Approximations of the measurement update and model likelihood for nonlinear spatiotemporal Cox processes

M. Rule

July 10, 2018

We are interested in approximations to the measurement update for a spatially-extended latent-variable point-process, where the latent variables are also spatial fields that undergo dynamics similar to chemical reaction-diffusion systems.

The latent variables or fields are concentrations, activations, or some similar physical quantity. They are therefore constrained to be non-negative, and also typically must obey conservation laws. Additionally, the observed point-process intensity field must also be constrained to be non-negative.

Such systems arise in chemical reaction diffusion systems, epidemiological models, and neural field models, where the measurement is a point-process that is coupled indirectly to the latent spatiotemporal system.

1 Problem statement

We will use a multivariate Gaussian approximation to model the joint distribution of latent variables. In the continuous filed case, this is a Gaussian process. In the numerical implementation, we project this process onto a finite basis to give a finite-dimensional multivariate Gaussian distribution. The filtering update therefore requires finding a multivariate Gaussian approximation to the non-conjugate update of a Poisson observation and a multivariate Gaussian prior.

Estimating this posterior involves an integral that becomes intractable in high-dimensions. There are numerous approximation methods to handle this, including moment-matching, variational Bayes, expected log-likelihoods, and the Laplace approximation. Often, the choice of link function furthers constrain which methods are applicable, as efficient algorithms may only be available in some classes. The update must be also constrained to prevent negative activity in the estimated latent states.

1.0.1 Multivariate Gaussian model

Let $\mathbf{A}(\mathbf{x})$ denote a latent vector of activity, defined over a spatial domain with coordinates $\mathbf{x} \in \Omega$ on some (likely bounded) domain Ω. We approximate the prior distribution over the latent activity vector $\mathbf{A}(\mathbf{x})$ in terms of its first two moments. We interpret these moments to reflect the mean and variance of a multivariate Gaussian distribution, for the purposes of both moment closure and the measurement update. This assumption is an approximation, as in general the tails of the Gaussian that extent to negative activation values are unphysical.

$$\Pr(\mathbf{A}(\mathbf{x})) \sim \text{Gaussian}\left(\mu_A(\mathbf{x}), \Sigma_A(\mathbf{x})\right) \tag{1}$$

$\mathbf{A}(x)$ is a field defined over a continuous spatial region, and so Eq. 1 denotes a Gaussian *process*. In practice, we project this continuous process onto a finite basis and work with N discrete activation variables corresponding to spatial regions, i.e. $\mathbf{A} = \{A_1, .., A_N\}$. In this numerical implementation, Eq. 1 represents a finite-dimensional multivariate Gaussian distribution.

1.0.2 Latent field definition and discretization

This activation field is mapped to point-process intensities via a link function $\lambda_0 = f(\mathbf{A})$. Furthermore, our observation model may include heterogeneity in terms of the density of agents or background level of activity, and we therefore incorporate a spatially inhomogeneous gain $\gamma(\mathbf{x})$ and bias $\beta(\mathbf{x})$ that adjust the observe point-process intensity. The intensity as a function of spatial coordinates $\mathbf{x}$ is then:

$$\lambda(\mathbf{x}) = f\left[\mathbf{A}(\mathbf{x})\right] \cdot \gamma(\mathbf{x}) + \beta(\mathbf{x}). \tag{2}$$

In practice, we project this continuous, infinite-dimensional process onto a finite set of discrete basis elements $\mathbf{B} = \{b_1, .., b_N\}$, where the expected firing rate for the n^{th} basis element is:

$$\lambda_n = \int_{\mathbf{x}\in\Omega} b_n(\mathbf{x}) \cdot \lambda(\mathbf{x}), \tag{3}$$

where $\int_{\mathbf{x}\in\Omega}$ denotes integration over the spatial domain Ω parameterized by $\mathbf{x}$. If the variations in the activity, gain, and bias, are small relative to the scale of the basis elements, we may approximate this integral as:

$$\begin{aligned} \lambda_n &\approx \lambda\left(\langle \mathbf{x} \rangle_{b_n}\right) \cdot v_n \\ v_n &= \int_{\mathbf{x}\in\Omega} b_n(\mathbf{x}), \end{aligned} \tag{4}$$

where $\langle \mathbf{x} \rangle_{b_n}$ is the centre of mass of the basis element b_n, and v_n is the volume of said basis element. We consider an especially simple case where basis functions all have identical volume v, so we may write the regional intensity as

$$\lambda_n \approx v \cdot \left[\gamma_n f(A_n) + \beta_n\right], \tag{5}$$

where v is the (uniform) volume of the basis elements, e.g. $v = \Delta_x^2 \Delta_t$ for a process with two spatial dimensions and one time dimensions, with a fixed region size. The spatially-varying intensity is then represented as a vector of per-region intensities, $\lambda = \{\lambda_1, .., \lambda_N\}$. Since the volume parameter v is redundant to the gain parameter γ, in the derivations that follow we assume that the bias and gain parameters have been premultiplied by the volume.

We introduce the gain and bias parameters to decouple inhomogeneity in the measurements from the underlying, latent spatiotemporal process. For example, in the case of retinal waves, different regions have differing densities of retinal ganglion cells, which amounts to a spatially inhomogeneous gain. Additionally, the amount of spontaneously-active background activity varies. In order to build up mathematical solutions that are immediately useful for numerical implementation, we carry-through these bias and gain parameters in the derivations that follow.

1.0.3 Count observations and the measurement posterior

Given the Gaussian prior, the Posterior estimate of the latent activations $\mathbf{A}$ is given by Bayes' rule:

$$\Pr(\mathbf{A}|\mathrm{Y}) = \frac{\Pr(\mathbf{A})}{\Pr(\mathrm{Y})} \cdot \Pr(\mathrm{Y}|\mathbf{A}) \tag{6}$$

We observe event counts over a finite number of spatial regions, and these observed counts are independent conditioned on the per-region intensity, so we can write:

$$\Pr(\mathbf{A}|\mathrm{Y}) = \frac{\Pr(\mathbf{A})}{\Pr(\mathrm{Y})} \prod_{n \in 1..N} \Pr(y_n|\mathbf{A}) \tag{7}$$

The dependence of the counts on the latent activation can be expanded to include the regional intensity λ_n as:

$$\Pr(y_n|A_n) = \Pr(y_n|\lambda_n) \cdot \Pr(\lambda_n \mid \mathbf{A}) \tag{8}$$

