Bottleneck Structure in Large Depth Networks

Mechanisms of Symmetry Learning

Arthur Jacot

New York University

November 2, 2023

Curse of Dimensionality

- Goal: learn a function $f^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ from N random observations $y_i = f(x_i)$.

Curse of Dimensionality

- Goal: learn a function $f^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ from N random observations $y_i = f(x_i)$.
- Linear Models / Kernel Methods:
  - Given data distribution, decompose $f^* = \sum_{k=1}^{\infty} \beta_k f_k$ (kernel PCA).
  - For large N: estimate $\hat{f} \approx \sum_{k=1}^{N} \beta_k f_k \implies \mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx \sum_{k=N+1}^{\infty} \beta_k^2$.

Curse of Dimensionality

- Goal: learn a function $f^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ from N random observations $y_i = f(x_i)$.
- Linear Models / Kernel Methods:
  - Given data distribution, decompose $f^* = \sum_{k=1}^{\infty} \beta_k f_k$ (kernel PCA).
  - For large N: estimate $\hat{f} \approx \sum_{k=1}^{N} \beta_k f_k \implies \mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx \sum_{k=N+1}^{\infty} \beta_k^2$.
- Polynomial basis: $f^*(x) = \sum_{m \in \mathbb{N}^{d_{in}}} \beta_m x_1^{m_1} \cdots x_{d_{in}}^{m_{d_{in}}}$.
  - Number of degree m monomial: $m^{d_{in}}$.

Curse of Dimensionality

- Goal: learn a function $f^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ from N random observations $y_i = f(x_i)$.
- Linear Models / Kernel Methods:
  - Given data distribution, decompose $f^* = \sum_{k=1}^{\infty} \beta_k f_k$ (kernel PCA).
  - For large N: estimate $\hat{f} \approx \sum_{k=1}^{N} \beta_k f_k \implies \mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx \sum_{k=N+1}^{\infty} \beta_k^2$.
- Polynomial basis: $f^*(x) = \sum_{m \in \mathbb{N}^{d_{in}}} \beta_m x_1^{m_1} \cdots x_{d_{in}}^{m_{d_{in}}}$.
  - Number of degree m monomial: $m^{d_{in}}$.
  - If $|\beta_m|^2 \sim \|m\|_1^{-\alpha - d_{in}}$ then k-largest coefficient $|\beta_k|^2 \sim k^{-\frac{\alpha + d_{in}}{d_{in}}}$.
  - Error $\mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx N^{-\frac{\alpha}{d_{in}}}$.

Curse of Dimensionality

- Goal: learn a function $f^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ from N random observations $y_i = f(x_i)$.
- Linear Models / Kernel Methods:
  - Given data distribution, decompose $f^* = \sum_{k=1}^{\infty} \beta_k f_k$ (kernel PCA).
  - For large N: estimate $\hat{f} \approx \sum_{k=1}^{N} \beta_k f_k \implies \mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx \sum_{k=N+1}^{\infty} \beta_k^2$.
- Polynomial basis: $f^*(x) = \sum_{m \in \mathbb{N}^{d_{in}}} \beta_m x_1^{m_1} \cdots x_{d_{in}}^{m_{d_{in}}}$.
  - Number of degree m monomial: $m^{d_{in}}$.
  - If $|\beta_m|^2 \sim \|m\|_1^{-\alpha-d_{in}}$ then k-largest coefficient $|\beta_k|^2 \sim k^{-\frac{\alpha+d_{in}}{d_{in}}}$.
  - Error $\mathbb{E}_x \left\| \hat{f}(x) - f^*(x) \right\|^2 \approx N^{-\frac{\alpha}{d_{in}}}$.
- How can DNN learn text and image tasks successfully?
  - DNNs capture a low-dim structure in these tasks.

Breaking the Curse of Dimensionality

Possible structures:

- The data lies on a d_{surf}-dimensional surface, with $d_{surf} \leq d_{in}$.
  - ‘asdjgoijdskjasry asifudh’ vs ‘Good morning!’

Breaking the Curse of Dimensionality

Possible structures:

- The data lies on a d_{surf}-dimensional surface, with $d_{surf} \leq d_{in}$.
  - 'asdjgoijdskjasry asifudh' vs 'Good morning!'
- Symmetries $f(g \cdot x) = f(x) \Rightarrow$ learn only ${}^{f}/_{G} : \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$:
  - Grammar rules: $p(\alpha|$'Ann left. She α'$) \approx p(\alpha|$'Ann left. Ann α'$)$.
  - Reasoning: $p(\alpha|$'It is raining. α'$) \approx p(\alpha|$'It is raining, the road is wet. α'$)$.

