A constructive method to minimize couple matchings

Pierre Bertrand, Michel Broniatowski, Jean-François Marcotorchino

Journal of Machine Learning Research 1 (2022) 1-48 Submitted 4/00; Published 10/00

A constructive method to minimize couple matchings

Pierre Bertrand PIERRE.BERTRAND@ENS-CACHAN.FR
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université
Paris, France

Michel Broniatowski MICHEL.BRONIATOWSKI@SORBONNE-UNIVERSITE.FR
CNRS UMR 8001
Sorbonne Université
Paris, France

Jean-François Marcotorchino JFMMARCO3@GMAIL.COM
Institut de statistique
Sorbonne Université
Paris, France

Editor: TODO

Abstract

The authors first recall the existence of a second equilibrium in addition to independence to reduce the information conveyed from the margins to the joined distribution: the so-called indeterminacy. They break down a drawing under indeterminacy into a mixture of three independent couplings. Leveraging on this decomposition they emphasis it is the best construction to reduce couple matchings, meaning, the expected number of equal couples drawn in a row. Besides they notice the Janson Vegelius coefficient is nothing but a deviation to indeterminacy and demonstrate it tends to 0 when the number of modalities increases. Eventually, they notice that the indeterminacy appears in two problems (Guessing and Task Partitioning) where couple matchings reduction is a key objective.

Keywords: Mathematical Relational Analysis, Optimal Transport, Logical Indeterminacy, Coupling Functions, Task Partitioning Problem, Guessing Problem

1. Introduction

In a precedent paper Bertrand et al. (2022), we highlighted a list of structural analogies between two discrete couplings namely independence and indeterminacy together with an application to graph clustering.

A discrete coupling is a function C operating on two discrete marginal laws $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ and which defines a probability law π on the product space:

$$\pi_{u,v} = C(\mu_u, \nu_v),\ \forall\ 1 \leq u \leq p,\ 1 \leq v \leq q$$

We respectively quote both those above mentioned couplings $C^{\times}$ (independence) and C^{+} (indeterminacy); this last notion has been initially introduced by J.-F. Marcotorchino in his seminal papers Marcotorchino (1984) and Marcotorchino and Conde-Céspedes (2013)) while their formula will be reintroduced and rediscussed later on in section 2.

Their usefulness arises in statistical applications: namely, most of our usual statistical deviation criteria for contingency analysis are expressed in terms of deviations from one of the two couplings (Conde-Céspedes (2013) gathers a classification of them, deviation to independence or deviation to indeterminacy). The most famous example for independence is the χ^2 index, widely used in practice which computes nothing but a deviation to the independence coupling of the empirical margins. Symmetrically the Janson-Vegelius coefficient, initially introduced in Janson and Vegelius (1977) as a contingency association index, measures also a deviation no longer to independence but rather to indeterminacy; we shall detail this point in subsection 4.1 before studying its global behavior.

Each criteria computes a deviation to a null hypothesis a so-called equilibrium. Furthermore, theoretical considerations lead to consider independence and indeterminacy as the only two possible "natural" equilibria: this fact being based upon a work of Csizar Csiszár et al. (1991), a summarized version of which is expressed in Bertrand et al. (2022).

While independence is commonly used and studied in the scientific literature, indeterminacy appears as a lesser known coupling, whose properties have been rarely presented in an explicit way. Section 3 of this paper, is precisely dedicated to the properties implied by indeterminacy.

The innovation of this paper can be stated as follows:

- We show that indeterminacy aims at minimizing couple matching occurrences (notion defined in definition 1): drawing two independent couples under indeterminacy, the probability to have both couples equal is minimized.
- We estimate the probability for a couple of margins uniformly and independently drawn to be eligible for an indeterminacy coupling (property 4).
- We decompose an indeterminacy coupling into a mixture of three independent couplings leading to a constructive drawing. This decomposition enables us to explain the couple matching minimization inherent to indeterminacy. In each of the three options, indeterminacy concentrates a margin on its mode while the other is uniformly and independently drawn (property 5).
- We analyze the Janson Vegelius correlation coefficient whose expression is nothing but a deviation to indeterminacy. Notably, we demonstrate that it tends to 0 in average when the number of modalities increases (property 8).
- We exhibit two applications of indeterminacy (Guessing Problem and Task Partitioning).

The paper is structured as follows.

Section 2 gathers a summarized version of the construction of indeterminacy. The construction is interpreted as a way to reduce couple matchings. In section 3, a first part computes the measure of the space of margins eligible for an indeterminacy coupling ; a second part is dedicated to the decomposition of indeterminacy. This decomposition is, to the best of our knowledge, new, and conveys an interpretation of the initial formula. Section 4 gathers an analysis of the Janson Vegelius coefficient and provides two information problems where the logical indeterminacy coupling appears. First, leveraging on the constructive drawing as well as on the reducing "couple matchings" property, we show that

indeterminacy naturally occurs in solving the so-called "guessing problem" Massey (1994) as well as the "task partitioning problem" Bunte and Lapidoth (2014).

2. Construction of indeterminacy

When we want to couple two marginal laws, the most common and straightforward way to proceed, consists in assuming independence and keep on computations. It is so well integrated in our mindset, that it naturally appears in real life applications, as soon as we want to build fast models up. In statistical analysis, the approach is quite the same: when we use a very classical and usual criterion like the χ^2 index, we are measuring nothing but a deviation to independence.

Thinking about how we first introduced independence, we immediately suggest empirical experiments: let us say if we roll a dice twice, how should we derive the resulting probabilities from a unique dice? Most of us will naturally apply independence coupling: it really relies on empirical experiments. We often considerate it as the null hypothesis in contingency table analysis.

Although being the most natural, it is not, by far, the only existing available coupling method; actually, as introduced by Sklar in Sklar (1973), any copula function will lead to a coupling function acting on two cumulative distribution functions.

In the discrete case, two probability measures $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ represent the initial margins we want to couple. The first one belongs to the simplex S_p of dimension p while the second belongs to S_q of dimension q.

A coupling π of μ and ν appears as an element of S_{pq} whose margins are μ and ν, meaning:

$$\sum_{u=1}^{p} \pi_{u,v} = \nu_v, \ \forall 1 \leq v \leq q \tag{1}$$

$$\sum_{v=1}^{q} \pi_{u,v} = \mu_u, \ \forall 1 \leq u \leq p \tag{2}$$

We quote $\mathcal{L}_{\mu,\nu}$ the subset of S_{pq} whose elements respect Equation (1) and Equation (2). It exactly corresponds to the space of couplings of μ and ν.

2.1 Reducing the information conveyed by the coupling

Among $\mathcal{L}_{\mu,\nu}$, some couplings π convey more information than others. We suppose we want to reduce the available information one can extract out of realizations from π. It remains to say we want π to be as uniform as possible.

The best way to hide information would be to use the uniform law $\mathbb{U}^{pq}$. Though, unless both margins are uniform, it does not belong to $\mathcal{L}_{\mu,\nu}$.

Hence, let us force π to belong to $\mathcal{L}_{\mu,\nu}$ while being as close as possible to $\mathbb{U}^{pq}$. Use of square distance is classical, actually motivated by the mean square error decomposition. We end up looking at

Problem 1 (Minimal Trade Model)**.**
$\min_{\pi \in \mathcal{L}_{\mu,\nu}} \sum_{u=1}^{p} \sum_{v=1}^{q} \left(\pi_{u,v} - \mathbb{U}^{pq}_{u,v}\right)^2$

It happens that we can compute the exact form of the solution (see Bertrand et al. (2022)). It is given by the so-called indeterminacy coupling quoted π^+ and whose formula is given below.

$$\pi^+_{u,v} = (\mu \oplus \nu)_{u,v} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq},\ \forall 1 \leq u \leq p,\ \forall 1 \leq v \leq q \tag{3}$$

This formula is positive if and only if the inequality is valid:

$$\frac{\mu_0}{q} + \frac{\nu_0}{p} - \frac{1}{pq} \geq 0 \tag{4}$$

where $\mu_0 = \min_{\forall 1 \leq u \leq p} \mu_u$ and $\nu_0 = \min_{\forall 1 \leq v \leq q} \nu_u$.

