# Deep Layer-wise Networks Have Closed-Form Weights

Accepted Paper
AIStats 2022

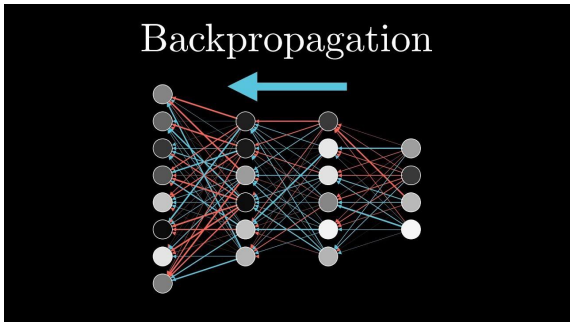
Chieh Wu, Aria Masoomi,
Arthur Gretton, Jennifer Dy

Feb/2/2022

# Neural Networks today are solved using **Backpropagation**

This triggered a debate over if the brain is also doing **Backpropagation?**

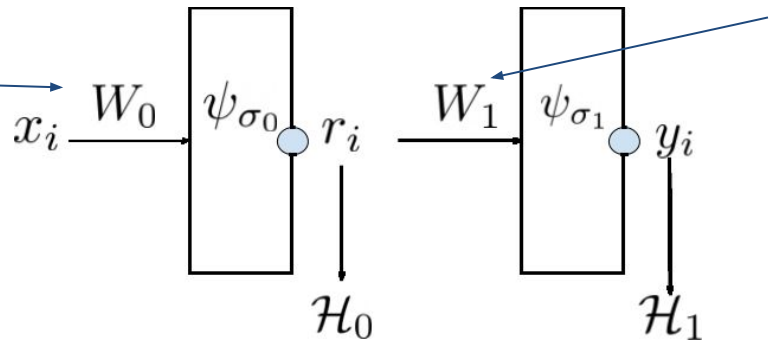
**Some said Yes, others said NO**




This debate has inspired the search for alternative optimization strategies without backpropagation.

# **Layer-wise Training** emerged from this debate to bypass backpropagation.

$W_0$ is first trained as a single layer network against an objective without $W_1$




Once $W_0$ is trained, we hold $W_0$ constant and train $W_1$ as a 2 layer network. This process repeats until L layers.

Since **Layer-wise Training** only uses a forward pass, *backpropagation is avoided* during training. But some theoretical questions remains.

1. Is the computation of gradients during optimization necessary? In fact, is it theoretically possible to use a simple and repetitive closed-form weight to optimize the network.

2. How do we know “when” to stop adding more layers?

# Our Contribution

1. **We discovered a closed-form solution to the weights of layer-wise networks.**
   a. By iteratively stacking layers with simple weights. Layer-wise networks can be automatically optimized to reach the global optimum.
   b. The closed-form weights turn out to be the **Kernel Mean Embedding**. (Computable via addition and without Gradient Descent).

2. **We discovered a strategy to identify the appropriate network depth.**
   a. We propose to stop adding more layers when the extra layers stop changing the network as a function.
   b. We discovered that using Kernel Mean Embedding as the weights, the network automatically converges to a kernel. We call it the **Neural Indicator Kernel**.

# Key Idea of our Approach

Our Empirical Risk

$$\mathcal{H} = \min_{f} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i)$$

We define the **kernel sequence** as

$$(f_1, f_{2\circ}, f_{3\circ}, ..., f_{L\circ})$$

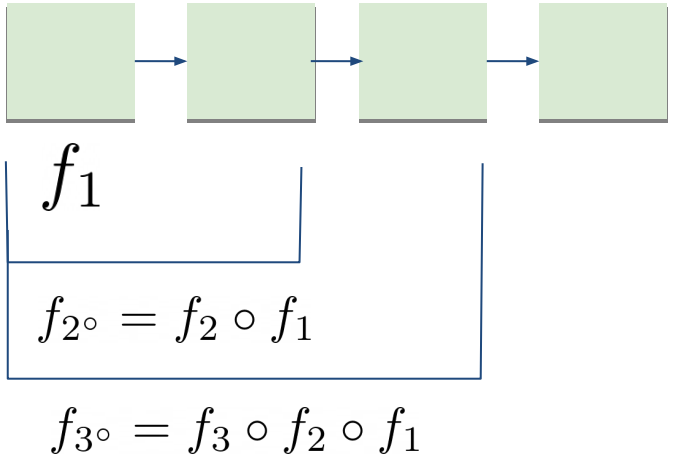
If we input each element of the kernel sequence into the Empircal Risk, we obtain another sequence called th **H-Sequence**.

