Communications in Advanced Mathematical Sciences
Vol. I, No. 1, 45-64, 2018

A Rosetta Stone for information theory and differential equations

Alessandro Selvitella[1]*

Abstract
In this paper, we propose a dictionary between Partial Differential Equations and Information Theory. As a model case, we will discuss in detail the example of the Schrödinger Equation and Shannon Information Theory. Comments will be made in both the continuous and discrete case and in both the noiseless and noisy case.

Keywords: Information Theory, PDEs, Nyquist Bit Rate, Shannon Capacity, Strichartz Estimates
2010 AMS: Primary 35Q94, Secondary 94A15, 94A17

[1] *Department of Mathematics and Statistics of University of Ottawa 585 King Edward Avenue (ON) K1N 7N5 Canada*
***Corresponding author**: aselvite@uottawa.ca
Received: 27 July 2018, **Accepted:** 16 September 2018, **Available online:** 30 September 2018

1. Introduction

The Rosetta Stone is a stele inscribed with a decree issued at Memphis, Egypt, in 196 BC on behalf of King Ptolemy V. The Rosetta Stone has writings in both Egyptian and Greek, using the three different scripts popular in Egypt at that time: hieroglyphic, demotic and Greek. This particular feature of having written essentially the same text on itself in the three scripts, provided a fundamental dictionary to understand Egyptian hieroglyphs.

The search of dictionaries is not a phenomenon confined to Humanistic Sciences and Linguistics: it is ubiquitous also in Technical Sciences, such as Mathematics and Physics. One of the most famous of these dictionaries is due to Wu and Yang [49], who published a paper containing a list of correspondences between the mathematical terminology related to the theory of Connections on Riemannian Manifolds, and the physical terminology concerning the Yang-Mills Theory. This dictionary translates, one to the other, physical and mathematical terminologies. This list is referred to in the literature by some as the *Wu-Yang Dictionary* (See for example [50]). The relationship between these two fields was most likely clear to many others before, at least implicitly. However, the *Wu-Yang Dictionary* played an important role. Since then, the interactions between mathematics and physics have become more natural and fruitful.

In a recent series of seminars, Weinstein [47], [48] proposed new applications of Gauge Theory beyond those familiar in the Natural Sciences. In his view, Neoclassical Economics is an example of Gauge Theory and, for this reason, he proposes a dictionary between Economics and Gauge Theory. The terminology "Rosetta Stone" in our title is inspired by his presentations.

In this paper, we propose a dictionary between Partial Differential Equations (PDEs) and Information Theory.

Information Theoretical methods have been already used to investigate the large time behavior of solutions of PDEs. Remarkable

is the use of entropy (Shannon Entropy, Kullback-Leibler Entropy, Von Neumann Entropy, etc..) to study the asymptotics of dissipative and heat-type equations, both in the case of classical and quantum systems. For a detailed discussion, we refer to [21].

Beyond the typical use of instruments from Information Theory being applied to PDEs, our dictionary underlines that many instruments of Information Theory already have natural counterparts in PDEs. For the use of some Information methods in the context of PDEs and Optimal Transport, we refer to [45], [44], [35], [5]. Our model examples will be the Schrödinger Equation and Shannon Theory of Communication.

Several authors have studied the relationship between the Schrödinger Equation and Information Theoretical tools, like the Fisher Information. In particular, there is a program due to Frieden and his collaborators (see for example [16], [17], [18], [19], [20] and the reference therein), based on what they called Extreme Physical Information (EPI). EPI states that scientific laws can be derived through the Fisher Information and are ruled by differential equations and probability. They extended their approach to encompass existing laws of biology, cancer growth, chemistry, and economics. In our paper, we do not argue in favor or against the claim that Information is the leading principle of all physical laws, but we underline that the vocabulary of Information Theory can have a natural translation into the PDE language and vice versa.

The two pioneering achievements of Classical Information Theory are the following theorems due to Shannon [40] (See Section 2 or directly [40] for the precise definitions and terminology involved in the theorems). The first theorem treats the case of a noiseless channel.

Theorem 1.1 (Fundamental theorem for a noiseless channel)**.** *Let a source S have an entropy H and a channel K have a capacity C. Then, it is possible to encode the output O of S in such a way to transmit at the average rate $C/H-\varepsilon$ over the channel K. Here $0<\varepsilon\ll 1$ is an arbitrary constant. It is not possible to transmit at an average rate greater than C/H.*

In this theorem, Shannon demonstrates the existence of a limit to the efficiency of *Source Coding*. The entropy of a source H corresponds to the minimum binary digits to be used for its coding. Any discrepancy from this limit translates into a growing complexity. A second theorem deals with the noisy case.

Theorem 1.2 (Fundamental theorem for a channel with noise)**.** *Let a source S have an entropy H and a channel K have a capacity C. If $H\leq C$, there exists a coding system such that the output O of S can be transmitted over K with an arbitrarily small frequency of errors (also called* equivocation*). If $H>C$, it is possible to encode S so that the* equivocation *is less than $H-C+\varepsilon$. Here $0<\varepsilon\ll 1$ is an arbitrarily small arbitrary constant. There is no method of encoding which gives an equivocation less than $H-C$.*

For the precise definitions of *Entropy* and *Capacity*, we refer to Section 2.

Remark 1.3. *Since the source is characterized by its information transmission rate (according to Shannon's definition of entropy), this theorem explains that the transmission of this information requires a channel with $C>H$. If we try to transmit a message through a channel of lower capacity, any excess of source entropy with respect to channel capacity will imply an increased rate of error for the receiver. Viceversa, a regime where $C\simeq H$ or $C\gg H$, translates to an increment of the complexity.*

Since then, a huge amount of literature has been developed. We refer to [40], [26], [9] and [33] for more details on Classical Information Theory. We refer to [32] , [11] for different uses of Information Theoretical tools in different areas of mathematics and statistics. In particular, the theory has enlarged to Quantum Information Theory, see for example [38], [37], [36], [27], [29], [30], [31], [23], [28] and [22] for a theoretical background and some connections to Fisher Information Inequalities. We specify that the main emphasis of our paper is on Classical more than Quantum Information Theory, even if our model example is the Schrödinger Equation.

A fundamental question in Information Theory is: how fast can we send data? Data rate depends on the bandwidth B, the level of the signal S and the level of the noise N.

For a noiseless channel, Nyquist's Theorem says that the theoretical maximum bit rate is given by

$$C=2\times B\times \log_2 L,$$

with L the number of signal levels used to represent data, and the bit rate C is measured bits/second.

The Shannon-Hartley's Theorem says that highest data rate for a noisy channel is given by

$$C = B \times \log_2 \left(1 + \frac{S}{N}\right).$$

We believe that the speed of data transmission plays the role of the speed of propagation in dispersive equations. This idea motivates our analogy and the dictionary.

The remaining part of the paper is organized as follows. In Section 2, we introduce some notation and give some preliminary results. In particular, we introduce Information Theoretic concepts, Strichartz Estimates and briefly treat the ODE case. In Section 2.2, we connect the terminologies of PDE and Information Theory, proposing a dictionary between the two fields and revisiting Shannon Code Sourcing Theorem and Strichartz Estimates from a common point of view. In particular, we introduce the Schrödinger Equation, Keel and Tao's Strichartz Estimates and explicitly illustrate the dictionary. In Section 4, we describe some possible further connections between PDEs and Shannon Information Theory, like the ones between maximizers of Entropy and Strichartz Norms, Time Recovery and the role of symmetries. We conclude with Section 5, in which we treat the discrete case, the noisy case, and give the example of the Kinetic Transport Equation.

2. Notation and preliminaries

In this section, we introduce the Informational Theoretic Concepts as presented in [40], we introduce Strichartz Estimates, as presented for example in [42] and [24], and, at the end, we briefly treat the ODE case.

2.1 Information theoretic concepts

The most important concepts employed by Shannon in [40] are the ones of *Entropy* and *Capacity*. In this subsection, we introduce the precise definitions of these two concepts.

Definition 2.1 (Entropy)**.** *Suppose that $X \in S := \{x_1, \dots, x_n\}$ is a discrete random variable with pmf $p(x) := P(X = x)$ for every $x \in S$ and $p(x) = 0$ otherwise. Then, the* Entropy $H(X)$ of the Discrete Random Variable X *is defined as follows*

$$H(X) := \mathbb{E}_X[I(x)] = -\sum_{x \in S} p(x) \log p(x). \tag{2.1}$$

Here, $I(x) := -\log p(x)$ is called Self-Information *and it is the entropy contribution of the individual message x. $\mathbb{E}_X$ is the expected value, taken with respect to $p(x)$.*

In an analogous manner, we define the Differential Entropy $H(X)$ of a Continuous Random Variable X *with support $x \in S \subset \mathbb{R}^n$ by:*

$$H(X) := \mathbb{E}_X[I(x)] = -\int_{x \in S} p(x) \log p(x) dx. \tag{2.2}$$

Again, $\mathbb{E}_X$ is the expected value, taken with respect to $p(x)$.

