Addendum to
"Stochastic Models, Information Theory, and Lie Groups, Vol. 1"
by G. S. Chirikjian
Birkhäuser, 2009

Version: 02/01/10

Contents

A S.M., I.T., L. G. © gregc@jhu.edu 02/01/10 5

A.1 General Questions about Notation 5

A.1.1 The Notation for a Parametric Family 5

A.1.2 Deviations from Standard Information-Theory Notation . 7

A.2 Addendum to Chapter 3: Probability and Information Theory . . 7

A.2.1 Conditional Expectation 7

A.2.2 Fisher Information and Convolution without Conditional Expectation 10

A.2.3 When Is Fisher Information Divergence Invariant Under Coordinate Changes ? 11

A.2.4 Variance and Entropy Powers 12

A.2.5 An Additional Closed-Form Density 13

A.3 Addendum to Chapter 4: Stochastic Processes 14

A.3.1 General Derivation of the Chapman-Kolmogorov Equation 14

A.3.2 Proofs of the Itô Fokker-Planck Equation 15

A.3.3 Integration over First vs. Second Argument of Conditional Probabilities 28

A.3.4 Inter-conversion between Itô and Stratonovich 29

A.3.5 SDEs in Cartesian and Polar Coordinates in the Plane . . 32

A.4 The Heat Equation in Curvilinear Coordinates 32

A.5 Addendum to Chapter 5 37

A.5.1 Increasing the Stability of Inverse Jacobian Iterations on pp. 150-1 37

A.5.2 Computing Normal Curvature 38

A.6 Addendum to Chapter 6 39

A.6.1 An Intuitive Introduction to Push-Forward Vector Fields 39

A.6.2 The Lagrange Identity 41

A.7 Addendum to Chapter 7 44

A.7.1 2D Manifolds Generated by Dividing the Plane by Groups 44

A.8 Addendum to Chapter 8: Stochastic Processes on Manifolds . . . 46

A.8.1 Compatibility of Stratonovich SDEs in Different Coordinate Charts 46

A.8.2 Compatibility of Itô SDEs in Different Coordinate Charts 48

A.8.3 Itô SDEs on Implicitly Defined Embedded Manifolds . . . 49

A.8.4 Example 1: An Itô SDE on the Circle 49
A.8.5 Example 2: An Itô SDE on the Sphere 50
A.8.6 Example 3: An Itô SDE on the Rotation Group 50
A.8.7 Example 4: A Class of Stratonovich SDEs on the Rotation Group 53
A.9 Supplemental Exercises 54

Appendix A

S.M., I.T., L. G. © gregc@jhu.edu 02/01/10

These notes are meant to clarify and accentuate certain points in the book "Stochastic Models, Information Theory, and Lie Groups. Vol. 1" by G. S. Chirikjian [3]. These notes are not meant for stand-alone use, as many definitions and symbols are defined in the book. Volume 1 of this two-volume set was published by Birkhäuser in 2009 and is available from the publisher as well as outlets such as amazon.com. Volume 2 is expected to appear January 2011. A separate addendum to Vol. 2 will be posted after it is published. These addenda will be updated periodically to include new material. Contact the author via email gregc@jhu.edu to report any errors. An up-to-date page of errata also is being maintained at the webpage of the author's lab.

The notes that follow are not meant as stand-alone course materials. For context and definition of notation see Volume 1.

A.1 General Questions about Notation

Some nonstandard notation was used in Volume 1. In some cases this was because of notational clashes that occurred when trying to span discussions across multiple fields that assign different meaning to the same symbols. In other cases it was to simplify the presentation. The main deviations from field-specific standard notation are summarized in this section.

A.1.1 The Notation for a Parametric Family

In this book $f(\mathbf{x};\boldsymbol{\theta})$ refers to a probability density function in the variable $\mathbf{x} \in D \subset \mathbb{R}^n$ *parameterized by* $\boldsymbol{\theta} \in B \subset \mathbb{R}^m$. It is assumed that D is a measurable space, but this need not always be the case for B. In many books this same pdf is written as $f(\mathbf{x}\,|\,\boldsymbol{\theta})$, indicating that this is a density in $\mathbf{x}$ *conditioned on*

$\boldsymbol{\theta}$. In order for the concept of conditioning to make sense, the space of values $\boldsymbol{\theta}$ should be measurable and a prior density $f(\boldsymbol{\theta})$ should exist such that

$$f(\mathbf{x},\boldsymbol{\theta}) = f(\mathbf{x}\,|\,\boldsymbol{\theta})f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}\,|\,\mathbf{x})f(\mathbf{x})$$

is a joint density on $D \times B \subset \mathbb{R}^{n+m}$.

If such a prior exists, then[1]

$$f(\mathbf{x}) = \int_{\boldsymbol{\theta}'\in\mathbb{R}^m} f(\mathbf{x}\,|\,\boldsymbol{\theta}')f(\boldsymbol{\theta}')d\boldsymbol{\theta}'$$

and

$$f(\boldsymbol{\theta}\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'\in\mathbb{R}^m} f(\mathbf{x}\,|\,\boldsymbol{\theta}')f(\boldsymbol{\theta}')d\boldsymbol{\theta}'}. \tag{A.1}$$

Therefore, in contexts when Bayesian calculations such as (A.1) are *not* calculated (which is most of Volume 1), parameterized families of pdfs are denoted as $f(\mathbf{x};\boldsymbol{\theta})$. A specific example of this are the multivariate Gaussian distribution,

$$\rho_{(\boldsymbol{\mu},\Sigma)}(\mathbf{x}) = \rho(\mathbf{x};\boldsymbol{\mu},\Sigma) \doteq \frac{1}{(2\pi)^{n/2}|\det\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Unless one is ready to discuss integration measures on the space of values of the form $(\boldsymbol{\mu},\Sigma)$, it is premature to denote the Gaussian as $\rho(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma)$ (which is how it is often denoted).

In addition to static Gaussian distributions, the issue of whether or not to use the conditional symbol arises in the study of diffusion processes. Solutions to Fokker-Planck equations such as[2]

$$\frac{\partial f}{\partial t} = -\sum_{i=1}^{n} a_i \frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{n} B_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} \quad \text{where} \quad f(\mathbf{x},0) = \delta(\mathbf{x})$$

are generally written as $f(\mathbf{x},t)$. The latter could be denoted as $f(\mathbf{x};t)$ (or $f(\mathbf{x};\mathbf{a},B,t)$), but since $\mathbf{x}\in\mathbb{R}^n$ is the only spatial variable, there is no confusion about mistaking $f(\mathbf{x},t)$ for a joint density on $\mathbb{R}^n\times\mathbb{R}_{\geq 0}$. In this context it would be confusing to use the notation $f(\mathbf{x}\,|\,t)$ in place of $f(\mathbf{x};t)$, since one often talks about conditioning on prior values of $\mathbf{x}$ (e.g., in discussions of Markov processes the notation $f(\mathbf{x}_1\,|\,\mathbf{x}_2)$ or $f(\mathbf{x}_1,t_1\,|\,\mathbf{x}_2,t_2)$ is used). Here the t_i's are not part of the domain on which the conditioning is being performed. And to say that the pdf is in addition conditioned on a specific instant of time would lead to too many conditioning symbols. This is one of a number of notational issues that occur when passing from discrete to continuous time. The latter is the emphasis in Volume 1, as this is the version that is most relevant to modeling physical phenomena such as Brownian motion.

[1] Each of the densities $f(\mathbf{x},\boldsymbol{\theta})$, $f(\mathbf{x})$, $f(\boldsymbol{\theta})$ can be extended over the whole of $\mathbb{R}^{n+m}$,$\mathbb{R}^n$,$\mathbb{R}^m$, respectively, by assigning them the value zero outside of $D\times B$, D, B, respectively.

[2] It was shown in Chapter 2 that in fact the solutions $f(\mathbf{x},t)$ to Fokker-Planck equations with constant coefficients $\mathbf{a}, B$ is a Gaussian.

Having said all this, in Volume 2, we will in fact discuss integration measures on spaces of symmetric positive definite matrices, and so it will make sense to define probability densities on spaces of covariance matrices (e.g., the Wishart distribution), and hence $\rho(\mathbf{x} \,|\, \boldsymbol{\mu}, \Sigma)$ will make sense in that context. But rather than leaving the reader wondering for hundreds of pages why a conditional symbol is used in the definition of the Gaussian, the innocuous semicolon was used.

A.1.2 Deviations from Standard Information-Theory Notation

The standard symbols for entropy, mutual information, etc., were not used in Volume 1. For example, the standard notation for the entropy of a random variable X is $H(X)$. In Volume 1, the notation used was $S(f)$ where $f(x)$ is the probability density function (pdf) corresponding to X. This notation reflects the computational nature of the book. Namely, if one wants to know the numerical value of entropy in some scenario, substitute $f(x)$ into the integral

$$S(f) = -\int_x f(x) \log f(x) dx.$$

The notation $S(f)$ also provides a degree of consistency with other definitions in information theory. For example, in the scalar case the Fisher information is

$$F(f) = \int_x \frac{1}{f} \left(\frac{df}{dx} \right)^2 dx$$

and the Kullback-Leibler divergence is

$$D_{KL}(f_1 \,\|\, f_2) = \int_x f_1(x) \log \frac{f_1(x)}{f_2(x)} dx.$$

The pdfs used in computing them are explicit in the symbols $F(f)$ and $D_{KL}(f_1 \,\|\, f_2)$. Why should f be hidden from sight when it comes to denoting entropy or mutual information ?

A.2 Addendum to Chapter 3: Probability and Information Theory

A.2.1 Conditional Expectation

This section serves as an addendum to p. 73 of Vol. 1, in which conditional expectation is discussed. The concept and notation for conditional expectation $\langle \cdot | \cdot \rangle$ are explained in the text. The usual expectation of a function $\phi(x_1, x_2)$ using a probability density $f(x_1, x_2)$ is

$$\langle \phi \rangle \doteq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x_1, x_2) f(x_1, x_2) dx_1 dx_2.$$

In contrast, the *conditional expectation* of any function $\phi(x_1)$ given x_2 is

$$\langle \phi(x_1) | x_2 \rangle \doteq \frac{1}{f_2(x_2)} \int_{-\infty}^{\infty} \phi(x_1) f(x_1, x_2) dx_1.$$

Here, of course,

$$f_2(x_2) = \int_{-\infty}^{\infty} \phi(x_1) f(x_1, x_2) dx_1.$$

This concept extends to higher dimensions (and even non-Euclidean settings) in a natural way. For example, given $\phi(x_1, x_2)$ and $f(x_1, x_2, x_3, x_4)$,

$$f_{3,4}(x_3, x_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_1 dx_2.$$

and

$$\langle \phi(x_1, x_2) | x_3, x_4 \rangle \doteq \frac{1}{f_{3,4}(x_3, x_4)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x_1, x_2) f(x_1, x_2, x_3, x_4) dx_1 dx_2.$$

Note that in the statement of a conditional expectation, such as one of the form $\langle \phi(x_1, x_2) | x_3, x_4 \rangle$, the master pdf used to perform the computation (in this case $f(x_1, x_2, x_3, x_4)$) is not explicitly stated. But its existence is necessary in order to perform the computation.

In (3.25), the following computation is performed

$$\begin{aligned}
\left(\frac{f_2'(x_2)}{f_2(x_2)}\right)^2 &= \left(\beta \left\langle \frac{\rho_1'(u)}{\rho_1(u)} \right| x_2 \right\rangle + (1-\beta) \left\langle \frac{\rho_2'(v)}{\rho_2(v)} \right| x_2 \right\rangle \Bigg)^2 \\
&= \left(\left\langle \beta \frac{\rho_1'(u)}{\rho_1(u)} + (1-\beta) \frac{\rho_2'(v)}{\rho_2(v)} \right| x_2 \right\rangle \Bigg)^2 && \text{(A.2)} \\
&\leq \left\langle \left(\beta \frac{\rho_1'(u)}{\rho_1(u)} + (1-\beta) \frac{\rho_2'(v)}{\rho_2(v)} \right)^2 \right| x_2 \right\rangle . && \text{(A.3)}
\end{aligned}$$

$$\begin{aligned}
\left\langle \left(\frac{f_2'(x_2)}{f_2(x_2)}\right)^2 \right\rangle &\leq \left\langle \left\langle \left(\beta \frac{\rho_1'(u)}{\rho_1(u)} + (1-\beta) \frac{\rho_2'(v)}{\rho_2(v)} \right)^2 \right| x_2 \right\rangle \right\rangle && \text{(A.4)} \\
&\leq \left\langle \left(\beta \frac{\rho_1'(u)}{\rho_1(u)} + (1-\beta) \frac{\rho_2'(v)}{\rho_2(v)} \right)^2 \right\rangle && \text{(A.5)} \\
&= \beta^2 \left\langle \left(\frac{\rho_1'(u)}{\rho_1(u)} \right)^2 \right\rangle + (1-\beta)^2 \left\langle \left(\frac{\rho_2'(v)}{\rho_2(v)} \right)^2 \right\rangle . && \text{(A.6)}
\end{aligned}$$

Since $f_2(x) = (\rho_1 * \rho_2)(x)$, it follows that

$$\boxed{F(\rho_1 * \rho_2) \leq \beta^2 F(\rho_1) + (1-\beta)^2 F(\rho_2)} \tag{A.7}$$

where $F(f)$ is the Fisher information of f. Optimizing over β gives

$$\boxed{F(\rho_1 * \rho_2) \leq \frac{F(\rho_1) \cdot F(\rho_2)}{F(\rho_1) + \cdot F(\rho_2)}} \tag{A.8}$$

Both the inequalities that are introduced in going from (A.2) to (A.3) and from (A.4) to (A.5) result from the application of Jensen's inequality for conditional expectation (3.24). However, since the conditional expectation of $\rho_1'(u)/\rho_1(u)$ is computed using the joint density $\rho_1(u)\rho_2(x_2 - u)$ and the conditional expectation of $\rho_2'(v)/\rho_2(v)$ is computed using the joint density $\rho_2(v)\rho_1(x_2 - v)$, care must be taken when combining these since the angle brackets hide the pdfs with which expectation is being computed. When both terms are combined resulting in (A.2), the expectation must be computed with respect to a trivariate pdf $F(u, v, x_2)$ with marginals

$$\begin{aligned}
\rho_2(v)\rho_1(x_2 - v) &= \int_{-\infty}^{\infty} F(u, v, x_2)du && \text{(A.9)}\\
\rho_1(u)\rho_2(x_2 - u) &= \int_{-\infty}^{\infty} F(u, v, x_2)dv && \text{(A.10)}\\
\rho_1(u)\rho_2(v) &= \int_{-\infty}^{\infty} F(u, v, x_2)dx_2 && \text{(A.11)}
\end{aligned}$$

(The last of these is simply a statement that the original random variables are independent. It is this marginal density that is used to compute the expectation in (A.5), and which separates to give the expectations with respect to u and v separately in the two terms in (A.6).) While not stated in the theorem of Blachman, the existence of such a pdf is necessary for this step to be valid.

It turns out that such an F can be constructed. And while it will exist (which is all we care about), it may not be unique. The general conditions under which the existence of trivariate pdfs are guaranteed that have certain marginals is addressed in [8] (provided to the author by Prof. Jim Fill). And in our particular problem it can be defined in terms of its Fourier transform as [5]

$$\hat{F}(\omega_u, \omega_v, \omega_{x_2}) = \hat{\rho}_1(\omega_u + \omega_{x_2})\hat{\rho}_2(\omega_v + \omega_{x_2}). \tag{A.12}$$

A.2.2 Fisher Information and Convolution without Conditional Expectation

Suppose we start the same as before, but do not use conditional expectation involving the joint distribution in u, v and x_2. Instead, just complete the square:

$$\begin{aligned}
\left(\frac{f_2'(x_2)}{f_2(x_2)}\right)^2 &= \left(\beta\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle + (1-\beta)\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle\right)^2 \\
&= \beta^2\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle^2 + (1-\beta)^2\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle^2 \\
&\quad +2\beta(1-\beta)\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle\cdot\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle
\end{aligned}$$

Now if we compute expectations we get

$$\begin{aligned}
\left\langle\left(\frac{f_2'(x_2)}{f_2(x_2)}\right)^2\right\rangle &= \beta^2\left\langle\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle^2\right\rangle + (1-\beta)^2\left\langle\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle^2\right\rangle \\
&\quad +2\beta(1-\beta)\left\langle\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle\cdot\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle\right\rangle \\
&\leq \beta^2\left\langle\left(\frac{\rho_1'(u)}{\rho_1(u)}\right)^2\right\rangle + (1-\beta)^2\left\langle\left(\frac{\rho_2'(v)}{\rho_2(v)}\right)^2\right\rangle \\
&\quad +2\beta(1-\beta)\left\langle\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle\cdot\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle\right\rangle
\end{aligned}$$

But we already established in Vol 1. that

$$\frac{f_2'(x_2)}{f_2(x_2)} = \left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle = \left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle$$

and so

$$\left\langle\left\langle \frac{\rho_1'(u)}{\rho_1(u)}\middle| x_2\right\rangle\cdot\left\langle \frac{\rho_2'(v)}{\rho_2(v)}\middle| x_2\right\rangle\right\rangle = \left\langle\left(\frac{f_2'(x_2)}{f_2(x_2)}\right)^2\right\rangle.$$

Therefore, bringing the cross term over to the other side, we have

$$[(1-\beta)^2+\beta^2]\left\langle\frac{f_2'(x_2)}{f_2(x_2)}\right\rangle \leq \beta^2\left\langle\left(\frac{\rho_1'(u)}{\rho_1(u)}\right)^2\right\rangle + (1-\beta)^2\left\langle\left(\frac{\rho_2'(v)}{\rho_2(v)}\right)^2\right\rangle,$$

or equivalently,

$$F(f_2) \leq \frac{\beta^2 F(\rho_1)}{(1-\beta)^2+\beta^2} + \frac{(1-\beta)^2 F(\rho_2)}{(1-\beta)^2+\beta^2}.$$

Comparing with (A.7) this is not the same. Choosing

$$\frac{\beta^2}{(1-\beta)^2+\beta^2} = \frac{F(\rho_2)}{F(\rho_1)+F(\rho_2)} \quad \text{and} \quad \frac{(1-\beta)^2}{(1-\beta)^2+\beta^2} = \frac{F(\rho_1)}{F(\rho_1)+F(\rho_2)}$$

then gives the inequality

$$F(f_2) \leq \frac{2 \cdot F(\rho_1) \cdot F(\rho_2)}{F(\rho_1)+F(\rho_2)} \tag{A.13}$$

Notice the factor of 2 here that does not exist in (A.8). Therefore, this bound is not as tight, and conditional expectation played a key role in obtaining the tighter bound.

A.2.3 When Is Fisher Information Divergence Invariant Under Coordinate Changes ?

In the middle of page 77 the chain-rule equation that is written as

$$\nabla^T_{\boldsymbol{\phi}} \tilde{f}(\boldsymbol{\phi}) = \left.(\nabla^T_{\mathbf{x}} f(\mathbf{x}))\right|_{\mathbf{x}(\boldsymbol{\phi})} J(\boldsymbol{\phi})$$

actually should be

$$\nabla^T_{\boldsymbol{\phi}} (f(\mathbf{x}(\boldsymbol{\phi}))) = \left.(\nabla^T_{\mathbf{x}} f(\mathbf{x}))\right|_{\mathbf{x}(\boldsymbol{\phi})} J(\boldsymbol{\phi}). \tag{A.14}$$

This error affects the resulting equations.

Since $\tilde{f}(\boldsymbol{\phi})$ is defined earlier on that page as

$$\tilde{f}(\boldsymbol{\phi}) = f(\mathbf{x}(\boldsymbol{\phi}))|J(\boldsymbol{\phi})|$$

it follows from the product rule that

$$\nabla^T_{\boldsymbol{\phi}} \tilde{f}(\boldsymbol{\phi}) = \nabla^T_{\boldsymbol{\phi}} \left\{ f(\mathbf{x}(\boldsymbol{\phi}))|J(\boldsymbol{\phi})| \right\} = \nabla^T_{\boldsymbol{\phi}} \left\{ f(\mathbf{x}(\boldsymbol{\phi})) \right\} |J(\boldsymbol{\phi})| + f(\mathbf{x}(\boldsymbol{\phi})) \nabla^T_{\boldsymbol{\phi}} |J(\boldsymbol{\phi})|.$$

Therefore

$$\frac{1}{\tilde{f}_1(\boldsymbol{\phi})} \nabla^T_{\boldsymbol{\phi}} \tilde{f}_1(\boldsymbol{\phi}) - \frac{1}{\tilde{f}_2(\boldsymbol{\phi})} \nabla^T_{\boldsymbol{\phi}} \tilde{f}_2(\boldsymbol{\phi}) =$$

$$\frac{1}{f_1(\mathbf{x}(\boldsymbol{\phi}))} \nabla^T_{\boldsymbol{\phi}} \left\{ f_1(\mathbf{x}(\boldsymbol{\phi})) \right\} + \frac{\nabla^T_{\boldsymbol{\phi}} |J(\boldsymbol{\phi})|}{|J(\boldsymbol{\phi})|} - \frac{1}{f_2(\mathbf{x}(\boldsymbol{\phi}))} \nabla^T_{\boldsymbol{\phi}} \left\{ f_2(\mathbf{x}(\boldsymbol{\phi})) \right\} - \frac{\nabla^T_{\boldsymbol{\phi}} |J(\boldsymbol{\phi})|}{|J(\boldsymbol{\phi})|} =$$

$$\frac{1}{f_1(\mathbf{x}(\boldsymbol{\phi}))} \nabla^T_{\boldsymbol{\phi}} \left\{ f_1(\mathbf{x}(\boldsymbol{\phi})) \right\} - \frac{1}{f_2(\mathbf{x}(\boldsymbol{\phi}))} \nabla^T_{\boldsymbol{\phi}} \left\{ f_2(\mathbf{x}(\boldsymbol{\phi})) \right\} =$$

$$\left(\frac{1}{f_1(\mathbf{x}(\boldsymbol{\phi}))} \left.(\nabla^T_{\mathbf{x}} f_1(\mathbf{x}))\right|_{\mathbf{x}(\boldsymbol{\phi})} - \frac{1}{f_2(\mathbf{x}(\boldsymbol{\phi}))} \left.(\nabla^T_{\mathbf{x}} f_2(\mathbf{x}))\right|_{\mathbf{x}(\boldsymbol{\phi})} \right) J(\boldsymbol{\phi}).$$

The first few equalities above are from direct substitution and the last equality uses (A.14).