The inequality (4) that margins have to satisfy considerably reduce their choice. To better understand its impact we describe in subsection 3.1 a method to transform any couple of margins into a couple of margins respecting Inequality (4). Furthermore we compute the probability that μ uniformly drawn within S_p and ν uniformly drawn among S_q and independent from μ respect Inequality (4); therefore, we measure the proportion of margins eligible for an indeterminacy coupling.

2.2 Couple matchings minimization

Developing the cost function of Problem 1 we observe it can be simplified such as minimizing:

$$\sum_{u=1}^{p}\sum_{v=1}^{q} \pi^2_{u,v} \tag{5}$$

A first remark is that substituting $\mathbb{U}^{pq}_{u,v}$ by any constant in Problem 1 would have led to the same simplification. Though, interpreting it as a probability measure requires the constant to be $\frac{1}{pq}$.

Equation (5) can be interpreted using the notion of couple matching:

Definition 1. *π being a probability law in the simplex S_{pq}, we draw under it two times independently. It leads to (U_1, V_1) and (U_2, V_2). A couple matching occurs when $U_1 = U_2$ and $V_1 = V_2$.*

Using this notion, Equation (5) is nothing but the probability of a couple matching. It means that π^+ corresponds to the coupling minimizing couple matchings for fixed margins.

We will decompose π^+ in subsection 3.2 to propose a constructive drawing that will actually explain the property of couple matching minimization.

3. Properties of indeterminacy

3.1 Measuring the subset of margins eligible to indeterminacy

The objective is simply to have an idea of the impact on the margins of the restrictions imposed by the constraints stated in Equation (4).

3.1.1 Coupling μ with μ

We have seen that if we give ourselves μ a probability law on p discrete values, therefore belonging to the simplex S_p, it is not always possible to construct π^+ coupling μ with itself. For this, the pair (μ,μ) must satisfy the hypothesis (4) which is written here:

$$\mu_0 \geq \frac{1}{2p}.$$

We want to estimate the probability that such an event happens. For this, we consider the uniform distribution on S_p, the simplex of all laws on p values. We compute the normalized Lebesque measure of the eligible subset of S_p.

Proposition 1.
The proportion of μ in S_p such that (μ, μ) respects Equation (4) is $\frac{1}{2^{p-1}}$.

Proof
We impose restricted bounds on the integrals constructing μ:

$$\int_{\frac{1}{2p}}^{1-\frac{p-1}{2p}} \int_{\frac{1}{2p}}^{1-\frac{p-2}{2p}-x_1} \dots \int_{\frac{1}{2p}}^{1-\frac{1}{2p}-\sum_{i=1}^{p-2} x_i} \mathrm{d}x_1 \dots \mathrm{d}x_{p-1}$$

With the change of variable: $x_1 \leftarrow x_1 + \frac{1}{2p}$

$$\int_{0}^{\frac{1}{2}} \int_{\frac{1}{2p}}^{1-\frac{p-1}{2p}-x_1} \dots \int_{\frac{1}{2p}}^{1-\frac{2}{2p}-\sum_{i=1}^{p-2} x_i} \mathrm{d}x_1 \dots \mathrm{d}x_{p-1}$$

If we continue with the successive changes of variables: $x_i \leftarrow x_i + \frac{1}{2p}$

$$\int_{0}^{\frac{1}{2}} \int_{0}^{\frac{1}{2}-x_1} \dots \int_{0}^{\frac{1}{2}-\sum_{i=1}^{p-2} x_i} \mathrm{d}x_1 \dots \mathrm{d}x_{p-1}$$

This is therefore exactly the definition of a probability law which would add to $\frac{1}{2}$ instead of 1, hence the result since each component is then multiplied by $\frac{1}{2}$ and the last is imposed by the sum at 1. ■

Remark 1.
The previous result is not surprising. A constructive method exists to build a valid μ. Indeed, by Inequality (4), μ_u is greater than $\frac{1}{2p}$ for all u. We deduce: $\mu_u = \frac{1}{2p} + \frac{r_u}{2}$ where r is an arbitrary probability law on p elements.

3.1.2 Coupling μ with an independent ν

So far, we have assumed to couple μ with itself, which simplified the computations but unnecessarily narrowed the problem. In fact, in the rest of the document, we couple μ with some other law ν. Following the same assumptions as before, we draw μ and ν uniformly and independently among the probability laws, therefore among S_p and S_q respectively. Defining

$$\alpha = p\mu_0, \tag{6}$$

Inequality (4) results from the following proposition.

Proposition 2 (Construction of eligible margins, discrete case)**.**
We can couple the margins (μ, ν) according to the indeterminacy if and only if there exists a positive real α such that:

$$\forall 1 \leq u \leq p,\ \mu_u \geq \frac{\alpha}{p}$$
$$\forall 1 \leq v \leq q,\ \nu_v \geq \frac{1-\alpha}{q}$$

As a first remark, α being the minimum of a set of p elements summing to 1, the numerator of the second inequality is indeed in $[0, 1]$, which implies that ν is greater than some $\beta = 1 - \alpha$ in $[0, 1]$. A second remark concerns the introduction of the variable p in the definition; it is there so as not to break the symmetry between μ and ν, ensuring that all the values α of $[0, 1]$ are eligible regardless of p or q.

Leveraging on Remark 1, we deduce the existence of two probability laws r and s on p and q elements such as:

$$\forall 1 \leq u \leq p,\ \mu_u = \frac{\alpha}{p} + (1-\alpha) r_u \tag{7}$$
$$\forall 1 \leq v \leq q,\ \nu_v = \frac{1-\alpha}{q} + \alpha s_v \tag{8}$$

Proposition 3 (Constructive eligible margins)**.**
A couple of probability laws $(\mu, \nu) \in S_p \times S_q$ respects Inequality (4) if and only it it exists a real $\alpha \in [0, 1]$ and a couple of probability laws $(r, s) \in S_p \times S_q$ such that Equations (7) and (8) are satisfied.

For fixed α, the eligible proportion of the space S_p of μ is therefore $(1 - \alpha)$, that of ν in S_q is α. Since the two laws are drawn independently, the eligible proportion is the product of both. Finally, the eligible proportion in the space $S_p \times S_q$ is given by:

$$\int_{\alpha=0}^{1} \alpha^{p-1}(1-\alpha)^{q-1}\,\mathrm{d}\alpha = \frac{(p-1)!(q-1)!}{(p+q-2)!} \tag{9}$$

The eligibility results are summarized in the following proposition:

Proposition 4 (Valid proportion)**.**
If μ is drawn in the simplex S_p uniformly, the probability that the pair (μ, μ) respects the

Inequality (4) is $\frac{1}{2^{p-1}}$*. Then, there exists a probability law* r *in the simplex* S_p *such that* μ *satisfies:*

$$\forall u, \mu_u = \frac{1}{2p} + \frac{r_u}{2}. \tag{10}$$

If additionally ν *is drawn in* S_q*, independently upon* μ *then, the probability that the pair* (μ, ν) *respects Inequality(4) is* $\frac{(p-1)!(q-1)!}{(p+q-2)!}$*. In this case, there exists a real* α*, a probability law* r *in the simplex* S_p *and a probability law* s *in the simplex* S_q *such that:*

$$\forall u,\ \mu_u = \frac{\alpha}{p} + (1-\alpha) r_u \tag{11}$$

$$\forall v,\ \nu_v = \frac{1-\alpha}{q} + \alpha s_v \tag{12}$$

In addition, the previous writings characterize compliance Inequality (4).

