Persuasion with limited communication capacity*

Maël Le Treust[†] and Tristan Tomala[‡]

December 15, 2017

Abstract

We consider a Bayesian persuasion problem where the persuader and the decision maker communicate through an imperfect channel which has a fixed and limited number of messages and is subject to exogenous noise. Imperfect communication entails a loss of payoff for the persuader. We establish an upper bound on the payoffs the persuader can secure by communicating through the channel. We also show that the bound is tight: if the persuasion problem consists of a large number of independent copies of the same base problem, then the persuader can achieve this bound arbitrarily closely by using strategies which tie all the problems together. We characterize this optimal payoff as a function of the information-theoretic capacity of the communication channel.

JEL Classification Numbers: C72, D82, D83.

*The authors thank James Best, Olivier Gossner, Frédéric Koessler, Marie Laclau, Daniel Martin, Ludovic Renou, Thomas Rivera, Jakub Steiner, Colin Stewart for stimulating discussions and comments. We also thank participants of: the 6th workshop on Stochastic Methods in Game Theory, Erice May 2017; the 13th European Meeting on Game Theory (SING13), Paris July 2017; the XXVI Colloque Gretsi, Juan-Les-Pins, September 2017; the 10th Transatlantic Theory Workshop, Paris September 2017; the 55th Allerton Conference, Monticello, Illinois, October 2017. We thank the Institute Henri Poincaré for hosting numerous research meetings.

†ETIS UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, F-95000, Cergy, France; mael.le-treust@ensea.fr; sites.google.com/site/maelletreust/. This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

‡HEC Paris and GREGHEC, 1 rue de la Libération, 78351 Jouy-en-Josas, France; tomala@hec.fr; studies2.hec.fr/jahia/Jahia/tomala. Tristan Tomala gratefully acknowledges the support the HEC foundation and ANR/Investissements d'Avenir under grant ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047.

1 Introduction

In modern internet societies, pieces of information are repeatedly and continuously disclosed by informed agents to decision makers. Information transmission is affected by at least two sources of friction. First, the sender and the receiver of a given message may have non-aligned incentives, in which case the sender might be unwilling to transmit truthful information. Second, communication between agents is often imperfect. The sender and the receiver may have time constraints to write or read messages, so that the sender has to summarize his arguments and cannot convey all the details. Further, there might be discrepancies between the informational content of a message that is intended by the sender and the one understood by the receiver, for instance if the mother tongue's of the sender and of the receiver are different, there are possible translation errors (see Blume, Board, and Kawamura, 2007). Also, messages travelling in a network of computers might be subject to random shocks, internal errors or protocol failures. Studying the effect of noise in communication channels is the starting point of information theory (Shannon, 1948).

How does imperfect communication reduce the possibilities of persuasion in a sender-receiver interaction and how the sender should optimally disclose information when facing limits on its capacity of communication?

In this paper, we consider a sender and a receiver who communicate over an imperfect channel and are engaged in a series of $n \geq 1$ persuasion problems. The sender observes n independent and identically distributed pieces of information and sends $k \geq 1$ messages to the receiver. Messages are sent through a channel which consists of two finite sets X, Y of respectively inputs and outputs messages and of a transition probability Q from X to Y: when the sender chooses input message x, the receiver receives output message y with probability $Q(y|x)$. Upon receiving k output messages from the channel, the receiver chooses n actions, one for each problem. Payoffs are additively separable across persuasion problems. We assume that the sender is able to commit to a disclosure strategy which maps sequences of pieces of information to sequences of input messages.

We analyze the optimal average payoff secured by the sender by committing to a strategy. We give an upper bound on this optimal payoff and show that this bound is approximately achieved when the numbers n and k grow large with fixed ratio n/k. We call this payoff the value of *the optimal splitting problem with information constraint*, which represents the best payoff that the sender can achieve by sending a message, subject to the constraint that the mutual information between the state and the message is no more than the capacity of the channel. We show that this value is given by the concave hull of the payoff function of the sender, subject to a constraint on the entropy of posterior beliefs. This is also the concave hull of a modified payoff function, where the sender pays a cost proportional to the mutual information between the state and the message.

Motivating example. Consider an innovating firm who has several projects to be financed by investors. The board of investors audits the firm who is given a limited amount of time to present all the projects. How to best organize the arguments in order to get a maximum number of projects approved?

To be specific, let's assume that all projects are ex-ante identical and equally likely to be of good or bad quality. When a project is approved, it yields a positive return of $+1$ to the investors if it is good, and a negative return of -7 if it is bad; rejecting a project yields a payoff of 0. The objective of the firm is to get a maximum number of projects approved. This example has the same structure as the main example of Kamenica and Gentzkow (2011), revisited by Bergemann and Morris (2017).

Suppose that the firm is able to commit to an information disclosure policy *à la* Kamenica and Gentzkow (2011) and has unlimited time to present arguments. In order to invest, the board of investors needs to be persuaded that the project is good with probability at least 7/8. Thus for each project, the firm would optimally draw a good message g or a bad message b with the following probabilities:

$$\mathbb{P}(g \mid \text{project is good}) = 1, \quad \mathbb{P}(g \mid \text{project is bad}) = 1/7.$$

This way, the belief that the project is good upon receiving the good message is:

$$\mathbb{P}(\text{project is good} \mid g) = 7/8,$$

and the project is accepted with probability 4/7 (details are in Section 3).

Now, suppose that the auditing board gives the firm only half the time it would require to talk about all projects. Namely, there is an even number n of projects, but the firm has to present $n/2$ messages. What is the optimal information disclosure policy of the firm?

A simple strategy the firm can adopt is to select half of the projects, focus on them, and communicate optimally for each of them. With this strategy, half of the projects are accepted with probability 4/7 each, so in expectation the average number of accepted projects is 2/7. This is not optimal, a better strategy would be to pair projects by two and to draw one message g, b for each pair in the following way.

$$\mathbb{P}(g \mid \text{both projects are good}) = 1, \quad \mathbb{P}(g \mid \text{both projects are bad}) = 0,$$

$$\mathbb{P}(g \mid \text{only one project is good}) = 1/6.$$

The total probability of g is 1/3 and upon observing this message, the beliefs about quality are:

$$\mathbb{P}(\text{both projects are good} \mid g) = 6/8,$$

$$\mathbb{P}(\text{only project 1 is good} \mid g) = \mathbb{P}(\text{only project 2 is good} \mid g) = 1/8.$$

Therefore each project is believed to be good with probability 7/8 and both projects are accepted when g is received. Thus, the expected average number of accepted projects is $1/3 > 2/7$. We show in Section 3 that this is the optimal way of pairing projects two by two. Is it possible to find a more complex strategy which improves the payoff further?

Our main result Theorem 4.3 gives an upper bound on the expected average number of accepted projects, when the number of messages is half the number of projects. The upper bound is tight: the optimal value approaches it as the number of project increases.

On this example, the upper bound is λ^* where (λ^*, p^*) is the unique solution in $[0,1]\times[0,\frac{1}{2}]$ of the system of equations,

$$\frac{1}{2} = \lambda^* \frac{7}{8} + (1-\lambda^*)p^*, \quad \frac{1}{2} = \lambda^* H\left(\frac{7}{8}\right) + (1-\lambda^*)H(p^*),$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$ is the entropy function. The first equation is Bayes plausibility (Kamenica and Gentzkow, 2011) coming from Bayes' rule, saying that the expected posterior belief is the prior belief. The second equation requires that the expected entropy of the posterior is $\frac{1}{2}$, which means that the *mutual information* between the quality of the project and the message sent to the receiver is equal to the number of messages per project that the firm can transmit.

Numerically $\lambda^* \approx 0.519 < \frac{4}{7} \approx 0.571$. So for large n, the sender can achieve a payoff better than $1/3$ but bounded away from the payoff obtained with perfect communication.

Related literature. We now describe the relationships between our contribution and the literature. This paper is at the junction of Bayesian persuasion and information theory.

The traditional game theoretic approach to strategic information disclosure assumes perfect communication and analyzes in isolation the problem of sending a single message. These are the well-known sender-receiver games where an informed player, the sender, communicates once with a receiver who takes an action. In the *cheap talk* version of this game, the message sent by the sender is costless and unverifiable, see for instance the seminal paper of Crawford and Sobel (1982). In the *Bayesian persuasion* game (Kamenica and Gentzkow, 2011), the sender chooses verifiably an information disclosure device, prior to learning his information. This model can be interpreted in several ways: (i) the sender has full commitment power and displays publicly the mechanism which links states and messages, (ii) the sender is not informed of the state parameter but is able to choose a statistical experiment whose distribution depends on the state, (iii) the sender is an *information designer* (Bergemann and Morris, 2016, 2017; Taneva, 2016) who chooses the information or signalling structure which will release information to the decision maker.

In parallel, information theory considers agents with perfectly aligned interests and analyzes the *rate* of information transmission over time. The sender observes an information *flow*, that is a stochastic process, and sends messages to the receiver over an imperfect channel represented by a transition probability from input to output messages. Truthful information transmission is the common goal of the sender and the receiver. The *rate of information transmission* is the average number of correct guesses made by the receiver over time. Shannon's theory (Shannon, 1948, 1959) determines whether a source of information can be transmitted over the channel with arbitrarily small probability of error, and shows that the rate of the source of information has to be smaller than *the capacity of the channel* defined as the maximal *mutual information* between input and output messages.

Our model of persuasion has two essential features. The sender and the receiver are engaged in a large number of identical copies of the same game and communication is restricted to an imperfect channel. As Kamenica and Gentzkow (2011), we consider the payoff obtained by the sender as a function of the belief of the receiver, when the receiver takes an optimal action given his belief.

With unrestricted communication, that is on a perfect channel with large set of inputs, the optimal payoff for the sender is given by the concave hull of this function. Then, solving any number of identical games amounts to solving each copy separately.

With a single copy, the game of persuasion with a noisy channel is studied by Tsakas and Tsakas (2017) who prove the existence of optimal solutions and show monotonicity of the sender's payoff with respect to the noise of the channel. We give a detailed example in section 3.2.

Considering many copies of the base game *and* restricted communication, we show that *linking independent problems together* yields a better payoff to the sender: the optimal strategy correlates all messages with the state parameters of all problems. In this respect, our work bears some similarity with Jackson and Sonnenschein (2007), who showed that a mechanism designer could achieve more outcomes in an incentive compatible manner by linking many identical problems together.

The optimal payoff that we characterize is related to models where the cost of information is measured by mutual information. Such measurements of information costs have been introduced in the literature on rational inattention by Sims (2003), (see e.g. Martin, 2017, for a strategic context). Matejka and McKay (2015) and Steiner, Stewart, and Matejka (2017) consider models where mutual information is either a direct cost paid by the agent, or where the agent is constrained to extract less information than some capacity. In the context of persuasion, Gentzkow and Kamenica (2014) consider a model where the sender gets his payoff from the game, minus a cost which is proportional to the mutual information between the state and the message. With Lagrangian methods, we find that the value of our optimal splitting problem with information constraint is the concave hull of the payoff function, net of such an information cost. Differently from those papers, the mutual information is not a primitive of our model. Our finding is that the noise and limitations in communication induce a *shadow cost* measured by the mutual information.

Entropy and mutual information appear endogenously in several papers on repeated games (Gossner and Tomala, 2006, 2007; Gossner and Vieille, 2002; Neyman and Okada, 1999, 2000). A related paper is Gossner, Hernández, and Neyman (2006), henceforth GHN, who also consider a sender-receiver game. In GHN, the sender and the receiver play an infinitely repeated game with common interests: both the sender and the receiver want to choose the action that matches the state. The sender knows the infinite sequence of states and can communicate with the receiver only through his actions. GHN characterize the best average payoff that the sender (and the receiver) can achieve. Their solution resembles ours: the optimal value is the payoff obtained when the sender can send a direct message to the receiver, subject to an information constraint.

There are important differences with our work. First, GHN study a cheap talk game with common interests. By contrast, we do not assume common interests and we assume commitment power for the sender. Second, GHN is truly a repeated game model: at any given time t both players choose actions and the information of the receiver at this time consists of past actions. In our case, the sender knows a finite sequence of states, chooses a finite sequence of input messages, the receiver observes the finite sequence of output messages and chooses a sequence of actions. This is why, rather than seeing our model as a repeated game of persuasion, we view it as a *spatial* model with identical copies of the same problem co-existing at the same time. This also explains why the number of copies n need not be equal to the number of messages k that the sender is able to input into the channel. Our result characterizes the optimal payoff as a function of the ratio of the number n of pieces of information to the number k of messages. In particular, this allows to analyze cases where the channel is perfect (*i.e.* not subject to random noise) but with limited input size: there are fewer messages than states or actions.

Cheap talk with a noisy channel has been studied by Blume, Board, and Kawamura (2007) who show that the presence of noise is possibly welfare improving. Such a phenomenon cannot happen in the persuasion context as the sender could commit to replicate the noise. Relatedly, Hernández and von Stengel (2014) consider a sender-receiver game with common interests over an imperfect channel. In that paper, there is only one state known by the sender and one action taken by the receiver, while the channel can be used a fixed number of times. Hernández and von Stengel (2014) characterize all the Nash equilibria of this game and study the differences with Shannon's coding methods. Again, we do not assume common interests and assume commitment power for the sender. More importantly, our focus is different and more in line with GHN: we do not treat a single persuasion problem but a large sequence of them and use information theory to study the asymptotics of the problem.

Our work is also related to some information theoretic literature. Following GHN, a line of papers study empirical coordination between a sender and a receiver (Cuff, Permuter, and Cover, 2010), by communicating over a perfect (Cuff and Zhao, 2011) or imperfect channel (Le Treust, 2017). Those papers implicitly assume common interest between the sender and the receiver and characterize the empirical distributions of (states, messages, actions) which are achievable, given the information structure and the noisy channel. These characterizations are related to the information theoretical problems of source coding (Wyner and Ziv, 1976) and channel coding (Gelfand and Pinsker, 1980) with partial information about the state (Le Treust and Bloch, 2016), whose solution is still not available for some simple cases. In a recent paper, Akyol, Langbort, and Başar (2017) have considered the problem of Bayesian persuasion for Gaussian state and channel. The authors calculate explicitly the optimal strategies for the quadratic cost functions considered by Crawford and Sobel (1982) and prove that they are linear.

The closest paper in this literature is Le Treust and Tomala (2016) where we have studied the empirical coordination between a persuader and a decision maker. In this proceeding, we have characterized the limit set of empirical distributions of states, messages and actions induced by approximate equilibria of the game with n copies and n messages, as n tends to infinity. There are several new contributions in the current paper. First, considering a large number of identical copies of a base persuasion game, we compare precisely the solution of the large game with the solution of the base game, with and without noisy channel. Second, rather than looking at approximate equilibria, we characterize the best payoff the sender can secure, given that the receiver chooses actions which are optimal for his Bayesian belief on the sequence of states. Third, to achieve this characterization, we introduce the optimal splitting problem under information constraint. The detailed study of this problem allows us to construct a strategy of the sender such that the best response of the receiver induces the target payoff. Fourth, the concavification under information constraint is easy to interpret and offers a nice interpretation of the mutual information as an information cost.

The paper is organized as follows. The model is described in section 2. In section 3, we provide benchmarks by studying examples where we calculate the solution of the single persuasion game with and without a noisy channel. In section 4, we consider large copies of identical problems, state our main result, and revisit the examples. In section 5, we study the concavification under information constraint. Section 6 discusses the cardinality of messages sets. Proofs are in the Appendix.

2 Model

2.1 Bayesian persuasion with restricted communication

We consider a Bayesian persuasion problem between two players, a sender (S) and a receiver (R). There is a finite state space Ω with a common prior $\mu \in \Delta(\Omega)$ and a finite set of actions A. Player $i = S, R$ cares about the state ω and the action taken a and has payoff $u_i(\omega, a)$.

In the *persuasion game with unrestricted communication*, the sender chooses a signaling structure, consisting of a finite set of messages M and a transition probability $\sigma : \Omega \to \Delta(M)$. Once chosen, the signaling structure is known to the receiver. Then, a state ω is drawn with probability $\mu(\omega)$, a message $m \in M$ is drawn with probability $\sigma(m|\omega)$, the message is observed by the receiver. The receiver then chooses an action $a \in A$.

In concrete settings, communication possibilities may be restricted, for instance messages may be subject to exogenous noise, or the number of possible messages may be smaller than the number of states or actions. We represent an imperfect communication channel by a transition probability $Q : X \to \Delta(Y)$, where X, Y are fixed finite sets of messages (words, letters or abstract symbols) and $\Delta(Y)$ is the set of probability distributions over Y. The set X represents the possible messages that the sender can input into the channel, the set Y is the set of messages that the receiver can possibly receive. When the sender chooses message x, message y is received with probability $Q(y|x)$.

Example 2.1. *Binary symmetric channel.* As an example, take binary sets of messages $X = \{x_0, x_1\}$, $Y = \{y_0, y_1\}$ and assume that the channel has a *noise level* $\varepsilon \in [0, \frac{1}{2}]$, that is $Q(y_j|x_i) = \varepsilon$ for $j \neq i$, see Figure 1. The generic case is $\varepsilon \in (0, \frac{1}{2})$ where the label of the message (0 or 1) is changed with positive probability, but observing a label 1 is still more likely when the input label is 1. When $\varepsilon = \frac{1}{2}$, the distribution of the output message is independent from the input message, so the channel completely disrupts the communication.

A special case is the *binary perfect channel* when $\varepsilon = 0$: identifying together the sets

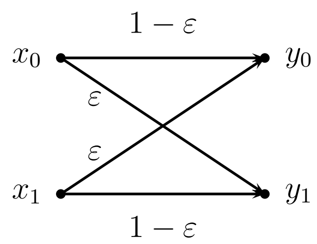


Figure 1: Binary symmetric channel.

X and Y, an input message x is received with certainty. Communication is then restricted only by the number of available messages, *i.e.* the cardinality of X.

In the *persuasion game with communication restricted by the channel*, the set of messages is fixed to be X and the sender can only choose a transition probability $\sigma : \Omega \to \Delta(X)$ which we will refer to as the strategy of the sender. Once chosen, it is known by the receiver. Then, in state ω, an input message x is drawn with probability $\sigma(x|\omega)$, an output message y is drawn from the channel with probability $Q(y|x)$ and announced to the receiver, who chooses an action a.

Our general objective is to characterize the best payoff that the sender can secure *in a robust way.* That is, what is the best payoff the sender can secure, for all optimal strategies of the receiver.

