Neural Information Processing Systems Foundation

NIPS 2012 Workshop on

Perturbations, Optimization, and Statistics

Learning robust representations through input perturbations.

From stochastic perturbations to analytic criteria and back.

Pascal Vincent

Université de Montréal

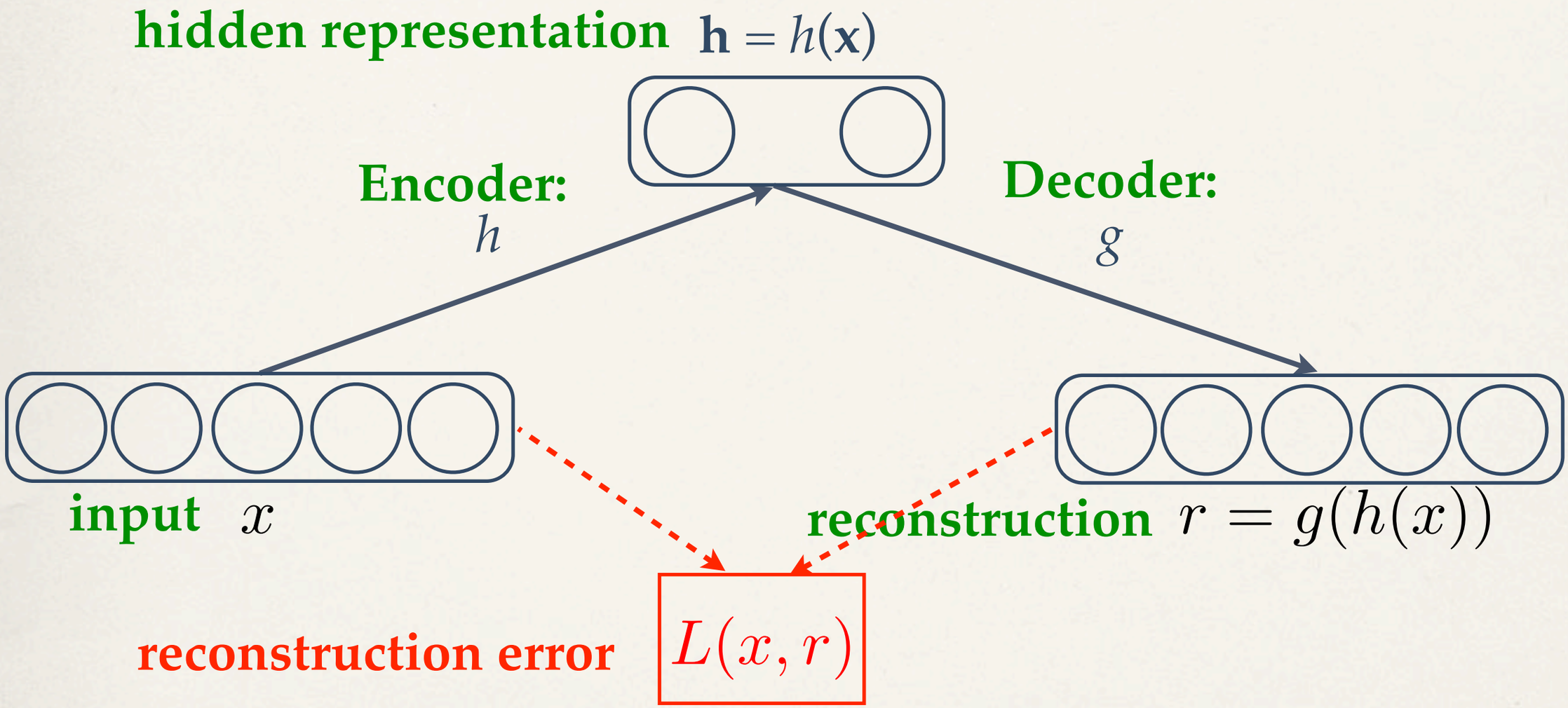
Département d'informatique et de recherche opérationnelle

LISA

Learning of a representation

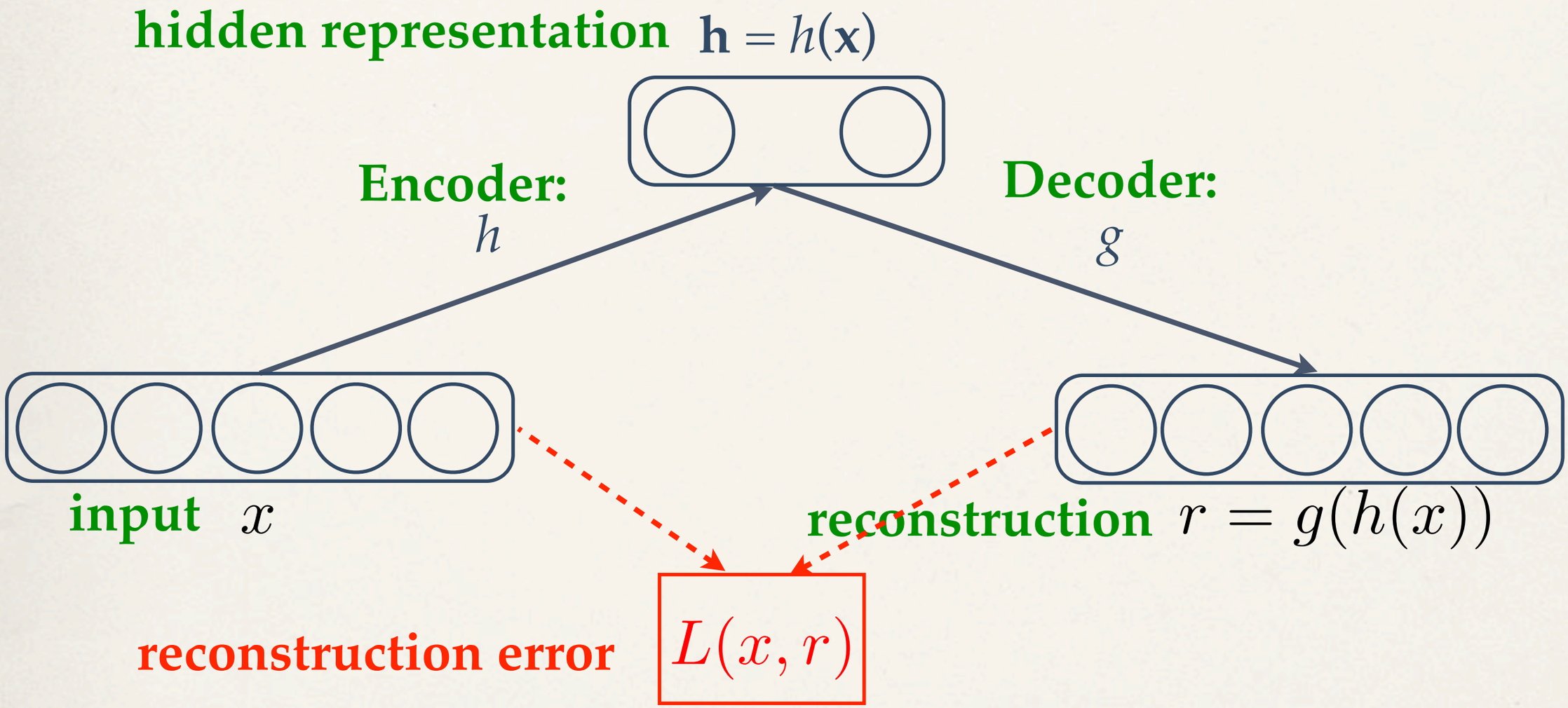
- We want to learn a **continuous deterministic** mapping from input **x** to hidden representation **h**.
- => Using **auto-encoders**
- **h(x)** should retain enough information to allow reconstructing **x**

Auto-Encoders (AE)



Auto-Encoders (AE)

