JMLR: Workshop and Conference Proceedings 30:1–17, 2015 ACML 2015

Expectation Propagation for Rectified Linear Poisson Regression

Young-Jun Ko YOUNGJUN.KO@EPFL.CH
École Polytechnique Fédérale de Lausanne, Switzerland

Matthias W. Seeger MSEEGER@GMAIL.COM

Abstract

The Poisson likelihood with rectified linear function as non-linearity is a physically plausible model to discribe the stochastic arrival process of photons or other particles at a detector. At low emission rates the discrete nature of this process leads to measurement noise that behaves very differently from additive white Gaussian noise. To address the intractable inference problem for such models, we present a novel efficient and robust Expectation Propagation algorithm entirely based on analytically tractable computations operating reliably in regimes where quadrature based implementations can fail. Full posterior inference therefore becomes an attractive alternative in areas generally dominated by methods of point estimation. Moreover, we discuss the rectified linear function in the context of other common non-linearities and identify situations where it can serve as a robust alternative.

Keywords: Expectation Propagation, Bayesian Poisson Regression, Cox Process, Poisson Denoising, Rectified Linear function

1. Introduction

Inhomogeneous Poisson processes with stochastic intensity functions, so called Cox process, have become an indispensable modeling framework to describe counts of random phenomena in various contexts. For example they are used to map the rate at which certain social, economical or ecological events occur in space and/or time (Vanhatalo et al., 2010; Diggle et al., 2013). In neuroscience, they motivate the Linear-Nonlinear-Poisson cascade model, widely used to describe neural responses to external stimuli (Pillow, 2007; Gerwinn et al., 2010; Park and Pillow, 2013; Park et al., 2014). Similar models have been applied to collaborative filtering tasks to understand user preferences from implicit feedback (Seeger and Bouchard, 2012; Ko and Khan, 2014). In image processing, the noise process in photon-limited acquisition scenarios, typical for astronomical- and biomedical imaging, is Poisson (Starck and Murtagh, 2002; Dupé et al., 2008; Carlavan and Blanc-Féraud, 2012), an observation that e.g. the Richardson-Lucy model for deconvolution is based on (Richardson, 1972; Lucy, 1974).

Common to all of these examples is the probabilistic setup using the Poisson likelihood in Eq. 1 to describe the generation of a vector of discrete observations $\boldsymbol{y} = [y_i]_{i=1,\dots,N} \in \mathbb{N}^N$.

$$P(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{N} \frac{1}{y_i!} \lambda_i^{y_i} e^{-\lambda_i} \tag{1}$$

	Exponential	Softplus	Rectified-Linear
$g(f)$	$\exp(f)$	$\log(1+\exp(f))$	$\max(0, f)$

Table 1: Typical non-linearities (see text)

The observations are independently sampled, given the latent intensities $\boldsymbol{\lambda} = [\lambda_i]_{i=1,\ldots,N} \in \mathbb{R}^N_{\geq 0}$, which are themselves considered to be stochastic quantities. They are parameterized as $\lambda_i = g(f_i)$, where $\boldsymbol{f} = [f_i]_{i=1,\ldots,N} \in \mathbb{R}^N$ is a real-valued multi-variate random variable and $g : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is a non-linear function that enforces non-negativity.

Several choices for g can be found in the literature, summarized e.g. in (Pillow, 2007), which we list in Table 1. In this work we are primarily concerned with studying the rectified-linear(RL) function. While clearly related, we avoid the terminology used for generalized linear models (McCullagh and Nelder, 1989) where the g^{-1} is referred to as the *link* function since the RL function is not invertible. Instead we adhere to terminology used in deep neural networks where the RL function as a replacement for sigmoidal non-linearities has sparked recent interest leading to several comparative studies (Glorot et al., 2011; Zeiler et al., 2013; Maas et al., 2013).

In all of the above examples posterior inference of the latent variable $\boldsymbol{f}$ is intractable and requires approximations due to the use of non-conjugate priors. Approximate inference in the presence of the RL function has, to the best of our knowledge, not been studied, in contrast to its differentiable alternatives, especially the exponential function. Thus, the problem that we address in this work is to devise a practical inference method for Poisson likelihood models with the RL function as non-linearity. We motivate the relevance of such models with two arguments: *robustness against outliers* and *physical plausibility* for certain applications.

Robustness against outliers. The non-linearity relates the outcome of the latent variable maps to the intensity of the Poisson process. We qualitatively illustrate the effect of the different non-linearities on the posterior mean intensity $\mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y}]$ in Figure 1, where we fitted Model 1 with a GP prior to the coal mining disaster dataset (Jarrett, 1979)[1]. The dataset consists of records of accidents over time, each of which is represented by a black line. Most notable is the different behavior in the high density area on the left, to which the exponential non-linearity responds strongest. While the exponential non-linearity was successfully used in various applications (Diggle et al., 2013; Vanhatalo et al., 2010; Gerwinn et al., 2007; Ko and Khan, 2014), it may have a potentially adverse effect on the robustness against outliers and thus could hurt generalization performance. This argument is not new, and was brought forward e.g. in the context of recommender systems (Seeger and Bouchard, 2012) and neuroscience (Park et al., 2014). In both cases the authors prefer the softplus non-linearity. Given that the RL function lower-bounds the softplus function and has the same asymptotic behavior, it is a viable alternative.

1. We took this example from (Vanhatalo et al., 2013), using the same setup, i.e. isotropic squared exponential kernel function and constant mean. Inference is done using Expectation Propagation. Hyper-parameters for mean and covariance function are learned. More information on the data can be found in Section 3.2.

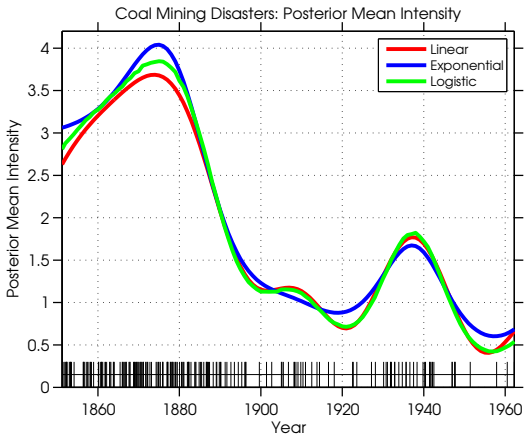


Figure 1: Coal Mining Disaster Data: posterior means of latent functions $\mathrm{E}\left[g(\boldsymbol{f})|\boldsymbol{y}\right]$. We recognize the stronger peaking behavior of the exponential non-linearity in high-density regions, while the other non-linearities are more sensitive in low density regions.