Breaking the Curse of Dimensionality

Possible structures:

- The data lies on a d_{surf}-dimensional surface, with $d_{surf} \leq d_{in}$.
  - 'asdjgoijdskjasry asifudh' vs 'Good morning!'
- Symmetries $f(g \cdot x) = f(x) \Rightarrow$ learn only $f/G : \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$:
  - Grammar rules: $p(\alpha|$'Ann left. She α'$) \approx p(\alpha|$'Ann left. Ann α'$)$.
  - Reasoning: $p(\alpha|$'It is raining. α'$) \approx p(\alpha|$'It is raining, the road is wet. α'$)$.
- Known symmetries: design specific features/kernels [Mallat, 2012].

Breaking the Curse of Dimensionality

Possible structures:

- The data lies on a d_{surf}-dimensional surface, with $d_{surf} \leq d_{in}$.
  - ‘asdjgoijdskjasry asifudh’ vs ‘Good morning!’
- Symmetries $f(g \cdot x) = f(x) \Rightarrow$ learn only $f/G : \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$:
  - Grammar rules: $p(\alpha|$‘Ann left. She α’$) \approx p(\alpha|$‘Ann left. Ann α’$)$.
  - Reasoning: $p(\alpha|$‘It is raining. α’$) \approx p(\alpha|$‘It is raining, the road is wet. α’$)$.
- Known symmetries: design specific features/kernels [Mallat, 2012].
- Shallow network learn functions of the form $f = h(Ax)$ with $\mathrm{Rank} A < d_{full}$ [Bach, 2017, Abbe et al., 2021].
  - Learns translation symmetries: $f(x + v) = f(x)$ for all $v \in \ker A$.

Breaking the Curse of Dimensionality

Possible structures:

- The data lies on a d_{surf}-dimensional surface, with $d_{surf} \leq d_{in}$.
  - 'asdjgoijdskjasry asifudh' vs 'Good morning!'
- Symmetries $f(g \cdot x) = f(x) \Rightarrow$ learn only $f/G : \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$:
  - Grammar rules: $p(\alpha|$'Ann left. She α'$) \approx p(\alpha|$'Ann left. Ann α'$)$.
  - Reasoning: $p(\alpha|$'It is raining. α'$) \approx p(\alpha|$'It is raining, the road is wet. α'$)$.
- Known symmetries: design specific features/kernels [Mallat, 2012].
- Shallow network learn functions of the form $f = h(Ax)$ with $\mathrm{Rank} A < d_{full}$ [Bach, 2017, Abbe et al., 2021].
  - Learns translation symmetries: $f(x + v) = f(x)$ for all $v \in \ker A$.
- Deep Networks learn functions $f = g \circ h$ with small inner dimension.
  - Learns general symmetries $f = \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$ (e.g. $f(Rx) = f(x)$ for rotations R).

Deep Neural Networks

Network with layers $\ell = 0, \ldots, L$ each containing w_ℓ neurons.

- Activations

$$\alpha_0(x) = x$$
$$\alpha_\ell(x) = \sigma\left(W_\ell \alpha_{\ell-1}(x) + b_\ell\right)$$
$$f_\theta(x) = W_L \sigma(\alpha_{L-1}) + b_L$$

- Parameters $\theta = (W_1, b_1, \ldots, W_L, b_L)$.
  - Initialized randomly $\theta \sim \mathcal{N}(0, \sigma^2)$.
  - Trained with gradient descent on the loss

$$\mathcal{L}_\lambda(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| f_\theta(x_i) - f^*(x_i) \right\|^2 + \lambda \left\| \theta \right\|^2 .$$

- Depth L, width $w = w_1 = \cdots = w_{L-1}$.