What this means is that

$$D_{FI}(\tilde{f}_1 \,\|\, \tilde{f}_2) =$$

$$\int_{\boldsymbol{\phi}\in\mathbb{R}^n} \left\| \left(\frac{1}{f_1(\mathbf{x}(\boldsymbol{\phi}))} \left. (\nabla_{\mathbf{x}}^T f_1(\mathbf{x})) \right|_{\mathbf{x}(\boldsymbol{\phi})} - \frac{1}{f_2(\mathbf{x}(\boldsymbol{\phi}))} \left. (\nabla_{\mathbf{x}}^T f_2(\mathbf{x})) \right|_{\mathbf{x}(\boldsymbol{\phi})} \right) J(\boldsymbol{\phi}) \right\|^2 f_1(\mathbf{x}(\boldsymbol{\phi}))|J(\boldsymbol{\phi})|d\boldsymbol{\phi}$$

and since $d\mathbf{x} = |J(\boldsymbol{\phi})|d\boldsymbol{\phi}$

$$D_{FI}(\tilde{f}_1 \,\|\, \tilde{f}_2) = D_{FI}(f_1 \,\|\, f_2)$$

if

$$J(\boldsymbol{\phi})J^T(\boldsymbol{\phi}) = \mathbb{I}$$

which is a little different than what is written.

A.2.4 Variance and Entropy Powers

The following relationship between entropy and variance is well-known. See, for example, [4]. Let $\hat{x}$ be an estimate of the mean of any pdf $f(x)$. Then

$$\begin{aligned} \langle (x-\hat{x})^2 \rangle &\geq \min_{\hat{x}\in\mathbb{R}} \langle (x-\hat{x})^2 \rangle \\ &= \langle (x-\mu_f)^2 \rangle \\ &= \sigma_f^2. \end{aligned}$$

On the other hand, we know that over all pdfs with mean μ_f and variance σ_f^2, the Gaussian is the one with maximum entropy. The entropy of a Gaussian with mean μ_f and variance σ_f^2 is $S(\rho_{\mu_f,\sigma_f^2}) = \frac{1}{2}\log(2\pi e\sigma_f^2)$. Since this entropy will be greater than $S(f)$, and since the function e^x is strictly increasing, we also have that $N(f) \leq N(\rho_{\mu_f,\sigma_f^2})$ where $N(f) = \exp(2S(f))/2\pi e$ is the entropy power of f. Therefore,

$$N(\rho_{\mu_f,\sigma_f^2}) = \frac{1}{2\pi e}\exp\left(2\cdot\frac{1}{2}\log(2\pi e\sigma_f^2)\right) = \sigma_f^2 \geq N(f).$$

Combining this with the inequalities above gives that the variance of any estimator of the mean (including the true variance itself) satisfies

$$\langle (x-\hat{x})^2 \rangle \geq N(f). \tag{A.15}$$

We can make some statements in the multidimensional case also. Let Σ_f be the covariance of a pdf $f(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^n$. Then the entropy power $N(f)$ will always be less than or equal to the entropy power of the multivariate Gaussian with covariance Σ_f. That is, if $|\Sigma_f|$ denotes the determinant of Σ_f,

$$N(f) \leq N(\rho_{\mu_f,\Sigma_f})$$

where

$$S(\rho_{\mu_f,\Sigma_f}) = \frac{1}{2}\log((2\pi e)^n|\Sigma_f|) = \frac{n}{2}\log\left((2\pi e)|\Sigma_f|^{\frac{1}{n}}\right)$$

and

$$N(\rho_{\mu_f,\Sigma_f}) = \frac{1}{2\pi e}\exp\left[\frac{2}{n}S(f)\right] = |\Sigma_f|^{\frac{1}{n}}.$$

Therefore,

$$N(f) \leq |\Sigma_f|^{\frac{1}{n}}.$$

This can be written in a form analogous to (A.15) by observing the algebraic-goemetric mean inequality (which holds for any set of positive real numbers):

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i \geq \left(\prod_{i=1}^{n}\lambda_i\right)^{\frac{1}{n}}$$

and substituting in for each λ_i the eigenvalues of Σ_f. This then gives

$$\frac{1}{n}\text{tr}(\Sigma_f) \geq |\Sigma_f|^{\frac{1}{n}}.$$

Then, mimicking the rest of the proof in the one-dimensional case,

$$\frac{1}{n}\text{tr}\left\langle(\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^T\right\rangle \geq N(f). \tag{A.16}$$

Furthermore, all k-dimensional marginals of the pdf $f(\mathbf{x})$ must satisfy this expression also, with k replacing n.

A.2.5 An Additional Closed-Form Density

The Gaussian distribution was investigated extensively in Chapter 2. Since the Gaussian is very special (e.g., it is the maximum entropy distribution under variance constraints, it satisfies the equality case of the entropy power inequality in the case when the two Gaussians have covariances that are scalar multiples of each other, is completely determined by its first two moments, etc.) it can be instructive to have other smooth pdfs that are not so special, to illustrate theorems in more general cases. In this section some of the properties of a class of smooth non-Gaussian pdfs are listed.

Let

$$f(x) = [a + bx^2]e^{-cx^2}.$$

If $a, b, c > 0$ and the integral of $f(x)$ over the real line is constrained to be 1, then this will be a pdf. Explicitly this is

$$\int_{-\infty}^{\infty}(a+bx^2)e^{-cx^2}dx = a\left(\frac{\pi}{c}\right)^{\frac{1}{2}} + \frac{\sqrt{\pi}}{2}\frac{b}{c^{\frac{3}{2}}} \doteq 1. \tag{A.17}$$

Furthermore, using integration by parts,

$$\int_{-\infty}^{\infty}x^2(a+bx^2)e^{-cx^2}dx = \frac{\sqrt{\pi}}{2}\frac{a}{c^{\frac{3}{2}}} + \frac{3\sqrt{\pi}}{4}\frac{b}{c^{\frac{5}{2}}} \doteq \sigma^2. \tag{A.18}$$

This means that we can define a pdf of the form

$$f_{(c,\sigma^2)}(x) \doteq [a(c,\sigma^2) + b(c,\sigma^2)x^2]e^{-cx^2} \tag{A.19}$$

where

$$b(c,\sigma^2) = \frac{2\sigma^2 c^{\frac{5}{2}} - c^{\frac{3}{2}}}{\pi^{\frac{1}{2}}} \quad \text{and} \quad a(c,\sigma^2) = \frac{\frac{3}{2}c^{\frac{1}{2}} - \sigma^2 c^{\frac{3}{2}}}{\pi^{\frac{1}{2}}}$$

A.3 Addendum to Chapter 4: Stochastic Processes

A.3.1 General Derivation of the Chapman-Kolmogorov Equation

Here a detailed derivation of the Chapman-Kolmogorov equation is given. The dependence on $t_1, ..., t_n$ is suppressed for notational convenience.

In general

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) &= p(\mathbf{x}_1 \,|\, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) p(\mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) \\ &= p(\mathbf{x}_1 \,|\, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) p(\mathbf{x}_2 \,|\, \mathbf{x}_3, ..., \mathbf{x}_n) p(\mathbf{x}_3, ..., \mathbf{x}_n). \end{aligned}$$

Dividing by $p(\mathbf{x}_3, ..., \mathbf{x}_n)$, in general we can write

$$p(\mathbf{x}_1 \,|\, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) p(\mathbf{x}_2 \,|\, \mathbf{x}_3, ..., \mathbf{x}_n) = \frac{p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n)}{p(\mathbf{x}_3, ..., \mathbf{x}_n)} = p(\mathbf{x}_1, \mathbf{x}_2 \,|\, \mathbf{x}_3, ..., \mathbf{x}_n).$$

In the special case of a Markov process, this expression reduces to

$$p(\mathbf{x}_1 \,|\, \mathbf{x}_2) p(\mathbf{x}_2 \,|\, \mathbf{x}_3) = p(\mathbf{x}_1, \mathbf{x}_2 \,|\, \mathbf{x}_3) = p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)/p(\mathbf{x}_3).$$

Integrating both sides over $\mathbf{x}_2$ then gives

$$\int_{\mathbb{R}^d} p(\mathbf{x}_1 \,|\, \mathbf{x}_2) p(\mathbf{x}_2 \,|\, \mathbf{x}_3) d\mathbf{x}_2 = \frac{p(\mathbf{x}_1, \mathbf{x}_3)}{p(\mathbf{x}_3)} = p(\mathbf{x}_1 \,|\, \mathbf{x}_3), \tag{A.20}$$

which is the Chapman-Kolmogorov equation for Markov processes in $\mathbb{R}^d$. It is easy to see that the restriction of the discussion to Euclidean space for the sake of concreteness is artificial in that pdfs on any domain on which an integration measure can be defined could have been used.

The Special Case of Stationary Processes

Now, if we reintroduce the dependence on time, the result of the previous subsection can be written as

$$\int_{\mathbb{R}^d} p(\mathbf{x}_1, t_1 \,|\, \mathbf{x}_2, t_2) p(\mathbf{x}_2, t_2 \,|\, \mathbf{x}_3, t_3) d\mathbf{x}_2 = p(\mathbf{x}_1, t_1 \,|\, \mathbf{x}_3, t_3). \tag{A.21}$$

A somewhat confusing thing to remember is that the tradition in the study of Markov processes is that $t_i > t_j$ when $i < j$ and therefore $\mathbf{x}_i$ is a more recent value of $\mathbf{x}$ than $\mathbf{x}_j$. In other words, as indices on the time variable increase, they denote times *further in the past.* This is the *opposite* of the convention used for stochastic differential equations.

For a stationary process,

$$p(\mathbf{x}_1, t_1 \,|\, \mathbf{x}_2, t_2) = p(\mathbf{x}_1, t_1 - t_2 \,|\, \mathbf{x}_2, 0) \doteq p(\mathbf{x}_1 \,|\, \mathbf{x}_2, t_1 - t_2)$$

and so using this definition on all terms in (A.21) gives

$$\int_{\mathbb{R}^d} p(\mathbf{x}_1 \,|\, \mathbf{x}_2, t_1 - t_2) p(\mathbf{x}_2 \,|\, \mathbf{x}_3, t_2 - t_3) d\mathbf{x}_2 = p(\mathbf{x}_1 \,|\, \mathbf{x}_3, t_1 - t_3). \qquad \text{(A.22)}$$

If we let $t_1 - t_2 = s$ and $t_2 - t_3 = t - s$ then $t_1 - t_3 = t$ and (4.16) results.

A.3.2 Proofs of the Itô Fokker-Planck Equation

In (4.58) and (4.59) the integrations over $\mathbf{y}$ should have been integrations over $\mathbf{x}$, and likewise the integration against $\epsilon(\mathbf{y})d\mathbf{y}$ should have been against $\epsilon(\mathbf{x})d\mathbf{x}$ This can be fixed by either: (a) swapping the roles of $\mathbf{x}$ and $\mathbf{y}$ in the conditional probability, (b) replacing the notation $p(\mathbf{x}|\mathbf{y}, t)$ with $p(\mathbf{x} \to \mathbf{y}, t)$ (which is equivalent to (a)), or (c) integrating over $\mathbf{x}$ and against $\epsilon(\mathbf{x})d\mathbf{x}$ rather than what was done.

In the following four subsections alternative proofs that make slightly different notational choices are presented. In the first, the Fokker-Planck equation is derived from Itô's rule rather than using the Chapman-Kolmogorov equation. In the second re-derivation, as in the book, $\mathbf{x}$ occurs after $\mathbf{y}$. The formulation is also modified by using transition probabilities that are not necessarily invariant under shifts in time. In the third re-derivation, $\mathbf{y}$ occurs after $\mathbf{x}$ (as in the book). In both of these two, the temporal relationship is allowed to change: in some instances it is dt, in others it is t, and in others it is $t + dt$. In the fourth rederivation, the temporal relationships are fixed. Before getting to those, some background is reviewed.

Recall that in general $p(\mathbf{a} \,|\, \mathbf{b}, \tau)$ denotes the probability of transitioning *from* $\mathbf{b}$ to $\mathbf{a}$ in a period of time $\tau > 0$ in the special case when $p(\mathbf{a}, t + \tau \,|\, \mathbf{b}, t)$ is independent of t. In some books this is written in a different notation as $p(\mathbf{b} \to \mathbf{a}, \tau)$, which adds to the confusion about the positions of the arguments.

The actual values of time do not matter, only their difference does. So, for example, it does not matter whether $\mathbf{a} = \mathbf{x}(t)$ and $\mathbf{b} = \mathbf{x}(t - dt)$ or $\mathbf{a} = \mathbf{x}(t + dt)$ and $\mathbf{b} = \mathbf{x}(t)$. The value of τ in $p(\mathbf{a} \,|\, \mathbf{b}, \tau)$ parameterizes the resulting conditional pdf in the position of the trajectory τ units of time after it was observed being at $\mathbf{b}$. *Whereas the time difference τ is relative, the position variables $\mathbf{a}$ and $\mathbf{b}$ are absolute.* The probabilistic relationship between $\mathbf{a}$ and $\mathbf{b}$ will change based on the value of τ.

Let $\mathbf{x}(t)$ satisfy the Itô SDE

$$d\mathbf{x}(t) = \mathbf{h}(\mathbf{x}(t), t)dt + H(\mathbf{x}(t), t)d\mathbf{w}(t) \qquad \text{(A.23)}$$

where $\mathbf{w} = [w_1, ..., w_2]^T$ is a vector of uncorrelated unit-strength Wiener processes. In component form (A.23) is written as

$$dx_i = h_i(\mathbf{x}, t)dt + \sum_{j=1}^{m} H_{ij}(\mathbf{x}, t)dw_j.$$

Since $\mathbf{x}(t + dt) = \mathbf{x}(t) + d\mathbf{x}(t)$, it follows that $\mathbf{x}(t)$ is a Markov process. This is because the distribution of its values at $t + dt$ is completely determined by the distribution at t and the transition probability describing the jump from t to $t + dt$.

Since sample paths generated by SDEs driven by white noise are continuous, $p(\mathbf{x}|\mathbf{y}, dt)$ is sharply peaked like $\delta(\mathbf{x}-\mathbf{y})$. Furthermore, while in general $p(\mathbf{x}|\mathbf{y}, \tau)$ is only a pdf in $\mathbf{x}$ (and not $\mathbf{y}$), and so

$$\int_{\mathbb{R}^d} p(\mathbf{x}|\mathbf{y}, \tau)d\mathbf{x} = 1 \quad \text{and} \quad \int_{\mathbb{R}^d} p(\mathbf{x}|\mathbf{y}, \tau)d\mathbf{y} \neq 1$$

for general values of $\tau > 0$, we have the special case

$$\lim_{\tau \to 0} \int_{\mathbb{R}^d} p(\mathbf{x}|\mathbf{y}, \tau)d\mathbf{y} = \int_{\mathbb{R}^d} \delta(\mathbf{x} - \mathbf{y})d\mathbf{y} = 1. \tag{A.24}$$

In the book, on p 121, the Taylor series for $\epsilon(\mathbf{y})$ expanded about $\epsilon(\boldsymbol{\xi})$, which is then multiplied by $p(\boldsymbol{\xi}|\mathbf{y}, \Delta t)$ inside an integral, was written (with $\partial\xi_j\partial\xi_k$ corrected as $\partial\xi_j\partial\xi_i$) as

$$\epsilon(\mathbf{y}) = \epsilon(\boldsymbol{\xi}) + \sum_{i=1}^{d}(y_i - \xi_i)\frac{\partial\epsilon}{\partial\xi_i} + \frac{1}{2}\sum_{i,j=1}^{d}(y_i - \xi_i)(y_j - \xi_j)\frac{\partial^2\epsilon}{\partial\xi_j\partial\xi_i} + ... \tag{A.25}$$

Where did this come from ? Well, in general if $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, then the Taylor series expansion of $f(\mathbf{x})$ when $\mathbf{x} = \mathbf{a} + \mathbf{v}$ for $\|\mathbf{v}\| << 1$ is

$$f(\mathbf{a} + \mathbf{v}) = f(\mathbf{a}) + \sum_{i=1}^{d} v_i \left(\frac{\partial f}{\partial v_i}\right)\Big|_{\mathbf{v}=\mathbf{a}} + \frac{1}{2}\sum_{i,j=1}^{d} v_i v_j \left(\frac{\partial^2 f}{\partial v_j \partial v_i}\right)\Bigg|_{\mathbf{v}=\mathbf{a}} + ...$$

On the right-hand-side of the above expression we could have replaced some of the $\mathbf{v}$'s and v_i's with $\mathbf{x}$'s and x_i's as

$$f(\mathbf{a} + \mathbf{v}) = f(\mathbf{a}) + \sum_{i=1}^{d} v_i \left(\frac{\partial f}{\partial x_i}\right)\Big|_{\mathbf{x}=\mathbf{a}} + \frac{1}{2}\sum_{i,j=1}^{d} v_i v_j \left(\frac{\partial^2 f}{\partial x_j \partial x_i}\right)\Bigg|_{\mathbf{x}=\mathbf{a}} + ...$$

since $\partial\mathbf{v}/\partial\mathbf{x}^T = \mathbb{I}$. The more familiar form may be

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{d}(x_i - a_i) \left(\frac{\partial f}{\partial x_i}\right)\Big|_{\mathbf{x}=\mathbf{a}} + \frac{1}{2}\sum_{i,j=1}^{d}(x_i - a_i)(x_j - a_j) \left(\frac{\partial^2 f}{\partial x_j \partial x_i}\right)\Bigg|_{\mathbf{x}=\mathbf{a}} + ...$$

When it is understood that the evaluation only applies to the partial derivatives, the parenthesis can be removed.

In the present context,

$$\epsilon(\mathbf{y}) = \epsilon(\boldsymbol{\xi} + (\mathbf{y} - \boldsymbol{\xi}))$$

and since Δt is small, and so $p(\boldsymbol{\xi}|\mathbf{y}, \Delta t)$ is sharply peaked like the delta function $\delta(\boldsymbol{\xi} - \mathbf{y})$, then the only values that contribute to the integral will be those for which $\|\mathbf{y} - \boldsymbol{\xi}\|$ is infinitesimally small. This justifies truncating the Taylor series as

$$\epsilon(\mathbf{y}) = \epsilon(\boldsymbol{\xi}) + \sum_{i=1}^{d}(y_i - \xi_i)\left.\frac{\partial \epsilon}{\partial y_i}\right|_{\mathbf{y}=\boldsymbol{\xi}} + \frac{1}{2}\sum_{i,j=1}^{d}(y_i - \xi_i)(y_j - \xi_j)\left.\frac{\partial^2 \epsilon}{\partial y_j \partial y_i}\right|_{\mathbf{y}=\boldsymbol{\xi}} + ...$$

And

$$\left.\frac{\partial \epsilon(\mathbf{y})}{\partial y_i}\right|_{\mathbf{y}=\boldsymbol{\xi}} = \frac{\partial \epsilon(\boldsymbol{\xi})}{\partial \xi_i}$$

Therefore, (A.25) is okay if we interpret all of the ϵ's on the right hand side to be $\epsilon(\boldsymbol{\xi})$.

In Versions 3 and 4 below, integrals of the form

$$I = \frac{1}{dt}\int_{\mathbb{R}^d} p(\boldsymbol{\xi}, t - dt|\mathbf{y}, 0)[\epsilon(\boldsymbol{\xi}) - \epsilon(\mathbf{x})]d\mathbf{x} = p(\boldsymbol{\xi}, t - dt|\mathbf{y}, 0)\int_{\mathbb{R}^d}\frac{1}{dt}[\epsilon(\boldsymbol{\xi}) - \epsilon(\mathbf{x})]d\mathbf{x}$$

need to be evaluated. Indeed, if dt is infinitesimally small, then $\mathbf{x}$ and $\boldsymbol{\xi}$ must be close to each other. This means that as $dt \to 0$,

$$\frac{1}{dt}[\epsilon(\boldsymbol{\xi}) - \epsilon(\mathbf{x})] = \frac{\partial \epsilon}{\partial \mathbf{x}^T}\left(\frac{\boldsymbol{\xi} - \mathbf{x}}{dt}\right).$$

Since in general

$$\int_{\mathbb{R}^d}\frac{\partial \epsilon}{\partial \mathbf{x}^T}d\mathbf{x} = \mathbf{0}^T,$$

it follows that $I = 0$.

