

Variational Information Maximization and Conditional Self-Supervised Training

Felix Agakov

University of Edinburgh, UK

May 31, 2006

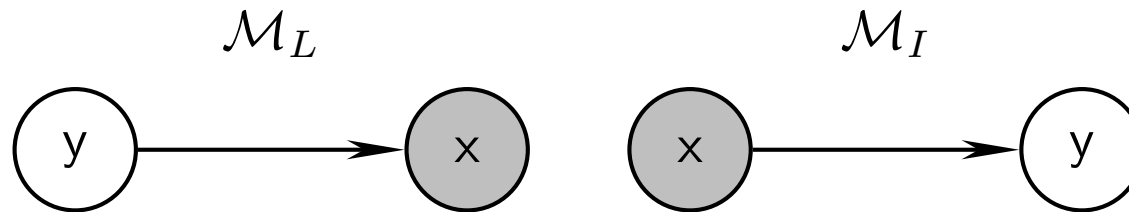
Overview

- Variational Information Maximization (IM) – variational extension of the Blahut-Arimoto algorithm for maximizing the channel capacity
- studying links between:
 - the variational EM for generative models
 - the variational EM for stochastic autoencoders(under the assumption of *equivalent inference*)
- comparing the objectives and the fixed points of the variational IM and EM algorithms
- deriving a simple way to train stochastic autoencoders, motivated by information theory
- motivating simple extensions of parametric encoding distributions
- performing empirical comparisons with some other approximate information-maximization methods

Extracting informative representations

- *Goal*: finding informative (unknown) representations $\{y\}$ of the visible training patterns $\{x\}$
- *Generative models*: explicit constraints on the generating distribution $p(x|y)$:
 - learning by fitting a constrained model to the observations (e.g. maximization of the marginal likelihood)
 - inference by applying probability rules
- *Encoder models*: explicit constraints on the encoding distribution $p(y|x)$
 - learning by optimizing other criteria (e.g. the mutual information $I(x, y)$)
 - extracting informative representations of the data **directly** from the dataset
 - (**motivation**: there may be little or no knowledge about the data-generating process; $p(x|y)$ may be difficult or expensive to parameterize; constraints on the posterior $p(y|x)$ may be physically-motivated)

Generative vs Encoder Models



Generative models:

- $\mathcal{M}_L \stackrel{\text{def}}{=} p(y)p(x|y)$ – maximizing the (marginal) likelihood
 $\mathcal{L} = \log p(x^{(1)}, \dots, x^{(M)})$

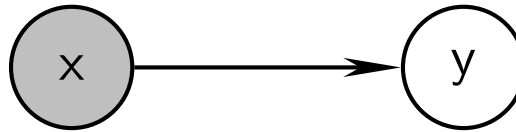
Encoder models:

- $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(x)\tilde{p}(y|x)$ – maximizing the mutual information
 $I(x, y) = H(y) - H(y|x)$, where $H(y) \stackrel{\text{def}}{=} -\langle \log \tilde{p}(y) \rangle_{\tilde{p}(y)}$,
 $H(y|x) \stackrel{\text{def}}{=} -\langle \log \tilde{p}(y|x) \rangle_{\tilde{p}(x)\tilde{p}(y|x)}$

How to relate these frameworks?

- equivalence for specific invertible encodings (Pearlmutter and Parra (1996), Cardoso (1997), MacKay (1999))
- what about noisy encoder models?

