A Novelty in Blahut–Arimoto Type Algorithms: Optimal Control over Noisy Communication Channels

1 author:

Makan Zamanipour
Researcher-ID: P-6298-2019; ORCID: 0000-0003-1606-9347; Scopus-ID: 56719734800

20 PUBLICATIONS   **15** CITATIONS

A Novelty in *Blahut-Arimoto* Type Algorithms: Optimal Control over Noisy Communication Channels

Makan Zamanipour, *Member, IEEE*

Abstract—A probability-theoretic problem under information constraints for the concept of optimal control over a noisy-memoryless channel is considered. For our *Observer-Controller* block, i.e., the lossy joint-source-channel-coding (JSCC) scheme, after providing the relative mathematical expressions, we propose a *Blahut-Arimoto*-type algorithm – which is, to the best of our knowledge, for the first time. The algorithm efficiently finds the probability-mass-functions (PMFs) required for
$$\min_{\mathcal{P}(i), i \in \{\mathcal{Y}, \hat{\mathcal{S}}, \mathcal{X}, \mathcal{S}, \hat{\mathcal{X}}\}} \phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}} | \mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}} | \mathcal{X}, \mathcal{S}).$$
This problem is an $NP-$hard and non-convex multi-objective optimisation (MOO) one, were the objective functions are respectively the distortion function $dim\Big(Null\big(\mathcal{I}(\hat{\mathcal{S}}; \mathcal{S})\big)\Big) \to \infty$ and the memoryless-channel capacity $dim\Big(Null\big(\mathcal{I}(\mathcal{X}; \hat{\mathcal{X}})\big)\Big) \to 0$. Our novel algorithm applies an *Alternating optimisation* method. Subsequently, a robust version of the algorithm is discussed with regard to the perturbed PMFs – parameter uncertainties in general. The aforementioned robustness is actualised by exploiting the *simultaneous-perturbation-stochastic-approximation* (SPSA). The principles of *detectability-and-stabilisability* as well as synchronisability are explored, in addition to providing the simulations - by which the efficiency of our work is shown. We also calculate the total complexity of our proposed algorithms respectively as $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K}\big)\Big)$ and $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K} + 0.33\mathcal{K}\big)\Big)$. Our methodology is generic which can be applied to other fields of studies which are optimisation-driven.

Index Terms—Alternating optimisation, Branch-and-Bound, control-theory, *detectability-and-stabilisability*, distortion function, information-theory, joint-source-channel-coding, multi-objective optimisation, rate-distortion-theory, synchronisability.

I. Introduction

A. Overview

COntrol-theory with an integration of information theory has received brilliant attention in the recent decade. There arises a totally important question, in this context, how to optimally control the system over a noisy communication channel with feedback [1], [2], [3], [4], [5], [6], [7], [8]. Undoubtedly, the principle of *distortion-function* as the criterion of estimation-error[1] plays an entirely central role.

In the pioneer work [1], [2], [3], the optimal control problem over noisy-memoryless communication channels was

explored. In other words, information-theoretic point-of-view in control theory was introduced for the first time in the aforementioned work.

In relation to the information-constraint based issues in control theory, [4], [5], [6], [7], [8] are brilliant. Some probability-theoretic procedures as well as information-theoretic ones were applied in [4], [5]. In [6], invariance properties of non-linear stochasticities for dynamical systems were taken into account with regard to information constraints. In [7], the problem of optimal zero-delay coding as well as the estimation of a stochastic dynamical system over a memoryless noisy communication-channel were explored. A lower-bound on the smallest channel capacity was obtained with an arbitrarily and small reconstruction-error[2]. In [8], a generalisation of *Bode's integral* formula was derived for a large range of nonlinear systems.

In particular, the part of the joint-source-channel-coding (JSCC) has given a huge amount of attention in the literature [9], [10], [11], [12]. It is totally obvious that the channel called memory-less due to the feedback, or generally speaking, discrete memory-less (DM). In [9], over a Gaussian interference channel, a zero-delay JSCC framework was explored for the transmission of correlated-Gaussian-sources. In [10], the optimal second-order rates in a novel JSCC scheme was achieved for a general discrete memoryless (DM) channel, while the source was an Ergodic and irreducible *Markov-process*. Over DM channels in [11], the transmission of DM sources using variable-length lossy source-channel codes was carefully investigated with feedback. The optimum-error-exponent was theoretically derived. Over an arbitrary additive channel in [12], the transmission of an arbitrary memoryless source for a JSCC problem was informationtheoretically explored at a infinite-blocklength regime. In the research papers expressed above, the existence and definition of an optimisation problem for a JSCC scheme defining the relative optimisation variables undeniably lack.

In order to design the system model from a mathematical-physical point-of-view, a multi-objective optimisation (MOO) problem is commonly determined. In an MOO [13], [14], [15], [16], the objective function is a vector set containing some metrics. The *Augmented Lagrangian methods* [14], [15], [16] such as *alternating-direction-method-of-multipliers* (ADMM) [14] as well as the Branch-and-Bound (BB) one [15], [16] may

Makan Zamanipour is with Lahijan University, Shaghayegh Street, Po. Box 1616, Lahijan, 44131, Iran, makan.zamanipour.2015@ieee.org.

[1]Reconstruction-error.

[2]Low-distortion regime.

be the totally applicable candidates for providing a globally acceptable solution to the MOO problems.

Why BB? In order to provide an entirely appropriate response to this question we can start with the following phrase "why not BB?" [17], [18], [19]:

- BB, with the given tolerance, is able to solve all test problems in order to find their approximate global optimal solutions with a totally acceptable amount of complexity among all the algorithms defined for $NP-$hard problems;
- BB saves either the storage space, or the computational time;
- BB, as a class of deterministic global optimisation algorithm, is a powerful global optimisation tool for $NP-$hard problems;
- In BB, the number of linear constraints in all solved linear programming problems fully remain as minimum as possible;
- In BB, the searching space moderately condenses which consequently results in efficiently finding the global optimisation;
- In BB, it is not tightly necessary for each affine function in the objective function to be positive; and
- In BB, instead of the decision variable space, the branching process totally occurs in the outer space of the affine functions.

Meanwhile, it is interesting to note that other algorithms have also some limitations [17], [18], [19] which are given in the following which can convince us to strongly use BB:

- A Greedy Search Algorithm never reverses it thus has only one shot to perform. Meanwhile, it sometimes fails to calculate the global optimum. Pen-ultimately, it may be able to create the unique worst possible solution. Finally, the correctness of the solution is more-or-less questionable.
- Dynamic Programming has no general formation in reality. In addition, several problems entail totally different methods. Finally, it requires an infinite amount of memory due to the fact that it partitions the main problem into innumerable sub-problems.
- A Genetic Algorithm is too heuristic which can be a bottleneck in practice. Although it always finds the best solution, by the way it is not always able to present the exact one. Pen-ultimately, it is hard to represent the operators for this type of algorithms. Finally, it is totally time-consuming.
- A Simulated Annealing algorithm is absolutely time-consuming oriented as well as the fact that its precision is not significantly acceptable.

After understanding the fact that BB can be a good candidate for the MOO problems, our discussions continues as follows. The optimisation problem however may not be convex over the set of the variable-vector [20], [21]. However, we can principally apply a class of methods traditionally named as the *Alternating minimisation* ones[3]. In the aforementioned methods, only a single block of variables are supposed to be updated per iteration while the remaining ones are kept fully-fixed.

In practice, moreover, there needs to be a relaxation mechanism according to which one can tackle the parameter uncertainties. The exact estimation of the objective function is indeed impossible in reality due to parameter uncertainties. Simultaneous perturbation stochastic approximation (SPSA) [22] is a pseudo-gradient descent stochastic optimisation algorithm[4] which is of a purely useful nature for the cases in which the objective function[5] and its gradients are unknown[6]. Specifically, SPSA is totally important when the function is not differentiable at the optimising point. SPSA works as follows: (*i*) In this method, first the error is applied with regard to the fact that in successive iterations, the current solution is simultaneously perturbed by random offsets; and (*ii*) it starts with a random solution, subsequently, moving towards the optimum value[7]. The literature has proven the excellence of SPSA compared with other algorithms e.g. genetic ones in terms of the convergence-speed and accuracy [22].

In relation to the entirely central principle of *stabilisation* over *control networked systems*, innumerable research papers were presented to both Academia and Industry [23], [24], [25], [26], [27], [28], [29], [30]. The problem of the output feedback stabilisation was considered in [23] in the presence of channel-input imperfectness − packet dropout as well as quantisation error − for discrete-time networked control systems. The problems of attacks and how to stabilise the relative control networked system against the attacks were explored in [24]. In [25], the stabilisation issue was taken into consideration while the mission was to transmit the feedback information over external erasure channels as well as the signaling ones. In [26], two scalar plants across a Gaussian multiple access channel was stabilised. The states of both plants were separately encoded while subsequently being transmitted over a common channel to a joint decoder − i.e., a controller. The problem of controlling an unstable linear-plant was considered in [27] over an additive-white-Gaussian-noise channel with Gaussian disturbances. While the signaling rate of communication could be totally different from the sampling rate of the underlying plant, the average transmit-power was constrained. The authors in [27] proved the achievability of the rate required to capture the dynamics of the plant. The stabilisation of noiseless nonlinear dynamic systems was considered in [28] over limited capacity communication channels. The authors in [28] showed the relevance of the stability with the positive equilibrium Lyapunov exponents of the dynamic system − the magnitude of the unstable eigenvalues of system Jacobian − in addition to the Shannon channel capacity. The problem of control and modelling of discrete-time stochastic systems was taken into account in [29], [30] in the presence of packet dropouts. In the research papers expressed above, the information theoretic standpoint inevitably lacks.