L_2-regularization

- L_2-Regularization is the ‘simplest’ regime that exhibits sparsity / symmetry learning.
  - Bias of DNNs is explicit (low parameter norm) instead of implicit (bias of GD/GD).
- Representation cost $R(f) = \min_{\theta : f_\theta = f} \|\theta\|^2$

$$\min_\theta C(f_\theta) + \lambda \|\theta\|^2 = \min_f C(f) + \lambda R(f).$$

L_2-regularization

- L_2-Regularization is the 'simplest' regime that exhibits sparsity / symmetry learning.
  - Bias of DNNs is explicit (low parameter norm) instead of implicit (bias of GD/GD).
- Representation cost $R(f) = \min_{\theta : f_\theta = f} \|\theta\|^2$

$$\min_\theta C(f_\theta) + \lambda \|\theta\|^2 = \min_f C(f) + \lambda R(f).$$

- Linear FCNN $A_\theta = W_L \cdots W_1$: L_p-Schatten norm $R(A) = L \sum_{i=1}^{\mathrm{Rank} A} s_i(A)^{\frac{2}{L}}$ [Dai et al., 2021].
  - Low-rank bias: learns translation symmetries $A(x+v) = Ax$ for $v \in \ker A$.

L_2-regularization

- L_2-Regularization is the ‘simplest’ regime that exhibits sparsity / symmetry learning.
  - Bias of DNNs is explicit (low parameter norm) instead of implicit (bias of GD/GD).
- Representation cost $R(f) = \min_{\theta : f_\theta = f} \|\theta\|^2$

$$\min_\theta C(f_\theta) + \lambda \|\theta\|^2 = \min_f C(f) + \lambda R(f).$$

- Linear FCNN $A_\theta = W_L \cdots W_1$: L_p-Schatten norm $R(A) = L \sum_{i=1}^{\mathrm{Rank} A} s_i(A)^{\frac{2}{L}}$ [Dai et al., 2021].
  - Low-rank bias: learns translation symmetries $A(x+v) = Ax$ for $v \in \ker A$.
- What is the rank of a nonlinear function?

Rank of nonlinear functions

There are multiple reasonable notions of rank for finite piecewise linear functions (FPLFs):

- Jacobian Rank: $\mathrm{Rank}_J(f;\Omega) = \max_{x\in\Omega} \mathrm{Rank}\,(Jf(x))$
- Bottleneck Rank $\mathrm{Rank}_{BN}(f;\Omega)$: the smallest k s.t. $f = \Omega \xrightarrow{g} \mathbb{R}^k \xrightarrow{h} \mathbb{R}^{d_{out}}$.

Rank of nonlinear functions

There are multiple reasonable notions of rank for finite piecewise linear functions (FPLFs):

- Jacobian Rank: $\mathrm{Rank}_J(f;\Omega) = \max_{x\in\Omega} \mathrm{Rank}\,(Jf(x))$
- Bottleneck Rank $\mathrm{Rank}_{BN}(f;\Omega)$: the smallest k s.t. $f = \Omega \xrightarrow{g} \mathbb{R}^k \xrightarrow{h} \mathbb{R}^{d_{out}}$.

Both satisfy

1. $\mathrm{Rank}(f \circ g) \leq \min\{\mathrm{Rank} f, \mathrm{Rank} g\}$,
2. $\mathrm{Rank}(f + g) \leq \mathrm{Rank} f + \mathrm{Rank} g$,
3. $\mathrm{Rank}\,(x \mapsto Ax + b) = \mathrm{Rank} A$.

Rank of nonlinear functions

There are multiple reasonable notions of rank for finite piecewise linear functions (FPLFs):

- Jacobian Rank: $\mathrm{Rank}_J(f;\Omega) = \max_{x\in\Omega} \mathrm{Rank}\left(Jf(x)\right)$
- Bottleneck Rank $\mathrm{Rank}_{BN}(f;\Omega)$: the smallest k s.t. $f = \Omega \xrightarrow{g} \mathbb{R}^k \xrightarrow{h} \mathbb{R}^{d_{out}}$.

Both satisfy

1. $\mathrm{Rank}(f \circ g) \leq \min\{\mathrm{Rank} f, \mathrm{Rank} g\}$,
2. $\mathrm{Rank}(f + g) \leq \mathrm{Rank} f + \mathrm{Rank} g$,
3. $\mathrm{Rank}\left(x \mapsto Ax + b\right) = \mathrm{Rank} A$.

$\Rightarrow \mathrm{Rank} f \leq \min\{d_{in}, d_{out}\}$
$\Rightarrow \mathrm{Rank} \phi \circ f \circ \psi = \mathrm{Rank} f$ for bijections ϕ, ψ.