Since the dependence of λ_n on $\mathbf{A}$ is deterministic (and vice-versa), we can therefore write

$$\Pr(y_n|A_n) = \frac{\Pr(\mathbf{A})}{\Pr(\mathrm{Y})} \prod_{n \in 1..N} \Pr(y_n|\lambda_n) \tag{9}$$

We observe regional counts $\mathrm{Y} = \{y_1, .., y_N\}$, which are Poisson distributed, with the observation likelihood

$$\Pr(y_n \mid \lambda_n) = \frac{\lambda_n^{y_n}}{y_n!} e^{-\lambda_n}. \tag{10}$$

1.0.4 The log-posterior

Consider the logarithmic form of the measurement update:

$$\log \Pr(\mathbf{A}|\mathrm{Y}) = \log \Pr(\mathbf{A}) - \log \Pr(\mathrm{Y}) + \sum_{n\in 1..N} \log \Pr(y_n|\lambda_n) \tag{11}$$

The prior $\mathbf{A}$ is approximated by a Gaussian distribution $\mathcal{N}(\mu_A, \Sigma_A)$, and so the log-prior on $\mathbf{A}$ is:

$$\log \Pr(\mathbf{A}) = -\tfrac{1}{2}\left[\log|2\pi\Sigma_A| + (\mathbf{A}-\mu_A)^\top \Sigma_A^{-1}(\mathbf{A}-\mu_A)\right]. \tag{12}$$

The conditional log-likelihood is given by the Poisson observation model with regional intensity $\lambda_n = \gamma_n f(A_n) + \beta_n$:

$$\begin{aligned}\log \Pr(\mathrm{Y} \mid \mathrm{A}) &= \sum_{n\in 1..N} \log \Pr(y_n|\lambda_n) \\ &= \sum_{n\in 1..N} \left[y_n \log(\lambda_n) - \lambda_n\right]. \\ &= \sum_{n\in 1..N} \left[y_n \log(\gamma_n f(A_n) + \beta_n) - (\gamma_n f(A_n) + \beta_n)\right].\end{aligned} \tag{13}$$

The marginal log-likelihood of the count observations $\log \Pr(\mathrm{Y})$ cannot be computed, except via an intractable integral. Approximating this integral will be a major challenge for computing the model likelihood, which we will address later. However, for fixed count observations Y, the $\Pr \mathrm{Y}$ term is constant, and so:

$$\begin{aligned}\log \Pr(\mathbf{A}|\mathrm{Y}) = &-\tfrac{1}{2}\left[\log|2\pi\Sigma_A| + (\mathbf{A}-\mu_a)^\top \Sigma_A^{-1}(\mathbf{A}-\mu_a)\right] \\ &+ \sum_{n\in 1..N} \left[y_n \log(\lambda_n) - \lambda_n\right] + \text{constant}\end{aligned} \tag{14}$$

2 Approximation methods

In this section, we explore various approaches to obtaining a Gaussian approximation to the posterior $\Pr(\mathbf{A}|\mathrm{Y}) \approx Q(\mathbf{A}) \sim \mathcal{N}(\hat{\mu}_A, \hat{\Sigma}_A)$. We examine three approaches: the Laplace approximation, the variational Bayes approach, and moment-matching.

2.1 Laplace approximation

For the Laplace approximation, we find the mode of the posterior and interpret the curvature at this mode as the inverse of the covariance matrix. Since we are dealing with spatiotemporal processes driven by physical agents (e.g. molecules, neurons, humans), we constrain the posterior mode to be non-negative. This departs slightly from the traditional Laplace approximation, in which the posterior mode is a non-extremal local maximum with zero slope. For this reason, the interpretation of the curvature at the mode as the inverse of the covariance must be treated with caution.

$$\hat{\mu_A} = \underset{\mathbf{A}}{\operatorname{argmax}} \left[\log \Pr(\mathbf{A} \mid \mathbf{Y})\right] \tag{15}$$

This can be solved with gradient descent or the Newton-Raphson method, which requires the gradient and Hessian of the log-posterior with respect to $\mathbf{A}$.

We introduce some abbreviations to simplify the notation in the derivations that follow. Denote log-probabilities "$\log \Pr(\cdot)$" as $\mathcal{L}_{(\cdot)}$, and denote the first and second derivatives of the log-measurement likelihood with respect to individual activation variables $A{\in}\{A_1, .., A_N\}$ as $\mathcal{L}'_{y|A}$ and $\mathcal{L}''_{y|A}$, respectively. Note that λ_n is synonymous with the locally adjusted rate $\gamma_n f(A_n){+}\beta_n$, such that $\lambda'_n{=}\gamma_n f'(A_n)$. We also omit indexing by the basis function number n when unambiguous. With these abbreviations, the gradient and Hessian of the log-posterior in A are:

$$\begin{aligned} \nabla_{\mathbf{A}} \mathcal{L}_{\mathbf{A}|\mathbf{Y}} &= (\mu_a - \mathbf{A})^\top \Sigma_A^{-1} + \nabla_{\mathbf{A}} \mathcal{L}_{\mathbf{Y}|\mathbf{A}} \\ \nabla^2_{\mathbf{A}} \mathcal{L}_{\mathbf{A}|\mathbf{Y}} &= \Sigma_A^{-1} + \nabla^2_{\mathbf{A}} \mathcal{L}_{\mathbf{Y}|\mathbf{A}}, \end{aligned} \tag{16}$$

where

$$\begin{aligned} \mathcal{L}'_{y|A} &= \frac{\gamma}{\lambda} f'(A)\,(y - \lambda) \\ \mathcal{L}''_{y|A} &= \frac{\gamma}{\lambda} \left[(y - \lambda)\, f''(A) - y \frac{\gamma}{\lambda} f'(A)^2 \right] \end{aligned} \tag{17}$$

2.1.1 The identity link function

In the case that $f(A_n) = A_n$, these gradients simplify to:

$$\begin{aligned} \mathcal{L}'_{y|A} &= (y - \lambda) \frac{\gamma}{\lambda} \\ \mathcal{L}''_{y|A} &= -y \left(\frac{\gamma}{\lambda} \right)^2 \end{aligned} \tag{18}$$

2.1.2 The exponential link function

In the case that $f(A) = \exp(A)$:

$$\begin{aligned} \mathcal{L}'_{y|A} &= (y - \lambda) \left(\frac{\gamma e^A}{\lambda} \right) \\ \mathcal{L}''_{y|A} &= \left(\frac{\gamma e^A}{\lambda} \right) \left[y \left(1 - \frac{\gamma e^A}{\lambda} \right) - \lambda \right] \end{aligned} \tag{19}$$