$$\{\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_L\}$$

Note the notation

$$f_{l\circ} = f_l \circ f_{l-1} \circ ... \circ f_1$$

then




Our strategy is to use closed-form weights to discover a **kernel sequence** that pushes the **H-sequences** to approach its global optimum.

# How did we pick the Empirical Risk?

We noticed that for classification

1. If our solution maps the data to labels {0, 1} instead of the true labels {-1, 1}, the solution is identical.

2. Therefore, enforcing f(x) = y doesn't make sense for classification.

3. We posit that by relaxing the constraint f(x) =y for classification, it would be easier during optimization to collide into a solution space.

4. Instead of MSE or Cross-Entropy, we decide to use an objective that focus on learning a mapping where similar and different classes become easily distinguishable.

But since there are so many ways to define similarity, how can an objective always choose the best similar for every possible situation?

We show that the **Hilbert Schmidt Independence Criterion** (HSIC) is ideal for this situation.

At each layer, we optimize the objective

$$\mathcal{H} = \begin{array}{ll} \max\limits_{W_l} & \mathrm{Tr}\left(\Gamma\left[\psi(R_{l-1}W_l)\psi^T(R_{l-1}W_l)\right]\right) \\ \text{s.t.} & W_l^T W_l = I, \Gamma = HYY^T H. \end{array} \tag{1}$$

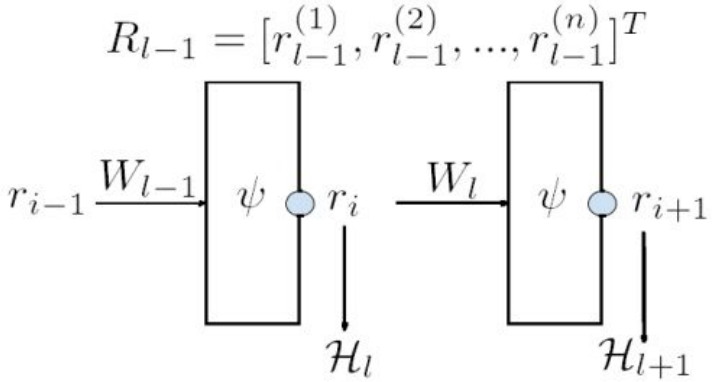



Fig. 1: Layer notation

At each layer, the input $r$ is multiplied by the weight $W$. It then passes through activation function $\psi$. This defines the feature map of a kernel.

We use the feature map of a Gaussian kernel as the activation function.

## HSIC automatically learns the ideal similarity measure

The HSIC objective can be rewritten as

$$\mathcal{H} = \max_{W_l} \sum_{i,j \in \mathcal{S}} \Gamma_{i,j} \mathcal{K}_{W_l}(r_i, r_j) - \sum_{i,j \in \mathcal{S}^c} |\Gamma_{i,j}| \mathcal{K}_{W_l}(r_i, r_j) \quad (4)$$
$$\text{s.t.} \quad W_l^T W_l = I.$$

Since HSIC can be rewritten in terms of similarity Kernel parameterized by the weights, the weights that maximize HSIC automatically discovers the ideal similarity.

