CSC2535: Computation in Neural Networks
Lecture 7:
Variational Bayesian Learning & Model Selection

(non-examinable material)

Matthew J. Beal

February 27, 2004

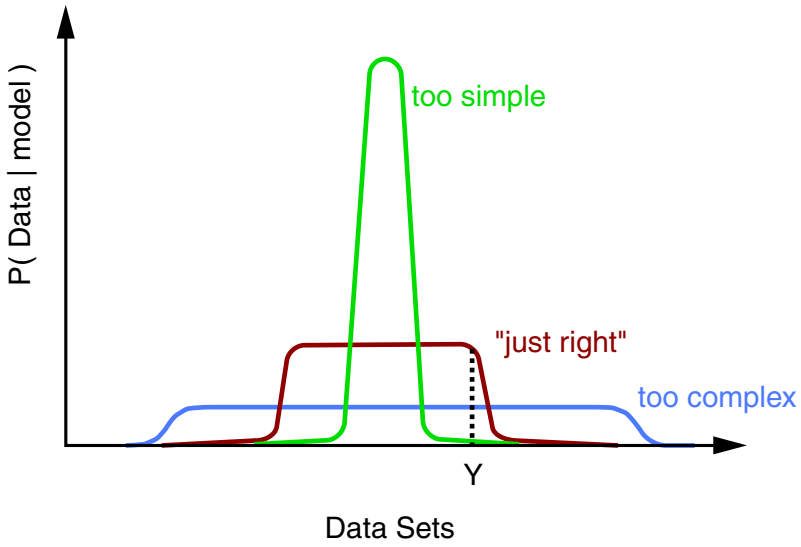
www.variational-bayes.org

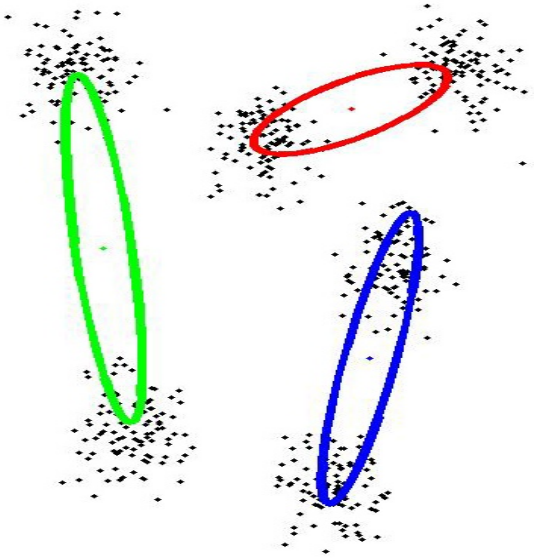
Bayesian Model Selection

Using Bayes' rule, select the model m_j with the highest probability given the data $\mathbf{y}$:

$$p(m_j \,|\, \mathbf{y}) = \frac{p(m_j)\, p(\mathbf{y} \,|\, m_j)}{p(\mathbf{y})}, \qquad \underbrace{p(\mathbf{y} \,|\, m_j) = \int d\boldsymbol{\theta}_j \; p(\boldsymbol{\theta}_j \,|\, m_j) p(\mathbf{y} \,|\, \boldsymbol{\theta}_j, m_j)}_{\text{marginal likelihood}}$$

(Sampling) interpretation of $p(\mathbf{y} \,|\, m_j)$: The probability that *randomly selected* parameter values from the model class m_j would generate data set $\mathbf{y}$.





- Model classes that are too simple are unlikely to generate the data set.
- Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.

Some examples of model selection

Structure learning The pattern of arcs between variables in a graphical model implies a set of conditional independence (CI) relationships; the structure learning problem is inferring the CI relationships that hold given a set of (complete or incomplete) observations of the variables. A related problem is learning the *direction* of the dependencies (i.e. $A \rightarrow B$, or $B \rightarrow A$).

Input dependence Selecting which input (i.e. explanatory) variables are needed to predict the output (i.e. response) variable in a regression/classification task can be equivalently cast as deciding whether each input variable is a parent (or, more accurately, an ancestor) of the output variable in the corresponding directed graph.

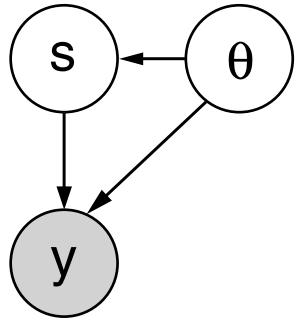
Cardinality Many statistical models contain discrete nominal latent variables, but their cardinalities are often unknown. Examples include deciding how many mixture components are required in a finite mixture model, or how many hidden states are needed in a hidden Markov model.

Dimensionality Other statistical models contain real-valued vectors of latent variables: model selection examples include choosing the intrinsic dimensionality in a probabilistic principal components analysis (pPCA) or factor analysis (FA) model, or the state-space dimensionality of a linear-Gaussian state-space model.

