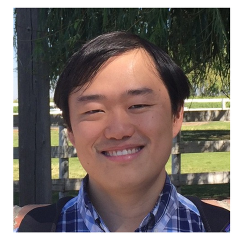
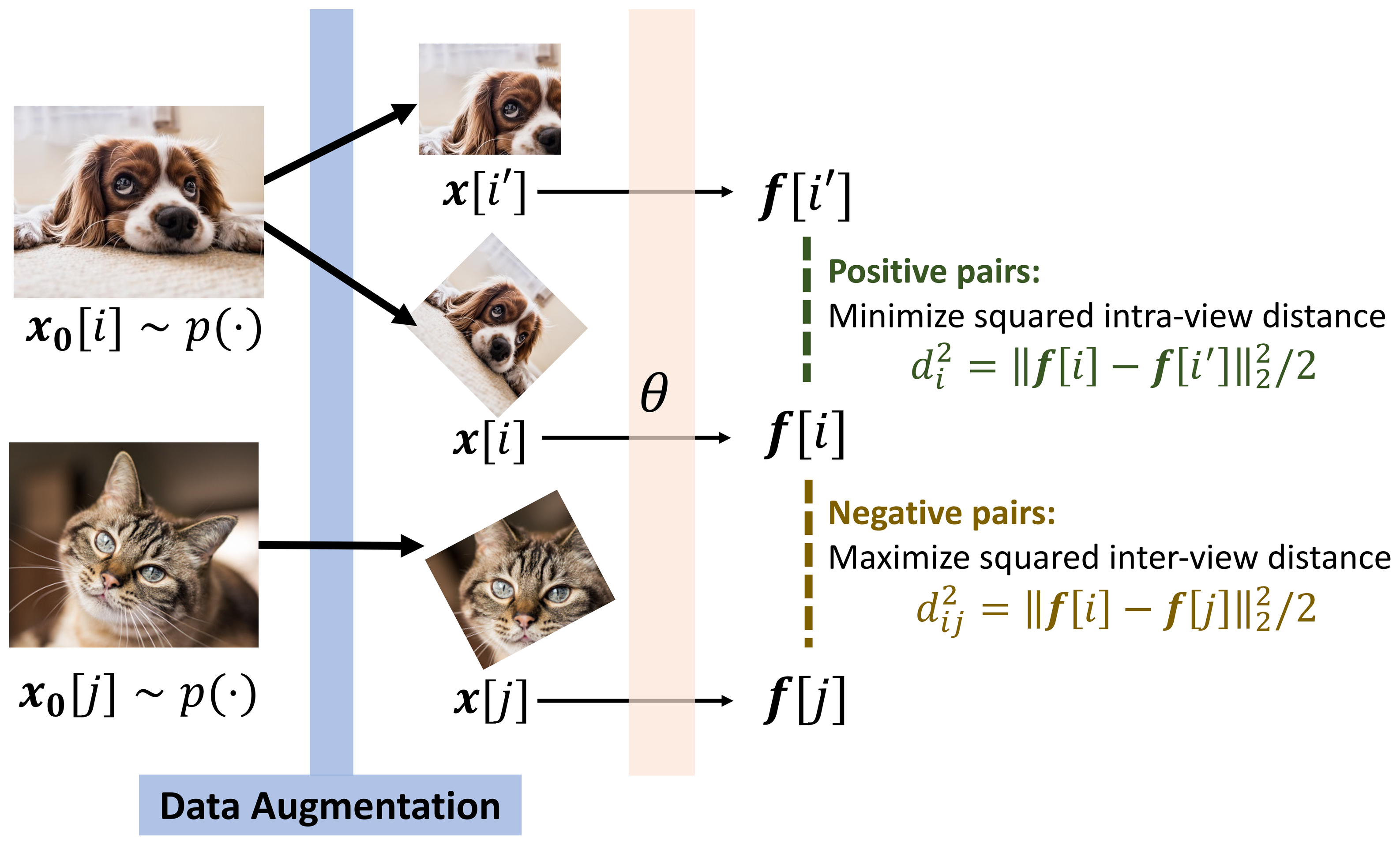
On the Role of Nonlinearity in Training Dynamics of Contrastive Learning on 1-layer Network