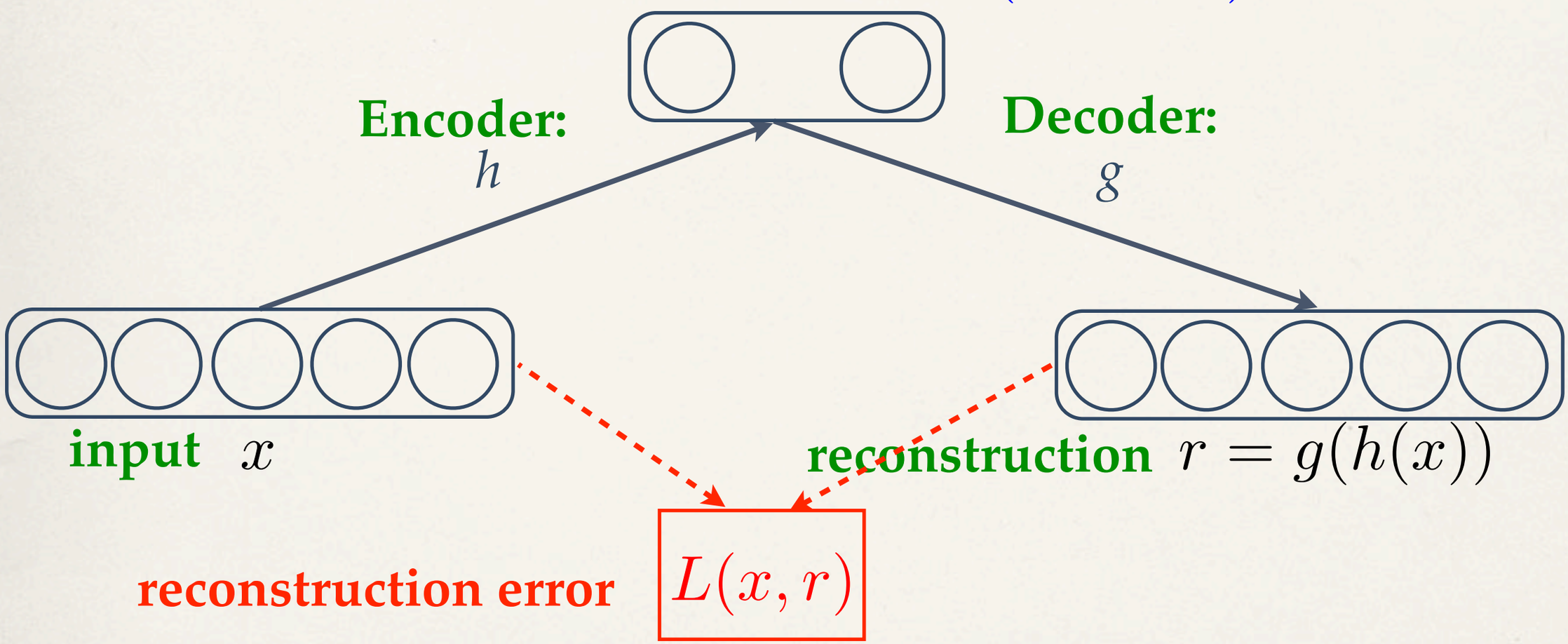
Typical form



Auto-Encoders (AE)

Typical form

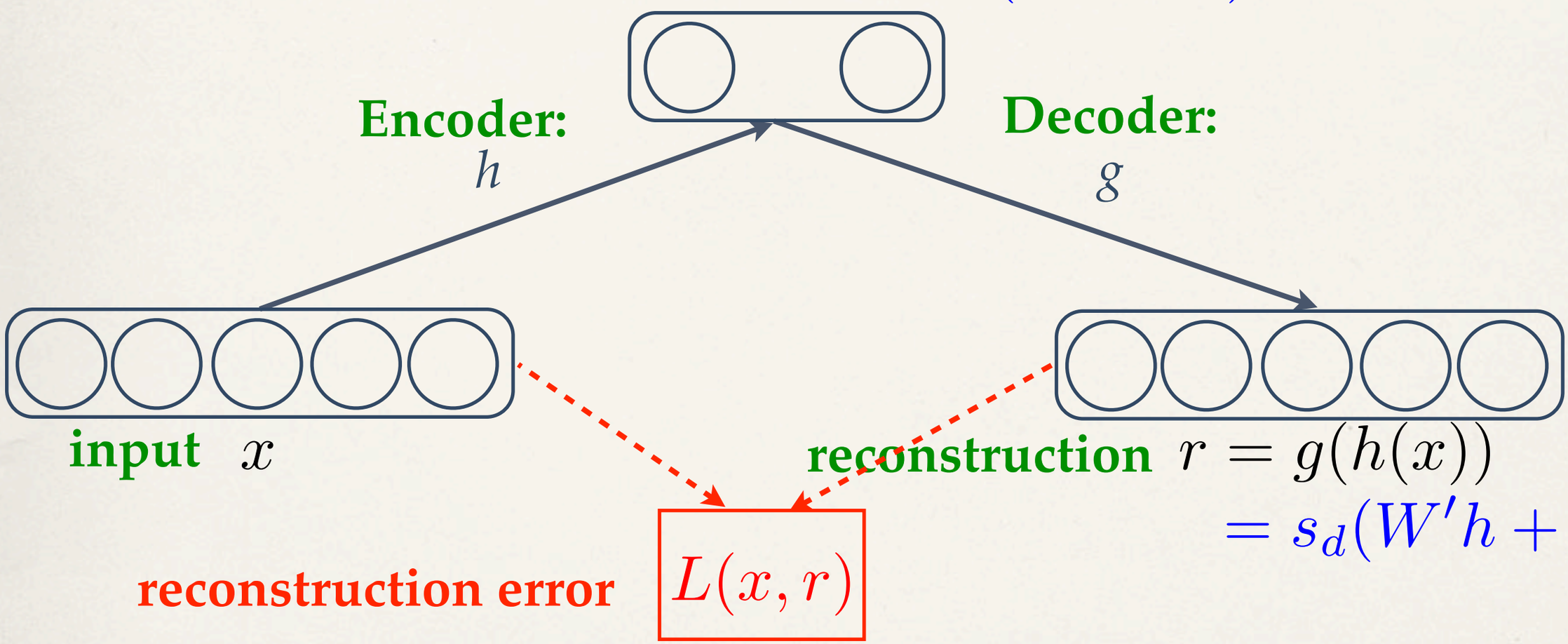
hidden representation $\mathbf{h} = h(\mathbf{x}) = s(Wx + b)$



Auto-Encoders (AE)

Typical form

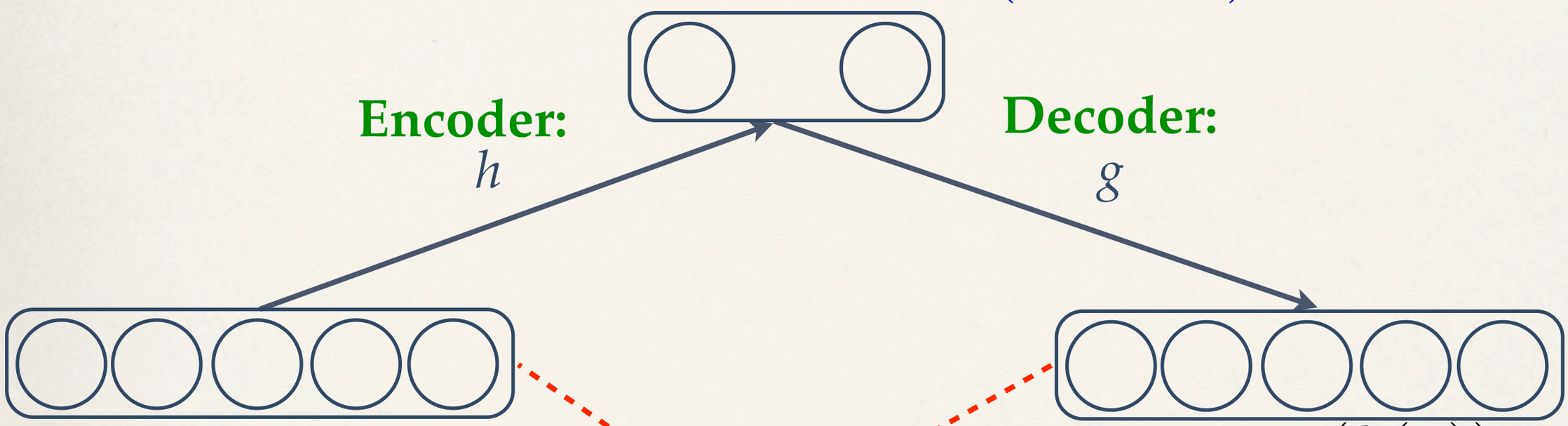
hidden representation $\mathbf{h} = h(\mathbf{x}) = s(Wx + b)$



