The Rational Speech Act, A Mathematical Review

B. PLUS, G. BEN ZENOU, P. PIANTANIDA

July 31, 2024

Contents

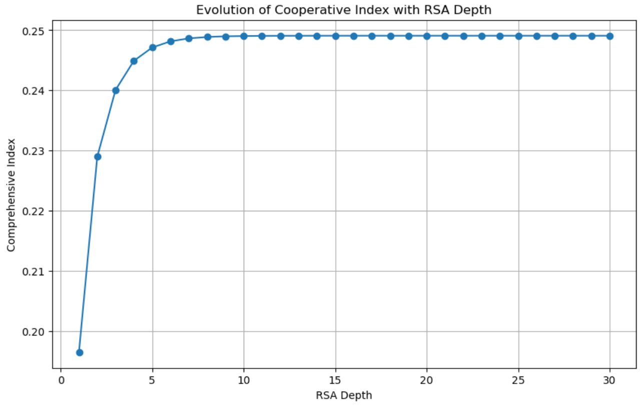


Figure 1: Evolution of CI as depth increases on a 10×10 L_0 matrix

1 Relevance of RSA iterations

In the original [?] paper which first introduced Rational Speech Act (RSA), only one iteration of Equation **??** was considered. But further research studied the effect of iterating *i.e* computing S_t and L_t for $t > 1$ in Equation **??**. As shown in Figure **??** the cooperative index (a simple measure of "how cooperative a pair speaker/listener is") increased with recursion depth. The relevance of considering multiple iteration RSA instead of classic 1-step RSA greatly depend on the problem under study. [?] showed that as matrix size and perturbation increases, speakers (or listeners) obtained after convergence of RSA performed better than when 1 step RSA.

For those reason we study in this paper the convergence of RSA algorithm.

2 RSA is the research of a single matrix

In former litterature, a classic idea is that RSA separetely find speaker and listener matrices S, L, or that RSA try to find a speaker given a specific listener. This formulation is relevant to understand the RSA model as a communication problem. In this paper, we are seeking a mathematical explanation of RSA. We want to find the problem whose unique solution is given by RSA. With this perspective in mind, we can look at RSA algorithm as the search of a single matrix L^* (or S^*) which is only built upon shared knowledge $R \in M_+(\mathcal{U} \times \mathcal{M})$, also called *lexicon*, and *prior* knowledge on meanings $P \in M_+(\mathcal{M})$. This problem is given by Equation **??** whereas classic RSA is more often represented by Equations **??**.

$$\begin{aligned} S_t(u \mid m) &= \frac{e^{\alpha \log L_{t-1}(m|u)}}{\Sigma_{u'} e^{\alpha \log L_{t-1}(m|u)}} \\ L_t(m \mid u) &= \frac{P(m) S_t(u \mid m)}{\Sigma_{m'} P(m') S_t(u \mid m')} \end{aligned} \tag{1}$$

Where $P(m)$ is the prior knowledge on meanings.

$$L_t(m \mid u) = \frac{P(m) \frac{e^{\alpha \log L_{t-1}(m|u)}}{\Sigma_{u'} P(m) e^{\alpha \log L_{t-1}(m|u')}}}{\Sigma_{m'} P(m') \frac{e^{\alpha \log L_{t-1}(m|u)}}{\Sigma_{u'} P(m) e^{\alpha \log L_{t-1}(m|u')}}} \tag{2}$$

This formulation of RSA algorithm highlights the fact that RSA is an iterative problem, defined by recurrence Equations **??** and initial conditions R and P. With that in mind, we will show in the following the existing link between RSA and Optimal Transport theory

3 RSA algorithm is equivalent to Sinkhorn-Knopp method

In this paper we aim at closing a gap in the understanding of RSA. RSA is given as an algorithm which has empirically been proven to mimic human pragmatism. But the algorithm does not come from a known objective function we try to miminize by computing pragmatic listener or speaker through **??**. In other words we do not know what RSA really try to optimize, but we know how to do it. The aim of this paper is to find the problem RSA tackles through iterations. We seek the problem of which RSA is the unique solution.

Sinkhorn-Knopp Algorithm **??** is an iterative method which consists of iteratively normalizing rows and columns of a given matrix A. It is extremely similar to RSA Algorithm **??**. If we look at RSA as previously presented in Equation **??** then RSA algorithm and SK algorithm are exactly the same at the end of each iteration of the loop. We consider here $\alpha = 1$. The case $\alpha \neq 1$ we be treated in other section.