For more flexibility, one might add another gain parameter *inside* the exponentiation, i.e. $f(A_n) = \exp(\delta A_n)$, which gives:

$$\begin{aligned}
\mathcal{L}'_{y|A} &= (y-\lambda)\left(\frac{\gamma\delta}{\lambda}e^{\delta A}\right) \\
\mathcal{L}''_{y|A} &= \gamma\delta^2\left[-\frac{y\gamma}{\lambda^2}e^{2\delta A} + \left(\frac{y}{\lambda}-1\right)e^{\delta A}\right] \\
&= \delta\,(y-\lambda)\left(\frac{\gamma\delta}{\lambda}e^{\delta A}\right) - y\left(\frac{\gamma\delta}{\lambda}e^{\delta A}\right)^2
\end{aligned} \tag{20}$$

2.1.3 The quadratic link function

Let $\lambda = A^2\gamma + \beta$. That is, $f(A) = A^2$, $f'(A) = 2A$, and $f''(A) = 2$.

$$\begin{aligned}
\mathcal{L}'_{y|A} &= \frac{\gamma}{2}A\,(y-\lambda) \\
\mathcal{L}''_{y|A} &= \frac{\gamma}{\lambda}\left[(y-\lambda)\,2 - y\frac{\gamma}{\lambda}4A^2\right]
\end{aligned} \tag{21}$$

A more flexible parameterization is $\lambda = \gamma(A+b)^2 + \beta$ gives:

$$\begin{aligned}
\mathcal{L}'_{y|A} &= \frac{\gamma}{\lambda}(2A+b)\,(y-\lambda) \\
\mathcal{L}''_{y|A} &= \frac{\gamma}{\lambda}\left[(y-\lambda)\,2 - y\frac{\gamma}{\lambda}(2A+b)^2\right]
\end{aligned} \tag{22}$$

2.1.4 Logistic link 1

We might want to consider the logistic link function, which maps the range $(-\infty, \infty)$ in activation A to $(0, 1)$, which then may be further adjusted to span a given range using the gain/bias parameters:

$$\begin{aligned}
f &= \frac{1}{1+e^{-\delta A}} \\
f' &= \delta\frac{e^{-\delta A}}{(1+e^{-\delta A})^2} = \delta[1-f(A)]f(A) \\
f'' &= \delta\,[1-2f]\,f'.
\end{aligned} \tag{23}$$

2.1.5 Logistic link 2

If the activation is bounded on $[0, \infty)$, it might make more sense to apply the logistic function to the log-activation, yielding the following link function:

$$\begin{aligned} f &= \frac{A}{\epsilon + A} \\ f' &= \frac{\epsilon}{(\epsilon + A)^2} \\ f'' &= -2\frac{\epsilon}{(\epsilon + A)^3}. \end{aligned} \tag{24}$$

where ϵ_n is an additional free parameter that acts a like an inverse gain.

2.2 Variational approximation

In the variational approximation, we find a Gaussian distribution $Q(\mathbf{A}) \sim \mathcal{N}(\hat{\mu}_A, \hat{\Sigma}_A)$ that approximates the true posterior by minimizing the KL divergence of the true posterior from the approximating distribution Q. This is conceptually equivalent to jointly maximizing the entropy of Q while also maximizing the expected log-probability of the true posterior under Q.

$$\underset{\mu_Q, \Sigma_Q}{\operatorname{argmin}} D_{\text{KL}}(Q \| P) = \underset{\mu_Q, \Sigma_Q}{\operatorname{argmin}} \int_{\mathbf{A}} Q(\mathbf{A}) \log \frac{Q(\mathbf{A})}{\Pr(\mathbf{A} \mid \mathbf{Y})} = \underset{\mu_Q, \Sigma_Q}{\operatorname{argmax}} \left[\mathrm{H}(Q) + \langle \log \Pr(\mathbf{A} \mid \mathbf{Y}) \rangle_Q \right] \tag{25}$$

To obtain a specific, analytically tractable form of the above, first expand $\log \Pr(\mathbf{A} \mid \mathbf{Y})$ using the logarithmic form of Bayes' rule:

$$\log \Pr(\mathbf{A} \mid \mathbf{Y}) = \log \Pr(\mathbf{Y} \mid \mathbf{A}) + \log \Pr(\mathbf{A}) - \log \Pr(\mathbf{Y}) \tag{26}$$

This gives convenient simplifications, as the prior $\log \Pr(\mathbf{A})$ is often Gaussian and has closed-form solutions, and the marginal data likelihood $\log \Pr(\mathbf{Y})$ is constant and can be dropped from the optimization. Expanding $\langle \log \Pr(\mathbf{A} \mid \mathbf{Y}) \rangle_Q$ in Eq. 25 gives:

$$\underset{\mu_Q, \Sigma_Q}{\operatorname{argmax}} \left[\mathrm{H}(Q) + \langle \log \Pr(\mathbf{Y} \mid \mathbf{A}) \rangle_Q + \langle \log \Pr(\mathbf{A}) \rangle_Q - \langle \log \Pr(\mathbf{Y}) \rangle_Q \right] \tag{27}$$

Dropping the constant $\langle \log \Pr(\mathbf{Y}) \rangle_Q$ term and recognizing that the remaining terms reflect the KL divergence of the approximating posterior from the prior, i.e. $D_{\text{KL}}(Q \| \Pr(\mathbf{A}))$, we get the following optimization problem:

$$\underset{\mu_Q, \Sigma_Q}{\operatorname{argmax}} \left[\langle \log \Pr(\mathbf{Y} \mid \mathbf{A}) \rangle_Q - D_{\text{KL}}(Q \| \Pr(\mathbf{A})) \right]. \tag{28}$$

The objective function for variational Bayes amounts to maximizing the data likelihood $\langle \log \Pr(\mathbf{Y} \mid \mathbf{A}) \rangle_Q$ under the approximation Q, while also minimizing the KL divergence of the prior from the approximating posterior. It can therefore be interpreted as a regularized maximum-likelihood approach. This form also connects to the objective functions often seen in variational autoencoders and in variational free energy.