Auto-Encoders (AE)

Typical form

hidden representation $\mathbf{h} = h(\mathbf{x}) = s(Wx + b)$



input x

reconstruction $r = g(h(x))$
$= s_d(W'h + b_d)$

reconstruction error $\boxed{L(x, r)}$

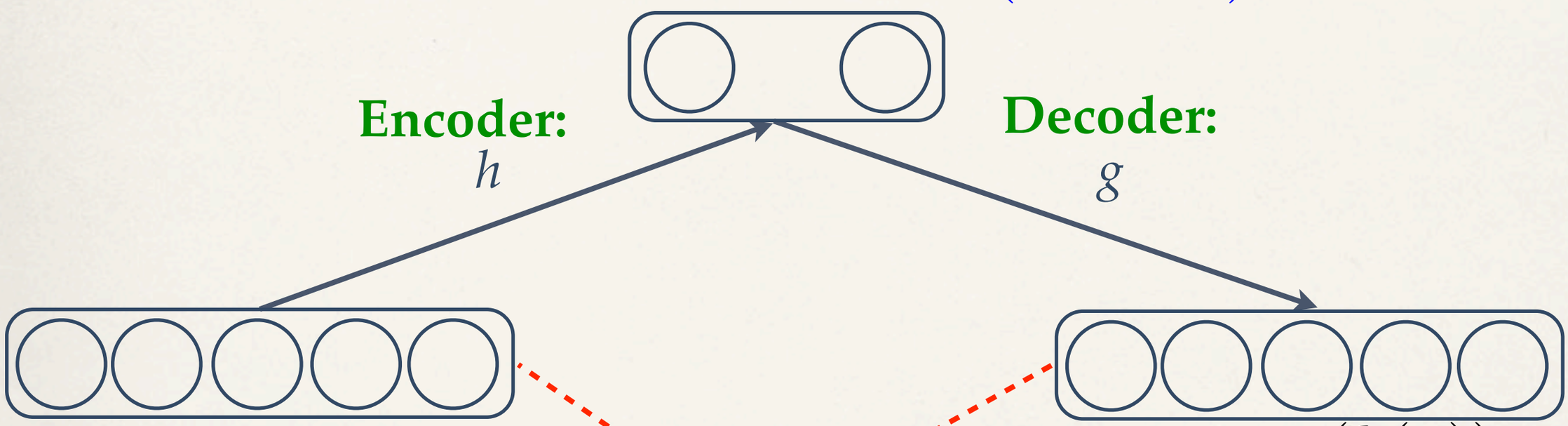
squared error: $\|x - r\|^2$

or Bernoulli cross-entropy

Auto-Encoders (AE)

Typical form

hidden representation $\mathbf{h} = h(\mathbf{x}) = s(Wx + b)$



reconstruction error $L(x, r)$

squared error: $\|x - r\|^2$

or Bernoulli cross-entropy

s is often a logistic sigmoid

Auto-Encoder training

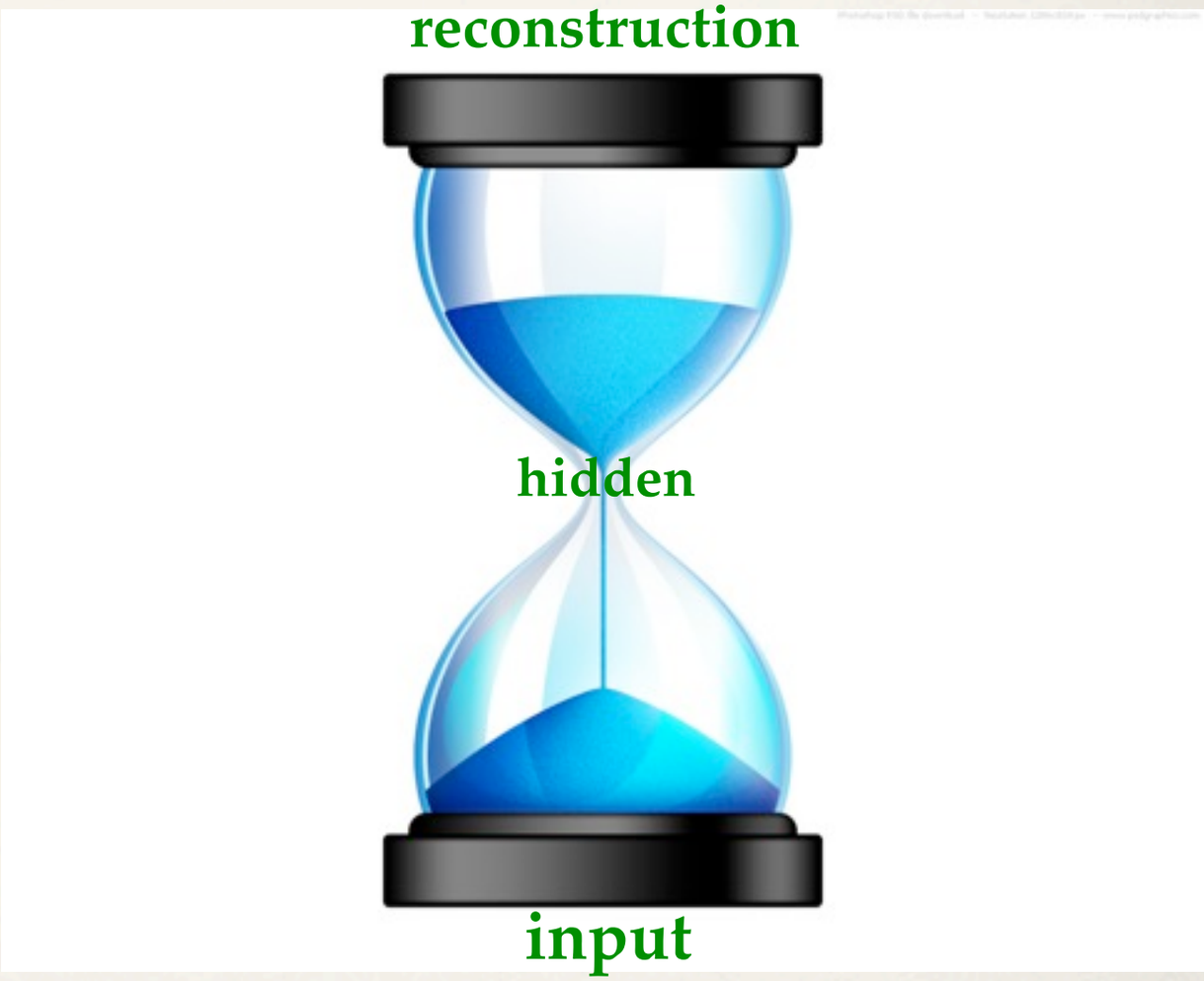
- Parameters $\theta = \{W, b, W', b_d\}$
- are learned from dataset D
- to minimize sum of reconstruction errors

$$\mathcal{J}_{\mathrm{AE}}(\theta) = \sum_{x \in D} L(x, g(h(\tilde{x}))$$

- with a gradient descent technique.

Over-complete representations?

- Traditional autoencoders used for dimensionality **reduction**.
- **Overcomplete** case yields useless solutions: **identity mapping**
- Need for alternative **regularization/constraining** (e.g. *sparsity*...)