Marginal likelihoods can be intractable to compute

The marginal likelihood is often a difficult integral to compute

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta})$$



because of the high dimensionality of the parameter space, analytical intractability, but also due to the presence of hidden variables, $\mathbf{s}$:

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \int d\mathbf{s}\; p(\mathbf{y}, \mathbf{s} \,|\, \boldsymbol{\theta}, m)$$

Example: A mixture model with K components, and we model $n = 100$ data points.

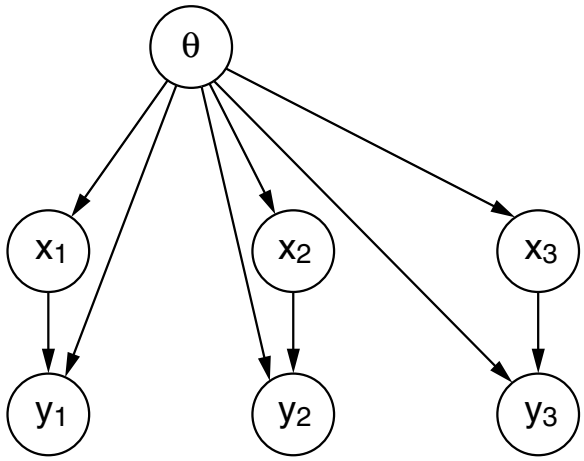
• The marginal likelihood for that model includes a sum over all possible joint settings of hidden variables (the component indicator variables), which is K^n terms.
• So, in a mixture model with even just 2 components, this becomes ridiculous.

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) \;, \quad \text{and} \quad p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{s}_i=1}^{K} p(\mathbf{s}_i \,|\, \boldsymbol{\theta})\, p(\mathbf{y}_i \,|\, \mathbf{s}_i, \boldsymbol{\theta})$$

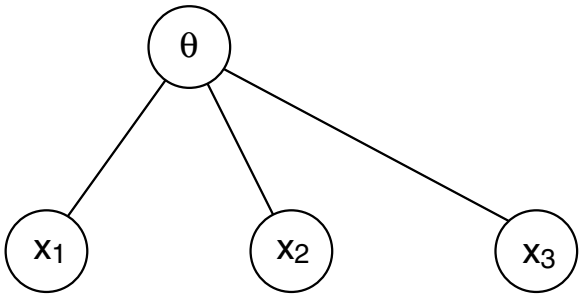
$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \sum_{\mathbf{s}_i=1}^{K} p(\mathbf{s}_i \,|\, \boldsymbol{\theta}) p(\mathbf{y}_i \,|\, \mathbf{s}_i, \boldsymbol{\theta}) \quad \ldots \quad \text{this has } 2^{100} \text{ terms!!}$$

Understanding why marginal likelihoods are intractable

Integrating out the parameters $\boldsymbol{\theta}$ couples the posterior distributions over the hidden variables for **every** data point:



(a) A generative graphical model for 3 i.i.d. data points, each with one hidden variable.



(b) Graph representing the exact posterior.

(a) Given the parameters, each $(\mathbf{x}_i, \mathbf{x}_j)$ pair $(i \neq j)$ are independent. Inference is simple and i.i.d. (for inference use your standard EM, or E-step constrained EM).

(b) If the parameters are uncertain quantities (unobserved), then the exact posterior couples $(\mathbf{x}_i, \mathbf{x}_j)$ pairs through $\boldsymbol{\theta}$. They are conditionally independent given $\boldsymbol{\theta}$, but marginally dependent.

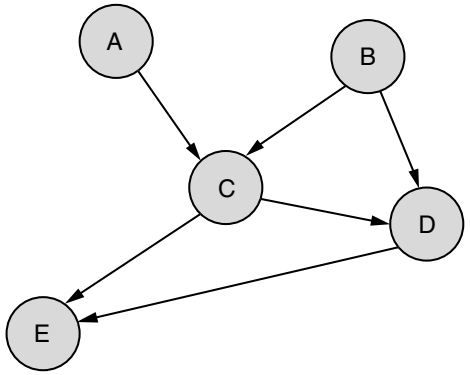
A structure model selection task

Which of the following graphical models is the data generating process?

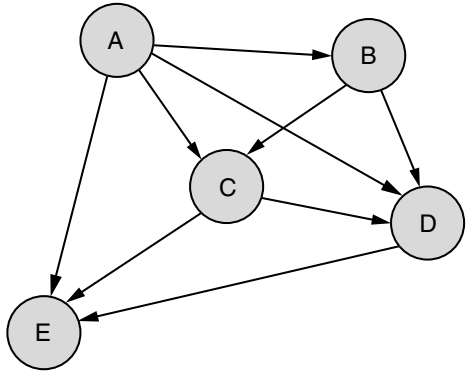
Discrete-valued directed acyclic graphical models: data $\mathbf{y} = (A, B, C, D, E)^n$

