Probabilistic Modelling and Reasoning

Sensitivity $P(y=1|x=1)$ TP. Specificity $P(y=0|x=0)$ TN.

d dimensional, each element takes N values, to specify $P(\mathbf{x},\mathbf{y},\mathbf{z})$ need $K^{3d}-1$ values. $p(\mathbf{z})=\prod p(z_i)$ needs $d(K-1)<K^d-1$. $p(x_d|x_1,...,x_{d-1})$ needs $K^{d-1}(K-1)$.

Ordered Markov property if, for each i, there is a minimal subset of variables $\pi_i \subseteq \mathrm{pre}_i$ st. $p(x)$ satisfies $x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \pi_i) \mid \pi_i$ ($\Leftrightarrow p(\mathbf{x})=\prod_{i=1}^d p(x_i|\pi_i)$).

$p(x,y|z)=p(x|z)p(y|z)$, $p(x|y,z)=p(x|z)$ ($p(y,z)>0$) $\Leftrightarrow x \perp\!\!\!\perp y \mid z$ ($p(z)>0$).

The sets X and Y are said to be d-separated by Z if every trail from any variable in X to any variable in Y is blocked by Z.

Directed local Markov property $x_i \perp\!\!\!\perp \mathrm{pre}_i \setminus \mathrm{pa}_i \mid \mathrm{pa}_i \Leftrightarrow x_i \perp\!\!\!\perp \mathrm{nondesc}(x_i) \setminus \mathrm{pa}_i \mid \mathrm{pa}_i$.

Factorisation $\Leftrightarrow$ ordered Markov property $\Leftrightarrow$ local directed Markov property $\Leftrightarrow$ global directed Markov property (all ind. from d-sep).

Markov blanket – minimal set of variables st. knowing their values makes x independent from the rest. Directed $MB(x)=$ {parents, children, co-parents}. Undirected $MB(x)=\mathrm{ne}(x)$.

Energy-based model $p(x_1,...,x_d)=\frac{1}{Z}\exp[-\sum_c E_c(\chi_c)]$.

Global Markov property – all independencies from graph separation. Let G be the undirected graph for $p(x_1,...,x_d)\propto \prod_c \phi_c(\chi_c)$ and X, Y, Z three disjoint subsets of $x_1,...,x_d$. If X and Y are separated by Z, then $X \perp\!\!\!\perp Y \mid Z$. Sound – graph separation does not indicate false ind. relations. Not complete – only allows to decide about independence, not about dependence.

Local Markov property relative to an undirected graph if p satisfies $\alpha \perp\!\!\!\perp X \setminus (\alpha \cup \mathrm{ne}(\alpha)) \mid \mathrm{ne}(\alpha)$ for all $\alpha \in X$.

Pairwise Markov property relative to an undirected graph if p satisfies $\alpha \perp\!\!\!\perp \beta \mid X \setminus \{\alpha,\beta\}$ for all non-neighbouring α, β.

p factorises according to $G \Rightarrow$ Global Markov $\Rightarrow$ Local Markov $\Rightarrow$ Pairwise Markov.

Intersection property holds for all distributions $p(\mathbf{x})>0$ for all values of $\mathbf{x}$ in its domain. Excludes deterministic relationships between the variables. If $A \perp\!\!\!\perp B \mid (C \cup D)$ and $A \perp\!\!\!\perp C \mid (B \cup D) \Rightarrow A \perp\!\!\!\perp (B \cup C) \mid D$.

A graph is said to be an independency map for a set of independencies I if the independencies asserted by the graph are part of I. For a directed graph G, let $I(G)$ be all independencies from d-separation. Distribution p satisfies independencies $I(p)$. $I(G)\subseteq I(p)$ for all p that factorise over G. $I(G)=I(p)$ graph is a perfect map for $I(p)$. A minimal I-map is a graph st. if you remove an edge, the graph is not an I-map any more.

Constructing undirected minimal I-maps using local Markov property for positive distribution $p>0$ (local independencies must imply global ones). For each node determine its Markov blanket $MB(x_i)$: minimal set of nodes U st. $x_i \perp\!\!\!\perp \mathrm{all} \setminus (x_i \cup U) \mid U$. Connect x_i to all nodes in $MB(x_i)$.

Why I-map? Local Markov by construction $\Rightarrow$ global Markov $\Rightarrow$ $X \perp\!\!\!\perp Y \mid Z$. Why minimal I-map? Remove $(x,u_1) \Rightarrow x_1 \perp\!\!\!\perp u_1 \mid u_2,u_3$, but then $MB(x_1)$ tells us this is not the case.