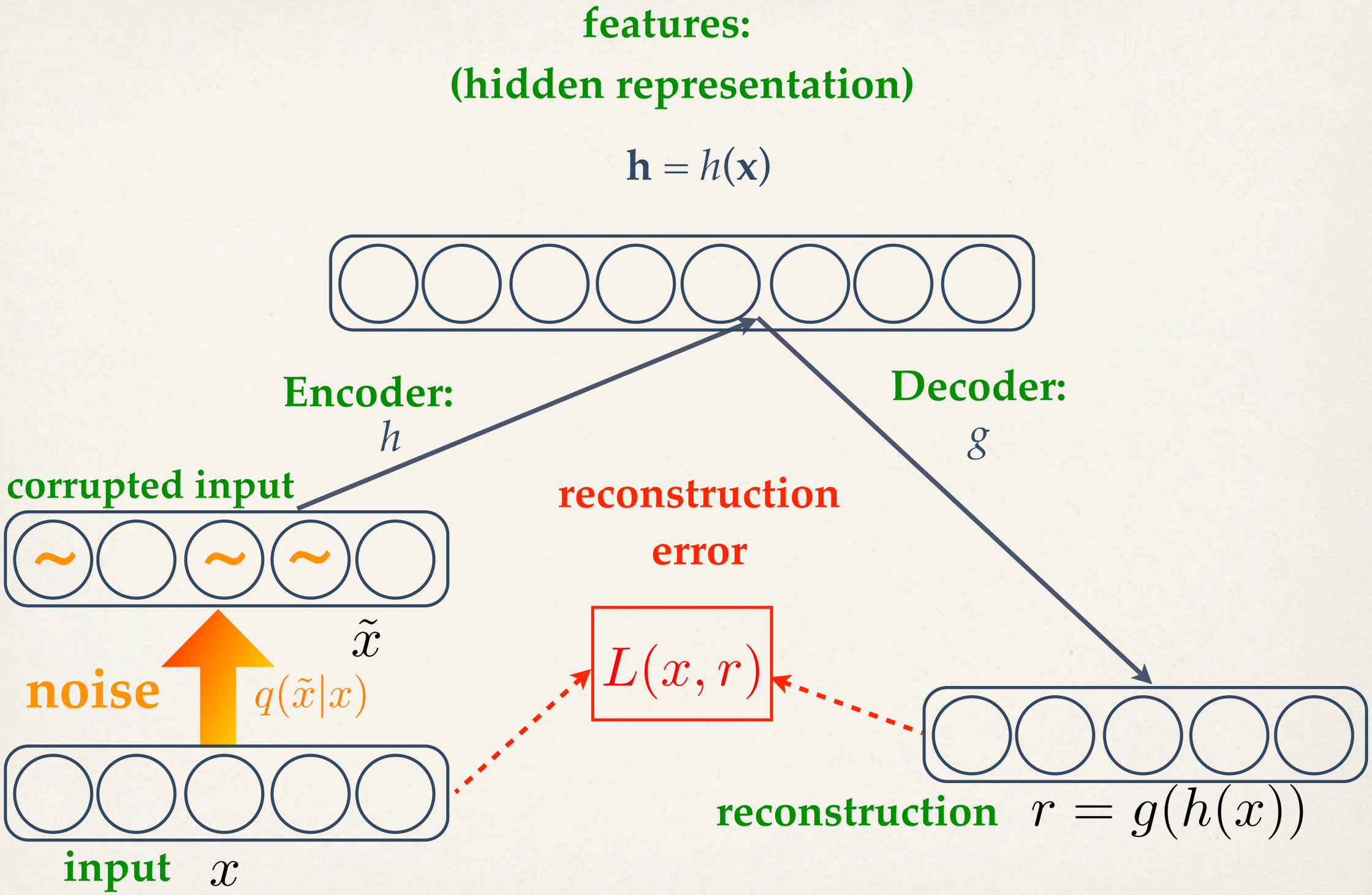


Denoising auto-encoders: motivation

(Vincent, Larochelle, Bengio, Manzagol, ICML 2008)

- Denoising corrupted input is a vastly **more challenging** task.
- Even in widely **over-complete case**...
  it **must** learn intelligent encoding/decoding.
- Will encourage **representation that is robust** to small perturbations of the input.

Denoising auto-encoder (DAE)



Denoising auto-encoder (DAE)

(Vincent, Larochelle, Bengio, Manzagol, ICML 2008)

- Autoencoder training minimizes:

$$\mathcal{J}_{\mathrm{AE}}(\theta) = \sum_{x \in D} L(x, g(h(\tilde{x}))$$

- Denoising autoencoder training minimizes

$$\mathcal{J}_{\mathrm{DAE}}(\theta) = \sum_{x \in D} \mathbb{E}_{q(\tilde{x}|x)} \left[L(x, g(h(\tilde{x})) \right]$$

Cannot compute expectation exactly
➪ use stochastic gradient descent,
sampling corrupted inputs $\tilde{x}|x$

Denoising auto-encoder (DAE)

(Vincent, Larochelle, Bengio, Manzagol, ICML 2008)

- Autoencoder training minimizes:

$$\mathcal{J}_{\mathrm{AE}}(\theta) = \sum_{x \in D} L(x, g(h(\tilde{x}))$$

- Denoising autoencoder training minimizes

$$\mathcal{J}_{\mathrm{DAE}}(\theta) = \sum_{x \in D} \mathbb{E}_{q(\tilde{x}|x)} \left[L(x, g(h(\tilde{x}))) \right]$$

Cannot compute expectation exactly
➪ use stochastic gradient descent,
sampling corrupted inputs $\tilde{x}|x$

Possible corruptions q:

- additive Gaussian noise
- zeroing pixels at random
- salt-and-pepper noise
- ...

Denoising auto-encoder (DAE)

(Vincent, Larochelle, Bengio, Manzagol, ICML 2008)

- Autoencoder training minimizes:

$$\mathcal{J}_{\mathrm{AE}}(\theta) = \sum_{x \in D} L(x, g(h(\tilde{x})))$$

- Denoising autoencoder training minimizes

$$\mathcal{J}_{\mathrm{DAE}}(\theta) = \sum_{x \in D} \mathbb{E}_{q(\tilde{x}|x)} [L(x, g(h(\tilde{x})))]$$

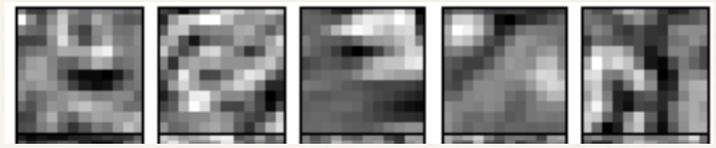
Cannot compute expectation exactly
➪ use stochastic gradient descent,
sampling corrupted inputs $\tilde{x}|x$

Possible corruptions q:

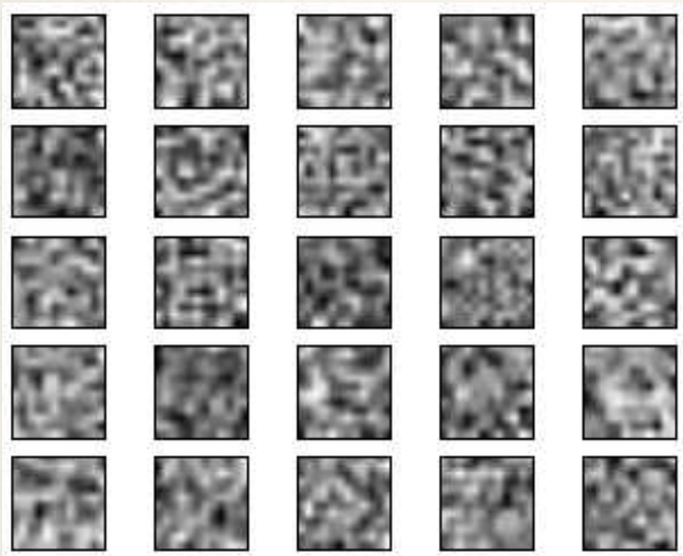
- additive Gaussian noise
- zeroing pixels at random
- salt-and-pepper noise
- ...

