Information Theory and Point Processes
Ronald Mahler, Random Sets LLC, Eagan, MN, U.S.A.

Abstract—This paper addresses theoretically correct vs. incorrect ways to apply information theory to point processes.

I. INTRODUCTION
Point processes (p.p.) theory addresses the statistical behavior of randomly varying finite ensembles of points in some state space \( \mathcal{X} \). There are multiple formulations of the theory; here the focus will be on the vector-based version. The purpose of the paper is to address theoretically correct vs. incorrect ways of devising information-theoretic functionals for vector-based p.p.s.

Several such functionals were recently introduced in [2]. In what follows it will be demonstrated that the theoretical approach in that paper is incorrect and that, as a consequence, Eqs. (42-54) of [2] are mathematically undefined when employed in typical engineering applications. In particular, in these applications the key Eqs. (42,47,51) involve summations of terms with incommensurable units of measurement; and this remains the case even if the theoretically correct approach is applied instead.

The paper is organized as follows: summary of the approach in [2] (Section II); description of the correct approach (Section III); and critique of the mathematically undefined formulas (Section IV).

II. BACKGROUND
A. Vector-Based Point Processes
Let \( \mathcal{X} \) be a topological space (hereafter referred to as the “base space”), endowed with a measure \( \lambda_X(B) \) (hereafter referred to as the “base measure”) defined on the Borel-measurable subsets \( B \) of \( \mathcal{X} \). In typical engineering applications \( \mathcal{X} \) is a region of a Euclidean space with unit measure, \( \lambda_X \), which is what will be assumed hereafter. In this case \( \lambda_X \) is Lebesgue measure and the unit of measurement is \( \epsilon \), which is what will be assumed hereafter.

The measure \( \lambda_X(B) \) is defined recursively by (1)

\[
\lambda_X(B) = \int_B \frac{dx}{f(x)}
\]

where the “test function” \( f(x) \) describes the realizations of \( \Phi \) from \( B \).

B. Probability Generating Function of a p.p.
According to [3], p. 2, the probability generating functional (p.g.f.) \( G_P(h) \) of a p.p. \( \Phi \) was introduced by Bartlett and Kendall in the late 1940s and is (3)

\[
G_P(h) = \sum_{n=0}^{\infty} \left( \prod_{i=1}^{n} h(x_i) \right) P^n_{\Phi}(dx_1, ..., dx_n)
\]

where the “test function” \( h : \mathcal{X} \to [0, 1] \) is unitless. Note that \( 0 \leq G_P(h) \leq G_P(1) = \lambda_X(\mathcal{X}) = 1 \).

C. Chain Differential of a p.g.f.
If \( F(h) \geq 0 \) is a functional on \( h : \mathcal{X} \to [0, 1] \), then its chain differential is (4)

\[
\delta F(h; \eta) = \lim_{\varepsilon \to 0} \frac{F(h + \varepsilon \eta) - F(h)}{\varepsilon}
\]

if the limits exist and are identical for any sequences \( \varepsilon \to 0 \)
and \( \eta \rightarrow (\eta) \) (pointwise).

Assume that \( F(h) \) is well-behaved enough that:
(a) \( \eta \rightarrow F(h; \eta) \) is linear and continuous in \( \eta \)
(b) \( \mu F, h \) is absolutely continuous w/r/t (with respect to) \( \lambda_X \).

Then \( \delta F(h; \delta_x) \) is shorthand for the Radon-Nikodým derivative \( \frac{d\mu F, h}{d\lambda_X}(x) \) (5).

P. \( \Phi \) and \( \mathcal{G}_P \) are related by (6)

\[
P^{(n)}_{\Phi}(B_1 \times ... \times B_n) = \frac{1}{n!} \delta^n \mathcal{G}_P(0; B_1, ..., B_n)
\]
for measurable \( B_1, \ldots, B_n \subseteq \mathcal{X} \), with \( p^{(0)}(\{\phi\}) = G_\Phi(0) \). For \( n \geq 0 \) define \((2), \text{Eq. (39)}:\)

\[
p^{(n)}_\Phi(x_1, \ldots, x_n) = \frac{1}{n!} \delta^n G_\Phi(0; \delta x_1, \ldots, \delta x_n),
\]

so that \( p^{(0)}_\Phi = G_\Phi(0) \). The left side of \((5)\) is the family of Janossy densities of \( \Phi \) indexed by \( n \geq 1 \). Because \( P^\Phi \) is symmetric, \( p^{(n)}_\Phi(x_1, \ldots, x_n) \) is symmetric w/r/t \( x_1, \ldots, x_n \) for each \( n \geq 2 \).

