Gradient and Hessian for variational inference in Poisson and probit Generalized Linear Models

M. Rule

March 30, 2020

These notes contain some derivations for variational inference in Poisson and probit Generalized Linear Models (GLMs) with a Gaussian prior and approximated Gaussian posterior. (see also here.)

0.0.1 Problem statement

Consider a population of neurons with firing-intensities $\boldsymbol{\lambda} = \rho(\boldsymbol{\theta})$, where $\rho(\cdot)$ is a firing-rate nonlinearity and $\boldsymbol{\theta}$ is a vector of synaptic activations (amount of input drive to each neuron). For stochastic models of spiking $\Pr(y|\theta)$ in the canonical exponential family, the probability of observing spikes $\mathbf{y}$ given $\boldsymbol{\theta}$ can be written as

$$\ln \Pr(\mathbf{y}|\mathbf{z}) = \mathbf{y}^\top \boldsymbol{\theta} - \mathbf{1}^\top A(\boldsymbol{\theta}) + \text{constant}, \tag{1}$$

where $A(x)$ is a known function whose derivative equals the firing-rate nonlinarity, i.e. $A'(\cdot) = \rho(\cdot)$.

Assume that the synaptic activations $\boldsymbol{\theta}$ are driven by shared latent variables $\mathbf{z}$ with a Gaussian prior $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \Sigma_z)$. Let $\boldsymbol{\theta} = \mathbf{B}\mathbf{z}$, where "$\mathbf{B}$" is a matrix of coupling coefficients which determine how the latent factors $\mathbf{z}$ drive each neuron.

We want to infer the distribution of $\mathbf{z}$ from observed spikes $\mathbf{y}$. The posterior is given by Bayes rule, $\Pr(\mathbf{z}|\mathbf{y}) = \Pr(\mathbf{y}|\mathbf{z})\Pr(\mathbf{z})/\Pr(\mathbf{y})$. However, this posterior does not admit a closed form if $A(\cdot)$ is nonlinear. Instead, one can use a variational Bayesian approach to obtain an approximate posterior.

0.0.2 Variational Bayes

In variational Bayes, the posterior on z is approximated as Gaussian, i.e. $\Pr(\mathbf{z}|\mathbf{y}) \approx Q(\mathbf{z})$, where $Q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)$. We then optimize $\boldsymbol{\mu}_q$ and Σ_q to minimize the Kullback-Leibler (KL) divergence from the true posterior $\Pr(\mathbf{z}|\mathbf{y})$ to $Q(\mathbf{z})$. This is equivalent to minimizing the KL divergenece $D_{\text{KL}}\left[Q(\mathbf{z}) \| \Pr(\mathbf{z})\right]$ from the prior to the posterior, while maximizing the expected log-likelihood $\langle \Pr(\mathbf{y}|\mathbf{z})\rangle$:

$$D_{\text{KL}}\left[Q(\mathbf{z}) \| \Pr(\mathbf{z}|\mathbf{y})\right] = D_{\text{KL}}\left[Q(\mathbf{z}) \| \Pr(\mathbf{z})\right] - \langle \ln \Pr(\mathbf{y}|\mathbf{z})\rangle + \text{constant}. \tag{2}$$

(In these notes, all expectations $\langle\cdot\rangle$ are taken with respect to the approximating posterior distribution.)

Since both $Q(\mathbf{z})$ and $\Pr(\mathbf{z})$ are multivariate Gaussian, the KL divergence $D_{\text{KL}}\left[Q(\mathbf{z}) \| \Pr(\mathbf{z})\right]$ has the closed form:

$$D_{\text{KL}}\left[Q(\mathbf{z}) \| \Pr(\mathbf{z})\right] = \tfrac{1}{2}\left\{(\boldsymbol{\mu}_z - \boldsymbol{\mu}_q)^\top \Sigma_z^{-1}(\boldsymbol{\mu}_z - \boldsymbol{\mu}_q) + \operatorname{tr}\left(\Sigma_z^{-1}\Sigma_q\right) + \ln\left|\Sigma_z^{-1}\Sigma_q\right|\right\} + \text{constant}. \tag{3}$$