Infinite Depth Limit

The infinite depth representation cost $R^{(0)}(f;\Omega) := \lim_{L\to\infty} \frac{R(f;\Omega,L)}{L}$ is a notion of rank

Theorem (*Jacot 2023a*)

For a bounded Ω, $R^{(0)}$ satisfies properties (1,2,3) and

$$\mathrm{Rank}_J(f;\Omega) \leq R^{(0)}(f;\Omega) \leq \mathrm{Rank}_{BN}(f;\Omega).$$

Infinite Depth Limit

The infinite depth representation cost $R^{(0)}(f;\Omega) := \lim_{L\to\infty} \frac{R(f;\Omega,L)}{L}$ is a notion of rank

Theorem (*Jacot 2023a*)

For a bounded Ω, $R^{(0)}$ satisfies properties (1,2,3) and

$$\mathrm{Rank}_J(f;\Omega) \leq R^{(0)}(f;\Omega) \leq \mathrm{Rank}_{BN}(f;\Omega).$$

Conjecture: $R^{(0)}(f;\Omega) = \mathrm{Rank}_{BN}(f;\Omega)$.
Proven for functions $f = \phi \circ A \circ \psi$ for bijections ϕ, ψ.

- Symmetries lead to low BN-rank: $f^* : \Omega \to \Omega/G \to \mathbb{R}^{d_{out}} \Rightarrow \mathrm{Rank}_{BN}(f^*;\Omega) \leq \dim \Omega/G$.
- Functions with symmetries require a small parameter norm.

Sketch of proof: Bottleneck Structure

Upper bound: For f of the form $\mathbb{R}^{d_{in}} \xrightarrow{g} \mathbb{R}^k_+ \xrightarrow{h} \mathbb{R}^{d_{out}}$ represent f as:

1. L_g layers representing g.
2. $L - L_g - L_h$ representing the identity on $\mathbb{R}^k_+$.
3. L_h layers representing h.

Sketch of proof: Bottleneck Structure

Upper bound: For f of the form $\mathbb{R}^{d_{in}} \xrightarrow{g} \mathbb{R}_+^k \xrightarrow{h} \mathbb{R}^{d_{out}}$ represent f as:

1. L_g layers representing g.
2. $L - L_g - L_h$ representing the identity on $\mathbb{R}_+^k$.
3. L_h layers representing h.

The identity layers each have parameter norm $\|W_\ell\|^2 = k$:

$$\frac{\|\theta\|^2}{L} = \frac{\|\theta_g\|^2 + \|\theta_h\|^2 + (L - L_g - L_h)k}{L} \xrightarrow{L\to\infty} k.$$

Sketch of proof: Bottleneck Structure

Upper bound: For f of the form $\mathbb{R}^{d_{in}} \stackrel{g}{\to} \mathbb{R}^k_+ \stackrel{h}{\to} \mathbb{R}^{d_{out}}$ represent f as:

1. L_g layers representing g.
2. $L - L_g - L_h$ representing the identity on $\mathbb{R}^k_+$.
3. L_h layers representing h.

The identity layers each have parameter norm $\|W_\ell\|^2 = k$:

$$\frac{\|\theta\|^2}{L} = \frac{\|\theta_g\|^2 + \|\theta_h\|^2 + (L - L_g - L_h)k}{L} \xrightarrow{L\to\infty} k.$$

Lower bound: For all $x \in \Omega$ let $L \to \infty$ in

$$\|Jf(x)\|_{2/L}^{2/L} = \|W_L D_{L-1}(x) \cdots D_1(x) W_1\|_{2/L}^{2/L} \leq \frac{\|W_L\|_F^2 + \cdots + \|W_1\|_F^2}{L} = \frac{\|\theta\|^2}{L}.$$

First correction

Theorem (Jacot 2023b)

At any point x where $\mathrm{Rank} Jf(x) = R^{(0)}(f;\Omega)$,

$$2\log|Jf(x)|_+ \leq R^{(1)}(f;\Omega),$$

1. $R^{(0)}(f\circ g) = R^{(0)}f = R^{(0)}g \Rightarrow R^{(1)}(f\circ g) \leq R^{(1)}f + R^{(1)}g$,
2. $R^{(0)}(f+g) = R^{(0)}f + R^{(0)}g \Rightarrow R^{(1)}(f+g) \leq R^{(1)}f + R^{(1)}g$,
3. *Under some cond. on* Ω, $R^{(1)}(x \mapsto Ax+b) = 2\log|A|_+$.

