Habilitation à diriger des recherches

Jonas KAHN

13 octobre 2016

# Remerciements

My first thanks are to the CNRS that gives me such comfortable working conditions.

My colleagues in Lille have given me a warm welcome after my PhD thesis. The workshop in stochastic geometry in particular has been a source of material and inspiration. Thanks all for the chocolate bars, and thank you, my office mate, for your warmth !

Also in Lille, I have spoken with biologists at IRI for several years. I still feel guilty about not having implemented the methods I had promised.

In Toulouse where I am now, I am deeply indebted to Pierre Weiss, who has never failed to help me refocus.

Finally, many thanks to my reviewers, who have accepted to read this late and clumsy memoir.

# Foreword

In this memoir, I only mention my works after my PhD thesis. Hence I will not reference any article on quantum statistics even if they were published after my defence.

The "habilitation à diriger des recherches" is usually a way to weave all the works of the former years into a narrative, to gather them into a single coherent tale, be it roundabout and unpredictable at times ; to summarise, in a global plan. There is none.

My work post-thesis stems from chance encounters. It might be just a problem I have heard of during a seminar, that I think I can solve. Or it might be a longer collaboration with someone whose subjects I like.

There were also many regrets during those years, unfulfilled projects, because of laziness or perfectionism. Ranging from a study on the effect of boarding on the success at competitive examination, to ideas on the Brownian map. The only such project I shall briefly mention in these notes is about data analysis and acquisition in FLIM-FRET, where I have never delivered on my promises... Those years have taught me to look for more collaborations so that stakes may be shared.

The first part of the manuscript dwells on two instances of my "mercenary" work, where I solve a problem and never look back : one is related to ancilla-driven quantum computing, the other to comparing Markov chains.

The second part gathers questions around imaging in science, and more generally spatially organised data. Thus, it has mainly a statistical flavour. I speak about fluorescence microscopy, and questions motivated by MRI. I have needed to correct the known (wrong) rates for estimating finite mixtures when working on microscopy. The MRI, on the other hand, has inspired investigations on compressed sensing with constraints, projection on measures spaces, and ultra-fast clustering algorithms that do not percolate.

The third part deals with stochastic geometry, in the widest sense. I view the geometry, and especially the metric structure, of any random object as a part of the field. The first work here consists in proving the existence of Gibbs measures for $T$-tessellations. The second shows that an improper Poisson line process has enough symmetries and hyperbolic behaviour to be a SIRSN as defined by Aldous, in any dimension.

I will start the memoir with short one-sentence summaries of each important result, with reference to the corresponding article.

# Table des matières

# Résultats

**Résultat 1**

La fidélité d'une porte quantique non-idéale dans le modèle de calcul quantique avec ancilla est bornée supérieurement par une fonction strictement décroissante de l'intrication des qubits concernés avec le reste du registre.

**Result 1**

The fidelity of an inaccurate quantum gate in ancilla-driven quantum computation is upper bounded by a decreasing function of the entanglement between the qubits that are acted on and the remainder of the register.

Morimae and J. Kahn (2010)

**Résultat 2**

Si $K$ et $L$ sont deux noyaux de Markov réversibles stochastiquement monotones sur un espace partiellement ordonné, avec la même distribution d'équilibre, et satisfont une *inégalité de comparaison* – une nouvelle relation d'ordre partiel – alors la chaîne de Markov associée à $K$ mélange plus vite que celle associée à $L$ *à tout instant* en variation totale, $L^2$ et séparation, entre autres, pour de bonnes conditions initiales.

**Result 2**

If $K$ and $L$ are two reversible stochastically monotone Markov kernels on a partially ordered space, and they satisfy a *comparison inequality* – a new partial order – then the Markov chain with kernel $K$ mixes faster than the Markov chain with kernel $L$ *at all times* in total variation and $L^2$ distances, and separation, among others.

J. Fill and J. Kahn (2013, Corollary 3.3)

**Résultat 3**

Parmi les chaînes de vie et de mort sur $\{0, \ldots, n\}$ qui convergent vers la distribution uniforme depuis l'état initial 0, celle uniforme – une chance sur deux d'aller à gauche ou à droite, ou de rester sur place aux extrémités – majorise *à tout instant* toutes les autres. Elle mélange donc plus vite dans de nombreux sens.

La seule exception classique est au sens du temps de mélange de Lovász-Winkler quand $n$ est impair, auquel cas la chaîne la plus rapide est aussi identifiée.

**Result 3**

The uniform birth-and-death chain – one chance in two to go right and left, or stay at endpoints – majorizes *at all times* any other birth-and-death chain on $\{0, \ldots, n\}$. Hence it mixes faster in many senses.

The only classical exception are Lovász-Winkler mixing times when $n$ is even, in which case the fastest chain is also given.

J. Fill and J. Kahn (2013, Theorems 4.3 and 6.5)

**Résultat 4**

Ajouter un pas à une chaîne de Markov ne ralentit pas le mélange, dans un certain nombre de cas particuliers.

**Result 4**

Extra updates do not delay mixing for several Markov chains.

J. Fill and J. Kahn (2013, Section 8)

**Résultat 5**

Le clustering récursif par plus proches voisins est extrêmement rapide et permet une réduction de dimension qui conserve le signal intéressant en IRM fonctionnelle.

**Result 5**

Recursive nearest neighbour clustering is extremely fast and reduces dimension while preserving the relevant signal in functional MRI.

Hoyos-Idrobo et al. (2016)

**Résultat 6**

Le test du maximum de vraisemblance est relativement efficace pour déterminer le nombre d'espèces en un pixel d'une image FLIM-FRET. La distance de transport permet de déterminer la similitude entre différents pixels.

**Result 6**

Maximum likelihood ratio test is comparatively efficient for finding the number of species in a pixel of a FLIM-FRET image. The transportation distance allows to measure similarity between pixels.

Heinrich, Jonas Kahn, et al. (2011) and Heinrich, Pisfil, et al. (2014)

**Résultat 7**

La vitesse minimax d'estimation d'une loi de mélange à $m$ composantes au plus, localement autour d'un mélange à $m_0$ composantes, est en $n^{-1/(4(m-m_0)+2)}$ sous des conditions de régularité et d'identifiabilité. Donc la vitesse globale est en $n^{-1/(4m-2)}$.

**Result 7**

The optimal local minimax rate of estimation of a finite mixture with at most $m$ components around a mixture with $m_0$ components is $n^{-1/(4(m-m_0)+2)}$, under sufficient regularity and identifiability conditions. Hence the global minimax rate of estimation of a finite mixture with at most $m$ components is $n^{-1/(4m-2)}$.

Heinrich and Jonas Kahn (2015, Theorems 3.2 et 3.3)

**Résultat 8**

Il existe des estimateurs qui convergent non uniformément à vitesse $n^{-1/2}$ vers toutes les lois de mélange fini.

**Result 8**

There are estimators that converge non-uniformly at rate $n^{-1/2}$ to all finite mixing distributions.

Heinrich and Jonas Kahn (2015, Theorem 3.5)

**Résultat 9**

Un voyageur de commerce reliant des points tirés proportionnellement à $\pi^{1-1/d}$ approche une densité $\pi$ par des courbes continues. Il est aussi possible d'approcher une probabilité $\pi$ par des tirages sous contraintes par un algorithme général de projection sur des mesures. Une application parmi d'autres est l'acquisition compressée sous contraintes, en visant une densité adaptée à la paire de bases acquisition/compression.

**Result 9**

A travelling salesman connecting points sampled according to $\pi^{1-1/d}$ converges to a density $\pi$ with continuous curves. A more general projection algorithm may be used to approach a probability $\pi$ by sampling under very general constraints. As a typical application, compressed sensing under acquisition constraints is considered, pairing acquisition/compression bases to determine the target $\pi$.

Chauffert, Ciuciu, Jonas Kahn, and P. Weiss (2014), Chauffert, Ciuciu, Jonas Kahn, and P. Weiss (2016), and Boyer et al. (2016)

**Résultat 10**

On ne peut construire que $o(a^k k^k)$ mosaïques en $T$ différentes sur $k$ droites données, pour tout $a$. Aussi les modifications Gibbsiennes de la mosaÏque en $T$ complètement aléatoires existent si l'énergie est bornée inférieurement par le nombre de segments.

**Result 10**

There are at most $o(a^k k^k)$ different $T$-tessellations on $k$ given lines, for any $a$. Hence Gibbsian modifications of the CRTT (Completely Random $T$-Tessellation) exist if the energy is bounded from below by the number of lines.

Jonas Kahn (2014)

**Résultat 11**

Les géodésiques du processus de droites de Poisson impropre génère un SIRSN en toute dimension.

**Result 11**

The geodesics of the improper Poisson line process are a SIRSN (scale-invariant random spatial network) in any dimension.

Jonas Kahn (2015)

# Presentation

## 1 Mercenary work

### 1.1 Ancilla

#### 1.1.1 Quantum computing basics

Quantum computing (Nielsen and Chuang, 2010) consists in using directly quantum objects and phenomena for computation. In some cases, the calculations can be much faster than with any classical computer.

One of the first ideas of possible uses for quantum computing has been simulating quantum systems (Feynman, 1982), which is likely exponentially slow on classical computers. Deutsch and Jozsa (1992) and Simon (1997) have devised quantum algorithms solving in polynomial time some problems with an oracle, where classical computers need exponential time under the same conditions.

But interest has really spiked with Shor's (1994) algorithm, that allows integer factorisation and discrete log calculation in polynomial time, and Grover's (1996) algorithm, that allows unstructured search in square root time.

There are several different models of quantum computing, that is sets of resources (the equivalent of bits in classical computing) and available operations (logic gates in classical), with which the computation must be run. Any physical realisation of the model "only" needs to implement these resources and operations. All those models are universal and equivalent, meaning that any algorithm written for one model may be translated up to a polynomial cost for use in another model, such as the quantum Turing machine (Deutsch, 1985).

The first model consists in a quantum gates circuit, where qubits are stored in a register and quantum gates may be applied directly on any qubit or qubit pair. Experimentally, it might be hard to maintain entanglement between all the qubits in the register.

More recently, one-way quantum computation (Raussendorf and Briegel, 2001) has been devised as another model. Here, the qubits are initially all entangled together in some *cluster state*, and the computation is carried out simply by measuring individual qubits. Entanglement propagates effects to the other qubits. This method clearly separates the preparation of the resource from the computation itself.

Ancilla-driven quantum computation (Anders et al., 2010) may be seen as

an intermediate model between the two. The qubits are stored in a register, as in a quantum circuit, but no gate is applied directly on any pair of qubits. Instead, each qubit may be entangled with an ancilla, and operating on the ancilla acts on the qubit in the register.

More formally, a quantum object is associated with a complex Hilbert space $\mathbb{H}$, and a pure state – we shall only use pure states – is a norm-one element on the Hilbert space. Any element of this space will be denoted with the ket notation $|\cdot\rangle$, and the adjoint linear form will be denoted with the bra notation $\langle\cdot|$.

For a qubit, the Hilbert space is $\mathbb{C}^2$. We write $|0\rangle$ and $|1\rangle$ for two vectors of a fixed orthonormal basis in $\mathbb{C}^2$. We call it the *computational basis*. Another orthonormal basis is defined as:

$$|+\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right),$$
$$|+\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right)$$

A quantum state may evolve unitarily, that is $|\psi\rangle \mapsto U|\psi\rangle$ with $U$ a unitary operator on $\mathbb{H}$. Alternatively, we may measure it and project it on an orthonormal basis of $\mathbb{H}$. If $\{|\phi_i\rangle\}_{1\leq i\leq d(\mathbb{H})}$ is such a basis, the quantum state $|\psi\rangle$ becomes $|\phi_i\rangle$ with probability $\|\langle\psi_i|\phi\rangle\|^2$.

We will use the following two unitary evolutions:

— the *Hadamard gate* $\hat{H}$ acts on a qubit via $\hat{H}|0\rangle = |+\rangle$ and $\hat{H}|1\rangle = |-\rangle$.

— the *Controlled-Z gate* (CZ) acts on a qubit pair via $\widehat{CZ} = |00\rangle\langle 00| + |01\rangle\langle 01| + |10\rangle\langle 10| - |11\rangle\langle 11|$.

We show on which qubit we act by adding indices to the operator. For example, $\hat{H}_A$ would be a Hadamard gate applied to the ancilla $A$.

Entanglement between two or several quantum objects mean that generally a system consisting of several quantum objects should be seen as a single big object. It is not enough to know each of the small objects to know the state of the system. This is because the Hilbert space associated to the big object is the tensor product of those of the small objects. Yet, many norm-1 vectors cannot be written as a direct product of states on the small objects. A typical example of a pair of entangled qubits is:

$$\frac{1}{\sqrt{2}}\left(|0\rangle|0\rangle + |1\rangle|1\rangle\right). \tag{1}$$

We may quantify entanglement through the *reduced density matrix* $\rho_A = \mathrm{Tr}_B\left(|\phi\rangle_{A\otimes B}\langle\phi|_{A\otimes B}\right)$, where $|\phi\rangle_{A\otimes B}$ is the quantum state of the two possibly

entangled subsystems $A$ and $B$, and where $\mathrm{Tr}_B$ is the partial trace on $B$. The reduced density matrix is a nonnegative trace-1 matrix. If there is no entanglement, that is if $|\phi\rangle_{A\otimes B} = |\phi\rangle_A \otimes |\phi\rangle_B$, then $\rho_A = |\phi\rangle_A \langle\phi|_A$. Maximally entangled states such as (1) yield $\rho_A = \frac{1}{\dim A} 1_A$. More generally we may evaluate entanglement between a qubit and another system as, on a scale from 0 to 1:

$$S = 2\left(1 - \mathrm{Tr}(\rho_A^2)\right). \tag{2}$$

### 1.1.2 Ancilla-driven quantum computation

The ancilla-driven quantum computation model consists of:

— a register with $N$ qubits.

— an ancilla of one qubit.

— being able to apply a Hadamard gate to any qubit in the register.

— being able to entangle any qubit $R$ in the register with the ancilla $A$ through the operator $\hat{E} = \hat{H}_A \hat{H}_R \widehat{CZ}_{AR}$.

— being able to measure the ancilla in all bases, or equivalently, being able to rotate the ancilla at will and to measure it in the computational basis.

Indeed, implementing all unitary transformations on a qubit, plus all entanglement operations on pair of qubits, yield a universal quantum computer (Anders et al., 2010). Ancilla-driven quantum computation allow both operations, as illustrated in Figure 1.

### 1.1.3 Entanglement and fidelity of inaccurate quantum gates

In practice, rotations, measurements and gates $E$ are never absolutely exactly implemented. How do the errors on the gates propagate to the states?

With Tomoyuki Morimae, we have studied the case when measurements (applied at stages (c) and (g) in Figure 1) are inaccurate: the projection axis is rotated by $\varepsilon$, that is we get the answer 1 for $\cos(\varepsilon/2)\,|1\rangle + \sin(\varepsilon/2)\,|0\rangle$ if we measure in the computational basis.

Let $F$ be the mean quantum fidelity of the gate, defined as $F = \mathbb{E}\left[|\langle\phi|\psi\rangle|^2\right]$, where $|\phi\rangle$ is the expected state in the register, and $|\psi\rangle$ the random state we really get. An ideal gate has fidelity 1, and fidelity is nonnegative.

This mean fidelity may depend on the initial state of the register. For a given $\varepsilon$, it may even be 1 if we are lucky. But there is no lucky case if the qubit on which we act is entangled with the remainder of the register:

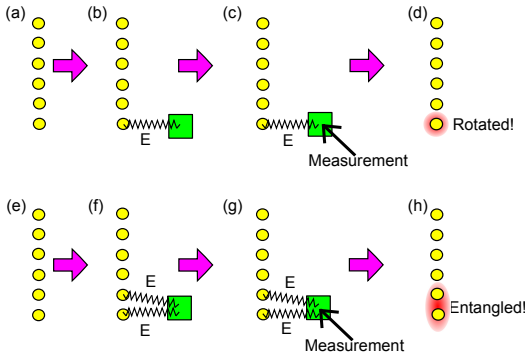



Figure 1 – Yellow circles are register qubits. Top line: a single-qubit rotation. (b) The ancilla (green square) is coupled to the qubit we want to rotate (say, the bottommost one) through the interaction $E$, which is represented by the black zigzag line. (c) After the interaction, the ancilla is projected onto a certain direction (represented by the solid black arrow). (d) The measurement backaction rotates the bottommost qubit of the register by the desired angle. Bottom line: a two-qubit entangling gate. (f) The ancilla (green square) is coupled to two qubits (say, the two bottommost ones) through the interaction $E$, which is *the same* as the interaction used in the top line. The interaction is represented by the black zigzag line. (g) After the interaction, the ancilla is projected (solid black arrow). (h) The measurement backaction causes the desired entangling gate between the two bottommost qubits of the register.

