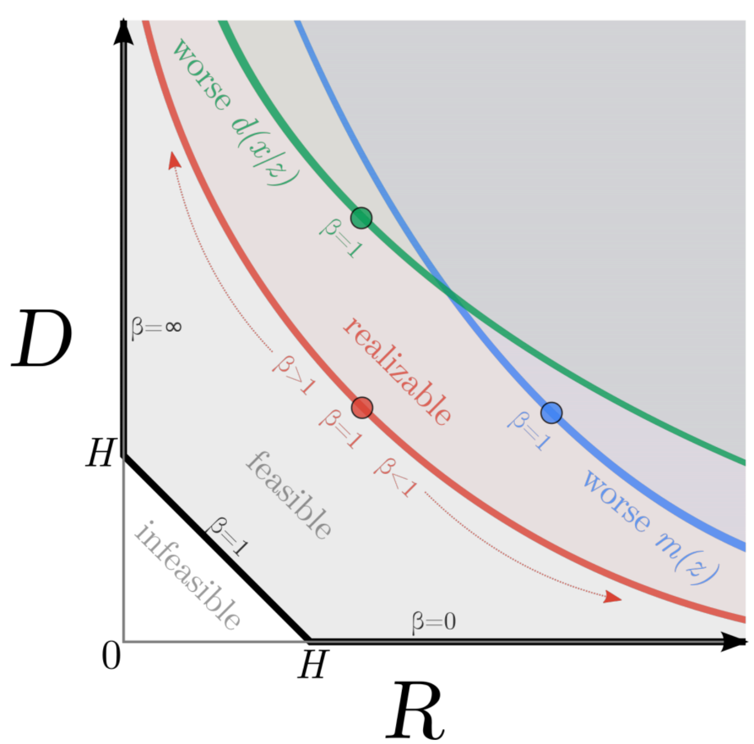
Information Theory for Deep Latent Variable Models

Gokul Swamy

Fixing a Broken ELBO

- Today, we're going to be talking about https://arxiv.org/pdf/1711.00464.pdf

- **Key Insight**: We can use rate and distortion to define a Pareto-optimal frontier for latent variable models. We can then select from this frontier based on our application.



Information Theory

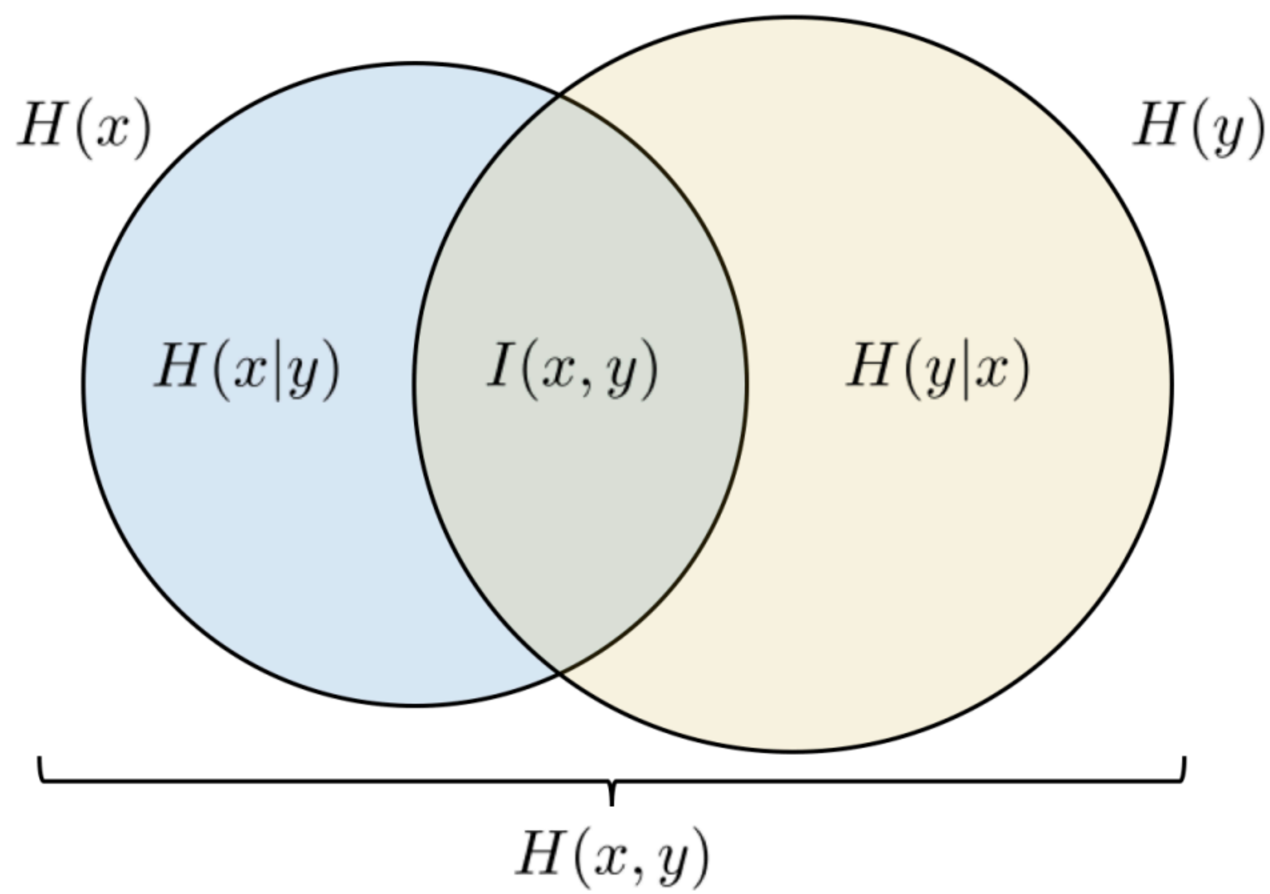
- *Information Theory* is a science of inequalities
- The goal is to define bounds on how well we can do, not figure out how to achieve them
  - That is *coding theory*, which is an almost orthogonal field
- Shannon's *A Mathematical Theory of Communication* outlines bounds for 2 problems:
  - **Reliable Communication**
  - **Lossy Compression**
- Most results will be stated without proof today, read Cover & Thomas for justification