## We discovered (as corollaries) that HSIC is simultaneously minimizing MSE and Cross-Entropy.

**Corollary 1.** *$\mathcal{H}_l \to \mathcal{H}^*$ leads to a minimization of MSE in preactivation via a translation of labels.*

**Corollary 2.** *$\mathcal{H}_l \to \mathcal{H}^*$ leads to a minimization of CE in activation via a change of bases.*

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

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

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

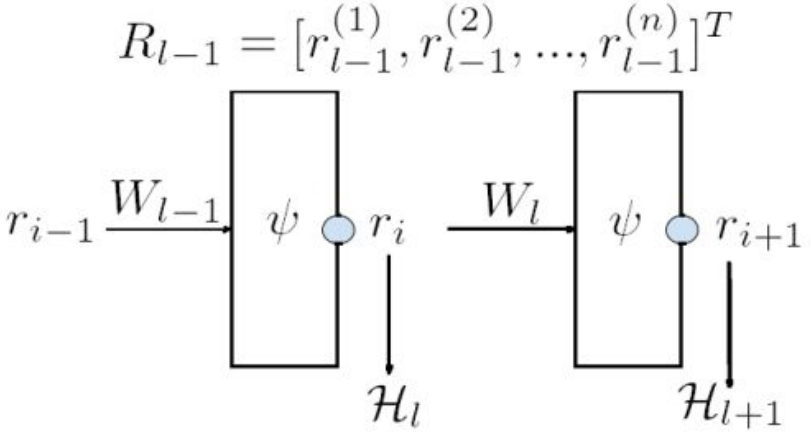



Fig. 1: Layer notation

$$W_s = \frac{1}{\sqrt{\zeta}}\left[\sum_\iota r_\iota^{(1)} \ \sum_\iota r_\iota^{(2)} \ \ldots \ \sum_\iota r_\iota^{(\tau)}\right] \tag{2}$$

**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.

## Our claim is based on our theoretical result of the following theorem.

**Theorem 1.** *From any initial risk value $\mathcal{H}_0$, there exists a set of bandwidths $\sigma_l$ and a Kernel Sequence $\{\phi_{l^\circ}\}_{l=1}^{L}$ parameterized by $W_l = W_s$ in Eq. (5) such that:*

*I. $\mathcal{H}_L$ can approach arbitrarily close to $\mathcal{H}^*$ such that for any $L > 1$ and $\delta > 0$ we can achieve*

$$\mathcal{H}^* - \mathcal{H}_L \leq \delta, \tag{6}$$

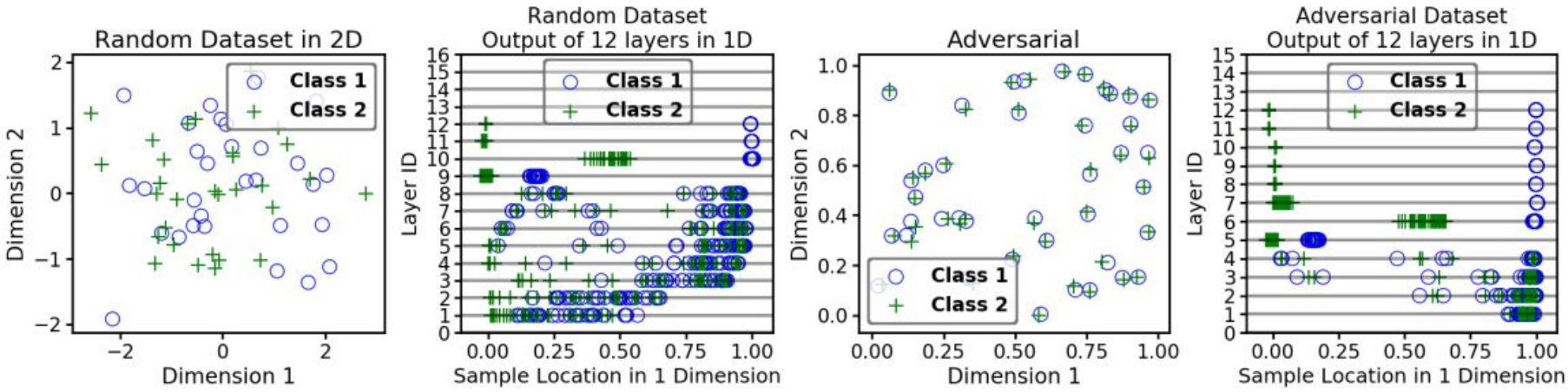
*II. as $L \to \infty$, the $\mathcal{H}$-Sequence converges to the global optimum where*

