

From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Networks with Affine Optimal Transport

Quentin Bouniot¹ Ievgen Redko² Anton Mallasto³ Charlotte Laclau¹
Oliver Struckmeier⁴ Karol Arndt⁴ Markus Heinonen⁴ Ville Kyrki⁴
Samuel Kaski^{4,5}

¹Telecom Paris

²Noah's Ark Lab

³Smarty.io

⁴Aalto University

⁵University of Manchester

Motivations

Non-linearity is at the heart of DNNs

- ▶ **Universal function approximators** thanks to non-linearity.
- ▶ Mainly introduced through **activation functions**, the common ingredients between architectures.

No such notion of quantifying non-linearity exists in the literature.

- ▶ Research mainly focus on quantifying **expressive power** of DNNs.

Goal: Measure non-linearity of activation functions **from data distribution**

Outline

- 1 Quantifying Non-linearity
- 2 Journey through DNNs History
- 3 Additional Results

Outline

- 1 Quantifying Non-linearity
- 2 Journey through DNNs History
- 3 Additional Results

General idea

Measure *non-linearity as lack of linearity* through *Optimal Transport (OT)*

- ▶ OT is a theoretically sounded tool when **comparing distributions**.
- ▶ We want a **bounded score** to be able to compare behaviors.

General idea

Measure *non-linearity* as *lack of linearity* through *Optimal Transport* (OT)

- ▶ We know the **closed-form solution** of the OT problem for random variables (RVs) following **normal distributions**.
- ▶ For any RVs X and Y , if $Y = TX$ with T Positive Semi-Definite (PSD) matrix, then **the solution of OT problem is exactly the one of their normal approximations** ($N_X \sim \mathcal{N}(\mu(X), \Sigma(X))$ and $N_Y \sim \mathcal{N}(\mu(Y), \Sigma(Y))$).
- ▶ We obtain an **upper bound** on the difference of the two OT problems.
- ▶ We can define the **affinity score** using this bound.

Identifiability

Theorem (Smith & Knott¹)

Let $X \sim \mu$ and $T(x) = \nabla \phi(x)$ for a convex function ϕ . Then T is the unique OT map between μ and $T_{\#}\mu$.

Corollary

If $Y = TX$ where T is a PSD affine transformation, then T is the unique OT map from X to Y .

- ▶ Uniqueness OT maps expressed as **gradients of convex functions**.
- ▶ Whenever two RVs are linked through an **affine transformation**, then the solution to the OT problem between them is **exactly the affine transformation**.

¹Cyril S Smith and Martin Knott. "Note on the optimal transportation of distributions". In: *Journal of Optimization Theory and Applications* 52.2 (1987), pp. 323–329.

Characterization

Theorem

For any RVs X, Y such that $Y = TX$ with T a PSD matrix. Let N_X and N_Y be their normal approximations. Then $W_2(N_X, N_Y) = W_2(X, Y)$ and $T = T_{\text{aff}}$ where T_{aff} is the OT map between N_X and N_Y and can be calculated in closed-form:

$$\begin{aligned} T_{\text{aff}}(x) &= Ax + b, \quad A = \Sigma(Y)^{\frac{1}{2}} \left(\Sigma(Y)^{\frac{1}{2}} \Sigma(X) \Sigma(Y)^{\frac{1}{2}} \right)^{-\frac{1}{2}} \Sigma(Y)^{\frac{1}{2}}, \\ b &= \mu(Y) - A\mu(X). \end{aligned} \tag{1}$$

- When X and Y differ by an **affine transformation**, the OT solution can be computed in closed-form using their normal approximations, **no matter how complicated X and Y are.**

Upper bound

Theorem (Gelbrich bound²)

Let X, Y be RVs and N_X, N_Y be their normal approximations. Then, $W_2(N_X, N_Y) \leq W_2(X, Y)$.

