T-61.6020 Machine Learning: Basic Principles: Chapter 10: Approximate inference – Part II

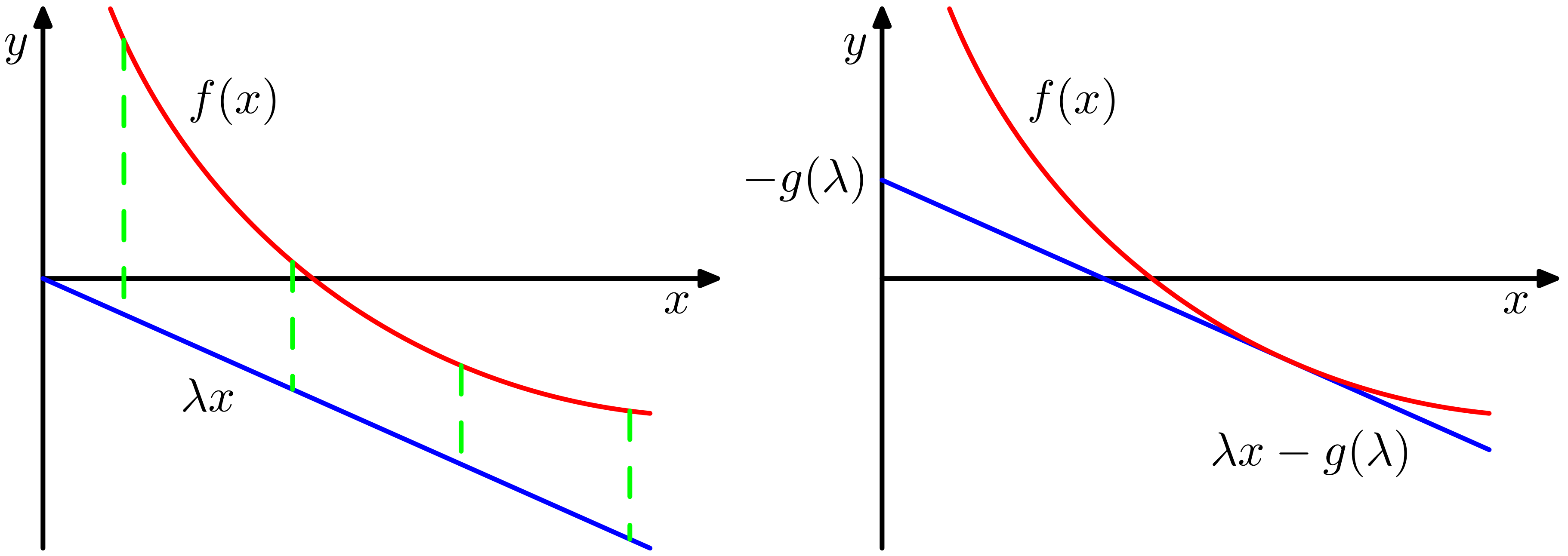
Ville Turunen

ville.t.turunen@tkk.fi

Local variational methods

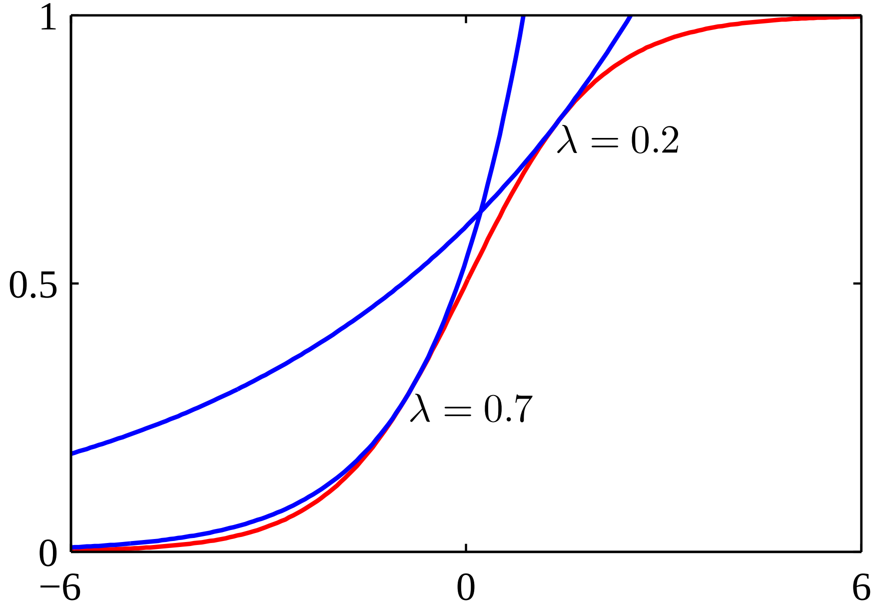
- consider only individual variables within a model
- finding bounds to simplify the distribution
- can be applied to multiple variables in turn

