Expectation Propagation

One thing to note about assumed density filtering is that our final approximation is dependent on the sequence in which we process the likelihood terms $\{t_i\}$. Note that we are always matching the moments between

$$q_{i-1}(\boldsymbol{\theta})t_i(\boldsymbol{\theta}),$$

the approximation using data up to the i^{th} term as well as the true i^{th} likelihood term, and the new approximation

$$q_i(\boldsymbol{\theta}) = \tilde{Z}_i q_{i-1}(\boldsymbol{\theta})\tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i).$$

This moment matching at time i therefore never considers data that appears in future terms. For this reason, we might accumulate errors and/or wish to later "revisit" a particular term and update the site parameters $(\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)$ in light of future data. This idea leads to **expectation propagation,** a refinement of assumed density filtering that can address some of these issues.

The idea is simple: once we have processed each of the likelihood terms, resulting in the approximation

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{n} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i),$$

we repeatedly revisit each term and update its site parameters. First, we select a site $1 \leq i \leq n$ to update, then form the so-called **cavity distribution,** which is our approximation using all but the i^{th} term in the product:

$$q_{-i}(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{j \neq i} \tilde{Z}_j \tilde{t}_j(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_j).$$

Another way of conceptualizing the cavity distribution is that we have divided the previous approximation by the old site term $\tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i)$. Next, we replace the removed site term with the true likelihood term t_i, forming the **tilted distribution**

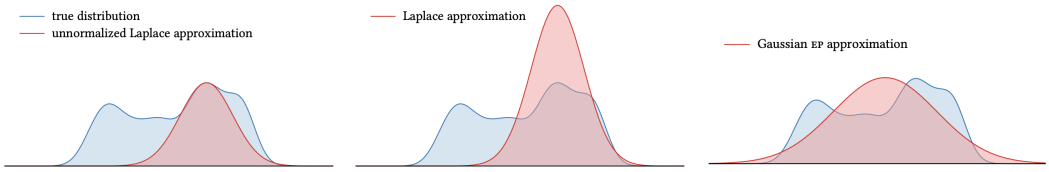
$$q_{-i}(\boldsymbol{\theta})t_i(\boldsymbol{\theta}).$$

Finally, we select new site parameters to (re)match the moments between the tilted distribution and our new approximation:

$$q_{\text{new}}(\boldsymbol{\theta}) = \tilde{Z}_i q_{-i}(\boldsymbol{\theta})\tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i).$$

We proceed continually updating site parameters in this manner until we reach convergence (i.e., none of the site parameters changes very much) or we expend a chosen computational budget.

Note that we do not need to use assumed density filtering to first initialize the site parameters, we can use basically any initialization scheme; a common alternative is to simply initialize all of the site parameters to $\tilde{Z}_i = 1$, $\tilde{\mu}_i = 0$ and $\tilde{\sigma}_i = \infty \; \forall \; 1 \leq i \leq n$, making the initial approximation $q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta})$.



(a) Unnormalized and normalized Laplace approximations (b) Expectation propagation

Figure 1: A comparison of the Laplace approximation and expectation propagation on a one-dimensional, intractable posterior.

Figures 1a and 1b compare the Laplace approximation and expectation propagation on a one-dimensional posterior: while the Laplace approximation does an okay job representing the posterior around its mode, expectation propagation is better at capturing the entirety of the distribution.

EP for Bayesian Probit Regression

Continuing our previous example, suppose that we have used assumed density filtering to approximate the intractable posterior on the latent function f for Bayesian probit regression with the standard normal CDF as our sigmoid function:

$$\frac{1}{Z}\mathcal{N}(\mathbf{f};\boldsymbol{\mu},)\prod_{i=1}^{n}\Phi(y_i f_i) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f};\boldsymbol{\mu},)\prod_{i=1}^{n}\tilde{Z}_i\mathcal{N}(f_i;\tilde{\mu}_i,\tilde{\sigma}_i^2).$$

Suppose we are updating the parameters for the i^{th} term and the current marginal belief about f_i using our approximation q is

$$q(f_i) = \mathcal{N}(f_i,\mu_i,\sigma_i^2).$$

We first compute the parameters of the cavity distribution:

$$q_{-i}(f_i) = \mathcal{N}(\mathbf{f};\boldsymbol{\mu},)\prod_{j\neq i}\tilde{Z}_j\mathcal{N}(f_j;\tilde{\mu}_j,\tilde{\sigma}_j^2) \propto \mathcal{N}(f_i,\mu_{-i},\sigma_{-i}^2)$$

where

$$\mu_{-i} = \sigma_{-i}^2\left(\frac{\mu_i}{\sigma_i^2} - \frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2}\right) \text{ and } \sigma_{-i}^2 = \left(\frac{1}{\sigma_i^2} - \frac{1}{\tilde{\sigma}_i^2}\right)^{-1}.$$

Next, we compute the moments of the tilted distribution,

$$q_{-i}(f_i)\Phi(y_i f_i),$$

which are

$$\hat{Z}_i = \Phi(z_i) \text{ where } z_i = \frac{y_i\mu_{-i}}{\sqrt{1+\sigma_{-i}^2}}$$

$$\hat{\mu}_i = \mu_{-i} + \frac{\mathcal{N}(z_i)}{\Phi(z_i)}s_i \text{ where } s_i = \frac{y_i\sigma_{-i}^2}{\sqrt{1+\sigma_{-i}^2}}$$

$$\hat{\sigma}_i^2 = \sigma_{-i}^2 - \frac{\mathcal{N}(z_i)}{\Phi(z_i)}s_i^2\left(z_i + \frac{\mathcal{N}(z_i)}{\Phi(z_i)}\right).$$

Note the similarity between these moments and the moments we computed during assumed density filtering. The derivation of these moments is almost identical to the derivation of the moments we matched in assumed density filtering, they only differ in the parameters of the Gaussian being multiplied by the standard normal CDF; for brevity, the derivations have been omitted here.

