Laplace Approximation

Thursday, September 11, 2008
Rice University
STAT 631 / ELEC 639: **Graphical Models**

Instructor:
Dr. Volkan Cevher

Scribe:
Ryan Guerra

Reviewers:
Beth Bower and Terrance Savitsky

Index Terms: *Laplace Approximation, Approximate Inference, Taylor Series, Chi Distribution, Normal Distribution, Probability Density Function*

When we left off with the *Joint Tree Algorithm* and the *Max-Sum Algorithm* last class, we had crafted "messages" to transverse a tree-structured graphical model in order to calculate marginal and joint distributions. We are interested in finding $p(z|x)$ when $p(x)$ is given as shown below.

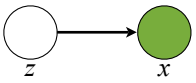

Figure 1. Graph representing both hidden (clear) and observed (shaded) variables with their conditional dependance indicated by the arrow.

In this case, x is our "observed variable" and z is our "given variable" for which we wish to make some inference. While we would normally wish to make some sort of exact inference about $p(z|x)$, this problem is often either impossible to solve or the required algorithm is intractable.

The next few lectures will focus on deterministic approximations to a *pdf* and then we will move on to stochastic approximations. The general hierarchy of approximation techniques is given here for reference.

Approximate Inference

- Deterministic Approximations
  (1) Laplace *(local)*
  (2) Variational Bayes *(global)*
  (3) Expectation Propagation *(both)*
- Stochastic Approximations

(1) Metropolis-Hastings/Gibbs

(2) SIS

1. Laplace Approximations to a PDF

1.1. **Motivation for Representation.** The idea here is that we wish to approximate any *pdf* such as the one given below with a nice, simple representation.

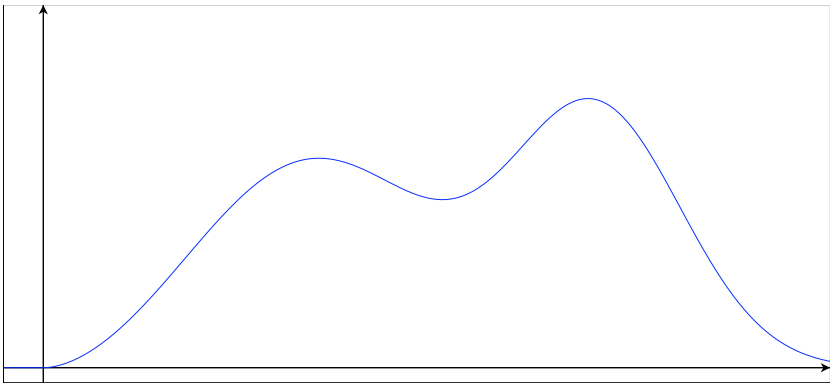

Figure 2. An example multi-modal distribution that we want to approximate.

The Laplace approximation is a method for using a Gaussian $\sim \mathcal{N}(\mu, \sigma^2)$ to represent a given *pdf*. This is obviously more effective for a single-mode[1] distribution, as many popular distributions could be roughly represented with a Gaussian.

As an example of what we mean by "represent," consider that we have some function $g(x)$ distributed according to the density function $p(x)$ and we wish to get the expected value $E\{g(x)\}$ through sampling.

$$E\{g(x)\} = \int g(x)p(x)\mathrm{d}x$$

We wish to calculate this expected value by sampling discrete values from $p(z)$, and thus get an estimate $\hat{E}[g(x)]$ for $E[g(x)]$ that can be calculated as such:

$$\hat{E}\{g(x)\} = \frac{1}{L}\sum_{i=1}^{L} g(x_i)$$

[1]A mode is a concentration of mass in a *pdf*. We could imagine a non-nice distribution like a U-quadratic that may not have its mass concentrated in a single area that would be poorly represented by the Laplace approximation.

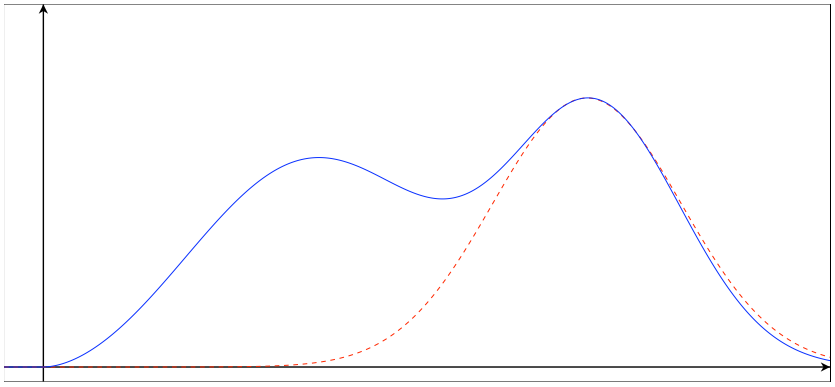

FIGURE 3. The original distribution we want to represent (blue), with its Gaussian approximation (red) obtained by using the *Laplace approximation method*. Note that we were only able to capture one of the original distribution's modes.

