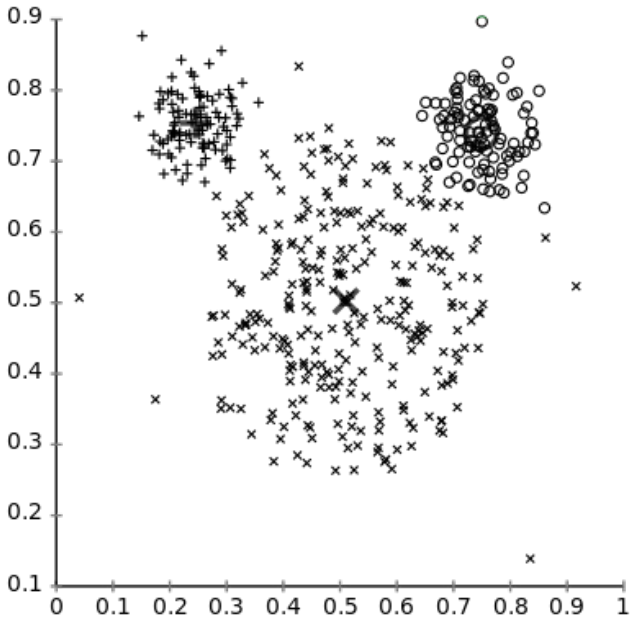
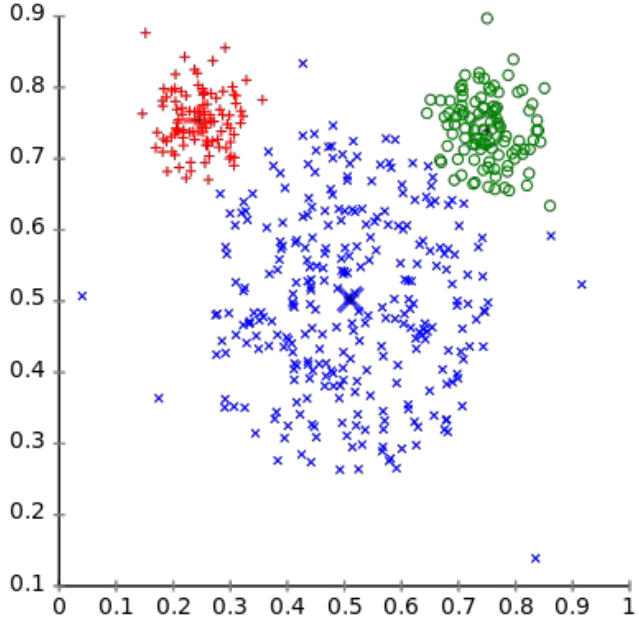
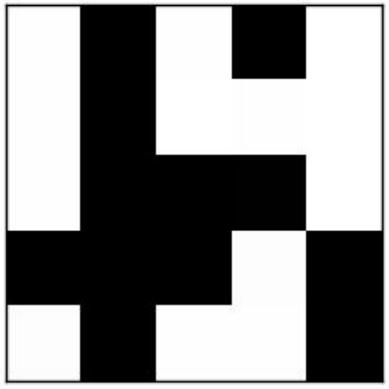
Generative models and EM algorithm