Yuandong Tian
yuandong@meta.com



Meta AI

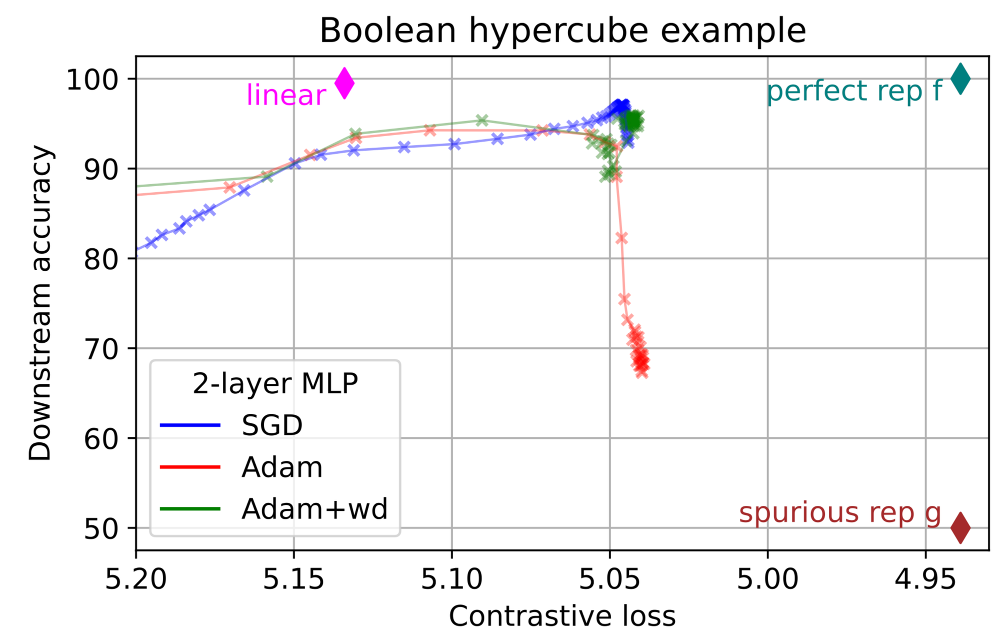
Background



$\boldsymbol{x}[i], \boldsymbol{x}[i'] \sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0[i])$

E.g., InfoNCE loss: $\mathcal{L}_{nce} := -\tau \sum_{i=1}^{N} \log \frac{e^{-d_i^2/\tau}}{\epsilon\, e^{-d_i^2/\tau} + \sum_{j \neq i} e^{-d_{ij}^2/\tau}}$

Is CL just loss + blackbox function family?



[Tables from N. Saunshi et al, Understanding Contrastive Learning Requires Incorporating Inductive Biases, ICML'22]

Inductive bias matters! (e.g., architecture, optimizer)

When the network is linear, CL can be shown to perform like PCA (α-CL)

How to understand the properties of CL with nonlinear network?

α-CL: a unified framework [Tian, NeurIPS'22]

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\text{sg}(\alpha_t)}(\boldsymbol{\theta}_t)$$

Pairwise importance $\alpha_t = \alpha(\boldsymbol{\theta}_t)$

The pairwise importance α can be

1. determined by $\alpha(\boldsymbol{\theta}) = \arg\min_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha(\boldsymbol{\theta}) - \mathcal{R}(\alpha)$, with different regularization $\mathcal{R}(\alpha)$.
2. *directly* specified (**α-CL-direct**)

Define $\mathcal{E}_\alpha$ as the *trace* of ***contrastive covariance*** $\mathbb{C}_\alpha[\cdot]$:

$$\mathcal{E}_\alpha(\boldsymbol{\theta}) := \frac{1}{2} \operatorname{tr} \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]$$

where the *contrastive covariance*

$$\mathbb{C}_\alpha[\boldsymbol{x}] := \frac{1}{2N^2} \sum_{i,j} \alpha_{ij} [(\boldsymbol{x}[i] - \boldsymbol{x}[j])(\boldsymbol{x}[i] - \boldsymbol{x}[j])^T - (\boldsymbol{x}[i] - \boldsymbol{x}[i'])(\boldsymbol{x}[i] - \boldsymbol{x}[i'])^T]$$

Goal: Analyze the local maxima of the energy function $\mathcal{E}_\alpha$

Setup

The Assumptions of Homogenous Activations

We assume the activation satisfies $h(x) = h'(x)x$
This includes Linear, ReLU and monomial activations (with additional constant)

