Variational Inference

Spring 2024

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing
Outline

• Gaussian Mixture Model and EM algorithm
• Global Variational Inference
  – Variational evidence lower bound
  – Mean-field variational inference
• Local variational inference
  – Convex conjugate
  – Logistic regression
  – Variational EM
• Variational message passing
Outline

• Gaussian Mixture Model and EM algorithm
• Global variational Inference
  – Mean-field variational inference
• Local variational inference
  – Logistic regression
• Variational message passing
Gaussian mixture model (GMM)

• A probabilistic version of the k-means clustering algorithm
• Given a set of data points and a cluster number $K$, how do you group the data points into $K$ clusters?
• Clustering is a fundamental data mining and pattern recognition task
K-means application

Figure 9.3: Two examples of the application of the K-means clustering algorithm to image segmentation showing the initial images together with their K-means segmentations obtained using various values of K. This also illustrates of the use of vector quantization for data compression, in which smaller values of K give higher compression at the expense of poorer image quality.

We can also use the result of a clustering algorithm to perform data compression. It is important to distinguish between lossless data compression, in which the goal is to be able to reconstruct the original data exactly from the compressed representation, and lossy data compression, in which we accept some errors in the reconstruction in return for higher levels of compression than can be achieved in the lossless case. We can apply the K-means algorithm to the problem of lossy data compression as follows. For each of the N data points, we store only the identity k of the cluster to which it is assigned. We also store the values of the K cluster centres µ_k, which typically requires significantly less data, provided we choose K ≪ N. Each data point is then approximated by its nearest centre µ_k. New data points can similarly be compressed by first finding the nearest µ_k and then storing the label k instead of the original data vector. This framework is often called vector quantization, and the vectors µ_k are called code-book vectors.
How do we use probabilistic modeling to represent the clustering procedure?

• Given $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and cluster number $K$

• For each data point $n$  
  
  \[ p(\mathbf{z}_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}} \]
  
  – Sample the cluster membership $\mathbf{z}_n : K \times 1$ one-hot vector, $z_{nk}=1$ means $\mathbf{x}_n$ belongs to cluster $k$

  \[ \mathbf{z}_n = [z_{n1}, \ldots, z_{nK}]^\top \]

  \[ \pi = [\pi_1, \ldots, \pi_K]^\top \]

  – Given $\mathbf{z}_n$, sample $\mathbf{x}_n$ from the cluster-specific Gaussian

  \[ p(\mathbf{x}_n | \mathbf{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)^{z_{nk}} \]
Graphical model representation

Task: Given \( X = \{x_1, \ldots, x_N\} \) and \( K \)
Infer: \( \pi \)
\( \{\mu_1, \ldots, \mu_K\} \)
\( \{\Sigma_1, \ldots, \Sigma_K\} \)
\( p(z_1, \ldots, z_N|X) \)
How to learn GMMs?

- Marginalize out $z$ and do MLE

$$p(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right\}$$

$$0 \leq \pi_k \leq 1$$

s.t. $$\sum_{k=1}^{K} \pi_k = 1$$
How to learn GMMs?

• Given the parameters, we calculate the posterior of the cluster membership

\[ p(z_n | X) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)} \]

Why? Leave it as your exercise
How to learn GMMs?

- Singularity issues

Suppose we use diagonal covariance, when the cluster center is close to a data point, it tends to collapse onto the point to increase the likelihood.

\[
\mathcal{N}(x_n | x_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}
\]

\[
\sigma_j \to 0.
\]
EM algorithm to learn GMM

- Can we get rid of the singularity issue?
- Can we jointly estimate the parameters and the posterior?

\[ p(\mathbf{z}_1, \ldots, \mathbf{z}_N | \mathbf{X}) \]
Let us look at a general case

Suppose we have a model governed by parameters $\theta$

$$p(X|\theta) = \int p(X, Z|\theta)\,dZ$$

Observations
Latent random variables

Question: what are $\theta$ and $Z$ for GMMs?
EM algorithm: how to estimate $\theta$

$$\log (p(X|\theta)) = \log \left( \int p(X, Z|\theta) dZ \right)$$

$$= \log \left( \int \frac{p(X, Z|\theta)}{q(Z)} q(Z) dZ \right)$$

$$\geq \int q(Z) \log \left( \frac{p(X, Z|\theta)}{q(Z)} \right) dZ$$

Jensen’s inequality

$$\log (p(X|\theta)) = \int q(Z) \log \left( \frac{p(X, Z|\theta)}{q(Z)} \right) dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X)} dZ$$