A strategy of the receiver is a mapping $\tau : Y \to A$. Knowing σ, the receiver chooses a best-reply τ which maximizes the expected payoff. That is, for each[1] y,

$$\tau(y) \in \arg\max_{a\in A} \sum\nolimits_{\omega,x} \mu(\omega)\sigma(x|\omega)Q(y|x)u_R(\omega, a).$$

Denote $BR(\sigma)$ the set of best replies of the receiver to the strategy σ.

Definition 2.2. *The* optimal robust payoff *of the sender is,*

$$U_S^*(\mu, Q) = \sup_{\sigma} \min_{\tau\in BR(\sigma)} \sum\nolimits_{\omega,x,y} \mu(\omega)\sigma(x|\omega)Q(y|x)u_S(\omega, \tau(y)).$$

This is the best payoff that the sender can achieve, provided that the receiver takes

[1]Since once chosen, σ is fixed and known, requiring optimality for all y's, or for all y's in the support, does not affect the solutions.

any optimal strategy. In case the receiver is indifferent between several actions, we want this quantity to be robust to the exact specification of the optimal action. Thus, we assume that if there are several optimal strategies, the receiver chooses the one which is the least preferred by the sender.

Note that this quantity depends on the prior and on the communication channel.

2.2 Linking independent problems

We consider now persuasion problems composed of a large number of independent identical copies of the same base problem. Communication is still restricted by the channel and by the number of times it can be used.

Precisely, the state space is now Ω^n for some positive integer n, so that a state is a sequence $\omega^n = (\omega_1, \ldots, \omega_n)$. We assume that the (ω_t)'s are independently and identically distributed, so that the prior probability μ^n on Ω^n is given by $\mu^n(\omega^n) = \prod_{t=1}^n \mu(\omega_t)$. The receiver chooses a sequence of actions $a^n = (a_1, \ldots, a_n)$ and the payoff for player $i = S, R$ is,

$$\bar{u}_i(\omega^n, a^n) = \frac{1}{n} \sum\nolimits_{t=1}^{n} u_i(\omega_t, a_t).$$

The communication resource available to the sender is the repeated use of the channel which is assumed to be memoryless. Precisely, the sender can choose a sequence of k messages $x^k = (x_1, \ldots, x_k)$ to input into the channel, and the receiver will observe $y^k = (y_1, \ldots, y_k)$ with probability $Q^k(y^k|x^k) = \prod_{t=1}^k Q(y_t|x_t)$.

A strategy of the sender is now a mapping $\sigma : \Omega^n \to \Delta(X^k)$, which is known by the receiver, once chosen. A strategy of the receiver is $\tau : Y^k \to A^n$. The optimal robust payoff of the sender in this problem is denoted:

$$U_S^*(\mu^n, Q^k) = \sup_{\sigma} \min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, x^k, y^k} \mu^n(\omega^n)\sigma(x^k|\omega^n)Q^k(y^k|x^k)\bar{u}_S(\omega^n, \tau(y^k)).$$

Our main goal is to provide a characterization of the optimal robust payoff for large problems, that is when n and k grow.

3 Benchmarks and examples

Before stating our main results, we recall as a benchmark what happens with unrestricted communication and examine the case of a single problem on a communication channel.

3.1 Persuasion with unrestricted communication

Take a simple persuasion problem with state space Ω, prior μ, action set A, payoffs u_i, assume that the receiver can choose messages in an arbitrarily large finite set M, and that messages are perfectly observed by the receiver.

The solution to this game is well-known (Kamenica and Gentzkow, 2011). Given a strategy $\sigma : \Omega \to \Delta(M)$, message m is received with total probability, $\mathbb{P}_\sigma(m) = \sum_\omega \mu(\omega)\sigma(m|\omega)$ and the posterior belief $\nu_\sigma(\cdot|m)$ upon receiving message m is given by $\nu_\sigma(\omega|m) = \frac{\mu(\omega)\sigma(m|\omega)}{\mathbb{P}_\sigma(m)}$ (for all messages m such that $\mathbb{P}_\sigma(m) > 0$). Bayes' rule dictates that $\mu = \sum_m \mathbb{P}_\sigma(m)\nu_\sigma(\cdot|m)$.

From the splitting lemma (Aumann and Maschler, 1995) or Bayes plausibility (Kamenica and Gentzkow, 2011), each decomposition of the prior into a convex combination of posteriors $\mu = \sum_m \lambda_m \nu_m$ is induced by the following strategy $\sigma(m|\omega) = \lambda_m \nu_m(\omega)/\mu(\omega)$.

Such a convex combination will be henceforth referred to as *a splitting of* μ. There is a one-to-one correspondance between the strategies of the sender and the splittings of the prior. From now on, we will use letter μ to denote the prior belief and letter ν to denote a generic belief or posterior of the receiver. With a slight abuse of notation, we identify the convex combination $\mu = \sum_m \lambda_m \nu_m$ with the *distribution of beliefs* where the receiver has belief ν_m with probability λ_m.

The optimal robust payoff of the sender is then easily found by the concavification method. For each belief $\nu \in \Delta(\Omega)$, denote $A^*(\nu)$ *the set of optimal actions* for the receiver with belief ν,

$$A^*(\nu) = \arg\max_{a \in A} \sum\nolimits_\omega \nu(\omega) u_R(\omega, a).$$

Then, τ is optimal given σ when for each m, the action $\tau(m)$ belongs to $A^*(\nu_\sigma(\cdot|m))$. Call the *robust payoff* $U_S(\nu)$ of the sender at the belief ν, the payoff he gets when the receiver chooses the optimal action which is *worst* for S. Denote,

$$U_S(\nu) = \min_{a \in A^*(\nu)} \sum\nolimits_{\omega} \nu(\omega) u_S(\omega, a).$$

With the same logic as in Kamenica and Gentzkow (2011), the optimal robust payoff is the concavification of U_S at μ,

$$\operatorname{cav} U_S(\mu) = \sup \left\{ \sum\nolimits_m \lambda_m U_S(\nu_m) : \sum\nolimits_m \lambda_m \nu_m = \mu \right\},$$

where the supremum is over the set of splittings of the prior: the numbers λ_m are non-negative summing up to 1 and $\nu_m \in \Delta(\Omega)$ for each m.

Observe that, contrary to Kamenica and Gentzkow (2011), we assume that in case of indifference, the receiver breaks ties in the worst way for the sender. This choice is motivated by robustness since any optimal action is legitimate for the receiver. Although for generic problems the choice of tie-breaking rule does not change the concavification function, in the above formula the supremum might not be reached exactly, but approximated arbitrarily closely, see Example 3.1.

Example 3.1. *Persuading to invest.* This example will be running throughout the paper and revisited in various contexts. The sender is a firm who persuades the receiver to invest in a risky project. If the receiver does not invest (action a_0), the payoff is 0 for both players. If the receiver invests (action a_1), the projects has return -7 in the bad state ω_0 and $+1$ in the good state ω_1. Both states are equally likely. The sender receives a fee of $+1$ only if the receiver invests. The payoff table is as follows, the entries are pairs of payoffs for the players $i = S, R$ depending on the state and action.

	a_0	a_1	μ
ω_0	$(0, 0)$	$(1, -7)$	$\frac{1}{2}$
ω_1	$(0, 0)$	$(1, 1)$	$\frac{1}{2}$

The receiver invests for sure only when he holds a belief ν such that $\nu(\omega_1) > 7/8$. If $\nu(\omega_1) = 7/8$ he is indifferent. Assuming that in case of indifference he does not invest, the robust payoff of the sender $U_S(\nu)$ is 1 if $\nu(\omega_1) > 7/8$ and 0 otherwise.

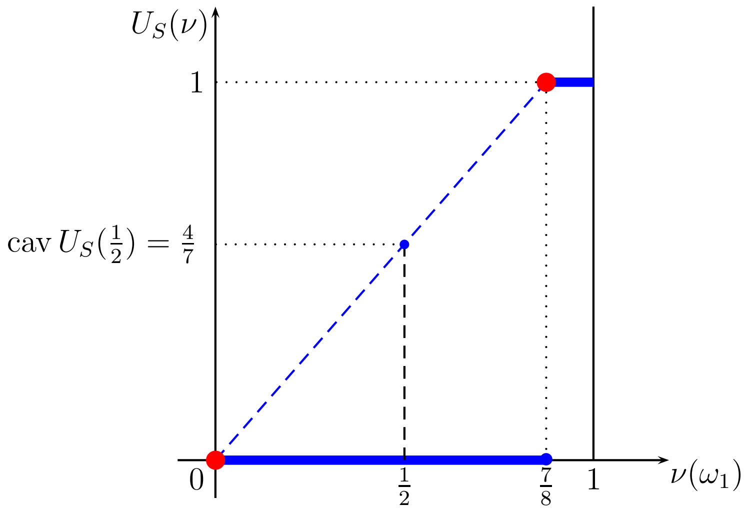


Figure 2: Concavification.

The concavification function $\text{cav}\, U_S(\nu)$ is continuous and equal to $\frac{8}{7}\nu(\omega_1)$ for $\nu(\omega_1) \leq \frac{7}{8}$ and 1 otherwise. It is easy to see that it does not depend on the action chosen by the receiver at $\nu(\omega_1) = \frac{7}{8}$, see Figure 2.

If the receiver would choose a_1 at the point of indifference, then the optimal splitting for the sender would be,

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{3}{7}(1,0) + \frac{4}{7}\left(\frac{1}{8}, \frac{7}{8}\right),$$

where a belief is denoted $\nu = (\nu(\omega_0), \nu(\omega_1))$. This yields a payoff of $\frac{4}{7}$ which is the highest that the sender can achieve given the uniform prior. For any small $\varepsilon > 0$, let's perturb the previous splitting a little bit to get,

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{3+8\varepsilon}{7+8\varepsilon}(1,0) + \frac{4}{7+8\varepsilon}\left(\frac{1}{8} - \varepsilon, \frac{7}{8} + \varepsilon\right),$$

which achieves the payoff $\frac{4}{7+8\varepsilon}$ irrespective of the tie-breaking rule. Letting ε tend 0, we see that the sender achieves a payoff arbitrarily close to $\frac{4}{7}$. This is indeed the optimal robust payoff.

Note that the tie-breaking rule might be relevant in non-generic cases. To see this, let's imagine that we change the payoffs of the receiver in order to push the indifference point from $\frac{7}{8}$ up to 1. In that case, the sender gets a payoff only when $\nu(\omega_1) = 1$ *and* the receiver chooses the sender-preferred action. In such a case, the concavification depends on the choice of the sender at the indifference point.

All examples in the paper will be generic, so that to calculate the concavification, it is without loss to assume that the receiver chooses the action preferred by the sender at indifference points.

Remark 3.2. *Copies of independent problems with unrestricted communication.* With unrestricted communication, the optimal robust payoff does not change if we take identical copies of the same persuasion problem. Indeed, the receiver treats each copy as a separate problem and takes an optimal action. Therefore, the sender cannot achieve more than $\operatorname{cav} U_S(\mu)$ for each copy, and he will thus also handle the problems separately.

3.2 Persuasion over the channel for a single problem

We consider again a simple persuasion problem but now, the set of messages for the sender is X, the set of messages for the receiver is Y and messages are filtered by the channel $Q : X \to \Delta(Y)$.

Also in this context, any strategy $\sigma : \Omega \to \Delta(X)$ translates into a splitting of the prior into posteriors which writes,

$$\mu = \sum\nolimits_{y \in Y} \lambda_y \nu_y,$$

where λ_y is the total probability of y and ν_y is the posterior belief, conditional on y. Obviously, the number of different posteriors is at most the cardinality of Y. Such a splitting is feasible if and only if there exists $\sigma : \Omega \to \Delta(X)$ such that,

$$\lambda_y = \sum\nolimits_{\omega,x} \mu(\omega)\sigma(x|\omega)Q(y|x)$$

and

$$\nu_y(\omega) = \mu(\omega) \sum\nolimits_{x} \sigma(x|\omega)Q(y|x)/\lambda_y.$$

The channel imposes severe restrictions on the set of feasible splittings which is studied in Tsakas and Tsakas (2017). Consider the following example.

Example 3.3. *Binary symmetric channel.* Consider the binary symmetric channel described in Example 2.1. Let a strategy σ be parametrized by $\sigma(x_0|\omega_0) = 1-\alpha$ and $\sigma(x_1|\omega_1) = 1-\beta$, see Figure 3.

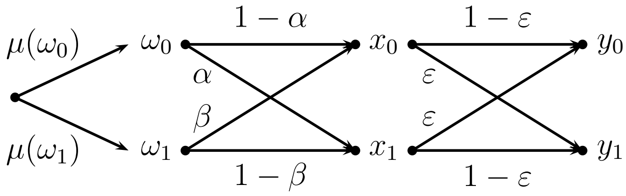


Figure 3: Strategy on the binary symmetric channel.

Then,

$$\mathbb{P}_\sigma(y_1|\omega_0) = \alpha(1-\varepsilon) + (1-\alpha)\varepsilon := \alpha \star \varepsilon,$$

$$\mathbb{P}_\sigma(y_0|\omega_1) = \beta(1-\varepsilon) + (1-\beta)\varepsilon := \beta \star \varepsilon.$$

It follows that $\mathbb{P}_\sigma(y_1) = \mu(\omega_0)\alpha \star \varepsilon + \mu(\omega_1)(1-\beta \star \varepsilon)$ and from Bayes' rule,

$$\mathbb{P}_\sigma(\omega_1|y_1) = \frac{\mu(\omega_1)(1-\beta\star\varepsilon)}{\mu(\omega_0)\alpha\star\varepsilon + \mu(\omega_1)(1-\beta\star\varepsilon)},$$

$$\mathbb{P}_\sigma(\omega_1|y_0) = \frac{\mu(\omega_1)\beta\star\varepsilon}{\mu(\omega_0)(1-\alpha\star\varepsilon) + \mu(\omega_1)\beta\star\varepsilon}.$$

It is easy to see that since $\alpha\star\varepsilon \in [\varepsilon, 1-\varepsilon]$, all the numbers $(\mathbb{P}_\sigma(y_1|\omega_0), \mathbb{P}_\sigma(y_0|\omega_1), \mathbb{P}_\sigma(y_0), \mathbb{P}_\sigma(\omega_1|y_1), \mathbb{P}_\sigma(\omega_1|y_0))$ belong to the interval $[\varepsilon, 1-\varepsilon]$. Let's characterize the feasible splittings.

A pair of posteriors (ν_0, ν_1) is *feasible* if there exists a number $\lambda \in [0,1]$ such that,

$$(\mu(\omega_0), \mu(\omega_1)) = \lambda(\nu_0(\omega_0), \nu_0(\omega_1)) + (1-\lambda)(\nu_1(\omega_0), \nu_1(\omega_1)).$$

Lemma 3.4. *A pair of posteriors* (ν_0, ν_1) *is feasible if and only if* $\nu_1 = \nu_0 = \mu$ *or,*

$$\varepsilon \leq \frac{\nu_0(\omega_1)(\nu_1(\omega_1) - \mu(\omega_1))}{\mu(\omega_1)(\nu_1(\omega_1) - \nu_0(\omega_1))} \leq 1 - \varepsilon$$

and

$$\varepsilon \leq \frac{(1 - \nu_0(\omega_1))(\mu(\omega_1) - \nu_0(\omega_1))}{(1 - \mu(\omega_1))(\nu_1(\omega_1) - \nu_0(\omega_1))} \leq 1 - \varepsilon.$$

The proof is in the Appendix (A.1.1). As an illustration, take the uniform prior $(\frac{1}{2}, \frac{1}{2})$ and a level of noise $\varepsilon = \frac{1}{4}$. The feasible posteriors are shown by the colored regions (green) on Figure 4.

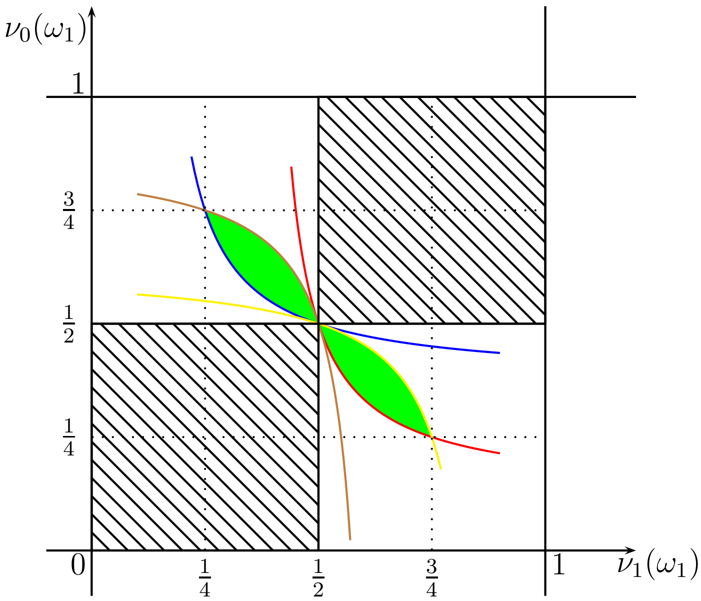


Figure 4: Feasible posteriors.

Example 3.5. *Persuading to invest over a noisy channel.* Consider the persuasion problem given in Example 3.1 and assume that communication is filtered through the binary symmetric channel with noise $\varepsilon = \frac{1}{4}$ studied in Example 3.3. From the previous discussion, it is impossible to induce beliefs with $\nu(\omega_1) > \frac{3}{4}$. Therefore, the receiver will never be confident enough to invest and the payoff is 0 for the sender.

This example demonstrates how exogenous noise in the communication limits the persuasion possibilities. Now, we consider the motivating example given in the introduction where the channel is noiseless and contains few messages, that is less than the number of states or the number of actions.

Example 3.6. *Persuading to invest with few words.* Consider two independent copies of the persuasion problem given in Example 3.1. The state space is $\{\omega_0, \omega_1\} \times \{\omega_0, \omega_1\}$, with uniform prior. The receiver has to choose two actions, one for each problem, so that the action set is $\{a_0, a_1\} \times \{a_0, a_1\}$. The payoff for each player is the average of payoffs in the two problems. With perfect communication, the sender can achieve $\frac{4}{7}$ in each problem, so $\frac{4}{7}$ on average.

Now, suppose that the channel is perfect but has only two messages $|X| = |Y| = 2$. The sender is able to send a perfect message but only from a binary set, whereas there are four states and four actions. How much can he achieve?

Achieving an average payoff of $\frac{2}{7}$ is easy. The sender focuses on the first state and communicates optimally about it, revealing nothing about the second state. This yields a payoff of $\frac{4}{7}$ for the first problem, and 0 for the second one.