The \( p^{(n)}_\Phi(x_1, \ldots, x_n) \) exist only if \( \Phi \) is “simple”—i.e., if \( x_1, \ldots, x_n \) are distinct for any realizations \( \Phi = (x_1, \ldots, x_n) \) with \( n \geq 2 \). Thus all \( \mu \)'s in \((2)\) have implicitly been assumed to be simple.

The chain differential, Gâteaux differential, and Frechét derivative of a p.g.f. \( G_\Phi \) exist and are equal; and are equivalent to the Volterra functional derivative\(^3\) \((\delta G_\Phi/\delta x)(h)\) of \( G_\Phi \) in the sense that

\[
\delta G_\Phi(h; \eta) = \int \eta(x) \cdot \frac{\delta G_\Phi(x)}{\delta x}(h) dx
\]

and thus that \( \delta G_\Phi(h; \delta_x) = (\delta G_\Phi/\delta x)(h) \).\(^4\) It is well known that the Frechét derivative admits a chain rule. The chain differential was introduced in 2005 in \((4)\) to permit a chain rule for functions that are not Frechét differentiable. The need for it in a p.g.f. context is therefore unclear.

For a more detailed critique of the vector-based p.p. framework, see \((6)\).

III. INFORMATION THEORY AND P.P.'S

The purpose of this section is to describe the theoretically correct way to apply information theory to p.p.'s. Given the mathematical formulation in \((2)\), “theoretically correct” means “correctly measure-theoretic.”

Remark 1: This section is not needed to understand why Eqs. (42-54) of \((2)\) are erroneous—see \((26,28,29)\). Rather, it is required for the demonstration in Remark 9 that these errors cannot be corrected by employing the theoretically correct approach described in this section.

It is organized as follows: Lebesgue integration (Section \(\text{III-A}\)); differential entropy as a simple example (Section \(\text{III-B}\)); measure and integration for p.p.'s (Section \(\text{III-C}\)); and optimal state estimation for p.p.'s (Section \(\text{III-D}\)).

A. Lebesgue Integration

Let \( \mu(B) \) be a unitless measure of the Lebesgue-measurable subsets \( B \subseteq \mathcal{X} \) and assume that it is absolutely continuous w/r/t \( \lambda_X \)—i.e., \( \lambda_X(B) = 0 \) implies \( \mu(B) = 0 \). Then by the Radon-Nikodým theorem, there is an almost everywhere unique Lebesgue-integrable function \( d\mu/d\lambda_X : \mathcal{X} \to \mathbb{R}^+ \)—the Radon-Nikodým derivative of \( \mu \) w/r/t the base measure \( \lambda_X \)—such that

\[
\mu(B) = \int_B \frac{d\mu}{d\lambda_X}(x) dx
\]

for all \( B \) and where \( \lambda_X(dx) \) has been abbreviated as \( dx \). Because \( \mu(B) \) is unitless and \( dx \) has unit \( 1 \), \( (d\mu/d\lambda_X(x)) \) must have unit \( \lambda^{-1} \).

Remark 2: Since \( G_\Phi \) and thus \( \mu_{G_\Phi,h} \) are unitless, it follows that \( \delta G_\Phi(h; \delta_x) = (d\mu_{G_\Phi,h}/d\lambda_X)(x) \) has unit \( \lambda^{-1} \) and \( \delta^n G_\Phi(h; \delta_x, \ldots, \delta_x) \) has unit \( \lambda^{-n} \).

Remark 3: From this it follows that \( p^{(n)}_\Phi(x_1, \ldots, x_n) \) has unit \( \lambda^{-n} \) for \( n \geq 2 \).