Enrico Guiraud
EP-SFT

Generative models: what and why

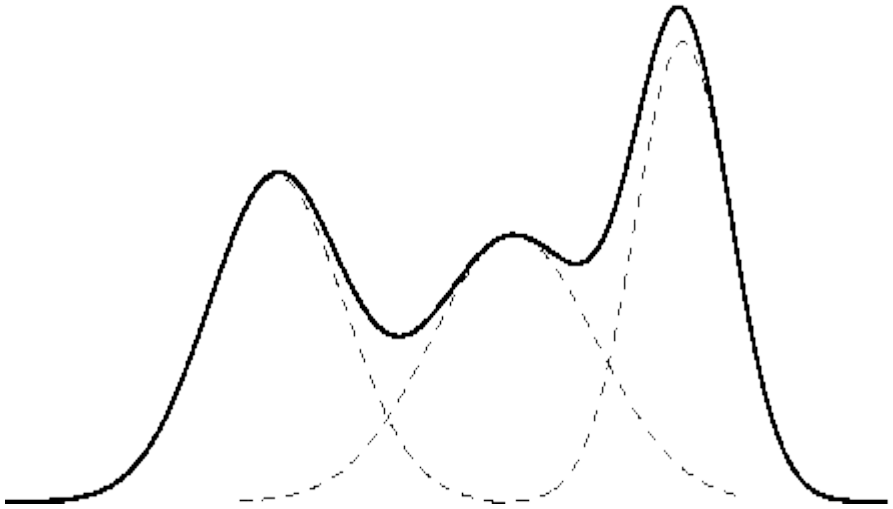






Generative models: the usual example

1. latent variables/causes
2. observable variables
3. model parameters



1. which gaussian
2. observed distribution
3. means and variances

Generative models and likelihood optimisation

Given a probabilistic model for the data generation process

hidden state

$$p\left(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\Theta}\right)$$

data-point parameters

our goal is maximising the likelihood

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_n p\left(\mathbf{y}^n \mid \boldsymbol{\Theta}\right) = \prod_n \sum_{\{\mathbf{s}\}} p\left(\mathbf{y}^n, \mathbf{s} \mid \boldsymbol{\Theta}\right)$$

ML parameters $\longleftrightarrow$ feature extraction

probabilistic inference $\longleftrightarrow$ $p\left(\mathbf{s} \mid \mathbf{y}^n, \boldsymbol{\Theta}\right)$

Generative models and likelihood optimisation

Given a probabilistic model for the data generation process

hidden state

$$p\left(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\Theta}\right)$$

data-point parameters

our goal is maximising the likelihood

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_n p\left(\mathbf{y}^n \mid \boldsymbol{\Theta}\right) = \prod_n \sum_{\{\mathbf{s}\}} p\left(\mathbf{y}^n, \mathbf{s} \mid \boldsymbol{\Theta}\right)$$

ML parameters ⟷ feature extraction

probabilistic inference ⟷ $p\left(\mathbf{s} \mid \mathbf{y}^n, \boldsymbol{\Theta}\right)$

EM learning algorithm

maximizing the likelihood is difficult

maximizing the log-likelihood is difficult

$$L(\mathbf{\Theta}) = \sum_n \log \left[\sum_{\{\mathbf{s}\}} p\left(\mathbf{y}^n, \mathbf{s} \mid \mathbf{\Theta}\right) \right] = \sum_n \log \left[\sum_{\{\mathbf{s}\}} q^n(\mathbf{s}) \frac{p\left(\mathbf{y}^n, \mathbf{s} \mid \mathbf{\Theta}\right)}{q^n(\mathbf{s})} \right]$$

$$\geq \sum_n \left\{ \sum_{\{\mathbf{s}\}} q^n(\mathbf{s}) \log \left[p\left(\mathbf{y}^n, \mathbf{s} \mid \mathbf{\Theta}\right) \right] - \sum_{\{\mathbf{s}\}} q^n(\mathbf{s}) \log \left(q^n(\mathbf{s}) \right) \right\} =: \mathcal{F}(\mathbf{q}, \mathbf{\Theta})$$

Shannon entropy

maximizing the free energy is usually possible

EM learning algorithm

$$L(\boldsymbol{\Theta}) \geq \mathcal{F}(\mathbf{q}, \boldsymbol{\Theta}) \text{ for any choice of } \mathbf{q}$$

Free energy optimised by **two-step coordinate ascent**

1. maximize $\mathcal{F}$ w.r.t. **q**, keeping **Θ** constant
2. maximize $\mathcal{F}$ w.r.t. **Θ**, keeping **q** constant

---

Common choice of **q** (mean-field): $q(\mathbf{s}) := \prod_h \widetilde{q}(s_h; \boldsymbol{\lambda}_h)$

1. maximize $\mathcal{F}$ w.r.t. λ, keeping Θ constant
2. maximize $\mathcal{F}$ w.r.t. Θ, keeping **q** constant

