Effects of noise on the grammar of languages

Gerrit Bauch*

May 29, 2024

Abstract

We study a signaling game of common interest in which a stochastic noise is perturbing the communication between an informed sender and an uninformed receiver. Despite this inhibiting factor, efficient communication is possible for any kind of noise and improves upon babbling unless the noisy channel is uninformative. Endowing a compositional message space with the Hamming distance, we explore the impact of a a well-known noise channel from information theory on the grammatical structure of efficient communication. Under noise, relabeling of cells cannot be arbitrary, but has to assign distant messages to the most distant states. The more noisy the channel, the less frequent messages are used that describe states closer to the pooling action. Efficient communication under noise can be learned through the forces of evolution, but not every equilibrium is stable.

Keywords: cheap talk, noisy communication, language formation, Voronoi language

1 Introduction

In many situations our communication is flawed by errors having various origins. A person may stammer or slip their tongue, background noise may make it harder to understand or the recipient may suffer from a hearing impairment. It is thus natural to assume that any kind of communication is imperfect and prone to error. This noise is to be taken into account by both the speaker and the recipient in order to come to a proper understanding and buffer minor errors. This chapter proposes

*For inspiring discussions and comments I thank Andreas Blume, Gerhard Jäger, Frédéric Koessler, Frank Riedel and Joel Sobel. Financial support by the DFG via grant Ri 1128-9-1 is gratefully acknowledged.
gerrit.bauch@uni-bielefeld.de

a simple cheap talk game of common interest with a stochastic noisy channel. The state space is infinite, but behavioral limitations restrict the agents to a finite message space. Irrespective of the noise, communication is helpful. Indeed, efficient languages exist for arbitrary noise channels and strictly reduce the joint expected loss of the agents below the one if they did not communicate. We analyze a natural class of noisy channels that captures the idea that close words are more likely to be mistaken. Bayesian updates are possible even for events with zero probability, explaining why slight stammers or spelling mistakes do not disturb a proper understanding. We are interested in the structural rules of efficient languages, i.e., their *grammar*. The following conclusions can be made under a quadratic loss on a Euclidean space. The sender wants to induce a maximal spread of optimal receiver actions. To minimize their loss, the sender assigns words to different convex clusters of states. These clusters have sharp boundaries, reducing vagueness in their language. Being highly *compositional*, common languages have a specific structure which makes some kind of errors more likely to happen than others. Endowing the message space with a metric that respects its compositionality, we employ a noisy channel from information theory. Close words are more easily confused. If noise is present, the sender ideally labels states that lead to a high loss when confused by distinct words. If necessary, the sender reduces their frequency of using words that describe average states in favor of stressing extreme ones. Agents can learn locally efficient communication by means of evolution. However, not all equilibria are stable.

This chapter contributes to the economic literature on cheap talk games with common interest and a stochastic noise. Our benchmark model follows the one of Voronoi languages, Jäger et al. (2011), featuring a sender who is restricted to a finite message space to describe a state out of a continuum. Messages are pooled into cells, giving them a geometric structure that can be interpreted as the grammar of a language, cf. Jäger (2007), Gärdenfors (2004). Noisy communication has already been studied as a generalization of many influential paper. For instance, Blume et al. (2007) extend the seminal work of Crawford & Sobel (1982) to a noisy talk. The authors show that noise can improve welfare under conflict of interest. A similar observation was made by Myerson (1991). Another example of adding noise to communication is Jeitschko & Normann (2012) who extend the famous labor market model of Spence (1978). They find that under stochastic signaling subjects' strategies are closer to equilibrium play. Nowak & Krakauer (1999) show that evolution favors restricting to finitely many messages if signals can be misunderstood. Deriving an equilibrium concept for codes, Hernández & von Stengel (2014) bridge the gap between classical information theory and game theory. Being limited to broad terms by bounded rationality, Cremer et al. (2007) study efficient communication in which the receiver faces decoding cost that in-

crease in the breadth of the word, i.e., the number of states covered. If this loss depends on the state and the action taken rather than just the breadth, Sobel (2015) recovers convexity of the states lumped together for each message. Martel et al. (2019) argue for coarse communication to arise even in the absence of conflict or bounded rationality. Rubinstein (1989) shows that optimal strategies can differ significantly if the common knowledge assumption is disturbed by errors in communication. Communication can also be impaired if agents are ignorant or do not share the same vocabulary, Blume & Board (2013).

The remaining structure of the paper is as follows. Section 2 introduces the formal model. The best reply of the receiver is analyzed in Section 3. The sender's best response and the existence of efficient equilibria are explored in Section 4. Section 5 presents a concrete noisy channel with desirable properties. Structural implications for efficiency under a quadratic loss are given in Section 6. Section 7 shows that efficient communication can be learned over time. A brief summary is given in Section 8. The appendix contains proofs and calculations.

2 Model and notation

We adapt the setting of Voronoi languages, cf. Jäger et al. (2011). There are two players, a sender and a receiver who engage in a cheap talk game. Let $T \subsetneq \mathbb{R}^L$, $L \in \mathbb{N}_{\geq 1}$, be a convex and compact set representing states of the world. We think of an element $t \in T$ as an observation the sender has made and wants to inform the receiver about. Nature draws the state according to a common prior distribution described by an atomless measure μ_0 on T that is absolutely continuous w.r.t. the Lebesgue measure with a strictly positive and continuous density function f_0. The sender can communicate by choosing a message v to be sent to the receiver. The message space W available to the sender is finite, making perfect revelation of the state impossible. Communication is further frustrated by introducing stochastic error or noise $\varepsilon\colon \mathrm{W} \to \Delta(\mathrm{W})$ that may confound the sent message. It is possible that not the intended word v is being received, but instead an erroneous message w with probability $\varepsilon(\mathrm{w} \,|\, \mathrm{v})$. The error admits the notion of a *Markov kernel* by interpreting $\varepsilon\colon 2^W \times \mathrm{W} \to [0,1]$ where W is endowed with the discrete σ-algebra. Having observed a message w, the receiver takes an action $\alpha(\mathrm{w}) \in T$, assigning it a representative type. Following the *cooperative principle* of Grice (1975), communication is first and foremost a cooperative effort. We thus assume that the agents have a common interest to match the interpretation $\alpha(\mathrm{w})$ and the state t. To this end, we endow T with a norm $\|.\|$ and weigh the norm difference between the state and the interpretation by a strictly convex and strictly increasing function $\ell\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$. If t is the state of the world and s is the action

taken by the receiver, the loss for both parties is thus $\ell(\|t-s\|)$. Since the sender does not know the state prior to the play, agents face an expected loss according to μ_0. A (pure) strategy of the sender is a (μ_0-measurable) $\pi\colon T \to \mathrm{W}$, called a *communication device*. The receiver's (pure) strategy is given by an *interpretation (map)* $\alpha\colon \mathrm{W} \to T$. The expected joint loss the agents face is thus

$$L(\pi,\alpha) := \mathbb{E}_{\mu_0}[\mathbb{E}_{\varepsilon(.\,|\,\pi(t))}[\ell(\|t-\alpha(\mathrm{w})\|)]] \tag{1}$$

$$= \int_T \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\pi(t)) \cdot \ell(\|t-\alpha(\mathrm{w})\|)\,\mu_0(\mathrm{d}t). \tag{2}$$

A strategy profile (π,α) is referred to as a *language*. A language describes how information is articulated and processed. The sender and the receiver aspire to use a language that minimizes the loss of communication. The solution concept employed is that of perfect Bayesian Nash equilibria which we subsequently also refer to as *noise equilibrium* following Blume et al. (2007). Altering π on null sets does not change the expected loss. Furthermore, if every message is flawlessly transmitted, i.e., $\varepsilon(\mathrm{w}\,|\,\mathrm{v}) = \mathbb{1}_{\mathrm{v}}(\mathrm{w})$ is the indicator function, the expected loss and hence the analysis reduces to the one in Jäger et al. (2011). We briefly motivate our assumptions. Compactness of T and continuity of ℓ ensure integrability. Convexity of T guarantees all best replies to be in T while convexity and monotonicity of ℓ makes the receiver's best replies unique.