Remark 2.2. *A property of the discrete entropy is that it is maximized when all the messages in the message space are equi-probable $p(x) = 1/n$ (most unpredictable case), which gives $H(X) = \log n$.*

Definition 2.3 (Joint Entropy)**.** *The* Joint Entropy of two Discrete Random Variables $X \in S := \{x_1, \dots, x_n\}$ and $Y \in T := \{y_1, \dots, y_n\}$ *is the entropy of their pairing (X, Y):*

$$H(X,Y) := \mathbb{E}_{X,Y}[I(x,y)] = -\sum_{x,y \in S \times T} p(x,y) \log p(x,y). \tag{2.3}$$

Here, $I(x,y) := -\log p(x,y)$ is the Joint Self-Information*, which is the entropy contribution of the individual joint message (x,y). $\mathbb{E}_{X,Y}$ is the expected value, taken with respect to the joint pmf $p(x,y)$.*

In an analogous manner, we define the Joint Entropy $H(X,Y)$ of two Continuous Random Variables X and Y *with supports* $x \in S \subset \mathbb{R}^n$ *and with support* $y \in T \subset \mathbb{R}^n$ *by:*

$$H(X,Y) := \mathbb{E}_{X,Y}[I(x,y)] = -\int \int_{(x,y)\in S\times T} p(x,y) \log p(x,y) dxdy, \tag{2.4}$$

with $p(x,y)$ *the joint pdf. Again,* $\mathbb{E}_{X,Y}$ *is the expected value, taken with respect to the joint pdf* $p(x,y)$.

Remark 2.4. *Note that if X and Y are independent, then their joint entropy is the sum of their individual entropies.*

From now on, we will skip specify the support in order to enlighten the notation.

Definition 2.5 (Conditional Entropy)**.** *The* Conditional Entropy $H(X|Y)$ of a Discrete Random Variable X given random variable Y *is defined by:*

$$H(X|Y) := \mathbb{E}_Y[H(X|y)] = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y) \log p(x|y) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)}. \tag{2.5}$$

The Conditional Entropy $H(X|Y)$ of a Continuous Random Variable X given random variable Y *is defined by:*

$$H(X|Y) := \mathbb{E}[H(X|y)] = -\int \int_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} dxdy, \tag{2.6}$$

where

$$p(y) = \int_x p(x,y) dx.$$

Remark 2.6. *The* Conditional Entropy $H(X|Y)$ of a Random Variable X given random variable Y *is the average conditional entropy of X over Y. It is also called* Equivocation of X about Y.

Remark 2.7. *Consider for example the continuous case. Then*

$$H(Y|X) := \mathbb{E}_X[H(Y|x)] = -\int \int_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} dxdy, \tag{2.7}$$

where

$$p(x) = \int_y p(x,y) dy.$$

In some sense, the conditional entropy is "almost symmetric", because the only difference between $H(Y|X)$ *and* $H(X|Y)$ *is in the denominator of the log. The complete symmetry is recovered in the case when* $p(x) = p(y)$ *both in the dependent or independent case.*

Remark 2.8. *A basic property of the conditional entropy is that:*

$$H(X|Y) = H(X,Y) - H(Y).$$

This means that the information produced by X given Y is the same as the information jointly produced by X and Y minus the information produced by Y alone.

Definition 2.9. *[Mutual Information] The* Mutual Information of two Discrete Random Variables X and Y *is defined as:*

$$I(X;Y) := \mathbb{E}_{X,Y}[SI(X,Y)] = \sum_y \sum_x p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right). \tag{2.8}$$

In the above formula, SI is called Specific Mutual Information*. Here* $p(x,y)$*,* $p(x)$ *and* $p(y)$ *are defined as in the previous definitions.*
The Mutual Information of two Continuous Random Variables X and Y *is defined as:*

$$I(X;Y) := \mathbb{E}_{X,Y}[SI(X,Y)] = \int_y \int_x p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) dx\, dy. \tag{2.9}$$

Again, SI is called Specific Mutual Information *and* $p(x,y)$*,* $p(x)$ *and* $p(y)$ *are defined as in the previous definitions.*

Remark 2.10. *The* Mutual Information $I(X;Y)$ *measures the amount of information that can be obtained about one random variable by observing another. It is a measure of the mutual dependence of two random variables and determines how similar the joint distribution* $p(X,Y)$ *is to the product of the marginal distributions* $p(X)p(Y)$*. It is important in communication theory because it can be used to maximize the amount of information shared between sent and received signals.*

Lemma 2.11 (Properties of the Discrete Entropy)**.** *Suppose X and Y are discrete random variables. Then, the following properties hold:*

- $I(X;Y) = H(X) - H(X|Y)$.
- $I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X,Y)$.

Lemma 2.12 (Properties of the Continuous Entropy)**.** *Suppose X and Y are continuous random variables. Then, the following properties hold:*

- *If X is limited to a certain volume V, then* $H(X)$ *is a maximum and equal to* $\log V$ *when* $p(x)$ *is constant* $(p(x) = 1/V)$ *in* V.
- *We have*

$$H(X;Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

and

$$H(Y|X) \leq H(Y).$$

- *Let* $X \in \mathbb{R}$ *be a random variable. The pdf* $p(x)$ *giving maximum entropy subject to the condition that the standard deviation of X is fixed to be* σ *is the Gaussian Distribution. Similarly in n dimensions, subject to the constraint of Variance-Covariance Matrix to be* Σ*. The entropy of a one-dimensional Gaussian distribution whose standard deviation is* σ *is given by* $H(x) = \frac{1}{2}\log\left[2\pi e\sigma^2\right]$*, while the n-dimensional counterpart is* $H(X) = \frac{1}{2}\log\left[(2\pi e)^n \det(\Sigma)\right]$.

We give the definition of discrete channel.

Definition 2.13. *We define a* channel *to be a triplet* $\{X, p(y|x), Y\}$ *consisting of an input random variable X, an output random variable Y and a conditional probability distribution* $p(y|x)$ *specifying the probability that we observe the output* $Y = y$ *given that* $X = x$*. The channel is said to be* memoryless *if the output distribution depends only on the input distribution and is conditionally independent of previous channel inputs and outputs.*

From now on, we will always consider memoryless channels.

Definition 2.14 (Capacity of a Channel)**.** *Consider the memoryless channel* $\{X, p(y|x), Y\}$*, as in Definition 2.13. Let* $I(X;Y)$ *be the* Mutual Information of Y and X *of Definition 2.9. The* Channel Capacity *is defined as*

$$C = \sup_{p_X(x)} I(X;Y), \tag{2.10}$$

where the supremum is taken over all possible pdfs $p_X(x)$ *of the input variable X.*

Remark 2.15. *The conditional distribution function of Y given X,* $p_{Y|X}(y|x)$ *is an intrinsic property of the channel. A single choice of* $p_X(x)$ *determines the joint pdf* $p_{X,Y}(x,y)$ *and so the* Mutual Information $I(X;Y)$*. Basically,* $I(X;Y)$ *depends on the channel and on the choice of the distribution of the input. The capacity then depends just on* $p_{Y|X}(y|x)$ *and so it is an intrinsic property of the channel.*

Remark 2.16. *There is one important difference between the continuous and discrete entropies. In the discrete case, the entropy measures in an absolute way the randomness of the chance variable. In the continuous case, the measurement is relative to the coordinate system.*

2.2 Strichartz estimates

In this section, we introduce the *Strichartz Estimates*. These estimates are very important in the context of PDEs and Harmonic Analysis, because they provide useful information concerning the dispersive behaviour of solutions to PDEs. Among the other things, one can give a more general characterization of the *Gaussian Distribution* by maximizing Strichartz Norms. We first introduce some characteristic quantities called *Admissible Exponents*.

Definition 2.17. *Fix $n \geq 1$. We call a set of exponents (q,r)* admissible *if $2 \leq q,r \leq +\infty$ and*

$$\frac{2}{q}+\frac{n}{r}=\frac{n}{2}.$$

Remark 2.18. *These exponents are characteristic quantities of certain norms, called* Strichartz Norms*, naturally arising in the context of Dispersive Equations and can vary from an equation to another equation. We refer to [43] for more details.*

Here is the precise characterization of the Multivariate Normal Distribution, through Strichartz Estimates.