EM learning algorithm

$$L(\boldsymbol{\Theta}) \geq \mathcal{F}(\mathbf{q}, \boldsymbol{\Theta}) \text{ for any choice of } \mathbf{q}$$

Free energy optimised by **two-step coordinate ascent**

1. maximize $\mathcal{F}$ w.r.t. **q**, keeping **Θ** constant
2. maximize $\mathcal{F}$ w.r.t. **Θ**, keeping **q** constant

---

Common choice of **q** (mean-field): $q(\mathbf{s}) := \prod_h \widetilde{q}(s_h;\, \boldsymbol{\lambda}_h)$

1. maximize $\mathcal{F}$ w.r.t. λ, keeping **Θ** constant
2. maximize $\mathcal{F}$ w.r.t. **Θ**, keeping **q** constant

Truncated variational EM

$$q(\mathbf{s}) := \frac{1}{Z} p\left(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\Theta}\right) \delta(\mathbf{s} \in \mathcal{K}\ )$$

1. find best sets $\mathcal{K}$
2. maximize $\mathcal{F}$ w.r.t. **Θ**, keeping **q** constant

---

- based on recent theoretical results*
- no factorization $\rightarrow$ correlations can be captured
- finding best $\mathcal{K}^n$ is non-trivial

*Jörg Lücke, "Truncated Variational Expectation Maximization", 2016 (in preparation)

Truncated variational EM

$$q(\mathbf{s}) := \frac{1}{Z} p\left(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\Theta}\right) \delta(\mathbf{s} \in \mathcal{K} \;\;)$$

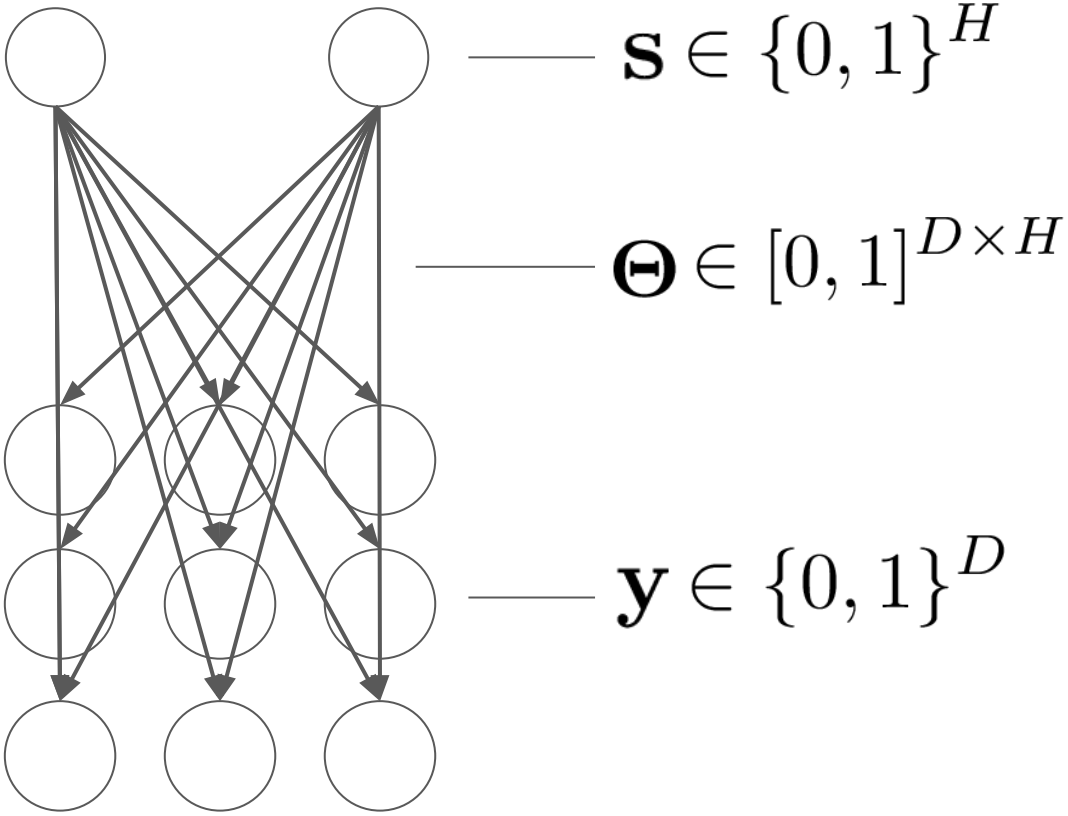
1. find best sets $\mathcal{K}$
2. maximize $\mathcal{F}$ w.r.t. **Θ**, keeping **q** constant

