THE UNIVERSITY OF CHICAGO

VISUALIZING NUCLEOSOME CLUSTER DYNAMICS WITH DENSE SINGLE MOLECULE LOCALIZATION MICROSCOPY

A THESIS SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
CLAYTON W. SEITZ

CHICAGO, ILLINOIS
SPRING 20XX

TABLE OF CONTENTS

ABSTRACT

CHAPTER 1

INTRODUCTION

CHAPTER 2

LITERATURE REVIEW

CHAPTER 3

MATERIALS AND METHODS

3.1 Single molecule localization microscopy

Single molecule localization microscopy (SMLM) is a type of super-resolution microscopy that allows the imaging of fluorescently-labeled molecules with high precision, well beyond the diffraction limit of light. SMLM techniques such as direct stochastic optical reconstruction microscopy (dSTORM), photoactivated localization microscopy (PALM), and related methods rely on the precise localization of single molecules by resolving them in time rather than space. By combining the precise localization of many individual molecules, SMLM can generate images with resolution down to a few nanometers. SMLM has been used in a variety of applications, including the imaging of subcellular structures such as synapses, mitochondria, and cytoskeletal elements, as well as the study of protein-protein interactions, molecular dynamics, and other biological processes at the nanoscale level.

Despite its successs and gaining popularity, the basic principle of SMLM is one of its primary limitations: the need for sparse activation leads to long acquisition times and expensive autofocusing equipment to actively correct for sample drift. This results in low throughput, poor time resolution when imaging dynamic processes, low labeling densities and a reduced choice of fluorophores. In addition, the need for sparse activation requires laborious optimization of dSTORM buffers containing oxygen scavenging systems and/or oxygen purging techniques. In response to these problems, a host software tools for SMLM have emerged, which permit the acquisition of emitters at higher densities. (Speiser 2021). In the multi-emitter setting, PSFs are no longer well-separated but may overlap, adding additional 1uncertainty into the localization process. Existing algorithms have explicitly modeled the point spread function as a mixture of single molecule PSFs or have utilized deep learning-based tools to estimate the parameters of each PSF embedded in the mixture.

Due to overlap, conventional detection strategies may undercount the emitters in a local neighborhood in some frames, localization uncertainty can increase for overlapping emitters, and some localizations may be missed entirely. These complications make conventional detection strategies inappropriate for reconstruction of super-resolution images from time-series. Perhaps most importantly, overlapping emitters can result in additional localization uncertainty, rendering discernment between the arrival of a new particle and a poorly localized existing one difficult. This issue poses a major bottleneck to super-resolution imaging acqusitions. Additional uncertainty can be partially alleviated by using pairwise or higher-order temporal correlations within a pixel neighborhood to deconvolve individual emitters. A similar idea is employed in super-resolution optical fluctuation imaging (SOFI) - a post-processing technique that uses image cumulants to deconvolve emitters.

3.1.1 Statistics of sCMOS cameras

We are now prepared to write the shot-noise limited signal, which is a vector with units of phototelectrons

$$\vec{S} = [\text{Poisson}(\mu_1), \text{Poisson}(\mu_2), ..., \text{Poisson}(\mu_N)] \tag{3.1}$$

However this noise model is incomplete, because detectors often suffer from dark noise, which may refer to readout noise or dark current, and contributes to a nonzero signal even in the absence of incident light. Dark current is due to statistical fluctuations in the photoelectron count due to thermal fluctuations. Readout noise is introduced by the amplifier circuit during the coversion of photoelectron charge to a voltage. Here, we use the Hamamatsu ORCA v3 CMOS camera, which is air cooled to -10C and has very low dark current - around 0.06 electrons/pixel/second - and can therefore be safely ignored for exposure times on the order of milliseconds. Readout noise has been often neglected in localization algorithms because its presence in EMCCD cameras is small enough that it can be ignored

within the tolerances of the localization precision. In the case of sCMOS cameras, however, the readout noise of each pixel is significantly higher and, in addition, every pixel has its own noise and gain characteristic sometimes with dramatic pixel-to-pixel variations.