Physical Plausibility. Established models for Poisson noise, e.g. for image deconvolution (Richardson, 1972; Lucy, 1974), explicitly require the use of the RL non-linearity, because the intensity of the noise process is non-negative and relates linearly to the underlying image intensity. Thus, the softplus function is not a suitable option, because low intensities can only be achieved by pushing the corresponding image intensity f_i towards $-\infty$.

Although image reconstruction problems are typically addressed by point estimation, there are compelling arguments for full posterior inference: Apart from benefits such as uncertainty quantification and a principled way for hyper-parameter learning, a practical inference method may be necessary to tackle difficult high-level problems in this context, such as blind deconvolution. For this severely ill-posed problem, where neither the blur kernel nor the original image are known, (Levin et al., 2011) show that joint MAP estimation tends to lead to degenerate solutions which can be avoided by using the marginal likelihood for learning.

A major issue in tackling approximate inference is the non-differentiability of the log-posterior when using the RL function as well as due to the use of common image priors (Seeger, 2008; Seeger and Nickisch, 2011b), making many popular gradient-based methods such as Laplace's method or Variational Bayes inconvenient or impossible to apply (Gerwinn et al., 2007). We therefore chose the Expectation Propagation algorithm (Minka, 2001; Opper and Winther, 2000), known to gracefully deal with a much larger variety of models while delivering practical accuracy and performance (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). Its greater generality however can come at the cost of a numerically more challenging implementation. Meeting these challenges is at the heart of our contributions presented here, which we summarize as follows.

Contributions. We derive an Expectation Propagation algorithm for the Poisson model with the rectified linear function based entirely on analytically tractable computations. We fully characterize the tilted distribution, the central object of the EP algorithm, in terms of its moments and provide an efficient and robust formula to compute them. We

demonstrate that in comparison to a quadrature based implementation (a) our formulation is more efficient to compute and (b) can operate in regimes where quadrature experiences numerical instabilities. We conduct a series of experiments that corroborate the utility of using the identity link: On the coal mining data set we show that compared to the RL function, using the exponential function can be harmful in terms of generalization performance. On a deconvolution problem on natural images, where the use of other non-linearities led to numerical instabilities and convergence issues, we show that taking into account the correct noise model significantly reduces the reconstruction error.

This paper is outlined as follows. In Section 2, we review prior models and derive the EP algorithm for the RL model. In Section 3 we present experimental results and conclude in Section 4.

2. Inference for the Poisson Likelihood Model

Before proceeding to discuss inference methods for the Poisson likelihood model, we briefly introduce the relevant priors on the latent variable $\boldsymbol{f}$. Here, we consider two classes of prior distributions, that are commonly encountered in practice: Gaussian process (GP) priors (Rasmussen and Williams, 2005) and sparse linear models (SLM) (Seeger, 2008; Seeger and Nickisch, 2011b).

Gaussian Processes. GP priors prominently feature in applications of spatio-temporal statistics to social or ecological questions (Diggle et al., 2013; Vanhatalo et al., 2010), where this model is often referred to as Gaussian Cox process, or to the analysis of neural spike counts (Pillow, 2007; Park and Pillow, 2013; Park et al., 2014) as the multivariate normal is well suited to represent dependencies and dynamics in the input domain. We use the following notation: Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a latent function distributed as $f \sim \mathrm{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$, where $m(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}')$ denote the mean- and covariance functions. For N inputs $\{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1,\dots,N}$, the prior over $\boldsymbol{f}$ can be written as a multi-variate normal distribution:

$$P(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}) \tag{2}$$

with mean vector $\boldsymbol{m}_i = m(\boldsymbol{x}_i)$ and covariance- or kernel matrix $\boldsymbol{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for all pairs of inputs.

Sparse Linear Models. In SLMs $\boldsymbol{f}$ itself is defined as a linear function $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ of a latent vector $\boldsymbol{u}$, where $\boldsymbol{u}$ exposes non-Gaussian, heavy-tailed statistics in an appropriately chosen linear transform domain $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. SLMs are often encountered in the context of inverse problems, e.g. in image processing, where the prior belief that image gradients or Wavelet coefficients of natural images are sparse (Simoncelli, 1999), has become a popular strategy to regularize ill-posed reconstruction problems. For example, for the deconvolution problem we define the linear operator $\boldsymbol{X}$ such that multiplying it with a vectorized image $\boldsymbol{u}$ amounts to convolving the image with a blur kernel $\boldsymbol{h}$, i.e. $\boldsymbol{f} = \boldsymbol{h} * \boldsymbol{u} = \boldsymbol{X}\boldsymbol{u}$. Assuming that $\boldsymbol{u}$ is well described by piecewise constant functions, one could be interested in penalizing the total variation of $\boldsymbol{u}$, such that $\boldsymbol{B} = [\nabla_x^T \ \nabla_y^T]^T$ consists of the gradient operators in x

Non-Linearity	Prior	Laplace	VB	EP
Exponential	GP SLM	Tract. N/A	Tract.	Approx.
Softplus	GP SLM	Tract. N/A	Approx.	Approx.
Rect.-Lin.	GP SLM	Constr. N/A	N/A	Tract. (**NEW**)

Table 2: Variational Inference methods for different non-linearities. We use the following abbreviations: ***Tract.*:** Computations are analytically *Tractable* (i.e. gradients/updates available in closed form). ***Approx.*:** Computations require additional *Approximations*, such as bounding techniques or numerical integration. ***Constr.*:** Constrained optimization is required.

and y direction. We model sparsity for $\boldsymbol{s}$ independently for each transform coefficient:

$$P(\boldsymbol{u}) = \prod_{j}^{M} l(s_j) \tag{3}$$

For simplicity we consider the Laplace potential $l(s_j) = e^{-\tau|s_j|}$ (Gerwinn et al., 2007; Seeger, 2008; Seeger and Nickisch, 2011b).