Proposition

Let X, Y be RVs and N_X, N_Y be their normal approximations. For T_{aff} as in (1),

$$W_2(T_{\text{aff}}X, Y) \leq \sqrt{2 \operatorname{Tr} [\Sigma(Y)]}. \quad (2)$$

- ▶ The difference in the cost between the two OT problems **increases when the map becomes non-linear**.
- ▶ We have equality iff RVs are linked through an **affine transformation**.

²Matthias Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Mathematische Nachrichten* 147.1 (1990), pp. 185–203.

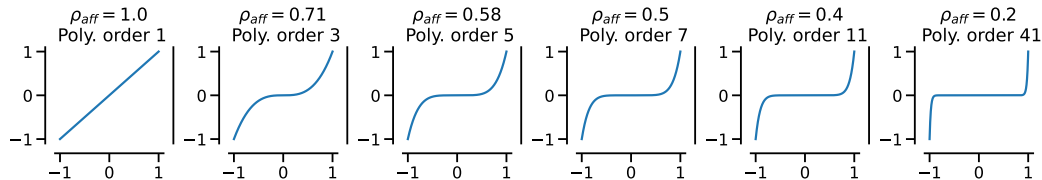
Affinity Score

$$\rho_{\text{aff}}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{W_2(T_{\text{aff}}\mathbf{X}, \mathbf{Y})}{\sqrt{2 \text{Tr}[\Sigma(\mathbf{Y})]}}$$

- ▶ ρ_{aff} describes how much Y differs from being a *PSD affine transformation* of X .
- ▶ $0 \leq \rho_{\text{aff}}(X, Y) \leq 1$, and $\rho_{\text{aff}}(X, Y) = 1 \Leftrightarrow Y = T_{\text{aff}}X$.

Toy Examples

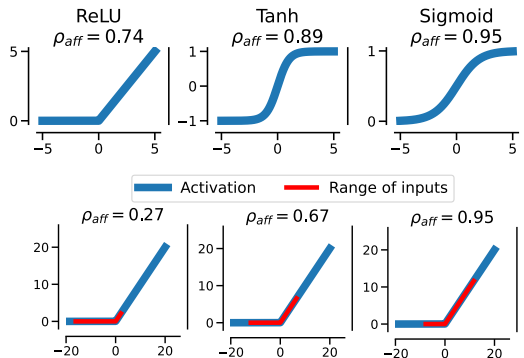
Affinity scores for simple polynomial functions



- Affinity score decreases as the transformation becomes **more non-linear**.

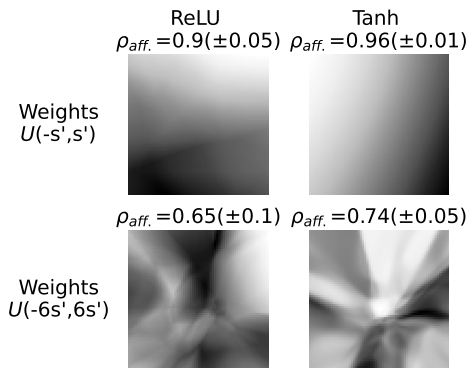
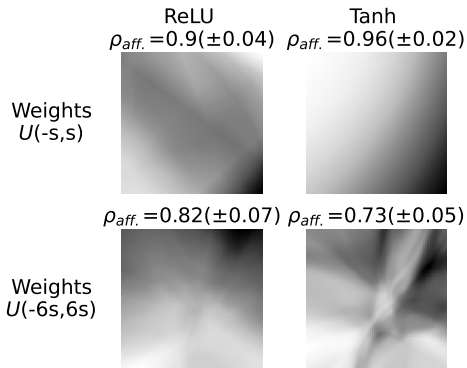
Activation functions

Affinity scores throughout the input domain of ReLU



- Affinity scores *vary* depending on the **activation function** and the **input domain** considered.
- For ReLU, **high** ρ_{aff} **values** in the linear part of the transformation.

Revisiting Neural Redshift³ phenomenon



(A) 3-layer MLP with ReLU or tanh activation following Glorot initialization.

(B) Same MLP with **shifted** $s' = s - 0.05$.

✓ Changing the domain of ReLU makes it *simplicity bias* vanish.