B. Simple Example: Differential Entropy

Let \( X \in \mathcal{X} \) be an r.v. with probability measure \( P_X(B) \) and probability density function (p.d.f.) \( f_X(x) = dP_X/\lambda_X(x) \), where \( dP_X/d\lambda_X \) is the Radon-Nikodým derivative of \( P_X \) w/r/t \( \lambda_X \). Then the differential entropy (DE) of \( X \) is:

\[
DE(f_X) = - \int f_X(x) \log f_X(x) dx = - \int \log f_X(x) P_X(dx).
\]

The DE has two well-known limitations as a quantifier of information. First, it is not invariant w/r/t change of coordinates. Second and in particular, it is undefined if \( X \) and thus \( f_X \) have units of measurement and thus \( \log f_X(x) \) is undefined. A minimum requirement for any quantifier of information (or entropy) should be that its numerical value does not change if (for example) one converts from metric to English units.

Csíszár information functionals, such as the Kullback-Leibler divergence

\[
KL(f_X; f_0) = \int \log \left( \frac{f_X(x)}{f_0(x)} \right) P_X(dx),
\]

do not have these limitations—see, e.g., Eq. (3) of \((7)\).

Nevertheless, the DE provides a simple illustration of how to correctly apply information theory to p.p.'s. Here we employ it to summarize this approach, assuming the measure-theoretic results to be established in Sections \(\text{III-C}\) and \(\text{III-D}\).

The DE of a p.p. \( \Phi \) must be

\[
DE(f_\Phi) = - \int f_\Phi(\varphi) \log f_\Phi(\varphi) d\varphi = - \int \log f_\Phi(\varphi) P_\Phi(d\varphi)
\]

where \( f_\Phi(\varphi) = (dP_\Phi/d\lambda_{\mathcal{X}^\infty})(\varphi) \) and where \( \lambda_{\mathcal{X}^\infty}(\mathcal{B}) \) is a measure on the p.p. state space \( \mathcal{X}^\infty \). But \( 10 \) is meaningless unless we answer the following question: What is \( \lambda_{\mathcal{X}^\infty} \)?

To be conceptually consistent, \( \lambda_{\mathcal{X}^\infty} \) must be an extension of \( \lambda_X \) to \( \mathcal{X}^\infty \). If \( X \) has unit of measurement \( \iota \) then the simplest extension is \( \lambda_{\mathcal{X}^\infty} = \lambda_X \) with \( c > 0 \) as given below in \((11)\). This leads, in \((12)\), to the formula \( f_\Phi(\varphi) = c^{\varphi} p_\Phi^{(1)}(\varphi) \) (which in turn, means that if \( X \) has a unit then \( p_\Phi^{(1)}(\varphi) \) is not the p.d.f. of \( \Phi \). That is:

Remark 4: Any p.p. quantifier of information (or entropy) that is measure-theoretic and generally applicable must be defined using \( f_\Phi(\varphi) \) rather than \( p_\Phi^{(1)}(\varphi) \).

C. Measure and Integration for p.p.'s

Just as the \( \sigma \)-algebra of \( \mathcal{X}^\infty \) is an extension of the \( \sigma \)-algebra of \( X \), so \( \lambda_{\mathcal{X}^\infty} \) is the following well-known extension of \( \lambda_X \) to \( \mathcal{X}^\infty \) \((5)\), p. 715:

\[
\lambda_{\mathcal{X}^n}(\mathcal{B}) = \mathbb{1}_\mathcal{X}(\phi) + \sum_{n \geq 1} \nabla^n_X(\mathcal{B} \cap \mathcal{X}^n) \frac{\lambda_X(x^n)}{c^n}.
\]

---

\(^3\)Functional derivative: see \((5)\), p. 375.

\(^4\)This all follows from the fact that a p.g.f. is a functional power series and thus is, in this sense, analytic \((6)\).
Here, \( c > 0 \) has unit \( \mu \); \( \lambda_X^1 = \lambda_X \); and for \( n \geq 2 \), \( \lambda_X^n \) is the extension of \( \lambda_X \) to \( X^n \), in which case \( \lambda_X^n(B_1 \times \ldots \times B_n) \) has unit \( \mu^n \). Because of \( c \), the summation in (11) is mathematically well-defined since its terms are unitless. Like \( \lambda_X \), \( \lambda_c \) is a possibly infinite-valued measure. Unlike \( \lambda_X \), it is unitless.