Better Proof (Version 0)

In this section Itô's rule is used rather than the Chapman-Kolmogorov equation to derive the Fokker-Planck equation. This has the same three essential steps as in other proofs. Namely

- The use of a Taylor series expansion of an arbitrary compactly supported function (in the current case this is embedded in the derivation of Itô's rule);
- Integration by parts;
- Localization of an expression from inside of an integral.

Since Itô's rule applies equally well to Itô SDEs in Cartesian and curvilinear coordinates, the latter (more general) case will be handled here.

Given the Itô SDE

$$d\mathbf{q} = \mathbf{a}(\mathbf{q},t)dt + B(\mathbf{q},t)d\mathbf{w} \quad \text{or} \quad dq_i = a_i(\mathbf{q},t)dt + \sum_{j=1}^{m} B_{ij}(\mathbf{q},t)dw_j \tag{A.26}$$

and given an arbitrary smooth compactly supported real-valued function $\epsilon(\mathbf{q})$, then Itô's rule is written as

$$d\epsilon = \left(\sum_{j=1}^{d} \frac{\partial \epsilon}{\partial q_j} a_j + \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^2 \epsilon}{\partial q_k \partial q_l} [BB^T]_{kl}\right) dt + \sum_{k=1}^{d} \sum_{l=1}^{m} \frac{\partial \epsilon}{\partial q_k} B_{kl} dw_l.$$

Now, recall that the expected value of any function $\phi(\mathbf{q},t)$ where $\mathbf{q}$ is a set of curvilinear coordinates in $\mathbb{R}^d$ is computed as

$$\langle \phi(\mathbf{q}(t)) \rangle \doteq \int_{\mathbb{R}^d} \phi(\mathbf{q},t) f(\mathbf{q},t) |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} \tag{A.27}$$

where $f(\mathbf{q},t)$ is the time-evolving probability density describing the ensemble behavior of all sample paths generated by the SDE in (A.26). Therefore,

$$\begin{aligned}
\langle d\epsilon \rangle &= \left\langle \left(\sum_{j=1}^{d} \frac{\partial \epsilon}{\partial q_j} a_j + \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^2 \epsilon}{\partial q_k \partial q_l} [BB^T]_{kl}\right) dt + \sum_{k=1}^{d} \sum_{l=1}^{m} \frac{\partial \epsilon}{\partial q_k} B_{kl} dw_l \right\rangle \\
&= \left\langle \sum_{j=1}^{d} \frac{\partial \epsilon}{\partial q_j} a_j + \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^2 \epsilon}{\partial q_k \partial q_l} [BB^T]_{kl} \right\rangle dt + \sum_{k=1}^{d} \sum_{l=1}^{m} \left\langle \frac{\partial \epsilon}{\partial q_k} B_{kl} \right\rangle \langle dw_l \rangle \\
&= \left\langle \sum_{j=1}^{d} \frac{\partial \epsilon}{\partial q_j} a_j + \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^2 \epsilon}{\partial q_k \partial q_l} [BB^T]_{kl} \right\rangle dt.
\end{aligned} \tag{A.28}$$

The expectation of the product becomes a product of expectations because of Itô's interpretation of the stochastic integral, and the corresponding SDE. And the second term disappears because $\langle dw_l \rangle$. Now referring back to (A.27) we can develop expressions that constrain the behavior of $f(\mathbf{q},t)$. First observe that

$$\begin{aligned}
\langle d\epsilon \rangle &= \langle \epsilon(\mathbf{q}(t+dt)) - \epsilon(\mathbf{q}(t)) \rangle = \langle d\epsilon(\mathbf{q}(t+dt)) \rangle - \langle d\epsilon(\mathbf{q}(t)) \rangle \\
&= \int_{\mathbb{R}^d} \epsilon(\mathbf{q}) f(\mathbf{q}, t+dt) |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} - \int_{\mathbb{R}^d} \epsilon(\mathbf{q}) f(\mathbf{q}) |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} \\
&= \int_{\mathbb{R}^d} \epsilon(\mathbf{q}) \frac{\partial f(\mathbf{q},t)}{\partial t} dt |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q}.
\end{aligned}$$

Expressing the right side of (A.28) in a similar way, and combining gives

$$\int_{\mathbb{R}^d} \epsilon(\mathbf{q}) \frac{\partial f(\mathbf{q},t)}{\partial t} dt |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} = \int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} \frac{\partial \epsilon}{\partial q_j} a_j + \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^2 \epsilon}{\partial q_k \partial q_l} [BB^T]_{kl}\right) f(\mathbf{q},t) |G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} dt.$$

Cancelling dt, and performing integration by parts on the right side of the form[3]

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\partial \epsilon}{\partial q_j} a_j f |G|^{\frac{1}{2}} d\mathbf{q} &= -\int_{\mathbb{R}^d} \frac{\partial}{\partial q_j}\left(a_j f |G|^{\frac{1}{2}}\right)\epsilon(\mathbf{q}) d\mathbf{q} \\ &= -\int_{\mathbb{R}^d} |G|^{-\frac{1}{2}} \frac{\partial}{\partial q_j}\left(a_j f |G|^{\frac{1}{2}}\right)\epsilon(\mathbf{q})|G|^{\frac{1}{2}} d\mathbf{q}\end{aligned}$$

(and doing so similarly twice for the other term) and then writing the result all on one side gives

$$\int_{\mathbb{R}^d} \left\{ \frac{\partial f(\mathbf{q},t)}{\partial t} + |G(\mathbf{q})|^{-\frac{1}{2}} \sum_{j=1}^{d} \frac{\partial}{\partial q_j}\left(|G(\mathbf{q})|^{\frac{1}{2}} a_j(\mathbf{q},t) f(\mathbf{q},t)\right)\right.$$

$$\left. -\frac{1}{2}|G(\mathbf{q})|^{-\frac{1}{2}} \sum_{k,l=1}^{d} \frac{\partial^2}{\partial q_k \partial q_l}\left([BB^T(\mathbf{q},t)]_{kl} f(\mathbf{q},t)|G(\mathbf{q})|^{\frac{1}{2}}\right)\right\} \epsilon(\mathbf{q})|G(\mathbf{q})|^{\frac{1}{2}} d\mathbf{q} = 0.$$

Localization then gives

$$\begin{aligned}\frac{\partial f(\mathbf{q},t)}{\partial t} &= -|G(\mathbf{q})|^{-\frac{1}{2}} \sum_{j=1}^{d} \frac{\partial}{\partial q_j}\left(|G(\mathbf{q})|^{\frac{1}{2}} a_j(\mathbf{q},t) f(\mathbf{q},t)\right) \\ &\quad +\frac{1}{2}|G(\mathbf{q})|^{-\frac{1}{2}} \sum_{k,l=1}^{d} \frac{\partial^2}{\partial q_k \partial q_l}\left([BB^T(\mathbf{q},t)]_{kl} f(\mathbf{q},t)|G(\mathbf{q})|^{\frac{1}{2}}\right).\end{aligned} \tag{A.29}$$

Better Proof (Version 1)

In what follows, $\mathbf{x} \in \mathbb{R}^d$ denotes the position of the stochastic trajectory at some time τ after being at $\mathbf{y} \in \mathbb{R}^d$. In some contexts the separation of time between these two states will be $\tau = dt$, in others it will be $\tau = t$, and in others it will be $\tau = t + dt$. If $\tau = dt$ then $\mathbf{x}$ might denote the Cartesian coordinates of $\mathbf{x}(t+dt)$ and $\mathbf{y}$ might denote the coordinates of $\mathbf{x}(t)$ (or, equivalently $\mathbf{x}$ might denote the Cartesian coordinates of $\mathbf{x}(t)$ and $\mathbf{y}$ might denote the coordinates of $\mathbf{x}(t-dt)$). In comparison, if $\tau = t$ then $\mathbf{x}$ might denote the Cartesian coordinates of $\mathbf{x}(t)$ and $\mathbf{y}$ might denote the coordinates of all possible initial states $\mathbf{x}(0)$.

The transition probability $p(\mathbf{x}, t+dt|\mathbf{y}, t)$ is exactly that which would be generated by making a histogram corresponding to an infinite number of sample paths of length dt generated from (A.23). Using the properties of $p(\mathbf{x}, t+dt|\mathbf{y}, t)$ in (4.52) and (4.53), it follows that

$$\int_{\mathbb{R}^d} (x_i - y_i) p(\mathbf{x}, t+dt|\mathbf{y}, t) d\mathbf{x} = \langle x_i - y_i \rangle = \langle dx_i \rangle = h_i(\mathbf{y}, t) dt \tag{A.30}$$

or, using different variable names,

$$\int_{\mathbb{R}^d} (\xi_i - x_i) p(\boldsymbol{\xi}, t+dt|\mathbf{x}, t) d\boldsymbol{\xi} = \langle \xi_i - x_i \rangle = \langle dx_i \rangle = h_i(\mathbf{x}, t) dt$$

[3] The surface terms disappear because $\epsilon(\mathbf{q})$ is taken to be compactly supported.

and

$$\int_{\mathbb{R}^d}(x_i{-}y_i)(x_j{-}y_j)p(\mathbf{x},t{+}dt|\mathbf{y},t)d\mathbf{x} = \langle (x_i{-}y_i)(x_j{-}y_j)\rangle = \langle dx_i dx_j\rangle = \sum_{k=1}^{m} H_{ik}(\mathbf{y},t)H^T_{kj}(\mathbf{y},t)dt \tag{A.31}$$

or

$$\int_{\mathbb{R}^d}(\xi_i{-}x_i)(\xi_j{-}x_j)p(\boldsymbol{\xi},t{+}dt|\mathbf{x},t)d\boldsymbol{\xi} = \langle (\xi_i{-}x_i)(\xi_j{-}x_j)\rangle = \langle dx_i dx_j\rangle = \sum_{k=1}^{m} H_{ik}(\mathbf{x},t)H^T_{kj}(\mathbf{x},t)dt.$$

In contrast to $p(\mathbf{x},t+dt|\mathbf{y},t)$ (or $p(\boldsymbol{\xi},t+dt|\mathbf{x},t)$), we now investigate $p(\mathbf{x},t+dt\,|\,\mathbf{y},0)$. The Chapman-Kolmogorov equation (A.21) can be written in the special case when $t_1 - t_3 = t + dt$ as

$$p(\mathbf{x},t+dt|\mathbf{y},0) = \int_{\mathbb{R}^d} p(\mathbf{x},t+dt|\boldsymbol{\xi},t)p(\boldsymbol{\xi},t|\mathbf{y},0)d\boldsymbol{\xi}. \tag{A.32}$$

Using the Chapman-Kolmogorov equation in the form of (A.32), together with the definition of partial derivative gives[4]

$$\begin{aligned}\frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t} &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[p(\mathbf{x},t+\Delta t\,|\,\mathbf{y},0) - p(\mathbf{x},t|\mathbf{y},0)\right] \qquad\text{(A.33)}\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[\int_{\mathbb{R}^n} p(\mathbf{x},t+\Delta t\,|\,\boldsymbol{\xi},t)p(\boldsymbol{\xi},t\,|\,\mathbf{y},0)d\boldsymbol{\xi} - p(\mathbf{x},t\,|\,\mathbf{y},0)\right].\end{aligned}$$

Let $\epsilon(\mathbf{x})$ be an arbitrary compactly supported function for which $\partial\epsilon/\partial x_i$ and $\partial^2\epsilon/\partial x_j\partial x_i$ are continuous for all $i,j,k = 1,...,n$. Then the projection of $\partial p/\partial t$ against $\epsilon(\mathbf{x})$ can be expanded as

$$\begin{aligned}\int_{\mathbb{R}^d}\frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[\int_{\mathbb{R}^d}\epsilon(\mathbf{x})d\mathbf{x}\int_{\mathbb{R}^d} p(\mathbf{x},t+\Delta t\,|\,\boldsymbol{\xi},t)p(\boldsymbol{\xi},t\,|\,\mathbf{y},0)d\boldsymbol{\xi}\right.\\ &\qquad\left. - \int_{\mathbb{R}^n} p(\mathbf{x},t|\mathbf{y},0)\epsilon(\mathbf{x})d\mathbf{x}\right].\end{aligned}$$

Now perform the following manipulations on the first term on the right hand side of the above equation:

$$\begin{aligned}\int_{\mathbb{R}^d}\epsilon(\mathbf{x})\left[\int_{\mathbb{R}^d} p(\mathbf{x},t+\Delta t\,|\,\boldsymbol{\xi},t)p(\boldsymbol{\xi},t\,|\,\mathbf{y},0)d\boldsymbol{\xi}\right]d\mathbf{x} &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\epsilon(\mathbf{x})p(\mathbf{x},t+\Delta t\,|\,\boldsymbol{\xi},t)p(\boldsymbol{\xi},t\,|\,\mathbf{y},0)d\mathbf{x}d\boldsymbol{\xi}\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\epsilon(\boldsymbol{\xi})p(\boldsymbol{\xi},t+\Delta t\,|\,\mathbf{x},t)p(\mathbf{x},t|\mathbf{y},0)d\boldsymbol{\xi}d\mathbf{x}\\ &= \int_{\mathbb{R}^d} p(\mathbf{x},t|\mathbf{y},0)\left[\int_{\mathbb{R}^d}\epsilon(\boldsymbol{\xi})p(\boldsymbol{\xi},t+\Delta t\,|\,\mathbf{x},t)d\boldsymbol{\xi}\right]d\mathbf{x}\end{aligned}$$

[4] A point of possible confusion is that the value of separation time in $p(\mathbf{x},t+\Delta t\,|\,\mathbf{y},0)$ and $p(\mathbf{x},t|\mathbf{y},0)$ are different, and so the resulting conditional pdfs are different, but the position variables $\mathbf{x}$ and $\mathbf{y}$ are the same in both.

where all that was done in the middle step was a swapping of the names of the variables of integration.

Therefore, (A.33) can be written as

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\mathbb{R}^d} p(\mathbf{x},t|\mathbf{y},0)\left[\int_{\mathbb{R}^d} p(\boldsymbol{\xi},t+\Delta t|\mathbf{x},t)\epsilon(\boldsymbol{\xi})d\boldsymbol{\xi} - \epsilon(\mathbf{x})\right]d\mathbf{x}. \tag{A.34}$$

In the above bracketed term, $\boldsymbol{\xi}$ denotes the possible positions of a stochastic trajectory dt units of time *after* it is observed at $\mathbf{x}$. This is different than the role of $\boldsymbol{\xi}$ in the prior equations, because we swapped the roles of $\mathbf{x}$ and $\boldsymbol{\xi}$ in order to isolate $p(\mathbf{x}|\mathbf{y},t)$.

Expanding the function $\epsilon(\boldsymbol{\xi})$ in its Taylor series about $\mathbf{x}$:

$$\epsilon(\boldsymbol{\xi}) = \epsilon(\mathbf{x}+(\boldsymbol{\xi}-\mathbf{x})) = \epsilon(\mathbf{x})+\sum_{i=1}^{d}(\xi_i-x_i)\frac{\partial\epsilon}{\partial x_i}+\frac{1}{2}\sum_{i,j=1}^{d}(\xi_i-x_i)(\xi_j-x_j)\frac{\partial^2\epsilon}{\partial x_i\partial x_j}+\ldots$$

and since here $\boldsymbol{\xi}$ is playing the same role as $\mathbf{y}$ in (A.30) and (A.31), substituting this series into the previous equation results in

$$\begin{aligned}
\int_{\mathbb{R}^d} p(\boldsymbol{\xi},t+dt\,|\,\mathbf{x},t)\epsilon(\boldsymbol{\xi})d\boldsymbol{\xi} \quad &= \quad \epsilon(\mathbf{x})\cdot\int_{\mathbb{R}^d} p(\boldsymbol{\xi},t+dt\,|\,\mathbf{x},t)d\boldsymbol{\xi} \\
&\quad +\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial x_i}\cdot\int_{\mathbb{R}^d}(\xi_i-x_i)p(\boldsymbol{\xi},t+dt\,|\,\mathbf{x},t)d\boldsymbol{\xi} \\
&\quad +\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial x_i\partial x_j}\cdot\int_{\mathbb{R}^d}(\xi_i-x_i)(\xi_j-x_j)p(\boldsymbol{\xi},t+dt\,|\,\mathbf{x},t)d\boldsymbol{\xi} \\
&= \quad \epsilon(\mathbf{x})+\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial x_i}h_i(\mathbf{x},t)dt+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial x_i\partial x_j}\sum_{k=1}^{m}H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)dt.
\end{aligned}$$

The first term on the last line above comes from the fact that

$$\int_{\mathbb{R}^d} p(\boldsymbol{\xi},t+dt\,|\,\mathbf{x},t)d\boldsymbol{\xi} = 1$$

and so this $\epsilon(\mathbf{x})$ will cancel with the one in (A.34). It follows that

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^d}\left[\sum_{i=1}^{n}\frac{\partial\epsilon}{\partial x_i}h_i(\mathbf{x},t)+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial x_i\partial x_j}\sum_{k=1}^{m}H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)\right]p(\mathbf{x},t|\mathbf{y},0)d\mathbf{y}$$

when (A.30) and (A.31) are observed.

The final step is to integrate the two terms on the right-hand side of the above equation by parts to generate

$$\int_{\mathbb{R}^d}\left\{\frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}+\sum_{i=1}^{d}\frac{\partial}{\partial y_i}(h_i(\mathbf{x},t)p(\mathbf{x},t|\mathbf{y},0))-\frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i\partial x_j}(H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)p(\mathbf{x},t|\mathbf{y},0))\right\}\epsilon(\mathbf{x})d\mathbf{x}=0 \tag{A.35}$$

Using the standard localization argument (4.56)⟹(4.57), and using $f(\mathbf{x},t)$ as shorthand for the transition probability $p(\mathbf{x},t|\mathbf{y},0)$, the term in braces becomes:

$$\boxed{\frac{\partial f(\mathbf{x},t)}{\partial t}+\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left(h_i(\mathbf{x},t)f(\mathbf{x},t)\right)-\frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i\partial x_j}\left(H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)f(\mathbf{x},t)\right)=0} \tag{A.36}$$

To be more precise about the relationship between $f(\mathbf{x},t)$ and $p(\mathbf{x},t|\mathbf{y},0)$, if $f(\mathbf{x},0)=p(\mathbf{x})$, then

$$f(\mathbf{x},t)=\int_{\mathbb{R}^d}p(\mathbf{x},t|\mathbf{y},0)p(\mathbf{y})d\mathbf{y}.$$

And so any pde consisting of partial derivatives in t and x that $p(\mathbf{x},t|\mathbf{y},0)$ satisfies for fixed $\mathbf{y}$ will also be satisifed by $f(\mathbf{x},t)$. And in the special case when $f(\mathbf{x},0)=\delta(\mathbf{x}-\mathbf{y}_0)$ then $f(\mathbf{x},t)=p(\mathbf{x},t|\mathbf{y}_0,0)$.

Better Proof (Version 2)

Here we will switch the roles of $\mathbf{x}$ and $\mathbf{y}$ in the conditional and perform the same proof. Only now, we will assume that the conditional probabilities are invariant under time shifts:

$$p(\mathbf{y},t_1+t|\mathbf{x},t)=p(\mathbf{y},t_1|\mathbf{x},0)\doteq p(\mathbf{y}|\mathbf{x},t_1).$$

In order to use this assumption, we restrict the discussion to the case when $h_i(\mathbf{x},t)=h_i(\mathbf{x})$ and $H_{ij}(\mathbf{x},t)=H_{ij}(\mathbf{x})$.