Often we will find that $p(x)$ cannot be easily sampled, and we wish to find an alternative way to draw samples from $p(x)$. This is the subject matter for chapter 11 from Bishop[2], but suffice to say we can instead draw samples from another, "nicer" distribution $q(x)$, where $q(x)$ is some known *pdf* and $q(x) \neq 0$.

$$\hat{E}\{g(x)\} = \sum_i g(x_i)p(x_i) = \sum_i g(x_i)\frac{p(x_i)}{q(x_i)}q(x_i)$$

So far, we haven't said anything about the choice of $q(x)$ we could use to represent our *pdf*, but we'd like to use something simple and computable because as the dimensions of the problem increase, the required computational memory increases dramatically. This is why we need approximations.

Thus far we have introduced the motivation behind approximation schemes, in particular the method of Laplace approximation. We will proceed by derving the Laplace Approximation using Taylor series expansions. Then we move to a paper by Tierney and Kadane [3] and describe the use of Laplace approximation to estimate a posterior mean. We conclude with an example of approximating the Chi distribution with a Normal distribution and demonstrate the quality of approximation through graphics.

1.2. **Derivation of the Laplace Approximation.** Suppose we wished to approximate $p(x) = \frac{f(x)}{z}, f(x) \geq 0$. Let's look at the Taylor series expansion[2] for the $\ln f(x)$:

[2]Definition: $f(x) = f(x)\Big|_{x=x_0} + \frac{f'(x)}{1!}\Big|_{x=x_0} \cdot (x-x_0) + \frac{f''(x)}{2!}\Big|_{x=x_0} \cdot (x-x_0)^2 + \ldots + \frac{f^{(n)}(x)}{n!}\Big|_{x=x_0} \cdot (x-x_0)^n + \ldots$

$$(1)\quad \ln f(x) = \ln f(x_0) + \underbrace{\left.\frac{\partial \ln f(x)}{\partial x}\right|_{x=x_0} \cdot (x - x_0)}_{\text{second term}} + \frac{1}{2}\left.\frac{\partial^2 \ln f(x)}{\partial x^2}\right|_{x=x_0} \cdot (x - x_0)^2 + h.o.t...$$

Let's assume that the higher-order terms are negligible and focus, for now, on the second term in (1):

$$(2)\qquad \left.\frac{\partial \ln f(x)}{\partial x}\right|_{x=x_0} \cdot (x - x_0) = \frac{1}{f(x)} \cdot \underbrace{\left.\frac{\partial f(x)}{\partial x}\right|_{x=x_0}}_{*} \cdot (x - x_0)$$

We notice that (*) is zero at local maxima of the *pdf*. If we find this local maxima and choose to expand our Taylor series around this point $x_{\max}$ we ensure that the second term in the RHS of (1) is always zero. This is done by setting $\frac{\partial f(x)}{\partial x}$ equal to zero and solving for $x_{\max}$, the local maxima of the *pdf*.

Taking the first three terms of the Taylor series expansion around $x_0 = x_{\max}$, then (1) becomes:

$$(3)\qquad \ln f(x) = \ln f(x_{\max}) + \frac{1}{2}\left.\frac{\partial^2 \ln f(x)}{\partial x^2}\right|_{x=x_{\max}} \cdot (x - x_{\max})^2$$

$$(4)\qquad \mathrm{e}^{\ln f(x)} = \exp\left[\ln f(x_{\max}) + \frac{1}{2}\left.\frac{\partial^2 \ln f(x)}{\partial x^2}\right|_{x=x_{\max}} \cdot (x - x_{\max})^2\right]$$

$$(5)\qquad \int \mathrm{e}^{\ln f(x)}\mathrm{d}x = \underbrace{\mathrm{e}^{\ln f(x_{\max})}}_{\text{constant}} \int \exp\left[\underbrace{\frac{1}{2}\left.\frac{\partial^2 \ln f(x)}{\partial x^2}\right|_{x=x_{\max}}}_{\text{constant}} \cdot (x - x_{\max})^2\right] \mathrm{d}x$$