³Damien Teney et al. "Neural Redshift: Random Networks are not Random Functions". In: CVPR. 2024.

Outline

- 1 Quantifying Non-linearity
- 2 Journey through DNNs History**
- 3 Additional Results

Non-linearity signature

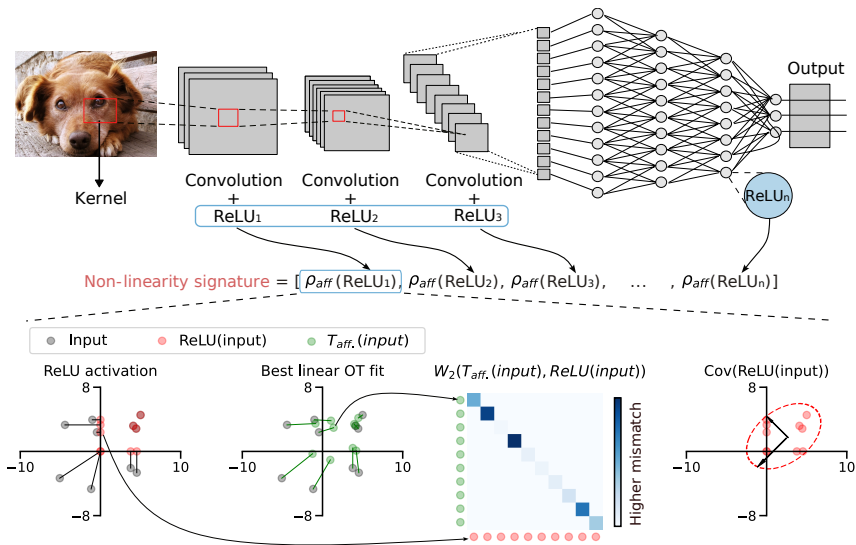
Notations

- ▶ Define a neural network N as a *composition of layers* F_i :
 $N = F_L \odot \dots \odot F_i \dots \odot F_1 = \bigodot_{k=1, \dots, L} F_k$ where $\odot$ stands for a composition.
- ▶ Each layer F_i is a function $F_i : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times c}$ whose outputs $F_i(\mathbf{X}_i)$ are inputs of the following layer F_{i+1} . Usual F_i include convolution, feedforward, pooling or activation functions.
- ▶ Define a *finite set of common activation functions* $\mathcal{A} := \{\sigma | \sigma : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times c}\}$

Non-linearity signature of N given $\mathbf{X}$:

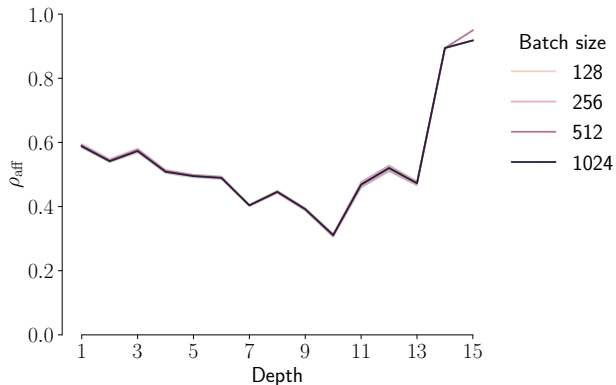
$$\rho_{\text{aff}}(N; \mathbf{X}) = \{\rho_{\text{aff}}(\mathbf{X}_i, \sigma(\mathbf{X}_i)), \forall \sigma \in F_i \cap \mathcal{A}, i \in \{1, \dots, L\}\}$$