Theorem 2.19. *[42], [24], [7], [39] Suppose $n=1$ or $n=2$. Then, for every (q,r) and $(\tilde{q},\tilde{r})$ admissible and for every $u_0 \in L^2_x(\mathbb{R}^n)$ such that $||u_0||^2_{L^2(\mathbb{R}^n)}=1$, we have*

$$\left|\left|e^{-it\Delta}u_0\right|\right|_{L^q_t L^r_x} \leq S(n,q,r), \tag{2.11}$$

where $S_h(n,q,r)=S_h(n,r)$ is the Sharp Homogeneous Strichartz Constant*, defined by*

$$S_h(n,r) := \sup\left\{||u||_{L^q_t L^r_x(\mathbb{R}\times\mathbb{R}^n)} : ||u||^2_{L^2_x(\mathbb{R}^n)}=1\right\}, \tag{2.12}$$

and given by

$$S_h(n,r)=2^{\frac{n}{4}-\frac{n(r-2)}{2r}} r^{-\frac{n}{2r}}. \tag{2.13}$$

Moreover, the inequality (2.11) becomes an equality if and only if $|u_0|^2$ is the pdf of a Multivariate Normal Distribution*.*

Recall that

$$\|f\|_{L^2(\mathbb{R}^n)} := \left(\int_{\mathbb{R}^n} |f|^2 dx\right)^{1/2}$$

and

$$\|F\|_{L^q_t L^r_x(\mathbb{R}\times\mathbb{R}^n)} := \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}^n}|F(t,x)|^r dx\right)^{q/r} dt\right)^{1/q}.$$

Analogously, we can define l^p spaces of integrable sequences by substituting integration with summation.

Remark 2.20. *This characterization does not need the restriction of fixed variance as the one achieved using the* Entropy Functional *and so it is more general. The result is conjectured to be true for any dimension $n \geq 1$. See for example [39], where the optimal constant has been computed in any dimension $n \geq 1$, under the hypothesis that the maximizers are* Gaussians *also in dimension $n \geq 3$.*

2.3 The ODE case

Consider the following ODE:

$$\dot{x}(t)=Lx(t), \quad x(0)=x_0.$$

Here $x : I \subset \mathbb{R} \to V$ where I is an interval containing the origin $t=0$, V is a vector space and x_0 represents the initial position. The operator $L : V \to V$ is taken to be linear and determines the behaviour of the solution $x(t)=e^{tL}x_0$ with its spectral properties. We define the operator norm of L as

$$\|L\|_{op} := \inf\{c>0 : \|Lv\|_V \leq c\|v\|_V\}$$

and also the operator norm of e^{tL} as

$$\|e^{tL}\|_{op} := \inf\{c > 0 : \|e^{tL}v\|_V \leq c\|v\|_V\},$$

for any $t \in I$. Note that in our case, it is natural to take $x_0 = (x_0^1, \ldots, x_0^n)$ with $n = dim(V)$ such that $x_0^i > 0$ for $i = 1, \ldots, n$ and $\|x_0\|_V = \|x_0\|_{l^1} = \sum_{i=1}^n x_0^i$. We give here a restricted dictionary for ODEs:

Information Theory	Differential Equations
Source	time $t = 0$
Source Encoding	map $0 \mapsto x_0$
Transmitter	Initial Datum x_0
Channel	Propagator e^{tL}
Receiver	Solution $x(t) = e^{tL}x_0$
Decoder	map $x(t) \mapsto t$
Inference of the source	time t
Nyquist Bit Rate	C_{ODE}

Here, we used the notation: $C_{ODE} := \sup\{\|e^{tL}x_0\|_{L_t^q l_x^r} \text{ s.t } \|x_0\|_{l^1} = 1\}$. As we will explain later in the PDE case, the role of the source and the transmitter in the case when the equation is deterministic (noiseless case) as in all this paper, are basically identical. Since the ODEs we are considering are linear, there exists a unique solution for any initial datum. Therefore, there is a bijection between t and $x(t)$. In the case in which we add to the ODE a random component (noisy case), this bijection disappears. The *source* is the position $t = 0$, to which the *encoder* assigns the initial datum x_0. The encoded message x_0 is transmitted by the channel e^{tL} to the receiver $x(t) = e^{tL}x_0$ which can deduce the position t by uniqueness. A quantity which characterizes the speed of transmission of the channel is C_{ODE}.

Example 2.21. *Consider the simplest case of a noiseless linear system of differential equations in a vector space $V = \mathbb{R}^2$ and $F(u) = Lu$, for some linear $L : V \to V$, given by $L = -Id_{2\times 2}$. Take $I = [0, +\infty)$. The ODE is then $\frac{d}{dt}x(t) = -x(t)$. The channel "input" is the initial condition $x(0)$, the channel is the fundamental matrix e^{tL} and the channel "output" is the solution $x(t) = e^{tL}x(0) = e^{-t}x(0)$. The corresponding of the Nyquist Bit Rate in the dictionary is therefore:*

$$C_{ODE} = \sup_{x_0 : \|x_0\|_{l^1}=1} \|x_0\|_{l^r} \int_0^{+\infty} e^{-qt} = \frac{1}{q},$$

when one component of x_0 is zero. Note that $1/q$ corresponds to the "rate" parameter of the corresponding exponential distribution.

Remark 2.22. *Consider the following ODE:*

$$\dot{x}(t) = F(x(t)), \quad x(0) = x_0.$$

Here $x : I \subset \mathbb{R} \to V$ where I is an interval containing the origin $t = 0$, V is a vector space and x_0 represents the initial position. Suppose that the operator $F : V \to V$ is nonlinear. This problem is more complicated than the case where $F = L$. One important point is that the existence is not guaranteed anymore for every time $t \in \mathbb{R}$. For example take $V = \mathbb{R}$ and $F(u) = u^2$. The corresponding ODE admits solutions blowing up in finite time. Even in the cases where the solutions exist for all times, uniqueness is not guaranteed. Consider for example $V = \mathbb{R}$ and $F(u) = \sqrt{|u|}$. The corresponding ODE admits multiple solutions with initial datum $x_0 = 0$. We will give further comments in Section 3.

3. The Rosetta Stone

In this section, we make explicit our proposed *Rosetta Stone* between Information Theory and Partial Differential Equations. We start with a general theorem of Keel and Tao on Strichartz Estimates [24], which generalizes Theorem 2.19 and then restrict our attention to a toy model, the Schrödinger Equation. For the purpose of translation, we rephrase their Strichartz Estimates into Information Theoretical terminology, connect maximizers of Entropy with Maximizers of Strichartz Norms and discuss a possible role of symmetries into encoding.

3.1 Keel and Tao's theorem

Let (X, dx) be a measure space and H be a Hilbert space. Consider the Banach space of functions $f : X \to \mathbb{C}$ with the following norm bounded:

$$\|f\|_p := \|f\|_{L^p(X)} = \left(\int_X |f(x)|^p dx \right)^{\frac{1}{p}}.$$

Consider the family of operators indexed by t given by

$$U(t) : X \to L^2(X)$$

and satisfying the following estimates:

- (*Energy Estimate*) for all t and all $f \in H$ we have:

$$\|U(t)f\|_{L^2(X)} \leq S\|f\|_H; \tag{3.1}$$

For some $\sigma > 0$, one of the following decay estimate holds

- for all $t \neq s$ and all $g \in L^1(X)$:

$$\|U(s)U(t)^* g\|_\infty \leq C|t-s|^{-\sigma}\|g\|_1 \tag{3.2}$$

or

- for all t, s and all $g \in L^1(X)$:

$$\|U(s)U(t)^* g\|_\infty \leq C\left(1+|t-s|\right)^{-\sigma}\|g\|_1. \tag{3.3}$$

Whenever one of these last two estimates holds together with the *Energy Estimate*, we have the following.

Definition 3.1. *We say that the exponent pair (q, r) is* σ-admissible *if $q, r \geq 2$, $(q, r, \sigma) \neq (2, \infty, 1)$ and*

$$\frac{1}{q} + \frac{\sigma}{r} \leq \frac{\sigma}{2}.$$

Theorem 3.2. *[24] If $U(t)$ obeys the estimates (3.1) and one between (3.2) and (3.3), then the estimates*

- $$\|U(t)f\|_{L_t^q L_x^r} \leq S\|f\|_H \tag{3.4}$$

- $$\left\| \int ds U(s)^* F \right\|_H \leq \|F\|_{L_t^{\tilde{q}'} L_x^{\tilde{r}'}} \tag{3.5}$$

- $$\left\| \int_{s<t} ds U(t)U(s)^* F \|_{L_t^q L_x^r} \right\| \leq \|F\|_{L_t^{q'} L_x^{r'}} \tag{3.6}$$

hold for all sharp admissible pairs (q, r) and $(\tilde{q}, \tilde{r})$.

3.2 The case of the Schrödinger equation

Keel and Tao's Theorem is abstract and holds for very general propagators $U(t)$. In the following, we will concentrate on the case of the Schrödinger Equation

$$i\partial_t u(t,x) = \Delta u(t,x), \quad (t,x) \in (0,\infty) \times \mathbb{R}^n, \tag{3.7}$$

and give some further comments on other PDEs in later sections.

