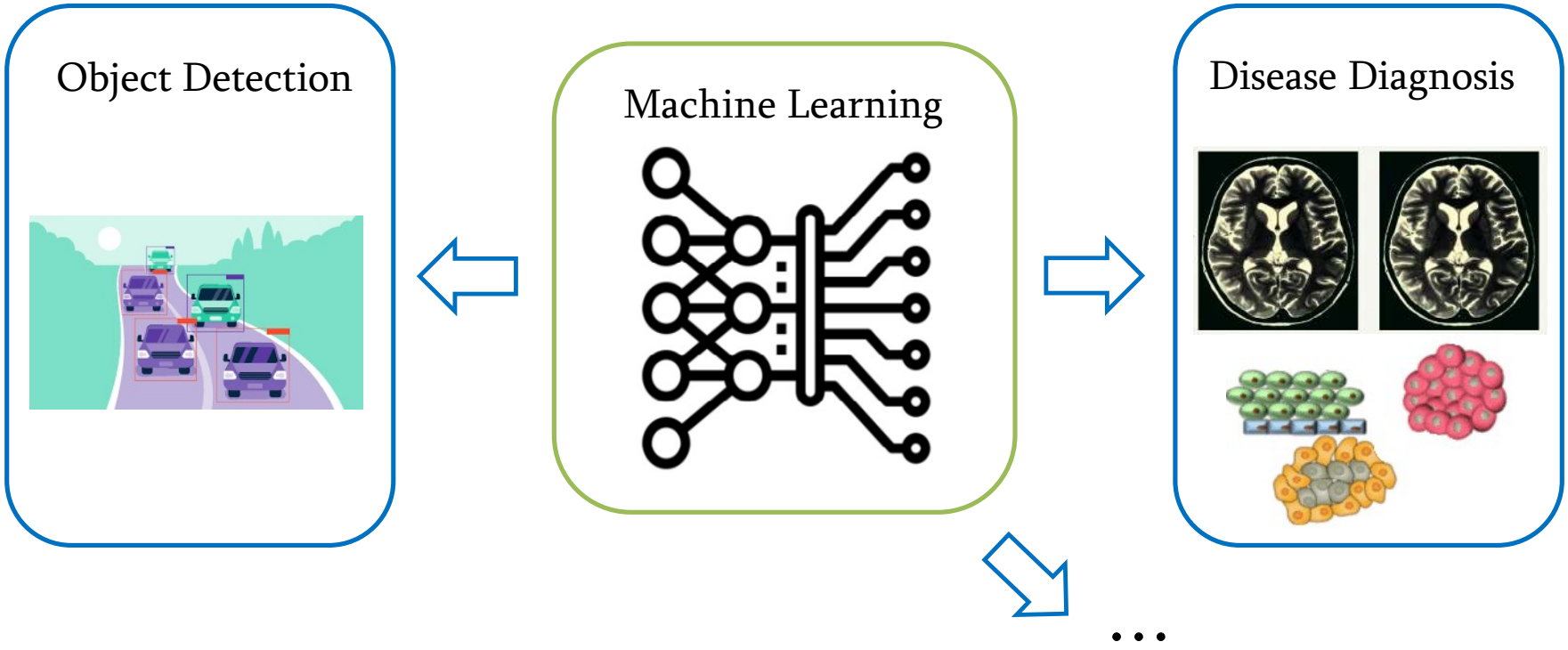
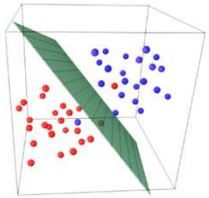
A Generalized Geometric Theoretical Framework of Centroid Discriminant Analysis for Linear Classification of Multi-dimensional Data

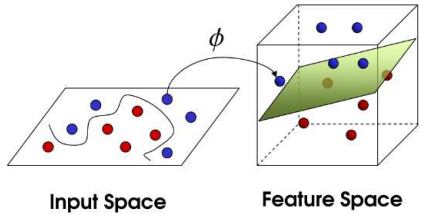
Yue Wu, Jialin Zhao, Carlo Vittorio Cannistraci
Center of Complex Network Intelligence (CCNI), Tsinghua University

§ 1. Background



§ 1. Background on classification





Aspect	Linear Classifiers	Nonlinear Kernel-based Classifiers	Other Nonlinear Methods
Performance	Good on linearly separable data	High accuracy on complex patterns	High potential if tuned well
Scalability	Good but usually inverse to performance	Poor (quadratic or cubic in N)	Varies (trees scale relatively well, deep nets need GPU)
Overfitting	Low (if regularized properly)	Moderate to high (if not regularized)	High risk, especially with deep nets
Explainability	High (easy to interpret weights)	Low (complex implicit decision surface)	Varies (trees are interpretable, nets are not)