$$\lim_{L \to \infty} \mathcal{H}_L = \mathcal{H}^*, \tag{7}$$

*III. the convergence is strictly monotonic where*

$$\mathcal{H}_l > \mathcal{H}_{l-1} \quad \forall l \geq 1. \tag{8}$$

## Experiment :



**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:

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

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

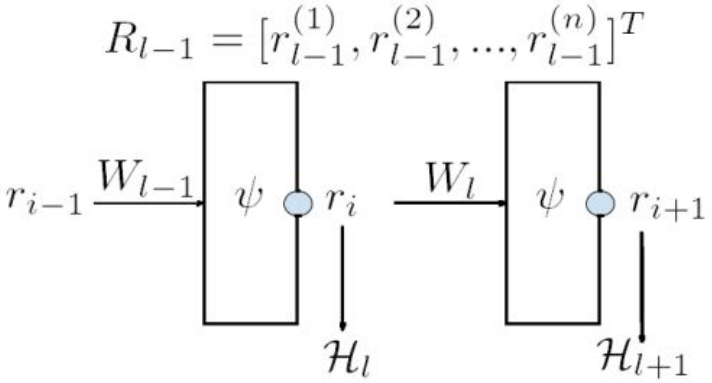



Fig. 1: Layer notation

The output of each layer after the activation function converges to the **Neural Indicator Kernel**.

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

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

As the number of layers approaches $\infty$, the trace ratio approach 0.

$$\lim_{l \to \infty} \frac{\mathrm{Tr}(S_w^l)}{\mathrm{Tr}(S_b^l)} = 0 \tag{5}$$

**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.

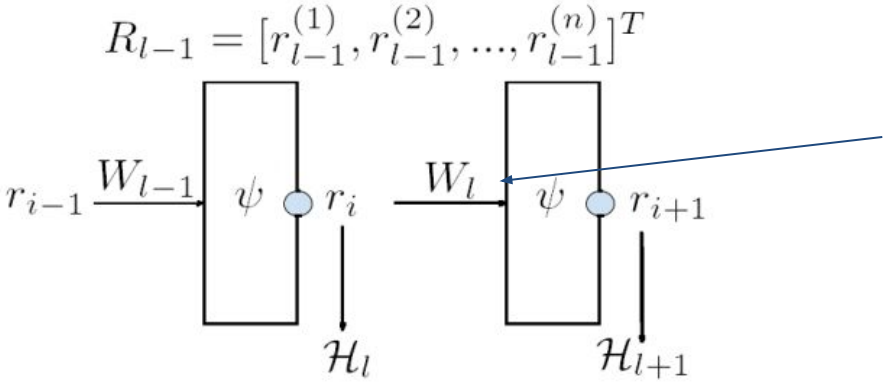



Fig. 1: Layer notation

The output of each layer **before** the activation function converges to a trace ratio of 0.