Claim 3.7. *The optimal robust payoff is $\frac{1}{3}$ for this example. It is achieved by the splitting*

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = \frac{2}{3}\left(\frac{6}{16}, \frac{5}{16}, \frac{5}{16}, 0\right) + \frac{1}{3}\left(0, \frac{1}{8}, \frac{1}{8}, \frac{6}{8}\right),$$

which corresponds to the following strategy,

$$\sigma(x_1|\omega_0, \omega_0) = 0, \sigma(x_1|\omega_0, \omega_1) = \sigma(x_1|\omega_1, \omega_0) = \frac{1}{6}, \sigma(x_1|\omega_1, \omega_1) = 1.$$

The intuition is the following. Since there are only two messages, any strategy induces two posteriors. Bayes' plausibility (or the splitting constraint) implies that one posterior must lie in the region where the receiver does not invest at all. So either the sender persuades the receiver to invest for only one of the two problems, or to invest for both of them. We show that it is optimal to persuade to invest for both problems. If the state is either the worst one (ω_0, ω_0) or the best one (ω_1, ω_1), it is fully disclosed. The strategy is the same in the two intermediary states (ω_0, ω_1) and (ω_1, ω_0) and both messages are sent with positive probability. The proof is in the Appendix (A.1.2).

The insight gained from this example is that the sender is better off by linking the two problems together, that is, the distribution of the message depends on both states. The advantage of linking problems together grows with the number of copies as our main result shows in the next section.

4 Main results

In this section we state our main result which is a characterization of optimal robust payoffs for large number of copies of the same problem. First, we introduce tools borrowed from information theory.

4.1 Mutual information and channel capacity

We start by recalling useful notions from information theory, the reader is referred to Cover and Thomas (2006). Let $\boldsymbol{x}$ be a random variable with values in some finite set with distribution p. The (Shannon) entropy of $\boldsymbol{x}$ is,

$$H(\boldsymbol{x}) = -\mathbb{E} \log p(\boldsymbol{x}) = -\sum\nolimits_x p(x) \log p(x),$$

where the logarithm has basis 2 and $0 \log 0 = 0$. Since this depends only on p, this is also denoted $H(p)$. Let $(\boldsymbol{x}, \boldsymbol{y})$ be a pair of finite random variables with distribution $\mathbb{P}(x, y)$. The conditional entropy of $\boldsymbol{y}$ given $\boldsymbol{x}$ is,

$$H(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}} H(\boldsymbol{y}|\boldsymbol{x} = x) = -\sum\nolimits_x \mathbb{P}(x) \sum\nolimits_y \mathbb{P}(y|x) \log \mathbb{P}(y|x).$$

The mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ is,

$$I(\boldsymbol{x}; \boldsymbol{y}) = H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) = H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y}).$$

Take a communication channel $Q : X \to \Delta(Y)$. If $\boldsymbol{y}$ is obtained from inputing a random variable $\boldsymbol{x}$ with distribution p into the channel, then the pair $(\boldsymbol{x}, \boldsymbol{y})$ has joint distribution $\mathbb{P}(x,y) = p(x)Q(y|x)$. The mutual information $I(\boldsymbol{x};\boldsymbol{y})$ depends only on this joint distribution and is thus a function of p and Q.

Definition 4.1. *The capacity of the channel $Q : X \to \Delta(Y)$ is*

$$C(Q) = \max_{p\in\Delta(X)} I(\boldsymbol{x};\boldsymbol{y}),$$

where the maximum is over the marginal distribution p of $\boldsymbol{x}$.

For instance, take $X = Y$ and assume that the channel is perfect so that $H(\boldsymbol{y}|\boldsymbol{x}) = 0$. The entropy of $\boldsymbol{x}$ is maximal and equal to $\log|X|$ when $\boldsymbol{x}$ is uniformly distributed. The capacity of the perfect channel is thus $\log|X|$. Intuitively, the capacity of the channel is the maximal number of bits of information that the channel can transmit. A perfect binary channel can transmit 1 bit of information. If $|X| = 2^m$, the channel can transmit m bits of information.

An another example, consider the binary symmetric channel with noise ε. Then the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is $(\varepsilon, 1-\varepsilon)$ or its permutation. Again, $H(\boldsymbol{x})$ is maximal when $\boldsymbol{x}$ is uniformly distributed. The capacity of the noisy binary channel is thus $C = 1 - H(\varepsilon)$, where with a slight abuse of notation, $H(\varepsilon)$ stands for the entropy of the binary distribution $(\varepsilon, 1-\varepsilon)$.

4.2 Splitting with information constraint

Consider a base persuasion problem with state space Ω, prior μ, action sets A and payoffs u_i, $i = S, R$.

A splitting of $\mu = \sum_m \lambda_m \mu_m$ can be seen as a joint distribution $\mathbb{P}$ of a random pair $(\boldsymbol{\omega}, \boldsymbol{m})$ in $\Omega \times M$ such that, the marginal distribution of $\boldsymbol{\omega}$ is $\mathbb{P}(\boldsymbol{\omega} = \omega) = \mu(\omega)$, the marginal distribution of $\boldsymbol{m}$ is $\mathbb{P}(\boldsymbol{m} = m) = \lambda_m$ and the conditional distribution of $\boldsymbol{\omega}$ given $\boldsymbol{m} = m$ is $\mathbb{P}(\boldsymbol{\omega} = \omega|\boldsymbol{m} = m) = \mu_m(\omega)$.

The *mutual information of the splitting* is the mutual information between $\boldsymbol{\omega}$ and $\boldsymbol{m}$:

$$I(\boldsymbol{\omega};\boldsymbol{m}) = H(\boldsymbol{\omega}) - H(\boldsymbol{\omega}|\boldsymbol{m}) = H(\mu) - \sum\nolimits_m \lambda_m H(\mu_m).$$

Let us consider an auxiliary optimisation problem where the sender has access only to the splittings whose mutual information is at most some given positive number C.

Definition 4.2. *For any $C \geq 0$, the optimal splitting problem with information constraint is:*

$$V(\mu, C) = \sup_{\boldsymbol{\omega},\boldsymbol{m}} \left\{ \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{m}} U_S : I(\boldsymbol{\omega}; \boldsymbol{m}) \leq C \right\} = \begin{array}{rl} \sup & \sum_m \lambda_m U_S(\mu_m) \\ \text{s.t.} & \sum_m \lambda_m \mu_m = \mu, \\ \text{and} & H(\mu) - \sum_m \lambda_m H(\mu_m) \leq C. \end{array}$$

If we interpret the mutual information as the cost of the signaling structure (Gentzkow and Kamenica, 2014; Sims, 2003), the value of this optimisation problem is the optimal payoff the sender can get with a signaling structure whose cost does not exceed the capacity C.

The mutual information constraint can be re-ordered as $\sum_m \lambda_m H(\mu_m) \geq H(\mu) - C$ which says that the expected entropy of the posteriors cannot be two low. That is, posteriors cannot be too precise, the precision being limited both by the entropy of the prior and the available capacity. Observe that if $H(\mu) \leq C$, the constraint is satisfied by all splittings. The value of the problem is thus the concavification of U_S in this case.

4.3 The characterization

We are now ready to state the main result of the paper. We are considering n identical copies of the persuasion problem with communication k times through the channel and recall that $U_S^*(\mu^n, Q^k)$ denotes the optimal robust payoff of the sender.

Theorem 4.3. *1. The optimal robust payoff of the sender is no more than the value of the optimal splitting with information constraint. For all k, n,*

$$U_S^*(\mu^n, Q^k) \leq V(\mu, \frac{k}{n} C(Q)).$$

2. The optimal robust payoff of the sender converges to the optimal splitting with infor-

mation constraint as n, k tend to infinity. For any rational number r and all $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that for all (k, n) such that $k = rn$ and $n \geq N(\varepsilon)$,

$$U_S^*(\mu^n, Q^{rn}) \geq V(\mu, rC(Q)) - \varepsilon.$$

To get some intuition, assume $n = k$ for the time being. The result says that $U_S^*(\mu_n, Q_n) \leq V(\mu, C(Q))$ for all n, and $U_S^*(\mu_n, Q_n) \to V(\mu, C(Q))$ as n tends to infinity.

The intuition is as follows. The optimal splitting problem with information constraint represents the best payoff the sender can achieve by sending a message whose mutual information with the state is no more than the capacity of the channel. The first clause of the theorem states that this is an upper bound on payoffs that the sender can reach by communicating over the channel. The proof of this necessary condition is simple. Indeed, the mutual information between the sequence of states and the sequence of messages to the receiver cannot exceed the capacity of the channel. Therefore, the upper bound derives naturally from properties of mutual information.

The second clause of the theorem states that the value of the optimal splitting problem with information constraint, can be obtained approximately for large problems. When n is large, the intuition that the capacity of the channel is the amount of information that the channel can transmit per unit of time, can be made concrete by appropriate use of laws of large numbers. More precisely, Shannon's coding theory says the following. Suppose that the mutual information between the random state $\boldsymbol{\omega}$ and a random message $\boldsymbol{m}$ is no more than the capacity $I(\boldsymbol{\omega}; \boldsymbol{m}) \leq C$. It is then possible for the sender to associate with the sequence of states $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$, a sequence of *intended messages* $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$ and a sequence of *actual input messages* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, such that upon receiving the *actual output messages* $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, the receiver is able to recover most intended messages with high probability. The coding scheme of our proof uses both of the original Shannon's source and channel coding schemes defined in Shannon (1948, 1959), the reader is also referred to the textbooks Cover and Thomas (2006, Chapters 7 and 10) and Gamal and Kim (2011, Chapter 3).

To complete the proof, we show that it is indeed optimal for the receiver to find out most of the intended messages from the messages actually received. Intuitively, it is optimal for the receiver to extract as much information as possible from the messages and thus to decode correctly the messages intended by the sender. We prove that the actual Bayesian beliefs of the receiver about the sequence of states $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$ are close to those induced by the intended messages $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$. The technical proof is deferred to the Appendix (A.3).

Now, there is no reason why the number n of pieces of information should be equal to the number k of times that the channel can be used. The result says that only the ratio $\frac{k}{n}$ matters. Indeed, when $k = rn$, then it is (asymptotically) equivalent to take $k = n$ and to multiply the capacity by r.

To be concrete, assume that $k = 2n$. This means intuitively that the channel can be used two times for each piece of information, so that the capacity is doubled. Alternatively, assuming $2k = n$ means that the channel can only be used once for every pair of problems. As in instance, consider $2k = n$ copies of Example 3.6 where the number of messages is half the number of states.

When the ratio $r = \frac{k}{n}$ is large, $rC \geq H(\mu)$ and the entropy constraint is automatically satisfied. Intuitively, if the channel could be used many times for each problem, the sender would be able to convey any message he wants.

4.4 Examples

Example 4.4. *Persuading to invest over a noisy channel.* Let us revisit Example 3.1 given by the following table.

	a_0	a_1	μ
ω_0	$(0, 0)$	$(1, -7)$	$\frac{1}{2}$
ω_1	$(0, 0)$	$(1, 1)$	$\frac{1}{2}$

Consider a large number n of independent copies with communication n times over a binary channel with noise $\varepsilon = \frac{1}{4}$. Recall that in the single problem, the receiver cannot be persuaded to invest and the payoff is 0.

Let us compute the optimal value of splitting with information constraint. The capacity of the channel is $1 - H(\frac{1}{4})$, the entropy of the uniform prior is 1, therefore the information constraint is $\sum_m \lambda_m H(\mu_m) \geq H(\frac{1}{4})$. Figure 5 shows the set of pairs of posteriors for the splittings which satisfy this constraint (green and blue region).

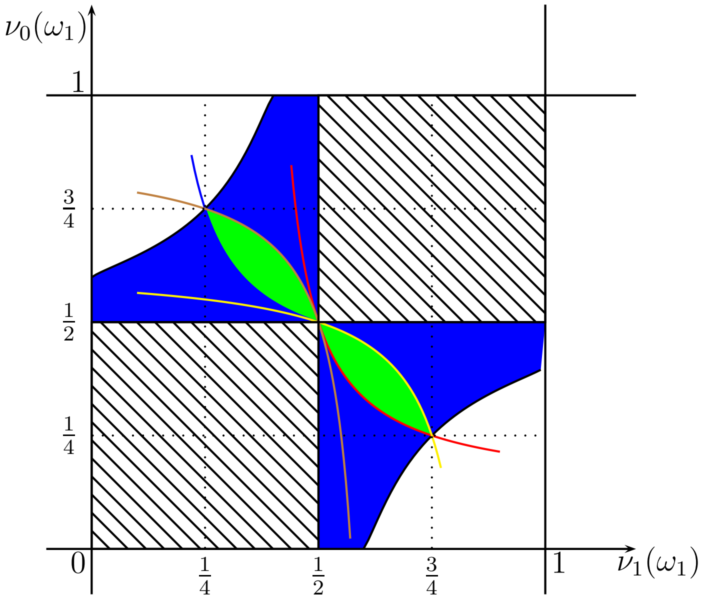


Figure 5: Feasible posteriors under information constraint.

Under this constraint the optimal splitting for the sender satisfies:

$$\left(\frac{1}{2},\frac{1}{2}\right) = \lambda\left(\frac{1}{8},\frac{7}{8}\right) + (1-\lambda)(\nu(\omega_0),\nu(\omega_1))$$

and

$$H\left(\frac{1}{4}\right) = \lambda H\left(\frac{7}{8}\right) + (1-\lambda)H(\nu(\omega_1)).$$

To see why it is optimal, first consider that the sender has to bring on some posterior ν with $\nu(\omega_1) > \frac{7}{8}$ in order to get some payoff. To get it with the highest probability, he should aim for $\nu(\omega_1) = \frac{7}{8}$. Among the posteriors that induce investment, this is also the one with highest entropy. Second, to maximize expected payoffs, the remaining posteriors must be as far away as possible from the prior, that is, the entropy constraint should bind. Also, note that only one posterior will be generated in the region $\nu(\omega_1) < \frac{7}{8}$. Since the entropy is strictly concave, replacing two posteriors on this region by their average does not change the payoff and increases the entropy.

Solving these two equations numerically we get, $\nu(\omega_1) \approx 0.340$ and $V(\mu, Q) = \lambda \approx 0.298$ which is about 52.1% of the unconstrained optimum $\frac{4}{7}$, see Figure 6.

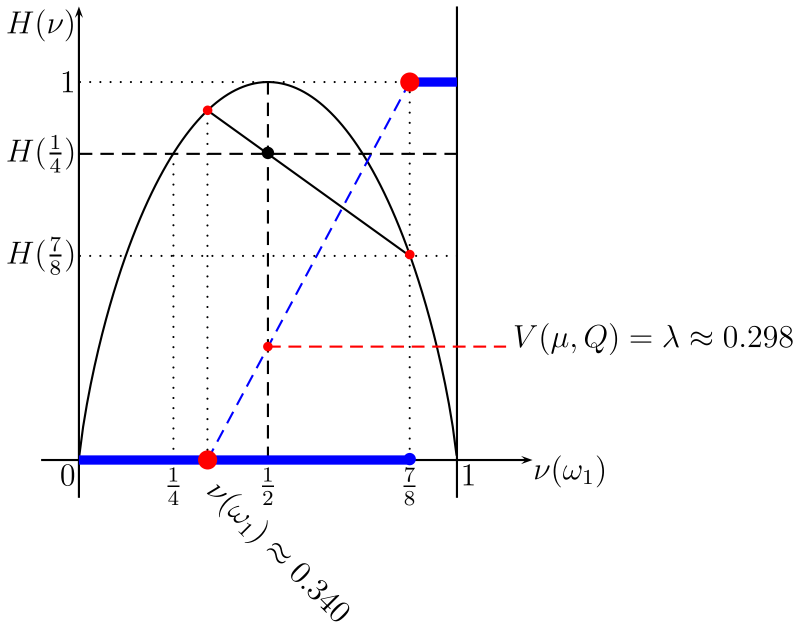


Figure 6: Optimal splitting with information constraint

Example 4.5. *Persuading to invest with few words.* Consider $n = 2k$ copies of the previous example where a perfect binary channel is used k times. That is, the number of messages is half the number of states. This can be seen as k copies of the problem with 4 states given in Example 3.6. The capacity of the binary perfect channel is 1, but since the channel is used half of the times, it is like the capacity is $\frac{1}{2}$. So from Theorem 4.3, we want to calculate the best payoff under the information constraint:

$$\sum_m \lambda_m H(\mu_m) \geq H\left(\frac{1}{2}\right) - \frac{1}{2} \cdot 1 = \frac{1}{2}.$$

Remark that this is the same constraint one would obtain (with $k = n$) on a noisy binary symmetric channel with ε such that $H(\varepsilon) = \frac{1}{2}$ so $\varepsilon \approx 0.11$. Under this constraint the optimal splitting for the sender satisfies:

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \lambda\left(\frac{1}{8}, \frac{7}{8}\right) + (1-\lambda)(\nu(\omega_0), \nu(\omega_1))$$

and

$$\frac{1}{2} = \lambda H\left(\frac{7}{8}\right) + (1-\lambda)H(\nu(\omega_1)).$$

Solving these equations numerically gives $\nu(\omega_1) \approx 0.095$ and

$$V(\mu, C(Q)) = \lambda = \frac{\frac{1}{2} - \nu(\omega_1)}{\frac{7}{8} - \nu(\omega_1)} \approx 0.519,$$

see Figure 7. This is about 90.8% of the unconstrained optimum $\frac{4}{7} \approx 0.571$.

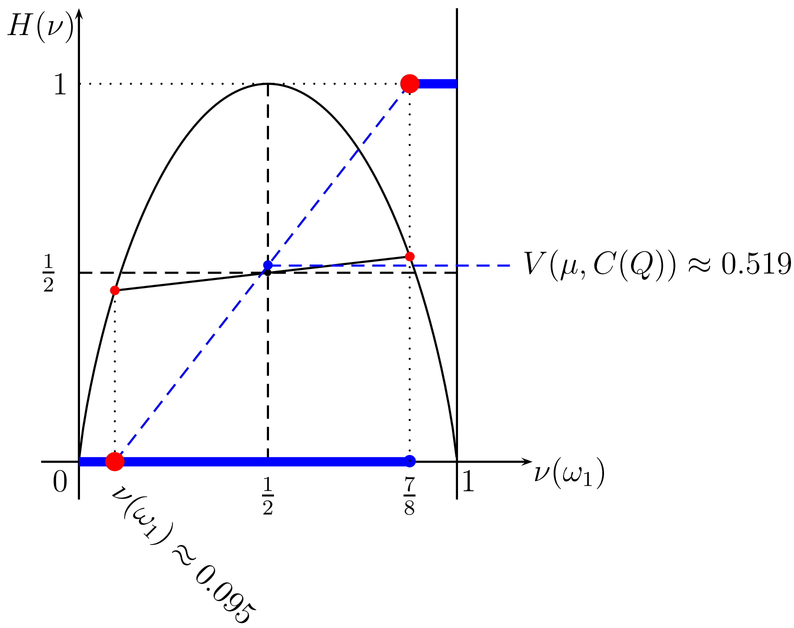


Figure 7: Optimal splitting with small perfect channel

5 Concavification with information constraint

In this section, we analyze the problem of maximizing the sender's payoff under the information constraint:

$$V(\mu, C) = \sup\Big\{\sum\nolimits_m \lambda_m U_S(\mu_m) : \sum\nolimits_m \lambda_m \mu_m = \mu, H(\mu) - \sum\nolimits_m \lambda_m H(\mu_m) \leq C\Big\}.$$

There are two ways to relate this problem to the concavification method. First, we show that this is the concavification of the extension of the payoff function on the hypograph of the entropy function. Second, we show that a Lagrangian method can be used, that is, we can express this value as the concavification of a Lagrangian function. These findings are presented in the next theorem.