First correction

Theorem (Jacot 2023b)

At any point x where $\mathrm{Rank} Jf(x) = R^{(0)}(f;\Omega)$,

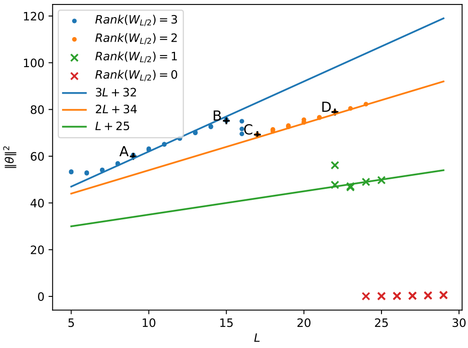
$$2\log|Jf(x)|_+ \leq R^{(1)}(f;\Omega),$$

1. $R^{(0)}(f \circ g) = R^{(0)}f = R^{(0)}g \Rightarrow R^{(1)}(f \circ g) \leq R^{(1)}f + R^{(1)}g$,
2. $R^{(0)}(f + g) = R^{(0)}f + R^{(0)}g \Rightarrow R^{(1)}(f + g) \leq R^{(1)}f + R^{(1)}g$,
3. *Under some cond. on* Ω, $R^{(1)}(x \mapsto Ax + b) = 2\log|A|_+$.

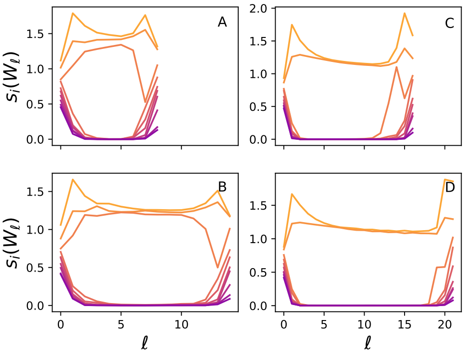
Balance between dimension reduction $R^{(0)}$ and regularity $R^{(1)}$:

$$\min_f C(f(X)) + \lambda L R^{(0)}(f) + \lambda R^{(1)}(f).$$

Parameter norm and depth



(a) Parameter norm and depth



(b) Bottleneck structure at different depths.

Impact of the Output Dim.

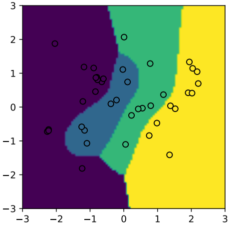
- General symmetries $\sim$: $f^*(x) = f^*(y)$ for all $x \sim y$.
  - $\mathrm{Rank}_{BN}(f^*; \Omega) \leq \dim \Omega/\sim$.
- Full Bottleneck $\dim \Omega/\sim \; < \min\{d_{in}, d_{out}\}$:
  - Inner dimension is smaller than input and output.
  - Non-generic: measure zero amongst functions.

Impact of the Output Dim.

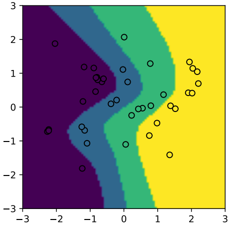
- General symmetries $\sim$: $f^*(x) = f^*(y)$ for all $x \sim y$.
  - $\mathrm{Rank}_{BN}(f^*; \Omega) \leq \dim \Omega/\sim$.
- Full Bottleneck $\dim \Omega/\sim < \min\{d_{in}, d_{out}\}$:
  - Inner dimension is smaller than input and output.
  - Non-generic: measure zero amongst functions.
- Half bottleneck $\dim \Omega/\sim \geq d_{out}$:
  - 'Full symmetry' $x \sim_{full} y \iff f^*(x) = f^*(y)$ vs 'True symetry' $(\sim) \prec (\sim_{full})$.
  - DNN learn $(\sim_{full})$ instead of $(\sim)$ in the bottleneck.
  - The true symmetry could be learned before the bottleneck.

Implications: Classification

- Class boundaries of a rank k classifier are topologically akin to dim. k classifications.
  - When $k = 1$, no tripoints (intersection of three classes)



(c) $L = 2, \lambda = 10^{-3}$



(d) $L = 9, \lambda = 10^{-3}$

Figure: Classification on 4 classes for two depths with L_2-regularization.

Symmetry overfitting?

- Finite data: always possible to fit with rank 1 $\Rightarrow$ rank underestimation!
  - Learns ‘spurious symmetries’.
  - Rank understimation is rare in practice. Why?

Symmetry overfitting?

- Finite data: always possible to fit with rank 1 $\Rightarrow$ rank underestimation!
  - Learns 'spurious symmetries'.
  - Rank understimation is rare in practice. Why?