Let $\mathbf{x}=\mathbf{x}(t)$ and $\mathbf{y}=\mathbf{x}(t+dt)$ where dt is an infinitesimal time increment. Using the properties of $p(\mathbf{y}|\mathbf{x},dt)$ it follows that

$$\int_{\mathbb{R}^d}(y_i-x_i)p(\mathbf{y}|\mathbf{x},dt)d\mathbf{y}=\langle y_i-x_i\rangle=h_i(\mathbf{x})dt \tag{A.37}$$

or

$$\int_{\mathbb{R}^d}(\xi_i-y_i)p(\boldsymbol{\xi}|\mathbf{y},dt)d\boldsymbol{\xi}=\langle \xi_i-y_i\rangle=h_i(\mathbf{y})dt$$

and

$$\int_{\mathbb{R}^d}(y_i-x_i)(y_j-x_j)p(\mathbf{y}|\mathbf{x},dt)d\mathbf{y}=\langle (y_i-x_i)(y_j-x_j)\rangle=\sum_{k=1}^{m}H_{ik}(\mathbf{x})H_{kj}^T(\mathbf{x})dt \tag{A.38}$$

or

$$\int_{\mathbb{R}^d}(\xi_i-y_i)(\xi_j-y_j)p(\boldsymbol{\xi}|\mathbf{y},dt)d\boldsymbol{\xi}=\langle (\xi_i-y_i)(\xi_j-y_j)\rangle=\sum_{k=1}^{m}H_{ik}(\mathbf{y})H_{kj}^T(\mathbf{y})dt.$$

Using the Chapman-Kolmogorov equation, together with the definition of partial derivative gives

$$\frac{\partial p(\mathbf{y}|\mathbf{x},t)}{\partial t}=\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[p(\mathbf{y}|\mathbf{x},t+\Delta t)-p(\mathbf{y}|\mathbf{x},t)\right]$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\int_{\mathbb{R}^n} p(\mathbf{y}|\boldsymbol{\xi}, \Delta t) p(\boldsymbol{\xi}|\mathbf{x}, t) d\boldsymbol{\xi} - p(\mathbf{y}|\mathbf{x}, t) \right].$$

Let $\epsilon(\mathbf{x})$ be an arbitrary compactly supported function for which $\partial\epsilon/\partial x_i$ and $\partial^2\epsilon/\partial x_j \partial x_i$ are continuous for all $i, j, k = 1, ..., n$. Then the projection of $\partial p/\partial t$ against $\epsilon(\mathbf{y})$ can be expanded as

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\partial p(\mathbf{y}|\mathbf{x}, t)}{\partial t} \epsilon(\mathbf{y}) d\mathbf{y} &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\int_{\mathbb{R}^d} \epsilon(\mathbf{y}) d\mathbf{y} \int_{\mathbb{R}^d} p(\mathbf{y}|\boldsymbol{\xi}, \Delta t) p(\boldsymbol{\xi}|\mathbf{x}, t) d\boldsymbol{\xi} \right. \\ &\qquad \left. - \int_{\mathbb{R}^n} p(\mathbf{y}|\mathbf{x}, t) \epsilon(\mathbf{y}) d\mathbf{y} \right]. \end{aligned} \tag{A.39}$$

Now perform the following manipulations on the first term on the right hand side of the above equation:

$$\begin{aligned} \int_{\mathbb{R}^d} \epsilon(\mathbf{y}) \left[\int_{\mathbb{R}^d} p(\mathbf{y} \,|\, \boldsymbol{\xi}, \Delta t) p(\boldsymbol{\xi} \,|\, \mathbf{x}, t) d\boldsymbol{\xi} \right] d\mathbf{y} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \epsilon(\mathbf{y}) p(\mathbf{y} \,|\, \boldsymbol{\xi}, \Delta t) p(\boldsymbol{\xi} \,|\, \mathbf{x}, t) d\mathbf{y} d\boldsymbol{\xi} \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \epsilon(\boldsymbol{\xi}) p(\boldsymbol{\xi} \,|\, \mathbf{y}, \Delta t) p(\mathbf{y} \,|\, \mathbf{x}, t) d\boldsymbol{\xi} d\mathbf{y} \\ &= \int_{\mathbb{R}^d} p(\mathbf{y} \,|\, \mathbf{x}, t) \left[\int_{\mathbb{R}^d} \epsilon(\boldsymbol{\xi}) p(\boldsymbol{\xi} \,|\, \mathbf{y}, \Delta t) d\boldsymbol{\xi} \right] d\mathbf{y} \end{aligned}$$

where all that was done in the middle step was a swapping of the names of the variables of integration. Therefore,

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{y}|\mathbf{x}, t)}{\partial t} \epsilon(\mathbf{y}) d\mathbf{y} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{\mathbb{R}^d} p(\mathbf{y}|\mathbf{x}, t) \left[\int_{\mathbb{R}^d} p(\boldsymbol{\xi}|\mathbf{y}, \Delta t) \epsilon(\boldsymbol{\xi}) d\boldsymbol{\xi} - \epsilon(\mathbf{y}) \right] d\mathbf{y}.$$

Expanding the function $\epsilon(\boldsymbol{\xi})$ in its Taylor series about $\mathbf{y}$:

$$\epsilon(\boldsymbol{\xi}) = \epsilon(\mathbf{y}) + \sum_{i=1}^{d} (\xi_i - y_i) \frac{\partial \epsilon}{\partial y_i} + \frac{1}{2} \sum_{i,j=1}^{d} (\xi_i - y_i)(\xi_j - y_j) \frac{\partial^2 \epsilon}{\partial y_j \partial y_k} + ...$$

and substituting this series into the previous equation results in

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{y}|\mathbf{x}, t)}{\partial t} \epsilon(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^d} \left[\sum_{i=1}^{n} \frac{\partial \epsilon}{\partial y_i} h_i(\mathbf{y}) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 \epsilon}{\partial y_i \partial y_j} \sum_{k=1}^{m} H_{ik}(\mathbf{y}) H_{kj}^T(\mathbf{y}) \right] p(\mathbf{y}|\mathbf{x}, t) d\mathbf{y}$$

when (A.37) and (A.38) are observed.

The final step is to integrate the two terms on the right-hand side of the above equation by parts to generate

$$\int_{\mathbb{R}^d} \left\{ \frac{\partial p(\mathbf{y}|\mathbf{x}, t)}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial y_i} (h_i(\mathbf{y}) p(\mathbf{y}|\mathbf{x}, t)) - \frac{1}{2} \sum_{k=1}^{m} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial y_i \partial y_j} (H_{ik}(\mathbf{y}) H_{kj}^T(\mathbf{y}) p(\mathbf{y}|\mathbf{x}, t)) \right\} \epsilon(\mathbf{y}) d\mathbf{y} = 0.$$

Using the standard localization argument, and using $f(\mathbf{y},t)$ as shorthand for $\int_{\mathbb{R}^d} p(\mathbf{y}|\mathbf{x},t)p(\mathbf{x})d\mathbf{x}$, (or, equivalently, $f(\mathbf{x},t)$ for $\int_{\mathbb{R}^d} p(\mathbf{x}|\mathbf{y},t)p(\mathbf{y})d\mathbf{y}$), the term in braces becomes:

$$\boxed{\frac{\partial f(\mathbf{x},t)}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(h_i(\mathbf{x})f(\mathbf{x},t)\right) - \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}\left(H_{ik}(\mathbf{x})H_{kj}^T(\mathbf{x})f(\mathbf{x},t)\right) = 0} \tag{A.40}$$

Better Proof (Version 3)

Here we will keep the meanings of $\mathbf{x} = \mathbf{x}(t)$, $\boldsymbol{\xi} = \mathbf{x}(t - dt)$ and $\mathbf{y} = \mathbf{x}(0)$ static and do not let them vary in meaning with context as in the previous proofs. Here $t > 0$ and dt is infinitesimally small.

The transition probability $p(\mathbf{x}|\boldsymbol{\xi}, dt)$ is exactly that which would be generated by making a histogram corresponding to an infinite number of sample paths of length dt generated from (A.23) with starting value of $\mathbf{x}(t)$ being $\boldsymbol{\xi}$. Using the properties of $p(\mathbf{x}|\boldsymbol{\xi}, dt)$ in (4.52) and (4.53) (with $\boldsymbol{\xi}$ taking the place of $\mathbf{x}$), it follows that

$$\int_{\mathbb{R}^d} (x_i - \xi_i)p(\mathbf{x}|\boldsymbol{\xi}, dt)d\mathbf{x} = \langle dx_i \rangle = \langle x_i - \xi_i \rangle = h_i(\boldsymbol{\xi})dt \tag{A.41}$$

and

$$\int_{\mathbb{R}^d} (x_i - \xi_i)(x_j - \xi_j)p(\mathbf{x}|\boldsymbol{\xi}, dt)d\mathbf{x} = \langle dx_i dx_j \rangle = \langle (x_i - \xi_i)(x_j - \xi_j) \rangle = \sum_{k=1}^{m} H_{ik}(\boldsymbol{\xi})H_{kj}^T(\boldsymbol{\xi})dt. \tag{A.42}$$

We now investigate $p(\mathbf{x}\,|\,\mathbf{y}, t)$. The Chapman-Kolmogorov equation, (4.16), or equivalently (A.22), can be written in the special case when $t_1 - t_3 = t$ as

$$p(\mathbf{x}\,|\,\mathbf{y}, t) = \int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi}, dt)p(\boldsymbol{\xi}\,|\,\mathbf{y}, t - dt)d\boldsymbol{\xi}. \tag{A.43}$$

Using the Chapman-Kolmogorov equation in the form of (A.43), together with the definition of partial derivative gives

$$\frac{\partial p(\mathbf{x}|\mathbf{y}, t)}{\partial t} = \frac{1}{dt}\left[\int_{\mathbb{R}^n} p(\mathbf{x}\,|\,\boldsymbol{\xi}, dt)p(\boldsymbol{\xi}\,|\,\mathbf{y}, t - dt)d\boldsymbol{\xi} - p(\boldsymbol{\xi}\,|\,\mathbf{y}, t - dt)\right]. \tag{A.44}$$

Here we have used the "backward difference" as opposed to the "forward difference" to compute the derivative. Since dt is infinitesimally small and all functions ρ, h, H are assumed to be smooth, the result should be the same (even though $\mathbf{x}(t)$ is not differentiable).

Let $\epsilon(\mathbf{x})$ be an arbitrary compactly supported function for which $\partial\epsilon/\partial x_i$ and $\partial^2\epsilon/\partial x_j \partial x_i$ are continuous for all $i, j, k = 1, ..., n$. Then the projection of $\partial p/\partial t$

against $\epsilon(\mathbf{x})$ can be expanded as

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} &= \frac{1}{dt}\left[\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)p(\boldsymbol{\xi}\,|\,\mathbf{y},t-dt)d\boldsymbol{\xi}\right)\epsilon(\mathbf{x})d\mathbf{x}\right. \\ &\left. -\int_{\mathbb{R}^n} p(\boldsymbol{\xi}\,|\,\mathbf{y},t-dt)\epsilon(\mathbf{x})d\mathbf{x}\right]. \qquad (A.45)\end{aligned}$$

Now changing the order of integration in the first term on the right hand side of the above equation:

$$\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)p(\boldsymbol{\xi}\,|\,\mathbf{y},t-dt)d\boldsymbol{\xi}\right)\epsilon(\mathbf{x})d\mathbf{x} =$$

$$\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)\epsilon(\mathbf{x})d\mathbf{x}\right)p(\boldsymbol{\xi}\,|\,\mathbf{y},t-dt)d\boldsymbol{\xi} \qquad (A.46)$$

Expanding the function $\epsilon(\mathbf{x})$ in its Taylor series about $\boldsymbol{\xi}$:

$$\epsilon(\mathbf{x}) = \epsilon(\boldsymbol{\xi}+(\mathbf{x}-\boldsymbol{\xi})) = \epsilon(\boldsymbol{\xi})+\sum_{i=1}^{d}(x_i-\xi_i)\frac{\partial\epsilon}{\partial\xi_i}+\frac{1}{2}\sum_{i,j=1}^{d}(x_i-\xi_i)(x_j-\xi_j)\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}+...$$

Here, of course, the following are equivalent

$$\frac{\partial\epsilon}{\partial\xi_i} = \left.\frac{\partial\epsilon}{\partial x_i}\right|_{\mathbf{x}=\boldsymbol{\xi}} \qquad \text{and} \qquad \frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j} = \left.\frac{\partial^2\epsilon}{\partial x_i\partial x_j}\right|_{\mathbf{x}=\boldsymbol{\xi}}.$$

Then

$$\begin{aligned}\int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)\epsilon(\mathbf{x})d\mathbf{x} &= \epsilon(\boldsymbol{\xi})\cdot\int_{\mathbb{R}^d} p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)d\mathbf{x} \\ &+\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial\xi_i}\cdot\int_{\mathbb{R}^d}(x_i-\xi_i)p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)d\mathbf{x} \\ &+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}\cdot\int_{\mathbb{R}^d}(x_i-\xi_i)(x_j-\xi_j)p(\mathbf{x}\,|\,\boldsymbol{\xi},dt)d\mathbf{x} \\ &= \epsilon(\boldsymbol{\xi})+\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial\xi_i}h_i(\boldsymbol{\xi},t)dt+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}\sum_{k=1}^{m}H_{ik}(\boldsymbol{\xi},t)H_{kj}^T(\boldsymbol{\xi},t)dt.\end{aligned}$$

Back-substituting into (A.46) and (A.45) gives

$$\begin{aligned}\int_{\mathbb{R}^d}\frac{\partial p(\mathbf{x}|\mathbf{y},t)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} &= \int_{\mathbb{R}^d}\left[\sum_{i=1}^{n}\frac{\partial\epsilon}{\partial\xi_i}h_i(\boldsymbol{\xi},t)+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}\sum_{k=1}^{m}H_{ik}(\boldsymbol{\xi},t)H_{kj}^T(\boldsymbol{\xi},t)\right]p(\boldsymbol{\xi}|\mathbf{y},t-dt)d\boldsymbol{\xi} \\ &+\frac{1}{dt}\int_{\mathbb{R}^d} p(\boldsymbol{\xi}|\mathbf{y},t-dt)[\epsilon(\boldsymbol{\xi})-\epsilon(\mathbf{x})]d\mathbf{x}\end{aligned}$$

when (A.41) and (A.42) are observed. Assuming that the second term can be made to vanish as $dt \to 0$, and $\boldsymbol{\xi} \to \mathbf{x}$, the above becomes

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x}|\mathbf{y},t)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^d}\left[\sum_{i=1}^{n} \frac{\partial \epsilon}{\partial x_i}h_i(\mathbf{x},t) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 \epsilon}{\partial x_i \partial x_j}\sum_{k=1}^{m} H_{ik}(\mathbf{x},t)H^T_{kj}(\mathbf{x},t)\right] p(\mathbf{x}|\mathbf{y},t)d\mathbf{x}.$$

The final step is to integrate the two terms on the right-hand side of the above equation by parts to generate

$$\int_{\mathbb{R}^d}\left\{\frac{\partial p(\mathbf{x}|\mathbf{y},t)}{\partial t} + \sum_{i=1}^{d}\frac{\partial}{\partial x_i}(h_i(\mathbf{x},t)p(\mathbf{x}|\mathbf{y},t)) - \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}(H_{ik}(\mathbf{x},t)H^T_{kj}(\mathbf{x},t)p(\mathbf{x}|\mathbf{y},t))\right\}\epsilon(\mathbf{x})d\mathbf{x} = 0 \tag{A.47}$$

Using the standard localization argument and using $f(\mathbf{x},t)$ as shorthand for the transition probability $p(\mathbf{x}|\mathbf{y},t)$, the term in braces becomes:

$$\boxed{\frac{\partial f(\mathbf{x},t)}{\partial t} + \sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left(h_i(\mathbf{x},t)f(\mathbf{x},t)\right) - \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}\left(H_{ik}(\mathbf{x},t)H^T_{kj}(\mathbf{x},t)f(\mathbf{x},t)\right) = 0.} \tag{A.48}$$

Better Proof (Version 4)

Here we will keep the meanings of $\mathbf{x} = \mathbf{x}(t)$, $\boldsymbol{\xi} = \mathbf{x}(t - dt)$ and $\mathbf{y} = \mathbf{x}(0)$ static and do not let them vary in meaning with context as in the previous proofs. Here $t > 0$ and dt is infinitesimally small.

The transition probability $p(\mathbf{x}, t|\boldsymbol{\xi}, t - dt)$ is exactly that which would be generated by making a histogram corresponding to an infinite number of sample paths of length dt generated from (A.23) with starting value of $\mathbf{x}(t)$ being $\boldsymbol{\xi}$. Using the properties of $p(\mathbf{x}, t|\boldsymbol{\xi}, t - dt)$ in (4.52) and (4.53) (with $\boldsymbol{\xi}$ taking the place of $\mathbf{x}$), it follows that

$$\int_{\mathbb{R}^d}(x_i - \xi_i)p(\mathbf{x}, t|\boldsymbol{\xi}, t - dt)d\mathbf{x} = \langle dx_i\rangle = \langle x_i - \xi_i\rangle = h_i(\boldsymbol{\xi}, t)dt \tag{A.49}$$

and

$$\int_{\mathbb{R}^d}(x_i - \xi_i)(x_j - \xi_j)p(\mathbf{x}, t|\boldsymbol{\xi}, t - dt)d\mathbf{x} = \langle dx_i dx_j\rangle = \langle(x_i - \xi_i)(x_j - \xi_j)\rangle = \sum_{k=1}^{m} H_{ik}(\boldsymbol{\xi}, t)H^T_{kj}(\boldsymbol{\xi}, t)dt. \tag{A.50}$$

We now investigate $p(\mathbf{x}, t\,|\,\mathbf{y}, 0)$. The Chapman-Kolmogorov equation, (4.16), or equivalently (A.22), can be written in the special case when $t_1 - t_3 = t$ as

$$p(\mathbf{x}, t\,|\,\mathbf{y}, 0) = \int_{\mathbb{R}^d} p(\mathbf{x}, t\,|\,\boldsymbol{\xi}, t - dt)p(\boldsymbol{\xi}, t - dt\,|\,\mathbf{y}, 0)d\boldsymbol{\xi}. \tag{A.51}$$

Using the Chapman-Kolmogorov equation in the form of (A.51), together with the definition of partial derivative gives

$$\frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t} = \frac{1}{dt}\left[\int_{\mathbb{R}^n} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)d\boldsymbol{\xi} - p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)\right]. \tag{A.52}$$

Here we have used the "backward difference" as opposed to the "forward difference" to compute the derivative. Since dt is infinitesimally small and all functions are assumed to be smooth, the result should be the same.

Let $\epsilon(\mathbf{x})$ be an arbitrary compactly supported function for which $\partial\epsilon/\partial x_i$ and $\partial^2\epsilon/\partial x_j\partial x_i$ are continuous for all $i,j,k=1,...,n$. Then the projection of $\partial p/\partial t$ against $\epsilon(\mathbf{x})$ can be expanded as

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} \;&=\; \frac{1}{dt}\left[\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)d\boldsymbol{\xi}\right)\epsilon(\mathbf{x})d\mathbf{x}\right.\\ &\qquad\left.-\int_{\mathbb{R}^n} p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)\epsilon(\mathbf{x})d\mathbf{x}\right]. \end{aligned}\tag{A.53}$$

Now changing the order of integration in the first term on the right hand side of the above equation:

$$\begin{aligned}&\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)d\boldsymbol{\xi}\right)\epsilon(\mathbf{x})d\mathbf{x} =\\ &\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)\epsilon(\mathbf{x})d\mathbf{x}\right)p(\boldsymbol{\xi},t-dt\,|\,\mathbf{y},0)d\boldsymbol{\xi}\end{aligned}\tag{A.54}$$

Expanding the function $\epsilon(\mathbf{x})$ in its Taylor series about $\boldsymbol{\xi}$:

$$\epsilon(\mathbf{x}) = \epsilon(\boldsymbol{\xi}+(\mathbf{x}-\boldsymbol{\xi})) = \epsilon(\boldsymbol{\xi})+\sum_{i=1}^{d}(x_i-\xi_i)\frac{\partial\epsilon}{\partial\xi_i}+\frac{1}{2}\sum_{i,j=1}^{d}(x_i-\xi_i)(x_j-\xi_j)\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}+...$$

Here, of course, the following are equivalent

$$\frac{\partial\epsilon}{\partial\xi_i} = \left.\frac{\partial\epsilon}{\partial x_i}\right|_{\mathbf{x}=\boldsymbol{\xi}} \quad\text{and}\quad \frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j} = \left.\frac{\partial^2\epsilon}{\partial x_i\partial x_j}\right|_{\mathbf{x}=\boldsymbol{\xi}}.$$

Then

$$\begin{aligned}\int_{\mathbb{R}^d} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)\epsilon(\mathbf{x})d\mathbf{x} \;&=\; \epsilon(\boldsymbol{\xi})\cdot\int_{\mathbb{R}^d} p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)d\mathbf{x}\\ &\quad+\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial\xi_i}\cdot\int_{\mathbb{R}^d}(x_i-\xi_i)p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)d\mathbf{x}\\ &\quad+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}\cdot\int_{\mathbb{R}^d}(x_i-\xi_i)(x_j-\xi_j)p(\mathbf{x},t\,|\,\boldsymbol{\xi},t-dt)d\mathbf{x}\\ &=\; \epsilon(\boldsymbol{\xi})+\sum_{i=1}^{d}\frac{\partial\epsilon}{\partial\xi_i}h_i(\boldsymbol{\xi},t)dt+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2\epsilon}{\partial\xi_i\partial\xi_j}\sum_{k=1}^{m}H_{ik}(\boldsymbol{\xi},t)H_{kj}^T(\boldsymbol{\xi},t)dt.\end{aligned}$$

Back-substituting into (A.54) and (A.53) gives

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} &= \int_{\mathbb{R}^d}\left[\sum_{i=1}^{n}\frac{\partial \epsilon}{\partial \xi_i}h_i(\boldsymbol{\xi},t) + \frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2 \epsilon}{\partial \xi_i \partial \xi_j}\sum_{k=1}^{m}H_{ik}(\boldsymbol{\xi},t)H_{kj}^T(\boldsymbol{\xi},t)\right]p(\boldsymbol{\xi},t-dt|\mathbf{y},0)d\boldsymbol{\xi} \\ &+ \frac{1}{dt}\int_{\mathbb{R}^d} p(\boldsymbol{\xi},t-dt|\mathbf{y},0)[\epsilon(\boldsymbol{\xi})-\epsilon(\mathbf{x})]d\mathbf{x}\end{aligned}$$

when (A.49) and (A.50) are observed. Assuming that the second term can be made to vanish as $dt \to 0$, and $\boldsymbol{\xi} \to \mathbf{x}$, the above becomes

$$\int_{\mathbb{R}^d} \frac{\partial p(\mathbf{x}|\mathbf{y},t)}{\partial t}\epsilon(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^d}\left[\sum_{i=1}^{n}\frac{\partial \epsilon}{\partial x_i}h_i(\mathbf{x},t) + \frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2 \epsilon}{\partial x_i \partial x_j}\sum_{k=1}^{m}H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)\right]p(\mathbf{x}|\mathbf{y},t)d\mathbf{x}.$$

The final step is to integrate the two terms on the right-hand side of the above equation by parts to generate

$$\int_{\mathbb{R}^d}\left\{\frac{\partial p(\mathbf{x},t|\mathbf{y},0)}{\partial t} + \sum_{i=1}^{d}\frac{\partial}{\partial x_i}(h_i(\mathbf{x},t)p(\mathbf{x},t|\mathbf{y},0)) - \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}(H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)p(\mathbf{x},t|\mathbf{y},0))\right\}\epsilon(\mathbf{x})d\mathbf{x} = 0 \tag{A.55}$$

Using the standard localization argument and using $f(\mathbf{x},t)$ as shorthand for the transition probability $p(\mathbf{x}|\mathbf{y},t)$, the term in braces becomes:

$$\boxed{\frac{\partial f(\mathbf{x},t)}{\partial t} + \sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left(h_i(\mathbf{x},t)f(\mathbf{x},t)\right) - \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}\left(H_{ik}(\mathbf{x},t)H_{kj}^T(\mathbf{x},t)f(\mathbf{x},t)\right) = 0.} \tag{A.56}$$

A.3.3 Integration over First vs. Second Argument of Conditional Probabilities

In the proof of the Fokker-Planck equation in the book is the integral over the second argument rather than the first an error or is it okay ?

