Derivation of Expectation Propagation for "Fast Gaussian Process Methods for Point Process Intensity Estimation"

John P. Cunningham
Department of Electrical Engineering,
Stanford University, Stanford, CA 94305
jcunnin@stanford.edu

Abstract

We derive the Expectation Propagation algorithm updates for approximating the posterior distribution on intensity in a conditionally inhomogeneous gamma interval process with a Gaussian Process prior (GP IGIP), a model which appeared in [1].

1 Expectation Propagation Algorithmic Details

Like the Laplace approximation, Expectation Propagation (EP) is a posterior approximation method [2] that creates a Gaussian approximation (or another exponential family distribution) to the true posterior. EP has been found to be superior to Laplace in many contexts (*e.g.* [3]). EP considers global posterior information via iterative local likelihood approximations, whereas Laplace uses information only at the mode of the posterior, setting that mode as the mean of the approximate posterior and the curvature at that point as the covariance. Thus, if the mode does not give an accurate summary of the posterior distribution, Laplace may be ineffective. We will not cover the details of EP here; see [4] for implementation notes and further explanation of EP applied to GP.

In the context of this problem, we have a GP prior on the intensity function $\{x(t)\}$ and the conditionally IGIP likelihood. For model selection (be that modal hyperparameter selection or approximate integration over hyperparameters), we are interested in the model posterior $p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta)$, which requires the intractable data evidence $p(\mathbf{y} \mid \theta) = \int_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x}, \theta)p(\mathbf{x} \mid \theta)d\mathbf{x}$. We would like to use EP to evaluate this data evidence. However, since EP makes iterative updates at each site x_i, running EP on the vector $\mathbf{x}$ is cumbersome and inherits the computational burdens previously discussed (for example, doing rank one updates to the full n-by-n covariance matrix are still necessary). Instead, we can exploit more problem specifics to make EP feasible on a much lower dimensional integral. To do so, we will step away from $\mathbf{x}$, the quantity of interest, and return to the original gamma interarrival distribution $f_z(z) \sim \Gamma(\gamma)$. We define $\mathbf{z}$ with $z_i = \int_{y_{i-1}}^{y_i} x(u)du$, and then the observed event times $\mathbf{y}$ have conditional distribution:

$$p(\mathbf{y} \mid \mathbf{z}) = \prod_{i=1}^{N} p(y_i \mid y_{i-1}, z_i) = \prod_{i=1}^{N} \frac{\gamma^\gamma}{\Gamma(\gamma)} z_i^{\gamma-1} \exp\{-\gamma z_i\}. \tag{1}$$

Then, we can equivalently write the data evidence, our quantity of interest, as $p(\mathbf{y} \mid \theta) = \int_{\mathbf{z}} p(\mathbf{y} \mid \mathbf{z}, \theta)p(\mathbf{z} \mid \theta)d\mathbf{z}$. Importantly, the data evidence is equivalent[1] , but the integral is over the N dimensional vector $\mathbf{z}$ (number of time events), not the n dimensional integral $\mathbf{x}$ (number of time points).

[1]Since the gamma likelihood truncates our distribution over the nonnegative orthant, there is a minor difference in the integral over $\mathbf{x}$ and the integral over $\mathbf{z}$. This difference arises because truncating $\mathbf{z}$ is not the same as individually truncating the elements of $\mathbf{x}$ that sum to $\mathbf{z}$. This minor discrepancy, we believe, is much

