ISIT'2013 1569719457

Memoryless Representation of Markov Processes

Amichai Painsky and Saharon Rosset
School of Mathematical Sciences, Tel Aviv University
Tel Aviv, Israel
amichaip@eng.tau.ac.il, saharon@post.tau.ac.il

Meir Feder
School of Electrical Engineering, Tel Aviv University
Tel Aviv, Israel
meir@eng.tau.ac.il

Abstract— Memoryless processes hold many theoretical and practical advantages. They are easy to describe, analyze, store and encrypt. They can also be seen as the essence of a family of regression processes, or as an innovation process triggering a dynamic system. The Gram-Schmidt procedure suggests a linear sequential method of whitening (decorrelating) any stochastic process. Applied on a Gaussian process, memorylessness (that is, statistical independence) is guaranteed. It is not clear however, how to sequentially construct a memoryless process from a non-Gaussian process. In this paper we present a non-linear sequential method to generate a memoryless process from any given Markov process under varying objectives and constraints. We differentiate between lossless and lossy methods, closed form and algorithmic solutions and discuss the properties and uniqueness of our suggested methods.

Keywords – Gram-Schmidt procedure, memoryless processes, optimal transportation problem, Markov procceses

I. INTRODUCTION

Several methods have been suggested to construct an uncorrelated or independent process from a given stochastic process. The Gram-Schmidt procedure suggests a simple sequential method which projects every new component on the linear span of the components previously observed. The difference between the current component and its projection is guaranteed to be orthogonal to all previous components. Applied on a Gaussian process, orthogonality results statistical independence and the subsequent process is therefore memoryless. Non-Gaussian processes on the other hand, do not hold this quality and a generalized form of sequentially generating a memoryless process from any given time dependent series is therefore required. Several non-sequential methods such as Principal Components Analysis and Independent Component Analysis [1-4] have received a great deal of attention, but we are aware of little previous work on sequential schemes for generating memoryless "innovation" processes.

The importance of innovation process representation spans a variety of fields. One example is dynamic system analysis in which complicated time dependent processes are approximated as independent processes triggering a dynamic system (human speech mechanism, for instance). Another major field for example is cryptography, where a memoryless language is easier to encrypt as it prevents an eavesdropper from learning the code by comparing its statistics with those of the serially correlated language.

Recently, Shayevitz and Feder presented the Posterior Matching (PM) scheme for communication with feedback [5]. It turns out that an essential part of their scheme is to produce statistical independence between every two consecutive transmissions. Inspired by this we suggest a general framework to sequentially construct memoryless processes from any given Markov process, for various types of desired distribution function, under different objective functions and constraints.

II. PROBLEM FORMULATION

For the remaining sections of this paper we use the following notation: we denote the input process at a time k as Y_k while Y^k refers to the vector $\{Y_i\}_{i=1}^k$. We use the same notation for our outcome process U. Therefore, for any process Y with a cumulative distribution function $F(Y^k)$ we would like to sequentially construct U^k such that:

(I) $F(U^k) = \prod_{i=1}^k F(U_i)$.
(II) Y^k *can be uniquely recovered from* U^k *for any k*

We show that the two constraints can always be met if we allow a U to take values on a continuous set and may need to be relaxed otherwise. The continuous case is discussed in the next section, followed by a comprehensive discussion on the discrete case in the remaining of this paper.

III. GENERALIZED GRAM-SCHMIDT

Following the footsteps of the PM scheme we define a generalized Gram-Schmidt method for the continuous case.

Lemma 1: Let X be any random variable $X \sim F_X(x)$ and $\Theta \sim \mathcal{U}nif[0,1]$ be statistically independent of it. In order to shape X to a uniform distribution (and vice versa) the following applies:

(i) $F_X^{-1}(\Theta) \sim F_X(x)$
(ii) Assume X is a non-atomic distribution ($F_X(x)$ is strictly increasing) then $F_X(X) \sim \mathcal{U}nif[0,1]$.
(iii) Assume X is discrete or a mixture probability distribution then $F_X(X) - \Theta \cdot P_X(x) \sim \mathcal{U}nif[0,1]$.

Proof: can be located in [7].

