1 Motivation

In the past few classes, we have explored the problem of finding tractable approximations to otherwise intractable pdfs. That is, given some probability distribution that is either difficult or impossible to work with, we wish to find an approximating pdf that is easier to work with and yet still yields good results.

Thus, far, we have considered both local approximation methods (Laplace Approximation) as well as global approximation methods (Variational Bayes). Global methods attempt to seek an approximation to the entire distribution over all variables while local methods attempt to find acceptable approximations for individual variables or groups of variables.

In this lecture, we consider a local variational method and apply this to variational logistic regression governed by the logistic sigmoid function.

This document is organized as follows: Section 2 details the bounding of the logistic sigmoid function which will be used in the final regression model. In Section 3, we give a general overview of the framework for using the bound on the logistic sigmoid derived in Section 2. Section 4 will develop the regression model for computing the posterior and how to optimize this model using the Expectation-Maximization (EM) algorithm. Section 5 will present some additional examples that use similar techniques for further illustration, and we give some concluding remarks in Section 6.
2 Variational lower bound on Logistic Sigmoid, \( \sigma(x) \)

We will utilize a lower bound on the logistic sigmoid (cdf) function, \( \sigma(x) = \frac{1}{1+\exp(-x)} \) to provide a normal-normal (quadratic in \( x \)) conjugate model when we derive variational logistic regression. Start with:

\[
\log \sigma(x) = -\log (1 + \exp(-x)) \tag{1}
\]
\[
= \log \left( \exp\left(\frac{-x}{2}\right) \left(\exp\left(\frac{-x}{2}\right) + \exp\left(\frac{-x}{2}\right)\right) \right) \tag{2}
\]
\[
= \frac{x}{2} - \log \left( \exp\left(\frac{x}{2}\right) + \exp\left(\frac{-x}{2}\right) \right) \tag{3}
\]

now, \( f(x) = -\log \left( \exp\left(\frac{\xi}{2}\right) + \exp\left(\frac{-\xi}{2}\right) \right) \) is a convex function of \( x^2 \) (which one may verify by taking the second derivative with respect to \( x^2 \) and noting it is everywhere positive). Therefore, we may construct a lower bound on \( f(x) \) which is a linear function of \( x^2 \),

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

which we maximize in \( x^2 \) to produce (using the chain rule of differentiation),

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

Denote the value of \( x \) corresponding to the tangent line for the given value of \( \lambda \) as \( \xi \), which gives us an expression for the variational parameter \( \lambda \) in terms of \( \xi \).

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

We’ll let \( \xi \) play the role of the (local) variational parameter i.l.o \( \lambda \) because we obtain a simpler expression for the lower bound.

\[
g(\lambda) = \lambda(\xi) \xi^2 - f(\xi) = \lambda(\xi) \xi^2 + \log \left( \exp\left(\frac{\xi}{2}\right) + \exp\left(\frac{-\xi}{2}\right) \right)
\]

so that we may state a lower bound on \( f(x) \) with

\[
f(x) \geq \lambda x^2 - g(\lambda) = \lambda x^2 - \lambda \xi^2 - \log \left( \exp\left(\frac{\xi}{2}\right) + \exp\left(\frac{-\xi}{2}\right) \right)
\]
We finally achieve a lower bound on $\sigma(x)$ by using equation (3), exponentiating $\log \sigma(x)$ and multiplying and dividing by $\exp(-\xi^2)$, as follows:

$$\log \sigma(x) \geq \frac{x}{2} + \lambda x^2 - \lambda \xi^2 - \log \left( \exp\left(\frac{\xi}{2}\right) + \exp\left(-\frac{\xi}{2}\right) \right)$$

$$\sigma(x) \geq \exp\left(\frac{x-\xi}{2}\right) \exp(\lambda(x^2 - \xi^2) \left[ \exp\left(-\frac{\xi}{2}\right) \left( \exp\left(\frac{\xi}{2}\right) + \exp\left(-\frac{\xi}{2}\right) \right) \right]^{-1}$$

So that we finally achieve our variational lower bound on $\sigma(x)$ as

$$\sigma(x) \geq \sigma(\xi) \exp \left( \frac{x - \xi}{2} + \lambda (x^2 - \xi^2) \right)$$

We see that this bound has the form of the exponential of a quadratic function in $x$, which gives us a local Gaussian approximation of the logistic cdf.

