BIOSTAT 527

5/31/23

Lecture 19

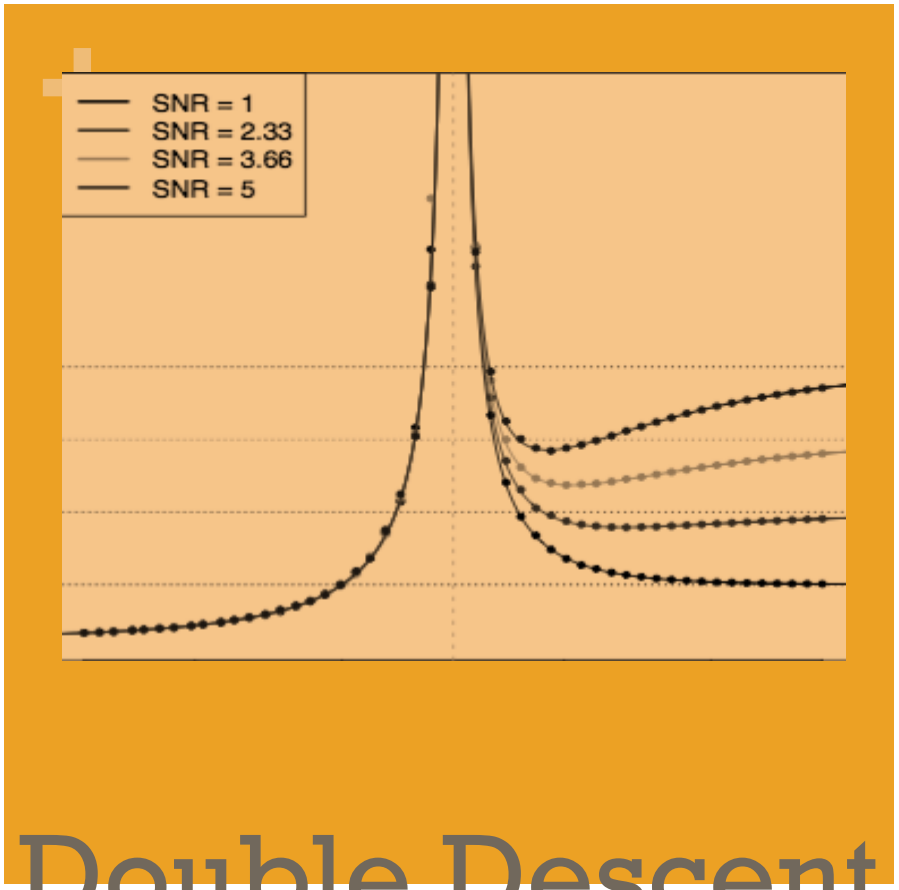
and last!

Have a nice summer!

Today: MNIST subset

- project results?
- Evaluations ✓
- DP Mixtures Models ←
- Manifold learning ←
- DD ←

Project < Report [Experimental 10%]



Double Descent

Beyond the Bias-Variance trade-off

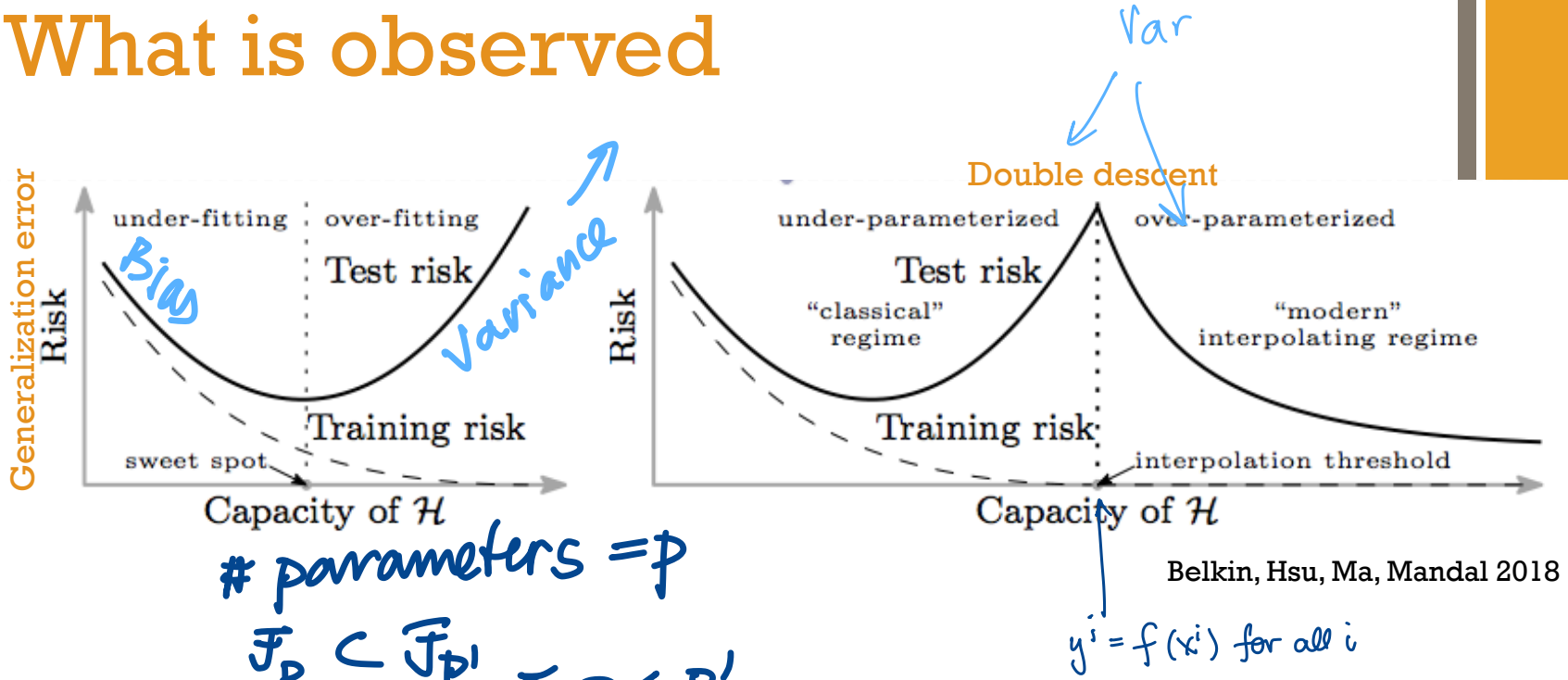
STAT 535+LPL2019

+ STAT 527

Marina Meila

University of Washington

What is observed



- Classical regime p < N
- Modern/Deep Learning/High dimensional regime N > n
  - Think N fixed, p increases, gamma=p/N
  - Training error = 0 (interpolation)
  - Test error decreases with p (or gamma)