In this application both the prior and approximating posterior are Gaussian, and the KL divergence term $D_{\mathrm{KL}}(Q\| \Pr(\mathbf{A}))$ has a closed-form solution reflecting the KL divergence between two multivariate Gaussian distributions:

$$D_{\mathrm{KL}}(Q\| \Pr(\mathbf{A})) = \frac{1}{2}\left[\log\frac{|\Sigma_A|}{|\hat{\Sigma}_A|} - D + \mathrm{tr}[\Sigma_A^{-1}\hat{\Sigma}_A] + (\mu_A - \hat{\mu}_A)^T\Sigma_A^{-1}(\mu_A - \hat{\mu}_A)\right], \tag{29}$$

where D is the dimensionality of the multivariate Gaussian. It remains then to calculate the expected log-likelihood, $\langle\log\Pr(\mathbf{Y} \mid \mathbf{A})\rangle_Q$. As discussed in the next section, this integral is not always tractable.

2.2.1 Challenges for the variational approximation in this application

The variational approximation integrates over the domain for A, which is truncated to $[0,\infty)$ since negative values for $\mathbf{A}$ are unphysical. Typically, this means that efficient algorithms are challenging to derive, as closed-form solutions for the relevant integrals do not exist, or at best involve the multivariate Gaussian cumulative distribution function, its inverses, and derivatives, which are numerically expensive to compute.

One may relax the constraint that $\mathbf{A}$ be non-negative, extending the domain of integration to $(-\infty,\infty)$, but then one must constrain optimization to return only positive means for the variational posterior. However, unless a rectifying (e.g. exponential, quadratic) link function is used, the inclusion of negative rates in the domain will make the Poisson observation likelihood undefined. For this reason, the variational update has been explored only for the exponential link function [[PARK]]. Because small changes in activation can lead to large fluctuations in rate owing to the amplification of the exponential link function, we have found that the exponential link is numerically unstable.

An implementation of variational optimization using the rectifying quadratic link function may be more numerically stable, and remains to be explored.

2.3 Moment-matching

Moment matching calculates or approximates the mean and covariance of the true posterior, and uses these moments to form a multivariate Gaussian approximation. When applied as a message-passing algorithm in a graphical model, moment matching is an important step of the expectation-propagation algorithm. Moment-matching can be performed explicitly by integrating the posterior moments, but in high dimensions there is no computationally tractable way to evaluate such integrals. Since spatial correlations are essential in spatiotemporal phenomena, we cannot discard this higher dimensional structure.

Another approach to moment-matching to note is that the the Gaussian distribution Q that minimizes KL divergence from Q to the true posterior will also match the moments of the posterior. We can therefore perform moment matching by minimizing this KL divergence:

$$\operatorname*{argmin}_{\mu_Q,\Sigma_Q} \int_{\mathbf{A}} \Pr\left(\mathbf{A} \mid \mathbf{Y}\right) \log \frac{\Pr\left(\mathbf{A} \mid \mathbf{Y}\right)}{Q(\mathbf{A})} = \operatorname*{argmax}_{\mu_Q,\Sigma_Q} \mathrm{H}\left[\Pr\left(\mathbf{A} \mid \mathbf{Y}\right)\right] + \langle \log Q \rangle_{\Pr(\mathbf{A}|\mathbf{Y})} \tag{30}$$

Note that the first term is the entropy of the (true) posterior distribution. It is constant for a given update, and therefore does not affect our optimization. We can focus on the second term, and optimize:

$$\operatorname*{argmax}_{\mu_Q,\Sigma_Q} \langle \log Q \rangle_{\Pr(\mathbf{A}|\mathbf{Y})} \tag{31}$$

The log-probability of a Gaussian approximation Q is with mean $\hat{\mu}_A$ and covariance matrix $\hat{\Sigma}_A$ is:

$$\log Q\left(A\right) = -\tfrac{1}{2}\left[\log\left|2\pi\Sigma_A\right| + (\mathbf{A} - \hat{\mu}_A)^\top \hat{\Sigma}_A^{-1} (\mathbf{A} - \hat{\mu}_A)\right] \tag{32}$$

We cannot calculate the true posterior $\Pr\left(\mathbf{A} \mid \mathbf{Y}\right)$, and so the integral $\langle \log Q \rangle_{\Pr(\mathbf{A}|\mathbf{Y})}$ cannot be computed directly. However, the normalization constant, although unknown, is constant with respect to this optimization, and is suffices to take the weighted expectation with respect to an un-normalized form of $\Pr\left(\mathbf{A} \mid \mathbf{Y}\right)$.

$$\Pr\left(\mathbf{A}|\mathbf{Y}\right) = \left|2\pi\Sigma_A\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{A}-\mu_a)^\top \Sigma_A^{-1}(\mathbf{A}-\mu_a)} \cdot \prod_{n\in 1..N} \left[\frac{(\gamma_n f(A_n) + \beta_n)^{y_n}}{\exp\left(\gamma_n f(A_n) + \beta_n\right)}\right] \tag{33}$$

This integral, however, remains essentially intractable, due to the product of Gaussian and Poisson-like terms. In high dimension there is no (to our knowledge) computationally efficient way to estimate this integral or its derivatives.

3 Expected log-likelihoods and variational Bayes

So far, we have explored three approaches to finding a Gaussian approximation to the measurement posterior: the Laplace approximation, variational Bayes, and moment matching. Moment matching is unsuitable because, to the best of our knowledge, there is no computationally tractable way to estimate the relevant moments in high-dimensions. The Laplace approximation and variational Bayes remain computationally tractable, with limitations. The Laplace approximation suffers from errors arising from the non-negativity constraint on activity levels, and the high skewness of the distributions causes the mode to be far from the mean. Errors in estimating the covariance are especially severe, as the covariance controls the trade-off between propagating past information, and incorporating new measurements, during the filtering procedure.

Variational Bayes also has a number of challenges. First, an efficient way of evaluating the relevant integrals and their derivatives is needed. In practice, this is simplest when we can tolerate the approximation of integrating over the full domain of the prior, including the unphysical negative activations. We must also use a rectifying link function, because if the predicted point-process intensity is negative, then the Poisson likelihood for our observations is not defined. To our knowledge, the only rectifying link function that has been explored to-date is the exponential link function, which suffers from unacceptable numerical instability in our application. There are a few other link functions that we might explore, but we will address another approximate solution in this section based on Laplace-approximated expected log-likelihoods.