---

- based on recent theoretical results*
- no factorization → correlations can be captured
- finding best $\mathcal{K}$ is non-trivial

*Jörg Lücke, “Truncated Variational Expectation Maximization”, 2016 (in preparation)

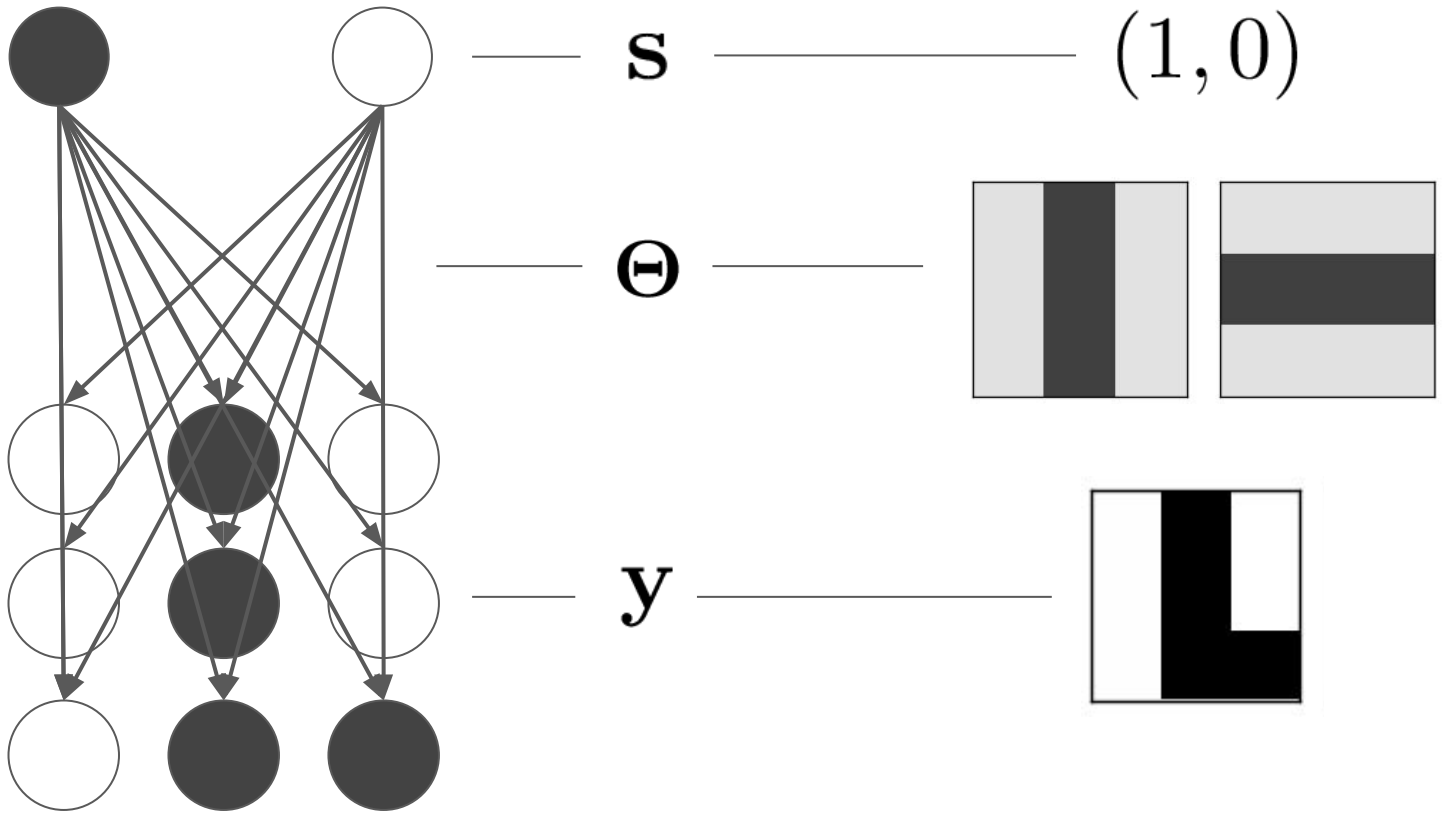
The noisy-OR model



Noisy $\rightarrow s_h$ activates y_d with probability Θ_{dh}