---

[3]See e.g. *coordinate-decent*.

[4]Totally independent from the dimensions of the search-parameter space.

[5]Call *loss-function*.

[6]Which is a fundamental beneficiary compared to traditional gradient decent algorithms.

[7]It achieves the optimal value via the local optima of the objective function successfully moving towards the global optimum.

Fig. 1: Structure of the paper.
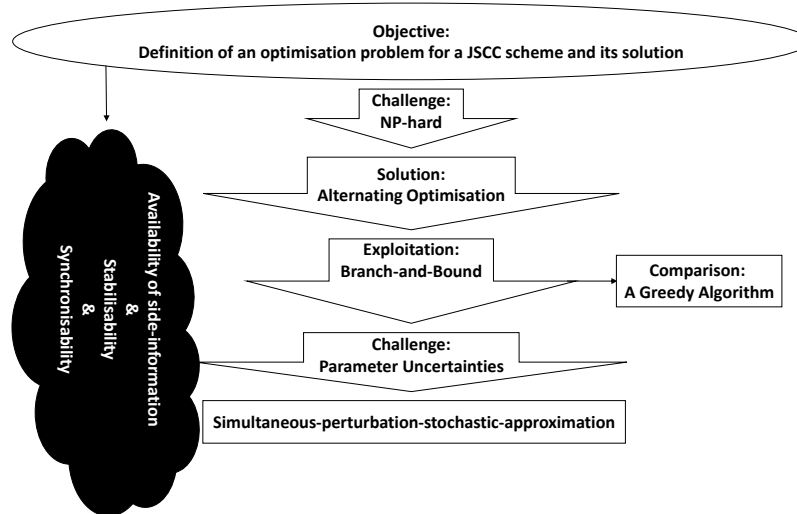

(a) General model of optimal control with feedback [1], [2], [3], [4], [5], [6], [7], [8].

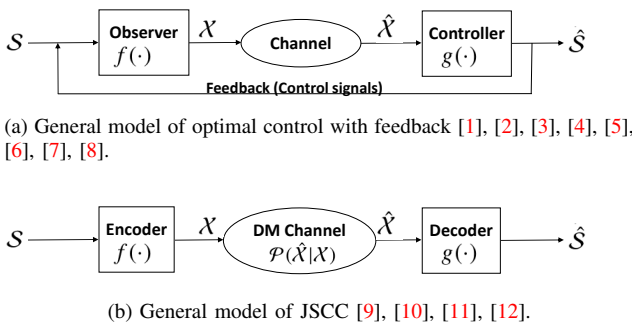(b) General model of JSCC [9], [10], [11], [12].

Fig. 2: System model.

B. Motivations and contributions

After carefully reviewing the literature, it seems that the definition of an optimisation problem for a JSCC scheme defining the relative optimisation variables still lacks. This is also in parallel with a lack of an information theoretic standpoint in optimal-control-theory. It is thus revealed that more acceptable versions of novelty and optimality as well as closed-form solutions are still required. These open trends motivate us to propose a new scheme relating to which our contribution is given in the following:

- (*i*) First, we define a JSCC scheme investigating its physical point-of-view, subsequently, we provide the mathematical expressions relating to this in the context of an optimisation problem. This type of optimisation problem, to the best of our knowledge is for the first time for a JSCC in-depth − for which the dual Lagrange is also derived.
- (*ii*) We technically propose a novel and efficient algorithm which conveniently finds the probability-mass-functions (PMFs) for lossy JSCC problems. To the best of our knowledge, this work, i.e., Blahut-Arimoto type algorithm has not been proposed in the literature for a JSCC scheme. Only [31] applied a Blahut-Arimoto algorithm

only in-practice without the theoretical background and the mathematical-physical implication of the problem, only from an analog-design viewpoint for JSCC schemes. In other words, the aforementioned work exploited the Blahut-Arimoto algorithm only-and-only for a practical design in relation to which the reader does not exactly know what the PMFs are. In our proposed algorithm, we apply BB method in order to deal with the MOO driven cost-function of JSCC. Our first proposed algorithm (Algorithm 1), indeed, finds the global optimum of the totally different objective-functions.

- (*iii*) Additionally, in order to tackle parameter uncertainties − e.g. outdated channel state information, channel-correlation etc − we also consider a perturbed version of PMFs in a SPSA based manner as Algorithm 2.
- (*iv*) Even for the case of side-information non-casually available at both transmit-and-receive sides, the problem is explored − and the required theoretical-mathematical expressions are derived.
- (*v*) Furthermore, we examine the stabilisability (controllability) for our framework according to the information-theoretic principle of *Data Processing Inequality*. Specifically, a game-theoretical point-of-view is derived to prove the principle of *detectability-and-stabilisability* in relation to the case of parameter uncertainties − that is, relating to Algorithm 2.
- (*vi*) Meanwhile, we explore the synchronisability of our proposed algorithms. And
- (*vii*) Finally, we evaluate the efficiency of our proposed algorithms by simulation results in addition to other relative topics such as complexity. We interestingly evaluate *detectability-and-stabilisability* for the case of correlated source-and-channel codebooks with the aid of the slope of the minimum square error (MSE) curve. We additionally add a comparison between our BB based procedure and a Greedy algorithm. We do this in order to convince the reader about the efficiency of our proposed BB based

framework compared to other possible algorithms.

We additionally re-emphasis that our procedure and proposed algorithm are generic which can be applied to other optimisation problems as well as the information-theoretic ones.[8].

C. General notation & preliminaries

Notations used throughout the paper are given in Table I. $\mathcal{Z}(\cdot; \theta)$ is the partition-function or Boltzmann (Gibbs) distribution.

TABLE I: List of notations.

Notation	Definition	Notation	Definition
$\mathcal{S}$	Source-symbol-set	$(\cdot)^\star$	Optimum
$\mathcal{X}$	Channel-symbol-set	$\mathcal{P}r(\cdot)$	Probability
$\mathcal{Y}$	Original Output	$\backslash$	*Set-minus*
$\hat{\mathcal{S}}$	Reconstructed $\mathcal{S}$	$\gamma \geq 0$	Learning-rate
$\hat{\mathcal{X}}$	Reconstructed $\mathcal{X}$	∇	Gradient
$\mathcal{R} + 1$	No. Branches	$dist(\cdot, \cdot)$	Distortion
$Null(\cdot)$	Null	$dim(\cdot)$	Dimension
$\lvert \cdot \rvert$	Absolute-value	Δ	Uncertainty
$\mathcal{D}_{kl}\{\cdot \lvert\rvert \cdot\}$	Kullback-Leibler	$exp(x)$	$\sum_{n=0}^{\infty} \frac{x^n}{n!}$
$\mathcal{I}(\cdot; \cdot)$	Mutual-Information	$\mathcal{H}(\cdot)$	Entropy
max	Maximum	min	Minimum
$\lfloor \cdot \rfloor$	Floor	$:=$	Defined by
$\mathcal{P}(\mathcal{A})$	Marginal PMF	$\mathcal{P}(\mathcal{A}, \mathcal{B})$	Joint PMF
$\mathcal{P}(\mathcal{A}\lvert\mathcal{B})$	Conditional PMF	$log(\cdot)$	Logarithm

ASSUMPTION 1. Anytime-CAPACITY ([1], [2], [3], [4]) *Throughout the paper, when we use the channel-capacity, we refer to the principle of anytime-capacity. The anytime-capacity of a channel is the least upper-bound of a given rate $\mathcal{R}$ according to which the channel can be used to construct a communication system. It must achieve a total stability over the time-horizon including the steady-state (relating to the Shannon-capacity).*

ASSUMPTION 2. NON-LINEARITY VS. LINEARITY ([1], [2], [3], [4]) *Generally speaking, both the Observer and the Controller can be linear or non-linear.*

ASSUMPTION 3. *In this paper, we do not talk about the feedback part since we consider the channel as a DM one.*

REMARK 1. CHANNEL-CORRELATION[9] & CORRELATION AMONG THE CHANNEL/SOURCE-CODEBOOKS [32], [33], [34], [35]. *(i) About the channel-correlation e.g. in multi-input multi-output control-networked-systems, we emphatically express that, generally speaking, this clearly momentous issue is considered in the context of parameter-uncertainties − for which the second algorithm is proposed. In other words, for the case of channel-correlation, a $\mathbb{Z}$−channel as a binary asymmetric channel is widely used [32] for which the sources and the received vectors are classified into two groups. Consequently, it is required to find two classes of PMFs over which the effect of the correlation appears. That is*

why we can only-and-only consider this case from a general point-of-view since we have nothing to do with this type of correlation coefficients since they cannot be our optimisation variables. (ii) In connection with the second issue, that is, correlation among the channel/source-codebooks [33], [34], [35], we should express that this issue may be able to have a positive knock-on effect on the performance − since it is called side-information. We go over this issue: we provide Propositions 3 and we investigate it in the section of numerical results.