In order to minimize D_{KL} in the variational Bayes approach, we must maximize the data log-likelihood under the approximating distribution Q, while simultaneously minimizing the KL divergence of the prior from this posterior approximation. Provided we interpret the multivariate Gaussian prior for $\mathbf{A}$ as having support over $(-\infty,\infty)^D$, the KL divergence term has a closed-form solution and well-defined derivatives. The challenge, then, is to calculate expected log-likelihood term:

$$\langle \log \Pr\left(\mathbf{Y} \mid \mathbf{A}\right)\rangle_Q \tag{34}$$

We now derive a general second-order approximation for the expected log-likelihood for a Gaussian latent-variable process with Poisson observations, where the latent variable may be linked to the Poisson intensity via an arbitrary link function. (See Zhou and Park, plus the expected log-likehood papers, for more detail). In the case of intractable $\langle \log \Pr\left(\mathbf{Y} \mid \mathbf{A}\right)\rangle_Q$, we approximate this integral via Laplace approximation. This yields and approximate variational inference method that is similar, but not identical, to the Laplace approximation.

3.1 Second-order approximations to the expected log-likelihood

To briefly review the notation, let $\mathbf{A} = \{A_1, .., A_N\}$ be a multivariate Gaussian latent variable reflecting our prior estimate for the distribution of latent activation, with mean μ_A and covariance Σ_A. Let $\lambda_n^0 = f(A_n)$ be an link function mapping the latent activity to a baseline intensity λ_n^0, which then might be further scaled and shifted due to spatially inhomogeneous gain γ_n or background activity β_n. We need to compute expected log-likelihoods under the approximating posterior distribution Q, that is:

$$\langle \mathcal{L}(y)n|A_n)\rangle = y_n \langle \log(\lambda_n)\rangle - \langle \lambda_n \rangle \tag{35}$$

Where $\langle\cdot\rangle$ denotes averaging over the posterior distribution $Q(\mathbf{A})$ with mean $\hat{\mu}_{\mathbf{A}}$ and covariance $\hat{\Sigma}_{\mathbf{A}}$. In certain cases, the expectations $\langle \log(\lambda_n)\rangle$ and $\langle \lambda_n \rangle$ may have closed-form solutions, for example in the log-Gaussian instance (cite Rule, Zhou). Here, however, we explore a general approach based on second-order Taylor expansions, which is accurate for small variances. If A_n is normally distributed with mean μ_{A_n} and variance ${\sigma_{A_n}}^2$, then out to second order:

$$\langle \mathcal{L}\left(y_n|A_n\right)\rangle \approx \mathcal{L}\left(y_n|\hat{\mu}_{A_n}\right) + \frac{\hat{\sigma}^2_{A_n}}{2}\mathcal{L}''\left(y_n|\hat{\mu}_{A_n}\right) \tag{36}$$

Gradient-based methods for optimizing the expected log-likelihood require derivatives of these approximated expectations. In general, derivatives with respect to the mean are:

$$\frac{d^n}{d\mu^m_{A_n}} \left\langle \mathcal{L}(y_{A_n}|A_n) \right\rangle \approx \mathcal{L}^{(m)}\left(y_n|\hat{\mu}_{A_n}\right) + \frac{\hat{\sigma}^2_{A_n}}{2}\mathcal{L}^{(m+2)}\left(y_n|\hat{\mu}_{A_n}\right) \tag{37}$$

A Newton-Raphson solver for optimizing the mean $\hat{\mu}_{\mathbf{A}}$ requires the Hessian of the objective function. Since the approximated expectations include second derivatives, the Hessian involves derivatives out to fourth order. The chain rule for higher-order derivatives of the logarithm is too cumbersome to state for the general case. Instead, we derive the equations for three versions of $f(A)$: A, A^2, and e^A. Optimizing the likelihood may also involve optimizing the variance σ^2 or in general, the covariance. We will address this in later sections.

3.1.1 For the case that $f(A) = A$

Interpreting the distribution of latent activations $\mathbf{A}$ as a multivariate Gaussian over $(-\infty, \infty)^D$ allows closed-form estimation of the D_{KL} contribution to the variational Bayes objective function. However, unless point-process intensities are artificially constrained to the domain $[0, \infty)$, the expected log-likelihood is undefined for the identity link function. This is because the Poisson measurement likelihood is not defined for negative intensities.

In this second-order approximation, we circumvent this issue by considering a locally-quadratic approximation of the likelihood function that continues the Poisson likelihood to negative intensities. Provided variance is small, and the posterior mean is constrained to be positive, this approximation may provide an accurate estimate of the expected log-likelihood. If $f(A) = A$ and so $\lambda = v \cdot \gamma \cdot (x + \beta/\gamma)$. Computing out to the 4th derivative.

$$\begin{aligned}
\mathcal{L}\left(y|A\right) &= y\log[\lambda] - \lambda \\
&= y\left[\log(v\cdot\gamma) + \log(\lambda/\gamma)\right] - v\cdot\gamma(x+\beta/\gamma) \\
&= y\log(\lambda/\gamma) - v\cdot\gamma A + \text{constant} \\
\mathcal{L}'\left(y|A\right) &= y(\lambda/\gamma)^{-1} - v\cdot\gamma \\
\mathcal{L}''\left(y|A\right) &= -y(\lambda/\gamma)^{-2} \\
\mathcal{L}^{(3)}\left(y|A\right) &= 2y(\lambda/\gamma)^{-3} \\
\mathcal{L}^{(4)}\left(y|A\right) &= -6y(\lambda/\gamma)^{-4}
\end{aligned} \tag{38}$$

As $\langle\lambda\rangle \longrightarrow 0$, the fourth derivative of the likelihood tends rapidly to infinity, which may create issues for numerical stability and accuracy. This behavior near $\langle\lambda\rangle \longrightarrow 0$ is similar to the issues that plague the Laplace approximation. In particular, the distribution may become highly skewed, which means that third or higher moments may be needed, and the second-order approximation may not be accurate. However, I have reason to suspect that the issues might be less severe for the expected log-likelihood compared to the Laplace approximation. In this case, we are using a quadratic expansion about an estimate of the posterior *mean*, whereas the Laplace approximation seeks the posterior *mode*. I expect that this will have a stabilizing effect, encouraging $\langle\lambda\rangle$ toward more positive values.