Before we begin the discussion of methods for approximate inference, we unify our notation. We would like to approximate an intractable distribution of the following form:

$$P(\boldsymbol{f}) = Z^{-1} \prod_{j=1}^{M} t_j(f_j)\ t_0(\boldsymbol{f}) \tag{4}$$

For GPs the optional coupled potential $t_0(\boldsymbol{f})$ is the prior defined in Eq. 2, and we have a product of the $M = N$ likelihood potentials $t_j(f_j) = P(y_j|\lambda_j)$. For SLMs $t_0(\boldsymbol{f}) = 1$, and we redefine $\boldsymbol{f} = [\boldsymbol{X}^T,\ \boldsymbol{B}^T]^T\boldsymbol{u}$. The potentials are $t_j(f_j) = P(y_j|\lambda_j)$ for $j \leq N$ and $t_j(f_j) = l(f_j)$ for $j > N$. We denote the approximation to $P(\boldsymbol{f})$ by $Q(\boldsymbol{f})$. Here, we seek to choose $Q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from the family of multivariate normal densities. This is justified by the fact that the likelihood for all non-linearities in Table 1 as well as the priors mentioned here are log-concave in $\boldsymbol{f}$, thus leading to a unimodal posterior (Paninski, 2004).

In Table 2 we list common variational inference techniques to find the parameters of the approximation. While Laplace's method is the preferred method in the GP setting (Park and Pillow, 2013; Diggle et al., 2013; Park et al., 2014), it cannot be applied to SLMs, because by design we expect many transform coefficients to be zero, where the Laplace potential is not differentiable. Another popular variational Bayesian (VB) technique is referred to as Variational Gaussian approximation (Opper and Archambeau, 2009) or KL method (Nickisch and Rasmussen, 2008; Challis and Barber, 2013). It is analytically tractable for the exponential function (Ko and Khan, 2014), whereas the softplus function requires approximations, e.g. quadrature or bounding techniques, as shown in (Seeger and Bouchard, 2012). For the RL function, however, this method is not even defined. This can be seen by

examining the VB objective which is the following Kullback-Leibler divergence:

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathrm{D}_{\mathrm{KL}}\left[Q(\boldsymbol{f}) \,||\, P(\boldsymbol{f})\right] \tag{5}$$

Expanding it as usual reveals that the logarithm of Eq. 1 needs to be integrated over the real line, which is infinite in case of the RL function:

$$\mathrm{D}_{\mathrm{KL}}\left[Q(\boldsymbol{f}) \,||\, P(\boldsymbol{f})\right] = \mathrm{E}_Q\left[\log \frac{Q(\boldsymbol{f})}{t_0(\boldsymbol{f})}\right] - \sum_{j=1}^{M} \mathrm{E}_Q\left[\log t_j(f_j)\right] \tag{6}$$

Next, we introduce Expectation Propagation which applies in spite of constrained and non-differentiable potentials.

2.1. Expectation Propagation

For this introduction we adopt the perspective and the notation of and refer to (Rasmussen and Williams, 2005) for a more detailed introduction. EP (Minka, 2001; Opper and Winther, 2000) approximates $P(\boldsymbol{f})$ in Eq. 4 by approximating each non-Gaussian potential $t_j(f_j)$ using unnormalized Gaussians $\tilde{t}_j(f_j) = \tilde{Z}_j \mathcal{N}(f_j|\tilde{\mu}_j, \tilde{\sigma}_j^2)$ to form a Gaussian approximation $Q(\boldsymbol{f})$ following the same factorization:

$$Q(\boldsymbol{f}) = Z_{\mathrm{EP}}^{-1} \prod_{j=1}^{M} \tilde{t}_j(f_j)\, t_0(\boldsymbol{f}) \tag{7}$$

The EP-approximation to the marginal likelihood is given by:

$$Z_{\mathrm{EP}} = \prod_{j=1}^{M} \tilde{Z}_j \int \prod_{j=1}^{M} \mathcal{N}(f_j|\tilde{\mu}_j, \tilde{\sigma}_j^2)\, t_0(\boldsymbol{f})\, \mathrm{d}\boldsymbol{f} \tag{8}$$

EP was devised to address the shortcomings of the assumed density filtering (ADF) method and can be motivated by and in special cases shown to minimize the KL-divergence $\mathrm{D}_{\mathrm{KL}}\left[P(\boldsymbol{f}) \,||\, Q(\boldsymbol{f})\right]$ (Minka, 2001). Note the order of the arguments in contrast to Eq. 6. Since this quantity is generally intractable, EP employs the following strategy to determine the variational parameters $\tilde{\mu}_j, \tilde{\sigma}_j^2$.

We define the i-th marginal *cavity* distribution by removing the i-th *approximate* potential $\tilde{t}_i(f_i)$ from $Q(\boldsymbol{f})$ and marginalizing over $\{f_j : j \neq i\}$, denoted as $\boldsymbol{f}_{\setminus i}$:

$$Q_{-i}(f_i) = \mathcal{N}\left(f_i|\mu_{-i}, \sigma_{-i}^2\right) \propto \int \prod_{j\neq i} \tilde{t}_j(f_j)\, t_0(\boldsymbol{f})\, \mathrm{d}\boldsymbol{f}_{\setminus i} \tag{9}$$

The so called *tilted* distribution replaces the approximate potential $\tilde{t}_i(f_i)$ in $Q(\boldsymbol{f})$ by the true non-Gaussian potential $t_i(f_i)$ by multiplying it with the cavity marginal:

$$\hat{P}(f_i) = \hat{Z}_i^{-1}\, t_i(f_i)\, Q_{-i}(f_i) \qquad \text{where} \quad \hat{Z}_i = \int t_i(f_i)\, Q_{-i}(f_i)\, \mathrm{d}f_i \tag{10}$$

The criterion to minimize in order to update the parameters of $\tilde{t}_i$ is the KL-divergence between the tilted- and the variational distribution $\mathrm{D}_{\mathrm{KL}}\left[\hat{P}(\boldsymbol{f}) \,||\, Q(\boldsymbol{f})\right]$. This operation can be shown to be expressed in terms of the following moment matching condition:

$$\mathrm{E}_Q\left[f_i\right] = \mathrm{E}_{\hat{P}}\left[f_i\right] \qquad \mathrm{Var}_Q\left[f_i\right] = \mathrm{Var}_{\hat{P}}\left[f_i\right] \tag{11}$$

The constant $\tilde{Z}_i$ is chosen such that the normalization constants of $\hat{P}(f_i)$ and $Q(f_i)$ match, i.e. we solve:

$$\tilde{Z}_i \int \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)\, Q_{-}(f_i)\, \mathrm{d}f_i = \hat{Z}_i \tag{12}$$

The EP update therefore consists of determining the first two moments and the normalization constant of the tilted distribution.

Once the parameters of a single $\tilde{t}_j$ are changed, we can update the representation of the full approximation $Q(\boldsymbol{f})$, which typically consists of $\boldsymbol{\mu}$ and $\mathrm{Var}_Q\left[\boldsymbol{f}\right]$. This process is repeated until convergence, i.e. until moment matching is achieved globally.