We define $\tilde{F}_X(x)$ as $\tilde{F}_X(x) = F_X(x)$ if $F_X(x)$ is strictly increasing and $\tilde{F}_X(x) = F_X(x) - \Theta \cdot P_X(x)$ otherwise. For a desired set of $F_{U_i}(u)$ we construct our process by setting:

$$U_1 = F_{U_1}^{-1}\left(\tilde{F}_{Y_1}(Y_1)\right) \quad (1)$$

$$U_k = F_{U_k}^{-1}\left(\tilde{F}_{Y_k|Y^{k-1}}(Y_k|Y^{k-1})\right) \quad \forall k > 1 \quad (2)$$

Lemma 1 guarantees that $\tilde{F}_{Y_k|Y^{k-1}}(Y_k|Y^{k-1})$ is uniformly distributed and applying $F_{U_k}^{-1}$ on it shapes it to the desired continuous distribution.

In other words, this method suggests that for every possible history of the process at a time k, the transformation $\tilde{F}_{Y_k|Y^{k-1}}(Y_k|Y^{k-1})$ shapes Y_k to the same (uniform) distribution. This ensures independence of its history. The method then reshapes it to the desired distribution. It is easy to see that U_k are statistically independent as every U_k is independent of Y^{k-1}. Moreover, since $F(U_i)$ is strictly increasing and $\tilde{F}_{Y_1}(Y_1)$ is uniformly distributed we can uniquely recover Y_1 from U_1 according to the construction of *lemma 1*. Simple induction steps show this is correct for every U_k $k > 1$. A discussion on the uniqueness of this method can be found in [6].

IV. Lossy Approximation in the discrete case

Let us assume now both Y and U take values on finite alphabet size of A and B respectively. Even in the simplest case, where both are binary and Y is a first order Markov chain, it is easy to see that no transformation can achieve both constraints mentioned above. We therefore relax the second constraint by replacing the unique recovery constraint with mutual information maximization $I(Y^k; U^k)$. Since we limit ourselves to sequential processing it is easy to show that this maximization problem is equivalent to maximizing $I(Y_k; U_k|Y^{k-1})$ $\forall k$. We also notice that the case where Y_k is uniquely recoverable from U_k given its past results with $I(Y_k; U_k|Y^{k-1})$ achieving its maximum as desired. Our problem can be reformulated as follows:

For any realization of Y_k, given any possible history Y^{k-1}, find a set of mapping functions to a desired distribution $P(U)$ such that the mutual information is maximal. For example, consider the binary case where Y is a first order Markov chain, and U is Bernoulli distributed

$$U_k \sim Ber(\beta)\ ,\ P_{Y_k}(Y_k = 0) = \gamma_k \quad (3)$$

$$P_{Y_k|Y_{k-1}}(Y_k = 0|Y_{k-1} = 0) = \alpha_1$$

$$P_{Y_k|Y_{k-1}}(Y_k = 0|Y_{k-1} = 1) = \alpha_2$$

Since Y is a first order Markov process it is easy to show that

$$I(Y_k; U_k|Y^{k-1}) = \gamma_{k-1} I(Y_k; U_k|Y_{k-1} = 0) + (1 - \gamma_{k-1}) I(Y_k; U_k|Y_{k-1} = 1). \quad (4)$$

In addition, we would like to find the distribution of U_k such that this mutual information is maximal. This distribution can be viewed as the closest approximation of the process Y as a memoryless process in terms of maximal mutual information with it.

This problem is actually a concave minimization over a convex polytope shaped set [7] and the minimum is guaranteed on to lie on one of the polytope's vertices. Unfortunately this is problem is NP hard and generally there is no closed form solution for this problem. Several approximations and exhaustive search solutions are available for this kind of problem, such as [8]. There are, however, several simple cases in which a closed form solution exists. One notable example is the binary case.

A. The binary case

Consider the following problem: Given two binary random variables X and Y and their marginal distributions $P_X(X = 0) = \alpha < 1/2$ and $P_Y(Y = 0) = \beta < 1/2$ we would like to find the conditional distributions $P_{Y|X}(y|x)$ such that the mutual information between X and Y is maximal. Simple derivation shows that the maximal mutual information is:

For $\beta > \alpha$:

$$I_{max}^{\beta>\alpha}(X;Y) = H_b(\beta) - (1-\alpha) H_b\left(\frac{\beta-\alpha}{1-\alpha}\right). \quad (5)$$

For $\beta < \alpha$:

$$I_{max}^{\beta<\alpha}(X;Y) = H_b(\beta) - \alpha H_b\left(\frac{\beta}{\alpha}\right). \quad (6)$$

Applying this result on the first order Markov process setup described above and assuming all parameters are smaller than $1/2$, the maximal average mutual information is just:

For $\beta < \alpha_1 < \alpha_2$:

$$I(Y_k; U_k|Y^{k-1}) = \gamma_{k-1} I_{max}^{\beta<\alpha_1}(X;Y) + (1-\gamma_{k-1}) I_{max}^{\beta<\alpha_2}(X;Y). \quad (7)$$

For $\alpha_1 \le \beta < \alpha_2$:

$$I(Y_k; U_k|Y^{k-1}) = \gamma_{k-1} I_{max}^{\beta>\alpha_1}(X;Y) + (1-\gamma_{k-1}) I_{max}^{\beta<\alpha_2}(X;Y). \quad (8)$$

For $\alpha_1 < \alpha_2 \le \beta$:

$$I(Y_k; U_k|Y^{k-1}) = \gamma_{k-1} I_{max}^{\beta>\alpha_1}(X;Y) + (1-\gamma_{k-1}) I_{max}^{\beta>\alpha_2}(X;Y). \quad (9)$$

It is easy to see that $\mathrm{I}(\mathrm{Y_k};\mathrm{U_k}|\mathrm{Y^{k-1}})$ is continuous in β. Simple derivation shows that for $\beta < \alpha_1$ the maximal mutual information is monotonically increasing in β and for $\alpha_2 < \beta$ it is monotonically decreasing in β. It can also be verified that all optimum points in the range $\alpha_1 < \beta < \alpha_2$ are local minima. All this leads to the conclusion that the maximum is located on the inner bounds of the range, $\beta = \alpha_1$ or $\beta = \alpha_2$. Comparing the two leads to a decision rule:

$$\gamma_{k-1} \begin{array}{c} \beta = \alpha_2 \\ \lessgtr \\ \beta = \alpha_1 \end{array} \frac{H_b(\alpha_2) - H_b(\alpha_1) + \alpha_2 H_b\left(\frac{\alpha_1}{\alpha_2}\right)}{\alpha_2 H_b\left(\frac{\alpha_1}{\alpha_2}\right) + (1-\alpha_1) H_b\left(\frac{\alpha_2-\alpha_1}{1-\alpha_1}\right)} \qquad (10)$$

Assuming the process Y is at its stationary state yields $\gamma = \frac{\alpha_2}{1-\alpha_1+\alpha_2}$. Applying this result to the inequality above, it is can be verified that for $\alpha_1 < \alpha_2 < 1/2$ we have:

$$\frac{\alpha_2}{1-\alpha_1+\alpha_2} \quad < \quad \frac{H_b(\alpha_2) - H_b(\alpha_1) + \alpha_2 H_b\left(\frac{\alpha_1}{\alpha_2}\right)}{\alpha_2 H_b\left(\frac{\alpha_1}{\alpha_2}\right) + (1-\alpha_1) H_b\left(\frac{\alpha_2-\alpha_1}{1-\alpha_1}\right)} \qquad (11)$$

which leads to the conclusion that $\beta_{opt} = \alpha_2$.
This result is easily generalized to all values of α_1 and α_2 which results with a decision rule stating β_{opt} equals the parameter closest to $1/2$

$$\beta_{opt} = argmax_{\theta \in \{\alpha_1, \alpha_2, 1-\alpha_1, 1-\alpha_2\}} \left(\frac{1}{2} - \theta\right). \qquad (12)$$

In other words, in order to best approximate a binary first order Markov process at its stationary state, we set the distribution of the binary memoryless process to be equal to the conditional distribution which holds the largest entropy.
Expanding this result to an m-order Markov process we have $M = 2^m$ Bernoulli distributions to be mapped to a single one. The maximization objective is therefore

$$I(Y_k; U_k | Y^{k-1}) = \sum_{i=0}^{M-1} \gamma_i \, I(Y_k; U_k | Y_{k-1}, \dots, Y_{k-M-1} = i) \qquad (13)$$

where γ_i is the probability of the vector $Y_{k-1}, \dots, Y_{k-M-1}$ to be equal to its i^{th} possible value, $\gamma_i = P(Y_{k-1}, \dots, Y_{k-M-1} = i)$ and $I(Y_k; U_k | Y_{k-1}, \dots, Y_{k-M-1} = i)$ is either $H_b(\beta) - \alpha_i H_b\left(\frac{\beta}{\alpha_i}\right)$ or $H_b(\beta) - (1-\alpha_i) H_b\left(\frac{\beta-\alpha_i}{1-\alpha_i}\right)$, depending on β and α_i as described above.
Simple calculus shows that as in the *M=2* case, $\mathrm{I}(\mathrm{Y_k};\mathrm{U_k}|\mathrm{Y^{k-1}})$ reaches its maximum on one of the inner bounds of β's range:

$$\beta_{opt} = argmax_{\beta \in \{\alpha_i\}} \left(H_b(\beta) - \sum_{\beta < \alpha_i} \gamma_i \alpha_i H_b\left(\frac{\beta}{\alpha_i}\right) - \sum_{\beta > \alpha_i} \gamma_i (1-\alpha_i) H_b\left(\frac{\beta-\alpha_i}{1-\alpha_i}\right) \right). \qquad (14)$$