**Claims 2 and 3 are based on the following Theorem.**

**Theorem 2.** *As $l \to \infty$ and $\mathcal{H}_l \to \mathcal{H}^*$, the following properties are satisfied:*

*I the scatter trace ratio $\mathcal{T}$ approaches 0 where*

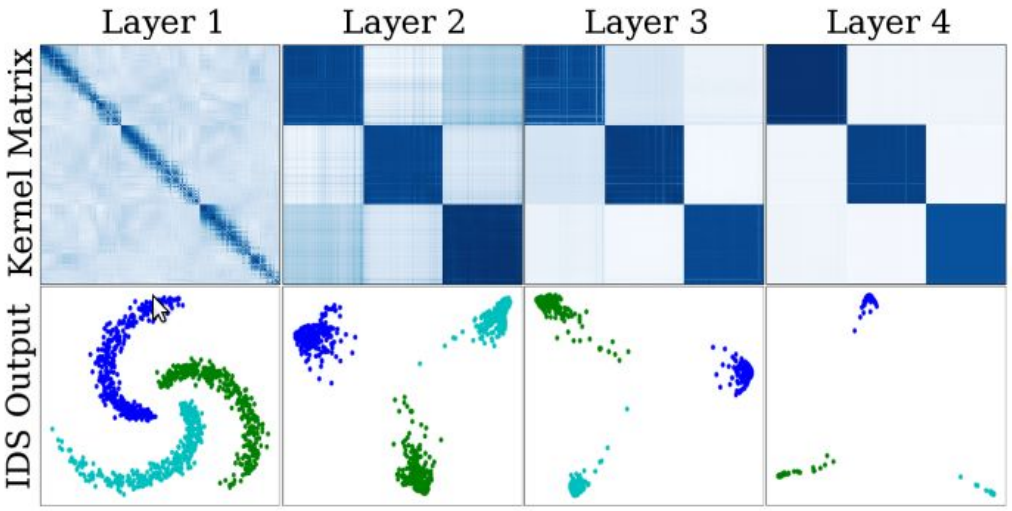
$$\lim_{l\to\infty} \frac{\mathrm{Tr}(S_w^l)}{\mathrm{Tr}(S_b^l)} = 0 \tag{10}$$

This is the **Neural Indicator Kernel**.

*II the Kernel Sequence converges to the following kernel:*

$$\lim_{l\to\infty} \mathcal{K}(x_i, x_j)^l = \mathcal{K}^*(x_i, x_j)^l = \begin{cases} 0 & \forall i, j \in \mathcal{S}^c \\ 1 & \forall i, j \in \mathcal{S} \end{cases} . \tag{11}$$

**Experiment :**



**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.

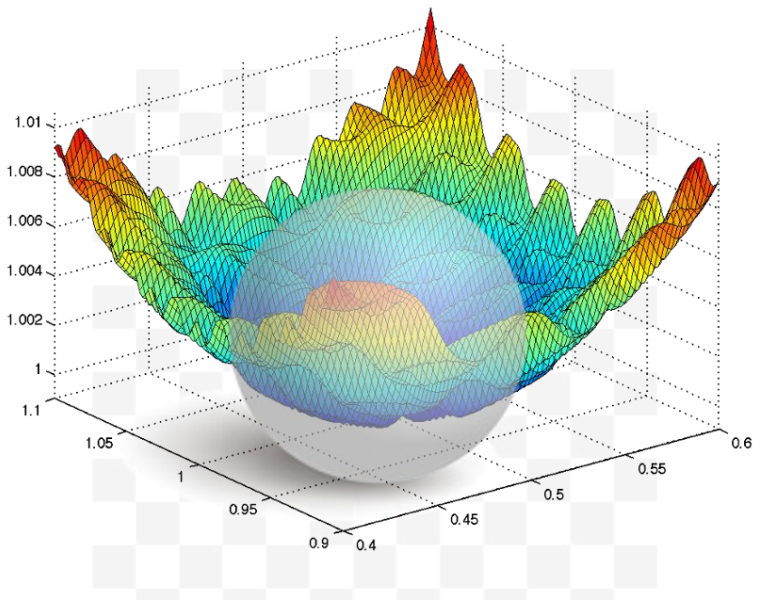
**Corollary 1.** *$\mathcal{H}_l \to \mathcal{H}^*$ leads to a minimization of MSE in preactivation via a translation of labels.*

**Corollary 2.** *$\mathcal{H}_l \to \mathcal{H}^*$ leads to a minimization of CE in activation via a change of bases.*

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

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

**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.




$$\max_{W} \quad \sum_{i,j}^{n} \Gamma_{i,j} e^{-\frac{W^T(x_i-x_j)(x_i-x_j)^T W}{2\sigma^2}}$$

$$\text{s.t} \quad W^T W = I$$

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

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

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

**Implication:** While $W_s$ is interesting from a neuroscience perspective, $W^*$ is the optimal solution to optimize the network. The matrix $\mathcal{Q}$ id $d \times d$, 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 $[D_1(W_l), ..., D_n(W_l)]$ 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