Algorithm 1 SK Method

1: **Given:** Initial matrix $R \in \mathbb{R}_+^{U\times M}$
2: $Q \leftarrow R$
3: **repeat**
4: $\quad \forall i, j \quad P_{ij} \leftarrow \frac{\mu_j Q_{ij}}{\sum_i Q_{ij}}$
5: $\quad \forall i, j \quad Q_{ij} \leftarrow \frac{\nu_j P_{ij}}{\sum_j P_{ij}}$
6: **until** convergence

Algorithm 2 RSA Algorithm

1: **Given:** Initial matrix $R \in \mathbb{R}_+^{U\times M}$
2: $L \leftarrow R$
3: **repeat**
4: $\quad \forall i, j \quad S_{ij} \leftarrow \frac{L_{ij}}{\sum_i L_{ij}}$
5: $\quad \forall i, j \quad L_{ij} \leftarrow \frac{P(m_j)S_{ij}}{\sum_j P(m_j)S_{ij}}$
6: **until** convergence

Figure 2: Comparison of the SK Method and RSA Algorithm

The Sinkhorn method (SK) is generally used to compute solution of the Entropic Optimal Transport problem under certain conditions. Under those conditions, explicited in Theorem **??**. *Entropic Optimal Transport* is another formulation of the so-called *Shrödinger problem* explicited in Definition **??** [**?**], [**?**]. As RSA and SK algorithm are the same, and SK solves Schrödinger bridging problem, it exists a formulation of RSA as a Schrödinger problem, explicited later in the paper. Finding this formualtion will allow us to unlock the rich litterature on both Schrödinger and Optimal Transport to study the Rational Speech Act. Optimal transport is a long-known problem studied in litterature, and the problem took several names and has been studied independently in different fields. We refer to [**?**] for a review of the different formulation of Optimal Transport problem.

4 The Shrödinger problem

We call $\Pi(\mu,\nu)$ the subset of $M_+(\mathcal{U}\times\mathcal{M})$ consisting of all those matrices $\bar{R}$ whose row and column sums give μ and ν respectively, that is, such that

$$\forall i = 1,\dots,U, \quad \sum_j \bar{R}_{ij} = \mu_i, \quad \text{and} \quad \forall j = 1,\dots M, \quad \sum_i \bar{R}_{ij} = \nu_j.$$

In the Shrödinger problem (or equivalently in Entropic Optimal Transport) we are searching for a matrix $\bar{R}$ "close" to R under marginals constraints. To measure how "close" $\bar{R}$ is to R we define the *relative entropy* also called *Kullback-Leibler divergence* or *I-projection* as $H(\bar{R} \mid R) := \sum_{ij}\left\{\bar{R}_{ij}\log\frac{\bar{R}_{ij}}{R_{ij}} + R_{ij} - \bar{R}_{ij}\right\}$. The Schrödinger problem is formulated as follow :

Definition 1 (Schrödinger Problem). *Let $R \in M_+(\mathcal{U}\times\mathcal{M})$ and $(\mu,\nu) \in M_+(\mathcal{U})\times M_+(\mathcal{M})$. We call the Schrödinger problem w.r.t. R between μ and ν the optimization problem consisting in minimizing among $\Pi(\mu,\nu)$ the relative entropy:*

$$\mathrm{Sch}(R;\mu,\nu) := \min_{\bar{R}} H(\bar{R} \mid R) \mid \bar{R} \in \Pi(\mu,\nu) \tag{3}$$

We remind the reader that we have no other assumptions on R and $\bar{R}$ but that they are non-negative. We use the conventions $a\log\frac{a}{0} = +\infty$ if $a > 0$, and $0\log 0 = 0\log\frac{0}{0} = 0$.

This problem may not admit a solution, depending on the chosen R, μ and ν. In addition the solution may not be unique. But if a solution $R^* = \arg\min \mathrm{Sch}(R;\mu,\nu)$ exists, it can be reached via Sinkhorn algorithm (see [**?**] [**?**] [**?**] or [**?**] theorem 4.1).

