THE POISSON TRANSFORM FOR UNNORMALISED STATISTICAL MODELS

SIMON BARTHELÉMÉ
COMPUTATIONAL COGNITIVE NEUROSCIENCE LABORATORY
UNIVERSITÉ DE GENÈVE
SIMON.BARTHELME@UNIGE.CH
&
NICOLAS CHOPIN
CREST-ENSAE
NICOLAS.CHOPIN@ENSAE.FR

ABSTRACT. Contrary to standard statistical models, unnormalised statistical models only specify the likelihood function up to a constant. While such models are natural and popular, the lack of normalisation makes inference much more difficult. Here we show that inferring the parameters of a unnormalised model on a space $\Omega$ can be mapped onto an equivalent problem of estimating the intensity of a Poisson point process on $\Omega$. The unnormalised statistical model now specifies an intensity function that does not need to be normalised. Effectively, the normalisation constant may now be inferred as just another parameter, at no loss of information. The result can be extended to cover non-IID models, which includes for example unnormalised models for sequences of graphs (dynamical graphs), or for sequences of binary vectors. As a consequence, we prove that unnormalised parameteric inference in non-IID models can be turned into a semi-parametric estimation problem. Moreover, we show that the noise-contrastive divergence of Gutmann and Hyvärinen (2012) can be understood as an approximation of the Poisson transform, and extended to non-IID settings. We use our results to fit spatial Markov chain models of eye movements, where the Poisson transform allows us to turn a highly non-standard model into vanilla semi-parametric logistic regression.

Unnormalised statistical models, and especially those based on exponential families, are popular throughout machine learning, statistics and computer vision (Wainwright and Jordan, 2008). The most iconic of these models is probably the Ising model (Ising, 1925), which applies to binary vectors $y \in \{0, 1\}^d$:

$$p(y|\theta, Q) \propto \exp(\theta^t y + y^t Q y).$$

The Ising model is useful in that it can describe correlated binary data thanks to the couplings induced by $Q$, but computing the normalisation constant involves a sum over all binary vectors of size $d$. That sum cannot be computed for $d$ much larger than 30, so that in most cases inference must proceed without the normalisation constant. Many techniques have been developed in recent years for such problems, including contrastive divergence (Hinton, 2002; Bengio and Delalleau, 2009), noise-contrastive divergence (Gutmann and Hyvärinen, 2012) and various forms of MCMC (Møller et al., 2006; Murray et al., 2012; Girolami et al., 2013).

The difficulty is compounded when unnormalised models are used for non-IID data, either sequential data, or data that include covariates. In neuroscience variants of the Ising model are used to describe the response of a neural population to a stimulus $s$ (Granot-Atedgi et al., 2013): here $y$ describes the state of $d$ neurons, where $y_i = 1$ indicates that neuron $i$ is active. The Ising model with covariates has the form:

$$p(y|\alpha(s), Q(s)) \propto \exp(\alpha(s)^t y + y^t Q(s) y).$$

If the data include $n$ levels of the stimulus $s$, there are now $n$ normalisation constants to approximate. In addition to models with covariates, there are also useful unnormalised sequential models. In our application we look at models of spatial Markov chains, where the transition density of the chain is specified up to a normalisation constant, and again one normalisation constant needs to be estimated per observation.

In this paper we show that unnormalised estimation is tightly related to the estimation of point process intensities, and formulate a Poisson transform that maps the log-likelihood of a model $L(\theta)$ into an equivalent cost function $M(\theta, \nu)$ defined in an expanded space, where the latent variables $\nu$ effectively estimate the normalisation constants. In the case of non-IID unnormalised models we show further that optimisation of $M(\theta, \nu)$ can be turned into a semi-parametric problem and addressed using standard kernel methods. In the second section, we show that the noise-contrastive divergence of Gutmann and Hyvärinen (2012) arises naturally as a tractable approximation of the Poisson transform, and that this new interpretation lets us extend its use to non-IID models. Finally, we apply these results to a class of unnormalised spatial Markov chains that are natural descriptions of eye movement sequences.
1. The Poisson transform

In this section we show how unnormalised likelihoods can be turned into Poisson process likelihoods at no loss of information. We call the procedure the Poisson transform, as it generalises the Poisson-multinomial transform [Baker 1994]. We give two interpretations, one in terms of upper-bound maximisation, and one in terms of generalised KL divergences. We begin with the IID case, with the generalisation to non-IID data treated further into the text.