D. Organisation

The rest of the paper is organised as follows. Problem formulation including the system set-up is given in Section II. Additionally, main results and our two proposed algorithms are discussed in Section III. Pen-ultimately, the evaluation of the framework by computer simulations are discussed in Section IV. Finally, Conclusions are listed in the last section as well as the fact that proofs are addressed in appendices. The structure of the paper is also give in Fig. 1.

II. SYSTEM MODEL

A JSCC scheme is depicted in Fig. 2 (subfigure b) with a complete unavailability of side-information[10] − of course we consider the availability of side-information in the next parts. In this figure, $\mathcal{S}$, $\mathcal{X}$, $\hat{\mathcal{S}}$ and $\hat{\mathcal{X}}$ are respectively the source-symbol-set and the channel-symbol one and their reconstructed versions. Additionally, let $\mathcal{Y}$ also be the original output of the system model. The main aim in a JSCC is, as discussed e.g. in [36], $\mathcal{P}r\left\{dist(\hat{\mathcal{S}}, \mathcal{S}) \geq \psi_1\right\} \leq \psi_2$ where the pair (ψ_1, ψ_2) is identified w.r.t. the given source-error-coding and channel-error-coding ψ_1 and ψ_2. Thus, the pair (f, g) is a (ψ_1, ψ_2) JSCC as well.

OBSERVER-CONTROLLER. Note that, the *source-channel separation-theorem*, indeed, allows us the following realisation. The Observer, that is, the total block of the encoder is spitted into two pieces: (*i*) a source encoder by which the source are encoded into bits; and (*ii*) a channel-encoder by which the aforementioned bits are converted into channel symbols.

III. MAIN RESULTS

A. Optimisation problem in order to guarantee $dim\left(Null\left(\mathcal{I}(\hat{\mathcal{S}}; \mathcal{S})\right)\right) \to \infty$ and $dim\left(Null\left(\mathcal{I}(\mathcal{X}; \hat{\mathcal{X}})\right)\right) \to 0$

OBJECTIVE. The main problem is about an optimisation over the distortion function $dim\left(Null\left(\mathcal{I}(\hat{\mathcal{S}}; \mathcal{S})\right)\right) \to \infty$, that is, $max \quad \mathcal{I}(\mathcal{S}; \hat{\mathcal{S}}\lvert\mathcal{X})$ and the channel capacity

---

[8]In information-theory and control-theory as well as computer science and applied science.

[9]Syndrome of the Channel.

[10]Relevant (but not necessarily perfect) information. One of the logical examples may be in relation to a Buyer who decides to purchase a used-good such as a car. Her success-rate will be fundamentally improved if she has knowledge about some relevant information over the aforementioned car. More specifically, she will purchase a better car with a higher quality if she is knowledgeable of e.g. its previous driver, whether he used to be an accident-maker teenager or an expert driver.

Algorithm 1 Our proposed algorithm.

$\textbf{Initialisation:}$ $t \leftarrow 0$, $x_{n,r}^{(t)} \leftarrow 0$, $\forall n \in \{1, \cdots, \mathcal{N}_0\}$, $\forall r \in \{0, \cdots, \mathcal{R}\}$, $\vartheta_1 \lll 1$, $\epsilon \ll 1$.

while $t < t_{max}$ **do**

    while $|x_{n,r}^{(t)} - x_{n,r}^{(t-1)}| \gg \vartheta_1$ **do**

        for $m \in \{1, \cdots, \mathcal{M}_0\}$ **do**

            for $n \in \{1, \cdots, \mathcal{N}_0\}$ **do**

                if $n \in \{\{1, \cdots, \mathcal{N}_0\}\} \setminus \{1, 2\}$ **do**

                Phase 1:

                    for $i \in \{1, \cdots, \mathcal{N}_0\}$, $i \neq n$ **do**

                        for $r \in \{0, \cdots, \mathcal{R}\}$ **do**

                          Compute $x_{n,r}^{(t)}$ by simultaneously:

                            (i) solving $\min_{x_{n,r}^{(t)}} \mathcal{F}_m(x_{n,r}^{(t)}, x_{i,r}^{(t)})$; and

                            (ii) fixing $x_{i,r}^{(t)} = 1 + \epsilon r$.

                Phase 2: $x_n^{(t)} \leftarrow x_{n,r^\star}^{(t)} = argmin \, \mathcal{F}_m(x_{n,r}^{(t)})$, $\forall r \in \{0, \cdots, \mathcal{R}\}$.

                elseif $n \in \{1, 2\}$ **do**

                $\min \, \alpha_1 + \alpha_2 \, s.t. \, 1; 2$.

                endif

            endfor

            endfor

        endfor

        endfor

    endwhile

    Update: $x_n^\star \leftarrow x_n^{(t)}$.

    $t \leftarrow t + 1$

endwhile

$\textbf{Output:}$ $x_n^\star$

end

$dim\left(Null\left(\mathcal{I}(\mathcal{X};\hat{\mathcal{X}})\right)\right) \rightarrow 0$, that is, $max \; \mathcal{I}(\mathcal{Y};\mathcal{X})$. We aim to find the optimum values for the PMFs according to which the aforementioned explications are realised. Towards this end, we need a *Blahut-Arimoto* type algorithm[11] which is the one that can find the optimum quantities for the PMFs.

Let us define a minimisation over the objective function $\underbrace{\phi_1 \mathcal{I}(\mathcal{S};\hat{\mathcal{S}}|\mathcal{X})}_{\mathcal{I}(\mathcal{S};\mathcal{Y})-\phi_1 \mathcal{I}(\mathcal{Y};\hat{\mathcal{S}}|\mathcal{X},\mathcal{S})} - \underbrace{\phi_2 \mathcal{I}(\mathcal{X};\hat{\mathcal{X}}|\mathcal{S})}_{\mathcal{I}(\mathcal{X};\mathcal{Y})-\phi_2 \mathcal{I}(\mathcal{Y};\hat{\mathcal{X}}|\mathcal{X},\mathcal{S})}$ – in which the first term is in relation to the source-coding part and the second term is in connection with the channel-coding part. Towards this end, the following PMFs are provisionally supposed to be defined as: (i) $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S},\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X},\mathcal{S})$; (ii) $\mathcal{P}(\hat{\mathcal{S}})$, $\mathcal{P}(\hat{\mathcal{X}})$; (iii) $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S})$, $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{S})$; (iv) $\mathcal{P}(\mathcal{Y}|\hat{\mathcal{X}})$, $\mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}})$; as well as (v) $\mathcal{P}(\mathcal{X})$. These PMFs are supposed to be defined with regard to $\mathcal{P}(\mathcal{A}|\mathcal{B}) = \sum_C \mathcal{P}(\mathcal{A}|C)\mathcal{P}(C|\mathcal{B})$ or $\mathcal{P}(\mathcal{A}|\mathcal{B}) = \frac{\mathcal{P}(\mathcal{A})}{\mathcal{Z}(\mathcal{B};\theta)} exp\{-\theta \mathcal{D}_{kl}\{\mathcal{P}(C|\mathcal{A})||\mathcal{P}(C|\mathcal{B})\}\}$ for the given $\mathcal{A}$, $\mathcal{B}$ and C while the $\mathcal{Z}(\cdot;\theta)$ operand is the partition-function according to the Boltzmann (Gibbs) distribution with the aid of the auxiliary value θ.

Now, having $\underline{\mathcal{P}}$: $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S},\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X},\mathcal{S})$, $\mathcal{P}(\hat{\mathcal{S}})$, $\mathcal{P}(\hat{\mathcal{X}})$, $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S})$, $\mathcal{P}(\hat{\mathcal{S}}|\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X})$, $\mathcal{P}(\hat{\mathcal{X}}|\mathcal{S})$, $\mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}})$, $\mathcal{P}(\mathcal{Y}|\hat{\mathcal{X}})$,

---

[11]See e.g. [37], [38] to understand what it is.

$\mathcal{P}(\mathcal{X})$, let us define the problem $\mathcal{P}_1$ as in Eq. (1)

$$\mathcal{P}_1 : \min_{\underline{\mathcal{P}}} \quad \underbrace{\phi_1 \mathcal{I}(\mathcal{S};\hat{\mathcal{S}}|\mathcal{X})}_{\mathcal{I}(\mathcal{S};\mathcal{Y})-\phi_1 \mathcal{I}(\mathcal{Y};\hat{\mathcal{S}}|\mathcal{X},\mathcal{S})} - \underbrace{\phi_2 \mathcal{I}(\mathcal{X};\hat{\mathcal{X}}|\mathcal{S})}_{\mathcal{I}(\mathcal{X};\mathcal{Y})-\phi_2 \mathcal{I}(\mathcal{Y};\hat{\mathcal{X}}|\mathcal{X},\mathcal{S})} \quad (1)$$
$$+ \alpha_1 + \alpha_2,$$

or

$$\max_{\underline{\mathcal{P}}} \; \phi_1 \mathcal{I}(\mathcal{Y};\hat{\mathcal{S}}|\mathcal{X},\mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y};\hat{\mathcal{X}}|\mathcal{X},\mathcal{S}) + \alpha_1 + \alpha_2,$$

subjected to

    (i)

$$1. \; ||\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S},\mathcal{X}) - \mathcal{P}(\hat{\mathcal{S}})|| \leq \alpha_1,$$

$$2. \; ||\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X},\mathcal{S}) - \mathcal{P}(\hat{\mathcal{X}})|| \leq \alpha_2;$$

    as well as (ii)

$$3. \; \mathcal{P}(\hat{\mathcal{S}}) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{S})\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S},\mathcal{X}),$$

which can be found with the aid of the $1-$st constraint

$$4. \; \mathcal{P}(\hat{\mathcal{X}}) = \sum_{\mathcal{X}} \mathcal{P}(\mathcal{X})\mathcal{P}(\hat{\mathcal{X}}|\mathcal{X},\mathcal{S}),$$

which can be found with the aid of the $2-$nd constraint in addition to the fact that we have to define $\mathcal{P}(\mathcal{X})$ – which is given in the following as the 11-th constraint; as well as (iii)

$$5. \; \mathcal{P}(\hat{\mathcal{S}}|\mathcal{S}) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{S};\theta_1)} exp\left\{-\theta_1 \mathcal{D}_{kl}\left\{\mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}})||\mathcal{P}(\mathcal{Y}|\mathcal{S})\right\}\right\},$$

Algorithm 2 Our proposed algorithm with SPSA.