Notations We note $\mathcal{U}$ the set of utterances, $\mathcal{M}$ the set of meanings, U and M their cardinalities. We note $R \in M_+(\mathcal{U}\times\mathcal{M})$ a lexicon, $\mu \in M_+(\mathcal{U})$ and $\nu \in M_+(\mathcal{M})$ prior knowledge on meanings and utterances called *marginals*. We also define $M(\mu) := \sum_i \mu_i$ and $M(\nu) := \sum_j \nu_j$ total *masses* of μ and ν. We say that R is *scalable* w.r.t μ and ν if $\mathrm{Sch}(R;\mu,\nu)$ defined by Definition **??** admits a solution.

To understand the link between Schrödinger problem and RSA we give the following exemple : when the lexicon R is a strictly positive squared matrix and prior knowledge is uniform *i.e* $P(m) = \frac{1}{U} = \frac{1}{M}$, then the problem $\mathrm{Sch}(R;\mathbf{1}_U,\mathbf{1}_U)$ defined by Definition **??** admits a single solution and RSA converges toward a single doubly stochastic matrix $L^* = S^* = R^*$ solution of $\mathrm{Sch}(R;\mathbf{1}_U,\mathbf{1}_U)$.

Theorem 1. *When the problem* $\mathrm{Sch}(R;\mu,\nu)$ *defined in Definition* **??** *admits a solution, then it is unique and the Sinkhorn algorithm* **??** *converges toward the solution*

We give here conditions for existence and uniqueness of the solution to the Schrödinger problem. First we define the following sets :

$$\begin{aligned} \forall A \subset \mathcal{U}, \quad M_R(A) \quad &:= \{m_j \in \mathcal{M} \mid \exists u_i \in A \text{ s.t. } R_{ij} > 0\}, \\ \forall B \subset \mathcal{M}, \quad U_R(B) \quad &:= \{u_i \in \mathcal{U} \mid \exists m_j \in B \text{ s.t. } R_{ij} > 0\}. \end{aligned}$$

Intuitively $M_R(A)$ corresponds to every meaning that utterances in $A \subset \mathcal{U}$ can describe. We also use the notation $\mu(A) = \sum_{a\in A} \mu_a$

Theorem 2 (Existence & uniqueness)**.** *Let $R \in M_+(\mathcal{U}\times\mathcal{M}), \mu \in M_+(\mathcal{U})$ and $\nu \in M_+(\mathcal{M})$. The three following assertions are equivalent (Th. 23 of [?]):*

1. *$M(\mu) = M(\nu)$ and for all $A \subset \mathcal{U}, \mu(A) \leq \nu\left(M_R(A)\right)$.*
2. *$M(\mu) = M(\nu)$ and for all $B \subset \mathcal{M}, \nu(B) \leq \mu\left(U_R(B)\right)$.*
3. *$\mathrm{Sch}(R;\mu,\nu)$ is scalable i.e has a unique solution.*

With $M(.)$ total masses introduced previously. If μ_i (resp. ν_j) describe prior knowledge on probabilities of obtaining u_i (resp. m_j) the condition "for all $A \subset \mathcal{U}, \mu(A) \leq \nu\left(M_R(A)\right)$" can be interpreted as "for all subset A of utterances, $Pr(u_i \in A) \leq Pr(m_j$ for m_j meanings for which u_i is true $(R_{ij} > 0)$"

5 Convergence of Sinkhorn algorithm

In this section we study the convergence of Sinkhorn Algorithm **??**. We are looking for the conditions under which the algorithm converges, the problem it solves and the limits of the convergence. We note $\mu_i^{\bar{R}} = \sum_j \bar{R}_{ij}$ and $\nu_j^{\bar{R}} = \sum_i \bar{R}_{ij}$ marginals associated with $\bar{R}$. We also introduce new marginals :

$$\begin{aligned} \mu^* &:= \arg\min\left\{H(\bar{\mu} \mid \mu) \mid \bar{\mu} = \mu^Q \text{ for some } Q \text{ with } H(Q \mid R) < +\infty \text{ and } \nu^Q = \nu\right\}, \\ \nu^* &:= \arg\min\left\{H(\bar{\nu} \mid \nu) \mid \bar{\nu} = \nu^P \text{ for some } P \text{ with } H(P \mid R) < +\infty \text{ and } \mu^P = \mu\right\}. \end{aligned} \tag{4}$$

Theorem 3. *Let $R \in M_+(\mathcal{U}\times\mathcal{M}), \mu \in M_+(\mathcal{U})$ and $\nu \in M_+(\mathcal{M})$ and the sequences $(P^n)_{n\in\mathbb{N}^*}$ and $(Q^n)_{n\in\mathbb{N}^*}$ given by Algorithm* **??***. If $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$, then (Th.11 of [?])*

$$P^n \underset{n\to+\infty}{\longrightarrow} P^* \quad \text{and} \quad Q^n \underset{n\to+\infty}{\longrightarrow} Q^*.$$

