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 iv

1 INTRODUCTION 1

2 SINGLE MOLECULE LOCALIZATION MICROSCOPY 2
- 2.0.1 Statistics of sCMOS cameras 3
- 2.0.2 Integrated isotropic Gaussian point spread function 5
- 2.0.3 Integrated astigmatic Gaussian point spread function 6
- 2.0.4 Localization microscopy as frequentist inference 7

3 THE NUCLEOSOME: LOST IN PHASE SPACE 9
- 3.1 The Fokker-Planck Equation 9
- 3.2 The Langevin Equation 10

4 DEEP LEARNING FOR LOCALIZATION MICROSCOPY 11

APPENDICES 12
- .0.1 Details of the Gaussian PSF 13
- .0.2 Fisher information for 2D integrated gaussian 15
- .0.3 Kramers-Moyal Expansion 15
- .0.4 The Multivariate Case 19
- .0.5 A Brief Note on Bayesian Nonparametrics 19

ABSTRACT

CHAPTER 1

INTRODUCTION

CHAPTER 2

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 emit-

ters, 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.

2.0.1 Statistics of sCMOS cameras

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

$$\vec{S} = [\text{Poisson}(\mu_1), \text{Poisson}(\mu_2), ..., \text{Poisson}(\mu_N)] \tag{2.1}$$

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:

$$\vec{H} = \vec{S} + \vec{\xi} \tag{2.2}$$

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,

$$P(H_k|\theta) = P(S_k) \circledast P(\xi_k) \tag{2.3}$$

$$= A \sum_{q=0}^{\infty} \frac{1}{q!} e^{-\mu_k} \mu_k^q \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(H_k - g_k q - o_k)}{2\sigma_k^2}} \tag{2.4}$$

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.

2.0.2 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:

$$\mu_k = i_0\lambda_k = i_0 \int_{\text{pixel}} G(x, y)dxdy \tag{2.5}$$

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

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

2.0.3 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

$$\mathrm{G}(x, y) = \frac{1}{2\pi\sigma_x(z)\sigma_y(z)} e^{-\frac{(x-x_0)^2}{2\sigma_x(z_0)^2} + \frac{(y-y_0)^2}{2\sigma_y(z_0)^2}} \tag{2.6}$$

2.0.4 Localization microscopy as frequentist inference

According to our image formation, molecules really do have an exact location in space. In pratice, this is only an approximation since molecules diffuse at physiological temperatures, and our exposure time would need to tend to zero for this to be exactly true. If we suppose that we can collect a sufficient amount of photons in a short enough time, the physical nature of the system suggests a frequentist inference procedure for localization. A natural choice is maximum likelihood estimation.

$$\theta_{\mathrm{MLE}} = \operatorname*{argmax}_{\theta} \prod_k P(H_k|\theta) = \operatorname*{argmin}_{\theta} - \sum_k \log P(H_k|\theta)$$

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

$$\ell(\vec{H}|\theta) = -\log \prod_k \frac{e^{-(\mu'_k)} \left(\mu'_k\right)^{n_k}}{n_k!} \tag{2.7}$$

$$= \sum_k \log n_k! + \mu'_k - n_k \log\left(\mu'_k\right) \tag{2.8}$$

When the log-likelihood is easy to compute, as it is under this approximation, 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 fundamental metric in information geometry.

In frequentist statistics, 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.

CHAPTER 3

THE NUCLEOSOME: LOST IN PHASE SPACE

3.1 The Fokker-Planck Equation

The Fokker-Planck equation is a central tool in non-equilibrium statistical mechanics, analagous to the master equation for discrete systems. It allows us to determine the time evolution of probability densities over continuous state spaces. Important examples in biophysics are the phase space of a particle or the membrane potential of a nerve cell.