$\textsc{Initialisation:}$ $t \leftarrow 0$, $x_{n,r}^{(t)} \leftarrow 0$, $\forall n \in \{1, \cdots, \mathcal{N}_0\}$, $\forall r \in \{0, \cdots, \mathcal{R}\}$, $\vartheta_1 \lll 1$, $\epsilon \ll 1$.

while $t < t_{max}$ **do**

    while $|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t-1)}| \gg \vartheta_1 \nabla \mathcal{F}_m(x_{n,r}^{(t)} \pm \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|)$

$$= \underbrace{\frac{\vartheta_1}{2\gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|} \left\{ \mathcal{F}_m(x_{n,r}^{(t)} + \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|) - \mathcal{F}_m(x_{n,r}^{(t)} - \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|) \right\}}_{SPSA} \textbf{ do}$$

$\cdots$ The same as Algorithm 1 $\cdots$

    endwhile

  Update: $x_n^\star \leftarrow x_n^{(t)}$.

  $t \leftarrow t + 1$

  endwhile

  $\textsc{Output:}$ $x_n^\star$

  end

---

or

$$5.\ \mathcal{P}(\hat{\mathcal{S}}|\mathcal{S}) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{S};\theta_1)} exp\left\{-\theta_1 \mathcal{D}_{kl}\left\{\mathcal{P}(X|\hat{\mathcal{S}})||\mathcal{P}(X|\mathcal{S})\right\}\right\},$$

or

$$5.\ \mathcal{P}(\hat{\mathcal{S}}|\mathcal{S}) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{S};\theta_1)} exp\left\{-\theta_1 \mathcal{D}_{kl}\left\{\mathcal{P}(\hat{X}|\hat{\mathcal{S}})||\mathcal{P}(\hat{X}|\mathcal{S})\right\}\right\},$$

– with the aid of $\mathcal{Y}$ since $\hat{X}$ or X would be of a partially useless nature for which some extra auxiliary variables should be defined –

$$6.\ \mathcal{P}(\hat{\mathcal{S}}|X) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(X;\theta_2)} exp\left\{-\theta_2 \mathcal{D}_{kl}\left\{\mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}})||\mathcal{P}(\mathcal{Y}|X)\right\}\right\},$$

or

$$6.\ \mathcal{P}(\hat{\mathcal{S}}|X) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(X;\theta_2)} exp\left\{-\theta_2 \mathcal{D}_{kl}\left\{\mathcal{P}(\mathcal{S}|\hat{\mathcal{S}})||\mathcal{P}(\mathcal{S}|X)\right\}\right\},$$

or

$$6.\ \mathcal{P}(\hat{\mathcal{S}}|X) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(X;\theta_2)} exp\left\{-\theta_2 \mathcal{D}_{kl}\left\{\mathcal{P}(\hat{X}|\hat{\mathcal{S}})||\mathcal{P}(\hat{X}|X)\right\}\right\},$$

while the first definition is selected in this paper since it entails less auxiliary variables, i.e., with the aid of $\mathcal{Y}$,

$$7.\ \mathcal{P}(\hat{X}|X) = \frac{\mathcal{P}(\hat{X})}{\mathcal{Z}(X;\theta_3)} exp\left\{-\theta_3 \mathcal{D}_{kl}\left\{\mathcal{P}(\mathcal{Y}|\hat{X})||\mathcal{P}(\mathcal{Y}|X)\right\}\right\},$$

or

$$7.\ \mathcal{P}(\hat{X}|X) = \frac{\mathcal{P}(\hat{X})}{\mathcal{Z}(X;\theta_3)} exp\left\{-\theta_3 \mathcal{D}_{kl}\left\{\mathcal{P}(\hat{\mathcal{S}}|\hat{X})||\mathcal{P}(\hat{\mathcal{S}}|X)\right\}\right\},$$

while the first definition is selected in this paper since it entails less auxiliary variables, i.e., with the aid of $\mathcal{Y}$,

$$8.\ \mathcal{P}(\hat{X}|\mathcal{S}) = \frac{\mathcal{P}(\hat{X})}{\mathcal{Z}(\mathcal{S};\theta_4)} exp\left\{-\theta_4 \mathcal{D}_{kl}\left\{\mathcal{P}(\mathcal{Y}|\hat{X})||\mathcal{P}(\mathcal{Y}|\mathcal{S})\right\}\right\},$$

or

$$8.\ \mathcal{P}(\hat{X}|\mathcal{S}) = \frac{\mathcal{P}(\hat{X})}{\mathcal{Z}(\mathcal{S};\theta_4)} exp\left\{-\theta_4 \mathcal{D}_{kl}\left\{\mathcal{P}(X|\hat{X})||\mathcal{P}(X|\mathcal{S})\right\}\right\},$$

or

$$8.\ \mathcal{P}(\hat{X}|\mathcal{S}) = \frac{\mathcal{P}(\hat{X})}{\mathcal{Z}(\mathcal{S};\theta_4)} exp\left\{-\theta_4 \mathcal{D}_{kl}\left\{\mathcal{P}(\hat{\mathcal{S}}|\hat{X})||\mathcal{P}(\hat{\mathcal{S}}|\mathcal{S})\right\}\right\},$$

while the first definition is selected in this paper since it entails less auxiliary variables, i.e., with the aid of $\mathcal{Y}$; as well as (*iv*)

$$9.\ \mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}}) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y}|\mathcal{S})\mathcal{P}(\mathcal{S}|\hat{\mathcal{S}}),$$

or

$$9.\ \mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}}) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y}|X)\mathcal{P}(X|\hat{\mathcal{S}}),$$

between which there seems to be no preference

$$10.\ \mathcal{P}(\mathcal{Y}|\hat{X}) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y}|\mathcal{S})\mathcal{P}(\mathcal{S}|\hat{X}),$$

or

$$10.\ \mathcal{P}(\mathcal{Y}|\hat{X}) = \sum_{X} \mathcal{P}(\mathcal{Y}|X)\mathcal{P}(X|\hat{X}),$$

between which there seems to be no preference; as well as (*v*)

$$11.\ \mathcal{P}(X) = \sum_{\hat{X}} \mathcal{P}(\hat{X})\mathcal{P}(X|\hat{X}).$$

As totally obvious, we need to additionally define the following

$$12.\ \sum_i \mathcal{P}(i) = 1, i \in \{\mathcal{Y}, \hat{\mathcal{S}}, X, \mathcal{S}, \hat{X}\},$$

which states that the sum of PMFs over all their relative values in the relating sample space is equal to 1. Meanwhile, ϕ_1 and ϕ_2 are Lagrange-multipliers. The constraints 1 and 2 are also interpreted by defining the slack variables $\alpha_{i \in \{1,2\}}$ according to the theory of *concentration-of-measure* being upper-bounded to $f(e^{-x})$.

REMARK 2. *The PMFs defined above being in cooperation with each other is strongly illustrated in Fig. 3. As a case in point, it is totally obvious that in order to define $\mathcal{P}(\hat{\mathcal{S}}|X)$, we need to know $\mathcal{P}(\mathcal{Y}|\hat{\mathcal{S}})$ – relating to the 6st constraint.*

PROPOSITION 1: *The physical point-of-view of the JSCC illustrated in Fig. 2 can be theoretically analysed by the problem $\mathcal{P}_1$ as in Eq. (1).* ∎