Where μ^* and ν^* are defined by Equations **??***. Moreover*

$$P^* := \arg\mathrm{Sch}(R;\mu,\nu^*) \quad \text{and} \quad Q^* := \arg\mathrm{Sch}(R;\mu^*,\nu) \tag{5}$$

In other words, the Sinkhorn algorithm converges and solves two different Schrödinger problem defined by Equations **??**. To find the Schrödinger problem associated with a given triple (R,μ,ν) we first solve the convex optimization problems Equations **??** to find μ^* and ν^*. The problem solved by Sinkhorn algorithm is then given by Equations **??**.

Remark 1. *The conditions $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$ are equivalent to $\mu_i^R = 0 \Rightarrow \mu_i = 0$ and $\nu_j^R = 0 \Rightarrow \nu_j = 0$ when using conventions specified in Section* **??**.

Remark 2. *If the problem is scalable (conditions of Theorem* **??** *are met) then Equations* **??** *give $\mu^* = \mu$ and $\nu^* = \nu$ and Sinkhorn algorithm converges toward the unique solution of* $\mathrm{Sch}(R;\mu,\nu)$.

6 Central problem

[**?**] studied Sinkhorn algorithm when the Schrödinger problem has no solution *i.e non-scalable* case. Problem may be unsolvable, Sinkhorn algorithm will still converge and give 2 different matrices P^* and Q^* desribed by Equations **??**. It is possible to explicit a relaxed and solvable version of the unsolvable initial problem. The link between the relaxed solution and matrices given by Algorithm**??** are explicited in this section.

We study here situations where conditions of Theorem **??** are not satisfied.

For non-scalable problem the relaxed version is defined by :

$$\mathrm{Sch}^{\varepsilon}(R;\mu,\nu) := \min\left\{\varepsilon H(\bar{R} \mid R) + H\left(\mu^{\bar{R}} \mid \mu\right) + H\left(\nu^{\bar{R}} \mid \nu\right) \mid \bar{R} \in \mathcal{M}_+(\mathcal{D}\times\mathcal{F})\right\} \tag{6}$$

The hard constraints on marginals are replaced a soft constraints making the problem solvable. The parameter ε quantify the level of relaxation. We note R^*_ε the solution of relaxed problem given by Equation **??**.

It can be shown that tR^*_ε converges toward the solution of a Schrödinger problem with balanced marginals.

Theorem 4. *Let $R \in M_+(\mathcal{U}\times\mathcal{M}), \mu \in M_+(\mathcal{U})$ and $\nu \in M_+(\mathcal{M})$ satisfy $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$. Considering P^* and Q^* given by Equations **??**, R^*_ε solution of Equation **??**, we have (Th. 17 of [**?**])*

$$R^*_\varepsilon \underset{\epsilon\to 0}{\to} R^*_{ij} := \sqrt{P^*_{ij}Q^*_{ij}} \tag{7}$$

Moreover R^* is solution of $\mathrm{Sch}(R;\sqrt{\mu^*\mu},\sqrt{\nu^*\nu})$ *with μ^* and ν^* defined by Equations **??***

Intuitively, the relaxed problem will approach the geometric mean of P^* and Q^* as $\epsilon \to 0$ where P^* and Q^* are obtained via Sinkhorn Algorithm **??**.

7 Differences between RSA and Sinkhorn

As shown in Figure **??** Rational Speech Act and Sinkhorn algorithm are really similar but are not exactly the same. If we consider the Schrödinger problem $\mathrm{Sch}(R;\mu,\nu^*)$ with $\mu = \mathbf{1}_U$ and $\nu^*_j = P(m_j)\times U$ then we easily obtain the listener of RSA by replacing μ_i and ν_j in Algorithm **??**. If we note $\mathbf{P}_U$ and $\mathbf{P}_M$ priors on utterances and meanings, the RSA is given by the following algorithm :