Stochastic perturbation

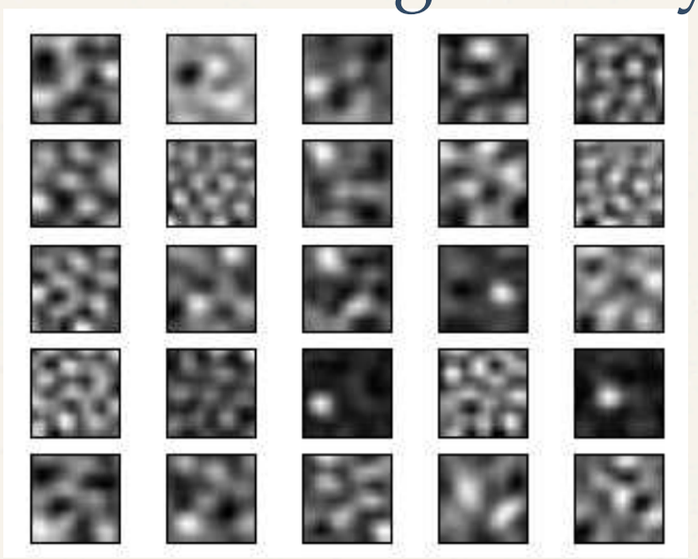
Learned filters

a) Natural image patches e.g.: 

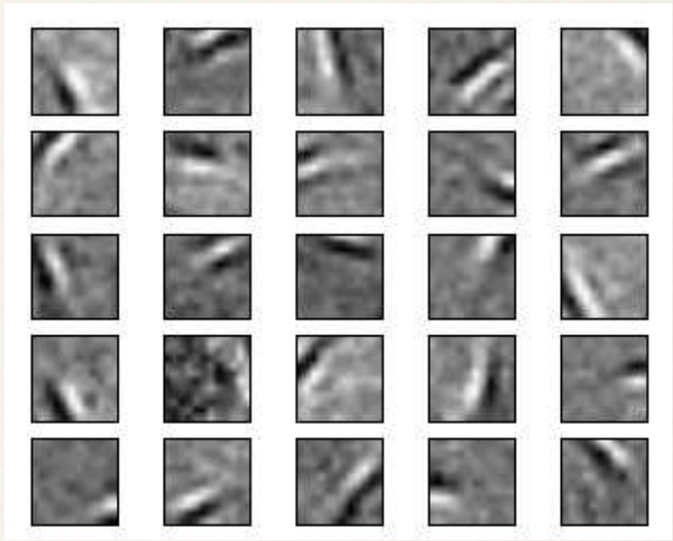
AE



AE with weight decay

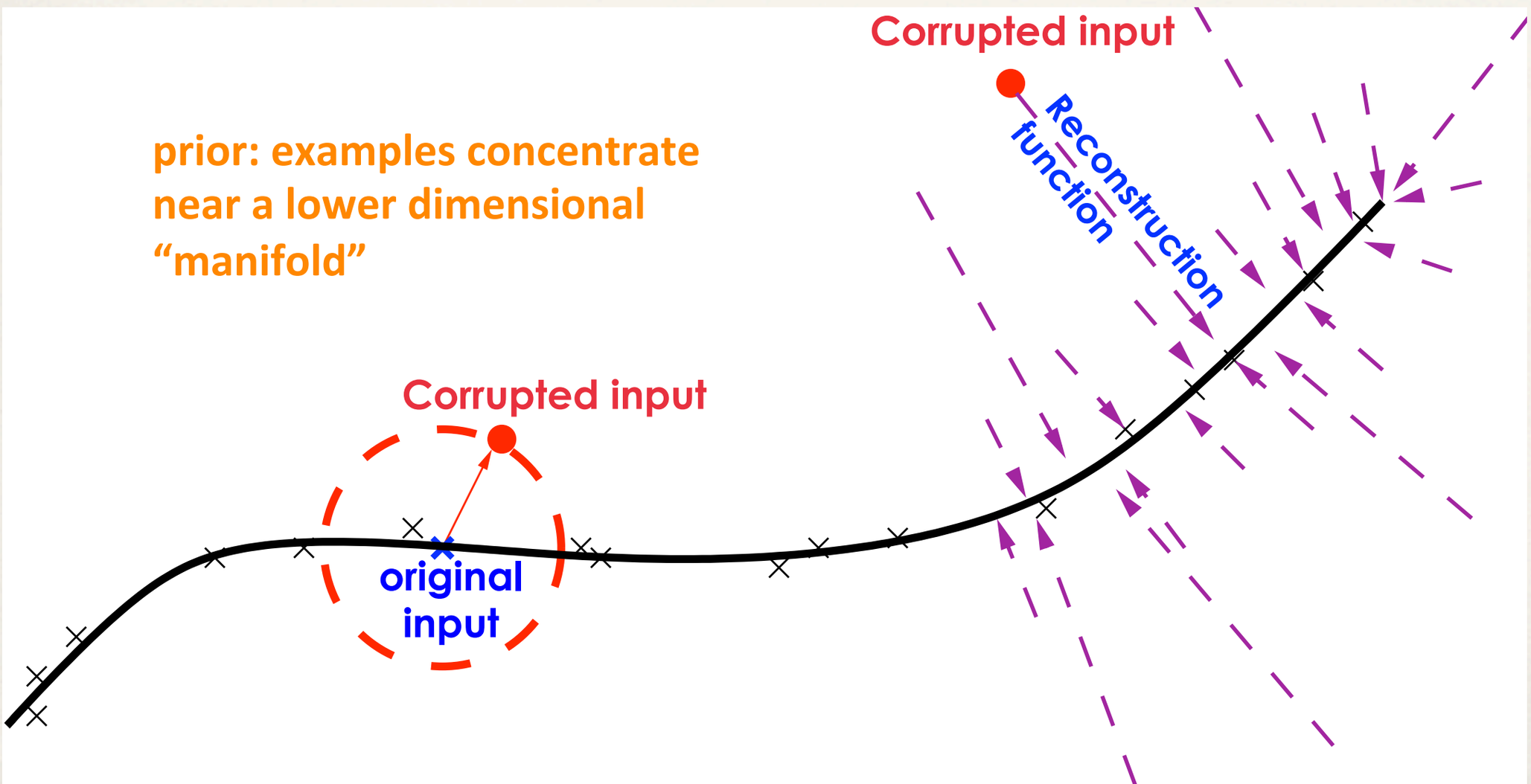


DAE



DAE: Geometric interpretation

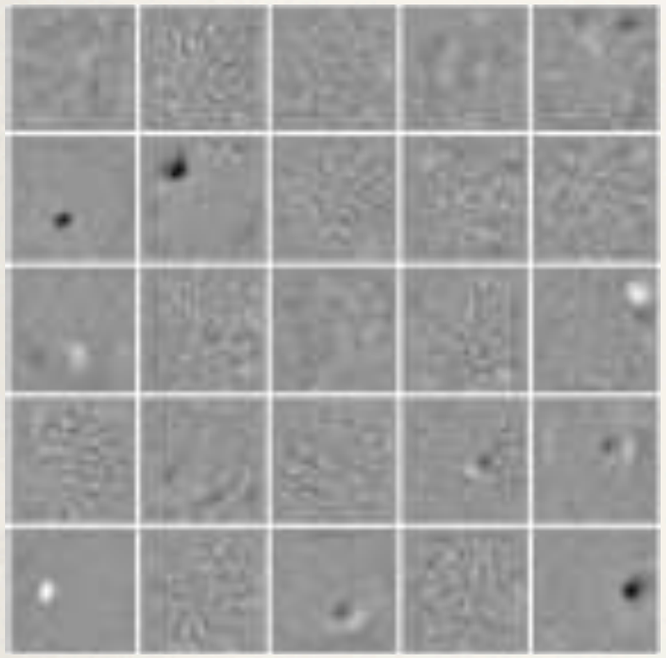
- DAE learns to «project back» corrupted input onto manifold.
- Representation $h \approx$ location on the manifold