Theorem (*A.J., 2023a*)

Given f^* with $\mathrm{Rank}_J(f^*; \Omega) = k^* > 1$, then for all ϵ there is a constant c_ϵ such that for any BN-rank 1 function $\hat{f}$ that fits $\hat{f}(x_i) = f^*(x_i)$ a dataset $x_1, \dots, x_N$ sampled i.i.d. from a distribution p with support Ω, we have $R^{(1)}(\hat{f}; \Omega, \sigma_a, L) > 2\left(1 - \frac{1}{k^*}\right) \log N + c_\epsilon$ with prob. at least $1 - \epsilon$.

Minima stability

Another possible explanation is that rank underestimating minima are unstable under reasonable learning rates $\eta \sim L^{-1}$:

Theorem (*A.J., 2023b*)

Given f^* with $\mathrm{Rank}_J(f^*;\Omega) = k^* > 1$, then with high probability over the sampling of a training set $x_1, \ldots, x_N$ (sampled from a distribution with support Ω), we have that for any parameters θ of a deep enough network that represent a BN-rank 1 function f_θ that fits the training set $f_\theta(x_i) = f^*(x_i)$ with norm $\|\theta\|^2 = L + c_1$ then there is a point $x \in \Omega$ where

$$\|J_\theta f_\theta(x)\|_F^2 \geq c'' L e^{-c_1} N^{4-\frac{4}{k^*}}.$$

GD with learning rate η cannot converge to a minima with $\frac{2}{N}\|J_\theta f_\theta(x_i)\|_{op}^2 \geq \eta^{-1}$.

Representation geodesics

- Representations $\alpha_\ell(x) = ((W_\ell \cdot + b_\ell) \circ \sigma \circ \cdots \circ \sigma \circ (W_1 \cdot + b_1))(x)$
- Infinite depth convergence of $\ell \mapsto \Sigma_\ell(x, y) = \alpha_\ell(x)^T \alpha_\ell(y)$?
  - Linear networks: $\Sigma_\ell(x, y) = x^T \left(A^T A\right)^{\frac{\ell}{L}} y$ ‘straight line in log space’.

Representation geodesics

- Representations $\alpha_\ell(x) = ((W_\ell \cdot + b_\ell) \circ \sigma \circ \cdots \circ \sigma \circ (W_1 \cdot + b_1))(x)$
- Infinite depth convergence of $\ell \mapsto \Sigma_\ell(x, y) = \alpha_\ell(x)^T \alpha_\ell(y)$?
  - Linear networks: $\Sigma_\ell(x, y) = x^T \left(A^T A\right)^{\frac{\ell}{L}} y$ ‘straight line in log space’.
- Limiting representations $K_p = \lim_{L\to\infty} \Sigma_\ell$ with $\frac{\ell}{L} \to p \in (0, 1)$ satisfy

$$R^{(0)}(f; \Omega) = R^{(0)}(id \to K_p; \Omega) = R^{(0)}(K_p \to f; \Omega),$$
$$R^{(1)}(f; \Omega) = R^{(1)}(id \to K_p; \Omega) + R^{(1)}(K_p \to f; \Omega).$$

- At any ratio $p \in (0, 1)$ with a continuous limit:

$$R^{(0)}(K_p \to K_p; \Omega) = R^{(0)}(f; \Omega),$$
$$R^{(1)}(K_p \to K_p; \Omega) = 0.$$

Identity cost

- $\mathrm{Rank}(id;\Omega)$ defines a notion of dimension of Ω.
- $\mathrm{Rank}_J(id;\Omega)$ is maximum local dimension.
- $\mathrm{Rank}_{BN}(id;\Omega)$ is embedding dimension.

Identity cost

- $\mathrm{Rank}(id;\Omega)$ defines a notion of dimension of Ω.
- $\mathrm{Rank}_J(id;\Omega)$ is maximum local dimension.
- $\mathrm{Rank}_{BN}(id;\Omega)$ is embedding dimension.

Proposition

For a domain Ω with $\mathrm{Rank}_J(id;\Omega) = \mathrm{Rank}_{BN}(id;\Omega) = k$, then $R^{(1)}(id;\Omega) = 0$ if and only if Ω is k-planar and completely positive.

- Piecewise continuous limit $\Sigma_p \Rightarrow k$-planar repr. at almost every ratio p.

Identity cost

- $\mathrm{Rank}(id;\Omega)$ defines a notion of dimension of Ω.
- $\mathrm{Rank}_J(id;\Omega)$ is maximum local dimension.
- $\mathrm{Rank}_{BN}(id;\Omega)$ is embedding dimension.

