The Rational Speech Act, A Mathematical Review

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

August 1, 2024

Contents

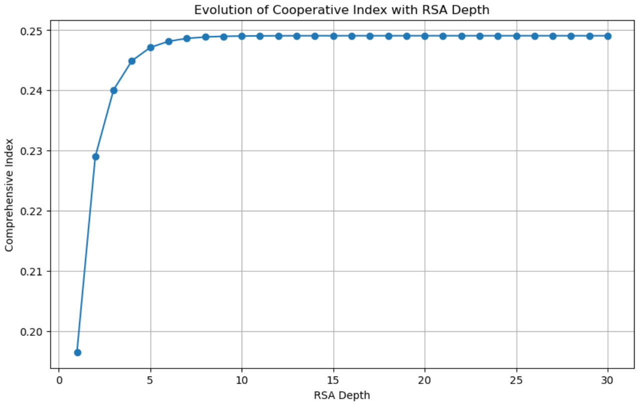


Figure 1: Evolution of CI as depth increases on a 10×10 L_0 matrix

1 Relevance of RSA iterations

In the original [Frank and Goodman, 2012] 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 1 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. [Wang et al., 2020] 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 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 fact, we barely know why RSA is interesting as we do not know from where formulas come from. 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.

Remark 1. *We consider in this paper the case $\alpha = 1$ in formula **??**. It has been showed that RSA more accurately models human behavior when $\alpha = 1$ and that $\alpha = 1$ is a limit case. When $\alpha > 1$ RSA will favour highly deterministic distributions (if $U = M$ and R is diagonalazible RSA will converge toward a full rank matrix). When $\alpha < 1$ RSA will favour more uniform distributions (giving uniform distribution when possible).*

Sinkhorn-Knopp Algorithm 1 is an iterative method which consists of iteratively normalizing rows and columns of a given matrix A. It is extremely similar to RSA Algorithm 2.

Algorithm 1 SK Method

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

Algorithm 2 RSA Algorithm

Given: Initial matrix $R \in \mathbb{R}_+^{U \times M}$
$L \leftarrow R$
repeat
  $\forall i, j \quad S_{ij} \leftarrow \frac{L_{ij}}{\sum_i L_{ij}}$
  $\forall i, j \quad L_{ij} \leftarrow \frac{P(m_j) S_{ij}}{\sum_j P(m_j) S_{ij}}$
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, explicited in Theoremic O 2. *Entropptimal Transport* is another formulation of the so-called *Shrödinger Bridging problem* formulated in Definition 1 [Gushchin et al., 2023], [Léonard, 2013]. As RSA and SK algorithm are the same, and SK solves Schrödinger bridging problem, we are looking for a formulation of RSA as a Schrödinger problem, explicited later in the paper in section 6. 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. The problem solved by matrix scaling algorithm took several names and has been studied independently in different fields. We refer to [Idel, 2016] for a review of the different formulation of Entropy Optimal Transport problem.

3 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

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

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 $H(\bar{R} \mid R) := \sum_{ij} \left\{ \bar{R}_{ij} \log \frac{\bar{R}_{ij}}{R_{ij}} + R_{ij} - \bar{R}_{ij} \right\}$.

Remark 2. *The relative entropy do not usually include the last terms $R_{ij} - \bar{R}_{ij}$ but we have no assumptions on R, so it may not be a probability matrix. In case R is a probability matrix, the last term is null and the relative entropy is the usual Kullback-Leibler divergence.*

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:*

$$\mathrm{Sch}(R; \mu, \nu) := \min_{\bar{R}} H(\bar{R} \mid R) \mid \bar{R} \in \Pi(\mu, \nu) \tag{1}$$

We use the conventions $a \log \frac{a}{0} = +\infty$ if $a > 0$, and $0 \log 0 = 0 \log \frac{0}{0} = 0$.

As the relative entropy is strictly convex w.r.t $\bar{R}$, if the Schrödinger problem admits a solution R^* then it is unique. Moreover, as $\Pi(\mu, \nu)$ is a compact set, the Schrödinger problem admits a solution if and only if there exists a $\bar{R} \in \Pi(\mu, \nu)$ such that $H(\bar{R} \mid R) < +\infty$.

Theorem 1. *If the Schrödinger problem $\mathrm{Sch}(R; \mu, \nu)$ admits a solution, then it is unique and the Sinkhorn algorithm 1 converges toward the solution :*

$$P^n \underset{n \to +\infty}{\longrightarrow} P^* \quad \text{and} \quad Q^n \underset{n \to +\infty}{\longrightarrow} Q^* \tag{2}$$

Where $P^* = Q^* = R^* = \arg \mathrm{Sch}(R; \mu, \nu)$

Proof. See [Knopp and Sinkhorn, 1967] [Krupp, 1979] [Sinkhorn, 1967] or [Idel, 2016] 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 utterances and meanings 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 1 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 marginals $(\mu, \nu) = (\mathbf{1}_U, \mathbf{1}_M)$ verifying $\sum_{ij} R_{ij} = \sum_i \mu_i = \sum_j \nu_j$, then the problem $\mathrm{Sch}(R; \mathbf{1}_U, \mathbf{1}_U)$ defined by Definition 1 admits a single solution and RSA converges toward a single doubly stochastic matrix $L^* = S^* = R^*$ solution of $\mathrm{Sch}(R; \mathbf{1}_U, \mathbf{1}_U)$.