Let $\mathbf{y}$ denote the Cartesian coordinates of the space in which $\mathbf{x}(t_1)$ moves and $\mathbf{x}$ denote the Cartesian coordinates of the space in which $\mathbf{x}(t_2)$ moves. Joint and conditional probability densities are related by the equalities

$$p(\mathbf{y},t_1|\mathbf{x},t_2)p(\mathbf{x},t_2) = p(\mathbf{y},t_1;\mathbf{x},t_2) = p(\mathbf{x},t_2|\mathbf{y},t_1)p(\mathbf{y},t_1).$$

Bayes' rule is then

$$p(\mathbf{y},t_1|\mathbf{x},t_2) = \frac{p(\mathbf{x},t_2|\mathbf{y},t_1)p(\mathbf{y},t_1)}{p(\mathbf{x},t_2)}. \tag{A.57}$$

Now, in the special case that $t_1 = t + dt$ and $t_2 = t$, then if

$$\int_{\mathbb{R}^d}(y_i - x_i)p(\mathbf{y},t+dt|\mathbf{x},t)d\mathbf{y} = h_i(\mathbf{x},t)dt$$

then we can also use (A.57) to write

$$\int_{\mathbb{R}^d} (y_i - x_i) \frac{p(\mathbf{x}, t|\mathbf{y}, t+dt)p(\mathbf{y}, t+dt)}{p(\mathbf{x}, t)} d\mathbf{y} = h_i(\mathbf{x}, t)dt.$$

Multiplying by -1, letting $dt \to -dt$, and observing that $p(\mathbf{x}, t \pm dt)/p(\mathbf{y}, t) \approx 1$ gives

$$\int_{\mathbb{R}^d} (x_i - y_i)p(\mathbf{x}, t|\mathbf{y}, t-dt)d\mathbf{y} = h_i(\mathbf{x}, t)dt.$$

So, in the case when $t_1 - t_2 = dt$ is infinitesimally small, it is okay to integrate over the second index in the conditional probability density.

But what about when $t_1 - t_2$ is not infinitesimal ?

Using Bayes' rule,

$$F(\mathbf{x}, t_1, t_2) \doteq \int_{\mathbb{R}^d} \epsilon(\mathbf{y})p(\mathbf{y}, t_1|\mathbf{x}, t_2)d\mathbf{y} = \int_{\mathbb{R}^d} \epsilon(\mathbf{y}) \frac{p(\mathbf{x}, t_2|\mathbf{y}, t_1)p(\mathbf{y}, t_1)}{p(\mathbf{x}, t_2)} d\mathbf{y}.$$

Therefore,

$$\int_{\mathbb{R}^d} \epsilon(\mathbf{y})p(\mathbf{x}, t_2|\mathbf{y}, t_1)p(\mathbf{y}, t_1)d\mathbf{y} = F(\mathbf{x}, t_1, t_2)p(\mathbf{x}, t_2)$$

This does not look promising.

Is (4.60) correct ? Suppose we start with (4.60), written as

$$\frac{\partial p(\mathbf{x}, t|\mathbf{y}, 0)}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial y_i} \left(h_i(\mathbf{y}, t)p(\mathbf{x}, t|\mathbf{y}, 0)\right) - \frac{1}{2} \sum_{k=1}^{m} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial y_i \partial y_j} \left(H_{ik}(\mathbf{y}, t)H_{kj}^T(\mathbf{y}, t)p(\mathbf{x}, t|\mathbf{y}, 0)\right) = 0.$$

Can we get from here using Bayes' rule to

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left(h_i(\mathbf{x}, t)p(\mathbf{x}, t)\right) - \frac{1}{2} \sum_{k=1}^{m} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} \left(H_{ik}(\mathbf{x}, t)H_{kj}^T(\mathbf{x}, t)p(\mathbf{x}, t)\right) = 0?$$

This does not look promising either.

In summary, Versions 0,1,4 are more favorable than the proof in the book, the last step of which is suspicious.

A.3.4 Inter-conversion between Itô and Stratonovich

Suppose we are given a Stratonovich SDE

$$d\mathbf{x} = \mathbf{h}^s(\mathbf{x}, t)dt + H^s(\mathbf{x}, t) \circledS \, d\mathbf{w},$$

or in component form

$$dx_i = h_i^s(\mathbf{x}, t)dt + \sum_j H_{ij}^s(\mathbf{x}, t) \circledS dw_j. \tag{A.58}$$

with coupling matrix entries $H^s_{ij}(\mathbf{x},t)$ and drift vector entries $h^s_i(\mathbf{x},t)$. Then how do we get the corresponding Itô equation of the form

$$dx_k = h_k(\mathbf{x},t)dt + \sum_l H_{kl}(\mathbf{x},t)dw_l\,? \tag{A.59}$$

And how do we relate the Itô and Stratonovich forms of the Fokker-Planck equation ?

First, recognize that if we make the choice $H^s_{ij}(\mathbf{x},t) = H_{ij}(\mathbf{x},t)$, then

$$\begin{aligned} H^s_{ij}\left(\mathbf{x}+\frac{1}{2}d\mathbf{x}, t+\frac{1}{2}dt\right) &= H_{ij}(\mathbf{x},t) + \frac{1}{2}\frac{\partial H_{ij}(\mathbf{x},t)}{\partial t}dt + \\ & \sum_k \frac{\partial H_{ij}}{\partial x_k}\cdot\left(\frac{1}{2}dx_k\right) + \frac{1}{2}\sum_{k,l}\frac{\partial^2 H_{ij}}{\partial x_k \partial x_l}\cdot\left(\frac{1}{2}dx_k\right)\cdot\left(\frac{1}{2}dx_l\right). \end{aligned} \tag{A.60}$$

Then substituting in (A.59) and using Itô's rules for expectations,[5]

$$\langle dt^2 \rangle = \langle dt dw_j \rangle = 0$$

and

$$\langle dw_i dw_j \rangle = dt\delta_{ij},$$

and all higher order terms disappear. Substituting these rules into (A.60) and summing gives

$$\sum_j H^s_{ij}\left(\mathbf{x}+\frac{1}{2}d\mathbf{x}, t+\frac{1}{2}t\right)dw_j = \sum_j H_{ij}\left(\mathbf{x},t\right)dw_j + \frac{1}{2}\sum_{k,j}\frac{\partial H_{ij}\left(\mathbf{x},t\right)}{\partial x_k}H_{kj}\left(\mathbf{x},t\right)dt.$$

Note that this would have been the same as if instead of $H^s_{ij}\left(\mathbf{x}+\frac{1}{2}d\mathbf{x}, t+\frac{1}{2}dt\right)$ in (A.60) we used $H^s_{ij}\left(\mathbf{x}+\frac{1}{2}d\mathbf{x}, t\right)$. In fact, both of these two choices are used in the literature. But since it does not affect the final outcome, there is no need to commit to one convention or the other.

Explicitly this is because

$$\frac{1}{2}\sum_{j,k}\frac{\partial H_{ij}}{\partial x_k}dx_k dw_j = \frac{1}{2}\sum_{j,k}\frac{\partial H_{ij}}{\partial x_k}\sum_l H_{kl}dw_l dw_j =$$

$$\frac{1}{2}\sum_{j,k}\frac{\partial H_{ij}}{\partial x_k}\sum_l H_{kl}\delta_{jl}dt = \frac{1}{2}\sum_{j,k}\frac{\partial H_{ij}}{\partial x_k}H_{kj}dt.$$

And so

$$dx_i = \left(h^s_i + \frac{1}{2}\sum_k\sum_j\frac{\partial H_{ij}}{\partial x_k}H_{kj}\right)dt + \sum_j H_{ij}dw_j.$$

[5] Here 0 is not truly zero, and these terms are not truly zero, rather, both are negligibly small and so are thought of as zero.

In other words,

$$\boxed{h_i = h_i^s + \frac{1}{2}\sum_k \sum_j \frac{\partial H_{ij}}{\partial x_k} H_{kj}.}$$

Substituting this SDE into the expression for the Itô version of the Fokker-Planck equation (4.61), gives

$$\frac{\partial f}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[\left(h_i^s + \frac{1}{2}\sum_k \sum_j \frac{\partial H_{ij}}{\partial x_k} H_{kj}\right) f\right] + \frac{1}{2}\sum_{i,j,k} \frac{\partial^2}{\partial x_i \partial x_j}\left[H_{ik}^s H_{jk}^s f\right] \tag{A.61}$$

where again, the choice $H_{ij} = H_{ij}^s$ was made, as explained in the text.

Let us split

$$\frac{\partial}{\partial x_i}\left[\left(h_i^s + \frac{1}{2}\sum_k \sum_j \frac{\partial H_{ij}}{\partial x_k} H_{kj}\right) f\right]$$

into

$$\frac{\partial}{\partial x_i}\left[h_i^s f\right] + \frac{1}{2}\sum_k \sum_j \frac{\partial}{\partial x_i}\left[\frac{\partial H_{ij}}{\partial x_k} H_{kj} f\right].$$

Now compare

$$\sum_{ijk} \frac{\partial}{\partial x_i}\left[\frac{\partial H_{ij}}{\partial x_k} H_{kj} f\right]$$

and

$$\sum_{ijk} \frac{\partial^2}{\partial x_i \partial x_j}\left[H_{ij} H_{kj} f\right].$$

Since i, j, k are dummy indices that get summed out, we can write

$$\sum_{ijk} \frac{\partial}{\partial x_i}\left[\frac{\partial H_{ij}}{\partial x_k} H_{kj} f\right] = \sum_{ijk} \frac{\partial}{\partial x_i}\left[\frac{\partial H_{ik}}{\partial x_j} H_{jk} f\right]$$

to make it more consistent in appearance with the other terms.

Now expanding out

$$\sum_{ijk} \frac{\partial^2}{\partial x_i \partial x_j}\left[H_{ik} H_{kj} f\right] = \sum_{ijk} \frac{\partial}{\partial x_i}\left[\frac{\partial}{\partial x_j}(H_{ik} H_{kj} f)\right]$$

we get the following terms inside of the summations:

$$\frac{\partial}{\partial x_i}\left[\frac{\partial H_{ik}}{\partial x_j} H_{kj} f + H_{ik}\frac{\partial}{\partial x_j}(H_{kj} f)\right].$$

The first of these terms cancels with part of the drift term in (A.61).

The result simplifies to (4.70):

$$\boxed{\frac{\partial f}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(h_i^s f\right) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left[\sum_{k=1}^{m} H_{ik}^s \frac{\partial}{\partial x_j}(H_{jk}^s f)\right].}$$

A.3.5 SDEs in Cartesian and Polar Coordinates in the Plane

This subsection is an addendum to p. 133. Note that in the unlabeled equation above (4.97) (corrected with x_1^{-1} replaced by x_1^{-2}), that this equation has already factored in the fact that $dx_i = dw_i$, and therefore $dx_1 dx_2 = 0$. More general versions of these equations are:

$$dr = \frac{x_1 dx_1 + x_2 dx_2}{(x_1^2 + x_2^2)^{\frac{1}{2}}} + \frac{x_2^2 (dx_1)^2 - 2x_1 x_2 dx_1 dx_2 + x_1^2 (dx_2)^2}{2(x_1^2 + x_2^2)^{\frac{3}{2}}}$$

and

$$d\phi = \frac{-x_2 dx_1 + x_1 dx_2}{x_1^2 + x_2^2} + \frac{x_1 x_2 (dx_1)^2 - x_1 x_2 (dx_2)^2 + (x_2^2 - x_1^2) dx_1 dx_2}{(x_1^2 + x_2^2)^2}$$

A.4 The Heat Equation in Curvilinear Coordinates

This addendum involves both stochastic processes in $\mathbb{R}^n$ and some geometry (through Jacobian matrices), and can be viewed as a bridge between Chapters 4, 5, and 6.

The heat equation on $\mathbb{R}^n$ is written in Cartesian coordinates as

$$\frac{\partial f}{\partial t} = \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}. \tag{A.62}$$

The summation on right hand side can be written as

$$\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} = \nabla^2 f = \text{div}(\text{grad} f).$$

Two ways to directly convert this to curvilinear coordinates are outlined below.

Method 1

Suppose $\mathbf{x} = \boldsymbol{\psi}(\mathbf{q})$. Then, much like what was done in Section 4.8.1 in the case of $\mathbb{R}^2$ and polar coordinates, we observe that

$$d\mathbf{x} = J(\mathbf{q}) d\mathbf{q} \quad \text{where} \quad J(\mathbf{q}) = \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{q}^T},$$

and if $\tilde{f}(\mathbf{q}) \doteq f(\boldsymbol{\psi}(\mathbf{q}))$ then

$$df = \frac{\partial f}{\partial \mathbf{x}^T} d\mathbf{x} = \frac{\partial \tilde{f}}{\partial \mathbf{q}^T} d\mathbf{q}$$

or

$$\frac{\partial f}{\partial \mathbf{x}^T} J(\mathbf{q}) d\mathbf{q} = \frac{\partial \tilde{f}}{\partial \mathbf{q}^T} d\mathbf{q}.$$

Taking the transpose of both sides and using the arbitrariness of $d\mathbf{q}$,

$$J^T(\mathbf{q})\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial \tilde{f}}{\partial \mathbf{q}}.$$

Therefore

$$\frac{\partial f}{\partial \mathbf{x}} = J^{-T}(\mathbf{q})\frac{\partial \tilde{f}}{\partial \mathbf{q}} \quad \text{or} \quad \frac{\partial f}{\partial x_i} = \sum_{j=1}^{n} J_{ij}^{-T}\frac{\partial \tilde{f}}{\partial q_j}. \tag{A.63}$$

Substituting (A.63) into the right-hand-side of (A.62) twice gives

$$\begin{aligned}
\sum_{i=1}^{n}\frac{\partial^2 f}{\partial x_i^2} &= \sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}\right)\\
&= \sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} J_{ij}^{-T}\frac{\partial \tilde{f}}{\partial q_j}\right)\\
&= \sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} J_{ij}^{-T}\frac{\partial \tilde{f}}{\partial q_j}\right)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} J_{ik}^{-T}\frac{\partial}{\partial q_k}\left(\sum_{j=1}^{n} J_{ij}^{-T}\frac{\partial \tilde{f}}{\partial q_j}\right).
\end{aligned}$$

Therefore, the heat equation on $\mathbb{R}^n$ can be written in curvilinear coordinates as

$$\frac{\partial f}{\partial t} = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} J_{ik}^{-T}\frac{\partial}{\partial q_k}\left(\sum_{j=1}^{n} J_{ij}^{-T}\frac{\partial f}{\partial q_j}\right) \tag{A.64}$$

where the tilde has been dropped since it is clear from the context that now $f = f(\mathbf{q}, t)$.

Method 2

The Laplace-Beltrami operator defined in (5.50),

$$\text{div}(\text{grad}f) \doteq |G|^{-\frac{1}{2}}\sum_{k=1}^{n}\frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\sum_{j=1}^{n} g^{kj}\frac{\partial f}{\partial q_j}\right),$$

means that the heat equation can be written as

$$\frac{\partial f}{\partial t} = \frac{1}{2}|G|^{-\frac{1}{2}}\sum_{k=1}^{n}\frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\sum_{j=1}^{n} g^{kj}\frac{\partial f}{\partial q_j}\right). \tag{A.65}$$

How does this compare with (A.64) ?

First, observe that

$$G = [g_{ij}] = J^T J \implies G^{-1} = [g^{ij}] = J^{-1}J^{-T}, \tag{A.66}$$

and so

$$\begin{aligned}
\sum_{k=1}^{n} \frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}} \sum_{j=1}^{n} g^{kj}\frac{\partial f}{\partial q_j}\right) &= \sum_{k=1}^{n} \frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}} \sum_{i,j=1}^{n} J_{ki}^{-1} J_{ij}^{-T}\frac{\partial f}{\partial q_j}\right) \\
&= \sum_{i,k=1}^{n} \frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\, J_{ki}^{-1} \sum_{j=1}^{n} J_{ji}^{-1}\frac{\partial f}{\partial q_j}\right) \\
&= \sum_{i,k=1}^{n} \frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\, J_{ki}^{-1}\right) \sum_{j=1}^{n} J_{ji}^{-1}\frac{\partial f}{\partial q_j} + \\
&\quad |G|^{\frac{1}{2}} \sum_{i,k=1}^{n} J_{ki}^{-1} \frac{\partial}{\partial q_k}\left(\sum_{j=1}^{n} J_{ji}^{-1}\frac{\partial f}{\partial q_j}\right)
\end{aligned}$$

Substituting this back into (A.65) gives

$$\frac{\partial f}{\partial t} = \frac{1}{2}|G|^{-\frac{1}{2}}\left\{\sum_{i,k=1}^{n} \frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\, J_{ki}^{-1}\right)\sum_{j=1}^{n} J_{ji}^{-1}\frac{\partial f}{\partial q_j} + |G|^{\frac{1}{2}}\sum_{i,k=1}^{n} J_{ki}^{-1}\frac{\partial}{\partial q_k}\left(\sum_{j=1}^{n} J_{ji}^{-1}\frac{\partial f}{\partial q_j}\right)\right\}.$$

Note that if the first term in the braces is zero, then the above equation will reduce exactly to (A.64). In particular, can we justify setting the following equality for each $j = 1, 2, ..., n$?

$$\sum_{i,k=1}^{n} J_{ji}^{-1}\frac{\partial}{\partial q_k}\left(|G|^{\frac{1}{2}}\, J_{ki}^{-1}\right) = 0$$

As explained in the following subsection, the answer is 'yes.'

Relationship to a Special Stratonovich-Fokker-Planck Equation

Given the Stratonovich SDE

$$d\mathbf{q} = \mathbf{a}^s(\mathbf{q})dt + J^{-1}(\mathbf{q}) \circledS d\mathbf{w},$$

the corresponding Fokker-Planck equation is

$$\frac{\partial f}{\partial t} = -|G|^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial}{\partial q_i}\left(a_i^s|G|^{\frac{1}{2}}f\right) + \frac{1}{2}|G|^{-\frac{1}{2}}\sum_{i,j,k=1}^{n}\frac{\partial}{\partial q_i}\left[J_{ik}^{-1}\frac{\partial}{\partial q_j}(J_{jk}^{-1}|G|^{\frac{1}{2}}f)\right] \tag{A.67}$$

As illustrated in Exercise 4.11, when

$$a_i^s(\mathbf{q}) = \frac{1}{2}|G|^{-\frac{1}{2}}\sum_{j,k=1}^{n} J_{ik}^{-1}\frac{\partial}{\partial q_j}\left(|G|^{\frac{1}{2}}J_{jk}^{-1}\right),$$

the heat equation results.