Constructing directed minimal I-maps. Assume an ordering. For each i, find a minimal subset of variables $\pi_i \subseteq \mathrm{pre}_i$ st $x_i \perp\!\!\!\perp \mathrm{pre}_i \setminus \pi_i \mid \pi_i$ holds in $I(p)$. Construct a graph with parents $\mathrm{pa}_i=\pi_i$.

Why I-map? Ordered Markov property $\Leftrightarrow$ factorisation. d-separation to detect independencies. Why minimal I-map? Remove an edge make indep. ass. that does not hold. Directed minimal I-maps are not unique, only a subset of indeps.

I-equivalence for directed graphs. Colliders without covering edge are called immoralities. Skeleton – which node is connected to which irrespective of direction. G_1 and G_2 are I-equivalent $\Leftrightarrow$ G_1 and G_2 have the same skeleton and the same set of immoralities.

Directed to undirected minimal I-map – moralisation. Form cliques for (x_i,pa_i). Undirected to directed minimal I-map – triangulation. Based on local Markov property. Independencies from the undirected graph with ordering.

Undirected graphs: unique I-maps, interactions are symmetrical, no natural ordering of vars, cannot represent 'explaining away' colliders. Directed graphs: I-equivalence, data generating process, ancestral sampling, forces directionality where there are none.

$\phi(x_1,...,x_d)$ has 2^d free parameters. $\prod_{i<j}\phi_{ij}(x_i,x_j)$ has $\binom{d}{2}2^2=\frac{d(d-1)}{2}2^2$ (num. edges).

Variable elimination. Heuristic to choose x^* is variable with least number of neighbours.

Sum-product message passing. Cost of marginal to messages: linear in number of variables d, exponential in maximum number of variables attached to a factor node. Recycling: most messages do not depend on χ_{target} and can be reused for computing $p(\chi_{\mathrm{target}})$. Messages correspond to effective factors obtained after marginalisation. Variables take K values and there are M elements in χ_i: $O(2dK^M)$.

Factor to variable messages. Eliminating variables $x_1,...,x_j$.
$\mu_{\phi\to x}(x)=\sum_{x_1,...,x_j}\phi(x_1,...,x_j,x)\prod_{i=1}^j \mu_{x_i\to\phi}(x_i)$.

Variable to factor messages. Multiplying effective factors.
$\mu_{x\to\phi}(x)=\prod_{i=1}^j \mu_{\phi_i\to x}(x)$.

Univariate marginals $p(x)=\prod_{i=1}^j \mu_{\phi_i\to x}(x)$.

Joint marginals $p(x_1,...,x_j)\propto \phi(x_1,...,x_j)\prod_{i=1}^j \mu_{x_i\to\phi}(x_i)$.

Can be used to compute conditionals, $\mathrm{argmax}_{\mathbf{x}}\, p(\mathbf{x})$. If not a tree, use variable elimination, condition on some variables st. the conditional is a tree.

Markov chain. If p satisfies ordered Markov property, the number of variables in the conditioning set can be reduced to a subset $\pi_i \subseteq \{x_1,...,x_{i-1}\}$.

If neither the transition nor emission distribution depend on i, we have a stationary (homogeneous) hidden Markov model.

d iid draws from a Gaussian mixture model with K mixture components. $h_i \perp\!\!\!\perp h_{i-1}$ and $\mathbf{v}_i \in \mathbb{R}^m$, $h_i \in \{1,...,K\}$. $p(h=k)=p_k$ and $p(\mathbf{v}|h=k)\sim\mathcal{N}$.

Filtering. Inferring the present. Marginal posterior. Alpha-recursion: $p(h_t|v_{1:t}) \propto \alpha(h_t)$.
$\phi_s(h_s, h_{s-1}) = p(h_s|h_{s-1})$, $f_s(h_s) = p(v_s|h_s)$, $\phi_1(h_1) = p(h_1)$.

$\alpha(h_1) = p(h_1)p(v_1|h_1) = p(h_1, v_1) \propto p(h_1|v_1)$
$\alpha(h_s) = \mu_{h_s \to \phi_{s+1}} = p(v_s|h_s) \sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}) = p(v_s|h_s)p(h_s, v_{1:s-1}) = p(v_s|h_s)p(h_s|v_{1:s-1}) \propto p(h_s|v_{1:s})$. Correction term updates the predictive distribution of h_s given $v_{1:s-1}$ to include the new data v_s.