It is important to note that we cannot measure the contribution by readout noise before amplification and therefore it must be expressed in units of ADU. This is in contrast to $\vec{S}$, which can be expressed in units of photoelectrons, because these statistical fluctuations can be predicted to be Poisson by quantum mechanics. Furthermore, the number of photoelectrons S_k is multiplied by a gain factor g_k which has units of [ADU/e^-], which generally must be measured for each pixel. Here, we will always assume that readout noise per pixel ξ_k is Gaussian with some pixel-specific offset o_k and variance σ_k^2. We will also assume that gain factors and pixel noise characteristics are constants and do not scale with the signal level S_k. Therefore our measurement, in units of ADU, is:

$$\vec{H} = \vec{S} + \vec{\xi} \tag{3.2}$$

What we are after is the joint distribution $P(\vec{H})$. A fundamental result in probability theory is that the distribution of H_k is the convolution of the distributions of S_k and ξ_k,

$$P(H_k|\theta) = P(S_k) \circledast P(\xi_k) \tag{3.3}$$

$$= A \sum_{q=0}^{\infty} \frac{1}{q!} e^{-\mu_k} \mu_k^q \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(H_k - g_k q - o_k)}{2\sigma_k^2}} \tag{3.4}$$

where $P(\xi_k) = \mathcal{N}(o_k, \sigma_k^2)$ and $P(S_k) = \text{Poisson}(g_k \mu_k)$. In practice, this expression is difficult to work with, so we look for an approximation. Notice that

$$\xi_k - o_k + \sigma_k^2 \sim \mathcal{N}(\sigma_k^2, \sigma_k^2) \approx \text{Poisson}(\sigma_k^2)$$

Since $H_k = S_k + \xi_k$, we transform $H'_k = H_k - o_k + \sigma_k^2$, which is distributed according to

$$H'_k \sim \text{Poisson}(\mu'_k)$$

where $\mu'_k = g_k \mu_k + \sigma_k^2$. This result can be seen from the fact the the convolution of two Poisson distributions is also Poisson.

3.1.2 Integrated isotropic Gaussian point spread function

Due to diffraction, any point emitter, such as a single fluorescent molecule, will be registered as a diffraction limited spot. It is common to describe the point spread function as a two-dimensional isotropic Gaussian:

$$\text{G}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}}$$

Modern cameras used in light microscopy, such as scientific complementary metal oxide semiconductor (sCMOS) cameras, are powered by the photoelectric effect. Electrons within each pixel, called photoelectrons, absorb enough energy from incoming photons to be promoted to the conduction band to give electrical current which can be detected. Integration of photoelectrons during the exposure time results in a monochrome image captured by a camera. The image of a single point particle, such as a fluorescent molecule, can be thought of as two-dimensional histogram of photon arrivals and a discretized form of the classical intensity profile $\text{G}(x, y)$. The value at a pixel approaches an integral of this density over the

pixel:

$$\mu_k = i_0 \lambda_k = i_0 \int_{\text{pixel}} G(x,y)dxdy \tag{3.5}$$

Let (x_k, y_k) be the center of pixel k. If a fluorescent molecule is located at (x_0, y_0), the probability of a photon arriving at pixel k per unit time reads

$$\lambda_k = \int_{x_k-\frac{1}{2}}^{x_k+\frac{1}{2}} G(x-x_0)dx \int_{y_k-\frac{1}{2}}^{y_k+\frac{1}{2}} G(y-y_0)dy$$

where $i_0 = g_k \eta N_0 \Delta$. The parameter η is the quantum efficiency and Δ is the exposure time. N_0 represents the number of photons emitted per unit time, which may be itself a Poisson random variable; however, this is inconsequential since it is a fixed value for a single image of a fluorescent molecule. We can then express the Gaussian integrals over a pixel by making use of the following property of the error function

$$\frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{2}\left(\text{erf}\left(\frac{b-\mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{a-\mu}{\sqrt{2}\sigma}\right)\right)$$

This gives a convenient expression for the fraction of photons which arrive at a pixel k

$$\lambda_k(x) = \frac{1}{2}\left(\text{erf}\left(\frac{x_k+\frac{1}{2}-x_0}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{x_k-\frac{1}{2}-x_0}{\sqrt{2}\sigma}\right)\right)$$
$$\lambda_k(y) = \frac{1}{2}\left(\text{erf}\left(\frac{y_k+\frac{1}{2}-y_0}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{y_k-\frac{1}{2}-y_0}{\sqrt{2}\sigma}\right)\right)$$

3.1.3 Integrated astigmatic Gaussian point spread function

In 2D SMLM simulations, a 2D PSF model with a z-dependent isotropic width σ can be used. For 3D, we could use that the isotropic Gaussian point spread function has a FWHM σ which is dependent on the axial coordinate. However, it can be shown that the error

around the focus is very large and negative and positive defocus cannot be distinguished given the symmetric dependence in z. Therefore, for 3D SMLM, a cost-effective approach is to introduce astigmatism into the detection path using a weak ($f \approx 10$m) cylindrical lens. This gives an anisotropic Gaussian point spread function which is elongated perpendicular to the optical axis, depending on the axial (z) position of the fluorescent emitter. A fairly simple model for $\sigma_x(z_0)$ and $\sigma_y(z_0)$ upon defocus of a fluorescent molecule might be

$$\sigma_x(z_0) = \sigma_0 + \alpha(z_0 + z_{min})^2 \quad \sigma_y(z_0) = \sigma_0 + \beta(z_0 - z_{min})^2$$

with the following continuous density over the pixel array

$$\mathrm{G}(x, y) = \frac{1}{2\pi\sigma_x(z)\sigma_y(z)} e^{-\frac{(x-x_0)^2}{2\sigma_x(z_0)^2} + \frac{(y-y_0)^2}{2\sigma_y(z_0)^2}} \tag{3.6}$$