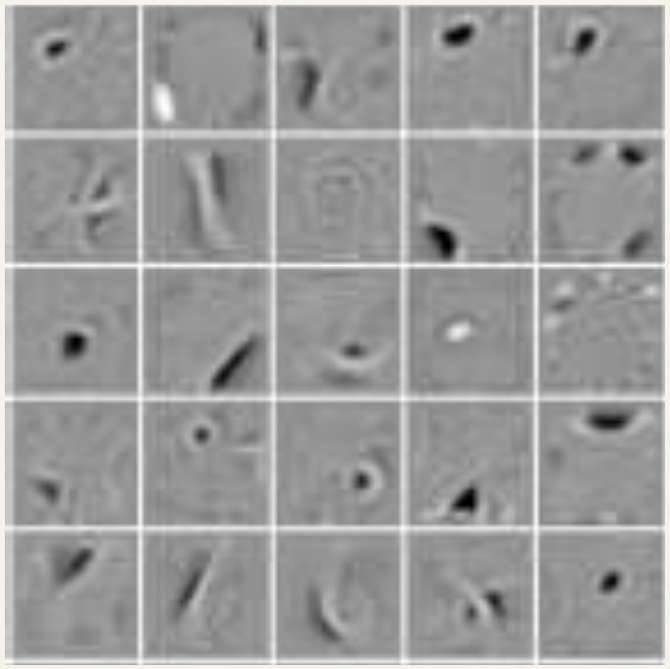
Learned filters

b) MNIST digits e.g.: 

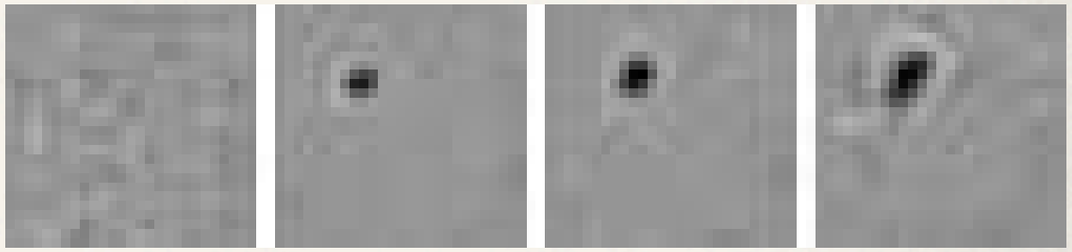
AE



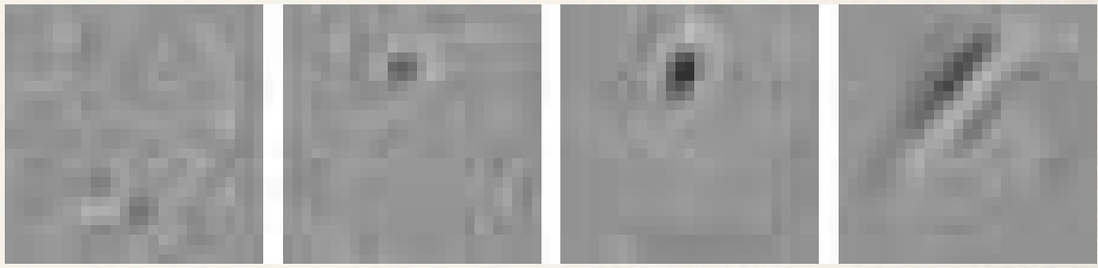
DAE



Increasing noise



(d) Neuron A (0%, 10%, 20%, 50% corruption)



(e) Neuron B (0%, 10%, 20%, 50% corruption)

Encouraging **representation** to be insensitive to corruption

- DAE encourages **reconstruction** to be insensitive to input corruption
- Alternative: encourage **representation** to be **insensitive**

$$\mathcal{J}_{\text{SCAE}}(\theta) = \sum_{x \in D} L(x, g(h(x))) + \lambda \mathbb{E}_{q(\tilde{x}|x)} \left[\|h(x) - h(\tilde{x})\|^2 \right]$$

Reconstruction error

stochastic regularization term

- Tied weights i.e. $W' = W^T$ prevent W from collapsing h to 0.

Encouraging **representation** to be insensitive to corruption

- DAE encourages **reconstruction** to be insensitive to input corruption
- Alternative: encourage **representation** to be **insensitive**

$$\mathcal{J}_{\text{SCAE}}(\theta) = \sum_{x \in D} L(x, g(h(x))) + \lambda \mathbb{E}_{q(\tilde{x}|x)}\left[\|h(x) - h(\tilde{x})\|^2\right]$$

Reconstruction error

stochastic regularization term

- Tied weights i.e. $W = W^T$ prevent W from collapsing h to 0.

Stochastic regularization

From stochastic to analytic penalty

- SCAE stochastic regularization term: $\mathbb{E}_{q(\tilde{x}|x)}\left[\|h(x) - h(\tilde{x})\|^2\right]$
- For small additive noise $\tilde{x}|x = x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$
- Taylor series expansion yields $h(x+\epsilon) = h(x) + \frac{\partial h}{\partial x}\epsilon + \ldots$
- It can be showed that

$$\underbrace{\mathbb{E}_{q(\tilde{x}|x)}\left[\|h(x) - h(\tilde{x})\|^2\right]}_{\substack{\text{stochastic} \\ \text{(SCAE)}}} \approx \underbrace{\sigma^2 \left\| \frac{\partial h}{\partial x}(x) \right\|_F^2}_{\substack{\text{analytic} \\ \text{(CAE)}}}$$

Contractive Auto-Encoder (CAE)

(Rifai, Vincent, Muller, Glorot, Bengio, ICML 2011)

- Minimize $$\mathcal{J}_{\mathrm{CAE}} = \sum_{x \in D}^{n} L(x, g(h(x))) + \lambda \left\| \frac{\partial h(x)}{\partial x} \right\|^2$$

  Reconstruction error — analytic contractive term

- For training examples, encourages both:
  - ➡small reconstruction error
  - ➡representation insensitive to small variations around example

Contractive Auto-Encoder (CAE)

(Rifai, Vincent, Muller, Glorot, Bengio, ICML 2011)

- Minimize $$\mathcal{J}_{\text{CAE}} = \sum_{x \in D}^{n} L(x, g(h(x))) + \lambda \left\| \frac{\partial h(x)}{\partial x} \right\|^2$$

Reconstruction error

analytic contractive term

Analytic regularization

- For training examples, encourages both:
  - ➡ small reconstruction error
  - ➡ representation insensitive to small variations around example

Computational considerations
CAE for a simple encoder layer

We defined $\mathbf{h} = h(\mathbf{x}) = s(\underbrace{Wx + b}_{a})$

Further suppose: s is an elementwise non-linearity

s' its first derivative.

Let $$J(x) = \frac{\partial h}{\partial x}(x)$$

$$J_j = s'(b + x^T W_j) W_j$$ where J_j and W_j represent j[th] row

CAE penalty is: $$\|J\|_F^2 = \sum_{j=1}^{d_h} s'(a_j)^2 \|W_j\|^2$$

Compare to L2 weight decay: $$\|W\|_F^2 = \sum_{j=1}^{d_h} \|W_j\|^2$$

Same complexity: $O(d_h\, d)$

Gradient backprop wrt parameters: $O(d_h\, d)$

Higher order Contractive Auto-Encoder (CAE+H)

(Rifai, Mesnil, Vincent, Muller, Bengio, Dauphin, Glorot; ECML 2011)

- CAE penalizes Jacobian norm
- We could also penalize higher order derivatives
- Computationally too expensive: second derivative is a 3-tensor, ...
- **Stochastic approach for efficiency:**
  Encourage Jacobian at x and at $x+\varepsilon$ to be the same.

$$\mathcal{J}_{\mathrm{CAE+H}} = \sum_{x \in D}^{n} L(x, g(h(x)) + \lambda \left\| \frac{\partial h}{\partial x}(x) \right\|^2$$

$$+\gamma \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)} \left[\left\| \frac{\partial h}{\partial x}(x) - \frac{\partial h}{\partial x}(x+\epsilon) \right\|^2 \right]$$

Higher order Contractive Auto-Encoder (CAE+H)

(Rifai, Mesnil, Vincent, Muller, Bengio, Dauphin, Glorot; ECML 2011)