The update of $Q(\boldsymbol{f})$, in particular obtaining $\mathrm{Var}_Q[\boldsymbol{f}]$ dominates the algorithm computationally, due to cubic scaling in the latent dimensionality. The cost can be reduced easily by doing a pass over all potentials before updating Q. This variant is often referred to as parallel EP (van Gerven et al., 2010). Convergence is not guaranteed in either case (Seeger and Nickisch, 2011a). But for log-concave models EP updates are known to be well-behaved, such that the algorithm converges reliably in practice (Seeger et al., 2007).

To the best of our knowledge, tilted moments for the exponential- and softplus functions are not available in closed form. Implementations based on quadrature are commonly found in the context of Gaussian processes (Vanhatalo et al., 2013; Rasmussen and Nickisch, 2010).

Computing tilted marginals is not a trivial task. E.g. plugging Eq. 1 into Eq. 10 shows that this quantity depends exponentially both on y and f. In Section 3 we illustrate that evaluating this expression directly during numerical integration can lead to problems.

So far, we have seen that popular methods, such as Laplace and VB approximations, are not particularly suitable for the RL function in contrast to EP, which in turn depends on the tractability of tilted moments. Next, we show that for the RL function these computations are indeed analytically tractable.

2.2. Tractable EP Updates for the Rectified-Linear Function

We drop indices and consider the update of a single approximate potential $\tilde{t}(s) = \tilde{Z}\,\mathcal{N}\left(\tilde{\mu}, \tilde{\sigma}^2\right)$. To obtain the first and second moments, it suffices to compute $\log \hat{Z}$, $\alpha := \frac{d}{d\mu_-} \log \hat{Z}$ and $\beta := -\frac{d^2}{d\mu_-^2} \log \hat{Z}$ since $\log \hat{Z}$ is related to the moment generating function. From these quantities, we can directly update the parameters of the approximate potential as shown e.g. in (Rasmussen and Williams, 2005)[2].

Our main result constructively shows how to compute $\hat{Z}$.

Proposition 1 *The tilted partition function $\hat{Z}$ can be computed in $O(y)$.*

2. See appendix

Proof The likelihood in Eq. 1 with the RL function the likelihood potential is $t(f) = \frac{1}{y!} f^y e^{-f} \mathrm{I}_{\{f \geq 0\}}$. Therefore, we can write the partition function of the tilted distribution as

$$\hat{Z} = \frac{1}{y!} \int_0^\infty f^y e^{-f} \mathcal{N}(f|\mu_-, \sigma_-^2) \, \mathrm{d}f \tag{13}$$

The exponential term results in a shift of the cavity mean and a constant factor:

$$\hat{Z} = \frac{1}{y!} e^{\frac{1}{2}\sigma_-^2 - \mu_-} \int_0^\infty f^y \mathcal{N}(f|\mu_- - \sigma_-^2, \sigma_-^2) \, \mathrm{d}f \tag{14}$$

Thus, computing $\hat{Z}$ boils down to computing the y-th moment of a truncated Gaussian. Define $m = \mu_- - \sigma_-^2$, $v = \sigma_-^2$, and $\kappa = -\frac{m}{\sqrt{v}}$. Let $I_y = \int_0^\infty f^y \mathcal{N}(f|m, v) \, \mathrm{d}f$. For $y \in \{0, 1\}$ the integral I_y can be readily evaluated as

$$I_0 = 1 - \Phi(\kappa) \qquad I_1 = m I_0 + \sqrt{v}\phi(\kappa) \tag{15}$$

where $\phi(x)$ is the standard normal density and $\Phi(x)$ its CDF. For $y > 1$ the application of integration by parts results in a recursion over y[3]:

$$I_y = \int_0^\infty f^{y-1} f \mathcal{N}\left(f|m, v\right) \, \mathrm{d}f = m I_{y-1} + v(y-1) I_{y-2} \tag{16}$$

where we have used that

$$f\mathcal{N}(f|m, v) = m\mathcal{N}(f|m, v) - v\frac{d}{df}\mathcal{N}(f|m, v) \tag{17}$$

For our implementation, we found it more convenient to express $\hat{Z}$ in terms of $L_y := \frac{d}{dm} \log I_y$. We can write it recursively as well[4] using Eq. 16:

$$L_y = \frac{y I_{y-1}}{I_y} = \frac{y I_{y-1}}{m I_{y-1} + v I_{y-2}} = \frac{y}{m + v L_{y-1}} \tag{18}$$

where the base cases are

$$L_0 = \phi(\kappa)/(\sigma_-(1 - \Phi(\kappa))) \qquad L_1 = I_0 / I_1 \tag{19}$$

Then, we can accumulate I_y recursively in the log-domain:

$$\log I_y = -\sum_{r=1}^{y} \log L_r + \log I_0 + \log(y!) \tag{20}$$

such that finally

$$\log \hat{Z} = \log I_y - \log(y!) + \frac{1}{2}\sigma_-^2 - \mu_- \tag{21}$$

3. There is an intimate relation between this form of $\hat{Z}$ and the solutions of certain differential equations. The first step of our recursion can be found (Gil et al., 2006), although motivated from a different perspective.
4. We provide a more detailed derivation in the supplementary material.

■

Since $\frac{d}{dm} f(m) = \frac{d}{d\mu_-} f(m)$, we conclude that $\alpha = L_y - 1$. Similar calculations, which we omit for readability[4], show that $\beta = L_y(L_y - L_{y-1})$, such that all relevant quantities can be computed based on L_y.

An alternative way to characterize moments of $\hat{P}$, which allows us to compute them to higher order is the following

Corrolary 1 *The first n moments of $\hat{P}(f)$ can be computed in $O(y+n)$.*

Proof The n-th moment can be written as follows, where we index the partition functions by y for clarity

$$\mathrm{E}_{\hat{P}}\left[f^n\right] \;=\; \frac{1}{y!\,\hat{Z}_y}\int_0^\infty f^{y+n} e^{-f} Q_-(f)\,\mathrm{d}f \;=\; \frac{(y+n)!\,\hat{Z}_{y+n}}{y!\,\hat{Z}_y} \tag{22}$$

Running the recursion up to $y+n$ evaluates all required partition functions in linear time. ■

This formulation can be evaluated in the log-domain by computing $\log \hat{Z}_y$ and $\log \hat{Z}_{y+n}$ and exponentiating their difference.