$\mathcal{H} \equiv \mathcal{F}$

risk = loss

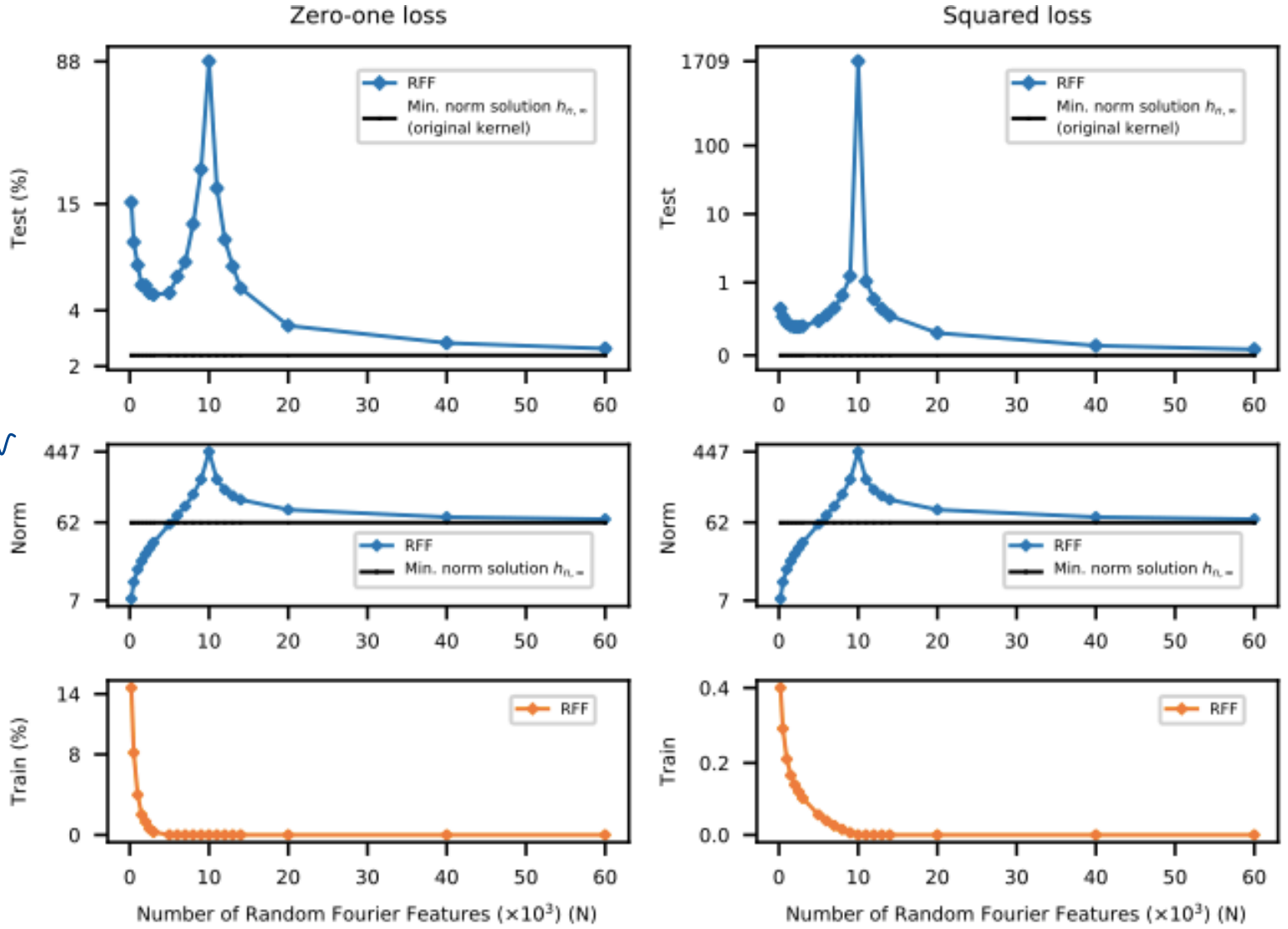
1. Thm : grad descent converges to $\hat{\theta}$ ⇒

Then $\hat{\theta} = \arg\min \|\theta\|_{RKHS}$

interpolators $\{f_\theta(x^{1:n}) = y^{1:n}\}$ in $\mathcal{F}$

$\hat{f}_\theta = \arg\min \|f_\theta\|_{RKHS}$

RFF



$\|w\|_2$

2. $\|f\|$ small $\Rightarrow$ f smooth

+ Theorem

If f^{true} smooth

$f^{true}(x^i) \approx f(x^i) \Rightarrow \|f - f^{true}\|$ small RKHS

Theorem 1. *Fix any $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed random variables, where x_i is drawn uniformly at random from a compact cube[2] $\Omega \subset \mathbb{R}^d$, and $y_i = h^*(x_i)$ for all i. There exists absolute constants $A, B > 0$ such that, for any interpolating $h \in \mathcal{H}_\infty$ (i.e., $h(x_i) = y_i$ for all i), so that with high probability*

$$\sup_{x \in \Omega} |h(x) - h^*(x)| < Ae^{-B(n/\log n)^{1/d}} \left(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty}\right).$$

4. Guarantee (of DD)

↑ estimated ↑ true

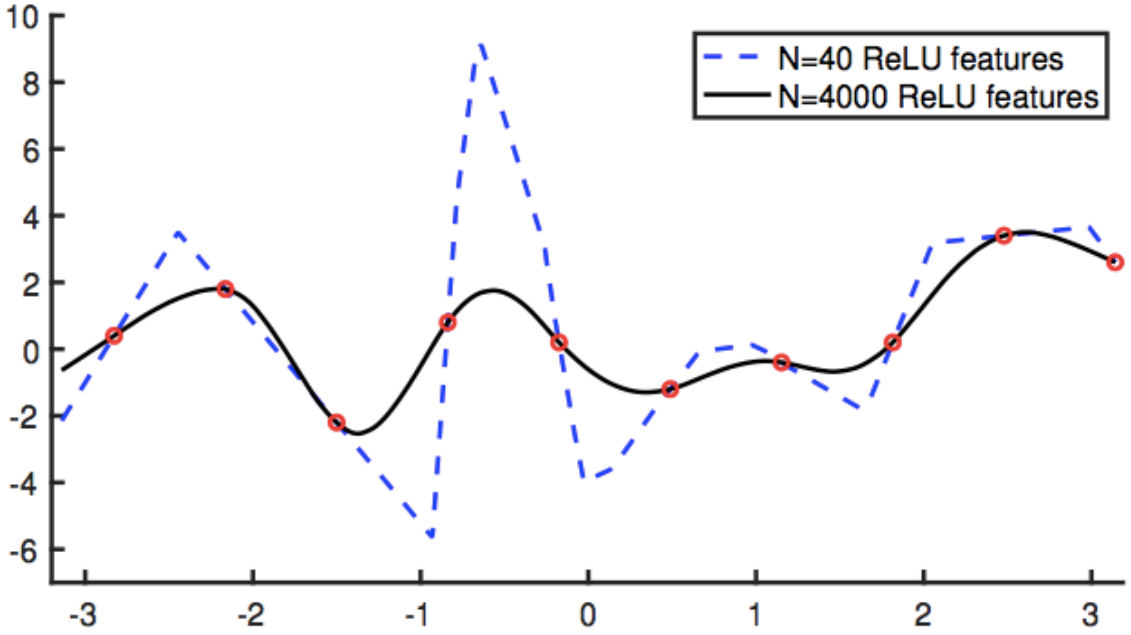
lower = smoother !!

3. $\mathcal{F}' \subset \mathcal{F}$

$\|\hat{f}'\| \geq \|\hat{f}\|$

smoothest interpolator becomes smoother

Main intuition [Belkin et al.]



