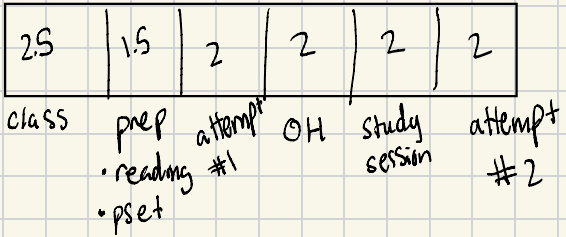
Week 10

↳ Midterms

- generally, reflect effort in class (OH + study sessions + psets)
- goal is learning:
  - ↳ redo one long question for up to 1/2 pts lost
  - ↳ I'll replace 1st midterm grade w/ 2nd, if better
- let's chat in Tom-Is!

↳ Guideline of 12 hours per class per week



↳ Anonymous eval!

Beyond Supervised Learning

Q: Lots of unlabelled data (images, text), how do we use it?

This week, we will explore *pretraining*

Autoencoders

Idea: Use data as its own label

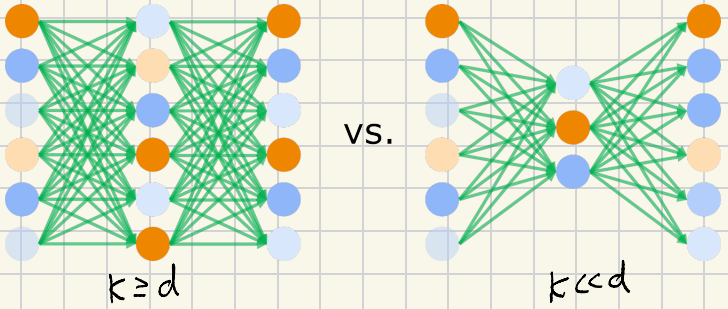
$f_0 : \mathbb{R}^d \to \mathbb{R}^k$ "Encoder"

$f_1 : \mathbb{R}^k \to \mathbb{R}^d$ "Decoder"

$z = f_0(x)$ is latent representation

$\tilde{x} = f_1(z)$ is reconstruction

$$\mathcal{L}_{recon} = \frac{1}{n} \sum_{i=1}^{n} \| x^{(i)} - \tilde{x}^{(i)} \|_2^2$$

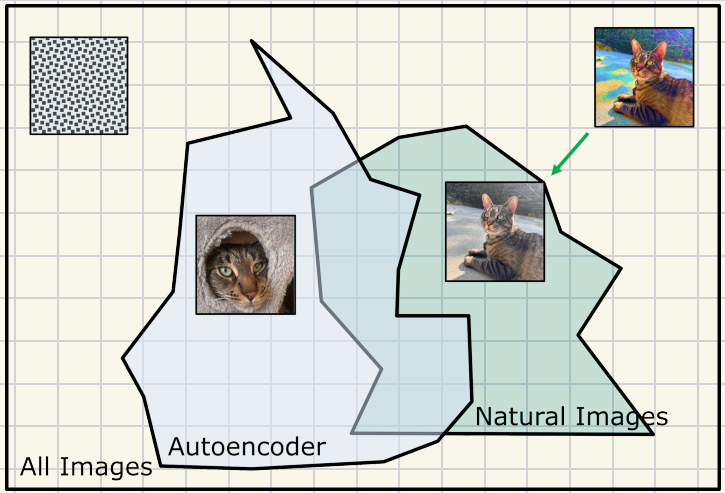


Applications include:

- data compression
- denoising
- inpainting
- representation learning

Data Manifold

Necessarily losing information, okay because we only want to represent meaningful images



Prefer working with latent rep.