While this statement is true, it resulted from the purely mechanical manipulation of symbols using the product rule for differentiation. In contrast, we can use a conceptual argument to study $a_i(\mathbf{q},t)$. Consider the Stratonovich SDE in Cartesian coordinates

$$d\mathbf{x} = d\mathbf{w}.$$

Obviously, the corresponding Fokker-Planck equation is the heat equation in (A.62). Now, if $\mathbf{x} = \boldsymbol{\psi}(\mathbf{q})$ then using the Stratonovich calculus, $d\mathbf{x} = J(\mathbf{q}) \textcircled{s}\, d\mathbf{q}$. Substitution then gives

$$J(\mathbf{q}) \textcircled{s}\, d\mathbf{q} = d\mathbf{w} \implies d\mathbf{q} = J^{-1}(\mathbf{q}) \textcircled{s}\, d\mathbf{w}.$$

Therefore, the heat equation in curvilinear coordinates should be (A.67) with $a_i^s(\mathbf{q}) = 0$. How is this paradox resolved ? It must be the case that due to the geometric structure of $\mathbb{R}^n$ that in any set of curvilinear coordinates

$$\boxed{\sum_{j,k=1}^{n} J_{ik}^{-1} \frac{\partial}{\partial q_j}\left(|G|^{\frac{1}{2}} J_{jk}^{-1}\right) = 0 \quad \text{for} \quad i = 1, 2, ..., n.} \tag{A.68}$$

Relationship to a Special Itô -Fokker-Planck Equation

Given the Itô SDE

$$d\mathbf{q} = \mathbf{a}(\mathbf{q})dt + J^{-1}(\mathbf{q})d\mathbf{w},$$

the corresponding Itô version of the Fokker-Planck equation in generalized coordinates is

$$\frac{\partial f}{\partial t} = -|G|^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\left(a_i |G|^{\frac{1}{2}} f\right) + \frac{1}{2}|G|^{-\frac{1}{2}} \sum_{i,j=1}^{n} \frac{\partial^2}{\partial q_i \partial q_j}\left[g^{ij}|G|^{\frac{1}{2}} f\right]. \tag{A.69}$$

where (A.66) has been used.

As illustrated in Exercise 4.10, (A.69) will become the heat equation under the special condition

$$a_i(\mathbf{q}) = \frac{1}{2}|G|^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial}{\partial q_j}\left(|G|^{\frac{1}{2}} g^{ij}\right).$$

While this statement is true, it resulted from the purely mechanical manipulation of symbols using the product rule for differentiation. In contrast, we can use a conceptual argument to evaluate $a_i(\mathbf{q},t)$. Consider the Itô SDE in Cartesian coordinates

$$d\mathbf{x} = d\mathbf{w}.$$

As with the Stratonovich interpretation, the corresponding Fokker-Planck equation is the heat equation in (A.62).

Now, if $\mathbf{x} = \boldsymbol{\psi}(\mathbf{q})$, then Itô 's rule as described in Section 4.5.5 can be applied in the current context, under the assumption that there is an underlying Itô SDE of the form

$$d\mathbf{q} = \mathbf{h}(\mathbf{q})dt + H(\mathbf{q})d\mathbf{w},$$

as

$$dx_i = \left(\sum_{j=1}^{n} \frac{\partial \psi_i}{\partial q_j} h_j + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial \psi_i^2}{\partial q_k \partial q_l} [HH^T]_{kl} \right) dt + \sum_{k,l} \frac{\partial \psi_i}{\partial q_k} H_{kl} dw_l = dw_i. \tag{A.70}$$

The second equality in the above expression forces

$$\sum_{k=1}^{n} \frac{\partial \psi_i}{\partial q_k} H_{kl} = \delta_{il} \implies H = J^{-1}$$

(where of course $J_{ik} = \partial \psi_i / \partial q_k$) and

$$\sum_{j=1}^{n} \frac{\partial \psi_i}{\partial q_j} h_j + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial \psi_i^2}{\partial q_k \partial q_l} [HH^T]_{kl} = 0.$$

This can be written as

$$\sum_{j=1}^{n} J_{ij} h_j + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial}{\partial q_k} (J_{il}) [J^{-1} J^{-T}]_{kl} = 0.$$

Multiplication by J_{pi}^{-1} and summation over i, together with the observation that $G^{-1} = J^{-1} J^{-T}$, gives

$$h_p = -\frac{1}{2} \sum_{i,k,l=1}^{n} J_{pi}^{-1} g^{lk} \frac{\partial}{\partial q_k} (J_{il}).$$

It will be convenient to switch the roles of p and i and write this as

$$h_i = -\frac{1}{2} \sum_{p,k,l=1}^{n} J_{ip}^{-1} g^{lk} \frac{\partial}{\partial q_k} (J_{pl}).$$

Clearly it must be the case that

$$a_i(\mathbf{q}) = h_i(\mathbf{q}). \tag{A.71}$$

Does enforcing this equality constrain J or G ?

Actually, no it does not. This can be observed by writing out explicitly

$$|G|^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial}{\partial q_j} \left(|G|^{\frac{1}{2}} g^{ij} \right) = - \sum_{p,k,l=1}^{n} J_{ip}^{-1} g^{lk} \frac{\partial}{\partial q_k} (J_{pl}).$$

Substituting $g^{ij} = \sum_k J_{ik}^{-1} J_{jk}^{-1}$ in the left side, and using the fact that $J^{-1} J = \mathbb{I}$ (and hence

$$-J^{-1} \frac{\partial J}{\partial q_k} = \frac{\partial (J^{-1})}{\partial q_k} J$$

as a result of differentiating with respect to q_k) gives

$$|G|^{-\frac{1}{2}}\sum_{j=1}^{n}\frac{\partial}{\partial q_j}\left(|G|^{\frac{1}{2}}\sum_{k}J_{ik}^{-1}J_{jk}^{-1}\right)=\sum_{p,k,l=1}^{n}\frac{\partial(J_{ip}^{-1})}{\partial q_k}J_{pl}g^{lk}.$$

But

$$\sum_{l=1}^{n}J_{pl}g^{lk}=\sum_{l=1}^{n}J_{pl}\left(\sum_{m}J_{lm}^{-1}J_{km}^{-1}\right)=J_{kp}^{-1}.$$

Therefore, (A.71) is equivalent to

$$|G|^{-\frac{1}{2}}\sum_{j,k=1}^{n}\frac{\partial}{\partial q_j}\left(J_{ik}^{-1}|G|^{\frac{1}{2}}J_{jk}^{-1}\right)=\sum_{p,k=1}^{n}\frac{\partial(J_{ip}^{-1})}{\partial q_k}J_{kp}^{-1}.$$

Using the product rule to expand the derivative on the left hand side, and multiplying both sides by $|G|^{\frac{1}{2}}$ then results in

$$\sum_{j,k=1}^{n}\frac{\partial}{\partial q_j}\left(J_{ik}^{-1}\right)|G|^{\frac{1}{2}}J_{jk}^{-1}+\sum_{j,k=1}^{n}J_{ik}^{-1}\frac{\partial}{\partial q_j}\left(|G|^{\frac{1}{2}}J_{jk}^{-1}\right)=|G|^{\frac{1}{2}}\sum_{p,k=1}^{n}\frac{\partial(J_{ip}^{-1})}{\partial q_k}J_{kp}^{-1}.$$

But the second term on the left is zero from (A.68), and what remains on both sides is exactly the same, only written in terms of different dummy indices.

Therefore, there are several equivalent ways to write the heat equation on $\mathbb{R}^n$ in curvilinear coordinates. The difference in the appearance of these expressions can be reduced by using (A.68) together with rules from matrix calculus.

A.5 Addendum to Chapter 5

A.5.1 Increasing the Stability of Inverse Jacobian Iterations on pp. 150-1

If the values of $\mathbf{q}$ and $\mathbf{k}$ form a valid pair, then it should be the case that $\mathbf{F}(\mathbf{q},\mathbf{k})=\mathbf{0}$. However, it can be the case that after several iterations numerical errors accumulate to cause this to be nonzero. This will then lead to numerical instabilities.

The algorithm stated on p. 151 and embodied by (5.16) does not incorporate a feedback mechanism to ensure that $\mathbf{F}(\mathbf{q},\mathbf{k})=\mathbf{0}$. This algorithm can be made more stable by interlacing a correction step that updates $\mathbf{k}$ after each update of $\mathbf{q}$. Basically, after we update $\mathbf{q}$, if $\|\mathbf{F}(\mathbf{q},\mathbf{k})\|$ is not zero, then holding fixed the value of $\mathbf{q}$ that was just computed, we should do Jacobian inverse iterations with $\mathbf{k}$ as the variable, updating $\mathbf{k}$ with an artificial trajectory connecting $\mathbf{F}(\mathbf{q},\mathbf{k})$ to the zero vector. Such iterations are completely analogous to the resolved-rate manipulator inverse kinematics algorithm described on p. 145 and 151 and in (5.17) (with $\mathbf{k}$ taking the place of $\mathbf{q}$, $\mathbf{F}(\mathbf{q},\mathbf{k})$ taking the place of $\mathbf{f}(\mathbf{q})$, and the artificial trajectory connecting $\mathbf{F}(\mathbf{q},\mathbf{k})$ to zero taking the place of $\mathbf{q}(t)$. Of course, updating $\mathbf{k}$ will cause it to deviate from the baseline linear path from $\mathbf{k}_0$ to $\mathbf{k}_{goal}$, and so the direction $\mathbf{k}_{goal} - \mathbf{k}$ (where $\mathbf{k}$ is always the most up-to-date value) should be used instead of $\mathbf{k}_{goal} - \mathbf{k}_0$ to drive the next iteration of $\mathbf{q}$.

A.5.2 Computing Normal Curvature

In Section 5.4.2, an arc-length parameterized curve $\mathbf{c}(s)$ is defined to be contained in a surface $\mathbf{x}(q_1, q_2)$. In other words,

$$\mathbf{c}(s) = \mathbf{x}(q_1(s), q_2(s)).$$

Application of the chain rule then gives

$$\mathbf{c}''(s) = \sum_{i=1}^{2} \frac{\partial \mathbf{x}}{\partial q_i} q_i''(s) + \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial^2 \mathbf{x}}{\partial q_i \partial q_j} q_i'(s) q_j'(s) = \kappa(s)\mathbf{n}_1(s) \tag{A.72}$$

where $' = d/ds$.

By the definition of curvature, $\mathbf{c}''(s) = \kappa(s)\mathbf{n}_1(s)$. Furthermore, two normals to the curve that are different than $\mathbf{n}_1(s)$ and $\mathbf{n}_2(s)$ are the surface normal $\mathbf{n}(q_1(s), q_2(s))$ and $\mathbf{m}(s) = \mathbf{n}(q_1(s), q_2(s)) \times \mathbf{c}'(s)$. Therefore, it is possible to resolve $\mathbf{c}''(s)$ as

$$\kappa(s)\mathbf{n}_1(s) = \kappa_n(s)\mathbf{n}(q_1(s), q_2(s)) + \kappa_g(s)\mathbf{m}(s).$$

A question raised was how do we get from (5.55), which is

$$\kappa_n = \mathbf{c}'' \cdot \mathbf{n},$$

to (5.56), which is

$$\kappa_n = \frac{d\mathbf{q}^T L d\mathbf{q}}{d\mathbf{q}^T G d\mathbf{q}}.$$

The answer lies in the fact that

$$\frac{\partial^2 \mathbf{x}}{\partial q_i \partial q_j} = L_{ij}\mathbf{n} + \sum_k \Gamma_{ij}^k \frac{\partial \mathbf{x}}{\partial q_k}$$

and since

$$\frac{\partial \mathbf{x}}{\partial q_k} \cdot \mathbf{n} = 0,$$

then using (A.72) gives

$$\kappa_n = \sum_{ij} L_{ij} q_i' q_j' = \frac{\sum_{ij} L_{ij} dq_i dq_j}{ds^2}.$$

But

$$ds^2 = \sum_{ij} G_{ij} dq_i dq_j,$$

and so (5.56) follows.

A.6 Addendum to Chapter 6

A.6.1 An Intuitive Introduction to Push-Forward Vector Fields

Rather than taking a top-down approach to push-forward vector fields in which the definition is stated and the properties follow, here some motivation is developed that will result in the definition as a natural outcome of intuitively desirable properties. Here the motivating example that is provided will be in the context of deformation of planar regions.

Consider the closed unit box $B = [0,1] \times [0,1] \subset \mathbb{R}^2$. To each point $\mathbf{x} \in B$, assign a vector using the function $\mathbf{a}(\mathbf{x}) \in \mathbb{R}^2$. Note that the values of $\mathbf{a}(\mathbf{x})$ need not be in B even though the values of $\mathbf{x}$ are. For concreteness, let us assume that

$$\mathbf{a}(\mathbf{x}) = \left(\begin{array}{c} \epsilon \\ -\epsilon \end{array} \right) \tag{A.73}$$

where ϵ is a small positive real number. It is not difficult to imagine drawing tiny little arrows on a Cartesian grid that overlays B to visualize this constant vector.

Now suppose that the original domain, B, is morphed, or transformed, by a mapping $\boldsymbol{\psi} : B \to D \subset \mathbb{R}^2$. In other words, the original unit square region is transformed into another shape. For concreteness, suppose that

$$\boldsymbol{\psi}(\mathbf{x}) = \left(\begin{array}{c} \frac{1}{2}x_1 \\ \\ 2x_2 \end{array} \right). \tag{A.74}$$

This would correspond to a uniform stretching of the original square along the x_2 direction, and a compression in the x_1 direction, each by a factor of 2. The resulting $D = \boldsymbol{\psi}(B)$ is a long and narrow rectangle of the same area as B. Let $\mathbf{y}$ denote Cartesian coordinates in the new domain. So $\mathbf{y} = \boldsymbol{\psi}(\mathbf{x})$ and $\mathbf{x} = \boldsymbol{\psi}^{-1}(\mathbf{y})$.

In the example in (A.74) $\boldsymbol{\psi}$ is an invertible mapping:

$$\boldsymbol{\psi}^{-1}(\mathbf{y}) = \left(\begin{array}{c} 2y_1 \\ \\ \frac{1}{2}y_2 \end{array} \right),$$

and more generally any $\boldsymbol{\psi}$ of interest will also be invertible.

Now, how should the original vector field (tiny little arrows drawn on B) transform under $\boldsymbol{\psi}$? In the limit as the length of these arrows become infinitely small, they will remain straight. And the base of each arrow originally at a point $\mathbf{x}$ should go to $\boldsymbol{\psi}(\mathbf{x})$. If we wanted to achieve this while keeping the original orientation of the arrows, then in the domain D we would just substitute $\mathbf{x} = \boldsymbol{\psi}^{-1}(\mathbf{y})$ into $\mathbf{a}(\mathbf{x})$ to get $\mathbf{a}(\boldsymbol{\psi}^{-1}(\mathbf{y})) = (\mathbf{a} \circ \boldsymbol{\psi}^{-1})(\mathbf{y})$ to get the representation of this vector field in the new coordinate system. But is this a desirable way to define the morphing of the vector field ?

For example, if the original square were made out of a rubber sheet, then as the sheet is stretched in one direction and compressed in the other, the tiny little arrows should rotate toward the direction of stretching and away from the direction of compression. And their length should change. This intuitive behavior would not be captured by the (straw-man) definition in the previous paragraph. Indeed, the way to update the transformation of the vector field in order to have it behave in the way it should behave intuitively, the Jacobian matrix of the transformation should enter the picture. Since the Jacobian describes how an infinitesimal vector $d\mathbf{x}$ is related to its counterpart $d\mathbf{y}$ as

$$d\mathbf{y} = \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}^T} d\mathbf{x},$$

this same relationship can be applied to $\mathbf{a}(\boldsymbol{\psi}^{-1}(\mathbf{y}))$ to make it transform in a natural way. Therefore we can write

$$\mathbf{a}_*(\mathbf{y}) = \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}^T} \mathbf{a}(\boldsymbol{\psi}^{-1}(\mathbf{y})). \tag{A.75}$$

This is the *push-forward* vector field associated with $\mathbf{a}(\mathbf{x})$. And we can define $\boldsymbol{\psi}_* : \mathbb{R}^2 \to \mathbb{R}^2$ to be the *push-forward* map that takes any vector field on the original domain and maps it to a vector field on the new domain.

One small aesthetic problem is that since $\boldsymbol{\psi} = \boldsymbol{\psi}(\mathbf{x})$, the right-hand-side of (A.75) is an expression that is written in terms of mixed variables ($\mathbf{x}$ and $\mathbf{y}$). This is easily addressed by converting everything to $\mathbf{y}$'s by rewriting every $\mathbf{x}$ as $\boldsymbol{\psi}^{-1}(\mathbf{y})$:

$$\boxed{\mathbf{a}_*(\mathbf{y}) = \left.\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}^T}\right|_{\mathbf{x}=\boldsymbol{\psi}^{-1}(\mathbf{y})} \mathbf{a}(\boldsymbol{\psi}^{-1}(\mathbf{y})).} \tag{A.76}$$

In more modern notation, the vector field $\mathcal{A}$ is

$$\mathcal{A} = \sum_i a_i(\mathbf{x}) \frac{\partial}{\partial x_i} \quad \text{and} \quad \mathcal{A}f = \sum_i a_i(\mathbf{x}) \frac{\partial f}{\partial x_i}$$

where $f(\mathbf{x})$ is an arbitrary differentiable function and

$$(d\boldsymbol{\psi}(\mathbf{x})\mathcal{A})f = \mathcal{A}f(\boldsymbol{\psi}(\mathbf{x})).$$

In modern notation $\mathbf{x}$ and $\boldsymbol{\psi}$ typically would not be written as bold. The differential $d\boldsymbol{\psi}(\mathbf{x})$ can be defined completely in terms of the push-forward map and differential of the mapping $\boldsymbol{\psi}$ such that it satisfies

$$\boxed{(\boldsymbol{\psi}_*\mathbf{a})(\mathbf{y}) = d\boldsymbol{\psi}|_{\mathbf{x}=\boldsymbol{\psi}^{-1}(\mathbf{y})} \mathbf{a}(\boldsymbol{\psi}^{-1}(\mathbf{y})).} \tag{A.77}$$

Returning to (A.76), this can be written in a different way (not in the book). Recall from multivariable calculus that Jacobians have the property

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^T} \frac{\partial \mathbf{x}}{\partial \mathbf{y}^T} = \frac{\partial \mathbf{y}}{\partial \mathbf{y}^T} = \mathbb{I},$$

and so

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^T} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}^T}\right)^{-1}.$$

Since $\mathbf{y} = \boldsymbol{\psi}(\mathbf{x})$ (or equivalently, $\mathbf{x} = \boldsymbol{\psi}^{-1}(\mathbf{y})$), this means that an alternative way to write the same thing is

$$\mathbf{a}_*(\mathbf{y}) = \left(\frac{\partial \boldsymbol{\psi}^{-1}}{\partial \mathbf{y}^T}\right)^{-1} (\mathbf{a} \circ \boldsymbol{\psi}^{-1})(\mathbf{y}). \tag{A.78}$$

A.6.2 The Lagrange Identity

Note that matrices $V \in \mathbb{R}^{n\times p}$ can be written as

$$V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p] \quad \text{where} \quad \mathbf{v}_i \in \mathbb{R}^n$$

or $V^T \in \mathbb{R}^{p\times n}$ where

$$V^T = [\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, ..., \tilde{\mathbf{v}}_n] \quad \text{where} \quad \tilde{\mathbf{v}}_i \in \mathbb{R}^p.$$

Then, it is easy to see that

$$\begin{bmatrix} \mathbf{v}_1\cdot\mathbf{w}_1 & \mathbf{v}_2\cdot\mathbf{w}_1 & \cdots & \mathbf{v}_p\cdot\mathbf{w}_1 \\ \mathbf{v}_1\cdot\mathbf{w}_2 & \mathbf{v}_2\cdot\mathbf{w}_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_1\cdot\mathbf{w}_p & \mathbf{v}_2\cdot\mathbf{w}_p & \cdots & \mathbf{v}_p\cdot\mathbf{w}_p \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^T\mathbf{w}_1 & \mathbf{v}_2^T\mathbf{w}_1 & \cdots & \mathbf{v}_p^T\mathbf{w}_1 \\ \mathbf{v}_1^T\mathbf{w}_2 & \mathbf{v}_2^T\mathbf{w}_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_1^T\mathbf{w}_p & \mathbf{v}_2^T\mathbf{w}_p & \cdots & \mathbf{v}_p^T\mathbf{w}_p \end{bmatrix} = V^T W \tag{A.79}$$

where W is defined in terms of rows and columns in complete analogy with the way that V was.

The determinant of the matrix in (A.79) is then

$$|V^T W| = |W^T V|.$$

When $n < p$ this determinant will be zero. Henceforth only the case $n \geq p$ is considered. Let V_k denote the k^{th} of the $\left(\begin{array}{c} n \\ p \end{array}\right)$ $p \times p$ square sub-matrices of V. Then

$$\boxed{|V^T W| = \sum_{k=1}^{\left(\begin{array}{c} n \\ p \end{array}\right)} |V_k| \cdot |W_k|.} \tag{A.80}$$

The right hand side can be written in several equivalent (though slightly different looking) ways using the fact that $|V_k| = |V_k^T|$ and $|W_k| = |W_k^T|$.