**Theorem 1.1.** *(Morimae and J. Kahn, 2010) With the above notations, a measurement inaccuracy $\varepsilon$ and with $S$ the entanglement* (2) *between one qubit and the remainder of the register, the fidelity of both the rotation of that qubit, and of its entanglement operation, is bounded by*

$$F \leq 1 - S \sin^2 \frac{\varepsilon}{2}.$$

Since quantum computing requires entanglement between qubits in the register to be efficient, this proves that inaccuracies in the measurement operators will have consequences. However, they may be mitigated in practice through the use of quantum error codes.

Let us mention that we have similar results for the other to implement ancilla-driven quantum computation with only one operation, that is the $\widehat{CZ + SWAP}$ gate instead of $\widehat{CZ}$. The relevant way to quantify entanglement is slightly harder in that case, and requires to consider entanglement between pairs and the remainder of the register.

## 1.2 Comparison inequalities for Markov chains

### 1.2.1 How to quantify convergence speed

Markov chains are random walks on a finite space with each step depending only on the present position. They are flexible, so they are often used as models in science. They are easy to analyse, so benefit from a large theory. They are easy to implement on a computer, and efficient for many problems.

A Markov chain is defined through a finite state space $\mathcal{X}$ and a Markov kernel $K$ on $\mathcal{X}$, also called *transition matrix*. A Markov kernel is a matrix with indices in $\mathcal{X}^2$, with nonnegative entries, and such that $\sum_{y \in \mathcal{X}} K(x, y) = 1$ for all $x \in \mathcal{X}$. The $K(x, y)$ entry is the probability that, if the Markov chain is on $x$, it will be on $y$ after the next step.

Unless otherwise specified, we assume the Markov chains are homogeneous, meaning that the kernel $K$ does not depend on time. Thus, if $\sigma$ is the initial probability distribution on $\mathcal{X}$, the probability distribution after $t$ steps is $\sigma K^t$. In a few cases, however, we shall work with inhomogeneous Markov chains, with kernel $K_t$ depending on time. We will always write $x_t$ for the random value the Markov chain attains at time $t$.

We may depict a Markov chain as a graph whose vertices are the elements of $\mathcal{X}$, and with directed edges from $x$ to $y$ with weight $K(x, y)$ if and only if it is nonzero. If the graph is connected and aperiodic, the Markov kernel has a unique stationary probability distribution $\pi$, that is $\pi K = \pi$. Moreover, the corresponding Markov chain converges to $\pi$ for all initial distribution.

The Markov kernel acts on the left on functions on $\mathcal{X}$ in $L^2(\pi)$. The reason why the $L^2$ norm is chosen with respect to $\pi$ is so that the adjoint $K^*$ becomes the time reversal for the Markov chain. In particular, a Markov chain with self-adjoint kernel is *reversible*. Locally, $\pi(x)K(x,y) = \pi(y)K(y,x)$.

A natural question is to quantify the speed of the convergence to $\pi$.

The usual way relies on the eigenvalues of $K$. The vector $(1,\dots,1)$ is a right eigenvector with eigenvalue 1. Perron-Frobenius ensures that all other eigenvalues have modulus at most one. If the Markov chain is aperiodic, these modulus are less than one. We write $1 = \beta_1 > |\beta_2| \geq |\beta_3| \geq \cdots \geq \left|\beta_{|\mathcal{X}|}\right|$ for the eigenvalues, and $V_i$ for the corresponding left eigenvectors. We may now write $\sigma = c_i(\sigma)V_i$ in this basis. So that $\sigma K^t = \pi + \sum_{i=2}^{|\mathcal{X}|} V_i c_i(\sigma)\beta_i^t$. Hence, for any reasonable metric, such as the total variation norm, we get $d(\sigma K^t, \pi) = O(\beta_2^t)$. The value $1 - |\beta_2|$ is the *spectral gap*. The higher the spectral gap, the faster the chain converges, asymptotically.

The main weakness of this estimate is that it is asymptotic. One of the main uses of estimating convergence speeds is knowing when to stop a Monte Carlo Markov chain (MCMC) sampling (Hastings, 1970): we follow the Markov chain for a while, and the final value should approximate sampling directly according to the stationary distribution $\pi$. Alternatively, we might want to estimate the expectation $\pi f$ of a function $f$ with respect to $\pi$, and we use $\frac{1}{T}\sum_{t=1}^{T} f(x_t)$ for a big enough $T$. But what does "big enough" mean? The asymptotic speed of convergence is no help here. A naive non-asymptotic form of the eigenvalue bound yields a multiplicative constant $|\mathcal{X}|$, that is, the size of the state space, which is usually very pessimistic. The more precise bounds require full knowledge of the spectrum and eigenvectors (Diaconis, 1988, for example).

The cutoff phenomenon Diaconis (1996) is a clear example of the pessimism of the eigenvalue bound. Quite often, especially for very symmetrical Markov chains, the total variation distance between the stationary distribution and the distribution after $t$ steps will be almost 1 until some $T$, after which it decreases exponentially and is close to 0 after $T + o(T)$ steps. In other words, the Markov chain attains the asymptotic regime at time $T$, and does not mix beforehand. Knowing this $T$ becomes the relevant practical question, rather than knowing the rate of exponential decay. That is the meaning of the famous result that seven riffle-shuffles are needed to shuffle 52 cards.

Joulin, Ollivier, et al. (2010), for example, give bounds for any time under hypotheses on the discrete Ricci curvature of the chain. These conditions can be checked for locally, hence easily. However, they constrain strongly the Markov chain.

To give a precise estimate of the mixing speed, we need a reference di-

vergence. Most often it will be the total variation distance $d_{TV}(\sigma, \pi) = \sup_{A \in \mathcal{B}(\mathcal{X})} \pi(A) - \sigma(A)$. Two common other choices would be $L^2(\pi)$, which works well with spectral methods, and the separation $\sup 1 - \frac{\sigma_i}{\pi_i}$, which is linked to strong stationary times (Diaconis and J. A. Fill, 1990).

When $\pi$ is uniform, we may use majorization (Marshall and Olkin, 1979) to deal simultaneously with all these divergences, and the some. If $v$ and $w$ are two sequences of $N$ numbers with the same sum, we say that $v$ *majorizes* $w$ if the sum of the $k$ largest numbers in $v$ is bigger than the sum of the $k$ largest numbers in $w$, for all $1 \leq k \leq N$. Indeed, many divergences between the uniform $\pi$ and $\sigma$ are Schur-convex functions of $\sigma$, meaning that if $\rho_1$ majorizes $\rho_2$, then $d(\pi, \rho_1) \geq d(\pi, \rho_2)$. These divergences include all the $L^p$ for $p \geq 1$, the separation, Hellinger distance, Kullback divergence $K(\sigma, \pi)$ and Kullback divergence $K(\pi, \sigma)$.

Lovász and Winkler (1995) have defined a very different measurement of mixing speed, motivated by the use of MCMC in sampling; namely, a stopping time. Specifically, let us consider all stopping times $T$ such that $X_T$ has distribution $\pi$. It can be shown that the expectation of $T$ is minimal if and only if there is a halting state $x$, that is a state such that if $X_t = x$, then $T = t$. This minimal expectation $\mathbb{E}[T] = T_{mix}$ is the *Lovász-Winkler* mixing time. There are several constructions of optimal stopping times, and there are practical ways to approach $\pi$ without knowing the whole Markov chain.

J. Fill and J. Kahn (2013) have designed a criterion showing that a Markov chain mixes faster than another with the very strong meaning that, *for any finite time*, the distance from $\pi$ to the distribution of the quick chain is lower than the distance to the distribution of the slow chain. We require some monotony properties.

### 1.2.2 Basic properties of comparison inequalities

From now on, the space $\mathcal{X}$ is endowed with a partial order. All scalar products are with respect to the stationary distribution $\pi$. Moreover, "$Y$ mixes faster than $Z$ for $d$" has the strong meaning: for all times $t$, $d(\pi, Y_t) \leq d(\pi, Z_t)$.

Let $\mathcal{K}$, $\mathcal{M}$ and $\mathcal{F}$ be respectively (i) the set of all Markov kernels on $\mathcal{X}$ with stationary distribution $\pi$, (ii) the set of nonnegative monotone functions on $\mathcal{X}$, (iii) the set of stochastically monotone kernels in $\mathcal{K}$, that is such that for any $f \in \mathcal{M}$, we still have $Kf \in \mathcal{M}$. Equivalently, if $x \leq y$, then the probability distribution of the Markov chain after one step starting at $y$ *stochastically dominates* that starting at $x$: $K(x, D) \geq K(y, D)$ for all down-sets $D$, that is such that if $z \leq w$ and $w \in D$, then $z \in D$.

*Comparison inequalities* are the partial order on $\mathcal{K}$ defined as $K \preceq L$ if $\langle Kf|g\rangle \leq \langle Lf|g\rangle$ for all $f, g \in \mathcal{M}$. Comparison inequalities are quite stable:

**Proposition 1.2** (J. Fill and J. Kahn, 2013, Propositions 2.3 et 2.9)**.**

1. *If $K \preceq L$, then $K^* \preceq L^*$.*
2. *If $K_n \preceq L_n$ for all $n$ and $K_n \to K$ and $L_n \to L$, then $K \preceq L$.*
3. *If $K_0 \preceq L_0$ et $K_1 \preceq L_1$, then for all $0 \leq \lambda \leq 1$:*
$$\lambda K_0 + (1-\lambda)K_1 \preceq \lambda L_0 + (1-\lambda)L_1.$$
4. *If $\mathcal{X}_0 \cup \mathcal{X}_1$ is a partition of $\mathcal{X}$ with the induced partial orders and stationary distributions, and if $K_i \preceq L_i$ on $\mathcal{X}_i$ for $i = 0, 1$, then if $K$ (resp. $Y$) is the direct sum of $K_0$ and $K_1$ (resp. $L_0$ and $L_1$), we have $K \preceq L$.*
5. *If $K_1, \ldots, K_t$ and $L_1, \ldots, L_t$, and their adjoints are all in $\mathcal{F}$, and if $K_i \preceq L_i$ for all $1 \leq i \leq t$, then $K_1 \ldots K_t \preceq L_1 \ldots L_t$ and both kernels are in $\mathcal{F}$.*

I have mentioned the last property to show that comparison inequalities can still be useful without reversibility. In a number of cases, we need that the time-reversed Markov chain be monotone, rather than the Markov chain itself. However reversible chains are easier to work with, and we state the following theorems under this hypothesis.

### 1.2.3 General speed comparisons

Comparison inequalities easily entail that if $Y$ has kernel $K$ and $Z$ has kernel $L$, both reversible in $\mathcal{F}$, with $K \preceq L$ and a shared initial distribution $\hat{\pi}$ such that $\hat{\pi}/\pi$ is nonincreasing, then $Y$ dominates $Z$. Indeed, the indicator of a down-set $D$ is a nonincreasing function, so that

$$\mathbb{P}[Y_t \in D] = \langle K^t \mathbf{1}_D | \frac{\hat{\pi}}{\pi} \rangle \leq \langle L^t \mathbf{1}_D | \frac{\hat{\pi}}{\pi} \rangle = \mathbb{P}[Z_t \in D].$$

We can notice that $\mathbb{P}[Y_t = \cdot]/\pi(\cdot)$ is nonincreasing for all $t$ under these conditions. Domination then implies that $\mathbb{P}[Z_t = \cdot]/\pi(\cdot)$ majorizes $\mathbb{P}[Y_t = \cdot]/\pi(\cdot)$. So that:

**Theorem 1.3** (J. Fill and J. Kahn, 2013, Corollaires 3.3 et 3.7)**.**

*If $K$ and $L$ are reversible, in $\mathcal{F}$, and $K \preceq L$, and if $\hat{\pi}/\pi$ is nonincreasing, then the Markov chain $Y$ with kernel $K$ mixes faster than the Markov chain $Z$ with kernel $L$, in total variation, in separation and in $L^2$. If moreover, $\pi$ is uniform, then $Y$ mixes faster than $Z$ in all $L^p$ with $p \geq 1$, in Hellinger distance and in Kullback divergence.*

A trick allows to get the latter result for $L^2$ with merely $K^2 \preceq L^2$ and both $K^2$ and $L^2$ stochastically monotone, rather than $K$ and $L$.

### 1.2.4 Adding steps do not delay mixing

(Peres and Winkler, 2013) have wondered whether adding steps could delay the mixing of a Markov chain. Intuitively, the answer should be negative, and they have proved it for the special case of monotone spin systems with total variation distance. On the other hand, Holroyd (2011) has provided counter-examples to the general case.

Adding means comparing an inhomogeneous Markov chain to another where some steps $K_t$ have been replaced with the identity kernel $I$, which translates as not moving.

Proposition 1.2 shows that comparison inequalities behave well with inhomogeneous Markov chains. Notice that skipping a step, that is the identity kernel $I$, is always in $\mathcal{F}$. Thus, it is enough to show that $I \preceq K_t$ and that these $K_t$ are all reversible and in $\mathcal{F}$. Under theses conditions, for good initial conditions, Theorem 1.3 immediately yields that extra steps can only speed up mixing in total variation, $L^2$ and separation.

We may carry out this strategy in the following cases, the last of which generalizes Peres and Winkler (2013):

**Proposition 1.4** (J. Fill and J. Kahn, 2013, Theorems 8.3, 8.5 and 8.6)**.**
*Adding extra steps do not delay mixing if the initial distribution $\hat{\pi}$ is such that $\hat{\pi}/\pi$ is nonincreasing and either:*

- *The state space $\mathcal{X}$ is totally ordered and $\pi$ is uniform.*
- *The state space $\mathcal{X}$ is the set of permutations with Bruhat order, and the steps are chosen among the following $K_i$: we sort the adjacent $i$ and $i+1$ with probability $p$, and anti-sort them with probability $1-p$.*
- *The state space $\mathcal{X}$ is a set of spin configurations: on each vertex of a graph $(V, E)$, there is an element of the partially ordered set $S$. The order on $\mathcal{X}$ is the product of the orders on each vertex. Moreover, we require the equilibrium distribution $\pi$ to be monotone, meaning that the distribution of the state of a vertex conditionally on the state of all other vertices is monotone in the state of all other vertices. We also require that those conditional laws $\pi_v$ satisfy $\langle f|g\rangle_{\pi_v} \geq \langle f|1\rangle_{\pi_v}\langle g|1\rangle_{\pi_v}$ for all $f, g \in \mathcal{M}$. Finally the steps have the following form: choose a vertex $v$ and set its state according to $\pi_v$.*

### 1.2.5 Birth-and-death chains

A birth-and-death chain is a Markov chain on $\mathcal{X} = \{0, 1, \ldots, n\}$ where on each step, either we do not move, or we move to an adjacent integer. All chains in this section start from zero, and are reversible.

The stationary distribution $\pi$ is uniform if and only if $K$ is symmetric. In such a case, $K^2$ is always monotone, even if $K$ is not.

We call *uniform chain* the symmetric Markokv chain with kernel $U$ given by $U(i, i+1) = 1/2$ and $U(0,0) = U(n,n) = 1/2$. A short calculation show that $U^2 \preceq K^2$. Hence the distribution $\sigma_t$ of the uniform chain is majorized by that of any symmetric birth-and-death chain.

*Ad hoc* methods allow to generalize the result:

**Theorem 1.5** (J. Fill and J. Kahn, 2013, Theorem 4.3)**.** *If $X$ is a symmetric Markov chain, then $\pi_t$, the distribution after $t$ steps starting from $0$, majorizes $\sigma_t$ the distribution after $t$ steps of the uniform Markov chain, for all $t$.*

Hence, the uniform chain mixes faster than any other chain for all the divergences controlled via majorization.