Variational IM for encoder models



$$\mathcal{M}_I = \tilde{p}(x)\tilde{p}(y|x)$$

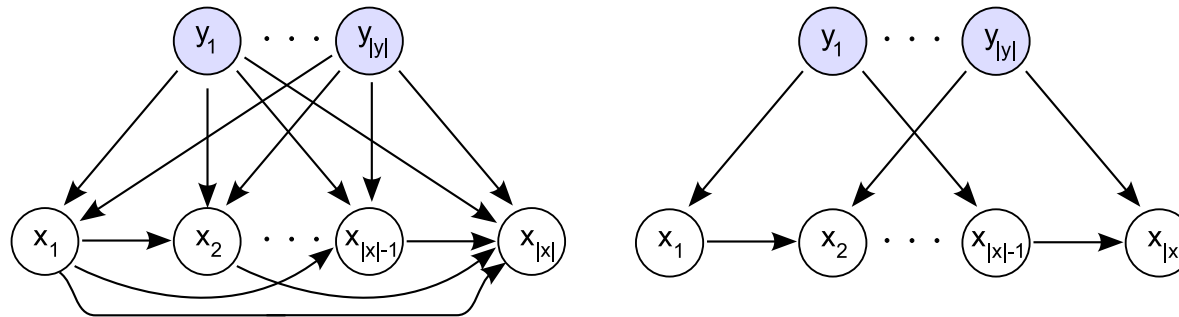
- $I(x, y) = \langle \log \tilde{p}(x|y) \rangle_{\tilde{p}(x)p(y|x)} - \langle \log \tilde{p}(x) \rangle_{\tilde{p}(x)}$
- assume that $\tilde{p}(x)$ is a known (e.g. empirical) distribution
- for large $|y|$, maximization of $I(x, y)$ is (generally) intractable
- instead, we maximize the lower bound on $I(x, y)$

$$I(x, y) \geq \tilde{I}(x, y) \stackrel{\text{def}}{=} \langle \log q(x|y) \rangle_{\tilde{p}(x)\tilde{p}(y|x)} - \langle \log \tilde{p}(x) \rangle_{\tilde{p}(x)}$$

- $q(x|y)$ is the *variational decoder* constrained to lie in a tractable family
- iterative maximization of $\tilde{I}(x, y)$ for $\tilde{p}(y|x)$ and $q(x|y)$ is guaranteed to maximize or leave unchanged a *lower bound* on $I(x, y)$ (cf Blahut-Arimoto algorithms) – the *variational IM algorithm*

Generalization of sparse approximations of $I(x, y)$

$q(x|y)$:



- Variational IM extends a number of common approximations of the mutual information $I(x, y)$

Sparse variational decoders:

- $I(x, y) = H(x) + \sum_{i=1}^{|x|} \langle \log p(x_i | x_1, \dots, x_{i-1}, y) \rangle_{p(x_1, \dots, x_{i-1}, y)}$
- choose a sparse decoder $q(x_i | x \setminus x_i, y) = q(x_i | \pi^x(x_i), \pi^y(x_i))$, where $\pi^x(x_i) \subset \{x_1, \dots, x_{i-1}\}$, $\pi^y(x_i) \subset \{y_1, \dots, y_{|y|}\}$, and run the IM
- optimally:

$$I(x, y) \geq H(x) - \sum_i H(x_i | \pi^x(x_i), \pi^y(x_i)),$$

i.e. the IM formally generalizes common sparse bounds on $I(x, y)$

Generalization of *as-if* Gaussian approximations

Parametric variational decoders:

- Linsker's *as-if Gaussian* approximation:

$$\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathbf{x})\tilde{p}(y|\mathbf{x}) \approx p_G(\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- this leads to

$$I_G(\mathbf{x}, y) \propto \log |\boldsymbol{\Sigma}_{xx}| - \log |\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}^T|,$$

where $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{xy}$, and $\boldsymbol{\Sigma}_{yy}$ are the partitions of $\boldsymbol{\Sigma}$

- the variational IM: choose $q(\mathbf{x}|y) \sim \mathcal{N}(W(y - \langle y \rangle), \sigma^2)$ and perform coordinate ascent on $\tilde{I}(\mathbf{x}, y) = \langle \log q(\mathbf{x}|y) \rangle_{\tilde{p}(\mathbf{x})\tilde{p}(y|\mathbf{x})} - \langle \log \tilde{p}(\mathbf{x}) \rangle_{\tilde{p}(\mathbf{x})}$
- then $\arg \max_W \tilde{I}(\mathbf{x}, y) \equiv I_G(\mathbf{x}, y)$
- the IM formally generalizes Linsker's *as-if Gaussian* criterion (1993)

Variational IM and EM for Generative Models

Can we relate the variational EM and IM algorithms?

