

Lecture 4: Pursuing Low-Dimensional Structures via Lossy Compression

Professor Yi Ma

School of Computing and Data Science
The University of Hong Kong

September 21, 2025

*“Everything should be made as simple as possible,
but not any simpler.”*

– Albert Einstein

Outline

① Objectives for Learning from Data

Precursors and Motivations

Linear and Discriminative Representation (LDR)

② Measure of Information Gain for Representations

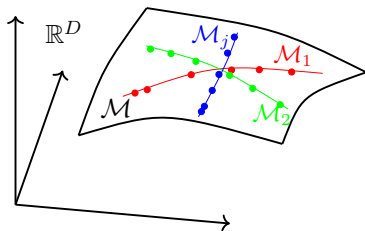
Principle of Maximizing Coding Rate Reduction (MCR²)

Experimental Verification

③ White-Box Deep Networks from Optimizing Rate Reduction

High-Dim Data with Mixed **Nonlinear** Low-Dim Structures

Figure: High-dimensional Real-World Data: data samples $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m]$ in \mathbb{R}^D lying on a mixture of low-dimensional submanifolds $\mathbf{X} \subset \cup_{j=1}^k \mathcal{M}_j \subset \mathbb{R}^D$.



The main objective of learning from (samples of) such real-world data:
seek a most compact and structured representation of the data.

Fitting Class Labels via a Deep Network

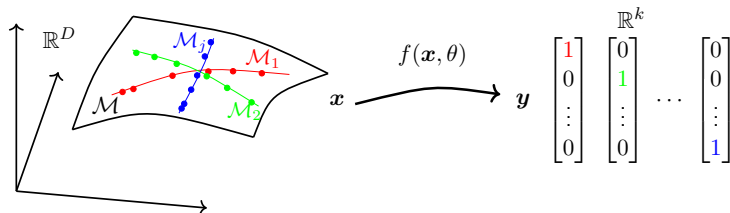


Figure: Black Box DNN for Classification: y is the class label of x represented as a “one-hot” vector in \mathbb{R}^k . To learn a nonlinear mapping $f(\cdot, \theta) : x \mapsto y$, say modeled by a deep network, using cross-entropy (CE) loss.

$$\min_{\theta \in \Theta} \text{CE}(\theta, x, y) \doteq -\mathbb{E}[\langle y, \log[f(x, \theta)] \rangle] \approx -\frac{1}{m} \sum_{i=1}^m \langle y_i, \log[f(x_i, \theta)] \rangle. \quad (1)$$

*Prevalence of **neural collapse** during the terminal phase of deep learning training,*
Papayan, Han, and Donoho, 2020.

Fitting Class Labels via a Deep Network

In a supervised setting, using cross-entropy (CE) loss:

$$\min_{\theta \in \Theta} \text{CE}(\theta, \mathbf{x}, \mathbf{y}) \doteq -\mathbb{E}[\langle \mathbf{y}, \log[f(\mathbf{x}, \theta)] \rangle] \approx -\frac{1}{m} \sum_{i=1}^m \langle \mathbf{y}_i, \log[f(\mathbf{x}_i, \theta)] \rangle. \quad (2)$$

Issues (an elephant in the room):

- A large deep neural networks can **fit arbitrary data and labels**.
- Statistical and geometric meaning of internal features **not clear**.
- Task/data-dependent and **not robust nor truly invariant**.

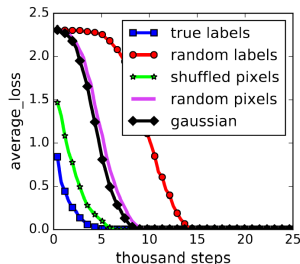


Figure: [Zhang et al, ICLR'17]

What did machines actually “learn” from doing this?

In terms of interpolating, extrapolating, or representing the data?

A Hypothesis: Information Bottleneck

[Tishby & Zaslavsky, 2015]

A feature mapping $f(\mathbf{x}, \theta)$ and a classifier $g(\mathbf{z})$ trained for downstream classification:

$$\mathbf{x} \xrightarrow{f(\mathbf{x}, \theta)} \mathbf{z}(\theta) \xrightarrow{g(\mathbf{z})} \mathbf{y}.$$

The IB Hypothesis: Features learned in a deep network trying to

$$\max_{\theta \in \Theta} \text{IB}(\mathbf{x}, \mathbf{y}, \mathbf{z}(\theta)) \doteq I(\mathbf{z}(\theta), \mathbf{y}) - \beta I(\mathbf{x}, \mathbf{z}(\theta)), \quad \beta > 0, \quad (3)$$

where $I(\mathbf{z}, \mathbf{y}) \doteq H(\mathbf{z}) - H(\mathbf{z}|\mathbf{y})$ and $H(\mathbf{z})$ is the entropy of \mathbf{z} .