3 Framework for how to use Variational Lower Bound on $\sigma(x)$

Suppose we wish to evaluate an integral of the form:

$$I = \int \sigma(a)p(a)da$$

$\sigma(a) =$ logistic cdf $\equiv$ likelihood function for the test data

$p(a) =$ Gaussian posterior where $a \equiv$ latent variables/parameters

We employ the variational lower bound for $\sigma(a) \geq f(a, \xi)$. Then the integral becomes the product of 2 Gaussians (exponential-quadratic function) and we can 'complete the square' in order to evaluate $I$, analytically. Note that we have introduced an additional parameter, $\xi$, the variational free parameter where

$$I \geq \int f(a, \xi)p(a)da = F(\xi)$$

We now find the value of $\xi^*$ that maximizes $F(\xi)$ to get the tightest bound, $F(\xi^*)$. Yet, $\sigma(a) \geq f(a, \xi)$, so that the required choice for $\xi^*$ depends on a specific $a$. Here, however, we marginalize over all values of $a$. So $\xi^*$ is a compromise, weighted by $p(a)$
4 Variational Logistic Regression

4.1 Derive the Regression Model from the class of Generalized Linear Models

Suppose we have an observed set of random samples, \( x = (x_1, \cdots, x_n) \), where \( x \) may be continuous, count or categorical data. We further suppose that the distribution for \( x \) derives from the exponential family of distributions, which has the following form:

\[
x \sim f(x|\eta) = \exp \left( \eta^T T(x) - \zeta(\eta) \right) h(x)
\]

For example, for \( x_i \sim (iid) \mathcal{N}(\mu, \sigma^2) \), we may factorize the joint distribution for \( x \) to demonstrate that the Gaussian distribution is a member of the exponential family, as:

\[
f(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right) = \frac{1}{\sqrt{2\pi}} \exp \left( \frac{\eta_1 T_1}{\sigma^2} - \frac{1}{2\sigma^2} \sum x_i^2 - \frac{\mu^2}{2\sigma^2} - \log \sigma \right) (h(x) = 1)
\]

where \( \eta_1 T_1 = \frac{n\mu}{\sigma^2} \bar{x} \) and \( \eta_2 T_2 = \frac{1}{2\sigma^2} \sum x_i^2 \) and \( \zeta(\eta) = \frac{\mu^2}{2\sigma^2} + \log \sigma \). As an aside, we note the important role played by \( T = (T_1, T_2) \) in the exponential family where the \( n \) observations, \((x_1, \cdots, x_n)\), are reduced to 2 statistics, in the case of the Gaussian distribution (independent of the data). So we refer to \( T \) as the sufficient statistics such that all the relevant information in the data useful for estimating our parameters, \( \eta \), are fully captured by the sufficient statistics.

Now we utilize the exponential family to regress \( \eta \) against a set of covariates, \( x \), as:

\[
g(\eta) = w^T \phi(x)
\]

where Shao (2003) notes \( g(\eta) \in \mathbb{R} \) is called the 'link function' and \( g(\eta) = \eta \) is called the canonical link function. \( w^T \phi(x) \) is a linear combination of transformed \( x \) (e.g. \( \sin(x) \)) and \( w \) may be viewed as slope parameters we estimate to determine the relationship of \( \phi(x) \) to \( g(\eta) \).
Now, let’s develop a regression model specific to count data. Let $t = (t_1, \cdots, t_k)$ for $t_n \in \{0, 1\}$, categorical, and $\pi_n = P(t_n = 1|w) \in [0, 1]$. Define the associated likelihood,

$$p(t_n|w) = \pi_n^{t_n} (1 - \pi_n)^{(1 - t_n)} \propto \exp \left[ t_n \log \left( \frac{\pi_n}{1 - \pi_n} \right) \right]$$

where $\eta_n = \log \left( \frac{\pi_n}{1 - \pi_n} \right) \in \mathbb{R}$ is our link function (which is a latent response). Let’s re-label $\eta_n$ as $a_n$ and define our regression function,

$$a_n = \log \frac{\pi_n}{1 - \pi_n} = w^T \phi(x_n) \quad (7)$$

We may invert this regression function and solve for $\pi_n$ with

$$\pi_n = \frac{1}{1 + \exp(-w^T \phi_n)} = \sigma(w^T \phi_n)$$

$$\sigma(w^T \phi_n) = \sigma(a_n) = \frac{1}{1 + \exp(-a_n)}$$

Then, returning to our likelihood,

$$p(t_n|w) = \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n} \quad (8)$$

$$= \left( \frac{1}{1 + \exp(-a_n)} \right)^{t_n} \left( 1 - \frac{1}{1 + \exp(-a_n)} \right)^{1-t_n} \quad (9)$$

$$= \exp(a_n t_n) \frac{\exp(-a_n)}{1 + \exp(-a_n)} = \exp(a_n t_n) \sigma(-a_n) \quad (10)$$

