Watch Our Video Presentation
https://youtu.be/usvePLCxEjs

For the 1st portion of the poster, visit poster 27.

Layer-wise Learning via Kernel Embedding

Aria Masoomi[1], Chieh Wu[1], Arthur Gretton[2], Jennifer Dy[1]

[1]Northeastern University / [2]University College London

Northeastern

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 formed 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. an we solve a network greedily (layer-wised), 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

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

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

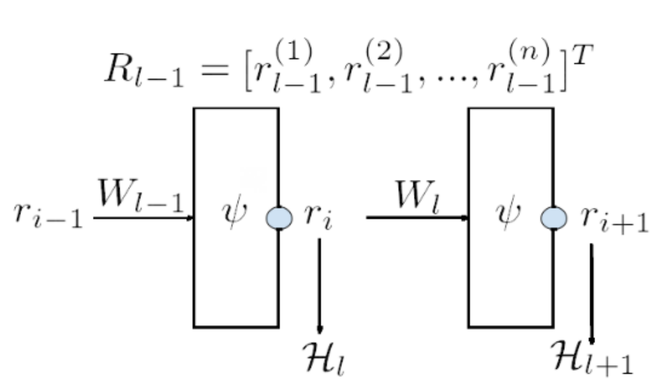


Fig. 1: Layer notation

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

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.

Theorem 1. *For any $\mathcal{H}_0$, there exists a set of bandwidths σ_l and a Kernel Sequence $\{\phi_{l^\circ}\}_{l=1}^L, \phi_l = \psi \circ \mathcal{W}_l$ parameterized by $W_l = W_s$ in Eq. (2) 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

$$\mathcal{H}^* - \mathcal{H}_L \leq \delta, \tag{3}$$

II. as $L \to \infty$, we have

$$\lim_{L\to\infty} \mathcal{H}_L = \mathcal{H}^*, \tag{4}$$

III. the convergence is strictly monotonic where

$$\mathcal{H}_l > \mathcal{H}_{l-1} \quad \forall l \geq 1. \tag{5}$$

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

$$\lim_{l\to\infty} \frac{\mathrm{Tr}(S_w^l)}{\mathrm{Tr}(S_b^l)} = 0 \tag{6}$$

II. the Kernel Sequence converges to the following kernel:

$$\lim_{l\to\infty} \mathcal{K}(x_i, x_j)^l = \mathcal{K}^*(x_i, x_j)^l = \begin{cases} 0 & \forall i,j \in \mathcal{S}^c \\ 1 & \forall i,j \in \mathcal{S} \end{cases}. \tag{7}$$

Maximize HSIC? What does it Accomplish?

Claim 2: **Implication:** The kernel is converging to a kernel that Maximize distance between the classes and Minimize distance within the classes.

Claim 3: We prove that maximizing HSIC Implicitly Minimizes Cross-Entropy (CE) and MSE for Classification.

Corollary 2.1. *Given $\mathcal{H}_l \to \mathcal{H}^*$, the network output in IDS solves MSE via a translation of labels.*

Corollary 2.2. *Given $\mathcal{H}_l \to \mathcal{H}^*$, the network output in RKHS solves CE via a change of bases.*

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

Optimally of the solution:

Claim 4: W_s is not the optimal layer-wise solution where at each layer

$$\frac{\partial}{\partial W_l}\mathcal{H}_l(W_s) \neq 0. \tag{8}$$

Claim 5: The optimal solution W^* where $\frac{\partial}{\partial W_l}\mathcal{H}_l(W_s) = 0$ is the eigenvector of the following matrix

$$\mathcal{Q}_{l^i} = R_{l-1}^T(\hat{\Gamma} - \mathrm{Diag}(\hat{\Gamma}1_n))R_{l-1}, \tag{9}$$

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

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?

Smaller section, more on generalization of W

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 6: 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

$$D_i(W_l) = \frac{1}{\sigma^2}\sum_{j\in\mathcal{S}_i}\Gamma_{i,j}\mathcal{K}_{W_l}(r_i, r_j) - \frac{1}{\sigma^2}\sum_{j\in\mathcal{S}_i^c}|\Gamma_{i,j}|\mathcal{K}_{W_l}(r_i, r_j). \tag{10}$$

