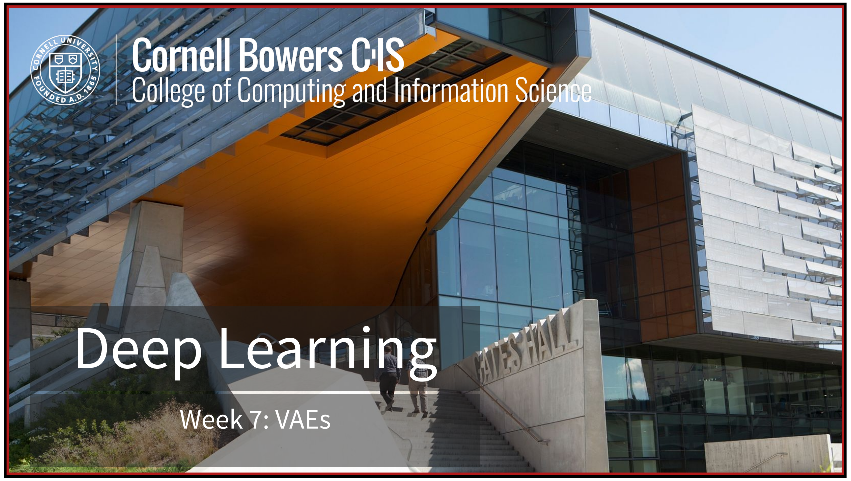


Cornell Bowers C·IS

Discriminative Models

typically supervised

Goal: model $p(Y|X)$
from samples of p(X,Y)
(* so that we can list most likely labels)

Generative Models

typically unsupervised

Goal: model $p(X)$
from samples
(* so that we can draw new random samples)

no labels Y!

Questions:

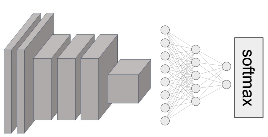
- Does one reduce to the other?
- Which is more difficult?

Cornell Bowers C·IS

Discriminative Models

typically supervised

Goal: model $p(Y|X)$
from samples of p(X,Y)
(* so that we can list most likely labels)



Generative Models

typically unsupervised

Goal: model $p(X)$
from samples
(* so that we can draw new random samples)

no labels Y!

Examples:

- GANs + variants
- Normalizing Flow Models
- Variational Autoencoders
    - Diffusion Models

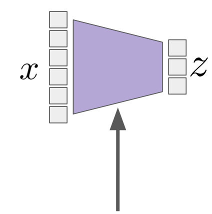
Cornell Bowers C·IS

Dimensionality Reduction

Want to compress image $x \in \mathbb{R}^D$
to code $z \in \mathbb{R}^d$

for the purposes of

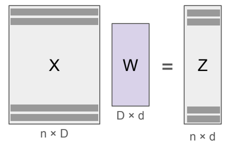
- visualization
- extracting important features (for downstream tasks)
- a more useful space, where geometry has semantic meaning



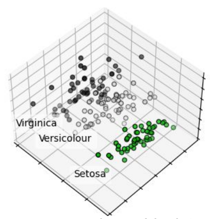
What properties should this mapping have?

Cornell Bowers C·IS

Principal Component Analysis (PCA)



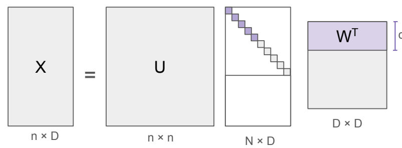
- a linear transformation
- that capture as much variance as possible
- the components of z are independent



`sklearn` demo: Iris dataset
(4 features, 3 classes)

Can be computed directly with linear algebra: take leading eigenvectors of X!

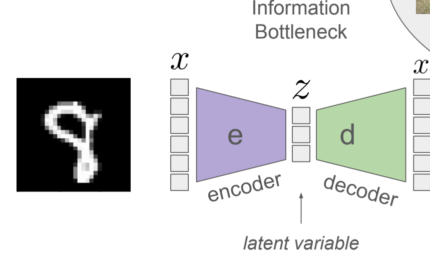
SVD:



Cornell Bowers C·IS

Autoencoders

Remember cycle consistency?



Question: What loss function should we use?