Smoothing. Inferring the past. $p(h_t|v_{1:u}), t < u$. Beta-recursion: $\beta(h_s) = \mu_{\phi_{s+1} \to h_s}(h_s) = \sum_{h_s+1} p(h_{s+1}|h_s)p(v_{s+1}|h_{s+1})\beta(h_{s+1}) = p(v_{s+1:u}|h_s)$ and $\beta(h_u) = 1$.
Alpha-beta recursion: $p(h_t|v_{1:u}) \propto \alpha(h_t)\beta(h_t)$.

Prediction. Inferring the future, $p(h_t|v_{1:u})$ and $p(v_t|v_{1:u}), t > u$.
Most likely hidden path. Viterbi alignment. $\operatorname{argmax}_{h_{1:t}} p(h_{1:t}|v_{1:t})$.

Probabilistic model is a probability distribution pdf/pmf. Statistical model is a set of probabilistic models indexed by parameters $\{p(\boldsymbol{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$. Learning is picking one element. Bayesian model is a statistical model with a prior p.d. on the parameters $\boldsymbol{\theta}$: $p(\boldsymbol{x}, \boldsymbol{\theta})$.

Ising model/Boltzmann machine $\tilde{p}(\boldsymbol{x};\boldsymbol{\theta}) = \exp(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})$ where $\boldsymbol{x} \in \{0, 1\}^m$. Partition function is sum.

Parameter estimation: use data to pick one element $p(\boldsymbol{x};\hat{\boldsymbol{\theta}})$ from the set of prob. models. Bayesian inference: use data to determine the posterior (plausibility of $\boldsymbol{\theta}$): $p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \to p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
Predict next $\boldsymbol{x}$: $p(\boldsymbol{x}|\mathcal{D}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$. Samples from posterior = from prior that produces data equal to observed.

Likelihood $L(\boldsymbol{\theta}) = p(\mathcal{D};\boldsymbol{\theta})$ probability that sampling from the model with $\boldsymbol{\theta}$ generates $\mathcal{D}$. MLE: $\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$. Establishes ordering of param. values. Ignores information in the data.

MLE: parameter config. for which some specific moments under the model are equal to the empirical moments.
$\int \boldsymbol{m}(\boldsymbol{x};\hat{\boldsymbol{\theta}})p(\boldsymbol{x};\hat{\boldsymbol{\theta}})d\boldsymbol{x} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{m}(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}})$.
Moments $\boldsymbol{m}(\boldsymbol{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\boldsymbol{x};\boldsymbol{\theta})$.

$p(x|\theta) = \mathcal{N}(x;\theta,\sigma^2)$; $p(\theta;\alpha_0) = \mathcal{N}(x;\mu_0,\sigma_0^2)$. Posterior $p(\theta|\mathcal{D}) = \mathcal{N}(\theta;\mu_n,\sigma_n^2)$. $\mu_n = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\mu_0$. $\frac{1}{\sigma_n^2} = \frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}$.
Beta distribution $\mathcal{B}(f;\alpha,\beta) \propto f^{\alpha-1}(1-f)^{\beta-1}$, $f \in [0, 1]$.
$p(x|\theta) = \theta^x(1-\theta)^{1-x}$. $p(\theta;\boldsymbol{\alpha}_0) = \beta(\theta;\alpha_0,\beta_0)$.
$p(\theta|\mathcal{D}) = \mathcal{B}(\theta;\alpha_0 + n_{x=1}, \beta_0 + n_{x=0})$.

Factor analysis. $H < D$. $p(\mathbf{h}) = \mathcal{N}(\mathbf{h};\mathbf{0},\mathbf{I})$. $p(\mathbf{v}|\mathbf{h};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{v};\mathbf{F}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})$. $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_H)$ $D \times H$. Columns – factors with factor loadings. $\boldsymbol{\Psi}$ diagonal. $\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon}$. $\mathbf{F}\mathbf{h}$ spans a H-dim subspace of $\mathbb{R}^D$. Same dist. $\mathbf{v} = (\mathbf{F}\mathbf{R})\tilde{\mathbf{h}} + \mathbf{c} + \boldsymbol{\epsilon}$. $\mathbf{F}$ is not unique, factors have little meaning by themselves, rotational ambiguity. PPCA $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$.

$\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_x,\mathbf{C}_x)$, $\mathcal{N}(\mathbf{z};\boldsymbol{\mu}_z,\mathbf{C}_z)$, $\mathbf{x} \perp\!\!\!\perp \mathbf{z}$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ has density $\mathcal{N}(\mathbf{y};\mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{\mu}_z, \mathbf{A}\mathbf{C}_x\mathbf{A}^\top + \mathbf{C}_z)$.
Orthonormal matrix $\mathbf{R}^\top = \mathbf{R}^{-1}$ or $\mathbf{R}^\top\mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}$ rotate points.