Connect $\mathbb{C}_\alpha[\cdot]$ with regular variance $\mathbb{V}[\cdot]$

If α satisfies $\alpha_{ij} = \mathcal{K}(\boldsymbol{x}_0[i], \boldsymbol{x}_0[j])$, where $\mathcal{K}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{l=0}^{+\infty} \phi_l(\boldsymbol{x})\phi_l(\boldsymbol{y})$

is a kernel, then for any function $\boldsymbol{g}(\cdot)$:

$$\mathbb{C}_\alpha[\boldsymbol{g}(\boldsymbol{x})] \to \sum_{l=0}^{+\infty} z_l^2 \mathbb{V}_{\boldsymbol{x}_0 \sim \tilde{p}_l(\cdot;\alpha)} [\mathbb{E}_{\boldsymbol{x} \sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}(\boldsymbol{x})]]$$

Where $\tilde{p}_l(\boldsymbol{x};\alpha) := \frac{1}{z_l(\alpha)} p_D(\boldsymbol{x})\phi_l(\boldsymbol{x};\alpha)$ is adjusted density of the data, and $z_l(\alpha)$ is the normalization constant.

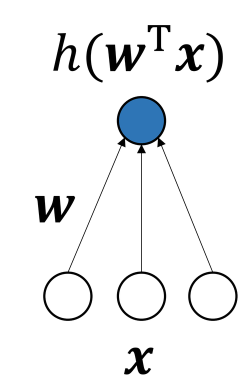
Example of Kernel-like α

Uniform $\alpha_u := 1$

Gaussian $\alpha_g := \exp\left(-\frac{\|h(\boldsymbol{w}^T\boldsymbol{x}_0[i]) - h(\boldsymbol{w}^T\boldsymbol{x}_0[j])\|_2^2}{2\tau}\right)$

1-layer nonlinear network

One node

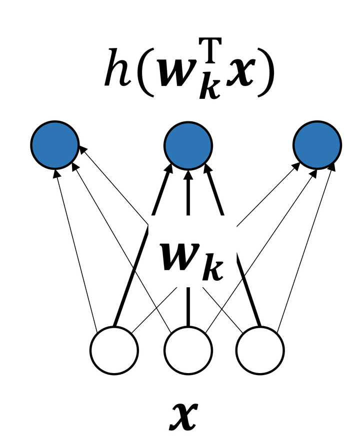


$\max_{\|\boldsymbol{w}\|_2=1} \mathbb{C}_\alpha[f_{\boldsymbol{\theta}}] = \mathbb{C}_\alpha[h(\boldsymbol{w}^T\boldsymbol{x})] = \boldsymbol{w}^T A(\boldsymbol{w})\boldsymbol{w}$

where $A(\boldsymbol{w}) := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]$

$\tilde{\boldsymbol{x}}^{\boldsymbol{w}} := \boldsymbol{x} \cdot h'(\boldsymbol{w}^T\boldsymbol{x})$ is the ***gated*** data point

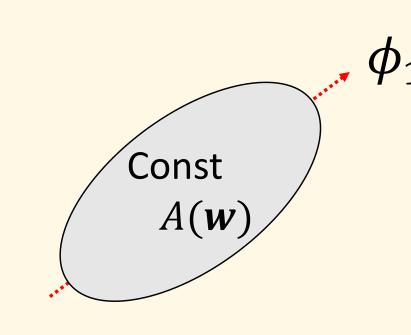
Multiple nodes



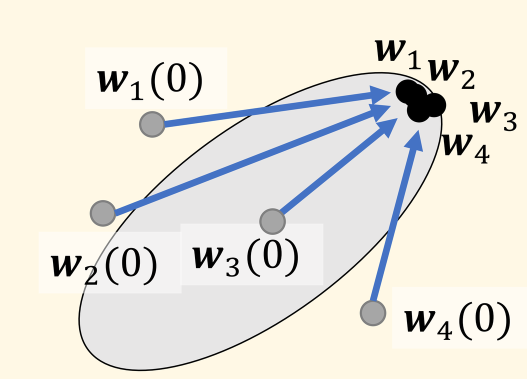
$\max_{\|\boldsymbol{w}_k\|_2=1,\ 1\le k\le K} \operatorname{tr} \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \max_{\|\boldsymbol{w}_k\|_2=1} \boldsymbol{w}_k^T A(\boldsymbol{w}_k)\boldsymbol{w}_k$