Non-linearity signature



Empirical Properties

Stability with respect to the choice of batch size

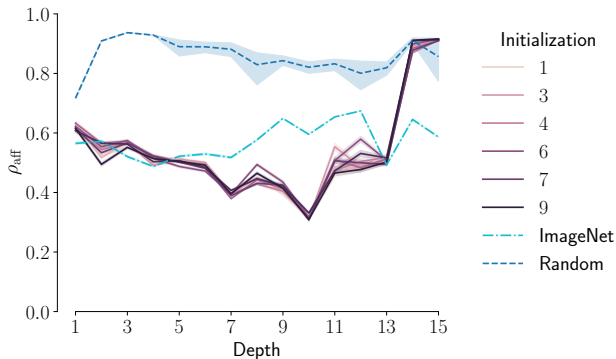


► VGG16 on CIFAR10, ρ_{aff} computed with *different batch size*

✓ Stable with respect to **batch size**

Empirical Properties

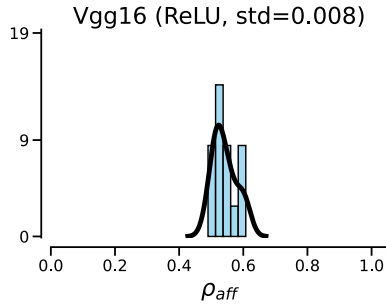
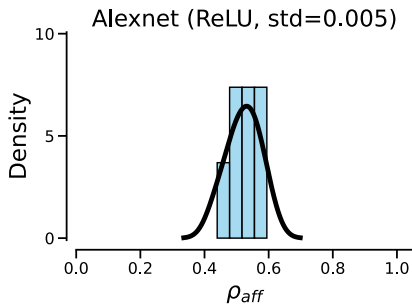
Impact of DNN's weights



► VGG16 with *different initialization* on CIFAR10

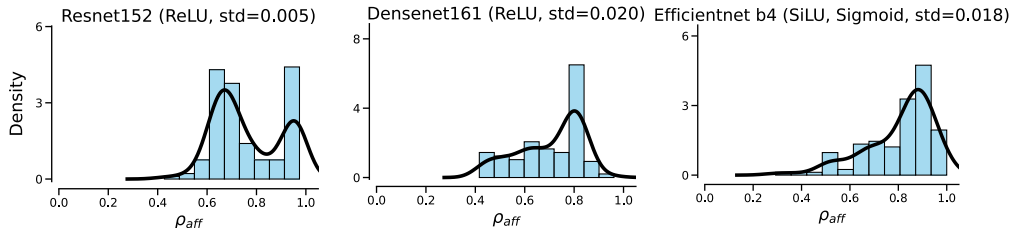
- ✓ Stable with respect to same training but **different random seeds**
- ✓ Difference between **learned weights**

Early Convnets



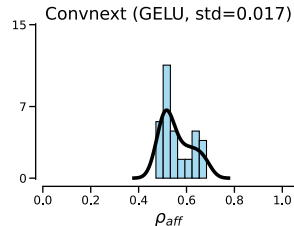
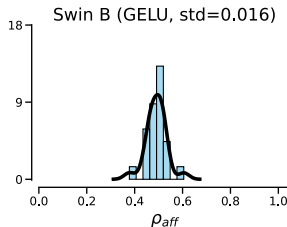
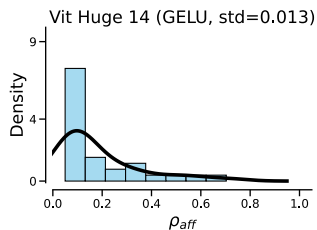
- Early convnets had **tiny variations** in non-linearity propagation.

Deeper Networks



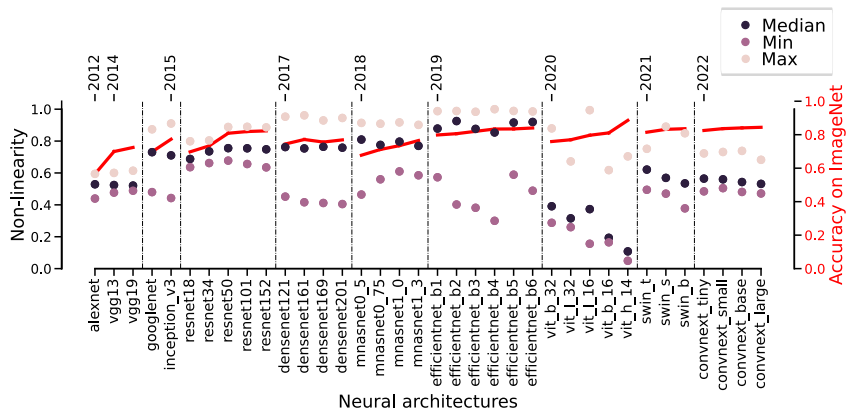
- Deeper networks introduce multiple ways for **increasing linearity** (higher scores), through, e.g., *residual connections*.