- **Minimal** informative features \mathbf{z} that most correlate with the label \mathbf{y}
- Task and label-dependent, consequently sacrificing generalizability, robustness, or transferability

Gap between Theory and Practice (a Bigger Elephant)

For high-dimensional real data,

many statistical and information-theoretic concepts such as entropy, mutual information, K-L divergence, and maximum likelihood:

- curse of **dimensionality** for computation.
- ill-posed for **degenerate** distributions.
- lack guarantees with **finite** (or non-asymptotic) samples.

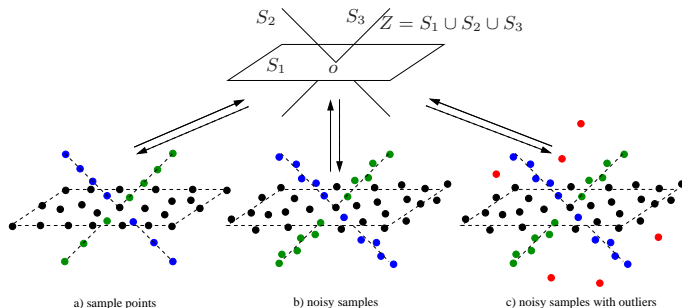
Reality check: principled formulations are replaced with approximate bounds, grossly simplifying assumptions, heuristics, even *ad hoc* tricks and hacks.

How to provide any performance guarantees at all?

A Principled Computational Approach

For high-dim data with **mixed linear** low-dimensional structures:

learn to compress, and compress to learn!

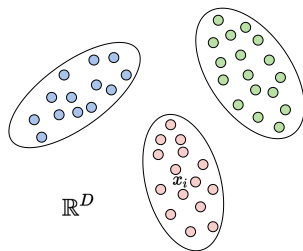


Generalized PCA for mixture of subspaces [Vidal, Ma, and Sastry, 2005]

1. Clustering Mixed Data (Interpolation)

Assume data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]$
are i.i.d. samples from a mixture
of distributions: $p(\mathbf{x}, \theta) = \sum_{j=1}^k \pi_j p_j(\mathbf{x}, \theta)$.

Classic approaches to cluster the data:
the maximum-likelihood (ML) estimate
via Expectation Maximization (EM):



$$\max_{\theta, \pi} \mathbb{E} \left[\log \left(\sum_{j=1}^k \pi_j p_j(\mathbf{x}, \theta) \right) \right] \approx \max_{\theta, \pi} \frac{1}{m} \sum_{i=1}^m \log \left(\sum_{j=1}^k \pi_j p_j(\mathbf{x}_i, \theta) \right).$$

Difficulties: ML is not well-defined when distributions are degenerate.

Clustering via Compression

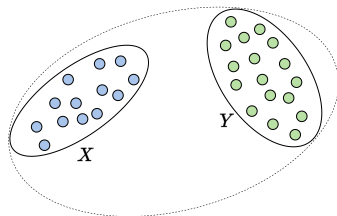
[Yi Ma, Harm Derksen, Wei Hong, and John Wright, TPAMI'07]

A Fundamental Idea:

Data belong to mixed low-dim structures should be compressible.

Cluster Criterion:

Whether the number of binary bits required to store the data is less (information gain):



$$\#bits(\mathbf{X} \cup \mathbf{Y}) \geq \#bits(\mathbf{X}) + \#bits(\mathbf{Y})?$$

"The whole is greater than the sum of the parts."
– Aristotle, 320 BC

Coding Length Function for Subspace-Like Data

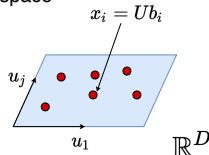
Theorem (Ma, TPAMI'07)

The number of bits needed to encode data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{D \times m}$ up to a precision $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \epsilon$ is bounded by:

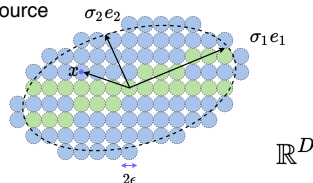
$$L(\mathbf{X}, \epsilon) \doteq \left(\frac{m + D}{2} \right) \log \det \left(\mathbf{I} + \frac{D}{m\epsilon^2} \mathbf{X} \mathbf{X}^\top \right).$$

This can be derived from constructively quantifying SVD of \mathbf{X} or by sphere packing $\text{vol}(\mathbf{X})$ as samples of a noisy Gaussian source.