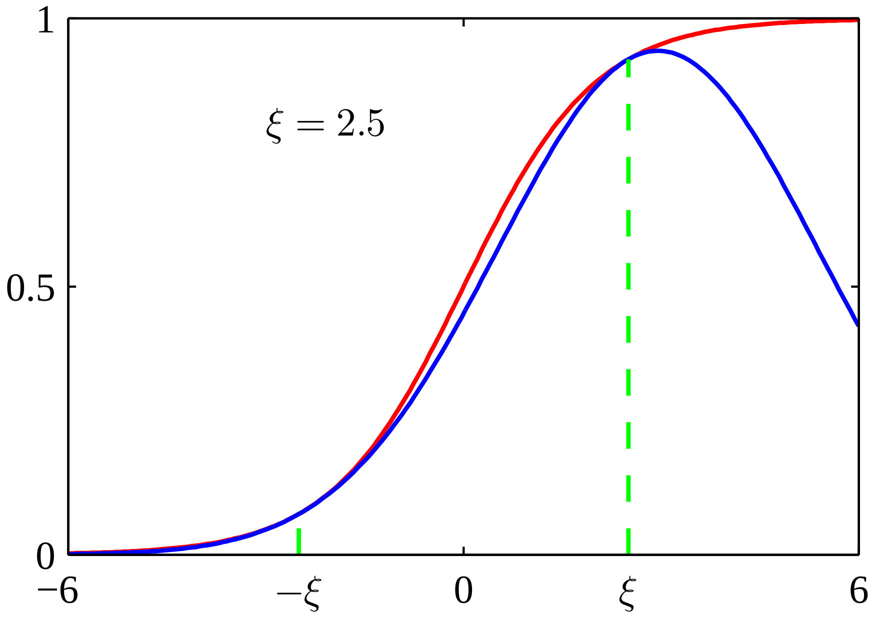
Linear bounds for convex functions



- minimize the discrepancy $f(x) - \lambda x$ with respect to λ
- *convex duality*
  - $g(\lambda) = \max_x \{\lambda x - f(x)\}$
  - $f(x) = \max_\lambda \{\lambda x - g(\lambda)\}$

Example: logistic sigmoid





- $\ln \sigma(x)$ is concave $\Rightarrow \sigma(x) \leq \exp(\lambda x - g(\lambda))$
- lower bound $f(a, \xi)$ is Gaussian
- $\int \sigma(a) p(a) da \geq \int f(a, \xi) p(a) da = F(\xi)$
- $F(\xi)$ maximized with respect to the variational parameter ξ

Variational logistic regression (1/2)

- variational approach to Bayesian logistic regression (Section 4.5)
- $p(C_1|\phi, \mathbf{t}) = \int p(C_1|\phi, \mathbf{w}) p(\mathbf{w}|\mathbf{t}) d\mathbf{w}$
- replace the posterior distribution $p(\mathbf{w}|\mathbf{t})$ with Gaussian approximation $q(\mathbf{w})$
- like the Laplace method, but more accurate
- based on finding the maximal lower bound for the marginal likelihood $p(\mathbf{t})$

Variational logistic regression (2/2)