$L(\theta, q(Z))$

$KL(q(Z)||p(Z|X, \theta)) \geq 0$
EM Algorithm

\[
\log (p(\mathbf{X}|\theta)) = \int q(\mathbf{Z}) \log \left( \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right) d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}
\]

\[
L(\theta, q(\mathbf{Z}))
\]

\[
\text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \theta)) \geq 0
\]

\[
\log (p(\mathbf{X}|\theta)) = L(\theta, q^*(\mathbf{Z})) \quad \text{when} \quad q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)
\]

Now fix \( q^*(\mathbf{Z}) \)

\[
\theta^{\text{new}} = \arg \max_{\hat{\theta}} L(\hat{\theta}, q^*(\mathbf{Z}))
\]

\[
\log (p(\mathbf{X}|\theta^{\text{new}})) \geq L(\theta^{\text{new}}, q^*(\mathbf{Z})) \geq L(\theta, q^*(\mathbf{Z})) = \log (p(\mathbf{X}|\theta))
\]

Like a bridge to improve the parameters!
EM Algorithm

- Choose an initial setting $\theta^{\text{new}}$
- Repeat
  \[
  \theta^{\text{old}} \leftarrow \theta^{\text{new}}
  \]
  Evaluate $q(Z) = p(Z|X, \theta^{\text{old}})$ \hspace{1cm} E step
  Fix $q(Z)$, $\theta^{\text{new}} = \arg \max_{\theta} L(\theta, q(Z))$ \hspace{1cm} M step
- Until $||\theta^{\text{old}} - \theta^{\text{new}}|| \leq \epsilon$ or reach the maximum # of iterations

\[
\log(p(X|\theta^{\text{new}}))
\]
\[
L(\theta^{\text{new}}, p(Z|X, \theta^{\text{old}}))
\]
\[
\log(p(X|\theta^{\text{old}}))
\]
\[
L(\theta^{\text{old}}, p(Z|X, \theta^{\text{old}}))
\]
GMM revisited

- Given \( X = \{x_1, \ldots, x_N\} \) and cluster number \( K \)
- For each data point \( n \)
  - Sample the cluster membership \( z_n : K \times 1 \) one-hot vector, \( z_{nk}=1 \) means \( x_n \) belongs to cluster \( k \)
  - Given \( z_n \), sample \( x_n \) from the cluster-specific Gaussian

\[
p(z_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}}
\]

\[
p(x_n | z_n) = \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}
\]
EM algorithm for GMM

\[ p(X, Z | \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{nk} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}} \]

\[ q(Z) = \prod_{n=1}^{N} p(z_n | X, \theta^{\text{old}}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \gamma_{nk} \]

\[ \gamma_{nk} \equiv p(z_{nk} = 1 | X, \theta^{\text{old}}) = \frac{\pi_{nk} \mathcal{N}(x_n | \mu_{nk}^{\text{old}}, \Sigma_{nk}^{\text{old}})}{\sum_{j=1}^{K} \pi_{jn} \mathcal{N}(x_n | \mu_{jn}^{\text{old}}, \Sigma_{jn}^{\text{old}})} \]
EM algorithm for GMM

\[ p(X, Z|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{k}^{z_{nk}} \mathcal{N}(x_{n}|\mu_{k}, \Sigma_{k})^{z_{nk}} \]

M step: \[ L(\theta, q(Z)) = \sum_{Z} q(Z) \log \left( \frac{p(X, Z|\mu, \Sigma, \pi)}{q(Z)} \right) \]

\[ = \sum_{Z} q(Z) \log \left( p(X, Z|\mu, \Sigma, \pi) \right) + \text{const} \]

\[ = \mathbb{E}_{q(Z)} \log \left( p(X, Z|\mu, \Sigma, \pi) \right) + \text{const} \]

\[ \mathbb{E}_{q(Z)} \left[ \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \pi_{k} + z_{nk} \log \left( \mathcal{N}(x_{n}|\mu_{k}, \Sigma_{k}) \right) \right] \]

\[ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \pi_{k} + \gamma_{nk} \log \left( \mathcal{N}(x_{n}|\mu_{k}, \Sigma_{k}) \right) \]
EM algorithm for GMM