3.1.2 For the case that $f(A) = e^A$

A closed-form solution for the expected log-likelihood exists under this link function, and closed-form expressions for the moments of log-Gaussian random variables are known (see Zhou and Park for application to log-Gaussian point processes). However, in this application the amplification of positive tails of the distribution by the exponential link function is numerically unstable and unphysical, indicating that the log-Gaussian model is inappropriate. In Rule et al. 2018, we noted that a second-order approximation to the expected likelihood was more stable and more accurate.

$$\begin{aligned}
\mathcal{L}\,(y|A) &= y\log[\lambda] - \lambda \\
&= y\log[v\cdot\gamma(e^A+\beta/\gamma)] - v\cdot\gamma(e^A+\beta/\gamma) \\
&= y\log[e^A+\beta/\gamma] - v\cdot\gamma e^A + \text{constant} \\
\mathcal{L}^{(n)}\,(y|A) &= yC^{(n-1)} - v\cdot\gamma e^A, \\
C &= \frac{e^A}{e^A+\beta/\gamma} = \frac{1}{1+\beta/\gamma e^{-A}} \\
C' &= C(1-C) \\
C'' &= C'(1-2C) \\
C^{(3)} &= C' - 6C'^2
\end{aligned} \tag{39}$$

3.1.3 For the case that $f(A) = A^2$

A quadratic link function is rectifying, so the Poisson likelihood remains well-defined even if the domain of **A** is extended to $(-\infty,\infty)^D$. However, to my knowledge there is no tractable closed-form for the expectation of logarithm of a generalized noncentral χ^2 distributed variable, and so the second-order approximation remains useful:

$$\begin{aligned}
\mathcal{L}\,(y|A) &= y\log[\lambda] - \lambda \\
&= y\log[v\cdot\gamma(A^2+\beta/\gamma)] - v\cdot\gamma(A^2+\beta/\gamma) \\
&= y\log[A^2+\beta/\gamma] - v\gamma A^2 + \text{constant} \\
\mathcal{L}'\,(y|A) &= yC - 2v\gamma A, \qquad C = \frac{2A}{A^2+\beta/\gamma} \\
\mathcal{L}''\,(y|A) &= yC' - 2v\gamma, \qquad C' = C(C^{-1}-C) \\
\mathcal{L}^{(3)}\,(y|A) &= yC'', \qquad C'' = C'(A^{-1}-2C) - CA^{-2} \\
\mathcal{L}^{(4)}\,(y|A) &= yC^{(3)}, \qquad C^{(3)} = 3(C/A)^2 - 6C'^2
\end{aligned} \tag{40}$$

3.2 Incorporating the prior

So far we have focused on the expected log-likelihood contribution to the variational posterior objective function. We also need to derive the gradients of the KL-divergence term. For the most part, this is identical to the derivation in Zhou and Park, so I present only some quick notes here. We are interested in the gradients (and hessians) for the KL divergence between two k dimensional multivariate Gaussians, which is:

$$D_{KL}\left(\mathcal{N}_0 \| \mathcal{N}_1\right) = \frac{1}{2}\left[\operatorname{tr}(\Pi_1\Sigma_0) + (\mu_1 - \mu_0)^\top \Pi_1 (\mu_1 - \mu_0) - \ln|\Pi_1| - \ln|\Sigma_0| - k\right] \tag{41}$$

Note that I have chosen to denote this in terms of the precision matrix $\Pi_1 = \Sigma_1^{-1}$, as it makes some of the derivations below more straightforward. The derivative with respect to μ_0 is the same as the derivative with respect to μ_1 and is:

$$\nabla_{\mu_1} D_{KL}\left(\mathcal{N}_0 \| \mathcal{N}_1\right) = \nabla_{\mu_0} D_{KL}\left(\mathcal{N}_0 \| \mathcal{N}_1\right) = \frac{1}{2}\left[(\mu_1 - \mu_0)^\top \Pi_1\right] \tag{42}$$

The derivative with respect to Σ_0 is:

$$\nabla_{\Sigma_0} D_{KL}\left(\mathcal{N}_0 \| \mathcal{N}_1\right) = \tfrac{1}{2}\left[\Pi_1 - \Pi_0\right] \tag{43}$$

The derivative with respect to Π_1 is:

$$\nabla_{\Pi_1} D_{KL}\left(\mathcal{N}_0 \| \mathcal{N}_1\right) = \tfrac{1}{2}\left[\Sigma_0 - \Sigma_1 + (\mu_1 - \mu_0)(\mu_1 - \mu_0)^\top\right] \tag{44}$$

For the variational interpretation we approximate the posterior P with approximation Q and minimize $D_{KL}\left(Q \| P\right)$. This involves taking the derivative with respect to Σ_0 above.

3.3 Optimizing the (approximate) variational approximation

We have derived approximations for the expected log-likelihood contribution to the variational Bayesian objective function, which must be optimized jointly over the posterior mean $\hat{\mu}_A$ and the posterior covariance $\hat{\Sigma}_A$. The above derivations provide gradients and Hessians for optimizing $\hat{\mu}_A$ for a fixed $\hat{\Sigma}_A$. In Zhou and Park, they explore the joint optimization for the (exact) objective function for a log-Gaussian variational approximation. They prove that $\hat{\Sigma}_A$ can be optimized using a fixed-point iteration.

In numerical experiments, I extended this approach by interleaving one-step of the Newton-Raphson optimization for $\hat{\mu}_A$ with one step of the fixed-point update for $\hat{\Sigma}_A$. In my experience this accelerated convergence. Does the fixed-point iteration for $\hat{\Sigma}_A$ convergs for the second-oder approximated expected log-likelihood? Could a similar approach be found for the other (non-exponential) link functions explored here.

3.3.1 The covariance update

To complete the variational approximation, we also need to optimize the posterior covariance. This involves the derivative of the expected log-likelihood with respect to Σ_0. In the second-order (Laplace-approximation-like) expected log-likelihood, the dependence on the covariance enters through the second-order terms, which are:

$$\frac{1}{2}\operatorname{diag}(\Sigma_0)\,\mathcal{L}''\,(y|\mu_0) \tag{45}$$

The derivative of the above with respect to Σ_0 is:

$$\frac{1}{2}\mathcal{L}''\,(y|\mu_0) \tag{46}$$

along the diagonal, and 0 elsewhere. The total gradient for Σ_0, incorporating both the D_{KL} and expected log-likelihood contributions, is:

$$\nabla_{\Sigma_0}\mathcal{L}(\text{A}|\text{Y}) = \frac{1}{2}\left[\Pi_1 - \Pi_0 + \mathcal{L}''\,(y|\mu_0)\right] \tag{47}$$

Setting this gradient to zero and solving for Σ_0 gives:

$$\Pi_0 = \Pi_1 + \mathcal{L}''\,(y|\mu) \tag{48}$$

Which amounts to adding the curvature of the log likelihood $\mathcal{L}''\,(y|\mu)$ (approximated as the curvature at the current posterior mean), to the prior precision matrix Π_1. This is similar to the Hessian observed for the Laplace update, with the exception that we use the curvature at the estimated posterior mean, rather than posterior mode.