Having access to all moments allows us to compute higher-order cumulants as well. Thus, for GP priors the techniques to correct the EP approximation described in (Opper et al., 2013) directly apply to this likelihood.

2.3. Implementation Details

We implemented the EP updates in C/C++ and used the GPML MATLAB toolbox for experiments (Rasmussen and Nickisch, 2010). We use parallel updating EP with the option of fractional updates (Seeger, 2008), which turned out to be unnecessary as EP converged reliably within 15 to 20 iterations. We used MATLAB's Parallel Computing Toolbox to compute variance for the SLM experiments with up to 2^{14} variables on NVIDIA Tesla C2070 GPUs with 6 GB device memory built into a workstation equipped with dual Intel Xeon X5670 CPUs (2.93 GHz), and 128 GB Memory. These computations were performed in single precision resulting in a large speedup without a negative impact on the convergence of EP, which can be quite sensitive to inaccurately approximated variances (Papandreou and Yuille, 2011). Thus with minimal effort and without further optimizations, we could run a single iteration of EP in about 30 seconds.

3. Experiments

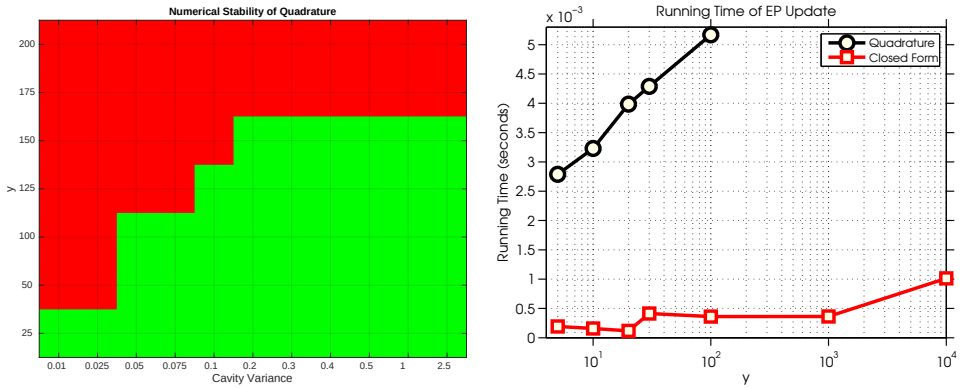
3.1. Synthetic Data

In the experiments on synthetic data we investigated the following two aspects: computational performance and numerical stability of our formulation in contrast to quadrature.

First, we investigate the numerical stability of quadrature by examining the behavior for a single EP update for the RL function[5]. We evaluate the unnormalized tilted density

5. We use MATLAB's Integral routine.

as $\hat{Z}\,\hat{P}(f) = e^{\log t(f) + \log Q_{-}(f)}$ and compute $\hat{Z}$ for different values of y and different cavity parameters. Since we can expect the cavity mean to be close to the observation, we set $\mu_{-} = y$ and vary the cavity variance. The outcome is shown in Figure 2a, where red shading denotes failure of quadrature resulting in an output which is infinite or not a number. In the green area the output matches our formulation. Our formulation works reliably in all of these cases.



(*a*) Numerical Stability (*b*) Running Time

Figure 2: **Left: Transition diagram of numerical stability of EP update using quadrature.** Red shading indicates failure due to numerical instability at a setting. Quadrature cannot handle large counts and is sensitive to small cavity variances. Our formulation works reliably in this regime and beyond. **Right: Running time.** We compare the scaling behavior of the running time of a single EP update using our formulation vs. adaptive quadrature as a function of the count y. As shown in Figure 2a, quadrature up to the same counts as our recursion.

Next, we compare the time to compute a single EP update, i.e. $\hat{Z}$ and the moments of $\hat{P}(f)$, using our recursion versus adaptive quadrature. Quadrature needs to be called three times to compute $\hat{Z}$ and the first and second moments of $\hat{P}(f)$, whereas we need to run our recursion only once. Quadrature can certainly be further optimized. But due to its complexity, this can be expected to be an error-prone undertaking.

Since our recursion scales linearly in the count y, we plot the time against y in Figure 2b. We see that our recursion is very efficient and runs robustly up to very large counts. As seen before, for quadrature, the computations cannot be run for counts beyond the order of 100. For the comparison, we performed 50 warm-up runs for both methods before averaging the running time over 150 calls to the respective implementations of the EP updates.

3.2. Cox Processes: Coal Mining Disaster Data

In this experiment we present a case where the use of the exponential link hurts generalization performance. We return to the introductory example of the coal mining disaster dataset and setup a prediction task using 10-fold cross validation. The dataset consists of 191 accidents in the period between 1851 and 1962, which we discretized into 100 equidistant bins. We compare inference for the three different link functions, using EP for all of them,

where the updates for the logistic and exponential links is implemented using quadrature[6]. As error measure we report the average of the negative log-predictive probabilities of the samples in the test fold, where the predictive probability for an unseen observation y^* given training data $\boldsymbol{y}$ is defined as:

$$P(y^*|\boldsymbol{y}) = \int P(y^*|g(f^*))Q(f^*, \boldsymbol{f}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}f^* \tag{23}$$

This can be computed as described in (Rasmussen and Williams, 2005) and amounts to evaluating $\hat{Z}$.

As in the demo in (Vanhatalo et al., 2013), we use a GP prior with a isotropic squared-exponential covariance function and a constant mean. We learn the kernel parameters as well as the mean by maximizing the marginal likelihood on the training fold. We repeatedly ran this experiment for 5 draws of the test folds. We report the cross validation errors in Table 3. This is an example where (asymptotically) linear behaviour seems to lead to better

	Exponential	Softplus	Rect. Linear
CV Error	1.63(±0.01)	1.61(±0.01)	**1.60**(±0.02)

Table 3: Coal Mining Disaster Data: Cross Validation Results

predictive performance. Softplus- and RL functions perform very similarly in this example, but better than exponential, consistently across different draws of the cross validation folds (Figure 3).

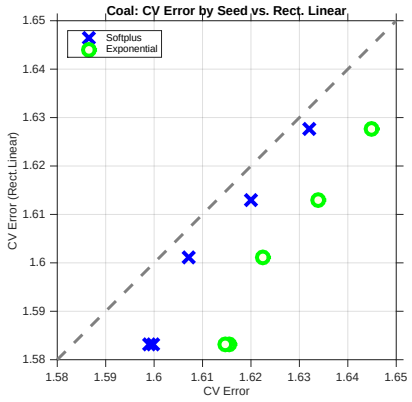


Figure 3: **Coal Mining Disaster Data:** Cross validation errors for different draws of folds. We show errors of Softplus vs. RL and Exponential vs. RL.