PROOF: See Appendix A for the proof. ∎

LEMMA 1: *Problem $\mathcal{P}_1$ has a dual-Lagrange with the duality-gap of zero. The dual problem is indeed obtained by finding the infimum of all the primal variables in the Lagrangian function.* ∎

$$\mathcal{L} = \phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}} | \mathcal{S}, \mathcal{X}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}} | \mathcal{S}, \mathcal{X}) + \alpha_1 + \alpha_2$$

$$+ \beta_1 \left\{ ||\mathcal{P}(\hat{\mathcal{S}} | \mathcal{S}, \mathcal{X}) - \mathcal{P}(\hat{\mathcal{S}})|| - \alpha_1 \le 0 \right\} + \beta_2 \left\{ ||\mathcal{P}(\hat{\mathcal{X}} | \mathcal{X}, \mathcal{S}) - \mathcal{P}(\hat{\mathcal{X}})|| - \alpha_2 \le 0 \right\}$$

$$+ \beta_3 \left\{ \mathcal{P}(\hat{\mathcal{S}}) - \sum_{\mathcal{S}} \mathcal{P}(\mathcal{S}) \mathcal{P}(\hat{\mathcal{S}} | \mathcal{S}, \mathcal{X}) \right\} + \beta_4 \left\{ \mathcal{P}(\hat{\mathcal{X}}) - \sum_{\mathcal{X}} \mathcal{P}(\mathcal{X}) \mathcal{P}(\hat{\mathcal{X}} | \mathcal{X}, \mathcal{S}) \right\}$$

$$+ \beta_5 \left\{ \mathcal{P}(\hat{\mathcal{S}} | \mathcal{S}) - \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{S}; \theta_1)} exp \left\{ -\theta_1 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{S}}) || \mathcal{P}(\mathcal{Y} | \mathcal{S}) \right\} \right\} \right\}$$

$$+ \beta_6 \left\{ \mathcal{P}(\hat{\mathcal{S}} | \mathcal{X}) - \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{X}; \theta_2)} exp \left\{ -\theta_2 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{S}}) || \mathcal{P}(\mathcal{Y} | \mathcal{X}) \right\} \right\} \right\} \qquad (2)$$

$$+ \beta_7 \left\{ \mathcal{P}(\hat{\mathcal{X}} | \mathcal{X}) - \frac{\mathcal{P}(\hat{\mathcal{X}})}{\mathcal{Z}(\mathcal{X}; \theta_3)} exp \left\{ -\theta_3 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{X}}) || \mathcal{P}(\mathcal{Y} | \mathcal{X}) \right\} \right\} \right\}$$

$$+ \beta_8 \left\{ \mathcal{P}(\hat{\mathcal{X}} | \mathcal{S}) - \frac{\mathcal{P}(\hat{\mathcal{X}})}{\mathcal{Z}(\mathcal{S}; \theta_4)} exp \left\{ -\theta_4 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{X}}) || \mathcal{P}(\mathcal{Y} | \mathcal{S}) \right\} \right\} \right\}$$

$$+ \beta_9 \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{S}}) - \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y} | \mathcal{S}) \mathcal{P}(\mathcal{S} | \hat{\mathcal{S}}) \right\} + \beta_{10} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{X}}) - \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y} | \mathcal{S}) \mathcal{P}(\mathcal{S} | \hat{\mathcal{X}}) \right\}$$

$$+ \beta_{11} \left\{ \mathcal{P}(\mathcal{X}) - \sum_{\hat{\mathcal{X}}} \mathcal{P}(\hat{\mathcal{X}}) \mathcal{P}(\mathcal{X} | \hat{\mathcal{X}}) \right\} + \beta_{12} \left\{ \sum_{i} \mathcal{P}(i) - 1, i \in \{\mathcal{Y}, \hat{\mathcal{S}}, \mathcal{X}, \mathcal{S}, \hat{\mathcal{X}}\} \right\}.$$
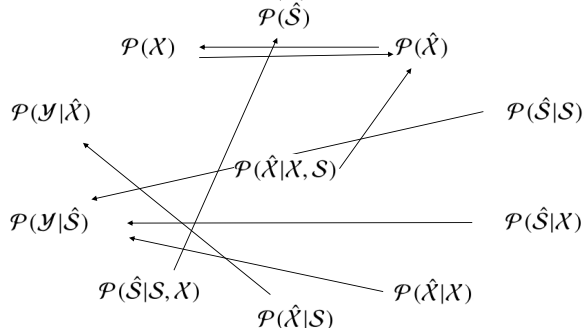

Fig. 3: The PMFs being in cooperation with each other.

PROOF: See e.g. [37], [38] for the proof. ∎

PROPOSITION 2: *Eq.* (2) *is the dual-problem for Eq.* (1) *where* $\beta_{k \in \{1, \cdots, 12\}}$ *are Lagrange-multipliers. For Eq.* (2), $\underline{\mathcal{P}^\star}$: $\mathcal{P}^\star(\hat{\mathcal{S}} | \mathcal{S}, \mathcal{X})$, $\mathcal{P}^\star(\hat{\mathcal{X}} | \mathcal{X}, \mathcal{S})$, $\mathcal{P}^\star(\hat{\mathcal{S}})$, $\mathcal{P}^\star(\hat{\mathcal{X}})$, $\mathcal{P}^\star(\hat{\mathcal{S}} | \mathcal{S})$, $\overline{\mathcal{P}^\star(\hat{\mathcal{S}} | \mathcal{X})}$, $\mathcal{P}^\star(\hat{\mathcal{X}} | \mathcal{X})$, $\mathcal{P}^\star(\hat{\mathcal{X}} | \mathcal{S})$, $\mathcal{P}^\star(\mathcal{Y} | \hat{\mathcal{S}})$, $\mathcal{P}^\star(\mathcal{Y} | \hat{\mathcal{X}})$, $\mathcal{P}^\star(\mathcal{X})$ *are obtained by applying* $\frac{\partial \mathcal{L}}{\partial \beta_{k \in \{1, \cdots, 12\}}} = 0$ *as well as* $\frac{\partial \mathcal{L}}{\partial \mathcal{P}} = 0$. ∎

PROOF: See Appendix B for the proof. ∎

PROPOSITION 3: *In the case of having a side-information at both sides for both source-coding and channel-coding* [33], *we have* $\underbrace{\phi_1 \mathcal{I}(\mathcal{S}; \hat{\mathcal{S}} | \mathcal{X}, \mathcal{U})}_{\mathcal{I}(\mathcal{S}; \mathcal{Y}) - \phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}} | \mathcal{X}, \mathcal{S}, \mathcal{U})} - \underbrace{\phi_2 \mathcal{I}(\mathcal{X}; \hat{\mathcal{X}} | \mathcal{S}, \mathcal{U})}_{\mathcal{I}(\mathcal{X}; \mathcal{Y}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}} | \mathcal{X}, \mathcal{S}, \mathcal{U})}$ *for which the* 1−*st and the* 2−*nd constraints are changed to* $||\mathcal{P}(\hat{\mathcal{S}} | \mathcal{X}, \mathcal{S}, \mathcal{U}) - \mathcal{P}(\hat{\mathcal{S}})|| \le \alpha_3$ *and* $||\mathcal{P}(\hat{\mathcal{X}} | \mathcal{X}, \mathcal{S}, \mathcal{U}) - \mathcal{P}(\hat{\mathcal{X}})|| \le \alpha_4$. *In addition, we also have the two extra constraints as follows* $\mathcal{P}(\hat{\mathcal{S}} | \mathcal{U}) = \frac{\mathcal{P}(\hat{\mathcal{S}})}{\mathcal{Z}(\mathcal{U}; \theta_5)} exp \left\{ -\theta_5 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{S}}) || \mathcal{P}(\mathcal{Y} | \mathcal{U}) \right\} \right\}$ *and* $\mathcal{P}(\hat{\mathcal{X}} | \mathcal{U}) = \frac{\mathcal{P}(\hat{\mathcal{X}})}{\mathcal{Z}(\mathcal{U}; \theta_6)} exp \left\{ -\theta_6 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \hat{\mathcal{X}}) || \mathcal{P}(\mathcal{Y} | \mathcal{U}) \right\} \right\}$ *– for the source-coding part – as well as another one as* $\mathcal{P}(\mathcal{X} | \mathcal{U}) = \frac{\mathcal{P}(\mathcal{X})}{\mathcal{Z}(\mathcal{U}; \theta_7)} exp \left\{ -\theta_7 \mathcal{D}_{kl} \left\{ \mathcal{P}(\mathcal{Y} | \mathcal{X}) || \mathcal{P}(\mathcal{Y} | \mathcal{U}) \right\} \right\}$ *– for*

the channel-coding part – which entail to define three new and consequent extra Lagrange-multipliers β_{13}, β_{14} *and* β_{15}, *respectively.* ∎

PROOF: See e.g. [33] for more investigation on the effect of the side-information. ∎

CHALLENGE. On the other hand, the problem $\mathcal{P}_1$ is non-convex and $NP-$hard. It is, more specifically, not jointly convex in all the PMF sets, however, it is separately convex in only one of them.

B. Blahut-Arimoto type Algorithm 1

Problem $\mathcal{P}_1$ is a MOO one from an optimisation-based perspective which is, generally speaking, expressed as $\mathcal{P}_2$ given as $\min_{x_1, \cdots, x_n} \mathcal{F}_1(x_1, \cdots, x_n), \cdots, \mathcal{F}_m(x_1, \cdots, x_n)$, $\forall n = 1, \cdots, \mathcal{N}_0, \forall m = 1, \cdots, \mathcal{M}_0$. BB as a branch of *Augment-Lagrangian methods*[12] can solve this problem.

Overall, in Algorithm 1, each iteration is divided into $\mathcal{N}_0$. Subsequently, each of the aforementioned time-slots is divided into $\mathcal{R} + 1$ *branches* according to the BB method (Phase 1). After bounding the $\mathcal{R} + 1$ optimal solutions in parallel of each other (Phase 2), the final result is achieved.