Suppose we have a random variable $\boldsymbol{x}$ and its joint distribution $P(\boldsymbol{x}, t)$, which is not necessarily stationary. Define a vector field $\vec{J}(\boldsymbol{x}, t)$ which is the probability current, which we will specify in a moment. The Fokker-Planck equation is by starting with a continuity equation for probability

$$\begin{aligned}\frac{d}{dt}\int_{V_0} P(\boldsymbol{x},t)dV &= \int_S P(\boldsymbol{x},t)(\vec{J}\cdot\hat{n})dS \\ &= -\int_{V_0} P(\boldsymbol{x},t)(\nabla\cdot\vec{J})dV\end{aligned}$$

Clearly this implies that

$$\frac{dP(\boldsymbol{x},t)}{dt} = -\left(\nabla\cdot\vec{J}\right)P(\boldsymbol{x},t)$$

We often call the divergence term, the Fokker-Planck operator $\mathcal{L}_{FP} = -\nabla\cdot\vec{J}$. A more rigorous derivation is given in the appendix, which tells us that, to second order

$$J(x_i,t) = \left(M_i^{(1)}(t) - \sum_j \frac{\partial}{\partial x_j} M_{ij}^{(2)}(t)\right) P(\boldsymbol{x},t)$$

where $M_i^n(t)$ is the nth moment of a transition kernel $T(x_i', t'|x_i, t)$ for variable i. The

first moment is essentially just the deterministic part of the Langevin dynamics. The second and higher moments will depend on these higher moments in the stochastic forcing terms.

3.2 The Langevin Equation

Consider a familiar Langevin dynamics on phase space $\boldsymbol{x} = (x, v)$, where a particle in a potential $V(x)$ experiences a viscous drag force and stochastic forcing $\xi(t)$ where $\xi(t) \sim \mathcal{N}(\mu, \sigma^2)$ and $\langle \xi(t)\xi(t+\tau)\rangle = \delta(t-\tau)$.

$$\begin{aligned}\dot{x} &= v \\ \dot{v} &= -\frac{\gamma}{m}v + \frac{1}{m}F(x) + \frac{1}{m}\xi(t)\end{aligned}$$

The moments of the transition kernel must be

$$M_x^{(1)} = v \;\; M_v^{(1)} = -\frac{\gamma}{m}v + \frac{1}{m}F(x) + \mu \;\; M_v^{(v)} = \sigma^2$$

To simplify the notation let us define $\nabla \cdot \vec{J} = \frac{\partial J_x}{\partial x} + \frac{\partial J_x}{\partial v} = \mathcal{L}_x + \mathcal{L}_v = \mathcal{L}_{FP}$. This gives the full Fokker-Planck equation $\frac{dP(\boldsymbol{x},t)}{dt} = -\mathcal{L}_{FP}P(\boldsymbol{x}, t)$.

$$\begin{aligned}\mathcal{L}_x P(\boldsymbol{x}, t) &= \frac{\partial}{\partial x}\left(vP(\boldsymbol{x}, t)\right) \\ \mathcal{L}_v P(\boldsymbol{x}, t) &= \frac{\partial}{\partial v}\left(-\frac{\gamma}{m}v + \frac{1}{m}F(x)\right)P(\boldsymbol{x}, t) + \sigma^2\frac{\partial^2}{\partial v^2}P(\boldsymbol{x}, t)\end{aligned}$$

CHAPTER 4

DEEP LEARNING FOR LOCALIZATION MICROSCOPY

Appendices

.0.1 Details of the Gaussian PSF

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

.0.2 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 $\frac{\partial \mu'_k}{\partial \theta_j}$.

.0.3 Kramers-Moyal Expansion

Given many instantiations of a stochastic variable x, we can construct a normalize histogram oxer all observations as a function of time $P(x,t)$. However, in order to systematically explore the relationship between the parameterization of the process and $P(x,t)$ we require

an expression for $\dot{P}(x,t)$. If we make a fundamental assumption that the evolution of $P(x,t)$ follows a Markov process i.e. its evolution has the memoryless property, then we can write

$$P(x',t) = \int T(x',t|x,t-\tau)P(x,t-\tau)dx \tag{2}$$

which is known at the Chapman-Kolmogorov equation. The factor $T(x',t|x,t-\tau)$ is known as the *transition operator* in a Markov process and determines the evolution of $P(x,t)$ in time. We proceed by writing $T(x',t|x,t-\tau)$ in a form referred to as the Kramers-Moyal expansion

$$\begin{aligned} T(x',t|x,t-\tau) &= \int \delta(u-x')T(u,t|x,t-\tau)du \\ &= \int \delta(x+u-x'-x)T(u,t|x,t-\tau)du \end{aligned}$$

If we use the Taylor expansion of the δ-function

$$\delta(x+u-x'-x) = \sum_{n=0}^{\infty} \frac{(u-x)^n}{n!}\left(-\frac{\partial}{\partial x}\right)^n \delta(x-x')$$