3.3. Sparse Linear Models

In this experiment, we consider a deconvolution problem of natural images under Poisson noise as described in Section 2. We generate noisy versions of the input as follows: We

6. The Laplace approximation for the logistic- and exponential links yielded similar results, so that we do not report them here.

rescale the maximum intensity of the input image $\boldsymbol{u}$ to a value $u_{\max} \in \{10, 20, 30\}$. We apply Gaussian blur to the image using a 3×3 blur kernel $\boldsymbol{h}$ with standard deviation 0.3 to obtain $\boldsymbol{f} = \boldsymbol{h} * \boldsymbol{u}$ and draw observations from $P(\boldsymbol{y}|g(\boldsymbol{f}))$ for 5 different initializations of the random number generator. For reconstruction, we use a total-variation prior with $\boldsymbol{B} = [\nabla_x^T \ \nabla_y^T]^T$ and Laplace potentials $l(s) = e^{-\tau|s|}$.

Here, we focus on comparing the correct Poisson noise model against a Gaussian noise assumption, which is often chosen based on convenience and familiarity. We use parallel EP for both models to infer the posterior mean as reconstruction. We used grid search to determine hyper-parameters using the marginal likelihood as criterion. We also tried to apply the softplus- and exponential non-linearities, but experienced numerical instability and convergence issues for a wide range of hyper-parameters.

We report relative ℓ_1 errors [7] of the reconstructions $\hat{\boldsymbol{u}} = \mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}]$ in Table 4. At lower intensities, the signal is much weaker leading generally to a higher error. It is this regime, where the correct likelihood yields the greatest improvements. As the intensity and thus the photon counts increase, the noise is better approximated by a Gaussian, such that both models perform similarly as expected.

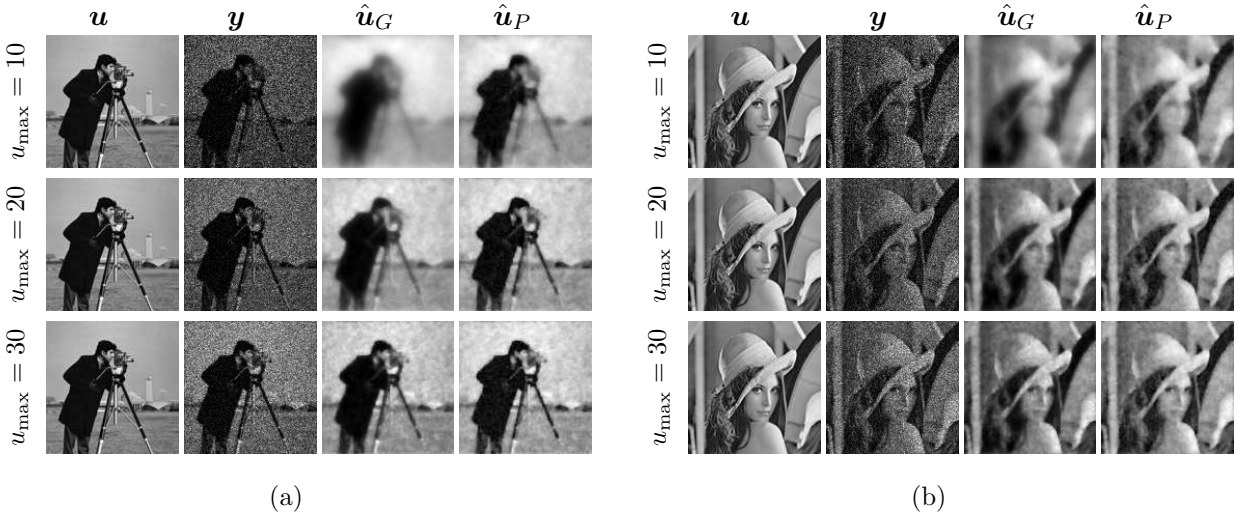
Image	$u_{\max}$	Gauss	Poisson RL
Face 32×32	10	0.488($\pm$0.005)	**0.317**($\pm$0.007)
	20	0.282($\pm$0.008)	**0.248**($\pm$0.026)
	30	0.245($\pm$0.007)	**0.207**($\pm$0.011)
Cam. Man 128×128	10	0.182($\pm$0.002)	**0.124**($\pm$0.001)
	20	0.113($\pm$0.001)	**0.094**($\pm$0.001)
	30	0.092($\pm$0.001)	**0.084**($\pm$0.001)
Lena 128×128	10	0.224($\pm$0.002)	**0.154**($\pm$0.003)
	20	0.128($\pm$0.001)	**0.111**($\pm$0.001)
	30	0.103($\pm$0.001)	**0.095**($\pm$0.001)

Table 4: Relative ℓ_1 errors for deconvolution with different likelihoods.

Apart from mere reconstruction errors, it is instructive to visually inspect the reconstructions for both models. We present exemplary reconstructions of the different input images in Figures 4 and 5. In Figure 4 each row corresponds to a different intensity level. We denote the reconstruction by $\hat{\boldsymbol{u}}_G$ and $\hat{\boldsymbol{u}}_P$, where a subscript "G" denotes the Gaussian likelihood and "P" the Poisson likelihood.