Proof. Sinkhorn proved in [**?**] that for strictly positive squared matrix R it corresponds a single doubly stochastic matrix. The sequences of matrices given by Algorithm 1 with $\mu = \mathbf{1}_U$ and $\nu = \mathbf{1}_M$ converges toward this doubly stochastic matrix with.

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 [Baradat and Ventre, 2023]):*

1. *$M(\mu) = M(\nu)$ and for all $A \subset \mathcal{U}, \mu(A) \leq \nu(M_R(A))$.*
2. *$M(\mu) = M(\nu)$ and for all $B \subset \mathcal{M}, \nu(B) \leq \mu(U_R(B))$.*
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 $A \subset \mathcal{U}, \mu(A) \leq \nu(M_R(A))$" 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 $(R_{ij} > 0)$"

4 Convergence of Sinkhorn algorithm

In this section we study the convergence of Sinkhorn Algorithm 1. We are looking for the conditions under which the algorithm converges, the problem it solves and the limits of the convergence. Conditions enonced in Theorem 2 are sufficient but not necessary to convergence of Algorithm 1. 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 :

$$\begin{aligned}\mu^* &:= \arg\min \left\{ H(\bar{\mu} \mid \mu) \mid \bar{\mu} = \mu^Q \text{ for some } Q \text{ with } H(Q \mid R) < +\infty \text{ and } \nu^Q = \nu \right\},\\ \nu^* &:= \arg\min \left\{ H(\bar{\nu} \mid \nu) \mid \bar{\nu} = \nu^P \text{ for some } P \text{ with } H(P \mid R) < +\infty \text{ and } \mu^P = \mu \right\}.\end{aligned} \tag{3}$$

Those marginals are associated with the problem $\mathrm{Sch}(R; \mu, \nu)$ and depend on R, μ and ν.

Theorem 3. *Let $R \in M_+(\mathcal{U} \times \mathcal{M}), \mu \in M_+(\mathcal{U})$, $\nu \in M_+(\mathcal{M})$ and the sequences $(P^n)_{n\in\mathbb{N}^*}$ and $(Q^n)_{n\in\mathbb{N}^*}$ given by Algorithm 1. If $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$, then (Th.11 of [Baradat and Ventre, 2023])*

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

Where μ^* and ν^* are defined by Equations 3. Moreover

$$P^* := \arg \mathrm{Sch}(R; \mu, \nu^*) \quad \text{and} \quad Q^* := \arg \mathrm{Sch}(R; \mu^*, \nu) \tag{4}$$

In other words, if the Sinkhorn algorithm converges it solves two different Schrödinger problem defined by Equations 5. To find the Schrödinger problem associated with a given triple (R, μ, ν) we first obtain μ^* and ν^* by solving the convex optimization problems of Equations 3 to find μ^* and ν^*. The problem solved by Sinkhorn algorithm is then given by Equations 5.

Remark 3. *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 3.*

Remark 4. *Since Equations 3 are convex optimization problems, μ^* and ν^* exists if and only if it exists Q (resp. P) such that $H(Q \mid R) < +\infty$ and $\nu^Q = \nu$ (resp. $H(P \mid R) < +\infty$ and $\mu^P = \mu$). If existence this solution is unique.*

Remark 5. *If the problem $\mathrm{Sch}(R; \mu, \nu)$ is scalable (i.e conditions of Theorem 2 are met) then $\mu^* = \mu$ and $\nu^* = \nu$ and Sinkhorn algorithm converges toward the unique solution.*

5 Differences between RSA and Sinkhorn

As shown in Figure 2 Rational Speech Act and Sinkhorn algorithm are really similar but are not exactly the same. 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: **for** $k = 1$ **to** n **do**
4: $\quad \bar{P}^k_{ij} \leftarrow \frac{\nu_j \bar{Q}^k_{ij}}{\sum_i \bar{Q}^k_{ij}}$
5: $\quad \bar{Q}^k_{ij} \leftarrow \frac{\mu_i \bar{P}^k_{ij}}{\sum_j \bar{P}^k_{ij}}$
6: **end for**
7: $S^n_{ij} \leftarrow \frac{\bar{P}^n_{ij}}{\sum_i \bar{P}^n_{ij}}$
8: $L^n_{ij} \leftarrow \frac{\bar{Q}^n_{ij}}{\sum_j \bar{Q}^n_{ij}}$

Remark 6.
- *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 7.*
- *Lines 1 to 6 exactly correspond to Sinkhorn algorithm 1 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.*
- *We chose μ and ν probability vectors but properties stand for any positive vectors.*

RSA algorithm consists of alternated projections of lexicon R to fit prior knowledge on utterances and meanings. Once at the end (line 7), we project the matrices on probability spaces to obtain probability vectors.