Information Measures

- These quantities will keep popping up so let's name them
- Entropy: $H(X) = \sum_{x \in X} p_X(x) * \log \frac{1}{p_X(x)} = E[\log \frac{1}{p_X(x)}]$
- Conditional Entropy: $H(X \mid Y) = E_{X,Y}[\log \frac{1}{p_{X|Y}(x)}] = H(X, Y) - H(Y)$
- Mutual Information: $I(X; Y) = I(Y; X) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$

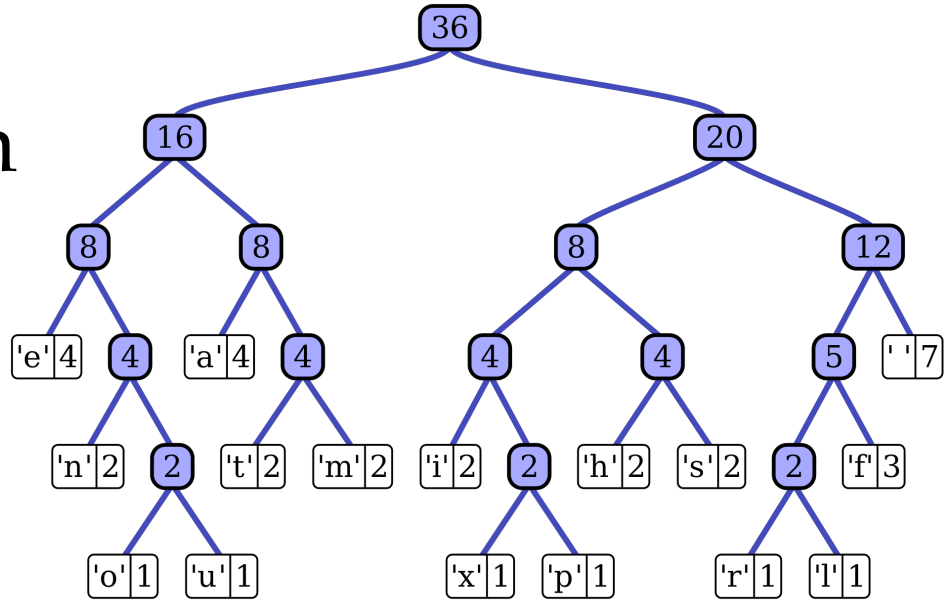
  $I(X; Y) = = H(X) + H(Y) - H(X, Y)$
- Relative Entropy: $D_{KL}(p \mid\mid q) = \sum_i p_i * \log \frac{p_i}{q_i} \neq D_{KL}(q \mid\mid p) \geq -\log 1 = 0$
  - Not a metric

Information Diagram



What is Entropy?

- Warm, fuzzy intuition is that it measures the randomness or unpredictability of a distribution.
- Minimum expected description length for a random variable
  - Achievable-ish using *Huffman Coding*: start out with singletons and construct tree by combining two nodes with lowest probability.
- Used in RL to incentivize exploration



Maximum Entropy

- What is the maximum entropy distribution over alphabet X?
  - Uniform: $0 \leq D(p\|u) \equiv \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = -H(X) + \log |\mathcal{X}|$.
- What is the maximum entropy distribution given a specific mean?
  - Exponential $f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$
- What is the maximum entropy distribution given a specific variance?
  - Gaussian
    - This is why the Central Limit Thm. Holds
    - Also why it is useful for physical noise ($\frac{1}{2}mv^2$)

Inverse Reinforcement Learning

- Reinforcement Learning: find actions that maximize reward
- Inverse Reinforcement Learning: find reward function that would have made actions taken optimal
  - Standard Recipe: write $R(s) = w^T \theta(s)$ and maximize over w
    - Guarantees match in expected feature counts after convergence
    - This requires optimality of demonstrator

MaxEnt Inverse Reinforcement Learning

- To relax this constraint, why don't we assume people behave as randomly as possible while still having the same expected feature counts (first moment constraint)
  - Boltzmann Rationality:

$$\mathbb{P}((s_i, a_i)|R) = \frac{1}{Z_i}\exp\{\alpha Q^*(s_i, a_i, R)\}$$

  - Then, we apply Bayesian Inference to recover reward function

$$\mathbb{P}(\tau|R) = \prod_{i=1}^{n}\mathbb{P}((s_i, a_i)|R) \quad \mathbb{P}(\tau|R) = \frac{1}{Z}\exp\{\alpha\sum_{i=1}^{n}Q^*(s_i, a_i, R)\}$$

$$\mathbb{P}(R|\tau) = \frac{\mathbb{P}(\tau|R)\mathbb{P}(R)}{\mathbb{P}(\tau)} = \frac{1}{Z}\exp\{\alpha\sum_{i=1}^{n}Q^*(s_i, a_i, R)\}\mathbb{P}(R)$$

Data Processing Inequality

- Assume we have 3 variables that form a Markov Chain. Then,

$$X \rightarrow Y \rightarrow Z \implies I(X;Z) \leq I(X;Y)$$

- Now, consider Y being your latent representation. What does this imply?

Reliable Communication

- Shannon's Block Diagram:

$$\text{Message} \longrightarrow \text{Encoder} \xrightarrow{\text{signal}} \text{Channel} \xrightarrow[\text{signal}]{\text{corrupted}} \text{Decoder} \longrightarrow \text{Message}$$

- Let **rate** of (M messages, n length)-block code be defined as

$$R \equiv \frac{1}{n} \log M$$

- Source Coding Theorem:

$$C \equiv \sup\{R | R \text{ is achievable}\}$$

$$C = \sup_{P_X} I(X;Y)$$

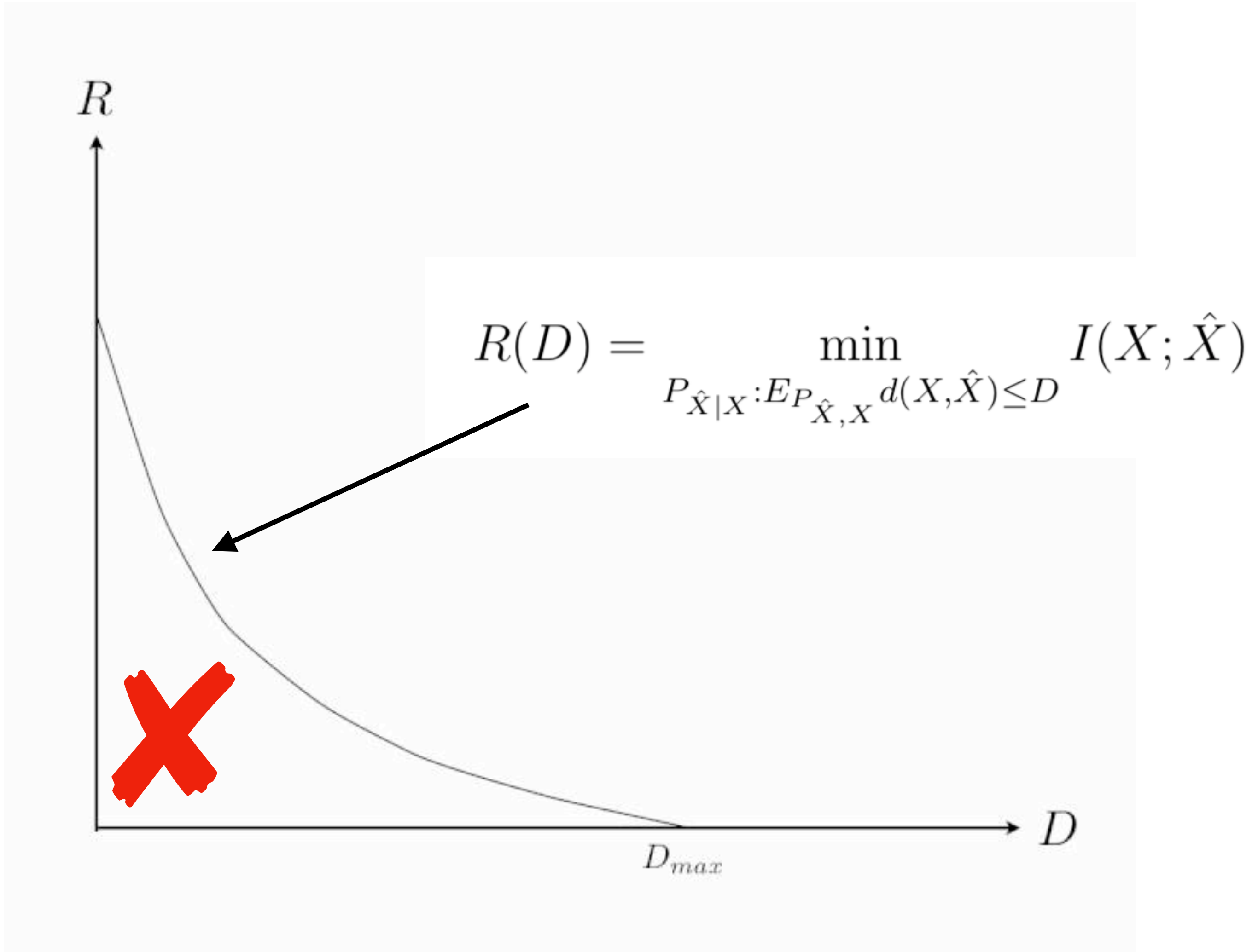
Lossy Compression

- We wanted perfect reconstruction before - what if we relax that constraint?
- We tolerate some level of distortion now. For example, Hamming distance: $d(X, \hat{X}) = (X - \hat{X})^2$
- Let f be our encoder, g be our decoder. Then, the expected distortion of this pair is D $= \sum_{x^n} p(x^n) \cdot d(x^n . g_n(f_n(x^n)))$
- Then, the rate-distortion theorem tells us that the lowest rate (best compression) we can send at is:

$$R(D) = \min_{P_{\hat{X}|X} : E_{P_{\hat{X},X}} d(X,\hat{X}) \le D} I(X; \hat{X})$$

Rate-Distortion Functions

- The rate-distortion defines the Pareto-optimal frontier for the problem:

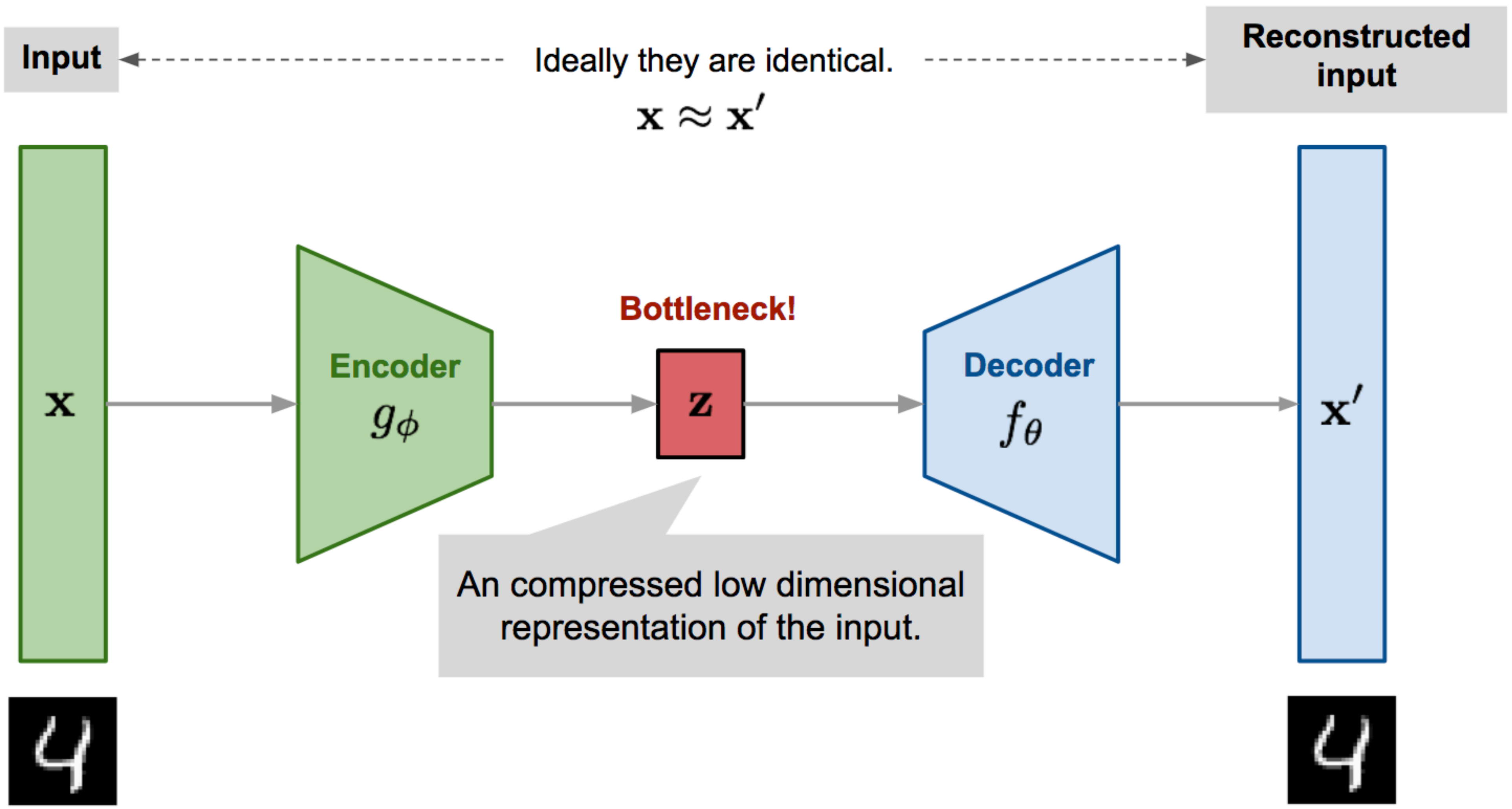


Source Coding vs. Rate-Distortion

- Source coding is effectively a *sphere packing problem* - how can we define a sphere around each symbol such that we have the minimum overlap (misinterpretations)
- Rate distortion is effectively a covering problem - how can we waste as little space (representations) as possible so we can compress as effectively as possible
- These are dual problems of each other

Vanilla Autoencoders

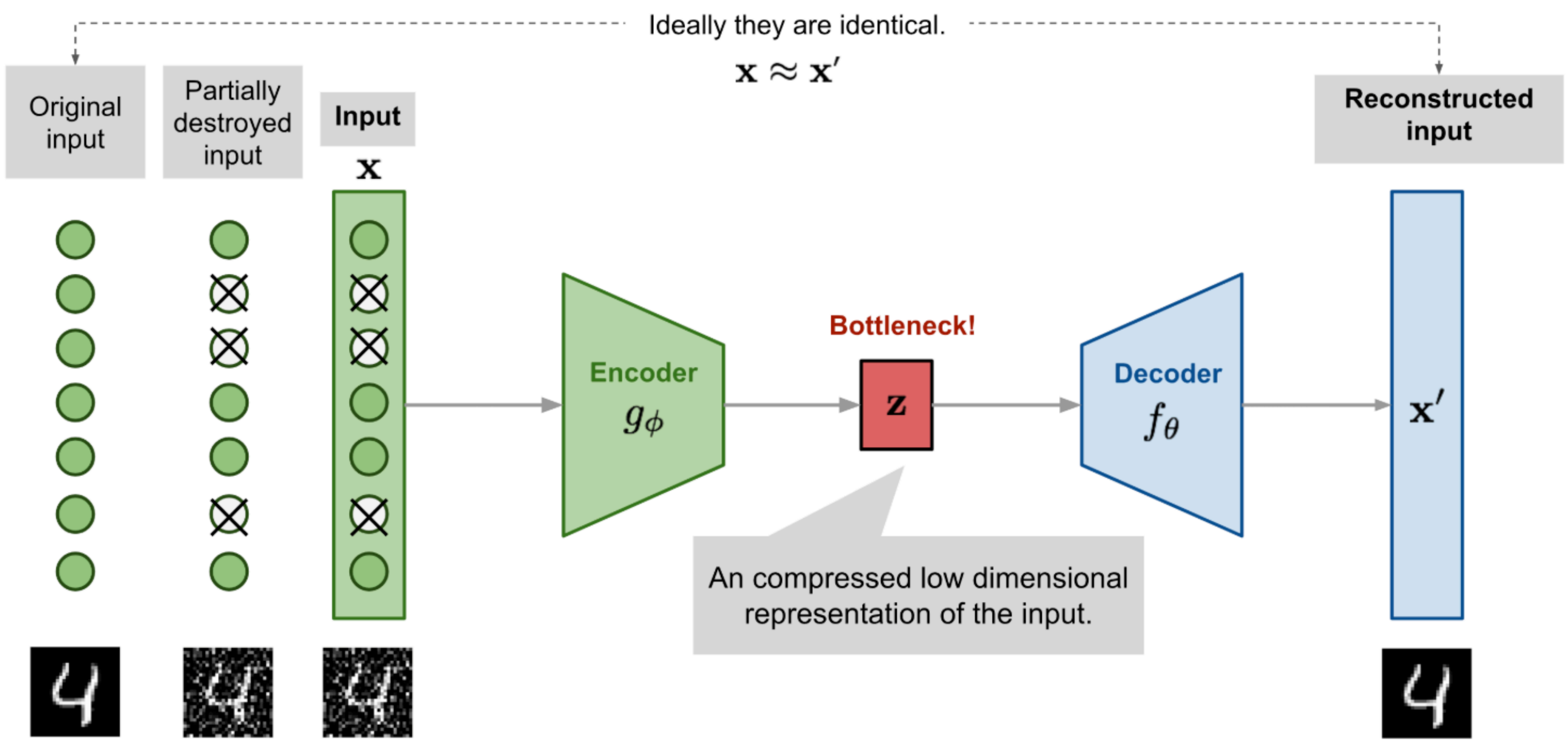
- "Unsupervised"