1.1. Background on Poisson point processes. Poisson point processes are described at length in [Kingman 1993], and we only give here the merest outline. A Inhomogeneous Poisson point process (IPP) with intensity function \( \lambda(y) \geq 0 \) over space \( \Omega \) defines a distribution over the set of countable subsets \( S \) of \( \Omega \), in such a way that, for any measurable subset \( A \subseteq \Omega \),

\[
\# \{ S \cap A \} \sim \text{Poi}(\lambda_A), \quad \lambda_A = \int_A \lambda(y) \, dy,
\]

assuming \( \lambda_A < +\infty \). In words, the number of points to be found in subset \( A \) has a Poisson distribution, with expectation given by the integral of the intensity function within \( A \); in discrete spaces the integral may of course be interpreted as a sum. In particular, provided \( \int \lambda(y) \, dy < +\infty \), the cardinality \( n \) of \( S \) is finite, and has a Poisson distribution with expectation equal to the integral of \( \lambda(y) \) over the domain (the fact follows from taking \( A = \Omega \)). Assuming again \( \int \lambda(y) \, dy < +\infty \), the log-likelihood of observing set \( S \) given the intensity function \( \lambda \) is given by:

\[
\log p(S|\lambda) = \sum_{y_i \in S} \log \lambda(y_i) - \int \lambda(y) \, dy.
\]

1.2. The Poisson transform in the IID case. The Poisson transform is simply stated: when we have \( n \) observations from an unnormalised model on \( \Omega \), we may treat them as the realisation of a certain point process at no loss of information. This results in a mapping from a likelihood function \( L(\theta) \) to another, which we note \( M(\theta, \nu) \), in an expanded space. \( M(\theta, \nu) \) has the same global maximum as \( L(\theta) \) and confidence intervals are preserved.

First, the log-likelihood function for \( n \) IID observations from an unnormalised model \( p(y|\theta) \propto \exp \{ f_\theta(y) \} \) can be written as:

\[
L(\theta) = \sum_{i=1}^{n} f_\theta(y_i) - n \log \left( \int_{\Omega} \exp \{ f_\theta(y) \} \, dy \right)
\]

and the ML estimate of \( \theta \) is the maximum of \( L(\theta) \). (For simplicity, we assume throughout that \( L(\theta) \) admits a unique maximiser.) We introduce the following alternative likelihood function:

\[
M(\theta, \nu) = \sum_{i=1}^{n} \{ f_\theta(y_i) + \nu \} - n \int_{\Omega} \exp \{ f_\theta(y) + \nu \} \, dy
\]

which by (1) is, up to additive constant \( n \log(n) \), the IPP likelihood on \( \Omega \) for intensity function \( \lambda(y) = \exp \{ f_\theta(y) + \nu + \log(n) \} \).

Our first theorem shows that estimating \( \theta \) via \( L(\theta) \) and via \( M(\theta, \nu) \) is equivalent.

Theorem 1. The set of points \( \theta^* \) such that \( \theta^* \in \arg \max_{\theta \in \Theta} L(\theta) \) matches the set of points \( \hat{\theta} \) such that \( \hat{\theta} \) is in \( \arg \max_{\theta \in \Theta, \nu \in \mathbb{R}} M(\theta, \nu) \). In particular, if \( \arg \max L(\theta) \) is a singleton, then so is \( \arg \max M(\theta, \nu) \).

Proof. For a fixed \( \theta \), \( M(\theta, \nu) \) admits a unique maximum in \( \nu \) at \( \nu^*(\theta) = -\log \int_{\Omega} \exp \{ f_\theta(y) \} \, dy \), hence \( M(\theta, \nu^*(\theta)) = L(\theta) - n \).

There are several remarks to make at this stage. First, since \( \nu^*(\theta) = -\log \int_{\Omega} \exp \{ f_\theta(y) \} \, dy \), maximising \( M(\theta, \nu) \) can be interpreted as estimating the normalisation constant along with the parameters. There is no estimation cost incurred in treating the normalisation constant as a free parameter, since the global maxima of \( L(\theta) \) and \( M(\theta, \nu) \) are the same.

Second, the usual way of computing a confidence interval for \( \theta \) is to invert the Hessian of \( L(\theta) \) at the mode. We show in the Appendix that the same confidence interval can be obtained from the Hessian of \( M(\theta, \nu) \) at the mode, so that the Poisson transform does not introduce any over or under-confidence. In addition, the Poisson-transformed likelihood can be used for penalised likelihood maximisation (see Application), does not introduce any spurious maxima, and in exponential families it can even be shown to preserve concavity (see Appendix).

Third, at this point we do not yet have a practical way of computing \( M(\theta, \nu) \), since we have assumed that integrals of the form \( \int_{\Omega} \exp \{ f_\theta(y) + \nu \} \, dy \) are intractable. The problem of approximating \( M(\theta, \nu) \) is dealt with in section [ ], where we will see that among other possibilities it can be approximated by logistic regression via noise-contrastive divergence.

Before we deal with practical ways of approximating \( M(\theta, \nu) \), we first generalise the Poisson transform to non-IID data.
1.3. The Poisson transform in the non-IID case. In the non-IID case we still have \( n \) datapoints \( y_1 \ldots y_n \in \Omega^n \) but their distribution is allowed to vary. For example the \( n \) datapoints might form a Markov chain with (unnormalised) transition density

\[
p_\theta (y_t | y_{t-1}) \propto \exp \{ f_\theta (y_t | y_{t-1}) \}
\]

which leads to the log-likelihood

\[
\mathcal{L} (\theta) = \sum_{t=1}^n \left[ f_\theta (y_t | y_{t-1}) - \log \int_\Omega \exp \{ f_\theta (y | y_{t-1}) \} \, dy \right].
\]

(The initial point \( y_0 \) is treated as a constant.) Models with covariates \( x_i \) are expressed as \( p(y_i | x_i, \theta) \propto \exp \{ f_\theta (y_i | x_i) \} \). These two cases are highly similar and for brevity we focus on the sequential case, which we use in our application.