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

Then Eq. (1) is equivalent to

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

**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.

**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 $\downarrow$ | $C \downarrow$ | $\mathcal{T} \downarrow$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| random | ISM | 0.38 | 3.30 ± 0.64 | **1.00 ± 0.00** | 0.38 ± 0.21 | **0.40 ± 0.37** | **1.00 ± 0.01** | **0.00 ± 0.01** | 0.05 ± 0.00 | **0.00 ± 0.06** | 0.02 ± 0.0 |
| | $W_s$ | 0.15 | 12 ± 0.66 | 0.99 ± 0.01 | 0.45 ± 0.20 | 0.52 ± 0.05 | 0.92 ± 0.01 | 2.37 ± 1.23 | 0.06 ± 0.13 | 0.05 ± 0.02 | 0.13 ± 0.01 |
| | CE | - | 3.30 ± 0.64 | **1.00 ± 0.00** | 0.48 ± 0.17 | 25.07 ± 5.55 | **1.00 ± 0.00** | 10.61 ± 11.52 | **0.0 ± 0.0** | **0.0 ± 0.0** | **0.0 ± 0.0** |
| | MSE | - | 3.30 ± 0.64 | 0.98 ± 0.04 | **0.63 ± 0.21** | 23.58 ± 8.38 | 0.93 ± 0.12 | 0.02 ± 0.04 | 0.74 ± 0.03 | 0.04 ± 0.04 | 0.08 ± 0.1 |
| adver | ISM | 0.5 | 3.60 ± 0.92 | **1.00 ± 0.00** | 0.38 ± 0.10 | **0.52 ± 0.51** | **1.00 ± 0.00** | **0.00 ± 0.00** | **0.04 ± 0.00** | **0.01 ± 0.08** | **0.01 ± 0.0** |
| | $W_s$ | 0.03 | 12.70 ± 1.50 | 0.90 ± 0.04 | **0.42 ± 0.18** | 2.82 ± 0.81 | 0.59 ± 0.19 | 15.02 ± 11.97 | 0.32 ± 0.15 | 0.30 ± 0.18 | 0.34 ± 0.19 |
| | CE | - | 3.60 ± 0.92 | 0.59 ± 0.04 | 0.29 ± 0.15 | 69.54 ± 24.14 | 0.10 ± 0.07 | 0.65 ± 0.16 | 0.63 ± 0.04 | 0.98 ± 0.03 | 0.92 ± 0.0 |
| | MSE | - | 3.60 ± 0.92 | 0.56 ± 0.02 | 0.32 ± 0.20 | 113.75 ± 21.71 | 0.02 ± 0.01 | 0.24 ± 0.01 | 0.70 ± 0.00 | 0.99 ± 0.02 | 0.95 ± 0.0 |
| spiral | ISM | 0.46 | 5.10 ± 0.30 | **1.00 ± 0.00** | **1.00 ± 0.00** | **0.87 ± 0.08** | 0.98 ± 0.01 | 0.01 ± 0.00 | 0.02 ± 0.01 | 0.04 ± 0.03 | 0.02 ± 0.0 |
| | $W_s$ | 0.93 | 4.00 ± 1.18 | 0.99 ± 0.01 | 0.96 ± 0.02 | 13.54 ± 5.66 | 0.88 ± 0.03 | 38.60 ± 25.24 | 0.06 ± 0.02 | 0.08 ± 0.04 | 0.08 ± 0 |
| | CE | - | 5.10 ± 0.30 | **1.00 ± 0** | **1.00 ± 0** | 11.59 ± 5.52 | **1.00 ± 0** | 57.08 ± 31.25 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 5.10 ± 0.30 | **1.00 ± 0** | 0.99 ± 0.01 | 456.77 ± 78.83 | **1.00 ± 0** | **0 ± 0** | 1.11 ± 0.04 | 0.40 ± 0.01 | **0 ± 0** |
| wine | ISM | 0.47 | 6.10 ± 0.54 | 0.99 ± 0 | **0.97 ± 0.05** | **0.28 ± 0.04** | 0.98 ± 0.01 | 0.01 ± 0 | 0.07 ± 0.01 | 0.04 ± 0.03 | 0.02 ± 0 |
| | $W_s$ | 0.98 | 3.00 ± 0 | 0.98 ± 0.01 | 0.92 ± 0.04 | 0.78 ± 0.09 | 0.93 ± 0.01 | 2.47 ± 0.26 | 0.06 ± 0.01 | 0.05 ± 0.01 | 0.08 ± 0.01 |