Picture from *https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f*

§ 1. Introduction of linear classifiers

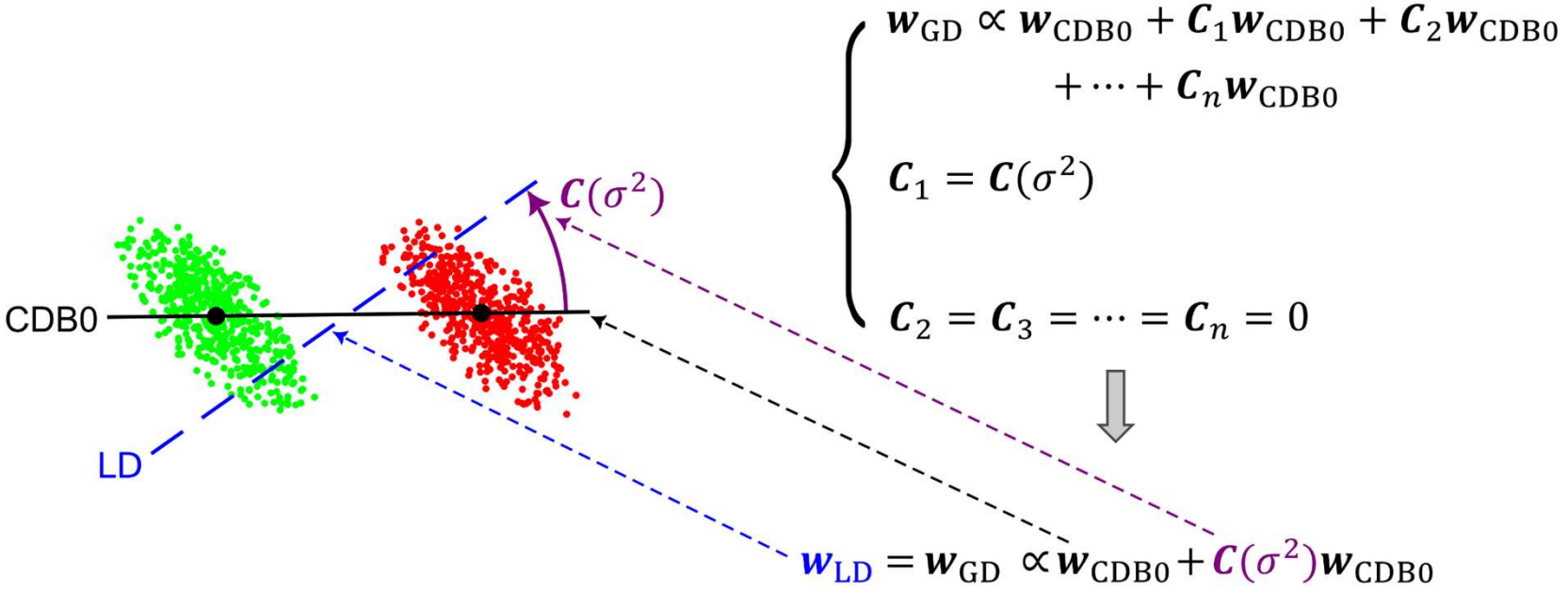
Principle	Method	Training time complexity
Variance	LDA	$O(NM^2 + M^3)$
	Fast LDA (SRDA)	$O(kNs)$ with conditions
Maximum-margin	SVM	$O(N^3)$
	Fast SVM (LIBLINEAR)	$O(kNM)$
maximal likelihood	Logistic Regression	$O(NM^2)$
Minimum-entropy	Perceptron	$O(kNM)$
Prototype	NMC	$O(NM)$

N – number of samples
M – number of features
k – number of iterations
s – average number of non-zero features in each sample

NMC: Nearest Mean Classifier
LDA: Linear Discriminant Analysis
SVM: Support Vector Machine

§ 2. Geometric Discriminant Analysis

Assumption: Centroids offer important information for discriminative power



GD: Geometric Discriminant
CDB0: initial Centroid Discriminant Basis
LD: Linear Discriminant

§ 2. Geometric Discriminant Analysis

General case

LDA maximizes: $S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right)^2}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}}$

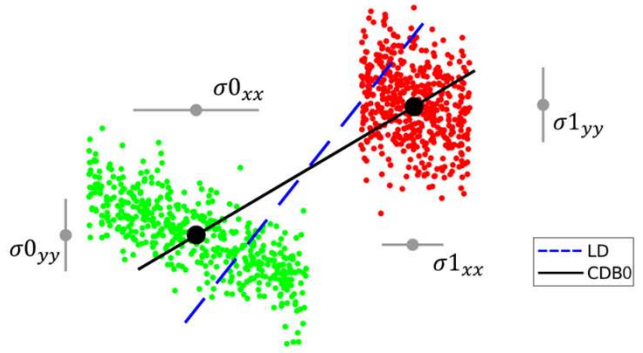
$$\boldsymbol{w}_{\text{LD}} \propto \left(\begin{bmatrix} 1 & \\ & 1 \end{bmatrix} + \begin{bmatrix} 0 & -\frac{\sigma^2_{xy}}{\sigma^2_{yy}} \\ -\frac{\sigma^2_{yx}}{\sigma^2_{yy}} & \frac{\sigma^2_{xx}}{\sigma^2_{yy}} - 1 \end{bmatrix} \right) \begin{bmatrix} \Delta\mu_x \\ \Delta\mu_y \end{bmatrix}$$