Remark 2 (Different shapes)**.**
We notice that the expression of the eligible proportion depends on whether we are interested in the coupling of μ *with itself or with a second and independent law* ν*: the second formula does not catch up with the first if* $p = q$*. The difference comes from independency only holding in the second case. Indeed* μ *imposes on itself* $\alpha = 1 - \alpha$ *generating a single case, backwards of the integration giving the general formula of the second case which relies on independence.*

3.2 Indetermincay as a mixture of three independent couplings

The formula which defines indeterminacy given in Equation (3) does not provide as such an efficient way to draw under indeterminacy nor any interpretation of its meaning. We propose to rewrite this formula so as to view indeterminacy as a classic mixture of three independent couplings. Our starting point is the usual form of an indeterminacy coupling.

$$\pi^+_{u,v} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq},\ \forall 1 \leq u \leq p,\ \forall 1 \leq v \leq q$$

Quoting $\mu_0 = \min_u \mu_u$ and $\nu_0 = \min_u \nu_u$ it rewrites:

$$\pi^+_{u,v} = \left[\frac{\mu_u - \mu_0}{q}\right] + \left[\frac{\nu_v - \nu_0}{p}\right] + \left[\frac{\mu_0}{q} + \frac{\nu_0}{p} - \frac{1}{pq}\right]$$

First let us remark that the three parts between square brackets are positive since Equation (4) is satisfied. Thus, we renormalize them to extract probability laws. Formally:

$$\pi^+_{u,v} = (1 - p\mu_0)\left[\frac{\mu_u - \mu_0}{q(1-p\mu_0)}\right] + (1 - q\nu_0)\left[\frac{\nu_v - \nu_0}{p(1-q\nu_0)}\right] + (p\mu_0 + q\nu_0 - 1)\left[\frac{1}{pq}\right] \tag{13}$$

Remark 3 (Tight case)**.** *In case any of the two first brackets equals* 0 *it means* μ *or* ν *is uniform. In that case indeterminacy and independence couplings are the same so that an interpretation of indeterminacy is trivial. In case the third bracket is null, it means Inequality (4) is sharp. Anticipating on the action of* T *defined below, it means it drops the whole uniform part of each margin leading to* $R = 3$ *never happening.*

We now define a transformation T on a probability law by:

Definition 2. *Given a probability law $s = s_1, \ldots, s_r$ on r elements, we quote s_0 its minimum. The transformation T^r generates a new law on the same elements by:*

$$\begin{aligned} T^r \quad : \quad & S_r \to S_r \\ & (s_i)_{1\leq i\leq r} \mapsto \left(\frac{s_i - s_0}{1 - r s_0}\right)_{1\leq i\leq r} \end{aligned}$$

We shall quote T the transformation acting on any S_r through $T|_{S_r} = T^r$.

We notice that T actually removes as much uniform part as possible from the probability law it operates on. $T(s)$ will tend to concentrate the realizations on the modes of s.

With this notation, Equation (13) rewrites:

$$\pi^+_{u,v} = (1 - p\mu_0)\frac{1}{q}T(\mu)_u + (1 - q\nu_0)\frac{1}{p}T(\nu)_v + (p\mu_0 + q\nu_0 - 1)\mathbb{U}^{pq}_{u,v} \tag{14}$$

Reading Equation (14), we are able to decompose an indeterminacy draw as stated in proposition 5.

Proposition 5 (Indeterminacy drawing decomposition)**.**
We introduce a random variable R on 3 modalities $1, 2, 3$ with respective probabilities $1-p\mu_0$, $1-q\nu_0$ and $p\mu_0+q\nu_0-1$. Realizations under indeterminacy eventually decomposes as follows:

1. *draw R;*
2. *if $R = 1$ then (u, v) is drawn under the independence coupling of $T(\mu)$ and $\mathbb{U}^q$;*
3. *if $R = 2$ then (u, v) is drawn under the independence coupling of $\mathbb{U}^p$ and $T(\mu)$;*
4. *if $R = 3$ then (u, v) is drawn under the independence coupling of $\mathbb{U}^p$ and $\mathbb{U}^q$ (it corresponds to $\mathbb{U}^{pq}$).*

Under this form, it appears that π^+ exhausts the uniform part of each margin. It is definitely coherent with indeterminacy being the projection of $\mathbb{U}^{pq}$ on $\mathcal{L}_{\mu,\nu}$.

$T(\mu)$ is more concentrated on the modes of μ than μ itself. Consequently when $R = 1$, U is concentrated on the mode of μ, far from the uniform: this is the concession to respect the margin on U. On any other value of R, U is uniformly drawn. Symmetrically, for V, the concentration on modes of ν happens when $R = 2$.

Eventually, Proposition 5 justifies the method induced by indeterminacy to reduce couple matchings. If, $R = 1$, a couple matching is rare since $U_1 = U_2$ is prevented by U being drawn uniformly under U^p; if, $R = 2$, V is drawn uniformly; if $R = 3$, both are drawn uniformly.

The decomposition enables us to interpret it as a mixture of three pretty straightforward drawings as well as to explain how it reduces couple matchings while respecting the forced margins. Indeed, transformation T defined in Definition 2 actually concentrates the probability law it is applied to on its modes. When R defined in Proposition 5 equals 1, $T(\mu)$ concentrates U on the modes of μ to be able to respect the margin μ while when $R = 2, 3$, U is drawn uniformly. This method leverages on V uniform to hide any disequilibrium of U while still avoiding couple matching.

4. Applications of indeterminacy

4.1 Janson Vegelius coefficient

In statistical analysis, given the values of two descriptive variables on n individuals, an usual and important problem is to use a coefficient or index, measuring the correlation between the two variables.

Formally, U represents a first variable which characterizes individuals among p modalities (for instance the city where they are living in, their socio-professional category, their ages, ...); a second variable V classifies them among q categories (or split them into q categories or classes).

Given independent realizations $(U_1, V_1), \ldots, (U_n, V_n)$, the categorization of n individuals, how do we measure the correlation between U and V? Correlation typically means that the value of V depends on the value of U. Expressing it with "dependence" notion, we naturally define a deviation-to-independence coefficient (*i.e.* a departure from independence index), for instance: the χ_2.

To do so, from the n realizations of (U, V), we deduce an empirical margin π counting the proportion of individuals in each couple of modalities:

$$\pi_{u,v} = \frac{\#\{i \ / \ U_i = u \ \& \ V_i = v\}}{n}, \ \forall 1 \leq u \leq p, \ \forall 1 \leq v \leq q \tag{15}$$

similarly, an empiric margin μ is deduced from the empiric π on the first variable and eventually an empiric second margin ν.

The empirical χ^2 index, denoted χ^2_n is defined through:

$$\chi^2_n(U, V) = \sum_{u=1}^{p} \sum_{v=1}^{q} \frac{(\pi_{u,v} - (\mu \otimes \nu)_{u,v})^2}{(\mu \otimes \nu)_{u,v}} \tag{16}$$

which obviously happens to be null if and only if the empirical distribution π of the observed data is an independence coupling of the empiric margins:

$$\pi^+_{u,v} = (\mu \otimes \nu)_{u,v} = \mu_u \nu_v \tag{17}$$

Obviously, such an event almost never occurs, even under independence.

