Random vectors

Let $(\Omega, \mathcal{B}, P)$ be a probability space. A real-valued random vector

$$\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$$

is a measurable map from the sample space Ω into $\mathbb{R}^n$, i.e.,

$$\boldsymbol{X} : \Omega \to \mathbb{R}^n. \tag{1}$$

Each component of $\boldsymbol{X}(\omega)$, say $X_i(\omega)$, is a real-valued random variable. As before, we can push forward the probability measure P from Borel sets of Ω to Borel sets of $\mathbb{R}^n$ via the mapping (1), i.e.,

$$P_{\boldsymbol{X}}(A) = \mathcal{B}(\mathbb{R}^n) \mapsto [0,1] \tag{2}$$

where $P_{\boldsymbol{X}}(A)$ is defined as[1]

$$P_{\boldsymbol{X}}(A) = P\left(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in A\}\right) \quad \text{for all} \quad A \in \mathcal{B}(\mathbb{R}^n). \tag{4}$$

Joint CDFs and PDFs. The *cumulative distribution function* (CDF) of a $\boldsymbol{X}(\omega)$ is defined as

$$F(x_1, \ldots, x_n) = P(\underbrace{\{\omega : X_1(\omega) \leq x_1\} \cap \cdots \cap \{\omega : X_n(\omega) \leq x_n\}}_{\text{event in } \mathcal{B}(\Omega) \text{ defined as intersection of } n \text{ events}}). \tag{5}$$

As before, if P is absolutely continuous with respect to the Lebesgue measure $dx_1 \cdots dx_n$ then there exists a (Lebesgue integrable) *probability density function*[2] (PDF) $p(x_1, \ldots, x_n)$ such that

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p(y_1, \ldots, y_n) dy_1 \cdots dy_n. \tag{6}$$

Equivalently, we can express $p(x_1, \ldots, x_n)$ as a (weak) derivative of $F(x_1, \ldots, x_n)$ as

$$p(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}. \tag{7}$$

The multivariate distribution function F and associated probability density function p satisfy similar properties as the properties we have seen for one one random variable (see [13] for details). For instance F is non-decreasing, with range in $[0,1]$, etc. Similarly, p is non-negative, and it allows us to compute the probability of the event

$$\{\omega \in \Omega : \boldsymbol{X}(\omega) \in A\}$$

as

$$P(A) = \int_A p(x_1, \ldots, x_n) dx_1 \cdots dx_n. \tag{8}$$

Frequency interpretation of the joint PDF: Consider a continuous random vector $\boldsymbol{X}(\omega)$ (i.e., a random vector with continuous CDF) with only two components, say $X_1(\omega)$ and $X_2(\omega)$. By using equations (5) and (6) we have

$$P(\{\omega : x_1 \leq X_1(\omega) \leq x_1 + \Delta x_1\} \cap \{\omega : x_2 \leq X_2(\omega) \leq x_2 + \Delta x_2\}) \simeq p(x_1, x_2) \Delta x_1 \Delta x_2. \tag{9}$$

[1] The set

$$\boldsymbol{X}^{-1}(A) = \{\omega \in \Omega : \boldsymbol{X}(\omega) \in A\} \tag{3}$$

is known as pre-image of A under the mapping $\boldsymbol{X}(\omega)$.

[2] Technically speaking, the joint probability density function $p(x_1, \ldots, x_n)$ is the Radon-Nikodym derivative of the probability measure P relative to the Lebesgue measure $dx_1 \cdots dx_n$.

Let us partition the tensor product space $\mathbb{R}^2$ with an evenly-spaced grid of width Δx_1 (along x_1) and Δx_2 (along x_2). Suppose we observe S realizations of $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$, and suppose that $n_A < S$ instances satisfy the condition

$$\{x_1 \leq X_1(\omega) \leq x_1 + \Delta x_1\} \quad \text{and} \quad \{x_2 \leq X_2(\omega) \leq x_2 + \Delta x_2\}. \tag{10}$$

Then from (9) we obtain the PDF estimate

$$p(x_1, x_2) \simeq \frac{1}{\Delta x_1 \Delta x_2} \frac{n_A}{n}. \tag{11}$$

Of course there are methods other than relative frequencies to estimate PDFs from data/observations. Among them, we have kernel methods [2] (see Figure 1) and generative modeling techniques based on diffusion [14].

Marginal CFDs and PDFs. Let $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$ be a random vector with joint distribution function function $F(x_1, x_2)$. The distribution of the random variable $X_1(\omega)$ can be obtained from $F(x_1, x_2)$ simply by sending x_2 to infinity, i.e.,

$$F(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2). \tag{12}$$

In fact,

$$\lim_{x_2 \to \infty} F(x_1, x_2) = P(\{\omega : X_1(\omega) \leq x_1\} \cap \{\omega : X_2(\omega) \leq \infty\}) = P(\{\omega : X_1(\omega) \leq x_1\}) = F(x_1). \tag{13}$$

We can write the last equation in terms of PDFs as

$$\lim_{x_2 \to \infty} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} p(y_1, y_2) dy_1 dy_2 = \int_{-\infty}^{x_1} p(y_1) dy_1. \tag{14}$$

Since x_1 is arbitrary, it follows from (14) that

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2 \qquad \text{(marginalization rule)}. \tag{15}$$

Moreover, we have $F(\infty, \infty) = 1$, i.e.,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2) dx_1 dx_2 = 1 \qquad \text{(normalization condition)}. \tag{16}$$

It is straightforward to extend these formulas to distribution functions and PDFs in more than two variables. For example, if $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega))$ is a four-dimensional random vector with distribution function $F(x_1, \ldots, x_4)$ and PDF $p(x_1, \ldots, x_4)$, then we can obtain the joint distribution function and the joint PDF of X_2 and X_3, respectively, as

$$F(x_2, x_3) = F(\infty, x_2, x_3, \infty), \qquad p(x_2, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2, x_3, x_4) dx_1 dx_4. \tag{17}$$

Example (Gaussian distribution): Consider the multivariate Gaussian PDF

$$p(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})/2}, \tag{18}$$

where

$$\boldsymbol{x}^T = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}, \tag{19}$$

$$\boldsymbol{\mu}^T = \begin{bmatrix} \mathbb{E}\{X_1\} & \dots & \mathbb{E}\{X_n\} \end{bmatrix} \qquad \text{(mean)}, \tag{20}$$

$$\Sigma_{ij} = \mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\,\mathbb{E}\{X_j\} \qquad \text{(covariance matrix)}. \tag{21}$$

It is straightforward to show that all marginal PDF and distribution functions are still Gaussians of the form (18).

Independence. Let $(\Omega, \mathcal{B}, P)$ be a probability space. Two events $A \in \mathcal{B}$ and $B \in \mathcal{B}$ are said to be *independent* if the probability of their intersection (i.e., the probability that both events A and B happen) equals the product of their probabilities, i.e.,

$$A, B \in \mathcal{B} \quad \text{independent} \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B). \tag{22}$$

Consider now a random vector $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$ with components $X_1(\omega)$ and $X_2(\omega)$. We say that the random variables $X_1(\omega)$ and $X_2(\omega)$ are statistically independent if

$$P(\underbrace{\{\omega : X_1(\omega) \leq x_1\}}_{\text{event } A} \cap \underbrace{\{\omega : X_2(\omega) \leq x_2\}}_{\text{event } B}) = P(\{\omega : X_1(\omega) \leq x_1\})P(\{\omega : X_2(\omega) \leq x_2\}), \tag{23}$$

for all $x_1, x_2 \in \mathbb{R}$. Equation (23) can be written in terms of the cumulative distribution function as

$$F(x_1, x_2) = F(x_1)F(x_2). \tag{24}$$

This also implies that the joint PDF of X_1 and X_2 (if it exists) is simply the product of the PDF of X_1 and the PDF of X_2, i.e.,

$$p(x_1, x_2) = p(x_1)p(x_2). \tag{25}$$

These formulas can be generalized to n independent random variables as

$$F(x_1, \dots, x_n) = F(x_1) \cdots F(x_n), \qquad p(x_1, \dots, x_n) = p(x_1) \cdots p(x_n). \tag{26}$$

Examples:

- *Jointly uniform random vector.* Let $\boldsymbol{X}$ be a n-dimensional random vector with zero-mean i.i.d. (independent identically distributed) uniform components in [-1,1]. The joint PDF of $\boldsymbol{X}$ is

$$p(x_1, \dots, x_n) = \begin{cases} \dfrac{1}{2^n} & (x_1, \dots, x_n) \in [-1, 1]^n \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

- *Jointly normal random vector.* Let $\boldsymbol{X}$ be a n-dimensional random vector with zero-mean i.i.d. Gaussian components with variance equal to one. The joint PDF of $\boldsymbol{X}$ is

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\boldsymbol{x}^T \boldsymbol{x}/2} \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{28}$$

  Clearly, from equation (18) we see that Gaussian random variables are independent if and only if

$$\mathbb{E}\{X_i X_j\} = \mathbb{E}\{X_i\}\mathbb{E}\{X_j\} \quad \text{for } i \neq j. \tag{29}$$

  In general, if (29) is satisfied then we say that X_i and X_j are *uncorrelated*. Lack of correlation is a much weaker statement than independence, yet sufficient to claim independence for Gaussian random variables.

Expectation, joint moments, and joint cumulants. Let $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ be a random vector defined on the probability space $(\Omega, \mathcal{B}, P)$. For any measurable function $g(X_1, \ldots, X_n)$ we define the expectation[3] as

$$\mathbb{E}\left\{g(X_1, \ldots, X_n)\right\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n) dx_1 \cdots dx_n. \tag{31}$$

In particular, if $g(X_1, \ldots, X_n) = X_1^{k_1} \cdots X_n^{k_n}$ then

$$\mathbb{E}\left\{X_1^{k_1} \cdots X_n^{k_n}\right\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{k_1} \cdots x_n^{k_n} p(x_1, \ldots, x_n) dx_1 \cdots dx_n \qquad \text{(joint moments)} \tag{32}$$

The correlation matrix[4] and the covariance matrix are defined as (see, e.g., (18))

$$\mathbb{E}\left\{X_i X_j\right\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p(x_i, x_j) dx_i dx_j \qquad \text{(correlation matrix)}, \tag{34}$$

$$\mathbb{E}\left\{(X_i - \mu_i)(X_j - \mu_j)\right\} = \mathbb{E}\left\{X_i X_j\right\} - \mu_i \mu_j \qquad \text{(covariance matrix)}. \tag{35}$$

where $\mu_i = \mathbb{E}\left\{X_i\right\}$ (mean of X_i).

Remark: We say that two random variables $X_i(\omega)$ and $X_j(\omega)$ are *uncorrelated* if

$$\mathbb{E}\left\{X_i X_j\right\} = \mathbb{E}\left\{X_i\right\} \mathbb{E}\left\{X_j\right\}. \tag{36}$$

Independent random variables are always uncorrelated. In fact, let $p(x_i, x_j)$ be the joint PDF of X_i and X_j. We know that if X_i and X_j are independent then $p(x_i, x_j)$ can be factorized as

$$p(x_i, x_j) = p(x_i) p(x_j). \tag{37}$$

A substitution of (37) into (34) immediately yields (36). However, uncorrelated random variables are not necessarily independent.