- M step:

\[
\max_{\pi_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \pi_k + \gamma_{nk} \log (N(x_n | \mu_k, \Sigma_k))
\]

\[
\pi_{k}^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk}}
\]

\[
\mu_{k}^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n
\]

\[
\Sigma_{k}^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T
\]

\[
N_k = \sum_{n=1}^{N} \gamma_{nk}
\]
EM algorithm for GMMs

• E step \( \gamma_{nk} \equiv p(z_{nk} = 1|X, \theta_{\text{old}}) = \frac{\pi_{k}^{\text{old}} \mathcal{N}(x_{n}|\mu_{k}^{\text{old}}, \Sigma_{k}^{\text{old}})}{\sum_{j=1}^{K} \pi_{j}^{\text{old}} \mathcal{N}(x_{n}|\mu_{j}^{\text{old}}, \Sigma_{j}^{\text{old}})} \)

• M step

\[
\pi_{k}^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk}}
\]

\[
\mu_{k}^{\text{new}} = \frac{1}{N_{k}} \sum_{n=1}^{N} \gamma_{nk} x_{n}
\]

\[
\Sigma_{k}^{\text{new}} = \frac{1}{N_{k}} \sum_{n=1}^{N} \gamma_{nk} (x_{n} - \mu_{k}) (x_{n} - \mu_{k})^{\top}
\]

We do not have any singularity issues!
EM algorithm for GMMs

Figure 9.8 Illustration of the EM algorithm using the Old Faithful set as used for the illustration of the $K$-means algorithm in Figure 9.1. See the text for details.

and the M step, for reasons that will become apparent shortly. In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibilities, given by (9.13). We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing coefficients using the results (9.17), (9.19), and (9.22). Note that in so doing we first evaluate the new means using (9.17) and then use these new values to find the covariances using (9.19), in keeping with the corresponding result for a single Gaussian distribution. We shall show that each update to the parameters resulting from an E step followed by an M step is guaranteed to increase the log likelihood function. In practice, the algorithm is deemed to have converged when the change in the log likelihood function, or alternatively in the parameters, falls below some threshold. We illustrate the EM algorithm for a mixture of two Gaussians applied to the rescaled Old Faithful data set in Figure 9.8. Here a mixture of two Gaussians is used, with centres initialized using the same values as for the $K$-means algorithm in Figure 9.1, and with precision matrices initialized to be proportional to the unit matrix. Plot (a) shows the data points in green, together with the initial configuration of the mixture model in which the one standard-deviation contours for the two
Practice

• Derive EM algorithm for mixture of Bernoulli distributions

• Derive EM algorithm for Bayesian linear regression
Outline

• Gaussian Mixture Model and EM algorithm
• Global variational Inference
• Local variational inference
• Variational message passing
Global variational inference

• Consider a general model

\[ p(X|\theta) = \int p(X, Z|\theta) dZ \]

Put aside the parameters first (either we use full Bayesian treatment to absorb \( \theta \) into \( Z \) or consider \( \theta \) as fixed constant)
Global variational inference

Question: how to compute the posterior $p(Z|X)$

$$p(Z|X) = \frac{p(X,Z)}{\int p(X,Z) dZ}$$

GMMs, Bayesian linear regression are a few exceptions...

In most cases, you cannot get an analytical result ....

$$p(z) = \mathcal{N}(z|0,1)$$

e.g.,

$$p(x|z) = \sigma(z)^x (1 - \sigma(x))^{1-z}$$
Global variational inference

Question: how to compute the posterior \( p(Z|X) \)

Idea: Now that the true posterior is complicated and tricky to compute, can we find a simple form of distribution (e.g., Gaussian) that approximates the true posterior? In other words, can we designate a family of simple distributions, from which we find the best member that is closest to the true posterior?

\[
p(z) = \mathcal{N}(z|0, 1) \quad p(x|z) = \sigma(z)^x (1 - \sigma(x))^{1-z}
\]

Let us use a Gaussian \( q(z) = \mathcal{N}(z|\mu, \sigma^2) \) to approximate the true posterior \( p(z|x) \)

The problem is how to determine the best \( \mu, \sigma^2 \)
Intuitive thoughts

• Suppose we assume the family (form) of approximate posterior $q(Z|\alpha)$

$$\alpha^* = \arg\min_{\alpha} \text{KL}(q(Z|\alpha) \| p(Z|X))$$

What is the issue?

