \text{and} \quad S_b^l = \sum_{i,j \in \mathcal{S}^c} W_l^T (r_i - r_j)(r_i - r_j)^T W_l. \quad (4)$$

As the number of layers approaches ∞ , the trace ratio approach 0.

$$\lim_{l \rightarrow \infty} \frac{\text{Tr}(S_w^l)}{\text{Tr}(S_b^l)} = 0 \quad (5)$$

Implication: While converging towards a kernel, the network via the HSIC objective pulls samples of the same class into a single point while pushing samples of different classes apart.

Why Maximize HSIC? What does it Accomplish?

Claim 4: We prove that maximizing HSIC Implicitly Minimizes Cross-Entropy (CE) and MSE for Classification.

Implication 1: Classification traditionally use MSE or CE. We refer to these objectives as **label-matching objectives**. We prove the HSIC is a **nonlabel-matching objective**. This implies that using our layer-wise construct for classification is equally flexible as a network trained by backpropagation.

Implication 2: This algorithm learns by identifying the average representation of a class and placing Gaussian Distributions around these average representations. Classification is accomplished via Bayes Optimal classifier in this feature space (RKHS).

Implication 3: Simple and repetitive patterns can also achieve Universality. Perhaps we should seek to explain the brain via this path.

Claim 5: W_s is not the optimal layer-wise solution where at each layer

$$\frac{\partial}{\partial W_l} \mathcal{H}_l(W_s) \neq 0. \quad (6)$$

Implication: If the network converges to the global optimum with W_s , there must exist a W^* that performs even better and converges with fewer layers.

Claim 6: The optimal solution W^* where $\frac{\partial}{\partial W_l} \mathcal{H}_l(W_s) = 0$ is the eigenvector of the following matrix

$$\mathcal{Q}_l = R_{l-1}^T (\hat{\Gamma} - \text{Diag}(\hat{\Gamma}1_n)) R_{l-1}, \quad (7)$$

where $\hat{\Gamma}$ is a function of $W_{\hat{l}}$ computed with $\hat{\Gamma} = \Gamma \odot K_{R_{l-1}W_{\hat{l}}}$.

Implication: While W_s is interesting from a neuroscience perspective, W^* is the optimal solution to optimize the network. The matrix \mathcal{Q} is $d \times d$, therefore, it does not depend on the size of the data.

What can we say about Generalization?

Observations:

1. W^* converges faster, requiring fewer layers.
2. W^* generalizes better?
3. W^* uses an infinitely wide network and have infinite complexity why does it generalize at all?

Claim 7: The solution yield by the eigenvector of \mathcal{Q} implicitly regularizes the HSIC objective.

The objective can be reformulated to isolate out n functions $[D_1(W_l), \dots, D_n(W_l)]$ that act as a penalty term during optimization. Let \mathcal{S}_i be the set of samples that belongs to the i_{th} class and let \mathcal{S}_i^c be its complement, then each function $D_i(W_l)$ is defined as

$$D_i(W_l) = \frac{1}{\sigma^2} \sum_{j \in \mathcal{S}_i} \Gamma_{i,j} \mathcal{K}_{W_l}(r_i, r_j) - \frac{1}{\sigma^2} \sum_{j \in \mathcal{S}_i^c} |\Gamma_{i,j}| \mathcal{K}_{W_l}(r_i, r_j). \quad (8)$$