- The target function h* is (mostly) smooth
  - i.e. $||h^*||_{RKHS}$ is small
- p > N, no noise, hence h_p interpolates data
- Train to minimize $||h_p||$ subject to 0 training error
- Then $||h_p||$ will decrease with p!

Theorem

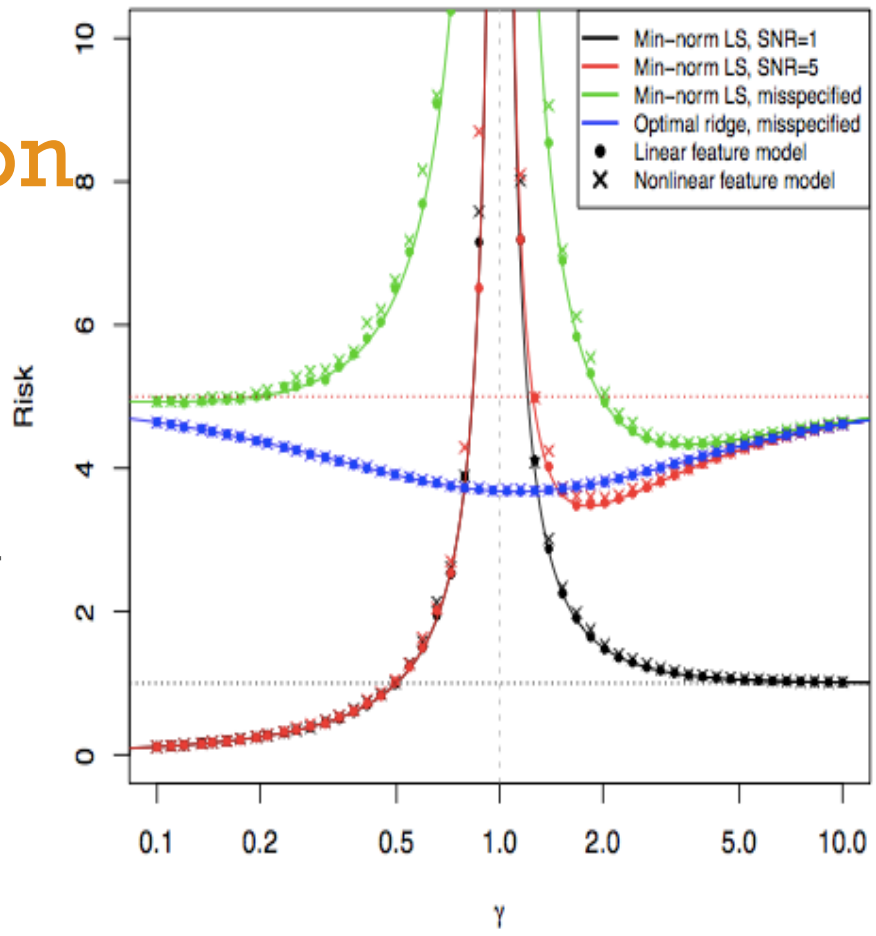
Theorem 1. *Fix any $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed random variables, where x_i is drawn uniformly at random from a compact cube[2] $\Omega \subset \mathbb{R}^d$, and $y_i = h^*(x_i)$ for all i. There exists absolute constants $A, B > 0$ such that, for any interpolating $h \in \mathcal{H}_\infty$ (i.e., $h(x_i) = y_i$ for all i), so that with high probability*

$$\sup_{x \in \Omega} |h(x) - h^*(x)| < Ae^{-B(n/\log n)^{1/d}} \left(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty} \right).$$

Linear regression

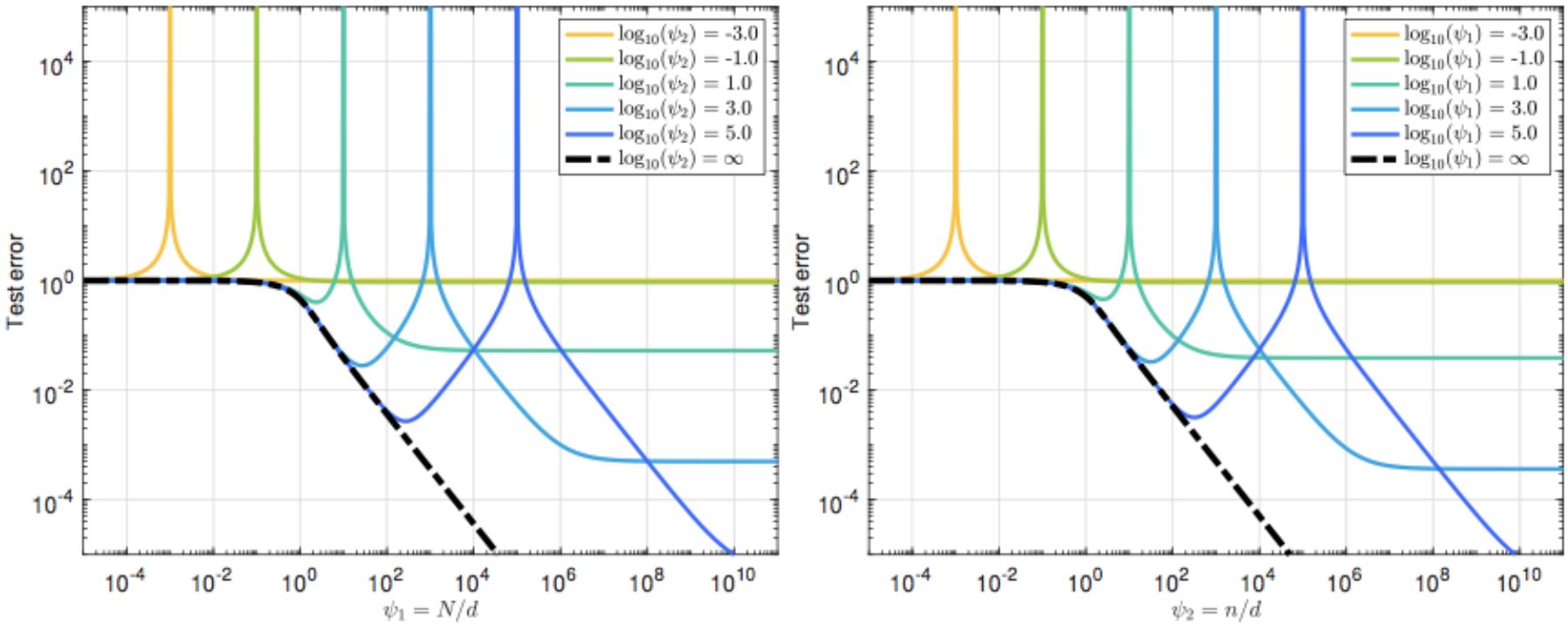
[Hastie, Montanari, Rosset, Tibshirani 2019]

- Linear, nonlinear features behave the same way
- Model correct, misspecified
- Noise level sigma affects asymptotic error
- and optimal N/n



- Double descent is not regularization

Figure 1: Asymptotic risk curves for the linear feature model, as a function of the limiting aspect ratio γ. The risks for min-norm least squares, when SNR $= 1$ and SNR $= 5$, are plotted in black and red, respectively. These two match for $\gamma < 1$ but differ for $\gamma > 1$. The null risks for SNR $= 1$ and SNR $= 5$ are marked by the dotted black and red lines, respectively. The risk for the case of a misspecified model (with significant approximation bias, $a = 1.5$ in (13)), when SNR $= 5$, is plotted in green. Optimally-tuned (equivalently, CV-tuned) ridge regression, in the same misspecified setup, has risk plotted in blue. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries. Meanwhile, the "x" points mark finite-sample risks for a nonlinear feature model, with $n = 200$, $p = [\gamma n]$, $d = 100$, and $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries, W has i.i.d. $N(0, 1/d)$ entries, and $\varphi(t) = a(|t| - b)$ is a "purely nonlinear" activation function, for constants a, b. The theory predicts that this nonlinear risk should converge to the linear risk with p features (regardless of d). The empirical agreement between these two—and the agreement in finite-sample and asymptotic risks—is striking.



