Decision-Theoretic Compression and Value-Aware Measures of Information and Noise

Tushant Jha, Arjun Pitchanathan and Kannan Srinathan
Center for Security, Theory and Algorithmic Research (C-STAR)
IIIT Hyderabad
Email: {tushant.jha, arjun.p}@research.iiit.ac.in, srinathan@iiit.ac.in

***Abstract*—The objective of this article is to explore the problem of compression under constraints that prohibit lossless communication. This is done by using the decision-theoretic notion of Value of Information (VoI) as the evaluative criteria amongst feasible lossy compressions.**

Particularly, we focus on lossy compression f of a random variable X that maximizes VoI of $(f(X))$, under constraints on the size of message space. We study the connection of variants of this problem to Bayesian persuasion games, and prove approximation results.

We also show the possibility of beneficial noise in strategic persuasion under communication constraint, where noisy welfare can be more than noiseless welfare.

This version of paper does not contain the appendices with proofs, and a full version of this paper is accessible at: https://github.com/particle-mania/dtcompress/blob/master/dtcompress_draft.pdf

I. Motivation

If Alice has access to an *information source*, ie. a discrete random variable X over domain $\{x_1, ...x_n\}$, and wishes to communicate the observed information to Bob, syntactic and statistical measures of information quantify the communicative resources required to send the information losslessly. For instance, communication complexity (in this case, $log_2 n$) quantifies the number of bits required for sending the message. Similarly, Shannon entropy provides a quantification of the asymptotic rate required by a channel between Alice and Bob for successful communication. However, *what should be done under inadequate resources?* Or, *when lossless communication is infeasible, how can we evaluate whether one lossy communication scheme is better than the other?*

It would seem that answering such a question would require a formal model of the semantic content of the *source*, since the notion of *'better'* can be intricately dependent on the context in which communication takes place. However, inspired by economics, it is possible to take a humbler approach to this quantification by recognizing all communicating parties as decision-theoretic agents who are interested in information for making better choices. This allows a more general approach to capture goal-oriented communication, by only assuming that communication is not an end in itself but an instrumental means to some other objective. (also discussed in [1]).

We are therefore interested in optimizing the Value of Information (a decision-theoretic notion of informational value, discussed in Section II-B) for lossy compressions. We study the connections of this problem with the Bayesian Persuasion setting, which has been well studied in economics and game theory literature [2]–[5]. While persuasion games and its' multi-agent generalizations have recently been employed in various strategically relevant settings, like auctions [6], ad placements [7], [8], network routing [9] and other similar contexts [4], [10], they have mostly focussed noiseless channels. Also, many of these problems make certain relevant assumptions regarding the action space of decision or about the nature of private information. However, we deal with the general model [4] that does not make any such assumptions, although we mostly focus on a non-strategic scenario, which provides closer model of *decision-theoretic compression*.

We review the Persuasion game setting in Section II-D with a communication-theoretic language. In Section III-A, we discuss the problem of optimizing VoI, and the associated quantity of value-proportional communication complexity, and show the hardness results and approximation algorithms for the same. We study some more variants of the problem, such as results for optimizing marginal value of information in Section III-B. In Section IV-A, we study the variant of the problem under stochastic noise, and in Section IV-B we use some of these insights to show how noisy channels may not always be *bad* for strategic persuasion. Finally, in Section V, we discuss the technical applications and conceptual implications of these results, and present new problems opened by this line of enquiry in Section VI.

II. Preliminaries

A. Notation for Probabilistic Functions

If X is a (discrete or continuous) random variable that takes value in the set $\mathcal{X}$, then by $dom(X)$ we denote $\mathcal{X}$ and by $spt(X)$ we denote the support of X in $\mathcal{X}$. By $\Delta(\mathcal{X})$, we denote the space of all probability measures over $\mathcal{X}$.

We denote the space of all probabilistic functions from a set $\mathcal{X}$ to a set $\mathcal{Y}$, ie. $(\Delta(\mathcal{Y}))^{\mathcal{X}}$, with $\mathbb{P}(\mathcal{X}, \mathcal{Y})$. For a $f \in \mathbb{P}(\mathcal{X}, \mathcal{Y})$ and a random variable X over $\mathcal{X}$, f can be can be interpreted as a collection of conditional probabilities $p(f(X) = y|X = x)$, and we abuse the notation by denoting with $f(X)$ the corresponding random variable over domain $\mathcal{Y}$. Also, given $f \in \mathbb{P}(\mathcal{X}, \mathcal{Y})$ and $g \in \mathbb{P}(\mathcal{Y}, \mathcal{Z})$, there exists a natural notion of function composition $g \circ f \in \mathbb{P}(\mathcal{X}, \mathcal{Z})$.

Similarly, by $\mathbb{D}(\mathcal{X}, \mathcal{Y})$, we denote all deterministic functions between $\mathcal{X}$ and $\mathcal{Y}$, ie. $\mathcal{Y}^{\mathcal{X}}$. $\mathbb{D}(\mathcal{X}, \mathcal{Y})$ can be interpreted as a subset of $\mathbb{P}(\mathcal{X}, \mathcal{Y})$, where the conditional probabilities are all either 0 or 1.

Given a random variable X with distribution $\mathcal{D}$ over $\mathcal{X}$, we can also recover conditional probabilities of the form $p(x|f(X)=y)$. This is equivalent to the posterior distribution for X, having observed $f(X)$.