Theorem 5.1. *For each $\mu \in \Delta(\Omega)$ and $C \geq 0$,*

1. *$V(\mu, C)$ is the concavification of the function $U_S^H : \Delta(\Omega) \times \mathbb{R} \to \mathbb{R}$,*

$$U_S^H(\nu, \eta) := \begin{cases} U_S(\nu) & \textit{if } \eta \leq H(\nu), \\ -\infty & \textit{otherwise,} \end{cases}$$

   calculated at $(\nu, \eta) = (\mu, H(\mu) - C)$.

2. *$V(\mu, C) = \inf_{t \geq 0} \Big\{\text{cav}\,(U_S + tH)(\mu) - t(H(\mu) - C)\Big\}$.*

The proof is in the Appendix (A.2). A direct implication of the second point is that there exists[2] $t^* = t^*(\mu, C)$ such that,

$$V(\mu, C) = \text{cav}\,(U_S + t^* H)(\mu) - t^*(H(\mu) - C).$$

[2] To see the existence of t^*, notice that $\text{cav}\,(U_S + tH)(\mu) - t(H(\mu) - C) \geq (U_S + tH)(\mu) - t(H(\mu) - C) = U_S(\mu) + tC$, which tends to $+\infty$ as $t \to +\infty$. Therefore, $t \mapsto \text{cav}\,(U_S + tH)(\mu) - t(H(\mu) - C)$ reaches a minimum at some t^*.

If $(\lambda_m^*, \nu_m^*)_m$ is an optimal splitting, let $\mathcal{I}^* = H(\mu) - \sum_m \lambda_m^* H(\nu_m^*)$ be its mutual information. We have,

$$V(\mu, C) = \sum\nolimits_m \lambda_m^* U_S(\nu_m^*) - t^*(\mathcal{I}^* - C). \tag{1}$$

We find then the usual Kuhn-Tucker slackness conditions. If $\mathcal{I}^* < C$, then $t^* = 0$, the unconstrained optimum is feasible.

If $t^* > 0$, the constraint is binding. The Lagrange multiplier t^* can be interpreted as the *shadow price of capacity*, that is, the marginal value of an extra unit of communication capacity.

This characterization has to be related with the *cost of information* considered in the literature on rational inattention (see Sims, 2003) where the agent pays a cost proportional to the mutual information between the state and the signal he observes. For persuasion games, Gentzkow and Kamenica (2014) assume that the sender pays a cost for choosing a disclosure strategy which is also related to the mutual information. In order to define this cost independently of the prior, they consider the mutual information between the state and the message *for a fixed exogenous distribution* of the state.

Our main result and its implication Equation (1) can be seen as a way to justify the use of mutual information as the information cost: we obtain it as a shadow cost. The optimal value of persuasion for a large number of copies of problems with communication over a noisy channel has the same value as a problem of persuasion with an information cost. There are some differences though. First, the information cost is not the mutual information, but the difference between the mutual information and the capacity of the channel. That is, a cost reduces the payoff only when the sender would like to send more information bits than the capacity. Second, the unit price of capacity is endogenous and given by the Lagrange multiplier of the information constraint.

6 Number of messages

In this section, we study the minimal number of messages required to achieve the optimal payoff. In unrestricted persuasion problems, it is known that the necessary number of messages to achieve the best payoff for the receiver is no more than the number of states (see Kamenica and Gentzkow, 2011).

With restricted communication, that is under information constraint, Theorem 5.1 shows that we are calculating the concavification of the payoff function with respect to an extra dimension, which suggests that an extra message might be needed.

Intuitively, it might be optimal to split on an extra posterior which does not yield a good payoff, but helps in satisfying the information constraint.

Lemma 6.1. *In the optimization problem,*

$$V(\mu, C) = \sup\left\{\sum_m \lambda_m U_S(\mu_m) : \sum_m \lambda_m \mu_m = \mu, H(\mu) - \sum_m \lambda_m H(\mu_m) \leq C\right\},$$

the number of posteriors can be restricted to $|\Omega| + 1$*. That is, without loss of generality, the supremum is taken over families* $(\lambda_m, \nu_m)_{m=1,\ldots,|\Omega|+1}$.

To make this intuition concrete, consider the following example.

Example 6.2. *Two-sided investment.* Consider the following payoff table.

	a_0	a_1	a_2	
ω_0	$(0,0)$	$(1,-7)$	$(1,1)$	$\frac{1}{2}$
ω_1	$(0,0)$	$(1,1)$	$(1,-7)$	$\frac{1}{2}$

There are two risky projects (a_1 and a_2) and the sender wants to persuade the receiver to invest in any of them. The sender invests only if $\nu(\omega_1) > 7/8$ or $\nu(\omega_1) < 1/8$.

With unrestricted communication, the solution is clear: the sender fully discloses the state and gets a payoff of 1. However, with a binary symmetric channel with noise $\varepsilon = 1/4$, the sender gets 0 in the single problem. Consider now n copies and assume that the channel can be used n times (n large).

The "one-sided" solution of Example 4.4 is feasible. Recall that this is the splitting such that,

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \lambda\left(\frac{1}{8}, \frac{7}{8}\right) + (1-\lambda)(\nu(\omega_0), \nu(\omega_1))$$

and

$$H\left(\frac{1}{4}\right) = \lambda H\left(\frac{7}{8}\right) + (1-\lambda)H(\nu(\omega_0), \nu(\omega_1)).$$

with $\nu(\omega_1) \approx 0.340$ and $\lambda \approx 0.298$. It is easy to see that this is optimal among the splittings with two posteriors. Indeed, it is not possible that the two posteriors induce investment while satisfying the information constraint.

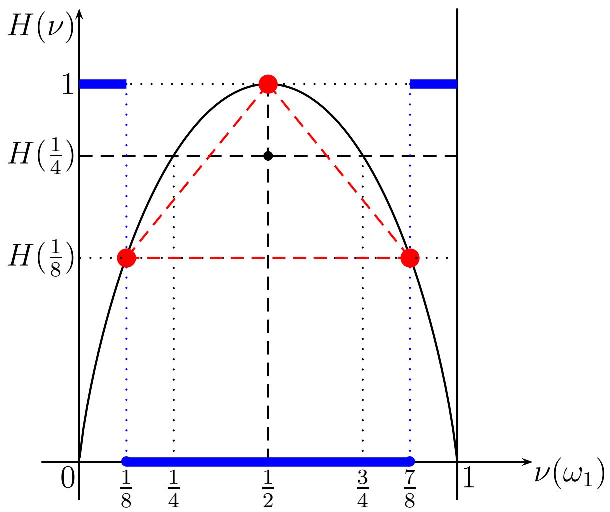
However, this is not optimal. The optimal splitting has three posteriors and is the following.

$$\left(\frac{1}{2}, \frac{1}{2}\right) = (1-\lambda)\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{\lambda}{2}\left(\frac{1}{8}, \frac{7}{8}\right) + \frac{\lambda}{2}\left(\frac{7}{8}, \frac{1}{8}\right)$$

with

$$H\left(\frac{1}{4}\right) = (1-\lambda)H\left(\frac{1}{2}\right) + \frac{\lambda}{2}H\left(\frac{1}{8}\right) + \frac{\lambda}{2}H\left(\frac{7}{8}\right).$$

This pins down a unique λ and solving numerically yields $\lambda \approx 0.413$. Since λ is the probability of investment, we get $V(\mu, Q) \approx 0.413$ which is about 38% better than what is achieved with a splitting with two points.



To see that this is optimal, first since there are two states, we know that three posteriors are sufficient. Second, it is not possible to have all posteriors in the investment region and to satisfy the information constraint. If there is only one posterior in the investment region, then the splitting is achieving no more than the "one-sided" solution. Therefore, it is optimal to have two posteriors in the investment region and one outside of the region. But then, it is optimal to choose the point in the middle region to be $(\frac{1}{2}, \frac{1}{2})$, since this is the one with the highest entropy.

Note that this example involves three actions. Indeed, the number of required messages can be bounded by the number of actions.

Lemma 6.3. *In the optimization problem,*

$$V(\mu, C) = \sup\left\{\sum\nolimits_m \lambda_m U_S(\nu_m) : \sum\nolimits_m \lambda_m \nu_m = \mu, \sum\nolimits_m \lambda_m H(\nu_m) \geq H(\mu) - C\right\},$$

the number of points can be restricted to $\min\{|A|, |\Omega| + 1\}$.

We have already seen that the number of points can be chosen less than or equal to $|\Omega| + 1$. Now intuitively, the number of actions is enough because two posteriors inducing the same action could be replaced by their average without changing payoffs and still satisfying the information constraint, see the Appendix (A.2) for the formal proof.

A Appendix

This appendix contains all the formal proofs. The proof of Theorem 4.3 appears last as it is the most involved and uses some auxiliary results from the proofs of the other results.

A.1 Proofs for Section 3

A.1.1 Proof of Lemma 3.4

For a, b in $[0, 1]$, consider the system,

$$\nu_1(\omega_1) = \frac{\mu(\omega_1)(1-b)}{\mu(\omega_0)a + \mu(\omega_1)(1-b)}, \quad \nu_0(\omega_1) = \frac{\mu(\omega_1)b}{\mu(\omega_0)(1-a) + \mu(\omega_1)b}. \tag{2}$$

If $\nu_1 = \nu_0 = \mu$, then it must be that $a = 1 - b$. Otherwise, $\nu_1(\omega_1) \neq \nu_0(\omega_1)$. It is easily verified that the system has a unique solution given by,

$$b = \frac{\nu_0(\omega_1)(\nu_1(\omega_1) - \mu(\omega_1))}{\mu(\omega_1)(\nu_1(\omega_1) - \nu_0(\omega_1))}$$

and

$$a = \frac{(1-\nu_0(\omega_1))(\mu(\omega_1) - \nu_0(\omega_1))}{(1-\mu(\omega_1))(\nu_1(\omega_1) - \nu_0(\omega_1))}.$$

Take a strategy σ defined by $\sigma(x_0|\omega_0) = 1 - \alpha$ and $\sigma(x_1|\omega_1) = 1 - \beta$ and a binary symmetric channel with noise ε. The posteriors ν_1, ν_0 are given by the system (2) for $a = \alpha \star \varepsilon$ and $b = \beta \star \varepsilon$. As α, β vary in $[0, 1]$, $\alpha \star \varepsilon$ and $\beta \star \varepsilon$ range freely over $[\varepsilon, 1-\varepsilon]$,

$$\{(\alpha \star \varepsilon, \beta \star \varepsilon) : (\alpha, \beta) \in [0, 1]^2\} = [\varepsilon, 1-\varepsilon]^2.$$

This concludes the proof. □

A.1.2 Proof of Claim 3.7

A generic belief over $\{\omega_0, \omega_1\} \times \{\omega_0, \omega_1\}$ is denoted ν. An action for the receiver is a pair in $\{a_0, a_1\} \times \{a_0, a_1\}$ and we denote it $a = (a(1), a(2))$.

The receiver has to choose actions in two separate decision problems. In each problem, he will invest if the probability of the high state is above $\frac{7}{8}$. For the sake of the calculation, we assume that the receiver invests in case of indifference (otherwise, we know that the optimal value is obtained with arbitrary precision).

The receiver with belief ν chooses:

$a(1) = a_1$ if $\nu(\omega_1, \omega_0) + \nu(\omega_1, \omega_1) \geq \frac{7}{8}$, $a(1) = a_0$ otherwise;

$a(2) = a_1$ if $\nu(\omega_0, \omega_1) + \nu(\omega_1, \omega_1) \geq \frac{7}{8}$, $a(2) = a_0$ otherwise.

Consider a splitting of the uniform prior $\mu = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) = \lambda\nu_0 + (1-\lambda)\nu_1$. We have,

$$\frac{1}{2} = \mu(\omega_1, \omega_0) + \mu(\omega_1, \omega_1) = \lambda(\nu_0(\omega_1, \omega_0) + \nu_0(\omega_1, \omega_1)) + (1-\lambda)(\nu_1(\omega_1, \omega_0) + \nu_1(\omega_1, \omega_1)).$$

Suppose that $\nu_0(\omega_1, \omega_0) + \nu_0(\omega_1, \omega_1) \geq \frac{7}{8}$, then it must be that $\nu_0(\omega_1, \omega_0) + \nu_0(\omega_1, \omega_1) < \frac{1}{2}$. This implies that for any splitting with two posteriors, the receiver chooses (a_0, a_0) at one of the two posteriors. Then, there are two possibilities. At the other posterior, either the receiver invests in only one of the problems and the average payoff is $\frac{1}{2}$ for the sender, or the receiver invests in both and the average payoff is 1 for the sender.

In the first case, by symmetry, say that the receiver invests in the first problem only. The sender then gets optimally $\frac{4}{7}$ in the first problem and 0 in the second, thus an average payoff of $\frac{2}{7}$.

In the second case, we look for the optimal way of splitting the uniform prior between ν_0 and ν_1 with $\nu_1(\omega_1, \omega_0) + \nu_1(\omega_1, \omega_1) \geq \frac{7}{8}$ and $\nu_1(\omega_0, \omega_1) + \nu_1(\omega_1, \omega_1) \geq \frac{7}{8}$.

First, let us remark that it is without loss of generality to consider posteriors with the following symmetry $\nu(\omega_0, \omega_1) = \nu(\omega_1, \omega_0)$. To see this, given a belief ν, define $\tilde{\nu}$ such that $\tilde{\nu}(\omega_i, \omega_j) = \nu(\omega_j, \omega_i)$. For a splitting $\mu = \lambda\nu_0 + (1-\lambda)\nu_1$, the symmetrized splitting $\mu = \lambda\tilde{\nu}_0 + (1-\lambda)\tilde{\nu}_1$ achieves the same payoff. Thus the sender gets the same payoff with,

$$\mu = \lambda\frac{\nu_0 + \tilde{\nu}_0}{2} + (1-\lambda)\frac{\nu_1 + \tilde{\nu}_1}{2}$$

which is symmetric. A symmetric posterior with $\nu_1(\omega_1, \omega_0) + \nu_1(\omega_1, \omega_1) \geq \frac{7}{8}$ and $\nu_1(\omega_0, \omega_1) +$

$\nu_1(\omega_1, \omega_1) \geq \frac{7}{8}$ can thus be written as,

$$(\nu(\omega_0, \omega_0), \nu(\omega_1, \omega_0), \nu(\omega_0, \omega_1), \nu(\omega_1, \omega_1)) = (1 - 2p - q, p, p, q),$$

with $p + q \geq \frac{7}{8}$ and $2p + q \leq 1$.

Second, among this set, it is optimal to split on a posterior such that $p + q = \frac{7}{8}$. Indeed, a line segment joining $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ to some $(1 - 2p' - q', p', p', q')$ with $p' + q' \geq \frac{7}{8}$, must contain some $(1 - 2p - q, p, p, q)$ with $p + q = \frac{7}{8}$. The optimal splitting is thus of the form

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = (1 - \lambda)(1 - 2\tilde{p} - \tilde{q}, \tilde{p}, \tilde{p}, \tilde{q}) + \lambda\left(\frac{1}{8} - p, p, p, \frac{7}{8} - p\right)$$

with $p \in [0, \frac{1}{8}]$, $\tilde{p} + 2\tilde{q} \leq 1$ (and necessarily $\tilde{p} + \tilde{q} \leq \frac{1}{2}$). Then optimally, we choose $(1 - 2\tilde{p} - \tilde{q}, \tilde{p}, \tilde{p}, \tilde{q})$ on the boundary of the probability simplex. Actually, we can choose $\tilde{q} = 0$. Precisely, for every $p \in [0, \frac{1}{8}]$, there exists $\lambda \in [0, 1]$ and $\tilde{p} \in [0, \frac{1}{2}]$ such that,

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = (1 - \lambda)(1 - 2\tilde{p}, \tilde{p}, \tilde{p}, 0) + \lambda\left(\frac{1}{8} - p, p, p, \frac{7}{8} - p\right).$$

Solving this equation yields,

$$\lambda = \frac{2}{7 - 8p} \quad \text{and} \quad \tilde{p} = \frac{\frac{7}{4} - 4p}{5 - 8p}.$$

It is easy to verify that $\lambda \in [0, 1]$ and $\tilde{p} \in [0, \frac{1}{2}]$. The payoff for this splitting is $\lambda = \frac{2}{7-8p}$ which is maximal for $p = \frac{1}{8}$, thus the optimal value of $\frac{1}{3}$. □

A.2 Proofs for Sections 5 and 6

A.2.1 Proof of Theorem 5.1

In this section, we prove a statement more general than Theorem 5.1. As a matter of fact, there is nothing specific to the entropy function, and a similar result holds for general functions.

Let $f : X \to \mathbb{R} \cup \{-\infty\}$ be a real-valued function defined on a convex domain X of $\mathbb{R}^d$.

The *concavification* of f is the smallest function $\text{cav}\, f : X \to \mathbb{R} \cup \{-\infty\}$ which is concave and majorizes f on X. This is the concave function whose hypograph is the convex hull of the hypograph of f. It is given by the following optimisation problem:

$$\text{cav}\, f(x) = \sup \left\{ \sum\nolimits_m \lambda_m f(x_m) : \sum\nolimits_m \lambda_m x_m = x \right\},$$

where the supremum ranges over all convex combinations $(\lambda_m, x_m)_m$, $x_m \in X$, $\lambda_m \geq 0$, $\sum_m \lambda_m = 1$ and $\sum_m \lambda_m x_m = x$ (see Rockafellar, 1970, p. 36).