The uniform chain also has the lowest Lovász-Winkler mixing time if $n$ is even. But not if $n$ is odd. As a counter-example, if $n = 1$, an optimal stopping time for the uniform chain is just to stop after the first step, with expectation 1. On the other hand, with the symmetric chain with $K(0,1) = 1$, we may stop at time zero with probability $1/2$, and stop at time 1 otherwise. The final $X_T$ does follow the uniform law, and $\mathbb{E}[T] = \frac{1}{2}$.

More generally, a stopping time is optimal if $X_T$ has law $\pi$ and there is a halting state. For birth-and-death chains, such conditions are met by $T$ defined as: draw $j$ according to $\pi$ and stop the chain when $j$ is hit. Indeed, $n$ is a halting state for $T$. Hence:

$$T_{mix} = \sum_{i=0}^{n} \pi_i T_i,$$

where $T_i$ is the hitting time of $i$ starting from 0. If $\pi$ is uniform, a few calculations yield

$$T_{mix} = \sum_{k=0}^{n-1} \frac{(k+1)(n-k)}{p_k},$$

with $K(i, i+1) = p_i$. After optimisation:

**Theorem 1.6.** *The birth-and-death chain on $\mathcal{X} = \{0, \dots, n\}$ starting from 0 with lowest Lovász-Winkler mixing time $T_{mix}$ is:*

— *the uniform chain if $n$ is even.*

— *if $n$ is odd, then:*

$$p_k = \begin{cases} 1 - \theta_n & \text{if } k \text{ is even,} \\ \theta_n & \text{if } k \text{ is odd,} \end{cases} \qquad (k = 0, \dots, n-1), \tag{3}$$

*with, for all $m$:*

$$\theta_{m-1} := \frac{1}{6}\left[\sqrt{(m^2+2)(m^2-4)} - (m^2-4)\right]. \tag{4}$$

Up to restriction to monotone kernels for some of them, part of these results can be generalized to non-uniform stationary distributions.

# 2 Images in science

Natural sciences generate many data with spatial organisation, often on a regular grid. They are not always images *stricto sensu*. Either because the spatial organisation is not that of the data, but that of the target, as in MRI where we observe a Fourier transform of the target image. Or because the data at a point in space (pixel or voxel) is not only an intensity, a single number. Such a pixel may either contain information relevant by itself, or data to be analysed further. An example of relevant data could be the absence or presence of molecules at that point, for several types simultaneously. Data yet to be analysed could be a sequence of arrival times whose characteristic times are the target information, as in FLIM-FRET fluorescence imaging. An intermediate case could be the time series of a brain voxel in functional MRI.

In all these cases, data analysis and optimisation of the acquisition yield original statistical questions. Relevant techniques may stem both from image analysis (after all, signal processing with non-artistic aims is a branch of statistics) and from more general parametric or non-parametric statistics.

What follows is a miscellanea of themes I have encountered when trying to work with experimentalists, be it their original problems or the mathematical problems they have generated.

Section 2.1 describes fluorescence lifetime imaging and fluorescence resonance energy transfer (FLUM-FRET), and the data it produces. We mention a few analysis methods and a few suggestions to analyse them. In particular, on each pixel, we have to estimate the parameters of a probability mixture. Section 2.2 deals with the optimal rates of estimation of these parameters, correcting the wrong rates in the literature.

In Section 2.4, we speak about compressed sensing in MRI, its physical constraints, and suggest further applications for the algorithms we have developed with that problem in mind.

In Section 2.3, we are interested in ultra-fast clustering algorithms when there spatial correlations, as in functional MRI.

## 2.1 Protein-protein interaction measurement with FLIM-FRET

The rise of fluorescence microscopy has been induced by Chalfie et al. (1994), who has managed to include the DNA of the GFP (green fluorescent protein) in the DNA of any protein of a cell. This allows to add a small fluorescent part to any protein the cell makes, and the experimentalist hopes it will not alter its function. It is then possible to see where these proteins go in a living cell.

Many tricks and improvements give access to other possibly subtle data (Lakowicz, 2013), such as the diffusion coefficients of the protein at a given scale. We will look into Förster (1948) fluorescence resonance energy transfer, or FRET.

FRET is an energy transfer between a fluorescent donor whose emission frequency is the excitation frequency of a fluorescent acceptor. This transfer happens only if the donor and acceptor are very close, less than 10 nanometres away. Hence if two proteins are conjugated respectively with the donor and the acceptor, and such a transfer is observed somewhere in the cell, the two proteins are very close at that point. This is an indicator of an interaction between the two proteins.

One of the more accurate ways to detect FRET is by observing fluorescence lifetime (FLIM), which may be done by time-correlated single photon counting (Duncan et al., 2004). More precisely, an excited fluorescent molecule may decay in several ways, one of which is emitting a photon: this is fluorescence. All decay modes are either strictly quantum, and hence have a half-life, or linked to a sufficiently random environment to have a half-life, too. Hence, if a fluorescence photon is emitted, it will be after an exponential time with natural parameter $\lambda$. FRET adds a decay mode. Hence, if $\lambda_D$ is the parameter without FRET, and $\lambda_F$ is the parameter of FRET itself at the distance between the two molecules, then the parameter with FRET is $\lambda_{DA} = \lambda_D + \lambda_F$.

If we use mean lifetimes instead of the natural parameters of exponentials, we get $\tau_{DA} = \frac{1}{\tau_D^{-1}+\tau_F^{-1}}$. The mean fluorescence lifetime of the donor decreases when FRET occurs. Its scale is $10^{-9}$ seconds. Experimentally, we send a laser pulse at the excitation frequency of the donor, and the microscope detects individually photons emitted by the donor, with a precision of $10^{-11}$ seconds (Waharte et al., 2006). Mathematically, we get samples from the lifetime distribution of the donor.

Observations are made with a confocal microscope: a small volume is observed, then we move the observation point. Moving on a grid, we get a

sequence of lifetimes at each pixel in an image. However, the volumes are big relatively to the molecules, so that there are never only interacting molecules. Specifically, the intensity at time $t$ after a laser pulse will be

$$I(t) = I_0 + \sum_i I\pi_i \exp\left(-\frac{t}{\tau_i}\right), \tag{5}$$

where $I_0$ is a known constant noise, $I$ is the total intensity of the donor immediately after the pulse, which is proportional to the number of donor molecules in the volume, and $\pi_i$ is the proportion of donor molecules in context $i$: typically, $i$ is either «far from the acceptor», or «close to the acceptor». But there may be other influences. The $\tau_i$ are the corresponding mean lifetimes.

Since the $\tau_i$ depend on the environment, they are not assumed to be known. The first relevant question is to know whether the lifetimes are generated by a mono-exponential (probably no interaction) or a multi-exponential. Heinrich, Jonas Kahn, et al. (2011) show that a simple likelihood ratio test is at least ten times more efficient than a chi-squared test for this task.

More generally, the aim is to estimate the proportions of the different mean lifetimes at each point. Without any *a priori* knowledge, like the mean lifetime without acceptor, this estimation of the parameters of a mixture is an extremely difficult problem. We determine optimal rates in the next section. Let us just note that if the mean lifetimes are ten times closer in a bi-exponential, then we need $10^6$ as many data to get the same relative precision...

Traditional estimation methods by biologists ($L^2$ fit of curves...) can certainly be improved. More fundamentally, we deal with an "image". We should be able to denoise the parameter estimation like a general image. Maybe even directly share date between different pixels. As soon as we have a good similarity measure between pixels, or small pixel patches, we may use non-local means between the patches, for example. This would greatly increase effective data per pixel.

Heinrich, Pisfil, et al. (2014) show that the transportation distance between empirical probabilities at each pixel is a somewhat robust similarity measure.

Other possible improvements stem from modifications to experimental methods allowed by better analysis, as Rebafka (2009) exemplifies. She had shown how to study lifetimes despite pile-up. Pile-up is the following: when two photons are emitted after the same laser pulse, the second is not detected. Hence the distribution of lifetimes change, and is no longer a mixture of exponentials, but in a known way. Before her work, experimentalists would

just use low intensities to ensure this case seldom happens, and use formulas with exponentials. Her methods allow the use of higher intensities, and hence quicker acquisition. Similarly a physical model photobleaching, which stops for a long time fluorescence of molecules that have been lit too much, would allow observing the cell longer.

## 2.2 Optimal estimation rates for finite mixtures

### 2.2.1 Mixtures

With the motivation of estimating fluorescence mean lifetimes (5), I have looked into the literature on estimating mixture parameters. It turns out the optimal rates are wrong. Here are the corrections.

A *mixture* is a probability law of the form:

$$P(\cdot, G) = \int P(\cdot, \theta) \mathrm{d}G(\theta), \tag{6}$$

where $G$ is the *mixing distribution*, a probability law on the space $\Theta$ of parameters $\theta$, and where $F(\cdot, \theta)$ are probability laws on the same space for all $\theta$.

Mixtures are mostly used in three cases. First, in classification, where each data point has to be labelled as belonging to a group (McLachlan and Peel, 2000), a generative approach consists in assuming that each group generates data with distribution $F(\cdot, \theta)$ for different $\theta$. Then the law of the unlabelled data is the mixture distribution.

Second, and this might be the most usual case, mixtures are just used because their flexibility allows a good representation of heterogeneous data. In such a case, the statistician wants to approach the mixture density $P(\cdot, G)$, the distribution of the data (Genovese and Wasserman, 2000, par exemple).

Third, and that is our case of interest, the process generating the data can be directly mapped to a mixture, and the parameters of interest are the parameters of the mixture itself: we want to know the mixing density $G$ rather than the mixture $P$.

This problem is much harder than the second. We can typically estimate $P$ at parametric rate $n^{-1/2}$, multiplied by a polylog if $G$ does not have finite support. On the other hand, an example of estimation of $G$ is the deconvolution, where $G$ is a true function, and $P(\cdot, \theta)$ is the law of the convolution noise shifted by $\theta$. In that case, estimation rate is logarithmic.

From now on, we deal only with finite mixtures, so that $G$ is a finite sum of Dirac delta. We write $\mathcal{G}_m$ for the set of mixing distributions with exactly $m$ components, and $\mathcal{G}_{\leq m}$ for the set of mixing distributions with at

most $m$ components. Unless otherwise stated, we assume that we know that the true mixture has at most $m$ components. Moreover, we assume that $P$ is a probability law on $\mathbb{R}$, and write $f(\cdot, \theta)$ for the corresponding density, and $F(\cdot, \theta)$ for the repartition function. Moreover, we assume that the set of parameters $\Theta$ is a compact in $\mathbb{R}$.

What follows may probably be generalised to spaces other than $\mathbb{R}$, or for a multi-dimensional $\Theta$. Non-compactness would require strengthening the identifiability conditions, however.

A good distance between mixing distributions is the ($L^1$) transportation distance, which removes any identifiability problems:

$$W(G_1, G_2) = \sup_{|f|_{\text{Lip}} \leq 1} \int_\Theta f(\theta) \mathrm{d}(G_1 - G_2)(\theta), \tag{7}$$

where $\|\cdot\|_{Lip}$ is the Lipschitz semi-norm. We write $\mathcal{W}_G(\varepsilon)$ for the $\varepsilon$-radius ball around $G$, in transportation distance.

Known rates were $n^{-1/4}$, which is strange since deconvolution, an infinite mixture identification, has logarithmic rate. Usually, when the rate does not depend on the number of parameters, it will be the same for an infinite number of parameters. We now solve the paradox.

### 2.2.2 Minimax rate

Let us give some intuition first. The observations are the empirical repartition function $F_n$. The Dvoretzky-Kiefer-Wolfowitz ensures that the distance to the true repartition function $F$ decreases like $n^{-1/2}$:

$$\|F_n - F\|_\infty \approx n^{-1/2}. \tag{8}$$

We consider $\widehat{G}_n$ the minimum distance estimator by Deely and Kruse (1968):

$$\|F(\cdot, \widehat{G}_n) - F_n\|_\infty = \inf_{G \in \mathcal{G}_{\leq m}} \|F(\cdot, G) - F_n\|_\infty. \tag{9}$$

The triangle inequality and remark (8) ensure that the mixture $F(\cdot, \hat{G}_n)$ converges to the true mixture $F$ at rate $n^{-1/2}$. Hence, we may bound the convergence rate of $\widehat{G}_n$ to $G$ if we can control $W(G_1, G_2)$ by a function of $\|F(G_1) - F(G_2)\|_\infty$.

Suppose now that $G$ is close to a given $G_0$ with $m_0$ components, so that

$G$ may be written as:

$$G = \sum_{i=1}^{m_0} \sum_{j=1}^{J_i} \pi_j \delta_{\theta_i + \varepsilon_n h_j},$$

$$\sum_{j=1}^{J_i} \pi_j = \pi_i + O(\varepsilon_n).$$

Then a Taylor expansion in $\theta$ yields

$$F(\cdot, G) = \sum_{k=0}^{K} \varepsilon_n^k \sum_{i=1}^{m_0} \left( \sum_{j=1}^{J_i} \frac{\pi_j h_j^k}{k!} \right) F^{(k)}(\cdot, \theta_i) + O(\varepsilon_n^{K+1}), \tag{10}$$

where the $F^{(k)}$ are derivatives with respect to $\theta$.

We see that we can expect the distance between $F(\cdot, G_1)$ and $F(\cdot, G_2)$ to scale like $\varepsilon_n^{K+1}$ if we can ensure that the following moments be equal: $\sum_{j=1}^{J_{i,1}} \frac{\pi_{j,1} h_{j,1}^k}{k!} = \sum_{j=1}^{J_{i,2}} \frac{\pi_{j,2} h_{j,2}^k}{k!}$ for all $i \leq m_0$ and all $k \leq K$.

We have $2J_i$ parameters for the component $i$, namely $\pi_j$ and $h_j$ for $j \in [1, J_i]$, we want at least two different solutions to a set of $(K+1)$ equations, namely $\sum \pi_j h_j^k = c_k$ for $k \in [0, K]$. This happens for $K + 2 \leq 2J_i$. Since at least one component of $G$ must be close to each component $\theta_i$ of $G_0$, there are at most $J_i = (m - m_0 + 1)$ components that are close to $\theta_i$. Hence the distance between the mixtures corresponding to $G_1$ and $G_2$ near $G_0$ will be at least of order $\varepsilon_n^{-(2(m-m_0)+1)}$.

Moreover, we may expect that the distance between $G_1$ and $G_2$ be of the same order as $h_j$, that is $\varepsilon_n$. Injecting the distance (8), we obtain an estimation rate $n^{-\frac{1}{4(m-m_0)+2}}$.

To make the above rigorous, we need to ensure that:

— the problem is smooth enough in $\theta$ to take the needed derivatives.

— a $k$-strong identifiability condition for $k$ big enough is satisfied, namely: for all finite set of distinct $\theta_j$, the equality

$$\left\| \sum_{p=0}^{k} \sum_{j} \alpha_{p,j} F^{(k)}(\cdot, \theta j) \right\| = 0$$

implies $\alpha_{p,j} = 0$ for all $p$ and $j$. Indeed, it ensures that (near)-equality of moments is necessary for (near)-equality of distributions in the expansion (10).

— If $W(G_1, G_2) \ll \varepsilon_n$, then $\|F(\cdot, G_1) - F(\cdot, G_2)\| \geq \delta W(G_1, G_2)^{2(m-m_0)+1}$ is still true for $\delta$ depending only on $G_0$. This allows to deal with $G_i$ that converge to $G_0$ along the same lines asymptotically. Taylor expansions require more care, and must be written by grouping components in a tree-like fashion. It works.

— No estimator is better, up to a multiplicative constant. This requires some regularity assumptions, such as having some expectations like $\mathbb{E}_{\theta_1}\left|\frac{f^{(p)}(\cdot,\theta_2)}{f(\cdot,\theta_3)}\right|^q$ be finite.