Our first step is to extend the Poisson transform \((eq. 3)\) to yield a function \( M (\theta, \nu) \) where \( \nu \) is now a vector of dimension \( n \) (one per conditional distribution)

\[
M (\theta, \nu) = \sum_{t=1}^n \left\{ f_\theta (y_t | y_{t-1}) + \nu_t - 1 \right\} - \int_\Omega \left\{ \sum_{t=1}^n \exp \{ f_\theta (y | y_{t-1}) + \nu_t - 1 \} \right\} \, dy.
\]

**Theorem 2.** The set of points \( \theta^* \) such that \( \theta^* \in \arg \max_{\theta \in \Theta} \mathcal{L} (\theta) \) matches the set of points \( \tilde{\theta} \) such that \( \left( \tilde{\theta}, \nu^* \right) = \arg \max_{\theta \in \Theta, \nu \in \mathbb{R}^n} M (\theta, \nu) \).

**Proof.** The proof is along the same lines as that of Theorem 1: maximising \( M (\theta, \nu) \) in \( \nu_t \) gives \( \nu_t (\theta) = -\log \int_\Omega \exp \{ f_\theta (y | y_{t-1}) \} \, dy \)

and \( M(\theta, \nu^* (\theta)) = \mathcal{L} (\theta) - n \).

Note that while \( \mathcal{L} (\theta) \) involves the sum of \( n \) separate integrals, \( M (\theta, \nu) \) involves a single integral over a sum. Further, since

\[
\nu_t (\theta) = -\log \left( \int_\Omega \exp \{ f_\theta (y | y_{t-1}) \} \, dy \right)
\]

the optimal value of \( \nu_t \) is a function of \( y_{t-1} \) only. This means that we can think of the integration constants as (hopefully smooth) functions of the previous point \( y_{t-1} \). This leads to the following result: let \( \mathcal{F} \) denote an appropriate function space that contains the function \( \chi : \Omega \to \mathbb{R} \) such that \( \chi (u) = -\log \int_\Omega \exp \{ f_\theta (y | u) \} \, dy \). We introduce the following functional

\[
\mathcal{M} (\chi, \chi) = \sum_{t=1}^n \left\{ f_\theta (y_t | y_{t-1}) + \chi (y_{t-1}) \right\} - \int_\Omega \left\{ \sum_{t=1}^n \exp \{ f_\theta (y | y_{t-1}) + \chi (y_{t-1}) \} \right\} \, dy.
\]

**Corollary 3.** The set of points \( \theta^* \) such that \( \theta^* \in \arg \max_{\theta \in \Theta} \mathcal{L} (\theta) \) matches the set of points \( \tilde{\theta} \) such that \( \left( \tilde{\theta}, \chi \right) \in \arg \max_{\theta \in \Theta, \chi \in \mathcal{F}} \mathcal{M} (\chi, \chi) \).

We can use this Corollary to turn inference on unnormalised models into a semiparametric problem, where \( \theta \) is estimated parametrically and the normalisation constants are estimated as a non-parametric function \( \chi (y_{t-1}) \).

2. Practical approximations for the Poisson transform

The Poisson transform gives us an alternative likelihood function for estimation, but one that still involves an intractable integral. In this section we briefly describe some practical approximations. One is based on importance sampling and leads to an unbiased estimate of the gradient (meaning that novel stochastic gradient and approximate Langevin sampling methods are possible). The second is based on logistic regression: we show that the noise-contrastive divergence of Gutmann and Hyvärinen (2012) approximates the Poisson-transformed likelihood. Using that connection, estimation in any non-IID setting can be turned into a semiparametric classification problem.

2.1. Unbiased estimation of the gradient. The first derivatives of \( M (\theta, \nu) \) (eq. 3) equal:

\[
\frac{1}{n} \frac{\partial}{\partial \theta} M (\theta, \nu) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} f_\theta (y_i) - \int_\Omega \frac{\partial}{\partial \theta} f_\theta (y_i) \exp \{ f_\theta (y) + \nu \} \, dy
\]

\[
\frac{1}{n} \frac{\partial}{\partial \nu} M (\theta, \nu) = 1 - \int_\Omega \exp \{ f_\theta (y) + \nu \} \, dy
\]

The expectations in the second terms can be estimated unbiasedly by Monte Carlo, which is not true in general for the untransformed likelihood. The availability of an unbiased estimator for the gradient means that stochastic gradient algorithms (and their MCMC counterpart, approximate Langevin sampling, Welling and Teh (2011)) can be applied directly. The resulting method has a straightforward interpretation, since we simply adjust \( \nu \) until \( \exp \{ f_\theta (y) + \nu \} \) normalises to 1 on average.
2.2. Logistic likelihoods as an approximation: IID case. In this section we show how to approximate Poisson-transformed likelihoods, see \[\text{[3]}\] and \[\text{[4]}\], using logistic regression. Reductions to logistic regression appear in many places in the statistical literature. In the context of estimation it is described in the well-known textbook of Hastie et al. (2003) and in detail in Baddeley et al. (2010). The use of logistic regression to estimate normalisation constants is described in Geyer (1994). Recently Gutmann and Hyvärinen (2012) introduced a more general theory which they call “noise-contrastive divergence”, and show that logistic regression can be used for joint estimation of parameters and normalisation constants.