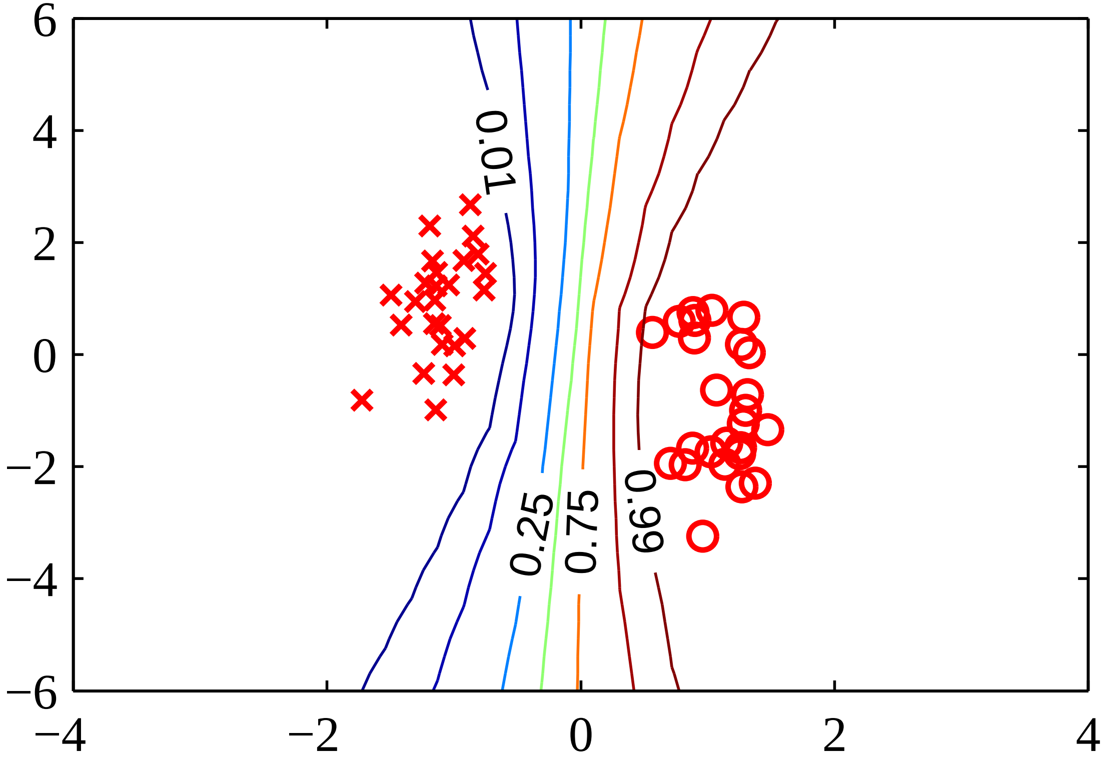
- find a local bound for each $p(t|\mathbf{w})$ separately
- a variational parameter ξ_n for each observarion (ϕ_n, t_n)
- combined: lower bound $h(\mathbf{w}, \xi) \leq p(\mathbf{t}|\mathbf{w})$

- result: Gaussian variational posterior $q(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$
  - $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m_0} + \sum_{n=1}^N (t_n - 1/2)\phi_n)$
  - $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2\sum_{n=1}^N \lambda(\xi_n)\phi_n\phi_n^T$

Optimizing the variational parameters

- parameters ξ_n determined by maximizing the lower bound on the marginal likelihood
- $\ln p(\mathbf{t}) \geq \ln \int h(\mathbf{w}, \xi) p(\mathbf{w}) d\mathbf{w} = L(\xi)$
- EM-algorithm
  - E-step: evaluate $q(w)$ using ξ^{old}
  - M-step: assume $q(w)$, maximize expectation to find ξ^{new}

Example of variational logistic regression



- predictive distribution

Expectation propagation (1/2)

- again, minimizes KL-divergence
- but now in reverse form: $\mathrm{KL}(p\|q)$
- optimum solution corresponds to *moment matching*