Using compactness of $\Theta$ to go from local to global, we finally get:

**Theorem 2.1** (Heinrich and Jonas Kahn, 2015, Theorems 3.2 et 3.3)**.** *Under (explicit) sufficient regularity conditions, the minimax estimation rate around $G_0$ with $m_0$ components in the space $\mathcal{G}_{\leq m}$ of finite mixtures with at most $m \geq m_0$ components is $n^{-\frac{1}{4(m-m_0)+2}}$. That is, denoting by $\widehat{G}_n$ any sequence of estimators, and $\varepsilon_n = n^{-\frac{1}{4(m-m_0)+2}+\kappa}$ pour un $\kappa > 0$:*

$$\infty > \liminf_{n\to\infty} \inf_{\hat{G}_n} \sup_{G_1 \in \mathcal{G}_m \cap \mathcal{W}_{G_0}(\varepsilon_n)} n^{1/(4(m-m_0)+2)}\, \mathbb{E}_{G_1}\left[W(G_1, \widehat{G}_n)\right] > 0.$$

*Under the same hypotheses, the global minimax rate of estimation on $\mathcal{G}_m$ is $n^{-1/(4m-2)}$.*

We may notice several points when reading the theorem. First, the rate gets worse when they are more components. Hence, it is natural that the rate for infinite mixtures is non-parametric.

Second, it is the uncertainty on the number of components that slows estimation. The worst case is when a mixture with many components is close to a single-component mixture. Many components can look like many different components by equalizing their moments. Note that the problem does not appear when the true mixture has few components, but when the true mixture is sufficiently close, for a given $n$, to a mixture with few components.

When the number of components is known, that is $m = m_0$, we obtain a $n^{-1/2}$ rate. This is not surprising, since locally we are in a classical parametric case. But this motivates the study of pointwise convergence rates.

### 2.2.3 Pointwise rate and interpretation

Under identifiability conditions, there are estimators of the number of components of a mixture, such that, for any fixed $G$, the estimator is exact with probability $1 - \varepsilon_n$, with $\varepsilon_n \ll n^{-1/2}$. Indeed, the asymptotic rate is

faster than any polynomial. An example of such an estimator is, with $G_{n,m}$ the minimum distance estimator in $\mathcal{G}_{\leq m}$ and some $\frac{1}{2} > \kappa > 0$:

$$\hat{m} = \hat{m}_n = \inf \left\{ m \geq 1 : \|F(\cdot, \widehat{G}_{n,m}) - F_n\|_\infty \leq n^{-1/2+\kappa} \right\}. \tag{11}$$

Now, if we know the number of components, the estimation can be done at rate $n^{-1/2}$. So that

**Theorem 2.2** (Heinrich and Jonas Kahn, 2015, Theorem 3.5)**.** *Under sufficient regularity conditions (less than in Theorem 2.1), there are estimators such that for any finite mixing distribution $G \in \mathcal{G}_{<\infty}$, the estimator converges to $G$ at rate $n^{-1/2}$:*

$$\mathbb{E}_G \left[ W(\widehat{G}_n, G) \right] = C(G) n^{-1/2}, \tag{12}$$

*where $C(G)$ depends only on $G$.*

At first sight, Theorems 2.1 and 2.2 may seem to contradict each other. Minimax and pointwise rates are seldom different in statistics, so that it may be worth dwelling on the meaning of this difference.

Minimax and pointwise differ by uniformity. Any point may be approached at rate $n^{-1/2}$, but the moment when the asymptotic regime is attained depends on the point. Hence the worst case among all points, or all points in a small ball, may not be $n^{-1/2}$. Another way to interpret this lack of uniformity is to notice that the constant $C(G)$ in the risk (12) explodes when $G$ gets close to a $G_0$ with fewer components.

The practical consequence is that, if two components of a mixture are very close, a huge number of observations are necessary to get a good estimate of the mixing distribution. It is easier to find the number of components, and the lower moments. But if $N$ data points are necessary to tell $G_1$ from $G_2$ with probability 0.9, and the two mixtures have three well-separated components, we need $10^{10}N$ to tell apart mixtures where the $\theta_i$ are divided by 10. Most often impossible in practice.

Hence, if an experimentalist has a choice, by choosing a marker, say, then it might be worth to ensure that the components are well-separated, even if there are less data points.

## 2.3 Functional MRI

The title of this section might be slightly misleading. It refers to my exchanges with Gaël Varoquaux, and I will mention another idea at the end. Most of the section, however, deals with Recursive Nearest neighbour

Agglomeration. The aim of the algorithm is quick dimension reduction for spatially organised data. It can certainly be applied outside MRI.

Functional MRI allows high-resolution observation of the brain, with about $10^5$ or $10^6$ voxels nowadays (Zalesky et al., 2014). The activity of the brain at each voxel is measured as a time series. The easiest analysis starts with replacing each time series by a standardised mean activity during the observation time.

We then get an image with $10^6$ voxels, for each subject and each repetition of each task. The final aim is to determine active areas for a given task, or to distinguish subjects according to their health, for example. In any case, the images are the input of other algorithms, such as an independent component analysis, or a classification algorithm.

Those algorithms get very slow and memory-hungry for such huge inputs. Reducing dimension is therefore useful. Here, we mean reducing to $10^4$ dimensions, and keeping most of the signal, not summarizing in three parameters. But we need dimension reduction algorithms that are fast and faithful.

A classical method that is *a priori* efficient relies on random projections. We project the state space with dimension $n$ on a random subspace with dimension $k$, with $k \ll n$. Projection is very fast, and Johnson and Lindenstrauss (1984) lemma ensures that the $L^2$ distances are approximately maintained. However, random projections do not take th structure of images into account: they are usually piecewise continuous, and using this information might enhance results. A yet bigger downside is the lack of interpretability: each projection axis is distributed over the whole image, and the image cannot be directly rebuilt from the projections.

Thirion et al. (2015) have thus tried to devise a clustering method on voxels. Voxels are gathered into "super-voxels". They are a way to compress data with little loss, and allow direct computation on the compressed data. The following heuristics give an idea of what is expected of such an algorithm.

We observe $Y = X + \varepsilon$, where $Y$ and $X$ are functions from $\Omega$ to $\mathbb{R}$ and $\Omega$ is typically a subset of $\mathbb{Z}^3$. We write $X$ and $Y$ as vectors of $\mathbb{R}^n$, with $X_\omega = X(\omega)$. The ground truth $X$ is assumed to be $L$-Lipschitz, say for the graph distance. The noise $\varepsilon$ is assumed to be Gaussian with variance $\sigma^2$, independent for each $\omega$. Assume that the clustering $\Phi$ yields a partition $(C_1, \dots, C_k)$ of $\Omega$ that is independent of the data. We write $i(\omega)$ for the $i$ such that $\omega \in C_i$, and $\bar{X}_i = \frac{1}{|C_i|} \sum_{\omega \in C_i} X_\omega$.

Then $\Phi : \mathbb{R}^n \to \mathbb{R}^k$ and $(\Phi Y)_i = \sqrt{|C_i|}\bar{Y}_i$. Hence $\Phi Y$ is a Gaussian vector with variance $\sigma^2$ and mean $\sqrt{|C|}\bar{X}$, coordinate-wise. We denote by $\Phi^\dagger$ the

pseudo-inverse of $\Phi$, and pull the projection back in the original space $\mathbb{R}^n$:

$$\begin{aligned}\Phi^\dagger \Phi Y_\omega &= \bar{Y}_{i(\omega)} \\ &= \bar{X}_{i(\omega)} + \frac{1}{|C_i|} \sum_{\chi \in C_{i(\omega)}} \varepsilon_\chi. \end{aligned} \tag{13}$$

We quantify fidelity by comparing $\|X\|_2^2$ and $\|\Phi Y\|_2^2$:

$$\begin{aligned}\|\Phi Y\|_2^2 &= \sum |C_i| \, \bar{X}_i^2 + \varepsilon_i^2 \\ &= \sum_\omega X_\omega^2 - \left(X_\omega^2 - \bar{X}_{i(\omega)}^2\right) + F \\ &= \|X\|_2^2 + F - \sum_i |C_i| \operatorname{Var}_i(X) \end{aligned} \tag{14}$$

where $F$ has a $\chi^2(k)$ distribution, and $\operatorname{Var}_i(X)$ is the intra-cluster variance of $X$, that is $\operatorname{Var}_i(X) = \frac{1}{|C_i|} \sum_{\omega \in C_i} (X(\omega) - \bar{X}_{i(\omega)})^2$.

We then notice that:

— The expression (13) highlights the super-voxels, and interpretation as local means.

— Consequently, the clustering has a denoising effect. It may be lower for the proposed algorithms, since they depend on the data.

— Many kernel methods (Rahimi and Recht, 2007) only depend on the distances between subjects. That is, if $Y^\alpha$ and $Y^\beta$ are observations corresponding to subjects $\alpha$ and $\beta$, we only need to maintain $\left\|Y^\alpha - Y^\beta\right\|$. Hence, only the differences between $Y$ must be $L$-Lipschitz, which is *a priori* less demanding.

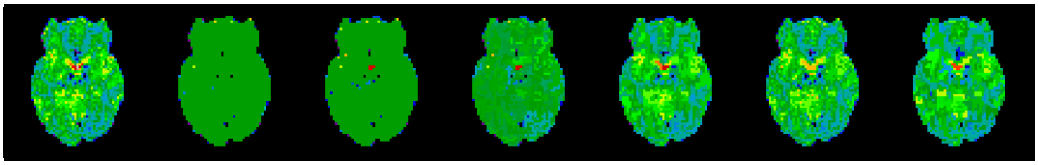
— All individuals $Y^\alpha$ are projected on the same clusters. Hence maybe a James-Stein estimator on all $\Phi Y_i^\alpha$ may improve denoising.

— A good algorithm must have a good fidelity (14), hence the intra-cluster variance must be small.

— Since this variance is bounded from above by $L^2 \operatorname{Diam}(C_i)^2$, and $\operatorname{Diam}(C_i) \leq |C_i|$, a good algorithm should probably have small clusters.

— Similarly, if the clusters are $d$-dimensional balls, then $\operatorname{Diam}(C_i)$ is of order $|C_i|^{1/d}$. A good algorithm should yield "compact" clusters.

— A good algorithm should gather similar voxels: this corresponds to lowering $L$ within the cluster.



(a) Original (b) single-linkage (c) average-linkage (d) complete-linkage (e) Ward (f) SLIC (g) ReNA

Figure 2 – **Approximation of an MRI image obtained with various feature grouping algorithms**

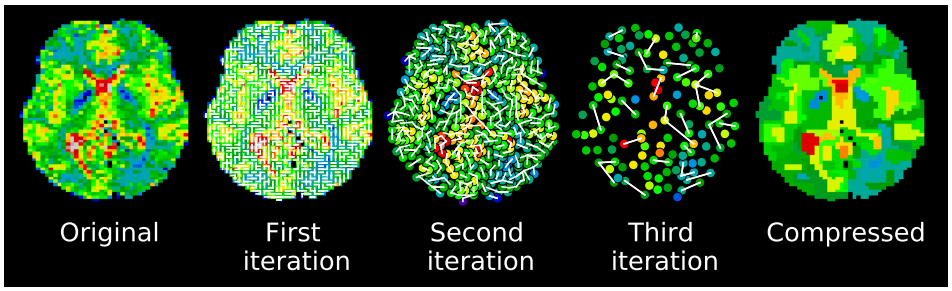



Figure 3 – **Illustration du principe de ReNA:** The white lines represent the edges of the graph. The algorithm considers each voxel in the original image as a cluster. From now on, for each iteration, the nearest clusters are merged, yielding a reduced graph, until the desired number of clusters is found.

Figure 2 depicts the result of several clustering algorithms; giant components are clearly bad. Percolation must be avoided.

Ward (1963) clustering directly aims at minimizing intra-cluster variance. Hence its good practical results are not surprising. However, its complexity is $O(n^2)$, and it is too slow for the targeted applications.

At the other extreme, single-linkage clustering (Gower and Ross, 1969) may be computed extremely fast. But there is a giant component and much information is lost when averaging.

Hence the proposition of Recursive Nearest Neighbour Algorithm (ReNA), illustrated in Figure 3. At each step, each cluster is merged with a neighbouring cluster, that is a cluster with which it shares an edge in $\mathbb{Z}^d$. This constraint ensures spatial coherence. The cluster with which it is merged is the closest datawise, that is the one that minimizes $\left|\bar{X}_i - \bar{X}_j\right|$.

The calculation is linear in the number of initial voxels, whatever the final number of clusters.

In practice, there is no percolation, and the clusters all have a similar size. The algorithm fulfills our intuitive specifications.

Experimentally, fidelity is high. As expected, further calculations are much faster. Thanks to the denoising, the results are sometimes even better than on the raw data.

The other methods in Hoyos-Idrobo et al. (2016) are either much slower, or lose too much information, at least for some uses. A striking example is independent component analysis with random projections. Since the "independent components" are spatially localised and the random projections are not, the results are awful.

### 2.3.1 Perspectives

I would like to prove theoretically that there is no percolation. The closest result I know is that by Teng and Yao (2007), on the percolation of the $k$-nearest neighbour graph on a Poisson point process.

Another classical use for functional MRI data is looking for functional connectivity: finding areas of the brain that work simultaneously. In such a case the time series on each voxel are not summarised with just a number, but rather the covariance matrix between regions of interest is computed, integrating over all the voxels in each region. A state, such as "healthy subject at rest", is modelled by a probability law on covariance matrices. The aim is to get a diagnostic by labelling the covariance matrix of a subject by a model. We thus need efficient models.

Traditionally, neurobiologists would use models like $\Sigma = \Sigma^* + \mathrm{d}\Sigma$, where $\mathrm{d}\Sigma$ is a random variable (Fair et al., 2007, par exemple). But the correlation matrix is a non-negative matrix, and this positivity is not guaranteed in such models. Among my projects with Gaël Varoquaux, I would like to see what we gain by modelling the random part as a tangent vector to $\Sigma$ on the positive semi-definite manifold.

## 2.4 Compressed sensing with physical constraints

### 2.4.1 Initial motivation: MRI acquisition

Most signals have a structure, meaning they can be sparsely represented in a basis (for example) known beforehand. That is, the biggest coefficients in the basis carry almost all the information, all the $L^2$ norm of the signal. However, we do not know beforehand which coefficients are big.

Candès, Romberg, and Tao (2006) have shown that it was possible to rebuild a sparse signal with only a few measurements. Those measurements

are linear forms along axes "without any relation" with the signal, like random projections. This is the start of compressed sensing theory. Lustig, Donoho, and Pauly (2007) have soon applied it to MRI.

However, the theory assumes that each measurement is chosen independently of the others. In real applications, physical constraints seldom allow such freedom. Most often, the measurements must be taken in a given basis. Moreover, and that is the main difference, successive measurements must follow a smooth path: if the measurement basis elements are indexed by an image, the chosen measurements must follow at least a continuous path, or its discretization.

Let us fix the notation and give a typical theorem.

We start with an orthogonal matrix e $\mathbf{A}_0 = [\mathbf{a}_1|\dots|\mathbf{a}_n]^*$, where the $\mathbf{a}_i$ are basis vectors. We extract from $A$ a random measurement matrix $\mathbf{A} = [\mathbf{a}_{J_1}|\dots|\mathbf{a}_{J_m}]^*$, where the $J_i$ are i.i.d. variables in $[1, n]$. We observe $\mathbf{y} = \mathbf{A}\mathbf{x}$ and want to find $\mathbf{x}$. Since the number of measurements $m$ is (much) smaller than the dimension of the space, the linear equation has many solutions. We use the following estimator:

$$\bar{\mathbf{x}} \in \underset{\mathbf{A}\mathbf{x}=\mathbf{y}}{\operatorname{Argmin}} \|\mathbf{x}\|_1 . \tag{15}$$

If there is noise or if the sparsity is not perfect, the equality should be relaxed, and we should use an estimator like LASSO, but the main ideas stay the same.