ALL OBSERVED

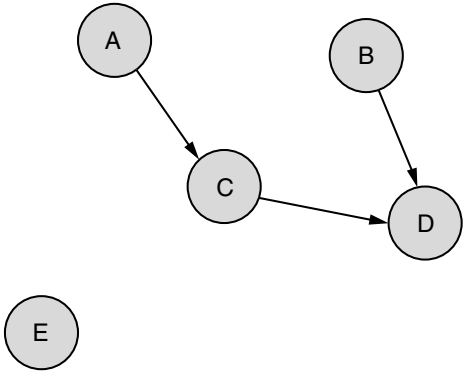
true



complex



simple

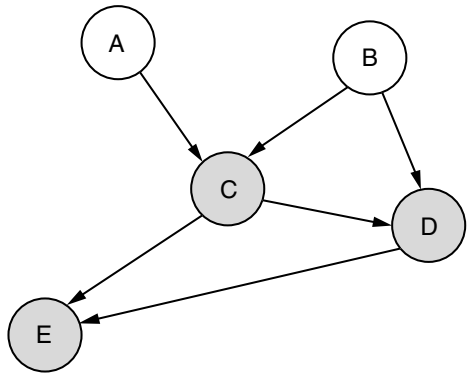


marginal likelihood tractable

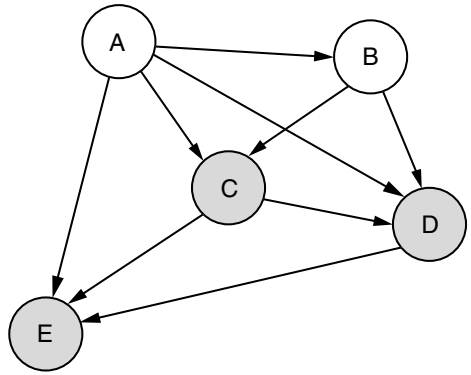
If the data are just $\mathbf{y} = (C, D, E)^n$, and $\mathbf{s} = (A, B)^n$ are **hidden** variables... ?

OBS.+HIDDEN

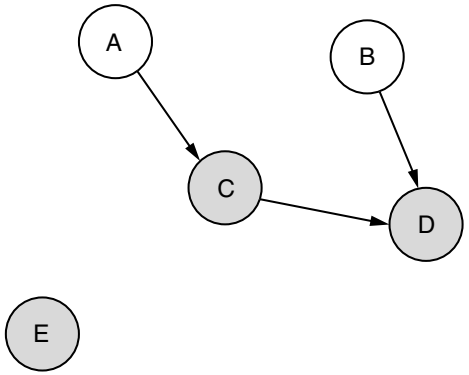
true



complex



simple



marginal likelihood intractable

Practical Bayesian methods

- Laplace approximations: appeal to Central Limit Theorem.
  - Makes a Gaussian approximation about a maximum *a posteriori* estimate, $\hat{\boldsymbol{\theta}}$.

$$\ln p(\mathbf{y} \,|\, m) \approx \ln p(\hat{\boldsymbol{\theta}} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) + \tfrac{d}{2} \ln 2\pi - \tfrac{1}{2} \ln |H|$$

- Large sample approximations: as $n \to \infty$, prior term vanishes.
  - Bayesian Information Criterion (BIC): $\ln p(\mathbf{y} \,|\, m) \approx \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \frac{d}{2} \ln n$
  - Cheeseman-Stutz (CS): $\ln p(\mathbf{y} \,|\, m) \approx \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}})$

- Markov chain Monte Carlo (MCMC): guaranteed to converge in the limit.
  - But many samples required for accurate results, hard to assess convergence.
  - Posterior is stored as a set of samples, which can be inefficient.

- **Variational approximations**: *this changes the objective function, from a marginal likelihood to a negative free energy.*
  - Here we construct a lower bound that is tractable to compute.
  - We also obtain tractable, efficient, and intuitive inference & learning steps.

(review, lecture 5): The lower bound interpretation of EM

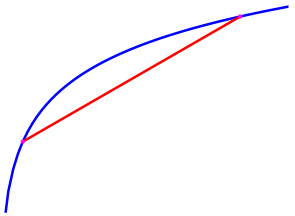
If $\mathbf{y}$ is the observed data, and $\mathbf{x}$ are hidden variables, then the log probability of the data is given by integrating out $\mathbf{x}$; also the ML parameter setting, $\boldsymbol{\theta}_{\text{ML}}$, is given by

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \ln \int d\mathbf{x}\; p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) \;, \qquad \boldsymbol{\theta}_{\text{ML}} \equiv \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}) \;.$$

Let's form a *free-energy*, by lower bounding the likelihood $\mathcal{L}(\boldsymbol{\theta})$ using Jensen's inequality