Independent one node objective

Linear activation

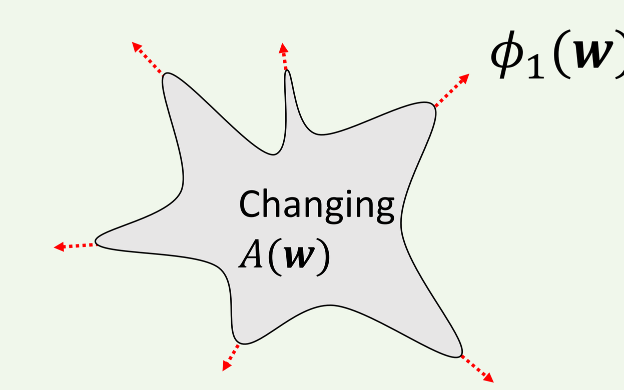


$\phi_1(\boldsymbol{w})$: Largest eigenvector of $A(\boldsymbol{w})$

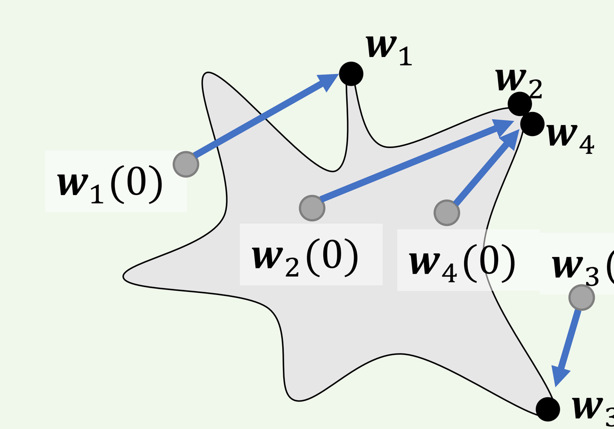


1. All $\boldsymbol{w}_k$ → global maximal eigenvector
2. More nodes do NOT help.

Homogenous nonlinear activation

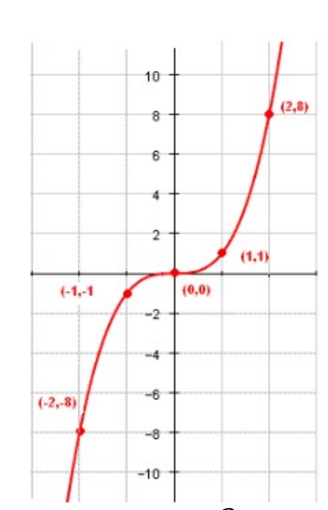


Multiple largest eigenvectors!



1. Each $\boldsymbol{w}_k$ can converge to different patterns
2. More nodes learn more patterns!

Critical Points != Local optima



$y = x^3$

Local roughness $\rho(\boldsymbol{w})$:

$$\|(A(\boldsymbol{v}) - A(\boldsymbol{w}))\boldsymbol{w}\|_2 \le \rho(\boldsymbol{w})\|\boldsymbol{v} - \boldsymbol{w}\|_2 + \mathcal{O}(\|\boldsymbol{v} - \boldsymbol{w}\|_2^2)$$

(for any $\boldsymbol{v}$ in the local neighborhood of $\boldsymbol{w}$)

[**Theorem**] if $A(\boldsymbol{w}_*)\boldsymbol{w}_* = \lambda_*\boldsymbol{w}_*$, and $\lambda_{\text{gap}}(\boldsymbol{w}_*) > \rho(\boldsymbol{w}_*)$, Then $\boldsymbol{w}_*$ is stable (i.e., local maximum)

[**Theorem**] Upper bound of $\rho(\boldsymbol{w})$ in Gaussian α_g $\ll$ Upper bound of $\rho(\boldsymbol{w})$ in Uniform α_u

	CIFAR-10	STL-10
Quadratic loss (uniform α)	73.58 ± 0.82	67.28 ± 1.21
InfoNCE loss (normalized Gaussian α)	87.86 ± 0.12	83.70 ± 0.12