B. Value of Information

Value of Information is a decision-theoretic quantification of information that was introduced in [11]. We can define Value of Information (VoI) in terms of an agent, say Bob, trying to make a choice that will maximize expected utility under some uncertainty. Working over a probability space over Ω and a random variable X defined in it, let us take $\mathcal{A}$ as a set of actions that Bob can choose from, and $U : \Omega \times \mathcal{A} \to \mathbb{R}^+$ as the objective function that Bob seeks to maximize.

Then the a priori choice that Bob makes would be $\underset{a \in \mathcal{A}}{argmax}\ \mathbf{E}[U(\omega, a)]$, with expected utility $\underset{a \in \mathcal{A}}{max}\ \mathbf{E}[U(\omega, a)]$.

However if Bob observed $X = x$, then his choice would be $\underset{a \in \mathcal{A}}{argmax}\ \mathbf{E}[U(\omega, a)|X = x]$. Therefore the expected utility for Bob in making the choice posterior to observing value of X is $\underset{X}{\mathbf{E}}\big[\underset{a \in \mathcal{A}}{max}\ \mathbf{E}[U(\omega, a)|X]\big]$. Following [11], the Value of Information for the random varaible X can then be defined as the difference in expected utility gained by observing a signal X, which is always nonnegative.

Definition 1: For a random variable X, and utility function $U : \Omega \times \mathcal{A} \to \mathbb{R}^+$, the value of information is defined as

$$V(X) = \underset{X}{\mathbf{E}}\big[\underset{a \in \mathcal{A}}{max}\ \mathbf{E}[U(\omega, a)|X]\big] - \underset{a \in \mathcal{A}}{max}\ \mathbf{E}[U(\omega, a)]$$

Finally, it is possible to define marginal value of information X wrt Y, as the value of a signal for an agent whose decision is already aware of Y. More formally, for random variables X and Y with joint distribution $\mathcal{D}$, $V(X|Y) = V(X, Y) - V(Y)$.

C. Syntactic Measures of Information

In this article, we are concerned with the communication complexity of transmitting a discrete random variable X, which is given by the Hartley function $C(X) = H_0(X) = log(|dom(X)|)$, which satisfies relevant properties [12], and denotes the number of bits required for this.

D. The Bayesian Persuasion Setup

Bayesian Persuasion is a model of strategic communication and signalling proposed in [2], and has been studied extensively in the economics and game theory literature [4], [5], and captures many important instances of economic communication under assymmetric information [5].

Informally, a simple 2-player persuasion can be understood in terms of a player, say *Bob* who chooses an action under some uncertainty, and the another player, say *Alice*, who has access to some relevant information, such that Bob's action impacts the utilities/welfare or payoffs of Alice and Bob. In this setting, the decision problem for Alice is *how* to reveal her private information to Bob, and *persuade* his action, in a strategically optimal manner. It is assumed that the Alice's strategy for revealing information, called *signalling scheme*, as well as Bob's decision policy with respect to Alice's signals, are both commonly known to Alice and Bob.

Alice's private information can be formally understood as a discrete random variable, over some space of *observations* or *states of the world*, hereafter referred as *state space*. And the utility functions of Alice and Bob, are nonnegative real-valued functions $U : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ for state space $\mathcal{X}$ and action space $\mathcal{A}$. For algorithmic purposes, we assume that this function is given to us as a $|\mathcal{X}| \times |\mathcal{A}|$ matrix of corresponding values. These concepts thereby formalize the informational aspect and the decision-theoretic aspects of the problem.

Definition 2: An *information source* is some random variable X, defined by $(\mathcal{X}, \mathcal{D})$, for *state space* $\mathcal{X}$, with distribution $\mathcal{D} \in \Delta(\mathcal{X})$.

Definition 3: Given an information source over a state space $\mathcal{X}$, a *decision context* is defined as a tuple $(\mathcal{A}, U)$, for *action space* $\mathcal{A}$ and *utility function* $U : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$.

In studying communication schemes in persuasion in this article, we are only concerned with the quantitative aspects of the *channel* over which signalling takes places (such as number of symbols or codewords available), and not necessarily with the concrete implementation of the codewords. Thus, we are only concerned with an abstraction of the mediating channel between Alice and Bob, without the underlying syntactic details. However, we extend the persuasion setting by considering channels with some known stochastic noise. Although, for Section III, we focus only on noiseless channels, where the noise shall be equivalent to a determinisitic identity function.

Definition 4: An abstract *channel* is defined as a 2-tuple $(\mathcal{M}, \eta)$, where $\mathcal{M}$ is the *message space* or the *symbol space*, and $\eta \in \mathbb{P}(\mathcal{M}, \mathcal{M})$ is the noise function.

Given these concepts, we can finally define a persuasion game, as a collection of source available to Alice, the decision contexts of Alice and Bob, and the mediating channel.

Definition 5: A *persuasion game* is defined as a tuple, $\mathcal{G} = (\mathcal{X}, \mathcal{D}, \mathcal{A}, U_A, U_B, \mathcal{M}, \eta)$, for *information source* $(\mathcal{X}, \mathcal{D})$, action space $\mathcal{A}$, Alice's and Bob's utility functions U_A and U_B respectively, and channel $(\mathcal{M}, \eta)$.

Given this problem, observe that the communication scheme can be seen as a combination of two separate components:

1) Alice's behavior $f \in \mathbb{P}(\mathcal{X}, \mathcal{M})$, determining which symbol $m \in \mathcal{M}$ to transmit given $X = x_i$.
2) and, Bob's behavior $g \in \mathbb{P}(\mathcal{M}, \mathcal{A})$ that characterizes which action to choose after receiving a symbol $m \in \mathcal{M}$.