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) = \ln \int d\mathbf{x}\; p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) &= \ln \int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta})}{q_{\mathbf{x}}(\mathbf{x})} \\ &\geq \int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta})}{q_{\mathbf{x}}(\mathbf{x})} \\ &\equiv \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y}) \end{aligned}$$

log function is *concave*

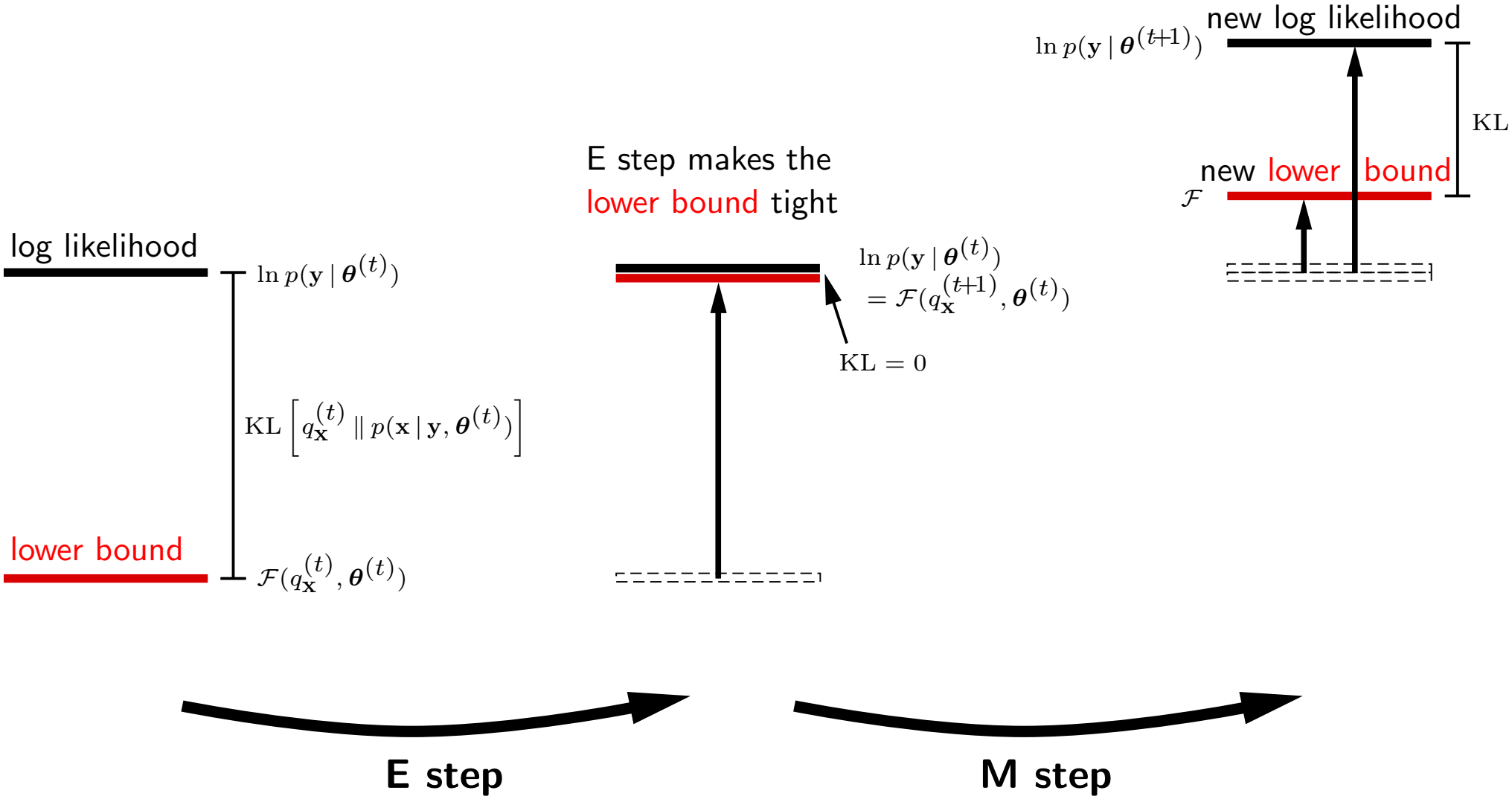


$\mathcal{F}(\cdot)$ is a lower bound on $\mathcal{L}$ — *for any* distribution we choose for $q_{\mathbf{x}}(\mathbf{x})$ — some $q(\cdot)$ will give tighter bounds than others.

$$\underbrace{\ln p(\mathbf{y} \,|\, \boldsymbol{\theta})}_{\substack{\text{desired} \\ \text{quantity}}} - \underbrace{\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}, \mathbf{y})}_{\text{computable}} = \underbrace{\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})}{p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta})} = \mathbf{KL}(q \| p)}_{\text{measure of inaccuracy of approximation}}$$