3.1.4 Localization microscopy as Bayesian inference

Armed with a model for image formation, we adopt the Bayesian view of localization microscopy, which provides a rigorous statistical framework for bounding our uncertainty in localization microscopy measurements. This involves defining a lower bound on the uncertainty in the localization parameters - typically lateral or lateral and axial coordinates of each of K emitters in the field of view. Some authors have estimated this lower bound using point spread functions in continuous space; however, the spatial binning process results in a loss of information during photon detection, which, in turn, can inflate our uncertainty. Therefore, in the remainder of this chapter, I make use of our generative image model rather than using calculus on continuous spaces to compute uncertainties. We begin with Bayes rule:

$$P(\theta|\vec{H}) = \frac{P(\vec{H}|\theta)P(\theta)}{\int P(\vec{H}|\theta)P(\theta)d\theta} \tag{3.7}$$

Under the Poisson approximation, the model negative log-likelihood is

$$\ell(\vec{H}|\theta) = -\log \prod_k \frac{e^{-(\mu'_k)} \left(\mu'_k\right)^{n_k}}{n_k!} \tag{3.8}$$

$$= \sum_k \log n_k! + \mu'_k - n_k \log\left(\mu'_k\right) \tag{3.9}$$

Clearly our task in Bayesian inference is to transform our prior knowledge about θ into a posterior distribution, using the observed data. However, when the log-likelihood is easy to compute (as it is here), we may ask: how much information does the data actually carry about parameters of our model? If the likelihood of the dataset is roughly constant for any parameter set we choose, we might expect our model cannot explain our observations. In other words, the data does not appear to carry much information about the parameters. After all, our posterior is being shaped in part by this likelihood. On the other hand, if ℓ has a number of bumps or inflection points, then we expect that maybe some parameter sets make our observed data more likely. The "bumpiness" of the likelihood surface is called the Fisher information - a fundmantal metric in information geometry.

3.1.5 The Fisher information metric

The Fisher information matrix $I(\theta)$ can be directly related to the curvature of the KL-Divergence over the parameter space

$$\begin{aligned}
\nabla^2_{\theta'} D_{KL}[\ell(H|\theta) \parallel \ell(H|\theta')] &= -\nabla_{\theta'} \int \ell(H|\theta) \nabla_{\theta'} \log \ell(H|\theta')\, dH \\
&= -\int \ell(H|\theta) \nabla^2_{\theta'} \log \ell(H|\theta')\, dH \\
&= -\mathbb{E}_\theta[\nabla^2_{\theta'} \log \ell(H|\theta')] \\
&= I(\theta)
\end{aligned}$$

We often call the Hessian matrix the *score*. The Fisher information is the result of averaging the score over the parameter space.

3.1.6 Fisher information for the 3D integrated gaussian

Estimating the Fisher information matrix in the 3D case is more challenging, since it cannot be written in the form of (1.10). A reasonable option is to turn to Monte Carlo methods to estimate the expectation. Classic Markov Chain Monte Carlo (MCMC) techniques like Metropolis-Hastings or Hamiltonian Monte Carlo could be used, but are notoriously slow and difficult to tune. Alternatives based on Langevin dynamics can be used when gradients of the likelihood are readily available. This class of methods use gradient information to guide an optimizer through the parameter space as well as stochastic terms to prevent the optimizer from becoming trapped in local modes. After burn-in Langevin dynamics can be interpreted as sampling from a stationary distribution, which in our case is the posterior distribution on PSF model parameters. Therefore, we will use Stochastic Gradient Langevin Dynamics (SGLD), an algorithm commonly used to sample from the parameter's posterior. By drawing samples, we can simultaneously estimate the Fisher information matrix (by computing the Hessian at each sample) while obtaining estimates of parameter uncertainty by computing the variance of a batch of samples.