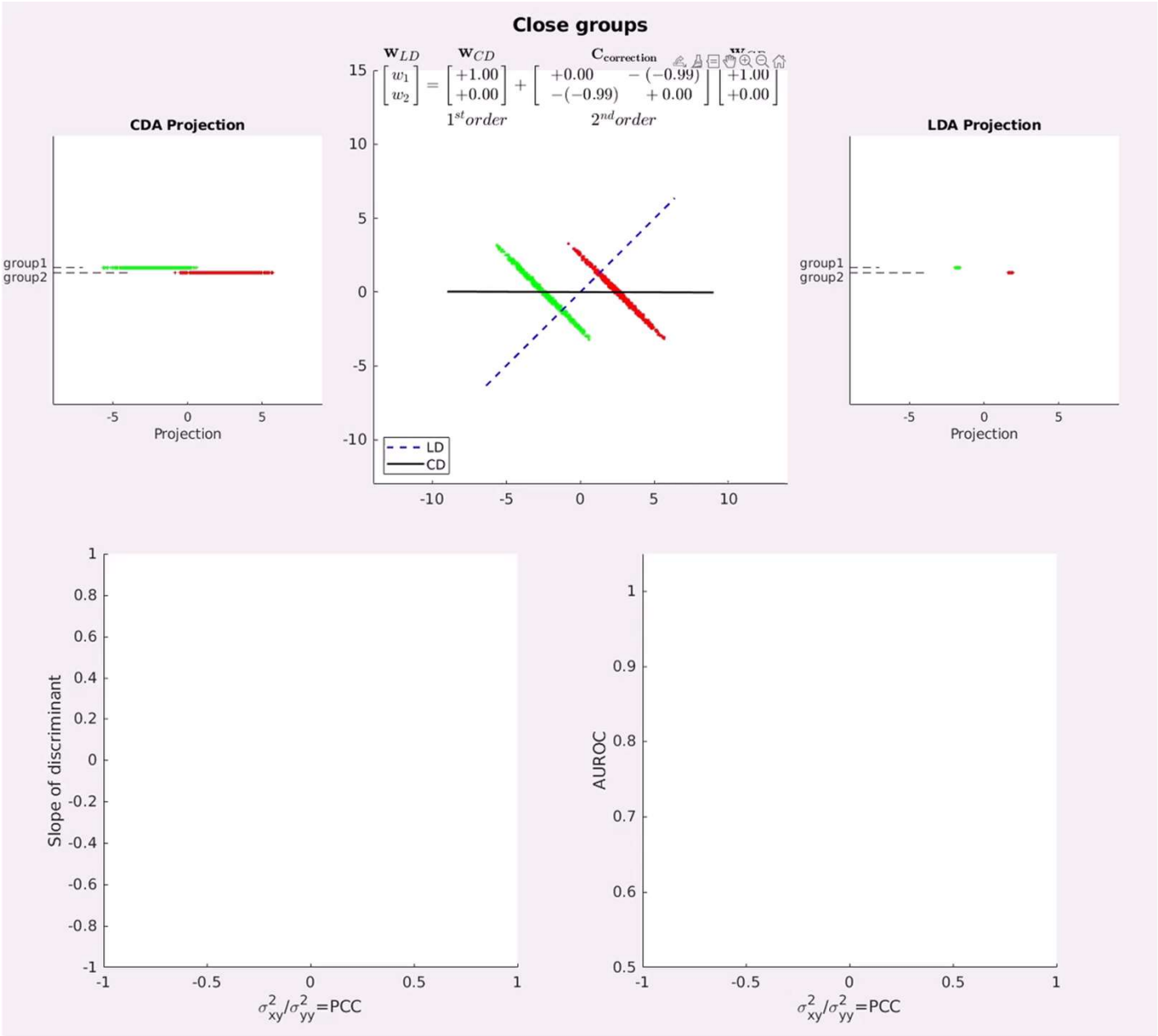
$$= \boldsymbol{w}_{\text{CDB0}} + \boldsymbol{C}_{\text{correction}}\, \boldsymbol{w}_{\text{CDB0}}$$

$$\underset{}{\boldsymbol{w}_{\text{LD}}} \quad \underset{}{\boldsymbol{w}_{\text{CDB0}}} \qquad \underset{}{\boldsymbol{C}_{\text{correction}}} \qquad \underset{}{\boldsymbol{w}_{\text{CDB0}}}$$

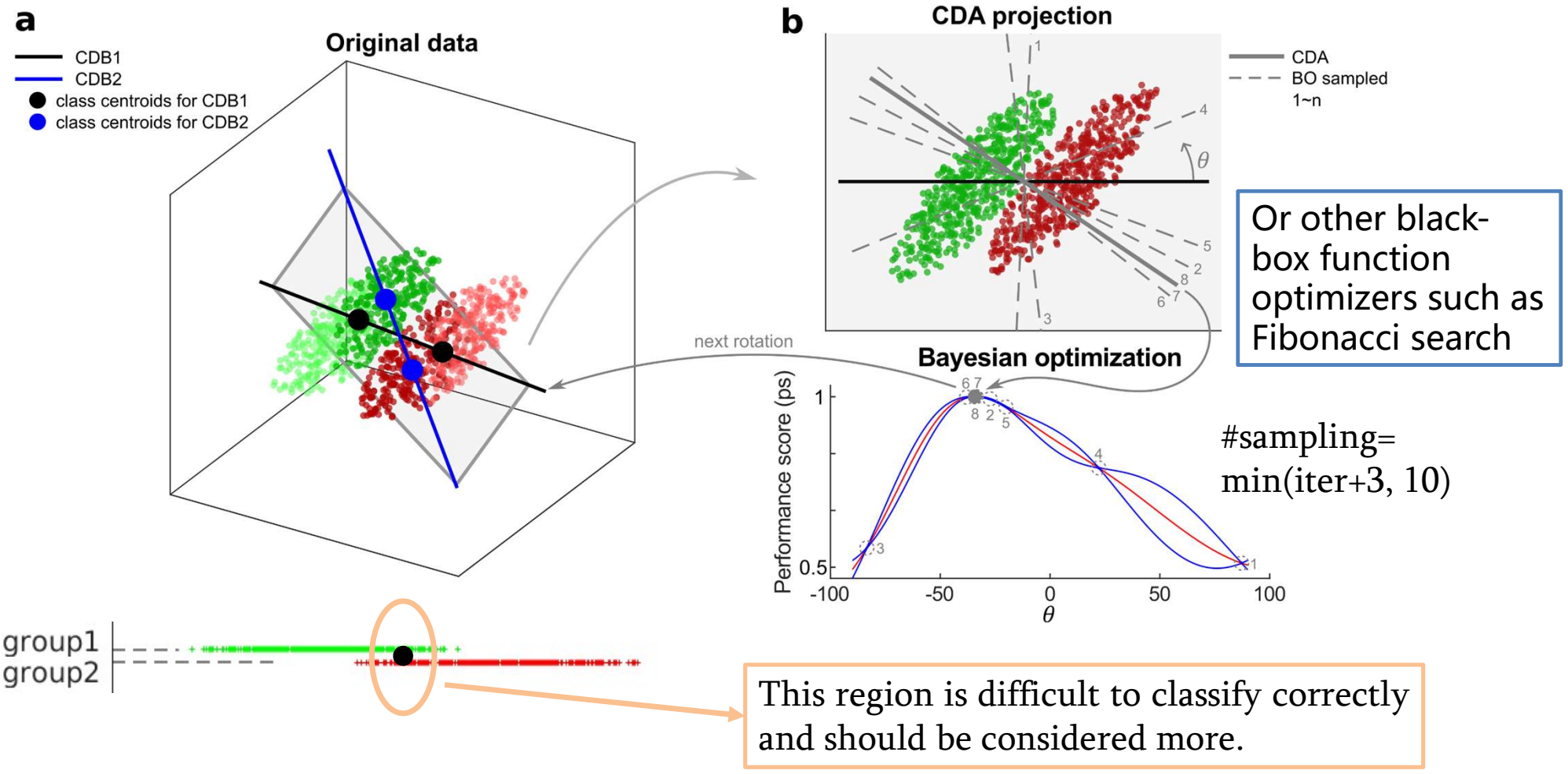
$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \propto \begin{bmatrix} +0.86 \\ +0.51 \end{bmatrix} + \begin{bmatrix} 0 & -(-0.73) \\ -(-0.73) & 0.8 \end{bmatrix} \begin{bmatrix} +0.86 \\ +0.51 \end{bmatrix}$$