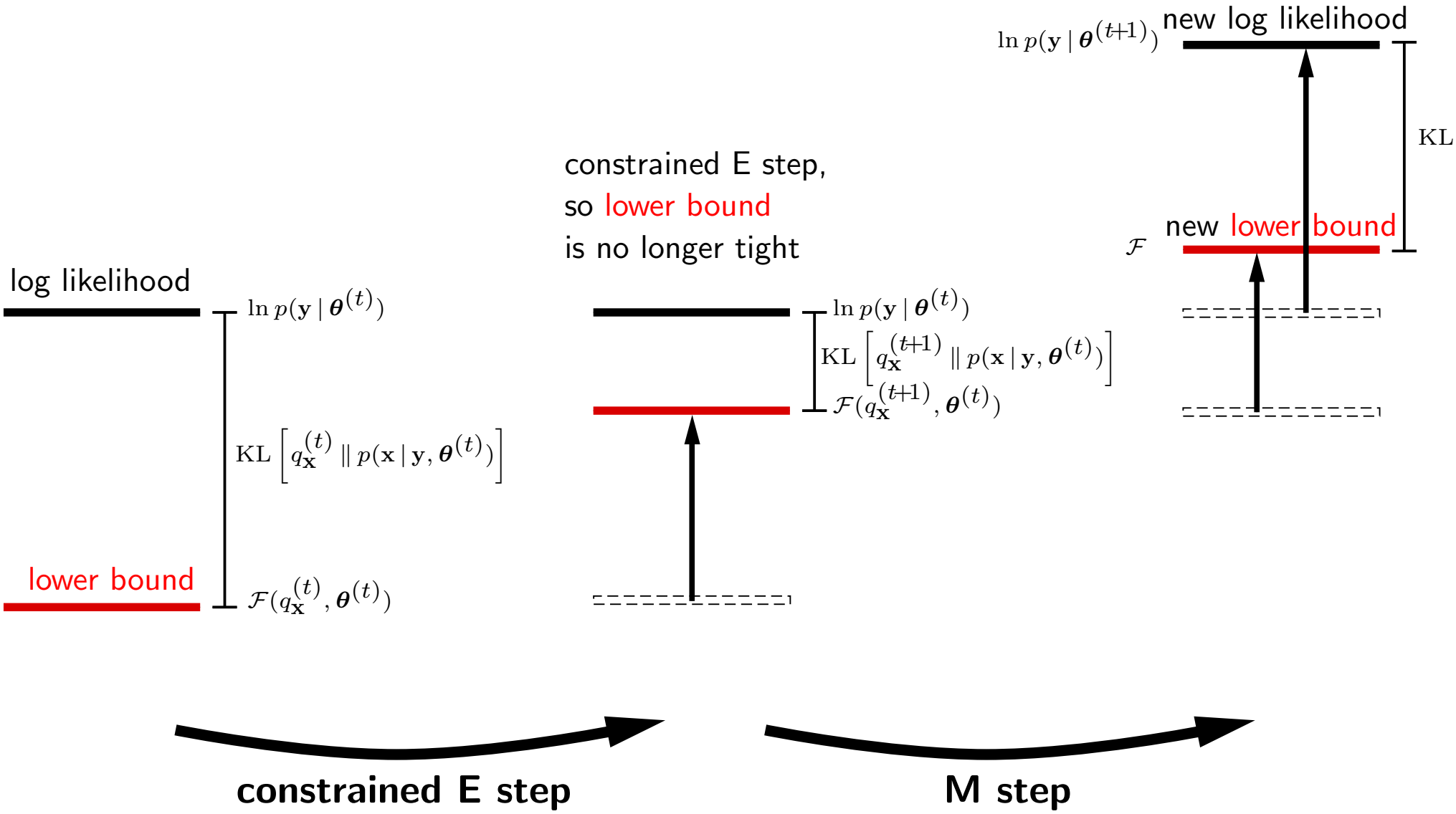
(review, lecture 5): Cartoon of ML EM learning using F

If the form of $q_{\mathbf{x}}(\mathbf{x})$ is not constrained to any particular family, then the bound can be made **tight** on every **E** step.



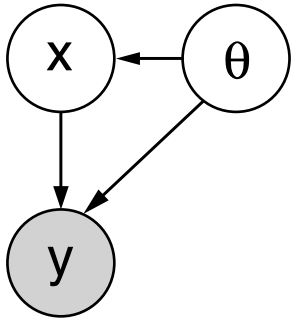
(review, lecture 5): Cartoon of ML EM learning using F

If the form of $q_{\mathbf{x}}(\mathbf{x})$ is not flexible enough to capture the hidden variable posterior distribution, then the bound is **loose** on the **E** step, by an amount which is the KL divergence between the approximate distribution $q_{\mathbf{x}}(\mathbf{x})$ and the true posterior $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$.



Lower Bounding the Marginal Likelihood

Variational Bayesian Learning

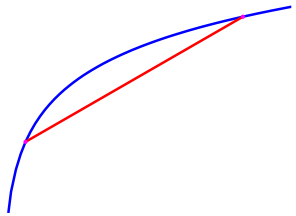


Let the hidden states be $\mathbf{x}$, data $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$.

