The Laplace Approximation and Variational Inference

José Miguel Hernández-Lobato

Department of Engineering, Cambridge University

April 10, 2013

The Laplace Approximation: Introduction

The **simplest deterministic method** for approximate inference. Restricted to GMs in which the variables of interest are continuous.

The factors for the continuous random variables will generally be some **continuous parametric functions**.

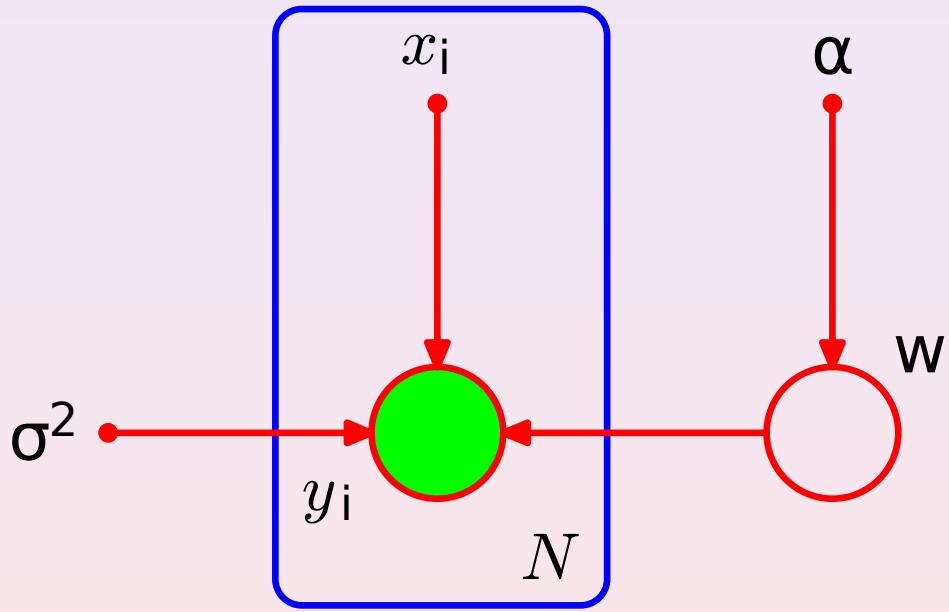
Probit regression model:

$$y_i = \begin{cases} 1 & \text{if} \quad \mathbf{w}^{\mathsf{T}}\mathbf{x}_i + \epsilon_i \geq 0 \\ -1 & \text{if} \quad \mathbf{w}^{\mathsf{T}}\mathbf{x}_i + \epsilon_i < 0 \end{cases}$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\alpha)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

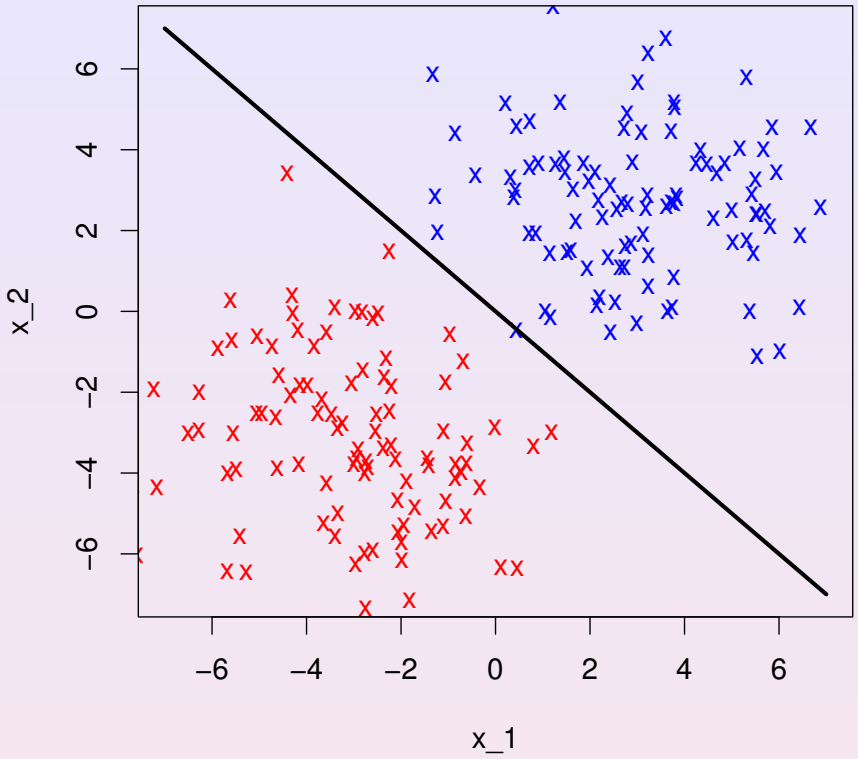
$$p(y_i|\mathbf{w}_i, \mathbf{x}_i) = \Phi(y_i \mathbf{w}^{\mathsf{T}}\mathbf{x}_i|0, \sigma^2)\,.$$



$\mathbf{w}$ are the variables of interest, and the evidence are the $\{y_i\}_{i=1}^N$. Furthermore,

$$p(\mathbf{y}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{w}, \mathbf{x}_i)p(\mathbf{w}) = \prod_{i=1}^N \Phi(y_i \mathbf{w}^{\mathsf{T}}\mathbf{x}_i|0, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}\alpha)\,.$$

The Laplace Approximation: Probit Regression Model



Sample data from the corresponding GM.

We want to make inference on $\mathbf{w}$ given some observed labels $\mathbf{y}$. For this, we can use Bayes theorem:

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}\,.$$

where $p(\mathbf{y})$ is a normalization constant which can be used for model selection.

We also want to compute a predictive distribution for new unlabeled instances:

$$p(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}) = \int p(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w})p(\mathbf{w}|\mathbf{y})d\mathbf{w}\,.$$

Unfortunately the required computations are intractable.