Using a symmetric idea, a lesser known criterion, called Janson-Vegelius Index, after the name of the inventors of this coefficient, who coined it in Janson and Vegelius (1977), Janson and Vegelius (1978) or Janson and Vegelius (1982) writes as a deviation to indeterminacy:

$$JV_n(U, V) = \sum_{u=1}^{p} \sum_{v=1}^{q} \frac{(\pi_{u,v} - (\mu \oplus \nu)_{u,v})^2}{\sqrt{\frac{p-2}{p}\left(\sum_{u=1}^{q} \mu_u^2 + 1\right)} \sqrt{\frac{q-2}{q}\left(\sum_{v=1}^{q} \nu_v^2 + 1\right)}} \tag{18}$$

and obviously is equal to zero if and only if the empirical π is an indeterminacy coupling of the empirical margins as defined in Equation (3).

We omit the subscript n in the following. JV index, although its formulation, using contingency notations appears as non trivial, is actually just a classical cosine, or a Pearson's like correlation coefficient when rewritten in the "Mathematical Relational Analysis" Space. A list of papers which gathers some of the most important key features about the subject

is Marcotorchino and Michaud (1979), Marcotorchino (1984), Messatfa (1990), Opitz and Paul (2005), Marcotorchino (1986), Marcotorchino (1991), Ah-Pine (2010).

The relational analysis space no longer encodes modalities but links between individuals. Two matrices X and Y of size $n \times n$ respectively associated to variables U and $V$are introduced as shown in Definition 3.

Definition 3 (Mathematical Relational Analysis notations)**.**
Let $(U_1, \ldots, U_n)$ and $(V_1, \ldots, V_n)$ be two n probabilistic draws of U and V respectively. We define two associated symmetric $n \times n$ matrices X and Y by

$$X_{i,j} = \mathbb{1}_{U_i = U_j},\ \forall 1 \leq i, j \leq n$$
$$Y_{i,j} = \mathbb{1}_{V_i = V_j},\ \forall 1 \leq i, j \leq n$$

Or in literal form:

- *$X_{i,j} = 1$, if i and j share the same modality of variable U, $X_{i,j} = 0$ if not;*
- *$Y_{i,j} = 1$, if i and j share the same modality of variable V, $Y_{i,j} = 0$ if not.*

To understand the notation, let us begin with some remarks about Definition 3. Basically, the two $\{0, 1\}$ matrices X and Y (which correspond in fact to two binary equivalence relations based on the drawn modalities) represent agreements and disagreements between the two variables on a same draw of size n; they are symmetric with 1 values on their diagonal.

As expected, one can pass from the relational encoding to the usual contingency encoding as well as in the reciprocal way; those transfer formulas are demonstrated in the mentioned articles. Coming back to the JV index, those formulas enable us to write JV as a cosine in the relational space:

$$JV(U,V) = JV(X,Y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)\left(Y_{i,j} - \frac{1}{q}\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{i,j} - \frac{1}{q}\right)^2}} \tag{19}$$

Calculations leading to Equation (19) from Equation (18) can be found in Marcotorchino and Michaud (1979) or Marcotorchino and El Ayoubi (1991).

4.1.1 Average value of JV through simulation

The idea here is to simulate random probability laws π uniformly in S_{p^2} then to calculate the values of the criterion JV on them in order to observe its distribution according to p.

We first propose Figure 1 which presents the distribution of the criterion. One element strikes immediately: values concentrate around 0 as p increases. It is so far an observation and it remains to show it in theory. We will start by demonstrating it in the case $\pi = \mu \otimes \mu$ for which the formula is simplified before demonstrating the general case.

For the moment, we propose to study Figure 2 precisely simulating the two cases to be treated. On the "General" curve, we draw uniformly a probability matrix in S_{p^2} and we calculate the value of JV; the operation is repeated 1000 times and the average is presented. On the "Independence" curve, we draw at random a probability law in S_p which we then couple with itself according to an independence relation; the same number of simulations is applied.

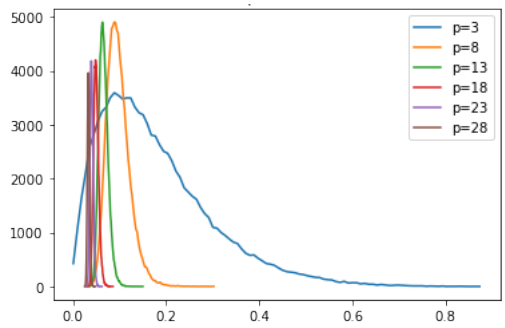

Figure 1: Distribution of the JV when π is uniform in S_{p^2}

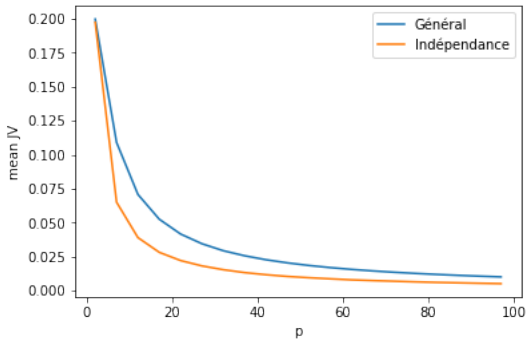

Figure 2: Mean value of JV when π is uniform in S_{p^2} (General) and when $\pi = \mu \otimes \mu$ with μ uniform in S_p

(Independence)

4.1.2 Average value of JV through computations

This section gives a proof (in the case of the independent coupling of μ with itself corresponding to "Independence" in Figure 2) of the limit property noted in the previous section. It is stated in Proposition 6.

Proposition 6 (JV limit, independence case)**.**
If μ is uniform in S_p then $\lim_{p\to\infty} \mathbb{E}_\mu [JV(\mu \otimes \mu)] = 0$

Proof
We start by using a sequence of inequalities allowing us to be interested in a reduced member

of JV.

$$\begin{aligned} JV(\mu \otimes \mu) &= \frac{p^2 \sum_{u=1}^{p} \left(\mu_u - \frac{1}{p}\right)^2 \sum_{u=1}^{p} \left(\mu_u - \frac{1}{p}\right)^2}{p(p-2)(\sum_{u=1}^{p} \mu_u^2) + 1} \\ &= \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p(p-2)(\sum_{u=1}^{p} \mu_u^2) + 1} \\ &\leq \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p^2(p-2)\frac{1}{p^2} + 1} \\ &\leq \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p-1} \\ &\leq 2p \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2 \qquad (20) \end{aligned}$$

$$(21)$$

To demonstrate the convergence, we recall Dirichlet's law as introduced in the Definition 4.

Definition 4 (Dirichlet's law)**.**
The density of Dirichlet's law $\mathcal{D}_p$ which expresses a uniform law on S_p is expressed as follows:

$$f(\mu_1, ..., \mu_p) \prod_{k=1}^{p} \mathrm{d}\mu_k = \frac{1}{B(p)} \prod_{k=1}^{p} \mu_k^0 \prod_{k=1}^{p} \mathrm{d}\mu_k = \frac{1}{B(p)} \prod k = 1^p \mathrm{d}\mu_k$$

where B is the mutinomial beta function.

This law is often presented as a uniform distribution over distributions. This is how we will use it here. Indeed, we notice that the uniform law on the simplex S_p that we have already used for the proof of the Proposition 1 is none other than the particular case $\alpha_1 = \ldots = \alpha_p = 1$.

We specify the moments of this law to deduce the exact calculation of the expectation of the upper bound.