The projection measures of \( \lambda_c \) are \( \lambda_c^{(0)}(B) = 1_B(\phi) \) and \( \lambda_c^{(n)}(B) = \lambda_X^n(B \cap X^n) / c^n \) for \( n \geq 1 \). In particular, \( \lambda_c^{(1)}(B) = \lambda_X(B \cap X) / c \).

Let \( f(\varphi) \) be a nonnegative unitless function of \( \varphi \). Then its integral w/r/t \( \lambda_c \) within \( B \subseteq X^n \) is

\[
\int_B f(\varphi) \lambda_c(d\varphi) = f(\phi) \cdot 1_B(\phi) + \sum_{n=1}^{\infty} \frac{1}{c^n} \int_{B \cap X^n} f(x_1, \ldots, x_n) dx_1 \cdots dx_n
\]

where \( \lambda_X^n(dx_1, \ldots, dx_n) \) has been abbreviated as \( dx_1 \cdots dx_n \). Thus \( \lambda_c(B) = \int_B \lambda_c(d\varphi) \).

D. Probability Density Function (p.d.f.) of a p.p.

If \( P_\Phi \) is absolutely continuous w/r/t \( \lambda_c \) then its p.d.f. is \( f_\Phi(\varphi) = dP_\Phi / d\lambda_c(\varphi) \), which is characterized by the Radon-Nikodym theorem

\[
P_\Phi(B) = \int_B f_\Phi(\varphi) \lambda_c(d\varphi) = \int_B \frac{dP_\Phi}{d\lambda_c}(\varphi) \lambda_c(d\varphi)
\]

(13)

\[
= \frac{dP_\Phi}{d\lambda_c}(\phi) \cdot 1_B(\phi) + \sum_{n=1}^{\infty} \frac{1}{c^n} \int_{B \cap X^n} \frac{dP_\Phi}{d\lambda_c}(x_1, \ldots, x_n) dx_1 \cdots dx_n
\]

(14)

where the restriction of \( dP_\Phi / d\lambda_c \) to \( X^n \) is equal to \( dP_\Phi^{(n)} / d\lambda_c \). By (14), (11), and (1), for \( n \geq 1 \) the projection measures \( P_\Phi^{(n)} \) of \( P_\Phi \) are given by

\[
\frac{1}{n!} \delta_\Phi(0; 1_{B_1}, \ldots, 1_{B_n}) = \frac{1}{c^n} \int_{B_1 \times \ldots \times B_n} \frac{dP_\Phi^{(n)}}{d\lambda_c}(x_1, \ldots, x_n) dx_1 \cdots dx_n
\]

(15)

\[
= \frac{1}{c^n} \int_{B_1 \times \ldots \times B_n} f_\Phi(x_1, \ldots, x_n) dx_1 \cdots dx_n.
\]

(16)

From this and (5) it follows that

\[
P_\Phi^{(n)}(x_1, \ldots, x_n) = \frac{1}{c^n} \cdot f_\Phi(x_1, \ldots, x_n)
\]

(17)

and thus that the p.d.f. of \( P_\Phi \) is the unitless function

\[
f_\Phi(\varphi) = c^{1/\alpha} \cdot P_\Phi^{(1/\alpha)}(\varphi).
\]

(18)

The application of advanced Gibbs statistical sampling techniques to exact closed-form approximations of \( f_\Phi(\varphi) \) or more precisely, of \( f_\Phi(X) \) as defined in Section III-B, has led to implementations of the generalized labeled multi-Bernoulli (GLMB) filter that are capable of simultaneous real-time tracking of over a million targets in significant clutter using off-the-shelf computing equipment [3].