$$1^{st} order \qquad 2^{nd} order$$

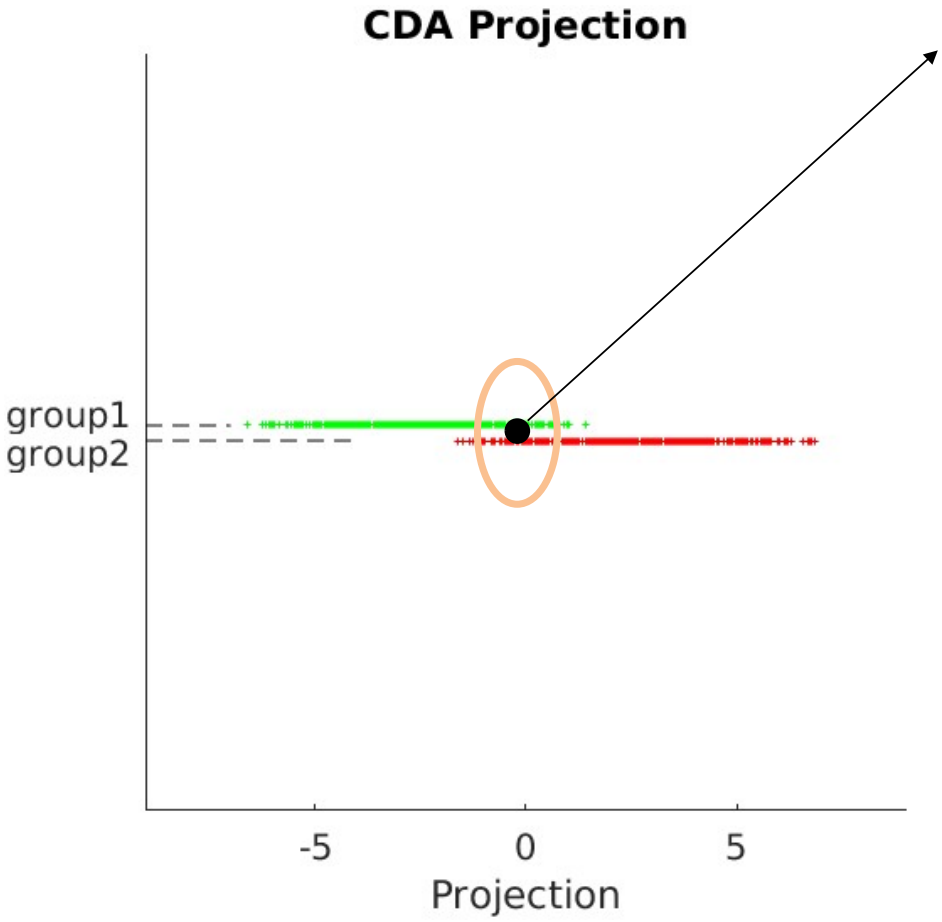



Close groups
$\mathbf{w}_{LD}$ $\mathbf{w}_{CD}$ $\mathbf{C}_{\text{correction}}$
$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} +1.00 \\ +0.00 \end{bmatrix} + \begin{bmatrix} +0.00 & -(-0.99) \\ -(-0.99) & +0.00 \end{bmatrix} \begin{bmatrix} +1.00 \\ +0.00 \end{bmatrix}$
$1^{st} order$ $2^{nd} order$
CDA Projection
group1
group2
Projection
LD
CD
LDA Projection
group1
group2
Projection
Slope of discriminant
$\sigma^2_{xy}/\sigma^2_{yy}$=PCC
AUROC
$\sigma^2_{xy}/\sigma^2_{yy}$=PCC