Where we took the exponent in (4) and integrated both sides in (5). We see that the RHS of (5) contains a bunch of constants and a single term inside the exponent that is quadratic with respect to x. For the sake of simplicity, let's let $\ln f(x) = F(x)$. Then if we let $\sigma^2 = -\frac{1}{L''(x_{\max})}$ we get a result that looks remarkably like a Gaussian!

$$(6)\qquad \int \mathrm{e}^{L(x)}\mathrm{d}x \approx \mathrm{e}^{L(x_{\max})} \int \exp\left[-\frac{(x - x_{\max})^2}{2\sigma^2}\right] \mathrm{d}x$$

This is the result of the *Laplace method for integrals*, though it is cited with an additional term n and with the traditional notation $x_{\max} = x^*$ in Tierney & Kadane [3] as:

$$\boxed{\int \mathrm{e}^{nL(x)}\mathrm{d}x \approx \mathrm{e}^{nL(x^*)} \int \exp\left[-\frac{n(x-x^*)^2}{2\sigma^2}\right] \mathrm{d}x = \sqrt{2\pi}\sigma n^{-\frac{1}{2}}\mathrm{e}^{nL(x^*)}} \tag{7}$$

1.3. **Application.** To elaborate further on what (7) gets us, I'm going to borrow from Tierney & Kadane's paper. Given a smooth, positive function g, we wish to approximate the posterior mean of $g(x)$.

$$E_n[g(x)] = E[g(x)|Y^{(n)}] = \frac{Pr[g, Y^{(n)}]}{Pr[Y^{(n)}]} = \frac{\int g(x)\mathrm{e}^{\mathcal{L}(x)}\pi(x)\mathrm{d}x}{\int \mathrm{e}^{\mathcal{L}(x)}\pi(x)\mathrm{d}x} \tag{8}$$

Where $\mathcal{L}(x)$ is the *log-likelihood function* $\ln p(x)$, $\pi(x)$ is the prior density, and $Y^{(n)}$ is the observed set of data. As we can see, the forms of the integrals in (8) are very similar to the forms seen in (7), and are then easily estimated by the *Laplace approximation of a pdf*. But you can read the paper if you're interested.

We now have a step-by-step process for using the Laplace approximation to approximate a single-mode *pdf* with a Gaussian:

(1) find a local maximum $x_{\max}$ of the given *pdf* $f(x)$

(2) calculate the variance $\sigma^2 = -\frac{1}{f''(x_{\max})}$

(3) approximate the *pdf* with $p(x) \approx \mathcal{N}[x_{\max}, \sigma^2]$

1.4. **Example: Chi distribution.** $p(x) = \frac{x^{k-1}\mathrm{e}^{-\frac{x^2}{2}}}{z}$, $z = 2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})$, $x > 0$

Note that z is a normalization constant that doesn't depend on x. Apparently most books don't even bother with it, and we'll ignore it here for the sake of convenience. Remember that we're working with the log-likelihood here, so $f(x) = \ln p(x)$

$$\begin{aligned}
f(x) = \ln p(x) &= \ln x^{k-1} + \ln \mathrm{e}^{-\frac{x^2}{2}} \\
\frac{\partial}{\partial x}\ln p(x) &= \frac{\partial}{\partial x}\left[\ln x^{k-1} - \frac{x^2}{2}\right] \\
&= \frac{1}{x^{k-1}} \cdot (k-1)x^{k-2} - x \\
&= \frac{k-1}{x} - x = 0 \\
x^* &= \sqrt{k-1}
\end{aligned}$$

Now that we've found the local maximum x^*, we compute the variance $\sigma^2 = -\frac{1}{f''(x^*)}$

$$\begin{aligned} f''(x^*) =& \frac{\partial^2 f(x)}{\partial x^2}\bigg|_{x=x^*} \\ =& -\frac{k-1}{x^2} - 1\bigg|_{x=x^*} \\ =& -2 \\ \sigma^2 =& \frac{1}{2} \end{aligned}$$