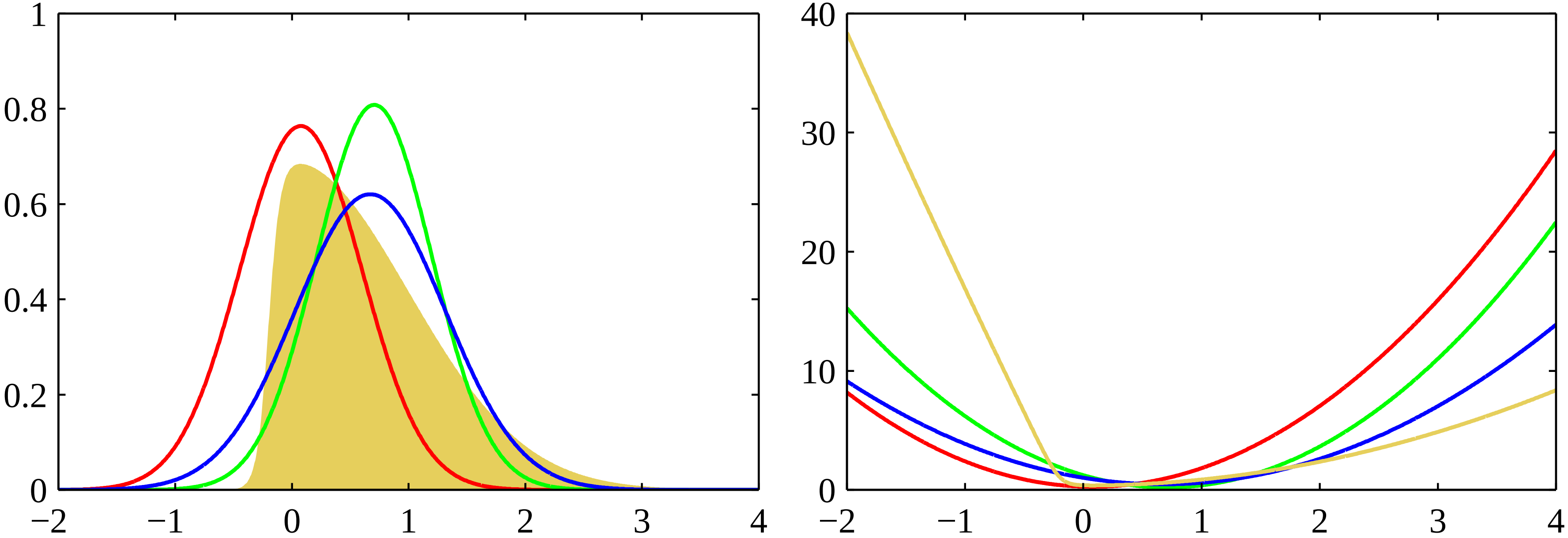
- consider joint distribution $p(D, \theta) = \prod_i f_i(\theta)$
- e.g. i.i.d. data
- the posterior $p(\theta|D) = \frac{1}{p(D)} \prod_i f_i(\theta)$
- approximate by $q(\theta) = \frac{1}{Z} \prod_i \tilde{f}_i(\theta)$
- constraint: $\tilde{f}_i(\theta)$ come from the exponential family

Expectation propagation (2/2)

- initialize $\tilde{f}_i(\theta)$ and $q(\theta) = \frac{1}{Z} \prod_i \tilde{f}_i(\theta)$
- update each $\tilde{f}_j(\theta)$ one at a time until convergence:

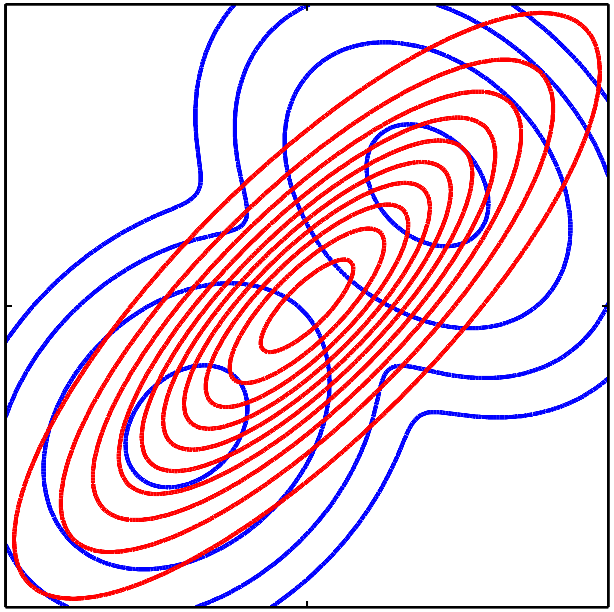
1. remove $\tilde{f}_j(\theta)$ from the posterior $q^{\setminus j}(\theta) = \prod_{i \neq j} \tilde{f}_i(\theta) = \frac{q(\theta)}{\tilde{f}_j(\theta)}$
2. evaluate new posterior $q^{new}(\theta)$ so that
   - KL$(\, f_j(\theta) q^{\setminus j}(\theta) / Z_j \; \| \; q^{new}(\theta) \,)$ is minimized
   - achieved by moment matching
3. evaluate the new factor
   - $\tilde{f}_j(\theta) = Z_j \frac{q^{new}(\theta)}{q^{\setminus j}(\theta)}$

EP: Illustration



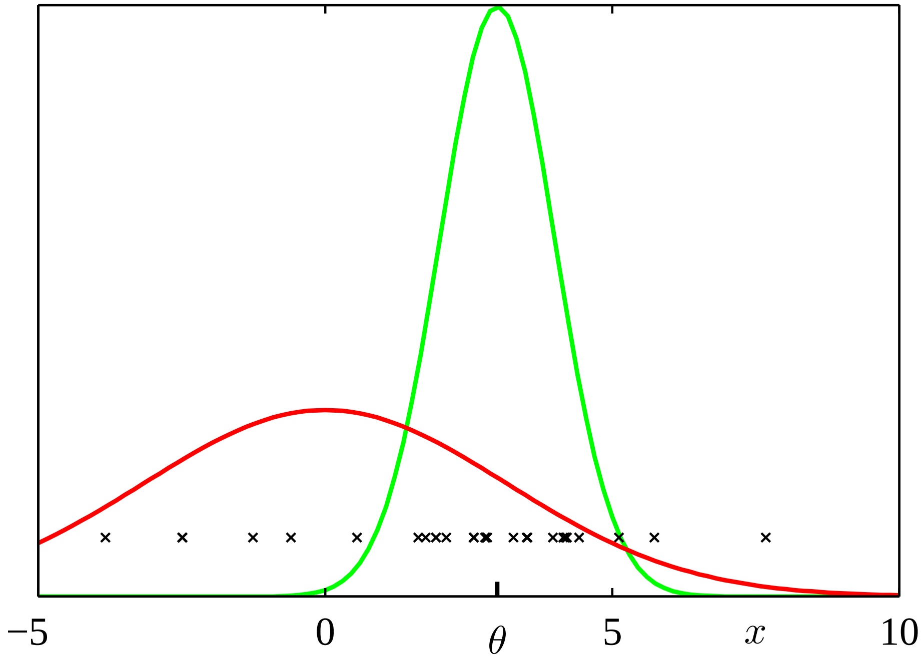
- original, laplace, global variational, EP