The essence of noise-contrastive divergence is to try and teach a logistic classifier to tell true data \(S = \{y_1, \ldots, y_n\}\), generated from \(p_\theta(y)\), from random reference data \(R = \{r_1, \ldots, r_m\}\), generated from some distribution with density \(q(r)\). Picking a point \(u\) at random from \(S \cup R\), and denoting \(z = 1\) (resp. \(z = 0\)) the event that \(u\) comes from \(S\) (resp. \(R\)), one obtains the following log odds ratio:

\[
\log \frac{p(z = 1|u)}{p(z = 0|u)} = \log p_\theta(u) - \log q(u) + \log(n/m).
\]

If we assume additionally that \(p_\theta(y)\) is unnormalised, \(p_\theta(y) \propto \exp \{f_\theta(y)\}\), one may replace above, in the same spirit as in our Poisson transform, the term \(\log p_\theta(u)\) by \(f_\theta(u) + \nu\), leading to

\[
\log \frac{p(z = 1|u)}{p(z = 0|u)} = f_\theta(u) + \nu - \log q(u) + \log(n/m).
\]

This leads to following simple recipe: generate reference data \(R\), then estimate jointly \((\theta, \nu)\) by fitting the logistic regression to the dataset \(S \cup R\), with points in \(S\) (resp. \(R\)) labelled as \(z_i = 1\) (resp. \(z_i = 0\)).

The obvious connection between our Poisson transform and noise-contrastive divergence is that in both cases the log normalising constant is treated as a free parameter. The following result reveals that this connection is actually deeper.

**Theorem 4.** For fixed \(\theta, \nu\), and \(S = \{y_1, \ldots, y_n\}\), and under the assumption that \(f_\theta(y) - \log q(y) \leq C(\theta)\) for all \(y \in \Omega\), the log-likelihood of the logistic regression defined above:

\[
\mathcal{R}^m(\theta, \nu) = \sum_{i=1}^n \log \left[ \frac{n \exp \{f_\theta(y_i) + \nu\}}{n \exp \{f_\theta(y_i) + \nu + mq(y_i)\}} \right] + \sum_{j=1}^m \log \left[ \frac{mq(r_j) + \log \prod_{i=1}^n f_\theta(r_j) + \nu + mq(r_j)}{n \exp \{f_\theta(r_j) + \nu \}} \right]
\]

is such that

\[
\mathcal{R}^m(\theta, \nu) + n \log(m/n) + \sum_{i=1}^n \log q(y_i) \to \mathcal{M}(\theta, \nu)
\]

almost surely as \(m \to +\infty\), relative to the randomness induced by the reference points \(\mathcal{R} = \{r_1, \ldots, r_m\}\).

**Proof.** See Appendix.

Note that the correcting term on the LHS of \[\text{[9]}\] does not depend on \(\theta\): the maximiser of \(\mathcal{R}^m(\theta, \nu)\) (in \(\theta\) and \(\nu\)) will converge as \(m \to +\infty\) to the maximiser of \(\mathcal{M}(\theta, \nu)\), and thus, by Theorem \[\text{[3]}\] to the MLE of the model. Similarly, the Hessian computed at that maximiser of \(\mathcal{R}^m(\theta, \nu)\) will converge to the Hessian of \(\mathcal{L}(\theta)\) at the MLE.

2.3. Logistic approximation: non IID models. Putting together Theorem 4 and the results in section \[\text{[3]}\] leads to the following extension of noise-contrastive divergence to non-IID problems. For an unnormalised Markov model \(p_\theta(y_i|y_{i-1}) \propto \exp \{f_\theta(y_i|y_{i-1})\}\), for data \(S = \{y_1, \ldots, y_n\}\), generate \(m\) reference datapoints \(r_j\) from kernel \(q(r_j|y_{i,j})\), (say, \(m = kn\)), and \(k\) points \(r_j\) are generated from ancestor \(y_{i-1}\), for each \(t\), then fit the semi-parametric logistic regression model that corresponds to the log odds ratio function:

\[
\log \frac{p(z = 1|u_{t-1}, u_t)}{p(z = 0|u_{t-1}, u_t)} = f_\theta(u_t|u_{t-1}) + \chi(u_{t-1}) - \log q(u_t|u_{t-1}) + \log(n/m)
\]

where \((u_{t-1}, u_t)\) represents a pair taken at random from \(\{(y_{i-1}, y_i)\} \cup \{(x_{i,j}, r_j)\}\). The parameters of this logistic model are vector \(\theta\), scalar \(\nu\), and function \(\chi : \mathcal{Y} \to \mathbb{R}\), which is why this model is indeed semi-parametric. In practice, fitting such a model is easily achieved using an appropriate regulariser (we use smoothing splines in our application).