We thus formally define a communication scheme for persuasion, and the associated welfares with it, as follows:

Definition 6: A commmunication scheme for a persuasion game $\mathcal{G}$ is a pair of probabilistic functions (f, g) where:

- encoding function,$f \in \mathbb{P}(\mathcal{X}, \mathcal{M})$,
- and, decoding function, $g \in \mathbb{P}(\mathcal{M}, \mathcal{A})$.

And *Alice's welfare* generated by this scheme is given by

$$\mathcal{W}_A(f, g) = \mathbf{E}[U_A(X, g \circ \eta \circ f(X)))]$$

while *Bob's welfare* can similarly be given by

$$\mathcal{W}_B(f, g) = \mathbf{E}[U_B(X, g \circ \eta \circ f(X)))]$$

A natural notion of a Nash equillibrium for communication schemes follows from this definition. A scheme (f^*, g^*) is at Nash equillibrium if

1) given f^*, $g^* \in \underset{g \in \mathbb{P}(\mathcal{M},\mathcal{A})}{argmax}\ \mathcal{W}_B(f^*, g)$
2) and, given g^*, $f^* \in \underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{argmax}\ \mathcal{W}_A(f, g^*)$

Observe that, given the prior distribution $\mathcal{D}$ of X, any encoding function f uniquely determines a posterior distribution $Pr(X = x_i|m_j)$ for each message symbol $m_j \in \mathcal{M}$. If Bob is a perfect Bayesian agent with utility function U, and knows the encoding function f and noise η, the optimal decoding policy will select the action that maximizes expected utility under the posterior distribution(s). Since, there is always at least one action that maximizes the expected utility under this posterior distribution, for a fixed encoding f there is always a deterministic decoding that optimizes Bob's welfare.

Lemma 1: For a given encoding $f \in \mathbb{P}(\mathcal{X}, \mathcal{M})$, there exists a deterministic decoding $g^* \in \mathbb{D}(\mathcal{M}, \mathcal{A})$, such that $g^* \in \underset{g \in \mathbb{P}(\mathcal{M},\mathcal{A})}{argmax}\ \mathcal{W}_B(f, g)$. In fact, for all $m \in \mathcal{M}$, g^* satisfies:

$$g^*(m) \in \underset{a \in \mathcal{A}}{argmax}\ \mathbf{E}[U_B(x, a))|f(X) = m]$$

It has been shown that no constant-factor approximation for the general persuasion game with noiseless signals exists, unless $P = NP$ [4]. We focus on a special case, of the noiseless setting (i.e. where $\eta : \mathcal{M} \to \mathcal{M}$) where both Alice and Bob have identical utilities (ie. $U_A = U_B$). In this case, Alice's choice for encoding scheme has a similar derandomization argument. (Proof in Appendix A).

Lemma 2: For noiseless persuasion game with identical utilities, ie. $U_A = U_B$, there exists a deterministic encoding $f^* \in \mathbb{D}(\mathcal{X}, \mathcal{M})$, such that there exists a $g^* \in \mathbb{D}(\mathcal{M}, \mathcal{A})$, such that (f^*, g^*) maximizes the welfare $\mathcal{W}_A = \mathcal{W}_B$.

III. Part 1: Value-Aware Lossy Compression under Noiseless Setting

A. Value-Proportional Communication Complexity

Problem 1: For a discrete random variable X over $\mathcal{X} = \{x_1, x_2, ...x_N\}$, and utility $U : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$, given a message space $\mathcal{M} = \{m_1, ...m_K\}$ of K symbols, finding

$$\underset{f:\mathcal{X}\to\mathcal{M}}{argmax}\ V(f(X))$$

In Problem 1, we want a compression scheme for X that preserves as much of it's VoI as possible. This problem also has a dual, with an information-theoretically relevant interpretation, ie. how many bits are required for communicating atleast ϵ fraction of $V(X)$.

Definition 7: For a discrete random variable X over $\mathcal{X}$, and utility function $U : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ for discrete set $\mathcal{A}$, for any $\epsilon \in (0, 1]$, the ϵ-value proportional communication complexity is given by

$$C^\epsilon(X) = \underset{\substack{K=|\mathcal{M}| \\ \underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{max} V(f(X)) \geq \epsilon V(X)}}{argmin} logK$$

Observe that when $|\mathcal{M}| \geq |\mathcal{X}|$, then for a bijective function f, $V(f(X)) = V(X)$. Furthermore, if the utility is an identity matrix, ie. for all i, $U(x_i, a_i) = 1$ and for all $i \neq j$, $U(x_i, a_j) = 0$, then for $\epsilon = 1$, we recover the standard definition of communication complexity of a random variable, ie. $C^1(X) = log|dom(X)|$.

It follows from the way we have defined Value of Information of $f(X)$, that it corresponds to the common welfare in the persuasion game with identical objectives, ie.

$$V(f(X)) = \underset{g \in \mathbb{P}(range(f),\mathcal{A})}{max} \mathcal{W}(f, g)$$

. It can be shown, using Lemmas 1 and 2, that finding the optimal encoder is equivalent to maximizing a submodular function over set of actions. Using hardness and approximation results for submodular maximization [13], [14] we can construct a $1 - 1/e$-approximation algorithm for optimal compression schemes, and algorithm to approximate C^ϵ under $o(log\ log|\mathcal{X}|)$ additive error. (Proofs in Appendix B).