In particular, in this algorithm, $\mathcal{N}_0$ stands for the number of β_k as the number of the optimisation variables, where $\mathcal{M}_0$ is 2 since there seems to be only two mutual-information functions as the objective functions. The **if** operand invokes the optimisation problem for those constraints that are supposed to be considered w.r.t. BB, prior of which the constraints 1 and 2 in terms of e.g. a bi-level (inner) sub-problem are considered. In relation to BB, in the first phase (Phase 1) the operation of branching is done, subsequently, in the next phase (Phase 2)

---

[12]See also *Douglas-Rachford splitting*: Applying Douglas-Rachford splitting method to the primal problem is equivalent to applying ADMM to the dual one with step-size equal to 1.
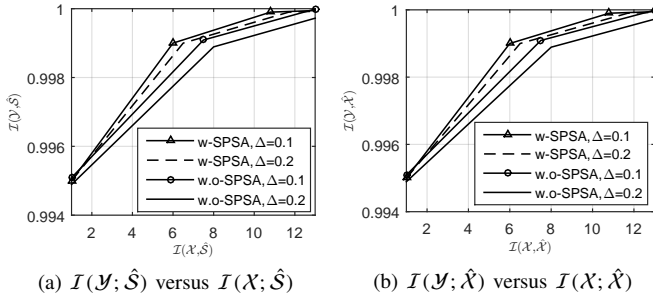
(a) $\mathcal{I}(\mathcal{Y};\hat{\mathcal{S}})$ versus $\mathcal{I}(\mathcal{X};\hat{\mathcal{S}})$ (b) $\mathcal{I}(\mathcal{Y};\hat{\mathcal{X}})$ versus $\mathcal{I}(\mathcal{X};\hat{\mathcal{X}})$

Fig. 4: Algorithm 1 versus Algorithm 2 while changing the amount of uncertainties, i.e., Δ.

(a) Iterations from 100 to 400.   (b) Iterations from 300 to 400.

Fig. 5: Algorithm 1 versus Algorithm 2: Accuracy vs. iterations while changing the amount of uncertainties, i.e., Δ.

the bounding is fulfilled. More specifically, in the first phase, only one optimisation variable is going to be optimally found while fixing the other ones. In the r−th branch, in fact, we assign the value of $1 + \epsilon r$ in which the quantities of 1 and ϵ are respectively arbitrary and very-small. Furthermore, x_n stands for the PMFs. Finally, it should be noted that the first *while-loop* is in relation to the iterations where the second one is in relation to the convergence.

C. How to deal with parameter uncertainties: Blahut-Arimoto type Algorithm 2

Meanwhile, there arises a clearly momentous question what if the PMFs are perturbed, or generally speaking, we have parameter uncertainty relating to them, i.e., $\mathcal{P} + \Delta$ holds. Algorithm 2 can solve the above-mentioned issue applying SPSA. The quantity of Δ, i.e, $|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|$ is widely taken into consideration as a Bernoulli distributed random variable. Indeed, the learning rate plays a vital role in the sense that the error quantity $\nabla \mathcal{F}_m(x_{n,r}^{(t)} \pm \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|) :=$ $\frac{\mathcal{F}_m(x_{n,r}^{(t)} + \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|) - \mathcal{F}_m(x_{n,r}^{(t)} - \gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|)}{2\gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|}$ could be minimised as less as possible. The term of $2\gamma_{n,r}^{(t)}|x_{n,r}^{(t)} - \hat{x}_{n,r}^{(t)}|$ indicates the difference[13] between the upper-bound and the lower-bound of the *simultaneous perturbations* over which the error quantity defined above is the sum of observations taken. In order to understand SPSA and for more justification over this see e.g. [22].

D. Stabilisability (controllability): See e.g. [23], [24], [25], [26], [27], [28], [29], [30]

THEOREM 1 *In relation to the case of having no parameter uncertainty, our JSCC is stabilisable.* ■

PROOF: See Appendix C for the proof. ■

THEOREM 2 *In relation to the case of having parameter uncertainty, our JSCC is stabilisable.* ■

PROOF: See Appendix D for the proof. ■

E. Synchronisability

THEOREM 3. *Synchronisation of our JSCC scheme is exponentially achievable.* ■
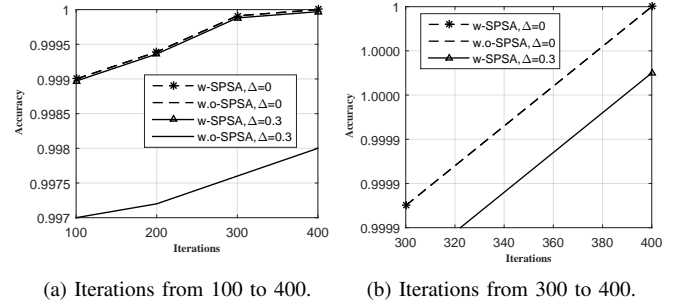
[13]Also called *velocity*.

PROOF: See Appendix E for the proof. ■

IV. NUMERICAL RESULTS

We in this part, show some comparisons. We tested our simulations through *CPLEX* and *C++* on a system with a RAM 8 G with the Processor of speed 2.4 GHz on *Ubuntu* 16.04. Some parameters are assumed as follows: $\Delta \in [0, 1]$ and the data sets are Bernoulli distributed.

REMARK 3. *Straightforwardly speaking, as there seems to be no work in the literature in relation to a Blahut-Arimoto type algorithm for JSCC, we thus do not have anything to compare with our proposed ones.*

In Fig. 4, $\mathcal{I}(\mathcal{Y};\hat{\mathcal{S}})$ is depicted versus the $\mathcal{I}(\mathcal{X};\hat{\mathcal{S}})$ regime in the subfigure (a). Moreover, $\mathcal{I}(\mathcal{Y};\hat{\mathcal{X}})$ is illustrated against the $\mathcal{I}(\mathcal{X};\hat{\mathcal{X}})$ regime in the subfigure (b). According to the aforementioned sub-figures, it is revealed that our proposed Algorithm 2, which is an extended version of Algorithm 1 adding SPSA, is extremely robust for the high amount of perturbation.

In Fig. 5, the accuracy of our proposed algorithm is also obvious. For this figure, we calculate the accuracy as $1 - \frac{|\mathcal{P}^\star - \mathcal{P}|}{\mathcal{P}}$. As also obvious in the subfigure (b), Algorithm 1 fails for the high range of parameter uncertainty; something that is compensated by adding SPSA in Algorithm 2.

In Fig. 6, 1 minus the accuracy − shown in Fig. 5 − that is, the minimum square error (MSE) is illustrated versus the iteration regime. Simply, our first proposed algorithm is used while $\Delta = 0$. As obvious, the performance, as discussed in *Remark 1*, with correlated source/channel-codebooks is considerably more adequate. Meanwhile, after taking a careful look at the two curves, it is revealed that the slope of the curve in relation to the with-correlation case − depicted as w-correlation − is 2 times higher than that of the other one, that is, without-correlation − shown as w.o-correlation. The aforementioned slopes are calculated as $1e − 09$ in comparison with $0.5e − 09$. This less slope for the without-correlation case proves that *detectability-and-stabilisability* needs to be more taken into account for it − as *Data Processing Inequality* theoretically imposes.

Let us here make an essential comparison between the MSE values achieved from the BB based Algorithm 1 with the following Greedy one. Simply, while $\Delta = 0$, we do this comparison in Fig. 7. The goal of this comparison is to see if our BB based procedure is efficient compared with other

possible methods. The response, as absolutely inevitable from Fig. 7, is an emphatically affirmative one, specifically for the higher iteration regime. In this figure, MSE_1 and MSE_3 stand for the MSE values obtained from respectively Algorithm1 and Algorithm 3 − Greedy one. In addition, the BB based Algorithm 2 − with-SPSA − is also compared to Algorithm 3 in the second sub-figure while $\Delta = 0.4$.

COMPLEXITY: The complexity of Algorithm 2 is calculated as follows. In each iteration $\mathcal{T}$, we experience the computational complexity of $O\big(\mathcal{K}\mathcal{M}_0\big)$ for the search processes where $\mathcal{K} = 1^{N_0}(\mathcal{R}_0 + 1)\mathcal{N}_0$. Furthermore, we have the complexity of $O\big(\mathcal{K}\log\mathcal{K}\big)$ in order to update the variables, in addition to $O\big(0.33\mathcal{K}\big)$ relating to the SPSA part in the algorithm. Therefore, the total complexity we experience is $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K} + 0.33\mathcal{K}\big)\Big)$. Take into account that the complexity of Algorithm 1 is $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K}\big)\Big)$ since there seems to be no SPSA part. Meanwhile, the complexity for the Greedy algorithm − that is Algorithm 3 − is $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K}\big)^2\Big)$.

Algorithm 3 A Greedy algorithm to find $\underline{\mathcal{P}}$.