In this chapter, we are interested in how the agents best respond to their peer's behavior, how they can achieve efficient communication and whether or not they can learn to reach a better understanding over time.

3 Induced beliefs and the receiver's best reply

A first step towards understanding a game is to pin down the best replies of both players aiming at characterizing equilibria. As in games of common interest a strategy profile leading to an efficient outcome entails mutual best replies and is thus an equilibrium, we get a better understanding of efficient languages. To this end assume that the receiver is aware of the sender's communication device $\pi\colon T \to \mathrm{W}$. The knowledge of π allows the receiver to form expectations about the distribution of the words they are going to receive. Formally, the number

$$\lambda^{\pi}(\mathrm{w}) := \mathbb{E}_{\mu_0}[\varepsilon(\mathrm{w}\,|\,\pi(t))] = \int_T \varepsilon(\mathrm{w}\,|\,\pi(t))\,\mu_0(\mathrm{d}t). \tag{3}$$

specifies the expected probability with which the receiver will observe the word w if the sender uses the communication device π.

A message w is received either if it was actually sent ($\varepsilon(\mathrm{w} \mid \mathrm{w}) > 0$) or by error ($\varepsilon(\mathrm{w} \mid \mathrm{v}) > 0$ for some v). If w is received with positive probability $\lambda^\pi(\mathrm{w}) > 0$, the receiver can use their knowledge about π to re-asses their informational environment. Formally, they use Bayes Rule to update their prior belief. The resulting posterior μ^π_{w} is characterized by its density function

$$f^\pi_{\mathrm{w}}(t) := \frac{f_0(t) \cdot \varepsilon(\mathrm{w} \mid \pi(t))}{\lambda^\pi(\mathrm{w})}. \tag{4}$$

Knowing π and receiving the signal w, the receiver re-evaluates the chances that the sender is of some type t by applying μ^π_{w}. If $\lambda^\pi(\mathrm{w}) = 0$, we set $\mu^\pi_{\mathrm{w}} := \mu_0$ by convention.

It is worth noting that the set of induced posterior beliefs $\{\mu^\pi_{\mathrm{w}}\}_{\mathrm{w}\in\mathrm{W}}$ can be interpreted as a decomposition of the prior belief μ_0. We can interpret λ^π as a distribution over posterior beliefs with support on the finite set $\{\mu^\pi_{\mathrm{w}}\}_{\mathrm{w}\in\mathrm{W}} \subset \Delta(\mathrm{W})$. It is well-known and useful that the weighted average of the posterior beliefs sums up to the prior belief, i.e.,

$$\sum_{\mathrm{w}\in\mathrm{W}} \lambda^\pi(\mu^\pi_{\mathrm{w}}) \cdot \mu^\pi_{\mathrm{w}} = \mu_0. \tag{5}$$

More precisely, for any random variable $X \colon T \to \mathbb{R}$ we have

$$\mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[X]] = \mathbb{E}_{\mu_0}[X]. \tag{6}$$

This property is referred to as *Bayes-Plausibility*, cf. Kamenica & Gentzkow (2011). Intuitively, the sender can only induce posterior beliefs that are in expectation (w.r.t. λ^π) the prior belief μ_0 by means of a communication device. Using the notion of λ^π and $\mu^{\pi_{\mathrm{w}}}$ allows us to write the expected loss in the following equivalent way

$$\begin{aligned}
L(\pi, \alpha) &= \mathbb{E}_{\mu_0}[\mathbb{E}_{\varepsilon(.\,\mid \pi(t))}[\ell(\|t - \alpha(\mathrm{w})\|)]] && (7)\\
&= \int_T \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(w \mid \pi(t)) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\, \mu_0(\mathrm{d}t) \\
&= \sum_{\mathrm{w}\in\mathrm{W}} \lambda^\pi(\mathrm{w}) \cdot \int_T \lambda^\pi(\mathrm{w})^{-1} \cdot \varepsilon(w \mid \pi(t)) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\, \mu_0(\mathrm{d}t) \\
&= \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\ell(\|t - \alpha(\mathrm{w})\|)]]. && (8)
\end{aligned}$$

Expression (7) describes the expected loss as a weighted sum of the deficits that occur due to the error for each realized type t. Instead, expression (8) aggregates the expected losses under each of the posteriors and weights them according to the

probability with which the posterior is induced. The latter expression proves useful in characterizing the receiver's loss minimizing interpretation for any (induced) posterior belief.

Lemma 3.1. *Having any belief $\mu \in \Delta(T)$ with positive density f, the receiver minimizes their expected loss by choosing the unique response*

$$\hat{s} \in \arg\min_{s \in T} \mathbb{E}_{\mu}[\ell(\|t - s\|)]. \tag{9}$$

Proof. All proofs are delegated to the appendix. □

Henceforth, denote by $\hat{\alpha}(\mu)$ the unique minimizer of a receiver holding the belief μ. We can thus define $\hat{\alpha}(\mathrm{w}) := \hat{\alpha}(\mu^{\pi}_{\mathrm{w}})$ if π is understood and refer to $\hat{\alpha}$ as the unique best reply of the receiver. There are thus two ways of thinking of the receiver's best reply $\hat{\alpha}(\mathrm{w})$ to a word. Firstly, their response is based on the heard word and its implicit meaning and interpretation. Secondly, the received word changes the listener's belief about the state of the world leading to their response. Furthermore, having a unique solution to the minimization problem, the receiver has no incentive to play a proper mixed strategy in equilibrium.

In a setting of common interest, we expect communication to serve the purpose of reducing misunderstandings, i.e., the expected loss. In particular, communication should foster understanding and result in a lower loss compared to the situation where individuals cannot exchange information. In the latter situation, the receiver does not receive a message from the sender and thus cannot update their belief. Sticking to the information at hand, the receiver seeks to minimize the loss given μ_0. We call $\alpha_{\mathrm{pool}} := \alpha(\mu_0) = \arg\min_{s \in T} \mathbb{E}_{\mu_0}[\ell(\|t - s\|)]$ the *pooling action*. If the receiver applies the pooling action, the *pooling loss* $L_{\mathrm{pool}} := \mathbb{E}_{\mu_0}[\ell(\|t - \alpha_{\mathrm{pool}}\|)]$ realizes. The following result characterizes precisely those communication devices that improve upon the pooling loss if the receiver plays a best response.

Proposition 3.2. *$L(\pi, \hat{\alpha}) \leq L_{\mathrm{pool}}$ for any communication device π. The inequality is strict if and only if there is a word $\mathrm{w} \in \mathrm{W}$ with $\lambda^{\pi}(\mathrm{w}) > 0$ and $\hat{\alpha}(\mathrm{w}) \neq \alpha_{\mathrm{pool}}$.*

As long as the receiver knows the communication device π, the presence of signals cannot be detrimental to the communication. If furthermore there is a word sent with positive probability that does want the receiver to not take the pooling action, communication strictly reduces the expected loss and vice versa. The following corollary states two readily verifiable conditions under which communication does not improve over the pooling loss.