3.3 Conservation of mass and flow on the space of probability measures

It is well known that if $p_0(x) = |u_0|^2$ defines a probability distribution, then also $p_t(x) = |e^{it\Delta}u_0|^2$ defines a probability distribution. This is mainly a consequence of the property of $e^{it\Delta}$ of being a unitary operator.

Theorem 3.3. *Consider $\mathscr{P}(\mathbb{R}^n)$, the set of all probability distributions on $\mathbb{R}^n$ and $u : (0,\infty) \times \mathbb{R}^n \to \mathbb{C}$ a solution to (3.7). Then u induces a flow in the space of probability distributions.*

Remark 3.4. *This observation justifies our choice of the Schrödinger Equation as a toy model. In fact, not all PDEs possess the property of conserving the charge/mass/number of particles/etc. For example, both the heat and the wave equation do not possess this property.*

3.4 Fundamental solution for the linear Schrödinger equation using fourier transform

The solution of the Linear Schrödinger Equation

$$i\partial_t u(t,x) = \Delta u(t,x), \quad (t,x) \in (0,\infty) \times \mathbb{R}^n,$$

with initial datum $u_0(x) = e^{-|x|^2} \in \mathscr{S}(\mathbb{R}^n)$ (Schwartz class) is given by

$$u(t,x) = (1-4it)^{-n/2} e^{-\frac{|x|^2}{1-4it}}. \tag{3.8}$$

This solution induces the probability density function:

$$p(t,x) = \left(\frac{\pi}{2}\right)^{-\frac{n}{2}} |1+16t^2|^{-n/2} e^{-\frac{2|x|^2}{1+16t^2}}. \tag{3.9}$$

Remark 3.5. *Note that if the initial datum is Gaussian, the solution is Gaussian for every time $t \in \mathbb{R}$. This implies that, as we see in later sections, the estimation of parameters from the final solutions can be naturally done in a parametric way. See Subsection 4.2.*

3.5 The information theoretic perspective of Strichartz estimates

In this section, we restate the Strichartz Estimate with Information Theoretical terminology. We propose the following dictionary:

Information Theory	**Differential Equations**
Source	time $t = 0$
Source Encoding	map $0 \mapsto u_0$
Transmitter	Initial Datum u_0
Channel	Propagator e^{tL}
Channel Encoding	External Potential
Receiver	Solution $u(t) = e^{tL}u_0$
Decoder	map $u(t) \mapsto t$
Inference of the source	time t
Nyquist Bit Rate	Strichartz Constant
Entropy	Strichartz Norms
Maximizer of the Entropy	Maximizer of Strichartz Norms
Linear Channel	$F = L$-Linear PDE
Nonlinear Channel	F Nonlinear-Nonlinear PDE
Gaussian	Gaussian

The *source* is the position $t = 0$, to which the *encoder* assigns the initial datum u_0. The message u_0 is transmitted by the channel e^{tL} to the receiver $u(t) = e^{tL}u_0$ which can deduce the position t by uniqueness. Before reaching the receiver, the message might be modified by the presence of an external potential (channel encoding). A quantity which characterizes the dispersion of the initial datum is the entropy of the source/a space-time norm, like the Strichartz Norm. The supremum over all possible sources gives the best Strichartz Constant, which is an intrinsic characteristic of the propagator and measures the maximal dispersive ability of the channel. Source encoding attempts to compress the data from a source in order to transmit it more efficiently. The best efficient way is when u_0 is a Gaussian pdf.

Remark 3.6. *Channel encoding adds extra data bits to make the transmission of data more robust to disturbances present on the transmission channel. This is basically the role played by a confining potential which tends to stabilize the wave.*

Remark 3.7. *This dictionary will be adjusted for the noisy case. First of all, the role played by the Nyquist Bit Rate is substituted by the Shannon Capacity. Then, for deterministic PDEs for which there is uniqueness, the map $t \to u(t)$ is a bijection, if we fix initial datum u_0. Therefore, the decoding is basically an identification map. In the noisy case, this cannot be true, because what the receiver gets is the noisy solution that the decoder needs to extract, so there cannot be a one to one correspondence between the noisy solution and the time t. In the noisy case, therefore, the noisy solution plays the role of the receiver and the denoised solution plays the role of the inference of the source. We refer to Section 5 for more details on the noisy case.*

We are now ready to restate Theorem 2.19 about Strichartz Estimates, using an Informational Theoretic point of view.

Theorem 3.8. *Let $t = 0$ be a source, whose transmitted signal u_0 has "entropy" H given by*

$$H := \left|\left|e^{-it\Delta}u_0\right|\right|_{L_t^q L_x^r}. \tag{3.10}$$

Consider the Channel $U(t) := e^{-it\Delta}$, whose Nyquist Bit Rate is

$$C := S(n,q,r) = \sup\left\{||u||_{L_t^q L_x^r(\mathbb{R}\times\mathbb{R}^n)} : ||u||^2_{L_x^2(\mathbb{R}^n)} = 1\right\}.$$

Then, it is possible to encode $t = 0$ by u_0 in such a way that the message can be transmitted at an average rate of at most C. The maximum rate $C = 2^{n/4-n(r-2)/(2r)}r^{n/(2r)}$ is reached when u_0 is Gaussian and measures the maximal speed of transmission of the channel $U(t)$. It is not possible to transmit at an average rate greater than the Strichartz Constant.

Some explanations are in order. Since we are in the noiseless case and by the uniqueness of the solution of the PDE (in this paper we are considering just linear PDEs), once you know $u(t,x)$ at any time t, then you know the solution at any previous and subsequent time. Therefore, any measure of relative entropy must be zero $H(X|Y) = H(Y|X) = 0$. No extra entropy is added during the flow, since the flow is deterministic. So, a good measure of mutual information must give $I(X,Y) = H(X) = H(Y)$. A reasonable such measure is indeed given by the Strichartz norms:

$$I(X,Y) = \left|\left|e^{-it\Delta}u_0\right|\right|_{L_t^q L_x^r}.$$

Therefore, a measure of speed of transmission is given by

$$S(n,q,r) = \sup_{u_0 \in L^2(X)} I(X,Y).$$

The Sharp Strichartz Constant plays then the role of *Nyquist Bit Rate of the Channel*.

Remark 3.9. *Our parallel between Strichartz Norms and Entropy is also justified by the result in [3], where the authors prove a monotonicity property of the Strichartz Norms under the evolution of a certain quadratic heat flow that looks like the monotonicity property of the entropy. The entropy measures a level of uncertainty and it always increases. Similarly, the Strichartz Norms measure dispersion and they are always increasing under the heat flow.*

Remark 3.10. *The Strichartz norm quantifies the dispersion of the propagator and "translates" to the Nyquist Bit Rate. It is actually an intrinsic property of the propagator and measures the information rate at which the input can travel. The Schrödinger Strichartz constant is the optimal upper bound on the rate at which information can be transmitted over the Schrödinger channel.*

Remark 3.11. *The Gaussian maximizes both the Shannon Entropy and the Strichartz Norms, but, differently from the usual capacity measure, there is no need of power upper bounds and the Gaussian is an absolute maximizer of the mutual information.*

Remark 3.12. *In quantum mechanics, loss of information corresponds to the violation of the unitary property, which has to do with the conservation of probability. Under the flow of the equation that we are considering, unitarity is preserved and in fact, the measure of information that we are using is constant along the flow as we showed in Subsection 3.3.*

Remark 3.13. *The Gaussian is added to the dictionary to show a connection between maximizers, as we will see in next section. It shows, also, a kind of centrality that that distribution plays in several fields. It is a kind of "fix point" of the dictionary.*

Remark 3.14. *Since the equations treated in this paper are all linear, we have uniqueness for free. This simplifies the decoding procedure. In the case of non-uniqueness, the situation can become more involved. Another complication appears if the transmitter can send just some information about the initial datum and not all. In this case, the presence of the symmetry of the equation might play a role in the reconstruction of the signal. See Section 4.3.*

Remark 3.15. *It might be interesting to extend this notion to the nonlinear case. In that situation for a certain range of nonlinearities, the long time behaviour of the solution is still linear (see [43]) and so a similar dictionary seem to be suitable.*

4. Further connections between PDEs and Shannon information theory

In this section, we describe some possible further parallels between PDEs and Information Theory.

4.1 Maximizers of the entropy are maximizers of the Strichartz norm?

The *Principle of Maximum Entropy* states that, among distributions belonging to a particular class (e.g. fixed variance, supported on the half-line, etc...), you should select the distribution with the maximum entropy, because it is the most uninformative. By doing this, you minimize the possibility of adding extra bias and you follow the physical principle that many systems tend to stabilize towards maximal entropy configurations. The following is a well-known theorem of Boltzmann on maximizers of the Entropy.