We can lower bound the marginal likelihood (Jensen's inequality):

$$\begin{aligned}
\ln p(\mathbf{y} \,|\, m) &= \ln \int d\mathbf{x}\, d\boldsymbol{\theta}\; p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m) \\
&= \ln \int d\mathbf{x}\, d\boldsymbol{\theta}\; q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} \\
&\geq \int d\mathbf{x}\, d\boldsymbol{\theta}\; q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})}.
\end{aligned}$$

log function is still *concave*!



Use a simpler, factorised approximation to $q(\mathbf{x}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\ln p(\mathbf{y} \,|\, m) \geq \int d\mathbf{x}\, d\boldsymbol{\theta}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} = \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}).$$

Optimising the VB lower bound using variational calculus

Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** updates:

$$\textbf{VBE} \qquad q^*_{\mathbf{x}}(\mathbf{x}) \;\propto\; \exp\left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta})\right] \qquad E-like\ \ step$$

$$\textbf{VBM} \qquad q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \;\propto\; p(\boldsymbol{\theta}) \exp\left[\int d\mathbf{x}\; q^*_{\mathbf{x}}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta})\right] \qquad M-like\ \ step$$

As before, maximizing $\mathcal{F}_m$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, q_{\mathbf{x}}(\mathbf{x})$ and the *true posterior*, $p(\boldsymbol{\theta}, \mathbf{x} \,|\, \mathbf{y}, m)$:

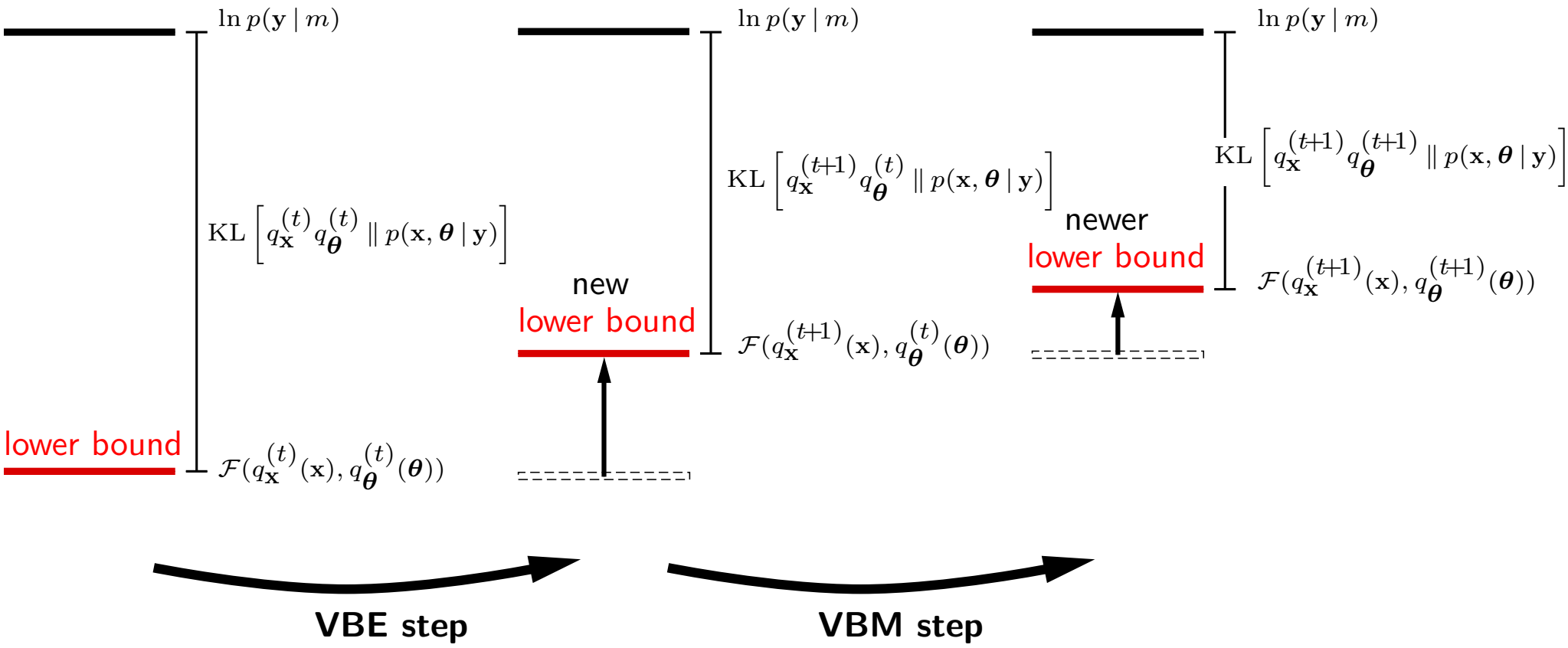
$$\underbrace{\ln p(\mathbf{y} \,|\, m)}_{\text{desired quantity}} - \underbrace{\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})}_{\text{computable}} = \underbrace{\int d\mathbf{x}\, d\boldsymbol{\theta}\; q_{\mathbf{x}}(\mathbf{x})\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)} = \mathbf{KL}(q \| p)}_{\text{measure of inaccuracy of approximation}}$$

In the limit as $n \to \infty$, for identifiable models, the variational lower bound approaches Schwartz's (1978) BIC criterion.