Proposition 7 (Dirichlet's law moments)**.**
Given $\mu \in S_p$ drawn according to Dirichlet's law of parameter $\alpha_1, \ldots, \alpha_p$, we write $\alpha_0 = \alpha_.$. So for all p-uplet $\beta_1, \ldots, \beta_p$ of positive integers, we have the formula (with $\beta_0 = \beta_.$):

$$\mathbb{E}\left(\prod_{u=1}^{p} \mu_u^{\beta_u}\right) = \frac{\Gamma\left(\sum_{u=1}^{p} \alpha_u\right)}{\Gamma\left(\sum_{u=1}^{p} \alpha_u + \beta_u\right)} \prod_{u=1}^{p} \frac{\Gamma(\alpha_u + \beta_u)}{\Gamma(\alpha_u)} = \frac{\Gamma\left(\alpha_0\right)}{\Gamma\left(\alpha_0 + \beta_0\right)} \prod_{u=1}^{p} \frac{\Gamma(\alpha_u + \beta_u)}{\Gamma(\alpha_u)}$$

Lemma 1 (Term tending to 0 and speed)**.**
If μ is uniform in S_p then $\mathbb{E}_\mu \left[\left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2\right] = o\left(\frac{1}{p}\right)$
This is enough to show the convergence in expectation of JV towards 0 according to the upper bound stated in Equation 20.

Remark 4 (Convergence and minimum speed)**.**
Lemma 1 shows that JV approaches its minimum 0 *and further exposes the speed of convergence. It becomes an equality if and only if* $\mu \otimes \mu = \mu \oplus \mu$ *which happens if and only if* $\mu = \mathbb{U}^p$.

Proof
We will essentially develop and use Proposition 7 in the case $\alpha_1 = \ldots = \alpha_p = 1$.

$$\mathbb{E}\left[\left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2\right] = \sum_{1 \leq u,v \leq p} \mathbb{E}(\mu_u^2 \mu_v^2) - \frac{2}{p}\sum_{u=1}^{p} \mathbb{E}(\mu_u^2) + \frac{1}{p^2}$$

For the case where the total power β is 4 separated into 2 in the first term ($u \neq v$):

$$\begin{aligned} \mathbb{E}(\mu_u^2 \mu_v^2) &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+\beta_0)} \frac{\Gamma(\alpha_u+2)\Gamma(\alpha_v+2)}{\Gamma(\alpha_u)\Gamma(\alpha_v)}; \\ &= \frac{4}{p(p+1)(p+2)(p+3)}. \end{aligned} \tag{22}$$

For the case where the total power β is 4 at once in the first term ($u = v$):

$$\begin{aligned} \mathbb{E}(\mu_u^4) &= \frac{2*3*4}{p(p+1)(p+2)(p+3)}; \\ &= \frac{24}{p(p+1)(p+2)(p+3)}. \end{aligned} \tag{23}$$

Finally, for the case where the total power β is 2 in the second term

$$\mathbb{E}(\mu_u^2) = \frac{2}{p(p+1)} \tag{24}$$

In the end, by combining Equations (22), (23) and (24), we get:

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2\right] &= \frac{4p(p-1)}{p(p+1)(p+2)(p+3)} + \frac{24p}{p(p+1)(p+2)(p+3)} - \frac{2*2p}{p(p+1)} + \frac{1}{p^2} \\ &= \frac{4(p-1)}{(p+1)(p+2)(p+3)} + \frac{24}{(p+1)(p+2)(p+3)} - \frac{4p}{(p+1)} + \frac{1}{p^2} \\ &= \frac{4p(p+5) - 4(p+2)(p+3)}{p(p+1)(p+2)(p+3)} + \frac{1}{p^2} \\ &= \frac{-24}{p(p+1)(p+2)(p+3)} + \frac{1}{p^2} \end{aligned}$$

Which allows us to conclude that even multiplied by p^r with $0 < r < 2$, the evaluated term always tends towards 0, in detail we have shown:

$$\mathbb{E}_\mu\left[JV(\mu \otimes \mu)\right] = \mathcal{O}(\frac{1}{p})$$

It thus ends the proof of Lemma 1 and Proposition 6. ■

■

Having shown that JV tends to 0 in the case $\pi = \mu \otimes \mu$ with μ uniform in S_p, we want to show the following more general Proposition.

Proposition 8 (JV limit)**.**
If π is uniform in S_{pq} then $\lim_{pq\to\infty} \mathbb{E}_\pi [JV(\pi)] = 0$

Proof
We show in a similar way to the independence case the inequality

$$JV(\pi) \leq \frac{4}{\sqrt{pq}} \left(pq \sum_{u,v} (\pi_{u,v}^2) - p \sum_u (\pi_{u,\cdot}^2) - q \sum_v (\pi_{\cdot,v}^2) + 1 \right)$$

Let us therefore calculate the expectation of the upper bound using Proposition 7 but this time with a Dirichlet law on the simplex S_{pq}. We thus consider $(\pi_{u,v})_{1\leq u\leq p, 1\leq v\leq q}$ as a vector of size pq.
For the first term

$$\mathbb{E}(\pi_{u,v}^2) = \frac{2}{pq(pq+1)}. \tag{25}$$

For the second term

$$\begin{aligned} \mathbb{E}(\pi_{u,\cdot}^2) &= \mathbb{E}\left(\sum_{v,v'} \pi_{u,v}\pi_{u,v'} \right) \\ &= (q^2 - q)\frac{1}{pq(pq+1)} + q\frac{2}{pq(pq+1)} \\ &= \frac{q+1}{p(pq+1)} \end{aligned} \tag{26}$$

By combining Equations (25) and (26) and playing with symmetry for the third term, we obtain:

$$\begin{aligned} \mathbb{E} & \left(pq \sum_{u,v} (\pi_{u,v}^2) - p \sum_u (\pi_{u,\cdot}^2) - q \sum_v (\pi_{\cdot,v}^2) + 1 \right) \\ &= p^2q^2 \frac{2}{pq(pq+1)} - p^2 \frac{q+1}{p(pq+1)} - q^2 \frac{p+1}{q(pq+1)} + 1 \\ &= \frac{2pq}{pq+1} - \frac{p(q+1)}{pq+1} - \frac{q(p+1)}{pq+1} + 1 \\ &= \frac{-p-q}{pq+1} + 1 \end{aligned}$$

So that we have:

$$\mathbb{E}_\pi [JV(\pi)] = \mathcal{O}\left(\frac{1}{\sqrt{pq}} \right)$$

■

Remark 5. *Since JV is positive we have also shown that it tends towards* 0 *in probability.*

The propensity of the JV to approach 0 assumes that we use it sparingly; if in an application p is large, matching 0 does not assume indeterminacy.

4.2 Guessing or spy problem

This two last subsections, illustrates the couple matching reduction under the guessing and the task partitioning problem. They further detail both than in the original version Bertrand et al. (2021).

The guessing problem appears in cryptography, when a spy screening a communication session tries to determine which message was sent making use of some partial information.

4.2.1 Original problem

In cryptography, a message u in a finite alphabet $\mathcal{U}$ of size p is typically sent from Alice to Bob while a spy whose name is Charlie tries to intercept it. A common strategy for Alice to communicate efficiently and secretly with Bob consists in encoding the message using a couple of keys (public, private) for each character or a symmetric encryption which only requires one shared key between Alice and Bob. The literature concerned with the encryption method to choose according to the situation is diverse, the most-used standard is Advanced Encryption Standard described in various articles. Possibly, Charlie observes an encrypted message V in a second finite alphabet $\mathcal{V}$ of size q which is a function of the message u.

Related to the cryptography situation, the guessing problem quoted hereafter as Problem 2 was first introduced in the article Massey (1994). While in cryptography Charlie tries to decode a sequence of messages, the guessing problem focuses on decoding a unique message. Furthermore, the initial version of Problem 2 is limited due to the lack of access to any prior knowledge by the spy. A second version described in subsection 4.2.2 will introduce a variable V correlated to the message; this second variable will code some information available to Charlie as for example the encrypted message.