Corollary 3.3. *If π is constant or if $\mathrm{v} \mapsto \varepsilon(\cdot \mid \mathrm{v})$ is constant, then $\mu^{\pi}_{\mathrm{w}} = \mu_0$ for all $\mathrm{w} \in \mathrm{W}$. Consequently, $\hat{\alpha} \equiv \alpha_{\mathrm{pool}}$ and thus $L(\pi, \hat{\alpha}) = L_{\mathrm{pool}}$.*

Although easy and intuitive, Corollary 3.3 demonstrates that there are two sources that can lead to non-beneficial communication. Firstly, the communication device may not be meaningful, i.e., received messages never offer additional information. Secondly, if the error channel is *uninformative*, i.e., does not convey any information: If everything is equally likely to be received, no matter what has been sent, the receiver cannot infer any additional information. While the latter problem is to be considered an exogenous problem of the environment, the first one lies within the power of the sender. On a warning notice, we stress that not all non-constant communication devices improve upon the pooling loss, not even if it changes the informational environment of the receiver.

Example 3.4. Let $T = [-\frac{1}{2}, \frac{1}{2}]$ with uniform prior μ_0. Assume the sender wants to inform the receiver about whether the state is close to the center $\alpha_{\text{pool}} = 0$ of the interval or not, using the words C and NC by following the communication device depicted in Figure 1. Assume the noisy channel confuses the two messages with probability $p < \frac{1}{2}$, i.e, if message C (NC) is sent, message NC (C) is received with probability p. If the receiver gets the message C, their posterior is given by its density

$$f_C^{\pi}(t) = 2 \cdot \begin{cases} 1-p & , t \in [-\frac{1}{4}, \frac{1}{4}], \\ p & , \text{otherwise} \end{cases}. \tag{10}$$

The noise channel is thus informative and makes the receiver believe that the state is more likely to lie in the center than outside of it, changing their posterior belief after the arrival of new information. However, if the expected loss is quadratic, i.e., $\ell \circ \|t - s\| = (t - s)^2$, the receiver best responds to both messages with the pooling action $\alpha_{\text{pool}} = 0$ which constitutes a noise equilibrium. The expected loss is L_{pool} even though posteriors other than the prior belief are induced. Consequently, merely changing the receiver's beliefs is not enough to reduce the expected loss below L_{pool}. One needs to make the receiver do different things for different messages. The situation is illustrated in Figure 1.

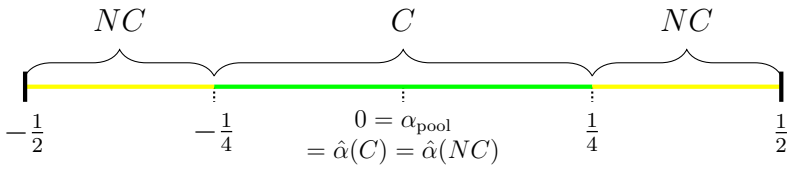

Figure 1: The language from Example 3.4 with two words that induces different posterior beliefs, but only leads to the constant pooling action. Changing beliefs is not sufficient to reduce expected loss below the pooling one.

In cheap talk games, there always are equilibria that attain the pooling loss. Assume the sender is using a constant communication device, thus provoking the constant pooling action on the receiver's part, resulting in the pooling loss. Since the interpretation map of the receiver is constant and thus independent of any message received, the sender does not have an incentive to deviate from their constant communication device either, establishing a so-called *babbling equilibrium*[1]

4 Efficient languages

In the previous section we have established that communication can only make the agents better off in equilibrium despite the inhibiting noise. However, the expected loss cannot be reduced to zero since full separation of states is impossible due to having more states than messages. That raises the questions of what is the most efficient way of communicating and (how) can it be achieved?

While the agents cannot re-negotiate their strategies during the play, say, after the state was revealed to the sender, we can think of the two players meeting before the game. Using a meta-language they discuss their strategies prior to the play in search of an efficient communication, cf. Jäger et al. (2011). A language (π, α) is called *efficient* if it minimizes the expected loss $L(\pi, \alpha)$ over all possible languages. In the presented setting, this requires both agents to minimize the occurring loss over their own action sets, i.e., to play a best response each. Any efficient language is thus a noise equilibrium. Unsurprisingly, noise equilibria may improve upon the pooling loss without being efficient, see Section 6.2. We now give a simple example of an efficient language.

Example 4.1. Let $T = [-\frac{1}{2}, \frac{1}{2}]$ with uniform prior μ_0. Assume the sender has two available messages, i.e., $\mathrm{W} = \{A, B\}$. The noisy channel confounds a sent message with probability p. Assume a quadratic loss, i.e., agents lose $(t-s)^2$ if the state is t and action s is taken. Figure 2 depicts an efficient language for every $p \leq \frac{1}{2}$. The sender uses their messages efficiently by cutting the interval into a left (sending A) and a right one (sending B). That way, the best responses of the receiver move away from the pooling action. If p increases, the best responses come closer to the pooling one as long as $p \leq \frac{1}{2}$. At $p = \frac{1}{2}$ the noisy channel becomes uninformative.

[1]The name "babbling equilibrium" is best explained with mixed strategies. Assume the sender randomizes according to a fixed distribution G over messages for all t. As the state of the world and the randomly picked message are completely independent from one another, the communication device is dubbed "babbling". The received messages on the receiver's end will thus also be independent of the state. Gaining no insights from any message, they optimally reply with the pooling action, in turn making the sender indifferent between all their communication devices.

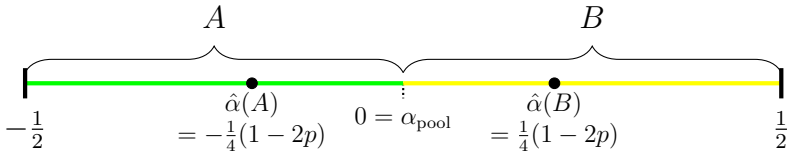

Figure 2: An efficient language in the setting of Example 4.1. The sender uses different words trying to reveal on which side of the interval the true state is situated. The closer the probability p of confusing the two words goes to $\frac{1}{2}$, the closer the optimal receiver responses become to the pooling action.

Efficient languages, such as the one in Example 4.1, can be explicitly computed by characterizing and internalizing a best response of the sender for any interpretation map. In contrast to the receiver's best response, the sender's best response is not unique. For instance, they can always perturb their strategy on a null set without altering the expected loss. More importantly for an economic interpretation however, the sender can be indifferent between sending two or more different words at their interim stage. The reason for this is that the corresponding actions taken by the receiver in response amount to the same loss. More precisely, fix any interpretation $\alpha\colon \mathrm{W} \to T$ of the receiver. Focusing on the sender's interim behavior[2] a type t-sender is indifferent between sending any word $\mathrm{v} \in \mathrm{W}$ out of the (non-empty) set

$$\arg\min_{\mathrm{v}'} \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w} \mid \mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|). \tag{11}$$

Applying reasonable choices in the presence of multiple minimizers (see the proof of Theorem 4.2 for details) we can derive a measurable partition $C^\alpha = \{C^\alpha_{\mathrm{v}}\}_{\mathrm{v}\in\mathrm{W}}$ of T, where each C^α_{v} consists only of types where v is a minimizing response of the sender given α, i.e., an element of (11). Any such partition defines a best reply communication device by letting $\pi^{C^\alpha}(t) = \mathrm{v}$ if and only if $t \in C^\alpha_{\mathrm{v}}$. It is worth mentioning that the set of types where different words can serve as a minimizing interim response, may not be a null set[3]. Hence, the two best replies constructed as explained above may differ perceptibly. Fortunately, the indifference sets have measure zero for the Euclidean norm as long as interpretations differ, cf. Proposition 6.4. Irrespective of the particular loss function or noise employed, we positively answer the question of existence of efficient languages.

Theorem 4.2. *Efficient languages exist.*

[2]Since the marginal distributions have full support, ex-ante and interim behavior coincide.

[3]For the maximum norm, the set of types that are indifferent between sending two different word might have positive measure, see, e.g., Figure 1 in Jäger et al. (2011).