Remark 6: It follows that (14) can be rewritten as

\[
P_\Phi(B) = \sum_{n=0}^{\infty} \int_{B \cap X^n} p_\Phi^{(n)}(x_1, \ldots, x_n) dx_1 \cdots dx_n.
\]

(19)

Remark 7: It also follows that if \( n \geq 1 \)

\[
P_\Phi^{(n)}(dx_1, \ldots, dx_n) = p_\Phi^{(n)}(x_1, \ldots, x_n) dx_1 \cdots dx_n.
\]

(20)

For, from (12) and (18) we have

\[
\int f(x_1, \ldots, x_n) P_\Phi^{(n)}(dx_1, \ldots, dx_n)
\]

(21)

\[
= \int f(x_1, \ldots, x_n) \cdot \frac{dP_\Phi^{(n)}}{d\lambda_c}(x_1, \ldots, x_n) \lambda_c(dx_1, \ldots, dx_n)
\]

(22)

\[
= \frac{1}{c^n} \int f(x_1, \ldots, x_n) \cdot c^n p_\Phi^{(n)}(x_1, \ldots, x_n) dx_1 \cdots dx_n
\]

(23)

E. Optimal State Estimation for p.p.'s

The most probable realization of \( \Phi \) is the maximum a posteriori (MAP) estimate extracted from (18):

\[
\hat{x}_n = \arg \sup_{x_1, x_2, \ldots, x_n} c^n p_\Phi^{(n)}(x_1, \ldots, x_n).
\]

(24)

Since each \( c > 0 \) determines a different most-probable estimate, (24) is essentially useless unless we answer the following question: What is the best choice for \( c \)?

To answer it, make the following changes from vector to finite-set notation. Assume that \( x_1, x_2, \ldots, x_n \) are distinct and let \( X = \{x_1, x_2, \ldots, x_n\} \). Write \( f_\Phi(X) = n! \cdot p_\Phi^{(n)}(x_1, \ldots, x_n) \) and \( |X| = n \), in which case (24) becomes

\[
\hat{X}_n = \arg \sup_{X \subseteq X^n} c^{|X|} f_\Phi(X).
\]

(25)

According to the analysis of [5], pp. 499-500, for this estimate to be accurate the magnitude of \( c \) should be approximately equal to the accuracy with which individual states \( x \in X \) are to be estimated.

Remark 8: Thus \( c \neq 1 \cdot t \) in general.

IV. MATHEMATICALLY UNDEFINED FORMULAS

We are now in a position to demonstrate that Eqs. (42-54) of [2] are mathematically undefined.

Begin by inspecting the key formula Eq. (42), the “Information generating functional for entropy”:

\[
G_\Phi(h) = \sum_{n=0}^{\infty} \left( \prod_{i=1}^{n} h(x_i) \right) \cdot \lambda \Phi(x_{n+1}) \cdot P_\Phi^{(n)}(dx_1, \ldots, dx_n).
\]

(26)
Since $P^{(n)}_\phi$ is a unitless measure and since by Remark 3 $p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha}$ has unit $e^{\alpha}$, then as long as $\alpha \neq 0$ the summation is mathematically undefined since its terms have different units of measurement for $n \geq 0$—and thus are incommensurable.

To see this more explicitly note that, by (26), we can rewrite (27) as:

$$G^\alpha_\phi(h) = \sum_{n \geq 0} \int \left( \prod_{i=1}^{n} h(x_i) \right) p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha} dx_1 \cdots dx_n$$ (27)

where $p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha} dx_1 \cdots dx_n$ has unit $e^{\alpha}$.

Remark 9: This error cannot be corrected by substituting $f_\phi(x_1, ..., x_n)$ in place of $p^{(n)}_\phi(x_1, ..., x_n)$—indeed, it becomes worse. For then $p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha} dx_1 \cdots dx_n$ becomes $e^{\alpha} p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha} dx_1 \cdots dx_n$ with unit $e^{\alpha}$, regardless of the values of both $\alpha$ and $c$. In particular, the error remains even when $\alpha = 0$.

Remark 10: It might be argued that $c = 1 \cdot t$ suffices as an engineering simplification or approximation, but this is not the case. By Remark 3 $c \neq 1 \cdot t$ in general if we are to find the most probable estimate of $\Phi$—i.e., find the best estimate for engineering purposes.