For our choice of the canonically-parameterized natural exponential family, the expected negative log-likelihood can be written as:

$$-\langle \ln \Pr(\mathbf{y}|\mathbf{z})\rangle = \mathbf{1}^\top\langle A(\boldsymbol{\theta})\rangle - \mathbf{y}^\top \mathbf{B}\boldsymbol{\mu}_q + \text{constant}. \tag{4}$$

Neglecting constants and terms that do not depend on $(\boldsymbol{\mu}_q, \Sigma_q)$, the overall loss function to be minimized is:

$$\mathcal{L}(\boldsymbol{\mu}_q, \Sigma_q) = \tfrac{1}{2}\left\{(\boldsymbol{\mu}_z - \boldsymbol{\mu}_q)^\top \Sigma_z^{-1}(\boldsymbol{\mu}_z - \boldsymbol{\mu}_q) + \operatorname{tr}\left(\Sigma_z^{-1}\Sigma_q\right) + \ln\left|\Sigma_z^{-1}\Sigma_q\right|\right\} + \mathbf{1}^\top\langle A(\boldsymbol{\theta})\rangle - \mathbf{y}^\top \mathbf{B}\boldsymbol{\mu}_q \quad . \tag{5}$$

0.0.3 Closed-form expectations

To optimize (5), we need to differentiate it in $\boldsymbol{\mu}_q$ and Σ_q. These derivatives are mostly straightforward, but the expectation $\langle A(\boldsymbol{\theta})\rangle$ poses difficulties when $A(\cdot)$ is nonlinear. We'll consider some choices of firing-rate nonlinearity for which the derivatives of $\langle A(\boldsymbol{\theta})\rangle$ have closed-form expressions when $\boldsymbol{\theta}$ is Gaussian.

Because we've assumed a Gaussian posterior on our latent state $\mathbf{z}$, and since $\boldsymbol{\theta} = \mathbf{B}\mathbf{z}$, the synaptic activations $\boldsymbol{\theta}$ are also Gaussian. The vectors $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\sigma}_\theta^2$ for the mean and variance of $\boldsymbol{\theta}$, respectively, are:

$$\begin{aligned} \boldsymbol{\mu}_\theta &= \mathbf{B}\boldsymbol{\mu}_q \\ \boldsymbol{\sigma}_\theta^2 &= \operatorname{diag}\left[\mathbf{B}\Sigma_q\mathbf{B}^\top\right] \end{aligned} \tag{6}$$

Consider a single, scalar $\theta \sim \mathcal{N}(\mu, \sigma^2)$. Using the chain rule and linearity of expectation, one can show that the partial derivatives $\partial_\mu\langle A(\theta)\rangle$ and $\partial_{\sigma^2}\langle A(\theta)\rangle$, with respect to μ and σ^2 respectively, are:

$$\begin{aligned} \partial_\mu\langle A(\theta)\rangle &= \langle A'(\theta)\rangle = \langle \rho(\theta)\rangle \\ \partial_{\sigma^2}\langle A(\theta)\rangle &= \tfrac{1}{2\sigma^2}\left\langle(\theta - \mu_\theta)A'(\theta)\right\rangle = \tfrac{1}{2\sigma^2}\left\langle(\theta - \mu)\rho(\theta)\right\rangle. \end{aligned} \tag{7}$$

For more compact notation, denote the expected firing rate as $\bar{\lambda} = \langle\rho(\theta)\rangle$, and denote the expected derivative of the firing-rate in θ as $\bar{\lambda}' = \langle\rho'(\theta)\rangle$. Note that $\bar{\lambda} = \partial_\mu\langle A(\theta)\rangle$ and $\frac{1}{2}\bar{\lambda}' = \partial_{\sigma^2}\langle A(\theta)\rangle$.

Closed-form expressions for $\bar{\lambda}$ and $\bar{\lambda}'$ exist only in some special cases, for example if the firing-rate function $\rho(\cdot)$ is a (rectified) polynomial. We consider two choices of firing-rate nonlinearity which admit closed-form expressions, "exponential" and "probit".

-Choosing $\rho = \exp$ corresponds to a Poisson GLM. In this case, $\bar{\lambda} = \bar{\lambda}' = \exp(\mu + \sigma^2/2)$. -Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution function, respectively, for a standard normal distribution. Choosing $\rho = \Phi$ corresponds to a probit GLM. In this case, $\bar{\lambda} = \Phi(\gamma\mu)$ and $\bar{\lambda}' = \gamma\phi(\gamma\mu)$, where $\gamma = (1 + \sigma^2)^{-1}$.