Variational EM:

- a standard way to handle intractability of the exact maximum likelihood learning in a generative model $\mathcal{M}_L \stackrel{\text{def}}{=} p(y)p(x|y)$ is to optimize the variational Jensen's lower bound on the likelihood
- for any $q(y|x)$, we get

$$\begin{aligned}\mathcal{L} &\geq \tilde{\mathcal{L}} \stackrel{\text{def}}{=} \langle \log p(x, y) / q(y|x) \rangle_{q(y|x)\tilde{p}(x)} \\ &= \langle \log p(x|y) \rangle_{q(y|x)\tilde{p}(x)} - \langle KL(q(y|x) || p(y)) \rangle_{\tilde{p}(x)}.\end{aligned}$$

- we assume that the patterns are i.i.d., and $\tilde{p}(x) \propto \sum_{i=1}^M \delta(x - x^{(i)})$ is the empirical distribution
- the variational EM optimizes \mathcal{L} for $p(y)$, $p(x|y)$, and the variational posterior $q(y|x)$ (typically chosen to facilitate computations of $\tilde{\mathcal{L}}$)

Variational IM and EM for Generative Models –2

- define an encoder model

$$\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} q(y|x)\tilde{p}(x).$$

- the parameterization satisfies the *equivalent inference* assumption
- i.e. for a fixed $q(y|x)$, the forward-inference in $\tilde{\mathcal{M}}_I$ is equivalent to the *variational* inference in \mathcal{M}_L (where the exact posterior of \mathcal{M}_L is approximated by $q(y|x)$)
- N.B.: we are typically interested in latent variables y which are predictive about the data x
- does optimization of $\tilde{\mathcal{L}}$ in \mathcal{M}_L indeed lead to an increase in the mutual information $I(x, y)$ in encoder $\tilde{\mathcal{M}}_I$?

Variational IM and EM for Generative Models –3

- it is straight-forward to show that:
 1. the Jensen's bound $\tilde{\mathcal{L}}$ on the likelihood of \mathcal{M}_L is in fact a proper lower bound on the mutual information in the corresponding encoder model $\tilde{\mathcal{M}}_I$
 2. we can easily find another lower bound on $I(x, y)$, which is at least as tight as $\tilde{\mathcal{L}}$
 3. this bound on $I(x, y)$ will be tractable whenever $\tilde{\mathcal{L}}$ is tractable
- optimization of this tighter bound gives rise to a specific instance of the variational IM algorithm

Variational IM and EM for Generative Models –4

Proposition 1: *For i.i.d. patterns $\{\mathbf{x}\}$, maximization of the standard variational lower bound on the likelihood in a generative model \mathcal{M}_L gives rise to maximization of a lower bound on the mutual information in $\tilde{\mathcal{M}}_I$. This bound is weaker or as tight as $\hat{I}_q(\mathbf{x}, y) = \langle \log p(\mathbf{x}|y) \rangle_{q(y|\mathbf{x})\tilde{p}(\mathbf{x})} + H_{\tilde{p}}(\mathbf{x})$, where $q(y|\mathbf{x})$ is the approximate posterior of the generative model.*

Sketch of the proof:

- by definition, the exact value of mutual information $I(\mathbf{x}, y)$ for model $\tilde{\mathcal{M}}_I$ is given by

$$I(\mathbf{x}, y) = H_{\tilde{p}}(\mathbf{x}) + \langle \log \tilde{p}(\mathbf{x}|y) \rangle_{q(y|\mathbf{x})\tilde{p}(\mathbf{x})},$$

where $\tilde{p}(\mathbf{x}|y) \propto \tilde{p}(\mathbf{x})q(y|\mathbf{x})$ is the exact decoder of $\tilde{\mathcal{M}}_I$

- from the non-negativity of the KL:

$$I(\mathbf{x}, y) \geq \hat{I}_q(\mathbf{x}, y) \stackrel{\text{def}}{=} H_{\tilde{p}}(\mathbf{x}) + \langle \log p(\mathbf{x}|y) \rangle_{q(y|\mathbf{x})\tilde{p}(\mathbf{x})}$$

- $I(\mathbf{x}, y) \geq \hat{I}_q(\mathbf{x}, y) \geq H_{\tilde{p}}(\mathbf{x}) + \tilde{\mathcal{L}}$

Sufficient Conditions for Equivalence

Equivalence condition:

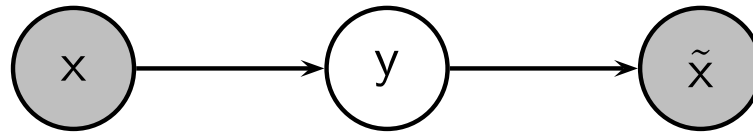
- when do maximizations of the lower bound $\hat{I}_q(x, y)$ on the mutual information and the variational Jensen's bound $\tilde{\mathcal{L}}$ on the likelihood lead to identical fixed points for $q(y|x)$ and $p(x|y)$?
- a sufficient condition: $\langle KL(q(y|x)||p(y)) \rangle_{\tilde{p}(x)} \sim const$
- e.g. $p(y)$ is flat, $q(y|x)$ is a Gaussian with a fixed noise

Generally:

- standard variational approaches to likelihood maximization in generative models may be viewed as a way to optimize a specific lower bound on the information content $I(x, y)$ in the corresponding encoder model $\tilde{\mathcal{M}}_I$
- a generally tighter (yet tractable) bound on the mutual information is given by $\hat{I}_q(x, y)$, which is optimized by a special case of the IM algorithm for $\tilde{\mathcal{M}}_I$

Variational IM and Feed-Forward Models

- consider a generalized chain $\mathcal{M}_C \stackrel{\text{def}}{=} p(y|x)p(\tilde{x}|y)$, where x and \tilde{x} are visible, and y 's are hidden:



- learning by optimizing the **conditional likelihood**:
 $\mathcal{L}_{\tilde{x}|x} = \langle \log \langle p(\tilde{x}|y) \rangle_{p(y|x)} \rangle_{\tilde{p}(x,\tilde{x})}$, where $\tilde{p}(\tilde{x}, x) \propto \sum_m \delta(\tilde{x} - \tilde{x}^{(m)}) \delta(x - x^{(m)})$
- exact inference: $p(y|x, \tilde{x}) \propto p(\tilde{x}|y)p(y|x)$
- define a recognition model $\mathcal{M}_{IC} \stackrel{\text{def}}{=} \tilde{p}(x, \tilde{x})p(y|x, \tilde{x})$, where $p(y|x, \tilde{x})$ is the exact posterior of the feed-forward model \mathcal{M}_C

Proposition 2: For i.i.d. patterns $\{x, \tilde{x}\}$, conditional likelihood learning in the feed-forward model \mathcal{M}_C gives rise to maximization of a lower bound on the conditional mutual information $I(\tilde{x}, y|x)$ in \mathcal{M}_{IC} . This bound is weaker or as tight as $\hat{I}_C(\tilde{x}, y|x) \stackrel{\text{def}}{=} \langle \log p(\tilde{x}|x, y) \rangle_{p(y|x,\tilde{x})\tilde{p}(x,\tilde{x})} + H_{\tilde{p}}(\tilde{x}|x)$.

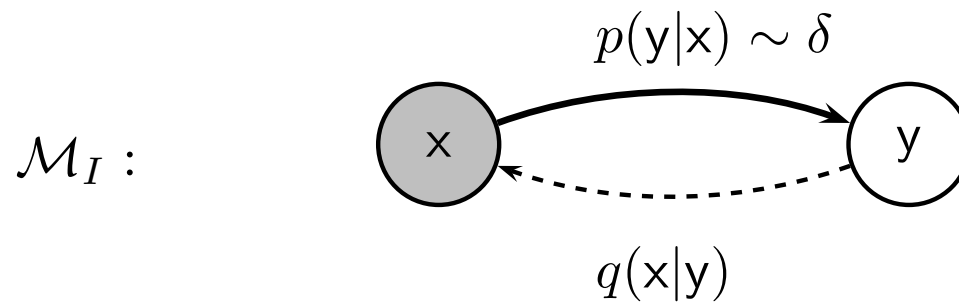
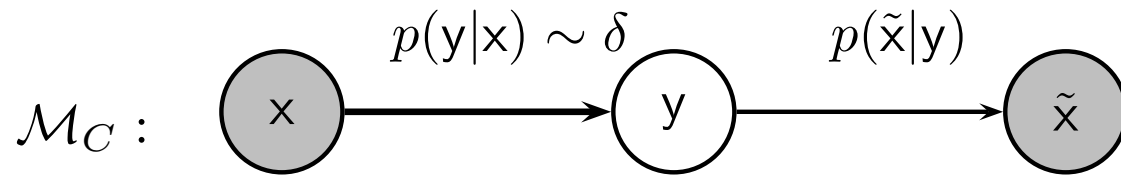
Variational IM and Fano's Inequality