**Theorem 2.3** (Chauffert, Ciuciu, Jonas Kahn, and P. Weiss, 2014; Bigot, Boyer, and P. Weiss, 2016, Theorem 3.1)**.** *Assume that* $\mathbf{x}$ *is* $s$*-sparse, i.e. there are only* $s$ *nonzero components among* $n$*. Let* $\boldsymbol{\pi}_k = \frac{\|\mathbf{a}_k\|_\infty^2}{\sum_{i=1}^n \|\mathbf{a}_i\|_\infty^2}$*. If the number of measurements* $m$ *satisfy:*

$$m \geq Cs \left( \sum_{i=1}^{n} \|\mathbf{a}_i\|_\infty^2 \right) \log\left(\frac{n}{\epsilon}\right),$$

*with* $C > 0$ *a universal constant, then* $\bar{\mathbf{x}} = \mathbf{x}$ *with probability* $1 - \varepsilon$*.*

In MRI, the signal should be sparse in a wavelets basis, with matrix $\boldsymbol{\Psi}$, and the measurement is made in the Fourier basis, with matrix $\mathbf{F}$. So that $A_0 = \mathbf{F}^*\boldsymbol{\Psi}$. We may then show that $\|\mathbf{a}_i\|_\infty^2 \propto \log(n)$. Hence $O(s\log(n)^2)$ Fourier coefficients are enough to rebuild exactly an $s$-sparse image.

Notice that the $L^1$ norm is a relaxation of the $L^0$ norm and allow a fast calculation of the minimizer.

Usually, those theorems are written with a uniform choice of the coordinates to be measured, and the number of necessary coefficients depends on

the coherence $n \max_{1 \leq k \leq n} \|\mathbf{a}_k\|^2$. However, the coherence is high between the Fourier and wavelets basis, of order $n$. The practical method of drawing more often coherent coordinates was known, and this theorem is a formalisation of it.

We would like to draw the samples according to the target density $\boldsymbol{\pi}$. As already stated, in MRI, we measure Fourier transforms of the image, or rather of a part of space, not necessarily at integer frequencies. Moreover, successive measurements are made along a curve in the Fourier space, with finite speed and acceleration. Therefore, we cannot draw samples according to independent coordinates. A first step toward a realistic acquisition model is to sample along a continuous path.

### 2.4.2 Travelling salesman as a variable density sampler

An easy way to get a continuous path when we know to draw random points is to connect them. Moreover, intuitively, each point brings new information, so that we would like to connect as many points as possible with as short a path as possible. Hence the idea of using the travelling salesman solution to connect points drawn according to a specific density.

Let $X_N = \{x_i\}_{1 \leq i \leq N}$ be an i.i.d. $N$-sample with law $\tilde{\boldsymbol{\pi}}$. Let $\gamma_N : [0,1] \to \Omega$ be the constant-speed parametrisation of the travelling salesman between the $N$ points, where $\Omega$ is a compact convex set in $\mathbb{R}^d$. Let $\Pi_N = (\gamma_N)_* \lambda_{[0,1]}$ be the pushforward measure on $\Omega$ of the Lebesgue measure on $[0,1]$. Intuitively, the weight of a volume is proportional to the length of its intersection with the travelling salesman. We want to approximate $\boldsymbol{\pi}$, that is, we want that $\Pi_N$ converges to $\boldsymbol{\pi}$ when $N$ goes to infinity.

How should we choose the distribution $\tilde{\boldsymbol{\pi}}$ of the points $X_N$? A first idea would be to draw them according to $\boldsymbol{\pi}$, but that would be wrong. We must take the distance between points into account.

Let us clarify the idea by looking at a small cube. The number of points in the small cube $N_c$ is proportional to $\tilde{\boldsymbol{\pi}}$, which is approximately a constant in the small cube. The typical distance between two points in the small cube scales like $N_c^{-1/d}$. The travelling salesman usually connects nearby points, hence the points within the small cube, and is scale-invariant. So that the expected length of the travelling salesman in the small cube scales like $N_c N_c^{-1/d} \propto \tilde{\boldsymbol{\pi}}^{1-1/d}$.

This heuristic could be applied to many algorithms building curves from points, such as a greedy algorithm. For the travelling salesman, it can be made rigorous by using subadditivity of an associated process and the asymptotic estimate of the length of a travelling salesman by Beardwood, Halton, and Hammersley (1959). We get:

**Theorem 2.4** (Chauffert, Ciuciu, Jonas Kahn, and P. A. Weiss, 2013, Theorem 3.1)**.** *Let $\boldsymbol{\pi}$ be a density on a compact set $\Omega$ in $\mathbb{R}^d$. Let $\tilde{\boldsymbol{\pi}} = \frac{\boldsymbol{\pi}^{(d-1)/d}}{\int_\Omega \boldsymbol{\pi}^{(d-1)/d}(x)\mathrm{d}x}$. Then, $\tilde{\boldsymbol{\pi}}^{\otimes n}$-almost surely with respect to the sequence of points $\{x_i\}_{i\in\mathbb{N}}$, the distribution of the travelling salesman converges weakly to $\boldsymbol{\pi}$:*

$$\Pi_N \xrightarrow[N\to\infty]{} \boldsymbol{\pi} \qquad\qquad \tilde{\boldsymbol{\pi}}^{\otimes n} p.s.$$

Notice that we require $\boldsymbol{\pi}$ to be a density. However, if the target density has atoms, we may simply spend some time there without moving.

Numerical simulations show good reconstructions. However, the physical constraints of MRI are not yet all taken into account: we have allowed infinite acceleration. A first idea would be to project the travelling salesman curve on the set of curves satisfying MRI constraints (Chauffert, P. Weiss, et al., 2016). This is easy numerically, but we might not converge to the target density any more.

Another approach is to directly project on the target density on the space of constraints. This is the subject of the next section.

### 2.4.3 Projection on measures sets

We broaden the framework. In MRI, we look for the pushforward measure of a curve with constraints, that is a subset of all possible curves. This yields a subset $\mathcal{M}_N$ of Radon measure on $\mathbb{R}^3$. We want to approximate $\boldsymbol{\pi}$ with an element in this set.

More generally, how can we project $\boldsymbol{\pi}$ on a subset $\mathcal{M}_N$ of the Radon measures on $\Omega \subset \mathbb{R}^d$ ? Chauffert, Ciuciu, Jonas Kahn, and P. Weiss (2016) define the projection as a solution to the following variational problem:

$$\mu_N^* \hat{=} \inf_{\mu\in\mathcal{M}_N} \|h \star (\mu - \boldsymbol{\pi})\|_2^2 , \tag{16}$$

where $h$ is a kernel in $L^2(\Omega)$, and $\star$ is the convolution. This formulation allows both theoretical results and a numerical implementation.

A good kernel $h$ must define a norm $\mathcal{N}_h(\mu) = \|h \star \mu\|_2^2$ on the space $\mathcal{M}$ of signed measures bounded in total variation on $\Omega$. Thus, if $\boldsymbol{\pi} \in \mathcal{M}_N$, then $\boldsymbol{\pi}$ is the only solution to the variational problem (16).

If $h$ is continuous, so that $h \star \mu \in L^2(\Omega)$, then the norm condition is equivalent to $h$ having all nonzero Fourier coefficients. In such a case, $\mathcal{N}_h$ is a metrisation of the weak topology on balls in $\mathcal{M}$, and in particular on the set of probability measures. Since we are dealing with probability measures, $\mathcal{M}_N$ is bounded in total variation. Hence, if $\mathcal{M}_N$ is also weakly closed, then

the problem (16) has at least one solution. In particular, if $\mathcal{M}_N$ is a set of pushforward measures $\{p_*\gamma : p \in \mathcal{P}\}$ with $\mathcal{P}$ a set of parametrisations $p : X \to \Omega$, and if $\mathcal{P}$ is compact for the pointwise convergence topology, then the problem (16) has at least one solution. This will be the case for all examples below.

To see why this definition of projection is convenient, let us write, with $\Omega$ the torus:

$$\begin{aligned}\langle h \star (\mu - \boldsymbol{\pi}) | h \star (\mu - \boldsymbol{\pi})\rangle_2^2 &= \langle H \star (\mu - \pi) | \mu - \boldsymbol{\pi}\rangle \\ &= \langle H\mu | \mu\rangle - 2\langle H\mu | \boldsymbol{\pi}\rangle + \langle H\boldsymbol{\pi} | \boldsymbol{\pi}\rangle,\end{aligned}$$

with $\hat{H}(\xi) = |\hat{h}|(\xi)^2$ for all $\xi \in \mathbb{Z}^d$. In particular, if $\mathcal{M}_N$ only contains measures on $N$ points $p_i$, with weight $1/N$ on each of them, minimising (16) is equivalent to minimising the following attraction-repulsion equation:

$$\min_{p \in \mathcal{M}_N} \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} H(p_i - p_j) - \frac{1}{N} \sum_{i=1}^{N} \int_\Omega H(x - p_i) \mathrm{d}x. \tag{17}$$

If $H$ has a minimum at zero, the first term is a repulsion term between points, and the second term is an attraction term to $\boldsymbol{\pi}$. Teuber et al. (2011) have suggested the use of this equation with $H(x) = \|x\|_2$ and $N$ independent points for a halftoning algorithm.

In general, it is hard to find a global minimum to such an equation. For example, if we set $\Omega$ to be the sphere $\mathbb{S}^2$ and $\boldsymbol{\pi}$ the uniform measure, we get Thomson's (1904) problem, consisting in finding the minimal energy configuration of $N$ electrons on a sphere. It is still an open problem.

However, finding a critical point is "easy". The attraction-repulsion equation (17) is continuous in $p_i$ if $H$ is nice, and stays continuous if we discretize the integral, yielding a function $J(p)$. So that we may use a projected gradient descent and reach a critical point, according to:

**Theorem 2.5** (Chauffert, Ciuciu, Jonas Kahn, and P. Weiss, 2016, Corollaire 2)**.** *If $H$ is definable in an o-minimal structure, and has a continuous L-Lipschitz gradient, and if $C$ is a closed set definable on an o-minimal structure, then the following sequence converges to a critical point of the attraction-repulsion functional $J(p)$:*

$$p^{(k+1)} \in P_C\left(p^{(k)} - \gamma \nabla J(p^{(k)})\right)$$

*with $0 < \gamma < \frac{N}{3L}$.*

This theorem is a special case of a very general theorem by Attouch, Bolte, and Svaiter (2013). I will not give a precise definition of functions and sets definable on an *o*-minimal structure (Coste, 2000), but they include all polynomial of elementary functions, their compositions, their level sets...

Ideally, what would we require of $\mathcal{M}_N$ to get good properties?

To approximate any measure $\boldsymbol{\pi}$ by a sequence of measures in the sequence of $\mathcal{M}_N$, a necessary and sufficient condition is that $\bigcup_N \mathcal{M}_N$ be weakly dense in the probability laws on $\Omega$, at least if the $\mathcal{M}_N$ are nested. The convergence rate may be bounded via control in transportation distance if $h$ is $L$-Lipschitz, that is $\mathcal{N}_h(\mu - \pi) \leq LW_1(\mu, \pi)$.

In particular, if the $\mathcal{M}_N$ are sets of $N$ Dirac deltas with weight $1/N$, the convergence rate is at worst $LN^{-1/d}$.

More generally, we need to approximate $\mathcal{M}_N$ by a set $\mathcal{A}_n$ of discrete measures if we want to apply the projected gradient algorithm. That is always possible for the Hausdorff distance. In practice, however, the set $\mathcal{A}_n$ might be hard to make explicit or compute with. In the examples below, it is well-behaved. Moreover, since $\mathcal{A}_n$ plays the part of $C$ in Theorem 2.5, it has to be definable on an *o*-minimal structure.

The projected gradient descent algorithm on $\mathcal{A}_n$ always finds a critical point of the original problem (16), and can run with $n = 10^5$ points in a few days on current computers. The results are much faster with $n = 10^3$ points.

We illustrate the flexibility of the algorithm by giving a few examples of projection sets $\mathcal{M}_N$ and a few possible applications.

— The set of $N$ Dirac delta $\mathcal{M}_N = \{\frac{1}{N}\sum \delta_{p_i} : p_i \in \Omega\}$.

— The curves with non necessarily isotropic constraints on speed, acceleration, upper derivatives, that can be encoded as Sobolev balls:

$$\mathcal{M}_N = \left\{ \frac{1}{N} p_* \lambda_{[0,N]} : p \in (W^{m,q}([0,N]))^d, p([0,N]) \subset \Omega, \forall 1 \leq j \leq m, \left\| p^{(j)} \right\|_q \leq \alpha_j \right\},$$

where $\lambda$ is the Lebesgue measure and $q \in [1, \infty]$ and the $\alpha_j$ are nonnegative real numbers. In this case, we write $\Delta t = \frac{N}{n}$, and use the discrete derivative operator $(Ds)_i = (s_i - s_{i-1})/\Delta t$, with adapted boundary conditions:

$$\mathcal{A}_n = \left\{ s \in \mathbb{R}^{n \cdot d}, \forall 1 \leq i \leq n, s_i \in \Omega \text{ and } \forall 1 \leq j \leq m, \left\| D^j s \right\|_q \Delta t \leq \alpha_j \right\}.$$

— The sets of $k$ segments covered at constant speeds, each in time $\frac{N}{k}$:

$$\mathcal{M}_N = \left\{ \frac{1}{N} \sum_{b=1}^{k} (l_b)_* \lambda_{[0, \frac{N}{k} \frac{N}{k}]} : l_b : [0, \frac{N}{k}] \to \Omega, l_b(\frac{N}{k} t) = t x_b^1 + (1-t) x_b^2 \right\}.$$

In this case, discretising each segment with $m = \frac{n}{N}$ points, we use:

$$\mathcal{A}_n = \Bigg\{ q \in \mathbb{R}^{n \cdot d} : \forall 1 \leq i \leq n, s_i \in \Omega \text{ and } q_j = q_i + \frac{j-i-1}{m-2}(q_{i+m-1} - q_i)$$
$$\forall i \in \left\{ 1, m+1, 2m+1, \ldots, \frac{N}{k}m + 1 \right\} \text{ and } i \leq j \leq i+m-1 \Bigg\}.$$

— Of course Equation (17) can be generalised to use weights on the $p_i$.

As for applications:

— The initial motivation was MRI. Projection on curves in Sobolev balls and on segments can both be applied, the latter being easy to implement on real MRI scanners. Figures 4 and 5 give an idea of the efficiency. Results for high resolution imaging are even more impressive, and Boyer et al. (2016) show figures.

— Just art. Halftoning is used by printers. Moreover, drawing a grey-level image with just one line is a promising technique. See the Girl with a Pearl Earring in Figure 6.

— The algorithm may be used to deconvolve isolated spikes. Candès and Fernandez-Granda (2014) give an algorithm for exact reconstruction of sufficiently distant spikes after convolution. However the algorithm is excruciatingly slow as soon as the underlying space is of dimension more than 1. We may alternate on gradient descent in Equation (17) and projection on the space of spikes to get solutions, in any dimension $d$.

### 2.4.4 Perspectives

I intend to go on collaborating with Pierre Weiss and those around him. A new PhD student has started working on those subjects.

Apart from the above applications, others would include singularity detection. Moreover, there are both purely numerical questions, like how to accelerate the algorithms, and a few purely mathematical questions.

An easy problem mixing both would be to project in transportation distance, which is a very natural distance for measures.

A more ambitious goal would be to find under what conditions a low transportation distance between the target density $\pi$ and the sample yields direct guarantees of compressed sensing using the sample. Obviously, the general case does not hold, since a regular grid is just downsampling.