We define the *moment generating function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ as

$$m(a_1, \ldots, a_n) = \mathbb{E}\left\{e^{a_1 X_1 + \cdots + a_n X_n}\right\}. \tag{38}$$

[3] Note that the expectation $\mathbb{E}\{\cdot\}$ is a linear operator from a space of functions, e.g., the space of real-valued functions that are measurable with respect $p(x_1, \ldots, x_n)$. Also, we do not need to assume the existence of the PDF to define the expectation operator. In fact, a more general expression for (31) is

$$\mathbb{E}\left\{g(X_1, \ldots, X_n)\right\} = \int_{\Omega} g(X_1(\omega), \ldots, X_n(\omega)) dP(\omega). \tag{30}$$

[4] Note that (34) follows from (32) using the marginalization property of the PDF. For instance

$$\begin{aligned}
\mathbb{E}\left\{X_1 X_2\right\} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 x_2 p(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_n) dx_3 \cdots dx_n \right) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2) dx_1 dx_2.
\end{aligned} \tag{33}$$

.

It is straightforward to show that

$$\frac{\partial m(0,\dots,0)}{\partial a_i} =\mathbb{E}\{X_i\}, \tag{39}$$

$$\frac{\partial^2 m(0,\dots,0)}{\partial a_j \partial a_i} =\mathbb{E}\{X_i X_j\} \tag{40}$$

$$\frac{\partial^3 m(0,\dots,0)}{\partial a_j \partial a_i \partial a_k} =\mathbb{E}\{X_i X_j X_k\},$$

$$\dots \tag{41}$$

Hence, the partial derivatives of the moment generating function evaluated at zero represent the joint moments of the components of random vector $\boldsymbol{X}$. Clearly, if $m(a_1,\dots,a_n)$ admits a convergent power series expansion at 0 then all joint moments exist.

The *cumulant generating function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega),\dots,X_n(\omega))$ is defined as

$$\Psi(a_1,\dots,a_n) = \log(m((a_1,\dots,a_n))). \tag{42}$$

It is straightfoward to show that

$$\frac{\partial \Psi(0,\dots,0)}{\partial a_i} =\mathbb{E}\{X_i\}, \tag{43}$$

$$\frac{\partial^2 \Psi(0,\dots,0)}{\partial a_j \partial a_i} =\mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j\}, \tag{44}$$

$$\frac{\partial^3 \Psi(0,\dots,0)}{\partial a_j \partial a_i \partial a_k} =\mathbb{E}\{X_i X_j X_k\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j X_k\} - \mathbb{E}\{X_j\}\mathbb{E}\{X_i X_k\} - \mathbb{E}\{X_k\}\mathbb{E}\{X_i X_j\}$$
$$+ 2\mathbb{E}\{X_i\}\mathbb{E}\{X_j\}\mathbb{E}\{X_k\},$$
$$\dots$$

The quantities at the right hand side are known as *joint cumulants* of the random variables $(X_1,\dots,X_n)$. The cumulants are often denoted as $\langle X_i X_j \cdots \rangle_c$ (see, e.g., [11])

$$\begin{aligned} \langle X_i \cdots \rangle_c &=\mathbb{E}\{X_i\}, \\ \langle X_i X_j \cdots \rangle_c &=\mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j\}, \\ &\dots. \end{aligned} \tag{45}$$

Characteristic function. The *characteristic function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega),\dots,X_n(\omega))$ is defined as

$$\phi(a_1,\dots,a_n) = \mathbb{E}\left\{ e^{i(a_1 X_1 + \dots + a_n X_n)} \right\}. \tag{46}$$

Note that the characteristic function is the *Fourier transform*[5] of the joint probability density function $p(x_1,\dots,x_n)$ and therefore it essentially carries the same information. The joint moments of $\boldsymbol{X}$ can be computed as

$$\mathbb{E}\left\{ X_1^{k_1} \cdots X_n^{k_n} \right\} = \frac{1}{i^{k_1+\dots+k_n}} \frac{\partial^{k_1+\dots+k_n} \phi(0,\dots,0)}{\partial^{k_1} a_1 \cdots \partial^{k_n} a_n}. \tag{47}$$

It is interesting to notice that the marginalization operation we have seen for the PDF, e.g.,

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2, \dots, x_n) dx_2 \cdots dx_n \tag{48}$$

[5]The Fourier transform (46) is taken in the appropriate function space, e.g., $L^2(\mathbb{R}^n)$ or in the space of tempered distributions $\mathcal{S}(\mathbb{R}^n)$.

turns out to be simplified substantially in Fourier space. Indeed

$$\phi(a_1) = \phi(a_1, 0, \ldots, 0) = \mathbb{E}\left\{e^{ia_1X_1 + i0X_2 \cdots + i0X_n}\right\}. \tag{49}$$

By using well-known series expansion of the complex exponential, it is possible to show that (see, e.g., [11])

$$\phi(a_1, a_2, \ldots, a_n) = \exp\left[\sum_{\nu_1,\ldots,\nu_n=0}^{\infty} \langle X_1^{\nu_1} \cdots X_n^{\nu_n}\rangle_c \prod_{k=1}^{n} \frac{(ia_k)^{\nu_k}}{\nu_k!}\right], \tag{50}$$

where the series at the exponent excludes the case $\nu_1 = \cdots = \nu_n = 0$. For example,

$$\phi(a_1, a_2) = \phi(a_1)\phi(a_2) \exp\left[\sum_{j,k=1}^{\infty} \left\langle X_1^j X_2^k \right\rangle_c \frac{(ia_1)^j (ia_2)^k}{j!k!}\right]. \tag{51}$$

Clearly, if X_1 and X_2 are independent we have $\left\langle X_1^j X_2^k \right\rangle_c = 0$ for all i and j and therefore (51) reduces to

$$\phi(a_1, a_2) = \phi(a_1)\phi(a_2). \tag{52}$$

Clearly, this equation is the Fourier transform of the PDF $p(x_1, x_2) = p(x_1)p(x_2)$, and shows that if X_1 and X_2 are independent both the joint PDF and the joint characteristic function can be factorized as a product of one-dimensional functions.

Joint PDF of m functions of n random variables. Let $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ be a random vector with joint probability density function $p(x_1, \ldots, x_n)$. Define

$$\begin{cases} Y_1 = g_1(X_1, \ldots, X_n) \\ \quad \vdots \\ Y_m = g_m(X_1, \ldots, X_n) \end{cases} \tag{53}$$

What is the joint probability density function of the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$? Note that m can be smaller, equal or larger than n. These cases need to be handled differently.

- If $n = m$ and $\{g_1, \ldots, g_m\}$ are distinct functions we proceed as in Theorem 1 below.
- If $m < n$ and $\{g_1, \ldots, g_m\}$ are distinct functions we can add $m - n$ equations to complement the system so that we have n independent equations in n variables:

$$\begin{cases} Y_1 = g_1(X_1, \ldots, X_n) \\ \quad \vdots \\ Y_m = g_m(X_1, \ldots, X_n) \\ Y_{m+1} = X_{m+1} \\ \quad \vdots \\ Y_n = X_n \end{cases} \tag{54}$$

Once the joint PDF of $Y_1, \ldots, Y_n$ is known (using Theorem 1 below) then we can marginalize it with respect to $(y_{m+1}, \ldots, y_n)$ to obtain $p(y_1, \ldots, y_m)$ as

$$p(y_1, \ldots, y_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(y_1, \ldots, y_m, y_{m+1}, \ldots y_n) dy_{m+1} \cdots dy_n. \tag{55}$$

- If we have more equations than variables (i.e. $m > n$) then the computation of the joint PDF of $(Y_1, \ldots, Y_m)$ is not as straightforward as above. Consider for example the mapping $Y_1(\omega) = X(\omega)$ and $Y_2(\omega) = X^2(\omega)$. Here we have two functions of the same random variable. Note also that $Y_2 = Y_1^2$. It can be shown that the joint PDF of $Y_1 = X$ and $Y_2 = X^2$ is

$$p(y_1, y_2) = p_X(y_1)\delta(y_2 - y_1^2), \tag{56}$$

where p_X is the PDF of X and $\delta(\cdot)$ is the Dirac delta function.

Theorem 1. Let $\boldsymbol{x}_k(\boldsymbol{y})$ $(k = 1, \ldots, r)$ be the zeros of the nonlinear system of equations $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ defined in (53) (for $n = m$) or in (54) (for $m < n$). The joint PDF of $Y_1, \ldots, Y_n$ is given by

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{i=1}^{r} \frac{p_{\boldsymbol{X}}(\boldsymbol{x}_i(\boldsymbol{y}))}{|J(\boldsymbol{x}_i(\boldsymbol{y}))|}, \tag{57}$$

where J is the Jacobian determinant[6] associated with the mapping $\boldsymbol{g}(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_i(\boldsymbol{y})$ (assumed non-zero).

The proof of this theorem is provided in [13, Chapter 8].

Example 1: Consider the mapping

$$Y_1 = X_1^2, \qquad Y_2 = X_1 + X_2. \tag{59}$$

Suppose we know the joint PDF of X_1 and X_2. What's the joint PDF of Y_1 and Y_2? The following mapping from (X_1, X_2) to (Y_1, Y_2) can be inverted as

$$\begin{cases} y_1 = x_1^2 \\ y_2 = x_1 + x_2 \end{cases} \quad \Rightarrow \quad \begin{cases} x_1 = \pm\sqrt{y_1} \\ x_2 = y_2 \mp \sqrt{y_1} \end{cases} . \tag{60}$$

The Jacobian determinant of (60) is easily obtained as

$$J(x_1, x_2) = \det \begin{bmatrix} 2x_1 & 0 \\ 1 & 1 \end{bmatrix} = 2x_1. \tag{61}$$

Hence, by applying Theorem 1, we obtain the following joint PDF of Y_1 and Y_2 is

$$p_{\boldsymbol{Y}}(y_1, y_2) = \frac{1}{2\sqrt{y}} \left[p_{\boldsymbol{X}}(\sqrt{y_1}, y_2 - \sqrt{y_1}) + p_{\boldsymbol{X}}(-\sqrt{y_1}, y_2 + \sqrt{y_1}) \right] \qquad y_1 \geq 0. \tag{62}$$

Example 2: Consider the mapping

$$Y_1(\omega) = X_1 \qquad Y_2(\omega) = 2\sin\left(2X_1(\omega) + X_2(\omega)\right), \tag{63}$$

where X_1 and X_2 and independent Gaussians with zero mean and variance one. In Figure 1 we estimate the joint PDF of Y_1 and Y_2 using the frequency approach approach, i.e., formula (11), and the 2D kernel density estimation method discussed in [2].