| | CE | - | 6.10 ± 0.54 | **1.00 ± 0.00** | 0.94 ± 0.06 | 3.30 ± 1.24 | **1.00 ± 0.00** | 40.33 ± 35.5 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 6.10 ± 0.54 | **1.00 ± 0** | 0.89 ± 0.17 | 77.45 ± 45.40 | **1.00 ± 0** | **0 ± 0** | 1.15 ± 0.07 | 0.49 ± 0.02 | **0 ± 0** |
| cancer | ISM | 0.39 | 8.10 ± 0.83 | 0.99 ± 0 | **0.97 ± 0.02** | **2.58 ± 1.07** | 0.96 ± 0.01 | 0.02 ± 0.01 | 0.04 ± 0.01 | 0.02 ± 0.04 | 0.04 ± 0.0 |
| | $W_s$ | 2.33 | 1.30 ± 0.46 | 0.98 ± 0.01 | 0.96 ± 0.03 | 6.21 ± 0.36 | 0.88 ± 0.01 | 41.31 ± 56.17 | 0.09 ± 0.01 | 0.09 ± 0.02 | 0.16 ± 0.03 |
| | CE | - | 8.10 ± 0.83 | **1.00 ± 0** | **0.97 ± 0.01** | 82.03 ± 35.15 | **1.00 ± 0** | 2330 ± 2915 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 8.10 ± 0.83 | **1.00 ± 0.00** | **0.97 ± 0.03** | 151.81 ± 27.27 | **1.00 ± 0** | **0 ± 0** | 0.66 ± 0.06 | **0 ± 0** | **0 ± 0** |
| car | ISM | 0.23 | 4.90 ± 0.30 | **1.00 ± 0** | **1.00 ± 0.01** | **1.51 ± 0.35** | 0.99 ± 0 | **0 ± 0** | 0.01 ± 0.00 | 0.04 ± 0.03 | 0.01 ± 0 |
| | $W_s$ | 1.56 | 2.70 ± 0.46 | 1.00 ± 0 | 1.00 ± 0 | 5.15 ± 1.07 | 0.93 ± 0.02 | 12.89 ± 2.05 | 0 ± 0 | 0.06 ± 0.02 | 0.08 ± 0.02 |
| | CE | - | 4.90 ± 0.30 | **1.00 ± 0** | **1.00 ± 0** | 25.79 ± 18.86 | **1.00 ± 0** | 225.11 ± 253 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 4.90 ± 0.30 | **1.00 ± 0** | **1.00 ± 0** | 504 ± 116.6 | **1.00 ± 0** | **0 ± 0** | 1.12 ± 0.07 | 0.40 ± 0 | **0 ± 0** |
| face | ISM | 0.44 | 4.00 ± 0 | **1.00 ± 0** | **0.99 ± 0.01** | **0.78 ± 0.08** | 0.97 ± 0 | **0 ± 0** | 0.17 ± 0 | 0.01 ± 0 | **0 ± 0** |
| | $W_s$ | 0.05 | 3.40 ± 0.66 | 0.97 ± 0.01 | 0.80 ± 0.26 | 11.12 ± 3.05 | 0.86 ± 0.04 | 2.07 ± 1.04 | 0.28 ± 0.51 | 0.04 ± 0.01 | 0.01 ± 0 |
| | CE | - | 4.00 ± 0 | **1.00 ± 0** | 0.79 ± 0.31 | 23.70 ± 8.85 | **1.00 ± 0** | 16099 ± 16330 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 4.00 ± 0 | 0.92 ± 0.10 | 0.52 ± 0.26 | 745.2 ± 282 | 0.94 ± 0.07 | 0.11 ± 0.12 | 3.50 ± 0.28 | 0.72 ± 0.01 | **0 ± 0** |
| divorce | ISM | 0.41 | 4.10 ± 0.54 | 0.99 ± 0.01 | 0.98 ± 0.02 | **0.71 ± 0.41** | 0.99 ± 0.01 | 0.01 ± 0.01 | 0.03 ± 0 | **0 ± 0.05** | 0.02 ± 0 |
| | $W_s$ | 2.10 | 2.30 ± 0.64 | 0.99 ± 0 | 0.95 ± 0.06 | 1.54 ± 0.13 | 0.91 ± 0.01 | 60.17 ± 70.64 | 0.04 ± 0.01 | 0.05 ± 0.01 | 0.08 ± 0 |
| | CE | - | 4.10 ± 0.54 | **1.00 ± 0** | **0.99 ± 0.02** | 2.62 ± 1.21 | **1.00 ± 0** | 14.11 ± 12.32 | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| | MSE | - | 4.10 ± 0.54 | **1.00 ± 0** | 0.97 ± 0.03 | 47.89 ± 24.31 | **1.00 ± 0** | **0 ± 0** | 0.73 ± 0.07 | **0 ± 0.01** | 0.01 ± 0 |