§ 3 Centroid Discriminant Analysis



§ 3 Centroid Discriminant Analysis



The operating point is sought to maximize:

$$performance\ score = \left(Fscore^{(\mathrm{pos})} + Fscore^{(\mathrm{neg})} + ACscore\right)/3$$

Sample weights update

Initialize sample weights by : $\boldsymbol{\alpha} \leftarrow [1; 1; \ldots; 1]^{N\times 1}/\sqrt{N}$

In each CDA rotation, find the 2nd line to create a 2D plane, by new sample weights:

$$d_i \leftarrow |q_i - oop| \quad \forall i \in \{1,2,\ldots,N\}$$

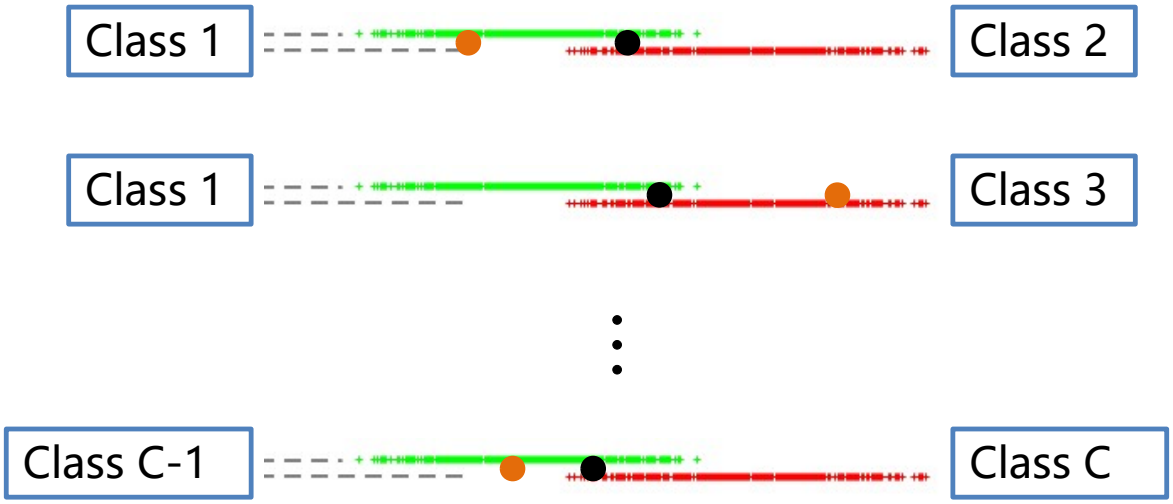
$$\boldsymbol{d}_{\mathrm{r}} \leftarrow |\boldsymbol{d} - \min(\boldsymbol{d}) - \max(\boldsymbol{d})|$$

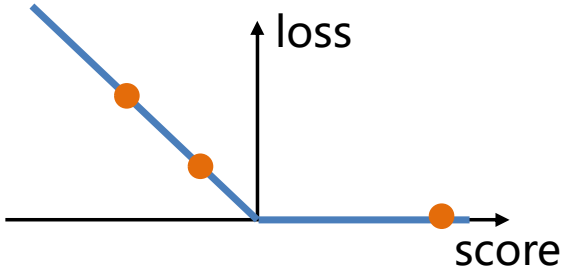
$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} \odot \boldsymbol{d}_{\mathrm{r}}/\|\boldsymbol{\alpha} \odot \boldsymbol{d}_{\mathrm{r}}\|_2$$

§ 3 Centroid Discriminant Analysis

Multiclass prediction via ECOC

To predict a new sample





$$\boldsymbol{CodingMatrix} = \begin{matrix} 1 & 1 & \dots & 0 \\ -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & 1 \\ 0 & 0 & \dots & -1 \end{matrix}$$

$$\boldsymbol{Loss} = \boldsymbol{CodingMatrix} \cdot \boldsymbol{q}$$

$$\hat{y} = \underset{c}{\operatorname{argmin}} \boldsymbol{Loss}_c$$

§ 3 Centroid Discriminant Analysis

27 Datasets

	Dataset	N	M	C	Balancedness	Modality/source	Classification task
Standard images	MNIST	70000	400	10	imbalanced	image	digits
	USPS	9298	256	10	imbalanced		digits
	EMNIST	145600	784	26	balanced		letters
	CIFAR10	60000	3072	10	balanced		objects
	SVHN	99289	3072	10	imbalanced		house numbers
	flower	3670	1200	5	imbalanced		flowers
	GTSRB	26635	1200	43	imbalanced		traffic signs
	STL10	13000	2352	10	balanced		objects
	FMNIST	70000	784	10	balanced		fashion objects
Chemical formula	bace	1513	198	2	imbalanced	chemical formula	BACE1 enzyme
	BBBP	2050	400	2	imbalanced		blood-brain barrier permeability
	clintox	1484	339	2	imbalanced		clinical toxicity
	HIV	41127	575	2	imbalanced		HIV drug activity