Independent component analysis. Non-Gaussian indep. latents $p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^D p_h(h_i)$. $p(\mathbf{v}|\mathbf{h};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{v};\mathbf{A}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})$. $H > D$ overcomplete. $H = D$. $\mathbf{v} = \mathbf{A}\mathbf{h} = \sum_{i=1}^D(\mathbf{a}_i\alpha_i)\frac{1}{\alpha_i}h_i$. Col. ordering and scaling ambiguities. Latent unit variance fixes scaling. No rotational for non-Gaussian latents.

Sub-Gaussian pdf less peaked at zero than a Gaussian with same variance (uniform). Super-Gaussian (Laplace).
$p(\mathbf{v};\mathbf{A}) = p_{\mathbf{h}}(\mathbf{B}\mathbf{v})|\det\mathbf{B}| = |\det\mathbf{B}|\prod_{j=1}^D p_h(\mathbf{b}_j\mathbf{v})$.
$\ell(\mathbf{B}) = \sum_{i=1}^n\sum_{j=1}^D \log p_h(\mathbf{b}_j\mathbf{v}_i) + n\log|\det\mathbf{B}|$.

Unobserved vars: hidden, missing data. $p(\mathbf{v};\boldsymbol{\theta}) = \int_{\mathbf{u}} p(\mathbf{u},\mathbf{v};\boldsymbol{\theta})d\mathbf{u}$. Marginal inference. $\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$.
$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}}\log p(\mathbf{u},\mathcal{D};\boldsymbol{\theta})|\mathcal{D};\boldsymbol{\theta}]$.

Intractable partition. $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) \propto \frac{1}{n}\sum_{i=1}^n \boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\theta}) - \mathbb{E}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{m}(\boldsymbol{x};\boldsymbol{\theta})]$. Gradient ascent, computing expectation.

Combined. $\ell(\boldsymbol{\theta}) = \log\int\tilde{p}(\mathbf{u},\mathcal{D};\boldsymbol{\theta})d\mathbf{u} - \log\int\tilde{p}(\mathbf{u},\mathbf{v};\boldsymbol{\theta})d\mathbf{u}d\mathbf{v}$.
$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})}[\mathbf{m}(\mathbf{u},\mathcal{D};\boldsymbol{\theta})|\mathcal{D};\boldsymbol{\theta}] - \mathbb{E}_{p(\mathbf{u},\mathbf{v};\boldsymbol{\theta})}[\mathbf{m}(\mathbf{u},\mathbf{v};\boldsymbol{\theta});\boldsymbol{\theta}]$.

Score matching. iid from p_*. $p(\boldsymbol{\xi};\boldsymbol{\theta})$ model pdf, known up to $Z(\boldsymbol{\theta})$.
Estimate the model. MLE $\log p(\boldsymbol{\xi};\hat{\boldsymbol{\theta}}) \approx \log p_*(\boldsymbol{\xi})$.
Slopes match $\nabla_{\boldsymbol{\xi}}\log p(\boldsymbol{\xi};\hat{\boldsymbol{\theta}}) \approx \nabla_{\boldsymbol{\xi}}\log p_*(\boldsymbol{\xi})$.
Model score $\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}}\log p(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}}\log\tilde{p}(\boldsymbol{\xi};\boldsymbol{\theta})$.
Data score $\boldsymbol{\psi}_*(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\xi}}\log p_*(\boldsymbol{\xi})$, cannot compute.
Estimate $\boldsymbol{\theta}$ by minimising dist. $J_{\text{sm}}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_*||\mathbf{x}(\boldsymbol{\xi};\boldsymbol{\theta}) - \boldsymbol{\psi}_*(\mathbf{x})||^2 = \mathbb{E}_*\sum_{j=1}^d[\partial_j\psi_j(\mathbf{x};\boldsymbol{\theta}) + \frac{1}{2}\psi_j^2(\mathbf{x};\boldsymbol{\theta})] + \text{const.}$ $\psi_j(\boldsymbol{\xi};\theta) = \frac{\partial\log\tilde{p}(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\xi_j}$. $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. $J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j^2(\mathbf{x}_i;\boldsymbol{\theta})]$.
Required: $[p_*(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})]_{a_j}^{b_j} = 0$, smooth and existing $\partial_j\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})$.