EP: Multimodal distribution



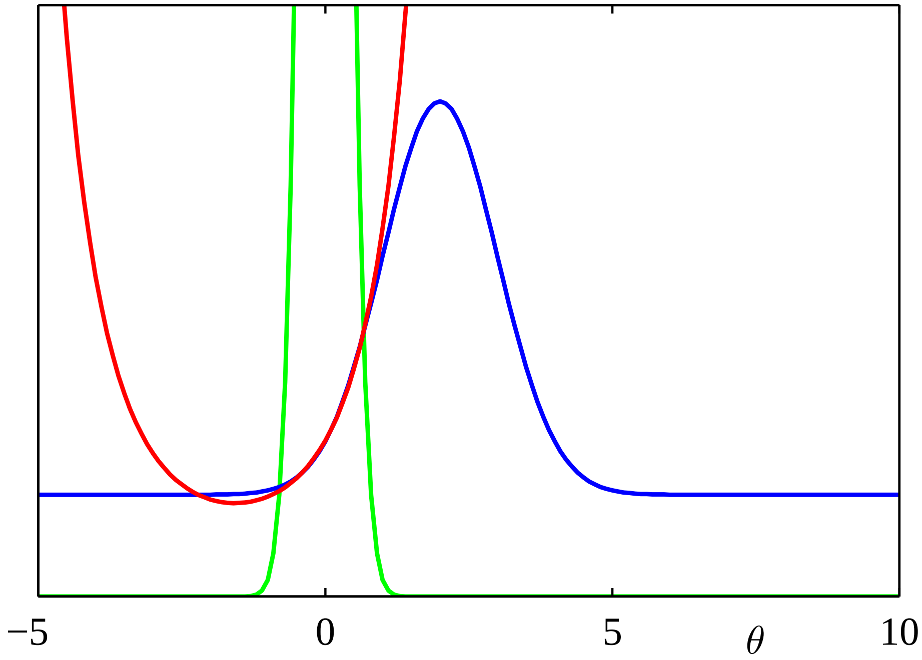
- minimizing $\mathrm{KL}(p\|q)$ tries to capture all of the modes

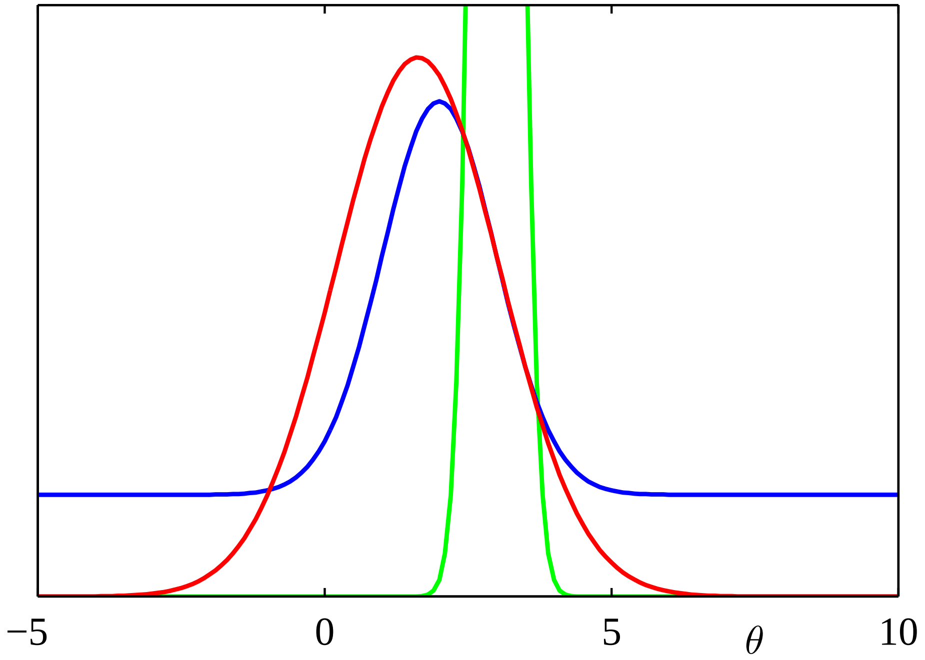
Example: The clutter problem



- $p(\mathbf{x}|\theta) = (1-w)N(\mathbf{x}|\theta, \mathbf{I}) + wN(x|\mathbf{0}, a\mathbf{I})$
- $p(D, \theta) = p(\theta) \prod_{n=1}^{N} p(\mathbf{x}_n|\theta)$
- $q(\theta) = N(\theta|\mathbf{m}, v\mathbf{I})$
- $\tilde{f}_n(\theta) = s_n N(\theta|\mathbf{m}_n, v_n\mathbf{I})$ (v_n can be negative)

Examples of approximations of factors





- $f_n(\theta)$, $\tilde{f}_n(\theta)$, $q^{\backslash n}(\theta)$

Performance