Linear subspace



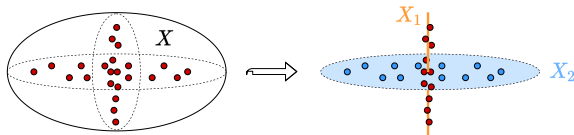
Gaussian source



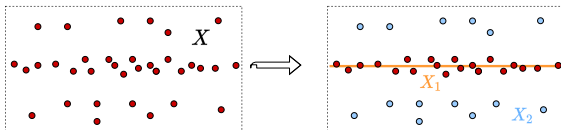
Cluster to Compress

$$L(\mathbf{X}) \geq L^c(\mathbf{X}) \doteq L(\mathbf{X}_1) + L(\mathbf{X}_2) + H(|\mathbf{X}_1|, |\mathbf{X}_2|)?$$

partitioning:



sifting:



A Greedy Algorithm

Seek a partition of the data $\mathbf{X} \rightarrow [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]$ such that

$$\min L^c(\mathbf{X}) \doteq L(\mathbf{X}_1) + \dots + L(\mathbf{X}_k) + H(|\mathbf{X}_1|, \dots, |\mathbf{X}_k|).$$

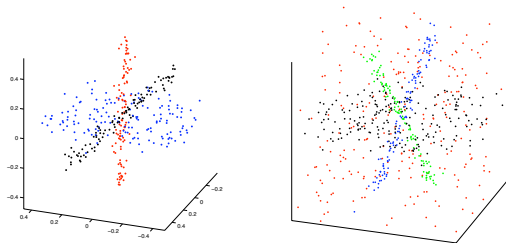
Optimize with a *bottom-up pair-wise* merging algorithm [Ma, TPAMI'07]:

- 1: **input:** the data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{D \times m}$ and a distortion $\epsilon^2 > 0$.
- 2: initialize \mathcal{S} as a set of sets with a single datum $\{S = \{\mathbf{x}\} \mid \mathbf{x} \in \mathbf{X}\}$.
- 3: **while** $|\mathcal{S}| > 1$ **do**
- 4: choose distinct sets $S_1, S_2 \in \mathcal{S}$ such that

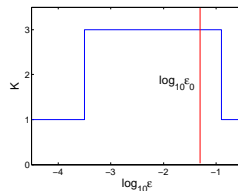
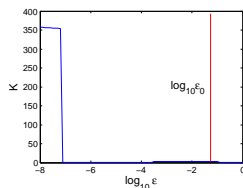
$$L^c(S_1 \cup S_2) - L^c(S_1, S_2)$$
 is minimal.
- 5: **if** $L^c(S_1 \cup S_2) - L^c(S_1, S_2) \geq 0$ **then** break;
- 6: **else** $\mathcal{S} := (\mathcal{S} \setminus \{S_1, S_2\}) \cup \{S_1 \cup S_2\}$.
- 7: **end**
- 8: **output:** \mathcal{S}

Surprisingly Good Performance

Empirically, **find global optimum** and **extremely robust to outliers**



A strikingly sharp **phase transition** w.r.t. quantization ϵ



Natural Image Segmentation [Mobahi et.al., IJCV'09]

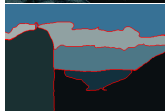
Compression alone, without any supervision, leads to **state of the art** segmentation on natural images (and many other types of data).



(a) Animals



(b) Buildings



(c) Landscape



(d) People



(e) Water

2. Classify Mixed Data (Extrapolation)

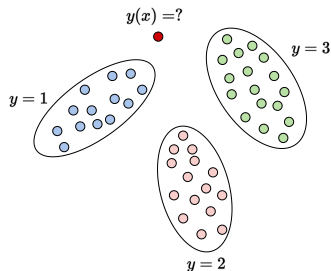
Assume data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]$ are i.i.d. samples from a mixture of distributions: $p(\mathbf{x}, \theta) = \sum_{j=1}^k \pi_j p_j(\mathbf{x}, \theta)$.