[6] In (57) it is assumed that

$$J(\boldsymbol{x}_i(\boldsymbol{y})) = \det \left[\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right]_{\boldsymbol{x} = \boldsymbol{x}_i(\boldsymbol{y})} \neq 0. \tag{58}$$

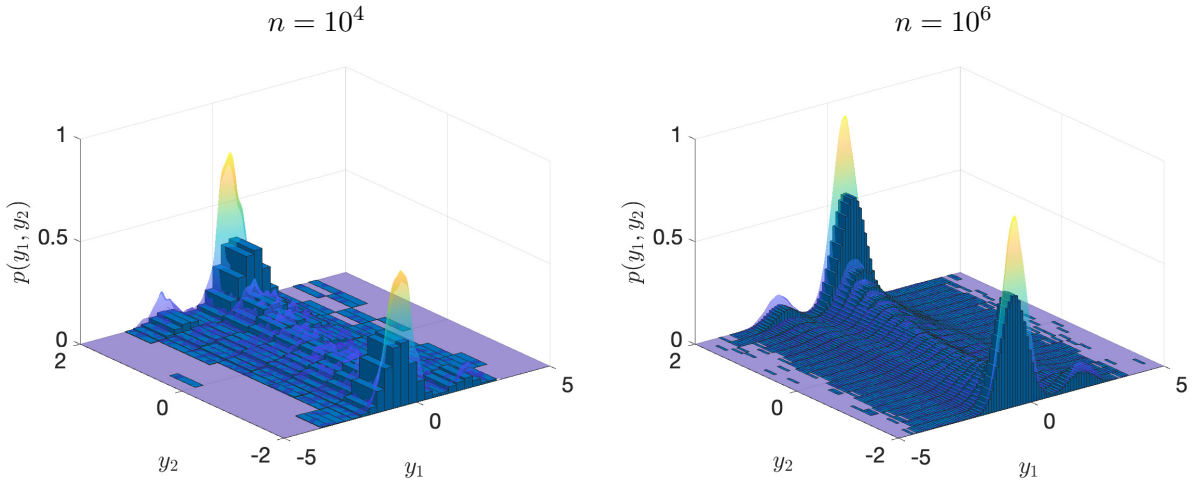

Figure 1: Estimation of the joint PDF of the random variables $Y_1 = X_1$ and $Y_2 = 2\sin(2X_1 + X_2)$ where X_1 and X_2 and independent Gaussians with zero mean and variance one. We show the results we obtain with the frequency approach, i.e., formula (11) and the 2D kernel density estimation method discussed in [2] (transparent surface plot). We plot results for a different number of samples n.

Other methods to compute the joint PDF of functions of random vectors. There are of course other methods to compute the joint PDF $(Y_1, \ldots, Y_m)$, given the joint PDF $(X_1, \ldots, X_n)$. For instance, methods based on the Dirac delta function [8], or methods based on the joint characteristic function. With reference to the previous example we have the joint characteristic function

$$\phi_{\boldsymbol{Y}}(a_1, a_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{ia_1 x_1^2 + ia_2(x_1+x_2)} p(x_1, x_2) dx_1 dx_2. \tag{64}$$

Clearly, if $\phi_{\boldsymbol{Y}}(a_1, a_2)$ can be computed then we can simply inverse Fourier transform it to obtain the joint PDF of (Y_1, Y_2). By using Dirac delta functions we can represent directly the joint PDF of the random variable

$$Y(\omega) = g(X_1(\omega), \ldots, X_n(\omega)), \tag{65}$$

as

$$p(y) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(y - g(x_1, \ldots, x_n)) p(x_1, \ldots, x_n) dx_1 \cdots dx_n \tag{66}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{ia(y - g(x_1, \ldots, x_n))} p(x_1, \ldots, x_n) dx_1 \cdots dx_n da. \tag{67}$$

Example 3: Let $Y_1 = X$ and $Y_2 = X^2$ (two functions of one random variable). What is the joint PDF of Y_1 and Y_2? The mapping (53) yields a Jacobian determinant that is zero, and therefore the mapping it is not invertible. This implies that theorem (1) cannot be applied. However, using the characteristic function approach we obtain

$$\phi(a_1, a_2) = \int_{-\infty}^{\infty} e^{ia_1 x + ia_2 x^2} p_X(x) dx. \tag{68}$$

Taking the inverse Fourier transform yields,

$$\begin{aligned} p(y_1,y_2) =& \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{ia_1(x-y_1)+ia_2(x^2-y_2)}p_X(x)dxda_1da_2 \\ =& \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta(x-y_1)e^{ia_1(x^2-y_2)}p_X(x)dxda_2 \\ =& \delta(y_1^2-y_2)p_X(y_1). \end{aligned} \tag{69}$$

Example 4: If $\boldsymbol{X}$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\boldsymbol{A}$ is invertible then $\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}$ is Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$. To show this, we notice that we have the unique solution $\boldsymbol{X}=\boldsymbol{A}^{-1}\boldsymbol{Y}$. Hence by applying Theorem 1 we obtain the following PDF of $\boldsymbol{Y}$

$$\begin{aligned} p_{\boldsymbol{Y}}(\boldsymbol{y}) =& \frac{p_{\boldsymbol{X}}\left(\boldsymbol{A}^{-1}\boldsymbol{y}\right)}{|\det(\boldsymbol{A})|} \\ =& \frac{1}{\sqrt{2\pi)^n\det(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)}}\exp\left[-\frac{1}{2}\left(\boldsymbol{A}^{-1}\boldsymbol{y}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{A}^{-1}\boldsymbol{y}-\boldsymbol{\mu}\right)\right] \\ =& \frac{1}{\sqrt{2\pi)^n\det(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)}}\exp\left[-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\right)^T\left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\right)\right]. \end{aligned} \tag{70}$$

Remark: If $(X_1,\ldots,X_n)$ are independent random variables and $(g_1,\ldots,g_n)$ are n functions from $\mathbb{R}$ into $\mathbb{R}$, then $Y_1=g_1(X_1)$, ..., $Y_n=g_n(X_n)$ are independent random variables. It is straightforward to prove this statement using the Dirac delta function representation (or the characteristic function) of PDF mapping [8]. To this end, let

$$Y_i(\omega)=g_i(X_i(\omega)). \tag{71}$$

We have

$$\begin{aligned} p(y_1,\ldots,y_n) =& \int_{-\infty}^{\infty}\prod_{j=1}^{n}\delta(y_j-g_j(x_j))p(x_1,\ldots,x_n)dx_1\cdots dx_n \\ =& \prod_{j=1}^{n}\int_{-\infty}^{\infty}\delta(y_j-g_j(x_j))p(x_j)dx_j \\ =& p(y_1)\cdots p(y_n). \end{aligned} \tag{72}$$

Sum of independent random variables. The PDF of the sum of independent random variables is the *convolution* the PDF of each variable. For example, let

$$Y=X_1+X_2+X_3 \tag{73}$$

be the sum of three independent random variables X_1, X_2 and X_3, with PDFs $p_1(x_1)$, $p_2(x_2)$ and $p_3(x_3)$ respectively. By using (66) we obtain

$$\begin{aligned} p(y) =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(y-x_1-x_2-x_3)p(x_1,x_2,x_3)dx_1dx_2dx_3 \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(x_1-y+x_2+x_3)p_1(x_1)p_2(x_2)p_3(x_3)dx_1dx_2dx_3 \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_1(x_2+x_3-y)p_2(x_2)p_3(x_3)dx_2dx_3 \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_1(x_1-y)p_2(x_1-x_3)p_3(x_3)dx_1dx_3. \end{aligned} \tag{74}$$

In the last equality we considered the mapping $x_1 = x_2 + x_3$ as a coordinate change from x_1 to x_2 with parameter x_3. Note that the process of computing the PDF of the sum of independent random variables can be also seen as a hierarchical process in which we proceed with two variables at a time To this end, we first compute the PDF of $Z = X_2 + X_3$ as

$$p_Z(z) = \int_{-\infty}^{\infty} p_2(z - x_3)p_3(x_3)dx_3. \tag{75}$$

Clearly, Z is independent of X_1 and therefore the PDF of $Y = Z + X_1$ is

$$p_Y(y) = \int_{-\infty}^{\infty} p_1(y - x_1)p_Z(x_1)dx_1. \tag{76}$$

A substitution of (75) into (76) yields (74). A more direct proof relies on the characteristic function, which is the Fourier transform of (74). In fact

$$\begin{aligned}\phi_Y(a) =& \mathbb{E}\left\{e^{ia(X_1+X_2+X_3)}\right\}\\ =& \mathbb{E}\left\{e^{iaX_1}e^{iaX_2}e^{iaX_3}\right\}\\ =& \mathbb{E}\left\{e^{iaX_1}\right\}\mathbb{E}\left\{e^{iaX_2}\right\}\mathbb{E}\left\{e^{iaX_3}\right\}.\end{aligned} \tag{77}$$

In the last step we used the fact that the expectation of a product of independent variables is the product of expectation. Equation (77) can be written as

$$\phi_Y(a) = \phi_{X_1}(a)\phi_{X_2}(a)\phi_{X_3}(a) \tag{78}$$

The inverse Fourier transform of (78) yields (74).

Example (sample average): Consider an experiment where we sample N independent realizations of a random variable $X(\omega)$ (like rolling a dice N times) and then take an average of all outcomes. Denote by $X_j(\omega)$ the outcome of the random variable $X(\omega)$ at the j-th sampling step and define

$$\overline{X}_N(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots X_N(\omega)}{N} \qquad \text{(sample average)}. \tag{79}$$

In this equation $X_j(\omega)$ are i.i.d. random variables with the same distribution as $X(\omega)$. Clearly if we repeat the sampling experiment multiple times, we have that $\overline{X}_N(\omega)$ attains different values (the samples of $\{X_1, \ldots, X_N\}$ are different from experiment to experiment). As we will see when we study Monte Carlo sampling methods, $\overline{X}_N(\omega)$ is an approximation of $\mathbb{E}\{X\}$, i.e.,

$$\mathbb{E}\{X\} = \int_{-\infty}^{\infty} xp_X(x)dx \simeq \overline{X}_N(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots X_N(\omega)}{N}. \tag{80}$$

By using (77) and (78) it is straightforward to compute the characteristic function of $\overline{X}_N(\omega)$, and therefore its PDF. Specifically, we obtain[7]

$$\phi_{\overline{X}_N}(a) = \mathbb{E}\left\{e^{iaX_j/N}\right\}^N \qquad \text{for any fixed } j. \tag{81}$$

Let us provide two illustrative examples.

[7] Recall that the variables X_j are i.i.d. and therefore they have the same PDF and same characteristic function.

- *Sample average of a Gaussian random variable*: Consider a Gaussian random variable $X(\omega)$ with PDF

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{82}$$

The sample average (79) in this case is a combination of i.i.d. variables X_j, all of which have PDFs (82). This implies that X_j/N are Gaussians with variance $1/N^2$, i.e.,

$$p_{X_j/N}(x) = \frac{N}{\sqrt{2\pi}} e^{-x^2N^2/2}. \tag{83}$$

The characteristic function of X_j/N, i.e., $\mathbb{E}\left\{e^{iaX_j/N}\right\}$ is obtained by taking the Fourier transform of (83). This yields

$$\mathbb{E}\left\{e^{iaX_j/N}\right\} = e^{-a^2/(2N^2)} \tag{84}$$

Taking the product in (81) we obtain

$$\phi_{\overline{X}_N}(a) = e^{-a^2/(2N)} \qquad \Leftrightarrow \qquad p_{\overline{X}_N}(x) = \sqrt{\frac{N}{2\pi}} e^{-x^2/(2N)}. \tag{85}$$

Therefore, the sample average $\overline{X}_N(\omega)$ is distributed as a Gaussian random variable with variance $1/N$. This means that as we sum up more and more terms in (79) we have a concentration of measure phenomenon such that the Gaussian 85 gets more and more concentrated nearby 0. Moreover,

$$\left|\overline{X}_N(\omega) - \mathbb{E}\{X\}\right| = O\left(\frac{1}{\sqrt{N}}\right) \tag{86}$$

i.e., the sample mean $\overline{X}_N(\omega)$ converges[8] to the mean of $X(\omega)$ at a rate $1/\sqrt{N}$.