- in the special case when the chain model \mathcal{M}_C and the empirical distribution $\tilde{p}(\mathbf{x}, \tilde{\mathbf{x}})$ define an **autoencoder**, the bound $\hat{I}_C(\tilde{\mathbf{x}}, \mathbf{y}|\mathbf{x})$ gives a *model-specific* upper bound on the probability of correct reconstructions
- $\mathcal{L}_{\tilde{\mathbf{x}}|\mathbf{x}} \leq \hat{I}_C(\tilde{\mathbf{x}}, \mathbf{y}|\mathbf{x}) - H_{\tilde{p}}(\tilde{\mathbf{x}}|\mathbf{x})$
- this contrasts with Fano's inequality (Fano (1961)):

$$p_e(\tilde{\mathbf{x}} \neq \mathbf{x}) \geq (H(\mathbf{x}) - I(\mathbf{x}, \mathbf{y}) - \log 2) / \log(|\mathbf{x}| - 1)$$

where $|\mathbf{x}|$ is the number of possible distinct reconstructions (size of the input alphabet)

Variational IM and Noiseless Autoencoders



$$\tilde{I}(x, y) = \mathcal{L}_{\tilde{x}|x} + H_{\tilde{p}}(x) \text{ for } q(x = s|y) = p(\tilde{x} = s|y)$$

- define a conditional model and the corresponding encoder as

$$\mathcal{M}_C \stackrel{\text{def}}{=} p(y|x)p(\tilde{x}|y), \mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(x)p(y|x) \text{ (N.B.: } p(y|x) \sim \delta)$$

Variational IM and Noiseless Autoencoders – 2

- *self-supervised* models (or autoencoders): the outputs $\{\tilde{x} \in \mathcal{R}_{\tilde{x}}\}$ are the exact uncorrupted copies of the sources $\{x \in \mathcal{R}_x\}$ for all the i.i.d. training patterns $(x^{(i)}, \tilde{x}^{(i)})$, $i \in \{1, \dots, M\}$
- the case of **noiseless autoencoders** ($p(y|x) \sim \delta$) is straight-forward

Proposition 3: *For i.i.d. patterns $\{x\}$, exact conditional likelihood learning in noiseless autoencoders \mathcal{M}_C is equivalent to maximizing the generic lower bound on mutual information $\tilde{I}(x, y) = \langle \log q(x|y) \rangle_{p(y|x)\tilde{p}(x)} + H_{\tilde{p}}(x)$ in noiseless channels \mathcal{M}_I , where the variational decoder $q(x|y)$ is constrained to be equivalent to the decoding distribution of the autoencoder.*

- optimization of the exact conditional likelihood $\mathcal{L}_{\tilde{x}|x}$ in conventional *noiseless* autoencoders is equivalent to a special case of the variational information-maximization

Exact IM and Noiseless Autoencoders

- the bound $I(x, y) \geq \tilde{I}(x, y) = \mathcal{L}_{\tilde{x}|x} + H_{\tilde{p}}(x)$ is saturated if $\forall s \in \mathcal{R}_x \equiv \mathcal{R}_{\tilde{x}}, y \in \mathcal{R}_y$

$$p_{\tilde{x}|y}(\tilde{x} = s|y) \propto p_{y|x}(y|x = s)\tilde{p}(x = s)$$

where we write $p_{\tilde{x}|y}$ and $p_{y|x}$ to explicitly refer to the decoding and encoding parts of autoencoder \mathcal{M}_C

- minimization of the *reconstruction error in noiseless autoencoders* \mathcal{M}_C and maximization of the *exact mutual information in noiseless encoder models* \mathcal{M}_I for **Bayes-optimal decoders** are equivalent
- this result changes for *stochastic autoencoders*