INITIALISATION: Lagrange multipliers $\phi_v, \forall v \in \{1,2\}$ are arbitrary; Iteration No. $\frac{1}{\varsigma}$ is arbitrary, Ground Set $\Xi \leftarrow \{0,1,2,\cdots,\frac{1}{\varsigma} - 1\}$, Feasible Set $\Omega \leftarrow \emptyset$, and $f(\Omega) := \phi_1 \mathcal{I}(\mathcal{Y};\hat{\mathcal{S}}|\mathcal{X},\mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y};\hat{\mathcal{X}}|\mathcal{X},\mathcal{S}) + \alpha_1 + \alpha_2$.
    while $|f(\Omega)| < f_{max}$ **do**
        $\Omega_{new} \leftarrow \max\limits_{\underline{\mathcal{P}} \in \Xi \setminus \Omega} f(\Omega \cup \{\underline{\mathcal{P}}\}) - f(\Omega), \ s.t. \ 1-12$
        $\Omega \leftarrow \Omega \cup \{\Omega_{new}\}$
    endwhile
OUTPUT: Ω
end

V. CONCLUSION

A novel optimisation problem was defined for a lossy JSCC framework in this paper. An efficient algorithm was consequently proposed aimed at finding the PMFs. Especially, our proposed BB based algorithm globally optimised an MOO based cost-function for JSCC. Additionally, a perturbed version of PMFs was taken into consideration in terms of parameter uncertainties. According to this, another algorithm was proposed adding SPSA. Novel mathematical solutions such as in the case of channel-correlation as well as detectability-and-stabilisability in addition to synchronisability were provided. The efficiency of our proposed algorithms by simulation results was characterised − as well as a validation over our BB based procedure comparing to a possible Greedy algorithm. Our procedure and proposed algorithms are generic which can be applied to every optimisation problem. The total complexity was calculated as $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K}\big)\Big)$ and $O\Big(\mathcal{T}\mathcal{K}\mathcal{M}_0\big(\mathcal{K}\log\mathcal{K} + 0.33\mathcal{K}\big)\Big)$ respectively for our two proposed algorithms.
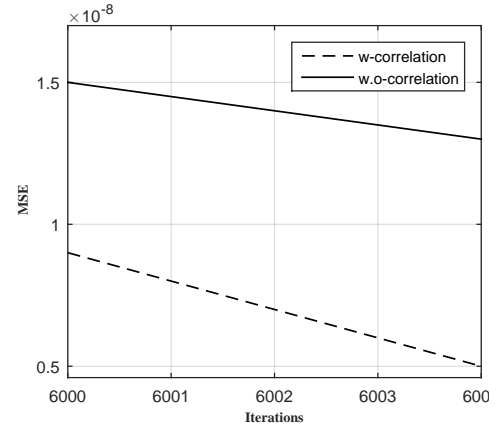

Fig. 6: MSE vs. iterations: with-correlation case compared to the with-out one.

APPENDIX A
PROOF OF PROPOSITION 1

The proof can be conveniently guaranteed as follows. It is required for us: (*i*) to maximise the channel capacity $\mathcal{I}(\hat{\mathcal{X}};\mathcal{X})$ in order to find $\mathcal{P}^\star(\mathcal{X})$ which is a concave function over it; and (*ii*) to minimise the distortion function $\mathcal{I}(\mathcal{S};\hat{\mathcal{S}})$ in order to find $\mathcal{P}^\star(\hat{\mathcal{S}})$ which is a convex function over it.

Therefore, the proof is completed. ∎

APPENDIX B
PROOF OF PROPOSITION 1

Since all the constraints are linear as $p - p_0$, all the $\beta_{k \in \{1,\cdots,12\}}$ are identified by 1.

Additionally, *Augmented Lagrangian function* has an extra term as $\frac{\beta}{2}||p - p_0||_2^2$ where β is the penalty-parameter; something that is ignored here for the sake of simplicity and presentation aimed at preserving the space.

This completes the proof ∎

APPENDIX C
PROOF OF THEOREM 1

According to the information-theoretic principle of *Data Processing Inequality*, the *loss in stabilisability* of the system has a positive value. Data Processing Inequality indeed indicates that the *detectability* of a system fully proves the stabilisability, but not necessarily vice versa.

A system, even if initially unstable, is stabilisable in a mean squared sense[14] if and only if $lim_{a\to\infty}\mathbb{E}\{|Y(a)|^2\} \leq \vartheta^2\mathbb{E}\{|X(a)|^2\}$. a is the time instance, ϑ^2 is the attenuation of the system − for which we have $\vartheta^2 \in [0,1]$ − and $\mathbb{E}\{\cdot\}$ stands for the Expected-value operand. In order to guarantee the detectability, consequently, the stabilisability (controllability) of the system, we say that: the system is *detectable-and-stabilisable in the mean squared* sense if and only if the *causality* − existence − of the encoder-decoder pair (f,g) or (ψ_1,ψ_2) − defined in Section II − is achievable, or correspondingly: (*i*) when $lim_{a\to\infty}\mathbb{E}\{|X(a)|^2\} < cte$ as well as

---

[14]See e.g. [23], [24], [25], [26], [27], [28], [29], [30] to understand what *mean-squared stabilisable* is.
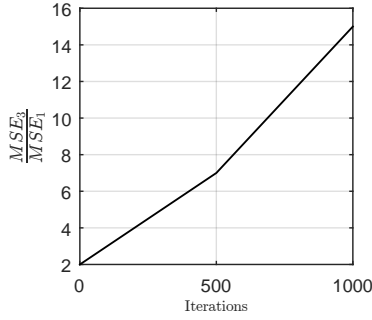
(a) MSE of Algorithm 3 divided to the MSE of Algorithm 1, $\Delta = 0$.
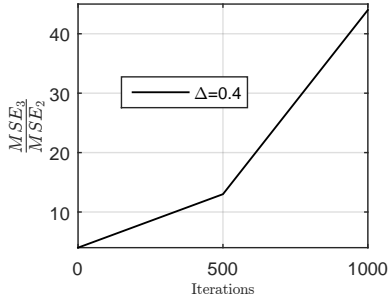
(b) MSE of Algorithm 3 divided to the MSE of Algorithm 2, $\Delta = 0.4$.

Fig. 7: MSE vs. iterations.

$lim_{a \to \infty} \mathcal{I}\left(\mathcal{S}(a); \hat{\mathcal{S}}(a)|\mathcal{X}(a)\right) < cte$ hold where *cte* stands for a constant value, as the necessary condition; and (*ii*) $max \; \phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}}|\mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}}|\mathcal{X}, \mathcal{S})$ is always feasible-and-solvable − since the *duality gap* was proven to be equated to zero in *Lemma 1* − as the necessary condition.

This completes the proof. ■

APPENDIX D
PROOF OF THEOREM 2

As discussed in Appendix C, for a system with $X_{a+1} \propto \eta X_a$, the detectability-and-stabilisability relies deeply upon the following: $0 \leq \log |\eta| < C_{max}$ where η and C_{max} stand respectively for the transmission-rate as well as the maximum-capacity of the system.

Let us use a game-theoretical point-of-view for the proof as follows[15]. Recall $\vartheta^2 \in [0, 1]$ defined in Appendix C and define $l(A, \Delta)) := \vartheta^2$ and consider two types of players as $A_i, \; \forall i \in [0, I - 1]$ − as the time-delay arose from the noisy communication channel − and $\Delta_j, \; \forall j \in [0, J - 1]$ − standing for the error Δ relating to parameter uncertainties. Over a 2-dimensional $i-$and$-j$ axis, of course it is not impossible to experience the optimal pair $(A^\star; \Delta^\star)$ as the saddle-point in which the *Nash-Equilibrium* is satisfied which realises $|l^\star(A^\star, \Delta^\star)|^2$ in accordance with $\phi_v^\star, \; \forall v \in \{1, 2\}$ − in relation to $\phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}}|\mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}}|\mathcal{X}, \mathcal{S})$.

COROLLARY 1: $\phi_v = 0, \; \forall v \in \{1, 2\}$, *in relation to* $\phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}}|\mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}}|\mathcal{X}, \mathcal{S})$, *is in accordance with the*

---

case of an unstable, but detectable-and-stabilisable sysetm. Furthermore, $\phi_v = \infty, \; \forall v \in \{1, 2\}$, *in relation to the ascending function* $\phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}}|\mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}}|\mathcal{X}, \mathcal{S})$ *over* $\phi_v, \; \forall v \in \{1, 2\}$ *is in correspondence to the case of a fully detectable-and-stabilisable system.* ■

This completes the proof. ■

APPENDIX E
PROOF OF THEOREM 3

DEFINITION 1. SYNCHRONISABILITY [40], [41], [42], [43]: *A function of the structure of* $e^{-\beta a}$ *is called exponentially synchronisable over the time-zone* $a \in [0, +\infty)$.

According to the principle of *joint-typicality-measure*[16], $max \; \phi_1 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{S}}|\mathcal{X}, \mathcal{S}) - \phi_2 \mathcal{I}(\mathcal{Y}; \hat{\mathcal{X}}|\mathcal{X}, \mathcal{S})$ has an error with the following structure: $e^{-\mathcal{H}(\cdot)a}$ over the time-zone $a \in [0, +\infty)$. So, it is exponentally synchronisable.