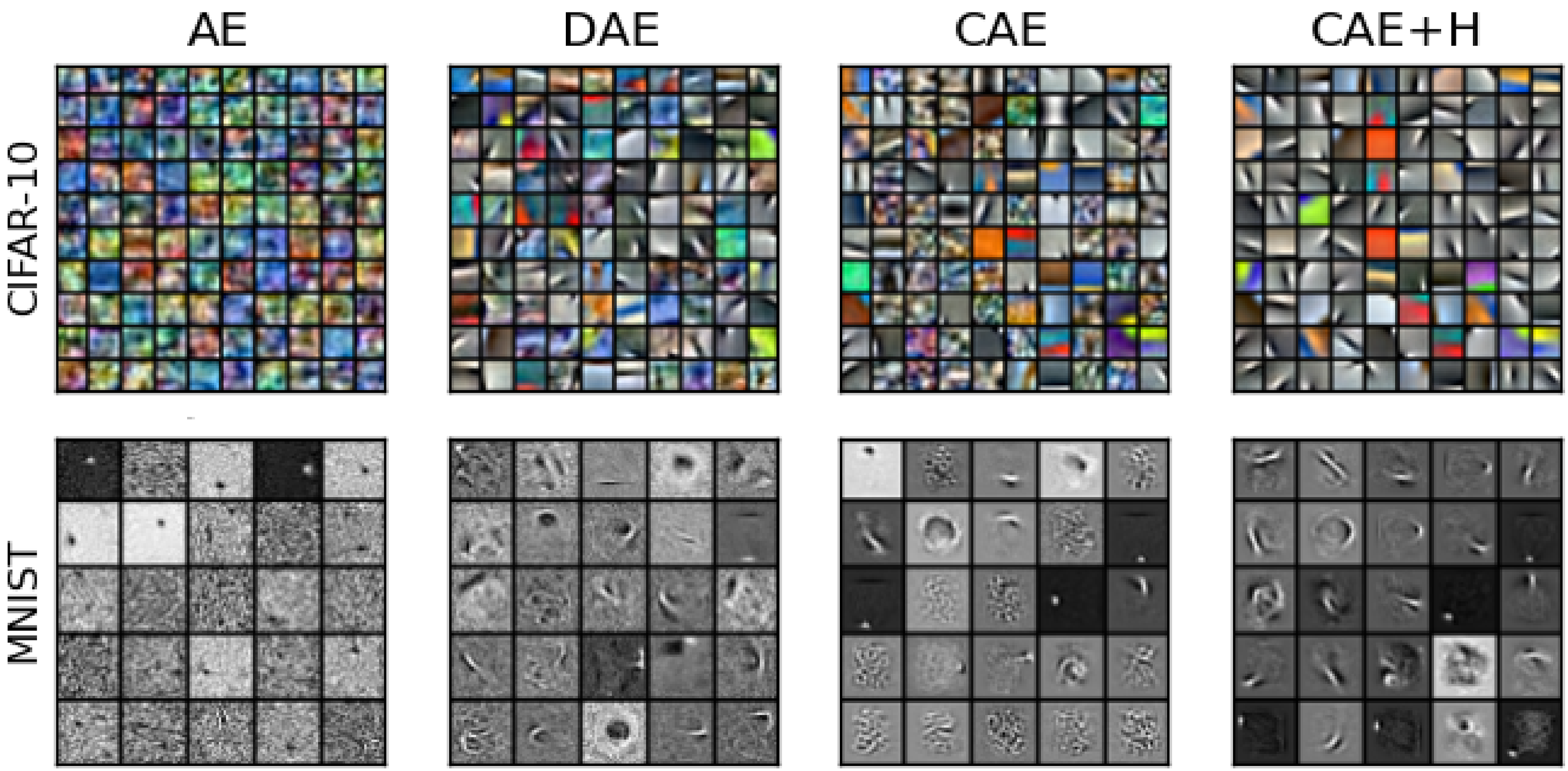
Stochastic & analytic regularization

- CAE penalizes Jacobian norm
- We could also penalize higher order derivatives
- Computationally too expensive: second derivative is a 3-tensor, ...
- **Stochastic approach for efficiency:**
  Encourage Jacobian at x and at $x+\varepsilon$ to be the same.

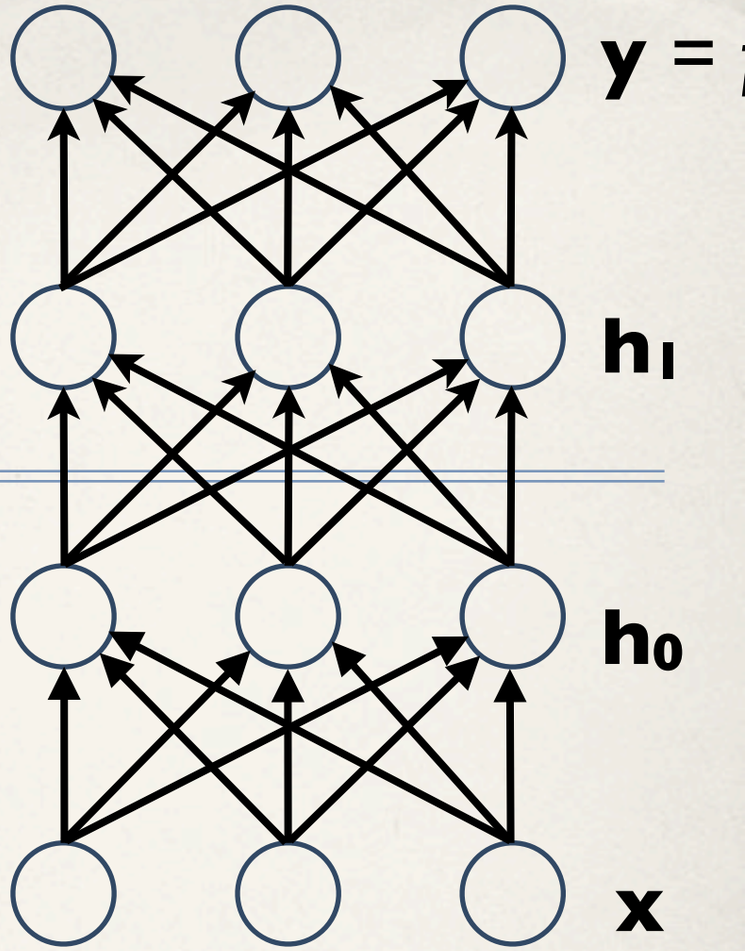
$$\mathcal{J}_{\mathrm{CAE+H}} = \sum_{x \in D} L(x, g(h(x))) + \lambda \left\| \frac{\partial h}{\partial x}(x) \right\|^2$$

$$+ \gamma \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)} \left[\left\| \frac{\partial h}{\partial x}(x) - \frac{\partial h}{\partial x}(x+\epsilon) \right\|^2 \right]$$

Learned filters



Usefulness of these regularized autoencoders



- Suitable for layerwise pretraining of deep networks (unsupevised)
- Alternative to RBM
- Won *Unsupervised and Transfer Learning Challenge* (ICML 2012)
- Allows **modeling** and leveraging data **manifold** tangent space (*Manifold Tangent Classifier*, Rifai et al. NIPS 2011)

  0.81% on MNIST
  no convolution
- DAE related to «dropouts» technique (Krizhevsky, Sutskever, Hinton 2012) that yields impressive object recognition performance.

DAE and CAE relationsip to RBMs

- Same functional form as RBM:
  $h(x)$ is expected hidden given visible
  $g(\mathbf{h})$ is expected visible given hidden

- With linear reconstruction and squared error,
  DAE amounts to learning the following energy

$$E(\mathbf{x}; \underbrace{\mathbf{W}, \mathbf{b}, \mathbf{c}}_{\theta}) = -\frac{\langle \mathbf{c}, \mathbf{x} \rangle - \frac{1}{2}\|\mathbf{x}\|^2 + \sum_{j=1}^{d_h} \text{softplus}\left(\langle \mathbf{W}_j, \mathbf{x} \rangle + \mathbf{b}_j\right)}{\sigma^2}$$

  using a stochastic variant of Hyvärinen's score matching inductive principle.