Variational IM and Stochastic Autoencoders

- for stochastic autoencoders ($p(y|x) \neq \delta$), maximization of $\mathcal{L}_{\tilde{x}|x} = \langle \log \langle p(\tilde{x}|y) \rangle_{p(y|x)} \rangle_{\tilde{p}(x,\tilde{x})}$ is no longer tractable
- variational EM for the conditional likelihood:

$$\mathcal{L}_{\tilde{x}|x} \geq \tilde{\mathcal{L}}_{\tilde{x}|x} = \langle \log p(\tilde{x}|x, y) \rangle_{q(y|x, \tilde{x})\tilde{p}(x, \tilde{x})} - \langle KL(q(y|x, \tilde{x}) || p(y|x)) \rangle_{\tilde{p}(x, \tilde{x})}$$

where $q(y|x, \tilde{x})$ is an arbitrary variational posterior, which we constrain to ensure the tractability of computations

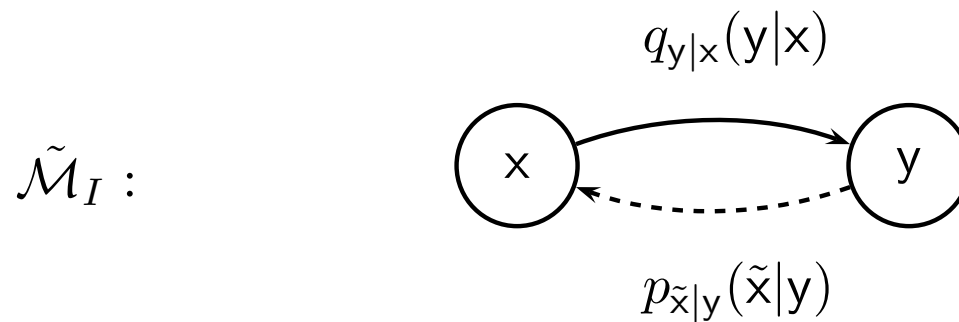
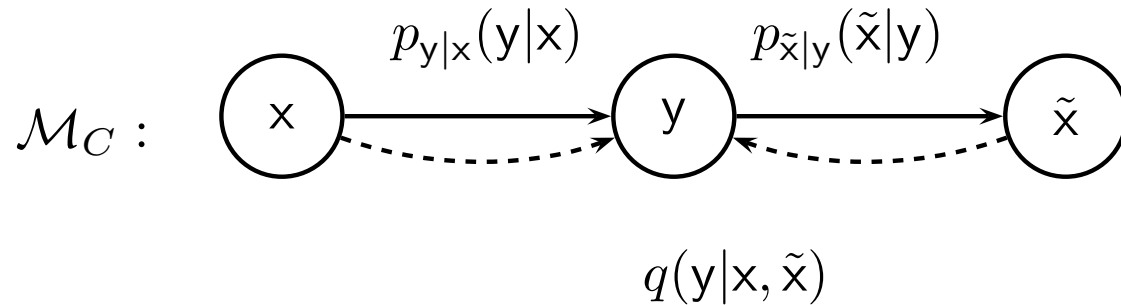
- define the corresponding encoder model $\tilde{\mathcal{M}}_I = q_{y|x}(y|x)\tilde{p}(x)$, where

$$q_{y|x}(y|x = s) \stackrel{\text{def}}{=} q(y|x = s, \tilde{x} = s)$$

for all $s \in \mathcal{R}_x = \mathcal{R}_{\tilde{x}}$

- (again, this is the *equivalent (variational) inference* assumption)

Variational IM and Stochastic Autoencoders – 2



$$\tilde{I}(x, y) = \tilde{\mathcal{L}}_{\tilde{x}|x} + H_{\tilde{p}}(x) \text{ for } q_{y|x}(y|x) = q(y|x = \tilde{x})$$