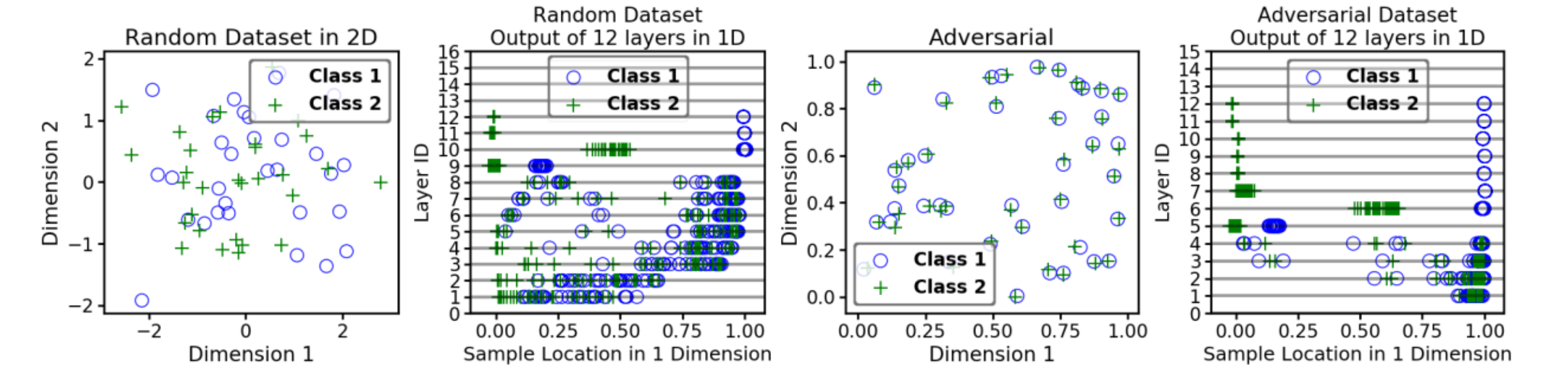
Then Eq. (1) is equivalent to

$$\max_{W_l} \sum_{i,j} \frac{\Gamma_{i,j}}{\sigma^2} e^{-\frac{(r_i - r_j)^T W W^T (r_i - r_j)}{2\sigma^2}} (r_i^T W_l W_l^T r_j) - \sum_i D_i(W_l) \|W_l^T r_i\|_2. \quad (9)$$

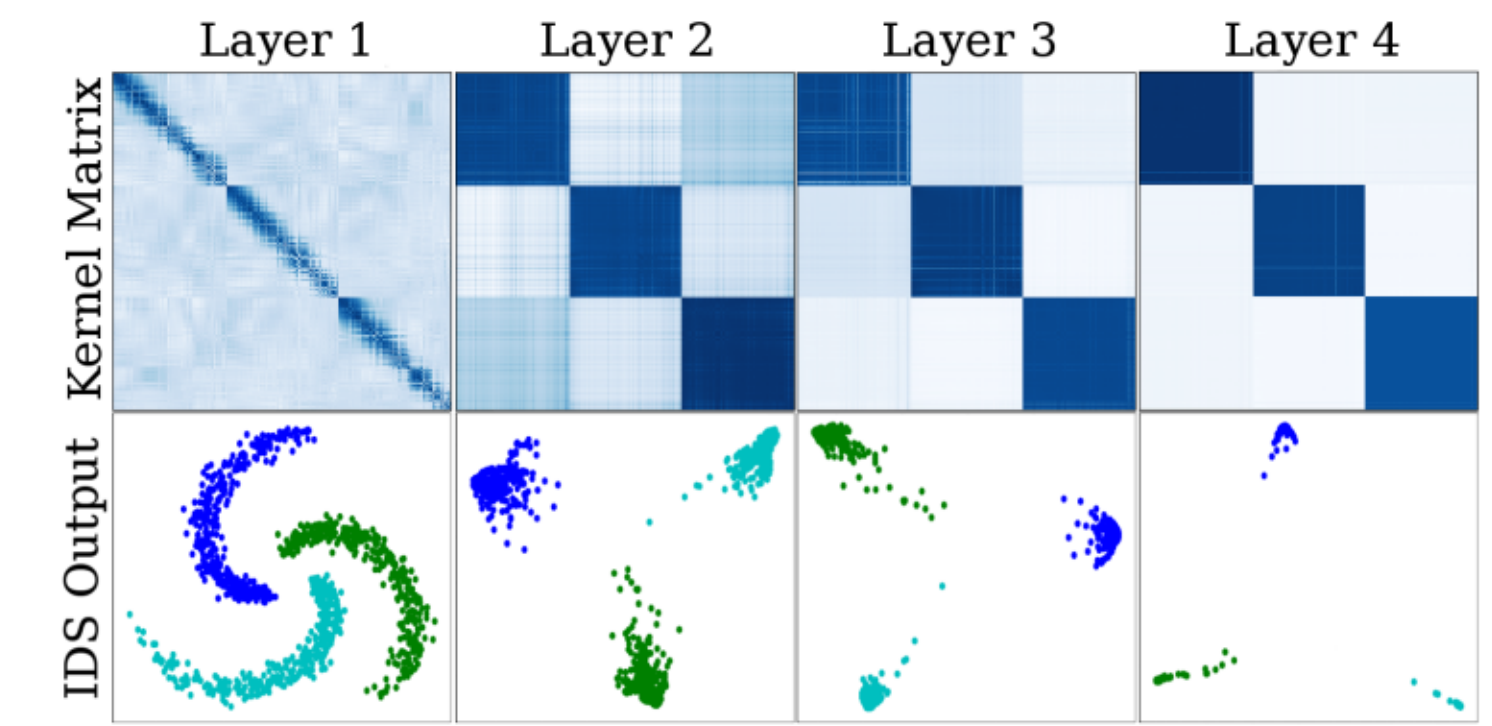
Implication: Based on this claim, $D_i(W_l)$ adds a negative variable cost to the sample norm in IDS, $\|W_l^T r_i\|_2$, describing how ISM implicitly regularizes HSIC. In fact, a better W_l imposes a heavier penalty on the objective where the overall HSIC value may actually decrease.