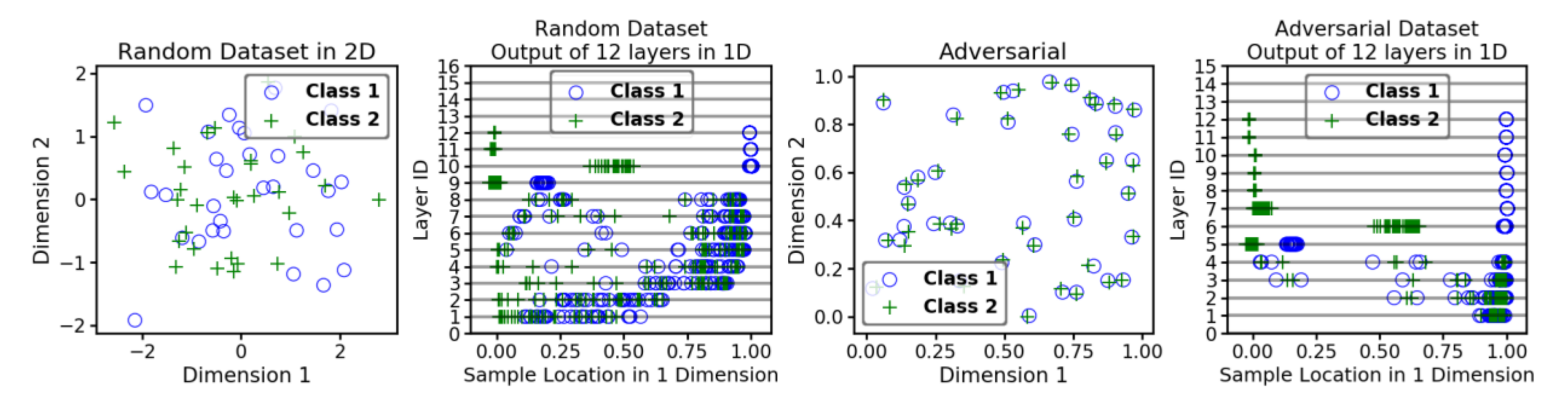
Then Eq. (1) is equivalent to

$$\max_{W_l} \sum_{i,j}\frac{\Gamma_{i,j}}{\sigma^2}e^{-\frac{(r_i-r_j)^TWW^T(r_i-r_j)}{2\sigma^2}}(r_i^TW_lW_l^Tr_j) - \sum_i D_i(W_l)||W_l^Tr_i||_2. \tag{11}$$

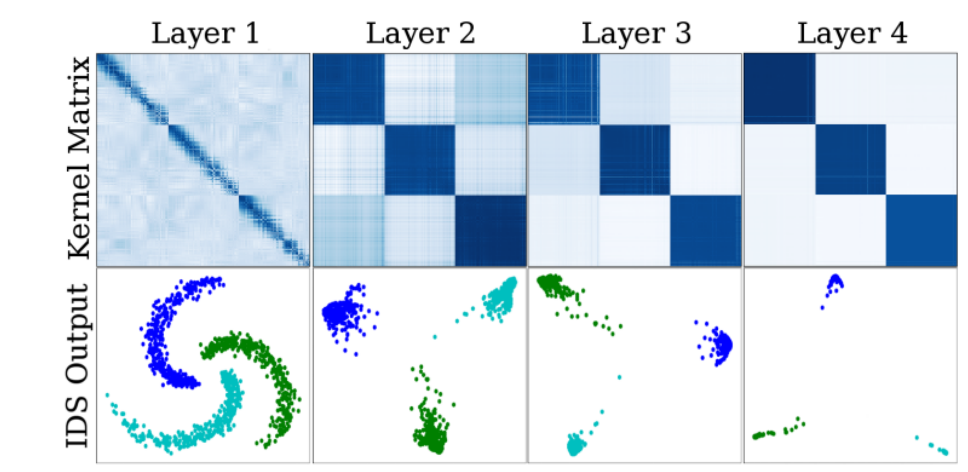
Implication: Based on this claim, $D_i(W_l)$ adds a negative variable cost to the sample norm in IDS, $||W_l^Tr_i||_2$, describing how ISM implicitly regularizes HSIC. In fact, a better W_l imposes a heavier penalty on the objective where the overall HSIC value may actually decrease.

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.



IDS and RKHS transition 3: Every layer, $\phi_l = \psi \circ \mathcal{W}_l$ send the data from previous RKHS to next RKHS, ψ is an activation function. $\mathcal{W}_l$ is linear map from RKHS to IDS.

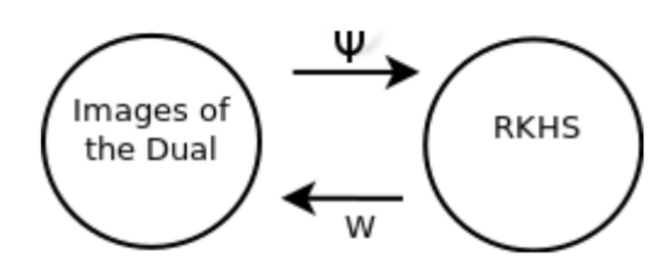


Fig. 4: Layer notation

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 ↑	Test Acc ↑	Time(s) ↓	$\mathcal{H}^* \uparrow$	MSE ↓	CE ↓	$C \downarrow$	$\mathcal{T} \downarrow$
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