The Laplace Approximation: Univariate Case I

The Laplace approximation will find a Gaussian approximation to the conditional distribution of a set of continuous variables:

Consider first a single scalar variable z:

$$p(z) = \frac{1}{Z} f(z),$$

where $f(z) = p(z, \mathbf{e})$, $\mathbf{e}$ are observed variables and $Z = \int f(z) dz$.

How do we set the parameters of $\mathcal{Q}(z)$, the Gaussian approximation, so that it is similar to $p(z)$ given that we **do not know** Z?

The first step is to find a mode (*i.e.*, a local maximum z_0) of $p(z)$

$$\left. \frac{df(z)}{dz} \right|_{z=z_0} = 0$$

Any optimization algorithm can be used for this purpose.

The Laplace Approximation: Univariate Case II

The logarithm of a Gaussian is a quadratic function of the variables.

We consider a truncated Taylor expansion of $\log f(z)$ center at the mode:

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} A (z - z_0)^2 , \qquad A = - \frac{d^2}{dz^2} \log f(z) \Big|_{z=z_0}$$

Taking the exponential we obtain:

$$f(z) \approx f(z_0) \exp \left\{ - \frac{A}{2} (z - z_0)^2 \right\} \qquad \mathcal{Q}(z) = \mathcal{N}(z | z_0, A^{-1})$$

The normalization constant Z can be approximated by $f(z_0)\sqrt{2\pi/A}$.

The **mean** of $\mathcal{Q}$ is z_0 and the **variance** is A^{-1}.

The Gaussian approximation will only be defined if $A > 0$, *i.e.*, z_0 must be a local maximum of $\log f$ and hence f with negative second derivative.

The Laplace Approximation: Multi-variate Case

The same principle can be applied to approximate a M-dimensional distribution $p(\mathbf{z}) = f(\mathbf{z})/Z$ defined over a vector of real values.

$$\log f(\mathbf{z}_0) \approx \log f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathsf{T}} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)\,, \qquad \mathbf{A} = - \bigtriangledown \bigtriangledown \log f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$$

Taking the exponential we have:

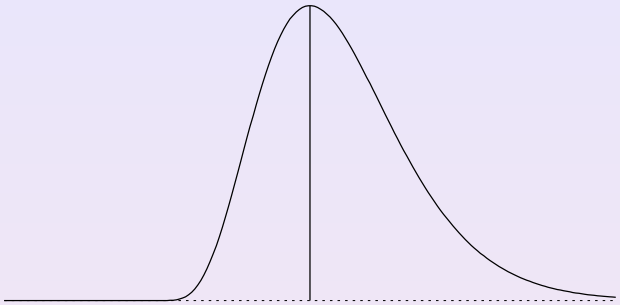
$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathsf{T}} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\}\,, \qquad \mathcal{Q}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

The normalization constant is approximated by $f(z_0)\sqrt{(2\pi)^M/|\mathbf{A}|}$.

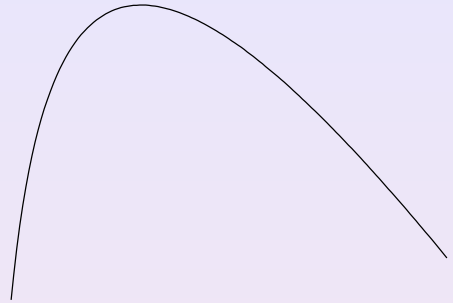
The **mean** of $\mathcal{Q}$ is $\mathbf{z}_0$ and the **covariance matrix** is $\mathbf{A}^{-1}$.

The Gaussian approximation will only be defined if $\mathbf{A}$ is positive semidefinite, *i.e.*, z_0 must be a local maximum not a minimum or a saddle point.

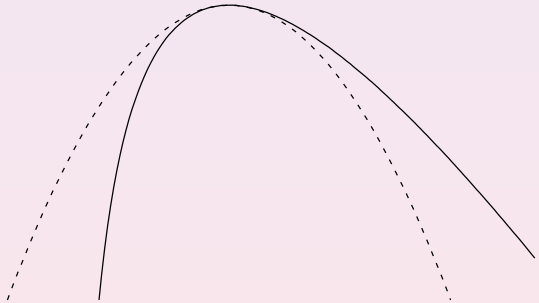
The Laplace Approximation: Example



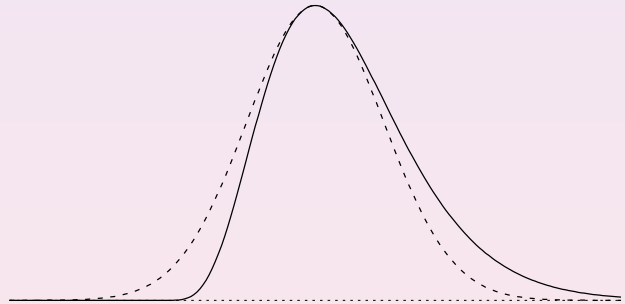
f(z)



log f(z)



log f(z) & log ~Q(z)



f(z) & ~Q(z)

The Laplace Approximation: Probit Regression I

For simplicity we consider that $\sigma^2 = 1$ and that $\alpha = 1$.