We introduce now a concavification with constraint. Let $f, g : X \to \mathbb{R} \cup \{-\infty\}$ be two functions defined on X. For $x \in X$ and $\gamma \in \mathbb{R}$ consider the problem:

$$\text{cav}_g\, f(x, \gamma) := \sup \left\{ \sum\nolimits_m \lambda_m f(x_m) : \sum\nolimits_m \lambda_m x_m = x, \sum\nolimits_m \lambda_m g(x_m) \geq \gamma \right\}.$$

The optimal splitting under information constraint is an instance of this problem:

$$\sup \left\{ \sum\nolimits_m \lambda_m U_S(\nu_m) : \sum\nolimits_m \lambda_m \nu_m = \mu, \sum\nolimits_m \lambda_m H(\nu_m) \geq H(\mu) - C \right\}.$$

Lemma A.1. *Let $f^g : X \times \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ defined by,*

$$f^g(x, \gamma) = \begin{cases} f(x) & \text{if } \gamma \leq g(x), \\ -\infty & \text{otherwise.} \end{cases}$$

Then for each $(x, \gamma) \in X \times \mathbb{R}$,

$$\text{cav}_g\, f(x, \gamma) = \text{cav}\, f^g(x, \gamma). \tag{3}$$

That is, the problem of optimal splitting with payoff function f under the constraint $\sum_m \lambda_m g(x_m) \geq \gamma$ is in fact the concavification of a bi-variate function, which is the extension of f to the hypograph of the constraint g.

Proof. [Lemma A.1] The function cav $f^g(x,\gamma)$ is given by the following program:

$$\begin{aligned} &\sup \sum\nolimits_m \lambda_m f(x_m) \\ &\text{s.t.} \sum\nolimits_m \lambda_m x_m = x, \sum\nolimits_m \lambda_m \gamma_m = \gamma \\ &\text{and } \forall m, \gamma_m \leq g(x_m). \end{aligned}$$

Take a family $(\lambda_m, x_m, \gamma_m)_m$ feasible for this program. We have $\sum_m \lambda_m g(x_m) \geq \gamma$, thus this family is feasible for $\text{cav}_g\, f(x,\gamma)$. Therefore, cav $f^g(x,\gamma) \leq \text{cav}_g\, f(x,\gamma)$.

Conversely, take a family $(\lambda_m, x_m)_m$ such that $\sum_m \lambda_m x_m = x$ and $\sum_m \lambda_m g(x_m) \geq \gamma$. Let $\bar{\gamma} = \sum_m \lambda_m g(x_m)$ and for each m, $\gamma_m = g(x_m) + \gamma - \bar{\gamma}$. Then, $\sum_m \lambda_m \gamma_m = \gamma$ and since $\bar{\gamma} \geq \gamma$, for each m, $\gamma_m \leq g(x_m)$. Thus, $(\lambda_m, x_m, \gamma_m)_m$ is feasible for cav $f^g(x,\gamma)$ and cav $f^g(x,\gamma) \geq \text{cav}_g\, f(x,\gamma)$. □

This characterization readily applies to the optimal splitting problem under information constraint. □

Now, we show that the Lagrangian approach is valid for the problem,

$$\text{cav}_g\, f(x,\gamma) = \sup \left\{ \sum\nolimits_m \lambda_m f(x_m) : \sum\nolimits_m \lambda_m x_m = x, \sum\nolimits_m \lambda_m g(x_m) \geq \gamma \right\}.$$

Proposition A.2.

$$\text{cav}_g\, f(x,\gamma) = \inf_{t \geq 0} \left\{ \text{cav}\,(f + tg)(x) - t\gamma \right\}.$$

That is, the concavification under constraint corresponds to the concavification of a Lagrangian.

Proof. [Proposition A.2] Recall that the Fenchel conjugate of $f : X \subseteq \mathbb{R}^d \to \mathbb{R}$ is $f^*(p) = \sup_x\{x \cdot p - f(x)\}$, where $x \cdot p$ denotes the inner product. Then, the largest convex function below f is equal to $(f^*)^*$ (Rockafellar, 1970, Corollary 12.1.1, p. 103), therefore $(f^*)^*(x) = -\text{cav}\,(-f)(x)$. Playing with signs, it follows that,

$$\text{cav}\, f(x) = \inf_p \left\{ x \cdot p + \sup_y \{f(y) - p \cdot y\} \right\}. \tag{4}$$

We apply this formula to the function:

$$f^g(x,\gamma) = \begin{cases} f(x) & \text{if } \gamma \le g(x), \\ -\infty & \text{otherwise.} \end{cases}$$

This gives,

$$\begin{aligned} \operatorname{cav} f^g(x,\gamma) &= \inf_{p,z} \Big\{ p\cdot x + z\gamma + \sup_{y,\eta}\{f^g(y,\eta) - p\cdot y - z\eta\} \Big\} \\ &= \inf_{p,z} \Big\{ p\cdot x + z\gamma + \sup_{y,\eta:\ \eta\le g(y)} \{f(y) - p\cdot y - z\eta\} \Big\}. \end{aligned}$$

If $z > 0$ then by letting $\eta \to -\infty$, the sup is $+\infty$. Therefore in the infimum we can restrict to $z \le 0$. Setting $t = -z \ge 0$ we get,

$$\begin{aligned} \operatorname{cav} f^g(x,\gamma) &= \inf_{t\ge 0,p} \Big\{ p\cdot x - t\gamma + \sup_{y,\eta:\ \eta\le g(y)} \{f(y) - p\cdot y + t\eta\} \Big\} \\ &= \inf_{t\ge 0,p} \Big\{ p\cdot x - t\gamma + \sup_{y}\{f(y) - p\cdot y + tg(y)\} \Big\} \\ &= \inf_{t\ge 0} \Big\{ \inf_{p} \Big\{ p\cdot x + \sup_{y}\{f(y) + tg(y) - p\cdot y\} \Big\} - t\gamma \Big\} \end{aligned}$$

where the second line holds since $t \ge 0$ and the third line is just re-organizing. The result follows by remarking that $\inf_p \Big\{ p\cdot x + \sup_y\{f(y) + tg(y) - p\cdot y\} \Big\} = \operatorname{cav}(f + tg)(x)$. □

This proves the second point of Theorem 5.1.

A.2.2 Proof of Lemma 6.1

Lemma 6.1 follows from a well-known fact about concavification.

Fact A.3. *In the optimisation problem,*

$$\operatorname{cav} f(x) = \sup\Big\{ \sum\nolimits_m \lambda_m f(x_m) : \sum\nolimits_m \lambda_m x_m = x \Big\},$$

where f is defined on $X \subseteq \mathbb{R}^d$, the number of points can be restricted to $d+1$. That is, *without loss of generality, the supremum is taken over families* $(\lambda_m, x_m)_{m=1}^{d+1}$.

The reader is referred to Rockafellar (1970, Corollary 17.1.5, p. 157). This implies that in a persuasion problem with unrestricted communication, the number of messages can be bounded by the dimension of $\Delta(\Omega)$ plus one, that is the number of states.

Corollary A.4. *In the optimisation problem,*

$$\text{cav}_g f(x,\gamma) = \sup\left\{\sum\nolimits_m \lambda_m f(x_m) : \sum\nolimits_m \lambda_m x_m = x, \sum\nolimits_m \lambda_m g(x_m) \geq \gamma\right\}$$

where f *is defined on* $X \subseteq \mathbb{R}^d$*, the number of points can be restricted to* $d+2$*.*

This follows from Lemma A.1 and Fact A.3, since the function f^g is defined on $X \times \mathbb{R} \subseteq \mathbb{R}^{d+1}$. Applying to the problem of optimal splitting under information constraint, gives a number of messages bounded by the dimension of $\Delta(\Omega)$ plus two, that is the number of states plus one. □

A.2.3 Proof of Lemma 6.3

Let $\widetilde{A}(\nu) = \text{argmin}\left\{\sum_\omega \nu(\omega)u_S(\omega,a) : a \in A^*(\nu)\right\}$ be the set of optimal actions of the receiver at ν which are worst for the sender.

Claim A.5. *For any action* a*, the set of* ν*'s such that* $a \in \widetilde{A}(\nu)$ *is convex.*

Proof. Observe first that the set of ν's such that $a \in A^*(\nu)$ is defined by linear inequalities, *i.e.* the optimality of a, therefore is convex. Consider now $a \in \widetilde{A}(\nu_1) \cap \widetilde{A}(\nu_2)$ and let's show that $a \in \widetilde{A}(t\nu_1 + (1-t)\nu_2)$ for $t \in (0,1)$. We have $a \in A^*(\nu_1) \cap A^*(\nu_2)$ and by the remark above, $a \in A^*(t\nu_1 + (1-t)\nu_2)$. Take $b \in A^*(t\nu_1 + (1-t)\nu_2)$. We have thus,

$$\sum\nolimits_\omega (t\nu_1(\omega) + (1-t)\nu_2(\omega))u_R(\omega,a) = \sum\nolimits_\omega (t\nu_1(\omega) + (1-t)\nu_2(\omega))u_R(\omega,b).$$

Since $a \in A^*(\nu_1) \cap A^*(\nu_2)$,

$$\sum\nolimits_\omega \nu_1(\omega)u_R(\omega,a) \geq \sum\nolimits_\omega \nu_1(\omega)u_R(\omega,b), \ \sum\nolimits_\omega \nu_2(\omega)u_R(\omega,a) \geq \sum\nolimits_\omega \nu_2(\omega)u_R(\omega,b).$$

Combined together, we get that $b \in A^*(\nu_1) \cap A^*(\nu_2)$. Since $a \in \widetilde{A}(\nu_1) \cap \widetilde{A}(\nu_2)$,

$$\sum_\omega \nu_1(\omega) u_R(\omega, a) \leq \sum_\omega \nu_1(\omega) u_R(\omega, b), \;\; \sum_\omega \nu_2(\omega) u_S(\omega, a) \leq \sum_\omega \nu_2(\omega) u_S(\omega, b).$$

Taking the convex combination of these two inequalities proves the claim. $\square$

Consider a feasible splitting (λ_m, μ_m) such that $\sum_m \lambda_m \nu_m = \mu$ and $\sum_m \lambda_m H(\nu_m) \geq H(\mu) - C$. For each action a, define $M(a) = \left\{m : \widetilde{A}(\nu_m) = \{a\}\right\}$. Denote $\tilde{\lambda}_a = \sum_{m \in M(a)} \lambda_m$ and

$$\tilde{\nu}_a = \sum_{m \in M(a)} \frac{\lambda_m}{\tilde{\lambda}_a} \nu_m.$$

We have,

$$\begin{aligned} \mu &= \sum_m \lambda_m \nu_m \\ &= \sum_a \tilde{\lambda}_a \sum_{m \in M(a)} \frac{\lambda_m}{\tilde{\lambda}_a} \nu_m \\ &= \sum_a \tilde{\lambda}_a \tilde{\nu}_a. \end{aligned}$$

This defines a splitting of μ with $|A|$ elements. We argue that the payoff is the same as the initial splitting. Let's calculate the expected payoff. From the previous claim, for each action a, $a \in \widetilde{A}(\tilde{\nu}_a)$. We have thus,

$$\begin{aligned} \sum_m \lambda_m U_S(\nu_m) &= \sum_a \tilde{\lambda}_a \sum_{m \in M(a)} \frac{\lambda_m}{\tilde{\lambda}_a} \sum_\omega \nu_m(\omega) U_S(\omega, a) \\ &= \sum_a \tilde{\lambda}_a \sum_\omega \tilde{\nu}_a(\omega) U_S(\omega, a) \\ &= \sum_a \tilde{\lambda}_a U_S(\tilde{\nu}_a). \end{aligned}$$

To conclude the proof, we check that the information constraint is satisfied. This follows from the concavity of entropy. Indeed,

$$H(\tilde{\nu}_a) \geq \sum\nolimits_{m\in M(a)} \frac{\lambda_m}{\tilde{\lambda}_a} H(\nu_m)$$

and thus,

$$\sum\nolimits_a \tilde{\lambda}_a H(\tilde{\nu}_a) \geq \sum\nolimits_m \lambda_m H(\nu_m) \geq H(\mu) - C.$$

□

A.3 Proof of Theorem 4.3

A.3.1 Proof of Theorem 4.3, point 1, the upper bound

1. *For all k, n, $U_S^*(\mu_n, Q_k) \leq V(\mu, \frac{k}{n}C(Q))$.*

Proof. Let us fix a strategy σ of the sender. This induces a probability distribution $\mathbb{P}_\sigma$ of sequences in $\Omega^n \times X^k \times Y^k$, the associated random sequences are denoted $(\boldsymbol{\omega}^n, \boldsymbol{x}^k, \boldsymbol{y}^k)$. Let $\boldsymbol{t}$ be a uniformly distributed random variable over $\{1, \ldots, n\}$, independent from $(\boldsymbol{\omega}^n, \boldsymbol{x}^k, \boldsymbol{y}^k)$ and denote $\boldsymbol{m} = (\boldsymbol{y}^k, \boldsymbol{t})$ taking values in $M = Y^k \times \{1, \ldots, n\}$.

We denote $\widetilde{\mathbb{P}}(\omega, m)$ the joint probability distribution of $(\boldsymbol{\omega}, \boldsymbol{m})$ defined by:

$$\begin{aligned}\widetilde{\mathbb{P}}(\omega, m) =& \widetilde{\mathbb{P}}\big(\boldsymbol{\omega} = \omega, (\boldsymbol{y}^k, \boldsymbol{t}) = m\big) \\ =& \widetilde{\mathbb{P}}(\boldsymbol{t} = t) \cdot \widetilde{\mathbb{P}}\big(\boldsymbol{\omega} = \omega, \boldsymbol{y}^k = y^k \big| \boldsymbol{t} = t\big) \\ =& \frac{1}{n} \cdot \mathbb{P}_\sigma\big(\boldsymbol{\omega}_t = \omega, \boldsymbol{y}^k = y^k\big).\end{aligned}$$

Note that the marginal distribution of $\widetilde{\mathbb{P}}(\omega, m)$ on Ω is equal to the prior μ:

$$\begin{aligned}\widetilde{\mathbb{P}}(\omega) =& \sum\nolimits_{t,y^k} \widetilde{\mathbb{P}}\big(\boldsymbol{\omega} = \omega, \boldsymbol{y}^k = y^k, \boldsymbol{t} = t\big) \\ =& \sum\nolimits_{t,y^k} \frac{1}{n} \cdot \mathbb{P}_\sigma\big(\boldsymbol{\omega}_t = \omega, \boldsymbol{y}^k = y^k\big) \\ =& \sum\nolimits_{t=1}^{n} \frac{1}{n} \cdot \mathbb{P}_\sigma\big(\boldsymbol{\omega}_t = \omega\big) \\ =& \mathbb{P}_\sigma\big(\omega\big) \cdot \sum\nolimits_{t=1}^{n} \frac{1}{n} = \mu(\omega).\end{aligned}$$

Fix now a strategy τ of the receiver $\tau : Y^k \to A^n$ and define $\tilde{\tau} : M \to A$ where $\tilde{\tau}(m) = \tilde{\tau}(y^k, t) = \tau_t(y^k)$, the t-th coordinate of $\tau(y^k)$. The expected average payoff of player $i = R, S$ writes:

$$\mathbb{E}_{\sigma,\tau}\left[\bar{u}_i\right] = \sum_{\omega^n,x^k,y^k} \mathbb{P}_\sigma(\omega^n, x^k, y^k)\left[\frac{1}{n}\sum_{t=1}^{n} u_i\Big(\omega_t, \tau_t(y^k)\Big)\right] \tag{5}$$

$$= \sum_{t=1}^{n}\sum_{\omega_t,x^k,y^k} \frac{1}{n} \cdot \mathbb{P}_\sigma(\omega_t, x^k, y^k) \cdot u_i\Big(\omega_t, \tau_t(y^k)\Big) \tag{6}$$

$$= \sum_{t=1}^{n}\sum_{\omega_t,y^k} \frac{1}{n} \cdot \mathbb{P}_\sigma(\omega_t, y^k) \cdot u_i\Big(\omega_t, \tau_t(y^k)\Big) \tag{7}$$

$$= \sum_{\omega,y^k,t} \widetilde{\mathbb{P}}(\omega, y^k, t) \cdot u_i\Big(\omega, \tilde{\tau}(y^k, t)\Big) \tag{8}$$

$$= \sum_{\omega,m} \widetilde{\mathbb{P}}(\omega, m) \cdot u_i\Big(\omega, \tilde{\tau}(m)\Big). \tag{9}$$

Equation (6) implies Equation (7) by summing over x^k which does not enter the payoff function. All other steps are re-orderings and change of variables.

A strategy τ is a best-reply to σ if and only if:

$$\begin{aligned}
&\tau(y^k) \in \arg\max_{a^n \in A^n} \sum_{\omega^n,x^k,y^k} \mu(\omega^n)\sigma(x^k|\omega^n)Q(y^k|x^k)\bar{u}_R(\omega^n, a^n) \\
\iff \quad &\tilde{\tau}(m) \in \arg\max_{a\in A} \sum_{\omega,m} \widetilde{\mathbb{P}}(\omega, m) \cdot u_R(\omega, a) \\
\iff \quad &\tilde{\tau}(m) \in \arg\max_{a\in A} \sum_{\omega} \tilde{\nu}_\sigma(\omega|m) \cdot u_R(\omega, a) \\
\iff \quad &\tilde{\tau}(m) \in A^*\Big(\tilde{\nu}_\sigma(\cdot|m)\Big)
\end{aligned}$$

where $\tilde{\nu}_\sigma(\omega|m) = \widetilde{\mathbb{P}}(\omega|m)$. We deduce for any strategy σ of the sender and any best-reply τ of the sender, the expected average payoffs are those induced by the splitting

$$\mu(\omega) = \sum_{m} \widetilde{\mathbb{P}}(m)\tilde{\nu}_\sigma(\omega|m).$$

Now, we bound the mutual information of this splitting. For any strategy σ, we have:

$$
\begin{aligned}
0 &\leq I(\boldsymbol{x}^k; \boldsymbol{y}^k) - I(\boldsymbol{\omega}^n; \boldsymbol{y}^k) && (10)\\
&= \sum_{t=1}^{k} H(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}) - \sum_{t=1}^{k} H(\boldsymbol{y}_t|\boldsymbol{x}^k, \boldsymbol{y}^{t-1}) - \sum_{t=1}^{n} H(\boldsymbol{\omega}_t|\boldsymbol{\omega}^{t-1}) + \sum_{t=1}^{n} H(\boldsymbol{\omega}_t|\boldsymbol{y}^k, \boldsymbol{\omega}^{t-1}) && (11)\\
&\leq \sum_{t=1}^{k} H(\boldsymbol{y}_t) - \sum_{t=1}^{k} H(\boldsymbol{y}_t|\boldsymbol{x}_t) - n \cdot H(\boldsymbol{\omega}) + \sum_{t=1}^{n} H(\boldsymbol{\omega}_t|\boldsymbol{y}^k) && (12)\\
&= \sum_{t=1}^{k} I(\boldsymbol{x}_t; \boldsymbol{y}_t) - n \cdot H(\boldsymbol{\omega}) + n \cdot \sum_{t=1}^{n} \mathbb{P}(\boldsymbol{t} = t) \cdot H(\boldsymbol{\omega}|\boldsymbol{y}^k, \boldsymbol{t} = t) && (13)\\
&\leq k \cdot \max_{\mathbb{P}(x)} I(\boldsymbol{x}; \boldsymbol{y}) - n \cdot H(\boldsymbol{\omega}) + n \cdot H(\boldsymbol{\omega}|\boldsymbol{y}^k, \boldsymbol{t}) && (14)\\
&= k \cdot \max_{\mathbb{P}(x)} I(\boldsymbol{x}; \boldsymbol{y}) - n \cdot H(\boldsymbol{\omega}) + n \cdot H(\boldsymbol{\omega}|\boldsymbol{m}) && (15)\\
&= k \cdot \max_{\mathbb{P}(x)} I(\boldsymbol{x}; \boldsymbol{y}) - n \cdot I(\boldsymbol{\omega}; \boldsymbol{m}). && (16)
\end{aligned}
$$

- Equation (10) holds since the triple $(\boldsymbol{\omega}^n, \boldsymbol{x}^k, \boldsymbol{y}^k)$ has the Markov chain property that is, its join distribution writes $\mu(\omega^n)\sigma(x^k|\omega^n)Q(y^k|x^k)$. This implies $I(\boldsymbol{x}^k; \boldsymbol{y}^k) \geq I(\boldsymbol{\omega}^n; \boldsymbol{y}^k)$, that is $\boldsymbol{x}^k$ is more informative that $\boldsymbol{\omega}^n$ about $\boldsymbol{y}^k$ (Cover and Thomas, 2006, Theorem 2.8.1, p. 34).