1. more efficient in lower dim
2. most meaningful features

Examples include:

- classification
- clustering
- generation

Variational Autoencoders

Goal: make latent space "nicely behaved"

Treat latent as random sample

$$z \sim \mathcal{N}(\mu_x, \Sigma_x)$$

μ_x: mean from encoder; Σ_x: variance from decoder

"Nicely behaved" $\Leftrightarrow \mathcal{N}(\mu_x, \Sigma_x) \approx \mathcal{N}(0, I)$

Distribution $P \approx$ Distribution Q

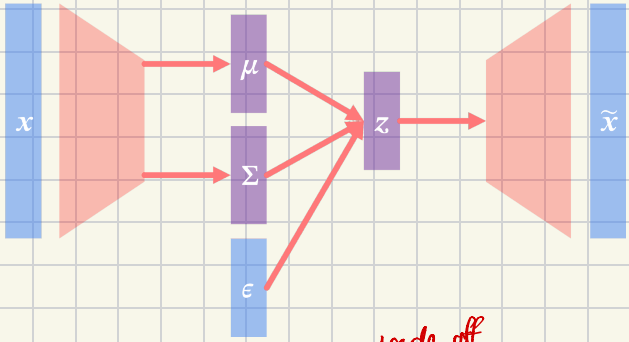
Loss: KL divergence

$$D_{KL}(P||Q) = E_{z\sim P}\left[\log \frac{P(z)}{Q(z)}\right]$$

Example in $\mathbb{R}$: $P = \mathcal{N}(\mu_A, 1)$, $Q = \mathcal{N}(\mu_B, 1)$

How do we backpropagate through a sample?

$$z = \mu_x + \Sigma_x^{1/2} \epsilon \quad \epsilon \sim \mathcal{N}(0, I) \text{ draw}$$



$$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \lambda D_{KL}(P||Q)$$

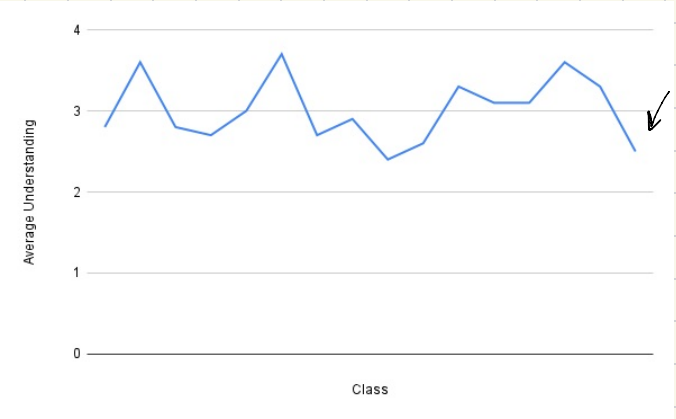
trade off

$\mathcal{L}_{recon}$: z encodes meaning of x

$D_{KL}(P||Q)$: z distributed similarly

Logistics

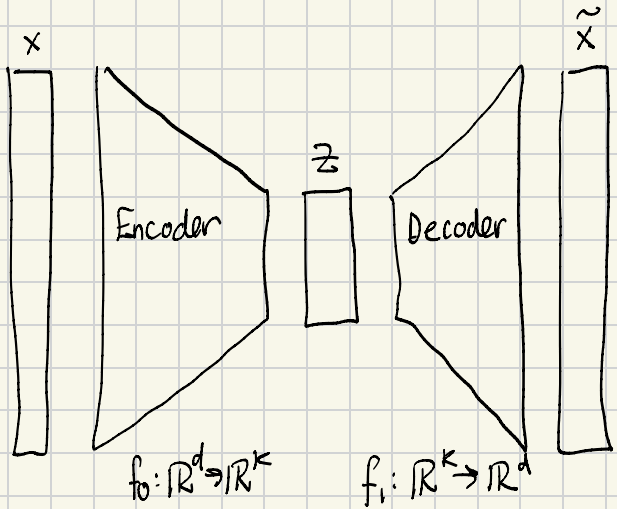
- Midterm ↺ Redo due 10/31
  - ↳ instructions on discord



first section had worse pacing + explanation, I'm sorry :-

Please do reading and come to office hours!