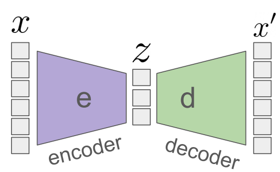
Cornell Bowers C·IS

Reconstruction Loss, first attempt

- "the obvious loss"

$$\sum_{x \in \mathcal{D}} (x - x')^2$$

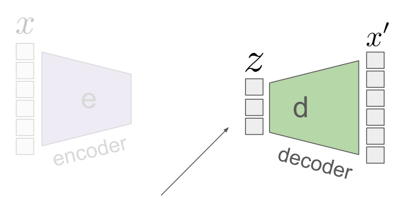
where $x' = e(d(x))$



The Result: an Autoencoder.
[Kramer, 1991]

Cornell Bowers C·IS

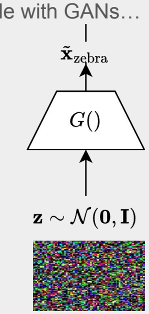
Sampling from an Autoencoder



$$z \sim \mathcal{N}(0, I)$$

feed decoder
(Gaussian) noise?

Recall how we could sample with GANs…



Autoencoder trained on MNIST: latent space

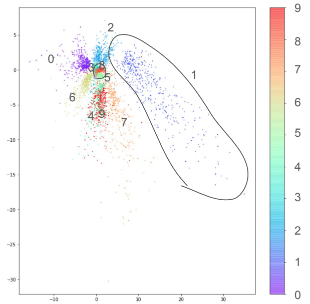


Figure 3-8. Plot of the latent space, colored by digit

Not a very nice representation…

- no symmetries between digit representations
- lots of empty space

Question:
What does this mean for sampling?

What's needed is some kind of "regularization"

to "encourage" the encoder to have "nice properties"...

- Contractive Autoencoders [2011]
- Sparse Autoencoder [2013]
- Variational Autoencoders [2014]

A Probabilistic Perspective

Building Blocks:

- Conditional and marginal probabilities
- Surprisal / Negative Log Likelihood
- Relative Entropy / KL Divergence

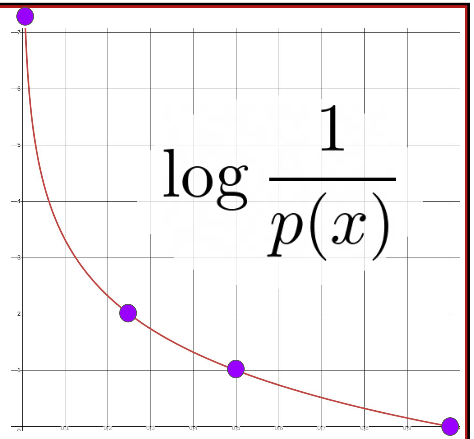
Conditional and Marginal Probabilities

$$p(X,Y) = p(Y|X)p(X)$$

$$p(X) = \int p(X,y)\,\mathrm{d}y$$

Cornell Bowers C·IS

Negative Log Probability
= “Surprisal”



$$\log \frac{1}{p(x)}$$

how surprised you are (bits)

probability p(x)

Cornell Bowers C·IS

KL Divergence (a.k.a. relative entropy)

$$\boldsymbol{D}(\, p \,\|\, q \,) := \underset{x \sim p}{\mathbb{E}} \left[\log \frac{p(x)}{q(x)} \right]$$

reality (e.g., dataset)

model

$$= \underbrace{\underset{x \sim p}{\mathbb{E}} \left[\log \frac{1}{q(x)} \right.}_{\text{Cross Entropy!}} - \underbrace{\left. \log \frac{1}{p(x)} \right]}_{\text{(constant; does not depend on model q)}}$$

- non-negative $\boldsymbol{D}(\, p \,\|\, q \,) \geq 0$
- zero means same $\boldsymbol{D}(\, p \,\|\, q \,) = 0 \iff p = q$
- not symmetric
- has many other, uniquely nice properties …

Cornell Bowers C·IS

KL Divergence

Justin’s Coin

Varsha’s Coin

× 2

Question:
Is it just as easy to mistake the output of Justin’s coin for that of Varsha’s coin, as vice versa?