For the probit firing-rate nonlinearity, we will also need to know $\partial_{\sigma^2}\langle\rho'(\boldsymbol{\theta})\rangle$ to calculate the Hessian-vector product. In this case, $\rho' = \phi$. We have from (7) that $\partial_{\sigma^2}\langle\phi(x)\rangle = \frac{1}{2\sigma^2}\langle\theta(\mu - \theta)\phi(\theta)\rangle$. This can be solved by writing the expectation as an integral and completing the square in the resulting Gaussian integral, yielding:

$$\partial_{\sigma^2}\langle\phi(x)\rangle = \frac{u - 1}{\sqrt{8\pi e^u(1 + \sigma^2)^3}}, \text{ where } u = \frac{\mu^2}{\sigma^2 + 1}. \tag{8}$$

0.0.4 Derivatives of the loss function

With these prelimenaries out of the way, we can now consider the derivatives of (5) in terms of $\boldsymbol{\mu}_q$ and Σ_q.

Derivatives in $\boldsymbol{\mu}_q$ The gradient and Hessian of $\mathcal{L}$ with respect to $\boldsymbol{\mu}_q$ are:

$$\begin{aligned} \partial_{\mu_q}\mathcal{L} &= \Sigma_z^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_z) + \mathbf{B}^\top\left(\bar{\boldsymbol{\lambda}} - \mathrm{y}\right) \\ \mathrm{H}_{\mu_q}\mathcal{L} &= \Sigma_z^{-1} + \mathbf{B}^\top \operatorname{diag}[\bar{\boldsymbol{\lambda}}']\mathbf{B} \end{aligned} \tag{9}$$

Gradient in Σ_q The gradient of (5) in Σ_q is more involved. The derivative of the term $\frac{1}{2}\{\operatorname{tr}(\Sigma_z^{-1}\Sigma_q) + \ln|\Sigma_z^{-1}\Sigma_q|\}$ can be obtained using identities provided in The Matrix Cookbook. The derivative of $\mathbf{1}^\top\langle A(\boldsymbol{\theta})\rangle$ can be obtained by considering derivatives with respect to individual elements of Σ_q, and is $\frac{1}{2}\mathbf{B}^\top \operatorname{diag}[\bar{\lambda}']\mathbf{B}$. Overall, we find that:

$$\partial_{\Sigma_q}\mathcal{L} = \frac{1}{2}\left\{\Sigma_z^{-1} + \Sigma_q^{-\top} + \mathbf{B}^\top \operatorname{diag}[\bar{\lambda}']\mathbf{B}\right\}. \tag{10}$$

Hessian-vector product in Σ_q Since Σ_q is a matrix, the Hessian of (5) in Σ_q is a fourth-order tensor. It is simpler to work with the Hessian-vector product. Here, the "vector" is a covariance matrix $\mathbf{M}$ to be optimized. The Hessian-vector product is given by the following identity:

$$\langle \mathbf{H}_{\Sigma_q}, \mathbf{M}\rangle = \partial_{\Sigma_q}\langle \mathbf{J}_{\Sigma_q}, \mathbf{M}\rangle = \partial_{\Sigma_q}\operatorname{tr}\left[\mathbf{J}_{\Sigma_q}^\top \mathbf{M}\right] \tag{11}$$

where $\langle\cdot,\cdot\rangle$ denotes the scalar (Frobenius) product. The Hessian-vector product for the terms $\Sigma_z^{-1} + \Sigma_q^{-\top}$ in (10) can be obtained using identities provided in The Matrix Cookbook:

$$\partial_{\Sigma_q}\operatorname{tr}\left[\left\{\Sigma_z^{-1} + \Sigma_q^{-\top}\right\}^\top \mathbf{M}\right] = -\Sigma_q^{-1}\mathbf{M}^\top\Sigma_q^{-1}. \tag{12}$$