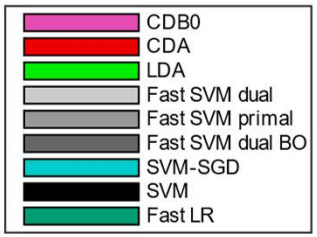
	Dataset	N	M	C	Balancedness	Modality/source	Classification task
Medical images	dermamnist	10015	2352	7	imbalanced	dermatoscope	dermal diseases
	pneumoniamnist	5856	784	2	imbalanced	chest X-Ray	pneumonia
	retinamnist	1600	2352	5	imbalanced	fundus camera	diabetic retinopathy
	breastmnist	780	784	2	imbalanced	breast ultrasound	breast diseases
	bloodmnist	17092	2352	8	imbalanced	blood cell microscope	blood diseases
	organamnist	58830	784	11	imbalanced	abdominal CT	human organs
	organcmnist	23583	784	11	imbalanced		
	organsmnist	25211	784	11	imbalanced		
	organmnist3d	1472	21952	11	imbalanced		
	nodulemnist3d	1633	21952	2	imbalanced	chest CT	nodule malignancy
	fracturemnist3d	1370	21952	3	imbalanced		fracture types
	adrenalmnist3d	1584	21952	2	imbalanced	shape from abdominal CT	adrenal gland mass
	vesselmnist3d	1908	21952	2	imbalanced	shape from brain MRA	aneurysm
	synapsemnist 3d	1759	21952	2	imbalanced	electron microscope	excitatory/inhibitory

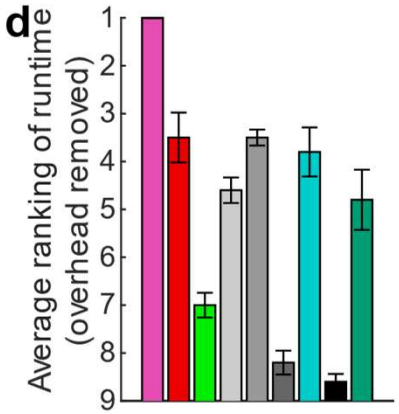
§ 3 Centroid Discriminant Analysis

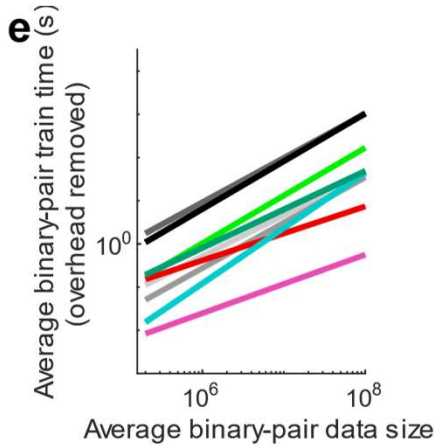
Time complexity and empirical speed on 27 datasets

Method	Training time complexity
CDB0	$O(NM + N\log N)$
CDA	$O(NM + N\log N)$
LDA	$O(NM^2 + M^3)$
SVM	$O(N^3)$
Fast SVM, SVM-SGD, Fast LR	$O(kNM)$