Weak law of large numbers: $\Pr(|\bar{x}_n - \mathbb{E}[x]| \geq \epsilon) \leq \frac{\mathbb{V}[x]}{n\epsilon^2}$. Chebyshev's inequality: $\Pr(|s - \mathbb{E}[s]| \geq \epsilon) \leq \frac{\mathbb{V}[s]}{\epsilon^2}$.

Importance sampling. $\int\frac{g(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})d\mathbf{x} = \mathbb{E}_{q(\mathbf{x})}[\frac{g(\mathbf{x})}{q(\mathbf{x})}]$. Good: $q(\mathbf{x})$ large when $|g(\mathbf{x})|$ large. Importance weights $w_i = \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}$, $\mathbf{x}_i \sim q(\mathbf{x})$. $\int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{\sum_{i=1}^n g(\mathbf{x}_i)w_i}{\sum_{i=1}^n w_i}$.
Inverse transform sampling. CDF F_x. Calculate F_x^{-1}. Sample n iid $y_i \sim \mathcal{U}(0, 1)$. Transform $x_i = F_x^{-1}(y_i)$.
Rejection sampling. Sample $\mathbf{x}_i \sim q(\mathbf{x})$. Draw Bernoulli. $p(y_i,\mathbf{x}_i) = q(\mathbf{x})f(\mathbf{x})^y(1 - f(\mathbf{x}))^{1-y}$. Accept $\mathbf{x}_i$ with $y_i = 1$. $\mathbf{x_i} \sim \frac{q(\mathbf{x})f(\mathbf{x})}{\int q(\mathbf{x})f(\mathbf{x})d\mathbf{x}}$.

Jensen's inequality $\log\mathbb{E}[g(\mathbf{x})] \geq \mathbb{E}[\log g(\mathbf{x})]$.
$\operatorname{argmin}_q \mathrm{KL}(q||p)$ optimal q avoids where p is small, local fit, mode seeking. $\operatorname{argmin}_q \mathrm{KL}(p||q)$ optimal q is nonzero where p is nonzero. MLE, global fit, moment matching. $q(\mathbf{y})$ variational distr.
$\log p(\mathbf{x}) = \log\mathbb{E}_{q(\mathbf{y})}\left[\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y})}\right] \geq \mathbb{E}_{q(\mathbf{y})}\left[\log\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y})}\right] = \mathcal{F}(\mathbf{x}, q)$ free energy. $\log p(\mathbf{x}) = \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x})) + \mathcal{F}(\mathbf{x}, q)$.
$\mathrm{KL} \geq 0 \Rightarrow \log p(\mathbf{x}) \geq \mathcal{F}(\mathbf{x}, q)$. $q(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}) \Rightarrow \max\mathcal{F}(\mathbf{x}, q)$.
Inference is optimisation $\log p(\mathbf{x}) = \max_{q(\mathbf{y})}\mathcal{F}(\mathbf{x}, q)$ and $p(\mathbf{y}|\mathbf{x}) = \operatorname{argmax}_{q(\mathbf{y})}\mathcal{F}(\mathbf{x}, q) = \operatorname{argmin}_{q(\mathbf{y})}\mathrm{KL}(q||p)$.
$\ell(\boldsymbol{\theta}_k) = \mathrm{KL}(q(\mathbf{h})||p(\mathbf{h}|\mathcal{D})) + J_{\mathcal{F}}(q,\boldsymbol{\theta}_k)$. Opt. $q^*(\mathbf{h}) = p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)$.
MLE $\max_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}}\max_{q(\mathbf{h})}J_{\mathcal{F}}(q,\boldsymbol{\theta})$. Maximising $J_{\mathcal{F}}$ we look for $\mathbf{h}$ st. maximally variable (large entropy) and compatible with $\mathcal{D}$.
Expectation step. $J_{\mathcal{F}}(q^*,\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)}[\log p(\mathbf{h},\mathcal{D};\boldsymbol{\theta})] - \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)}[\log p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)]$ (does not depend on $\boldsymbol{\theta}$). Maximisation step. $\operatorname{argmax}_{\boldsymbol{\theta}} J_{\mathcal{F}}(q^*,\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}}\mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)}[\log p(\mathbf{h},\mathcal{D};\boldsymbol{\theta})]$.

$\int\mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(y|Ax,B^2)dx \propto \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2)$.
$\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2) \propto \mathcal{N}(x|m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(m_2 - m_1), \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$.

68–95–99.7.

$f(x)(\log f(x))' = f'(x)$.