Theorem 4.1 (Boltzmann's Theorem)**.** *Consider the following subset of* $\mathbb{R}^n$*:*

$$\mathscr{C} := \{g : \mathbb{R}^n \to \mathbb{R} \text{ such that } \mathrm{E}[g(x)] = c \text{ with } x \in S \subset \mathbb{R}^n\},$$

where $S \subset \mathbb{R}^n$ *is a closed subset,* $g(x) = (g_1(x), \dots, g_n(x))$ *and* $c = (c_1, \dots, c_n) \in \mathbb{R}^n$*. Suppose there exists* $g \in \mathscr{C}$*, such that* $supp(g) = S$ *and such that*

$$g \in argmax_{\mathscr{C}} H(X),$$

then

$$g(x) = c \exp\left(\sum_{j=1}^{n} \lambda_j g_j(x)\right) \quad \text{for all } x \in S$$

with c, λ_j *such that* $\int_S g(x) = 1$ *and such that* $\mathrm{E}[g(x)] = c$*. A viceversa holds.*

Remark 4.2. *A similar version of this theorem works also in the discrete case.*

We can list several cases in which maximizers of the entropy are known:

- $\mathbb{R}^n$ *case*: The Univariate $N(\mu, \sigma^2)$ or Multivariate $N(\mu, \Sigma)$ Normal Distribution has maximum entropy among all real-valued distributions with specified mean μ and standard deviation σ and Var-Cov matrix $|\Sigma|$ respectively. Therefore, the assumption of normality imposes the minimal prior structural constraint beyond these moments. We saw that the entropy of the Univariate Normal Distribution whose standard deviation is σ is given by $H(x) = \frac{1}{2}\log\left[2\sigma^2\pi e\right]$, while the entropy of the Multivariate Normal Distribution with fixed $|\Sigma|$ is $H(X) = \frac{1}{2}\log\left[(2\pi e)^n \det(\Sigma)\right]$
- *Interval case*: The uniform distribution on the interval $[a,b]$ is the maximum entropy distribution among all continuous distributions which are supported in the interval $[a,b]$. The entropy of the uniform distribution in the interval $[a,b]$ is given by $\log[b-a]$.
- *Circular case*: The von Mises distribution [34] has pdf given by

$$f(\theta; \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)}.$$

The function $I_0(\kappa)$ denotes the modified Bessel function of order 0, the angle $\theta \in [0, 2\pi]$ and μ and κ are the scale and concentration parameter, respectively. The entropy of the Von Mises distribution is given by

$$H = -\int_0^{2\pi} f(\theta;\mu,\kappa)\ln(f(\theta;\mu,\kappa))d\theta = \ln(2\pi I_0(\kappa)) - \kappa\frac{I_1(\kappa)}{I_0(\kappa)},$$

with the function $I_1(\kappa)$ denoting the modified Bessel function of order 1.

- *Half-line case*: If $X \in \mathbb{R}^+$, namely $p(x) = 0$ for $x \leq 0$, and the first moment of X is fixed to be a, $a = \int_0^\infty p(x)xdx$,, then the maximum entropy distribution is the exponential distribution with pdf $p(x) = \frac{1}{a}e^{-x/a}$ for $x > 0$ and $p(x) = 0$ otherwise. In this case, the entropy is equal to $\log[ea]$.

Remark 4.3. *Some classes of distributions do not contain a maximum entropy distribution. It is possible that a class contain distributions of arbitrarily large entropy (for example, the class of continuous distributions on $\mathbb{R}$ with mean 0 but arbitrary standard deviation), or that the Entropy is bounded above, but there is no distribution which attains the maximal entropy (for example, if you add too many constraints -See [9]-).*

In the case of $\mathbb{R}^n$, Strichartz Norms are maximized by Gaussians (conjectured in $n \geq 3$ and proved $n = 1, 2$, see [39]). As far as we know, no work has been done to describe the maximizers of the Strichartz Norms on other types of domains. We have the following question.

Question:

We are wondering if the connection between Strichartz Norms and Entropy that we are proposing here is reflected also in the maximizers of these norms on domains different from $\mathbb{R}^n$. More precisely. If we fix the domain Ω, is it true that the class of functions which maximize some Strichartz Norms on Ω maximize also the Entropy? For which admissible exponents is this true?

Differently from the Entropy, computing the Strichartz Norm of distribution functions, apart from few particular cases, like the Gaussian (see [39]) is not exactly simple. Just for the sake of illustration, we discuss the case of the pdf of a uniform random variable between -1 and 1. We take as initial datum u_0 the function $u_0 = \frac{1}{2}\chi_{[-1,1]}$. If we compute its Fourier Transform, we get:

$$\hat{u}(0,\xi) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} u_0(x)e^{ix\xi}dx = \frac{1}{2\sqrt{2\pi}}\int_{-1}^{1} e^{ix\xi}dx = \frac{1}{2\sqrt{2\pi}}\frac{e^{ix\xi}}{i\xi}\Big|_{-1}^{1} = \frac{1}{2\sqrt{2\pi}}\frac{e^{+i\xi} - e^{-i\xi}}{i\xi}.$$

Using the propagator (similarly to what we did in Section 3.4), we get

$$u(t,x) = \int_{\mathbb{R}} e^{i|\xi|^2 t + ix\xi}\frac{1}{2\sqrt{2\pi}}\frac{e^{-i\xi} - e^{i\xi}}{-i\xi}d\xi.$$

If now we take the space derivative of this function, we get:

$$\frac{\partial}{\partial x}u(t,x) = \int_{\mathbb{R}} e^{i|\xi|^2 t + ix\xi}\frac{1}{2\sqrt{2\pi}}\left(e^{i\xi} - e^{-i\xi}\right)d\xi$$

$$= \frac{1}{2\sqrt{2\pi}}\int_{\mathbb{R}} e^{i|\xi|^2 t + i(x+1)\xi}d\xi - \frac{1}{2\sqrt{2\pi}}\int_{\mathbb{R}} e^{i|\xi|^2 t + i(x-1)\xi}d\xi.$$

Each of the two terms represent the Schrödinger Evolution of a Delta Function, with center at $+1$ and -1 respectively. Therefore, we get:

$$\frac{\partial}{\partial x}u(t,x) = \frac{\pi^{\frac{1}{2}}}{4t^{\frac{1}{2}}}e^{-i\frac{t}{|t|}\frac{\pi}{4}}e^{-i\frac{(x+1)^2}{4t}} - \frac{\pi^{\frac{1}{2}}}{4t^{\frac{1}{2}}}e^{-i\frac{t}{|t|}\frac{\pi}{4}}e^{-i\frac{(x-1)^2}{4t}}.$$

Now, by integrating in x, we find:

$$u(t,x) \propto e^{-i\frac{t}{|t|}}\left(\Phi\left(\frac{x+1}{2}\left(\frac{i}{t}\right)^{\frac{1}{2}}\right) - \Phi\left(\frac{x-1}{2}\left(\frac{i}{t}\right)^{\frac{1}{2}}\right)\right).$$

Here $\Phi(x)$ is the cumulative distribution function of the Standard Normal. Now, we should compute the Strichartz Norm of this function $u(t,x)$, but since already this cannot be computed in closed form, neither any space time integral can be. So, it seems hard to verify that an optimal bound is attained by a certain distribution, just by direct computation.

4.2 Time recovery: estimation of the source

One of the main goals of the *Theory of Communication* is to transmit a message and be able to recover it entirely or at least with the highest possible precision. In this section, we rephrase this in the context of PDEs, using the dictionary.

Given n output observations, it is simple to estimate the time t at which the signal has been sent. Suppose the input is again a Gaussian distribution $u_0(x) = e^{-|x|^2}$ and so that $p_0 \propto e^{-2|x|^2}$. Then

$$u(t,x) = (1-4it)^{-n/2} e^{-\frac{|x|^2}{1-4it}} \tag{4.1}$$

and so

$$p(t,x) = \left(\frac{\pi}{2}\right)^{-\frac{n}{2}} |1+16t^2|^{-n/2} e^{-\frac{2|x|^2}{1+16t^2}}, \tag{4.2}$$

as we showed in Subsection 3.4. A random process X_t with this distribution is a Normal Random Variable $X_t \simeq \mathcal{N}(0,\sigma_t^2)$ with $\sigma_t^2 = \frac{1+16t^2}{4}$. This automatically implies that: $t^2 = \frac{4\sigma_t^2-1}{16}$. Therefore the MLE estimator of the time t^2 is a linear transformation of the MLE of the variance: $\hat{t^2}_{MLE} = \frac{4\hat{\sigma_t^2}_{MLE}-1}{16}$, with $\hat{\sigma_t^2}_{MLE} := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Remark 4.4. *The correspondence between a solution $u(t,x)$ of the Schrödinger equation and $p(t,x)$ is not a bijection. In fact, once you know $p(t,x)$, every function $e^{i\theta}u(t,x)$ with $\theta \in [0,2\pi)$ gives rise to the same $p(t,x)$. This does not prevent to recover the time t, since the parameter θ does not have any effect on the variance.*