N – Number of samples
M – Number of features
k – Number of iterations

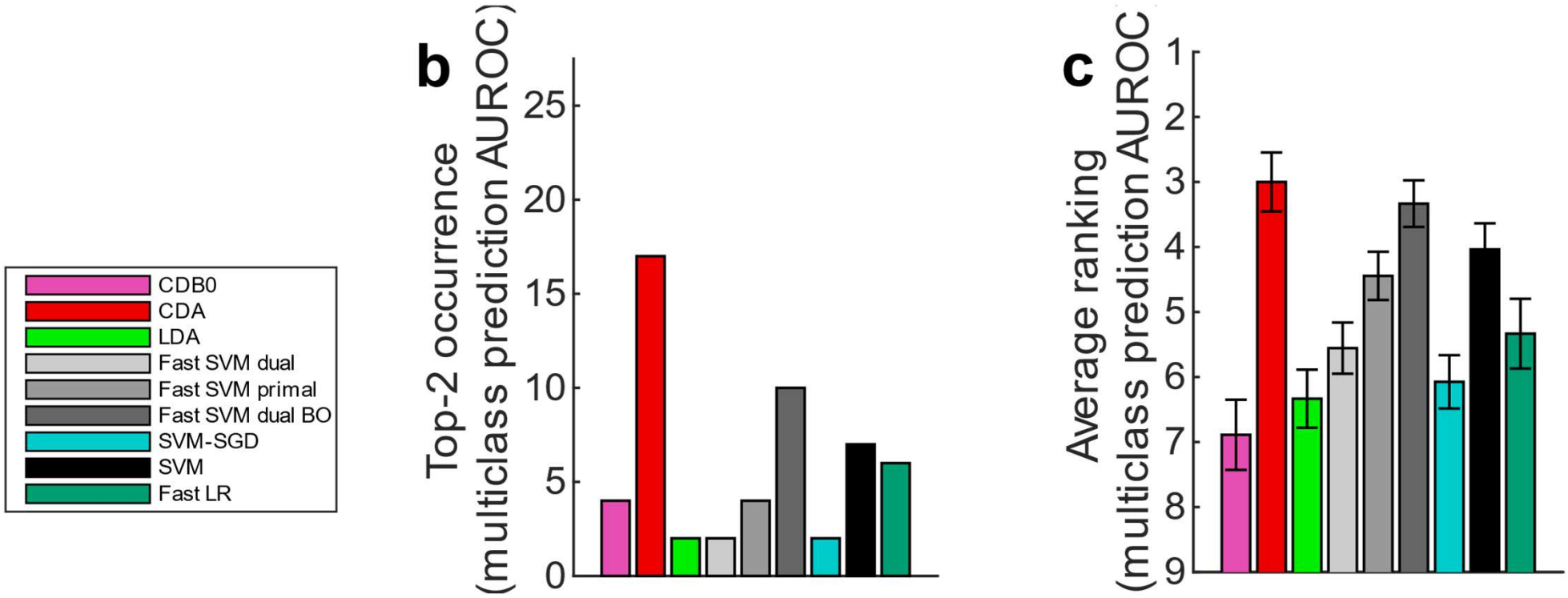






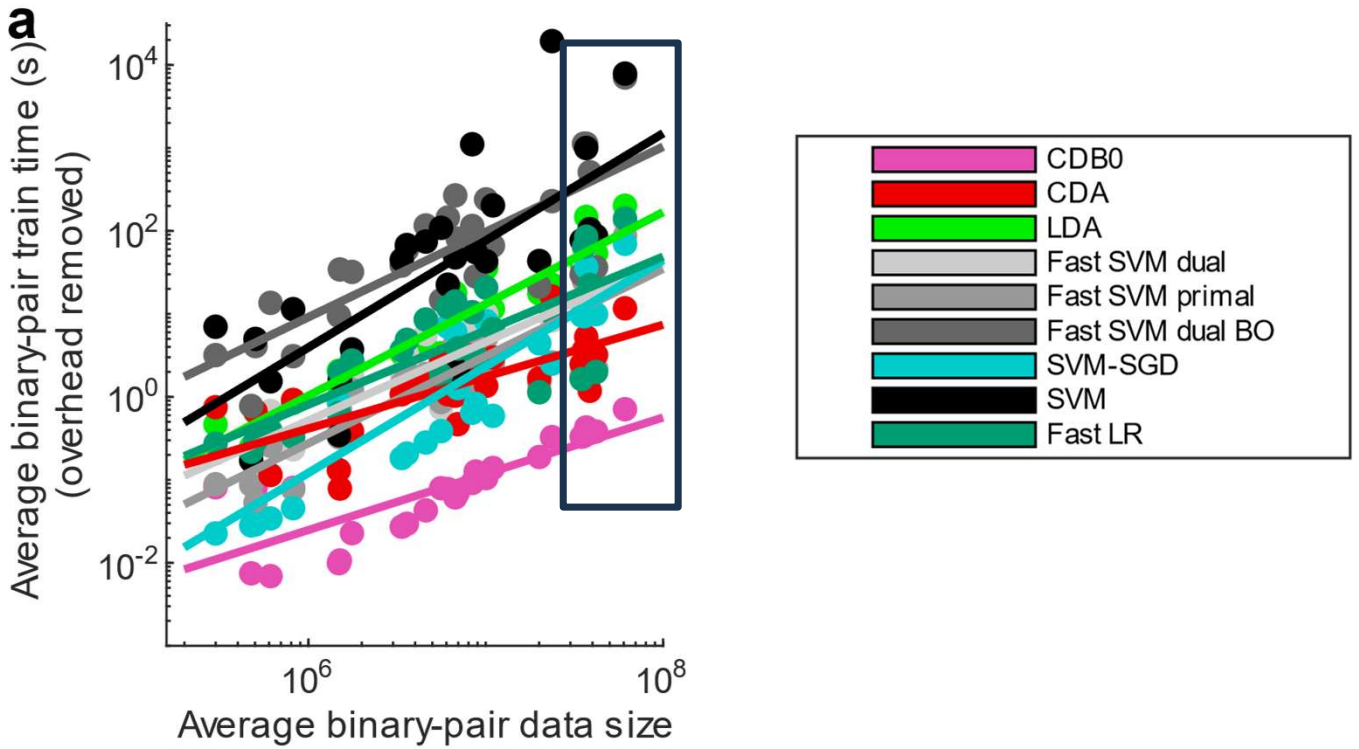
§ 3 Centroid Discriminant Analysis

Multiclass prediction performance AUROC on 27 real datasets



§ 3 Centroid Discriminant Analysis

Training time



The weighted average CDA iterations (rotations) for each dataset is **29.33**

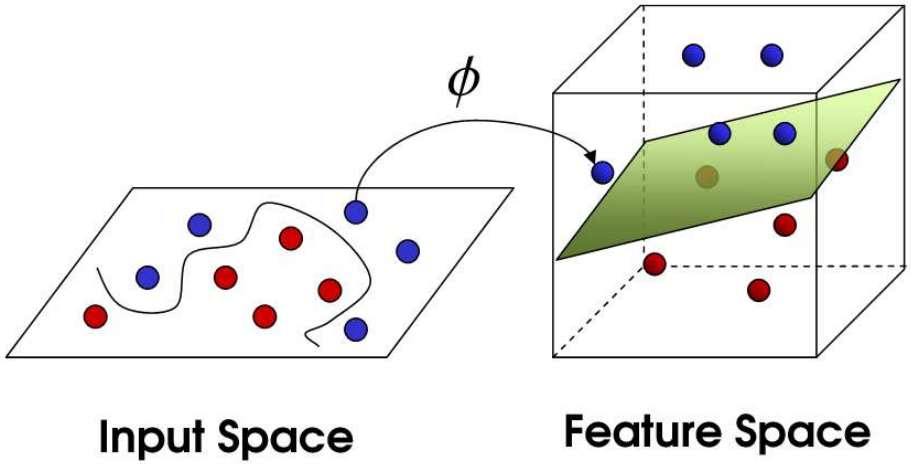
- **29.33** projection $\boldsymbol{X}\boldsymbol{w}_{\mathrm{CDB}}^{\mathrm{T}}$ $(O(NM))$
- ~300 OOP search $(O(N\log N))$