Again, $\{x(t)\}$ is a GP in continuous time with a fixed mean and stationary kernel $K_x(\tau)$, that is $\{x(t)\} \sim \mathcal{N}(\mu, K_x(\tau))$. Conveniently, the vector $\mathbf{z}$ is also Gaussian distributed (since each z_i is a linear transformation of $\{x(t)\}$). Then, we have $\mathbf{z} \sim \mathcal{N}(\mathbf{m}, \Pi)$, where $\mathbf{m}_i = (y_{i-1} - y_i)\mu$. If we choose the squared exponential (SE) kernel for $K_x(\tau)$, namely:

$$K(t_i - t_j) = \sigma_f^2 \exp\left\{-\frac{\kappa}{2}(t_i - t_j)^2\right\} + \sigma_v^2 \delta_{ij}, \tag{2}$$

then Π will have the form:

$$\begin{aligned}\Pi = \left\{K_z(i,j)\right\}_{i,j\in\{1,\ldots,N\}} \text{ where } K_z(i,j) &= \int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{j_i} K_x(u-v)dudv \\ &= \sigma_f^2 \int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} \exp\left\{-\frac{\kappa}{2}(u-v)^2\right\} + \sigma_v^2\delta_{u-v}dudv.\end{aligned} \tag{3}$$

Define $\tilde{y}_i = y_i\sqrt{\frac{\kappa}{2}}$ and $\mathrm{erf}(u) = \int_0^u \frac{2}{\sqrt{\pi}}\exp(-v^2)dv$. By this definition, $\int \mathrm{erf}(u)du = u\mathrm{erf}(u) + \frac{1}{\sqrt{\pi}}\exp(-u^2)$, which can be carried through Eq. 3 to yield the (lengthy but computationally simple) expression:

$$\begin{aligned}K_z(i,j) = \frac{\sigma_f^2\sqrt{\pi}}{\kappa}\Big[&(\tilde{y}_i - \tilde{y}_j)\mathrm{erf}(\tilde{y}_i - \tilde{y}_j) + \frac{1}{\sqrt{\pi}}\exp\{-(\tilde{y}_i - \tilde{y}_j)^2\} - (\tilde{y}_i - \tilde{y}_{j-1})\mathrm{erf}(\tilde{y}_i - \tilde{y}_{j-1}) \\ &-\frac{1}{\sqrt{\pi}}\exp\{-(\tilde{y}_i - \tilde{y}_{j-1})^2\} - (\tilde{y}_{i-1} - \tilde{y}_j)\mathrm{erf}(\tilde{y}_{i-1} - \tilde{y}_j) - \frac{1}{\sqrt{\pi}}\exp\{-(\tilde{y}_{i-1} - \tilde{y}_j)^2\} \\ &+(\tilde{y}_{i-1} - \tilde{y}_{j-1})\mathrm{erf}(\tilde{y}_{i-1} - \tilde{y}_{j-1}) + \frac{1}{\sqrt{\pi}}\exp\{-(\tilde{y}_{i-1} - \tilde{y}_{j-1})^2\}\Big] + \\ &\sigma_v^2\Big[(y_i - y_j)_+ - (y_{i-1} - y_j)_+ - (y_i - y_{j-1})_+ + (y_{i-1} - y_{j-1})_+\Big],\end{aligned} \tag{4}$$

where the notation $(\cdot)_+ \triangleq \max(\cdot, 0)$. It is important to note in the details above that only the event times y_i appear, and thus this covariance matrix is calculated in $\mathcal{O}(N^2)$ time and memory, and the larger n is never required.

We have now constructed the distributions $p(\mathbf{y} \mid \mathbf{z})$ and $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{m}, \Pi)$. EP approximates the true posterior $p(\mathbf{z} \mid \mathbf{y})$ with the normal distribution

$$q(\mathbf{z} \mid \mathbf{y}) \triangleq \frac{1}{Z_{EP}} p(\mathbf{z}) \prod_{i=1}^{N} t_i(z_i) = \mathcal{N}(\mu, \Sigma) \text{ where } t_i(z_i) \triangleq \tilde{Z}_i \mathcal{N}\left(\tilde{\mu}_i, \tilde{\sigma}_i^2\right). \tag{5}$$