Lemma 3: For a given $|\mathcal{M}| = K$, it is NP-hard to compute the optimal value compression $\underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{argmax}\ V(f(X))$. However, there exists a $(1-1/e)$-approximation algorithm that computes a compression scheme $f^* \in \mathbb{D}(\mathcal{X}, \mathcal{M})$, such that

$$V(f^*(X)) \geq (1 - 1/e) \underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{max} V(f(X))$$

Proposition 1: For a given random variable X, utility U, and ϵ, it is NP-hard to compute $C^\epsilon(X)$. However, there exists an approximation algorithm with additive error in $o(log\ log|\mathcal{X}|)$.

B. Marginal Value of Information

It is possible to consider a slightly general version of Problem 1, by replacing the VoI with marginal VoI with respect to some other discrete random variable. Following the equivalence between value-optimal compression and persuasion with identical utilities from previous section, this leads to a generalization of the basic persuasion model where Alice and Bob are both endowed with some (potentially correlated) private information source, and utility $U : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^+$ is dependent on the joint random variable.

Problem 2: For a discrete random variables X and Y, over $\mathcal{X}$ and $\mathcal{Y}$ respectively, with joint distribution $\mathcal{D}$, and utility $U : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^+$, given a message space $\mathcal{M}$ of K symbols, finding $\underset{f:\mathcal{X}\to\mathcal{M}}{argmax}\ V(f(X)|Y)$

Since Problem 2 is a generalization of Problem 1, the hardness results follow. However, a similar equivalence to submodular optimization fails, since the decoder is no longer a deterministic function over messages, but depends on Bob's private information (ie. Y) as well. However, it is possible to show a correspondence with results similar to Lemma 1 and 2 for this generalization of persuasion games, and use them to provide a O($\sqrt{K}$)-approximation algorithm. The proof involves an equivalence of general persuasion with combinatorial auctions, and is postponed till Appendix C.

Proposition 2: For a given $|\mathcal{M}| = K$, there exists a O($\sqrt{K}$)-approximation algorithm that finding a compression scheme $f^* \in \mathbb{P}(\mathcal{X}, \mathcal{M})$, to maximize $V(f^*(X)|Y)$.

IV. Part 2: Welfare Effects of Exogenous Noise

A. Value-Aware Compression under Noise

It is also possible to study the generalization of Problem 1 with the introduction of stochastic noise, by asking to find an f such that $V(\eta(f(X)))$ is maximized.

Problem 3: For a discrete random variable X, and utility $U : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^+$, given a message space $\mathcal{M}$ of K symbols and stochastic noise over them $\eta \in \mathbb{P}(\mathcal{M}, \mathcal{M})$, finding

$$\underset{f:\mathcal{X}\rightarrow\mathcal{M}}{\textit{argmax}}\ V(\eta(f(X)))$$

As before, observe that the welfare achieved by an encoding in the corresponding noisy persuasion game is equivalent to the value sought to be maximized in Problem 3. It is important to notice that with the introduction of noise, the problem is no longer equivalent to selecting a subset of actions, since some actions could have optimal utility for the states they are assigned to, but may behave very badly when they are selected *'incorrectly'* due to noise.

Example 1: Consider a persuasion game with random variable X with uniform distribution, and the following utility function and noise function.

$$U : \begin{bmatrix} & a_1 & a_2 & a_3 & a_4 \\ x_1 & 100 & 0 & 0 & 0 \\ x_2 & 90 & 100 & 0 & 90 \\ x_3 & 0 & 0 & 100 & 80 \end{bmatrix}$$

$$\eta : \begin{bmatrix} & m_1 & m_2 \\ m_1 & 0.75 & 0.25 \\ m_2 & 0.25 & 0.75 \end{bmatrix}$$

Observe that in the noiseless situation, assigning $\{x_1, x_2\}$ and a_1 to m_1, and $\{x_3\}$ and a_3 to m_2 provides the optimal scheme, and generates welfare $\frac{290}{3}$.

If we borrow the encoding scheme from the above scheme, we observe that it leads to the posteriors, (for Bob), $\{x_1 : \frac{3}{8}, x_2 : \frac{3}{8}, x_3 : \frac{1}{4}\}$ over m_1 and $\{x_1 : \frac{1}{8}, x_2 : \frac{1}{8}, x_3 : \frac{3}{4}\}$ over m_2. Still, the optimal decoder for these posteriors would assign: 1) a_1, that provides $\underset{a}{\textit{max}}\ \mathbb{E}[U(x,a)|\phi(x) = m_1] = \frac{285}{4}$, to m_1; 2) and, a_3, that provides $\underset{a}{\textit{max}}\ \mathbb{E}[U(x,a)|\phi(x) = m_2] = 75$, to m_1. This leads to a welfare, under noise, of $\frac{290}{4}$.

Whereas if we take the encoding scheme to assign $\{x_1\}$ to m_1, and $\{x_2, x_3\}$ to m_2, the posteriors Bob receives are $\{x_1 : \frac{3}{4}, x_2 : \frac{1}{8}, x_3 : \frac{1}{8}\}$ over m_1 and $\{x_1 : \frac{1}{4}, x_2 : \frac{3}{8}, x_3 : \frac{3}{8}\}$ over m_2. This would lead to assigning a_1 to m_1, but a_4 to m_2, and generates welfare, under noise, as 75, which is more than the above scheme taken from the noiseless version.

However, it is still possible, in the case of uniform noise structures η, to construct an equivalent submodular optimization problem. The exact construction of equivalent submodular function and proof are presented in Appendix D.