So we note that our likelihood will be approximated by a Gaussian distribution when we introduce the variational parameters, $\xi_n$.

Now we complete the specification of our logistic regression model with a Gaussian prior on our slope parameters, $w$, $p(w) = \mathcal{N}(w|m_0, S_0)$, where $m_0$ and $S_0$ are fixed hyperparameters.

### 4.2 Develop Variational Lower bound for Posterior

We are moving towards defining some tractable approximation for our posterior (in $w$) to achieve a normal-normal model. Recall, we derive a posterior on $w$ as

$$p(w|t) \propto p(w)p(t|w) = p(t, w) \quad (11)$$
As an aside, we might then use the posterior to conduct variable selection by thresholding (based on the False Discovery Rate) the posterior values on \( w_n \).

Returning to our likelihood from equation (10), we introduce a variation parameter, \( \xi_n \), corresponding to each observation, \((\phi_n, t_n)\) to approximate \( \sigma(-a_n) \) with the variational lower bound from equation (5) such that,

\[
p(t, w) = p(t|w)p(w) \geq h(w, \xi)p(w)
p(t_n|w) = \exp(at)\sigma(-a) \geq \exp(at)\sigma(\xi) \exp \left( -\frac{(a_n + \xi_n)}{2} - \lambda(\xi_n)(a_n^2 - \xi_n^2) \right)
\]

plugging in for \( a_n = w^T\phi_n \)

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

Returning to equation (11), we now take the log of both sides. The inequality is maintained due to the monotonicity of the logarithm. We then insert our lower bound for \( p(t|w) \) to obtain:

\[
\log p(t|w) \geq \log h(w, \xi) + \log p(w) \equiv q(w)
\]

\[
= \log p(w) + \sum_{n=1}^{N} \left( \log \sigma(\xi_n) + \frac{(w^T\phi_n + \xi_n)}{2} - \lambda(\xi_n)((w^T\phi_n)^2 - \xi_n^2) \right)
\]

Substituting in the Gaussian prior on \( p(w) \), supplies

\[
\log p(t|w) \propto \log q(w) \tag{12}
\]

\[
\doteq \frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + \sum_{n=1}^{N} \left( w^T\phi_n(t_n - \frac{1}{2}) - \lambda(\xi_n)w^T(\phi_n\phi_n^T)w \right) \tag{13}
\]

Our variational approximation \( q(w) \) is quadratic in \( w \), allowing us to complete the square in \( w \) to produce our variational approximation for the posterior,

\[
q(w, \xi) = N(w|m_N, S_N), \text{ where,}
\]

\[
m_N = S_N \left( S_0^{-1}m_0 + \sum_{n=1}^{N} \left( t_n - \frac{1}{2} \right) \phi_n \right) \tag{14}
\]

\[
S_N^{-1} = S_0^{-1} + 2 \sum_{n=1}^{N} \lambda(\xi_n)\phi_n\phi_n^T \tag{16}
\]
4.3 Optimize Approximation over Variational Parameters, $\xi_n$

We have obtained a Gaussian approximation for our posterior distribution and can now optimize over $\xi_n$ to make this a tight bound, starting with

$$\log p(t) = \log \int p(t|w)p(w)$$

$$\geq \log \int h(w,\xi)p(w)dw = \mathcal{L}(\xi)$$ (18)

Since $\mathcal{L}(\xi)$ is defined as an integration over latent, $w$, we may use the EM algorithm to maximize the partial likelihood ($\xi$) by maximizing the full likelihood ($\xi, w$).

We implement the EM algorithm from Bishop (2007) for this application with the following steps,

1. **E (Estimation) Step** - Choose initial values for $\xi_{\text{old}}$ and use these values to determine the variational posterior distribution, $q(w, \xi_{\text{old}})$.