Experimental Results

Experiment 1: We designed an Adversarial and Random dataset to trick the network using W_s . It was able to achieve perfect classification on the 12th layer.



Experiment 2: Verification of Claims 2 and 3.



Experiment 3: We plot out all key metrics during training at each layer. Here, the objective is clearly monotonic and converging towards a global optimal of 1. Moreover, the trends for \mathcal{T} and \mathcal{C} indicate an incremental clustering of samples into separate partitions. Corresponding to low \mathcal{T} and \mathcal{C} values, the low MSE and CE errors at convergence further reinforces claim 4. This experiment indicates that the behavior of a layer-wise MLP is predictable via the cyclic transition model.

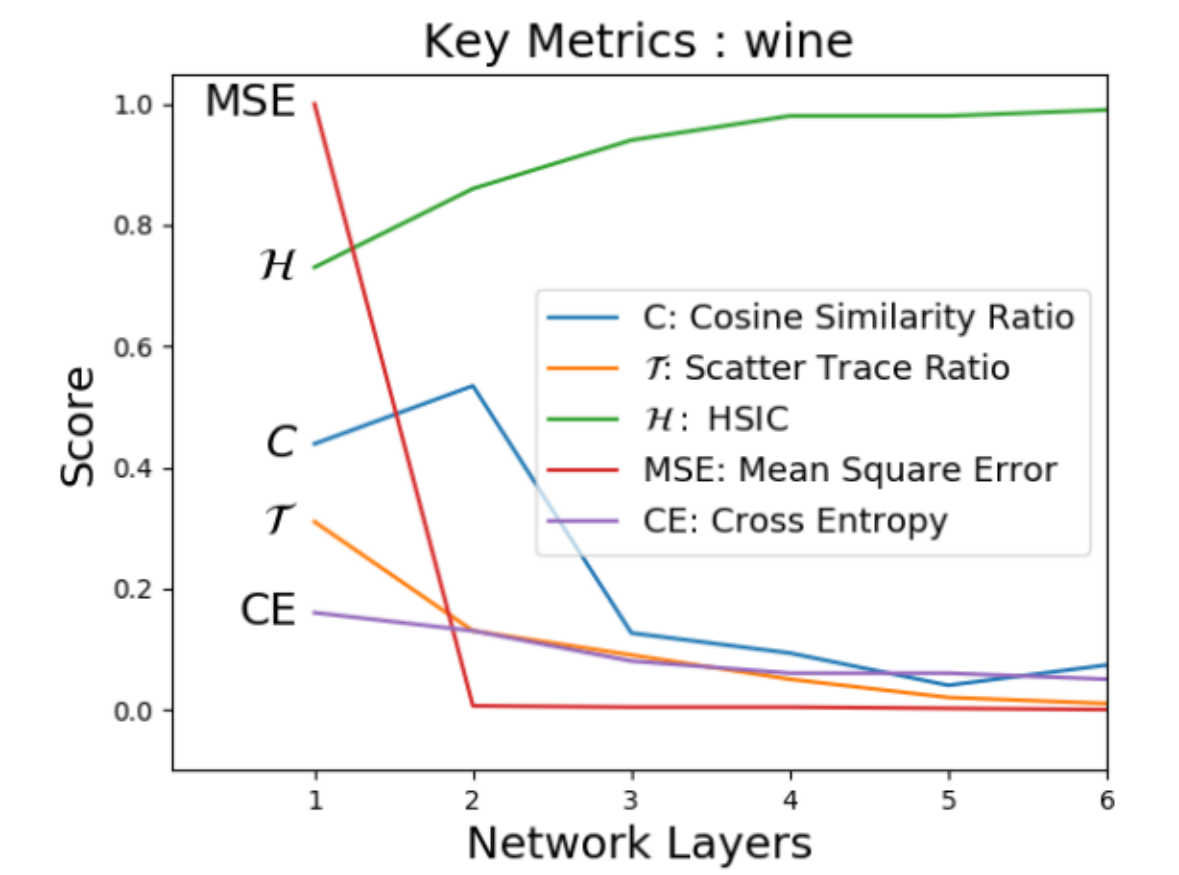


Fig. 4: Layer notation

Experiment 4: Using both W_s and W^* , we conduct 10-fold cross-validation across all 8 datasets and report their mean and the standard deviation for all key metrics. We compare our MLP against MLPs of the same size trained via SGD, where instead of HSIC, MSE and CE are used as the empirical risk.