[Link to visualization]

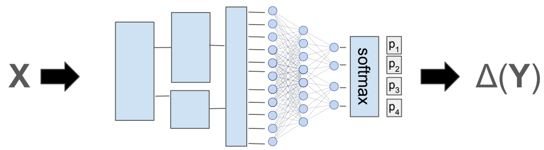
Cornell Bowers C·IS

Building Blocks:

✅ Conditional and marginal probabilities
✅ Surprisal / Negative Log Likelihood
✅ Relative Entropy / KL Divergence

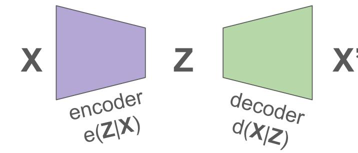
Neural Networks as Conditional Probabilities

A network with a softmax encodes a conditional probability distribution



p(**Y**|**X**) : **X** → Δ(**Y**)

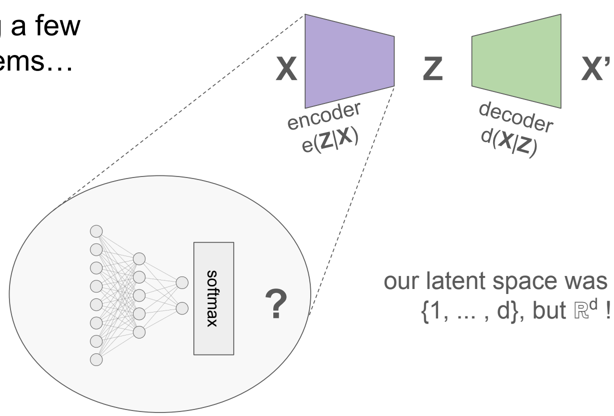
Reconstruction Loss, using surprisal



How surprising would it be to encode x, decode the result, and recover x?

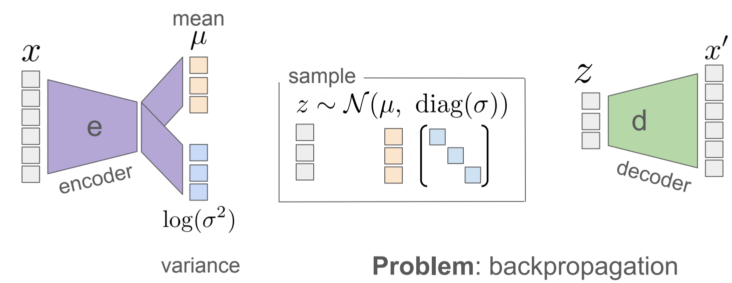
$$\mathop{\mathbb{E}}_{z \sim e(x)} \left[\log \frac{1}{d(x|z)} \right]$$

Fixing a few problems…



our latent space was not {1, ... , d}, but $\mathbb{R}^d$!

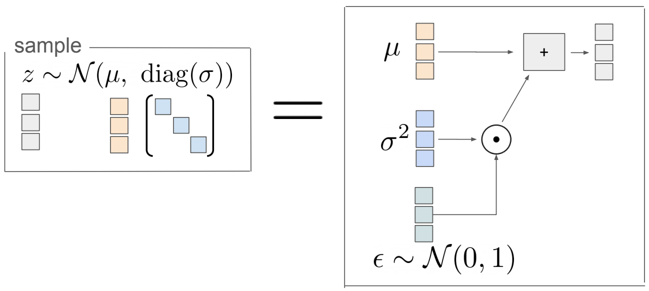
An Architecture for Gaussians



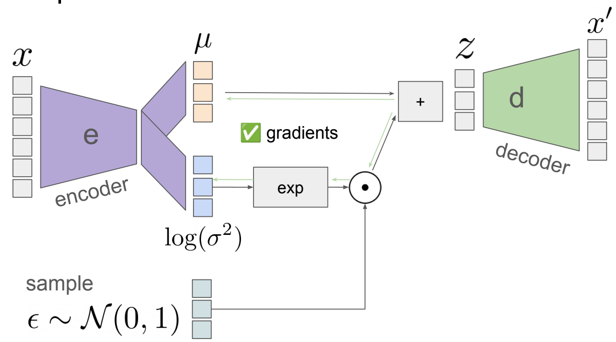
Problem: backpropagation through sampling process?