Post-ViT Era



- ▶ ViTs are **highly non-linear** compared to convnets.
- ▶ Following architectures (Swin, Convnext), combining best practices from both ViTs and deep convnets, have signatures closer to early convnets.

Throughout DNNs Architectures

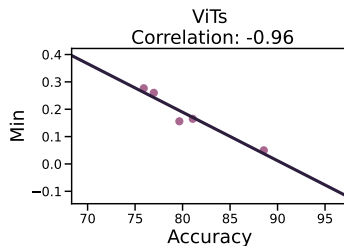
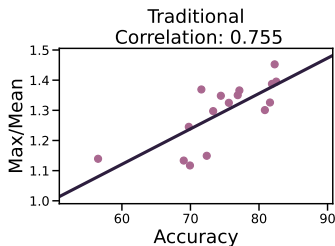


- **Affinity scores statistics and Accuracy** (in red) throughout DNNs architectures.
- Before ViTs: max and median values are increasing, also gap between min and max.
- Within ViTs: Trend of decreasing min values

Outline

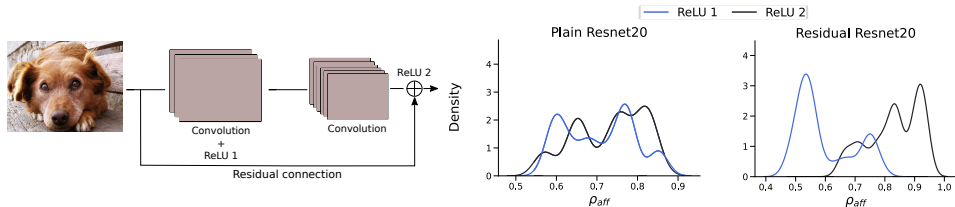
- 1 Quantifying Non-linearity
- 2 Journey through DNNs History
- 3 Additional Results**

Correlation with Accuracy



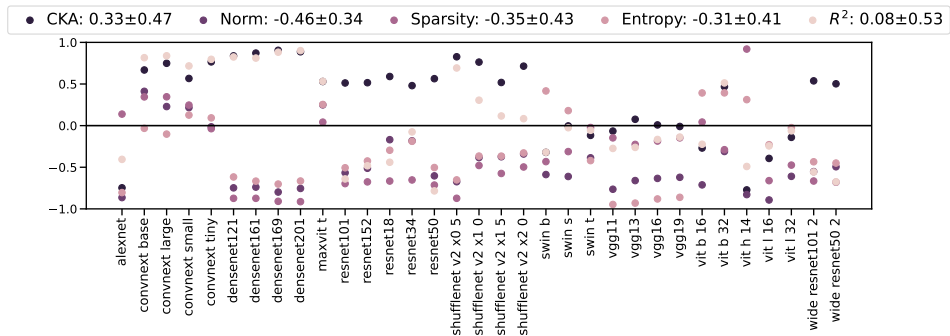
- ▶ We separate architectures into semantically meaningful groups: **Traditional architectures** (Alexnet, VGGs, ResNets and DenseNets) and **ViTs**.
- ▶ Confirms **shaking effect** for traditional models.
- ▶ Clear trend toward **more non-linearity in ViTs**.