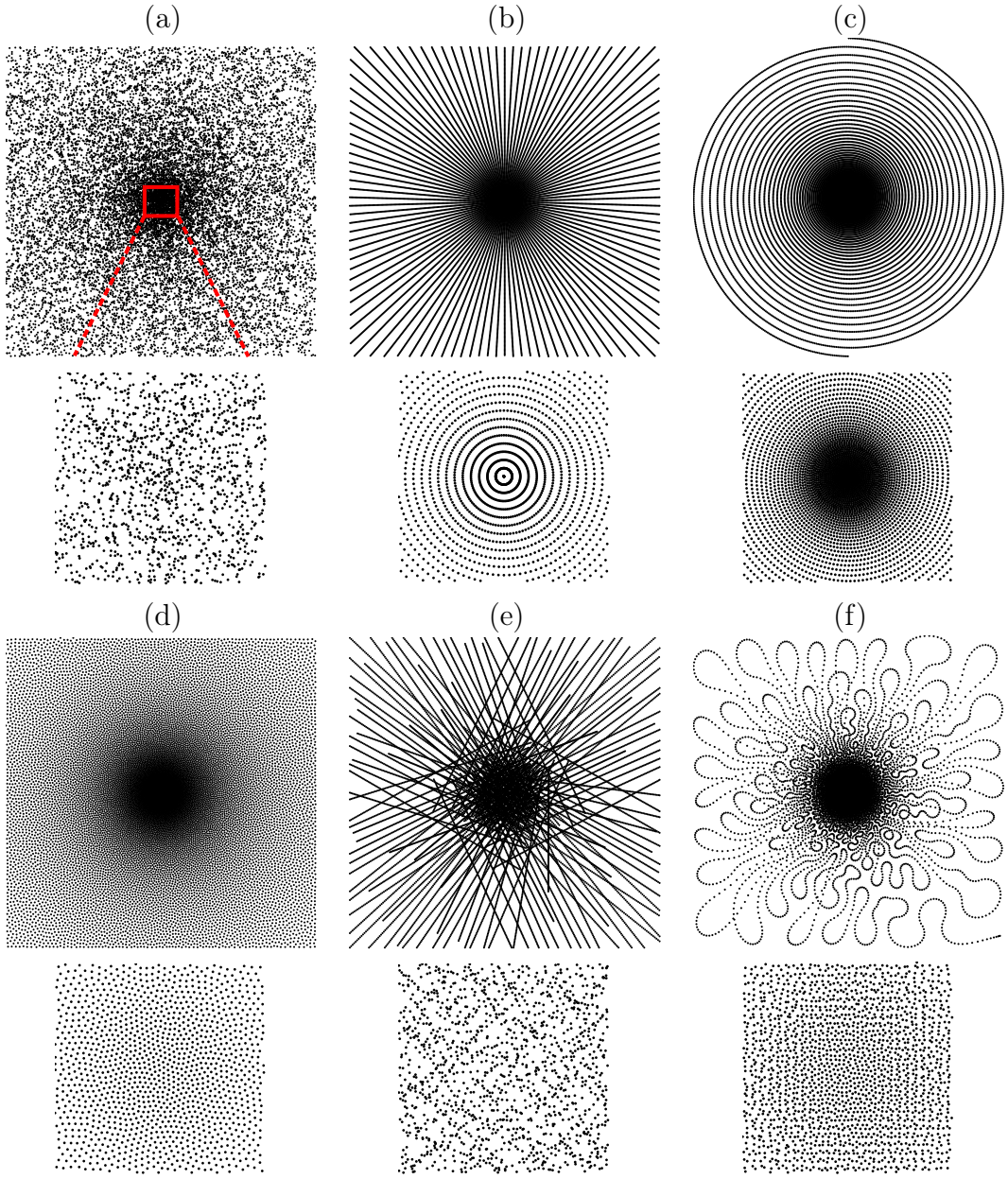



Figure 4 – Classical sampling schemes (a-c) and sampling schemes obtained with the projection algorithm (d-f). Top: (a): independent drawing; (b): radial lines; (c): spiral. Second line: centre zoom. Third ligne: (d): isolated points; (e): segments of variable length; (f): admissible curves fo MRI. Bottom: centre zoom. Corresponding reconstruction results in Figure 5.

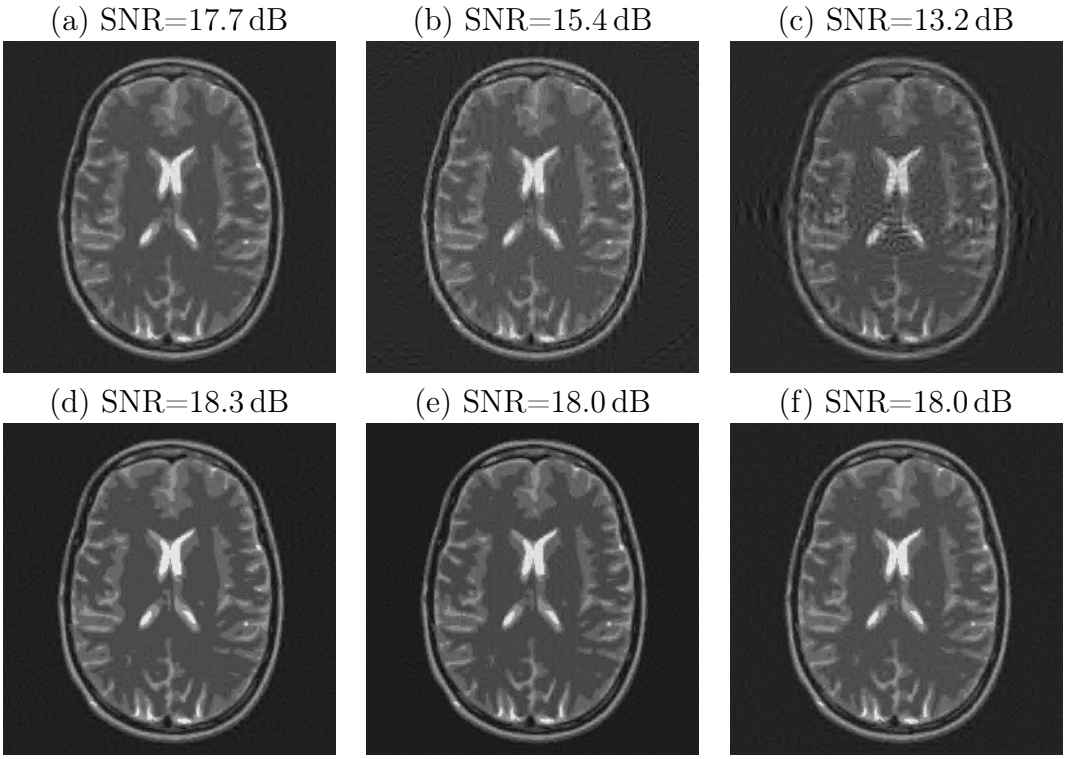



Figure 5 – Reconstructions for the sampling schemes in Figure 4 on a classical test image.

Figure 6 – *Girl with a Pearl Earring* by VERMEER (1665). Projection on a Sobolev ball $W^{1,\infty}$, with $N = 150000$ points. Result after 10000 iterations

Moreover, all critical points attained in practice are similar, meaning that the distance to the target density is about the same, for all starting points of the algorithm. Understanding the phenomenon looks hard and significant: when the number of points in the sample increases, do all critical points have a value close to the minimum value? Or at least, do their basins of attraction cover almost all the space?

# 3 Stochastic geometry and random metric spaces

Stochastic geometry is the study of random objects in a space with a geometry, often in their geometric aspects. It might be studying the shape of a Voronoi cell in the tessellation generated by a Poisson point process in $\mathbb{R}^d$, or wonder whether a point process with marks that are radius-1 circles percolates, for example. Knowing how information can be transmitted between sensors in a network or characterise a rock by its pores are two examples of practical problems that stochastic geometry is well-suited to formalise. Stoyan, Kendall, and Mecke (1996) give a good introduction to the field.

Some objects of stochastic geometry generate their own geometry: think of a graph between points connected if they are close enough. The graph yields a metric space on its vertices. In general, stochastic geometry gives easy ways to build geometric metric spaces. This point of view allows to include in the field of stochastic geometry the study of random metric spaces generated by topological or combinatorial means, like the Brownian map. We will only mention it briefly for comparison purposes.

In Section 3.1, we count $T$-tessellations on a fixed number of lines, so that we may prove existence of some Gibbs measures on these tessellations. They allow flexible modelisation of landscapes such as plots of land.

In Section 3.2, we explain the notion of a SIRSN (scale-invariant random spatial network) that Aldous (2014) introduced to model road networks. The improper Poisson line process is a SIRSN. Moreover it generates a random metric space that is interesting in itself.

## 3.1 $T$-tessellations

$T$-tessellations are tessellations of a (convex compact to make it easy) subset of the plane $W$ whose vertices are all of degree three with a flat angle, that is, are $T$-shaped. They look like plots of land, and Kiêu et al. (2013) have made the model with this motivation.

The first models of $T$-tessellations are special cases of polygonal Markov fields by Arak, Clifford, and Surgailis (1993). Those fields are a very general

model of a random geometric planar graph, whose nice properties allow exact sampling. However Kiêu et al. (2013) wished for a more flexible model, with Gibbs measures on a good underlying model.

The underlying model they choose is the completely random $T$-tessellation. We may intuitively define it as: adding a segment does not depend on the tessellation, as long as we still have a $T$-tessellation. Slightly more precisely, if $T$ is a $T$-tessellation, if $s$ is a segment and if $T + \{s\}$ is still a $T$-tessellation, then the ratio of the probability densities $\frac{p(T+\{s\})}{p(T)}$ does not depend on $T$. The clearest way to write the definition relies on the Papangelou kernel, and is similar to a Poisson process.

A consequence is that the probability density of a tessellation is given by the probability density of the lines that support the segments of the tessellation, and this density is given by of a Poisson line process. We give a complete description of this Poisson line process in Section 3.2.1 and give here only the intuition: it's throwing lines completely at random. The number of lines that hit the finite volume $W$ is a Poisson variable of parameter $\lambda$, the intensity.

It is not obvious that this measure $\mu$ on tessellations is finite, however, so it might be impossible to normalise it and get a probability measure. Indeed, each configuration of Poisson lines has a weight multiplied by the number of different $T$-tessellations it supports. We say that $k$ lines support a $T$-tessellation $T$ if $T$ is a union of segments that are all supported by the lines, and exactly one segment is supported by each line. This number of $T$-tessellations on $k$ lines must not be too high on average. We now bound the worst case by combinatorial means, and it turns out it is sufficient.

More precisely, suppose that for any set of $k$ lines in generic position, there are at most $\mathcal{N}(k) = o(a^k k^k)$ different $T$-tessellations supported by those lines, for all $a > 0$. Then the total variation of the non-normalised measure we have described is:

$$\begin{aligned} |\mu| &\leq \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \mathcal{N}(k) \\ &= \sum_{k=0}^{\infty} o((a\lambda e)^k) \qquad\qquad < \infty. \end{aligned}$$

We may prove it by combinatorial means:

**Theorem 3.1** (Jonas Kahn, 2014, Theorem 2.1)**.** *The number of $T$-tessellations supported on $k$ lines is bounded from above thus, for any $\varepsilon > 0$:*

$$\mathcal{N}(k) \leq C^k \left( \frac{k}{(\ln k)^{1-\varepsilon}} \right)^{k - k/(\ln k)},$$

*where $C$ depends only on $\varepsilon$.*

*Idea of proof.* The idea is to label each line (in the support) of a $T$-tessellation with enough information to rebuild the tessellation. The number of different $T$-tessellations on those lines is then bounded by the number of different sets of labels.

We give a simplified version. We choose an axis that is not parallel to any line in $T$, nor to the borders of the domain $W$, and we call it the time axis. We say that a segment of a tessellation is born at its minimum time point and dies at its maximum time point. It is now enough to label each line with the birth time of its segment, and the number of murders committed by its segment, that is the number of other segments that die when they intersect this segment. The rebuilding process is illustrated in Figure 7.

Now, a segment can only be born on the border or when crossing an other segment, hence a line: that is at most $k$ possibilities for each segment, hence $k^k$ total. Moreover, each segment only dies once, and it can be killed only be the border or a line. The numbers of murders for all lines then correspond to putting $k$ objects in $(k+1)$ boxes, hence $\binom{2k+1}{k} \leq 4^k$ possibilities. So that $\mathcal{N}(k) \leq (4k)^k$.

This bound is not enough for our needs. Getting Theorem 3.1 requires smarter labelling, that we merely mention now. We need to forget a well-chosen subset of the birth times, add other information similar to the number of murders, and go back and forth along the time axis. □

This combinatorial bound is essentially optimal: if we choose a grid with $k/(\ln k)$ vertical lines and $k - k/(\ln k)$ horizontal lines, it is easy to see that there are at least $\left(\frac{k}{\ln k}\right)^{k-k/(\ln k)}$ different $T$-tessellations on those lines. Just take the vertical segments to have maximal length, and there are $k/(\ln k)$ different possible positions for each horizontal segment.

However, we may hope that an analysis that takes into account the typical position and geometry of random lines could give a much lower mean number of $T$-tessellations on $k$ lines. Indeed, imagine the following heuristic: suppose we already have $k+1$ segments. In how many ways can the last segment be added? One for each interval between two segments that its line crosses, that is one plus the number $n$ of segments the line crosses. Hence the segment length should be of order $1/n$, with 1 the diameter of $W$. Now the probability that a segment crosses a line is proportional to the length of the segment. Hence typically $1/n$ for each line. So that each line should cross about $k/n$ segments. Hence $n = \sqrt{k}$. There are about $\sqrt{k}$ ways to set the last segment, hence we expect about $\sqrt{k}^k$ different $T$-tessellations.

Making the above rigorous is probably very hard...

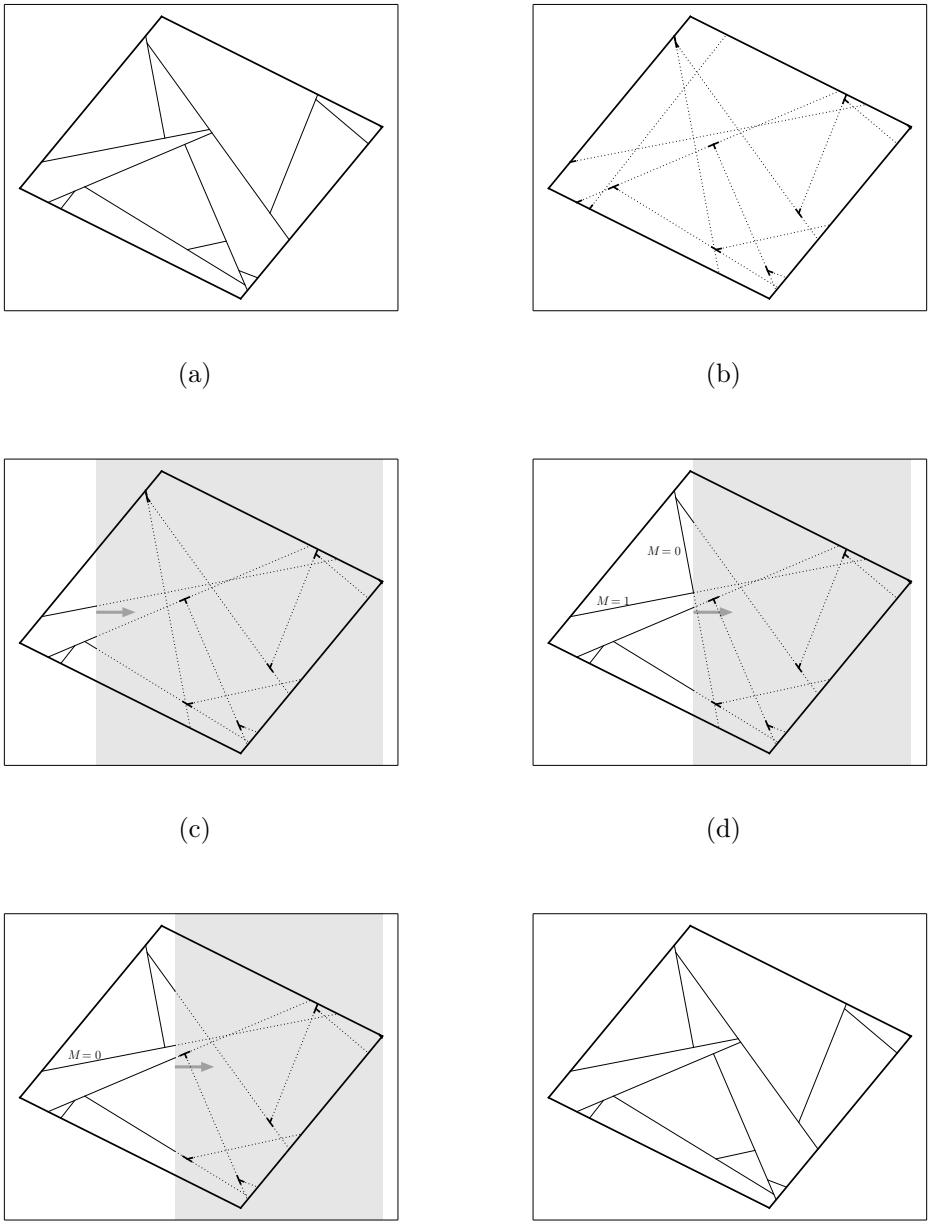



Figure 7 – (a) is the tessellation to be rebuilt. We start knowing the lines, the birth times (b) and the numbers of murders. We move along the time axis and prolongate the live segments (c). When two segments cross (d), we kill the one with zero murders left and decrease the other's counter by 1 (e). At the end of the pass, we get the tessellation (f).