Definition 8: A stochastic noise function η is *uniform noise* if for some δ, for all $m_i \in \mathcal{M}$: $\eta(m_i|m_i) = 1 - (K-1)\delta$, and $\eta(m_j|m_i) = \delta$ for any $j \neq i$.

Corollary 1: For uniform noise η, there exists a $(1 - 1/e)$-approximation algorithm for finding $f \in \mathbb{D}(\mathcal{X}, \mathcal{M})$ that maximizes $V(\eta(f(X)))$.

B. Can Noise have Value?

The introduction of Noise in the channel can also have surprising welfare consequences in the persuasion games where the objective are not identical. While intuitively, noise would negatively effect welfare by restricting the flow of information, it remains an interesting question to ask whether it can ever have a contrary effect.

Observe, from Definition 6 and Lemma 1, that Bob's optimal policy is governed by the Bayes-optimal response to posteriors generated by $\eta \circ f$, for noise η and Alice's encoding f. Therefore, if $\mathcal{G}^0$ and $\mathcal{G}^\eta$ were persuasion games with same utilities and distribution, but the former was noiseless and latter had some nontrivial noise, then it is easy to observe that $\mathcal{W}^e ta_A(f, g) = \mathcal{W}^0_A(\eta \circ f, g$. This implies that introduction of noise cannot be helpful for Alice, since if the noise was helpful then Alice could simply have simulated the noise in the noiseless setting to generate equivalent posteriors and welfares.

However, Alice's loss due to noise can sometimes lead to Bob's gain. The space of posteriors that can result from encodings in $\{\eta \circ f : f \in \mathbb{P}(\mathcal{X}, \mathcal{M})\} \subseteq \mathbb{P}(\mathcal{X}, \mathcal{M})$ can be much more restricted than those in $\mathbb{P}(\mathcal{X}, \mathcal{M})$. Therefore, it is possible to construct cases where the encoding that generates optimal welfare for Alice in the noisy setting uses a different map from $\mathcal{X}$ to $\mathcal{M}$ that is more valuable for Bob's welfare, than Alice's optimal encoding for noiseless version. An example to demonstrate this is presented in Appendix E.

Proposition 3: There exists a noisy persuasion game $\mathcal{G}$ with uniform noise η, and a noiseless persuasion game $\mathcal{G}'$ with same information source and decision contexts but a noiseless version of $\mathcal{G}$'s channel, such that if (f, g) is a Nash scheme with optimal welfare for Alice in $\mathcal{G}$ and (f', g') for $\mathcal{G}'$, then Bob's welfare in noisy game is more than Bob's welfare in noiseless, ie. $\mathcal{W}_B(f, g) = \mathcal{W}'_B(\eta \circ f, g) > \mathcal{W}'_B(f', g\prime)$.

V. Conceptual Discussion

In this paper, in general, we have brought to attention to the study of communication under syntactic constraints, by using the objective of communication as a yardstick. While attempts at escaping syntactic realm in information theory are often semantic theories [15]–[18], humbler approaches have focussed on the *pragmatic* or *utilitarian* context (or *means-and-end* aspects) of communication (such as recent work in universal communication protocols [1], [19]–[21]). The theoretical approach of this article belongs to the latter trend, focussing on VoI as an abstraction of the *ends*, in the language of expected utlities. The subsequent class of quantitative and algorithmic problems lead to various important insights.

a) Economics of Information Storage: Persuasion under identical utilities offers a model for studying the information-theoretic limit of how Bayesian agents would optimally utilitise memory with limited storage capacity. When Alice and Bob are not necessarily separate agents, but merely time-separated, then persuasion can be interpreted as an agent deciding what information to store, and thereby communicate to its own future *version* for some later decision-making. This understanding lies at the core of the foundations of decision-theoretic compression.

b) Applications to Pragmatic Clustering, or "All reductions lose information, but some retain are useful information": The problem of clustering also deals with finding a mapping from a large cloud of points (ie. $\mathcal{X}$) to some small number of clusters in a meaningful way. Any such mapping would naturally lose some information present in the overall point cloud, but would reduce an intractable amount of data to a tractable number of clusters, thereby simplifying any further data processing. This presents a natural trade-off between the loss of information and number of clusters. While one approach could be to maximize mutual information between the point cloud and clusters, as discussed in [22], decision-theoretic compression offers an alternative approach to minimizing loss of valuable information and dealing with the trade-off whenever the loss function is well-understood. In this sense, decision-theoretic compression offers a foundation for understanding how pragmatic reduction of high-dimension information to small number of classes and clusters.

c) Game-Theoretic Implications of Propostion 3: The receiver/decoder can benefit from channel noise in persuasion with communication constraints, is an observation that leads to interesting insights for information structure design (or "(economic) mechanism design for information" [3]). In the noiseless setting, if Bob (ie. the receiver) chooses to simulate the noise in his behavior, it would lead to a suboptimal decoding policy. However, in cases where noise is valuable, if Bob precommits to a noisy strategy and Alice accounts for Bob's suboptimality, it can lead to an outcome that is overall (in expectation) better for Bob. In such cases, if the channel is chosen prior to emergence of an equillblrium signalling, then the choice itself leads to different economic outcomes.