$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

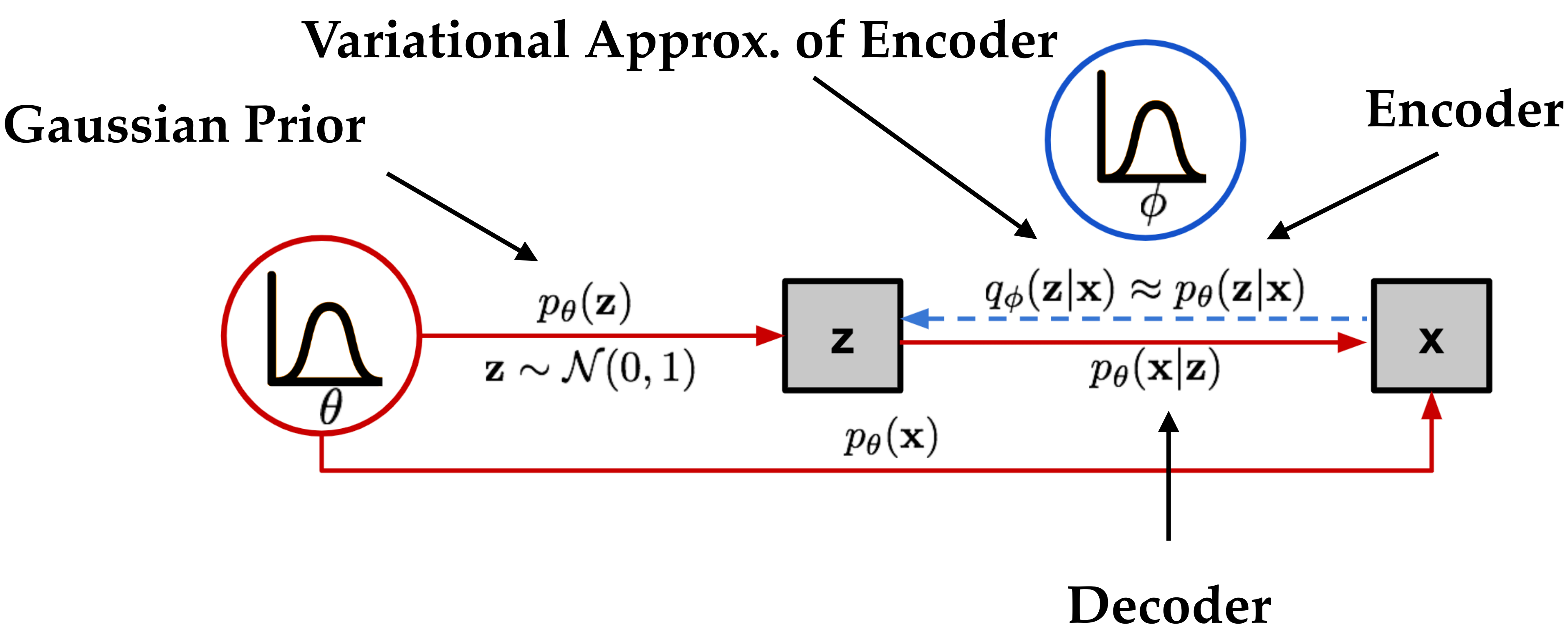
Robustifying Autoencoders

- Just add noise?



Variational Autoencoders

- What if instead of simply compressing data, we want to be able to generate new data
  - Idea: cast generation as a problem of sampling from a tractable (ideally high entropy) distribution and then transforming sample
  - This gives us the probabilistic encoder-decoder structure used in rate-distortion theory



Learning a VAE

- What optimization problem are we trying to solve?

$$\theta^* = \arg\max_\theta \prod_{i=1}^{n} p_\theta(\mathbf{x}^{(i)})$$

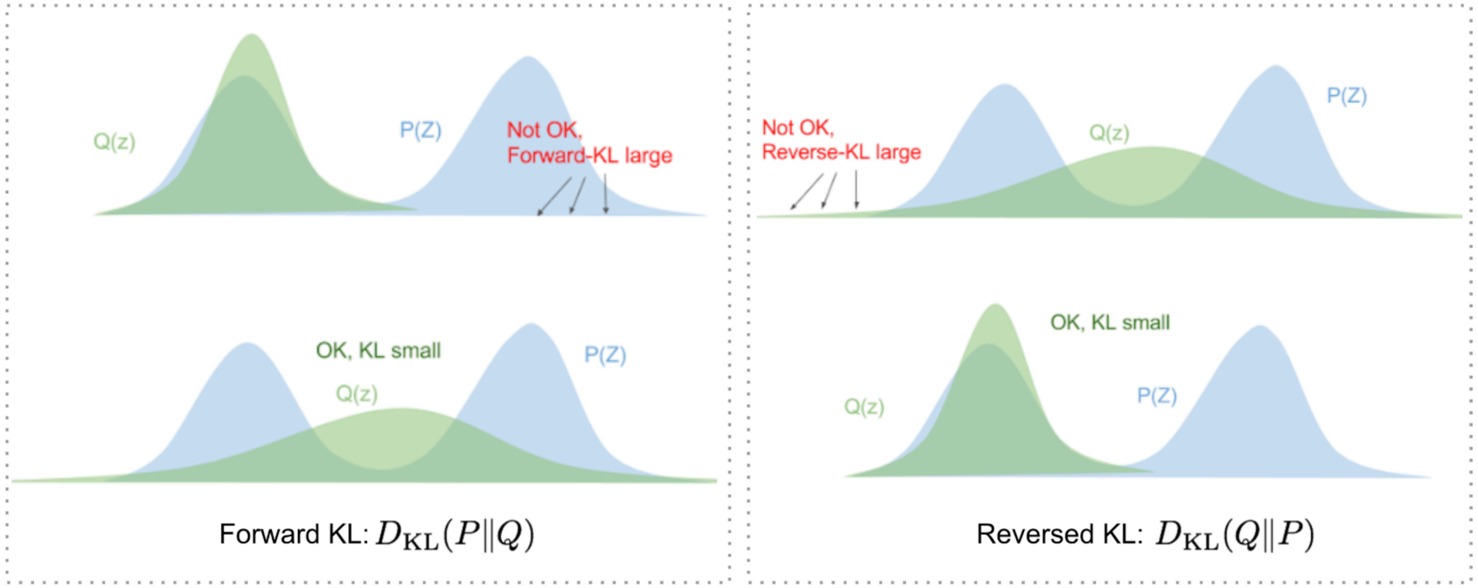
$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(\mathbf{x}^{(i)})$$

$$p_\theta(\mathbf{x}^{(i)}) = \int p_\theta(\mathbf{x}^{(i)}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$$

- This is very hard because we need to evaluate an integral for every possible decoder
    - Idea: reduce search space through a variational approximation

Choosing a Variational Approximation

- Effectively, we are performing a projection onto some space of nicely parameterizable distributions
- What distance metric should we use to define closest?
  - Reversed KL divergence: incentivizing covering of p



Evidence Lower Bound

- Now, we have a function to minimize:

$$q_\lambda^*(z \mid x) = \arg\min_\lambda \mathbb{KL}(q_\lambda(z \mid x) \mid\mid p(z \mid x)).$$

- How do we make it tractable to compute?