| | obj | $\sigma \uparrow$ | $L \downarrow$ | Train Acc \uparrow | Test Acc \uparrow | Time(s) \downarrow | $\mathcal{H}^* \uparrow$ | MSE \downarrow | CE \downarrow | $\mathcal{C} \downarrow$ | $\mathcal{T} \downarrow$ |
|---------|-------|-------------------|------------------|----------------------|---------------------|----------------------|--------------------------|-------------------|-----------------|--------------------------|--------------------------|
| random | ISM | 0.38 | 3.30 \pm 0.64 | 1.00 \pm 0.00 | 0.38 \pm 0.21 | 0.40 \pm 0.37 | 1.00 \pm 0.01 | 0.00 \pm 0.01 | 0.05 \pm 0.00 | 0.00 \pm 0.06 | 0.02 \pm 0.0 |
| | W_s | 0.15 | 12 \pm 0.66 | 0.99 \pm 0.01 | 0.45 \pm 0.20 | 0.52 \pm 0.05 | 0.92 \pm 0.01 | 2.37 \pm 1.23 | 0.06 \pm 0.13 | 0.05 \pm 0.02 | 0.13 \pm 0.01 |
| | CE | - | 3.30 \pm 0.64 | 1.00 \pm 0.00 | 0.48 \pm 0.17 | 25.07 \pm 5.55 | 1.00 \pm 0.00 | 10.61 \pm 11.52 | 0.0 \pm 0.0 | 0.0 \pm 0.0 | 0.0 \pm 0.0 |
| | MSE | - | 3.30 \pm 0.64 | 0.98 \pm 0.04 | 0.63 \pm 0.21 | 23.58 \pm 5.38 | 0.93 \pm 0.12 | 0.02 \pm 0.04 | 0.74 \pm 0.03 | 0.04 \pm 0.04 | 0.08 \pm 0.1 |
| adver | ISM | 0.5 | 3.60 \pm 0.92 | 1.00 \pm 0.00 | 0.38 \pm 0.10 | 0.52 \pm 0.51 | 1.00 \pm 0.00 | 0.00 \pm 0.00 | 0.04 \pm 0.00 | 0.01 \pm 0.08 | 0.01 \pm 0.0 |
| | W_s | 0.03 | 12.70 \pm 1.50 | 0.90 \pm 0.04 | 0.42 \pm 0.18 | 2.82 \pm 0.81 | 0.59 \pm 0.19 | 15.02 \pm 11.97 | 0.32 \pm 0.15 | 0.30 \pm 0.18 | 0.34 \pm 0.19 |
| | CE | - | 3.60 \pm 0.92 | 0.59 \pm 0.04 | 0.29 \pm 0.15 | 69.54 \pm 24.14 | 0.10 \pm 0.07 | 0.65 \pm 0.16 | 0.63 \pm 0.04 | 0.98 \pm 0.03 | 0.92 \pm 0.0 |
| | MSE | - | 3.60 \pm 0.92 | 0.56 \pm 0.02 | 0.32 \pm 0.20 | 113.75 \pm 21.71 | 0.02 \pm 0.01 | 0.24 \pm 0.01 | 0.70 \pm 0.00 | 0.99 \pm 0.02 | 0.95 \pm 0.0 |