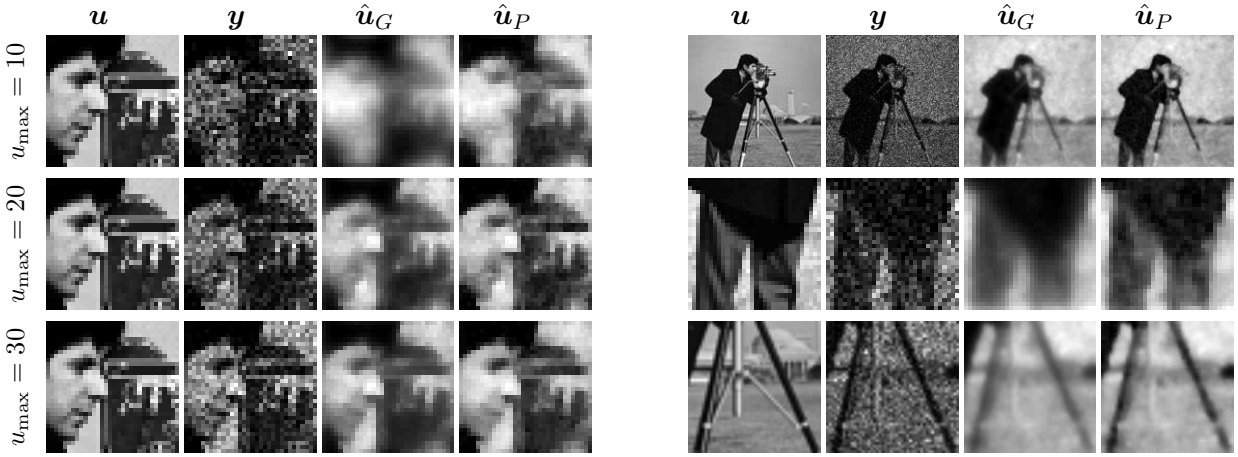
Poisson noise is difficult to deal with, especially for natural images such as in Figure 4b, since fine details become are very hard to distinguish from noise. The Gaussian noise model explains the data at very low intensities by an overly smooth image. Noise is removed, but so is also much of the high-frequency content which is crucial for recognizing details. Thus, fine structures tend to be blurred and contrast diminished. Using the Poisson likelihood instead captures edges much better. We illustrate this effect by showing a cross section of Figure 5a in Figure 6 and magnified sub-images in Figures 5b.

7. The relative ℓ_1 error of a reconstruction $\hat{\boldsymbol{u}}$ of an image $\boldsymbol{u}$ is defined as $\ell_{\boldsymbol{u}}(\hat{\boldsymbol{u}}) = \|\hat{\boldsymbol{u}} - \boldsymbol{u}\|_1 / \|\boldsymbol{u}\|_1$



(a) (b)

Figure 4: Denoising results at different maximum intensity levels. $\boldsymbol{u}$: input image. $\boldsymbol{y}$: noisy image. $\hat{\boldsymbol{u}}_G$: Gaussian likelihood. $\hat{\boldsymbol{u}}_P$: Poisson likelihood. **Left:** Cameraman 128×128. **Right:** Lena 128×128.



(a) (b)

Figure 5: **Left:** Denoising results on high-resolution 32×32 sub-image at different maximum intensity levels. **Right:** Zoom-in comparison at $u_{\max} = 20$ for Cameraman 128×128. $\boldsymbol{u}$: input image. $\boldsymbol{y}$: noisy image. $\hat{\boldsymbol{u}}_G$: Gaussian likelihood. $\hat{\boldsymbol{u}}_P$: Poisson likelihood. The correct noise model helps to recover contrast and distinguish image features from noise.

4. Conclusion

We studied inference in Poisson models using the rectified linear function as non-linearity. This function stands out in that it imposes a hard positivity constraint on the underlying latent variable. This function is the natural and physically plausible choice for models of Poisson noise in image processing, but is challenging to deal with in practice.

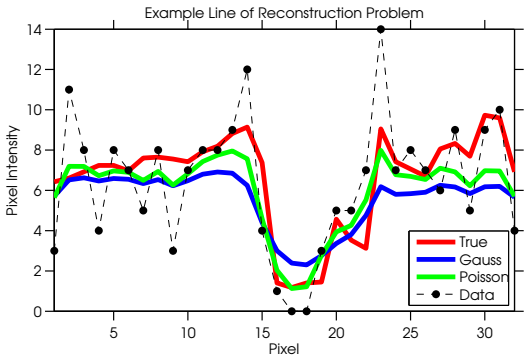


Figure 6: Cameraman Face: Example cross section of $\boldsymbol{u}, \boldsymbol{y}, \hat{\boldsymbol{u}}_G$ and $\hat{\boldsymbol{u}}_P$. We see that modeling the Poisson noise correctly helps recovering contrast and edges, which is crucial for image quality.

Here, we derived an analytically tractable Expectation Propagation algorithm for approximate inference in Poisson likelihood models using the RL function. We showed that in contrast to quadrature, computations required by our formulation are more efficient and numerically stable.

Equipped with this method, we demonstrated that the identity link is useful in situations where a non-linear link hurts generalization and that taking into account non-Gaussian noise statistics in a Poisson deconvolution problem leads to superior performance at no extra cost.

There are three avenues we would like to pursue to extend this work: To improve scalability, we would like to study the effect of factorized Gaussian approximations. For greater flexibility, we would like to investigate the compatibility with sparsity priors that are not log-concave such as spike and slab mixtures (Hernández-Lobato et al., 2014). Finally, we would like to apply this to high-level applications such as blind deconvolution under Poisson noise, in areas such as neuroscience and biomedical imaging.

References

Mikael Carlavan and Laure Blanc-Féraud. Sparse Poisson noisy image deblurring. *IEEE Trans Image Process*, 21(4):1834–46, April 2012.

Edward Challis and David Barber. Gaussian Kullback-Leibler Approximate Inference. *J. Mach. Learn. Res.*, 14(1):2239–2286, January 2013. ISSN 1532-4435.

Peter J. Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M. Taylor. Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. pages 542–563, December 2013.

François-Xavier Dupé, Mohamed-Jalal Fadili, and Jean-Luc Starck. Deconvolution of confocal microscopy images using proximal iteration and sparse representations. In *ISBI*, pages 736–739. IEEE, 2008.

Sebastian Gerwinn, Jakob H. Macke, Matthias Seeger, and Matthias Bethge. Bayesian Inference for Spiking Neuron Models with a Sparsity Prior. In *NIPS*. Curran Associates, Inc., 2007.

Sebastian Gerwinn, Jakob H Macke, and Matthias Bethge. Bayesian inference for generalized linear models for spiking neurons. *Front Comput Neurosci*, 4:12, 2010.

Amparo Gil, Javier Segura, and Nico M. Temme. Computing the real parabolic cylinder functions U(a, x), V(a, x). *ACM Trans. Math. Softw.*, 32(1):70–101, 2006.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 315–323. JMLR.org, 2011.

José Miguel Hernández-Lobato, Daniel Hernández-Lobato, and Alberto Suárez. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, page 1, 2014. ISSN 1573-0565.

R. G. Jarrett. A Note on the Intervals Between Coal-Mining Disasters. *Biometrika*, 66(1):pp. 191–193, 1979. ISSN 00063444.

Young-Jun Ko and Mohammad Emtiyaz Khan. Variational Gaussian Inference for Bilinear Models of Count Data. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, 2014.