# Authors

Chieh Wu

ch.wu@northeastern.edu
Electrical & Computer Engineering
Dept. Northeastern University

Aria Masoomi

masoomi.a@husky.neu.edu
Electrical & Computer Engineering
Dept. Northeastern University

Jennifer Dy
jdy@ece.neu.edu
Electrical & Computer Engineering
Dept. Northeastern University

Arthur Gretton
arthur.gretton@gmail.com
Gatsby Neural Science Unit
University College of London.

Chieh Wu

http://chiehwu.com
ch.wu@northeastern.edu
Electrical & Computer Engineering Dept
Northeastern University

**Publications:**

- **Wu, Chieh**, Aria Masoomi, Arthur Gretton, Jennifer Dy. "Deep Layer-wise Networks Have Closed-Form Weights."***AIStats 2022***.
- **Wu, Chieh**, Aria Masoomi et al. "Instance-wise Feature Grouping." *Advances in neural information processing systems*. ***NeurIPS 2020***.
- **Wu, Chieh**, Aria Masoomi et al. "Learning via Dependence Network." *Advances in neural information processing systems*. ***NeurIPS Workshop 2020*** (Beyond Backpropagation).
- **Wu, Chieh**, Aria Masoomi et al. "Layer-wise Network Training via ISM." *Advances in neural information processing systems*. ***NeurIPS Workshop 2020*** (Beyond Backpropagation).
- **Wu, Chieh**, et al. "Solving Interpretable Kernel Dimension Reduction." *Advances in neural information processing systems*. ***NeurIPS 2019***.
- **Wu, Chieh, et al.** "Deep kernel Learning for Clustering." *SIAM International Conference on Data Mining*. ***SDM 2019***.
- **Wu, Chieh**, et al. "Iterative Spectral Method for Alternative Clustering." *International Conference on Artificial Intelligence and Statistics*. ***AIStats 2018***.
- **Wu, Chieh**, et al. "Spectral Non-Convex Optimization for Dimension Reduction with Hilbert-Schmidt Independence Criterion." arXiv preprint https://arxiv.org/abs/1909.05097, 2019.
- **Wu, Chieh**, et al. "Layer-wise Learning of Kernel Dependence Networks." arXiv preprint arXiv:2006.08539 (2020).
- Shi Dong, **Chieh Wu** et al. "Using Undersampling with Ensemble Learning to Identify Factors Contributing to Preterm Birth" ***ICMLA 2020***.
- Li, Xiangyu, **Chieh Wu**, Shi Dong, Jennifer Dy, and David Kaeli. "Interactive Kernel Dimension Alternative Clustering on GPUs." In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 885-892. ***IEEE, 2018***.
- Dong, Shi, **Chieh Wu**, et al. "A Hybrid Approach to Identifying Key Factors in Environmental Health Studies." In 2018 IEEE International Conference on Big Data (Big Data), pp. 2855-2862. ***IEEE, 2018***.