Classic approach to classify the data is via maximum a posteriori (MAP) classifier:

$$\hat{y}(\mathbf{x}) = \arg \max_j \log p_j(\mathbf{x}, \theta) + \log \pi_j.$$

Difficulties: distributions p_j are hard to estimate and log likelihood is not well-defined when distributions are degenerate.

(probably why SVMs or deep networks prevail instead...)



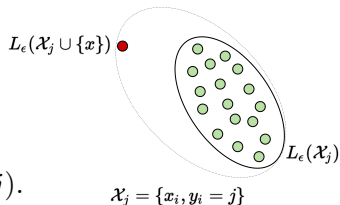
Classify to Compress

[Wright, Tao, Lin, Shum, and Ma, NIPS'07]

A Fundamental Idea:

Count additional #bits needed to encode a query sample \mathbf{x} with data in each class \mathbf{X}_j :

$$\delta L_\epsilon(\mathbf{x}, j) \doteq L_\epsilon(\mathbf{X}_j \cup \{\mathbf{x}\}) - L_\epsilon(\mathbf{X}_j) + L(j).$$



Classification Criterion: Minimum Incremental Coding Length (MICL):

$$\hat{y}(\mathbf{x}) = \arg \min_j \delta L_\epsilon(\mathbf{x}, j).$$

Law of Parsimony: *"Entities should not be multiplied without necessity."*
–William of Ockham

Asymptotic Property of MICL

Theorem (Wright, NIPS'07)

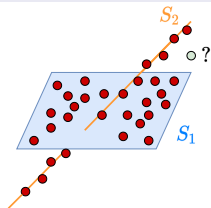
As the number of samples m goes to infinity, the MICL criterion converges at a rate of $O(m^{-1/2})$ to the following criterion:

$$\hat{y}_\epsilon(\mathbf{x}) = \arg \max_j \underbrace{L_G\left(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j + \frac{\epsilon^2}{D} \mathbf{I}\right) + \log \pi_j + \frac{1}{2} D_\epsilon(\boldsymbol{\Sigma}_j)}_{\text{Regularized MAP}},$$

where $D_\epsilon(\boldsymbol{\Sigma}_j) \doteq \text{tr}\left(\boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}_j + \frac{\epsilon^2}{D} \mathbf{I})^{-1}\right)$ is the effective dimension.

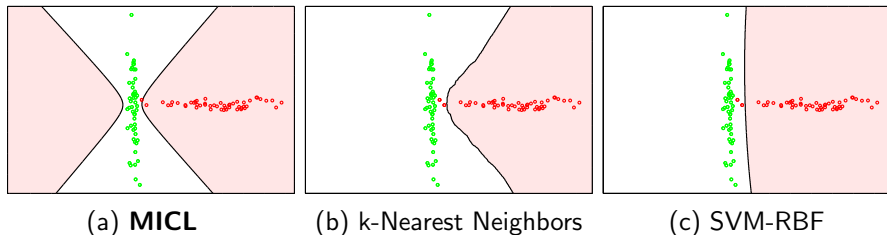
Everything else equal, MICL prefers a class with higher effective dimension.

Err on the side of caution!



Extrapolation of Low-Dim Structure for Classification

Figure: A truly extrapolating (nearest subspace) classifier!



Difficulty in practice: inference computationally costly (non-parametric) and possibly need a kernel (nonlinearity).

Go beyond (non-parametric) data interpolation and extrapolation?

Represent Multi-class Multi-dimensional Data

Given samples

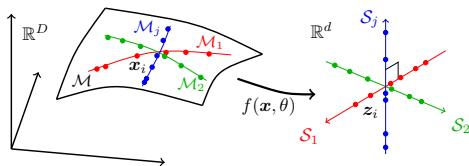
$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m] \subset \cup_{j=1}^k \mathcal{M}_j,$$

seek a good representation

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_m] \subset \mathbb{R}^d$$

through a continuous mapping:

$$f(\mathbf{x}, \theta) : \mathbf{x} \in \mathbb{R}^D \mapsto \mathbf{z} \in \mathbb{R}^d.$$



Goals of “**re-present**” the data:

- **compression**: from high-dimensional samples to compact features.
- **linearization**: from nonlinear structures $\cup_{j=1}^k \mathcal{M}_j$ to linear $\cup_{j=1}^k \mathcal{S}_j$.
- **sparsity**: from separable components \mathcal{M}_j 's to incoherent \mathcal{S}_j 's.
- **self-consistent**: from compact structured \mathbf{Z} back to the data \mathbf{X} .