Variational IM and Stochastic Autoencoders – 3

Proposition 4: For i.i.d. patterns $\{x, \tilde{x}\}$, optimization of the standard variational Jensen's bound on the conditional likelihood $\tilde{\mathcal{L}}_{\tilde{x}|x}$ in stochastic autoencoders $\mathcal{M}_C = p(y|x)p(\tilde{x}|y)$ reduces to maximization of a specific variational lower bound on the mutual information $\tilde{I}(x, y)$ in the stochastic memoryless channel $\tilde{\mathcal{M}}_I = q_{y|x}(y|x)\tilde{p}(x)$, where $q_{y|x}(y|x) \stackrel{\text{def}}{=} q(y|x = \tilde{x})$

Sketch of the proof:

- since $\tilde{p}(\tilde{x}|x) \sim \delta(\tilde{x} - x)$, $q(y|x \neq \tilde{x})$ do not affect the bound $\tilde{\mathcal{L}}_{\tilde{x}|x}$
- express the fixed point updates of the variational EM on the bound $\tilde{\mathcal{L}}_{\tilde{x}|x}$ and compare them with the variational IM for the encoder $\tilde{\mathcal{M}}_I = q_{y|x}(y|x)\tilde{p}(x)$
- this may be carried out by iterative substitutions of the likelihood-optimal ($\tilde{\mathcal{L}}_{\tilde{x}|x}$) variational parameters back into $\tilde{\mathcal{L}}_{\tilde{x}|x}$

Variational IM and Stochastic Autoencoders – 4

Sketch of the proof (continued):

- fixed points for the t^{th} iteration of the variational EM algorithm:

1. $p_{y|x}^{(t-1)}(y|x = s^{(m)}) = q^{(t-1)}(y|x = \tilde{x} = s^{(m)})$

2. $q^{(t)}(y|x = \tilde{x} = s^{(m)}) = \arg \max_q \langle \log p_{\tilde{x}|y}^{(t-1)}(x = s^{(m)}|y) \rangle_{q(y|x=\tilde{x}=s^{(m)})}$
(denote as $\dots \stackrel{\text{def}}{=} r^{(t)}(y|x = s^{(m)})$)

3. $p_{\tilde{x}|y}^{(t)}(\tilde{x} = s^{(m)}|y) = \arg \max_{p_{\tilde{x}|y}} \langle \log p_{\tilde{x}|y}(x = s^{(m)}|y) \rangle_{r^{(t)}(y|x=s^{(m)})}$

- combine (1) and (2) to express the optimal encoder for the next iteration:

$$p_{y|x}^{(t)}(y|x = s^{(m)}) = q^{(t)}(y|x = s^{(m)}, \tilde{x} = s^{(m)}) \in \mathcal{F}_{q_{y|x}} \subseteq \mathcal{F}_{p_{y|x}},$$

where $\mathcal{F}_{q_{y|x}} \subseteq \mathcal{F}_{p_{y|x}}$ are families of variational and exact posteriors respectively

Variational IM and Stochastic Autoencoders – 5

- the fixed points of $\tilde{\mathcal{L}}_{\tilde{x}|x}$ for the autoencoder's decoding and encoding mappings $p_{\tilde{x}|y}^{(t)}, p_{y|x}^{(t)}$ are equivalent to the ones obtained by the iterative maximization of

$$\tilde{I}(x, y) = \langle \log p_{\tilde{x}|y}(x|y) \rangle_{q_{y|x}(y|x)\tilde{p}(x)} + H_{\tilde{p}}(x)$$