In the case when $p = n$ (A.80) reduces to $|V^T W| = |V| \cdot |W|$. In the case when $n > p$, (A.80) becomes more interesting.

For example, if $p = 2$ and $n = 3$ then

$$V = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{32} \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}.$$

By brute-force evaluation, it can be shown that

$$|V^T W| = \begin{vmatrix} w_{11}v_{11} + w_{21}v_{21} + w_{31}v_{31} & w_{11}v_{12} + w_{21}v_{22} + w_{31}v_{32} \\ w_{12}v_{11} + w_{22}v_{21} + w_{32}v_{31} & w_{12}v_{12} + w_{22}v_{22} + w_{32}v_{32} \end{vmatrix}$$

is equal to

$$\begin{vmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{vmatrix} \cdot \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} + \begin{vmatrix} v_{11} & v_{12} \\ v_{31} & v_{32} \end{vmatrix} \cdot \begin{vmatrix} w_{11} & w_{12} \\ w_{31} & w_{32} \end{vmatrix} + \begin{vmatrix} v_{21} & v_{22} \\ v_{31} & v_{32} \end{vmatrix} \cdot \begin{vmatrix} w_{21} & w_{22} \\ w_{31} & w_{32} \end{vmatrix}.$$

If $p = 1$ and $n = 3$ then

$$V = \begin{pmatrix} v_{11} \\ v_{21} \\ v_{31} \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} w_{11} \\ w_{21} \\ w_{31} \end{pmatrix}$$

and

$$|V^T W| = w_{11}v_{11} + w_{21}v_{21} + w_{31}v_{31} = |V_1| \cdot |W_1| + |V_2| \cdot |W_2| + |V_3| \cdot |W_3|.$$

Given these examples, it should not be difficult to see that (A.80) is true, and induction would be a natural way to prove it in general.

The summation and label k in (A.80) are not very descriptive. Instead, suppose that $1 \leq k_1 < k_2 < \cdots < k_p \leq n$. Then using $\mathbf{e}_k$ to denote the k^{th} natural unit basis vector in $\mathbb{R}^n$,

$$|V_{(k_1,k_2,...,k_p)}| \doteq |\,[\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, ..., \tilde{\mathbf{v}}_n]\,[\mathbf{e}_{k_1}, \mathbf{e}_{k_2}, ..., \mathbf{e}_{k_p}]\,| = \left| \begin{pmatrix} \mathbf{e}_{k_1}^T \\ \mathbf{e}_{k_2}^T \\ \vdots \\ \mathbf{e}_{k_p}^T \end{pmatrix} V \right| \tag{A.81}$$

will be the one of the $\begin{pmatrix} n \\ p \end{pmatrix}$ minors of V in which the p rows k_1, k_2, ..., k_p are chosen. Then (A.80) will be written in a more descriptive way as

$$\boxed{|V^T W| = \sum_{1 \leq k_1 < k_2 < \cdots < k_p \leq n} |V_{(k_1,k_2,...,k_p)}| \cdot |W_{(k_1,k_2,...,k_p)}|.} \tag{A.82}$$

When using $\det[a_{ij}]$ to denote the determinant of $A = [a_{ij}]$ it follows from (A.81) that

$$\begin{aligned} |V_{(k_1,k_2,...,k_p)}| &= |V^T_{(k_1,k_2,...,k_p)}| = |\tilde{\mathbf{v}}_{k_1}, \tilde{\mathbf{v}}_{k_2}, ..., \tilde{\mathbf{v}}_{k_p}| \\ &= \det[\mathbf{e}_{k_i} \cdot \mathbf{v}_j] = \det[\mathbf{v}_i \cdot \mathbf{e}_{k_j}]. \end{aligned} \tag{A.83}$$

Now if $\pi \in \Pi_n$ is a permutation with $\pi(1) < \pi(2) < \cdots < \pi(p)$, then we can set $k_i = \pi(i)$ and this permutation can be used to represent a particular sequence $1 \leq k_1 < k_2 < \cdots < k_p \leq n$. However, since Π_n can contain more than one permutation with this property, for any well-behaved function $f : \mathbb{R}^p \to \mathbb{R}$ we will have

$$\sum_{\pi \in \Pi \,|\, \pi(1)<\pi(2)<\cdots<\pi(p)} f(\pi(1), \pi(2), ..., \pi(p)) = c(n,p) \cdot \left(\sum_{1 \leq k_1 < k_2 < \cdots < k_p \leq n} f(k_1, k_2, ..., k_p) \right) \tag{A.84}$$

where $c(n,p) \geq 1$ is an integer multiplier related to the properties of permutations. The value of $c(n,p)$ is easy to assess. Fixing $\pi(i) = k_i$ for $i = 1, ..., p$ partially constrains π. But $\pi(i)$ for $i = p+1, ..., n$ is completely unconstrained. The group of all possible permutations of these $n-p$ symbols is isomorphic to Π_{n-p}. Therefore there are $|\Pi_{n-p}| = (n-p)!$ possible ways for $\pi(p+1), ...\pi(n)$ to behave, and the mapping defined by the action $\pi \cdot (1, ..., p) \to (k_1, ..., k_p)$ is therefore $(n-p)!$-to-one. And so

$$c(n,p) = (n-p)!$$

Stated in a slightly different way, for a fixed set of p numbers $\{k_1, ..., k_p\}$ ordered as $1 \leq k_1 < k_2 < \cdots < k_p \leq n$, we should expect that in the set of permutations Π_n there should be $(n-p)!$ different ways to map the remaining numbers $n-p$ numbers $\{p+1, ..., n\}$ into $\{1, ..., n\} - \{k_1, ..., k_p\}$.

For example, suppose that $p = 1$ and $n = 2$. Π_2 consists of $2! = 2$ elements, π_1 for which $(\pi_1(1), \pi_1(2)) = (1, 2)$ and π_2 for which $(\pi_2(1), \pi_2(2)) = (2, 1)$. Summing over $f(k_1)$ from $1 \leq k_1 \leq 2$ gives $f(1) + f(2)$ for the right hand side of (A.84). Similarly, summing $f(\pi(1))$ over $\pi \in \Pi_2$ gives $f(\pi_1(1)) + f(\pi_2(1)) = f(1) + f(2)$, and so $c(2, 1) = 1$.

As a second example, if $p = 2$ and $n = 3$, the elements of Π_3 are denoted as

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}; \quad \pi_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}; \quad \pi_3 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix};$$

$$\pi_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}; \quad \pi_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}; \quad \pi_6 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

Here they are labeled from 1 to 6 whereas on p. 324 of Vol. 1 they were labeled from 0 to 5, but this is immaterial. The point is that for any two ordered numbers (k_1, k_2) where $1 \leq k_1 < k_2 \leq 3$, there will be exactly 1 permutation from Π_3 that will map $(1, 2)$ to that ordered set. For example $\pi_1 \cdot (1, 2) = (1, 2)$, $\pi_6 \cdot (1, 2) = (1, 3)$, and $\pi_2 \cdot (1, 2) = (2, 3)$. And so in this case the left hand side of (A.84) gives $f(\pi_1 \cdot (1,2)) + f(\pi_2 \cdot (1,2)) + f(\pi_6 \cdot (1,2))$ while the right hand side of (A.84) gives $f(1,2) + f(1,3) + f(2,3)$. Since these are the same, $c(3,2) = 1$.

On the other hand, if $n = 3$ and $p = 1$, then $1 \leq k_1 \leq 3$ and the sum on the right hand side of (A.84) becomes $f(1) + f(2) + f(3)$. But on the left hand side $\pi_1(1) = \pi_6(1) = 1$, $\pi_2(1) = \pi_4(1) = 2$, and $\pi_3(1) = \pi_5(1) = 3$ indicating that the sum over $\pi \in \Pi_3$, which can be written as $f(\pi_1(1)) + f(\pi_6(1)) + f(\pi_2(1)) + f(\pi_4(1)) + f(\pi_3(1)) + f(\pi_5(1))$ will give twice the value as the sum on the right, and so $c(3,1) = 2$.

Therefore, using permutation notation and (A.83), (A.82) and (A.84) can be written together as the *Lagrange identity* from Exercise 6.17.

$$|V^T W| = [(n-p)!]^{-1} \left(\sum_{\pi \in \Pi_n \mid \pi(1) < \cdots < \pi(p)} \det[\mathbf{w}_i \cdot \mathbf{e}_{\pi(j)}] \det[\mathbf{v}_i \cdot \mathbf{e}_{\pi(j)}] \right). \tag{A.85}$$

Alternatively, instead of summing over the whole of Π_n subject to the constraint $\pi(1) < \cdots < \pi(p)$, π can be chosen to be a coset representative of $\sigma \in \tilde{\Pi}_{n-p} \backslash \Pi_n$ where $\tilde{\Pi}_{n-p} \cong \Pi_{n-p}$ is the isotropy subgroup of Π_n that leaves the first p entries of any permutation in Π_n fixed. Then the factor of $[(n-p)!]^{-1}$ will vanish. (Often the distinction between $\tilde{\Pi}_{n-p}$ and Π_{n-p} is blurred and the tilde is dropped.) The reason why the factor $[(n-p)!]^{-1}$ disappears in this case is that

$$\begin{aligned}
\sum_{\pi \in \Pi_n} f(\pi(1), ...\pi(p)) &= \sum_{\pi' \in \Pi_{n-p}} \sum_{\pi \in \sigma \in \Pi_{n-p} \backslash \Pi_n} f((\pi' \circ \pi)(1), ..., (\pi' \circ \pi)(p)) \\
&= \sum_{\pi' \in \Pi_{n-p}} \sum_{\pi \in \sigma \in \Pi_{n-p} \backslash \Pi_n} f(\pi(1), ..., \pi(p)) \\
&= (n-p)! \cdot \sum_{\pi \in \sigma \in \Pi_{n-p} \backslash \Pi_n} f(\pi(1), ..., \pi(p))
\end{aligned}$$

because $\pi' \in \tilde{\Pi}_{n-p}$ does not affect the first p entries of any $\pi \in \Pi_n$. Therefore, (A.85) can be written in the alternative form

$$|V^T W| = \sum_{\pi \in \sigma \in \Pi_{n-p} \backslash \Pi_n \mid \pi(1) < \cdots < \pi(p)} \det[\mathbf{w}_i \cdot \mathbf{e}_{\pi(j)}] \det[\mathbf{v}_i \cdot \mathbf{e}_{\pi(j)}]. \tag{A.86}$$

A.7 Addendum to Chapter 7

A.7.1 2D Manifolds Generated by Dividing the Plane by Groups

In Section 7.2, the sixth class of examples included the torus, Klein bottle, and real projective plane. These manifolds can be associated with tilings of the

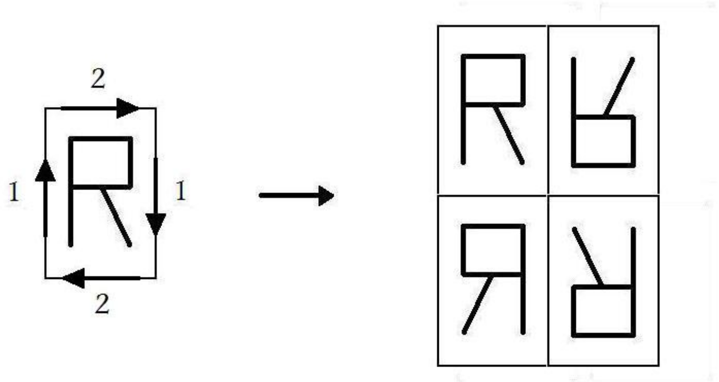

Figure A.1: A Pattern on the Real Projective Plane Transferred to the Euclidean Plane

plane in which the tiles are flipped out of plane and rotated in plane before being pasted down. Transformations were given that can be used to generate all such motions for these three tilings by repeated application of the transformations in different orders.

The set of basic transformations that produces the lattice generated by the asymmetric unit and unit cells shown in Figure A.1 (which is the corrected version of Figure 7.3) are

$$b_1' = \begin{pmatrix} 1 & 0 & w \\ 0 & -1 & h \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad b_2' = \begin{pmatrix} -1 & 0 & w \\ 0 & 1 & -h \\ 0 & 0 & 1 \end{pmatrix}.$$

(Where this is the corrected b_2'.) The set of generators is not unique. For example, it is also possible to use

$$c_1 = \begin{pmatrix} -1 & 0 & w \\ 0 & 1 & h \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad c_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The equivalence of the set of transformations generated by $\{b_1', b_2'\}$ vs. $\{c_1, c_2\}$ follows when we observe that each b_i' can be written as a product of integer powers of c_j's, and vice versa.

A.8 Addendum to Chapter 8: Stochastic Processes on Manifolds

This section supplements the discussion in Chapters 7 and 8. Chapter 7 discussed the classical differential-geometric machinery for describing how to move from one coordinate chart to an overlapping one. Chapter 8 described stochastic differential equations on manifolds from two perspectives: (1) Stratonovich SDEs within a single coordinate chart; (2) Itô SDEs for implicitly defined manifolds embedded in Euclidean space. In this section these two situations are explained in more detail and supplemental examples are given.

A.8.1 Compatibility of Stratonovich SDEs in Different Coordinate Charts

Suppose that two overlapping coordinate charts of an n-dimensional manifold, M are given: $\phi(U) \in \mathbb{R}^n$ and $\tilde{\phi}(V) \in \mathbb{R}^n$ where U and V are connected open subsets of M, and $U \cap V \neq \emptyset$. Then the composite mapping

$$(\tilde{\phi} \circ \phi^{-1}) : \phi(U \cap V) \to \tilde{\phi}(U \cap V) \tag{A.87}$$

is a mapping between open subsets of $\mathbb{R}^n$. If $\mathbf{q}$ are coordinates in $\phi(U)$ and $\tilde{\mathbf{q}}$ are coordinates in $\tilde{\phi}(V)$, then on the overlap $U \cap V$ the mapping $\tilde{\phi} \circ \phi^{-1}$ in (A.87) can be written as an invertible change of coordinates of the form $\tilde{\mathbf{q}} = \tilde{\mathbf{q}}(\mathbf{q})$.

Given a Stratonovich SDE on $\phi(U)$,

$$d\mathbf{q} = \mathbf{a}^s(\mathbf{q}, t)dt + H^s(\mathbf{q}, t) \circledS\, d\mathbf{w}, \tag{A.88}$$

and given a Stratonovich SDE on $\tilde{\phi}(V)$,

$$d\tilde{\mathbf{q}} = \tilde{\mathbf{a}}^s(\tilde{\mathbf{q}}, t)dt + \tilde{H}^s(\tilde{\mathbf{q}}, t) \circledS\, d\mathbf{w}, \tag{A.89}$$

then if these describe the same stochastic process, they must bear some relationship to each other on regions where both coordinate systems apply. In order to obtain this relationship, first note that regular (and Stratonovich) calculus gives

$$d\tilde{\mathbf{q}} = \frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T} d\mathbf{q}. \tag{A.90}$$

Substituting this into (A.89) and inverting the Jacobian gives

$$d\mathbf{q} = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{\mathbf{a}}^s(\tilde{\mathbf{q}}(\mathbf{q}), t))dt + \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{H}^s(\tilde{\mathbf{q}}(\mathbf{q}), t) \circledS\, d\mathbf{w}.$$

A sufficient condition for this to be the same as the process in (A.88) is

$$\boxed{\mathbf{a}^s(\mathbf{q}, t) = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{\mathbf{a}}^s(\tilde{\mathbf{q}}(\mathbf{q}), t)) \quad \text{and} \quad H^s(\mathbf{q}, t) = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{H}^s(\tilde{\mathbf{q}}(\mathbf{q}), t).} \tag{A.91}$$

These compatibility conditions for the SDEs can be used together with the compatibility condition for the metric tensor,

$$G(\mathbf{q}) = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^T \tilde{G}(\tilde{\mathbf{q}}(\mathbf{q}))\, \frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T},$$

to write the Fokker-Planck equation in two different coordinates in a consistent way. In particular, only the square root of the determinant of this metric tensor shows up in the Fokker-Planck equation, and so the key thing to observe is that

$$\boxed{|G(\mathbf{q})|^{\frac{1}{2}} = |\tilde{G}(\tilde{\mathbf{q}}(\mathbf{q}))|^{\frac{1}{2}} \cdot \left|\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right|.}$$

As an example, consider the sphere $S^2 \subset \mathbb{R}^3$. Let U be the open "northern" hemisphere parameterized as

$$\begin{aligned} x_1 &= q_1 \\ x_2 &= q_2 \\ x_3 &= \sqrt{1 - q_1^2 - q_2^2} \end{aligned}$$

and let V be the open "eastern" hemisphere parameterized as

$$\begin{aligned} x_1 &= \tilde{q}_2 \\ x_2 &= \sqrt{1 - \tilde{q}_1^2 - \tilde{q}_2^2} \\ x_3 &= \tilde{q}_1 \end{aligned}$$

The coordinates q_i are simply Cartesian coordinates in the $x_1 - x_2$ plane, and $\tilde{q}_i$ are Cartesian coordinates in the $x_3 - x_1$ plane. In the region where these two charts overlap (which corresponds to one quarter of the whole sphere) we must have

$$\begin{aligned} q_1 &= \tilde{q}_2 \\ q_2 &= \sqrt{1 - \tilde{q}_1^2 - \tilde{q}_2^2} \end{aligned}$$

and

$$\begin{aligned} \tilde{q}_1 &= \sqrt{1 - q_1^2 - q_2^2} \\ \tilde{q}_2 &= q_1. \end{aligned}$$

And so, for example, any Stratonovich SDE written in the coordinates $\tilde{q}_i$ can be converted into one in coordinates q_i by using this explicit $\tilde{\mathbf{q}}(\mathbf{q})$ and the associated Jacobian matrix

$$\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T} = \left(\begin{array}{cc} \frac{-q_1}{\sqrt{1-q_1^2-q_2^2}} & \frac{-q_2}{\sqrt{1-q_1^2-q_2^2}} \\ & \\ 1 & 0 \end{array}\right).$$

A.8.2 Compatibility of Itô SDEs in Different Coordinate Charts

Consider the same manifold and coordinate charts as in the previous subsection. Given an Itô SDE on $\phi(U)$,

$$d\mathbf{q} = \mathbf{a}(\mathbf{q}, t)dt + H(\mathbf{q}, t)\, d\mathbf{w}, \tag{A.92}$$

and given an Itô SDE on $\tilde{\phi}(V)$,

$$d\tilde{\mathbf{q}} = \tilde{\mathbf{a}}(\tilde{\mathbf{q}}, t)dt + \tilde{H}(\tilde{\mathbf{q}}, t)\, d\mathbf{w}, \tag{A.93}$$

it is possible to define compatibility conditions that are analogous with the Stratonovich case. However, these conditions are more complicated because instead of (A.90), Itô's rule must be used. Explicitly, this gives

$$d\tilde{\mathbf{q}} = \frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T} d\mathbf{q} + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial^2 \tilde{\mathbf{q}}}{\partial q_k \partial q_l} dq_k dq_l.$$

Substituting the components of (A.92) into the right side and evaluating expectations in the usual way (i.e., $dw_i dt = 0$ and $dw_i dw_j = \delta_{ij} dt$) gives

$$d\tilde{\mathbf{q}} = \frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T} d\mathbf{q} + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial^2 \tilde{\mathbf{q}}}{\partial q_k \partial q_l} \sum_{s=1}^{m} H_{ks}(\mathbf{q}, t) H_{ls}(\mathbf{q}, t) dt \tag{A.94}$$

where $H \in \mathbb{R}^{n \times m}$.

Substituting this into (A.93), and inverting the Jacobian matrix gives the Itô SDE

$$d\mathbf{q} = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \left[\tilde{\mathbf{a}}(\tilde{\mathbf{q}}(\mathbf{q}), t) - \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial^2 \tilde{\mathbf{q}}}{\partial q_k \partial q_l} \sum_{s=1}^{m} H_{ks}(\mathbf{q}, t) H_{ls}(\mathbf{q}, t)\right] dt + \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{H}(\tilde{\mathbf{q}}(\mathbf{q}), t) \circledS\, d\mathbf{w}.$$

In other words, sufficient conditions for compatibility are

$$\mathbf{a}(\mathbf{q}, t) = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \left[\tilde{\mathbf{a}}(\tilde{\mathbf{q}}(\mathbf{q}), t) - \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial^2 \tilde{\mathbf{q}}}{\partial q_k \partial q_l} \sum_{s=1}^{m} H_{ks}(\mathbf{q}, t) H_{ls}(\mathbf{q}, t)\right] \tag{A.95}$$

and

$$H(\mathbf{q}, t) = \left(\frac{\partial \tilde{\mathbf{q}}}{\partial \mathbf{q}^T}\right)^{-1} \tilde{H}(\tilde{\mathbf{q}}(\mathbf{q}), t). \tag{A.96}$$

Note that (A.95) can be written in terms of $\tilde{H}$ rather than H by back substituting (A.96). These conditions for conversion between coordinates should not be confused with the conversion between Itô and Stratonovich forms. That conversion was discussed in Section A.3.4 and is independent of the current discussion, which transforms one Itô SDE into another. Note also that the discussion of how the metric tensor transforms under coordinate changes is not affected by the use of Itô or Stratonovich forms of SDEs; regular calculus is used in the context of all computations associated with Fokker-Planck equations.