2. **M (Maximization) Step**
   - Recall
     $$\mathcal{L}(q, \xi) = \sum_w q(w) \log \frac{p(t,w|\xi)}{q(w)}$$
     $$\geq \sum_w q(w) \log p(t,w|\xi)$$ (20)
     $$\geq \sum_w q(w) \log h(w,\xi)p(w)$$
     $$= \mathbb{E}_{q(w)} [\log h(w,\xi)p(w)]$$ (22)
   - So then we define our optimization equation for the M-step as,
     $$Q(\xi, \xi_{\text{old}}) = \mathbb{E}_{q(w)} [\log h(w,\xi)p(w)]$$ (23)
     $$= \mathbb{E}_{q(w)} [\log h(w,\xi)]$$ (24)
   - Plugging in for $\log h(w,\xi)$,
     $$Q(\xi, \xi_{\text{old}}) = \sum_{n=1}^{N} \left( \log \sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n) (\phi_n^T \mathbb{E}[ww^T] \phi_n - \xi_n^2) \right)$$ (25)
• Now plug in our expressions for $\sigma(\xi_n)$, $\lambda(\xi_n)$ and differentiate $Q$ to obtain,

$$0 = \lambda'(\xi_n) \left( \phi_n^T \mathbb{E} [ww^T] \phi_n - \xi_n^2 \right)$$  \hspace{1cm} (26)

• Since $\lambda'(\xi_n) \neq 0$, we finally obtain the results for one iteration of the EM algorithm,

$$(\xi_{\text{new}})^2 = \phi_n^T \mathbb{E} [ww^T] \phi_n$$  \hspace{1cm} (27)

$$= \phi_n^T (S_N + m_Nm_N^T) \phi_n$$  \hspace{1cm} (28)

5 Additional Examples

5.1 with S-L approximation

In Jaakkola and Jordan (2000), variational logistic regression is compared to a sequential approximation method by Spiegelhalter and Lauritzen, which is then referred to as the S-L approximation. It utilizes a Laplace approximation. The variational methods perform quite well in comparison, with similar results at low variance and better results at high variance. Figure 1 shows the error in posterior mean and standard deviation, as well as the K-L divergence between the true and approximated distributions for each method to demonstrate the effectiveness of variational techniques.

5.2 Bayesian Logistic Regression for Image Database Queries

Ksantini and Ziou (2008) apply variational lower bands and logistic regression to the dual problems of image classification and retrieval for image databases. They compare this to classical image classification techniques and find that the variational methods achieve superior results.

Their method, dubbed Bayesian Logistic Regression Model (BLRM) uses an approximation method based on the same variation method employing the logistic sigmoid described in this document. This model is used to create feature vectors that are stored in the image database. When the system is queried for an image, a target feature vector is created for the query and comparison metrics are calculated for each of the images in the system. The system then returns a fixed number of images that have the highest metric value.

To evaluate the system, a study was conducted on a group of individuals who would enter image search terms into both the BLRM and CLRM databases.
Figure 1: (a)-(d) Errors in the posterior means and standard deviation as a function $g(\mu') = (1 + e^{-\mu'})^{-1}$ of the prior mean $\mu'$ for (a)-(b) prior standard deviation $\sigma = 1$ and (c)-(d) $\sigma = 2$ (e)-(f) K-L divergence between true and approximate posterior distribution for (e) $\sigma = 2$ and (f) $\sigma = 3$ Jaakkola and Jordan (2000)

The user would then evaluate the precision of all images returned. The precision score for both methods is plotted in 2 against the scope (total number of images returned), which shows that the BLRM model greatly outperforms the classical model.

Additionally, the authors note that the variational approximation used in the BLRM also drastically reduces both the training and query time relative to the CLRM system.

6 Conclusion

Variational methods like the one described here rely on bounding functions of interest with (often simpler) convex or concave bounds. Convex functions are important, because these do not have local minima distinct from global minima, and thus can be solved using a wide variety of methods that require fewer iterations and have strong convergence/error bounds. Concave func-
Figure 2: (a)-(b) Comparison between the Classic and Bayesian models for image retrieval and classification. Precision is scored by the user and is based on the relevance of the images returned by the database. Scope is the number of images returned by the database. Ksantini and Ziou (2008)

tions can be converted to convex functions and thus are also sufficient. EM explored in this paper is one technique for solving for specific arguments of the resultant maximization/minimization problems.

References