Here however it is not possible to conclude that β equals the parameter closest to $1/2$. Simple counter example shows it is necessary to search over all possible parameters as a result of the nature of our concave minimization problem over a convex polytope.

V. Lossless Transformation in the Discrete Case

The lossy approximation may not be adequate in applications where unique recovery of the original process is required. It is therefore necessary to increase the alphabet so that every marginal distribution of Y, given any possible history of the process can be accommodated.
This problem can be formulated as follows:
Assume Y is a multinomial distributed m-order Markov process, taking values on finite alphabet size of *A*. Given its past, we denote its conditional probability as $Y_m \sim multnom(\alpha_{1m}, \alpha_{2m}, \dots, \alpha_{Am})$ where $1 \le m \le M = A^m$. We would like to find a distribution $U \sim multnom(\beta_1, \beta_2, \dots, \beta_B)$ where $B \ge A$ is unknown, and *M* sets of conditional probabilities between every Y_m and U, such that every instance of Y_m can be uniquely recovered from U. In addition, we would like the entropy of U to be as small as possible. Without loss of generality we assume that $\alpha_{am} \le \alpha_{(a+1)m} \ \forall a < A$, since we can always place them in such order. We also order the sets such that $\alpha_{1m} \le \alpha_{1(m+1)}$. Notice we have $\alpha_{1m} \le 0.5 \ \forall m$.
For example, for *A=2* and *M=2*, it is easy to verify that $B \ge 3$ is a necessary condition for Y_m to be uniquely recoverable from U, $\forall m \le M$. Simple calculus shows the conditional probabilities which achieve the minimal entropy are $\beta_1 = \alpha_1$, $\beta_2 = \alpha_2 - \alpha_1$ and $\beta_3 = 1 - \alpha_2$.

A. Minimizing B

Let us start with finding the minimal *B* such that the process *Y* is guaranteed to be uniquely recoverable from *U*.
Looking at the free parameters of our problems we have:
Defining β_i: *B-1* parameters .
Defining M conditional probability distributions between A and B variables: *M(A-1)(B-1)* parameters.
In order for Y_m to be uniquely recoverable from *U*, each instance of *U* needs to be at most assigned to a single instance of Y_m. This means that for each of the *M* sets, we have *B(A-1)* constraints. Therefore, in order to have more free parameters than constraints we demand *(B-1)+M(A-1)(B-1)≥MB(A-1).* Rearranging this inequality leads to:

$$B \geq M(A-1)+1. \quad (15)$$

For example, assuming a binary alphabet on Y we get $B \geq M+1$. Several special cases exist in which it is possible to go under this lower bound, like cases in which some parameters are a result of addition or subtraction of other parameters such as $\alpha_2 = 1-\alpha_1$ in the binary case. We focus however, on solving the most general case.

B. The Optimization Problem

The problem stated above can be formulated as the following optimization problem:

$$min\ H(U)\ s.t.\ H(Y_m|U=u_b) \leq 0 \quad \forall m,b \quad (16)$$

Unfortunately this is a concave minimization problem over a non-convex set. However, we show this problem can also be formulated as a mixed integer problem.

C. Mixed Integer Problem Formulation

In order to formulate our problem as a mixed integer problem we first notice the free parameters are all conditional probabilities, as they fully determine the outcome distribution. We use the notation p_{mab} to describe the conditional probability $P(U=u_b|Y_m=y_a)$. A necessary and sufficient condition for zero conditional entropy is that for every value $U=u_b$, in every set $m \leq M$, there is only a single $Y_m = y_a$ such that $p_{mab} > 0$. This is accomplished by adding Boolean variables X_{mab} and demanding:

$$\begin{gathered} p_{mab} - X_{mab} \leq 0 \\ \sum_{a=1}^{A} X_{mab} = 1\ , X_{mab} \in \{0,1\} \end{gathered} \quad (17)$$

Additionally, we add inequality constraints to make sure that $\beta_i \leq \beta_{i+1}\ \forall i$. This will become handy in our lower bound analysis later this section. Therefore, our optimization problem can be written as follows:

Define a vector of parameters $z = [p_{mab}\ \ X_{mab}]^T$.