§ 4. Nonlinear kernel-based CDA

$$\boldsymbol{w}_{\mathrm{CDB}} = (\boldsymbol{\alpha} \odot \boldsymbol{c} \odot \boldsymbol{y} \cdot n)^{\mathrm{T}} \boldsymbol{X} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}$$

$$\boldsymbol{q} = \boldsymbol{X}\boldsymbol{w}_{\mathrm{CDB}}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}\right)^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$$

$$\boldsymbol{q} = \boldsymbol{X}\boldsymbol{w}_{\mathrm{CDB}}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}\right)^{\mathrm{T}} = Ker(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{\beta}$$



$\boldsymbol{w}_{\mathrm{CDB}}$: classification discriminant vector
$\boldsymbol{\alpha}$: sample weights
$\boldsymbol{c}$: weight sum division for each class
n: factor to normalize CDB as a unit vector
$\boldsymbol{y}$: labels of 1 or -1
$\boldsymbol{\beta}$: a short notation for $\boldsymbol{\alpha} \odot \boldsymbol{c} \odot \boldsymbol{y} \cdot n$
$\boldsymbol{q}$: data projection vector onto CDB

Mercer's conditions for kernels:

- Kernel matrix $Ker(\cdot,\cdot)$ is positive semi-definite

Picture from *https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f*

§ 4. Nonlinear kernel-based CDA - preliminary result

Table 1: Test set classification performance.

Dataset	Method	AUROC	AUPR	Fscore	ACscore
SVHN subset (image)	CDA	0.615±0.02	0.63±0.02	0.619±0.02	0.423±0.05
	nCDA	0.777±0.01	0.782±0.01	0.78±0.01	0.731±0.02
	SVM	0.555±0.01	0.568±0.007	0.551±0.006	0.273±0.05
	nSVM	0.736±0.02	0.776±0.009	0.756±0.008	0.654±0.03
	nLDA	**0.786±0.01**	**0.79±0.01**	**0.789±0.01**	**0.743±0.02**
ClinTox (chemical)	CDA	0.567	0.561	0.56	0.351
	nCDA	**0.625**	**0.627**	**0.627**	**0.46**
	SVM	0.565	0.578	0.575	0.294
	nSVM	0.500	0.481	0.480	0.000
	nLDA	0.605	0.612	0.611	0.409
Fracture 3D (medical image)	CDA	0.518±0.01	0.526±0.02	0.486±0.02	0.279±0.05
	nCDA	**0.625±0.04**	0.62±0.04	**0.607±0.04**	**0.577±0.08**
	SVM	0.576±0.009	0.579±0.008	0.577±0.008	0.505±0.05
	nSVM	0.608±0.06	0.591±0.07	0.586±0.07	0.44±0.2
	nLDA	0.608±0.03	**0.626±0.02**	0.605±0.02	0.491±0.1

§ 4. CDA as a pretraining method for NN initialization

Idea: Put CDA binary model weights to the weights for each hidden layer unit, provided dimension alignment.

Table 19: Test set multiclass prediction performance on the SVHN dataset (ResNet-18 pretrained feature extractor + classifier).

Method	AUROC	AUPR	Fscore	ACscore	Acc	AAE
CDA initialized linear layer	**0.79**	**0.797**	**0.795**	**0.748**	**0.664**	**0.637**
Randomly initialized linear layer	0.78	0.795	0.789	0.73	0.653	0.621
Gaussian CDA	**0.815**	**0.816**	**0.815**	**0.783**	**0.683**	-
CDA initialized MLP	0.796	0.799	0.798	0.757	0.674	0.824
Randomly initialized MLP	0.795	0.798	0.798	0.755	0.666	0.797

Hyperparameters: 150 epochs; Batch size = 128; Initial LR = 0.005 with 50% decay every 40 epochs; L2 regularization = 10^{-4}. For nonlinear method tests (the last 3), we take a subset=24000/99289 as total data due to time limit to perform kernel method on full dataset

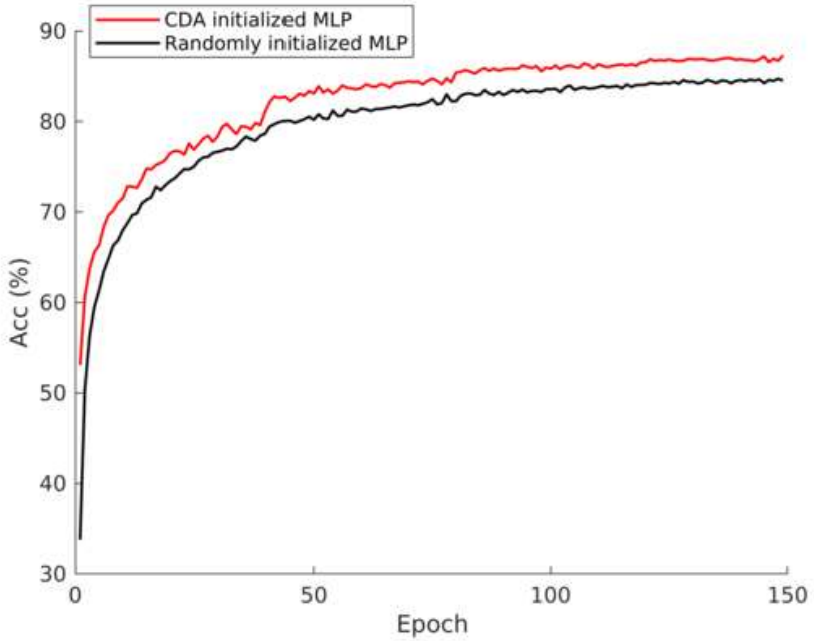


Figure 15: Training curve comparison between CDA initialized MLP and randomly initialized MLP.