Residual Connections



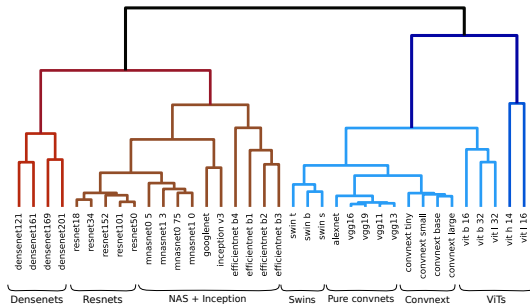
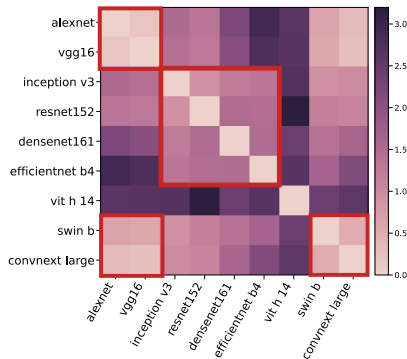
- ▶ Comparing a Resnet20 with (Residual) and without (Plain) **residual connections**.
- ▶ Residual connections introduce a clear trend toward a **bi-modal distribution** of affinity scores.

Unique Measure



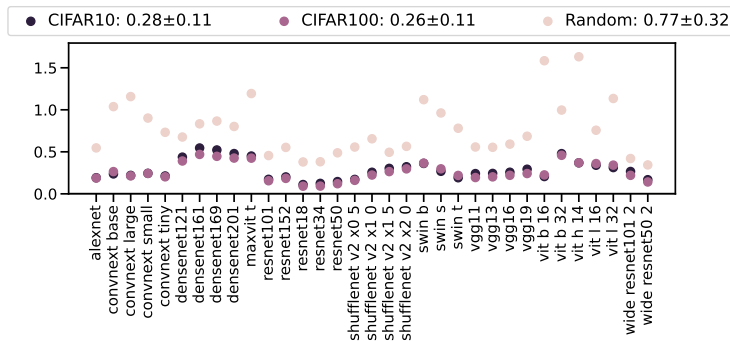
- No other criterion consistently correlates with the affinity score across 33 architectures used in our test.

Clustering of architectures



- Clustering of the architectures using the **pairwise Dynamic Time Window (DTW) distances** between non-linearity signatures.

Deviations between datasets



- Deviations to ImageNet of different datasets (CIFAR10, CIFAR100, random data), for each architecture.

Take-Home Message

Understanding Deep Neural Networks Through the Lens of their Non-Linearity⁴

- ✓ First theoretical sound tool to measure non-linearity in DNNs
- ✓ Different developments in Deep Learning can be understood through the prism of non-linearity
- ✓ Variety of potential applications

⁴Quentin Bouniot et al. "From alexnet to transformers: Measuring the non-linearity of deep neural networks with affine optimal transport". In: CVPR. 2025.

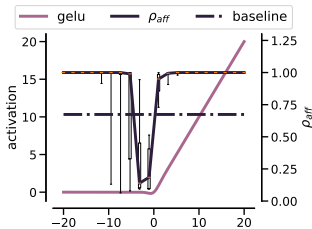
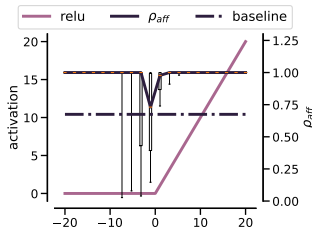
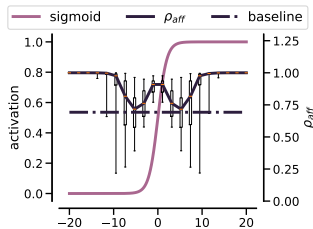
Thank you for listening !

Do not hesitate to contact us if you have questions.

- [4] Quentin Bouniot et al. “From alexnet to transformers: Measuring the non-linearity of deep neural networks with affine optimal transport”. In: *CVPR*. 2025.