- Equation (11) comes from the chain rule of entropy $H(\boldsymbol{y}^k) = \sum_{t=1}^{k} H(\boldsymbol{y}_t|\boldsymbol{y}^{t-1})$.

- Equation (12) follows since the channel is memoryless $H(\boldsymbol{y}_t|\boldsymbol{x}^k, \boldsymbol{y}^{t-1}) = H(\boldsymbol{y}_t|\boldsymbol{x}_t)$, the sequence of states is i.i.d. $H(\boldsymbol{\omega}_t|\boldsymbol{\omega}^{t-1}) = H(\boldsymbol{\omega}_t)$, and conditioning reduces entropy $H(\boldsymbol{\omega}_t|\boldsymbol{y}^k, \boldsymbol{\omega}^{t-1}) \leq H(\boldsymbol{\omega}_t|\boldsymbol{y}^k)$.

- Equation (13) is a simple re-writing with the introduction of the uniform random variable $\boldsymbol{t} \in \{1, \ldots, n\}$.

- Equation (14) comes from taking the maximum over the marginal distribution $\mathbb{P}(x)$.

- Equation (15) comes from the change of variable $\boldsymbol{m} = (\boldsymbol{y}^k, \boldsymbol{t})$.

Then, Equation (16) is equivalent to:

$$
\begin{aligned}
& k \cdot \max_{\mathbb{P}(x)} I(\boldsymbol{x}; \boldsymbol{y}) - n \cdot I(\boldsymbol{\omega}; \boldsymbol{m}) \geq 0 \\
\iff \quad & H(\boldsymbol{\omega}|\boldsymbol{m}) \geq H(\boldsymbol{\omega}) - \frac{k}{n} \cdot \max_{\mathbb{P}(x)} I(\boldsymbol{x}; \boldsymbol{y}) \\
\iff \quad & \sum_{m} \lambda_m H(\mu_m) \geq H(\mu) - \frac{k}{n} \cdot C(Q).
\end{aligned}
$$

Therefore, for any strategy σ and all n, k, we have:

$$
\begin{aligned}
&\min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, x^k, y^k} \mu(\omega^n)\sigma(x^k|\omega^n)Q(y^k|x^k)\bar{u}_S(\omega^n, \tau(y^k)) \\
=&\min_{\tilde{\tau} \in BR(\sigma)} \sum\nolimits_{\omega, m} \widetilde{\mathbb{P}}(\omega, m) \cdot u_S\Big(\omega, \tilde{\tau}(m)\Big) \\
=&\sum\nolimits_m \widetilde{\mathbb{P}}(m) \min_{\tilde{\tau}(m) \in A^*(\tilde{\nu}_\sigma(\cdot|m))} \sum\nolimits_\omega \tilde{\nu}_\sigma(\cdot|m) \cdot u_S\Big(\omega, \tilde{\tau}(m)\Big) \\
=&\sum\nolimits_m \widetilde{\mathbb{P}}(m) \cdot U_S\big(\tilde{\nu}_\sigma(\cdot|m)\big) \\
\leq&\sup_\sigma \Big\{ \sum\nolimits_m \widetilde{\mathbb{P}}(m) \cdot U_S\big(\tilde{\nu}_\sigma(\cdot|m)\big) \\
&\qquad \text{s.t. } \sum\nolimits_m \widetilde{\mathbb{P}}(m) \cdot \tilde{\nu}_\sigma(\cdot|m) = \mu, \\
&\qquad \text{and } \sum\nolimits_m \widetilde{\mathbb{P}}(m) \cdot H\big(\tilde{\nu}_\sigma(\cdot|m)\big) \geq H(\mu) - \frac{k}{n} \cdot C(Q) \Big\} \\
=&\sup \Big\{ \sum\nolimits_m \lambda_m \cdot U_S\big(\nu_m\big) \\
&\qquad \text{s.t. } \sum\nolimits_m \lambda_m \nu_m = \mu, \\
&\qquad \text{and } \sum\nolimits_m \lambda_m H\big(\nu_m\big) \geq H(\mu) - \frac{k}{n} \cdot C(Q) \Big\} \\
=& V(\mu, \frac{k}{n} C(Q)).
\end{aligned}
$$

This proves that for all n and k we have:

$$
\begin{aligned}
U_S^*(\mu_n, Q_k) =& \sup_\sigma \min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, x^k, y^k} \mu(\omega^n)\sigma(x^k|\omega^n)Q(y^k|x^k)\bar{u}_S(\omega^n, \tau(y^k)) \\
\leq& V(\mu, \frac{k}{n} C(Q))
\end{aligned}
$$

as desired. □

A.3.2 Proof of Theorem 4.3, point 2, the limit value

2. For any rational number r and all $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that for all (k, n) such that $k = rn$ and $n \geq N(\varepsilon)$, $U_S^*(\mu^n, Q^{rn}) \geq V(\mu, rC(Q)) - \varepsilon$.

Zero capacity. First, let's investigate the case $C(Q) = 0$.

Lemma A.6. *If the channel capacity is equal to zero:* $\max_{p(x)} I(\boldsymbol{x};\boldsymbol{y}) = 0$*, then for all* k, n*, we have:*

$$U_S^*(\mu^n, Q^k) = V(\mu, \frac{k}{n}C(Q)).$$

Proof. [Lemma A.6] Let $(\boldsymbol{x}, \boldsymbol{y})$ be a pair of random variables such that the conditional probability of $\{\boldsymbol{y} = y\}$ given $\{\boldsymbol{x} = x\}$ is $Q(y|x)$. If the capacity of the channel is 0, then $I(\boldsymbol{x}, \boldsymbol{y}) = H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) = 0$ which implies that $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent: no information can be sent through the channel.

This implies that for any splitting which satisfies the information constraint, the random variables $\boldsymbol{\omega}$ and $\boldsymbol{m}$ are independent, and for all $m \in M$ we have $\nu_m = \mu$. Hence:

$$V(\mu, \frac{k}{n}C(Q)) = U_S(\mu).$$

Moreover, for any strategy σ, the sequence of messages $\boldsymbol{y}^k$ of the receiver is independent from the sequence of states $\boldsymbol{\omega}^n$. It follows that,

$$\begin{aligned}
U_S^*(\mu^n, Q^k) =& \sup_{\sigma} \min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, x^k, y^k} \mu^n(\omega^n)\sigma(x^k|\omega^n)Q^k(y^k)\bar{u}_S(\omega^n, \tau(y^k)) \\
=& \min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, y^k} \mu^n(\omega^n)Q^k(y^k)\left[\frac{1}{n}\sum\nolimits_{t=1}^{n} u_S(\omega_t, \tau_t(y^k))\right] \\
=& \frac{1}{n}\sum\nolimits_{t=1}^{n} \min_{a_t \in A^*(\mu)} \sum\nolimits_{\omega_t} \mu(\omega_t)u_S(\omega_t, a_t) \\
=& \min_{a \in A^*(\mu)} \sum\nolimits_{\omega} \mu(\omega)u_S(\omega, a) = U_S(\mu),
\end{aligned}$$

which concludes the proof. □

Positive capacity. We assume from now on $C(Q) > 0$. The goal is to take a splitting of the prior which satisfies the information constraint, and to show that the associated payoff can be approximately achieved by strategy σ of the sender and a best-reply $\tau \in BR(\sigma)$ of the receiver. The next lemma states that we can focus on splittings such that the information constraint is satisfied with strict inequality and where the action of the receiver is unique for each posterior in the splitting. Concretely, we prove that such splittings are dense in the set of feasible splittings.

Recall that we denote $\widetilde{A}(\nu)$ the set of worst optimal actions when the belief is $\nu \in \Delta(\Omega)$,

$$\widetilde{A}(\nu) = \text{argmin}\left\{ \sum\nolimits_{\omega} \nu(\omega) u_S(\omega, a) : a \in A^*(\nu) \right\}.$$

Consider the following program:

$$\begin{aligned}\widehat{V}(\mu, \frac{k}{n}C(Q)) = \sup \Bigg\{ & \sum\nolimits_m \lambda_m U_S(\nu_m) \\ & \text{s.t.} \sum\nolimits_m \lambda_m \nu_m = \mu, \\ & \text{and } \ H(\mu) - \sum\nolimits_m \lambda_m H(\nu_m) < \frac{k}{n} C(Q) \\ & \text{and } \ \forall m, \ \widetilde{A}(\nu_m) \text{ is a singleton } \Bigg\}.\end{aligned}$$

Lemma A.7. *For all integers (k, n), $\mu \in \Delta(\Omega)$ and Q such that $C(Q) > 0$ we have:*

$$V(\mu, \frac{k}{n}C(Q)) = \widehat{V}(\mu, \frac{k}{n}C(Q)). \tag{17}$$

Proof. [Lemma A.7]

Remark A.8. From Lemma 6.3, we know that we can restrict the number of messages, *i.e.* the number of posteriors to $K = \min\{|A|, |\Omega|+1\}$. Therefore, from now on a splitting $(\lambda_m, \nu_m)_m$ will be understood to be a composed of $\lambda = (\lambda_1, \dots, \lambda_K) \in \Delta(\{1, \dots, K\})$ and $(\nu_m)_m \in (\Delta(\Omega))^K$. The set of splittings of μ is thus a convex and compact subset of

$$\Delta(\{1, \dots, K\}) \times (\Delta(\Omega))^K$$

which itself is a compact and convex set in some finite dimension space. All statements below about closed or open sets of splittings relate to the topology induced by the Euclidean topology on this finite dimension space.

We consider the following sets:

$$\begin{aligned}
\mathcal{S}_1 = \Big\{(\lambda_m,\nu_m)_m, \quad & \text{s.t.} \quad \sum\nolimits_m \lambda_m \nu_m = \mu, \\
& \text{and} \quad \sum\nolimits_m \lambda_m H(\nu_m) \geq H(\mu) - \frac{k}{n} C(Q)\Big\}, \\
\mathcal{S}_2 = \Big\{(\lambda_m,\nu_m)_m, \quad & \text{s.t.} \quad \sum\nolimits_m \lambda_m \nu_m = \mu, \\
& \text{and} \quad \forall m,\ \widetilde{A}(\nu_m) \text{ is a singleton } \Big\}, \\
\mathcal{S}_3 = \Big\{(\lambda_m,\nu_m)_m, \quad & \text{s.t.} \quad \sum\nolimits_m \lambda_m \nu_m = \mu, \\
& \text{and} \quad \sum\nolimits_m \lambda_m H(\nu_m) > H(\mu) - \frac{k}{n} C(Q)\Big\}.
\end{aligned}$$

We will prove that the set $\mathcal{S}_2 \cap \mathcal{S}_3$ is dense in $\mathcal{S}_1$, which implies that Equation (17) is satisfied. We first argue that $\widetilde{A}(\nu)$ is a singleton for an open and dense set of posteriors ν.

Definition A.9. *Two actions a and b are equivalent $a \sim_i b$ for player $i = S, R$, if for all $\omega \in \Omega$, $u_i(\omega, a) = u_i(\omega, b)$.*

We say that two actions a and b are *completely equivalent* if they are equivalent for both players. Without loss of generality, we assume that no two actions are completely equivalent. Otherwise, we can merge them into one single action and work on the reduced problem.

Denote $F_i \subseteq \Delta(\Omega)$ the set of beliefs for which player $i \in \{S, R\}$ is indifferent between two actions which are not equivalent:

$$F_i = \Big\{\nu \in \Delta(\Omega) : \ \ \exists a, b,\ a \nsim_i b, \sum\nolimits_\omega \nu(\omega) u_i(\omega, a) = \sum\nolimits_\omega \nu(\omega) u_i(\omega, b)\Big\}.$$

Let $F^c = \Delta(\Omega) \setminus \Big(F_R \cup F_S\Big)$ be the set of beliefs where at least one player is not indifferent between any two actions.

Claim A.10. *The set F^c is open and dense in $\Delta(\Omega)$ and for each $\nu \in F^c$, $\widetilde{A}(\nu)$ is a singleton.*

Proof. [Claim A.10] For each i and each pair of actions a, b with $a \not\sim_i b$, the set,

$$F_i(a,b) = \left\{ \nu \in \Delta(\Omega) : \ \sum_{\omega} \nu(\omega) u_i(\omega, a) = \sum_{\omega} \nu(\omega) u_i(\omega, b) \right\}$$

is a closed hyperplane of dimension $\dim(F_i(a,b)) \leq |\Omega| - 2$. Thus, F_R and F_S are closed and $F_R \cup F_S$ is included in a finite union of hyperplanes of dimension at most $|\Omega| - 2$. The complementary set is thus open and dense in $\Delta(\Omega)$.

Then, if $\widetilde{A}(\nu)$ contains two distinct actions $a \neq b$, both players are indifferent between a and b at ν. So if $\nu \in F^c$, $\widetilde{A}(\nu)$ is a singleton. □

It follows that $\mathcal{S}_2$ is open and dense in $\mathcal{S}_1$.

Claim A.11. *If the channel capacity is strictly positive $C(Q) > 0$, the set $\mathcal{S}_3$ is non-empty, open and dense in $\mathcal{S}_1$.*

Proof. [Claim A.11] Take a feasible splitting $(\lambda_m, \nu_m)_m$ in $\mathcal{S}_1$:

$$\sum_{m} \lambda_m H(\nu_m) \geq H(\mu) - \frac{k}{n} C(Q).$$

For $\varepsilon > 0$, consider the perturbed splitting $(\lambda_m, (1-\varepsilon)\nu_m + \varepsilon\mu)_m$. From concavity of the entropy,

$$\begin{aligned}
\sum_{m} \lambda_m H((1-\varepsilon)\nu_m + \varepsilon\mu) &\geq (1-\varepsilon) \sum_{m} \lambda_m H(\nu_m) + \varepsilon H(\mu), \\
&\geq H(\mu) - \frac{k}{n} C(Q) + \varepsilon \frac{k}{n} C(Q) \\
&> H(\mu) - \frac{k}{n} C(Q),
\end{aligned}$$

thus the information constraint is satisfied with strict inequality for $\varepsilon > 0$. It follows that $\mathcal{S}_3$ is non-empty and dense in $\mathcal{S}_1$. By continuity of the entropy, $\mathcal{S}_3$ is open in $\mathcal{S}_1$. □

Since $\mathcal{S}_2$ and $\mathcal{S}_3$ are open and dense, we conclude that $\mathcal{S}_2 \cap \mathcal{S}_3$ is dense in $\mathcal{S}_1$ and that $V(\mu, \frac{k}{n} C(Q)) = \widehat{V}(\mu, \frac{k}{n} C(Q))$ as desired. □

Given a strategy σ of the sender, we denote the induced expected payoff as follows:

$$\begin{aligned}\widehat{U}_{S,\sigma}(\mu^n, Q^k) &= \min_{\tau \in BR(\sigma)} \sum\nolimits_{\omega^n, x^k, y^k} \mu(\omega^n)\sigma(x^k|\omega^n)Q(y^k|x^k)\bar{u}_S(\omega^n, \tau(y^k)), \\ &= \min_{\tau \in BR(\sigma)} \mathbb{E}_{\boldsymbol{\omega}^n, \boldsymbol{x}^k, \boldsymbol{y}^k}\left[\bar{u}_S(\boldsymbol{\omega}^n, \tau(\boldsymbol{y}^k))\right].\end{aligned}$$

Our goal now is to prove the following.

Proposition A.12. *For any rational number r and all $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that for all (k, n) such that $k = rn$ and $n \geq N(\varepsilon)$, there exists a strategy σ such that:*

$$\left|\widehat{U}_{S,\sigma}(\mu^n, Q^k) - \widehat{V}(\mu, rC(Q))\right| \leq \varepsilon. \tag{18}$$

Proof. [Proposition A.12] Let us fix from now on a splitting $(\lambda_m, \nu_m)_m$ which satisfies the three conditions:

$$\sum\nolimits_m \lambda_m \nu_m = \mu, \tag{19}$$

$$H(\mu) - \sum\nolimits_m \lambda_m H(\mu_m) < rC(Q), \tag{20}$$

$$\forall m,\ \widetilde{A}(\nu_m) \text{ is a singleton.} \tag{21}$$

Let $M = \{1, \ldots, |M|\}$ be the set of messages associated to this splitting. We will consider a strategy of the sender as a mapping

$$\sigma : \Omega^n \to \Delta(M^n \times X^k).$$

This means that conditional on the sequence of states ω^n, the sender chooses a sequence of messages $m^n \in M^n$ and a sequence of symbols $x^k \in X^k$ which he inputs into the channel. Observe that any strategy, *i.e.* any mapping from Ω^n to $\Delta(X^k)$, can be defined in this way. The messages m^n are immaterial and can be seen as a pure mental construct of the sender. Nevertheless, they are important in our construction. These are the messages that the sender *intends* to send to the receiver through the symbols x^k, and that the receiver should decoded from the sequence y^k.