3.2 Markovian photoswitching dynamics

The number of molecules within the diffraction limit is $K\left(\frac{\lambda}{2\mathrm{NA}}\right)$. If α is the *detection probability*, then $\alpha K\left(\frac{\lambda}{2\mathrm{NA}}\right)$ are detected, on average. We want to minimize

$$\mathcal{L} = \alpha K\left(\frac{\lambda}{2\mathrm{NA}}\right) + \gamma\left(\Delta_{\mathrm{SR}} + \frac{2N}{\log(1-\alpha)}\right)^2$$

The second term contains $\frac{2N}{\log(1-\alpha)}$, which is the minimum number of frames needed to detect 99 percent of N molecules (which can be obtained from the geometric distribution). If we assume a two-state generator, then

$$P(t) = P(0)e^{Gt}$$

and $G\pi = 0$ gives $\pi = (\alpha, \beta) = \frac{1}{k}(k_{12}, k_{21})$ where $k = k_{12} + k_{21}$.

The true photo-switching behavior of the fluorophore is a continuous time stochastic phenomenon. However, an experimenter can only ever observe a discretized manifestation of this by imaging the fluorophore in a sequence of frames. These frames are regarded as a set of sequential exposures of the fluorophore and the observed discrete time signal indicates whether the fluorophore has been observed in a particular frame. It is the continuous time process on which we wish to draw inference based on the observed discrete-time process indicating whether the fluorophore was observed in a frame. In this section we first present the continuous time Markov model of the true (hidden) photo-switching behavior, and then derive the observed discrete time signal, together with key results on its statistical properties.

We model the photoswitching behavior of a fluorophore as a continuous time Markov process. It is a stochastic process, which satisfies the Markov property

$$P(X_t|X_{t-1}, X_{t-2}, ..., X_{t-N}) = P(X_t|X_{t-1})$$

We consider a general model for $X(t)$ that can accommodate the numerous mechanisms of

photo-switching utilized in standard SMLM approaches such as (F)PALM and (d)STORM. Specifically, this model consists of a photon emitting (On) state 1, $m+1$ non photon emitting (dark/temporary off) states 0_0, 0_1, ..., 0_m, where $m \in \mathbb{Z}\geq 0$, and a photobleached (absorbing/permanently off) state 2. In order to accommodate for the $m = 0$ case when we have a single dark state, we use the notational convention that state $0_0 \equiv 0$. The model, allows for transitions from state 1 to the multiple dark states (from a photochemical perspective, these can include triplet, redox and quenched states). These dark states are typically accessed via the first dark state 0 (reached as a result of inter-system crossing of the excited $S1$ electron to the triplet $T1$ state). Further dark states 0_{i+1}, $i = 0, \ldots, m-1$, are accessible by previous dark states 0_i (by, for example, the successive additions of electrons forming radical anions (Van de Linde et al., 2010)). We allow the On state 1 to be accessible by any dark state and we consider the most general model in which the absorption state 2 is accessible from any combination of other states (Vogelsang et al., 2010; Van de Linde and Sauer, 2014; Ha and Tinnefeld, 2012).

To capture $P(X_t|X_{t-1})$ for all possible pairs X_t and X_{t-1}, we define a square generator matrix $G \in \mathcal{R}^{N\times N}$ where N denotes the number of states. As such, the elements of G represent the probability of a transition from a state ω_j to ω_i in an infinitesimal time interval

$$G_{ij} = \Pr\left(X(t+dt) = \omega_i, \mid X(t) = \omega_j\right)$$

Let the state space for the process $X(t)$ be $\Omega = \{0_0, 0_1, 0_2, 1, 2\}$. The generator matrix for such a process reads

$$G = \begin{pmatrix} \lambda_{00} & \lambda_{00_1} & 0 & \lambda_{01} & \mu_0 \\ 0 & \lambda_{0_1 0_1} & \lambda_{0_1 0_2} & \lambda_{0_1 1} & \mu_1 \\ 0 & 0 & \lambda_{0_2 0_2} & \lambda_{0_2 1} & \mu_2 \\ \lambda_{10} & 0 & 0 & \lambda_{11} & \mu_0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Of considerable practical interest is using the matrix G to determine the probability the system is in a particular state at a time t. This can be achieved by solving the master equation:

$$\frac{\partial P(\omega_i)}{\partial t} = \sum_j G_{ji} P(\omega_j, t) - G_{ij} P(\omega_i, t)$$

Our notational convention is the G_{ij} is a transition $i \to j$ while G_{ji} is $j \to i$. The above equation and its solution is very powerful in this context and in the broader context of non-equilibrium dynamics, so I will devote a short section to its derivation. The reader can safely skip this section if time is scarce.