Even if agents were unable to find or agree on an efficient equilibrium play, they should not use mixed strategies, i.e., randomizing their behavior. Intuitively, any randomization blurs proper understanding under common interest, thereby increasing the expected loss. If the receiver mixed between responses to a message w, the loss in communication would increase for states close to one of these receiver actions. Likewise, if the sender mixed between words that do not induce the same expected loss, the overall expected loss would increase. We give a formal explanation for why introducing randomness into the signaling or interpretation procedure makes coordination harder to achieve. Firstly note that the receiver always strictly favors a pure strategy $\alpha\colon \mathrm{W} \to T$ over a mixed one $\tau\colon \mathrm{W} \to \Delta(T)$ a posteriori by Lemma 3.1. Secondly, let $\sigma\colon T \to \Delta(\mathrm{W})$ be a mixed strategy of the sender, specifying a probability $\sigma(\mathrm{w} \mid t)$ of playing $\mathrm{w} \in \mathrm{W}$ if they are of type $t \in T$. The expected loss is thus given by

$$L(\sigma, \alpha) = \int_T \sum_{\mathrm{v} \in \mathrm{W}} \sigma(\mathrm{v} \mid t) \cdot \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\ \mu_0(\mathrm{d}t). \tag{12}$$

Sending any pure $\mathrm{v} \in \arg\min_{\mathrm{v}'} \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|)$ if the state is t is weakly decreasing the expected loss, even strictly if $\sigma(. \mid t)$ assigns a positive measure to any $\widetilde{\mathrm{v}}$ not in (11). We thus have proven the following proposition.

Proposition 4.3. *For any mixed language (σ, τ) there is a pure language (π, α) with weakly smaller loss $L(\pi, \alpha) \leq L(\sigma, \tau)$.*
The inequality can be made strict if τ is not a pure strategy or the support of $\sigma(. \mid t)$ contains a word not in (11) *for positive mass of states t.*

5 A metric-dependent noise channel

So far, the stochastic noise could be arbitrarily chosen. In the following we study a noise that perturbs communication in a natural way. Messages that are similar are more easily mistaken than ones that are dissimilar from one another. For instance, the English words "flank" and "plank" can be more easily misunderstood than "flank" and "igloo"[4]. In order to formalize this idea, we have to to endow the message space with sufficient structure and a measure of how similar its messages are. Our approach follows the models used in information theory that are inspired by the following theory in linguistics. Human languages are *compositional*[5], i.e.,

[4]The words taken for this example are inspired by the solutions to the *Wordle web-based word game* from July 20, May 17, May 23 2023, respectively.

[5]While implicitly being used in the seminal work of Frege (1892), cf. Miller (2007) for a modern reference, and even sometimes being called *Frege's principle of compositionality*, Frege arguably never explicitly formalized the concept himself, cf. Pelletier (2001). Although com-

they are constructed from smaller building blocks. For instance, text is made up of sentences, sentences consist of words, words are sequences of letters, syllables, phonems. The agents have a set $\mathcal{A}$ with at least two elements at hand from which they can construct a sequence of length n to form their message space $\mathrm{W} = \mathcal{A}^n$. For easier reference, we call $\mathcal{A}$ an *alphabet* with elements that could represent *letters*, syllables or phonems. A sequence of letters $\mathrm{w} \in \mathrm{W}$ is a *word*. Restricting to a fixed length of words is common in information theory and has proven a solid baseline for coding theory Roth (2006). As a measure of distance between two words we use the *Hamming distance*, defined by

$$d\colon \mathrm{W} \times \mathrm{W} \to \mathbb{N}_0, ((\mathrm{w}_k)_k, (\mathrm{v}_k)_k) \mapsto \#\{k \in \{1, \dots, n\} \mid \mathrm{w}_k \neq \mathrm{v}_k\}. \tag{13}$$

Words are considered farther away from one another the more letters in the order of appearance differ. The Hamming distance was first studied by Hamming (1950) and is in fact a metric on W. It plays a crucial role in many applied fields related to information theory, especially coding theory and telecommunication, cf. Roth (2006).

The noisy channel we are now introducing behaves well with the Hamming distance and is best understood defining it on letters first. Let $\widetilde{\varepsilon}\colon \mathcal{A} \to \Delta(\mathcal{A})$ denote the function

$$a \mapsto \widetilde{\varepsilon}(. \mid a), \quad \widetilde{\varepsilon}(b \mid a) := \begin{cases} 1-p & , b = a, \\ \frac{p}{\#A-1} & , b \neq a \end{cases}, \tag{14}$$

where the exogenous parameter $p \in [0, 1]$ is the *crossover probability*, i.e., the probability of wrongly transmitting one intended letter a. In case of an error each of the other $\#\mathcal{A} - 1$ symbols is assumed to be equally likely received. It is a well-known noise that is used to model error transitions in telecommunication, data storage, but also finds application in DNA heritage of cell-divisions, cf. MacKay (2002) and Cover & Thomas (2006)).

Depending on the concrete scenario, errors can feature different levels of correlation. We follow the branch of literature that assumes independent occurrences of errors for each letter. To this end we can extend the $\#\mathcal{A}$-ary symmetric error channel with crossover probability p to $\mathrm{W} = \mathcal{A}^n$ by gluing n independent copies of $\widetilde{\varepsilon}$ together. The result is called *$\#\mathcal{A}$-ary symmetric channel of length n with crossover probability p*. Using the Hamming distance, the transition probabilities are given by

$$\varepsilon(\mathrm{w} \mid \mathrm{v}) := (1-p)^{n - d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{\#\mathcal{A} - 1}\right)^{d(\mathrm{w},\mathrm{v})}. \tag{15}$$

pelling, deriving meaning from compositionality alone is a debatable practice among linguists and challenged by other concepts, such as *contextuality*, cf. Szabó (2022).

The probability $\varepsilon(\mathrm{w} \mid \mathrm{v})$ that w is received if v is sent only depends on the Hamming distance $d(\mathrm{w}, \mathrm{v})$ and the crossover probability p. Especially, $\varepsilon(\mathrm{w} \mid \mathrm{v}) = \varepsilon(\mathrm{v} \mid \mathrm{w})$. For fixed $\mathrm{v} \in \mathrm{W}$ and $d \in \{0, \dots, n\}$ there are precisely $\binom{n}{d} \cdot (\#\mathcal{A} - 1)^d$ different words w with $d = d(\mathrm{w}, \mathrm{v})$ in W. Hence, ε follows a binomial distribution on the set of families $\{\mathrm{w} \in \mathrm{W} \mid d(\mathrm{w}, \mathrm{v}) = d\}_{d=0}^{n}$. Being mainly interested in the transition probability from v to a particular w rather than to such a family of words, ε is itself the natural probability distribution to consider. For convenience, we write $m := \#\mathcal{A} - 1$ and $\widetilde{p} := \frac{p}{(1-p)m}$ and can rewrite $\varepsilon(\mathrm{w} \mid \mathrm{v}) = (1-p)^n \cdot \widetilde{p}^{d(\mathrm{w},\mathrm{v})}$ which is often convenient. The following immediate statements capture the nice interplay between the noise channel ε and the Hamming-distance for different crossover probabilities p.

Remark 5.1. Suppose $\mathrm{v} \in \mathrm{W}$ is the word sent.

(i) If $p = \widetilde{p} = 0$, there is no noise, i.e., v is received with probability 1.

(ii) If $0 < p < \frac{m}{m+1}$, i.e., $0 < \widetilde{p} < 1$, $\varepsilon(\mathrm{w} \mid \mathrm{v})$ is decreasing in $d(\mathrm{w}, \mathrm{v})$. Especially it is most likely to receive v.

(iii) If $p = \frac{m}{m+1}$, i.e., $\widetilde{p} = 1$, ε is uninformative.

(iv) If $\frac{m}{m+1} < p < 1$, i.e., $1 < \widetilde{p} < \infty$, $\varepsilon(\mathrm{w} \mid \mathrm{v})$ is increasing in $d(\mathrm{w}, \mathrm{v})$. Especially, words with maximum distance $d(\mathrm{w}, \mathrm{v}) = n$ are most likely received.