VI. Conclusion and Open Problems

The space of problems and models discussed in this article further opens up many more relevant mathematical problems. We conclude by discussing three such problems.

a) Optimal Noise Structures: While Proposition 3 discusses the possibility of positive welfare gains due to noise, the algorithmic problem of computing the optimal noise structure for a persuasion game has not been addressed in this article.

b) Value-Proportional Shannon Entropy: Similar to the Definition 7, we can also define ϵ-value proportional Shannon entropy, as the minimum entropy required for a compression to retain atleast ϵ fraction of the VoI of X, ie.

$$H^{\epsilon}(X) = \underset{\substack{f \\ V(f(X)) \geq \epsilon V(X)}}{\mathit{argmin}} \quad H(f(X))$$

Computing entropy-constrained value-optimal lossy compression is a natural extension of problems studied here.

c) Decision-Theoretic Compression of Cooperative Interaction: While in Section III-B we discuss a generalization of persuasion where both Alice and Bob have access to some private information, we only restricted our attention to one-way communication from Alice to Bob. A natural generalization is to consider two-way interactive computation, where Alice and Bob are required to cooperate over an action and the utility ($U : \mathcal{Z} \times \mathcal{A} \rightarrow \mathbb{R}^{+}$) depends on a random variable $Z(X, Y)$ which is a function of their private variables, X and Y respectively. Informally, it is possible to ask, for some K, which protocol Π leads to an interactive computation of $Z'(X, Y)$ that achieves the maximal value of information $V(Z')$, among all protocols that exchange at most K bits.

References

[1] M. Sudan, "Communication amid uncertainty," *2012 IEEE Information Theory Workshop*, pp. 158–161, 2012.

[2] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, October 2011. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590

[3] S. Dughmi, D. Kempe, and R. Qiang, "Persuasion with limited communication," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, ser. EC '16. New York, NY, USA: ACM, 2016, pp. 663–680. [Online]. Available: http://doi.acm.org/10.1145/2940716.2940781

[4] S. Dughmi and H. Xu, "Algorithmic bayesian persuasion," in *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 412–425. [Online]. Available: http://doi.acm.org/10.1145/2897518.2897583

[5] S. Dughmi, "Algorithmic information structure design: a survey," *SIGecom Exchanges*, vol. 15, pp. 2–24, 2017.

[6] Y. Emek, M. Feldman, I. Gamzu, R. P. Leme, and M. Tennenholtz, "Signaling schemes for revenue maximization," *ACM Trans. Economics and Comput.*, vol. 2, pp. 5:1–5:19, 2012.

[7] A. Badanidiyuru, K. Bhawalkar, and H. Xu, "Targeting and signaling in ad auctions," in *SODA*, 2018.

[8] Z. Li, Z. Zheng, F. Wu, and G. Chen, "On designing optimal data purchasing strategies for online ad auctions," in *AAMAS*, 2018.

[9] U. Bhaskar, Y. Cheng, Y. K. Ko, and C. Swamy, "Hardness results for signaling in bayesian zero-sum and network routing games," in *EC*, 2016.

[10] Y. Babichenko and S. Barman, "Algorithmic aspects of private bayesian persuasion," in *ITCS*, 2017.

[11] R. A. Howard, "Information value theory," *IEEE Transactions on Systems Science and Cybernetics*, vol. 2, no. 1, pp. 22–26, Aug 1966.

[12] R. V. L. Hartley, "Transmission of information1," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1928.tb01236.x

[13] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM J. Comput.*, vol. 40, pp. 1740–1766, 2011.

[14] R. K. Iyer and J. A. Bilmes, "Submodular optimization with submodular cover and submodular knapsack constraints," in *NIPS*, 2013.

[15] Y. Bar-Hillel and R. Carnap, "An outline of the theory of semantic information," 1964.

[16] J. Hintikka, "On semantic information," in *In Information and Inference*, 1970, pp. 3–27.

[17] L. Floridi, "Outline of a theory of strongly semantic information," *Minds and Machines*, no. 14, p. 2004.

[18] ——, "Semantic conceptions of information," in *The Stanford Encyclopedia of Philosophy*, spring 2017 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2017.

[19] B. Juba and M. Sudan, "Universal semantic communication ii: A theory of goal-oriented communication," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 15, 2008.

[20] ——, "Efficient semantic communication via compatible beliefs," in *Electronic Colloquium on Computational Complexity*, 2010.

[21] O. Goldreich, B. Juba, and M. Sudan, "A theory of goal-oriented communication," in *Electronic Colloquium on Computational Complexity*, 2009.

[22] F. Cicalese, L. Gargano, and U. Vaccaro, "H(x) vs. h(f (x))," *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 51–55, 2017.

[23] S. Dobzinski, N. Nisan, and M. Schapira, "Approximation algorithms for combinatorial auctions with complement-free bidders," *Math. Oper. Res.*, vol. 35, pp. 1–13, 2005.

[24] U. Feige, "On maximizing welfare when utility functions are subadditive," *SIAM J. Comput.*, vol. 39, pp. 122–142, 2006.

Appendix A
Proof for Lemma 2

Lemma 2: For noiseless persuasion game with identical utilities, i.e. $U_A = U_B$, there exists a deterministic encoding $f^* \in \mathbb{D}(\mathcal{X}, \mathcal{M})$, such that there exists a $g^* \in \mathbb{D}(\mathcal{M}, \mathcal{A})$, such that (f^*, g^*) maximizes the welfare $\mathcal{W}_A = \mathcal{W}_B$.

Proof: Consider any encoder $f \in \mathbb{P}(\mathcal{X}, \mathcal{M})$. Notice that the corresponding optimal decoder, g^* assigns a single action to every message symbol. Let the action associated with m_j be a_j. Suppose there exists some state $x \in \mathcal{X}$ such that the output of f given state x is supported on multiple messages $m_{j_1}, m_{j_2}, ...m_{j_k} \in \mathcal{M}$. Let m_{j_i} be the message whose corresponding action a_{j_i} maximizes the utility $U_A(x, a_{j_i})$.