- More refined analysis includes noise, non-linearity, data dimension n, ridge regularization lambda [Mei, Montanari 2019]
- When is global minimum in overparametrized regime?
- Enough data N/n > 1
- lambda → 0 (or min-norm LS)
- p >> N
- SNR || beta ||/noise > 1
- Bias, Variance strictly decreasing with p/N to > 0 limit

CSE 547/STAT 548

Non-linear dimension reduction: an introduction

Marina Meilă

Department of Statistics
University of Washington

January 2022

Outline

Manifold learning / Non-lin dimension reduction

"Manifold" in machine learning — Neural Networks - autoencoders
"data manifold" - hidden layers
≃ representation

- t-SNE (UMAP) → visualisation - no guar-
clustering for dim red.

- Manifold learning (ML) ⇐ use mathe-
matical def
- estimates with guarantees

t SNE - perplexity → can emulate UMAP & other algorithms
- ϱ "learning rate" ⇒ granularity (for clustering)
smoothness

- better understood
for clustering

Who needs manifold learning?

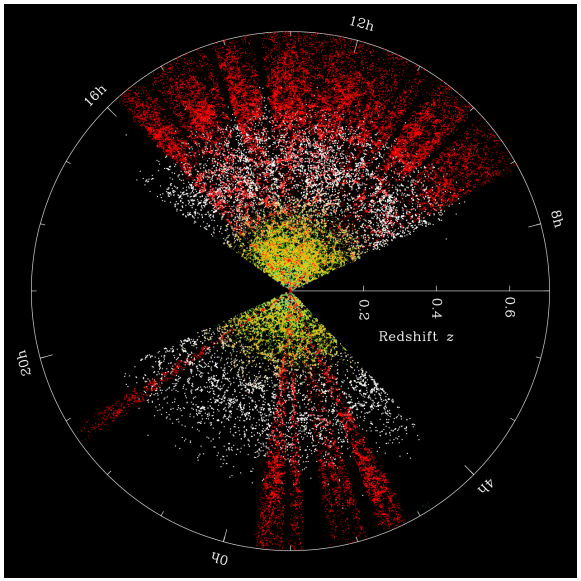
- What is PCA good for?
  - eigenvalues → analyze structure : d=? high low
  - reduce dim $D \to d$
  - eigenvectors → interpret
  - remove noise (smooth)
  - preprocessing → passed on to other inferences prediction ...

ML

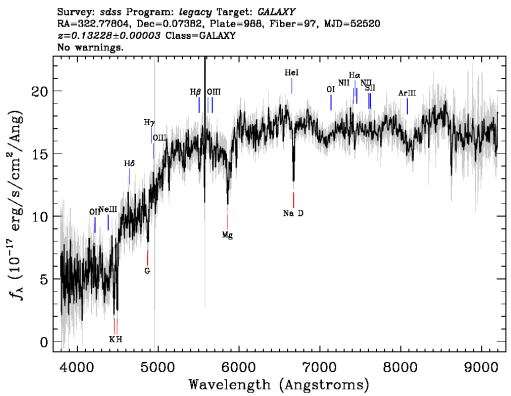
First step in data pipeline

Spectra of galaxies measured by the Sloan Digital Sky Survey (SDSS)

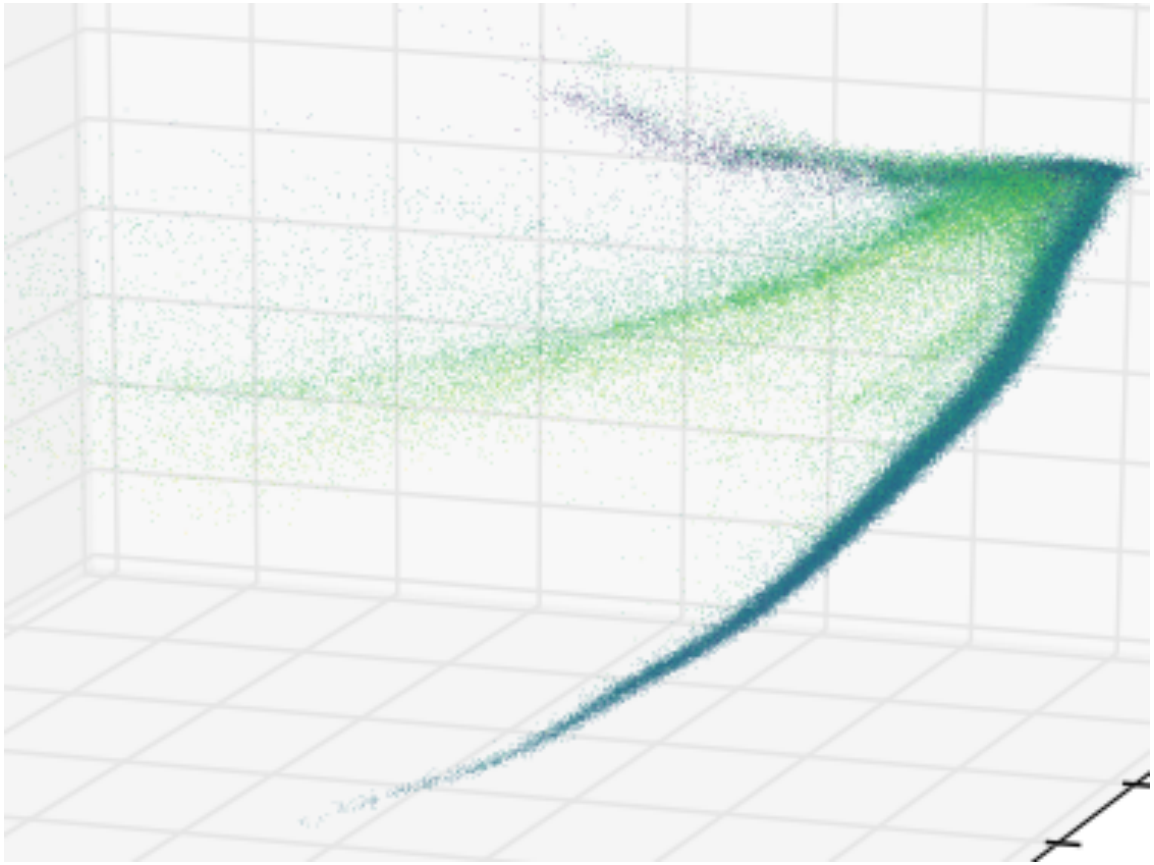
www.sdss.org



www.sdss.org



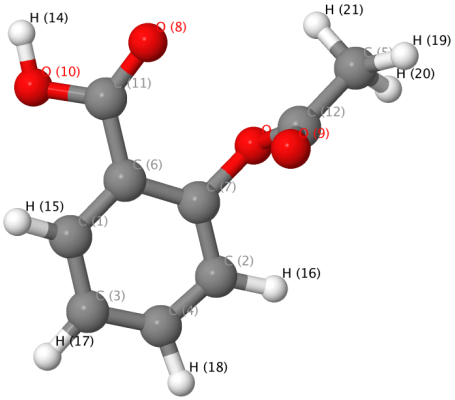
- Preprocessed by Jacob VanderPlas and Grace Telford
- $n = 675,000$ spectra $\times D = 3750$ dimensions