$$\mathbb{KL}(q_\lambda(z \mid x) \mid\mid p(z \mid x)) =$$

$$\mathbf{E}_q[\log q_\lambda(z \mid x)] - \mathbf{E}_q[\log p(x, z)] + \log p(x)$$

  - p(x) is the problem (this is what we're trying to find a tractable approximation to in the first place so it can't be in our lost function)

Evidence Lower Bound

- Define the ELBO as follows:

$$ELBO(\lambda) = \mathbf{E}_q[\log p(x, z)] - \mathbf{E}_q[\log q_\lambda(z \mid x)]$$

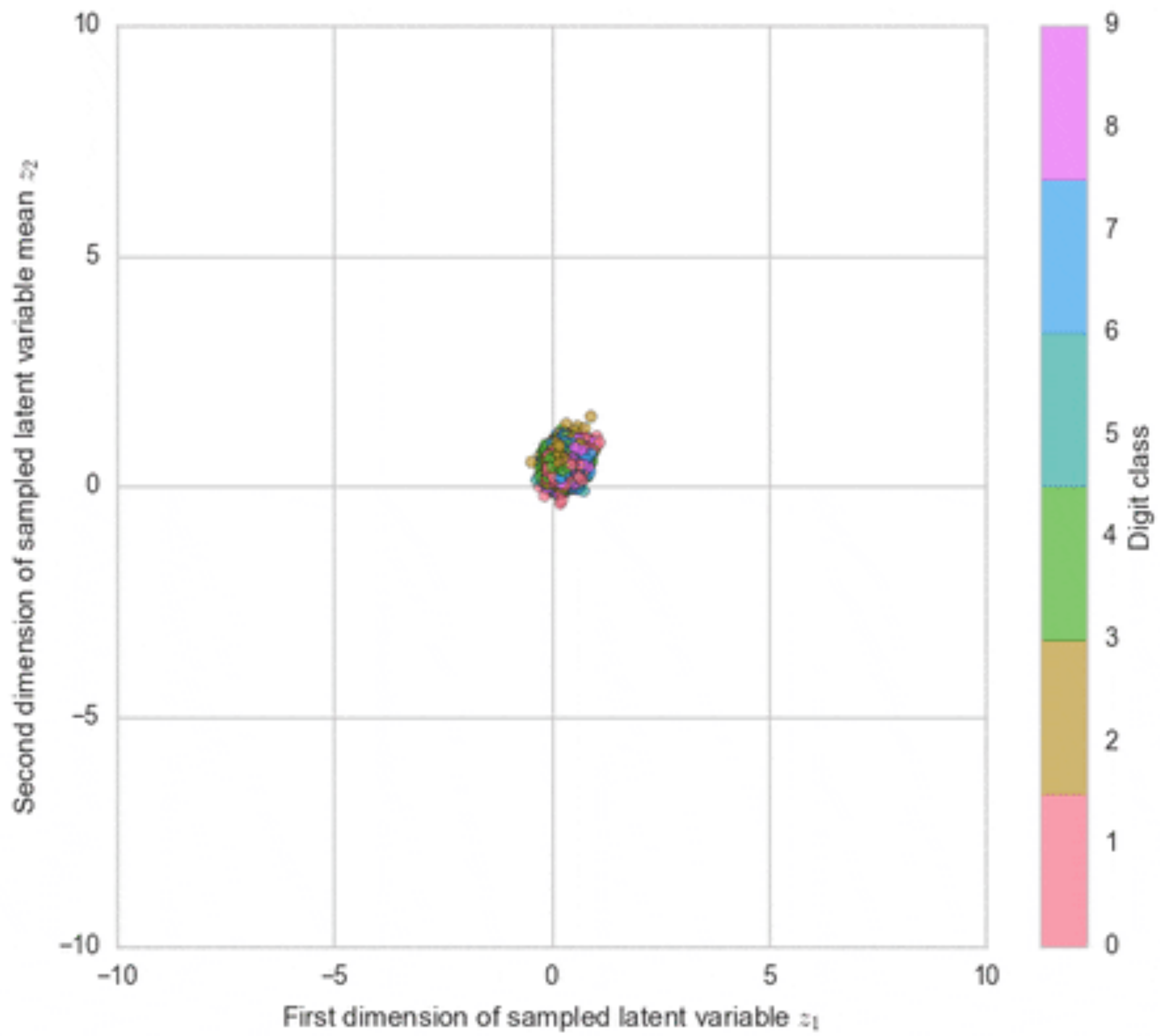
- Then, we can rewrite p(x) as

$$\log p(x) = ELBO(\lambda) + \mathbb{KL}(q_\lambda(z \mid x) \mid\mid p(z \mid x))$$