The posterior distribution is:

$$p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) = f(\mathbf{w}), \qquad \log f(\mathbf{w}) = \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{w}).$$

Furthermore, we have that:

$$\log f(\mathbf{w}) = \sum_{i=1}^{N} \log \Phi(y_i \mathbf{w}^{\mathsf{T}} \mathbf{x}_i) - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}.$$

Let $\mathbf{w}_0$ a maximum of f. Computing the negative Hessian at $\mathbf{w}_0$:

$$\mathbf{A} = -\nabla\nabla \log f(\mathbf{w}_0) = \sum_{i=1}^{N} \left[v_i(s_i + v_i)\mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \right] + \mathbf{I}, \quad v_i = \frac{\mathcal{N}(s_i|0,1)}{\Phi(s_i)}, \quad s_i = y_i \mathbf{w}_0^{\mathsf{T}} \mathbf{x}_i.$$

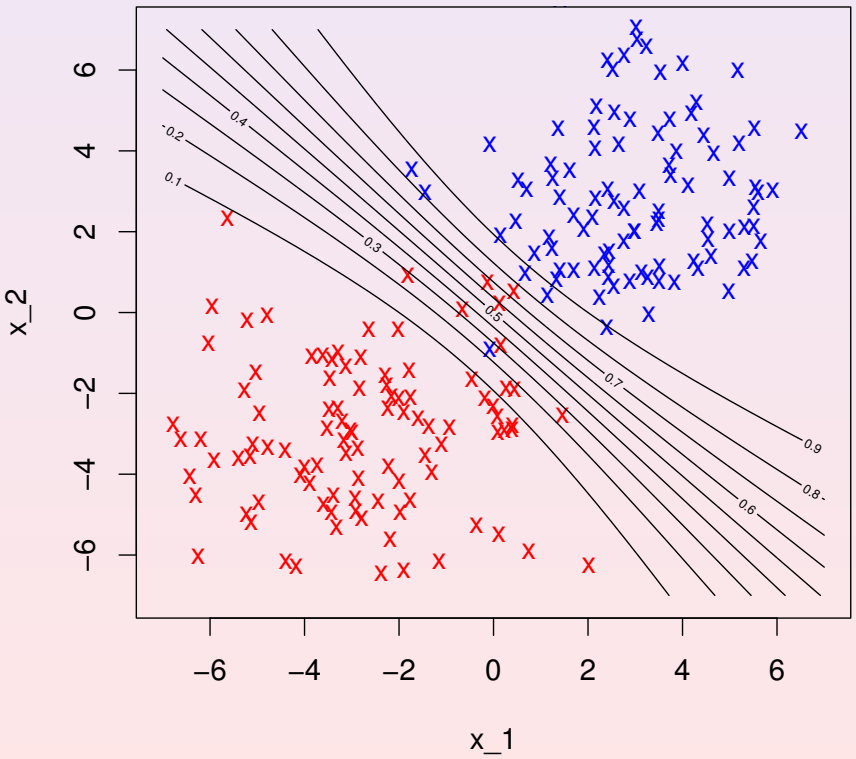
The approximate posterior is $\mathcal{Q}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{A}^{-1})$. The normalization constant is $Z \approx f(\mathbf{w}_0)\sqrt{\frac{(2\pi)^M}{|\mathbf{A}|}}$, where M is the dimensionality of $\mathbf{w}$.

The Laplace Approximation: Probit Regression II

It is also possible to compute an approximate predictive distribution:

$$p(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}) \approx \int p(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w})\mathcal{Q}(\mathbf{w})d\mathbf{w} = \int \Phi(y_{\text{new}}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}})\mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{A}^{-1})d\mathbf{w}\,,$$

$$= \Phi\left(\frac{y_{\text{new}}\mathbf{w}_0^{\mathsf{T}}\mathbf{x}_{\text{new}}}{\sqrt{\mathbf{x}_{\text{new}}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{x}_{\text{new}} + 1}}\right).$$



Uncertainty is high near the decision boundary and progressively decreases as we move away from it.

Uncertainty is significantly larger in regions where there is no data.

MAP solutions do not consider this uncertainty.

The Laplace Approximation: Considerations

- The mode of $\log f$ can be found using a numerical optimization method. The Hessian can be approximated by differences.
- Many distributions can be multi-modal, what leads to many different Laplace approximations, depending on the mode.
- In many cases, the posterior distribution of $\mathbf{z}$ will converge to a Gaussian as the number of observations (evidence) increases.
- Only applicable on real variables.
- Only focuses around the mode and can fail to capture global properties.
- No need to know Z. Furthermore, it provides an estimate of Z.
- Depends on the basis. If z is mapped to $u(z)$, the density is transformed to $p(u) = p(z)|dz/du|$ and the approximation will be different.

Variational Inference: Introduction

Based on the calculus of variations, *i.e.*, a generalization of standard calculus. Deals with functionals, functions and derivatives of functionals rather than functions, variables and derivatives. Similar rules apply.

Can be applied to models of either continuous or discrete random variables.

Approximates both the posterior distribution and its normalization constant: $p(\mathbf{z}|\mathbf{e})$ and $p(\mathbf{e})$.

It is based on the following **decomposition**:

$$\log p(\mathbf{e}) = \mathcal{L}(\mathcal{Q}) + \mathrm{KL}(\mathcal{Q}||p)$$

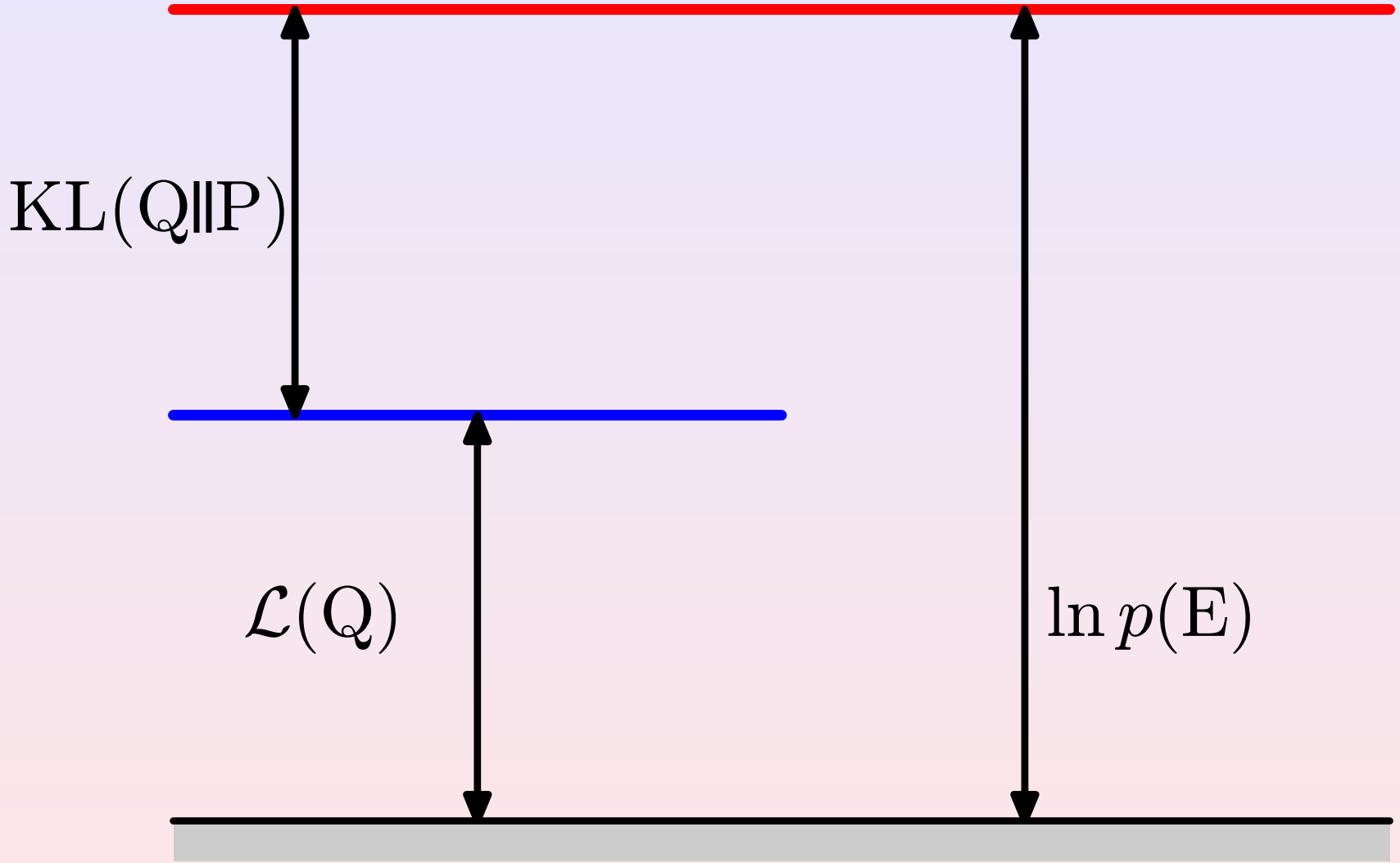
where

$$\mathcal{L}(\mathcal{Q}) = \sum_{\mathbf{z}} \mathcal{Q}(\mathbf{z}) \log \left\{ \frac{p(\mathbf{z}, \mathbf{e})}{\mathcal{Q}(\mathbf{z})} \right\}, \quad \mathrm{KL}(\mathcal{Q}||p) = -\sum_{\mathbf{z}} \mathcal{Q}(\mathbf{z}) \log \left\{ \frac{p(\mathbf{z}|\mathbf{e})}{\mathcal{Q}(\mathbf{z})} \right\} \geq 0$$

The Kullback Leibler divergence measures the **fit** of $\mathcal{Q}$ to $p(\mathbf{z}|\mathbf{e})$.

$\mathcal{L}(\mathcal{Q})$ approximates $\log p(\mathbf{e})$.

Decomposition of the Marginal Likelihood



Variational Inference: Choosing the approximation $\mathcal{Q}$

$\mathcal{Q}$ minimizes $\mathrm{KL}(\mathcal{Q}||p)$ (maximizes $\mathcal{L}(\mathcal{Q})$) when $\mathcal{Q} = p(\mathbf{z}|\mathbf{e})$.

In practice, one selects $\mathcal{Q}$ to be a parametric distribution $\mathcal{Q}(\mathbf{z}|\boldsymbol{\theta})$, for which $\mathcal{L}(\mathcal{Q})$ can be computed analytically, *e.g.*, a Gaussian.

The lower bound then becomes a function of $\boldsymbol{\theta}$ and can be optimized using non-linear optimization techniques such as gradient descent.

An alternative is to assume that $\mathcal{Q}$ factorizes with respect to a partition of $\mathbf{z}$ into M disjoint groups $\mathbf{z}_i$, with $i = 1, \ldots, M$:

$$\mathcal{Q}(\mathbf{z}) = \prod_{i=1}^{M} \mathcal{Q}_i(\mathbf{z}_i)$$

and **no further assumptions are made** about $\mathcal{Q}$.

This approach is known in the literature as variational mean field.

Variational Inference: Variational Mean-Field

Substituting $\mathcal{Q}$ in $\mathcal{L}(\cdot)$ and looking for the dependence with respect to $\mathcal{Q}_j$:

$$\begin{aligned}\mathcal{L}(\mathcal{Q}) &= \sum_{\mathbf{z}} \prod_{i=1}^{M} \mathcal{Q}_i(\mathbf{z}_i) \left\{ \log p(\mathbf{z}, \mathbf{e}) - \sum_{i=1}^{M} \log \mathcal{Q}_i(\mathbf{z}_i) \right\} \\ &= \sum_{\mathbf{z}_j} \left[\mathcal{Q}_j(\mathbf{z}_j) \left\{ \sum_{\mathbf{z}_{i \neq j}} \log p(\mathbf{z}, \mathbf{e}) \prod_{i \neq j}^{M} \mathcal{Q}_i(\mathbf{z}_i) \right\} - \mathcal{Q}_j(\mathbf{z}_j) \log \mathcal{Q}_j(\mathbf{z}_j) \right] + \text{const} \\ &= \sum_{\mathbf{z}_j} \left[\mathcal{Q}_j(\mathbf{z}_j) \log \hat{p}(\mathbf{z}_j, \mathbf{e}) - \mathcal{Q}_j(\mathbf{z}_j) \log \mathcal{Q}_j(\mathbf{z}_j) \right] + \text{const}\end{aligned}$$