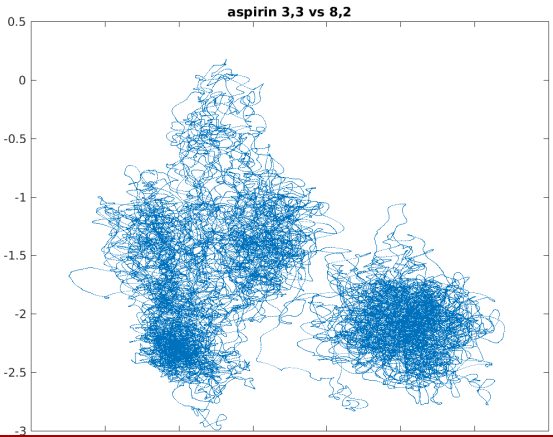


embedding by James McQueen

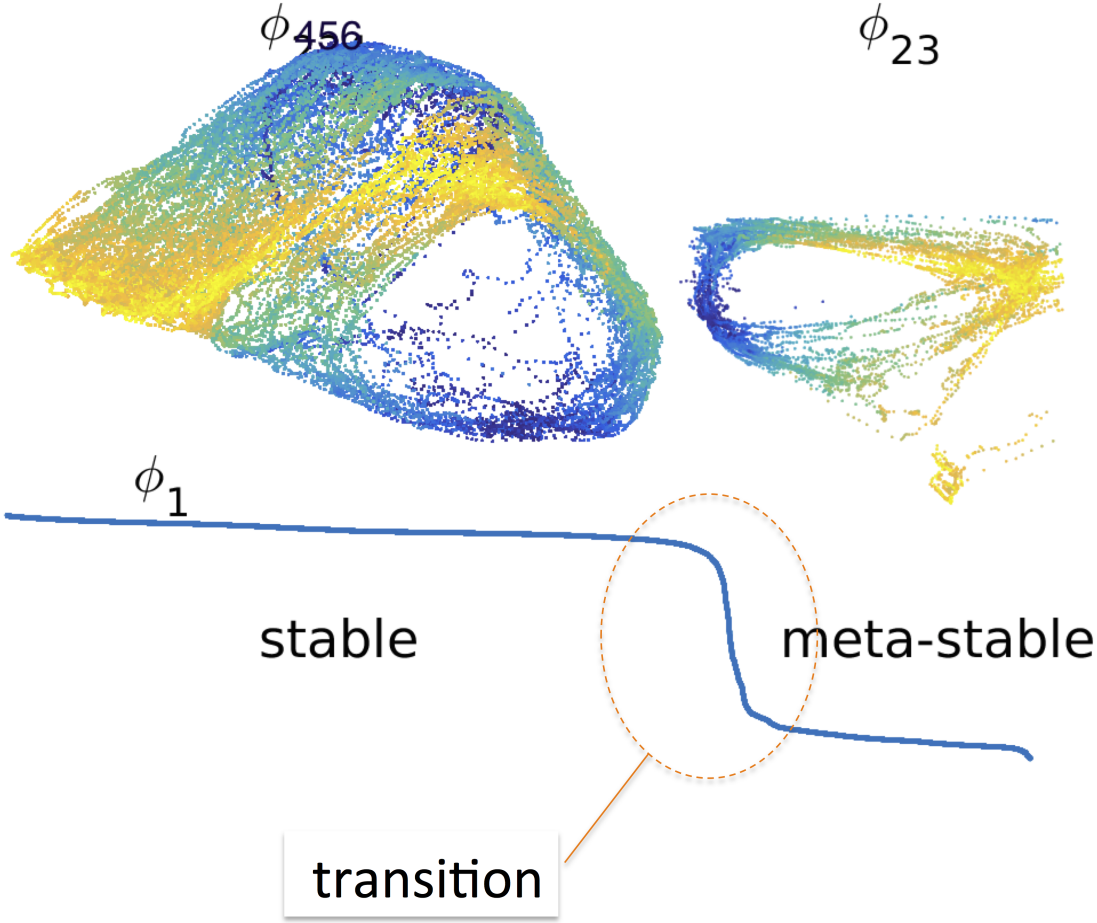
Molecular configurations

aspirin molecule



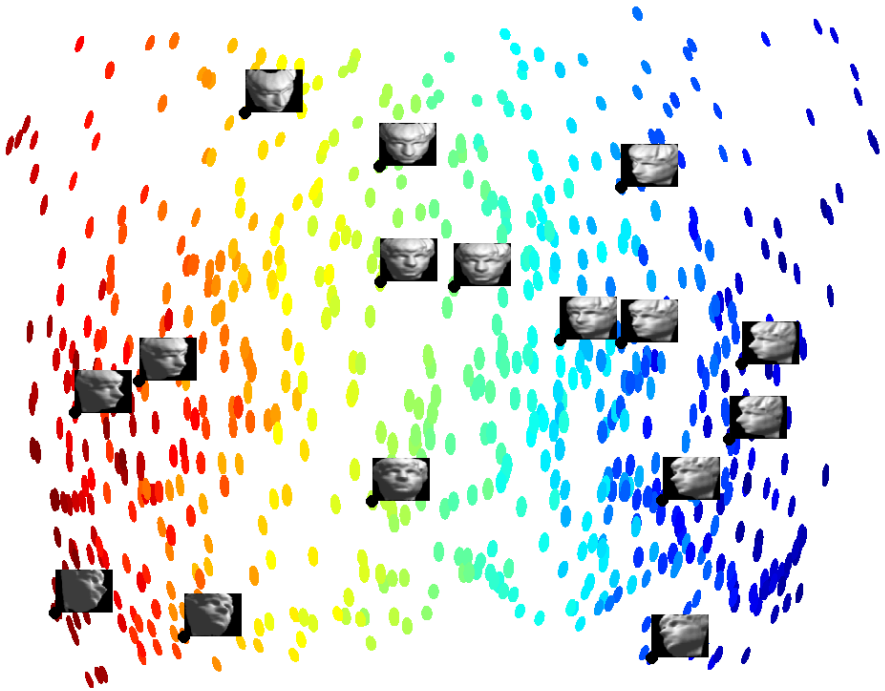


- Data from Molecular Dynamics (MD) simulations of small molecules by [Chmiela et al. 2016]
- $n \approx 200,000$ configurations $\times$ $D \sim 20 - 60$ dimensions



When to do (non-linear) dimension reduction

- $n = 698$ gray images of faces in $D = 64 \times 64$ dimensions
- head moves up/down and right/left
- With only two degrees of freedom, the faces define a 2D manifold in the space of all 64×64 gray images