Now all we need is to create the Normal distribution:

$$\hat{p}(x) \sim \mathcal{N}\left[x \middle| \sqrt{k-1}, \frac{1}{2}\right]$$

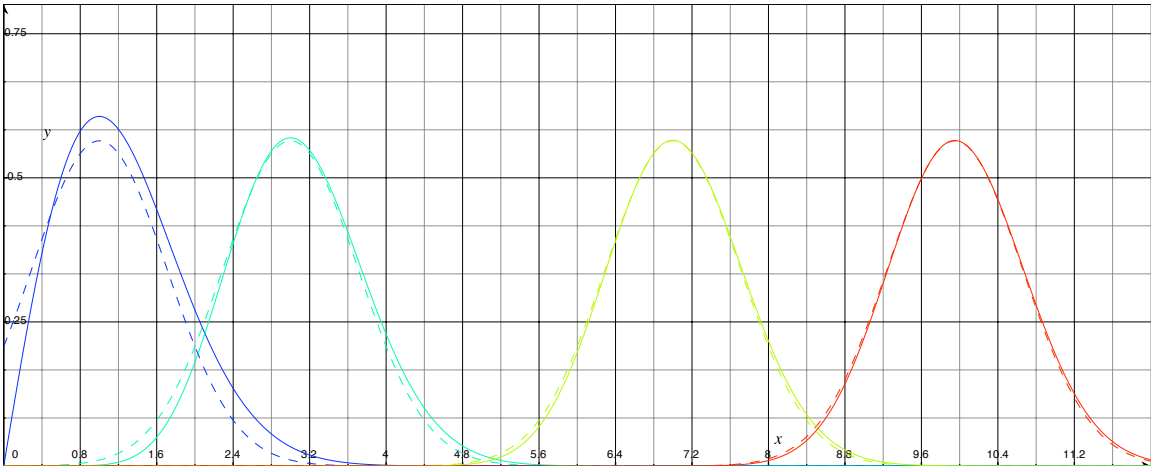

FIGURE 4. A plot of four Chi distributions (solid) and the corresponding Normal approximations (dashed)

1.5. **Application: A Medical Inference Problem.** In the paper "Laplace's Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables" [1], the authors introduce a medical inference problem. The figure below shows the Bayesian graphical model for the problem. There are two different experimental treatments for a disease. The goal is to estimate the posterior mean of the increase in one year survival probability.

In previous work, the authors ran a Monte Carlo simulation to sample from the posterior to determine the posterior mean. In this example, we compare the Laplace approximation to the posterior to the Monte Carlo sampling method. Below is a graph of the Monte Carlo sampling and the Laplace approximation superimposed.

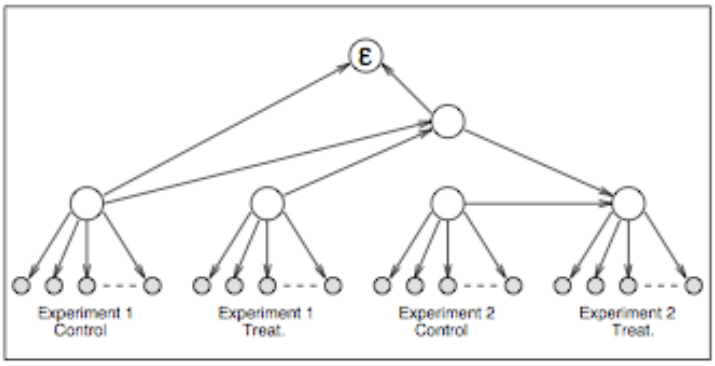

FIGURE 5. The Graphical Model

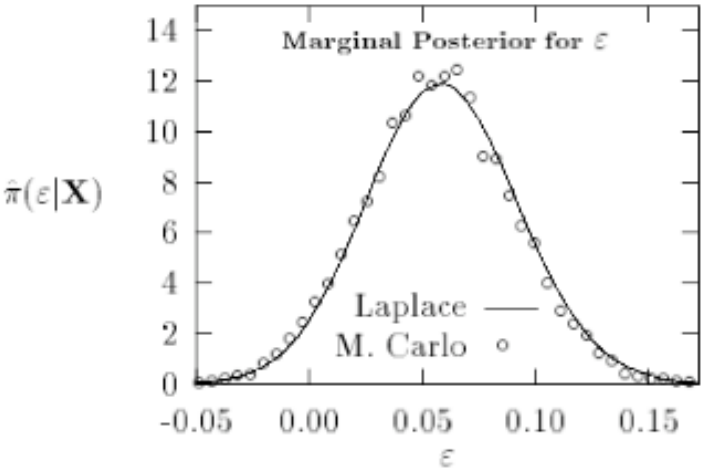

FIGURE 6. Laplace Approximation v. Monte Carlo sampling

The Laplace approximation appears to do well approximating the posterior. The authors note that the Monte Carlo method takes 20 times longer computationally than the Laplace approximation, making Laplace approximation suitable for this example.

REFERENCES

[1] Adriano Azevedo-Filho and Ross D. Shachter. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. *Uncertainty in Artificial Intelligence*, 1994.
[2] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[3] Luke Tierney and Joseph Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), 1986.