Lastly, we update the site parameters such that the moments of

$$q_{\text{new}}(f_i) = \tilde{Z}_i q_{-i}(f_i)\mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2)$$

to match the moments above. To do this, we set the site parameters to be

$$\begin{aligned}
\tilde{\mu}_i &= \tilde{\sigma}_i^2 \left(\frac{\hat{\mu}_i}{\hat{\sigma}_i^2} - \frac{\mu_{-i}}{\sigma_{-i}^2} \right) \\
\tilde{\sigma}_i^2 &= \left(\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_{-i}^2} \right)^{-1} \\
\tilde{Z}_i &= \hat{Z}_i \sqrt{2\pi \left(\sigma_{-i}^2 + \tilde{\sigma}_i^2 \right)} \exp\left(- \frac{\left(\mu_{-i} - \tilde{\mu}_i \right)^2}{2\left(\sigma_{-i}^2 + \tilde{\sigma}_i^2 \right)} \right).
\end{aligned}$$

We then recompute the approximation $q(\mathbf{f})$ and resample another term to update our belief about.

EP for truncation

Another common use case for expectation propagation is to approximately truncate a Gaussian random variable at some threshold. For a univariate Gaussian, x, we can represent this truncation using a single (non-Gaussian) "likelihood" term:

$$t(x) = \mathbb{1}(x < a)$$

where a is the threshold and $\mathbb{1}$ is the indicator function, which takes value 1 when the argument is true and 0 otherwise.

Again, suppose our initial approximation of the truncation term is $\tilde{t}(x) = \mathcal{N}(x; \tilde{\mu}, \tilde{\sigma}^2)$, giving rise to the approximate belief $q(x) = \mathcal{N}(x; \mu, \sigma^2)$. From here, we can calculate the mean and variance of the cavity distribution as before:

$$\mu_- = \sigma_-^2 \left(\frac{\mu}{\sigma^2} - \frac{\tilde{\mu}}{\tilde{\sigma}^2} \right) \text{ and } \sigma_-^2 = \left(\frac{1}{\sigma^2} - \frac{1}{\tilde{\sigma}^2} \right)^{-1}.$$

Omitting the tedious derivations, the resulting updates to the site parameters are

$$\begin{aligned}
\tilde{\mu} &= \mu_- + \sigma_- \left(\frac{\phi(z)}{\Phi(z)} + z \right)^{-1} \quad \text{where } z = \frac{a - \mu_-}{\sigma_-} \\
\tilde{\sigma}^2 &= \frac{\Phi(z)}{\phi(z)} \left(\frac{\phi(z)}{\Phi(z)} + z \right)^{-1} - \sigma_-^2 \\
\tilde{Z} &= \Phi(z) \sqrt{2\pi \left(\tilde{\sigma}^2 + \sigma_-^2 \right)} \exp\left(- \frac{1}{2\left(1 + \frac{\Phi(z)}{\phi(z)} z \right)} \right).
\end{aligned}$$

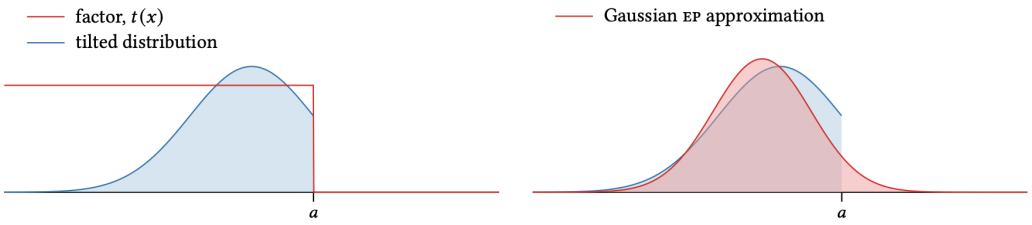


Figure 2: An EP approximation to a truncated Gaussian.

Figure 2 shows the effect of truncating a 1-dimensional Gaussian to be less than a threshold a. As an approximation, there is still some mass above the threshold but it accounts for only roughly 5% of the total probability mass.

Theoretical motivation

We conclude with one brief note about the theoretical motivation behind the moment matching used in assumed density filtering and expectation propagation. The Kullback–Leibler (KL) divergence (also called *relative entropy*) is a notion of "distance" between probability distributions, defined by

$$d_{\text{KL}}\big(p(\boldsymbol{\theta}) \parallel q(\boldsymbol{\theta})\big) = \int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta}.$$

Notice that KL divergence is not a true distance, as it is not symmetric, but it does satisfy $d_{\text{KL}}(p \parallel q) \geq 0$ with $d_{\text{KL}}(p \parallel q) = 0$ if and only if $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$ almost everywhere.

A well-known result is that the KL divergence between an arbitrary probability distribution $p(\boldsymbol{\theta})$ and a multivariate Gaussian distribution $q(\boldsymbol{\theta})$ (in the direction $d_{\text{KL}}(p \parallel q)$) is minimized when q is chosen to match the moments of p. Therefore, in the case of Gaussian approximations, these methods can be seen as iteratively building up an approximate posterior in the desired family by minimizing the KL divergence at every step.

Minimizing KL divergence in "the other direction," $d_{\text{KL}}(q \parallel p)$, gives rise to another family of approximation techniques known as *variational Bayesian inference*, which we will learn more about later in the semester.