The EP implementation, and from that the calculation of data evidence, is typical for GP (see [4]). The only step particular to this problem is that of fitting a local (unnormalized) Gaussian to the product of the ith cavity distribution $q_{-i}(z_i)$ and the ith likelihood $p(y_i \mid y_{i-1}, z_i)$. By the standard Kullback-Leibler minimization, we must match the first and second moments (and zeroth, for good measure) of $\hat{q}(z_i) \triangleq \hat{Z}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\sigma}_i^2\right)$ to the moments of $q_{-i}(z_i)p(y_i \mid y_{i-1}, z_i)$. In detail:

$$\hat{Z}_i = \int_0^\infty q_{-i}(z_i)p(y_i \mid y_{i-1}, z_i)dz_i \tag{6}$$

$$\hat{\mu}_i = \frac{1}{\hat{Z}_i}\int_0^\infty z_i q_{-i}(z_i)p(y_i \mid y_{i-1}, z_i)dz_i \tag{7}$$

$$\hat{\sigma}_i^2 = \frac{1}{\hat{Z}_i}\int_0^\infty z_i^2 q_{-i}(z_i)p(y_i \mid y_{i-1}, z_i)dz_i - \hat{\mu}_i^2. \tag{8}$$

smaller than the error introduced by including density outside the nonnegative orthant, as does the Laplace approximation.

Considering the first in detail:

$$\begin{aligned}\hat{Z}_i &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_{-i}}\exp\Big\{-\frac{1}{2\sigma_{-i}^2}(z_i-\mu_{-i})^2\Big\}\frac{\gamma^\gamma}{\Gamma(\gamma)}z_i^{\gamma-1}\exp\Big\{-\gamma z_i\Big\}dz_i \\ &= \frac{\gamma^\gamma}{\Gamma(\gamma)}\exp\Big\{\frac{1}{2}\sigma_{-i}^2\gamma^2-\mu_{-i}\gamma\Big\}\int_0^\infty z_i^{\gamma-1}\frac{1}{\sqrt{2\pi}\sigma_{-i}}\exp\Big\{-\frac{1}{2\sigma_{-i}^2}\big(z_i-\big(\mu_{-i}-\gamma\sigma_{-i}^2\big)\big)^2\Big\}dz_i \\ &= \frac{\gamma^\gamma}{\Gamma(\gamma)}\exp\Big\{\frac{1}{2}\sigma_{-i}^2\gamma^2-\mu_{-i}\gamma\Big\}\bar{E}_{r(z_i)}\big(z_i^{\gamma-1}\big), \end{aligned} \tag{9}$$

where $\bar{E}$ represents the truncated expectation (integrating over the nonnegative half-line instead of the real line), and $r(z_i) \sim \mathcal{N}\big(\mu_{-i}-\gamma\sigma_{-i}^2, \sigma_{-i}^2\big)$. In words, the normalizing constant $\hat{Z}_i$ is the product of a constant and a truncated higher order moment (the $(\gamma-1)$th moment) of a univariate normal distribution $r(z_i)$. By the same logic as Eq. 9, and substituting in for $\hat{Z}_i$,

$$\hat{\mu}_i = \frac{\bar{E}_{r(z_i)}\big(z_i^\gamma\big)}{\bar{E}_{r(z_i)}\big(z_i^{\gamma-1}\big)} \qquad \text{and} \qquad \hat{\sigma}_i^2 = \frac{\bar{E}_{r(z_i)}\big(z_i^{\gamma+1}\big)}{\bar{E}_{r(z_i)}\big(z_i^{\gamma-1}\big)} - \hat{\mu}_i^2. \tag{10}$$

Thus, the only difficult step in calculating an EP update is that of calculating high order truncated moments of a univariate normal distribution. There is no closed form expression for non-integer moments, so we here restrict ourselves to the case of integer values of γ only. If, in a particular application, it is essential to have non-integer values of γ, these moments can be empirically calculated, at the cost of both accuracy and speed. For many applications, however, an integer γ should be adequate.