- *Sample average of a Cauchy random variable*: Consider a Cauchy random variable $X(\omega)$ with PDF

$$p_X(x) = \frac{1}{\pi\left(1+x^2\right)} \tag{87}$$

We have seen that all moments of X (including the mean) are undefined, e.g.,

$$\int_{-\infty}^{\infty} x p_X(x) dx = \infty - \infty. \tag{88}$$

Does that mean that if we sample the Cauchy variable X and compute the sample mean (79) we may not converge to anything even for a very large number of samples N? Yes! To show this, recall that the PDF of X_j/N in this case is

$$p_{X_j/N}(x) = \frac{N}{\pi\left(1+N^2x^2\right)}. \tag{89}$$

The Fourier transform of this PDF is

$$\phi_{X_j/N}(a) = \mathbb{E}\left\{e^{iaX_j/N}\right\} = e^{|a|/N}. \tag{90}$$

Therefore the characteristic function and PDF of the sample average (79) is

$$\phi_{\overline{X}_N}(a) = e^{|a|} \qquad \Leftrightarrow \qquad p_{\overline{X}_N}(x) = \frac{1}{\pi\left(1+x^2\right)}, \tag{91}$$

independently of the number of samples N! Stated in different terms: no matter how many samples we consider, we have that the sample average of a Cauchy random variable is always a Cauchy random variable.

[8] As we will see, there are different modes of convergence of sequences of random variables, e.g., converge in probability, convergence in distribution, mean-square convergence, etc. In this case we have mean square convergence, which implies convergence in probability (thanks to the Markov inequality), and convergence in distribution.

Mapping random vectors to random vectors with desired distributions. It is always possible to flow a multivariate PDF $p_0(\boldsymbol{x})$ into another multivariate (target) PDF $p_1(\boldsymbol{x})$ using a time-dependent vector field, i.e., a system of ODEs of the form

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{f}(\boldsymbol{x},t) \qquad \boldsymbol{x}(0) = \boldsymbol{X}_0(\omega). \tag{92}$$

Indeed, as shown in Appendix A, the PDF of the state vector of the dynamical system (92) satisfies the Lioville equation

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} + \nabla \cdot [\boldsymbol{f}(\boldsymbol{x},t)p(\boldsymbol{x},t)] = 0. \tag{93}$$

Of course, if we set $p(\boldsymbol{x},0) = p_0(\boldsymbol{x})$ and $p(\boldsymbol{x},1) = p_1(\boldsymbol{x})$ we can design (via optimization) a vector $\boldsymbol{f}(\boldsymbol{x},t)$ that transports $p_0(\boldsymbol{x})$ to $p_1(\boldsymbol{x})$. This concept has been recently leveraged to develop samplers for arbitrary PDFs $p_1(\boldsymbol{x})$. The idea is to sample a known PDF $p_0(\boldsymbol{x})$, i.e., sample $\boldsymbol{X}_0(\omega)$ and then flow such samples to samples of $\boldsymbol{p}_1(\boldsymbol{x})$ using (92) and an appropriate (optimized) vector field $\boldsymbol{f}(\boldsymbol{x},t)$. Techniques such as continuous normalizing flows and *flow-matching* approaches [12, 6] are notable examples of methods in this category. Note that the target distribution can also be the distribution of a vector with independent components. In this sense, we can flow an arbitrary PDF into a fully separated one, hence transforming the components of an arbitrary random vector into independent random variables. In addition, the vector field $\boldsymbol{f}(\boldsymbol{x},t)$ can be chosen to minimize a Wasserstein metric or other metrics (e.g. the KL divergence), in which case we talk about optimal mass transport [10].

Rosenblatt transformation. Another transformation that allows us to map a PDF $p_0(\boldsymbol{x})$ into another (arbitrary) PDF $p_1(\boldsymbol{x})$ is the *Rosenblatt transformation* [17]. Essentially, one can show that there exists a unique monotone increasing transformation of the form

$$\boldsymbol{T}(\boldsymbol{x}) = (T_1(x_1), T_2(x_1,x_2), \ldots, T_n(x_1,x_2,\ldots,x_n)) \tag{94}$$

such that

$$p_0(\boldsymbol{T}(\boldsymbol{x}))\det(\nabla \boldsymbol{T}(\boldsymbol{x})) = p_1(\boldsymbol{x}). \tag{95}$$

In this setting, the samples of $p_1(\boldsymbol{x})$ are obtained by simply mapping the samples of $p_0(\boldsymbol{x})$ via the transformation $\boldsymbol{T}(\boldsymbol{x})$ (when available). For the computation of $\boldsymbol{T}(\boldsymbol{x})$ see [17] and the references therein.

Mapping correlated Gaussian vectors to independent ones. To transform a correlated Gaussian vector into an independent one we just need a linear map. Essentially, if $\boldsymbol{X}$ is Gaussian with zero mean and identity covariance then

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu} \tag{96}$$

has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. This means that if we are interested in sampling a correlated Gaussian we can simply decompose the covariance $\boldsymbol{\Sigma}$ using the Cholesky factorization, sample $\boldsymbol{X}$ and map the samples as in (96), where $\boldsymbol{A}$ is the lower-triangular factor of the Cholesky decomposition. Similarly, if we to "de-correlate" a correlated Gaussian vector $\boldsymbol{Y}$ we can apply invert the Cholesky factor of the covariance matrix and construct the vector

$$\boldsymbol{X} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}). \tag{97}$$

Lebesgue spaces of random variables. The expectation operator $\mathbb{E}\{\cdot\}$ is a linear integral operator over a probability measure. Such an operator can be used to define norms and inner products in spaces of random variables. For example,

$$\mathbb{E}\left\{|X|^q\right\} = \int_\Omega |X(\omega)|^q dP(\omega) \qquad q \in \mathbb{N} \tag{98}$$

is essentially a weighted q norm. The space of random variables satisfying $\mathbb{E}\{|X|^q\} < \infty$ is denoted as $L^q(\Omega, \mathcal{B}, P)$, in analogy with the classical Lebesgue space for functions. The case $q = 2$ is of particular importance as it has the structure of a Hilbert space. Specifically, for any two random variables in $L^2(\Omega, \mathcal{B}, P)$ we have the inner product

$$\mathbb{E}\{XY\} = \int_\Omega X(\omega)Y(\omega)dP(\omega) \tag{99}$$

and the norm

$$\mathbb{E}\left\{X^2\right\} = \int_\Omega X(\omega)^2 dP(\omega). \tag{100}$$

The inner product (99) allows us to define orthogonal random variables. Specifically, $X(\omega)$ and $Y(\omega)$ are orthogonal in $L^2(\Omega, \mathcal{B}, P)$ if they are uncorrelated, i.e., $\mathbb{E}\{XY\} = 0$. Also, $X(\omega)$ and $Y(\omega)$ are orthonormal if they are orthogonal and have norm equal to one, i.e., $\mathbb{E}\left\{X^2\right\} = \mathbb{E}\left\{Y^2\right\} = 1$.

Application to dynamical systems

Consider the following linear dynamical system

$$\begin{cases} \dot{x}(t) + \xi(\omega)x(t) = 0 \\ x(0) = x_0(\omega) \end{cases} \tag{101}$$

where $\xi(\omega)$ and $x_0(\omega)$ are independent random variables. Specifically $\xi(\omega)$ is uniformly distributed in $[0, 1]$, while $x_0(\omega)$ is Gaussian random variable with mean zero and variance one. As is well-known, the analytical solution of (101) is

$$x(t;\omega) = x_0(\omega)e^{-t\xi(\omega)}. \tag{102}$$

Let us compute the mean, the second-order moment and the auto-correlation function of the solution $x(t;\omega)$, i.e., $\mathbb{E}\{x(t;\omega)\}$, $\mathbb{E}\{x(t;\omega)^2\}$, and $\mathbb{E}\{x(t;\omega)x(t'\omega)\}$ versus time. We have

$$\mathbb{E}\{x(t;\omega)\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0 e^{-x_0^2/2}dx_0 \int_0^1 e^{-t\xi}d\xi = 0, \tag{103}$$

$$\mathbb{E}\left\{x(t;\omega)^2\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0^2 e^{-x_0^2/2}dx_0 \int_0^1 e^{-2t\xi}d\xi = \frac{1}{2t}\left(1 - e^{-2t}\right), \tag{104}$$

$$\mathbb{E}\left\{x(t;\omega)x(t';\omega)\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0^2 e^{-x_0^2/2}dx_0 \int_0^1 e^{-(t+t')\xi}d\xi = \frac{1}{t+t'}\left(1 - e^{-(t+t')}\right). \tag{105}$$

The one-time probability density function of $x(t;\omega)$ can be easily computed by using the Dirac delta function approach [8]. Indeed,

$$\begin{aligned} p(x,t) =& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_0^1 \delta\left(x - x_0 e^{-\xi t}\right) e^{-x_0^2/2}dx_0 d\xi \\ =& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_0^1 \frac{\delta\left(x_0 - xe^{\xi t}\right)}{e^{-\xi t}} e^{-x_0^2/2}dx_0 d\xi \qquad (106) \\ =& \frac{1}{\sqrt{2\pi}}\int_0^1 e^{\xi t - (xe^{\xi t})^2/2}d\xi. \qquad (107) \end{aligned}$$

Now consider the change of variables from ξ to u defined as

$$u = \frac{xe^{\xi t}}{\sqrt{2}} \quad \Rightarrow \quad d\xi = \frac{\sqrt{2}}{xt}e^{-\xi t}du. \tag{108}$$

A substitution of (108) into (107) yields

$$p(x,t) = \frac{1}{xt\sqrt{\pi}} \int_{x/\sqrt{2}}^{xe^t/\sqrt{2}} e^{-u^2} du \tag{109}$$

$$= \frac{1}{xt\sqrt{\pi}} \left[\text{erf}\left(\frac{xe^t}{\sqrt{2}}\right) - \text{erf}\left(\frac{x}{\sqrt{2}}\right) \right]. \tag{110}$$

Liouville equation approach: We can transform the linear system (101) involving one random variable at the right hand side to an equivalent 2D linear system evolving from a random initial state (an no random variables at the right hand side). To this end, we notice that

$$\begin{cases} \dot{x}(t) + yx(t) = 0 \\ \dot{y}(t) = 0 \\ x(0) = x_0(\omega) \\ y(0) = \xi(\omega) \end{cases} \tag{111}$$

is completely equivalent to (101). In this setting, we can derive a linear transport equation for the joint PDF of $x(t;\omega)$ and $y(t;\omega)$, i.e., $x(t;\omega)$ and $\xi(\omega)$. Such PDF equation takes the form

$$\begin{cases} \dfrac{\partial p(x,y,t)}{\partial t} = \dfrac{\partial}{\partial x}\left(xyp(x,y,t)\right) + \dfrac{\partial}{\partial y}\left(xyp(x,y,t)\right) \\ p(x,y,0) = p_{x_0}(x) p_\xi(y) \end{cases} \tag{112}$$

It can be verified by a direct substitution that the solution the initial value problem (112) is

$$p(x,y,t) = \frac{1}{\sqrt{2\pi}} e^{yt - (xe^{yt})^2/2}. \qquad y \in [0,1], \qquad x \in \mathbb{R}. \tag{113}$$

Note that the joint PDF (113) was already obtained in equation (107), right before marginalizing with respect to ξ.