Malte Kuss and Carl Edward Rasmussen. Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

Anat Levin, Yair Weiss, Frédo Durand, and William T. Freeman. Understanding Blind Deconvolution Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2354–2367, 2011.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astron. J.*, 79: 745+, June 1974. ISSN 00046256.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML (3)*, volume 28 of *ICML 2013 Workshop on Deep Learning for Audio*, 2013.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, London, 1989.

Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. In *UAI*, pages 362–369. Morgan Kaufmann, 2001. ISBN 1-55860-800-1.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

Manfred Opper and Cédric Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792, 2009.

Manfred Opper and Ole Winther. Gaussian Processes for Classification: Mean-Field Algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

Manfred Opper, Ulrich Paquet, and Ole Winther. Perturbative Corrections for Approximate Inference in Gaussian Latent Variable Models. *Journal of Machine Learning Research*, 14:2857–2898, 2013.

Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network*, 15(4):243–62, November 2004.

George Papandreou and Alan Yuille. Efficient variational inference in large-scale Bayesian compressed sensing. page Proc. IEEE Workshop on Information Theory in Computer Vision and Pattern Recognition, September 2011. doi: 10.1109/ICCVW.2011.6130406.

Mijung Park and Jonathan W. Pillow. Bayesian inference for low rank spatiotemporal neural receptive fields. In *NIPS*, pages 2688–2696, 2013.

Mijung Park, J Patrick Weller, Gregory D Horwitz, and Jonathan W Pillow. Bayesian active learning of neural firing rate maps with transformed gaussian process priors. *Neural Computation*, 26(8): 1519–41, August 2014.

Jonathan Pillow. Likelihood-based approaches to modeling the neural code. *Bayesian brain: Probabilistic approaches to neural coding*, pages 53–70, 2007.

Carl Edward Rasmussen and Hannes Nickisch. Gaussian Processes for Machine Learning (GPML) Toolbox. *J. Mach. Learn. Res.*, 11:3011–3015, December 2010. ISSN 1532-4435.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

William Hadley Richardson. Bayesian-Based Iterative Method of Image Restoration. *J. Opt. Soc. Am.*, 62(1):55–59, Jan 1972.

Matthias Seeger and Guillaume Bouchard. Fast Variational Bayesian Inference for Non-Conjugate Matrix Factorization Models. In *AISTATS*, volume 22 of *JMLR Proceedings*, pages 1012–1018. JMLR.org, 2012.

Matthias Seeger, Sebastian Gerwinn, and Matthias Bethge. Bayesian Inference for Sparse Generalized Linear Models. In *ECML*, volume 4701 of *Lecture Notes in Computer Science*, pages 298–309. Springer, 2007.

Matthias W. Seeger. Bayesian Inference and Optimal Design for the Sparse Linear Model. *Journal of Machine Learning Research*, 9:759–813, 2008.

Matthias W. Seeger and Hannes Nickisch. Fast Convergent Algorithms for Expectation Propagation Approximate Bayesian Inference. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 652–660. JMLR.org, 2011a.

Matthias W. Seeger and Hannes Nickisch. Large Scale Bayesian Inference and Experimental Design for Sparse Linear Models. *SIAM J. Imaging Sciences*, 4(1):166–199, 2011b.

Eero P. Simoncelli. Modeling the Joint Statistics of Images in the Wavelet Domain. In *IN PROC SPIE, 44TH ANNUAL MEETING*, pages 188–195, 1999.

Jean-Luc Starck and Fionn Murtagh. *Astronomical Image and Data Analysis*. Springer, 2002.

Marcel A. J. van Gerven, Botond Cseke, Floris P. de Lange, and Tom Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1):150–161, 2010.

Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in medicine*, 29(15):1580–1607, 2010.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(1):1175–1179, 2013.

Matthew D. Zeiler, Marc'Aurelio Ranzato, Rajat Monga, Mark Z. Mao, K. Yang, Quoc Viet Le, Patrick Nguyen, Andrew W. Senior, Vincent Vanhoucke, Jeffrey Dean, and Geoffrey E. Hinton. On rectified linear units for speech processing. In *ICASSP*, pages 3517–3521. IEEE, 2013.

Appendix A. Additional Derivations

The parameters of the approximate potential depend on α and β from Section 2.2 in the following way:

$$\tilde{\sigma}^2 = \frac{1-\beta\sigma_-^2}{\beta} \qquad \tilde{\mu} = \tilde{\sigma}^2 \frac{\alpha+\beta\mu_-}{1-\beta\sigma_-^2} \tag{24}$$

Next, we derive the expression for $L_y = \frac{d}{dm}\log I_y = I_y^{-1}\frac{d}{dm}I_y$. By using a symmetry argument we note that $\frac{d}{dm}\mathcal{N}(f|m,v) = -\frac{d}{df}\mathcal{N}(f|m,v)$. Thus,

$$\begin{aligned}
\frac{d}{dm}I_y &= \int_0^\infty f^y \frac{d}{dm}\mathcal{N}(f|m,v)\,\mathrm{d}f && (25)\\
&= -\int_0^\infty f^y \frac{d}{df}\mathcal{N}(f|m,v)\,\mathrm{d}f && (26)\\
&= \left[f^y\mathcal{N}(f|m,v)\right]_0^\infty + y\int_0^\infty f^{y-1}\mathcal{N}(f|m,v)\,\mathrm{d}f && (27)\\
&= yI_{y-1} && (28)
\end{aligned}$$

where we have used integration by parts.

Next, we show that $\beta = L_y(L_y - L_{y-1})$. We have $\beta = -\frac{d}{dm}\alpha = -\frac{d}{dm}L_y$. For convenience, we denote $\frac{d}{dm}\cdot$ by $(\cdot)'$. Then,

$$\begin{aligned}
L_y' &= \left(\frac{I_y'}{I_y}\right)' && (29)\\
&= \frac{I_y''}{I_y} - \left(\frac{I_y'}{I_y}\right)^2 && (30)\\
&= \frac{y(y-1)I_{y-2}}{I_y} - (L_y)^2 && (31)\\
&= \frac{y(y-1)I_{y-2}}{mI_{y-1}+v(y-1)I_{y-2}} - (L_y)^2 && (32)\\
&= \frac{yL_{y-1}}{m+vL_{y-1}} - (L_y)^2 && (33)\\
&= L_yL_{y-1} - (L_y)^2 && (34)\\
&= -L_y(L_y - L_{y-1}) && (35)
\end{aligned}$$

In Eq. 32 we used the recursion for I_y (Eq. 16) and in Eq. 34 the recursion for L_y (Eq. 18) from the main text.