which is a negative KL divergence and we have defined

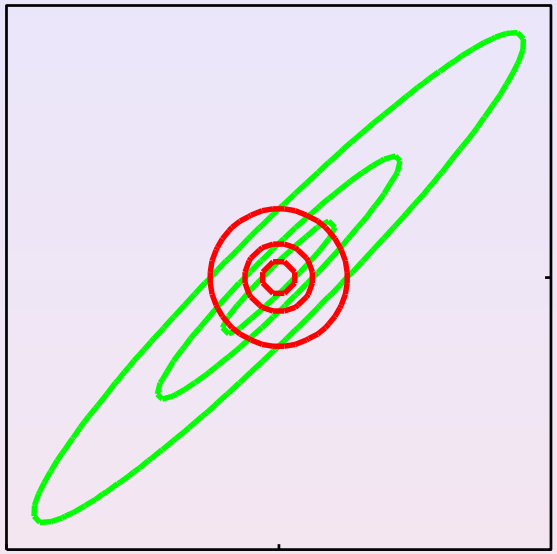
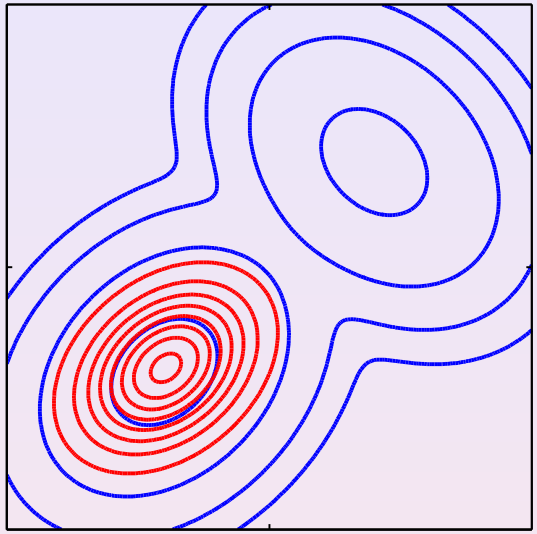
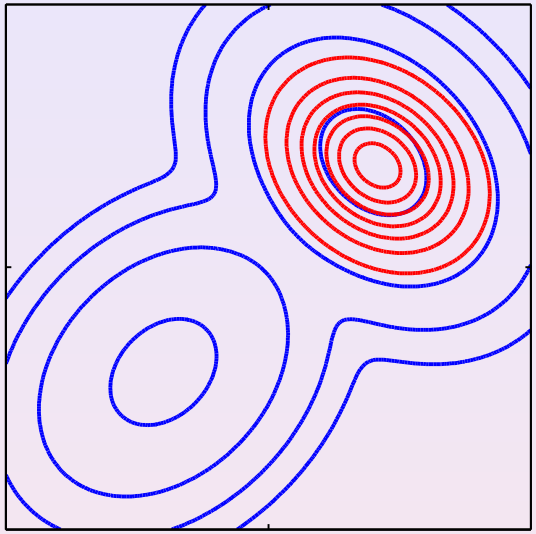
$\log \hat{p}(\mathbf{z}_j, \mathbf{e}) = \mathbb{E}_{i \neq j} \left[\log p(\mathbf{z}, \mathbf{e}) \right] + \text{const}$.

The optimal $\mathcal{Q}_j$ given that the other factors are kept fixed is:

$$\log \mathcal{Q}_j^{\star}(\mathbf{z}_j) = \mathbb{E}_{i \neq j} \left[\log p(\mathbf{z}, \mathbf{e}) \right] + \text{const} \tag{1}$$

To optimize $\mathcal{Q}$ we iterate, optimizing each $\mathcal{Q}_j$ using (1) .

Properties of Variational Approximations

The KL divergence $\mathrm{KL}(\mathcal{Q}||p)$ favors solutions that take high probability where p takes high probability, but can ignore important regions.

The optimization problem is not convex and can have multiple local optima.

Variational Inference Example: 2D Ising Model I

Ising models (ferromagnetic or anti-ferromagnetic) are arrays of spins, *e.g.*, atoms that can take states ± 1, that are magnetically coupled to each other.

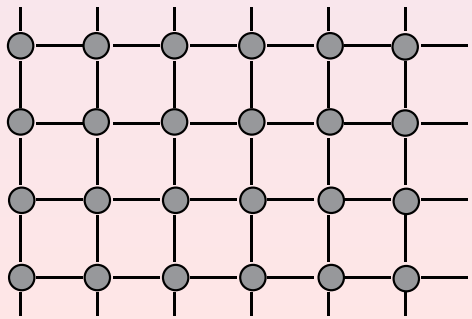
At temperature T, the probability of a spin configuration $\mathbf{z} \in \{-1, 1\}^N$ is

$$p(\mathbf{z}|\beta, J, H) = \frac{1}{Z(\beta, J, H)} \exp\left\{-\beta E(\mathbf{z}; J, H)\right\} ,$$

where $\beta = 1/(k_B T)$, k_B is Boltzmann's constant and Z is a normalizer,

$$E(\mathbf{z}; J, H) = -\left[\frac{1}{2}\sum_{m,n} J_{m,n} z_m z_n + \sum_n H z_n\right]$$