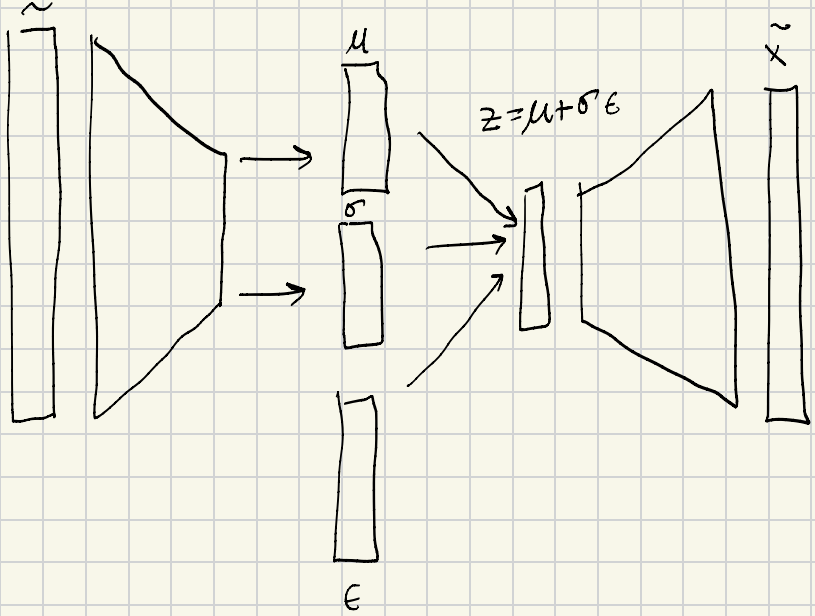
Autoencoder



$$\mathcal{L}_{recon} = \sum_{i=1}^{n} \| x^{(i)} - \tilde{x}^{(i)} \|_2^2$$

Variational Autoencoder

Distribution of latent is "nice"



$$\mathcal{L}_{dist} = \sum_{i=1}^{n} D_{KL}\left(\mathcal{N}(\mu_{x^{(i)}}, \sigma^2_{x^{(i)}}) \,\|\, \mathcal{N}(0,1)\right)$$

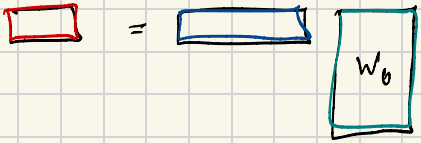
$$\mathcal{L} = \mathcal{L}_{recon} + \lambda \mathcal{L}_{dist}$$

Principal Component Analysis

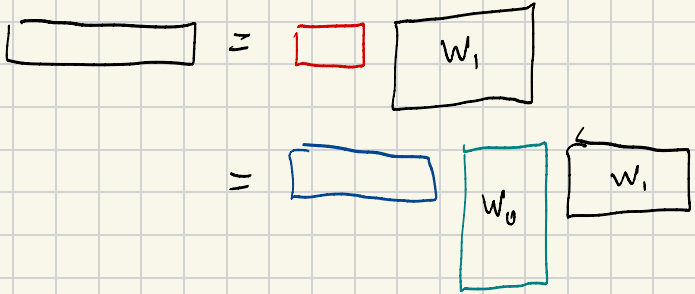
The "linear regression" of autoencoders

$f_0 = W_0 \in \mathbb{R}^{k \times d}$ $\quad f_1 = W_1 \in \mathbb{R}^{d \times k}$

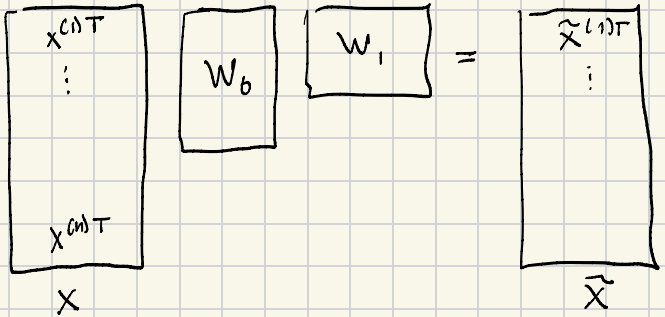
$z^T = x^T W_0$



$\tilde{x}^T = x^T W_0 W_1 = z^T W_1$



$X \in \mathbb{R}^{n \times d}$



$$\mathcal{L}_{recon} = \sum_{i=1}^{n} \| x^{(i)} - \tilde{x}^{(i)} \|_2^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d} (x_j^{(i)} - \tilde{x}_j^{(i)})^2$$