KL divergence is commonly used to measure the difference between distributions
Detour: go back to what we have derived before

\[ \log(p(X)) = \int q(Z) \log \left( \frac{p(X, Z)}{q(Z)} \right) dZ + \int q(Z) \log \left( \frac{q(Z)}{p(Z|X)} \right) dZ \]

Evidence

\[ \mathcal{L}(q) \]

Variational Lower Bound

\[ \text{KL}(q(Z)||p(Z|X)) \geq 0 \]

\[ \max \mathcal{L}(q) \iff \min \text{KL}(q(Z)||p(Z|X)) \]

Key: Maximize the variational lower bound is equivalent to minimizing the KL divergence!
Global variational inference

• Given a family $S$ of the approximate posterior $q(Z)$,

$$q^*(Z) = \arg\max_{q \in S} \mathcal{L}(q) = \int q(Z) \log \frac{p(X, Z)}{q(Z)} \, dZ$$

$$\mathbb{E}_q \log \frac{p(X, Z)}{q(Z)}$$

Usually there is a trade-off: The larger the family $S$, the better the approximation quality, but the harder the optimization.
Mean-field variational inference

- Assume the approximate posterior is factorized:

\[ q(Z) = \prod_i q(Z_i) \quad \text{Z = \{Z_1, Z_2, \ldots\} nonoverlapping} \]

Each \( q(Z_i) \) is a free form distribution

\[
\max \mathcal{L}(q) = \int \prod_i q(Z_i) \log \left\{ \frac{p(X, Z)}{\prod_i q(Z_i)} \right\} dZ
\]

Solve this optimization by alternative updating
Mean-field variational inference

\[
\max \mathcal{L}(q) = \int \prod_i q(Z_i) \log \left\{ \frac{p(X, Z)}{\prod_i q(Z_i)} \right\} dZ
\]

Update \( q(Z_j) \) giving \( \{q(Z_i)\}_{i \neq j} \) fixed

\[
\mathcal{L}(q(Z_j)) = \int q(Z_j) \prod_{i \neq j} q(Z_i) \log (p(X, Z)) dZ - \int q(Z_j) \log (q(Z_j)) dZ_j + \text{const}
\]

\[\mathbb{E}_{q(Z_{-j})} \log (p(X, Z)) \quad q(Z_{-j}) = \prod_{i \neq j} q(Z_i)\]

Solve this

\[
q(Z_j) \propto \exp \left\{ \mathbb{E}_{q(Z_{-j})} \log (p(X, Z)) \right\}
\]

\[
q(Z_j) = \frac{\exp \left\{ \mathbb{E}_{q(Z_{-j})} \log (p(X, Z)) \right\}}{\int \exp \left\{ \mathbb{E}_{q(Z_{-j})} \log (p(X, Z)) \right\} dZ_j}
\]
Mean-field variational Inference: algorithmic framework

- Choose a factorized posterior form \( q(\mathbf{Z}) = \prod_i q(Z_i) \)
- Repeat
  - For each \( j \)
    - Fixed \( \{q(Z_i)\}_{i\neq j} \) and update \( q(Z_j) \propto \exp \{ \mathbb{E}_{q(Z_{-j})} \log (p(\mathbf{X}, \mathbf{Z})) \} \)
  - End for
- Until each \( q(Z_j) \) changes little or reach maximum # of iterations
Variational Linear Regression

\[ p(\alpha) = \text{Gam}(\alpha|a_0, b_0) \]

\[ p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) \]

\[ p(t|w) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T \phi_n, \beta^{-1}) \]

\[ p(t, w, \alpha) = p(t|w)p(w|\alpha)p(\alpha) \]

Observed data \( t = [t_1, \ldots, t_N]^T \)

Inference task: \( p(w, \alpha|t) \)
Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

\[ q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha) \]

Alternating update

\[ q(\alpha) \propto \exp\{\mathbb{E}_{q(\mathbf{w})} \log p(\mathbf{t}, \mathbf{w}, \alpha)\} \]

\[ q(\alpha) = \text{Gam}(\alpha|a_N, b_N) \]

\[ a_N = a_0 + \frac{d}{2} \quad \text{d: dimension of } \mathbf{w} \]

\[ b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^\top \mathbf{w}] \]
Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

\[ q(\mathbf{w}, \alpha) = q(\mathbf{w}) q(\alpha) \]