The strategy σ induces a joint probability distribution $\mathbb{P}_\sigma$ over $\Omega^n \times M^n \times X^k \times Y^k$,

$$\mathbb{P}_\sigma\big(\omega^n, m^n, x^k, y^k\big) = \prod_{t=1}^n \mu(\omega_t) \times \sigma(m^n, x^k|\omega^n) \times \prod_{t=1}^n Q(y_t|x_t)$$

Let us denote $\nu^\sigma_{t,y^k} \in \Delta(\Omega)$ the posterior belief on $\boldsymbol{\omega}_t$ conditional on the sequence y^k. That is,

$$\nu^\sigma_{t,y^k}(\omega) = \mathbb{P}_\sigma\big(\boldsymbol{\omega}_t = \omega \mid y^k\big).$$

For each sequence y^k of messages and for each t, the receiver chooses an optimal action $a_t \in A^*(\nu^\sigma_{t,y^k})$. In the worst case (for the sender), this action a_t belongs to $\widetilde{A}(\nu^\sigma_{t,y^k})$. It follows that,

Claim A.13.

$$\widehat{U}_{S,\sigma}(\mu^n, Q^k) = \sum_{m^n,y^k} \mathbb{P}_\sigma(m^n, y^k)\frac{1}{n}\sum_{t=1}^n U_S(\nu^\sigma_{t,y^k}).$$

Now, we will define an event $B \subseteq M^n \times Y^k$ such that for every $(m^n, y^k) \in B$, $\frac{1}{n}\sum_{t=1}^n U_S(\nu^\sigma_{t,y^k})$ is close to $\sum_m \lambda_m U_S(\nu_m)$. We need some notations. For $\nu_1, \nu_2 \in \Delta(\Omega)$, the ℓ^1 distance is $\|\nu_1 - \nu_2\| = \sum_\omega |\nu_1(\omega) - \nu_2(\omega)|$. The Kullback-Leibler (KL) divergence is,

$$D(\nu_1\|\nu_2) = \sum_\omega \nu_1(\omega)\log\frac{\nu_1(\omega)}{\nu_2(\omega)}.$$

These two distances are related by Pinsker's inequality (Cover and Thomas, 2006, Lemma 11.6.1, p. 370):

$$\|\nu_1 - \nu_2\| \leq \sqrt{2\ln 2\, D(\nu_1\|\nu_2)}.$$

We will introduce several positive parameters α, γ, δ, to be thought of as small.

Notation A.14. *For a sequence (m^n, y^k) and $\alpha > 0$, denote*

$$T_\alpha(m^n, y^k) = \Big\{t \in \{1, \dots, n\} : D(\nu^\sigma_{t,y^k}\|\nu_{m_t}) \leq \frac{\alpha^2}{2\ln 2}\Big\}.$$

This is the set of indices $t = 1, \ldots, n$ such that the posterior belief ν^{σ}_{t,y^k} about $\boldsymbol{\omega}_t$ is close to the theoretical belief ν_{m_t}. Intuitively, this is the set of indices where *the message* m_t is approximately transmitted.

Remark A.15. *Since the set of posteriors ν such that $\widetilde{A}(\nu)$ is a singleton is open, there exists $\alpha_0 > 0$ such that for all m,*

$$D(\nu \| \nu_m) \leq \alpha_0 \Rightarrow \widetilde{A}(\nu) = \widetilde{A}(\nu_m).$$

Whenever $\widetilde{A}(\nu)$ is a singleton, denote $\widetilde{A}(\nu) = \{\tilde{a}(\nu)\}$ the unique (worst) optimal action. From now on, we assume that $\alpha \in (0, \alpha_0)$. With the remark above, this implies that for each $t \in T_\alpha(m^n, y^k)$, the action chosen by the receiver for problem t is $\tau_t(m^n, y^k) = \tilde{a}(\nu_{m_t})$. So precisely, $T_\alpha(m^n, y^k)$ is the set of indices t such that the receiver plays the action $\tilde{a}(\nu_{m_t})$ which corresponds to the message m_t. So in this sense, this is the set of indices for which the information transmission is successful.

Notation A.16. *For a sequence (m^n, y^k) and $m \in M$, denote*

$$\operatorname{freq}_m(m^n, y^k) = \frac{1}{n}\Big|\{t = 1, \ldots, n : m_t = m\}\Big|$$

the empirical frequency of message m in the sequence m^n.

For $\alpha, \gamma, \delta > 0$, let

$$B_{\alpha,\gamma,\delta} = \Big\{(m^n, y^k) : \frac{|T_\alpha(m^n, y^k)|}{n} \geq 1 - \gamma \text{ and } \sum\nolimits_m |\lambda_m - \operatorname{freq}_m(m^n, y^k)| \leq \delta\Big\}$$

Lemma A.17. *For each $(m^n, y^k) \in B_{\alpha,\gamma,\delta}$,*

$$\Big|\frac{1}{n}\sum\nolimits_{t=1}^{n} U_S(\nu^{\sigma}_{t,y^k}) - \sum\nolimits_m \lambda_m U_S(\nu_m)\Big| \leq (\alpha + 2\gamma + \delta)\|u\|,$$

where $\|u\| = \max_{\omega,a} |u_S(\omega, a)|$ is the largest absolute value of payoffs for the receiver.

Proof. Denote $u^* = \sum_m \lambda_m U_S(\nu_m)$. We have

$$\begin{aligned}\Big|\frac{1}{n}\sum\nolimits_{t=1}^{n} U_S(\nu^{\sigma}_{t,y^k}) - u^*\Big| &\leq \Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - u^*)\Big| + \Big|\frac{1}{n}\sum\nolimits_{t\notin T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - u^*)\Big| \\ &\leq \Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - u^*)\Big| + \gamma\|U\|\end{aligned}$$

Then,

$$\begin{aligned}\Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - u^*)\Big| &\leq \Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - U_S(\nu_{m_t}))\Big| \\ &+ \Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu_{m_t}) - u^*)\Big|\end{aligned}$$

Since $\alpha \leq \alpha_0$, for each $t \in T_\alpha(m^n, y^k)$, $\tilde{a}(\nu^{\sigma}_{t,y^k}) = \tilde{a}(\nu_{m_t})$. Therefore, for $t \in T_\alpha(m^n, y^k)$

$$\Big|U_S(\nu^{\sigma}_{t,y^k}) - U_S(\nu_{m_t})\Big| \leq \sum\nolimits_{\omega} |\nu^{\sigma}_{t,y^k}(\omega) - \nu_{m_t}(\omega))| \cdot |u_S(\omega, a)| \leq \|\nu^{\sigma}_{t,y^k} - \nu_{m_t}\| \cdot \|u\| \leq \alpha\|u\|,$$

where the latter inequality comes from Pinsker's inequality and the definition of $T_\alpha(m^n, y^k)$. It follows,

$$\Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu^{\sigma}_{t,y^k}) - u^*)\Big| \leq \alpha\|u\| + \Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu_{m_t}) - u^*)\Big|$$

Now from $\frac{|T_\alpha(m^n,y^k)|}{n} \geq 1 - \gamma$, we have,

$$\Big|\frac{1}{n}\sum\nolimits_{t\in T_\alpha(m^n,y^k)}(U_S(\nu_{m_t}) - u^*)\Big| \leq \Big|\frac{1}{n}\sum\nolimits_{t=1}^{n}(U_S(\nu_{m_t}) - u^*)\Big| + \gamma\|u\|.$$

Then,

$$\begin{aligned}\Big|\frac{1}{n}\sum\nolimits_{t=1}^{n}(U_S(\nu_{m_t}) - u^*)\Big| &= \Big|\sum\nolimits_{m}(\text{freq}_{\,m}(m^n, y^k) - \lambda_m)U_S(\nu_m)\Big| \\ &\leq \sum\nolimits_{m}\Big|\text{freq}_{\,m}(m^n, y^k) - \lambda_m\Big| \cdot \Big|U_S(\nu_m)\Big| \\ &\leq \|u\|\delta.\end{aligned}$$

Collecting all inequalities together yields the desired conclusion. □

We have the direct consequence:

Corollary A.18.

$$\left|\widehat{U}_{S,\sigma}(\mu^n, Q^k) - \widehat{V}(\mu, rC(Q))\right| \leq (\alpha + 2\gamma + \delta)\|u\| + (1 - \mathbb{P}_\sigma(B_{\alpha,\gamma,\delta}))\|u\|.$$

We see from this inequality that estimating the probability of the set $B_{\alpha,\gamma,\delta}$ is key. The last step of the proof is the actual construction of the strategy. The idea is that, since the information constraint is satisfied *i.e.* $I(\boldsymbol{\omega}; \boldsymbol{m}) < rC(Q)$, there is enough capacity to transmit $\boldsymbol{m}$ over the channel. More precisely, we construct a strategy such that the set $B_{\alpha,\gamma,\delta}$ has probability close to 1. This way, for most sequences $(\omega^n, m^n, x^k, y^k)$, the receiver *gets the right message* in most stages. That is, at most stages the receiver plays the action corresponding to the message.

We turn now to the actual construction. We use standard information theoretic techniques for Channel Coding (Gamal and Kim, 2011, Chap. 3.1, p. 38) and Lossy Source Coding (Gamal and Kim, 2011, Chap. 3.6, p. 56). Using information theoretic language, the sender is viewed as an *encoder* who encrypts his intended m^n messages in sequences of inputs x^k. The encoding is such that a *decoder* who reads the sequence y^k, is able to find out the correct m^n with high probability. This is described as follows.

For $\delta > 0$, we define the set of *typical sequences* A_δ as follows:

$$A_\delta = \Big\{(\omega^n, m^n, x^k, y^k), \quad \text{s.t.} \quad \sum_{\omega,m} \left|\lambda_m \mu_m(\omega) - \text{freq}_{\omega,m}(\omega^n, m^n)\right| \leq \delta, \tag{22}$$

$$\text{and} \quad \sum_{x,y} \left|\mathbb{P}(x) \times Q(y|x) - \text{freq}_{x,y}(x^k, y^k)\right| \leq \delta \Big\}. \tag{23}$$

A pair of sequences (ω^n, m^n) which satisfies Equation (22) will be called *jointly typical*. Similarly, pair of sequences (x^k, y^k) which satisfies Equation (23) will be called *jointly typical*. With a slight abuse of notation, we will write $(\omega^n, m^n) \in A_\delta$ or $(x^k, y^k) \in A_\delta$ to indicate jointly typical sequences.

Since condition (20) is satisfied with strict inequality, there exists a small parameter

$\eta > 0$ and a "rate" $\mathsf{R} \geq 0$, such that:

$$\begin{aligned} \mathsf{R} =& H(\mu) - \sum\nolimits_m \lambda_m H(\mu_m) + \eta, && (24) \\ \mathsf{R} \leq& rC(Q) - \eta. && (25) \end{aligned}$$

Moreover, we can assume that $n\mathsf{R}$ is an integer for n large enough.

- *Random codebook.* A *codebook* is a family c of $|J| = 2^{n\mathsf{R}}$ sequences $m^n(j)$ and $x^k(j)$ indexed by $j \in J$. A *random codebook* is the draw of a codebook from the marginal i.i.d. probability distributions $(\lambda_m)^{\otimes n}$ and $\mathbb{P}(x)^{\otimes n}$. The selected codebook is known by the encoder and the decoder.

- *Encoding function.* The encoder observes the sequence of states $\omega^n \in \Omega^n$. It finds an index $j \in J$ such that the sequences $(\omega^n, m^n(j)) \in A_\delta$ are jointly typical, *i.e.* satisfy Equation (22). The encoder sends the sequence $x^k(j)$ corresponding to the index $j \in J$.

- *Decoding function.* The decoder observes the sequence of channel output $y^k \in Y^k$. It finds an index $\hat{j} \in J$ such that the sequences $\left(x^k(\hat{j}), y^k\right) \in A_\delta$ are jointly typical, *i.e.* satisfy Equation (23). The decoder decodes the sequence $m^n(\hat{j})$.

- *Error Event.* We introduce the indicator of error $E_\delta \in \{0, 1\}$ defined as follows:

$$E_\delta = \begin{cases} 0 \text{ if } & j = \hat{j} \quad \text{and} \quad \left(\omega^n, m^n, x^k, y^k\right) \in A_\delta, \\ 1 \text{ if } & j \neq \hat{j} \quad \text{or} \quad \left(\omega^n, m^n, x^k, y^k\right) \notin A_\delta. \end{cases} \quad (26)$$

  An error $E_\delta = 1$ occurs in the coding process if: 1) the indexes $j \in J$ and $\hat{j} \in J$ are not equal or 2) the sequences of symbols $\left(\omega^n, m^n, x^k, y^k\right) \notin A_\delta$, *i.e.* are not jointly typical.

An important result in information theory is that the expected probability of error over the random codebook is small.

Expected error probability. For all $\varepsilon_2 > 0$, for all $\eta > 0$, there exists a $\bar{\delta} > 0$, for all $\delta \leq \bar{\delta}$ there exists $\bar{n}$ such that for all $n \geq \bar{n}$ and $k = r \cdot n$, the expected probability of the following error events are bounded by ε_2:

$$\mathbb{E}\bigg[\mathbb{P}_c\bigg(\forall j \in J, \quad \big(\omega^n, m^n(j)\big) \notin A_\delta\bigg)\bigg] \leq \varepsilon_2, \tag{27}$$

$$\mathbb{E}\bigg[\mathbb{P}_c\bigg(\exists j' \neq j, \text{ s.t. } \big(y^k, x^k(j')\big) \in A_\delta\bigg)\bigg] \leq \varepsilon_2. \tag{28}$$

- Equation (27) comes from Equation (24) and the Covering Lemma A.19, (Gamal and Kim, 2011, Lemma 3.3, p. 62).
- Equation (28) comes from Equation (25) and the Packing Lemma A.20, (Gamal and Kim, 2011, Lemma 3.1, p. 46).

If the expected probability of error is small over the codebooks, then it has to be small for at least one codebook. Following a standard analysis of the error probability, (Gamal and Kim, 2011, pp. 42–43, 60–61), Equations (27), (28) imply that:

$$\forall \varepsilon_2 > 0,\ \forall \eta > 0,\ \exists \bar{\delta} > 0,\ \forall \delta \leq \bar{\delta},\ \exists \bar{n} > 0,\ \forall n \geq \bar{n}, \quad \exists c^\star, \qquad \mathbb{P}_{c^\star}\big(E_\delta = 1\big) \leq \varepsilon_2. \tag{29}$$

The strategy σ of the sender consists in using this codebook $c^\star$ in order to find the sequence $m^n(j)$ which is jointly typical with ω^n, and in sending the sequence $x^k(j)$. By construction, this satisfies Equation (29), *i.e.* it has a low probability of error.

Control of the Beliefs. This construction has the property that the decoder who uses the decoding schemes, makes an error with small probability. Now, the receiver needs not use the decoding scheme. Actually, the receiver calculates the posterior belief on the sequence of states ω^n, given y^k. The next step shows that those beliefs are close to the prescribed beliefs ν_m at most stages. We have the following chain of inequalities.

$$\mathbb{E}_\sigma\left[\frac{1}{n}\sum\nolimits_{t=1}^n D\Big(\nu^\sigma_{t,y^k}\Big\|\nu_{m_t}\Big)\;\Big|\;E_\delta=0\right]$$

$$=\sum\nolimits_{m^n,y^k}\mathbb{P}_\sigma(m^n,y^k|E_\delta=0)\cdot\frac{1}{n}\sum\nolimits_{t=1}^n D\Big(\nu^\sigma_{t,y^k}\Big\|\nu_{m_t}\Big) \tag{30}$$

$$=\frac{1}{n}\sum\nolimits_{(\omega^n,m^n,y^k)\in A_\delta}\mathbb{P}_\sigma(\omega^n,m^n,y^k|E_\delta=0)\cdot\log_2\frac{1}{\prod_{t=1}^n\nu_{m_t}(\omega_t)}-\frac{1}{n}\sum\nolimits_{t=1}^n H(\boldsymbol{\omega}_t|\boldsymbol{y}^k,E_\delta=0) \tag{31}$$

$$\leq\frac{1}{n}\sum\nolimits_{(\omega^n,m^n,y^k)\in A_\delta}\mathbb{P}_\sigma(\omega^n,m^n,y^k|E_\delta=0)\cdot\log_2\frac{1}{\prod_{t=1}^n\nu_{m_t}(\omega_t)}-\frac{1}{n}\sum\nolimits_{t=1}^n H(\boldsymbol{\omega}_t|\boldsymbol{m}^n,\boldsymbol{y}^k,E_\delta=0) \tag{32}$$

$$\leq\frac{1}{n}\sum\nolimits_{(\omega^n,m^n,y^k)\in A_\delta}\mathbb{P}_\sigma(\omega^n,m^n,y^k|E_\delta=0)\cdot n\cdot\Big(H(\boldsymbol{\omega}|\boldsymbol{m})+\delta\Big)-\frac{1}{n}H(\boldsymbol{\omega}^n|\boldsymbol{m}^n,\boldsymbol{y}^k,E_\delta=0) \tag{33}$$

$$\leq\frac{1}{n}I(\boldsymbol{\omega}^n;\boldsymbol{m}^n,\boldsymbol{y}^k|E_\delta=0)-I(\boldsymbol{\omega};\boldsymbol{m})+\delta+\frac{1}{n}+\log_2|\Omega|\cdot\mathbb{P}_\sigma\big(E_\delta=1\big) \tag{34}$$

$$\leq\frac{1}{n}I(\boldsymbol{\omega}^n;\boldsymbol{m}^n|E_\delta=0)-I(\boldsymbol{\omega};\boldsymbol{m})+\delta+\frac{2}{n}+2\log_2|\Omega|\cdot\mathbb{P}_\sigma\big(E_\delta=1\big) \tag{35}$$

$$\leq\eta+\delta+\frac{2}{n}+2\log_2|\Omega|\cdot\mathbb{P}_\sigma\big(E_\delta=1\big). \tag{36}$$

- Equation (30) comes from the definition of the expected K-L divergence.

- Equation (31) comes from the conditioning by $E_\delta=0$, since the support of $\mathbb{P}_\sigma(\omega^n,m^n,y^k|E_\delta=0)$ is included in A_δ.

- Equation (32) comes from the property of the entropy $H(\boldsymbol{\omega}_t|\boldsymbol{m}^n,\boldsymbol{y}^k,E_\delta=0)\leq H(\boldsymbol{\omega}_t|\boldsymbol{y}^k,E_\delta=0)$.