Though, the original version provides a collection of results that easily transpose themselves to the more realistic one. Let us formalize this simplest situation: U is a random variable which takes its values in a finite alphabet $\mathcal{U}$ and follows the probability law $\mathbb{P}_U = \mu$. A sender "Alice" generates a sequence of independent messages under μ.

Problem 2 (Original Guessing Problem or Spy Problem)**.**
When Alice sends a message $U = u$ to Bob, the spy Charlie must find out the value u of the realization. He has access to a sequence of formatted questions for any guess $\tilde{u}$ he may have: "Does u equal $\tilde{u}$?" for which the binary answer is limited to "yes/no".

Definition 5 (Original Strategy)**.**
A strategy $S = \sigma$ of Charlie is defined by an order on $\mathcal{U}$ representing the first try, the second and so on until number p. It can be deterministic or random: we quote $\mathbb{P}_S$ its probability law.

Besides, for a given position $i \in [1,p]$, $\sigma[i]$ is the element in $\mathcal{U}$ corresponding to the i-th try.

In Massey (1994), a measure of performance is associated to any fixed strategy σ of Charlie. It basically computes the ρ moment of G which counts the number of trials needed by Charlie to find out which message u was sent. We shall add another performance measure later on.

Definition 6 (Performance measure)**.**
The function $G(\sigma, u)$ is defined as the number of questions required to eventually obtain a "yes" in Problem 2 when Charlie proposed the order $S = \sigma$ and Alice generated the message $U = u$. It can be a random variable even for a fixed u as soon as S is. $G(S,U)$ is a random variable and whose formal definition is:

$$G(\sigma, u) = \sum_{i=1}^{p} i \mathbb{1}_{\sigma[i]=u}$$

We eventually define the efficiency of a strategy S by a measure of the $\rho-$moment of $G(S,U)$ under the independent coupling of $S \sim P_S$ and $U \sim P_U$.

$$||G(S,U)||_\rho = \left[\mathbb{E}_{(S,U,V)\sim\mathbb{P}_{S,U,V}} \left(G(S,U)^\rho\right)\right]$$

The definition of $G(\sigma, u)$ precisely codes the number of trials before Charlie discovers the message u. For instance, with an alphabet $\mathcal{U} = \{a, b, c, d\}$, if the message is $u = c$ and the strategy σ of the spy consists in the order (b, c, a, d) (meaning he first proposes message b then c,...) we have:

$$\begin{aligned} G(\sigma, u) &= \sum_{i=1}^{p} i \mathbb{1}_{u=\sigma[i]} \\ &= 1 \cdot \mathbb{1}_{u=b} + 2 \cdot \mathbb{1}_{u=c} + 3 \cdot \mathbb{1}_{u=a} + 4 \cdot \mathbb{1}_{u=d} \\ &= 2 \cdot \mathbb{1}_{u=c} \\ &= 2 \end{aligned}$$

It has been proven in the same article Massey (1994) a natural result: provided $\mathbb{P}_U = \mu$ is known, the best strategy consists in proposing answers under the deterministic order σ of decreasing probabilities. That is to say we first propose the message which appears most often, then the second most probable and so on:

$$\mu_{\sigma[p]} \leq \ldots \leq \mu_{\sigma[1]}$$

Besides they demonstrated a lower bound on the average number of questions which no strategy can break as it is specified in Theorem 2.

Theorem 2 (Lower bound on the efficiency)**.**
The minimal expected number of questions to solve Problem 2 verifies the inequality:

$$\min_S ||G(S,U)||_\rho \geq (1 + \log(p))^{-\rho} \left[\sum_{u\in\mathcal{U}} \mathbb{P}(U = u)^{\frac{1}{1+\rho}}\right]^{1+\rho}$$

A practical application of Theorem 2 is to provide a guarantee on the average time a spy will take to guess a message. The sender, on its side, is motivated by maximizing the lower bound.

4.2.2 Extended problems

As announced beforehand, Charlie has now access to an observed random variable V correlated with the sent message U. In the common cryptography problem it would be the encrypted message that Charlie observes when Alice sends a message, hence a deterministic function of the message U. Here, we generalize and suppose it can also contain, for instance, the size of the message, the frequency channel used, the sender's location, the receiver, or any physical information a spy can have access to. Finally, the added information, more or less useful, is encoded into a random variable V whose values belong to a finite alphabet $\mathcal{V}$ of size q. Obviously, V is correlated with the message U but we do not suppose their link is deterministic as it would be for an encryption.

As mentioned in the article Arikan (1996), Charlie now chooses its strategy according to the value taken by the observed second variable V: he typically adapts himself to the conveyed encryption. The probability law of the couple (U, V) is quoted $\mathbb{P}_{U,V} = \pi$ while its margins are $\mathbb{P}_U = \mu$ and $\mathbb{P}_V = \nu$.

The gain function now expresses as $G(S, U|V)$: we purposely use the notation symbol "knowing V" to insist on the fact that V is known when the spy decides the strategy he uses. Eventually, for any observed value $V = v$, an original strategy S_v (see Definition 5) is built up leading to an original gain function $G(S_v, U)$ that is to say:

$$G(S, U|V) = \sum_{v \in \mathcal{V}} G(S_v, U) \mathbb{1}_{V=v}$$

The same article comes up with a generalization of Proposition 2 that we report here:

Theorem 3 (Generalized lower bound on the efficiency)**.**
For any strategy, the average time to reconstruct the message always respects the lower bound:

$$\mathbb{E}_{(S,U,V)\sim\mathbb{P}_{S,U,V}}[G(S, U|V)^\rho] \geq (1 + \log(p))^{-\rho} \sum_{v \in \mathcal{V}} \left[\sum_{u \in \mathcal{U}} (\pi_{u,v})^{\frac{1}{1+\rho}} \right]^{1+\rho}$$

Proof [Proof]
The result is plain given that, as we already noticed, S decomposes into original strategies S_v for any fixed v. Hence, for any v, the local or original assigned strategy obeys Proposition 2 which directly leads to the result. ■

4.2.3 Indeterminacy as a lower bound

Let us move away from the literature and measure Charlie's performance by its probability to find out after one trial the message u Alice sent. It is a reasonable measure as, if a sequence of messages is sent, we may have to jump from one to the following after only one trial.

Definition 7 (one-shot performance).
For a given strategy S, we define the following performance measure as the probability to find out the value u after one trial, formally:

$$\begin{aligned} M(S,U,V) &= \mathbb{P}_{(S,U,V)\sim\mathbb{P}_{S,U,V}}\left(S[1]=U\right) \\ &= \sum_{u=1}^{p}\sum_{v=1}^{q}\pi_{u,v}\mathbb{P}_{S_v}\left(S_v[1]=u\right). \end{aligned} \tag{27}$$

Remark 6 (Generalized one-shot performance).
One could easily introduce a measure whose name could be "k-shots performance" evaluating the probability to guess after up to k trials. We would hence notice that if $k \geq p$ then the "k-shots performance" equals 1 for any sensitive strategy. We will not detail it further here.

We suppose as for the original optimal strategy that the spy has access to the distribution $\mathbb{P}_{U,V} = \pi$. We can imagine he previously observed the non-encrypted messages in a preliminary step.