Alternating update

\[ q(\mathbf{w}) \propto \exp \{ \mathbb{E}_{q(\alpha)} \log p(t, \mathbf{w}, \alpha) \} \]

\[ q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N) \]

\[ \mathbf{m}_N = \beta \mathbf{S}_N \phi^T \mathbf{t} \]

\[ \mathbf{S}_N = (\mathbb{E}[\alpha] \mathbf{I} + \beta \phi^T \phi)^{-1} \]
Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

\[ q(w, \alpha) = q(w)q(\alpha) \]

The required moments

\[
\mathbb{E}[\alpha] = a_N/b_N \\
\mathbb{E}[ww^T] = m_Nm_N^T + S_N
\]

Predictive distribution

\[
p(t|x, t) = \int p(t|x, w)p(w|t)\,dw \simeq \int p(t|x, w)q(w)\,dw \\
= \mathcal{N}(t|m_N^T\phi(x), \sigma^2(x))
\]
Exponential family

\[ p(\eta | \nu_0, \nu_0) = f(\nu_0, \chi_0)g(\eta)^{\nu_0} \exp \{ \nu_0 \eta^T \chi_0 \} \]

\[ p(X, Z | \eta) = \prod_{n=1}^{N} h(x_n, z_n)g(\eta) \exp \{ \eta^T u(x_n, z_n) \} \]

Task: \[ p(\eta, Z | X) \]

Assume: \[ q(Z, \eta) = q(Z)q(\eta) \]
Exponential family

The updates are analytical

\[ q^*(z_n) = h(x_n, z_n) g(E[\eta]) \exp \left\{ E[\eta^T] u(x_n, z_n) \right\} \]

\[ q^*(\eta) = f(\nu_N, \chi_N) g(\eta)^\nu_N \exp \left\{ \eta^T \chi_N \right\} \]

\[ \nu_N = \nu_0 + N \]

\[ \chi_N = \nu_0 \chi_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n} [u(x_n, z_n)] \]
Outline

- Gaussian Mixture Model and EM algorithm
- Global variational Inference
- Local variational inference
- Variational message passing
Local Variational Inference

• Seeks an bound for a factor function of individual variables or a subset of variables

• Convex conjugate

\[ f(x) = \max_{\lambda} \lambda x - g(\lambda) \]

\[ g(\lambda) = \max_{x} \lambda x - f(x) \]

• Key idea: if a factor is convex, use the convex conjugate obtain an bound (easier form)
Local Variational Inference

In general

\[
\log(p(X, \theta)) = \log(p(\theta)) + \sum_n \log(p(x_n|\theta))
\]

If it is convex to \( \theta \),

\[
f(\theta) \geq \lambda^T \theta - g(\lambda)
\]
Variational logistic regression

- Let us consider the sigmoid function in the likelihood

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

\[ \log(\sigma(x)) = -\log(1 + e^{-x}) \] is concave

Let's verify it
Variational logistic regression

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

\[ \log(\sigma(x)) = - \log(1 + e^{-x}) \]

\[ = - \log \left\{ e^{-x/2} (e^{x/2} + e^{-x/2}) \right\} \]

\[ = \frac{x}{2} - \log \left\{ e^{x/2} + e^{-x/2} \right\} \]

\[ f(x) \]
Variational logistic regression

\[ f(x) = -\ln(e^{x/2} + e^{-x/2}) \]  
Symmetric

\[ f(x) = f(|x|) = f(\sqrt{x^2}) = \tilde{f}(x^2) \]

We can show that \(\tilde{f}\) is convex to \(x^2\), so we can use convex conjugate

\[ g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f(\sqrt{x^2}) \right\} \]
Variational logistic regression

\[ g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f \left( \sqrt{x^2} \right) \right\} \]

\[ 0 = \lambda - \frac{dx}{dx^2} \frac{d}{dx} f(x) = \lambda + \frac{1}{4x} \tanh \left( \frac{x}{2} \right) \]

\[ \xi \quad \text{is the optimal } x \text{ corresponding to } \lambda \]

\[ \lambda(\xi) = -\frac{1}{4\xi} \tanh \left( \frac{\xi}{2} \right) = -\frac{1}{2\xi} \left[ \sigma(\xi) - \frac{1}{2} \right] \]