- Equation (33) comes from the property of typical sequences $(\omega^n,m^n)\in A_\delta$, stated in Lemma A.21 and in Gamal and Kim (2011, Property 1, pp. 26), and the chain rule for entropy:

$$H(\boldsymbol{\omega}^n|\boldsymbol{m}^n,\boldsymbol{y}^k,E_\delta=0)\leq\sum\nolimits_{t=1}^n H(\boldsymbol{\omega}_t|\boldsymbol{m}^n,\boldsymbol{y}^k,E_\delta=0).$$

- Equation (34) comes from Lemma A.23 (see section A.4), which implies,

$$\frac{1}{n}H(\boldsymbol{\omega}^n|E_\delta=0)-\frac{1}{n}H(\boldsymbol{\omega}^n)+\frac{1}{n}+\log_2|\Omega|\cdot\mathbb{P}_\sigma(E_\delta=1)\geq 0.$$

Adding this expression to Equation (33) yields Equation (34).

- Equation (35) comes from Lemma A.23 (see section A.4) which implies that,

$$I(\boldsymbol{\omega}^n;\boldsymbol{y}^k|\boldsymbol{m}^n, E_\delta = 0) \leq I(\boldsymbol{\omega}^n;\boldsymbol{y}^k|\boldsymbol{m}^n)+1+n\cdot\log_2|\Omega|\cdot\mathbb{P}_\sigma(E_\delta = 1) = 1+n\cdot\log_2|\Omega|\cdot\mathbb{P}_\sigma(E_\delta = 1),$$

where $I(\boldsymbol{\omega}^n;\boldsymbol{y}^k|\boldsymbol{m}^n) = 0$, from the Markov chain property of the triple $(\boldsymbol{\omega}^n, \boldsymbol{m}^n, \boldsymbol{y}^k)$.

- Equation (36) comes from the cardinality of the codebook:

$$I(\boldsymbol{\omega}^n;\boldsymbol{m}^n|E_\delta = 0) \leq H(\boldsymbol{m}^n) \leq \log_2|J| = n\cdot \mathsf{R} = n\cdot(I(\boldsymbol{\omega};\boldsymbol{m}) + \eta).$$

The last argument is inspired by Merhav and Shamai (2007, Equation (23)) for the problem of "Information Rates Subject to State Masking".

Then we have:

$$\begin{aligned}
&1 - \mathbb{P}_\sigma(B_{\alpha,\gamma,\delta}) := \mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta})\\
=&\mathbb{P}_\sigma(E_\delta = 1)\mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta}|E_\delta = 1) + \mathbb{P}_\sigma(E_\delta = 0)\mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta}|E_\delta = 0)\\
\leq&\mathbb{P}_\sigma(E_\delta = 1) + \mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta}|E_\delta = 0)\\
\leq&\varepsilon_2 + \mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta}|E_\delta = 0). \qquad (37)
\end{aligned}$$

Moreover:

$$
\begin{aligned}
&\mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta}|E_\delta=0)\\
=&\sum\nolimits_{m^n,y^k}\mathbb{P}_\sigma\Big((m^n,y^k)\in B^c_{\alpha,\gamma,\delta}\Big|E_\delta=0\Big) && (38)\\
=&\sum\nolimits_{m^n,y^k}\mathbb{P}_\sigma\left((m^n,y^k)\quad\text{s.t.}\quad\frac{|T_\alpha(m^n,y^k)|}{n}<1-\gamma\middle|E_\delta=0\right) && (39)\\
=&\mathbb{P}_\sigma\left(\frac{\#}{n}\left\{t,\ \text{s.t.}\ D\Big(\nu^\sigma_{t,y^k}\Big\|\nu_{m_t}\Big)\le\frac{\alpha^2}{2\ln 2}\right\}<1-\gamma\middle|E_\delta=0\right) && (40)\\
=&\mathbb{P}_\sigma\left(\frac{\#}{n}\left\{t,\ \text{s.t.}\ D\Big(\nu^\sigma_{t,y^k}\Big\|\nu_{m_t}\Big)>\frac{\alpha^2}{2\ln 2}\right\}\ge\gamma\middle|E_\delta=0\right) && (41)\\
\le&\frac{2\ln 2}{\alpha^2\gamma}\cdot\mathbb{E}_\sigma\left[\frac{1}{n}\sum\nolimits_{t=1}^n D\Big(\nu^\sigma_{t,y^k}\Big\|\nu_{m_t}\Big)\right] && (42)\\
\le&\frac{2\ln 2}{\alpha^2\gamma}\cdot\left(\eta+\delta+\frac{2}{n}+2\log_2|\Omega|\cdot\mathbb{P}_\sigma\big(E_\delta=1\big)\right). && (43)
\end{aligned}
$$

- Equations (38) to (41) are simple reformulations.

- Equation (42) comes from a use of Markov's inequality, detailed in Lemma A.22 (see section A.4).

- Equation (43) comes from equation (36).

Combining equations (29), (37), and (43) we obtain the following statement:

$\forall\varepsilon_3>0$, $\forall\gamma>0$, $\exists\bar{\eta}$, $\forall\eta\le\bar{\eta}$, $\exists\bar{\delta}$, $\forall\delta\le\bar{\delta}$, $\exists\bar{n}$, $\forall n\ge\bar{n}$, $\exists\sigma$, such that:

$$\mathbb{P}_\sigma(B^c_{\alpha,\gamma,\delta})\le 2\cdot\mathbb{P}_\sigma\big(E_\delta=1\big)+\frac{2\ln 2}{\alpha^2\gamma}\cdot\left(\eta+\delta+\frac{2}{n}+2\log_2|\Omega|\cdot\mathbb{P}_\sigma\big(E_\delta=1\big)\right)\le\varepsilon_3$$

To conclude the proof of Proposition A.12, we take the inequality of Corollary A.18. We choose the parameters $\alpha,\gamma,\eta,\delta$ small and then n large, in order to get:

$$\left|\widehat{U}_{S,\sigma}(\mu^n,Q^k)-\widehat{V}(\mu,rC(Q))\right|\le(\alpha+2\gamma+\delta)\|u\|+(1-\mathbb{P}_\sigma(B_{\alpha,\gamma,\delta}))\|u\|\le\varepsilon.$$

This ends the proof. □

A.4 Additional lemmas

The next three lemmas are standard results in information theory. They are recalled for the convenience of the reader.

Lemma A.19 (Covering lemma: compression of information source, Lemma 3.3, p. 62 in Gamal and Kim (2011)). *Consider a random sequence ω^n with i.i.d. distribution $\mathbb{P}^{\otimes n}(\omega)$ and a family of 2^{nR} sequences $\big(m^n(j)\big)_{j\in\{1,\ldots,2^{nR}\}}$ independently drawn from the i.i.d. distribution $\mathbb{P}^{\otimes n}(m)$. Assume that $\mathsf{R} = I(\boldsymbol{\omega};\boldsymbol{m}) + \eta$ with $\eta > 0$.*

For all $\varepsilon > 0$, there exists $\bar{\delta} > 0$, such that for all $\delta \leq \bar{\delta}$, there exists $\bar{n}$, such that for all $n \geq \bar{n}$:

$$\mathbb{P}\Big(\forall j \in J, \quad \big(\omega^n, m^n(j)\big) \notin A_\delta\Big) \leq \varepsilon.$$

Lemma A.20 (Packing lemma: transmission over a noisy channel, Lemma 3.1, p. 46 Gamal and Kim (2011)). *Consider a random sequence y^k drawn with i.i.d. distribution $\mathbb{P}^{\otimes k}(y)$ and a family of 2^{kR} sequences $\big(x^k(j)\big)_{j\in\{1,\ldots,2^{kR}\}}$ independently drawn from the i.i.d. distribution $\mathbb{P}^{\otimes k}(x)$. Assume that $\mathsf{R} = I(\boldsymbol{x};\boldsymbol{y}) - \eta$ with $\eta > 0$.*

For all $\varepsilon > 0$, there exists $\bar{\delta} > 0$, such that for all $\delta \leq \bar{\delta}$, there exists $\bar{k}$, such that for all $k \geq \bar{k}$:

$$\mathbb{P}\Big(\exists j \in J, \quad \big(x^k(j), y^k\big) \in A_\delta\Big) \leq \varepsilon.$$

Lemma A.21 (Typical sequences, Property 1, p. 26 in Gamal and Kim (2011)). *The typical sequences $(\omega^n, m^n) \in A_\delta$ satisfy:*

$$\forall \delta_2 > 0,\ \exists \bar{\delta}_2 > 0,\ \forall \delta \leq \bar{\delta}_2,\ \forall n,\ \forall (\omega^n, m^n) \in A_\delta,$$
$$\left| \frac{1}{n} \cdot \log_2 \frac{1}{\prod_{t=1}^n \mathbb{P}(\omega_t|m_t)} - H(\boldsymbol{\omega}|\boldsymbol{m}) \right| \leq \delta_2,$$

where $\bar{\delta}_2 = \delta_2 \cdot H(\boldsymbol{\omega}|\boldsymbol{m})$.

The next two lemmas are easy ancillary results that were used in the proofs and were omitted in the previous section to ease the reading.

Lemma A.22 (Markov's inequality)**.** *For all $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ we have:*

$$\mathbb{E}_\sigma\left[\frac{1}{n}\sum_{t=1}^{n} D\Big(\mathbb{P}_\sigma(\boldsymbol{\omega}_t|\boldsymbol{y}^n, E_\delta = 0)\Big\|\mathbb{P}(\boldsymbol{\omega}_t|\boldsymbol{m}_t)\Big)\right] \leq \varepsilon_0 \tag{44}$$

$$\Longrightarrow \mathbb{P}_{m^n,y^n}\left(\frac{\#}{n}\left\{t,\ s.t.\ D\Big(\mathbb{P}_\sigma(\boldsymbol{\omega}_t|\boldsymbol{y}^n, E_\delta = 0)\Big\|\mathbb{P}(\boldsymbol{\omega}_t|\boldsymbol{m}_t)\Big) > \varepsilon_1\right\} > \varepsilon_2\right) \leq \frac{\varepsilon_0}{\varepsilon_1 \cdot \varepsilon_2}. \tag{45}$$

Proof. [Lemma A.22] We denote by $D_t = D\big(\mathbb{P}_\sigma(\boldsymbol{\omega}_t|\boldsymbol{y}^n, E_\delta = 0)\big\|\mathbb{P}(\boldsymbol{\omega}_t|\boldsymbol{m}_t)\big)$ and $D^n = \{D_t\}_t$ the K-L divergence. We have that:

$$\mathbb{P}\left(\frac{\#}{n}\left\{t,\ \text{s.t. } D_t > \varepsilon_1\right\} > \varepsilon_2\right) = \mathbb{P}\left(\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbb{1}\left\{D_t > \varepsilon_1\right\} > \varepsilon_2\right) \tag{46}$$

$$\leq \frac{\mathbb{E}\left[\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbb{1}\left\{D_t > \varepsilon_1\right\}\right]}{\varepsilon_2} \tag{47}$$

$$= \frac{\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\left\{D_t > \varepsilon_1\right\}\right]}{\varepsilon_2} \tag{48}$$

$$= \frac{\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbb{P}\Big(D_t > \varepsilon_1\Big)}{\varepsilon_2} \tag{49}$$

$$\leq \frac{\frac{1}{n}\cdot\sum_{t=1}^{n}\frac{\mathbb{E}[D_t]}{\varepsilon_1}}{\varepsilon_2} \tag{50}$$

$$= \frac{1}{\varepsilon_1\cdot\varepsilon_2}\cdot\mathbb{E}\left[\frac{1}{n}\cdot\sum_{t=1}^{n} D_t\right] \leq \frac{\varepsilon_0}{\varepsilon_1\cdot\varepsilon_2}. \tag{51}$$

Equations (46), (48), (49), (51) are reformulations of probabilities and expectations. Equations (47), (50), come from Markov's inequality $\mathbb{P}(X \geq \alpha) \leq \mathbb{E}[X]/\alpha$. □

Lemma A.23. *Consider an i.i.d. random sequence $\boldsymbol{\omega}^n$. For all $\varepsilon > 0$, there exists $\bar{n} \in \mathbb{N}$ such that for all $n \geq \bar{n}$ we have:*

$$H(\boldsymbol{\omega}^n|E_\delta = 0) \geq n\cdot\Big(H(\boldsymbol{\omega}) - \varepsilon\Big). \tag{52}$$

Proof. [Lemma A.23]

$$H(\boldsymbol{\omega}^n|E_\delta = 0) = \frac{1}{\mathbb{P}(E_\delta = 0)} \cdot \Big(H(\boldsymbol{\omega}^n|E_\delta = 1) - \mathbb{P}(E_\delta = 1) \cdot H(\boldsymbol{\omega}^n|E_\delta = 1)\Big) \tag{53}$$

$$\geq H(\boldsymbol{\omega}^n|E_\delta) - \mathbb{P}(E_\delta = 1) \cdot H(\boldsymbol{\omega}^n|E_\delta = 1) \tag{54}$$

$$\geq H(\boldsymbol{\omega}^n) - H(E_\delta) - \mathbb{P}(E_\delta = 1) \cdot H(\boldsymbol{\omega}^n|E_\delta = 1) \tag{55}$$

$$\geq H(\boldsymbol{\omega}^n) - n \cdot \varepsilon. \tag{56}$$

Equation (53) comes from the definition of the conditional entropy.

Equation (54) comes from the property $\mathbb{P}(E_\delta = 0) \leq 1$.

Equation (55) comes from the property $H(\boldsymbol{\omega}^n|E_\delta) = H(\boldsymbol{\omega}^n, E_\delta) - H(E_\delta) \geq H(\boldsymbol{\omega}^n) - H(E_\delta)$.

Equation (56) comes from the i.i.d. property of the state ω and the definition of the error event $E_\delta = 1$. Hence for all ε, there exists a $\bar{n} \in \mathbb{N}$ such that for all $n \geq \bar{n}$ we have: $H(\mathbb{P}(E_\delta = 1)) + \mathbb{P}(E_\delta = 1) \cdot \log_2 |\Omega| \leq \varepsilon$. □

References

Akyol, E., C. Langbort, and T. Başar (2017): "Information-Theoretic Approach to Strategic Communication as a Hierarchical Game," *Proceedings of the IEEE*, 105(2), 205–218.

Aumann, R., and M. Maschler (1995): *Repeated Games with Incomplete Information.* MIT Press, Cambrige, MA.

Bergemann, D., and S. Morris (2016): "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium," *American Economic Review Papers and Proceedings*, 106(5), 586–591.

——— (2017): "Information Design: a Unified Perspective," *Cowles Foundation Discussion Paper No 2075.*

Blume, A., O. J. Board, and K. Kawamura (2007): "Noisy Talk," *Theoretical Economics*, 2, 395–440.

Cover, T. M., and J. A. Thomas (2006): *Elements of information theory.* 2nd. Ed., Wiley-Interscience, New York.

Crawford, V. P., and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50(6), 1431–1451.

Cuff, P., H. Permuter, and T. Cover (2010): "Coordination Capacity," *IEEE Transactions on Information Theory*, 56(9), 4181–4206.

Cuff, P., and L. Zhao (2011): "Coordination using implicit communication," *Proceedings of the IEEE Information Theory Workshop (ITW)*, pp. 467– 471.

Gamal, A. E., and Y.-H. Kim (2011): *Network Information Theory.* Cambridge University Press.

Gelfand, S. I., and M. S. Pinsker (1980): "Coding for channel with random parameters," *Problems of Control and Information Theory*, 9(1), 19–31.

Gentzkow, M., and E. Kamenica (2014): "Costly Persuasion," *American Economic Review*, 104, 457 – 462.

Gossner, O., P. Hernández, and A. Neyman (2006): "Optimal Use of Communication Resources," *Econometrica*, 74(6), 1603–1636.

Gossner, O., and T. Tomala (2006): "Empirical distributions of beliefs under imperfect observation," *Mathematics of Operation Research*, 31(1), 13–30.

——— (2007): "Secret Correlation in Repeated Games with Imperfect Monitoring," *Mathematics of Operation Research*, 32(2), 413–424.

Gossner, O., and N. Vieille (2002): "How to play with a biased coin?," *Games and Economic Behavior*, 41(2), 206–226.

Hernández, P., and B. von Stengel (2014): "Nash codes for noisy channels," *Operations Research*, 62(6), 1221–1235.

Jackson, M. O., and H. F. Sonnenschein (2007): "Overcoming Incentive Constraints by Linking Decisions," *Econometrica*, 75(1), 241 – 257.

Kamenica, E., and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590 – 2615.

Le Treust, M. (2017): "Joint Empirical Coordination of Source and Channel," *IEEE Transactions on Information Theory*, 63(8), 5087–5114.

Le Treust, M., and M. Bloch (2016): "Empirical Coordination, State Masking and State Amplification: Core of the Decoder's Knowledge," *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*.

Le Treust, M., and T. Tomala (2016): "Information Design for Strategic Coordination of Autonomous Devices with Non-Aligned Utilities," *Proceedings of the IEEE 54th Allerton conference, Monticello, Illinois*, pp. 233–242.

Martin, D. (2017): "Strategic Pricing with Rational Inattention to Quality," *Games and Economic Behavior*, 104, 131–145.

Matejka, F., and A. McKay (2015): "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 105(1), 272–98.

Merhav, N., and S. Shamai (2007): "Information Rates Subject to State Masking," *IEEE Transactions on Information Theory*, 53(6), 2254–2261.

Neyman, A., and D. Okada (1999): "Strategic Entropy and Complexity in Repeated Games," *Games and Economic Behavior*, 29(1–2), 191–223.

——— (2000): "Repeated Games with Bounded Entropy," *Games and Economic Behavior*, 30(2), 228–247.

Rockafellar, R. (1970): *Convex Analysis*, Princeton landmarks in mathematics and physics. Princeton University Press.

Shannon, C. (1948): "A mathematical theory of communication," *Bell System Technical Journal*, 27, 379–423.

——— (1959): "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record, Part 4*, pp. 142–163.

Sims, C. (2003): "Implication of Rational Inattention," *Journal of Monetary Economics*, 50(3), 665–690.

Steiner, J., C. Stewart, and F. Matejka (2017): "Rational Inattention Dynamics: Inertia and Delay in Decision-Making," *Econometrica*, 84(2), 521–553.

Taneva, I. (2016): "Information Design," *Manuscript, School of Economics, The University of Edinburgh.*

Tsakas, E., and N. Tsakas (2017): "Noisy persuasion," *Working Paper.*

Wyner, A. D., and J. Ziv (1976): "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, 22(1), 1–11.