Similarly, suppose that different subsets of the data are observed at different instants: independently, we observe X_i, $i = 1,\dots,n$ at t_X and we observe Y_j, $j = 1,\dots,m$ at t_Y. We can estimate the signed distance T between the emission times of the signals X and Y in the following way. We first estimate the time when X_i's have been emitted:

$$\hat{t^2}_{X\,MLE} = \frac{4\hat{\sigma^2}_{X\,MLE}-1}{16},$$

with

$$\hat{\sigma^2}_{X\,MLE} := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

and

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Then, we estimate the time at which the Y_j's have been emitted:

$$\hat{t^2}_{Y\,MLE} = \frac{4\hat{\sigma^2}_{Y\,MLE}-1}{16},$$

with

$$\hat{\sigma^2}_{Y\,MLE} := \frac{1}{m-1}\sum_{j=1}^m (Y_j - \bar{Y})^2$$

and

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^m Y_i.$$

Then,

$$\hat{T} = \hat{t_{Y\,MLE}} - \hat{t_{X\,MLE}} = \left(\frac{4\hat{\sigma^2}_{Y\,MLE}-1}{16}\right)^{1/2} - \left(\frac{4\hat{\sigma^2}_{Y\,MLE}-1}{16}\right)^{1/2}.$$

We can also estimate the period if the same signal is sent in a periodic way at instants t_k, $k = 1,\dots,N$. First, we can estimate for each k:

$$\hat{t^2_{k\,MLE}} = \frac{4\hat{\sigma^2_{t_k MLE}} - 1}{16},$$

with

$$\hat{\sigma^2_{t_k MLE}} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then,

$$\hat{T} := \frac{1}{N-1}\sum_{k=1}^{N-1}(t_{i+1} - t^i) = \frac{1}{N-1}(t_N - t_1).$$

Remark 4.5. *In the case that the signal is not Gaussian, using the Central Limit Theorem, we can prove that these estimators are asymptotic estimators, for $t \in [0,T]$ with $0 < T < +\infty$.*

4.3 Symmetries and encoding/decoding

The purpose of source encoding is to decrease the dimension of the source data. In the context of the Schrödinger Equation, the source coding can be seen as a way to organize the particles. For example in the case of Gaussian initial data, this means giving to the Variance-Covariance Matrix a particular structure. The main point here is that, due to its symmetries and since it is possible to transmit the message at the optimal rate, the source encoding for the Schrodinger Equation, based on domain, co-domain and structure invariance do not affect the optimal transmission rate. This is because the symmetries of the propagator have counterparts in the symmetries of the Strichartz Norms. In this section, we summarize these symmetries, to make explicit what we mean. As explained in [15] (see also [39] following [15]), Strichartz Estimates are invariant by the following set of symmetries.

Lemma 4.6. *[15] Let $\mathcal{G}$ be the group of transformations generated by:*

- *space-time translations: $u(t,x) \mapsto u(t+t_0, x+x_0)$, with $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$;*
- *parabolic dilations: $u(t,x) \mapsto u(\lambda^2 t, \lambda x)$, with $\lambda > 0$;*
- *change of scale: $u(t,x) \mapsto \mu u(t,x)$, with $\mu > 0$;*
- *space rotations: $u(t,x) \mapsto u(t,Rx)$, with $R \in SO(n)$;*
- *phase shifts: $u(t,x) \mapsto e^{i\theta}u(t,x)$, with $\theta \in \mathbb{R}$;*
- *Galilean transformations:*

$$u(t,x) \mapsto e^{\frac{i}{4}\left(|v|^2 t + 2v\cdot x\right)} u(t, x+tv),$$

with $v \in \mathbb{R}^n$.

Then, if u solves equation (3.7) and $g \in \mathcal{G}$, also $v = g \circ u$ solves equation (3.7). Moreover, the constants $S_h(n,q,r)$ are left unchanged by the action of $\mathcal{G}$.

Not all these symmetries leave invariant the set of probability distributions $\mathscr{P}(\mathbb{R}^n)$. Therefore, we need to reduce the set of symmetries in our treatment and, in particular, we need to combine the scaling and the parabolic dilations in order to have all the family inside the space of probability distributions $\mathscr{P}(\mathbb{R}^n)$.

Lemma 4.7. *Consider $u_{\mu,\lambda} = \mu u(\lambda^2 t, \lambda x)$ such that $u(t,x) \in \mathscr{P}(\mathbb{R}^n)$ maximizes (2.11), then $\mu = \lambda^{n/2}$.*

Proof.

$$1 = ||u_\lambda||^2_{L^2(\mathbb{R}^n)} = \mu^2 \int_{\mathbb{R}^n} |u(\lambda^2 t, \lambda x)|^2 dx = \mu^2 \lambda^{-n} ||u||^2_{L^2(\mathbb{R}^n)} = \mu^2 \lambda^{-n},$$

so $\mu = \lambda^{n/2}$. □

Remark 4.8. *We notice that some of the symmetries can be seen just at the level of the generator of the family u, but not by the family of probability distributions $p_t(x)$. For example the phase shifts $u(t,x) \mapsto e^{i\theta} u(t,x)$, with $\theta \in \mathbb{R}$ give rise to the same probability distribution function because $|e^{i\theta} u(t,x)|^2 = |u(t,x)|^2$ and, partially, the Galilean transformations $u(t,x) \mapsto e^{\frac{i}{4}\left(|v|^2 t + 2v\cdot x\right)} u(t, x+tv)$, with $v \in \mathbb{R}^n$ reduces to a space translation with $x_0 = vt$, since $\left| e^{\frac{i}{4}\left(|v|^2 t + 2v\cdot x\right)} u(t, x+tv) \right|^2 = |u(t,x+tv)|^2$. In some sense, the parameter θ can be seen as a* latent variable.

Therefore, we have the complete set of probability distributions induced by the generator $u(t,x)$.

Theorem 4.9. *Consider $p_t(x) = |u(t,x)|^2$ a probability distribution function generated by $u(t,x)$ (see Subsection 3.4). Let $\mathscr{S}$ be the group of transformations generated by:*

- *inertial-space translations and time translations: $p(t,x) \mapsto p(t+t_0, x+x_0+vt)$, with $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$;*
- *scaling-parabolic dilations: $u(t,x) \mapsto \lambda^n u(\lambda^2 t, \lambda x)$, with $\lambda > 0$;*
- *space rotations: $u(t,x) \mapsto u(t, Rx)$, with $R \in SO(n)$;*

Then, if u solves equation (3.7) and $g \in \mathscr{S}$, also $v = g \circ u$ solves equation (3.7), $q_t(x) = |v(t,x)|^2$ is still a probability distribution for every $g \in \mathscr{S}$ and the constant $S_h(n,q,r)$ is left unchanged by the action of $\mathscr{S}$.

Remark 4.10. *Optimality and symmetry are two strictly related concepts in the calculus of variations. Usually, extremizers possess some extra structure (see for example [6]). This might suggest to use symmetries/properties of the channel to optimize a code. Strichartz Inequalities, Soboloev Inequalities and several others possess radial extremizers (see again [6]).*

5. Some final remarks

In this section, we collect some further comments about the discrete case, the noisy case and the Kinetic Transport Equation.

5.1 The discrete case

Up to now, we mainly treated the case of a continuous channel and so continuous PDEs. One can think about extending this machinery to the discrete case. A possible approach would be to consider the Continuous Schrödinger Equation and use as initial data a sum of delta functions. This approach fails for the following reason. Consider for instance the case of a single delta function $u_0(x) = \delta(x)$. This gives rise to the solution:

$$u(t,x) = \left(\frac{1}{\pi t}\right)^{\frac{1}{2}} e^{i\frac{x^2}{t}}.$$

ans so to the pdf

$$|u(t,x)|^2 = \left(\frac{1}{\pi t}\right)^{\frac{1}{2}},$$

which is constant in space. So, every Strichartz Norm of this function would be infinity. Moreover, the estimates cannot work even locally in space because of the following reason. Consider a subset of $\mathbb{R}^n$ with volume V. Then,

$$\|u(t,x)\|_{L^q_t L^r_x} = \int_t \left(\frac{1}{\pi t}\right)^{\frac{q}{2}} V^{\frac{1}{r}} dt,$$

while

$$\|u(t,x)\|_{L^2_x} = \left(\frac{1}{\pi t}\right)^{\frac{1}{2}} V^{\frac{1}{2}}.$$

But an inequality of the type

$$\int_t \left(\frac{1}{\pi t}\right)^{\frac{q}{2}} V^{\frac{1}{r}} dt \leq C \left(\frac{1}{\pi t}\right)^{\frac{1}{2}} V^{\frac{1}{2}},$$

cannot hold for a constant C independent of time, because the time variable appears at different powers in the left and right hand side of the inequality. The main reason for all of this is that the mass is conserved just for L^2 solutions and $\delta(x)$ has a regularity lower than L^2.