3.2.1 Solving the master equation

The master equation is a first order differential equation and its solution is straightforward when the dimensionality of the state space is small. The solution is found easily by massaging the master equation into something that has a simple exponential solution. Define a matrix W s.t. $W_{ij} = T_{ij}$ and $W_{ii} = -\sum_{j \neq i} G_{ij}$. This operator acts on $P(\omega)$ and gives a vector of probability currents $\dot{P}(\omega, t) = \mathcal{J}(\boldsymbol{\omega}) = WP(\boldsymbol{\omega})$.

$$P(\boldsymbol{\omega}, t) = \exp(WP(\omega))$$

The matrix W for the 4-state system presented before reads

$$W = \begin{pmatrix} -\sigma_0 & \lambda_{00_1} & 0 & \lambda_{01} & \mu_0 \\ 0 & -\sigma_{0_1} & \lambda_{0_1 0_2} & \lambda_{0_1 1} & \mu_1 \\ 0 & 0 & -\sigma_{0_2} & \lambda_{0_2 1} & \mu_2 \\ \lambda_{10} & 0 & 0 & -\sigma_1 & \mu_0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where we have introduced shorthands $\sigma = -\sum_{j \neq i} G_{ij}$.

3.2.2 Modeling photoswitching dynamics

The imaging procedure requires taking a series of successive frames. Frame n is formed by taking an exposure over the time interval $[n\Delta, (n+1)\Delta)$, where $n \in \mathbb{Z} \geq 0$. The constant Δ corresponds to the exposure time for a single frame, also known as the frame length. We define the discrete time observed process $Y_n : n \in \mathbb{Z} \geq 0$, with state space $S_Y = 0, 1$, as $Y_n = 1$ if the fluorophore (characterized by $X(t)$) is observed in frame n and equal to 0 otherwise. For the fluorophore to be observed in the time interval $[n\Delta, (n+1)\Delta)$, it must be in the On state 1 for a minimum time of $\delta \in [0, \Delta)$. The value of δ is unknown and is a result of background noise and the imaging system's limited sensitivity. We note that if $X(t)$ exhibits multiple jumps to state 1 within a frame, then a sufficient condition for observing the fluorophore is that the total time spent in the On state exceeds δ. The $\delta = 0$ case is the idealistic scenario of a noiseless system and perfect sensitivity such that the fluorophore is detected if it enters the On state for any non-zero amount of time during the exposure time Δ.

$$Y_n = \mathbb{1}_{[\delta, \Delta)} \left(\int_{n\Delta}^{(n+1)\Delta} \mathbb{1}_1(X(t)) dt \right)$$

Further, the distribution on Δt will depend on the hidden state of the system before integration begins.

$$\Delta t = \int_{n\Delta}^{(n+1)\Delta} \mathbb{1}_1(X(t))dt \qquad \Delta t \sim P(\Delta t | X_t = i)$$

where $P(\Delta t|X_t = i)$ will be an exponential distribution with parameters dependent on the starting state. Following (Patel 2019), we write down the joint probability of a transition from state i to state j during an interval Δ and the observation $Y_n = \ell$.

$$\begin{aligned} b^{\ell}_{ij,\Delta} &= \mathbb{P}(Y_n = \ell, X(\Delta) = j | X(0) = i) \\ &= \sum_k q_{ij}(k, \Delta)\xi_{ij}(\ell, k, \Delta) \end{aligned}$$

where we have made the following definitions

$$\begin{aligned} q_{ij}(k, \Delta) &= \mathbb{P}(N(\Delta) = k, X(\Delta) = j | X(0) = i) \\ \xi_{ij}(\ell, k, \Delta) &= \mathbb{P}(Y_n = \ell | N(\Delta) = k, X(\Delta) = j, X(0) = i) \end{aligned}$$

We focus first on $q_{ij}(k, \Delta)$, which represents the probability distribution over k transitions of type ij occuring in the interval Δ. This is analagous to considering the time-evolution of the probability mass over states in our solution of the master equation. However, here we instead find the time-evolution of the probability mass over the number of transitions k of a certain type q_{ij}, and then integrate out the variable k. As an example, the distribution over k for transition of type q_{i0} (from $i \to 0$), is determined by the distribution over the number of $i \to 1$ transitions and the rate of a $1 \to 0$ transition over a finite time interval Δt

$$q_{i0}(k, t + \Delta t) = (1 - \sigma_0 \Delta t)q_{i0}(k, t) + \lambda_{10} q_{i1}(k, t)\Delta t + \mathcal{O}(\Delta t)$$

The negative contribution represents the distribution over the number of transitions $i \to 0$ that have already occured by time t and will leave the 0 state in the interval Δt. The full coupled system of differential equations reads

3.3 A deep learning framework for super-resolution microscopy

3.4 Bayesian clustering of chromatin nanodomains

3.4.1 A Brief Note on Bayesian Nonparametrics

In parametric modeling, it is assumed that data can be represented by models using a fixed, finite number of parameters. Examples of parametric models include clusters of K Gaussians and polynomial regression models. In many problems, determining the number of parameters a priori is difficult; for example, selecting the number of clusters in a cluster model, the number of segments in an image segmentation problem, the number of chains in a hidden Markov model, or the number of topics in a topic modelling problem before the data is seen can be problematic.