This completes the proof. ■

REFERENCES

[1] A. Sahai, S. Mitter, "The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link-Part I: Scalar Systems," *IEEE Trans. Info. Theory*, Vol. 52, no. 8, pp. 3369-3395, 2006.
[2] S. Tatikonda, A. Sahai, S. Mitter, "Stochastic linear control over a communication channel," *IEEE Trans. Auto. Control* Vol. 49, no. 9, pp. 1549-1561, 2004.
[3] S. Tatikonda, S. Mitter, "Control over noisy channels," *IEEE Trans. Auto. Control* Vol. 49, no. 7, pp. 1196-1201, 2004.
[4] S. Yuksel, T. Basar, "Control Over Noisy Forward and Reverse Channels," *IEEE Trans. Auto. Control* Vol. 56, no. 5, pp. 1014-1029, 2011.
[5] C. Kawan, S. Yuksel, "Entropy bounds on state estimation for stochastic non-linear systems under information constraints," https://arxiv.org/abs/1612.00564, 2018.
[6] C. Kawan, S. Yuksel, "Invariance Properties of Nonlinear Stochastic Dynamical Systems under Information Constraints," https://arxiv.org/abs/1901.02825, 2019.
[7] C. Kawan, Se. Yuksel, "Metric and topological entropy bounds for optimal coding of stochastic dynamical systems," *IEEE Trans. Auto. Control* Vol. PP, no. 99, pp. 1-1, 2019.
[8] S. Yuksel, "Stationary and Ergodic Properties of Stochastic NonLinear Systems Controlled over Communication Channels," *SIAM J. Control Optim.*, Vol. 54, no. 5, pp. 2844-2871, 2016.
[9] X. Chen, "Zero-Delay Gaussian Joint Source-Channel Coding for the Interference Channel," *IEEE Commun. Letters*, Vol. 22, no. 4, pp. 712715, 2018.
[10] R. Yaguchi, M. Hayashi, "Second Order Analysis for Joint Source-Channel Coding With General Channel and Markovian Source," *IEEE Trans. Info. Theory*, Vol. 65, no. 9, pp. 5750-5770, 2019.
[11] L. V. Truong, V. Y. F. Tan, "The Reliability Function of Variable-Length Lossy Joint Source-Channel Coding With Feedback," *IEEE Trans. Info. Theory*, Vol. 65, no. 8, 5028-5042, 2019.
[12] L. Zhou, V. Y. F. Tan, M. Motani, "The Dispersion of Mismatched Joint Source-Channel Coding for Arbitrary Sources and Additive Channels," *IEEE Trans. Info. Theory*, Vol. 65, no. 4, pp. 2234-2251, 2019.
[13] Zhouchen. Lin, M. Chen, Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," https://arxiv.org/abs/1009.5055, 2013.
[14] Y. Yu, X. Bu, K. Yang, Z. Wu, Z. Han, "Green Large-Scale Fog Computing Resource Allocation Using Joint Benders Decomposition, Dinkelbach Algorithm, ADMM, and Branch-and-Bound," *IEEE Internet of Things J.*, Vol. 6, no. 3, pp. 4106-4117, June 2019.
[15] S. Tanaka, K. Tierney, C. Parreno-Torres, R. Alvarez-Valdes, R. Ruiz, "A branch and bound approach for large pre-marshalling problems," *Eu. J. Operational Research*, Vol. 278, no. 1, pp. 211-225, 2019.
[16] J. Niebling and G. Eichfelder, "A Branch-and-Bound-Based Algorithm for Nonconvex Multiobjective Optimization," *SIAM J. Optim.*, Vol. 29, no. 1, pp. 794-821, 2019.

---

[15]Please do not be confused with the principle of $\epsilon-$Nash Equilibria, or $\alpha-$stable [39].

[16]See e.g. [44] to see what joint-typicality-measure is.

[17] A. E. Ezugwu, V. Pillay, D. Hirasen, K. Sivanarain, P. Govender, "A Comparative Study of Meta-Heuristic Optimization Algorithms for $0 - 1$ Knapsack," *IEEE Access*, Vol. 7, pp. 43979-44001, 2019.

[18] J. Yin, H. Jiao, Y. Shang, "Global Algorithm for Generalized Affine Multiplicative Programming Problem," *IEEE Access*, Vol. 7, pp. 162245-162253, 2019.

[19] X. Li, Y. Liu, M. Wang, Z. Song, "GO-APSR: A Globally Optimal Affine Point Set Registration Method," *IEEE Access*, Vol. 7, pp. 137232-137240, 2019.

[20] Y. Shi, X. Mao, C. Zhao, Y. Liu, "A Block Alternating Optimization Method for Direction-of-Arrival Estimation With Nested Array," *IEEE Access*, Vol. 7, pp. 76659-76668, 2019.

[21] R. Jyothi, P. Babu, R. Bah, "Novel Algorithms Based on Majorization Minimization for Nonnegative Matrix Factorization," *IEEE Access*, Vol. 7, pp. 115682-115695, 2019.

[22] Z. D. Yenice, N. Adhikari, Y. K. Wong, V. Aksakalli, A. T. Gumus, B. Abbasi, "SPSA-FSR: Simultaneous Perturbation Stochastic Approximation for Feature Selection and Ranking," https://arxiv.org/abs/1804.05589. 2018.

[23] Y. Feng, X. Chen, G. Gu, "Output Feedback Stabilization for Discrete-Time Systems Under Limited Communication," *IEEE Trans. Automatic Control*, Vol. 62, no. 4, pp. 1927-1932, 2017.

[24] J. Liu, Z. Wu, D. Yue, J. H. Park, "Stabilization of Networked Control Systems With Hybrid-Driven Mechanism and Probabilistic Cyber Attacks," *IEEE Trans Sys. Man. Cyber. Sys.*, Vol. PP, no. 99, pp. 1-11.

[25] J. Liu, V. Gupta, "On Stabilization of Decentralized Systems Across Analog Erasure Links," *IEEE Trans. Automatic Control*, Vol. 62, no. 3, pp. 1411-1416, 2017.

[26] J. Liu, V. Gupta, "Stabilizability Conditions for Linear Time Invariant Systems Across a Gaussian MAC Channel,"*IEEE Trans. Automatic Control*, Vol. 64, no. 6, pp. 2310-2323, 2019.

[27] A. Khina, E. R. Garding, G. M. Pettersson, V. Kostina, and B. Hassibi, "Control over Gaussian Channels With and Without Source-Channel Separation," *IEEE Trans. Automatic Control*, Vol. 64, no. 9, pp. 3690-3705, 2019.

[28] V. Sanjaroon, A. Farhadi, A. S. Motahari and B. H. Khalaj, "Stabilization of Nonlinear Dynamic Systems over Limited Capacity Communication Channels," *IEEE Trans. Automatic Control*, Vol. PP, no. 99, pp. 1-1, 2020.

[29] Z. Hu, P. Shi, J. Zhang, F. Deng, "Control of Discrete-Time Stochastic Systems With Packet Loss by Event-Triggered Approach," *IEEE Trans Sys. Man. Cyber. Sys.* , Vol. PP, no. 99, pp. 1-10, 2020.

[30] Z. Hu, F. Deng, P. Shi, C. Lim, "A New Approach to Characterize Successive Packet Losses in Stochastic Networked Systems," *IEEE Trans Sys. Man. Cyber. Sys.*, Vol. PP, no. 99, pp. 1-1, 2020.

[31] M. Hassanin, and J. Garcia-Frias, "Analog Mappings for Non-Linear Channels with Applications to Underwater Channels," *IEEE Trans. Commun.*, Vol. 68, no. 1, pp. 445-455, 2020.

[32] F. Behnamfar, F. Alajaji, T. Linder, "Image transmission over the polya channel via channel-optimized quantization," *IEEE Trans. Signal Process.* Vol. 53, no. 2, pp. 728-733, 2005.

[33] T.M. Cover, M. Chiang, "Duality Between Channel Capacity and Rate Distortion With Two-Sided State Information," *IEEE Trans. Info. Theory* Vol. 48, no. 6, pp. 1629-1638, 2002.

[34] F. Hekland, G. E. Oien and T. A. Ramstad, "Quantifying Performance Losses in Source-Channel Coding," http://citeseerx.ist.psu.edu, 2007.

[35] D. Gunduz, E. Erkip, A. Goldsmith, H. Vincent Poor, "Transmission of Correlated Sources Over Multi-user Channels," *IEEE 2007 Info. Theory Apps. Workshop*, Jan-Feb 2007, CA, USA.

[36] L. Yu, V. Y. F. Tan, "Asymptotic Coupling and Its Applications in Information Theory," *IEEE Trans. Info. Theory* Vol. 65, no. 3, pp. 1321-1344, 2019.

[37] H. Li and N. Cai, "A Blahut-Arimoto Type Algorithm for Computing Classical-Quantum Channel Capacity," https://arxiv.org/pdf/1904.11188.pdf, 2019.

[38] N. Ramakrishnan, R. Iten, V. B. Scholz, M. Berta, "Non-Commutative Blahut-Arimoto Algorithms," https://arxiv.org/abs/1905.01286, 2019.

[39] N. K. Chada, S. Lasanen, L. Roininen, "Posterior Convergence Analysis of α-Stable Sheets," https://arxiv.org/abs/1907.03086v5, 2019.

[40] P. Yu and D. V. Dimarogonas, "Explicit computation of sampling period in periodic event-triggered multi-agent control under limited data rate," IEEE Transactions on Control of Network Systems," Vol. 6, no. 4, pp. 1366-1378, 2019.

[41] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, "*Linear matrix inequalities in system and control theory*," *SIAM Applied Math.*, QA, USA, 1994.

[42] P. S. Maybeck, "*Stochastic Models, Estimation and Control (I and II)*", Academic Press, New York, 1982.

[43] D. E. Kirk, "*Optimal control theory: an introduction*," Courier Corporation, 2012.

[44] T. M. Cover, J. A. Thomas, "*Elements of Information Theory*," Wiley, USA, 2005.