(v) If $p = 1$, i.e., $\widetilde{p} = \infty$, only words with maximum distance $d(\mathrm{w}, \mathrm{v}) = n$ are received.

For $p = 0$ there is no noise in communication and our model coincides with Jäger et al. (2011). If $0 < p < \frac{m}{m+1}$ the $\#\mathcal{A}$-ary symmetric channel of length n captures our intuition that words that are close in the Hamming distance are more likely to be confounded. This property is most pronounced for $p \approx 0$ and vanishes completely at the *uninformativeness bound* $p = \frac{m}{m+1}$. In fact, we find a continuous and monotonic loss in the amount of information that can be transmitted through the noisy channel for increasing p measured by *Shannon entropy*, cf. Section 5.2. At the uninformativeness bound, $\varepsilon(. \mid \mathrm{v})$ is the constant uniform distribution on W for all v, making the sent and received word independent, recall also Corollary 3.3.

Interestingly, for $\frac{m}{m+1} < p$ informative communication can take place again, even less efficient than before. The receiver now suspects the received word to stem most likely from one among those having the maximal distance to the it. Since this set contains more than one word if $m > 1$, the receiver cannot pin down a single most probable word and is thus less confident about their guess of

the originally sent word.[6] Our discussion of entropy in Section 5.2 quantifies this observation.

5.1 Bayesian updates and limit cases

In many realistic scenarios, errors in communication may be present, but not too abundant. In the following, we will thus focus on the case $0 < p < \frac{m}{m+1}$ for which errors occur with positive probability, the noisy channel is not uninformative and has the property that close words in the Hamming distance are more easily misunderstood.[7] In the following, we will derive the receiver's posterior beliefs under ε. We also study how the noisy channel behaves in the limit cases if the crossover probability p goes to zero or the uninformativeness bound and find suitable updates even for events of probability zero.

Since $\varepsilon(\mathrm{w} \,|\, \mathrm{v}) > 0$ for all w, v the receiver can always use Bayes rule to update their prior belief. A short calculation yields posteriors beliefs with densities of the form

$$f_{\mathrm{w}}^{\pi}(t) = f_0(t) \cdot \left(\int_T \widetilde{p}^{\,d(\mathrm{w},\pi(t'))-d(\mathrm{w},\pi(t))} \; \mu_0(\mathrm{d}t') \right)^{-1} \tag{16}$$

for a received word w. Note that the integrand is continuous in $p \in [0, 1)$. The employed error channel allows us to study the limit cases for the receivers posterior belief for $p \to 0$ and $p \to \frac{m}{m+1}$.

Proposition 5.2. *Let π be known to the receiver who observes $\mathrm{w} \in \mathrm{W}$. Then the following properties hold.*

(i) $\lim\limits_{p \to \frac{m}{m+1}} f_{\mathrm{w}}^{\pi}(t) = f_0(t)$.

(ii) *Let* $d^* := \min\{d \in \{0, \dots, n\} \,|\, \mu_0(\{t' \,|\, d(\mathrm{w}, \pi(t')) = d\}) > 0\}$.

(a) *If* $d^* < d(\mathrm{w}, \pi(t))$ *then* $\lim\limits_{p \to 0} f_{\mathrm{w}}^{\pi}(t) = 0$.

(b) *If* $d^* = d(\mathrm{w}, \pi(t))$ *then* $\lim\limits_{p \to 0} f_{\mathrm{w}}^{\pi}(t) = f_0(t) \cdot \mu_0(\{t' \,|\, d(\mathrm{w}, \pi(t')) = d(\mathrm{w}, \pi(t))\})^{-1}$.

(c) *If* $d^* > d(\mathrm{w}, \pi(t))$ *the limit of the posterior belief for* $p \to 0$ *is not defined.*

[6] For binary channels, i.e., $m = 1$, there is exactly one word with maximal distance n to a fixed v. The roles of the two letters (bits) simply switch and entropy is symmetric around the uninformativeness bound $p = \frac{1}{2}$.

[7] We restrict to the intuitive case $p < \frac{m}{m+1}$ where errors not abound. However, an analysis for $p \to 1$ can be deducted with analogous results.

The first statement simply says that there is a smooth transition of the beliefs towards the common prior if the error channel gets uninformative. The second part deals with the behavior of the posteriors if the crossover probability goes to zero. In the presented continuous case, it is important to keep track of null sets but can still be intuitively explained. To start with, when receiving w the receiver determines the closest words to w that are sent the sender with positive probability. Let the according distance be d^* and let receiver contemplate about the state of the world being t.
If $d^* < d(\mathrm{w}, \pi(t))$, the likelihood that the state is t goes to zero if $p \to 0$ since there are events with positive probability in which words with $d(\mathrm{w}, v) = d^*$ are sent. If $d^* = d(\mathrm{w}, \pi(t))$, then there is no event with positive probability in which words strictly closer to w are sent than $\pi(t)$. Taking the limit $p \to 0$, the receiver will discard states in which words even farther away from w than d^* are sent. Consequently, the receiver concludes that the true state t' fulfills $d^* = d(\mathrm{w}, \pi(t'))$. The posterior is consequently given as stated in Proposition 5.2 (ii)(b). Remarkably, this is true even if w is not expected to be sent with positive probability and the Bayesian update for $p = 0$ is undefined.[8] Our intuitive interpretation is as follows. A receiver who hears the word "orunge" concludes that "orange" must have been the word sent even if they believe the error to be arbitrarily small. However, the case $d^* > d(\mathrm{w}, \pi(t))$ makes it impossible to apply Bayes rule as the receiver neither expects w to be sent with positive probability, nor do they believe that another word with distance $d(\mathrm{w}, \pi(t))$ has been sent.[9]

5.2 Entropy

Ever since the seminal work of Shannon (1948), the most important measure for the "amount of information" a communication channel can transport is given by *(Shannon) entropy*. In the following we determine the entropy of the q-ary symmetric channel of length n and crossover probability p. Intuitively, the more noise the harder it is for the receiver to confidently decode an observed message. We confirm the natural guess that the noise is maximal at the uninformativeness bound $p = \frac{m}{m+1}$ and strictly monotonically increasing for both $p \nearrow \frac{m}{m+1}$ and $p \searrow \frac{m}{m+1}$. While there is no noise for $p = 0$, information cannot become perfect for $p = 1$ unless for a binary alphabet, i.e., $m = 1$.

Formally, the entropy of a discrete probability measure P on a finite set X is

[8] Ortoleva (2012) provide a theoretical model to update given null events.

[9] If one considers the analogue model with a finite state space and the prior belief has full support, the case (c) disappears. Especially, the limit for $p \to 0$ is always defined.

defined as

$$H(P) = -\sum_{x \in X} P(x) \cdot \log(P(x)), \tag{17}$$

with the convention $0 = P(x) \cdot \log(P(x))$ if $P(x) = 0$. The base choice of the logarithm is a question of normalization and usually chosen to be $\#X$ which is convenient for our setting. The value $H(P)$ is interpreted as the average of the *information content* $-\log(P(x))$ and attributed to describe how surprising the observation of an element x is given its probability $P(x)$. Some important properties of the entropy function H include non-negativity, strict concavity in the probability distribution P with the maximum being attained at the uniform distribution on X and symmetry in the order of the elements.

Turning to our metric-dependent error channel, $\varepsilon(.\,|\,\mathrm{v})$ defines a probability distributions on the set W for any fixed sent word v and any error probability $p \in [0, 1]$ (that is suppressed by the notation). As the choice of v only leads to a permutation of the probabilities across W and by symmetry of H, $H(\varepsilon(.\,|\,\mathrm{v}))$ does not depend on v and $H_\varepsilon(p) = H(\varepsilon(.\,|\,\mathrm{v}))$ is well-defined. The following proposition summarizes characterizing properties of $H_\varepsilon(p)$.