## 3.2 Improper Poisson line process

### 3.2.1 SIRSN and definitions

In what follows, $B(x, r)$ is the Euclidean ball with centre $x$ and radius $r$.

Aldous (2014) has introduced the notion of a SIRSN (scale-invariant random spatial network) as an abstraction of road networks and maps as they can be used on the Internet. Hence the definition:

1. A user wants to go from a point $A$ to a point $B$. Hence a SIRSN consists in a route $\mathcal{R}(x_1, x_2)$ between each pair of points $x_1, x_2$ in $\mathbb{R}^d$, such that $\mathcal{R}(x_1, x_2) = \mathcal{R}(x_2, x_1)$. Each route is random and almost surely unique. A route is a finite-length path.
2. When we move a map, or rotate it, its appearance stays roughly the same. Similarly, zooming makes smaller roads appear. Hence a SIRSN must be statistically equivariant under translation, rotation and scaling: if $\mathcal{N}(x_1, \dots, x_k)$ is the network of routes between pairs of $x_i$ in a finite set $x_1, \dots, x_k \in \mathbb{R}^d$, and if $\mathfrak{R}$ is Euclidean similarity, then the networks $\mathcal{N}(\mathfrak{R}x_1, \dots, \mathfrak{R}x_k)$ and $\mathfrak{R}\mathcal{N}(x_1, \dots, x_k)$ follow the same law.
3. A good network allows to join two points without too many detours. Hence, if $D_1$ is the length of the route between two points at Euclidean distance 1, then $\mathbb{E}[D_1] < \infty$.
4. A good road network is built without too many roads at each scale. Moreover, routes follow the bigger roads, except near the parking slots. Therefore, let us consider $\{\Xi_n\}_{n\in\mathbb{N}}$ nested Poisson point processes (defined below) with intensity $n$, that is such that $\Xi_n \subset \Xi_{n+1}$. We write $\Xi = \bigcup_n \Xi_n$. Then the following long-distance network $\mathcal{F}$ has finite intensity $p(1)$, that is, finite mean length per unit volume:

$$\mathcal{F} = \bigcup_{x_1, x_2 \in \Xi} \big(\mathcal{R}(x_1, x_2) \setminus (B(x_1, 1) \cup B(x_2, 1))\big). \tag{18}$$

Let us make the last requirement clearer. The use of Poisson point processes is a technical device to work with only countably many routes. Morally, we require that all the routes between all the pairs of points in space, use only exactly the same (motor)ways except maybe when they are close to their starting or final point.

Those few properties already have a huge effect on any model that satisfy them. For example, Aldous (2014) has shown that there are singly-infinite paths, whose every subset is part of a route, but there are no doubly-infinite paths, that is paths with no end whose every subset is part of a route. On the other hand, it is not easy to build an explicit SIRSN. Aldous (2014) has

built a suitable hierarchical model in $\mathbb{R}^2$, but it is not very natural since invariance is added by symmetrising the whole process as a last step, so that each realisation has long-range dependences. It was the only known example.

The definition suggests a general way to produce a potential SIRSN: we start from a good random metric on $\mathbb{R}^d$, and the routes are the corresponding geodesics. If the random metric is invariant under similarities, we obtain automatically Property 2. Property 4 is a kind of hyperbolicity: the geodesics get closer. Now, random metric objects are most often hyperbolic (Gromov, 2003, for example).

Aldous (2014) has suggested two such random metrics. We will dwell on one of them: the improper Poisson line process.

Intuitively, we throw lines uniformly at random in $\mathbb{R}^d$, and mark them with a well-chosen random speed. There are more and more "slow" lines, so that those lines are dense in $\mathbb{R}^d$. The metric is the minimum time to drive from one point to the other while respecting the speed limits. Strangely enough, there are such continuous paths connecting any two points even in dimension 3 or more, even if the lines do not cross.

To state the above rigorously, we start by recalling that a Poisson point process on a measured space $(\mathbb{X}, \mu)$ is a random set of points in $\mathbb{X}$ such that, if we write $N(B)$ for the number of points in $B$:

— If the $B_i$ are measurable and disjoint, then the $N(B_i)$ are independent.

— $N(B_i)$ is a Poisson variable with parameter $\mu(B_i)$.

We may also parametrise the set of lines in $\mathbb{R}^d$ by a $\mathcal{H}$, so that the Lebesgue measure on the parameter space of any set of lines be invariant by any isometry of $\mathbb{R}^d$. For example in two dimensions, this space of parameters is $\mathcal{H} = [0, \pi[\times\mathbb{R}$, where the first parameter is the angle of the line with the abscissas, and the second parameter is the algebraic distance of the line to the origin of a specified frame.

We may then see a line $l$ with a speed $v$ as an element $(l, v)$ of $\mathcal{H}\times\mathbb{R}^+$. The improper Poisson line process $\Pi$ is the image of the Poisson point process on $\mathcal{H} \times \mathbb{R}^+$ with the measure with density $(\gamma - 1)v^{-\gamma}$ with respect to Lebesgue. We require $\gamma > d$.

With this notation, there are almost surely locally Lipschitz paths $\xi : [0, T] \to \mathbb{R}^d$ between each pair of points $a$ and $b$ in $\mathbb{R}^d$ such that:

— $\xi(0) = a$ and $\xi(T) = b$.

— For almost all $t \in [0, T]$, either $\xi'(t) = 0$, or $\xi$ follows a line in $\Pi$, that is there is a $v \geq |\xi'(t)|$ such that$(\xi(t) + \xi'(t)\mathbb{R}, v) \in \Pi$.

We call such a path a $\Pi$-path.

We may then define the metric on $\mathbb{R}^d$ given by the infimum of the $T$ of such paths connecting $a$ and $b$. We call it the $\Pi$-distance, or the time $T_{ab}$ needed to go from $a$ to $b$, to pursue the road analogy.

Similarly, we speak of $\Pi$-ball, $\Pi$-diameter, etc.

We have thus defined a rnadom metric space. The power law on speeds ensures scale invariance up to a multiplicative constant. Hence the law of the geodesics stay the same when zooming.

Some further notation: $\mathcal{L}_\xi$ is the support of a $\Pi$-path $\xi$, that is the set of lines in $\Pi$ that $\xi$ follows for a nonzero Euclidean length: $\mathcal{L}(\xi) = \{l \in \Pi : m_1(l \cap \xi(0,T)) > 0\}$ where $m_1$ is the one-dimensional Hausdorff measure, and $\xi : [0,T] \to \mathbb{R}^d$.

### 3.2.2 Properties

Kendall (2014) has obtained important results on this process: first, it is indeed a geodesic metric space. Moreover, in dimension 2, the geodesics are almost everywhere unique, the geodesics are locally of finite mean-length, and the subnetwork obtained from the routes connecting points of an independent Poisson point process has finite length in a compact set. The latter properties establish a weak form of Property 4 of a SIRSN.

We first further our understanding of the process by comparing the radius of balls in the Euclidean metric and the random metric:

**Theorem 3.2** (Jonas Kahn, 2015, simplified version of Theorems 3.1 and 5.1 and their proofs)**.** *There is a $T_1$ such that, for all $\frac{1}{2} > \varepsilon > 0$, for all $x \in \mathbb{R}^d$, for all radii $r$, with probability at least $1-\varepsilon$:*

$$\begin{aligned} T_{x,r} &\hat{=} \sup_{y,z \in B(x,r)} T_{y,z} \\ &\leq T_r \left(\ln \frac{1}{\varepsilon}\right)^{\frac{1}{\gamma-1}} && \text{with} \\ T_r &= r^{\frac{\gamma-d}{\gamma-1}} T_1. \end{aligned}$$

*In particular, the $\Pi$-diameter of a Euclidean ball with radius $r$ has all exponential moments, and more. For all $\delta < T_r^{\frac{1}{\gamma-1}}$:*

$$\mathbb{E}\left[\exp\left(\delta T_{x,r}^{\gamma-1}\right)\right] < \infty.$$

*The exponent $\gamma - 1$ in the moment cannot be improved.*

*On the other hand, the Euclidean diameter of a $\Pi$-ball with radius $r$ has finite expectation.*

Property 3 of a SIRSN is a consequence of the last part of the theorem.

Moreover, the two controls together entail that the random metric space generated by $\Pi$ is homeomorphic to $\mathbb{R}^d$.

We notice that Euclidean balls are very tightly controlled in $\Pi$-distance, whereas the control of $\Pi$-balls in Euclidean distance is much looser. Indeed, since the probability that a line with speed $v$ is close to a point $x$ scales like $v^{-(\gamma-1)}$, the maximum polynomial moment of the Euclidean diameter of the $\Pi$-ball cannot be better than $\gamma - 1$. This is linked to the fractal nature of the process.

Almost sure uniqueness of geodesics in dimension 2 has been obtained by Kendall (2014). Notice that the "almost sure" is not for all points simultaneously: the cut-locus is not empty.

In dimension at least 3, we get uniqueness via the technical notion of *many directions*, which formalises the idea that a set of lines close to a given point $x$ have so many different unit vectors that the only way to touch them all with a finite path is to touch them near $x$.

**Definition 3.3** (Jonas Kahn, 2015, Definitions 4.3 and 4.4)**.** *For a set of lines $\mathcal{L} = \{l_j\}_{j\in\mathcal{J}}$ and a subset $X$ of $\mathbb{R}^d$, a* $\mathcal{L}$-tour *in $X$ is a continuous curve $f$ in $X$ such that for all $j \in \mathcal{J}$, there is a $t_j$ such that $f(t_j) \in l_j$. The tour is* finite *if $f$ is rectifiable, else it is* infinite*.*

*A $\Pi$-path $\xi$ has* many directions *near a point $x$ if, for all $\varepsilon > 0$, all $\mathcal{L}_\xi$-tours in $\mathbb{R}^d \setminus B(x, \varepsilon)$ are infinite.*

The point of the notion is that, if $\xi$ has many directions near $x$, then all closed finite $\mathcal{L}_\xi$-tours contain $x$. Moreover:

**Lemma 3.4** (Jonas Kahn, 2015, Lemme 4.6)**.** *Let $d \geq 3$. Let $l$ be a fixed line independent on $\Pi$. Almost surely, for all $x \in l$, for all $y \notin l$, the geodesics $g_{xy}$ all have many directions near $x$.*

The proof of the lemma is quite technical since it applies simultaneously to uncountably many points whereas it is not true for all points in $\mathbb{R}^d$.

Since the lines in $\Pi$ come from a Poisson point process, each line $l$ is independent of $\Pi \setminus \{l\}$, which follows the same law as $\Pi$. Noticing that almost surely all geodesics between $x$ and $y$ are supported by the same lines, Lemma 3.4 implies that the ends of each segment of a geodesic are contained in all the other geodesics, and hence that they are the same. This yields Property 1 of a SIRSN.

The last property of a SIRSN, Property 4, follows from the pigeonhole principle.

To establish Equation (18), we bound the total intersection length of all geodesics between pair of points of $\Xi$, minus radius-1 balls around their ends, with the ball $B(0, \frac{1}{3})$. This length is $\ell = m_1(\mathcal{F} \cap B(0, \frac{1}{3}))$, where $m_1$ is the Hausdorff measure.

If $x$ or $y$ are in the ball $B(0, \frac{2}{3})$, the intersection of $g_{xy} \setminus (B(x, 1) \cup B(y, 1))$ with $B(0, \frac{1}{3})$ is empty. Hence the geodesics $g_{xy}$ that contribute to $\ell$ have the following form, illustrated in Figure 8:

— $x$ and $y$ are outside $B(0, \frac{2}{3})$.

— They hit $B(0, \frac{2}{3})$ for the first time at a point $s$ on the corresponding sphere.

— They hit $B(0, \frac{1}{3})$ for the first time at a point $u$ on the corresponding sphere.

— They hit $B(0, \frac{1}{3})$ for the last time at a point $v$ on the corresponding sphere.

— They hit $B(0, \frac{2}{3})$ for the last time at a point $z$ on the corresponding sphere.

Hence the geodesic $g_{xy}$ must cross twice the *border*, that is $B(0, \frac{2}{3}) \setminus B(0, \frac{1}{3})$, once between $s$ and $u$, and once between $v$ and $z$. Each of these crossings has Euclidean length at least $\frac{1}{3}$.

On the other hand, Theorem 3.2 gives a tight probabilistic bound on the $\Pi$-distance $T$ between $s$ and $z$. We also can bound with high probability the number of lines faster than $\frac{1}{6T}$ that hit the ball $B(0, \frac{2}{3})$. Let us write $\mathcal{V}$ for the intersection of those lines with $B(0, \frac{2}{3})$, which is a set of dimension 1. The intersection of $g_{xy}$ with lines slower than $\frac{1}{6T}$ has length at most $\frac{1}{6}$. So that $g_{xy}$ intersects the fast lines between $s$ and $u$ on a length at least $\frac{1}{6T}$, and again between $v$ and $z$. As a consequence, the set of pair of points $(t, w) \in \mathcal{V}^2$ with $t$ on the geodesic $g_{xy}$ between $s$ and $u$ and $w$ on $g_{xy}$ between $v$ and $z$, has a two-dimensional Hausdorff measure of $\frac{1}{36}$ at least.

Since we can bound the measure of $\mathcal{V}^2$, we can find a finite set $\{g^i\}$ of geodesics such that any other geodesic $g_{xy}$ intersects a $g^i$ between $s$ and $u$, and between $w$ and $z$. By uniqueness of the geodesics, it coincides with $g^i$ between $u$ and $v$. So that $\mathcal{F} \cap B(0, \frac{1}{3})$ is the union of the intersection of the finite number of geodesics $g^i$ with $B(0, \frac{1}{3})$. Precise calculations yield a moment with exponential form:

**Theorem 3.5** (Jonas Kahn, 2015, Theorem 6.1)**.** *Let $\gamma > d \geq 2$. Let $\ell = m_1(\mathcal{F} \cap B(0, \frac{1}{3}))$ be the length of the long-distance network in $B(0, \frac{1}{3})$. For all $\varepsilon < \varepsilon_{max}$, with probability $1 - \varepsilon$, this length is less than $C(\ln(C_1/\varepsilon))^2$, where the constants $\varepsilon_{max}$, $C$ and $C_1$ only depend on $\gamma$ and $d$. Hence the*

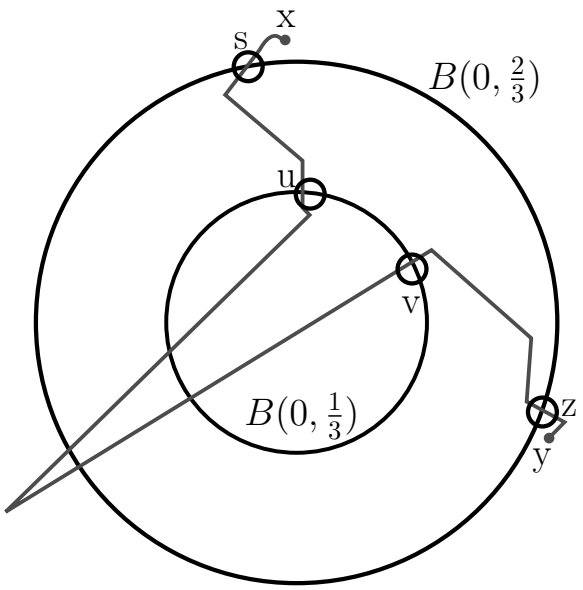



Figure 8 – The geodesic from $x$ to $y$ hits $B(0, \frac{2}{3})$ for the first time at $s$ and the last time at $z$. It hits $B(0, \frac{1}{3})$ for the first time at $u$ and the last time at $v$.