More Examples of Activation Functions

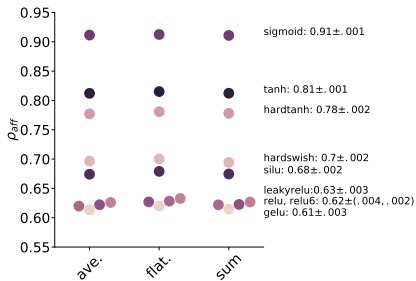
Affinity scores over input domain of activation functions



- ▶ $\mathbf{X} \sim \mathcal{N}(\mu, \sigma)$, with μ sliding over the domain and multiple σ for each μ .
- ▶ $\rho_{aff}(\mathbf{X}, f(\mathbf{X}))$ for popular activation functions f .
- ▶ Activation functions can be characterized by **the lowest score achieved** and **the range of non-linearity**.

Robustness to Pooling

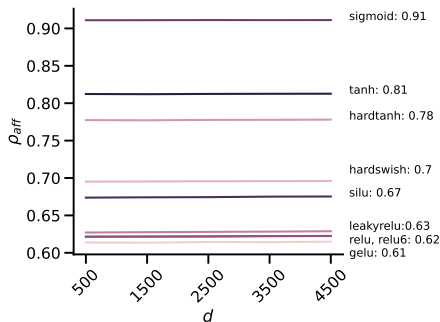
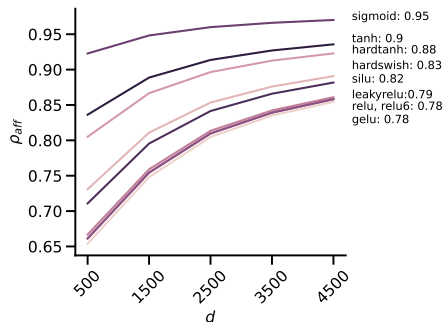
Affinity scores are robust to pooling



- Manipulating 4-order tensor is computationally expensive
- ✓ Averaging over a dimension preserve affinity scores

Covariance estimation

Shrinkage of the covariance makes it robust to sample size

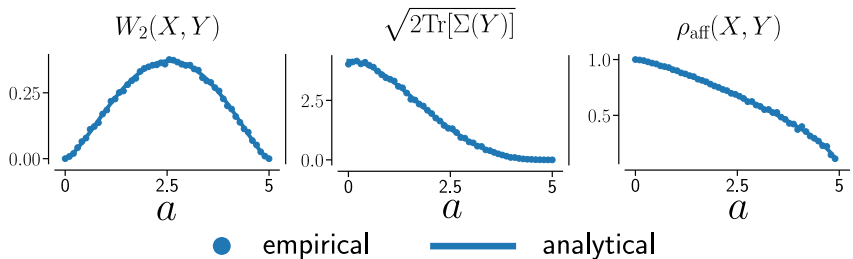


✓ Ledoit-Wolfe shrinkage⁵ of the covariance gives stable results for affinity scores.

⁵Olivier Ledoit and Michael Wolf. "Honey, I shrunk the sample covariance matrix". In: *Journal of Portfolio Management* 30.4 (2004), pp. 110–119.

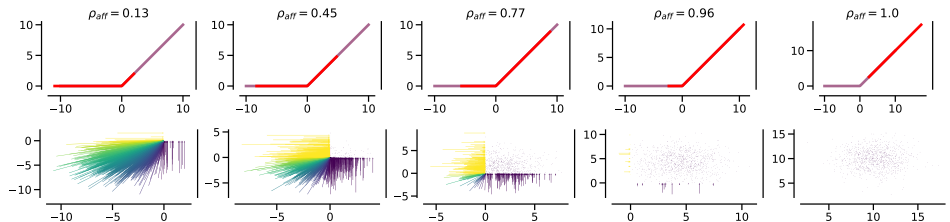
On perfect linearity and Asymmetry

$$X \sim \mathcal{U}([-a, -a + 5]), Y = \text{ReLU}(X)$$



- Affinity score is not symmetric around the origin because of the difference in behavior of the numerator and denominator.
- Numerator is roughly symmetric and main source of asymmetry comes from the denominator.

Displacement of mass with ReLU



- Effect of ReLU is not symmetric around the origin.
- Moving everything to a single point has more impact