- Above energy closely related to free energy of Gaussian-binary RBM (identical for σ=1)

Score matching
(Hyvärinen 2005)

We want to learn a p.d.f. : $p_\theta(x) = \frac{1}{Z(\theta)} e^{-E_\theta(x)}$

with **intractable** partition function Z

Score matching: alternative inductive principle to max. likelihood

Find parameters that minimize objective:

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

$$\|J_E(x)\|^2$$

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

$$\|J_E(x)\|^2$$

First derivative encouraged to be small: ensures training points stay close to local minima of E

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

$\|J_E(x)\|^2$

Laplacian

$\mathrm{Tr}(H_E(x))$

First derivative encouraged to be small: ensures training points stay close to local minima of E

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

$\|J_E(x)\|^2$

First derivative encouraged to be small: ensures training points stay close to local minima of E

Laplacian

$\mathrm{Tr}(H_E(x))$

Encourage large positive curvature in all directions

Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

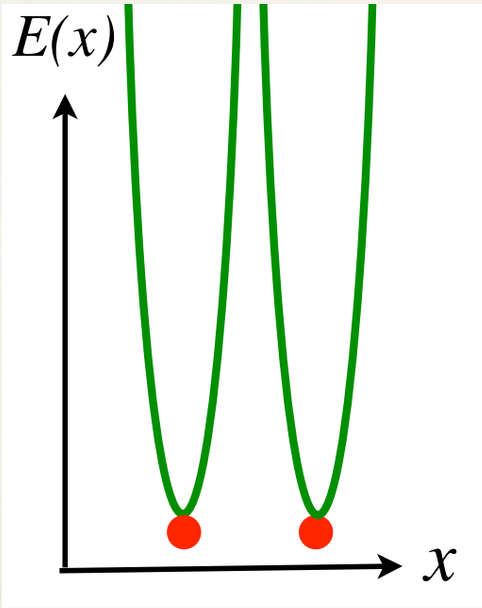
$\|J_E(x)\|^2$

Laplacian

$\mathrm{Tr}(H_E(x))$

First derivative encouraged to be small: ensures training points stay close to local minima of E

Encourage large positive curvature in all directions



Score matching

my geometric interpretation

$$J_{SM}(\theta) = \sum_{x \in D} \left(\left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

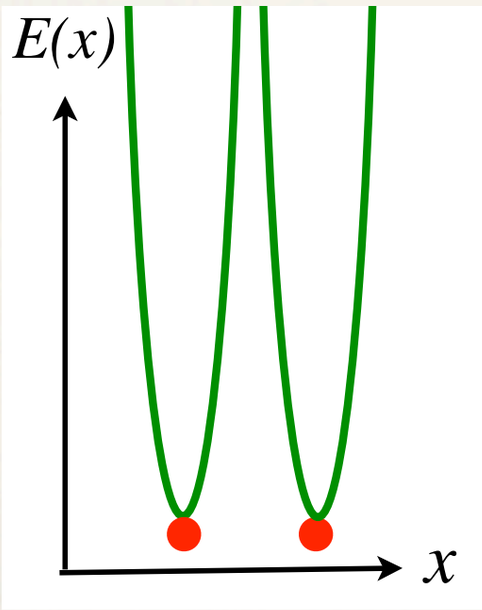
$\|J_E(x)\|^2$

First derivative encouraged to be small: ensures training points stay close to local minima of E

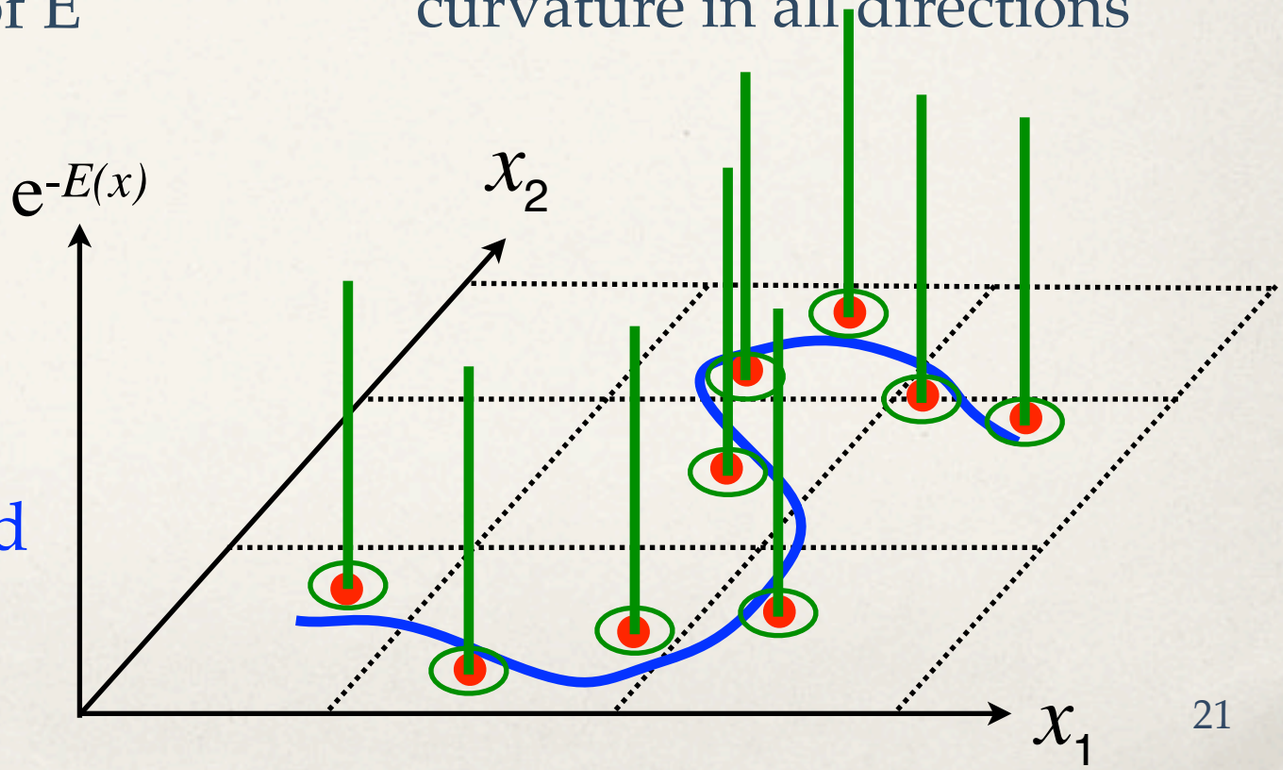
Laplacian

$\mathrm{Tr}(H_E(x))$

Encourage large positive curvature in all directions



sharply peaked density



Score matching variants

Original score matching (Hyvärinen 2005):

$$J_{SM}(\theta) \quad = \quad \sum_{x \in D} \left(\frac{1}{2} \left\| \frac{\partial E}{\partial x}(x) \right\|^2 - \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x) \right)$$

Analytic

Regularized score matching (Kingma & LeCun 2010):

$$J_{SMreg,\lambda}(\theta) \quad = \quad J_{SM} + \sum_{x \in D} \lambda \sum_{i=1}^{d} \frac{\partial^2 E}{\partial x_i^2}(x)$$

Analytic

Denoising score matching (Vincent 2011)

$$J_{DSM,\sigma} = \sum_{x \in D} \left(\mathbf{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \left[\frac{1}{2} \left\| \frac{\partial E}{\partial x}(x+\epsilon) - \frac{1}{\sigma^2}\epsilon \right\|^2 \right] \right)$$

Stochastic

Analytic v.s. stochastic ?

a) Analytic approximation of stochastic perturbation

- - Equiv. to tiny perturbations: does not probe far away
- + Potentially more efficient. Ex:
  CAE's Jacobian penalty probes sensitivity in all d directions in $O(d_h\, d)$
  With DAE or SCAE it would require encoding d corrupted inputs: $O(d_h\, d^2)$

b) Stochastic approximation of analytic criterion

- + can render practical otherwise computationally infeasible criteria
  Ex: CAE+H
- - less precise, more noisy

CAE+H actually leverages both

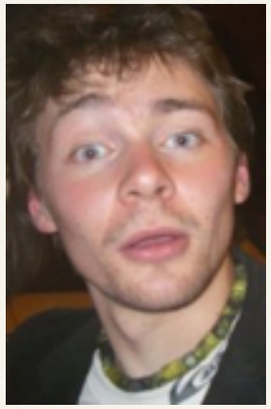
Thank you to past and current students who did most of the hard work


Hugo Larochelle


Isabelle Lajoie


Salah Rifai


Xavier Muller


Grégoire Mesnil


Yann Dauphin

and to my colleague and mentor


Yoshua Bengio

Questions ?