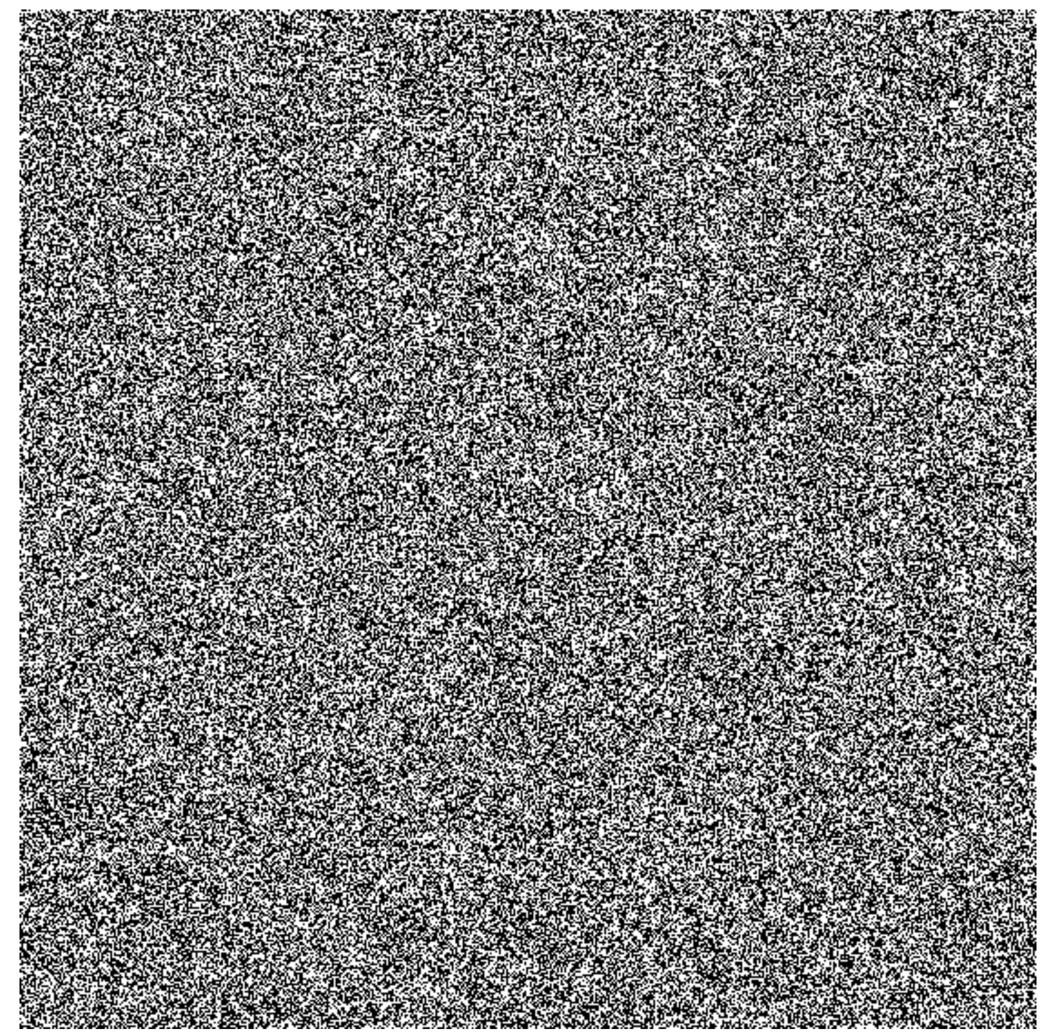
- KL is always non-negative so we can minimize it by maximizing the ELBO because of the above constraint.
- Thus, rearranging terms, the loss for a single data point becomes the negative of:

$$ELBO_i(\theta, \phi) = \mathbb{E}q_\theta(z \mid x_i)[\log p_\phi(x_i \mid z)] - \mathbb{KL}(q_\theta(z \mid x_i) \mid\mid p(z))$$

Hype GIFs



10
5
0
−5
−10
−10
−5
0
5
10
Second dimension of sampled latent variable mean z_2
First dimension of sampled latent variable z_1
9
8
7
6
5
4
3
2
1
0
Digit class



Fixing a Broken ELBO

- Our old friends return:

$$H \equiv -\int dx\, p^*(x) \log p^*(x)$$
$$D \equiv -\int dx\, p^*(x) \int dz\, e(z|x) \log d(x|z)$$
$$R \equiv \int dx\, p^*(x) \int dz\, e(z|x) \log \frac{e(z|x)}{m(z)}$$

  - H is the entropy of the data
  - D is negative log likelihood of reconstruction (not Hamming)
  - R is excess number of bits to encode samples from encoder using a code designed for m(z)

- This gives us the following bound:

$$H - D \leq \mathrm{I_e}(X; Z) \leq R$$

Beta VAEs

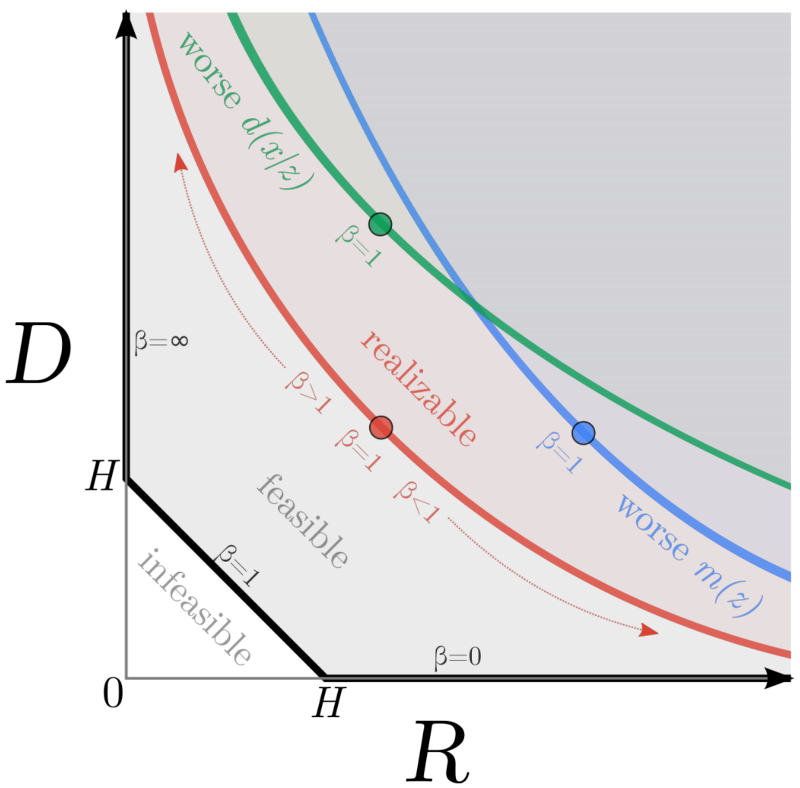
- Let us define a slightly more general version of ELBO:

$$\min_{e(z|x),m(z),d(x|z)} \int dx\, p^*(x) \int dz\, e(z|x) \left[-\log d(x|z) + \beta \log \frac{e(z|x)}{m(z)} \right].$$

- If we set Beta = 1, then we get the traditional ELBO and can see that ELBO = -(D + R)
  - Thus, there are a wide variety of encoder-decoder pairs with the same value of ELBO loss.