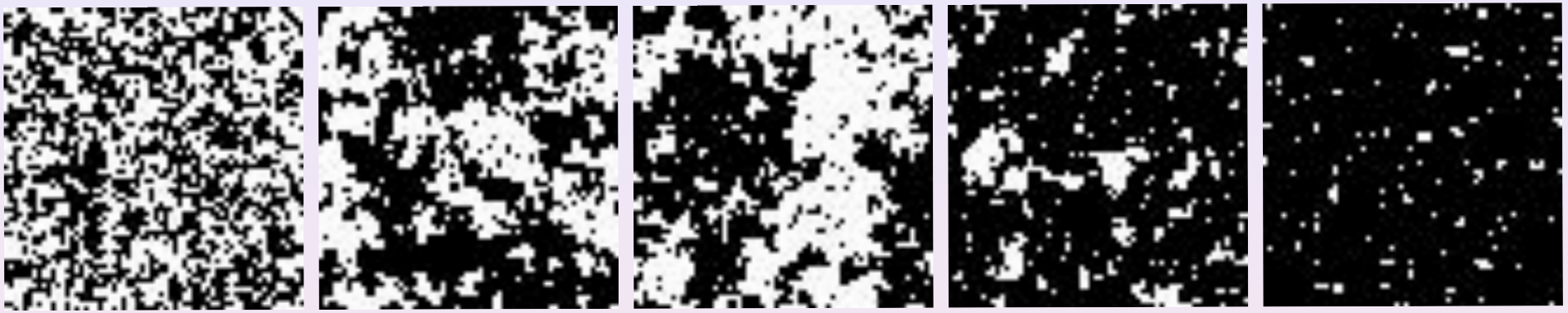
H is the applied field, $J_{m,n} = J$ is m and n are neighbors and 0 otherwise.



Can be described by an un-directed GM.

The evaluation of $p(\mathbf{z}|\beta, J, H)$ is intractable since computing Z requires summing out $\mathbf{z}$.

Variational Inference Example: 2D Ising Model II



As the temperature increases the probability of any state becomes uniform.

For low temperatures it is likely to find all the spins in the same position.

2D Ising Model: Mean Field Approximation I

We choose a factorizing approximation:

$$\mathcal{Q}(\mathbf{z}) = \prod_{i=1}^{N} \mathcal{Q}_i(z_i)\,, \qquad \log \mathcal{Q}_j(z_j) \propto \mathbb{E}_{i\neq j}[\log \tilde{p}(\mathbf{z}|\beta, J, H)]\,.$$

which gives a closed form solution for $\mathcal{Q}_j$:

$$\mathcal{Q}_j(z_j = 1) = \frac{\exp(a_j)}{exp(a_j) + exp(-a_j)} = \frac{1}{1 + exp(-2a_j)}\,, \quad a_j = \beta \left(J \sum_{n\in \mathrm{Nb}_j} \overline{z}_n + H \right)\,.$$

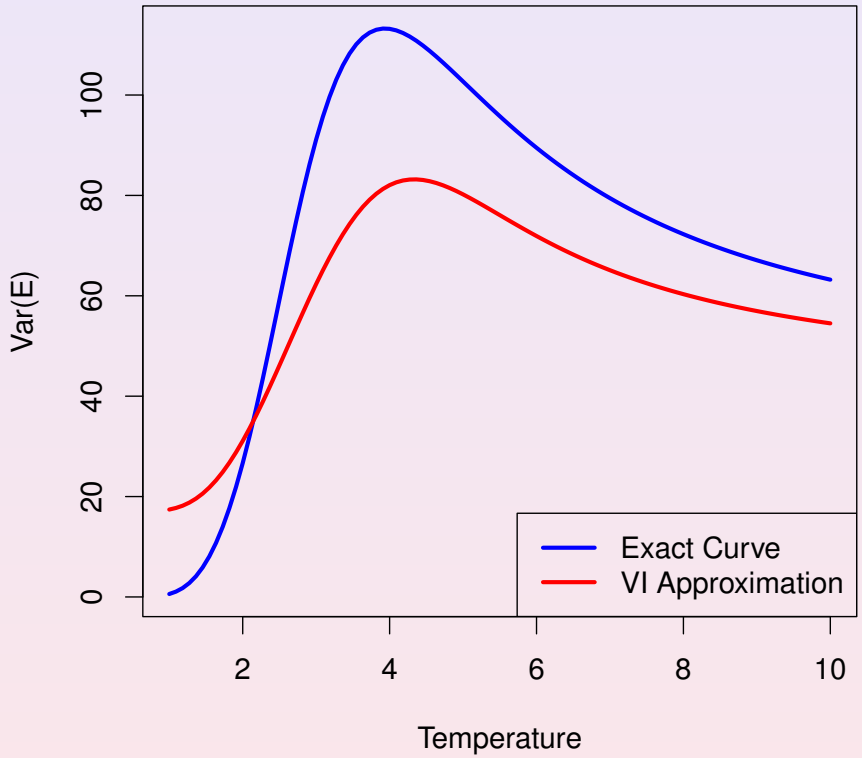
where $\overline{z}_n = 1 \cdot \mathcal{Q}_n(z_n = 1) - 1 \cdot \mathcal{Q}_n(z_n = -1) = \tanh(a_n)$.

This process has to be iterated for $j = 1, \ldots, N$ until convergence of $\mathcal{Q}$.

Given $\mathcal{Q}$ it is easy to evaluate the lower bound on $Z(\beta, J, H)$, which can be used to approximate the value of this constant.

2D Ising Model: Mean Field Approximation II

The VI approximation can be used to analyze interesting properties of the ising model related to phase transitions.



Phase transition: transformation of a system **from one state of matter to another**.

During a phase transition **certain properties of the system change** often **discontinuously**.

The variance of the energy is related to the **heat capacity** of the system.

Variance of the energy as a function of the temperature.
Exact and as computed by VI.