4 Computing the model likelihood

Once the Gaussian approximation is computed, how should we estimate the *likelihood* of the data given the model parameters θ, $\Pr(\mathbf{Y}|\theta)$ (or just $\Pr(\mathbf{Y})$ for short)? There are numerous approximation methods available, and it remains unclear to me which is best.

4.1 Integration via Laplace approximation

This likelihood is the integral of the prior $\Pr(\mathbf{A})$ times the measurement likelihood $\Pr(\mathbf{Y}|\mathbf{A})$, and is also the normalization constant for the posterior distribution:

$$\Pr(\mathbf{Y}) = \int_{\mathbf{A}} \Pr(\mathbf{Y}|\mathbf{A})\Pr(\mathbf{A}) = \langle \Pr(\mathbf{Y}|\mathbf{A})\rangle_{\Pr(\mathbf{A})} \tag{49}$$

Given a Gaussian approximation Q for the posterior $\Pr(\mathbf{A}|\mathbf{Y})$, we can approximate this integral using the posterior mean $\hat{\mu}_A$ and covariance $\hat{\Sigma}_A$. This Gaussian approximation can be obtained from any of the previously mentioned approximation methods, Laplace approximation, or exact or approximated variational inference.

Under this approximation, we evaluate $\Pr(\mathbf{Y}|\mathbf{A})\Pr(\mathbf{A})$ at $\hat{\mu}_A$, and also compute the curvature at this point *(which should match $\hat{\Sigma}_A$ if everything has gone as planned!).*

$$\log \Pr(\mathbf{A}|\mathbf{Y}) \approx -\tfrac{1}{2}\left[\log|2\pi\hat{\Sigma}_A| + (\mathbf{A}-\hat{\mu}_a)^\top \hat{\Sigma}_A^{-1}(\mathbf{A}-\hat{\mu}_a)\right] \tag{50}$$

We use the following logarithmic relationship derived from Bayes' rule

$$\log \Pr(\mathbf{Y}) = \log\left[\Pr(\mathbf{A})\Pr(\mathbf{Y}|\mathbf{A})\right] - \log \Pr(\mathbf{A}|\mathbf{Y}) \tag{51}$$

Substituting the forms for the above

$$\begin{aligned}\log \Pr(\mathbf{Y}) \approx& -\tfrac{1}{2}\left[\log|2\pi\Sigma_A| + (\mathbf{A}-\mu_A)^\top \Sigma_A^{-1}(\mathbf{A}-\mu_A)\right]\\ &+ \sum_{n\in 1..N}\left[y_n\log(\gamma_n f(A_n)+\beta_n) - (\gamma_n f(A_n)+\beta_n)\right]\\ &+ \tfrac{1}{2}\left[\log|2\pi\hat{\Sigma}_A| + (\mathbf{A}-\hat{\mu}_A)^\top \hat{\Sigma}_A^{-1}(\mathbf{A}-\hat{\mu}_A)\right]\end{aligned} \tag{52}$$

Evaluating the above at the posterior mean $\hat{\mu}_A$ amounts to a Laplace approximation of the integral for the likelihood.

$$\begin{aligned}\log \Pr(\mathbf{Y}) \approx& \log\left[\Pr(\hat{\mu}_A)\Pr(\mathbf{Y}|\hat{\mu}_A)\right] - \log \Pr(\hat{\mu}_A|\mathbf{Y})\\ =& -\tfrac{1}{2}\left[\log|2\pi\Sigma_A| + (\hat{\mu}_A-\mu_A)^\top \Sigma_A^{-1}(\hat{\mu}_A-\mu_A)\right]\\ &+ \sum_{n\in 1..N}\left[y_n\log(\gamma_n f(\hat{\mu}_{A_n})+\beta_n) - (\gamma_n f(\hat{\mu}_{A_n})+\beta_n)\right]\\ &+ \tfrac{1}{2}\left[\log|2\pi\hat{\Sigma}_A| + (\hat{\mu}_A-\hat{\mu}_A)^\top \hat{\Sigma}_A^{-1}(\hat{\mu}_A-\hat{\mu}_A)\right]\end{aligned} \tag{53}$$

Cleaning things up, and denoting $\hat{\lambda}_n = \gamma_n f(\hat{\mu}_{A_n}) + \beta_n$, we get:

$$\begin{aligned}\log \Pr(\mathbf{Y}) \approx& -\tfrac{1}{2}\left[\log\frac{|\Sigma_A|}{|\hat{\Sigma}_A|} + (\hat{\mu}_A-\mu_A)^\top \Sigma_A^{-1}(\hat{\mu}_A-\mu_A)\right]\\ &+ \sum_{n\in 1..N}\left[y_n\log(\hat{\lambda}_n) - \hat{\lambda}_n\right]\end{aligned} \tag{54}$$

4.1.1 On the similarity between the Laplace-approximated likelihood and D_{KL}

Note the similarity to KL divergence

$$D_{\mathrm{KL}}(Q\|\Pr(\mathbf{A})) = \frac{1}{2}\left[\log\frac{|\Sigma_A|}{|\hat{\Sigma}_A|} - D + \operatorname{tr}\left[\Sigma_A^{-1}\hat{\Sigma}_A\right] + (\mu_A-\hat{\mu}_A)^T\Sigma_A^{-1}(\mu_A-\hat{\mu}_A)\right] \tag{55}$$

Which gives the relation

$$-\frac{1}{2}\left[\log\frac{|\Sigma_A|}{|\hat{\Sigma}_A|} + (\mu_A-\hat{\mu}_A)^T\Sigma_A^{-1}(\mu_A-\hat{\mu}_A)\right] = \frac{1}{2}\left[\operatorname{tr}\left[\Sigma_A^{-1}\hat{\Sigma}_A\right] - D\right] - D_{\mathrm{KL}}(Q\|\Pr(\mathbf{A})) \tag{56}$$

Which allows the likelihood to be written as

$$
\begin{aligned}
\log \Pr(\mathbf{Y}) \approx{}& \frac{1}{2}\left[\operatorname{tr}\left[\Sigma_A^{-1}\hat{\Sigma}_A\right] - D\right] - D_{\mathrm{KL}}(Q \| \Pr(\mathbf{A})) \\
&+ \sum_{n \in 1..N} \left[y_n \log(\hat{\lambda}_n) - \hat{\lambda}_n\right]
\end{aligned}
\tag{57}
$$