In nonparametric modeling, the number of parameters is not fixed, and often grows with the sample size. Kernel density estimation is an example of a nonparametric model. In Bayesian nonparametrics, the number of parameters is itself considered to be a random variable. One example is to do clustering with k-means (or mixture of Gassuians) while the number of clusters k is unknown. Bayesian inference addresses this problem by treating k itself as a random variable. A prior is defined over an infinite dimensional model space, and inference is done to select the number of parameters. Such models have infinite capacity, in that they include an infinite number of parameters a priori; however, given finite data, only a finite set of these parameters will be used. Unused parameters will be integrated out.

3.4.2 Variational Bayes

A complete description of the generative model in the Bayesian framework includes the prior distribution $P(\theta)$ which describes the spatial distribution and temporal dynamics of fluorophores, and the likelihood $P(\vec{H}|\theta)$ - a distribution of images generated by the microscope for a given configuration of fluorophores. According to Bayes rule, the posterior on θ can be constructed as

$$P(\theta|\vec{H}) = \frac{P(\vec{H}|\theta)P(\theta)}{\int P(\vec{H}|\theta)P(\theta)d\theta} \tag{3.10}$$

The posterior $P(\theta|\vec{H})$ is commonly approximated using Markov Chain Monte Carlo (MCMC) methods. MCMC is asymptotically exact, but can be slow and hyperparameters can be difficult to tune. On the other hand, variational inference (VI) attempts to approximate the true posterior distribution $p(\theta|x)$ with a variational distribution $q(\theta)$, which is often a model distribution which can be easily sampled from e.g., a multivariate Gaussian. One method to fit the variational distribution is to minimize the KL-divergence between the variational distribution and the true posterior:

$$\begin{aligned}
D_{\mathrm{KL}}(q(\theta)||p(\theta|x)) &= \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)}{p(\theta|x)}\right) \\
&= \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)p(x)}{p(\theta,x)}\right) \\
&= \log p(x) + \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)}{p(x|\theta)p(\theta)}\right) \\
&= \log p(x) + \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log q(\theta) - \log p(x|\theta) - \log p(\theta)\right)
\end{aligned}$$

We can treat $\log p(x)$ as a constant here, so the KL-divergence is minimized by minimizing this expectation. Equivalently, we can maximize its negative, which is often called the evidence lower bound (ELBO).

$$\text{ELBO} = -\mathop{\mathbb{E}}_{\theta\sim q(\theta)}(\log q(\theta) - \log p(x|\theta) - \log p(\theta)) \tag{3.11}$$

Let $\theta = (\theta_1, ..., \theta_K)$ and $\theta_k = (x_k, y_k)$ be the pointillist coordinates

$$\begin{aligned}
\nabla_\phi \text{ELBO} &= -\nabla_\phi \left(\mathop{\mathbb{E}}_{\theta\sim q_\phi(\theta)} \big(\log q_\phi(\theta) - \log p(x|\theta)\big) \right) \\
&= \nabla_\phi \int_\theta q_\phi(\theta) \log p(x|\theta) d\theta - \nabla_\phi \int_\theta q_\phi(\theta) \log q_\phi(\theta) d\theta \\
&= \int_\theta \nabla_\phi q_\phi(\theta) \log p(x|\theta) d\theta - \nabla_\phi H(q_\phi(\theta))
\end{aligned}$$

where $H(q_\phi(\theta))$ is the differential entropy of the variational posterior. If the variational posterior is simple, like a multivariate Gaussian, gradients of the differential entropy can be computed analytically. However, the first term is intractable and often must be estimated using Monte Carlo methods

$$\int_\theta \nabla_\phi q_\phi(\theta) \log p(x|\theta) d\theta \approx \mathop{\mathbb{E}}_{\theta\sim q_\phi(\theta)} \nabla_\phi q_\phi(\theta) \log p(x|\theta)$$

CHAPTER 4

RESULTS

CHAPTER 5

DISCUSSION

Appendices

We approximate the log-likelihood function to second order in terms of the Jacobian and Hessian matrices.