Define A_{eq} , b_{eq} as the equality constraints in a matrix and vector forms respectively and A_{ineq} , b_{ineq} as the inequality constraints in a matrix and vector forms respectively.

We have:

$$\begin{gathered} min\ f(z) \\ s.t. \quad A_{eq}z = b_{eq}\ ,\ A_{ineq}z \leq b_{ineq}\ ,\ 0 \leq z \leq 1 \\ z(boolean\ indicators) \in \{0,1\} \end{gathered} \quad (18)$$

Where $f(z)$ is the entropy of the random variable U, in terms of p_{mab}.

D. Mixed Integer Problem Discussion

Mixed integer problems are studied broadly in the computer science community. There are well established methodologies for convex minimization in a mixed integer problem and specifically in the linear case [9-10]. The study of non-convex optimization in mixed integer problem is also growing quite rapidly, though there is less software available yet. The most broadly used mixed integer optimization solver is the CPLEX, developed by IBM. CPLEX provides a mixed integer linear programming (MILP) solution, based on a branch and bound oriented algorithm. We use the MILP in lower bounding our objective function as described in the following sections.

E. An Exhaustive Solution

As shown above, the problem we are dealing with is a hard one and therefore we present a combinatorial search method to find the minimal entropy.

We notice that each of the known parameters α_{am} can be expressed as a convex combination of the free parameters β_b such that $A\underline{\beta} = \underline{\alpha}$, where A represents the convex coefficients. However, if we are to ensure the conditional entropy constraint as stated above it is easy to show that matrix A must be a Boolean matrix. In addition, a necessary condition for the recovery of $\underline{\beta}$ from A and $\underline{\alpha}$ is that A is full rank. This however is not a sufficient condition since there is no guarantee that $\underline{\beta}$ is a valid probability distribution.

We would therefore search over all Boolean matrices A of a full rank, and for each of these matrices we check if the resulting $\underline{\beta}$ is a probability distribution. If so, we calculate its entropy and proceed. This process grows exponentially with K and M (and thus N) but may be feasible for smaller values of these figures.

F. Greedy Solution

The entropy minimization problem can also be viewed as an attempt to minimize the entropy of a random variable $U \sim multinom(\beta_1, \beta_2, \ldots, \beta_B)$ on a set of discrete points, representing valid solutions to the problem we defined.

Let us remember that $\beta_b \leq \beta_{b+1}\ \forall b < B$ as stated in the previous sections.

Proposition: β_B is not greater than $max_m\{max_a\{\alpha_{am}\}\}$.

Proof: Assume $\beta_B > max_m\{max_a\{\alpha_{am}\}\}$. Than for this m there must be at least two values y_i and y_j for which $p_{mab} > 0$. This is a contradiction to the zero conditional entropy constraint ■

A greedy algorithm would therefore like to "squeeze" all the distribution to the values which are less constrained from above, so that it is as large as possible. We then suggest that in every step of the algorithm we set $\beta_B = max_m\{max_a\{\alpha_{am}\}\}$ which leaves us with a $B-1$ problem. Updating the remaining probabilities and repeating this maximization step, will ensure that in each step we increase the least constrained β as much as possible.

It is easy to notice, however, that this solution is not optimal through a simple counter example.

G. Lowest Entropy Bound

As discussed above, we are dealing with entropy minimization over a discrete set of valid solutions. This problem can be viewed as a mixed integer non convex minimization, which is a hard problem.

However, we can find boundaries on each of the parameters β_i and see the lowest entropy we can hope for. This way, we relax the search over a set of valid solutions to a search in a continuous space, bounded by a polytope.

We find the boundaries of each β_i by changing our minimization objective to a simpler linear one (minimize/maximize β_i). This way we find a valid solution for which β_i is at its bound. This problem is a simple MILP as shown above. By looking at all these boundaries together and minimizing the entropy in this continuous space we can find a lower bound for the minimal entropy one can expect.

We notice that this bound is not tight, and we even do not know how far it is from the valid minima, as it is not necessarily a valid solution. However, it does give us a benchmark to compare our greedy algorithm with and decide if we are satisfied with it or need more powerful tools.