Two strategies immediately stand out:

1. S_{max}: systematically returns at v fixed (observed by hypothesis), the u associated with the maximal probability on the margin $\mathbb{P}_{U|V=v}$

2. S_{margin}: returns at v fixed a random realization of x under the law $\mathbb{P}_{U|V=v}$

Similarly as in see Massey (1994) where they prove S_{max} is the best strategy in case the performance measure is G, we can show it also maximizes the one-shot performance. Actually, since $\mathbb{P}_{S_v}$ only depends on v, M is maximal under S_{max}, when $\mathbb{P}_{S_v} = \delta_{u_v}$ where $u_v = \text{argmax}_u \pi_{u,v}$ so that:

$$M(S,U,V) \geq M(S_{max},U,V) = \sum_{v=1}^{q}\pi_{u_v,v}. \tag{28}$$

Eventually, we quote $u_1 = \text{argmax}_u \mu_u$ and notice that

$$\sum_{v=1}^{q}\pi_{u_v,v} \geq \sum_{v=1}^{q}\pi_{u_1,v} = \mu_{u_1}$$

leading to the proposition 9.

Proposition 9 (Charlie's best performance).
We suppose that the margins μ and ν are fixed. Then, for any coupling probability π between messages U and ciphers V, the best one-shot performance Charlie can perform always happens under S_{max}. Furthermore it admits a fixed lower-bound μ_{u_1} independent on π; to summarize:

$$M(S,U,V) \geq M(S_{max},U,V) = \sum_{v=1}^{q}\pi_{u_v,v} \geq \mu_{u_1}. \tag{29}$$

Let us suppose, commendable task if any, that Alice wants to minimize Charlie's best one-shot performance. We also suppose that the margins μ on U and ν on V are fixed. It is a common hypothesis: the alphabet $\mathcal{U}$ in which the messages are composed typically respects a distribution on letters; variable V on its own, if it represents frequencies for instance may have to satisfy occupation weights on each channel. Eventually, Alice can only leverage on the coupling between U and V.

Precisely, let us now compute the corresponding value for two canonical couplings (independence and indeterminacy). Both are optimal (for Alice).

$$M^{\times} = M(S_{max}, \mu \otimes \nu) = \mu_{u_1} \qquad\qquad M^{+} = M(S_{max}, \mu \oplus \nu) = \mu_{u_1}$$

Regarding the second strategy S_{margin} we know it is less efficient for Charlie in term of one-shot performance. Yet, it is by far harder to cope with for the sender who cannot easily prevent random conflicts. Consequently we come back to the reduction of couple matchings (here a success for Charlie), whose indeterminacy coupling, we know, prevents us against. Let us unfold this remark hereafter.

Replacing $\mathbb{P}_{S_v}$ by its value under the second strategy in Equation (27) allows us to estimate the one-shot performance of S_{margin} which is given by:

$$M(S_{margin}, U, V) = \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \nu_v (\pi_{u|V=v})^2 = \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \frac{\pi_{u,v}^2}{\nu_v}. \tag{30}$$

Concerning the strategy S_{margin} we have the two bounds:

$$\frac{||\pi||_2^2}{\min_{v\in\mathcal{V}} \nu_v} \geq M(S_{margin}, U, V) \geq \frac{||\pi||_2^2}{\max_{v\in\mathcal{V}} \nu_v}, \tag{31}$$

with

$$||\pi||_2 = \sqrt{\sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \pi_{u,v}^2}.$$

Depending on ν the bounds are more or less tight. If ν_v goes to the uniform, the situation becomes similar to S_{max}: both couplings are equal and optimal.

In any case, Equation (31) shows that studying the guessing problem brings us back to Problem 1 associated a square-deviation cost whose solution is given by the indeterminacy coupling of the margins. It guarantees an efficient reduction of conflicts (see section 2.2) and eventually a controlled one-shot performance under S_{margin} as expressed in Equation (31) and an optimal one under S_{max}.

4.3 Tasks partitioning

The task partitioning problem occurs in manufacturing process to optimize the way to assign tasks to production teams, or to machines in job-shop scheduling.

Task partitioning problem is originally introduced in Bunte and Lapidoth (2014) where the authors provide a lower bound on the moment of the number of tasks to perform. Let us follow the gathering work of Kumar et al. (2019) where they also coin a generalized task partitioning problem basically adapting it as a special case of the guessing problem.

Formally, we begin with the original problem of tasks partitioning: a finite set $\mathcal{U}$ of tasks size of which is quoted p is given together with an integer $q \leq p$. The problem consists in creating a partition $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_q)$ of $\mathcal{U}$ in q classes to minimize the number of tasks to perform, knowing that if one needs to perform a task $u \in \mathcal{A}_i$, it is mandatory to launch simultaneously the whole subset of tasks included within $\mathcal{A}_i$.

Practically, a task $U = u$ to perform is randomly drawn from $\mathcal{U}$ under a probability distribution $\mathbb{P}_U = \mu$ representing the tasks frequencies. As any task, the task u to perform is assigned to a unique class $\mathcal{A}_{i(u)}$ of the arbitrary partition. Hence, $A(u) = |\mathcal{A}_{i(u)}|$ counts the number of tasks to perform. Precisely, one plays on the partition knowledge to perform, in average, as few tasks as possible.

Similarly to the guessing problem, the performance of a partition $\mathcal{A}$ is estimated through the $\rho-$moment of $A(U)$, formally $\mathbb{E}_U\left[A(U)^\rho\right]$. Moreover, the authors show in Kumar et al. (2019), quite similarly as for Theorem 2 that we have:

$$\min_{\mathcal{A}} \mathbb{E}_{U\sim\mu}\left[A(U)^\rho\right] \geq \frac{1}{q}\left[\sum_{u\in\mathcal{U}} (\mu_u)^{\frac{1}{1+\rho}}\right]^{1+\rho} \tag{32}$$

which expresses a minimum average number of tasks to perform whatever the partition is.

Inspired by the general guessing problem, they extend the task partitioning problem. Let us introduce here this generalized version, in which we are no longer interested in minimizing the number of tasks to perform but rather in reducing the number of tasks before a selected (or a chosen) task u.

Indeed, in the first version, as soon as u is drawn, an arbitrary rule imposes to perform the whole subset $\mathcal{A}_{i(u)}$ leading to realize $A(u)$ tasks. In the new version, tasks are performed sequentially in $\mathcal{A}_{i(u)}$ according to a global strategy S that can be deterministic or random.

Typically, tasks may consist in a signatures flow which an administration requires while q would be the number of workers dedicated to perform those signatures on incoming documents. A worker can be given the entitlement to perform several signatures, assistants usually do. In that case, the partition encodes the assignments of tasks to workers. When a worker $V = v$ is assigned a document, the depositor waits until the signature. Then the worker follows his own strategy S_v to sign his assigned documents, meaning he can always follow the same order leading to a deterministic strategy or change every day leading to a random strategy.

With a global strategy S which gathers the workers' strategies S_v, $\forall 1 \leq v \leq q$ and for a task u to perform, the performance of a partition $\mathcal{A}$ is measured using

$$N_{S,\mathcal{A}}(u)$$

which represents the number of tasks performed before the intended task u (u included). A lower bound is provided in the paper Kumar et al. (2019).

Let us now suppose the keys $1 \leq v \leq q$ are associated with offices that must perform a proportion ν_v of the incoming tasks which still follow a distribution μ. It actually appears as a sensible problem where a manager would have to distribute in advance tasks among teams according to the usual observed distribution of tasks and a list of available teams with their capacities.

Besides, we suppose each team uses the strategy S_{margin} to perform tasks, meaning they randomly perform one according to their margin theoretical distribution; for a document signing, they randomly sign one.

Remark 7 (Concrete estimated distribution)**.**
In any of the previous applications, for spy as well as for tasks, we are dealing with probabilities. Actually, we send a finite and integer number n of messages and we similarly distribute a finite number n of tasks.