Seeking a Linear Discriminative Representation (LDR)

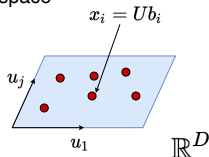
Desiderata: Representation $z = f(x, \theta)$ have the following properties:

- ① *Within-Class Compressible:* Features of the same class/cluster should be highly compressed in a **low-dimensional** linear subspace.
- ② *Between-Class Discriminative:* Features of different classes/clusters should be in highly **incoherent** linear subspaces.
- ③ *Maximally Informative:* Dimension (or variance) of features for each class/cluster should be **the same as that of the data**.

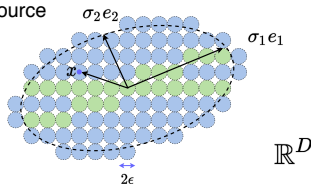
Is there a principled objective for all such properties, together?

Compactness Measure for Linear/Gaussian Representation

Linear subspace



Gaussian source



Theorem (Coding Length, Ma & Derksen TPAMI'07)

The number of bits needed to encode data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{D \times m}$ up to a precision $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \epsilon$ is bounded by:

$$L(\mathbf{X}, \epsilon) \doteq \left(\frac{m + D}{2} \right) \log \det \left(\mathbf{I} + \frac{D}{m\epsilon^2} \mathbf{X} \mathbf{X}^\top \right).$$

This can be derived from constructively quantifying SVD of \mathbf{X} or by sphere packing $\text{vol}(\mathbf{X})$ as samples of a noisy Gaussian source.

Compactness Measure for Linear/Gaussian Representation

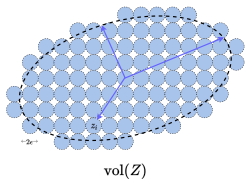
If \mathbf{X} is not (piecewise) linear or Gaussian, consider a **nonlinear** mapping:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{D \times m} \xrightarrow{f(\mathbf{x}, \theta)} \mathbf{Z}(\theta) = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m] \in \mathbb{R}^{d \times m}.$$

The average coding length per sample (rate) subject to a distortion ϵ :

$$R(\mathbf{Z}, \epsilon) \doteq \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{m\epsilon^2} \mathbf{Z} \mathbf{Z}^\top \right). \quad (4)$$

Rate distortion is an intrinsic measure for the volume of all features.



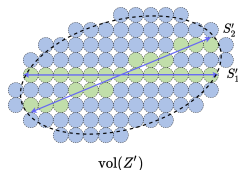
Compactness Measure for Mixed Linear Representations

The features \mathbf{Z} of **multi-class** data

$$\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2 \cup \cdots \cup \mathbf{X}_k \subset \cup_{j=1}^k \mathcal{M}_j.$$

may be partitioned into **multiple** subsets:

$$\mathbf{Z} = \mathbf{Z}_1 \cup \mathbf{Z}_2 \cup \cdots \cup \mathbf{Z}_k \subset \cup_{j=1}^k \mathcal{S}_j.$$



W.r.t. this partition, the **average coding rate** is:

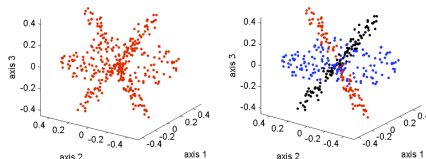
$$R^c(\mathbf{Z}, \epsilon \mid \mathbf{\Pi}) \doteq \sum_{j=1}^k \frac{\text{tr}(\mathbf{\Pi}_j)}{2m} \log \det \left(\mathbf{I} + \frac{d}{\text{tr}(\mathbf{\Pi}_j) \epsilon^2} \mathbf{Z} \mathbf{\Pi}_j \mathbf{Z}^\top \right), \quad (5)$$

where $\mathbf{\Pi} = \{\mathbf{\Pi}_j \in \mathbb{R}^{m \times m}\}_{j=1}^k$ encode the membership of the m samples in the k classes: the diagonal entry $\mathbf{\Pi}_j(i, i)$ of $\mathbf{\Pi}_j$ is the probability of sample i belonging to subset j . $\Omega \doteq \{\mathbf{\Pi} \mid \sum \mathbf{\Pi}_j = \mathbf{I}, \mathbf{\Pi}_j \geq \mathbf{0}\}$