Inserting this into the result from above, pulling out terms independent of u and swapping the order of the sum and integration gixes

$$T(x',t|x,t-\tau) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\partial}{\partial x}\right)^n \delta(x-x') \int (u-x)^n T(u,t|x,t-\tau)du \tag{3}$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\partial}{\partial x}\right)^n \delta(x-x') M_n(x,t) \tag{4}$$

noticing that $M_n(x,t) = \int (u-x)^n T(u,t|x,t-\tau)du$ is just the nth moment of the

transition operator T. Plugging (2.6) back in to (2.4) gixes

$$P(x,t) = \int \left(1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n M_n(x,t)\right) \delta(x - x') P(x, t-\tau) dx \tag{5}$$

$$= P(x', t-\tau) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{6}$$

Approximating the derixatixe as a finite difference and taking the limit $\tau \to 0$ gixes

$$\dot{P}(x,t) = \lim_{\tau \to 0} \left(\frac{P(x,t) - P(x, t-\tau)}{\tau}\right) \tag{7}$$

$$= \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{8}$$

which is formally known as the Kramers-Moyal (KM) expansion. The Fokker-Planck equation is a special case of (2.10) where we neglect terms $n > 2$ in the *diffusion approximation.*

Consider the following Ito stochastic differential equation

$$d\vec{x} = F(\vec{x}, t) + G(\vec{x}, t) dW$$

The SDE gixen aboxe corresponds to the Kramers-Moyal expansion (KME) of a transition density $T(x', t' | x, t)$ see (Risken 1989) for a full derixation.

$$\frac{\partial P(x,t)}{\partial t} = \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{9}$$

where M_n is the nth moment of the transition density. In the diffusion approximation, the KME becomes the Fokker-Planck equation (FPE) (Risken 1989). For the sake of demonstration, consider the univariate case with random variable x and the form of $T(x', t'|x, t)$ is a Gaussian with mean $\mu(t)$ and variance $\sigma^2(t)$. In this scenario, the FPE applies because $M_n = 0$ for all $n > 2$. Given that the drift $M_1(x, t) = \mu(t)$ and the diffusion $M_2(x, t) = \sigma^2(t)$, the FPE reads

$$\frac{\partial P(x,t)}{\partial t} = \left(-\frac{\partial}{\partial x} M^{(1)}(t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} M^{(2)}(t) \right) P(x,t) \tag{10}$$

We can additionally define the term in parentheses as a differential operator acting on $P(x, t)$

$$\hat{\mathcal{L}}_{FP} = \left(-\frac{\partial}{\partial x} M^{(1)}(t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} M^{(2)}(t) \right) \tag{11}$$

It is common to additionally define the probability current $J(x, t)$ as

$$J(x,t) = \left(M^{(1)}(t) - \frac{1}{2} \frac{\partial}{\partial x} M^{(2)}(t) \right) P(x,t) \tag{12}$$

This definition provides some useful intuition. The value of $J(x, t)$ is the net probability flux into the interval between x and $x + dx$ at at time t. This also allows us to write the FPE as a continuity equation

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x} \tag{13}$$

.0.4 The Multivariate Case

If we now generalize the above equation to a case where we are faced with many variables $\boldsymbol{x} = (x_1, x_2, ..., x_n)$. The continuity equation becomes

$$\frac{\partial P(\vec{x}, t)}{\partial t} = -\vec{\nabla} \cdot J(\vec{x}, t) \tag{14}$$

where the multivariate probability current now has the interpretation of the net flux into or out of a volume dx^n centered around $\boldsymbol{x}$. If we consider each dimension,

$$J(x_i, t) = \left(M_i^{(1)}(t) - \sum_j \frac{\partial}{\partial x_j} M_{ij}^{(2)}(t) \right) P(\vec{x}, t) \tag{15}$$

The full Fokker-Planck equation then reads

$$\frac{\partial P(\vec{x}, t)}{\partial t} = \vec{\nabla} \cdot J(\vec{x}, t) \tag{16}$$

$$= \sum_{i=1}^{N} \left(-\frac{\partial}{\partial x_i} M_i^{(1)}(t) + \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j} M_{ij}^{(2)}(t) \right) P(\vec{x}, t) \tag{17}$$

.0.5 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.