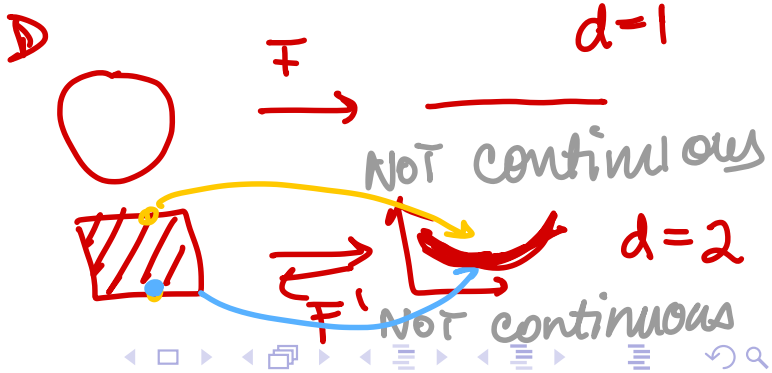


F = embedding alg

$x \in \mathbb{R}^D \longmapsto y \in \mathbb{R}^d \in \mathbb{R}^m$

$D >> m \geq d$

F embedding = map to lower dim so that F smooth C^1 and F^{-1} smooth



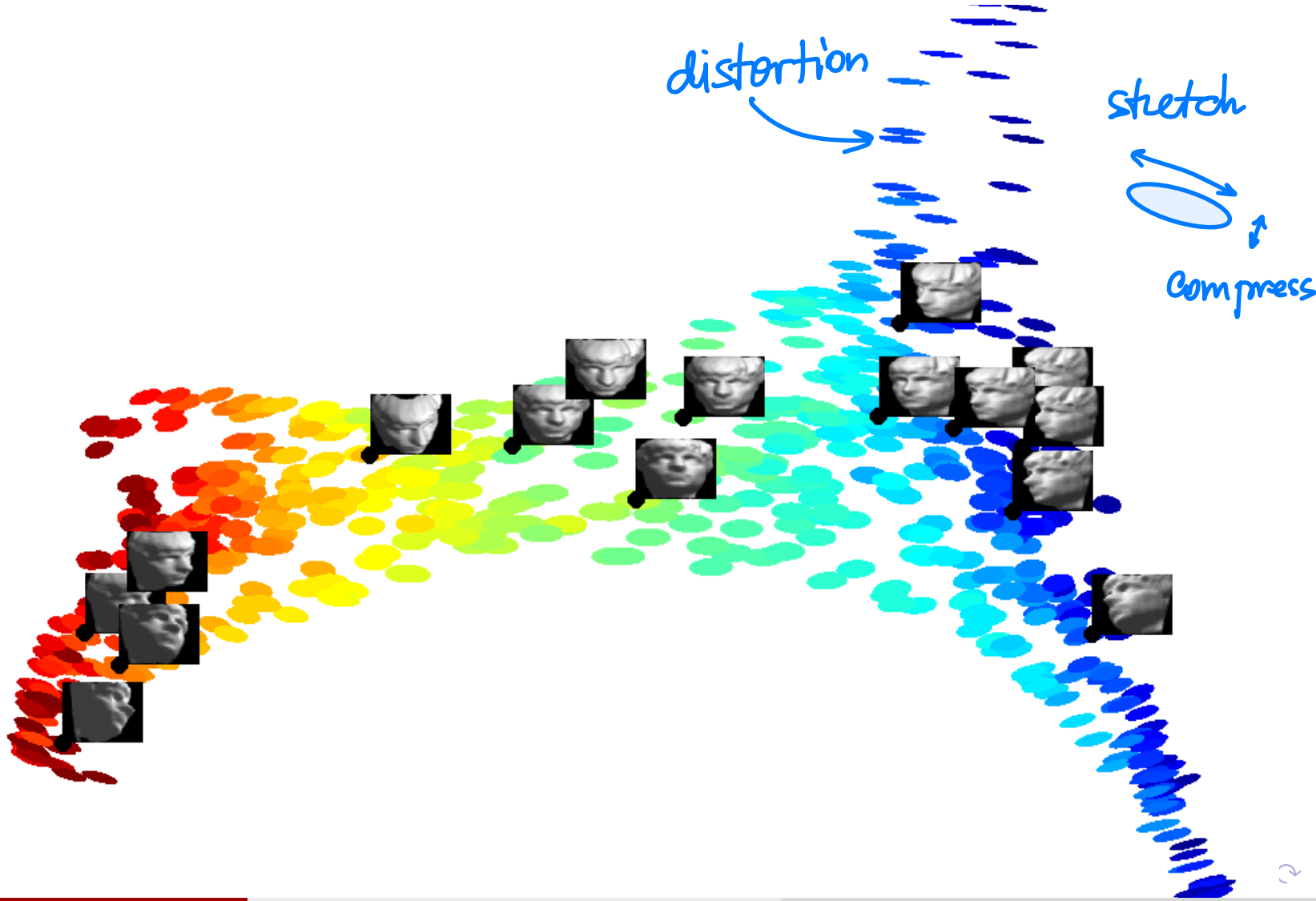
Wanted from ML algorithm

1. $X \xrightarrow{F} y$ $\quad F$ smooth, $\exists F^{-1}$ smooth

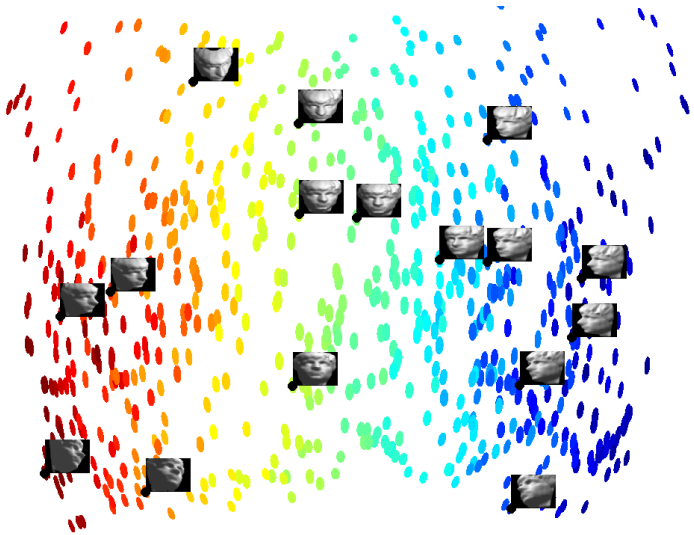
2. isometry $\quad F$ preserves distances