Note: \[ \lambda(\xi) = \lambda(-\xi) \]
Variational logistic regression

\[ g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f \left( \sqrt{x^2} \right) \right\} \]

\[ g(\lambda) = \lambda(\xi)\xi^2 - f(\xi) = \lambda(\xi)\xi^2 + \ln(e^{\xi}/2 + e^{-\xi}/2) \]

\[ f(x) \geq \lambda x^2 - g(\lambda) = \lambda x^2 - \lambda \xi^2 - \ln(e^{\xi}/2 + e^{-\xi}/2) \]

\[ \log(\sigma(x)) = x/2 + f(x) \]

\[ \sigma(x) \geq \sigma(\xi) \exp \left\{ (x - \xi)/2 + \lambda(\xi)(x^2 - \xi^2) \right\} \]
The left-hand plot shows the logistic sigmoid function for this particular value of \( \lambda \). If we denote this value of \( \lambda \), denoted by the dashed green lines. 

The stationarity condition leads to lower bound on \( \sigma(x) \) having the functional form of \( \frac{\ln(1 + e^{\lambda x})}{\lambda} \), which is plotted for two values of \( \lambda \). We can also obtain a lower bound on the sigmoid having the functional form of \( \frac{\ln(1 + e^{\lambda x})}{\lambda} \).

\( \xi = 2.5 \)

\( \xi = 2.5 \)
Variational logistic regression

- Given an arbitrary feature vector \( \phi \), the binary response \( t \) is sampled from

\[
p(t|w) = \sigma(a)^t \{1 - \sigma(a)\}^{1-t}
\]

\[
a = w^T \phi
\]

\[
= \left(\frac{1}{1 + e^{-a}}\right)^t \left(1 - \frac{1}{1 + e^{-a}}\right)^{1-t}
\]

\[
= e^{at} \frac{e^{-a}}{1 + e^{-a}} = e^{at} \sigma(-a)
\]
Variational logistic regression

From the previous result

\[
\sigma(z) \geq \sigma(\xi) \exp \left\{ (z - \xi)/2 - \lambda(\xi)(z^2 - \xi^2) \right\}
\]

where \( \lambda(\xi) = \frac{1}{2\xi} \left[ \sigma(\xi) - \frac{1}{2} \right] \)

Note: We omit – in the previous symbol.

\[
p(t|w) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}
\]
Variational logistic regression

\[ p(t|w) = e^{at}\sigma(-a) \geq e^{at}\sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\} \]

Given the design matrix (features after appropriate (nonlinear) transformations)

\[ \Phi = [\phi_1, \ldots, \phi_N]^T \]

Binary output \( t = [t_1, \ldots, t_N]^T \)

Each \( t_n \in \{0, 1\} \)

\[ p(t) = \int p(t|w)p(w) \, dw = \int \left[ \prod_{n=1}^{N} p(t_n|w) \right] p(w) \, dw \]

\[ p(t, w) = p(t|w)p(w) \geq h(w, \xi)p(w) \]

\[ h(w, \xi) = \prod_{n=1}^{N} \sigma(\xi_n) \exp \left\{ w^T \phi_n t_n - (w^T \phi_n + \xi_n)/2 - \lambda(\xi_n)([w^T \phi_n]^2 - \xi_n^2) \right\} . \]
Variational logistic regression

• Consider approximate posterior $q(w)$

$$\max \mathbb{E}_{q(w)} \log \left\{ \frac{p(t, w)}{q(w)} \right\}$$

$$\min \text{KL}(q(w) \| p(w|t))$$

Infeasible! Also, if you directly optimize w.r.t a free
from, you obtain the true posterior

Solution: We maximize its variational lower bound!

$$p(t, w) = p(t|w)p(w) \geq h(w, \xi)p(w)$$

Why called "variational LB"?