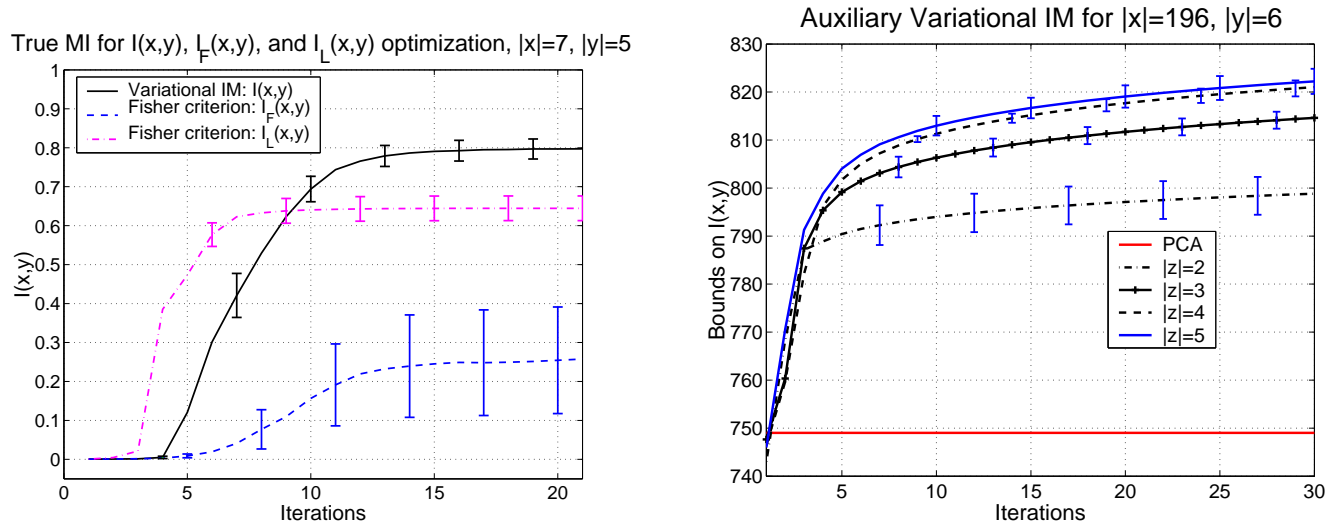
where $q_{y|x}(y|x) \equiv q(y|x = \tilde{x}) \in \mathcal{F}_{q_{y|x}}$ for all $x, \tilde{x} \in \mathcal{R}_x$

- if $\langle \log p_{\tilde{x}|y}(x|y) \rangle_{p_{y|x}(y|x)}$ is tractable, we may choose $\mathcal{F}_{q_{y|x}} = \mathcal{F}_{p_{y|x}}$, i.e.

$$\tilde{I}(x, y) = \langle \log p_{\tilde{x}|y}(x|y) \rangle_{p_{y|x}(y|x)\tilde{p}(x)} + H_{\tilde{p}}(x), \quad p_{y|x}(y|x) \in \mathcal{F}_{q_{y|x}} = \mathcal{F}_{p_{y|x}}$$

- the variational conditional likelihood training in autoencoders may be viewed as a special case of the variational IM algorithm for the stochastic channel $\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} \tilde{p}(x)p_{y|x}(y|x)$ (or $\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} \tilde{p}(x)q_{y|x}(y|x)$)
- naive optimization of $\tilde{\mathcal{L}}_{\tilde{x}|x}$ is more computationally demanding than the variational IM

Variational IM: Further Observations



- **Variational IM:** favorable empirical evidence compared with some other approximations, e.g. based on the Fisher Information criterion (Brunel and Nadal (1998), Corduneanu and Jaakkola (2003))
- *left plot:* the exact $I(x, y)$ for $p(y|x) \stackrel{\text{def}}{=} \prod_i \sigma((2y_i - 1)(v_i^T x + b_i))$, $q(x|y) \sim \mathcal{N}(Wx, \sigma^2)$, $v_{i,j} \in [-1, 1]$, $M = 20$, $x \sim \mathcal{N}_x$
- *right plot:* IM bounds on $I(x, y)$ for $p(y|x) \sim \mathcal{N}(Wx, \sigma^2)$ and $q(x|y) \sim \sum_z \mathcal{N}(U_z Y, s_z^2)p(z|y)$ for different choices of $|z|$, $M = 30$, $x \sim$ handwritten digits (scaled MNIST) – conditional mixture bounds

Variational IM: Summary

- we looked at a relation of the **variational Blahut-Arimoto (Information Maximizing)** algorithm for training encoder models with the maximum likelihood learning in generative models and self-supervised learning in stochastic autoencoders
- in contrast to previous studies, focused on stochastic mappings (rather than invertible special cases)
- assumption: **equivalence of the (exact or variational) inference**
- derived sufficient conditions for equivalence of the variational EM and IM for generative and encoder models
- showed the equivalence of the variational EM and IM for conditional and encoder models (N.B.: practically, the IM formulation is simpler)
- showed that some of the common approximations of $I(x, y)$ are special cases of the variational IM
- outlined further comparisons and extensions