The Hessian-vector product for the term $\mathbf{B}^\top \operatorname{diag}[\bar{\lambda}']\mathbf{B}$ in (10) is more complicated. We can write

$$\begin{aligned}\partial_{\Sigma_q}\operatorname{tr}\left[\left\{\mathbf{B}^\top \operatorname{diag}[\bar{\lambda}']\mathbf{B}\right\}^\top \mathbf{M}\right] &= \partial_{\Sigma_q}\operatorname{tr}\left[\mathbf{B}\mathbf{M}\mathbf{B}^\top \operatorname{diag}[\bar{\lambda}']\right] \\ &= \mathbf{B}^\top \operatorname{diag}[\mathbf{B}\mathbf{M}\mathbf{B}^\top]\operatorname{diag}\left[\partial_{\sigma^2_{\boldsymbol{\theta}}}\langle\rho'(\boldsymbol{\theta})\rangle\right]\mathbf{B}.\end{aligned} \tag{13}$$

The first step in (13) uses the fact that the trace is invariant under cyclic permutations. The second step follows from Lemma 1 (Appendix, below), with $\mathbf{C} = \mathbf{B}\mathbf{M}\mathbf{B}^\top$ and using the fact that $\bar{\lambda}' = \langle\rho'(\theta)\rangle$. In general, the Hessian-vector product in Σ_q is

$$\langle \mathbf{H}_{\Sigma_q}, \mathbf{M}\rangle = \frac{1}{2}\left\{-\Sigma_q^{-1}\mathbf{M}^\top\Sigma_q^{-1} + \mathbf{B}^\top \operatorname{diag}[\mathbf{B}\mathbf{M}\mathbf{B}^\top]\operatorname{diag}\left[\partial_{\sigma^2_{\boldsymbol{\theta}}}\langle\rho'(\boldsymbol{\theta})\rangle\right]\mathbf{B}\right\} \tag{14}$$

For the exponential firing-rate nonlinearity, $\partial_{\sigma^2_{\boldsymbol{\theta}}}\langle\rho'(\boldsymbol{\theta})\rangle = \frac{1}{2}\bar{\lambda}$. The solution for the probit firing-rate nonlinearity is given in (8).

0.0.5 Conclude

That's all for now! I'll need to integrate these with the various other derivations (e.g. see also here.).

0.0.6 Appendix

Lemma 1 (We use Einstein summation to simplify the notation)

$$\begin{aligned}
\partial_{\Sigma_{q,ij}} \operatorname{tr}\left[\mathbf{C} \operatorname{diag}\left[\langle f(\boldsymbol{\theta})\rangle\right]\right] &= \partial_{\Sigma_{q,ij}} \left[\mathbf{C} \operatorname{diag}\left[\langle f(\boldsymbol{\theta})\rangle\right]\right]_{kk} \\
&= \partial_{\Sigma_{q,ij}} \left[\mathbf{C}_{lm} \operatorname{diag}\left[\langle f(\boldsymbol{\theta})\rangle\right]_{mn}\right]_{kk} \\
&= \partial_{\Sigma_{q,ij}} \left[\mathbf{C}_{kk} \operatorname{diag}\left[\langle f(\boldsymbol{\theta})\rangle\right]_{k}\right] \\
&= \mathbf{C}_{kk} \langle \partial_{\Sigma_{q,ij}} f(\theta_k)\rangle \\
&= \mathbf{C}_{kk} \mathbf{B}^{\top}_{ik}\, \partial_{\sigma^2_{\theta}} \langle f(\theta_k)\rangle \mathbf{B}_{kj} \\
&= \mathbf{B}^{\top}_{ik} \mathbf{C}_{kk}\, \partial_{\sigma^2_{\theta}} \langle f(\theta_k)\rangle \mathbf{B}_{kj} \\
&= \left\{\mathbf{B}^{\top} \operatorname{diag}[\mathbf{C}] \operatorname{diag}\left[\partial_{\sigma^2_{\boldsymbol{\theta}}} \langle f(\boldsymbol{\theta})\rangle\right] \mathbf{B}\right\}_{ij}
\end{aligned} \tag{15}$$

$$\partial_{\Sigma_q} \operatorname{tr}\left[\mathbf{C} \operatorname{diag}\left[\langle f(\boldsymbol{\theta})\rangle\right]\right] = \mathbf{B}^{\top} \operatorname{diag}[\mathbf{C}] \operatorname{diag}\left[\partial_{\sigma^2_{\boldsymbol{\theta}}} \langle f(\boldsymbol{\theta})\rangle\right] \mathbf{B}$$