Consider a new encoder f' which is the same as f, except that when the state is x, f' deterministically outputs m_{j_i}. Note that the welfare of (f', g^*) is at least as much as that of (f, g^*). Since it is possible to derandomize an encoding function without decreasing its welfare, there must exist an optimal encoding function which is deterministic. ■

Appendix B
Proof for Lemma 3 and Proposition 1

Lemma 3: For a given $|\mathcal{M}| = K$, it is NP-hard to compute the optimal value compression $\underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{argmax}\ V(f(X))$. However, there exists a $(1-1/e)$-approximation algorithm that computes a compression scheme $f^* \in \mathbb{D}(\mathcal{X}, \mathcal{M})$, such that

$$V(f^*(X)) \geq (1 - 1/e) \max_{f \in \mathbb{P}(\mathcal{X},\mathcal{M})} V(f(X))$$

Proof: The problem of finding optimal deterministic scheme, using Lemma 1 and 2, can be viewed as the problem of selecting K actions $\mathcal{A}' \subseteq \mathcal{A}$, and such that the utility generated by assigning states $x \in \mathcal{X}$ to their maximal action in $\mathcal{A}'$ is maximized.

More formally, define a set function q over $\mathcal{A}$, such that for any $S \subseteq \mathcal{A}$, $q(S) = \sum_{x \in \mathcal{X}} \max_{a \in S} U(x, a)$. Observe, that this is a monotone submodular set function over $\mathcal{A}$, and the problem of finding optimal compression schemes is equivalent to optimizing this set function $q(S)$ under the constraint $|S| \leq K$. The hardness result and approximation algorithm follows from results in submodular optimization [13], where it has been shown that $1 - 1/e$ approximation is the best approximation possible. ■

Proposition 1: For a given random variable X, utility U, and ϵ, it is NP-hard to compute $C^\epsilon(X)$. However, there exists an approximation algorithm with additive error in $o(log\ log|\mathcal{X}|)$.

Proof: The problem of computing $\underset{\substack{K=|\mathcal{M}| \\ \max_{f \in \mathbb{P}(\mathcal{X},\mathcal{M})} V(f(X)) \geq \epsilon V(X)}}{argmin} K$ is dual of the submodular optimization problem that corresponds to problem 1. Therefore, a $1 - 1/e$-approximate algorithm for problem 1 can be used to construct a $logN$-approximate algorithm for computing minimal K for achieving ϵ fraction of VoI, using duality correspondence discussed in [14]. This translates to an approximation algorithm for computing C^ϵ with additive error in $o(loglogN)$ ■

Appendix C
Proof for Proposition 2

We first present an alternative proof for Lemma 3, using reduction to Combinatorial Auctions with XOS Utilities, and then build upon that reduction to prove Proposition 2.

Lemma 3: For a given $|\mathcal{M}| = K$, it is NP-hard to compute the optimal value compression $\underset{f \in \mathbb{P}(\mathcal{X},\mathcal{M})}{argmax}\ V(f(X))$. However, there exists a $(1-1/e)$-approximation algorithm that computes a compression scheme $f^* \in \mathbb{D}(\mathcal{X}, \mathcal{M})$, such that

$$V(f^*(X)) \geq (1 - 1/e) \max_{f \in \mathbb{P}(\mathcal{X},\mathcal{M})} V(f(X))$$

Proof: The problem of finding the optimal deterministic encoding scheme (f, g) is equivalent to finding a partition of $\mathcal{X}$ into sets $S_1, S_2, ...S_{\mathbf{m}}$, where $S_j = f^{-1}(m_j)$, in a way that maximizes

$$\sum_{m \in \mathcal{M}} \sum_{x \in f^{-1}(m)} p(x)U(x, g(m)) = \sum_{m \in \mathcal{M}} \max_{a \in \mathcal{A}} \sum_{x \in f^{-1}(m)} p(x)U(x, a)$$

. (Where $p(x)$ denotes probability of x).

Now let v be a set function over $\mathcal{X}$, such that for a $S \subseteq \mathcal{X}$, $v(S) = \max_{a \in \mathcal{A}} \sum_{x \in S} p(x)U(x, a)$. Observe that if the partition corresponding to symbol m_j is S_j, then $v(S_j)$ denotes the contribution of that symbol to the overall welfare of the scheme. It is therefore possible to view the problem of finding optimal encoding as the problem of distributing items in $\mathcal{X}$ to $\mathbf{m}$ players with individual player utilities over their bundles given by v (common for all players).

In combinatorial allocation and welfare maximization literature, such set functions which can be represented as maximum over some additive set functions (or linear sums of weights) are called XOS utilities or fractionally subadditive. And it has been shown, in [23] and [24], that given these weighted sum expressions, it is possible to find allocations with $\frac{e-1}{e}$ approximation guarantee in polynomial time. ■

Proposition 2: For a given $|\mathcal{M}| = K$, there exists a O$(\sqrt{K})$-approximation algorithm that finding a compression scheme $f^* \in \mathbb{P}(\mathcal{X}, \mathcal{M})$, to maximize $V(f^*(X)|Y)$.

Proof: Maximizing welfare is equivalent to finding the f that maximizes

$$\mathbb{E}[U(X, Y, g(f(X), Y))] = \sum_{m \in \mathcal{M}} \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} p(x, y)U(x, y, a)$$

Observe that this equivalent to finding an allocation of items in $\mathcal{X}$ among K players with valuations over bundles $\sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} p(x, y)U(x, y, a)$, which is a sum of XOS or fractionally subadditive valuations.