The interpretation of the above procedure follows the same lines as in the previous section: for \(m \to +\infty\), the log-likelihood of this logistic model converges to that of the semi-parametric Poisson model defined in Theorem \[\text{[3]}\] in particular, \(\chi\) must be seen as an estimator of the (typically smooth) function \(\chi_{t-1} \to -\log \int \exp \{f_\theta(y|y_{i-1})\} dy\).

More generally, one may extend this approach to other non-IID models. For instance, if \(p_\theta(y_i) \propto \exp \{f_\theta(y_i|x_i)\}\), where \(x_i\) are covariates, then fit the same type of semi-parametric logistic regression as above, but with \(\chi\) a function of covariates \(x_i\).
The presence of dependencies motivates the introduction of models of eye movements as spatial Markov chains. Here we note $y_t$, the fixation location at time $t$, and use a log-linear form for the kernel:

$$
\log p(y_t|y_{t-1}) = s(y_t) + r(y_t, y_{t-1}) + \nu
$$

where $s(y_t)$ represents purely spatial factors, and $r(y_t, y_{t-1})$ is an interaction term that represents spatial dependencies. A well-known factor affecting fixation locations is the centrality bias (Tatler and Vincent, 2009), a preference for looking at central locations, and we take $s(y_t)$ to be a smooth function of $|y_t|$ (the distance to the center): $s(y_t) = s(\|y_t\|)$. Potential interactions between successive locations include a tendency not to stray too far from the current location (Engbert et al., 2014), and a tendency for making movements along the cardinal axes (vertical and horizontal, Foulsham et al., 2008). We therefore further decompose $r(y_t, y_{t-1})$ into

$$
r(y_t, y_{t-1}) = r_{dist}(\|y_t - y_{t-1}\|) + r_{ang}(\angle(y_t - y_{t-1}))
$$

the sum of a distance and an angular component. We model the unknown functions $s$, $r_{dist}$ and $r_{ang}$ non-parametrically, using smoothing splines. The corresponding estimators are therefore obtained by penalised likelihood maximisation, and the Poisson transform extends straightforwardly to this case: replace the maximisation of $\hat{L}(\theta) - \text{pen}(\theta)$ by the maximisation of $\mathcal{M}(\theta, \chi) - \text{pen}(\theta)$, where $\theta = (s, r_{dist}, r_{ang})$, and $\chi$ is a non-parametric function used to estimate the normalising constant, as explained in the previous section.

We use the data of Kienzle et al. (2009), who recorded eye movements while subjects where exploring a set of photographs (Fig. 1). There are 14 subjects, each contributing between 600 and 2,000 datapoints. Thanks to the techniques described above, the model described by eq. [10] can be turned into a logistic regression, and the R package mgcv (Wood, 2006) can be used to estimate the different components using smoothing splines. We used a uniform, IID reference kernel $q(y_t|y_{t-1}) = 1/\Omega$ to produce negative examples, with 20 times as many negative examples as positive. Although the logistic approximation introduces Monte Carlo variance, the estimates are very stable (see Appendix). We fit separate functions for each subject to account for interindividual variability. The results are shown on Fig. 2. We replicate known effects from the literature: central locations dominate (although some subjects may display an off-center bias), and dependencies include both an inhibitory effect of distance and a preference for movements along cardinal orientations.

Once the data have been put into a suitable format, model fitting can be performed in one line of R code and takes around 5 minutes on a normal desktop. The Poisson transform thus turns an otherwise highly non-standard model into a convenient Generalised Additive Model.

4. Discussion

The Poisson transform suggests a new way of thinking about inference in unnormalised models: if we think of the data as coming from a point process, the integration constant becomes just another parameter to estimate. We have shown that the same idea extends

---

1Specifically: gam(class ~ s(delta,k=10)+s(dcenter,k=40)+s(fxc.prev,fyc.prev,k=40)+s(angle,bs="cc",k=20),data=data,family=binomial,method="REML")
to unnormalised models in the sequential context and in the presence of covariates, in which case parametric estimation may be turned into a semi-parametric problem. Practical approximations of Poisson-transformed likelihoods can be computed using Monte Carlo or using logistic likelihoods that follow from a reinterpretation of noise-contrastive divergence.

Part of the challenge in applying the Poisson transform to models with high-dimensional covariates or dependencies on a high-dimensional vector of past values will be in the design of appropriate kernels for the non-parametric part, which corresponds to conditional normalisation constants. The great advantage of the reduction to logistic regression is that we will be able to leverage the existing literature on nonlinear classification and dimensionality reduction, including recent developments in hashing (Li and König, 2011). Inference in unnormalised models will probably always remain challenging, but we believe the Poisson transform should alleviate some of the difficulties.