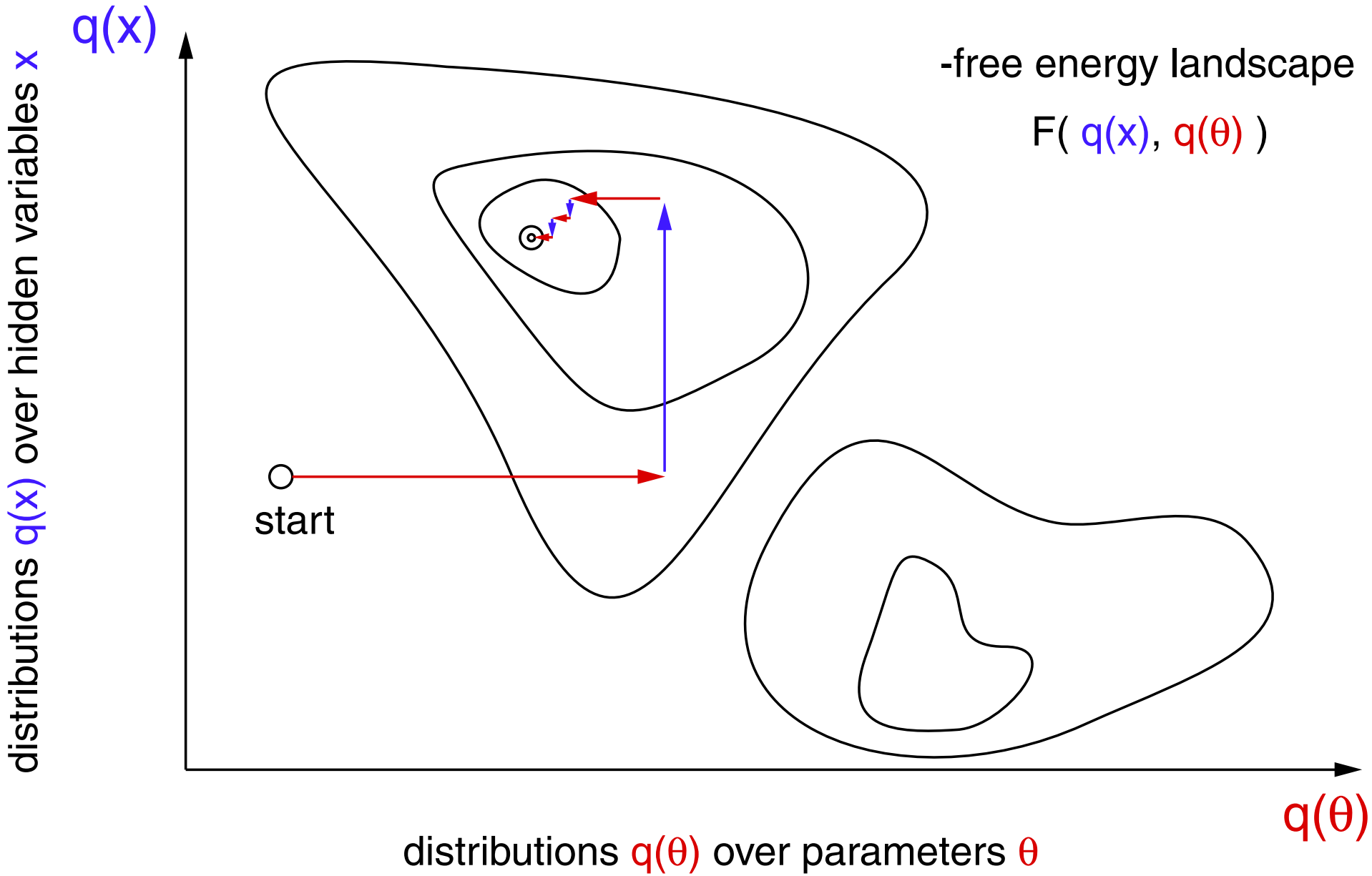
Cartoon of VB EM learning using F

Using the factorisation $q(\mathbf{x}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, we know the bound can't be tight bound for either **VBE** or **VBM** steps.

NB: log marginal likelihood $\ln p(\mathbf{y} \mid m)$ is constant (if the hyperparameters remain fixed).



VB-EM is just coordinate ascent in $(q(x), q(\theta))$ space



The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta} \,|\, \mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{y} \,|\, m)$

VB-E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \,|\, \mathbf{y}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\right]$$

Properties of VB-EM:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\boldsymbol{\phi}}$.

