# Autoencoders

Machine Learning: Jordan Boyd-Graber
University of Maryland
SLIDES ADAPTED FROM IAN GOODFELLOW

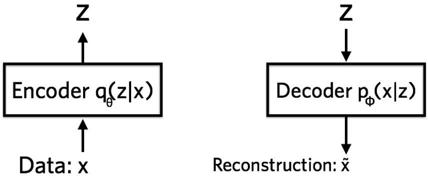
## Problems of Autoencoders

- Unsupervised
  - Lots of data
  - Need priors / regularization
- Probabilistic loss function
  - does not work well for discrete data (more later)
  - hard to explain hidden layer probabilistically

## Problems of Autoencoders

- Unsupervised
  - Lots of data
  - Need priors / regularization
- Probabilistic loss function
  - does not work well for discrete data (more later)
  - hard to explain hidden layer probabilistically
- So let's use variational inference

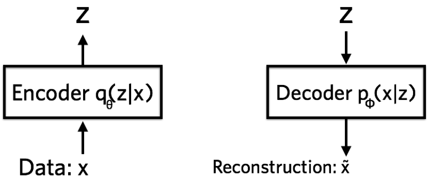
## Loss Function




$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \quad (1)$$

- Reconstruction error
- Variational representation distribution
- Regularization

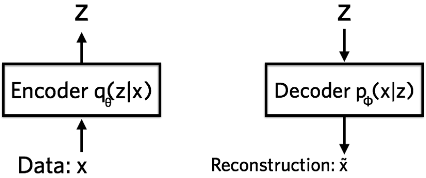
## Loss Function




$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{1}$$

- Reconstruction error
- Variational representation distribution
- Regularization

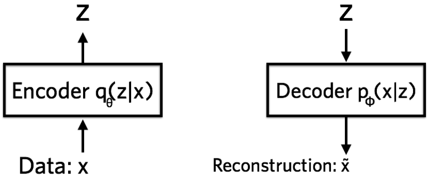
## Loss Function




$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \qquad (1)$$

- Reconstruction error
- Variational representation distribution
- Regularization

## Loss Function




$$\ell_i \equiv -\mathbb{E}_{z \sim q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] + \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \qquad (1)$$

- Reconstruction error
- Variational representation distribution
- Regularization

### Interpretation

- Lower bound on reconstruction of decoder
- Keep representation constrained
- Probabilistic parameterization

## Make this Concrete

- $\text{KL}(q_\theta(z \mid x_i) \,||\, p(z))$
- $q(z \mid x_i)$: normal distribution with output of NN as mean [variational distribution]
- $p(z)$: standard normal distribution
- Decoder $p_\phi(x \mid z)$ depends on model / data:
  - Grayscale Image? Bernoulli distribution for each pixel
  - Words? Multinomial over vocabulary

## Make this Concrete

- $\text{KL}(q_\theta(z \mid x_i) \,||\, p(z))$
- $q(z \mid x_i)$: normal distribution with output of NN as mean [variational distribution]
- $p(z)$: standard normal distribution
- Decoder $p_\phi(x \mid z)$ depends on model / data:
    - Grayscale Image? Bernoulli distribution for each pixel
    - Words? Multinomial over vocabulary

## Make this Concrete

- $\text{KL}(q_\theta(z \mid x_i) \,||\, p(z))$
- $q(z \mid x_i)$: normal distribution with output of NN as mean [variational distribution]
- $p(z)$: standard normal distribution
- Decoder $p_\phi(x \mid z)$ depends on model / data:
  - Grayscale Image? Bernoulli distribution for each pixel
  - Words? Multinomial over vocabulary

## Variational Inference Story

$$\ell_i(\theta) = \mathbb{E}_{q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{2}$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$

## Variational Inference Story

$$\ell_i(\theta) = \mathbb{E}_{q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \quad (2)$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$
- Actually simpler than LDA
  - No global latent variables (only $z$)
  - Can minibatch the data

## Variational Inference Story

$$\ell_i(\theta) = \mathbb{E}_{q_\theta(z \mid x_i)}\left[\log p_\phi(x_i \mid z)\right] - \mathrm{KL}(q_\theta(z \mid x_i) \,||\, p(z)) \tag{2}$$

- Want to optimize $p_\phi(x \mid z)$ (likelihood)
- ELBO remains lower bound
- Difference is KL between variational distribution and $p(z)$
- Actually simpler than LDA
  - No global latent variables (only $z$)
  - Can minibatch the data
  - But what about $\phi$? (encoder)

## Variational EM

- Learn variational parameters
- Update $\phi$ using supervised backprop

## Variational EM

- Learn variational parameters
- Update $\phi$ using supervised backprop
- What if $x$ is discrete? (Later)