$$= \| X - \tilde{X} \|_F^2$$

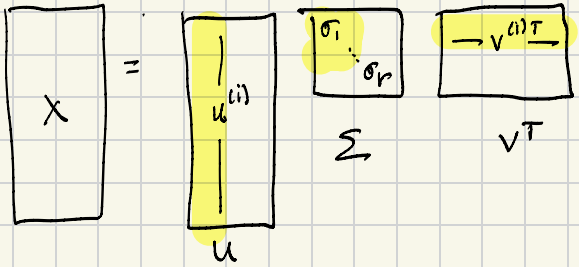
↑ sum of every entry squared

Q: What is the best W_0 and W_1?

Singular Value Decomposition

$$X = U \Sigma V^T \quad \leftarrow r = \text{rank}$$

$n \times d \quad n \times r \quad r \times r \quad r \times d$



$$X = \sum_{i=1}^{r} u_i \sigma_i v_i^T$$

u_i orthonormal $\quad U^T U = I_{r \times r}$

v_i orthonormal $\quad V^T V = I_{r \times r}$

Fact: $\|X\|_F^2 \overset{\text{HW}}{=} \sum_{i=1}^{r} \sigma_i^2$

Q: What is the best rank k (latent dim) approximation to X?

Assume $k \leq r$

$$X_k = \sum_{i=1}^{k} u_i \sigma_i v_i^T$$

$$\|X - X_k\|_F^2 = ?$$

Back to PCA

Goal: $\tilde{X} = X_k$

Q: What are W_0 and W_1?

A: $W_0 = V_k$ $\quad W_1 = V_k^T$

$d \times k$ ✓ $\quad k \times d$ ✓

$$\tilde{X} = X W_0 W_1^T$$

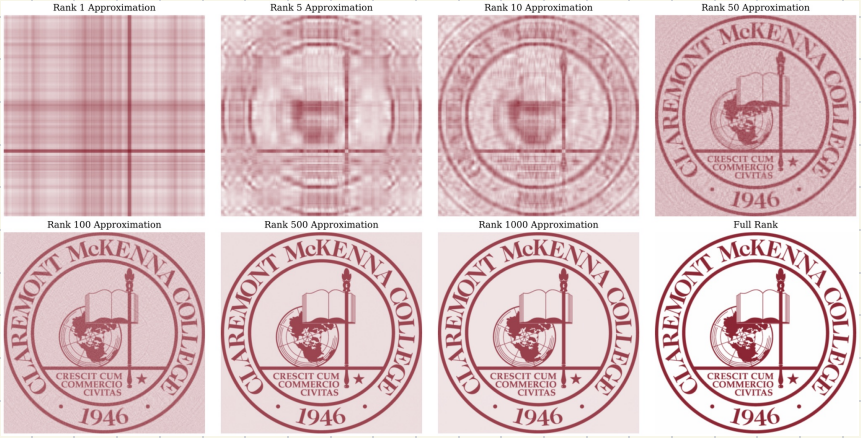
$$V_k V_k^T = \sum_{i=1}^{k} v_i v_i^T$$

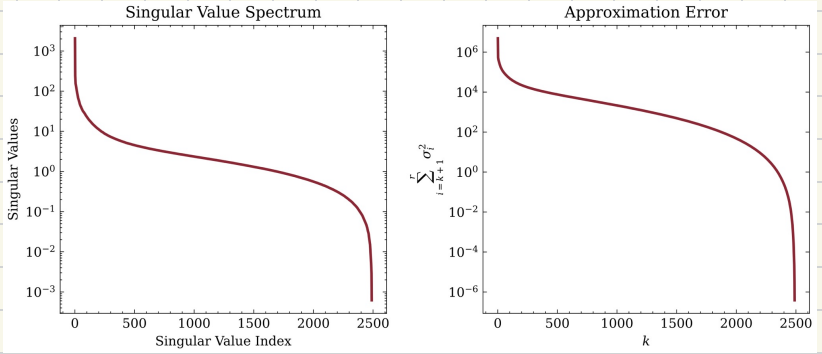
$$\begin{bmatrix} | \\ v_i \\ | \end{bmatrix} \begin{bmatrix} - v_i^T - \end{bmatrix}$$