A second approach is to consider the *Discrete Schrödinger Equation*:

$$i\partial_t u_k(t) + h^{-2}\left(u_{k+h}(t) + u_{k-h}(t) - 2u_n(t)\right) = 0,$$

with $k \in \mathbb{Z}$ and $u_0 \in l^2$. This equation resembles the continuous model, when the step size h is small $0 < h \ll 1$. Now, the domain is itself discrete and so we cannot do anything but choosing an initial datum defined on a discrete set. Moreover, the quantity which is now conserved for this equation is the $\|\cdot\|_{l^2}$, namely the discrete version of the $\|\cdot\|_{L^2}$ norm. Using the main theorem of [24], the authors in [41] have been able to prove Strichartz Estimates:

$$\|u_n(t)\|_{L_t^q l_x^r} \leq C\|u_n(0)\|_{l_x^2},$$

for $(q,r) \geq 2$ and

$$\frac{1}{q} + \frac{n}{3r} \leq \frac{n}{6}.$$

In this setting, we can use the Rosetta Stone and translate this discrete PDE in the context of Information Theory, in particular using the Discrete Schrödinger Equation as a toy model for *Discrete Channels*.

Remark 5.1. *For the* Discrete Schrödinger Equation*, the problem of finding the Sharp Strichartz Constant is still open, and the problem of finding the maximizer is open, as well. Possibly, the dictionary will be somehow helpful to answer this question.*

5.2 The noisy case

In this subsection, we consider the *Stochastic Linear Schrödinger Equation*

$$i\partial_t u(t,x) = \Delta u(t,x) \circ d\beta, \quad (t \geq s, x) \in (0,\infty) \times \mathbb{R}^n,$$

with initial datum $u(s,x) = u_s(x)$. This equation has an explicit solution (see [10] for details).

Proposition 5.2. *[10] For any $s \leq T$ and $u_s(x) \in \mathscr{S}'(\mathbb{R}^n)$, there exists a unique solution of the* Stochastic Linear Schrödinger Equation*, almost surely in $C([s,T];\mathscr{S}'(\mathbb{R}^n))$. Its Fourier Transform in space is given by*

$$\hat{u}(t,\xi) = e^{-i|\xi|^2(\beta(t)-\beta(s))}\hat{u}_s(\xi), \quad t \geq s, \quad \xi \in \mathbb{R}^n.$$

Moreover, if $u_s \in H_x^\sigma$ for some $\sigma \in \mathbb{R}$, then $u(\cdot) \in C([0,T];H_x^\sigma)$ a.s. and $\|u(t)\|_{H^\sigma} = \|u_s\|_{H^\sigma}$, a.s. for $t \geq s$. If $u_s \in L_x^1$, then the explicit solution takes the form:

$$u(t) = U(t,s)u_s := \frac{1}{(4\pi i(\beta(t)-\beta(s)))^{n/2}} \int_{\mathbb{R}^n} exp\left(\frac{i|x-y|^2}{(4(\beta(t)-\beta(s)))}\right) u_s(y)dy$$

for $t \in [s,T]$.

Strichartz Estimates have been proved also in the stochastic case.

Theorem 5.3. *Let $2 \leq r < +\infty$ and $2 \leq p \leq +\infty$ be such that $\frac{2}{r} > n\left(\frac{1}{2} - \frac{1}{p}\right)$ or $r = +\infty$ and $p = 2$. Let ρ be such that $r' \leq \rho \leq r$. Then, there exists a constant $c = c(\rho,r,p) > 0$ such that, for any $s \in \mathbb{R}$, $T \geq 0$ and $f \in L^\rho_{\mathscr{P}}(\Omega;L^{r'}(s,s+T;L_x^{p'}))$, the following estimate holds:*

$$\left|\int_s U(\cdot,\sigma)f(\sigma)d\sigma\right| \leq c(\rho,r,p)T^\beta |f|_{L^\rho_{\mathscr{P}}(\Omega;L^{r'}(s,s+T;L_x^{p'}))}$$

with $\beta = \frac{2}{r} - \frac{n}{2}\left(\frac{1}{2} - \frac{1}{p}\right)$.

Remark 5.4. *For the precise definition of the function spaces in the previous theorem we refer to [10].*

Concerning the optimal Strichartz Constant in the stochastic case $c(\rho, r, p)$ and the function which realizes it, as far as we know, there are no results. Nevertheless, we can propose a dictionary in the same flavour of the deterministic case.

Information Theory	**Differential Equations**
Source	u_0
Source Encoding	map $u_0 \mapsto u_0 + noise$
Transmitter	Noisy Initial Datum $u_0 + noise$
Channel	Propagator e^{tL}
Channel Encoding	External Potential
Receiver	Noisy Solution $u(t) + noise = e^{tL}(u_0 + noise)$
Decoder	map $u(t) + noise \mapsto u(t)$
Inference of the source	$E[u(t) + noise]$
Shannon Capacity	Strichartz Constant
Entropy	Strichartz Norms
Maximizer of the Entropy	Maximizer of Strichartz Norms
Linear Channel	$F = L$-Linear PDE
Nonlinear Channel	F Nonlinear-Nonlinear PDE
Gaussian	Gaussian

The main difference with respect to the deterministic case is that Nyquist Bit Rate is substituted by Shannon Capacity. Moreover, in the stochastic case, we do not have a one to one correspondence between the encoded signal and the time $t = 0$, as well as between what the decoder sees and the time t. For this reason, the dictionary cannot be simplified using the bijection between t and $u(t)$, as in the deterministic case.

Remark 5.5. *Each couple of admissible exponents give a different measure of "capacity", non necessarily equivalent. For this reason, there exist different notions of capacity that might not be equivalent.*

If we try to compute the mutual information for a joint pdf $p_\lambda(x,y) = \lambda^a p(\lambda^b x, \lambda^c y)$, using the constraints that also $p(x)$, $p(y)$, $p(x,y)$ need to be pdfs as well with the marginals of $p_\lambda(x,y)$, we conclude that I is invariant under this rescaling if and only if $a = n(b+c)$.

Consider a generalized version of the mutual information:

$$I_{\alpha,\beta,\gamma,r}(X;Y) = \int_y \int_x p(x,y)^r \log\left(\frac{p(x,y)^\alpha}{p(x)^\beta p(y)^\gamma}\right) dxdy. \tag{5.1}$$

The scale invariance implies $r = 1$ and so from now on $I_{\alpha,\beta,\gamma}(X;Y) := I_{\alpha,\beta,\gamma,r}(X;Y)$. The argument of the log needs to be scale free and so $\alpha(b+c) = (\beta b + \gamma c)$. This condition is trivially satisfied when $\alpha = \beta = \gamma$, which gives the original $I(X;Y)$ up to a constant scale. But for example, in the symmetric case $c = b$ and so $a = 2nb$, for $\alpha = \frac{\beta+\gamma}{2}$ we have the corresponding scale invariance of $I_{\alpha,\beta,\gamma}(X;Y)$. Note that fixed $q := \frac{b}{b+c}$, we have scale invariance whenever $\alpha = q\beta + (1-q)\gamma$. The choice of q is strictly related to the choice of the channel $p(y|x)$. This is reminiscent of the admissible exponents, which are strictly related to the linear propagator of the corresponding PDE.

It would be interesting to study more deeply the properties of $I_{\alpha,\beta,\gamma}(X;Y)$ in the context of Information Theory and see if the corresponding capacity

$$C_{\alpha,\beta,\gamma} := \sup_{p_X(x)} I_{\alpha,\beta,\gamma}(X;Y)$$

plays any important role.

5.3 The example of the kinetic transport equation

In this subsection, we give another example of our proposed relationship between Information Theory and PDEs. We consider the *Kinetic Transport Equation*:

$$\partial_t f(t,x,v) + v \cdot \nabla_x f(t,x,v) = 0, \quad f(0,x,v) = f^0(x,v)$$

for $(t,x,v) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$. This equation satisfies a similar set of Strichartz Estimates:

$$\|f\|_{L_t^q L_x^p L_v^r} \leq C \|f^0\|_{L^a_{x,v}}$$

for a set of admissible exponents;

$$\frac{2}{q} = n\left(\frac{1}{r} - \frac{1}{p}\right), \quad \frac{1}{a} = \frac{1}{2}\left(\frac{1}{r} + \frac{1}{p}\right), \quad q > a, \quad p \geq a.$$

The distribution function $f(t,x,v)$ is a non negative function $f(t,x,v) \geq 0$ depending on the time $t \in \mathbb{R}$, on the position $x \in \mathbb{R}^n$ and on the velocity $v \in \mathbb{R}^n$ and it is required to be integrable in x and v, $\iint_{x,v} f(x,v,)dxdv < +\infty$ for every $t \in \mathbb{R}$. From a physical point of view, the density f describes the evolution of the system of particles and $\int_C f(t,x,v)dxdv$ represents the probability of finding particles in the position-velocity space region C, at a fixed instant t. Furthermore, the *Kinetic Transport Equation* admits several conservation laws and in particular the conservation of the number of the particles (see for example [12]). These properties put ourselves in the same framework of the Schrödinger Equation and therefore a similar Rosetta Stone might be produced in the case of the Kinetic Transport Equation, as well.