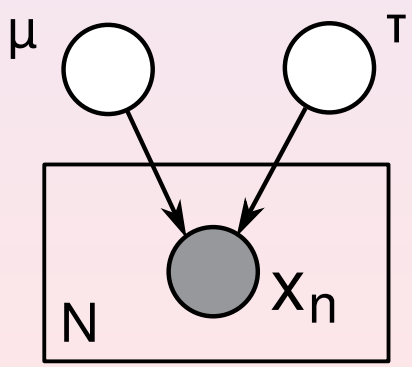
Example: Unknown Mean and Variance of a Gaussian

Goal: infer the posterior distribution of the mean μ and precision τ of a Gaussian distribution given a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ of independent samples.

The log likelihood of μ and τ is:

$$\log p(\mathcal{D}|\mu, \tau) = -\frac{N}{2} \log 2\pi\tau^{-1} - \frac{\tau}{2} \sum_{n=1}^{N} (x_n - \mu)^2$$

$$= \frac{N}{2} \log \tau - \frac{\tau}{2} \left[N(\mu - \overline{x})^2 + S \right] + \text{const}\,,$$



where $S = \sum_n (x_n - \overline{x})^2$ and $\overline{x}$ is the empirical mean.

The priors for μ and τ are uniform and conjugate:

$$p(\mu) = 1/\sigma_\mu\,, \qquad p(\tau) = 1/\tau$$

Mean Field: Unknown Mean and Variance of a Gaussian I

We enforce that the posterior approximation factorizes $\mathcal{Q}(\mu, \tau) = \mathcal{Q}_\mu(\mu)\mathcal{Q}_\tau(\tau)$ and solve for the optimal factors.

$$\log \mathcal{Q}_\mu(\mu) = \mathbb{E}_{\mathcal{Q}_\tau}\left[\log p(\mathcal{D}, \mu, \tau)\right] , \qquad \log \mathcal{Q}_\tau(\tau) = \mathbb{E}_{\mathcal{Q}_\mu}\left[\log p(\mathcal{D}, \mu, \tau)\right] ,$$

This gives the following optimal factors given that the other factor is fixed:

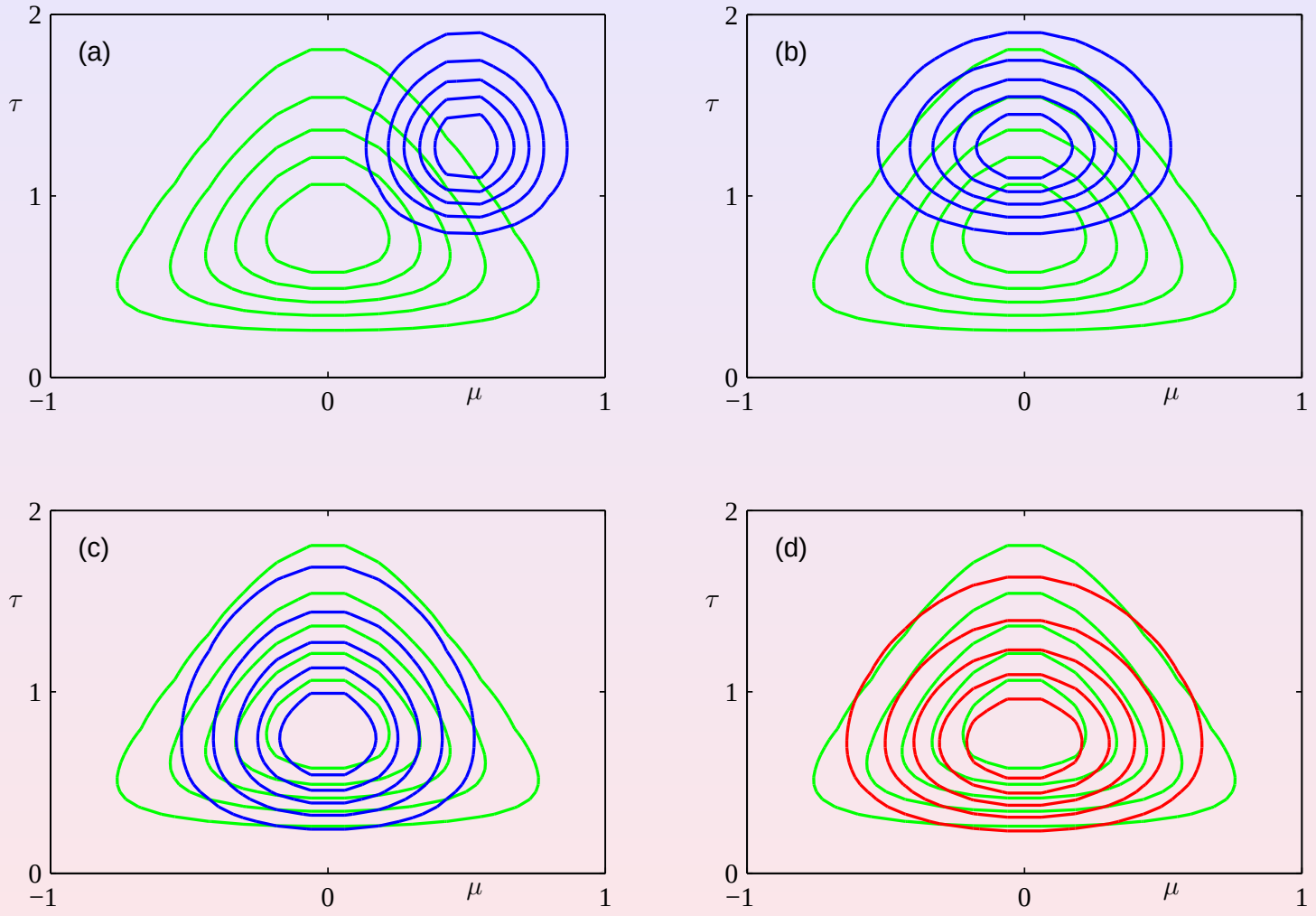
$$\mathcal{Q}_\mu(\mu) = \mathcal{N}(\mu|\overline{x}, \lambda^{-1}) , \qquad \mathcal{Q}_\tau(\tau) = \text{Gamma}(\tau|a, b)$$

$$\mathcal{Q}_\tau(\tau) = b^a \frac{1}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\} ,$$

where $\lambda = N\mathbb{E}_{\mathcal{Q}_\tau}[\tau] = Na/b$, $a = N/2$ and $b = N/2(\lambda^{-1} + S)$.