$$\ell(\theta + \Delta\theta) \approx \ell(\theta) + \nabla\ell^T \Delta\theta + \frac{1}{2}\theta^T H\theta$$

and

$$\nabla\ell(\theta + \Delta\theta) = \nabla\ell(\theta) + H\Delta\theta = 0$$

and so $\Delta\theta = -H^{-1}\nabla\ell(\theta)$. We will make use of both of these in the following sections, so we derive their analytical form now. We will derive the gradients for the integrated astigmatic Gaussian, since it is the more general case. As before, define $i_0 = g_k\gamma\Delta t N_0$ such that $\mu'_k = i_0\lambda_k$

$$J_{x_0} = \beta_k\lambda_y\frac{\partial\lambda_x}{\partial x_0} \quad J_{y_0} = \beta_k\lambda_x\frac{\partial\lambda_y}{\partial y_0} \quad J_{z_0} = \frac{\partial\mu'_k}{\partial\sigma_x}\frac{\partial\sigma_x}{\partial z_0} + \frac{\partial\mu'_k}{\partial\sigma_y}\frac{\partial\sigma_y}{\partial z_0}$$

$$\begin{aligned}
J_{x_0} &= \beta_k\lambda_y\frac{\partial\lambda_x}{\partial x_0} \\
&= \frac{\beta_k\lambda_y}{2}\frac{\partial}{\partial x_0}\left(\operatorname{erf}\left(\frac{x_k + \frac{1}{2} - x_0}{\sqrt{2}\sigma_x}\right) - \operatorname{erf}\left(\frac{x_k - \frac{1}{2} - x_0}{\sqrt{2}\sigma_x}\right)\right) \\
&= \frac{\beta_k\lambda_y}{\sqrt{2\pi}\sigma_x}\left(\exp\left(\frac{(x_k - \frac{1}{2} - x_0)^2}{2\sigma_x^2}\right) - \exp\left(\frac{(x_k + \frac{1}{2} - x_0)^2}{2\sigma_x^2}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
J_{y_0} &= \beta_k \lambda_x \frac{\partial \lambda_y}{\partial y_0} \\
&= \frac{\beta_k \lambda_x}{2} \frac{\partial}{\partial y_0} \left(\operatorname{erf}\left(\frac{y_k + \frac{1}{2} - y_0}{\sqrt{2}\sigma_y} \right) - \operatorname{erf}\left(\frac{y_k - \frac{1}{2} - y_0}{\sqrt{2}\sigma_y} \right) \right) \\
&= \frac{\beta_k \lambda_x}{\sqrt{2\pi}\sigma_y} \left(\exp\left(\frac{(y_k - \frac{1}{2} - y_0)^2}{2\sigma_y^2} \right) - \exp\left(\frac{(y_k + \frac{1}{2} - y_0)^2}{2\sigma_y^2} \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
J_{\sigma_x} &= \beta_k \lambda_y \frac{\partial \lambda_x}{\partial \sigma_x} \\
&= \frac{\beta_k \lambda_y}{2} \frac{\partial}{\partial \sigma_x} \left(\operatorname{erf}\left(\frac{x_k + \frac{1}{2} - x_0}{\sqrt{2}\sigma_x} \right) - \operatorname{erf}\left(\frac{x_k - \frac{1}{2} - x_0}{\sqrt{2}\sigma_x} \right) \right) \\
&= \frac{\beta_k \lambda_y}{\sqrt{2\pi}} \left(\frac{\left(x - x_0 - \frac{1}{2} \right) e^{-\frac{\left(x - x_0 - \frac{1}{2} \right)^2}{2\sigma_x^2}}}{\sigma_x^2} - \frac{\left(x - x_0 + \frac{1}{2} \right) e^{-\frac{\left(x - x_0 + \frac{1}{2} \right)^2}{2\sigma_x^2}}}{\sigma_x^2} \right)
\end{aligned}
$$

$$
\begin{aligned}
J_{\sigma_y} &= \beta_k \lambda_x \frac{\partial \lambda_y}{\partial \sigma_y} \\
&= \frac{\beta_k \lambda_x}{2} \frac{\partial}{\partial \sigma_y} \left(\operatorname{erf}\left(\frac{y_k + \frac{1}{2} - y_0}{\sqrt{2}\sigma_y} \right) - \operatorname{erf}\left(\frac{y_k - \frac{1}{2} - y_0}{\sqrt{2}\sigma_y} \right) \right) \\
&= \frac{\beta_k \lambda_x}{\sqrt{2\pi}} \left(\frac{\left(y - y_0 - \frac{1}{2} \right) e^{-\frac{\left(y - y_0 - \frac{1}{2} \right)^2}{2\sigma_y^2}}}{\sigma_y^2} - \frac{\left(y - y_0 + \frac{1}{2} \right) e^{-\frac{\left(y - y_0 + \frac{1}{2} \right)^2}{2\sigma_y^2}}}{\sigma_y^2} \right)
\end{aligned}
$$