We also note that as we increase B the number of valid solutions grows exponentially. This leads to a more packed set of solutions which tightens the suggested lower bound as we converge to a polytope over a continuous set.

H. Entropy Minimization on a Polytope Set

The following proposition suggests an optimal method of entropy minimization over a polytope (thus convex) set.

Proposition: Assume a random variable X with a multinomial distribution, $X \sim multnom(\beta_1, \beta_2, \ldots, \beta_N)$, where β_i are parameters bounded such that:

- $a_i \leq \beta_i \leq b_i \; \forall i$
- $\sum_i a_i \leq 1$ and $\sum_i b_i \geq 1$ (to ensure a feasible solution)
- $a_i \leq a_{i+1}$ and $b_i \leq b_{i+1} \; \forall i$.
- $a_i \leq b_i \; \forall i$

The minimal entropy is achieved iff $\exists k > 0$ s.t.

- $\beta_i = b_i \;\; \forall i > k$
- $\beta_i = a_i \;\; \forall i < k$
- $\beta_k = 1 - \sum_{i \neq k} \beta_i$

proof: can be located in [11].

VI. CONCLUSTIONS

In this paper we presented a sequential non-linear method to generate a memoryless process from any Markov process under different objectives and constraints. We showed there exists a simple closed form solution if we allow the outcome process to take values on a continuous set. Restricting the alphabet however, may cause lossy recovery of the original process. Two solutions are presented in the face of two possible objectives in the discrete case. First, assuming the alphabet size is too small to allow lossless recovery we aim to maximize the mutual information with the original process. The second objective focuses on finding a minimal alphabet size so that a unique recovery is guaranteed, while minimizing the entropy of the resulting process. In both cases the problem is shown to be hard and several approaches are discussed. In addition, a simple closed-form solution is provided for a binary first order Markov process.

The problem of finding a single marginal distribution function to be fitted to multiple ones under varying costs functions can be viewed as a multi-marginal generalization of the well-studied optimal transportation problem [12]. In other words, we suggest that the optimal transportation problem can be generalized to a design problem in which we are given not a single but multiple source distribution functions. We are then interested not only in finding conditional distributions to minimize a cost function, but also in finding the single target distribution that minimizes the cost. We conjecture that this problem has multiple applications in the fields of Economics, Engineering and others. The complete study of the designed multi-marginal optimal transportation problem and additional extensions is considered as our current research path.

ACKNOWLEDGMENT

This research was funded in part by Israeli Science Foundation grant 634-09 and by a grant to Amichai Painsky from the Israeli Center for Absorption in Science.

REFERENCES

[1] I. Jolliffe "Principal component analysis", John Wiley & Sons, 2005.

[2] A. Hyvärinen, "Independent component analysis for time-dependent stochastic processes," in Proc. Int. Conf. Artificial Neural Networks (ICANN'98), Sweden, 1998, pp. 541–546

[3] L. B. Almeida, "MISEP-linear and nonlinear ICA based on mutual information,"*Signal Process.*, vol. 84, no. 2, pp. 231–245, 2004.

[4] B. Scholkopf, A. Smola and K.R. Mulller, "Nonlinear component analysis as a kernel eigenvalue problem". Neural Computation, 10(5), 1299–1319, 1998

[5] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," IEEE Trans. Info. Theory. vol 57, no. 3, pp. 1186-1222, 2011.

[6] A. Painsky, S. Rosset and M. Feder, "On the Uniqueness of the Posterior Matching Feedback Communication Scheme", http://www.math.tau.ac.il/~amichaip

[7] M. Kovačević, I. Stanojević and V. Šenk, "On the Hardness of Entropy Minimization and Related Problems", arXiv:1207.1238, 2012.

[8] K. Takahito and Y. Shiguro. "A Simplicial Algorithm for Concave Minimization and Its Performance as a Heuristic Tool." Technical Report of Department of Computer Science, 1-18, 2007.

[9] C. Floudas and Christodoulos A, "Nonlinear and mixed-integer optimization: fundamentals and applications", Oxford University Press, 1995.

[10] Tawarmalani, Mohit, and Nikolaos V. Sahinidis. "Global optimization of mixed-integer nonlinear programs: A theoretical and computational study."Mathematical programming 99.3, 2004

[11] A. Painsky, S. Rosset and M. Feder, "Memoryless Representation of Markov Processes: Appendices and Proofs", http://www.math.tau.ac.il/~amichaip/

[12] L. Kantorovich, "On the translocation of masses," Compt. Rend. Akad. Sei 7, 199-201, 1942.