We iteratively optimize $\mathcal{Q}_\mu$ and $\mathcal{Q}_\tau$ until convergence.

Mean Field: Unknown Mean and Variance of a Gaussian II



Variational Inference: Local Methods I

Bounds over specific factors can be useful to deal with intractable expectations.

Consider the logistic function:

$$\sigma(z) = \frac{1}{1+e^{-z}}\,, \qquad \log\sigma(z) = \log\left(1+e^{-z}\right) = \frac{z}{2} - \log\left(e^{\frac{z}{2}} + e^{-\frac{z}{2}}\right)\,.$$

The function $f(z) = -\log\left(e^{\frac{z}{2}} + e^{-\frac{z}{2}}\right)$ is convex in terms of z^2. We can hence approximate this function by a tangent line at ξ which is a global lower bound:

$$f(z) \geq f(\xi) + \left.\frac{df(z)}{d(z^2)}\right|_{z=\xi}(z^2-\xi^2) = -\frac{\xi}{2} + \log\sigma(\xi) + \lambda(\xi)(z^2-\xi^2)\,.$$

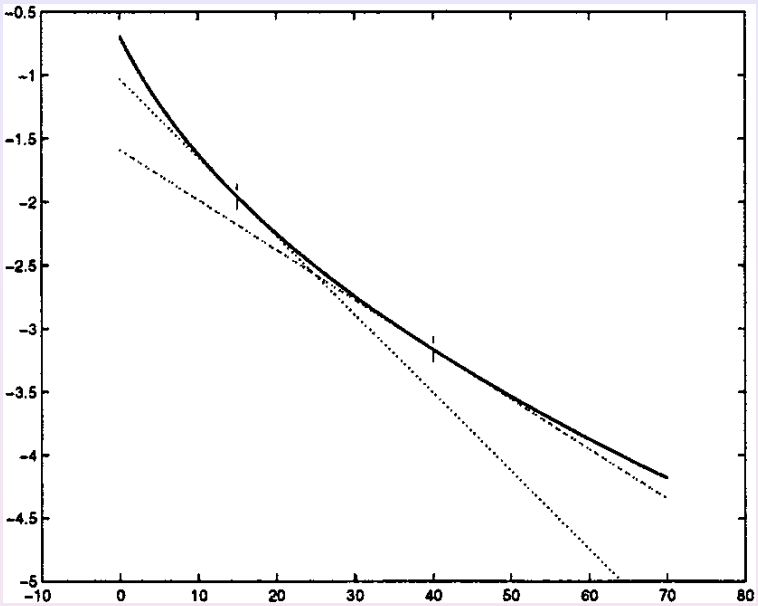
which is tight for $z^2 = \xi^2$ and gives

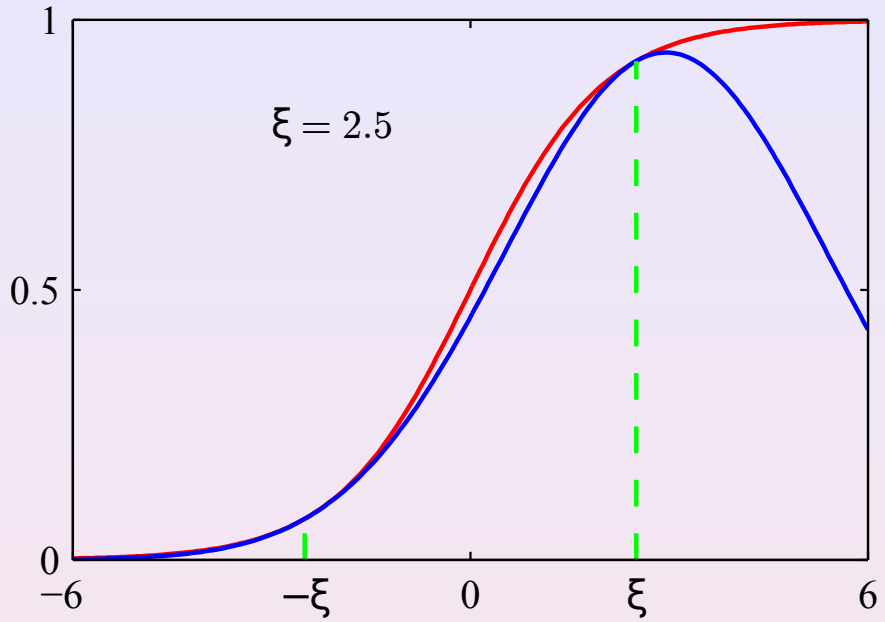
$$\sigma(z) \geq \sigma(\xi)\exp\left\{\frac{z-\xi}{2} - \lambda(\xi)(z^2-\xi^2)\right\} = \underline{\sigma}(z,\xi)\,, \quad \lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]\,.$$

The lower bound is Gaussian with respect to z.

Can be **maximized** with respect to ξ to find the best fit.

Variational Inference: Local Methods II





A tangent line is a global lower bound of any convex function.

The bound on the right is tight for $z = \xi$ and $z = -\xi$ (green dashed lines).

Variational Inference: Considerations

- The VI approximation $\mathcal{Q}$ tends to be more compact than the exact p.
- Takes into account global properties of the exact distribution, unlike the Laplace approximation.
- Independent of the basis used. The KL divergence is invariant to any variable transformation.
- Computing each factor given that the others are kept fixed is a convex optimization problem.
- However, the global optimization problem need not be convex and there could be local optima.
- Only suitable when the the logarithm of p is tractable.
- If this is not the case, sometimes further approximations can be made.