Parsimony: Clustering by Minimizing Coding Rate/Length

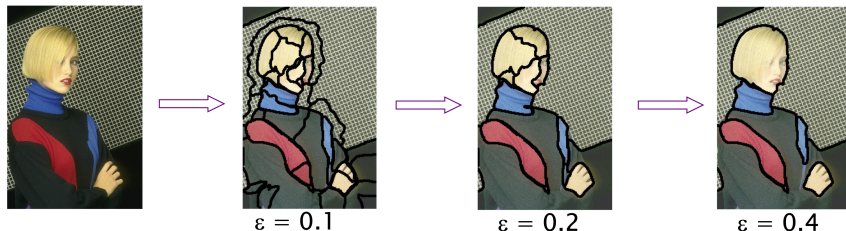
Segmentation of Multivariate Mixed Data via Lossy Coding and Compression,

Yi Ma et. al., TPAMI, 2007.



$$\min_{\Pi} R^c(\mathbf{Z} \mid \Pi, \epsilon) = \sum_{j=1}^k \frac{\text{tr}(\Pi_j)}{2m} \log \det \left(\mathbf{I} + \frac{d}{\text{tr}(\Pi_j) \epsilon^2} \mathbf{Z} \Pi_j \mathbf{Z}^\top \right).$$

State of the art unsupervised image segmentation (IJCV 2011):



Measure for Linear Discriminative Representation (LDR)

A fundamental idea: maximize the **difference** between the coding rate of all features and the average rate of features in each of the classes:

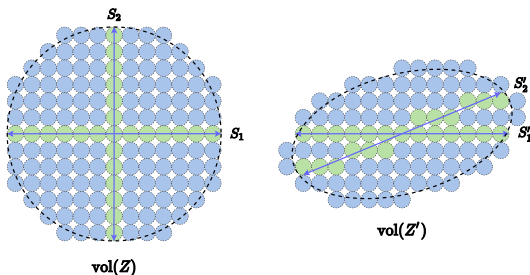
$$\Delta R(\mathbf{Z}, \mathbf{\Pi}, \epsilon) = \underbrace{\frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{m\epsilon^2} \mathbf{Z} \mathbf{Z}^\top \right)}_R - \underbrace{\sum_{j=1}^k \frac{\text{tr}(\mathbf{\Pi}_j)}{2m} \log \det \left(\mathbf{I} + \frac{d}{\text{tr}(\mathbf{\Pi}_j)\epsilon^2} \mathbf{Z} \mathbf{\Pi}_j \mathbf{Z}^\top \right)}_{R^c}.$$

This difference is called **rate reduction** (measuring **information gain**):

- Large R : **expand** all features \mathbf{Z} as **large** as possible.
- Small R^c : **compress** each class \mathbf{Z}_j as **small** as possible.

Slogan: similarity contracts and dissimilarity contrasts!

Interpretation of MCR^2 : Sphere Packing and Counting



Example: two subspaces S_1 and S_2 in \mathbb{R}^2 .

- $\log \#(\text{green spheres} + \text{blue spheres}) = \text{rate of span of all samples } R.$
- $\log \#(\text{green spheres}) = \text{rate of the two subspaces } R^c.$
- $\log \#(\text{blue spheres}) = \text{rate reduction } \Delta R.$

Comparison to Contrastive Learning

[Hadsell, Chopra, and LeCun, CVPR'06]

When k is large, a randomly chosen **pair** (x_i, x_j) is of high probability belonging to different classes. Minimize the **contrastive loss**:

$$\min -\log \frac{\exp(\langle z_i, z'_i \rangle)}{\sum_{j \neq i} \exp(\langle z_i, z_j \rangle)}.$$

The learned features of such pairs of samples together with their augmentations \mathbf{Z}_i and \mathbf{Z}_j should have large rate reduction:

$$\max_{ij} \sum \Delta R_{ij} \doteq R(\mathbf{Z}_i \cup \mathbf{Z}_j, \epsilon) - \frac{1}{2}(R(\mathbf{Z}_i, \epsilon) + R(\mathbf{Z}_j, \epsilon)).$$

MCR² contrasts triplets, quadruplets, or any number of sets.