It’s possible to take equality
Variational logistic regression

\[
\max \mathbb{E}_{q(w)} \log \left\{ \frac{p(w)h(w, \xi)}{q(w)} \right\}
\]

The same as Mean-Field

\[
q(w) \propto \exp \left( \log \left\{ p(w)h(w, \xi) \right\} \right)
\]

\[
p(w) = \mathcal{N}(w|m_0, S_0)
\]

\[
-\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0)
\]

\[
+ \sum_{n=1}^{N} \left\{ w^T \phi_n (t_n - 1/2) - \lambda(\xi_n) w^T (\phi_n \phi_n^T) w \right\} + \text{const}
\]

Complete squares

\[
q(w) = \mathcal{N}(w|m_N, S_N)
\]

\[
m_N = S_N \left( S_0^{-1} m_0 + \sum_{n=1}^{N} (t_n - 1/2) \phi_n \right)
\]

\[
S_N^{-1} = S_0^{-1} + 2 \sum_{n=1}^{N} \lambda(\xi_n) \phi_n \phi_n^T.
\]
Are we done?

- No, we haven’t identified the variational parameters $\xi$

$$
\max_{\xi} \mathcal{L}(q, \xi) = \mathbb{E}_{q(w)} \log \left\{ \frac{p(w)h(w, \xi)}{q(w)} \right\}
$$

$$
\sum_{n=1}^{N} \left\{ \ln \sigma(\xi_n) - \xi_n/2 - \lambda(\xi_n)(\phi_n^T \mathbb{E}[ww^T]\phi_n - \xi_n^2) \right\} + \text{const}
$$

$$
0 = \lambda'(\xi_n)(\phi_n^T \mathbb{E}[ww^T]\phi_n - \xi_n^2)
$$

$$
(\xi_{n}^{\text{new}})^2 = \phi_n^T \mathbb{E}[ww^T]\phi_n = \phi_n^T (S_N + m_N m_N^T) \phi_n
$$
Variational logistic regression

• We conduct an EM procedure

\[
\max \quad \mathcal{L}(q, \xi) = \mathbb{E}_{q(w)} \log \left\{ \frac{p(w)h(w, \xi)}{q(w)} \right\}
\]

E step: update \( q(w) \)

M step: update \( \xi \)

Alternatively maximize the variational lower bond
Why is it called variational bound

• The variational bound is variational transformation, it means, if you do NOT restrict the range of the variational parameters, they always have settings that reach equality

\[
\log(p(t)) \geq \mathbb{E}_{q(w)} \log \left\{ \frac{p(t, w)}{q(w)} \right\}
\]

\[
\mathbb{E}_{q(w)} \log \left\{ \frac{p(t, w)}{q(w)} \right\} \geq \mathcal{L}(q, \xi) = \mathbb{E}_{q(w)} \log \left\{ \frac{p(w)h(w, \xi)}{q(w)} \right\}
\]

In practice, we often restrict the family/range of the variational parameters to gain the computational easiness.
Variational EM algorithm

- In general, if we also need to estimate hyperparameters.

\[
p(X, Z|\theta)
\]

\[
\max_{q(Z), \theta} \mathcal{L}(q(Z), \theta) = \mathbb{E}_{q(Z)} \log \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\}
\]

**E step:** \( q(Z) \leftarrow \arg\max_{q \in S} \mathcal{L}(q(Z), \theta) \) \( \text{fix } \theta \)

**M step:** \( \theta \leftarrow \arg\max_{\theta} \mathcal{L}(q(Z), \theta) \) \( \text{fix } q(Z) \)
### Variational Message Passing

- Consider a Bayesian network

\[ p(x) = \prod_i p(x_i | \text{pa}(x_i)) \]

- Assume a factorized posterior over the nodes

\[ q(x) = \prod_i q_i(x_i) \]
Variational Message Passing

• Consider the update on each node

\[ q(x_j) \propto \exp\left\{ \mathbb{E}_{q(x_{\neg j})} [\log p(x)] \right\} \]

Questions: which factors involve \( x_j \) ?

The conditional probabilities where \( x_j \) is a parent/child

\[ q(x_j) \propto \exp\left\{ \mathbb{E} \left[ \log p(x_j | \text{pa}(x_j)) \right] \right\} \]

\[ + \sum_{x_j \in \text{pa}(x_t)} \mathbb{E} \left[ \log p(x_t | x_j, \text{pa}(x_t) \setminus \{x_j\}) \right] \]
Variational Message Passing

- Given a graphical model, the mean-field variational update only requires a local computation on the graph
What you need to know

• What is EM algorithm
• Log(Evidence) = Variational Lower Bound + KL
• What is EM algorithm
• Global variational inference, mean-field
• General update in exponential family
• Local variational inference, convex conjugate
• Variational message passing
• Being able to derive and implement variational inference