Conditional probability. Conditional probability is a measure of the probability of an event A occurring, given that another event B has already occurred. Such probability defined as[9]

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{114}$$

Note that the conditional probability is non-zero if A and B are intersecting. Also note that if B is a subset of A then $P(A|B) = 1$.

Clearly, if A and B are independent events then by equation (22) we have that $P(A \cap B) = P(A)P(B)$. This implies that if A and B are independent then $P(A|B) = P(A)$. In other words, B has no effect whatsoever on the probability of A occurring. Moreover, $P(A \cap B) \leq P(B)$ (also $P(A \cap B) \leq P(A)$) and therefore we always have that

$$P(A|B) \leq 1. \tag{115}$$

[9] An example of conditional probability could be the following:

- Event A: "Daniele's team scores a goal".
- Event B: "Daniele takes a shot".

The conditional probability $P(A|B)$, i.e., the probability that Daniele's team scores a goal, conditional to Daniele taking a shot equals the probability that Daniele takes a shot and scores a goal, divided by the probability that Daniele takes a shot.

In the context of random vectors with multiple components, we may be interested in determining the conditional probability of an event involving one component, given that an event involving another component has already occurred. This yields the concept of *conditional CDF* and *conditional PDF*. Let us first clarify these concepts for a random vector with only two components $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$. By using the definition of the cumulative distribution function (5) we define (see [13, Ch. 7])

$$F(x_1|x_2) = \frac{F(x_1,x_2)}{F(x_2)} \quad \Leftrightarrow \quad F(x_1,x_2) = F(x_1|x_2)F(x_2). \tag{116}$$

Conditioning on particular outcomes. The determination of the conditional density of $X_1(\omega)$ given $X_2(\omega) = x_2$, i.e., when $X_2(\omega)$ attains a specific deterministic value x_2, is of particular interest. This density cannot be derived directly from (114) or (116) because, as we know, the event $X_2(\omega) = x_2$ has zero probability (if X_2 is a continuous random variable). However, one can make sense of such conditional probability by taking a suitable limit. Specifically, consider

$$P(\underbrace{\{X_1(\omega) \leq x_1\}}_{A} \cap \underbrace{\{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\}}_{B}) = F(x_1, x_2 + \Delta x_2) - F(x_1, x_2) \tag{117}$$

and

$$P(\{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\}) = F(x_2 + \Delta x_2) - F(x_2). \tag{118}$$

In (117) it is understood that $F(x_1, x_2)$ is the joint distribution function of (X_1, X_2), while in (118) $F(x_2)$ denotes the distribution function of X_2 alone. Clearly, for small Δx_2

$$F(x_1, x_2 + \Delta x_2) - F(x_1, x_2) \simeq \Delta x_2 \int_{-\infty}^{x_1} p(y_1, x_2) dy_1, \tag{119}$$

and

$$F(x_2 + \Delta x_2) - F(x_2) \simeq p(x_2)\Delta x_2. \tag{120}$$

Using (117)-(120) we can calculate the conditional probability

$$\frac{P(\{X_1(\omega) \leq x_1\} \cap \{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\})}{P(\{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\})} \simeq \frac{\Delta x_2 \int_{-\infty}^{x_1} p(y_1, x_2) dy_1}{\Delta x_2 p(x_2)}. \tag{121}$$

By sending Δx_2 to zero gives

$$F(x_1|X_2 = x_2) = \frac{\int_{-\infty}^{x_1} p(y_1, x_2) dy_1}{p(x_2)} \qquad \text{(conditional CDF)}. \tag{122}$$

Finally, by differentiating the previous equation with respect to x_1 we obtain

$$p(x_1|X_2 = x_2) = \frac{p(x_1, x_2)}{p(x_2)} \qquad \text{(conditional PDF)}. \tag{123}$$

In summary, to compute the conditional PDF, $p(x_1|X_2 = x_2)$ we literally take a section of the joint $p(x_1, x_2)$ for some fixed value of x_2 and then rescale the function we obtain by the number $p(x_2)$, i.e., the one-dimensional PDF of $p(x)$ of $X_2(\omega)$ evaluated at $x = x_2$. This procedure is illustrated in Figure 2 for a PDF represented in terms of a point cloud. Equation (123) can be written as

$$p(x_1, x_2) = p(x_1|x_2)p(x_2) = p(x_2|x_1)p(x_1) \tag{124}$$

which yields the identities

$$p(x_2) = \int_{-\infty}^{\infty} p(x_2|x_1)p(x_1)dx_1, \qquad p(x_1) = \int_{-\infty}^{\infty} p(x_1|x_2)p(x_2)dx_2. \tag{125}$$

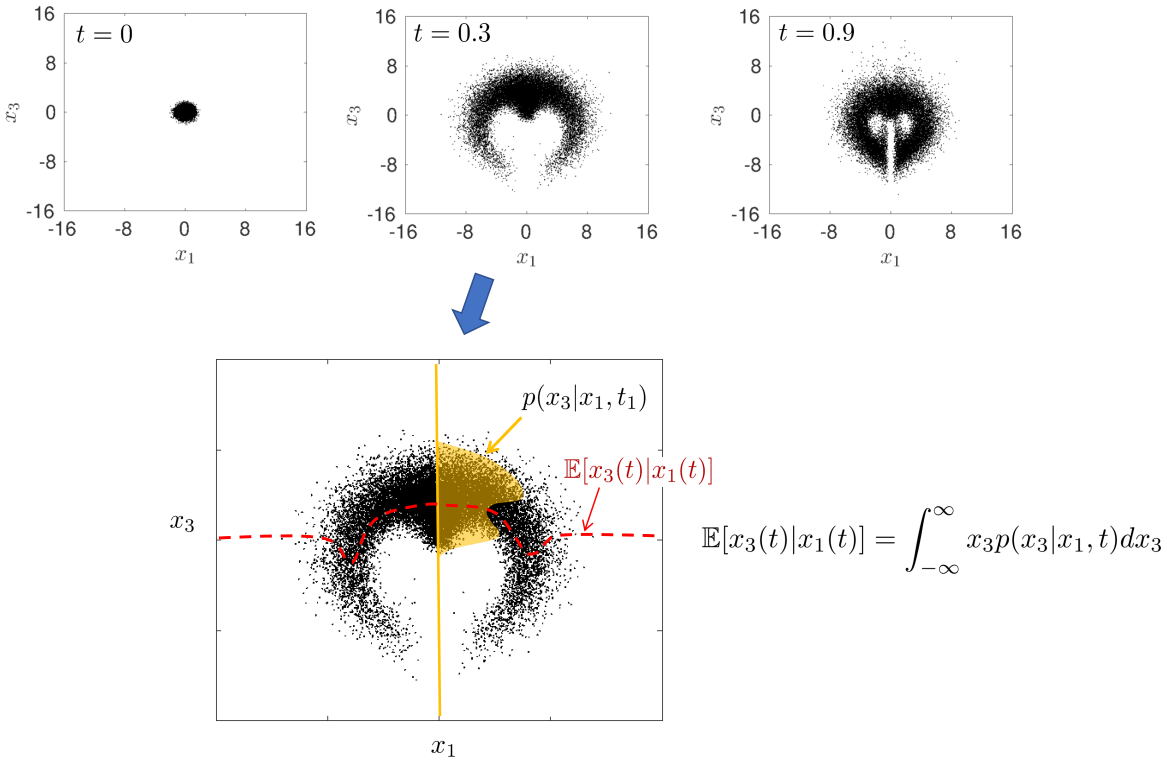

Figure 2: Point clouds representing the joint PDF of the phase variables $x_1(t)$ and $x_3(t)$ of the Kraichnan-Orzag system at different times, i.e., $p(x_3, x_1, t)$. Shown is the procedure to compute the conditional PDF $p(x_3|x_1, t)$ and the corresponding conditional mean $\mathbb{E}\{X_3|X_1 = x_1\}$.

The conditional probability density rule can be generalized to multiple random variables. For instance, if $p(x_1, x_2, x_3, x_4)$ denotes the joint PDF of four random variables then

$$p(x_1, x_2, x_3, x_4) = p(x_1|x_2, x_3, x_4)p(x_2, x_3, x_4) = p(x_1|x_2, x_3, x_4)p(x_2|x_3, x_4)p(x_3|x_4)p(x_4). \tag{126}$$

Moreover, conditional probability densities satisfy the *marginalization rule*. For instance

$$p(x_1, x_3|x_4, x_5) = \int_{-\infty}^{\infty} p(x_1, x_2, x_3|x_4, x_5) dx_2. \tag{127}$$

This property follows directly from the definition of conditional probability density (123).

Conditional expectation. Let $\boldsymbol{X}(\omega)$ and $\boldsymbol{Y}(\omega)$ be two random vectors defined on the probability space $(\Omega, \mathcal{B}, P)$. The *conditional mean* of $g(\boldsymbol{X}(\omega))$ (g is a measurable function) assuming $\boldsymbol{Y}(\omega) = \boldsymbol{y}$ is defined as[10]

$$\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}, \tag{128}$$

where

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})} \tag{129}$$

is the conditional PDF of $\boldsymbol{X}(\omega)$ given $\boldsymbol{Y}(\omega) = \boldsymbol{y}$. Note that the $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\}$ is a function of $\boldsymbol{y}$. The conditional mean defined in equation (128) allows us to write the conditional moments of a random

[10]The conditional mean in equation (128) is often written as $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y}\}$.

variable or a random vector, given information on another random vector. For example, the conditional mean and conditional correlation of $\boldsymbol{X}$ given $\boldsymbol{Y}(\omega)=\boldsymbol{y}$ are defined as

$$\mathbb{E}\{X_i|\boldsymbol{Y}=\boldsymbol{y}\}=\int_{-\infty}^{\infty}x_i p(x_i|\boldsymbol{y})dx_i, \tag{130}$$

$$\mathbb{E}\{X_iX_j|\boldsymbol{Y}=\boldsymbol{y}\}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_ix_j p(x_i,x_j|\boldsymbol{y})dx_idx_j. \tag{131}$$

The conditional mean of a system with two random variables is visualized in Figure 2. By combining (129), (128) and (31) we see that

$$\mathbb{E}\{g(\boldsymbol{X})\}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\}p(\boldsymbol{y})d\boldsymbol{y}. \tag{132}$$

In this sense, $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\}$ can be interpreted as a random variable, i.e., a scalar function of the random variable $\boldsymbol{Y}$ which, if averaged over $p(\boldsymbol{y})$, yields exactly $\mathbb{E}\{g(\boldsymbol{X})\}$.