We can play with this further, bringing in the second-order expected log-likelihood:

$$
\mathcal{L}\left(y_n|\hat{\mu}_{A_n}\right) = \sum_{n \in 1..N} \left[y_n \log(\hat{\lambda}_n) - \hat{\lambda}_n\right]
\tag{58}
$$

Which, from the second-order approximation, is:

$$
\mathcal{L}\left(y_n|\hat{\mu}_{A_n}\right) \approx \left\langle \mathcal{L}\left(y_n|\hat{\mu}_{A_n}\right)\right\rangle - \frac{\hat{\sigma}^2_{A_n}}{2}\mathcal{L}''\left(y_n|\hat{\mu}_{A_n}\right)
\tag{59}
$$

So...

$$
\begin{aligned}
\log \Pr(\mathbf{Y}) \approx{}& \frac{1}{2}\left[\operatorname{tr}\left[\Sigma_A^{-1}\hat{\Sigma}_A\right] - D\right] - D_{\mathrm{KL}}(Q \| \Pr(\mathbf{A})) \\
&+ \langle \mathcal{L}\left(y|\hat{\mu}_A\right)\rangle - \sum_{n \in 1..N} \frac{\hat{\sigma}^2_{A_n}}{2}\mathcal{L}''\left(y_n|\hat{\mu}_{A_n}\right) \\
&= -D_{\mathrm{KL}}(Q \| \Pr(\mathbf{A}|\mathbf{Y})) + \frac{1}{2}\operatorname{tr}\left(\Sigma_A^{-1}\hat{\Sigma}_A\right) - \sum_{n \in 1..N} \frac{\hat{\sigma}^2_{A_n}}{2}\mathcal{L}''\left(y_n|\hat{\mu}_{A_n}\right) \\
&+ \text{constant}
\end{aligned}
\tag{60}
$$

Which is to say that a point estimate of the log-likelihood is *very* similar to the (negative) KL-divergence penalty, differing only by the trace term and the curvature correction (and the dimensionality constant D).

Empirically, these terms are small and do not dominate the likelihood. As we shall see in the following sections, this similarity is not a coincidence: the Laplace approximation is connected to the Evidence Lower Bound (ELBO) in the variational Bayesian approach, especially when a second-order approximation is used to evaluate the expected log-likelihood.

4.2 The expected log-likelihood

One can consider the expected value of the log-likelihood relative to the prior distribution $\Pr(\mathbf{A})$. Starting from the logarithmic form of Bayes' rule, we have:

$$
\log \Pr(\mathbf{Y}) = \log \Pr(\mathbf{A}) + \log \Pr(\mathbf{Y}|\mathbf{A}) - \log \Pr(\mathbf{A}|\mathbf{Y})
\tag{61}
$$

For the true posterior $\Pr(\mathbf{A}|\mathbf{Y})$, this equality holds for all $\mathbf{A}$, as we could recover the log-likelihood by evaluating this expression at any point. In the Laplace approximation, we evaluated this quantity at the posterior mean. For the expected log-likelihood approach here, we take the expectation with respect to the prior distribution $\Pr(\mathbf{A})$:

$$\begin{aligned} \langle \log P_Y \rangle_{P_A} &= \langle \log P_A \rangle_{P_A} + \left\langle \log P_{Y|A} \right\rangle_{P_A} - \left\langle \log P_{A|Y} \right\rangle_{P_A} \\ &= \left\langle \log P_{Y|A} \right\rangle_{P_A} + D_{\mathrm{KL}}(P_A \| P_{A|Y}) \end{aligned} \tag{62}$$

We cannot compute the above exactly, because we do not have access to the true posterior $P_{A|Y}$, but we do have access to an approximating posterior $Q{\approx}P_{A|Y}$, which we can use to approximate the expected log-likelihood. Note that the additional D_{LK} term *increases* the expected log likelihood, opposite of its role in the variational approach.

$$\langle \log P_Y \rangle_{P_A} \approx \left\langle \log P_{Y|A} \right\rangle_{P_A} + D_{\mathrm{KL}}(P_A \| Q) \tag{63}$$

To estimate the expected log-likelihood term $\left\langle \log P_{Y|A} \right\rangle_{P_A}$, we may either use the second-order approach that we derived for variational Bayes', or simply evaluate $\log P_{Y|A}$ at the prior mean for a faster approximation.

4.3 ELBO and variational Bayes

When deriving the variational update, we omitted the (constant) data likelihood term. The full form for $D_{\mathrm{KL}}(Q\|P)$ is:

$$\begin{aligned} D_{\mathrm{KL}}(Q\|P) = \int_A Q_A \log \frac{Q_A}{P_{A|Y}} &= \langle \log Q \rangle_Q - \left\langle \log P_{A|Y} \right\rangle_Q \\ &= -H_Q - \left\langle \log P_{Y,A} \right\rangle_Q + \log P_Y \\ &= - \left\langle \log P_{Y|A} \right\rangle_Q + D_{\mathrm{KL}}(Q \| P_A) + \log P_Y, \end{aligned} \tag{64}$$

which implies the following identity for the log-likelihood:

$$\begin{aligned} \log P_Y &= D_{\mathrm{KL}}(Q\|P) + \left\langle \log P_{Y|A} \right\rangle_Q - D_{\mathrm{KL}}(Q \| P_A) \\ &= D_{\mathrm{KL}}(Q\|P) + H_Q + \left\langle \log P_{Y,A} \right\rangle_Q . \end{aligned} \tag{65}$$

That is: the data likelihood is the expected log-likelihood under the variational posterior, minus the KL divergence of the posterior form the prior, plus the KL divergence of the variational posterior from the *true* posterior. Since this last term is always positive, the following bound holds:

$$\begin{aligned} \log P_Y &\geq \left\langle \log P_{Y|A} \right\rangle_Q - D_{\mathrm{KL}}(Q \| P_A) \\ &= H_Q + \left\langle \log P_{Y,A} \right\rangle_Q \end{aligned} \tag{66}$$

This inequality is sometimes called the Evidence Lower Bound (ELBO). The more accurately we can approximate the posterior as a Gaussian, the smaller $D_{\mathrm{KL}}(Q\|P)$ becomes and the tight this bound becomes.

$$ \tag{67}$$

5 Going forward

In practice, we use the Laplace approximation for obtaining the approximate posterior Q. This avoids needing to jointly optimize the posterior covariance, trading accuracy for speed. We have explore the Laplace, ELBO, and expected log-likelihood approaches to estimating the likelihood, and found them to be broadly similar. Note, however, that the accuracy of expressions involving second-order expansions is expected to break down if the variance of the latent state becomes large.