Proposition 5.3. *The entropy of the $\#\mathcal{A}$-ary symmetric error channel of length n with crossover probability p is*

$$H_\varepsilon(p) = -n \cdot (p \log(p) + (1-p)\log(1-p)) + n \cdot p \log(m). \tag{18}$$

It is a concave in $p \in [0, 1]$ with its unique maximum being attained at $p = \frac{m}{m+1}$ with value $\log(\#\mathrm{W})$. *Moreover, $H_\varepsilon(0) = 0$ and $H_\varepsilon(1) = n \log(m)$.*

The proposition gives a quantitative view on the observations made in Remark 5.1. Choosing the base #W for the logarithm, we can interpret $H_\varepsilon(p)$ as the percentage of noise of the considered channel. For $p = \frac{m}{m+1}$ entropy is maximal and equal to 1, which we interpret as each message being equally likely received. The error channel thus conveys no information. For $p = 0$ entropy is zero, showing that there is no noise. Each piece of information is perfectly transmitted and can be correctly decoded. If $p = 1$ entropy is $\log(m)/\log(m+1)$, telling us how much information is lost. For $m = 1$, i.e., a binary alphabet, this expression is again zero which makes sense since the roles of the letters simply swap, see also footnote 6 on page 13. In contrast, if $m > 1$, although information can be recovered, there will be noise left nevertheless. Received words cannot be unambiguously decoded. In between those extreme cases, due to concavity, we have a monotonic increase of entropy towards the uninformativeness bound $p = \frac{m}{m+1}$ from both sides. For $p \nearrow$ communication gets hindered more and more on $[0, \frac{m}{m+1}]$, while afterwards communication gets facilitated again.

Entropy is the quantitative measure for the amount of information that can be achieved through communication under noise. The following example illustrates the relation between an increasing noise in terms of entropy and the corresponding expected loss for a fixed communication device with a non-binary alphabet.

Example 5.4. Let $T = [-\frac{1}{2}, \frac{1}{2}]$ with uniform μ_0. Let $\mathrm{W} = \{L, M, R\}$ and consider $\pi\colon T \to \mathrm{W}, \pi([-\frac{1}{2}, 0]) = L, \pi((0, \frac{1}{2}]) = R$. Figure 3 depicts the best responses of the receiver, the associated expected loss and the entropy of ε. At the uninformativeness bound $p = \frac{2}{3}$ actions are pooling and expected loss and entropy are maximal. The receiver can perfectly decode L and R as R and L respectively if $p = 1$. However, for $p = 1$, they now also received the erroneous message M with positive probability, but cannot perfectly recover the originally sent message in that case. Consequently, communication under $p = 1$ is worse than for $p = 0$. In numerical terms, we find that $L(1)/L(2/3) = 0.625 \approx 0.63 \approx H_\varepsilon(1)$, thus entropy roughly captures the expected loss.

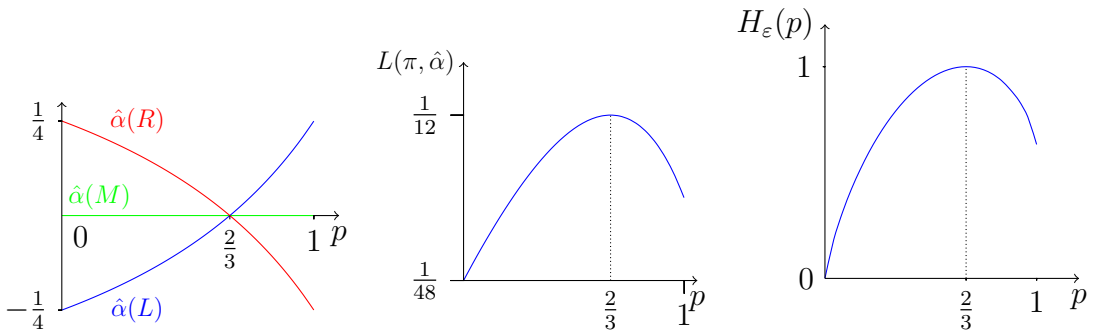

Figure 3: Optimal responses, expected loss and entropy for the communication device in Example 5.4. For three letters the uninformativeness bound is $p = \frac{2}{3}$. Actions coincide with the pooling action $\alpha_{\text{pool}} = 0$, expected loss and entropy are maximal. For $p = 1$, optimal receiver responses to the messages used with positive probability can be recovered up to re-labeling. Observing L can be perfectly decoded to have been the original message R. However, M will now be received with positive probability, leading to a loss in information in comparison to $p = 0$.

6 The grammar of efficient languages

The most prominent loss function in the study of communication games is the quadratic loss. It is used as the prime example in the field of cheap talk games ever

since the seminal of Crawford & Sobel (1982), allowing for analytical tractability. We analyze the effects of noise on the geometric structure of languages under a quadratic loss. In the field of linguistics, the set of structural rules of a language is more prominently known as its *grammar*. Seeking for efficiency, we replicate the following grammatical patterns of the agents' best responses under noise. A speaker distributes their words in a way that maximizes a certain spread of their induced interpretations. The range of meanings that a word represents is convex. Words must not be vague. Noise necessitates to reserve distant words to states that result in a huge loss if confused. The more noise the clearer the speaker will stress words that should not be confused, reducing or even forsaking the use of words that are easily confused.

For now, let ε be an arbitrary noisy channel. Throughout this section we consider a quadratic loss, i.e, we let $\|.\| = \|.\|_2$ be the Euclidean norm induced by the scalar product $\langle .\,,\, .\rangle$ and $\ell(x) = x^2$. Under this assumption, the best response of the receiver can be written down explicitly.

Lemma 6.1. *Having any (posterior) belief μ, with continuous density $f > 0$, the receiver's unique best interpretation is given by*

$$\hat{\alpha}(\mu) = \mathbb{E}_{\mu}[t]. \tag{19}$$

Given a communication device π and the induced best responses, the expected loss is

$$L(\pi, \hat{\alpha}) = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\left\|t - \mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t']\right\|_2^2]] = \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2]. \tag{20}$$

Furthermore, the receiver on average plays the pooling action

$$\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \alpha_{\mathrm{pool}}. \tag{21}$$

Lemma 6.1 proves useful for pinning down analytical properties of efficient languages. As a first application we show that profitable communication is always possible, as long as the error channel is not uninformative.

Proposition 6.2. *If ε is not uninformative there is (π, α) with $L(\pi, \alpha) < L_{\mathrm{pool}}$.*

In the following subsections, we want to elicit grammatical, i.e., geometric and structural, properties of efficient communication.

6.1 Convexity of meaning

Employing a quadratic loss provides a new tractable view on what makes up an efficient language. By expression (20), the expected loss is the difference between a term that depends only on the state space and its measure and a weighted sum of

the square norms of the induced interpretations. Minimizing the loss of a language is thus equivalent to maximizing the spread of induced squared interpretations. The expected loss is the difference of the pooling loss and the (non-negative) difference of this spread and the square of the pooling action, precisely

$$L(\pi, \hat{\alpha}) = L_{\text{pool}} - \left(\mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2] - \|\alpha_{\text{pool}}\|_2^2\right). \tag{22}$$

Since $\alpha_{\text{pool}} = \sum_{\mathrm{w}} \lambda^\pi(\mathrm{w})\hat{\alpha}(\mathrm{w})$, an efficient language is a decomposition of α_{pool} that maximizes the weighted sum of induced square norm interpretations among all communication devices. Summarized as a structural consequence, efficient languages are *as separating as possible*.