G for Sculpture Faces

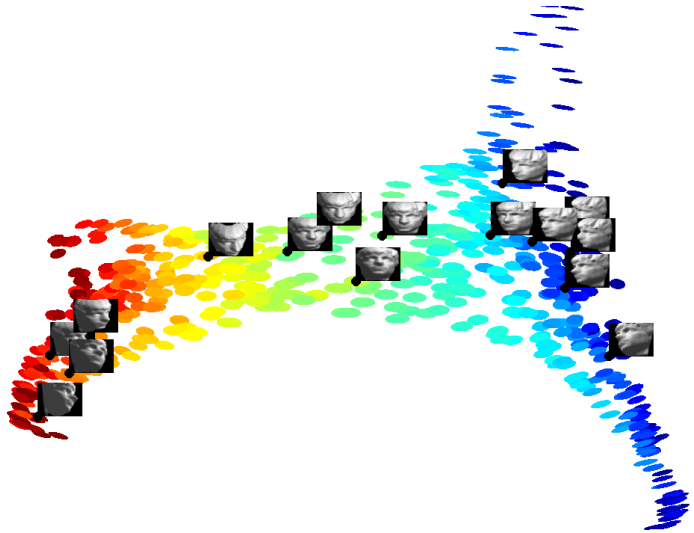
- $n = 698$ gray images of faces in $D = 64 \times 64$ dimensions
- head moves up/down and right/left



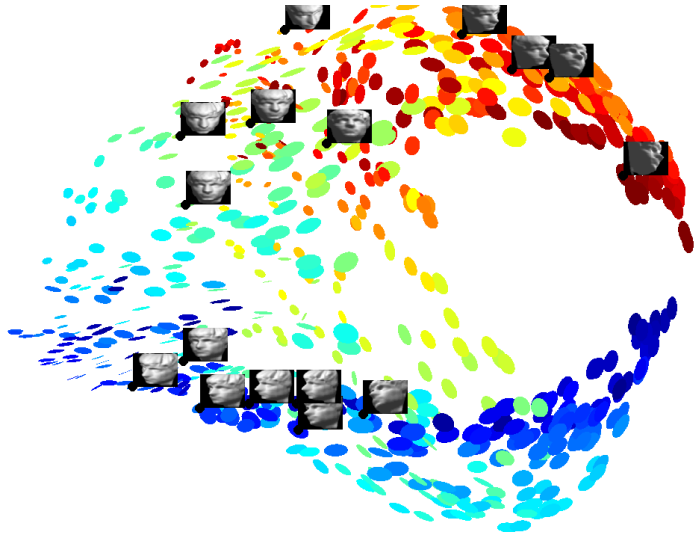
Corrections for 3 embeddings of the same data



Isomap

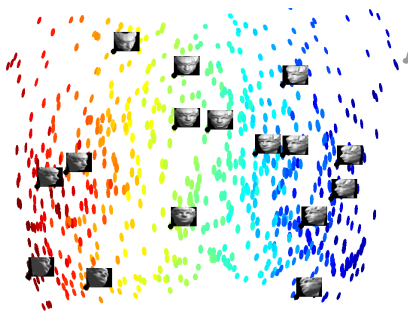


LTSA



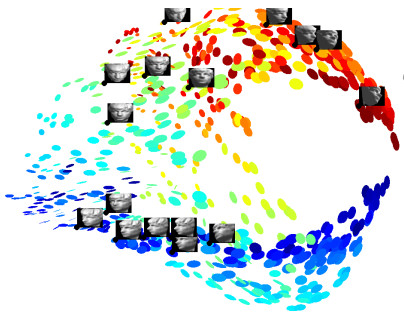
Laplacian Eigenmaps

Isomap vs. Diffusion Maps



Isomap

- Preserves geodesic distances
  - but only when $\mathcal{M}$ is flat and "data" convex
- Computes all-pairs shortest paths $\mathcal{O}(n^3)$
- Stores/processes dense matrix



DiffusionMap

most guarantees known

- Distorts geodesic distances
- Computes only distances to nearest neighbors $\mathcal{O}(n^{1+\epsilon})$
- Stores/processes sparse matrix

LTSA

- t-SNE, UMAP visualization algorithms

Metric Manifold Learning

Wanted

- eliminate distortions for any "well-behaved" $\mathcal{M}$
- and any any "well-behaved" embedding $\phi(\mathcal{M})$
- in a tractable and statistically grounded way

Idea

Given data $\mathcal{D} \subset \mathcal{M}$, some embedding $\phi(\mathcal{D})$ that preserves topology (true in many cases)

- Estimate distortion of ϕ and correct it! (see prev/next slide)
- The correction is called the **pushforward Riemannian Metric** g

NP clustering = ∞ mixtures

DPMM ↕ (Finite) Mixtures of Gaussians

Lecture 14: Dirichlet Process Mixtures in a nutshell

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

March 9, 2018

- The Chinese Restaurant Process → generates cluster labels & cluster parameters μ_k, Σ_k

  Dirichlet Process Mixture → generates samples

The Chinese Restaurant Process

$\Pr[\text{new table}] = \frac{\alpha}{n+\alpha} \propto \frac{1}{n}$

$\mathbb{E}[\#\text{tables}] \approx \alpha\left(1+\frac{1}{2}+\cdots\frac{1}{n}\right) \approx \alpha \ln n$

- **Given** parameters $\alpha > 0$, G_0, with G_0 a continuous measure on measurable space $(\Theta, \mathcal{B})$.
- Assume we already have samples $\theta_{1:n} \in \Theta$.
- The probability of θ_{n+1} is then

$\theta_{k+1} \sim G_0$ ← prior

$$\theta_{n+1} \mid \theta_{1:n} \sim \sum_{k=1}^{K} \frac{n_k}{n+\alpha}\delta_{\theta_k} + \frac{\alpha}{n+\alpha} G_0. \quad (1)$$

(K = # tables; n_k = # customers at table k)

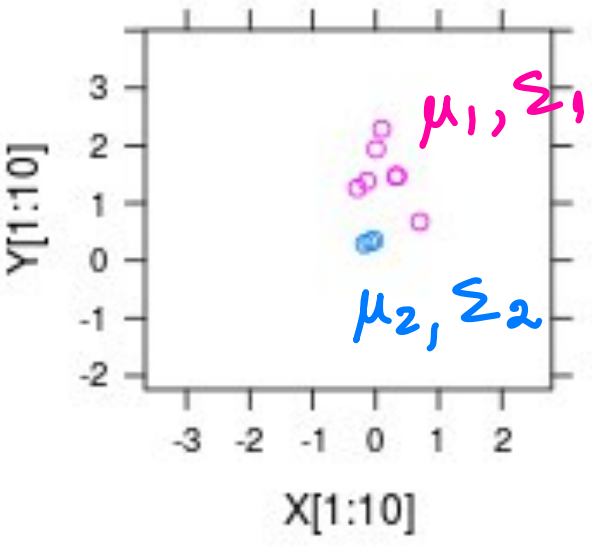
In the above, K represents the number of distinct values among the n samples $\theta_{1:n}$. Note: all distinct θ's are sampled from G_0!

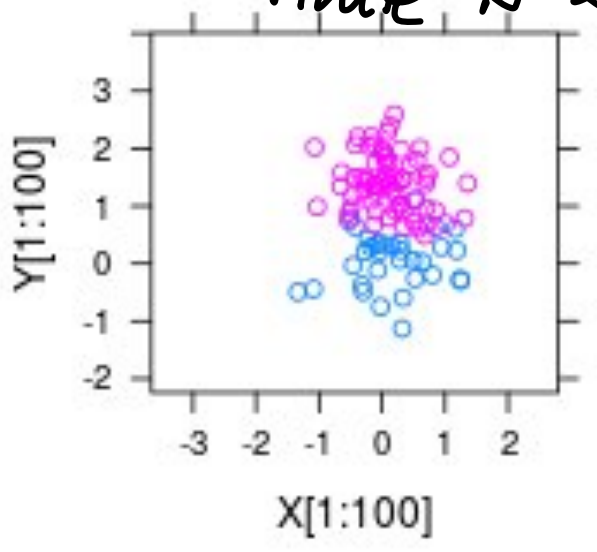
- This defines a **Chinese Restaurant Process (CRP)**. It is easy to see that the process is exchangeable. $\Leftrightarrow$ likelihood invariant to ordering $1{:}n$
- One can also prove that for $n \to \infty$, $\theta_{1:n} \to G$ where $G \sim DP(\alpha, G_0)$.