APPENDIX

A.1 DERIVATIVES OF POISSON-TRANSFORMED LIKELIHOODS

The first and second derivatives of $L(\theta)$ and $M(\theta, \nu)$ are needed in the proofs and we collect them here.

Derivatives of $L(\theta)$:

$$L(\theta) = \sum_{i=1}^{n} f_\theta(y_i) - n \log \left( \int \exp \{ f_\theta(s) \} \, ds \right) := \sum_{i=1}^{n} f_\theta(y_i) - n \phi(\theta)$$

$$\frac{d}{d\theta} \phi(\theta) = \int \frac{\partial}{\partial \theta} f(s; \theta) \exp \{ f_\theta(s) - \phi(\theta) \} \, ds = E_\theta \left( \frac{\partial}{\partial \theta} f \right)$$

$$\frac{d^2}{d\theta^2} \phi(\theta) = E_\theta \left( \frac{\partial^2}{\partial \theta^2} f \right) + E_\theta \left( \left( \frac{\partial}{\partial \theta} f \right) \left( \frac{\partial}{\partial \theta} f \right)^t \right) - E_\theta \left( \frac{\partial}{\partial \theta} f \right) E_\theta \left( \frac{\partial}{\partial \theta} f \right)^t$$

$$\frac{d}{d\theta} L = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} f_\theta(y_i) - n \frac{d}{d\theta} \phi(\theta)$$

$$\frac{d^2}{d\theta^2} L = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} f_\theta(y_i) - n \frac{d^2}{d\theta^2} \phi(\theta)$$

where we have used $E_\theta$ as shorthand for the expectation with respect to density $\exp \{ f_\theta(s) - \phi(\theta) \}$.

Derivatives of $M(\theta, \nu)$:
We can show that the same confidence intervals can be obtained from tates inference. The Poisson transform preserves this log-concavity.

\[ \mathcal{M}(\theta, \nu) = \sum_{i=1}^{n} \{ f_\theta(y_i) + \nu \} - n \int \exp \{ f_\theta(s) + \nu \} \, ds \]

\[ \frac{\partial}{\partial \theta} \mathcal{M}(\theta, \nu) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} f_\theta(y_i) - n E_{\theta, \nu} \left( \frac{\partial}{\partial \theta} f \right) \]

\[ \frac{\partial}{\partial \nu} \mathcal{M}(\theta, \nu) = -n \int \exp \{ f_\theta(s) + \nu \} \, ds \]

\[ \frac{\partial^2}{\partial \theta^2} \mathcal{M}(\theta, \nu) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} f_\theta(y_i) - n \left( E_{\theta, \nu} \left( \frac{\partial^2}{\partial \theta^2} f \right) + E_{\theta, \nu} \left( \left( \frac{\partial}{\partial \theta} f \right) \left( \frac{\partial}{\partial \theta} f \right)^t \right) \right) \]

\[ \frac{\partial^2}{\partial \nu^2} \mathcal{M}(\theta, \nu) = -n \int \exp \{ f_\theta(s) + \nu \} \, ds \]

\[ \frac{\partial}{\partial \theta \partial \nu} \mathcal{M}(\theta, \nu) = -n E_{\theta, \nu} \left( \frac{\partial}{\partial \theta} f \right) \]

where we have used \( E_{\theta, \nu} \) as shorthand for the linear operator \( E_{\theta, \nu}(\varphi) = \int \varphi(s) \exp \{ f_\theta(s) + \nu \} \, ds \) (which is not an expectation in general).

### A.2 Further properties of the Poisson transform

#### A.2.1 The Poisson transform preserves confidence intervals.

The usual method for obtaining confidence intervals for \( \theta \) is to invert the Hessian matrix of \( \mathcal{L}(\theta) \) at the mode, \( \theta^* \):

\[ C_\mathcal{L} = \left( -\frac{\partial^2}{\partial \theta \partial \theta} \mathcal{L}_{|\theta=\theta^*} \right)^{-1} \]

We can show that the same confidence intervals can be obtained from \( \mathcal{M}(\theta, \nu) \) at the joint mode, \( \theta^*, \nu^* \).

At the joint maximum, \( \nu^* \) normalises the intensity function, and the Hessian of \( \mathcal{M} \) equals:

\[
H = \begin{bmatrix}
H_{aa} & H_{ba} \\
H_{ab} & H_{bb}
\end{bmatrix} = \begin{bmatrix}
\frac{\partial^2}{\partial \theta^2} \mathcal{M}(\theta, \nu) & \frac{\partial}{\partial \nu} \frac{\partial}{\partial \theta} \mathcal{M}(\theta, \nu) \\
\frac{\partial}{\partial \theta} \frac{\partial}{\partial \nu} \mathcal{M}(\theta, \nu) & \frac{\partial^2}{\partial \nu^2} \mathcal{M}(\theta, \nu)
\end{bmatrix}
= \begin{bmatrix}
\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} f_\theta(y_i) - n E_{\theta} \left( \frac{\partial^2}{\partial \theta^2} f \right) - n E_{\theta} \left( \left( \frac{\partial}{\partial \theta} f \right) \left( \frac{\partial}{\partial \theta} f \right)^t \right) & -n E_{\theta} \left( \frac{\partial}{\partial \theta} f \right)^t \\
-n E_{\theta} \left( \frac{\partial}{\partial \theta} f \right) & -n
\end{bmatrix}
\]

where again \( E \) denotes the expectation with respect to density \( \exp \{ f_\theta(s) - \phi(\theta) \} \).