OR $\rightarrow y_d$ is active if at least one s_h activates it

The noisy-OR model



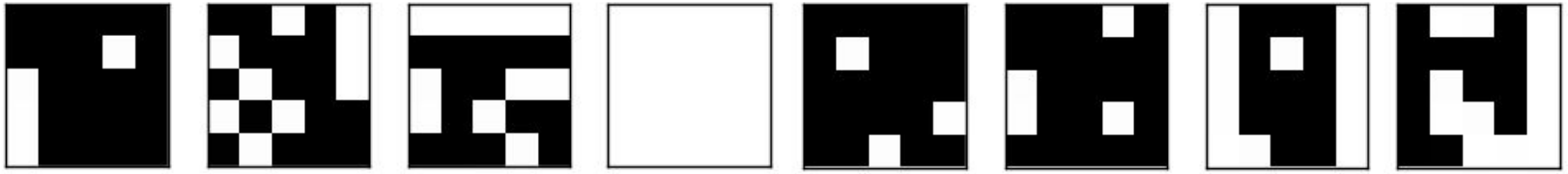
Noisy $\rightarrow s_h$ activates y_d with probability Θ_{dh}

OR $\rightarrow y_d$ is active if at least one s_h activates it

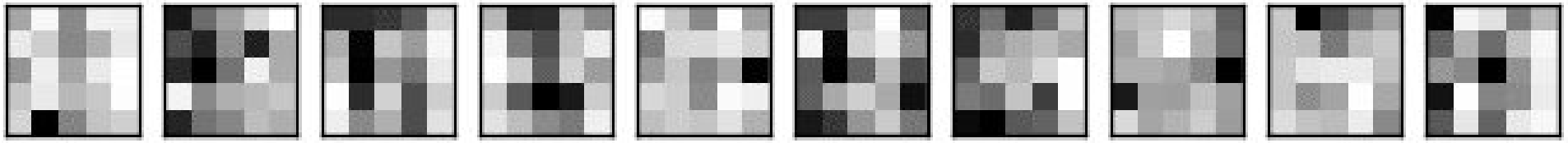
A first test: bars test

True parameters Θ

Sample data-points (8 of 800)