Moreover, $\mathcal{U}$ and $\mathcal{V}$ are finite. Eventually, for any $u \in \mathcal{U}$ and $v \in \mathcal{V}$, an integer number $n_{u,v}$ of tasks is associated corresponding (in the spy problem) to the number of same letters u sent using channel v. To convert $n_{u,v}$ into a probability measure, one would use Equation 15.

Reciprocally, given a probability measure, one will draw n messages according to π. As n increases, it will approximate the theoretical distribution better and better.

From now on, we can rewrite our task partitioning problem under the form of a guessing problem:

- $V = v$, formerly corresponds to a worker, now it represents the information the spy has access to ;
- $U = u$, formerly represents a task to perform, now it represents a sent message ;
- $S = \sigma$, formerly represents the order in which tasks are performed, now it represents the order in which Charlie proposes his guesses.

Under this formalism, we are interested in measuring the probability $M(S, U, V)$ of executing u first as an extended application of the one-shot performance of Definition 7 and we have:

$$M(S, U, V) \geq \frac{||\pi||_2^2}{\max_{1 \leq v \leq q} \nu_v} \geq \frac{||\pi^+||_2^2}{\max_{1 \leq v \leq q} \nu_v} \tag{33}$$

This inequality provides a lower bound for any distribution of the tasks among the team, no distribution can generate a worst "one-shot probability" of satisfying the intended task.

In task partitioning actually, each u is uniquely associated to a worker $v = i(u)$ so that the random variable representing the worker is deterministic conditionally to U. Yet, it is a reducing case of the guessing problem where V is random.

Remark 8 (Splitting mass)**.**
Eventually, we notice that having V random conditionally on U is a generalization of task partitioning along the same lines the Monge-Kantorovith problem was an extension of the one dimension Monge problem: we allow the mass splitting possibility, since a task may be randomly assigned among several workers.

In task partitioning problem using a partition $\mathcal{A}$ instead of V (hence allowing no mass splitting), we notice $\mathbb{P}_Y = \nu$ is not properly defined. Let us extract it from the partition $\mathcal{A}$ by providing each worker with a probability which sums up the probabilities of the tasks he has to perform. Formally, we define $V(\mathcal{A})$ to be a random variable whose probability is:

$$\nu_v^{\mathcal{A}} = \mathbb{P}_{V(\mathcal{A})}(V(\mathcal{A}) = v) = \sum_{u \in \mathcal{A}_v} \mu_u$$

together with the couple probability:

$$\pi_{u,v}^{\mathcal{A}} = \mathbb{P}_{U,V(\mathcal{A})}\left(U = u, V(\mathcal{A}) = v\right) = \mu_u \mathbb{1}_{u \in \mathcal{A}_v}$$

It enables us to deduce that the one-shot probability of satisfying the intended task for a partitioning problem accepts as a lower bound:

$$M(S,U,\mathcal{A}) = M\left(S,U,V(\mathcal{A})\right) \geq \frac{||\pi^{\mathcal{A}}||_2^2}{\max_{1 \leq v \leq q} \nu_v^{\mathcal{A}}} \geq \frac{||C^+(\mu,\nu^{\mathcal{A}})||_2^2}{\max_{1 \leq v \leq q} \nu_v^{\mathcal{A}}}$$

Indeed, a partition problem appears as a particular coupling of U and $V(\mathcal{A})$ (where $V(\mathcal{A})|U$ is deterministic) and no coupling can be worse than $C^+\left(U, V(\mathcal{A})\right)$.

A direct application is that no office affectation should provide a worse one-shot performance...

The efficiency of indeterminacy coupling in guessing problem as well as in task partitioning directly comes from its ability to reduce couple matchings. Either it prevents the spy from discovering the message or it provides a worst strategy by preventing a task from being performed.

5. Conclusion

The main innovation of this paper is the decomposition of indeterminacy. It enables us, first to efficiently generate a drawing, second to interpret it as a mixture of three pretty straightforward drawings and last but not least to explain how it reduces couple matchings while respecting the forced margins. Since indeterminacy cannot be defined on all margins, the paper also computes the proportion of eligible margins. Furthermore, it proposes a constructive method to transform any couple into an eligible couple. Besides, the limit we have show of the Janson Vegelius coefficient helps us to mind when defining a threshold to conclude to indeterminacy. Eventually, the two applications, already presented in proceedings Bertrand et al. (2021) are interpreted to the light of the new decomposition.

Acknowledgments

TODO

References

Julien Ah-Pine. On aggregating binary relations using 0-1 integer linear programming. In *ISAIM*, pages 1–10, 2010.

Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42, 1996.

Pierre Bertrand, Michel Broniatowski, and Jean-François Marcotorchino. Minimization with respect to divergences and applications. In *International Conference on Geometric Science of Information*, pages 818–828. Springer, 2021.

Pierre Bertrand, Michel Broniatowski, and Jean-François Marcotorchino. Independence versus indetermination: basis of two canonical clustering criteria. *Advances in Data Analysis and Classification*, pages 1–25, 2022.

Christoph Bunte and Amos Lapidoth. Encoding tasks and Rényi entropy. *IEEE Transactions on Information Theory*, 60(9):5065–5076, Sep 2014. ISSN 1557-9654. doi: 10.1109/tit.2014.2329490. URL `http://dx.doi.org/10.1109/TIT.2014.2329490`.

Patricia Conde-Céspedes. *Modélisations et extensions du formalisme de l'analyse relationnelle mathématique à la modularisation des grands graphes.* PhD thesis, Paris 6, 2013.

Imre Csiszár et al. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The annals of statistics*, 19(4):2032–2066, 1991.

Svante Janson and Jan Vegelius. Correlation coefficients for nominal scales. *Uppsala: Department of Statistics*, 1977.

Svante Janson and Jan Vegelius. On the applicability of truncated component analysis based on correlation coefficients for nominal scales. *Applied Psychological Measurement*, pages 135–145, 1978.

Svante Janson and Jan Vegelius. The J-index as a measure of nominal scale response agreement. *Applied Psychological Measurement*, pages 111–121, 1982.

M. Ashok Kumar, Albert Sunny, Ashish Thakre, and Ashisha Kumar. A unified framework for problems on guessing, source coding and task partitioning, 2019.

Jean-François Marcotorchino and Patricia Conde-Céspedes. Optimal transport and minimal trade problem, impacts on relational metrics and applications to large graphs and networks modularity. In *International Conference on Geometric Science of Information*, pages 169–179. Springer, 2013.

Jean-François Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences. *Publication du Centre Scientifique IBM de Paris et Cahiers du Séminaire Analyse des Données et Processus Stochastiques Université Libre de Bruxelles*, pages 1–57, 1984.

Jean-François Marcotorchino. Maximal association theory as a tool of research. *Classification as a tool of research , W.Gaul and M. Schader editors, North Holland Amsterdam*, 1986.

Jean-François Marcotorchino. Seriation problems:an overview. *Applied Stochastic Models and Data Analysis*, 7:139–151, 1991.

Jean-François Marcotorchino and Najoi El Ayoubi. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée*, 39:25–46, 1991.

Jean-François Marcotorchino and Pierre Michaud. *Optimisation en Analyse Ordinale des Données*. Ed Masson, Paris, 1979.

James L Massey. Guessing and entropy. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 204. IEEE, 1994.

Hammou Messatfa. Maximal association for the sum of squares of a contingency table. *Revue RAIRO, Recherche Opérationnelle*, 24:29–47, 1990.

Otto Opitz and Henning Paul. Aggregation of ordinal judgements based on condorcet's majority rule. *Data Analysis and Decision Support. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg*, 2005.

Abe Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6): 449–460, 1973.