| spiral | ISM | 0.46 | 5.10 \pm 0.30 | 1.00 \pm 0.00 | 1.00 \pm 0.00 | 0.87 \pm 0.08 | 0.98 \pm 0.01 | 0.01 \pm 0.00 | 0.02 \pm 0.01 | 0.04 \pm 0.03 | 0.02 \pm 0.0 |
| | W_s | 0.93 | 4.00 \pm 1.18 | 0.99 \pm 0.01 | 0.96 \pm 0.02 | 13.54 \pm 5.66 | 0.88 \pm 0.03 | 38.60 \pm 25.24 | 0.06 \pm 0.02 | 0.08 \pm 0.04 | 0.08 \pm 0.0 |
| | CE | - | 5.10 \pm 0.30 | 1.00 \pm 0 | 1.00 \pm 0 | 11.59 \pm 5.52 | 1.00 \pm 0 | 57.08 \pm 31.25 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 5.10 \pm 0.30 | 1.00 \pm 0 | 0.99 \pm 0.01 | 456.77 \pm 78.83 | 1.00 \pm 0 | 0.0 \pm 0 | 1.11 \pm 0.04 | 0.40 \pm 0.01 | 0.0 \pm 0 |
| wine | ISM | 0.47 | 6.10 \pm 0.54 | 0.99 \pm 0 | 0.97 \pm 0.05 | 0.28 \pm 0.04 | 0.98 \pm 0.01 | 0.01 \pm 0 | 0.07 \pm 0.01 | 0.04 \pm 0.03 | 0.02 \pm 0 |
| | W_s | 0.98 | 3.00 \pm 0 | 0.98 \pm 0.01 | 0.92 \pm 0.04 | 0.78 \pm 0.09 | 0.93 \pm 0.01 | 2.47 \pm 0.26 | 0.06 \pm 0.01 | 0.05 \pm 0.01 | 0.08 \pm 0.01 |
| | CE | - | 6.10 \pm 0.54 | 1.00 \pm 0.00 | 0.94 \pm 0.06 | 3.30 \pm 1.24 | 1.00 \pm 0.00 | 40.33 \pm 35.5 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 6.10 \pm 0.54 | 1.00 \pm 0 | 0.89 \pm 0.17 | 77.45 \pm 45.40 | 1.00 \pm 0 | 0.0 \pm 0 | 1.15 \pm 0.07 | 0.49 \pm 0.02 | 0.0 \pm 0 |
| cancer | ISM | 0.39 | 8.10 \pm 0.83 | 0.99 \pm 0 | 0.97 \pm 0.02 | 2.58 \pm 1.07 | 0.96 \pm 0.01 | 0.02 \pm 0.01 | 0.04 \pm 0.01 | 0.02 \pm 0.04 | 0.04 \pm 0.0 |
| | W_s | 2.33 | 1.30 \pm 0.46 | 0.98 \pm 0.01 | 0.96 \pm 0.03 | 6.21 \pm 0.36 | 0.88 \pm 0.01 | 41.31 \pm 56.17 | 0.09 \pm 0.01 | 0.09 \pm 0.02 | 0.16 \pm 0.03 |
| | CE | - | 8.10 \pm 0.83 | 1.00 \pm 0 | 0.97 \pm 0.01 | 82.03 \pm 35.15 | 1.00 \pm 0 | 2330 \pm 2915 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 8.10 \pm 0.83 | 1.00 \pm 0.00 | 0.97 \pm 0.03 | 151.81 \pm 27.27 | 1.00 \pm 0 | 0.0 \pm 0 | 0.66 \pm 0.06 | 0.0 \pm 0 | 0.0 \pm 0 |