Principle of Maximal Coding Rate Reduction (MCR²)

[Yu, Chan, You, Song, Ma, NeurIPS2020]

Learn a mapping $f(\mathbf{x}, \theta)$ (for a given partition Π):

$$\mathbf{X} \xrightarrow{f(\mathbf{x}, \theta)} \mathbf{Z}(\theta) \xrightarrow{\Pi, \epsilon} \Delta R(\mathbf{Z}(\theta), \Pi, \epsilon) \quad (6)$$

so as to **Maximize the Coding Rate Reduction (MCR²)**:

$$\begin{aligned} \max_{\theta} \quad & \Delta R(\mathbf{Z}(\theta), \Pi, \epsilon) = R(\mathbf{Z}(\theta), \epsilon) - R^c(\mathbf{Z}(\theta), \epsilon \mid \Pi), \\ \text{subject to} \quad & \|\mathbf{Z}_j(\theta)\|_F^2 = m_j, \Pi \in \Omega. \end{aligned} \quad (7)$$

Since ΔR is *monotonic* in the scale of \mathbf{Z} , one needs to:

normalize the features $\mathbf{z} = f(\mathbf{x}, \theta)$ so as to compare $\mathbf{Z}(\theta)$ and $\mathbf{Z}(\theta')$!

Batch normalization, Sergey Ioffe and Christian Szegedy, 2015.

Layer normalization'16, instance normalization'16; group normalization'18...

Theoretical Justification of the MCR² Principle

Theorem (Informal Statement [Yu et.al., NeurIPS2020])

Suppose $\mathbf{Z}^ = \mathbf{Z}_1^* \cup \dots \cup \mathbf{Z}_k^*$ is the optimal solution that maximizes the rate reduction (7). We have:*

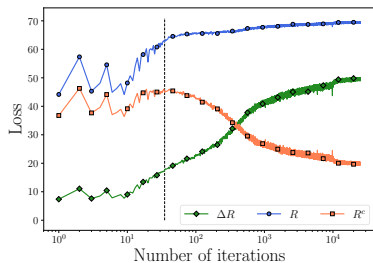
- *Between-class Discriminative: As long as the ambient space is adequately large ($d \geq \sum_{j=1}^k d_j$), the subspaces are all orthogonal to each other, i.e. $(\mathbf{Z}_i^*)^\top \mathbf{Z}_j^* = \mathbf{0}$ for $i \neq j$.*
- *Maximally Informative Representation: As long as the coding precision is adequately high, i.e., $\epsilon^4 < \min_j \left\{ \frac{m_j}{m} \frac{d^2}{d_j^2} \right\}$, each subspace achieves its maximal dimension, i.e. $\text{rank}(\mathbf{Z}_j^*) = d_j$. In addition, the largest $d_j - 1$ singular values of \mathbf{Z}_j^* are equal.*

A new slogan, beyond Aristotle:

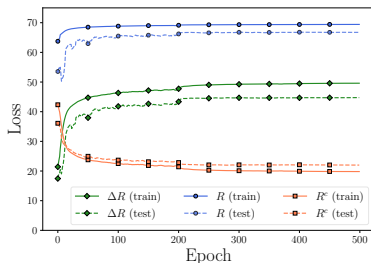
The whole is to be maximally greater than the sum of the parts!

Experiment I: Supervised Deep Learning

Experimental Setup: Train $f(x, \theta)$ as ResNet18 on the CIFAR10 dataset, feature z dimension $d = 128$, precision $\epsilon^2 = 0.5$.



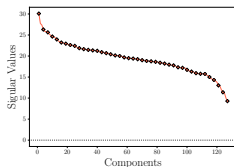
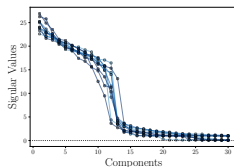
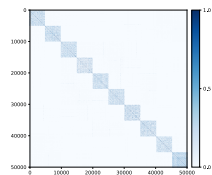
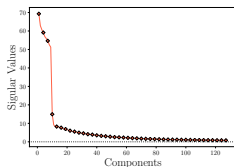
(a)



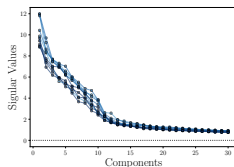
(b)

Figure: (a). Evolution of $R, R^c, \Delta R$ during the training process; (b). Training loss versus testing loss.