Proposition

For a domain Ω with $\mathrm{Rank}_J(id;\Omega) = \mathrm{Rank}_{BN}(id;\Omega) = k$, then $R^{(1)}(id;\Omega) = 0$ if and only if Ω is k-planar and completely positive.

- Piecewise continuous limit $\Sigma_p \Rightarrow k$-planar repr. at almost every ratio p.
- But Σ_ℓ does not converge in general!

Bottleneck Structure on the Weights

The weights of almost all layers are approximately rank k:

Theorem

Given parameters θ of a depth L network, with $\|\theta\|^2 \leq kL + c_1$ and a point x such that $\mathrm{Rank} Jf_\theta(x) = k$, then there are $w_\ell \times k$ (semi-)orthonormal V_ℓ such that

$$\sum_{\ell=1}^{L} \left\| W_\ell - V_\ell V_{\ell+1}^T \right\|_F^2 \leq c_1 - 2 \log |Jf_\theta(x)|_+$$

thus for any $p \in (0, 1)$ there are at least $(1 - p)L$ layers ℓ with

$$\left\| W_\ell - V_\ell V_{\ell-1}^T \right\|_F^2 \leq \frac{c_1 - 2 \log |Jf_\theta(x)|_+}{pL}.$$

Convergence of the representations

The representations $\alpha_\ell(x)$ of almost all layers converge, assuming a stable network (so that GD with learning rate $\eta \sim L^{-1}$ can converge to it):

Theorem

If furthermore $\|J_\theta f_\theta(x)\|_F^2 \leq cL$, *then* $\sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x)\|_2^2 \leq \frac{cLe^{\frac{2}{k}c_1}}{k|Jf_\theta(x)|_+^{2/k}}$ *and thus for all* $p \in (0,1)$ *there are at least* $(1-p)L$ *layers such that*

$$\|\alpha_{\ell-1}(x)\|_2^2 \leq \frac{1}{p}\frac{ce^{\frac{2}{k}c_1}}{k\,|Jf_\theta(x)|_+^{2/k}}.$$

$\implies$ Symmetries are learned in the first $o(L)$ layers as $L \to \infty$.

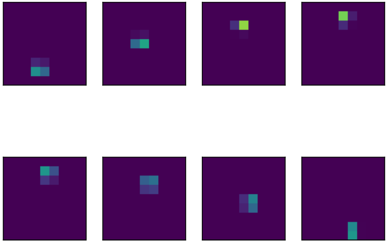
Convolutional Networks

- Inputs x and activations $\alpha_\ell(x)$ are $n \times n$ images with w_ℓ channels.
- Weights W_ℓ are multi-channel convolutions.
- Can represent a general translation equivariant functions f_θ.

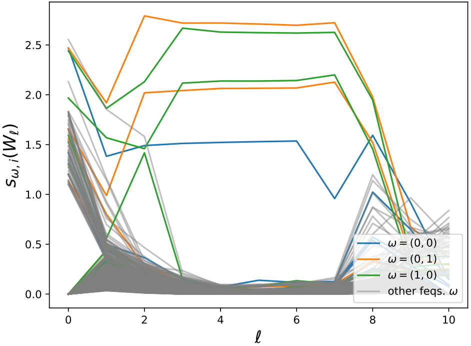
Convolutional Networks

- Inputs x and activations $\alpha_\ell(x)$ are $n \times n$ images with w_ℓ channels.
- Weights W_ℓ are multi-channel convolutions.
- Can represent a general translation equivariant functions f_θ.
- Bottleneck structure:
  - The singular $s_{\omega,i}(W_\ell)$ are indexed by frequency $\omega \in [0, n-1]^2$ and channel i.
  - In the bottleneck, only a few singular values are close to 1.

Learning Newtonian Mechanics



(a) Learning the trajectory of a 'ball' under gravity.



(b) Singular values of W_ℓ colored by frequency. The network keeps position and velocity in two freq. 1 pairs.

Conclusion

- Botleneck structure appears in L_2-regularized DNNs.
- Relations between:
  - Dimensionality inside the bottleneck.
  - Large depth L parameter norm.
  - Dimensionality of the symmetries of the task.
- To show: This breaks the curse of dimensionality!

Bibliography I

Emmanuel Abbe, Enric Boix-Adserà, Matthew Stewart Brennan, Guy Bresler, and Dheeraj Mysore Nagaraj. The staircase property: How hierarchical structure can guide deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=fj6rFciApc.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Q38D6xxrKHe.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=3oQyjABdbC8.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=6iDHce-0B-a.

Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff, 2023b.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Implications: Summary Statistics

- Complex system: $\partial_t x(t) = F(x(t))$.
- 'Macroscopic description': $u : \mathbb{R}^P \to \mathbb{R}^D$ for $D \ll P$ s.t.

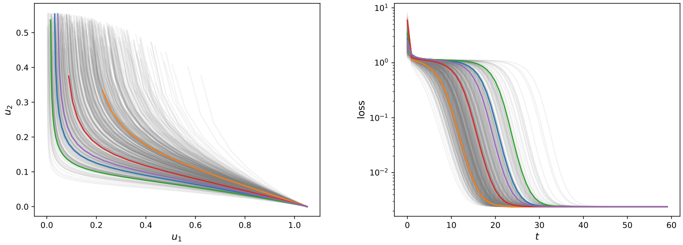
$$\partial_t u(x(t)) \approx G(u(x(t))).$$

Implications: Summary Statistics

- Complex system: $\partial_t x(t) = F(x(t))$.
- 'Macroscopic description': $u : \mathbb{R}^P \to \mathbb{R}^D$ for $D \ll P$ s.t.

$$\partial_t u(x(t)) \approx G(u(x(t))).$$

- Rank 1 Matrix Factorization $\mathcal{L}(\theta) = \left\| ww^T - \theta\theta^T \right\|_F^2$.
  - Invariant under rotation of θ around w.
- Summary statistics [Arous et al., 2022]: $u(\theta) = \left(\left| w^T\theta \right|, \left\| (I - ww^T)\theta \right\| \right)$.



Implications: Summary Statistics

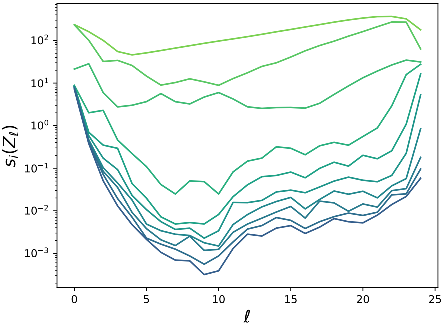
Use a depth $L = 25$ DNN to learn:

$$\theta_0 \mapsto (\mathcal{L}(\theta_0), \mathcal{L}(\theta_1), \ldots, \mathcal{L}(\theta_T))$$

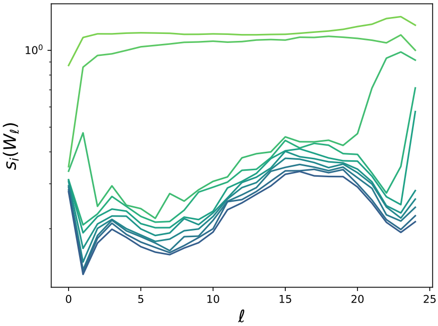
Implications: Summary Statistics

Use a depth $L = 25$ DNN to learn:

$$\theta_0 \mapsto u(\theta_0) \mapsto (\mathcal{L}(\theta_0), \mathcal{L}(\theta_1), \ldots, \mathcal{L}(\theta_T))$$

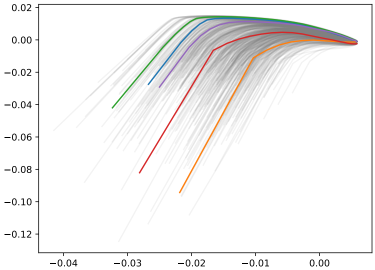


(i) Singular values of $\alpha_\ell(X)$.

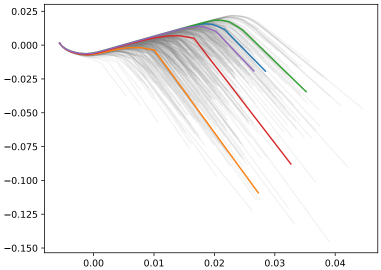


(j) Singular values of W_ℓ.

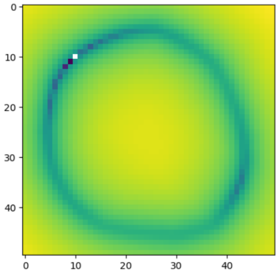
Summary Statistics



(k) PCA of Z_6.



(l) PCA of Z_{15}.



(m) Rotation symmetry at layer $\ell = 2$.