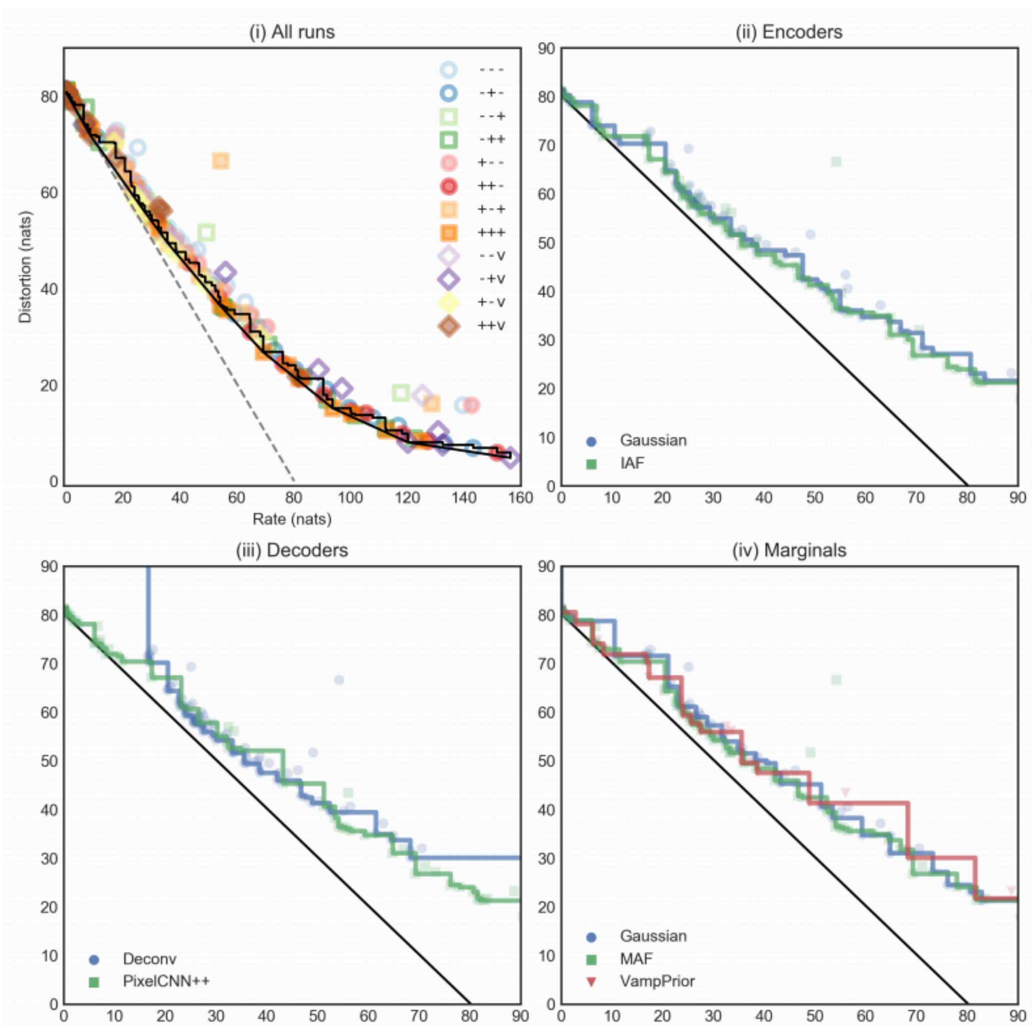
Fixing a Broken ELBO

- We can define a feasible set by using standard rate-distortion theory
- We can define a realizable set by considering our parametric family
- We can control where we fall in the realizable set through manipulation of Beta

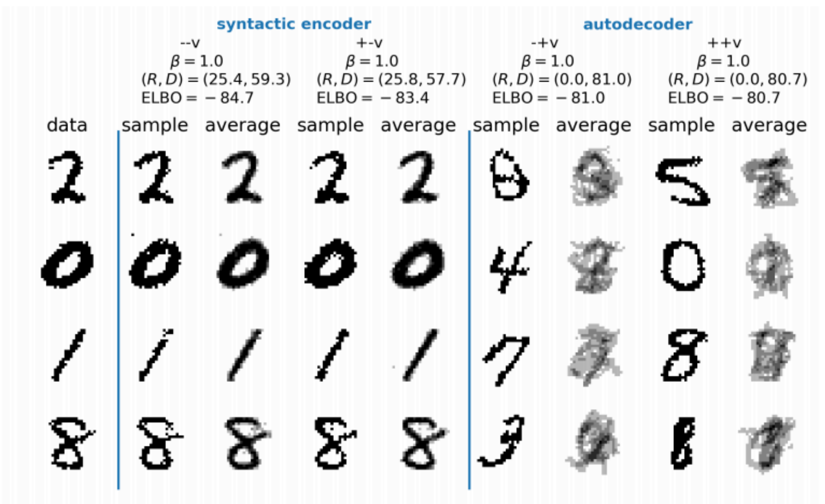


Praxis

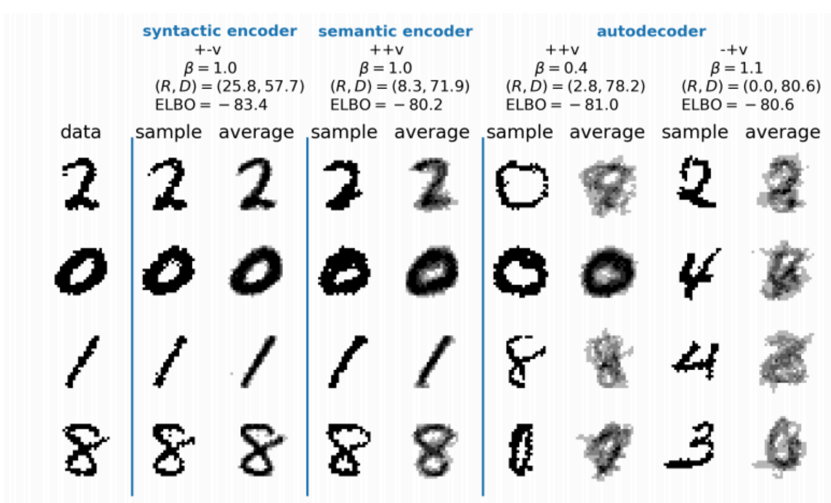
- We can see these tradeoffs experimentally too:



(a) Distortion vs Rate



(c) Reconstructions from 4 VAE models with $\beta = 1$.



(d) Reconstructions from models with the same ELBO.

Information Bottleneck

- Another cool use of mutual information is to make sure that we learn an efficient embedding input

- By the data processing inequality, the maximally informative embedding is the identity

- However, we want a less complex representation so we can constrain the mutual information

$$\max_{\boldsymbol{\theta}} I(Z, Y; \boldsymbol{\theta}) \text{ s.t. } I(X, Z; \boldsymbol{\theta}) \leq I_c$$

$$R_{IB}(\boldsymbol{\theta}) = I(Z, Y; \boldsymbol{\theta}) - \beta I(Z, X; \boldsymbol{\theta})$$

- Here, Beta controls tradeoff between compression and reconstruction.

Questions?

Thanks :)