6 RSA as a Schrödinger problem

We can now describe RSA with Schrödinger problems.

Theorem 4. *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 classic RSA. If $H(\mu \mid \mu^R) < \infty$ and $H(\nu \mid \nu^R) < \infty$ the sequences $(S^n)_{n\in\mathbb{N}^*}$ and $(L^n)_{n\in\mathbb{N}^*}$ given by RSA Algorithm 3 verify :*

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

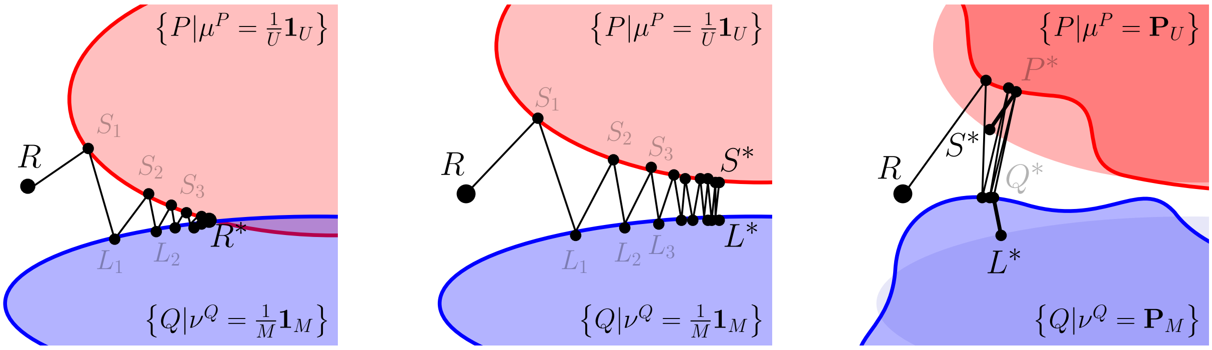
Where $S^*_{ij} = \frac{1}{\nu_j}P^*_{ij}$ and $L^*_{ij} = \frac{1}{\mu_i}Q^*_{ij}$ with μ^* and ν^* are defined by Equations 3 and

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

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

Remark 7. *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 we add the condition $U = M$, RSA directly solves $\mathrm{Sch}(R;\mathbf{1}_U,\mathbf{1}_U)$.*

Proof : The proof is a direct application of Theorem 3. Lines 1 to 6 of RSA Algorithm 3 are Sinkhorn algorithm so by applying Theorem 3 to the triple (R,μ,ν) we obtain P^* and Q^*. The last line of RSA Algorithm 3 is a normalization step, as matrices P^* and Q^* are already normalized up to a factor μ_i and ν_j respectively, we divide by those factors to obtain S^* and L^*.



(a) Uniform prior, $U = M$ (b) Uniform prior, $U \neq M$ (c) Non uniform priors, $U \neq M$

Figure 3: RSA iterations in different settings

7 Relaxed problem

[Baradat and Ventre, 2023] 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 5. 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 Algorithm1 are explicited in this section.

We focus situations where conditions of Theorem 2 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 $\nu^{\bar{R}} = \nu$ and $\mu^{\bar{R}} = \mu$ are replaced by soft constraints making the problem solvable. The parameter ε quantify the level of relaxation. We note R^*_ε the solution of relaxed problem 6.

As $\varepsilon \to 0$, R^*_ε converges toward the solution of a Schrödinger problem with balanced marginals.

Theorem 5. *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 5, R^*_ε solution of Equation 6, we have (Th. 17 of [Baradat and Ventre, 2023])*

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

Moreover R^* is solution of $\mathrm{Sch}(R;\sqrt{\mu^*\mu},\sqrt{\nu^*\nu})$ with μ^* and ν^* defined by Equations 3

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 1.

References

[Baradat and Ventre, 2023] Baradat, A. and Ventre, E. (2023). Convergence of the sinkhorn algorithm when the schrödinger problem has no solution.

[Frank and Goodman, 2012] Frank, M. and Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science (New York, N.Y.)*, 336:998.

[Gushchin et al., 2023] Gushchin, N., Kolesov, A., Mokrov, P., Karpikova, P., Spiridonov, A., Burnaev, E., and Korotin, A. (2023). Building the bridge of schrödinger: A continuous entropic optimal transport benchmark.

[Idel, 2016] Idel, M. (2016). A review of matrix scaling and sinkhorn's normal form for matrices and positive maps.

[Knopp and Sinkhorn, 1967] Knopp, P. and Sinkhorn, R. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343 – 348.

[Krupp, 1979] Krupp, R. S. (1979). Properties of krulthof's projection method. *The Bell System Technical Journal*, 58(2):517–538.

[Léonard, 2013] Léonard, C. (2013). A survey of the schrödinger problem and some of its connections with optimal transport.

[Sinkhorn, 1967] Sinkhorn, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. ii.

[Wang et al., 2020] Wang, P., Wang, J., Paranamana, P., and Shafto, P. (2020). A mathematical theory of cooperative communication.

Appendix

Proofs and Demonstrations