A.8.3 Itô SDEs on Implicitly Defined Embedded Manifolds

Below several Itô SDEs in $\mathbb{R}^n$ in Cartesian coordinates are given that that describe processes that evolve on embedded manifolds. This approach is very different than the SDEs defined in coordinate patches in the previous subsection. Whereas the Stratonovich form was more convenient in the context of transitioning between coordinate charts, the Itô form is more suitable for implicitly defined embedded manifolds.

A.8.4 Example 1: An Itô SDE on the Circle

The unit circle in the plane is defined by the condition

$$f(x_1, x_2) \doteq x_1^2 + x_2^2 = 1. \tag{A.97}$$

In (8.25) the Itô SDE

$$\begin{aligned} dx_1 &= h_1(\mathbf{x},t)dt + H_1(\mathbf{x},t)dw \\ dx_2 &= h_1(\mathbf{x},t)dt + H_2(\mathbf{x},t)dw. \end{aligned} \tag{A.98}$$

was given,

$$h_1(\mathbf{x},t) = -\frac{1}{2}x_1, \; H_1(\mathbf{x},t) = -x_2, \; h_2(\mathbf{x},t) = -\frac{1}{2}x_2, \; H_2(\mathbf{x},t) = +x_1 \tag{A.99}$$

and it was shown by introducing the coordinate θ, where $x_1 = \cos\theta$ and $x_2 = \sin\theta$ that this SDE evolves on the unit circle. Here the same fact is shown without introducing the coordinate θ.

Suppose that the initial conditions are such that $f(x_1(0), x_2(0)) = 1$, as must be the case if the starting condition is on the circle. The condition that the solution will stay on the circle is $df = 0$. However, since the governing SDE is of Itô type, the evaluation of this condition requires the use of Itô rule. In particular,

$$df = \left(\sum_j \frac{\partial f}{\partial x_j} h_j(\mathbf{x},t) + \frac{1}{2}\sum_{k,l} \frac{\partial^2 f}{\partial x_k \partial x_l}[H(\mathbf{x},t)H^T(\mathbf{x},t)]_{kl}\right) dt + \sum_k \frac{\partial f}{\partial x_k} H_k(\mathbf{x},t)dw \tag{A.100}$$

Computing the derivatives of $f(x_1, x_2)$ in (A.97) and substituting these and (A.99) into (A.100) gives

$$df|_{\mathbf{x}=\mathbf{x}(t)} = 0,$$

indicating that the trajectory stays on the circle.

A.8.5 Example 2: An Itô SDE on the Sphere

The following Itô SDE appears as (8.35) in Vol. 1:

$$d\mathbf{x} = \mathbf{h}(\mathbf{x},t)dt + H(\mathbf{x},t)d\mathbf{w}$$

where

$$\mathbf{h}(\mathbf{x},t) = -\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \text{and} \quad H(\mathbf{x},t) = \begin{pmatrix} x_2 & x_3 & 0 \\ -x_1 & 0 & x_3 \\ 0 & -x_1 & -x_2 \end{pmatrix}$$

and $\mathbf{x}, \mathbf{w} \in \mathbb{R}^3$. This defines stochastic paths $\mathbf{x}(t)$.

Using a similar procedure as in Example 1, if

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$$

and $f(x_1(0), x_2(0), x_3(0)) = 1$, then showing that $df = 0$ means that the trajectory will evolve on the unit sphere. Substituting all of the above quantities into Itô's rule, which in the present case is

$$df = \left(\sum_j \frac{\partial f}{\partial x_j} h_j(\mathbf{x},t) + \frac{1}{2} \sum_{k,l} \frac{\partial^2 f}{\partial x_k \partial x_l} [H(\mathbf{x},t)H^T(\mathbf{x},t)]_{kl} \right) dt + \sum_{k,l} \frac{\partial f}{\partial x_k} H_{kl}(\mathbf{x},t) dw_l, \tag{A.101}$$

verifies that $df|_{\mathbf{x}=\mathbf{x}(t)} = 0$.

A.8.6 Example 3: An Itô SDE on the Rotation Group

Consider the following Itô SDE studied by Brockett [1, 2]:

$$dR = -\frac{(n-1)}{2} R\, dt + \sum_{i',j'=1}^{n} (E_{i'j'} - E_{j'i'}) R\, dw_{i'j'} \tag{A.102}$$

where dw_{ij} for $i, j = 1, ..., n$ are n^2 uncorrelated unit-strength white noises and $E_{i'j'}$ is the matrix with the number $1/\sqrt{2}$ in the $i'j'^{th}$ entry and zero everywhere else. In other words, the ij^{th} entry of the matrix $E_{i'j'}$ is

$$(E_{i'j'})_{ij} = \frac{1}{\sqrt{2}} \delta_{ii'} \delta_{jj'}.$$

In this way $E_{i'j'} - E_{j'i'}$ is a unit basis vector for the Lie algebra $so(n)$ in the sense that $\|E_{i'j'} - E_{j'i'}\|_F = 1$.

In order to show that this SDE on $\mathbb{R}^{n \times n}$ describes a process $R(t)$ that evolves on $SO(n)$ (viewed as an implicitly defined embedded manifold), we must show that when $R(0) \in SO(n)$

$$dF|_{R=R(t)} = \mathbb{O} \quad \text{where} \quad F(R) = R^T R - \mathbb{I}. \tag{A.103}$$

The statement $F(R) = \mathbb{O}$ is the constraint of orthogonality, and $\det R(t) = +1$ is naturally satisfied if the same condition holds at $t = 0$ while orthogonality is enforced.

Two ways to address this problem are: (1) to use the $\vee : \mathbb{R}^{n\times n} \rightarrow \mathbb{R}^{n^2}$ operation and convert (A.102) into the standard form of vector-valued stochastic processes and use the standard form of Itô's formula; (2) to develop a variation of Itô's formula specifically to handle matrix-valued stochastic processes. The second approach is taken here, and the first is left as an exercise.

The key to extending Itô's formula to matrix-valued stochastic processes such as (A.102), or more generally

$$dR_{ij} = A_{ij}dt + \sum_{i'j'} B_{ij;i'j'}dw_{i'j'}, \tag{A.104}$$

where

$$A_{ij} = A_{ij}(R,t) \quad \text{and} \quad B_{ij;i'j'} = B_{ij;i'j'}(R,t),$$

is to observe that

$$\langle dw_{ij}dw_{i'j'}\rangle = \delta_{ii'}\delta_{jj'} \quad \text{and} \quad \langle dw_{ij}dt\rangle = 0$$

and the usual rules for computing expectations of more complicated expressions apply. For example,

$$\left\langle B_{ij;i'j'}dw_{i'j'}B_{i_1j_1;i'_1j'_1}dw_{i'_1j'_1}\right\rangle = B_{ij;i'j'}B_{i_1j_1;i'_1j'_1}\delta_{i'i'_1}\delta_{j'j'_1}.$$

The evaluation of the left-hand-side of (A.103) in terms of components using Itô's rule becomes

$$\begin{aligned} dF_{rs} &= \left(\sum_{ij}\frac{\partial F_{rs}}{\partial R_{ij}}A_{ij} + \frac{1}{2}\sum_{i,j,k,l}\frac{\partial^2 F_{rs}}{\partial R_{ij}\partial R_{kl}}\sum_{i'j'}B_{ij;i'j'}B_{kl;i'j'}\right)dt \\ &\quad + \sum_{i,j,k,l}\frac{\partial F_{rs}}{\partial R_{ij}}B_{ij;kl}dw_{kl} \end{aligned} \tag{A.105}$$

Returning to (A.102) and fitting it into the component form in (A.104) we see that

$$\begin{aligned} dR_{ij} &= -\frac{(n-1)}{2}R_{ij}dt + \sum_{l,i',j'=1}^{n}(E_{i'j'} - E_{j'i'})_{il}R_{lj}dw_{i'j'} \\ &= -\frac{(n-1)}{2}R_{ij}dt + \frac{1}{\sqrt{2}}\sum_{l,i',j'=1}^{n}(\delta_{ii'}\delta_{lj'} - \delta_{j'i}\delta_{i'l})R_{lj}dw_{i'j'} \\ &= -\frac{(n-1)}{2}R_{ij}dt + \frac{1}{\sqrt{2}}\sum_{i',j'=1}^{n}(\delta_{ii'}R_{j'j} - \delta_{j'i}R_{i'j})dw_{i'j'}. \end{aligned}$$

Therefore,

$$A_{ij} = -\frac{(n-1)}{2}R_{ij} \quad \text{and} \quad B_{ij;i'j'} = (\delta_{ii'}R_{j'j} - \delta_{j'i}R_{i'j})/\sqrt{2}.$$

The component form of (A.103) is

$$F_{rs} = \sum_{p=1}^{n} R_{pr} R_{ps} - \delta_{rs}.$$

From this it is clear (from regular calculus) that

$$\begin{aligned}
\frac{\partial F_{rs}}{\partial R_{ij}} &= \sum_p \left\{ \frac{\partial R_{pr}}{\partial R_{ij}} R_{ps} + R_{pr} \frac{\partial R_{ps}}{\partial R_{ij}} \right\} \\
&= \sum_p \left\{ \delta_{ip}\delta_{jr} R_{ps} + R_{pr}\delta_{pi}\delta_{sj} \right\} \\
&= \delta_{jr} R_{is} + R_{ir}\delta_{sj}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 F_{rs}}{\partial R_{ij}\partial R_{kl}} &= \frac{\partial}{\partial R_{kl}} \left(\delta_{jr} R_{is} + R_{ir}\delta_{sj}\right) \\
&= (\delta_{jr}\delta_{ls} + \delta_{sj}\delta_{lr})\delta_{ik}.
\end{aligned}$$

Substituting these into (A.105), we find

$$\begin{aligned}
\sum_{ij} \frac{\partial F_{rs}}{\partial R_{ij}} A_{ij} &= -\frac{(n-1)}{2} \sum_{ij} R_{ij}(\delta_{jr} R_{is} + R_{ir}\delta_{sj}) \\
&= -\frac{(n-1)}{2} \left(\sum_i R_{ir} R_{is} + \sum_i R_{is} R_{ir} \right) && \text{(A.106)} \\
&= -\frac{(n-1)}{2} (\delta_{rs} + \delta_{sr}) \\
&= -(n-1)\delta_{rs} && \text{(A.107)}
\end{aligned}$$

where at (A.106) we substitute $F(R) = \mathbb{O}$.

Next observe that

$$\begin{aligned}
\sum_{i'j'} B_{ij;i'j'} B_{kl;i'j'} &= \frac{1}{2} \sum_{i'j'} (\delta_{ii'} R_{j'j} - \delta_{j'i} R_{i'j})(\delta_{ki'} R_{j'l} - \delta_{j'k} R_{i'l}) \\
&= \frac{1}{2} \left[\delta_{ik} \sum_{j'} R_{j'j} R_{j'l} + \delta_{ik} \sum_{i'} R_{i'j} R_{i'l} - R_{kj} R_{il} - R_{kj} R_{il} \right] \\
&= \delta_{ik}\delta_{jl} - R_{kj} R_{il} && \text{(A.108)}
\end{aligned}$$

where to get to (A.108) we inserted $F(R) = \mathbb{O}$ and so

$$\begin{aligned}
\frac{1}{2} \sum_{i,j,k,l} \frac{\partial^2 F_{rs}}{\partial R_{ij}\partial R_{kl}} \sum_{i'j'} B_{ij;i'j'} B_{kl;i'j'} &= \frac{1}{2} \sum_{i,j,k,l} (\delta_{jr}\delta_{ls} + \delta_{sj}\delta_{lr})\delta_{ik}(\delta_{ik}\delta_{jl} - R_{kj} R_{il}) \\
&= (n-1)\delta_{sr}. && \text{(A.109)}
\end{aligned}$$

Therefore, combining (A.107) and (A.109) the term in parenthesis in (A.105) vanishes. Furthermore,

$$\begin{aligned}\sum_{i,j}\frac{\partial F_{rs}}{\partial R_{ij}}B_{ij;kl} &= \sum_{i,j}(\delta_{jr}R_{is}+R_{ir}\delta_{sj})(\delta_{ii'}R_{j'j}-\delta_{j'i}R_{i'j})\\ &= \sum_{i,j}\{\delta_{jr}R_{is}\delta_{ii'}R_{j'j}+R_{ir}\delta_{sj}\delta_{ii'}R_{j'j}-\delta_{jr}R_{is}\delta_{j'i}R_{i'j}-R_{ir}\delta_{sj}\delta_{j'i}R_{i'j}\}\\ &= R_{i's}R_{j'r}+R_{i'r}R_{j's}-R_{j's}R_{i'r}-R_{j'r}R_{i's}\\ &= 0.\end{aligned}$$

This verifies that $dF|_{R=R(t)} = \mathbb{O}$ when $R(t)$ is defined by (A.102) with $R(0) \in SO(n)$.

A.8.7 Example 4: A Class of Stratonovich SDEs on the Rotation Group

Since the Stratonovich calculus behaves in the same way as usual calculus, the condition $R^T R = \mathbb{I}$ can be enforced as $R^T dR + (dR)^T R = \mathbb{O}$, or $R^T dR = -(R^T dR)^T$. This means that

$$dR = \sum_{i,j=1}^{n}(E_{ij}-E_{ji})R\,\omega_{ij}\,dt \tag{A.110}$$

will evolve on the rotation group $SO(n)$ where

$$\omega_{ij}(t)dt = a_{ij}(R,t)dt + \sum_{i'j'} b_{ij;i'j'}(R,t) \circledS\, dw_{i'j'}.$$

In the special case when $a_{ij}(R,t) = 0$ and $b_{ij;i'j'}(R,t) = \delta_{ii'}\delta_{jj'}$ this Stratonovich equation will be equivalent to the Itô equation in Example 3.

A.9 Supplemental Exercises

E.1 Consider the ratio $r(k,n,p) \doteq f(k;n,p)/\rho(k;np,np(1-p))$ in (2.23). Let $n \in \{2,4,10,20\}$. Choose $p \in \{0,1/2,1\}$. Then $k = np$ will be an integer. For each of these choices, plot $r(k,n,p)$ as a function of n. Does $r(k,n,p) \to 1$ as n increases ?

E.2 Evaluate the entropy power inequality for multivariate Gaussian distributions of the form $\rho(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)$ and $\rho(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)$ as an inequality relating their covariance matrices. Does anything special happen in the case when $\Sigma_1 = c\Sigma_2$?

E.3 In the one-dimensional case show that over all distributions with variance σ^2, the Fisher information is minimized by the Gaussian distribution $\rho_{(\mu,\sigma^2)}(x)$. Is there a multi-dimensional extension of this, and if so, what is it ?

E.4 Verify (A.17) and (A.18) and compute the following for the distribution $f_{(c,\sigma^2)}(x)$ defined in (A.19): (a) the convolution $(f_{(c_1,\sigma_1^2)} * f_{(c_2,\sigma_2^2)})(x)$; (b) The Fisher information $F(f_{(c,\sigma^2)})$; (c) The entropy $S(f_{(c,\sigma^2)})$.

E.5 Using the results of E.4, verify the Fisher information inequality

$$F(f_{(c_1,\sigma_1^2)} * f_{(c_2,\sigma_2^2)}) \leq \frac{f_{(c_1,\sigma_1^2)} \cdot f_{(c_2,\sigma_2^2)}}{f_{(c_1,\sigma_1^2)} + f_{(c_2,\sigma_2^2)}}$$

and the entropy-power inequality

$$N(f_{(c_1,\sigma_1^2)} * f_{(c_2,\sigma_2^2)}) \geq N(f_{(c_1,\sigma_1^2)}) + N(f_{(c_2,\sigma_2^2)}).$$

E.6 Substitute (A.12) into the Fourier inversion formula

$$F(x,y,z) = \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{F}(\omega_x,\omega_y,\omega_z)\exp i(x\omega_x+y\omega_y+z\omega_z)d\omega_x d\omega_y d\omega_z$$

and using the convolution theorem show that (A.9)-(A.11) are satisfied.

E.7 Show for both the Itô or Stratonovich forms of the Fokker-Planck equation given in coordinates $\mathbf{q}$ that generate a solution $f(\mathbf{q},t)$ that these equations will also hold in coordinates $\mathbf{s}$ and generate $\tilde{f}(\mathbf{s},t) = f(\mathbf{q}(\mathbf{s}),t)$ where the relationships $\mathbf{q} = \mathbf{q}(\mathbf{s})$ and $\mathbf{s} = \mathbf{s}(\mathbf{q})$ are invertible.

E.8 Suppose that the Itô and Stratonovich SDEs for a given stochastic process are not the same in a particular set of coordinates. Under what conditions will it be possible to define a new set of coordinates in which the Itô and Stratonovich SDEs are the same ?

E.9 Consider the function $\phi(\mathbf{x}) = -\rho(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1) - \rho(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)$. If $\boldsymbol{\mu}_i$ are relatively well separated and Σ_i are relatively small, this will be a function with two minima close to (but, due to interaction of the tails, not exactly at) $\boldsymbol{\mu}_i$. Pick some values of $\boldsymbol{\mu}_i$ and Σ_i. (a) Write the Fokker-Planck equation corresponding to

$$d\mathbf{x} = -\nabla\phi dt + \sigma(t)d\mathbf{w}.$$

(b) How should you select $\sigma(t)$ if you want with high probability $\mathbf{x}(t)$ to settle in to the global (deeper) minimum from any starting point ? (Hint: If $\sigma(t) \rightarrow 0$ as $t \rightarrow \infty$ the solution will settle down somewhere.) (c) Use your "cooling schedule" $\sigma(t)$ in 100 numerical simulations of the SDE with random initial values of $\mathbf{x}(0) = \mathbf{x}_0$. Where does the solution settle most of the time using your cooling schedule ? For other stochastic search techniques see [7].

E.10. Let $\rho(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1)$ and $\rho(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)$ be Gaussians on $\mathbb{R}^n$. Let

$$\phi(\boldsymbol{\mu}_1, \Sigma_1; \boldsymbol{\mu}_2, \Sigma_2) = \int_{\mathbb{R}^n} |\rho(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1) - \rho(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)|^2 \, d\mathbf{x}.$$

First, compute $\phi(\boldsymbol{\mu}_1, \Sigma_1; \boldsymbol{\mu}_2, \Sigma_2)$ as a closed-form expression.

E.11. Let $\mathbf{y} = \boldsymbol{\psi}(\mathbf{x}) = g \cdot \mathbf{x} = R\mathbf{x} + \mathbf{t}$ be a rigid-body transformation on $\mathbb{R}^n$. Let $\mathbf{a}(\mathbf{x})$ be a vector field on $\mathbb{R}^n$. Calculate the following: $\boldsymbol{\psi}^{-1}(\mathbf{y})$, $\partial\boldsymbol{\psi}/\partial\mathbf{x}^T$, $\partial\boldsymbol{\psi}^{-1}/\partial\mathbf{y}^T$, $\boldsymbol{\psi}_*$, $d\boldsymbol{\psi}$, and $\mathbf{a}_*(\mathbf{y})$ using the different variants in the expressions in (A.75)-(A.78).

E.12. When defining transformation laws for vector fields using the push forward, are they contravariant or covariant ? And how do these compare with the vector fields and Cartesian tensors used in engineering ?

E.13. Use the $\vee : \mathbb{R}^{n\times n} \rightarrow \mathbb{R}^{n^2}$ operation and convert (A.102) into the standard form of vector-valued stochastic processes and use the standard form of Itô's formula to show that when $R(0) \in SO(n)$ then so too is $R(t)$. Hint: Use the properties of the Kronecker product.

E.14. Compare the SDE for the kinematic cart, which is a stochastic process on the motion group of the plane, $SE(2)$, when viewed as an Itô equation and when viewed as Stratonovich. Compare the resulting Fokker-Planck equations. Write this as an implicit equation viewing $SE(2)$ as being embedded in $\mathbb{R}^{3\times 3}$.

Bibliography

[1] Brockett, R.W., "Lie algebras and Lie groups in control theory," in *Geometric Methods in System Theory*, (D.Q. Mayne and R.W. Brockett, eds.), Reidel Publishing Company, Dordrecht-Holland, 1973.

[2] Brockett, R.W., "Notes on Stochastic Processes on Manifolds," in *Systems and Control in the Twenty-First Century* (C.I. Byrnes et al eds.), Birkhäuser, Boston, 1997.

[3] Chirikjian, G.S., *Stochastic Models, Information Theory, and Lie Groups, Vol. 1*, Birkhäuser, Boston, 2009.

[4] Cover, T.M., Thomas, J.A., *Elements of Information Theory*, Wiley-Interscience, 2^{nd} ed., Hoboken, NJ, 2006.

[5] Fill, J., Email communication, Sunday, October 11, 2009 3:17 am

[6] Pinkser, M.S., *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco, 1964.

[7] Spall, J.C., *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, John Wiley and Sons, 2003

[8] Strassen, V., "The Existence of Probability Measures with Given Marginals," *The Annals of Mathematical Statistics*, Volume 36, Number 2 (1965), 423-439.