These remarks apply with full force to:

1) Key formula Eq. (47), the “Laplace information functional for entropy”:

$$L^\alpha_\phi(f) = \sum_{n \geq 0} \int \exp \left( -\sum_{i=1}^{n} f(x_i) \right) p^{(n)}_\phi(x_1, ..., x_n)^{-\alpha} dx_1 \cdots dx_n$$ (28)

(for functions $f : \mathcal{X} \to \mathbb{R}^+$, which has the general form $L^\alpha_\phi(f) = G^\alpha_\phi(e^{-f})$ where $e^{-f}(x) = e^{-f(x)}$.

2) Key formula Eq. (51), the “cumulant information functional”:

$$W^\alpha_\phi(f) = \log L^\alpha_\phi(f).$$ (29)

3) Any formula defined in terms of $G^\alpha_\phi(h)$, $L^\alpha_\phi(h)$, or $W^\alpha_\phi(h)$—i.e., Eqs. (42-54).

Additional errors should be noted. Consider the “Shannon entropy,” Eq. (43), which from Eq. (42) can be written as

$$\left[ \frac{\partial}{\partial \alpha} G^\alpha_\phi(1) \right]_{\alpha=0} = -\sum_{n \geq 0} \int \log p^{(n)}_\phi(x_1, ..., x_n) p^{(n)}_\phi(dx_1, ..., dx_n)$$ (30)

$$= -\int \log p^{(1)}_\phi(\varphi) p^{(1)}_\phi(\varphi)(d\varphi).$$ (31)

By Remark 3 this is a theoretically erroneous version of the differential entropy formula since $p^{(1)}_\phi(\varphi) \neq f_\phi(\varphi)$. It is also mathematically undefined since, by Remark 3 $p^{(n)}_\phi(x_1, ..., x_n)$ has unit $e^{-n}$ and thus $\log p^{(1)}_\phi(\varphi)$ is mathematically undefined.

The same is true of the “Shannon entropy moments,” Eq. (48):$^5$

$$E_\phi \left[ \log p^{(m)}_\phi \right] = (-1)^m \left[ \frac{\partial^m}{\partial \alpha^m} G^\alpha_\phi(0) \right]_{\alpha=0}$$ (32)

$$= \int \left( \log p^{(|\varphi|)}_\phi(\varphi) \right)^m p^{(|\varphi|)}_\phi(\varphi)(d\varphi).$$ (33)

Remark 11: One could patch up these particular errors by substituting $f_\phi(\varphi)$ in place of $p^{(|\varphi|)}_\phi(\varphi)$. But this would not change the fact that they are secondary errors inherited from the inherently erroneous $G^\alpha_\phi(h)$ and $L^\alpha_\phi(h)$.

REFERENCES

[1] D. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes, Springer-Verlag, New York, 1988.

[2] D. Clark, “Local entropy statistics for point processes,” IEEE Trans. Info. Theory, Vol. 66, No. 2, pp. 1155-1163, 2020.

[3] J. Moyal, “The general theory of stochastic population processes,” Acta Mathematica, Vol. 108, pp. 1-31, 1962.

[4] P. Bernard, “Chain differentials with an application to the mathematical operator,” Nonlinear Analysis: Theory, Methods and Applications, Vol. 62, No. 7, pp. 1225-1233, 2005.

[5] R. Mahler, Statistical Multisource-Multitarget Information Fusion, Artech House, Norwood, MA, 2007.

[6] R. Mahler, “Statistics 103 for multitarget tracking,” Sensors, 19(1): 202, 2019, open source: https://doi.org/10.3390/s19010202.

[7] Yu Qiao and Nobuki Minematsu, “A study on invariance of f-divergence and its application to speech recognition,” IEEE Trans. Sign. Proc., Vol. 58, No. 7, pp. 3884-3890, 2010.

[8] M. Beard, B.-T. Vo, and B.-N. Vo, “A solution for large-scale multi-object tracking,” IEEE Trans. Sign. Proc., Vol. 68, pp. 2754-2769, 2020.

$^5$As originally written, Eq. (48) had a typo: $L^\alpha_\phi(f)$ should have been $L^\alpha_\phi(0)$.