Inverting \(-H\) also yields confidence intervals. By the inversion rule for block matrices, the approximate covariance for \( \theta \) using \( \mathcal{M}(\theta, \nu) \) equals

\[
C_{\mathcal{M}}^{-1} = \left( H_{aa} - H_{ba} H_{bb}^{-1} H_{ab} \right) = \left( H_{aa} - \frac{1}{n} n^2 E_{\theta} \left( \frac{\partial}{\partial \theta} f \right)^t \right) E_{\theta} \left( \frac{\partial}{\partial \theta} f \right)^t
\]

\[
= \left( \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} f_\theta(y_i) - n E_{\theta} \left( \frac{\partial^2}{\partial \theta^2} f \right) - n E_{\theta} \left( \left( \frac{\partial}{\partial \theta} f \right) \left( \frac{\partial}{\partial \theta} f \right)^t \right) \right.
\]

\[
\left. -n E_{\theta} \left( \frac{\partial}{\partial \theta} f \right)^t \right) + n E_{\theta} \left( \frac{\partial}{\partial \theta} f \right) E_{\theta} \left( \frac{\partial}{\partial \theta} f \right)^t
\]

\[
= C_{\mathcal{L}}^{-1}
\]

### A.2.1 Preservation of log-concavity in exponential families.

In exponential families, the log-likelihood is concave, which facilitates inference. The Poisson transform preserves this log-concavity.

In the natural parameterisation, exponential-family models are given by:

\[
\mathcal{L}(\theta) = \exp \left\{ \sum_{i=1}^{n} s(y_i)^t \theta - \phi(\theta) \right\}
\]
with \( s(y) \) a vector of sufficient statistics. The second derivative of \( \mathcal{L}(\theta) \) simplifies to:

\[
- \frac{1}{n} \frac{\partial}{\partial \theta} \mathcal{L}(\theta) = \int s(y)s(y)^t \exp \left( s(y)^t \theta - \phi(\theta) \right) = E_\theta \{ s(y)s(y)^t \}
\]

a p.s.d. matrix, which establishes concavity.

The second derivatives of \( M(\theta, \nu) \) (section 4) also simplify:

\[
- \frac{1}{n} \frac{\partial^2}{\partial \theta^2} M(\theta, \nu) = \exp \{ \nu - \nu^*(\theta) \} \int s(y)s(y)^t \exp \left( s(y)^t \theta - \phi(\theta) \right) \\
- \frac{1}{n} \frac{\partial^2}{\partial \nu \partial \theta} M(\theta, \nu) = \exp \{ \nu - \nu^*(\theta) \} \int s(y) \exp \left( s(y)^t \theta - \phi(\theta) \right) \\
- \frac{1}{n} \frac{\partial^2}{\partial \nu^2} M(\theta, \nu) = \exp \{ \nu - \nu^*(\theta) \}
\]

so that the full Hessian \( H \) can be written in block-form as:

\[
- \frac{1}{n} \exp \{ \nu^*(\theta) - \nu \} H = \begin{bmatrix} E_\theta \{ s(y)s(y)^t \} & E \{ s(y) \} \\ E_\theta \{ s(y)^t \} & 1 \end{bmatrix} = A
\]

and \( H \) is n.s.d if and only if for all \( x, c \) such that \( (x, c) \neq 0 \):

\[
\begin{bmatrix} x^t & c \end{bmatrix} A \begin{bmatrix} x \\ c \end{bmatrix} > 0
\]

which the following establishes:

\[
\begin{bmatrix} x^t & c \end{bmatrix} \begin{bmatrix} E_\theta \{ s(y)s(y)^t \} & E \{ s(y) \} \\ E_\theta \{ s(y)^t \} & 1 \end{bmatrix} \begin{bmatrix} x \\ c \end{bmatrix} = \begin{bmatrix} x^t & c \end{bmatrix} \begin{bmatrix} E_\theta \{ s(y)s(y)^t \} x + cE_\theta \{ s(y) \} \\ E_\theta \{ s(y)^t \} x + c \end{bmatrix}
\]

\[
= E_\theta \left( x^t s(y) s(y)^t x \right) + 2E_\theta \{ x^t s(y) \} c + c^2
\]

\[
= E_\theta \left( \left( s(y)^t x + c \right)^2 \right) > 0
\]

assuming \( E_\theta \{ s(y)s(y)^t \} \) is p.s.d. for all \( \theta \).