The Reparameterization Trick

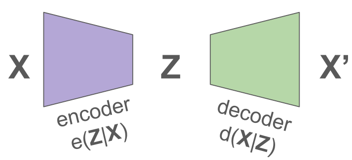
$$\mathcal{N}(\mu, \mathrm{diag}(\sigma)) = \mu + \sigma^2 \odot \mathcal{N}(0, I)$$



The Reparameterization Trick



Reconstruction Loss, using surprisal



How surprising would it be to encode x, decode it, and recover the same sample?

$$\mathop{\mathbb{E}}_{z\sim e(x)}\left[\cancel{\log}\,\frac{\cancel{1}}{d(x|z)}\right]$$

$$\propto \cancel{\exp}\left(\cancel{-}\frac{1}{2}(x - f(z))^2\right)$$

Essentially MSE, again!

We're back at an autoencoder, but probabilistic

The upshot: we can now add a regularization term

$$\boldsymbol{D}(\ e(Z|x)\ \|\ p(Z)\ )$$

Want each encoding … to match a prior (e.g., a standard Gaussian)

Questions:
Does this have a connection to PCA?
Is there a conceptual problem with this regularization?

Cornell Bowers C·IS

Variational Inference

Cornell Bowers C·IS

Motivating VAEs

- Have joint model $p(X, Z)$
- observe x (but not z);
- want to calculate posterior $p(Z|x) = \dfrac{p(x, Z)}{p(x)}$,
- which requires $p(x) = \displaystyle\int p(x, z)\, \mathrm{d}z$
    - i.e., the "evidence".
    - But the integral is often intractable!

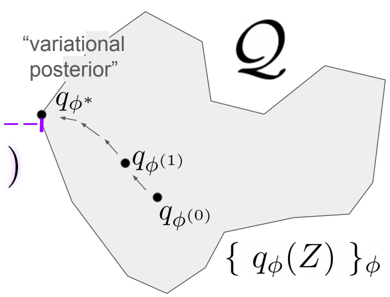
- so, instead …

Cornell Bowers C·IS

Optimize!

"true posterior"

$p(Z|x)$



$$\min_\phi \ \boldsymbol{D}(\, q_\phi(Z) \,\|\, p(Z|x)\,)$$

$$= \mathop{\mathbb{E}}_{Z\sim q_\phi}\left[\log \frac{q_\phi(Z)}{p(Z|x)}\right]$$

$$= \mathop{\mathbb{E}}_{Z\sim q_\phi}\left[\log \frac{q_\phi(Z)}{p(Z|x)} \cdot \frac{p(x)}{p(x)}\right]$$

$$= \underbrace{\mathop{\mathbb{E}}_{Z\sim q_\phi}\left[\log \frac{q_\phi(Z)}{p(x, Z)}\right]}_{\text{Tractable; -ELBO}(x,\phi)} + \underbrace{\boxed{\log p(x)}}_{\text{(constant; does not depend on } \phi\text{)}}$$

Cornell Bowers C·IS

Variational Bound

$$\boldsymbol{D}(\, q_\phi(Z) \,\|\, p(Z|x)\,) = \underbrace{\mathop{\mathbb{E}}_{Z\sim q_\phi}\left[\log \frac{q_\phi(Z)}{p(x, Z)}\right]}_{\text{- ELBO(x)}} + \log p(x)$$

$$\mathrm{ELBO}_{p,\phi}(x) + \underbrace{\boldsymbol{D}(\, q_\phi(Z) \,\|\, p(Z|x)\,)}_{\text{non-negative}} = \log p(x)$$

$$\mathrm{ELBO}_{p,\phi}(x) \leq \underbrace{\log p(x)}_{\text{"(log) evidence"}}$$

Cornell Bowers CIS

What does this have to do with autoencoders?

Replace:

$$q(Z) \rightsquigarrow e(Z|\boxed{X})$$

$$p(X,Z) \rightsquigarrow \underbrace{p(Z)}_{\text{prior / regularizer}} d(Z|X)$$

$$-\mathrm{ELBO}_{p,d,e}(x) = \mathop{\mathbb{E}}_{Z\sim e_\phi(x)}\left[\log \frac{e_\phi(Z|x)}{p(Z)d(x|Z)}\right]$$

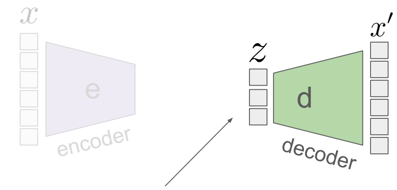
$$\xrightarrow{p(Z)} z \xrightarrow{d(X|Z)} x$$

Questions:

1. What is the evidence, in this case?
2. Is there something strange about it?

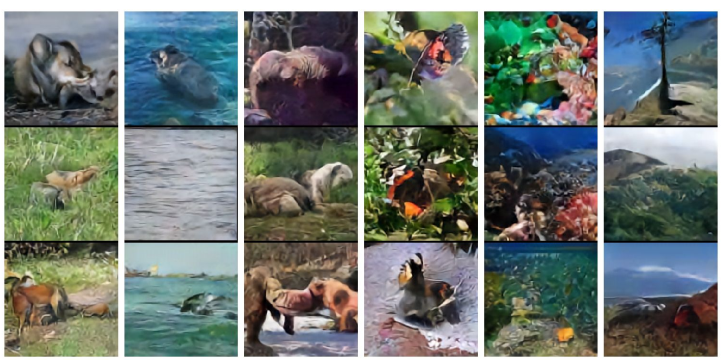
Cornell Bowers CIS

Sampling from a VAE



$$z \sim p_{\text{rior}} = \mathcal{N}(0,1)$$

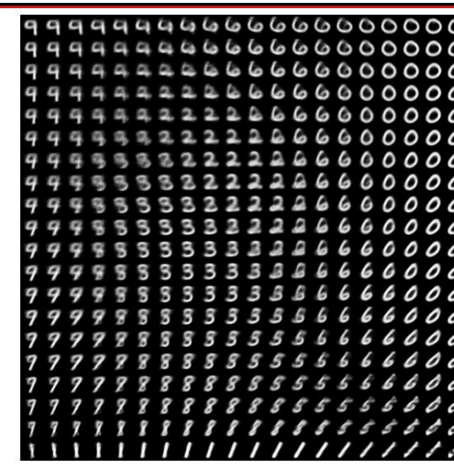
Cornell Bowers CIS



Cornell Bowers CIS

a much nicer space…

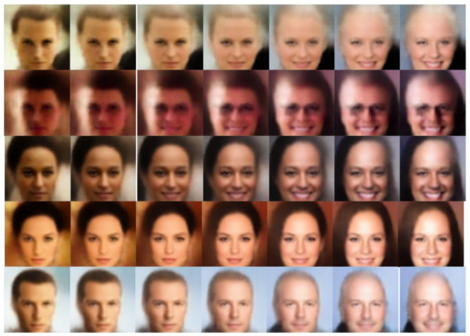
can smoothly interpolate digits in a meaningful, digit-y kind of way



Cornell Bowers C·IS

a much nicer space

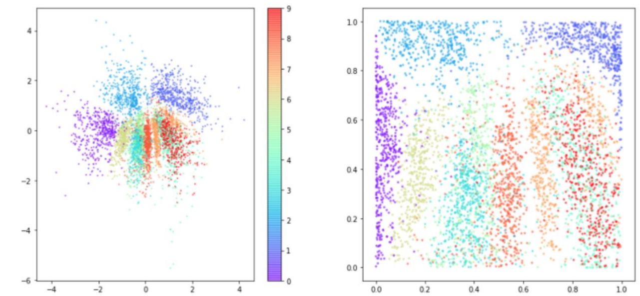
dimensions in latent space correspond to meaningful concepts, like sentiment and orientation



Cornell Bowers C·IS

Back to MNIST: Visualizing latent space again

VAE Latent space, note the distribution is centered, and each digit has an equal portion

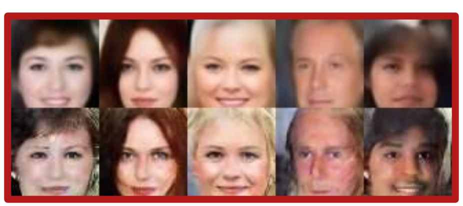


Cornell Bowers C·IS

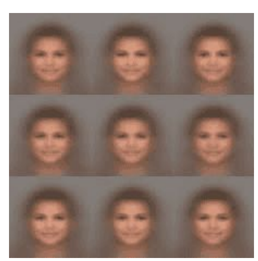
The Biggest Drawback of VAEs

- Out of the box, generated images can be blurry.

  Question: Why?



VAE v. GAN



https://borisburkov.net/2022-12-31-1/

Cornell Bowers C·IS

Hierarchical VAEs

The generative process is modeled as a Markov chain, where each latent z_t is generated only from the previous latent z_{t+1}