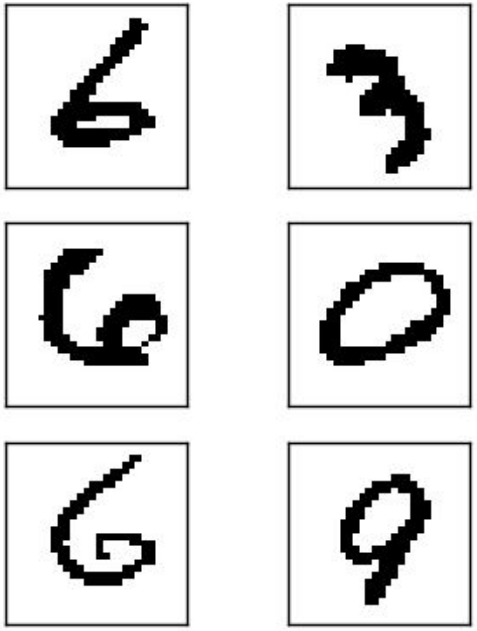


Learning process of parameters Θ

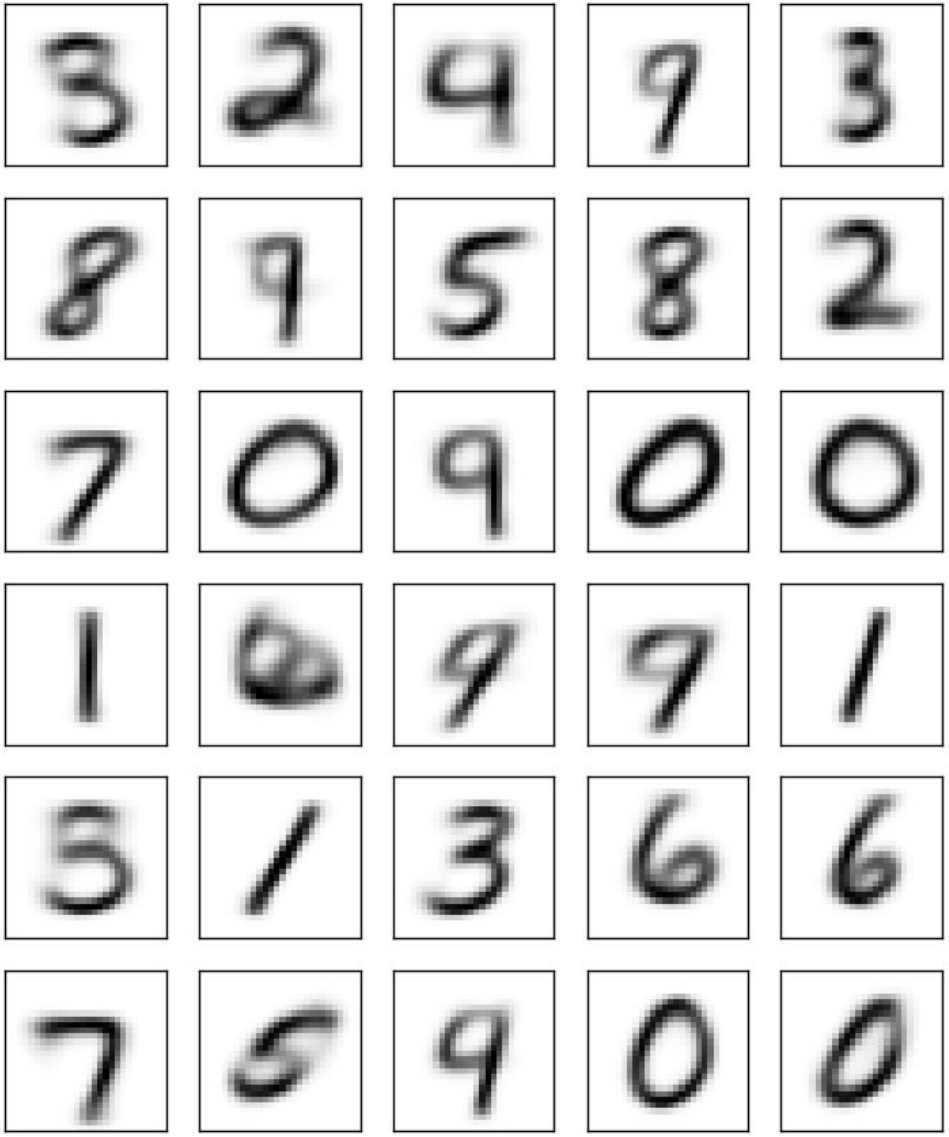


MNIST digits data-set

Sample data-points



Learned parameters



30 hidden variables
784 observables
60’000 images
22 hours runtime
on 32 cores

Thank you!

Mean-field vs TVEM

$$q(\mathbf{s}) := \prod_h \widetilde{q}(s_h;\, \boldsymbol{\lambda}_h) \qquad\qquad q(\mathbf{s}) := \frac{1}{Z} p\left(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\Theta}\right) \delta(\mathbf{s} \in \mathcal{K}^n)$$

- does not require to choose the shape of q
- does not require the addition of free parameters
- provides an efficiently computable free energy
- first step is generic (i.e. model independent)
- captures correlations among hidden variables
- depending on the model, might require more processing power
- depending on the model, might require longer execution times