**A.3 Noise-contrastive divergence approximates the Poisson transform**

We assume that

\[
f_\theta(y) - \log q(y) \leq C(\theta)
\]

for a certain constant \( C(\theta) \) that may depend on \( \theta \), and all \( y \in \Omega \). We rewrite the log-odds ratio as \( h(y) = - \log(m) \) where

\[
h(y) := f_\theta(y) + \nu - \log q(y) + \log(n)
\]

does not depend on \( m \); note \( h(y) \leq \bar{h} := C(\theta) + \eta + \log(n) \). One has:

\[
\mathcal{R}_m(\theta, \nu) + \log(m/n) = \sum_{i=1}^n \log \left[ \frac{m \exp \{ f_\theta(y_i) + \nu \}}{n \exp \{ f_\theta(y_i) + \nu \} + m q(y_i)} \right] + \sum_{j=1}^m \log \left[ \frac{m q(r_j)}{n \exp \{ f_\theta(r_j) + \nu \} + m q(r_j)} \right]
\]

where the first term trivially converges (as \( m \to +\infty \)) to

\[
\sum_{i=1}^n \{ f_\theta(y_i) + \nu - \log q(y_i) \}.
\]

Regarding the second term, one has:

\[
\log \left[ \frac{m q(r_j)}{n \exp \{ f_\theta(r_j) + \nu \} + m q(r_j)} \right] = \log \left[ 1 - \frac{1}{1 + m \exp \{ -h(r_j) \}} \right]
\]

where

\[
0 \leq \frac{1}{1 + m \exp \{ -h(r_j) \}} \leq \frac{1}{m \exp(\bar{h})}.
\]
Since $|\log(1-x) + x| \leq x^2$ for $x \in [0, 1/2]$, we have, for $m$ large enough, that

$$\left| \log \left[ \frac{m(q(r_j))}{n \exp \{ f_{\theta}(r_j) + \nu \} + m(q(r_j))} \right] + \frac{1}{1 + m \exp \{ -h(r_j) \}} \right| \leq \frac{\exp(2\tilde{h})}{m^2}$$

and

$$\left| \frac{1}{1 + m \exp \{ -h(r_j) \}} - \frac{1}{m} \exp \{ h(r_j) \} \right| \leq \frac{\exp(2\tilde{h})}{m^2}$$

and since, by the law of large numbers,

$$\frac{1}{m} \sum_{j=1}^{m} \exp \{ h(r_j) \} \to \mathbb{E}_q[\exp \{ h(r_j) \}] = n \int \exp \{ f_{\theta}(y) + \nu \} \, dy < +\infty$$

almost surely as $m \to +\infty$, one also has:

$$\sum_{j=1}^{m} \log \left[ \frac{m(q(y_i))}{n \exp \{ f_{\theta}(y_i) + \eta \} + m(q(y_i))} \right] \to -n \int \exp \{ f_{\theta}(y) + \nu \} \, dy$$

almost surely, since the difference between the two sums is bounded deterministically by $\exp(2\tilde{h})/m$.

### A.4 Additional information on the application

In our application we fit a spatial Markov chain model using logistic regression. Since the procedure involves the generation of a random set of reference points, we incur some Monte Carlo error in the estimates. Estimating the magnitude of the Monte Carlo error is just a matter of running the procedure several times to look at variability in the estimates. We did so over 5 repetitions and report the results in Fig. 3. For each repetition we plot the estimated smooth effect of saccade angle $r_{ang}$, along with a 95% confidence band. Since smoothing splines are used, smoothing hyperparameters had to be inferred from the data (using REML, Wood 2011), and the reported confidence band is conditional on the estimated value of the smoothing hyperparameters. The fits and confidence bands are extremely stable over independent repetitions.
REFERENCES

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., Shah, R., and Turner, R. (2010). Spatial logistic regression and change-of-support in poisson point processes. *Electronic Journal of Statistics*, 4(0):1151–1201.

Baker, S. G. (1994). The Multinomial-Poisson transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(4):495–504.

Barthelmé, S., Trukenbrod, H., Engbert, R., and Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of vision*, 13(12).

Bengio, Y. and Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural computation*, 21(6):1601–1621.

Engbert, R., Trukenbrod, H. A., Barthelmé, S., and Wichmann, F. A. (2014). Spatial statistics and attentional dynamics in scene viewing.

Foulsham, T., Kingstone, A., and Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, 48(17):1777–1790.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. Technical Report 568, School of Statistics, University of Minnesota.

Girolami, M., Lyne, A.-M., Simpson, D., and Atchade, Y. (2013). Playing russian roulette with intractable likelihoods. *arxiv 1306.4032*.

Granot-Atedgi, E., Tkačik, G., Segev, R., and Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population codes. *PLoS Comput Biol*, 9(3):e1002922+.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(1):307–361.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*, Springer, corrected edition.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Kingman, J. F. C. (1993). *Poisson Processes (Oxford Studies in Probability)*. Oxford University Press.

Li, P. and König, A. C. (2011). Theory and applications of b-bit minwise hashing. *Commun. ACM*, 54(8):101–109.

Møller, J., Pettitt, A. N., Reeves, R., and Berthélsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions.

Tatler, B. and Vincent, B. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6):1029–1054.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference (Foundations and Trends(r) Machine Learning)*. Now Publishers Inc.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 1 edition.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.