Algorithm 3 RSA Algorithm

1: **Given:** $R \in \mathbb{R}_+^{U\times M}, \mu = \mathbf{P}_U, \nu = \mathbf{P}_M$
2: $L \leftarrow R$
3: **repeat**
4: $\quad \bar{P}_{ij} \leftarrow \frac{\nu_j \bar{Q}_{ij}}{\sum_i \bar{Q}_{ij}}$
5: $\quad \bar{Q}_{ij} \leftarrow \frac{\mu_i \bar{P}_{ij}}{\sum_j \bar{P}_{ij}}$
6: **until** convergence
7: $S_{ij} \leftarrow \frac{\bar{P}_{ij}}{\sum_i \bar{P}_{ij}} \quad L_{ij} \leftarrow \frac{\bar{Q}_{ij}}{\sum_j \bar{Q}_{ij}}$

Remark 3.
- *We introduced prior knowledge on utterances which is usually not used. If not wanted, we can take uniform prior $\mu = \frac{1}{U}\mathbf{1}_U$ to recover classic RSA. This would simplify μ_i in line 5.*
- *Lines 1 to 6 exactly correspond to Sinkhorn algorithm **??** with $\mu = \mathbf{P}_U$ and $\nu = \mathbf{P}_M$.*
- *The last line is a normalization step to obtain probabilities. It can be seen as a projection on probability space.*

RSA algorithm basically consist of alternated projections of lexicon R to fit prior knowledge on utterances and meanings. Then once at the end, we project the matrix on probability space to obtain probability matrices.

8 RSA as a Schrödinger problem

The RSA can now be fully described as Schrödinger problems.

Theorem 5. *Let $R \in M_+(\mathcal{U} \times \mathcal{M}), \mu = \mathbf{P}_U$ and $\nu = \mathbf{P}_M$ defining base lexicon and prior knowledge in the RSA Algorithm **??**. If $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$ the sequences $(P^n)_{n\in\mathbb{N}^*}$ and $(Q^n)_{n\in\mathbb{N}^*}$ given by Algorithm **??** verify :*

$$S^n \underset{n\to+\infty}{\longrightarrow} S^* \quad \textit{and} \quad L^n \underset{n\to+\infty}{\longrightarrow} L^*$$

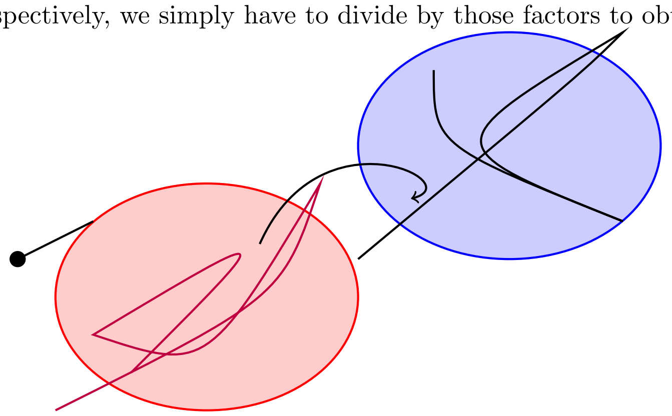
*Where $S^*_{ij} = \frac{1}{\nu_j} P^*_{ij}$ and $L^*_{ij} = \frac{1}{\mu_i} Q^*_{ij}$ with μ^* and ν^* are defined by Equations **??** and*

$$P^* := \arg\mathrm{Sch}(R; \mu, \nu^*) \quad \textit{and} \quad Q^* := \arg\mathrm{Sch}(R; \mu^*, \nu) \tag{8}$$

In other words, the RSA Algorithm gives the normalized matrices found when solving the Schrödinger problem with balanced marginals.

Remark 4. *If the problem is scalable then balanced marginals are the priors $\mu^* = \mu = \mathbf{P}_U$ and $\nu^* = \nu = \mathbf{P}_M$. If in addition priors are uniform, then the normalization step can be removed. In the case of $U = M$ (i.e R is a squared matrix) and uniform priors, RSA directly solves $\mathrm{Sch}(R; \mathbf{1}_U, \mathbf{1}_U)$.*

Proof : The proof is a direct application of Theorem **??**. Lines 1 to 6 of RSA Algorithm **??** are Sinkhorn algorithm so by applying Theorem **??** to the triple (R, μ, ν) we obtain P^* and Q^*. The last line of RSA Algorithm **??** is a normalization step, as matrices P^* and Q^* are already normalized up to a factor μ_i and ν_j respectively, we simply have to divide by those factors to obtain S^* and L^*.



article [utf8]inputenc tikz
[active,tightpage]preview tikzpicture

- $$\frac{\pi^2}{6} = \lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{k^2}$$