Luckily, computing the Hessian matrix for (2.9) is tractable, and is actually quite simple when one takes advantage of the chain rule for Hessian matrices. Looking at (2.9), the likelihood is a hierarchical function that maps a vector space Θ to a vector space Λ to a

scalar value. Formally, we define $T : \Theta \to \Lambda$ and $W : \Lambda \to \mathbb{R}$. The parameter vector $(x_0, y_0, z_0, \sigma_0, N_0) \in \Theta$, the Poisson rate vector $\vec{\lambda} \in \Lambda$ and $\ell \in \mathbb{R}$. Note that we choose to optimize σ_x and σ_y directly and compute z_0 to simplify the computation of the Hessian. To get the Hessian, we need the chain-rule for Hessian matrices, which can be quickly computed in terms of the jacobian and hessian of T and W.

$$H_\ell = J_\mu^T H_\ell J_\mu + (J_\ell \otimes I_n) H_\mu$$

where we have used J_μ to represent the jacobian of T and J_ℓ for the jacobian of W. Similar notation is used for the corresponding Hessian matrices. In the 3D case, the Hessian matrix is not directly separable since $\mu \propto \lambda_x(x_0, \sigma_0, \sigma_x)\lambda_y(y_0, \sigma_0, \sigma_y)$. To see this, an abstract representation of the Hessian reads

.0.1 A Newton-Raphson method for maximum likelihood estimation

Newton-Raphson can be advantageous compared to gradient-descent by taking advantage of second-order derivatives when they are available. However, computing second-order derivatives can be cumbersome, and often intractable in a number of machine-learning contexts.

.0.2 Brief derivation of the master equation

Each row of the generator matrix G_i represents the a conditional probability distribution on the future state given the present state $P(\omega|X(t) = \omega_j)$:

$$\sum_j G_{ij} = \sum_j P(X(t + dt) = \omega_j | X(t) = \omega_i) = 1$$

The generator is not necessarily symmetric. Also, one should note that the columns G_j *do not* define a probability distribution and do not necessarily sum to unity. The columns of G have no meaning in this context, since we have defined the rows to represent a probability

of the future given the present. The first order dynamics for a particular state ω_i is given by

$$P(\omega_i, t + dt) = P(\omega_i, t) + \mathcal{J}_i dt$$

where the probability current $\mathcal{J}_i$ must be the net probability flux into the state ω_i

$$\mathcal{J}_i = \sum_i G_{ij} P(\omega_j, t) - \sum_j G_{ij} P(\omega_i, t)$$

The first is a sum on a column (in) and the second a sum on a row (out). This can be simplified further by noticing that the normalization condition implies

$$G_{ij} = 1 - \sum_j G_{ij}(1 - \delta_{ij}) = 1 - \sum_j G_{ij} + \sum_j G_{ij}\delta_{ij}$$

Inserting this into the probability current gives

$$\begin{aligned}
\mathcal{J}_i &= \sum_i G_{ij} P(\omega_j, t) - \sum_j G_{ij} P(\omega_i, t) \\
&= \sum_i \left(1 - \sum_j G_{ij} + \sum_j G_{ij}\delta_{ij} \right) P(\omega_j, t) - \sum_j G_{ij} P(\omega_i, t) \\
&= |\Omega| - |\Omega| + \sum_i \sum_j G_{ij} P(\omega_j, t)\delta_{ij} - \sum_j G_{ij} P(\omega_i, t) \\
&= \sum_j G_{ji} P(\omega_j, t) - G_{ij} P(\omega_i, t)
\end{aligned}$$

which is RHS of the master equation. Notice that the Kronecker delta effectively just swaps the index.

.0.3 *Fisher information for 2D integrated gaussian*

For the 2D integrated gaussian point spread function, the Hessian only contains separable second order derivatives, so the Fisher information matrix takes on a convenient form

$$I_{ij}(\theta) = \mathop{\mathbb{E}}_{\theta}\left(\frac{\partial \ell}{\partial \theta_i}\frac{\partial \ell}{\partial \theta_j}\right) \tag{1}$$

For an arbitrary parameter then we have

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\sum_k x_k \log x_k + \mu'_k - x_k \log\left(\mu'_k\right) \\ &= \sum_k \frac{\partial \mu'_k}{\partial \theta_i}\left(\frac{\mu'_k - x_k}{\mu'_k}\right)\end{aligned}$$

$$I_{ij}(\theta) = \mathop{\mathbb{E}}_{\theta}\left(\sum_k \frac{\partial \mu'_k}{\partial \theta_i}\frac{\partial \mu'_k}{\partial \theta_j}\left(\frac{\mu'_k - x_k}{\mu'_k}\right)^2\right) = \sum_k \frac{1}{\mu'_k}\frac{\partial \mu'_k}{\partial \theta_i}\frac{\partial \mu'_k}{\partial \theta_j}$$

To compute the bound, it turns out all we need is the jacobian $\frac{\partial \mu'_k}{\partial \theta_j}$.