However, while the class of fractionally subadditive valuations is closed under addition, there is no straightforward way to simulate demand queries for valuations expressed as sum of such valuations.

Nevertheless, since value queries can simulated, ie. evaluating $v(S)$ given any set S, we can use algorithms for welfare maximizing combinatorial allocation with value oracles that provide a $O(\sqrt{K})$-approximation guarantee. [23]

■

APPENDIX D
PROOF FOR COROLLARY 1

Corollary 1: For uniform noise η, there exists a $(1-1/e)$-approximation algorithm for finding $f \in \mathbb{D}(\mathcal{X}, \mathcal{M})$ that maximizes $V(\eta(f(X))$.

Proof: Let us borrow the definition of monotone submodular function $q : 2^{\mathcal{A}} \to \mathbf{R}^{+}$, that denotes the maximum welfare of a scheme that decodes to a subset S of actions. Now, if define $q', q^* : 2^{\mathcal{A}} \to \mathbf{R}^{+}$, as $q'(S) = \sum\limits_{a \in S} \sum\limits_{x \in \mathcal{X}} p(x)U(x,a)$ and $q^*(S) = \delta q'(S) + (1-(K-2)\delta)q(S)$, we can see that q^* is also a monotone submodular function.

Also, $q^*(S)$ for $S \subseteq \mathcal{A}$, is equal to welfare under noise of a scheme that assigns actions in S to message symbols. This allows us to use existing approximation methods for submodular maximization under cardinality constraint of $\mathbf{m}$ from [13]. ■

APPENDIX E
EXAMPLE TO ILLUSTRATE PROPOSTION 3

Proposition 3: There exists a noisy persuasion game $\mathcal{G}$ with uniform noise η, and a noiseless persuasion game $\mathcal{G}'$ with same information source and decision contexts but a noiseless version of $\mathcal{G}$'s channel, such that if (f, g) is a Nash scheme with optimal welfare for Alice in $\mathcal{G}$ and (f', g') for $\mathcal{G}'$, then Bob's welfare in noisy game is more than Bob's welfare in noiseless, ie. $\mathcal{W}_B(f,g) = \mathcal{W}'_B(\eta \circ f, g) > \mathcal{W}'_B(f', g\prime)$.

Proof: Let $\mathcal{G}$ be a persuasion game with utility and noise functions as described below, and message set $\mathcal{M} = \{m_1, m_2\}$.

$$U_A : \begin{bmatrix} & a_1 & a_2 & a_3 & a_4 \\ x_1 & 100 & 0 & 99 & 90 \\ x_2 & 100 & 0 & 90 & 99 \\ x_3 & 0 & 100 & 90 & 99 \end{bmatrix}$$

,

$$U_B : \begin{bmatrix} & a_1 & a_2 & a_3 & a_4 \\ x_1 & 100 & 80 & 150 & 0 \\ x_2 & 100 & 80 & 0 & 95 \\ x_3 & 80 & 100 & 0 & 90 \end{bmatrix}$$

$$\eta : \begin{bmatrix} & m_1 & m_2 \\ m_1 & 0.98 & 0.02 \\ m_2 & 0.02 & 0.98 \end{bmatrix}$$

Let $\mathcal{G}'$ be the noiseless version of $\mathcal{G}$. Let f and f' be encoders that have outputs as summarized in the figure below.

	x_1	x_2	x_3
f	m_1	m_2	m_2
f'	m_1	m_1	m_2

Intuitively, the other possible encoder, which outputs m_1 for $\{x_1, x_3\}$ and m_2 for x_2, will perform badly for Alice in both the noisy and noiseless cases, and indeed this can be verified to be true. Let us explore what happens when we use the encoder f in the noiseless case. If the message is m_1, Bob knows that the state was x_1. If the message is m_2, x_2 and x_3 both have a posterior probability of half. With this, we can calculate the expected utility (for Bob) for each action given the message received. This is summarized in the table below.

	a_1	a_2	a_3	a_4
m_1	100	80	150	0
m_2	90	90	0	92.5

Thus the optimal decoder for Bob assigns a_3 to m_1 and a_4 to m_2. This gives a welfare of $335/3$ for Bob and 99 for Alice. In the noisy case, the posteriors become $\{x_1 : 0.98/1.02, x_2 : 0.02/1.02, x_3 : 0.02/1.02\}$ over m_1 and $\{x_1 : 0.02/1.98, x_2 : 0.98/1.98, x_3 : 0.98/1.98\}$ over m_2. The corresponding expected utilities for each action are summarized below. In what follows, we give the values as decimals rounded to two places rather than fractions, as this is more illuminating.

	a_1	a_2	a_3	a_4
m_1	99.61	80.39	144.12	3.63
m_2	90.10	89.90	1.52	91.57

Thus the optimal decoder for Bob in the noisy case is the same as in the noiseless case. The corresponding welfare is 109.43 for Bob and 98.82 for Alice. Similarly we can compute the welfare for Alice in the noisy and noiseless cases for both choices of encoder, as summarized in the table below.

	$\mathcal{G}$	$\mathcal{G}'$
f	98.82	99
f'	98	100

We can see that the optimal encoder for Alice is f in the noisy case and f' in the noiseless case. By performing a similar computation to that above we see that the welfare for Bob in $\mathcal{G}'$ when the encoder is f' is 100, which is lower than his welfare of 109.43 in $\mathcal{G}$.

■