It is worthwhile to discuss this insight by revisiting Example 3.4 and Example 4.1. Even though the communication device in Example 3.4 results in different posterior believes, they do not accomplish a spread of induced actions. In contrast, the best responses in Example 4.1 achieve the maximal spread that the noisy channel allows. An apparent difference of the two communication devices is that the cells in the efficient example are convex, while this is not always the case for the inferior one. Indeed, non-convexity of cells is an indicator that the sender does not play a best response to the receivers actions under a quadratic loss, as the next proposition proves.

Proposition 6.3. *The set of states for which sending* $\mathrm{v} \in \mathrm{W}$ *is a (the unique) best reply given* α *is a closed (open) convex set.*

Proposition 6.3 implies further structural rules on efficient languages. If the sender plays a best reply to the receiver's interpretation map α, an induced cell corresponds to the indicative meaning of the corresponding word. Each cell thus forms a category in the sense of Lewis (1969). Our result thus generalizes the linguistic idea that *(simple) words have convex categories* to communication under noise, cf. Gärdenfors (2004), Jäger (2007) and Jäger et al. (2011).

Starting at which temperature does on say that it is "hot"? Certainly, most people agree that 100°F (38°C) is hot, but there is hardly a precise threshold separating "hot" from "not hot" temperatures. Words like "hot", "tall" or "many" are *vague*, they lack a precisely definition, cf. Sorensen (2023) and Lipman (2009).[10] Under a quadratic loss and the metric-dependent noise from Section 5, the sender does not use vague words. If they say "hot" this has a crisp meaning. Formally, if two words v, v' are no *synonyms*, i.e., $\alpha(\mathrm{v}) \neq \alpha(\mathrm{v}')$, sending one of the two words is almost surely strictly preferred. As a result, the induced cells have sharp boundaries.

[10]Vagueness is often associated with the so-called *sorites paradox*: When makes taking away grains of sand from a heap stop us referring to the remaining sand as a heap?

Proposition 6.4. *Let ε by the q-ary symmetric channel of length n with crossover probability p and α an interpretation map of the receiver.*
If $\alpha(\mathrm{v}) \neq \alpha(\mathrm{v}')$, the set of states for which the sender is indifferent between sending v *and* v' *is a null set for all but at most n values of $p \in [0,1]$.*
If in addition $p = 0$, the sender is almost surely never indifferent, while for the uninformativeness bound $p = \frac{m}{m+1}$ they always are.

6.2 The shape of cells and their labeling

In this section, we study an extensive example of a two-dimensional state space with words of length two. Four classes of noise equilibria are given, each characterized by two different properties. Cells shapes can be either quadratic or triangular and the labeling can respect far away states by using distinct words or not. Our results suggest that efficient communication under noise needs to be clear in distinguishing distinct states by words that are not easily confused and use a cell structure minimizing the length of the boundary between cells.

Let $T = [-\frac{1}{2}, \frac{1}{2}]^2$ be endowed with the uniform distribution $\mu_0 \sim \mathcal{U}(T)$, especially $\alpha_{\text{pool}} = (0,0)$. The sender uses an alphabet with two letters and can send words of length two $\mathrm{W} = \{A, B\}$. Communication is frustrated by a binary symmetric channel of length two and crossover probability p. Assume $p < \frac{1}{2}$ so that the error function is not uninformative and it is less likely to confound words that are farther away from one another. Consider the communication devices depicted in Figure 4 which constitute in fact noise equilibria. A detailed derivation of all formulae in Table 1 and explanation of the claims is given in appendix Section 9.3.

Languages 1a and 1b employ quadratic tessellations, while 2a and 2b use triangular shaped cells. Note that neighboring cells in 1a and 2a are associated with words that are close to one another (Hamming distance 1) and words that are less easily confused are used for cells that are far away from one another. In contrast, each cell in 1b and 2b has a neighbor the assigned word of which has distance 1 and one with distance 2.

Figure 4 plots the loci of best responses of the receiver are drawn from $p = 0$ to the uninformativeness bound $p = \frac{1}{2}$. The precise formulae are summarized in Table 1. As is clear from Proposition 5.2, the receiver's interpretations start at the center of each cell and continuously move to the pooling action. Interestingly, the convergence to the pooling action differs significantly. While the loci in 1a and 2a are straight lines, the ones for 1b and 2b bend towards the cell that uses the word with distance 1 to the own one. The intuition for this is straightforward: Look w.l.o.g. at $\hat{\alpha}(AA)$: Within the type-a languages, the cells of AB and BA absorb the same amount of mistakes from and towards AA, linearly in p. The word BB does so quadratically but also point-symmetrically w.r.t. the pooling

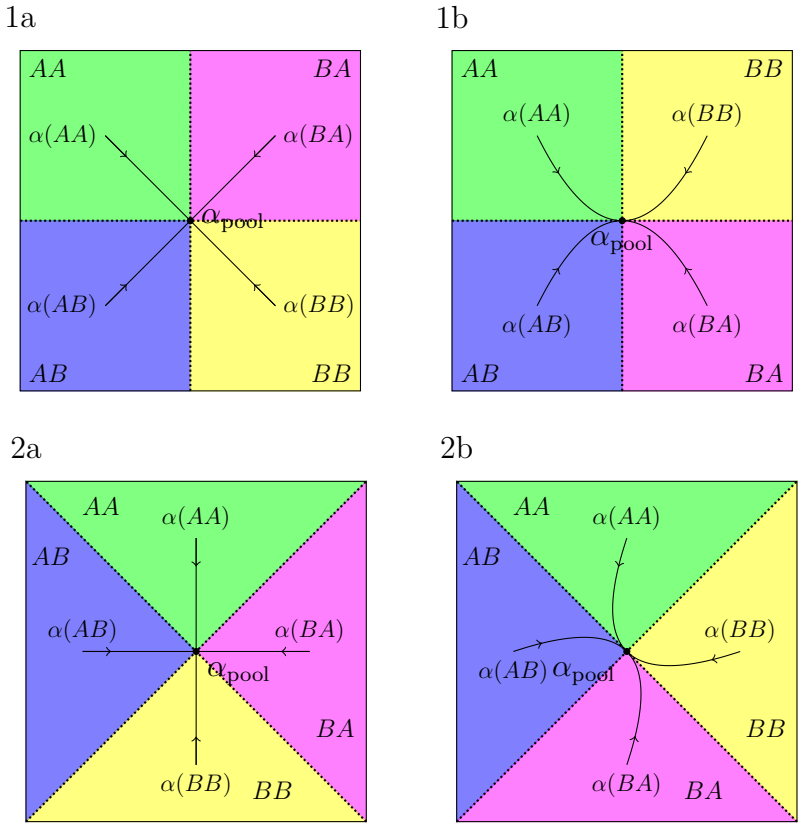

Figure 4: Four communication devices on T. The black lines depict the loci of the receiver's best responses for $p \in [0, \frac{1}{2}]$, moving from the center of the resp. cells to the pooling action. For every p they constitute a noise equilibrium respectively. While the best responses in 1a and 2a move to α_{pool} in a straight line, they are pulled closer to the cell of their neighbor with closer words in 1b and 2b.