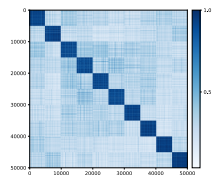
Visualization of Learned Representations \mathbb{Z}

(a) MCR^2 (overall)(b) MCR^2 (PCA of every class)(c) MCR^2 (cosine similarity)

(d) CE (overall)



(e) CE (PCA of every class)

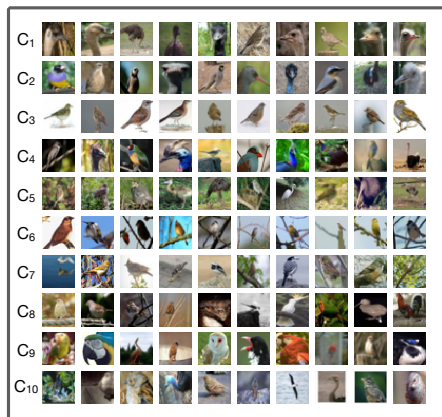


(f) CE (cosine similarity)

Figure: PCA of learned representations from MCR^2 and cross-entropy.

No neural collapse!

Visualization - Samples along Principal Components



(a) Bird



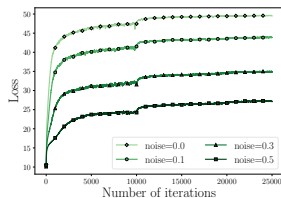
(b) Ship

Figure: Top-10 “principal” images for class - “Bird” and “Ship” in the CIFAR10.

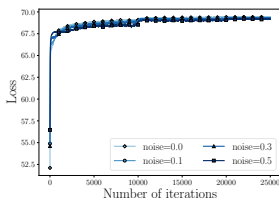
Experiment II: Robustness to Label Noise

	RATIO=0.0	RATIO=0.1	RATIO=0.2	RATIO=0.3	RATIO=0.4	RATIO=0.5
CE TRAINING	0.939	0.909	0.861	0.791	0.724	0.603
MCR ² TRAINING	0.940	0.911	0.897	0.881	0.866	0.843

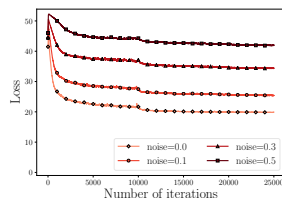
Table 1: Classification results with features learned with labels corrupted at different levels.



(a) $\Delta R(\mathbf{Z}(\theta), \Pi, \epsilon)$



(b) $R(\mathbf{Z}(\theta), \epsilon)$



(c) $R^c(\mathbf{Z}(\theta), \epsilon | \Pi)$

Figure: Evolution of $R, R^c, \Delta R$ of MCR² during training with corrupted labels.

Represent only what can be jointly compressed.

Deep Networks from Optimizing Rate Reduction

$$\mathbf{X} \xrightarrow{f(\mathbf{x}, \theta)} \mathbf{Z}(\theta); \quad \max_{\theta} \Delta R(\mathbf{Z}(\theta), \mathbf{\Pi}, \epsilon).$$

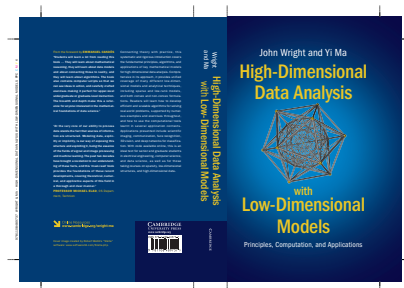
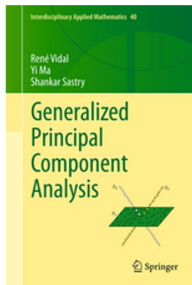
Final features learned by MCR² are more interpretable and robust, **but**:

- The borrowed deep network (e.g. ResNet) is still a “black box”!
- Why is a “deep” architecture necessary, and how wide and deep?
- What are the roles of the “linear and nonlinear” operators?
- Why “multi-channel” convolutions?
- ...

Replace black box networks with entirely “white box” networks?

References

- ① **Clustering** via Lossy Coding and Compression (TPAMI 2007):
<http://people.eecs.berkeley.edu/~yima/psfile/Ma-PAMI07.pdf>
- ② **Classification** via Minimal Incremental Coding Length (NeurIPS 2007):
http://people.eecs.berkeley.edu/~yima/psfile/MICL_SJIS.pdf
- ③ **Representation** via Maximal Coding Rate Reduction (NeurIPS 2020):
<https://arxiv.org/abs/2006.08558>



Lecture 4: Pursuing Low-Dimensional Structures via Lossy Compression

Thank you!
Questions, please?

"We compress to learn, and we learn to compress."
– John Wright & Yi Ma



SIMONS
FOUNDATION