| car | ISM | 0.23 | 4.90 \pm 0.30 | 1.00 \pm 0 | 1.00 \pm 0.01 | 1.51 \pm 0.35 | 0.99 \pm 0 | 0.0 \pm 0 | 0.01 \pm 0.00 | 0.04 \pm 0.03 | 0.01 \pm 0 |
| | W_s | 1.56 | 2.70 \pm 0.46 | 1.00 \pm 0 | 1.00 \pm 0 | 5.15 \pm 1.07 | 0.93 \pm 0.02 | 12.89 \pm 2.05 | 0.0 \pm 0 | 0.06 \pm 0.02 | 0.08 \pm 0.02 |
| | CE | - | 4.90 \pm 0.30 | 1.00 \pm 0 | 1.00 \pm 0 | 25.79 \pm 18.86 | 1.00 \pm 0 | 225.11 \pm 253 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 4.90 \pm 0.30 | 1.00 \pm 0 | 1.00 \pm 0 | 504 \pm 116.6 | 1.00 \pm 0 | 0.0 \pm 0 | 1.12 \pm 0.07 | 0.40 \pm 0 | 0.0 \pm 0 |
| face | ISM | 0.44 | 4.00 \pm 0 | 1.00 \pm 0 | 0.99 \pm 0.01 | 0.78 \pm 0.08 | 0.97 \pm 0 | 0.0 \pm 0 | 0.17 \pm 0 | 0.01 \pm 0 | 0.0 \pm 0 |
| | W_s | 0.05 | 3.40 \pm 0.66 | 0.97 \pm 0.01 | 0.80 \pm 0.26 | 11.12 \pm 3.05 | 0.86 \pm 0.04 | 2.07 \pm 1.04 | 0.28 \pm 0.51 | 0.04 \pm 0.01 | 0.01 \pm 0 |
| | CE | - | 4.00 \pm 0 | 1.00 \pm 0 | 0.79 \pm 0.31 | 23.70 \pm 8.85 | 1.00 \pm 0 | 10699 \pm 16330 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 4.00 \pm 0 | 0.92 \pm 0.10 | 0.52 \pm 0.26 | 745.2 \pm 282 | 0.94 \pm 0.07 | 0.11 \pm 0.12 | 3.50 \pm 0.28 | 0.72 \pm 0.01 | 0.0 \pm 0 |
| divorce | ISM | 0.41 | 4.10 \pm 0.54 | 0.99 \pm 0.01 | 0.98 \pm 0.02 | 0.71 \pm 0.41 | 0.99 \pm 0.01 | 0.01 \pm 0.01 | 0.03 \pm 0 | 0.0 \pm 0.05 | 0.02 \pm 0 |
| | W_s | 2.10 | 2.30 \pm 0.64 | 0.99 \pm 0 | 0.95 \pm 0.06 | 1.54 \pm 0.13 | 0.91 \pm 0.01 | 60.17 \pm 70.64 | 0.04 \pm 0.01 | 0.05 \pm 0.01 | 0.08 \pm 0 |
| | CE | - | 4.10 \pm 0.54 | 1.00 \pm 0 | 0.99 \pm 0.02 | 2.62 \pm 1.21 | 1.00 \pm 0 | 14.11 \pm 12.32 | 0.0 \pm 0 | 0.0 \pm 0 | 0.0 \pm 0 |
| | MSE | - | 4.10 \pm 0.54 | 1.00 \pm 0 | 0.97 \pm 0.03 | 47.89 \pm 24.31 | 1.00 \pm 0 | 0.0 \pm 0 | 0.73 \pm 0.07 | 0.0 \pm 0.01 | 0.01 \pm 0 |