Bayes' theorem. Let $p(\boldsymbol{x},\boldsymbol{y})$ the joint probability density of two random vectors. Using conditional probabilities we have

$$p(\boldsymbol{x},\boldsymbol{y})=p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})=p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \qquad \Leftrightarrow \qquad p(\boldsymbol{x}|\boldsymbol{y})=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}. \tag{133}$$

This can be written, e.g., as

$$p(\boldsymbol{x}|\boldsymbol{y})=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\displaystyle\int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}}. \tag{134}$$

Note that by using the marginalization rule we also have

$$p(\boldsymbol{x})=\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}. \tag{135}$$

Sampling high-dimensional PDFs using Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) refers to a class of methods that allow us to sample high-dimensional probability density functions [3]. In MCMC we construct a discrete Markov process that has a stationary PDF that coincides with the PDF of interest, i.e., the PDF we'd like to sample from. Hence, simulations of the Markov chain provide samples of the high-dimensional PDF we are interested in, once a transient called *burn-in phase* of the chain, is completed. To simulate the Markov chain it is common to use the Monte Carlo method, hence the name Markov Chain Monte Carlo. There are many different MCMC algorithms available to sample from high-dimensional PDFs. Perhaps, the simplest ones are

- Gibbs sampling (briefly described hereafter);
- Metropolis-Hastings algorithm (see [3]).

Gibbs sampling. Suppose we are given a three-dimensional PDF $p(x_1,x_2,x_3)$ and that the conditional PDFs $p(x_1|x_2,x_3)$, $p(x_2|x_1,x_3)$ and $p(x_3|x_1,x_2)$ are all available[11]. To sample from $p(x_1,x_2,x_3)$ we proceed as follows:

[11] Recall that to compute the conditional PDF $p(x_1|x_2,x_3)$ we literally set x_2 and x_3 in $p(x_1,x_2,x_3)$ equal to some number, say $x_2=x_2^*$ and $x_3=x_3^*$ and then normalize the one-dimensional function $p(x_1,x_2^*,x_3^*)$ so that the integral with respect to x_1 equals one.

1. Initialize $x_2 = x_2^{(i)}$ and $x_3 = x_3^{(i)}$. Here $x_2^{(i)}$ and $x_3^{(i)}$ are two real numbers. The superscript "i" is an integer number that labels the sample

$$\boldsymbol{X}^{(i)}(\omega) = \begin{bmatrix} x_1^{(i)}(\omega) & x_2^{(i)}(\omega) & x_3^{(i)}(\omega) \end{bmatrix} \qquad i \in \mathbb{N}. \tag{136}$$

2. Sample a new $x_1^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_1|x_2^{(i)}, x_3^{(i)}\right)$ (e.g., using the inverse CDF approach).

3. With the sample $x_1^{(i+1)}$ available, sample a new $x_2^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_2|x_1^{(i+1)}, x_3^{(i)}\right)$.

4. With the sample $x_2^{(i+1)}$ available, sample a new $x_3^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_3|x_1^{(i+1)}, x_2^{(i+1)}\right)$.

5. Update $x_j^{(i)} \leftarrow x_j^{(i+1)}$ for $j = 1, 2, 3$ and go back to point 2.

This algorithm allows us to compute $\boldsymbol{X}^{(i+1)}$ from $\boldsymbol{X}^{(i)}$ by sampling known one-dimensional conditional transition densities. To sample from such arbitrary one-dimensional transition densities we can use different methods. If the inverse cumulative distribution of each conditional PDF is known (or computable), then we are all set. In fact we can just sample a uniform PDF in $[0, 1]$ and then map such samples using the inverse cumulative distribution function. The mapping $\boldsymbol{X}^{(i)} \to \boldsymbol{X}^{(i+1)}$ defines a *random walk* in $\mathbb{R}^3$. It can be shown that the stationary distribution of such random walk coincides with $p(x_1, x_2, x_3)$. In other words, after the *burn-in phase* is completed, i.e., for sufficiently large i, we have that $\boldsymbol{X}^{(i)}(\omega)$ are samples of the joint PDF $p(x_1, x_2, x_3)$.

Example (Gibbs's sampling): Suppose we are interested in sampling the joint PDF

$$p(x_1, x_2) = K \sin^2(x_1 x_2), \qquad (x_1, x_2) \in [0, 1]^2 \tag{137}$$

where

$$K = \frac{4}{2 - \mathrm{Si}(2)} \qquad \mathrm{Si}(x) = \int_0^x \frac{\sin(t)}{t} dt \tag{138}$$

using Gibbs' sampling. To this end, we first compute the marginals

$$p(x_1) = \frac{K}{2}\left(1 - \frac{\sin(2x_1)}{2x_1}\right), \tag{139}$$

$$p(x_2) = \frac{K}{2}\left(1 - \frac{\sin(2x_2)}{2x_2}\right), \tag{140}$$

and the conditionals

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{4x_2 \sin^2(x_1 x_2)}{2x_2 - \sin(2x_2)}, \tag{141}$$

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p(x_1)} = \frac{4x_1 \sin^2(x_1 x_2)}{2x_1 - \sin(2x_1)}. \tag{142}$$

In Figure 3 we plot the PDF (137) and the samples we obtain from the Gibbs' algorithm. To sample from (141)-(142) we use the 1D inverse CDF algorithm with numerically computed CDFs.

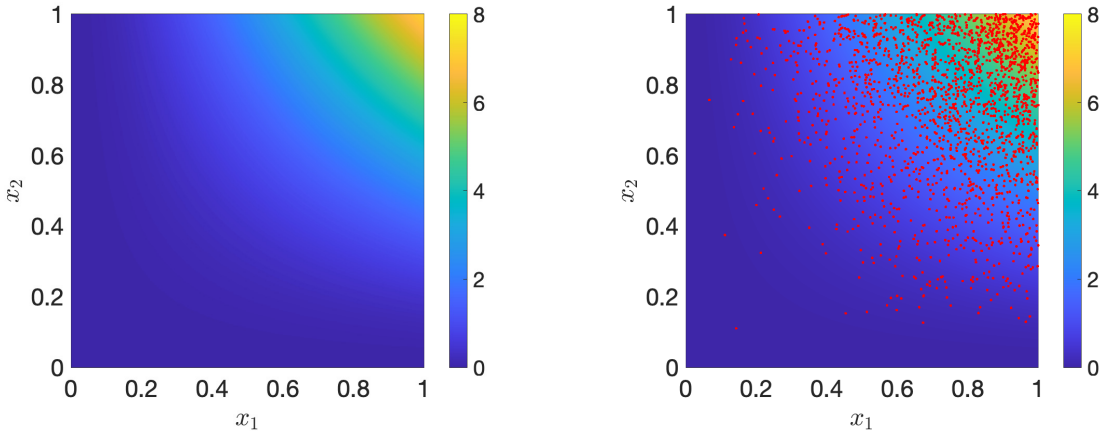

Figure 3: Plot of the 2D PDF (137) and samples of such PDF obtained by using Gibbs sampling.

Sampling approximate high-dimensional PDFs using copulas

A copula is a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform on the interval $[0,1]$. Copulas are used to model the dependence between random variables. Given a random vector

$$\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \tag{143}$$

we know can transform each component to a uniform random variable via the probability mappings

$$U_i(\omega) = F_i(X_i(\omega)) \qquad i = 1, \ldots, n. \tag{144}$$

A *copula* is defined as the cumulative distribution function of $(U_1, \ldots, U_n)$, i.e.,

$$F_{\boldsymbol{U}}(u_1, \ldots, u_n) = P\left\{\{\omega : U_1(\omega) \leq u_1\} \cap \cdots \cap \{\omega : U_n(\omega) \leq u_n\}\right\}. \tag{145}$$

The copula $F_{\boldsymbol{U}}$ is a CDF on the unit cube $[0,1]^n$, and it contains all information on the dependence structure between the components of the random vector $\boldsymbol{X}(\omega)$. Note that by using the definition (144)

$$\begin{aligned} F_{\boldsymbol{U}}(u_1, \ldots, u_n) =& P\left\{\{\omega : U_1(\omega) \leq u_1\} \cap \cdots \cap \{\omega : U_n(\omega) \leq u_n\}\right\} \\ =& P\left\{\{\omega : F_1(X_1(\omega)) \leq u_1\} \cap \cdots \cap \{\omega : F_n(X_n(\omega)) \leq u_n\}\right\} \\ =& P\left\{\{\omega : X_1(\omega) \leq F_1^{-1}(u_1)\} \cap \cdots \cap \{\omega : X_n(\omega) \leq F_n^{-1}(u_n)\}\right\} \\ =& F_{\boldsymbol{X}}\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right). \end{aligned} \tag{146}$$

Moreover, that equation (146) can be reversed to obtain

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = F_{\boldsymbol{U}}\left(F_1(x_1), \ldots, F_n(x_n)\right). \tag{147}$$

Hence, it appears that it is always possible write the joint CDF $F_{\boldsymbol{X}}(x_1, \ldots, x_n)$ of a random vector $\boldsymbol{X}$ in terms of a copula $F_{\boldsymbol{U}}(u_1, \ldots, u_n)$ and the individual CDFs F_i of each component $X_i(\omega)$. This result is known as *Sklar's theorem* (see, e.g., [5, Theorem 1.9]).

By differentiating (147) with respect to $(x_1, \ldots, x_n)$ yields the copula representation of the PDF of $\boldsymbol{X}$, i.e.,

$$p_{\boldsymbol{X}}(x_1, \ldots, x_n) = p_{\boldsymbol{U}}(F_1(x_1), \ldots, F_n(x_n)) p_1(x_1) \cdots p_n(x_n), \tag{148}$$

where $p_{\boldsymbol{U}}(u_1, \ldots, u_n)$ is the PDF of the copula.

Gaussian copula. The Gaussian copula is a distribution over the unit cuve $[0,1]^n$. It is constructed from a multivariate normal distribution by using simple one-dimensional probability mappings. For a given correlation coefficient matrix $\boldsymbol{R}$ (with entries in $[-1,1]$) the Gaussian copula is defined as

$$p_{\boldsymbol{U}}(u_1,\ldots,u_n) = \frac{1}{\sqrt{\det(\boldsymbol{R})}} \exp\left[\frac{1}{2}\boldsymbol{z_u}\left(\boldsymbol{I}-\boldsymbol{R}^{-1}\right)\boldsymbol{z_u}^T\right] \tag{149}$$

where

$$\boldsymbol{z_u} = \left(F_g^{-1}(u_1),\ldots,F_g^{-1}(u_n)\right) \tag{150}$$

and F_g denotes the Gaussian cumulative distribution function

$$F_g(u) = \frac{1}{2}\left[1+\text{erf}\left(\frac{u}{\sqrt{2}}\right)\right] \tag{151}$$

Sampling from Gaussian copulas. Suppose we are interested in sampling a random vector with given marginals and correlation function

$$C_{ij} = \mathbb{E}\left\{X_i X_j\right\} \tag{152}$$

In other words, we are not interested in sampling the full PDF of $\boldsymbol{X}$ but rather create an approximate model, i.e., a *surrogate*, that allows us to sample the vector given only its marginals and the correlation matrix $\boldsymbol{C}$.