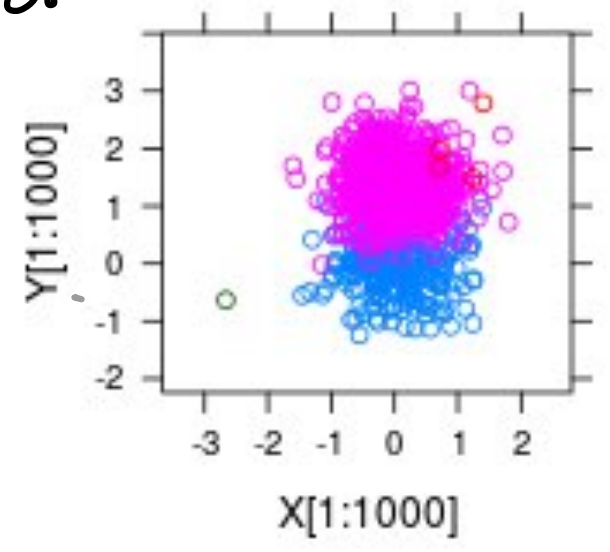
$\theta_k = (\mu_k, \Sigma_k)$ cluster params

$x_i \sim p_{\theta_i}$

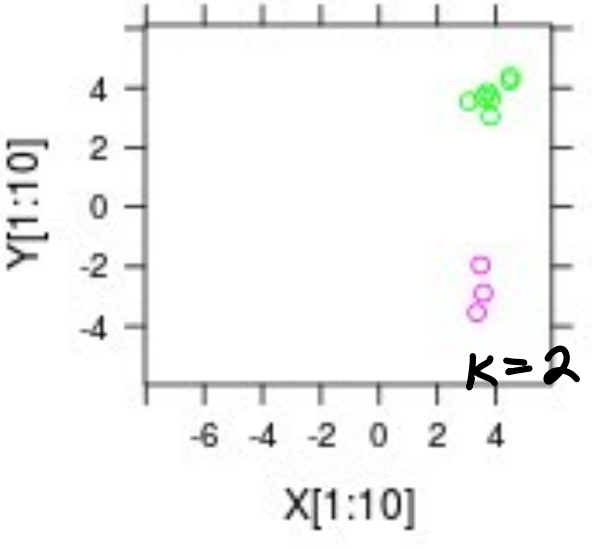
Finite $K=2$ or $\alpha=0$

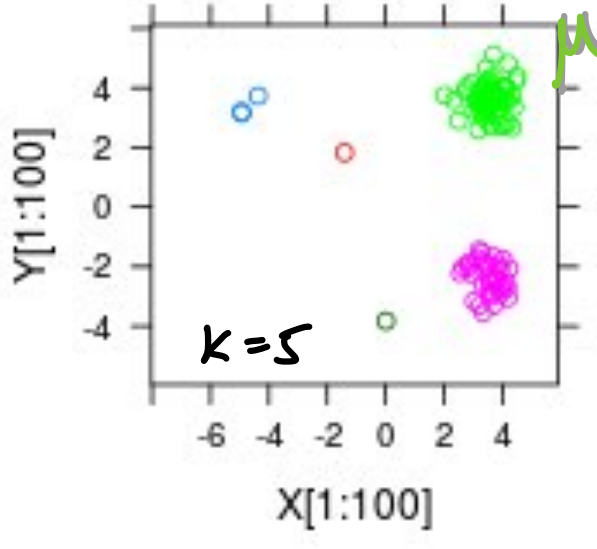


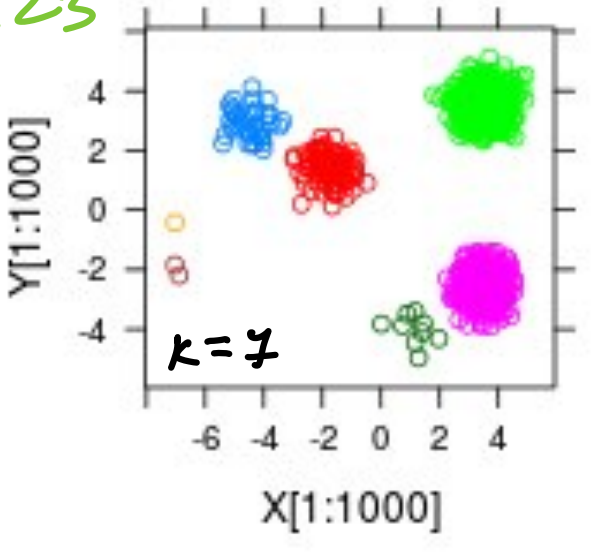




$\alpha > 0$ DPMM







Dirichlet Process

- A **Dirichlet Process (DP)** is distribution over measures.
- Let $(\Theta, \mathcal{B})$, α, G_0 be as above.
- We say that the random function G is drawn from $DP(\alpha, G_0)$ iff

$$\text{for any partition } B_{1:K} \subset \mathcal{B} \text{ of } \Theta, G(B_{1:K}) \sim Dirichlet(\alpha G_0(B_{1:K})). \quad (2)$$

Dirichlet Process Mixture

- Given: $DP(\alpha, G_0)$, family of distributions $\{f_\theta\}$ on $\mathcal{X}$.
- For $i = 1, 2, \ldots n$

$$\theta_i \sim CRP(\alpha, G_0, \theta_{1:i-1}) \tag{3}$$

$$x_i \sim f_{\theta_i} \tag{4}$$

Estimation of DP Mixture by Gibbs sampling

Input $\alpha, G_0, \{f\}, \mathcal{D} = \{x_1, \ldots x_n\} \subseteq \mathcal{X}$

State cluster assignments c_i, $i = 1 : n$,
parameters θ_k for all distinct k

Iterate

1. for $i = 1 : n$(reassign data to clusters)
   - 1.1 if $n_{c_i} = 1$ delete this cluster and its θ_{c_i}
   - 1.2 resample c_i by

$$c_i = \begin{cases} \text{existing } k & \text{w.p } \propto \frac{n_k}{n-1+\alpha} f(x_i, \theta_k) \\ \text{new cluster} & \text{w.p } \frac{\alpha}{n-1+\alpha} \int f(x_i, \theta) G_0(\theta) d\theta \end{cases} \tag{5}$$

   - 1.3 if c_i is new label, sample a new θ_{c_i} from f_θ $G_0(\theta)$
2. (resample cluster parameters)
   for $k \in \{c_{1:n}\}$
   - 2.1 sample θ_k from posterior $f_{\theta_k} \propto G_0(\theta) \prod_{i \in C_k} f(x_i, \theta)$

   this can be computed in closed form if G_0 is conjugate prior

Output a state with high posterior