Examples of where Variational Bayesian EM is being used

The Variational Bayesian EM algorithm has been used to approximate Bayesian learning in a wide range of models, such as:

- probabilistic PCA and factor analysis (Bishop, 1999)
- mixtures of Gaussians (Attias, 1999)
- mixtures of factor analysers (Ghahramani & Beal, 1999)
- state-space models (Ghahramani & Beal, 2000; Beal, 2003)
- ICA, IFA (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- mixtures of experts (Ueda & Ghahramani, 2000)
- hidden Markov models (MacKay, 1995; Beal, 2003)

The main advantage is that it can be used to **automatically do model selection** and does not suffer from overfitting to the same extent as ML methods do.

Also it is about as computationally demanding as the usual EM algorithm.

See: www.variational-bayes.org

More to VB than we can cover

- Empirically VB seems to do well on model selection problems.
- But how can we be sure the bound is equally tight for different models?
  - We can't! — by the very intractability of the integral.
  - But it is feasible to enumerate all possibilities when n is small.
- We can use clever sampling techniques to get a handle on the marginal likelihood
  - importance sampling using the variational approximation as importance dist$^{\mathrm{n}}$, and
  - unbiased marginal likelihood estimates using Neal's Annealed Importance Sampling.
- The VB algorithm can be analysed closely for *Conjugate-Exponential* models.
- Theoretical guarantees of improvement over some methods, e.g. *Cheeseman-Stutz*.
- There are more sophisticated variational methods available to us, e.g. *Bethe*, *Kikuchi*, and *generalised belief propagation*. But these have not yet been successfully applied to *Bayesian* integrals — only to the E-steps of standard EM algorithms.

Extra I — Variational calculus, aka freeform extremisation of $\mathcal{F}$

Optimal forms of $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ are obtained by taking functional derivatives of $\mathcal{F}$ wrt to each, keeping the other constant, and finding where in q-space the derivative is zero.

A Langrange multiplier constraint is required to ensure the variational distribution is properly normalised. For example, for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, a Langragian is

$$\mathcal{R}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}) + \lambda \left(1 - \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) \;.$$

So taking the functional derivative:

$$\begin{aligned}\frac{\partial}{\partial q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\mathcal{R}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) &= \frac{\partial}{\partial q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x})\; \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m) + \ln \frac{p(\boldsymbol{\theta} \,|\, m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\right] - \lambda \\ &= \int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x})\; \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta} \,|\, m) - \ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \lambda \;.\end{aligned}$$

Setting this to 0, and rearranging produces

$$\ln q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta} \,|\, m) + \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) - \ln \mathcal{Z}_{\boldsymbol{\theta}}^{(t+1)} \;,$$

where $\mathcal{Z}_{\boldsymbol{\theta}}$ is the normalisation constant, or *partition function* (here it is exactly λ).

Extra II — Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) models, which satisfy **(1)** and **(2)**:

Condition (1). The joint probability over *variables* is in the exponential family:

$$p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})\; g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}, \mathbf{y})\right\}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, $\mathbf{u}$ are *sufficient statistics*

Condition (2). The prior over *parameters* is conjugate to this joint probability:

$$p(\boldsymbol{\theta} \,|\, \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\; g(\boldsymbol{\theta})^\eta \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}\right\}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- η: number of pseudo-observations
- $\boldsymbol{\nu}$: values of pseudo-observations

Extra III — Conjugate-Exponential examples

Some models in the **CE** family:

- Gaussian mixtures
- factor analysis, probabilistic PCA
- hidden Markov models and factorial HMMs
- linear dynamical systems and switching models
- discrete-variable belief networks

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Some models not in the **CE** family:

- Boltzmann machines, MRFs (no conjugacy)
- logistic regression (no conjugacy)
- sigmoid belief networks (not exponential)
- independent components analysis (not exponential)

One can often approximate these models with models in the **CE** family e.g. IFA (Attias, 1998).

Extra IV — A very useful result in CE models

Theorem Given an iid data set $\mathbf{y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model is **CE** then:

(a) $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also conjugate, *i.e.*

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}\right\}$$

(b) $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

$$q_{\mathbf{x}_i}(\mathbf{x}_i) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \tilde{\boldsymbol{\theta}})$$

where $\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \stackrel{?}{=} \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$

KEY points:

(a) the approximate parameter posterior is of the same form as the prior;

(b) the approximate hidden variable posterior, *averaging over all parameters*, is of the same form as the *exact* hidden variable posterior under $\tilde{\boldsymbol{\theta}}$.

further reading on variational Bayesian methods

Variational methods

- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to variational methods in graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.
- R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–369. Kluwer Academic Publishers, 1998.

Variational Bayesian methods

- H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, 1999.
- M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, UK, 2003.
- C. M. Bishop. Variational PCA. In *Proc. Ninth Int. Conf. on Artificial Neural Networks. ICANN*, 1999.
- Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.
- Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems 13*, Cambridge, MA, 2001. MIT Press.
- J. W. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, December 2000.
- N. Ueda and Z. Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 2002.

Advanced variational methods (for the E-step)

- J. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, MA, 2001. MIT Press.
- A. L. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. Technical report, Smith-Kettlewell Eye Research Institute, 2001.