Though no simple closed form can be derived for truncated higher order integer moments of a normal distribution, we can recursively calculate these moments. We begin with the truncated moment generating function from [5]. Given a normal distribution $u \sim \mathcal{N}\big(a, b^2\big)$, and letting $c = -\frac{a}{b\sqrt{\pi}}$, we write:

$$\bar{E}(u^M) = \frac{1}{2}\left[\sum_{k=0}^{M}\binom{M}{k}a^{M-k}(b\sqrt{2})^k\int_c^\infty\Big(\frac{2}{\sqrt{\pi}}\Big)v^k\exp\big\{-v^2\big\}dv\right]. \tag{11}$$

As suggested in [5], the integral in Eq. 11 can be solved for any k in closed form using integration by parts and the $\mathrm{erf}(\cdot)$ function as previously defined. However, it is tedious and impractical to detail the result of this integral for all reasonable integers that could be assigned to γ. Instead, for any k, we can recursively solve this integral (via two consecutive integrations by parts), and we see that:

$$\begin{aligned}(k=0)\quad &\left[\int_c^\infty\Big(\frac{2}{\sqrt{\pi}}\Big)v^0\exp\big\{-v^2\big\}dv\right] &&= 1-\mathrm{erf}(c), \\ (k=1)\quad &\left[\int_c^\infty\Big(\frac{2}{\sqrt{\pi}}\Big)v^1\exp\big\{-v^2\big\}dv\right] &&= \frac{1}{\sqrt{\pi}}\exp\big\{-c^2\big\}, \\ (k>1)\quad &\left[\int_c^\infty\Big(\frac{2}{\sqrt{\pi}}\Big)v^k\exp\big\{-v^2\big\}dv\right] &&= c^{k-1}\frac{1}{\sqrt{\pi}}\exp\big\{-c^2\big\} \\ &&&\quad + \frac{1}{2}(k-1)\left[\int_c^\infty\Big(\frac{2}{\sqrt{\pi}}\Big)v^{k-2}\exp\big\{-v^2\big\}dv\right].\end{aligned} \tag{12}$$

The integral for any order k can be calculated using only a simple calculation and the $(k-2)$th order of the same integral. For the EP updates, we need the $(\gamma-1)$th, γth, and $(\gamma+1)$th truncated moments as in Eqs. 9,10. By Eqs. 11,12, we can calculate these moments precisely and in $\mathcal{O}(\gamma)$

time, which should for all reasonable purposes be instantaneous. We have shown that this detail of the EP update is exact and computationally simple (though, as is often the case with EP, care must be taken to ensure numerical stability). Since all the other details are standard for EP, the entire EP update then has computational and memory complexity typical for EP, which is $\mathcal{O}(N^3)$ due to the Cholesky factorization required in updating the posterior covariance. Further, since the gamma likelihood is log concave in z_i, EP has only positive site updates (avoiding a known pitfall of EP, see [6]) and has attractive convergence properties (it has been conjectured that EP with a log concave likelihood will always converge [4]). Thus, we have developed a stable EP implementation that operates only on the number of events N instead of the larger n.

Accordingly, we can make a fast approximation of $p(\mathbf{y} \mid \theta)$ using either a Laplace approximation or EP on the transformed variable $\mathbf{z}$. In particular cases, one estimate may do better than others. In our specific application, we find that EP and Laplace perform similarly when the majority of the prior mass is in the nonnegative orthant, and that EP sometimes outperforms when this does not hold. More study is required to understand if EP offers a meaningful improvement in this setting.

References

[1] J. P. Cunningham, M. Sahani, and K. V. Shenoy. In *Proceedings of the 25th International Conference on Machine Learning*. 2008.

[2] T. Minka. *Ph.D. Thesis, Massachusetts Institute of Technology*, 2001.

[3] M. Kuss and C. Rasmussen. *J Machine Learng Research*, 6:1679–1704, 2005.

[4] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[5] J. Jawitz. *Advances in Water Resources*, 27:269–281, 2004.

[6] M. Seeger. *Technical Report, University of Edinburgh*, 2002.