Acknowledgements

I thank my family for their constant support. I thank also my supervisor Prof. Narayanaswamy Balakrishnan for his constant help and inspiring guidance. I thank the referees for their useful comments and the time spent reviewing my manuscript.

References

[1] S. Amari and H. Nagaoka, Methods of Information Geometry, *Translations of Mathematical Monographs* Vol. **191** Am. Math. Soc. (2000).

[2] C. Atkinson and A. F. S. Mitchell, Rao's Distance Measure, *Samkhyã-The Indian Journal of Statistics* **43** (1981) 345-365.

[3] J. Bennett, N. Bez, A. Carbery and D. Hundertmark, Heat-flow monotonicity of Strichartz norms, *Anal. PDE 2* no. **2** (2009) 147-158.

[4] P. Billingsley, Probability and Measure, (2012) Wiley.

[5] S. Bobkov, I. Gentil and M. Ledoux, Hypercontractivity of Hamilton-Jacobi equations, *J. Math. Pures Appl.* **80** (2001) 669-696.

[6] A. Burchard, A Short Course on Rearrangement Inequalities, `http://www.math.toronto.edu/almut/rearrange.pdf`

[7] M. Christ and Q. René, Gaussians rarely extremize adjoint Fourier restriction inequalities for paraboloids, *Proc. Amer. Math. Soc.* **142** no. 3 (2014) 887-896.

[8] K. Conrad, Probability Distributions and Maximum Entropy, `http://www.math.uconn.edu/~kconrad/blurbs/analysis/entropypost.pdf`

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (1991) *Wiley*.

[10] A. De Bouard and A. Debussche, The nonlinear Schrödinger equation with white noise dispersion, *J. Funct. Anal.* **259** no. 5 (2010) 1300-1321.

[11] A. Dembo, T. M. Cover and J. A. Thomas, Information Theoretic Inequalities, *IEEE Transactions on Information Theory* Vol. **37** no. 6 (1991) 1501-1518.

[12] J. Dolbeault, An introduction to kinetic equations: the Vlasov-Poisson system and the Boltzmann equation, *Discrete Contin. Dyn. Syst.* **8** no. 2 (2002) 361-380.

[13] S. Flego, A. Plastino and A. R. Plastino, Information Theory Consequences of the Scale-Invariance of Schrödinger's Equation, *Entropy* **13** (2011) 2049-2058.

[14] D. Foschi, Inhomogeneous Strichartz estimates, *J. Hyperbolic Differ. Equ.* **2** no. 1 (2005) 1-24.

[15] D. Foschi, Maximizers for the Strichartz inequality, *J. Eur. Math. Soc.* **9** no. 4 (2007) 739-774.

[16] B. R. Frieden, Fisher Information as the basis for the Schrödinger wave equation, *Am. J. Phys.* **57** (1989) 1004-1008.

[17] B. R. Frieden, *Physics from Fisher Information: A Unification*, 1st Ed. (1998) Cambridge University Press.

[18] B. R. FRIEDEN, *Science from Fisher Information: A Unification*, 2nd Ed. (2004) Cambridge University Press.

[19] B. R. FRIEDEN AND R. A. GATENBY, Principle of maximum Fisher information from Hardy's axioms applied to statistical systems, *Phys. Rev. E* **88**, 042144 (2013).

[20] B. R. FRIEDEN, A. PLASTINO, A. R. PLASTINO AND B. H. SOFFER, Schrödinger link between nonequilibrium thermodynamics and Fisher information, *Physical Review E* **66** 046128 (2002)

[21] P. GARBACZEWSKI, Differential entropy and time Review, *Entropy* **7** (2005) 253-299.

[22] P. GIBILISCO, F. HIAI AND D. PETZ, Quantum covariance, quantum Fisher information, and the uncertainty relations, *IEEE Trans. Inform. Theory* **55** no. 1 (2009) 439-443.

[23] M. HAYASHI, S. ISHIZAKA, A. KAWACHI, G. KIMURA, GEN AND T. OGAWA, *Introduction to quantum information science*, (2015) Springer.

[24] M. KEEL, T. TAO, Endpoint Strichartz estimates, *Amer. J. Math.* **120** no. 5 (1998) 955-980.

[25] M. KUNZE, On the existence of a maximizer for the Strichartz inequality, *Comm. Math. Phys.* **243** no. 1 (2003) 137-162.

[26] R. LANDAUER, Information is Physical, *Proc. Workshop on Physics and Computation PhysComp'92* IEEE Comp. Sci. Press (1993) 1-4.

[27] S. LUO, On Covariance and Quantum Fisher Information, *Theory Prob. Appl.* Vol. **53** No.2 (2009) 329-334.

[28] S. LUO, Quantum Fisher Information and Uncertainty Relations, *Lett. Math. Phys.* Volume **53** Issue 3 (2000) 243-251.

[29] S. LUO AND Q. ZHANG, On skew information, *IEEE Trans. Inform. Theory* **50** (2004) no. 8 1778-1782.

[30] S. LUO AND Q. ZHANG, Correction to "On skew information", *IEEE Trans. Inform. Theory* **51** no. 12 (2005) 4432.

[31] S. LUO AND Q. ZHANG, An informational characterization of Schrödinger's uncertainty relations, *J. Statist. Phys.* **114** no. 5-6 (2004) 1557-1576.

[32] E. LUTWAK, S. LV, D. YANG AND G. ZHANG, Extensions of Fisher information and Stam's inequality *IEEE Trans. Inform. Theory* **58** no. 3 (2012) 1319-1327.

[33] D. J. C. MACKAY, *Information Theory, Inference, and Learning Algorithms*, (2003) Cambridge University Press.

[34] K. MARDIA AND P. E. JUPP, *Directional Statistics*, (1999) Wiley.

[35] K. MARTON, An inequality for relative entropy and logarithmic Sobolev inequalities in Euclidean spaces, *Journal of Functional Analysis* **264** (2013) 34-61.

[36] H. NAGAOKA, On Fisher information of quantum statistical models, *Proc. 10th Symposium on Information Theory and Its Applications* (1987) 241-246 (in Japanese).

[37] D. PETZ, Covariance and Fisher information in quantum mechanics, *J. Phys. A* **35** no. 4 (2002) 929-939.

[38] D. PETZ AND C. GHINEA Introduction to quantum Fisher information, *Quantum probability and related topics* QP–PQ: Quantum Probab. White Noise Anal. **27** (2011) 261-281.

[39] A. SELVITELLA, Remarks on the sharp constant for the Schrodinger Strichartz estimate and applications, *Electronic Journal of Differential Equations* Vol. **2015** No. 270 (2015) 1-19.

[40] C. E. SHANNON, A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol. **27** 379-423/623-656 (July/October 1948).

[41] A. STEFANOV AND P. KEVREKIDIS, Asymptotic behaviour of small solutions for the discrete nonlinear Schrödinger and Klein-Gordon equations, *Nonlinearity* **18** no. 4 (2005) 1841-1857.

[42] R.S. STRICHARTZ, Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations, *Duke Math. J.* **44** no. 3 (1977) 705-714.

[43] T.TAO, *Nonlinear Dispersive Equations: Local and Global Analysis*, (2006) CBMS Number 106.

[44] F. OTTO AND C. VILLANI, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality, *J. Funct. Anal.* **173** (2000) 361-400.

[45] C. VILLANI, A short proof of the "concavity of entropy power", *IEEE Trans. Inform. Theory* **46** no.4 (2000) 1695-1696.

[46] A. WEHRL, General properties of entropy, *Rev. Mod. Phys.* **50** (1978) 221-260.

[47] E. WEINSTEIN, Gauge Theory and Inflation: Enlarging the Wu-Yang Dictionary to a unifying Rosetta Stone for Geometry in Application, `https://www.youtube.com/watch?v=h5gnATQMtPg`

[48] E. WEINSTEIN, Gauge Theory and Inflation: Enlarging the Wu-Yang Dictionary to a unifying Rosetta Stone for Geometry in Application, http://pirsa.org/pdf/files/7c58ac48-fe90-425d-96c0-42fedcde51b7.pdf

[49] T. T. WU AND C. N. YANG, Concept of nonintegrable phase factors and global formulation of gauge fields, *Phys. Rev. D* (3) **12** no. 12 (1975) 3845-3857.

[50] J. ZHOU, Derivatives in Mathematics and Physics,http://faculty.math.tsinghua.edu.cn/~jzhou/Connection04.pdf