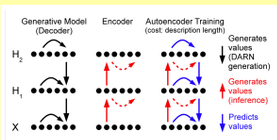
1310.8499

Introduction

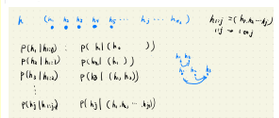
- Two directions
  - observation to representation
  - representation to observation
- probabilistic versions of regularised autoencoders
  - What are auto encoders
  - Before probabilistic version what do we have
  - iterative
  - producing correlated approximate samples from previous approximate samples
- Deep AutoRegressive Networks
  - generate independent, exact samples via ancestral sampling
  - viewing autoencoders through the lens of MDL
    - equivalence of compression and prediction
  - deep autoencoders
  - Minimising the description length
    - minimising the Helmholtz variational free energy
    - encoder plays the role of the variational distribution
  - stochastic gradient descent method
    not an expectation maximisation algorithm
  - stochastic hidden units
    - autoregressive
      each unit receives input both from the preceding layer and the preceding units within the same layer
      - Autoregressive structure
        little computational cost during both learning and generation
  - deterministic nonlinearities

Model Architecture

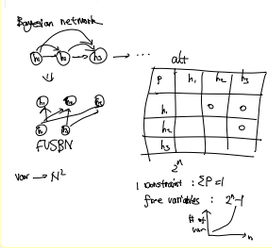
- Architecture
  - 

    - $q(H \mid X)$ is the encoder
    - $p(H)$ is the decoder prior
    - $p(X\mid H)$ is the decoder conditionals

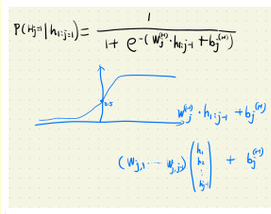
    decode: $P(X|H)P(H)$
    encode: $P(H|X)P(X)$
  - decoder prior representation h

    $$p(h) = \prod_{j=1}^{n_h} p(h_j|h_{1:j-1})$$

  - parameterise the conditional probability mass of hj

    $$p(H_j = 1|h_{1:j-1}) = \sigma(W_j^{(H)} \cdot h_{1:j-1} + b_j^{(H)})$$

    $$p(x|h) = \prod_{j=1}^{n_x} p(x_j|x_{1:j-1}, h)$$

    $$q(h|x) = \prod_{j=1}^{n_x} q(h_j|h_{1:j-1}, x)$$
  - decoder
    $p(X|H)$
    - $p(x|h) = \prod_{j=1}^{n_x} p(x_j|x_{1:j-1}, h)$
      parameterise the conditional probability mass of hj
      - $p(X_j = 1|x_{1:j-1}, h) = \sigma(W_j^{(X|H)} \cdot (x_{1:j-1}, h) + b_j^{(X|H)})$
      - $(x_{1:j-1}, h)$
        concatenation of the vector x1:j-1 with the vector h
        - $q(H_j = 1|h_{1:j-1}, x) = \sigma(W_j^{(H|X)} \cdot x + b_j^{(H|X)})$
  - encoder
    $q(H|X)$
    - $q(h|x) = \prod_{j=1}^{n_x} q(h_j|h_{1:j-1}, x)$
      parameterise the conditional probability mass of hj
      - $q(H_j = 1|h_{1:j-1}, x) = \sigma(W_j^{(H|X)} \cdot x + b_j^{(H|X)})$
        - $(x_{1:j-1}, h)$

Deeper Architectures

- boosting DARN's representational power in three ways
  - Additional stochastic hidden layers
    - $p(H^{(l)}|H^{(l+1)}) = \prod_{j=1}^{n_l^{(l)}} q(H_j^{(l)}|H_{1:j-1}^{(l)}, H^{(l+1)})$
    - $q(H^{(l)}|H^{(l-1)}) = \prod_{j=1}^{n_l^{(l)}} q(H_j^{(l)}|H_{1:j-1}^{(l)}, H^{(l-1)})$
  - Additional deterministic hidden layers
    - $$d^{(l)} = \tanh(U_l h^{(l+1)})$$
      $$p(H_j^{(l)} = 1|h_{1:j-1}^{(l)}, h^{(l+1)}) = \sigma(W_j^{(l)} \cdot h_{1:j-1}^{(l)} + b_j^{(l)}),$$
  - Alternate kinds of autoregressivity

Local connectivity

- scale
  by restricting connectivity
- Local connectivity
  - fully convolutional
  - less weight sharing

Sampling

- A DARN without a stochastic hidden layer but with an autoregressive visible layer is a fully visible sigmoid belief network

Minimum Description Length and Autoencoders

- MDL principle
  finding parameters that try to maximally compress the training data
  - Shannon's source coding theorem
    - the description length is equal to the negative logarithm of the probability of the random variable taking that particular value
  - representation h, the description length
    $L(x|h) = -\log_2 p(x|h).$
  - the expected description length
    $L(x) = \sum_h q(h|x)(L(h) + L(x|h)),$
    - q(h|x) is the encoder probability of the representation h.
    - $L(h) = -\log_2 p(h) + \log_2 q(h|x).$
  - Helmholtz variational free energy
    $L(x) = -\sum_h q(h|x)(\log_2 p(x,h) - \log_2 q(h|x))$