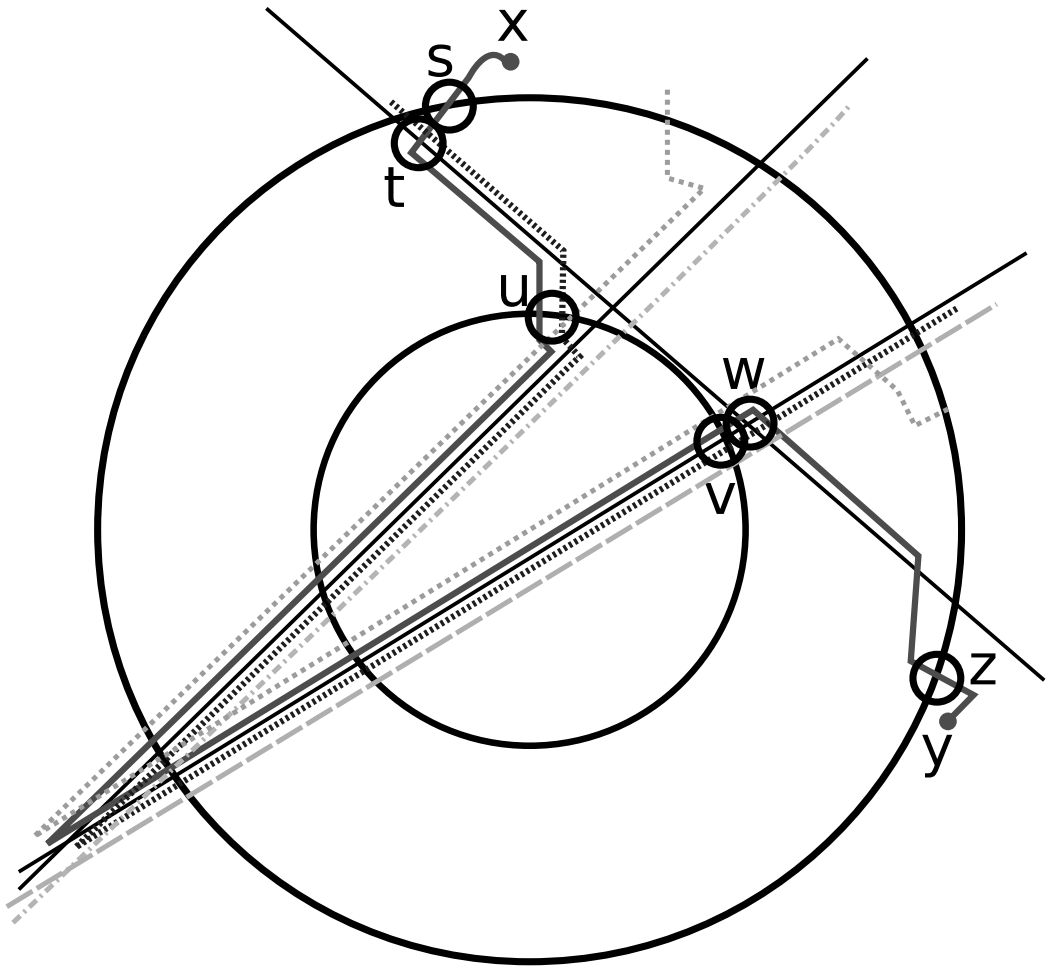



Figure 9 – Close parallel curves in the figure agree. They are separated to make the figure more readable. The three thin black lines are the fast lines. The light dashed lines and black dotted curves are a family of geodesics $\{g^i\}$. The solid curve connecting $x$ and $y$ is the geodesic $g_{xy}$.
The geodesic $g_{xy}$ has a common point $t$ with the black dotted geodesic on a fast (black) line when first crossing the border, and another $w$ when last crossing the border. Hence thay agree between $u$ and $v$.
Any other geodesic contributing to $\ell$ would meet one of the geodesics in the family $\{g^i\}$ in the same way.

*moment, for all* $\delta < \sqrt{C}$,

$$\mathbb{E}\left[\exp\left(\delta\sqrt{\ell}\right)\right] < \infty. \tag{19}$$

*In particular,* $\ell$ *has finite mean, and the improper Poisson line process is a SIRSN.*

### 3.2.3 Connections

Since the improper Poisson line process is a random metric space, it might be worth to compare it with the Brownian map (Le Gall, 2014, for example).

The Brownian map is a random metric on the sphere $\mathbb{S}^2$. It is still homeomorphic to $\mathbb{S}^2$, but has Hausdorff dimension 4. Its behaviour is very hyperbolic: the set of all points in the inside of all its geodesics has Hausdorff dimension 1. Moreover its cut-locus starting from a point is a tree with dimension 2.

The improper Poisson line process is a random metric on $\mathbb{R}^d$. It is homeomorphic to $\mathbb{R}^d$. Its Hausdorff dimension is $(d\gamma - d)/(\gamma - d)$, which is larger than $d$. Notice that for $d = 2$ and $\gamma = 3$, we get the dimensions of the Brownian map. Moreover, if we can prove that any geodesic can be approximated by geodesics between points of the process $\Xi$, the set of all points in the inside of all its geodesics has Hausdorff dimension 1. I have no idea about the cut-locus.

### 3.2.4 Perspectives

I would like to check that the Brownian plane, that is the Gromov-Hausdorff tangent cone to the Brownian map, is a SIRSN. Specifically, since the definition of a SIRSN requires a map from $\mathbb{R}^2$ into the space of interest, I would like to prove it for a well-chosen limit of the random triangulations mapped to the plane through circle packing. Invariances in particular should be easy to get.

Besides, what happens if we drop something else than lines, with scaled sizes. Do we still get a SIRSN?

Is there a universal SIRSN?

If there is, what happens if the elements are dropped in other spaces than the Euclidean space, such as a sphere? What kind of metric space do we obtain if we drop scale invariance, and change the proportion of fast lines?

# References


Aldous, D. J. (2014). "Scale-invariant random spatial networks". In: *Electronic Journal of Probability* 19.15, pp. 1–41. URL: arxiv:1204.0817.

Anders, Janet et al. (2010). "Ancilla-driven universal quantum computation". In: *Physical Review A* 82.2, p. 020301.

Arak, T., P. Clifford, and D. Surgailis (1993). "Point-based polygonal models for random graphs". In: *Advances in Applied Probability* 25, pp. 348–372.

Attouch, Hedy, Jérôme Bolte, and Benar Fux Svaiter (2013). "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical Programming* 137.1-2, pp. 91–129.

Beardwood, Jillian, John H Halton, and John Michael Hammersley (1959). "The shortest path through many points". In: *Proc. Cambridge Philos. Soc.* Vol. 55. 4. Cambridge Univ Press, pp. 299–327.

Bigot, Jérémie, Claire Boyer, and Pierre Weiss (2016). "An analysis of block sampling strategies in compressed sensing". In: *IEEE Transactions on Information Theory* 62.4, pp. 2125–2139.

Boyer, Claire et al. (2016). "On the generation of sampling schemes for Magnetic Resonance Imaging". In: *Accepté à SIAM Imaging Science*.

Candès, Emmanuel J and Carlos Fernandez-Granda (2014). "Towards a Mathematical Theory of Super-resolution". In: *Communications on Pure and Applied Mathematics* 67.6, pp. 906–956.

Candès, Emmanuel J, Justin Romberg, and Terence Tao (2006). "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *Information Theory, IEEE Transactions on* 52.2, pp. 489–509.

Chalfie, Martin et al. (1994). "Green fluorescent protein as a marker for gene expression". In: *Science* 263.5148, pp. 802–805.

Chauffert, Nicolas, Philippe Ciuciu, Jonas Kahn, and Pierre Weiss (2014). "Variable density sampling with continuous trajectories. Application to MRI." In: *SIAM Journal of Imaging Sciences* 7.4, pp. 1962–1992. URL: https://hal.inria.fr/hal-00908486.

— (2016). "A projection algorithm on measure sets". In: *Constructive Approximation*, pp. 1–29.

Chauffert, Nicolas, Philippe Ciuciu, Jonas Kahn, and Pierre Armand Weiss (2013). "Travelling salesman-based variable density sampling". In: *10th international conference on Sampling Theory and Applications (SampTA 2013)*. Bremen, Germany, pp. 509–512.

Chauffert, Nicolas, Pierre Weiss, et al. (2016). "Gradient waveform design for variable density sampling in Magnetic Resonance Imaging". In: *IEEE Transactions on Medical Imaging* 35.9, pp. 2026–2039.

Coste, Michel (2000). *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa.

Deely, J. J. and R. L. Kruse (1968). "Construction of Sequences Estimating the Mixing Distribution". In: *The Annals of Mathematical Statistics* 39.1, pp. 286–288.

Deutsch, David (1985). "Quantum theory, the Church-Turing principle and the universal quantum computer". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 400. 1818. The Royal Society, pp. 97–117.

Deutsch, David and Richard Jozsa (1992). "Rapid solution of problems by quantum computation". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 439. 1907. The Royal Society, pp. 553–558.

Diaconis, Persi (1988). "Group representations in probability and statistics". In: *Lecture Notes-Monograph Series* 11, pp. i–192.

— (1996). "The cutoff phenomenon in finite Markov chains". In: *Proceedings of the National Academy of Sciences* 93.4, pp. 1659–1664.

Diaconis, Persi and James Allen Fill (1990). "Strong stationary times via a new form of duality". In: *The Annals of Probability*, pp. 1483–1522.

Duncan, RR et al. (2004). "Multi-dimensional time-correlated single photon counting (TCSPC) fluorescence lifetime imaging microscopy (FLIM) to detect FRET in cells". In: *Journal of microscopy* 215.1, pp. 1–12.

Fair, Damien A et al. (2007). "Development of distinct control networks through segregation and integration". In: *Proceedings of the National Academy of Sciences* 104.33, pp. 13507–13512.

Feynman, Richard P (1982). "Simulating physics with computers". In: *International journal of theoretical physics* 21.6/7, pp. 467–488.

Fill, J. and J. Kahn (2013). "Comparison Inequalities and Fastest-Mixing Markov Chains". In: *Annals of Applied Probability* 23.5, pp. 1778–1816. URL: `arxi:1109.6075`.

Förster, Th (1948). "Zwischenmolekulare energiewanderung und fluoreszenz". In: *Annalen der physik* 437.1-2, pp. 55–75.

Genovese, Christopher R. and Larry Wasserman (2000). "Rates of convergence for the Gaussian mixture sieve". In: *Ann. Statist.* 28.4, pp. 1105–1127.

Gower, John C and GJS Ross (1969). "Minimum spanning trees and single linkage cluster analysis". In: *Applied statistics*, pp. 54–64.

Gromov, Mikhail (2003). "Random walk in random groups". In: *Geometric and Functional Analysis* 13.1, pp. 73–146.

Grover, Lov K (1996). "A fast quantum mechanical algorithm for database search". In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. ACM, pp. 212–219.

Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109.

Heinrich, Philippe and Jonas Kahn (2015). "Optimal rates for finite mixture estimation". In: *arXiv preprint arXiv:1507.04313; submitted to Annals of Statistics*.

Heinrich, Philippe, Jonas Kahn, et al. (2011). *Remarks on the statistical study of protein-protein interaction in living cells*. URL: arXiv:1105.5738.

Heinrich, Philippe, Mariano Gonzalez Pisfil, et al. (2014). "Implementation of Transportation Distance for Analyzing FLIM and FRET Experiments". In: *Bulletin of mathematical biology* 76.10, pp. 2596–2626.

Holroyd, Alexander E (2011). "Some circumstances where extra updates can delay mixing". In: *Journal of Statistical Physics* 145.6, pp. 1649–1652.

Hoyos-Idrobo, Andrés et al. (2016). "Recursive nearest agglomeration (ReNA): fast clustering for approximation of structured signals". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. URL: https://hal.inria.fr/hal-01366651.

Johnson, William B and Joram Lindenstrauss (1984). "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary mathematics* 26.189-206, p. 1.

Joulin, Aldéric, Yann Ollivier, et al. (2010). "Curvature, concentration and error estimates for Markov chain Monte Carlo". In: *The Annals of Probability* 38.6, pp. 2418–2442.

Kahn, Jonas (2014). "How many T-tessellations on k lines? Existence of associated Gibbs measures on bounded convex domains". In: *Random Structures & Algorithms*, n/a–n/a. ISSN: 1098-2418. DOI: 10.1002/rsa.20557. URL: http://dx.doi.org/10.1002/rsa.20557.

— (2015). "Improper poisson line process as SIRSN in any dimension". In: *arXiv preprint arXiv:1503.03976; submitted to Annals of Probability*.

Kendall, W. S. (2014). "From random lines to metric spaces". In: *Annals of Probability, to appear*. URL: arxiv:1403.1156v1.

Kiêu, K. et al. (2013). "A completely random T-tessellation model and Gibbsian extensions". In: *ArXiv e-prints*. arXiv: 1302.1809 [math.ST].

Lakowicz, Joseph R (2013). *Principles of fluorescence spectroscopy*. Springer Science & Business Media.

Le Gall, J.-F. (2014). "Random geometry on the sphere". In: *ArXiv e-prints*. arXiv: 1403.7943 [math.PR].

Lovász, László and Peter Winkler (1995). "Mixing of Random Walks and Other Diffusions on a Graph". In: *Survey in Combinatorics*. Vol. 218. Lecture Note Series. Cambridge University Press, pp. 119–154.

Lustig, Michael, David Donoho, and John M Pauly (2007). "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic resonance in medicine* 58.6, pp. 1182–1195.

Marshall, Albert W. and Ingram Olkin (1979). *Inequalities: theory of majorization and its applications*. New York: Academic Press Inc. [Harcourt Brace Jovanovich Publishers], pp. xx+569. ISBN: 0-12-473750-1.

McLachlan, G. and D. Peel (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Morimae, T. and J. Kahn (2010). "Entanglement-fidelity relations for inaccurate ancilla-driven quantum computation". In: *Phys. Rev. A* 82, p. 052314. URL: `arXiv:1011.3806`.

Nielsen, Michael A and Isaac L Chuang (2010). *Quantum computation and quantum information*. Cambridge university press.

Peres, Yuval and Peter Winkler (2013). "Can extra updates delay mixing?" In: *Communications in Mathematical Physics* 323.3, pp. 1007–1016.

Rahimi, Ali and Benjamin Recht (2007). "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*, pp. 1177–1184.

Raussendorf, Robert and Hans J Briegel (2001). "A one-way quantum computer". In: *Physical Review Letters* 86.22, p. 5188.

Rebafka, Tabea (2009). "ESTIMATION DANS LE MODÈLE D'EMPILEMENT AVEC APPLICATION AUX MESURES DE LA FLUORESCENCE RÉSOLUE EN TEMPS". PhD thesis. Télécom ParisTech.

Shor, Peter W (1994). "Algorithms for quantum computation: Discrete logarithms and factoring". In: *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*. IEEE, pp. 124–134.

Simon, Daniel R (1997). "On the power of quantum computation". In: *SIAM journal on computing* 26.5, pp. 1474–1483.

Stoyan, D., W. S. Kendall, and Joseph Mecke (1996). *Stochastic Geometry and Its Applications*. WILEY.

Teng, Shang-Hua and Frances F Yao (2007). "K-nearest-neighbor clustering and percolation theory". In: *Algorithmica* 49.3, pp. 192–211.

Teuber, Tanja et al. (2011). "Dithering by differences of convex functions". In: *SIAM Journal on Imaging Sciences* 4.1, pp. 79–108.

Thirion, B. et al. (2015). "Fast clustering for scalable statistical analysis on structured images". In: *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*.

Thomson, Joseph John (1904). "XXIV. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 7.39, pp. 237–265.

Waharte et al., François (2006). "Setup and Characterization of a Multi-photon FLIM Instrument for Protein-Protein Interaction Measurement in Living Cells". In: *Cytometry Part A* 69A, pp. 299–306.

Ward, Joe H (1963). "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301, pp. 236–244.

Zalesky, Andrew et al. (2014). "Time-resolved resting-state brain networks". In: *Proceedings of the National Academy of Sciences* 111.28, pp. 10341–10346.