Gaussian copulas can be used for that as they have sufficient degrees of freedom to enforce a correlation structure while being consistent with marginals. The procedure to sample a Gaussian copula is very simple:

1. We first sample realizations of a Gaussian vector with zero mean and unit variance, say $\boldsymbol{Z}^{(j)}$.
2. Such sample vectors are then transformed to a vector with correlation $\boldsymbol{R}$ (copula correlation) using the Cholesky factor[12] of $\boldsymbol{R}$. Call such samples $\boldsymbol{Y}^{(j)}$.
3. Each component of the rotated vectors is then mapped to a uniform distribution via the Gaussian probability mapping F_g (which is the same for all components), i.e., $\boldsymbol{U}^{(i)} = F_g(\boldsymbol{Y}^{(j)})$. The samples $\boldsymbol{U}^{(j)}$ are samples of the Gaussian copula.
4. At this point we generate a sample of the random vector of interest as $\boldsymbol{X}$ by using the marginal CDF mapping as

$$\boldsymbol{X}^{(i)} = \left(F_1^{-1}\left(U_1^{(i)}\right),\ldots,F_n^{-1}\left(U_n^{(i)}\right)\right). \tag{154}$$

Nataf transformation. It is important to emphasize that the correlation matrix $\boldsymbol{R}$ is not the correlation matrix $\boldsymbol{C}$ of the vector $\boldsymbol{X}$ we are interested in sampling, but rather the "copula correlation matrix". There is of course a relation between the two. Indeed, by using the the copula PDF we have

$$C_{ij} = \int_{[0,1]^2} F_i^{-1}(u_i)F_j^{-1}(u_j)p_{\boldsymbol{U}}(u_i,u_j;\rho_{ij})du_i du_j \tag{155}$$

where

$$p_{\boldsymbol{U}}(u_i,u_j;\rho_{ij}) = \frac{1}{\sqrt{1-\rho_{ij}^2}}\exp\left[\frac{1}{2(1-\rho_{ij}^2)}\left[2\rho_{ij}F_g^{-1}(u_i)F_g^{-1}(u_j) - \rho_{ij}^2\left(F_g^{-1}(u_j)^2 + F_g^{-1}(u_j)^2\right)\right]\right] \tag{156}$$

[12]The matrix $\boldsymbol{R}$ is symmetric and positive definite. Hence it admits the Cholesky factorization $\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^T$ where L is lower triangular with positive diagonal entries. If $\boldsymbol{Z}$ is a Gaussian vector with zero mean and unit covariance then $\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Z}$ is a Gaussian vector with zero mean and covariance $\boldsymbol{R}$. In fact

$$\mathbb{E}\left\{\boldsymbol{Y}\boldsymbol{Y}^T\right\} = \boldsymbol{L}\mathbb{E}\left\{\boldsymbol{Z}\boldsymbol{Z}^T\right\}\boldsymbol{L}^T = \boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{R}. \tag{153}$$

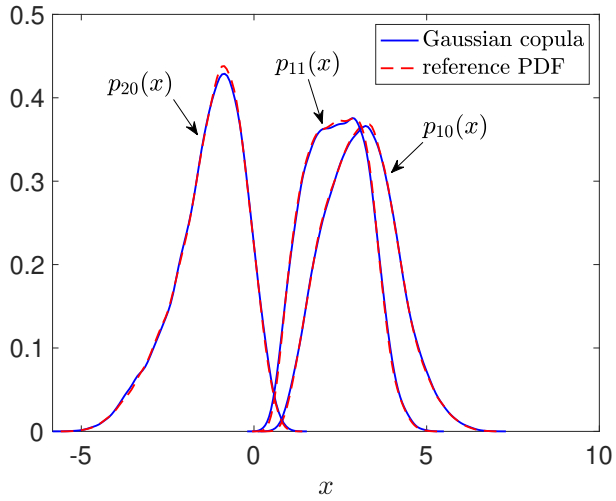

Figure 4: Marginals of the random function at points x_{10}, x_{11} and x_{20}. Comparison between reference marginals and marginals obtained from the Gaussian copula approximation. Note also that the marginals p_{10} and p_{11} overlap significantly due to the proximity of the spatial points x_{10} and x_{11} and the smoothness of the random function (159).

is the marginal of (149). Given a target correlation matrix C_{ij}, we can campute the copula correlation coefficients ρ_{ij} by solving the following sequence of 1D optimization problem

$$\min_{\rho_{ij}\in[-1,1]} \left| C_{ij} - \int_{[0,1]^2} F_i^{-1}(u_i) F_j^{-1}(u_j) p_{\boldsymbol{U}}(u_i, u_j; \rho_{ij}) du_i du_j \right| \qquad j > i. \tag{157}$$

Once all entries of $\boldsymbol{R}$, i.e., $\rho_{ij} \in [-1, 1]$ $(j > i)$ are computed in this way, we may end up with a symmetric matrix $\boldsymbol{R}$ that may not positive definite. The closest (in the Frobenious norm) positive definite matrix to $\boldsymbol{R}$ is easily obtained by doing an spectral decomposition and chopping of all negative eigenvalues.

$$\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \quad \Rightarrow \quad \boldsymbol{R}^+ = \boldsymbol{Q}\boldsymbol{\Lambda}^+\boldsymbol{Q}^T \quad \boldsymbol{\Lambda}^+ = \max\{\boldsymbol{\Lambda}, \boldsymbol{0}\}. \tag{158}$$

The transformation between correlated Gaussian copula variable (with correlation $\boldsymbol{R}$) and the vector $\boldsymbol{X}$ with given marginal distributions and correlation $\boldsymbol{C}$ is called *Nataf transformation*.

Example (Gaussian copulas): Let us use Gaussian copulas to sample a random function $f(x;\omega)$ evaluated at $N = 100$ evenly-spaced grid points x_j. This yields a joint PDF in 100 dimensions. For this test problem, consider

$$f(x;\omega) = \sin(3x) e^{\xi_1(\omega)+\xi_2(\omega)} + e^{\cos(2x-2)}\xi_3(\omega) + \frac{2\xi_4(\omega)}{2\sin(2x)+3} \qquad x \in [0, 3\pi], \tag{159}$$

where $(\xi_1, \ldots, \xi_4)$ are four i.i.d uniform random variables in $[0, 1]$. The random vector representing $f(x;\omega)$ on the spatial grid has components

$$X_j = f(x_j;\omega), \qquad x_j = \frac{3\pi(j-1)}{N-1}, \qquad j = 1, \ldots, 100. \tag{160}$$

We are given the marginals of $f(x;\omega)$, i.e., the PDFs of $f(x_i;\omega)$ at each spatial point x_i (see Figure 4), and the two-point correlation

$$C_{ij} = \mathbb{E}\left\{ f(x_i;\omega) f(x_j;\omega) \right\}. \tag{161}$$

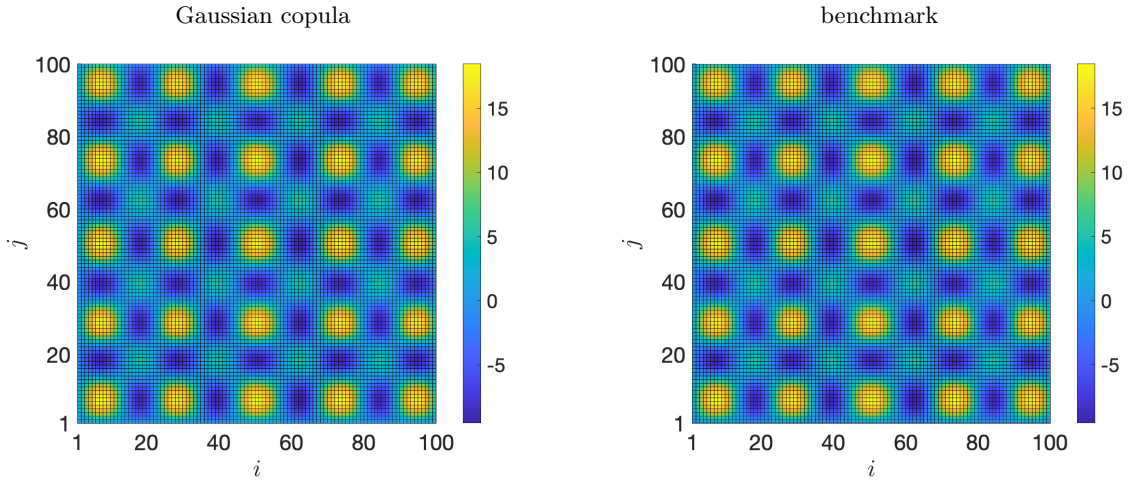

Figure 5: Comparison between the correlation matrix we obtain from Gaussian copula and the benchmark correlation (161).

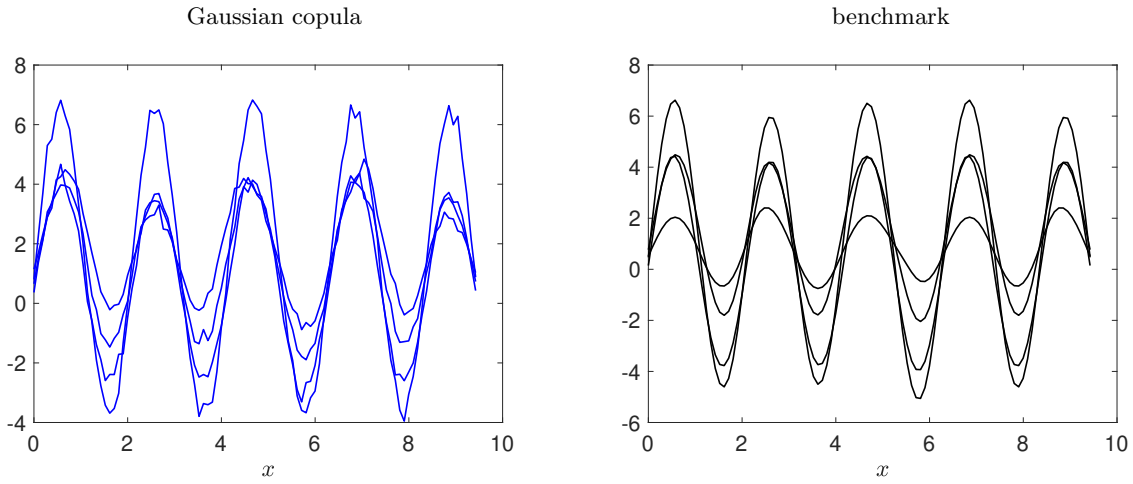

Figure 6: Samples of the Gaussian copula approximation (left) and samples from (159).

In Figure 5 we compare the correlation of the samples obtained from the Gaussian copula with the benchmark correlation. It is seen that the Gaussian copula has basically the same correlation and also the same marginals (see Figure 4) as the benchmark model.

In Figure 6 we plot a few samples of the Gaussian copula approximation and compare them with samples of the (159). Note that there exists a strong correlation structure in the random function (159) due to smoothness. Hence sampling each marginal distribution independently would completely destroy such correlation structure and result in sample patterns that are essentially Brownian motion (zig-zag) sort of noise. On the other hand, the sample patterns shown in Figure 6 capture the overall trend of the original patterns, with some minor deviations.