$$\tilde{X} = \sum_{j=1}^{r} u_j \sigma_j v_j^T \sum_{i=1}^{k} v_i v_i^T$$

$$= \sum_{j=1}^{k} u_j \sigma_j v_j^T = X_k$$

$X W_0 W_1 = X_k$ reconstruction



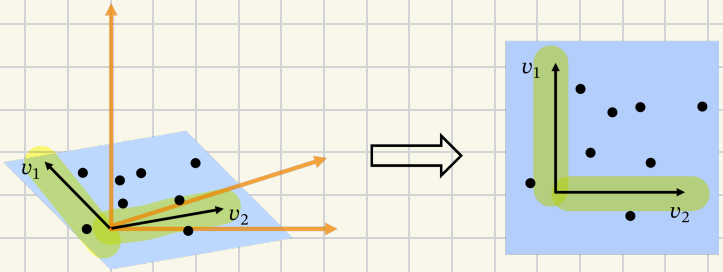


Latent Representation

$X W_0 = X V_k$ latent representation

$n \times d \quad d \times k \qquad n \times k$

$$z^{(i)} = x^{(i)T} V_k = \sum_{j=1}^{k} \langle x^{(i)}, v_j \rangle$$



V_k principal components

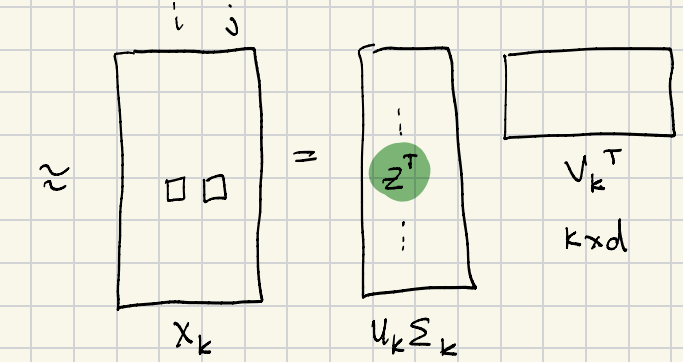
Semantic Embeddings

$X \in \mathbb{R}^{n \times d}$ "document-word" matrix

d words and n documents

$X_{ij} = \mathbb{1}[$ word j appears in doc $i]$

	cat	dog	kitten	puppy	great	good
doc_1	1		1	1		1
doc_2		1			1	
doc_3		1		1		
⋮						



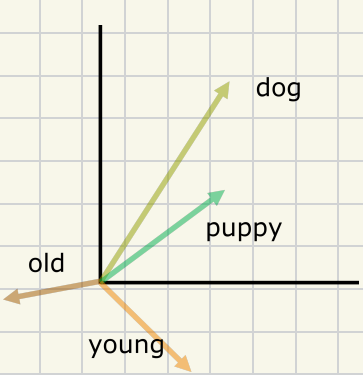
z^T latent representation of document

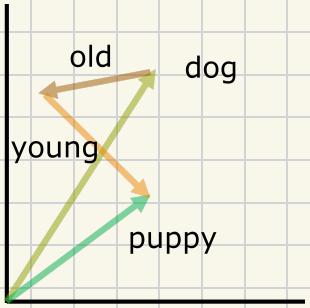
v_i, v_j principal components of words

If $\langle z, v_i \rangle \approx 1$ and $\langle z, v_j \rangle \approx 1$, v_i and v_j are close!

If v_i and v_j are in the same doc, v_i and v_j are close!

Word Math





$$\|Q\,a\,\|_2^2 = a^T Q^T Q\,a$$

Unsupervised Translation