Learning high-dimensional PDFs from data

This is an active area of research today. There are two mainstreams commonly used for learning high-dimensional PDFs from data:

- Learning the functional form of the PDF using high-dimensional function representations such as tensors [18, 15], mixture PDF models, kernel density approaches, etc. These approaches can be

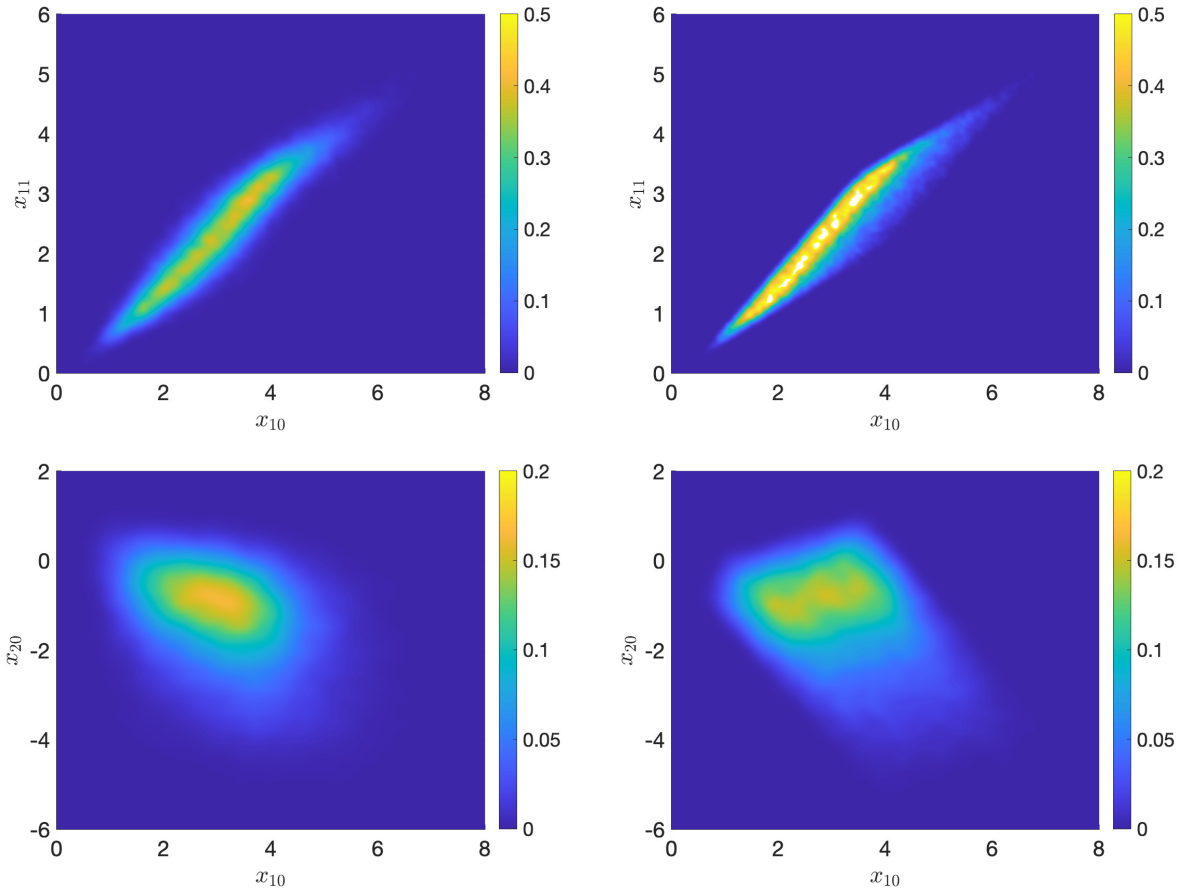

Figure 7: Joint PDFs from the Gaussian copula model (left), and benchmark PDFs (right). Note that the joint PDF of (159) at two neighboring nodes (x_{10} and x_{11}) is rather thin, due to the strong correlation between such random variables induced by the smoothness of $f(x;\omega)$. Note also that the Gaussian copula model does not reproduce exactly the joint PDF, but rather it approximates it.

computationally expensive due to due to high-dimensionality of the function $p(x_1,\ldots,x_n)$, which requires a representation.

- Developing a sampler for the PDF we are estimating from data, rather that attempting to compute the functional form of the PDF. These methods are computationally much more efficient that learning the functional form of the PDF. There are many different techniques that have been proposed/developed in the past decade or so to learn a sampler for a high-dimensional PDF that is learned from data.

  a) **Generative Adversarial Networks (GANs)**. GANs consist of two components: a generator that produces fake data samples and a discriminator that tries to distinguish between real data and generated samples. The generator learns to create samples that mimic the real data distribution. GANs do not provide an explicit PDF but can generate samples from the high-dimensional distribution. They are widely used in tasks like image generation, video synthesis, and more (see [7, 1]).

  b) **Variational Autoencoders (VAEs)**. VAEs combine neural networks with probabilistic models. The encoder maps input data into a latent space, and the decoder reconstructs the data from this latent space. The VAE is trained to maximize a lower bound on the likelihood of the data, thus learning an approximate distribution. The model does not produce an explicit PDF but provides a way to sample from the learned distribution.

c) **Diffusion Models**. Diffusion models gradually convert data samples into noise through a diffusion process, then learn to reverse this process to generate new samples. Like GANs, they can implicitly learn high-dimensional distributions but do so through iterative denoising steps. In this class of models we find, e.g., DDPM [14], score-based models [16], and INDM [9].

d) **Continuous normalizing flows (CNFs)**. In CNFs, the goal is to transform a simple base distribution (like a Gaussian) into a complex target distribution through a sequence of invertible transformations, governed by differential equations (ODEs). A particular case of CNFs is flow matching [12], which simplifies this process by directly learning the optimal velocity field, which describes how data points flow from the base distribution to the target. Unlike traditional approaches that require learning complex dynamics over continuous time, flow matching trains the model to match the flow of data at discrete steps, making it more efficient. This method allows for the accurate estimation of high-dimensional PDFs, while also ensuring that the transformation remains invertible. Flow approaches are provably convergent [6, 4].

Appendix A: Derivation of the Liouville equation

Consider the nonlinear dynamical system

$$\begin{cases} \dfrac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{f}(\boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0(\omega) \end{cases} \tag{162}$$

where $\boldsymbol{x}_0(\omega)$ is a random vector with known joint probability density function $p_0(\boldsymbol{x})$. We know that if $\boldsymbol{f}(\boldsymbol{x})$ is continuously differentiable in $\boldsymbol{x}$ then (162) admits a smooth flow $\boldsymbol{x}(t, \boldsymbol{x}_0(\omega))$, which is at least continuously differentiable in $\boldsymbol{x}_0$. The flow is also continuously differentiable in t, i.e., $x(t, \boldsymbol{x}_0(\omega))$ is a diffeomorphism in t. We are interested in determining an evolution equation for $p(\boldsymbol{x}, t)$, i.e., the probability density function of $\boldsymbol{x}(t, \boldsymbol{x}_0)$ at time t. To this end, consider the characteristic function representation of the PDF $p(\boldsymbol{x}, t)$

$$\phi(\boldsymbol{a}, t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)} p(\boldsymbol{x}_0) d\boldsymbol{x}_0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i\boldsymbol{a}\cdot\boldsymbol{x}} p(\boldsymbol{x}, t) d\boldsymbol{x} \tag{163}$$

Differentiating with respect to t yields

$$\begin{aligned} \frac{\partial \phi(\boldsymbol{a}, t)}{\partial t} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} i\boldsymbol{a} \cdot \frac{\partial \boldsymbol{x}(t, \boldsymbol{x}_0)}{\partial t} e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)} p(\boldsymbol{x}_0) d\boldsymbol{x}_0 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} i\boldsymbol{a} \cdot \boldsymbol{f}\left(\boldsymbol{x}(t, \boldsymbol{x}_0)\right) e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)} p(\boldsymbol{x}_0) d\boldsymbol{x}_0 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} i\boldsymbol{a} \cdot \boldsymbol{f}\left(\boldsymbol{x}\right) e^{i\boldsymbol{a}\cdot\boldsymbol{x}} p(\boldsymbol{x}, t) d\boldsymbol{x} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{x}} \left(e^{i\boldsymbol{a}\cdot\boldsymbol{x}}\right) \cdot \boldsymbol{f}\left(\boldsymbol{x}\right) p(\boldsymbol{x}, t) d\boldsymbol{x} \\ &= -\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i\boldsymbol{a}\cdot\boldsymbol{x}} \nabla \cdot \left(\boldsymbol{f}\left(\boldsymbol{x}\right) p(\boldsymbol{x}, t)\right) d\boldsymbol{x}. \quad \text{(integrating by parts)} \end{aligned} \tag{164}$$

By using (163) and (164) we obtain

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i\boldsymbol{a}\cdot\boldsymbol{x}} \left[\frac{\partial p(\boldsymbol{x}, t)}{\partial t} + \nabla \cdot \left(\boldsymbol{f}\left(\boldsymbol{x}\right) p(\boldsymbol{x}, t)\right)\right] d\boldsymbol{x} = 0, \quad \text{for all } \boldsymbol{a} \in \mathbb{R}^n, \tag{165}$$

which implies that the function between square bracket must be equal to zero for all $\boldsymbol{x}$ and all t, i.e.,

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} + \nabla \cdot \left(\boldsymbol{f}\left(\boldsymbol{x}\right) p(\boldsymbol{x}, t)\right) = 0 \qquad \text{(Liouville equation)}. \tag{166}$$

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, pages 1–30, 2017.

[2] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957, 2010.

[3] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2011.

[4] X. Cheng, J. Lu, Y. Tan, and Y. Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *arXiv:2310.17582*, 2024.

[5] C. Czado. *Analyzing dependent data with vine copulas*. Springer, 2019.

[6] Y. Gao, J. Huang, Y. Jiao, and S. Zheng. Convergence of continuous normalizing flows for learning probability distributions. *arXiv:2404.00551*, 2024.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Su, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[8] A. I. Khuri. Applications of Dirac's delta function in statistics. *Int. J. Math. Educ. Sci. Technol.*, 35(2):185–195, 2004.

[9] D. Kim, B. Na, S. J. Kwon, D. Lee, W. Kang, and I.-C. Moon. Maximum likelihood training of implicit nonlinear diffusion models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024.

[10] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[11] R. Kubo. Generalized cumulant expansion method. *Journal of the Physical Society of Japan*, 17(7):1100–1120, 1962.

[12] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2023.

[13] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill, third edition, 1991.

[14] Y. Ren, H. Zhao, Y. Khoo, and L. Ying. High-dimensional density estimation with tensorizing flow. *Res. Math. Sci.*, 10:1–25, 2023.

[15] Y. Ren, H. Zhao, Y. Khoo, and L. Ying. High-dimensional density estimation with tensorizing flow. *Res. Math. Sci.*, 10:1–25, 2023.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, pages 1–36, 2021.

[17] A. Spantini, D. Bigoni, and Y. Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19:1–71, 2018.

[18] X. Tang and L. Ying. Solving high-dimensional Fokker-Planck equation with functional hierarchical tensor. *Journal of Computational Physics*, 511:113110, 2024.