Case	$\alpha(AA)$	$\alpha(AB)$	$L(\pi,\alpha)$
1a	$\frac{1}{4}(-1+2p, 1-2p)$	$\frac{1}{4}(-1+2p, -1+2p)$	$\frac{1}{6}-\frac{1}{8}(1-2p)^2$
1b	$\frac{1}{4}(-1+2p, 1-4p+4p^2)$	$\frac{1}{4}(-1+2p, -1+4p-4p^2)$	$\frac{1}{6}-\frac{1}{8}(1-2p)^2(1-2p+2p^2)$
2a	$\frac{1}{3}(0, -1+2p)$	$\frac{1}{3}(-1+2p, 0)$	$\frac{1}{6}-\frac{1}{9}(1-2p)^2$
2b	$\frac{1}{3}(-p+2p^2, 1-3p+2p^2)$	$\frac{1}{3}(-1+3p-2p^2, p-2p^2)$	$\frac{1}{6}-\frac{1}{9}(1-2p)^2(1-2p+2p^2)$

Table 1: Explicit formulae for the optimal actions and the expected loss of the languages from Figure 4. The remaining formulae can be found in the appendix Section 9.3 or by symmetry.

action. Increasing p thus shifts the interpretation on a straight line towards α_{pool}. In contrast, for the type-b languages, the locus of $\hat{\alpha}(AA)$ is again pulled linearly by the cells of AB and BA, while quadratically by the one of BB, which is the weaker force since $p < \frac{1}{2} < 1$. As it is less likely to confuse AA and BB, the border between their cells is less permeable for mistakes.

Let us take a look at their performance and focus on the nomination of the cells first. Each type-a language outperforms their type-b ones for each noise level, see Figure 5. The labeling of the cells thus is an important determinant of the efficiency of a language. In order to reduce the harm from miscommunication, the sender should use similar words to describe similar states and use distinct words to describe states that should not be mistaken.

Turning towards the structure of the cells employed, we find that the quadratic ones outperform the triangular shaped ones, i.e., type-1 languages have a lower expected loss than type-2 ones. The squares provide a more compact structure and have less points near and on the indifference levels than the triangle shapes.[11] As a result the interpretations for type-1 languages stay farther away from the pooling action than their resp. type-2 peers, leading to a more separable decomposition of the prior belief.

Interestingly, language 2a results in a lower loss than 1b for a crossover probability exceeding $p = \frac{1}{2} - \sqrt{2} \approx 6\%$, suggesting that the labeling of the cells can be more important than the choice of stable cell structures. However, when analyzing language formation we find both type-1 languages to be stable outcomes of evolution whereas type-2 languages are unstable, cf. Section 7.

[11] The total interior border length of type-1 languages is $1 + 1 = 2$, while for type-2 languages this is $\sqrt{2} + \sqrt{2} > 2.8$.

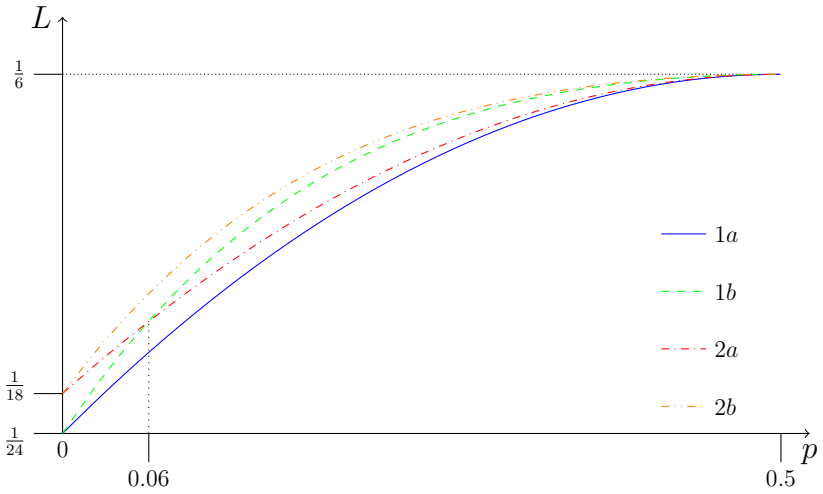

Figure 5: Expected loss of the four languages from Figure 4 for different crossover probabilities p. For $p = 0$ the assignments of words to cells is irrelevant and 1a and 1b have a lower loss than 2a and 2b. At the uninformativeness bound $p = \frac{1}{2}$ the pooling loss realizes for all languages. For all noise levels, 1a (2a) has a lower loss than 1b (2b). Interestingly, for $p > 6\%$ 2a has a lower loss than 1b.

6.3 Efficient languages with four words

Only a few words are really necessary to get the main idea of a sentence. The remaining words are either decorative, concerned with details or redundant. Emphasizing key words becomes more important the more likely communication is noisy. The following example indicates that the sender has an incentive to stress the words that describe the states which are farthest away from one another and would thus lead to a high loss if confused. This property is achieved by enlarging the cells on the outskirts. As a consequence, the cells close to the pooling action shrink. In a sense, the sender is willing to give up precision over the whole state space to ensure that extreme states are not mistaken.

Let $T = [-\frac{1}{2}, \frac{1}{2}]$ be endowed with the uniform distribution. The word space is given by $\mathrm{W} = \{A, B\}^2$ and the noise is given by a binary symmetric channel of length two with crossover probability p. For different values of p, we can categorize efficient languages. Some of them are depicted in Figure 6. Again, note that extreme states are articulated by sending words that are not easily mistaken. However, in contrast to the example in Section 6.2, the low dimension of the current state does not allow to sustain its structure for an increasing noise.

Imagine a speaker describes their friend the height of a person they just have met. They use words in the scheme "very tiny" (AA), "not too tiny" (AB), "not

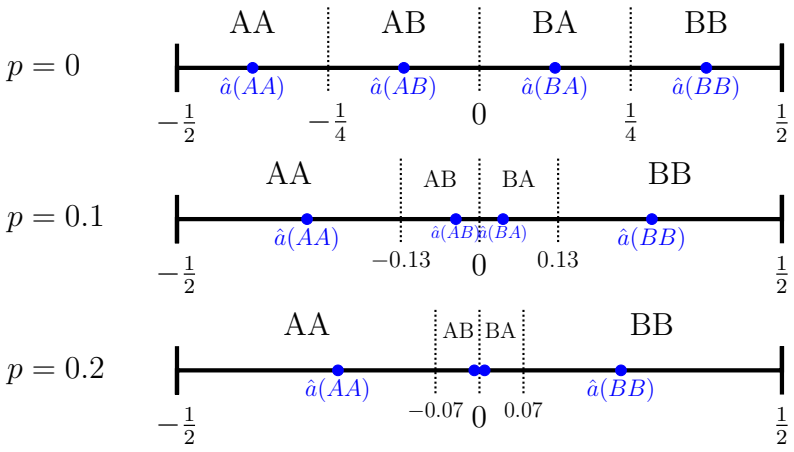

Figure 6: Efficient languages in the setting of Section 6.3 for different crossover probabilities p. Words of maximal distance are used to articulate the boundaries and their cells enlarge with increasing error.

too tall" (BA) and "very tall" (BB) to set their height in relation to the average height (indicated by the state 0). Note that we can interpret the letter A as indicating "tiny" and B as "tall". If a letter is repeatedly used, its indicative meaning is stressed, if both are used, its meaning is attenuated by the second letter.[12] Perhaps in contrast to the reader's intuition, our rational sender does not like vagueness and thus always has precise definition of the ranges of height addressed when using a word. If there is no noise, the speaker efficiently communicates the height by equally splitting the state space (if heights are equally distributed). If noise increases, the speaker puts more emphasis describing the extreme cases than on average sized persons. In the limit, the speaker will not use words to describe an average person at all and the meaning of the opposite words AA and BB becomes simply "tiny" and "tall".

7 Evolution

Natural languages have formed over time by the laws of evolution and are perpetually changing. The development from indistinct shrieks to elaborate speech has continuously improved coordination among human beings. Evolutionary game theory has proved useful in deriving qualitative properties of evolution in biology by means of simple models, cf. Smith & Price (1973), Hofbauer & Sigmund (1998). Its main idea is as follows. The population of a species at each point in time is

[12] Of course, exchanging the words AB and BA in the example does not qualitatively change the example.

a distribution over groups of different characteristics. Individuals are randomly paired and play a game. Their expected excess payoff is proportional to their fitness. If a group performs better than the average their subsequent share in the population increases. In the following, we apply evolutionary game theory to communication under noise. As it turns out, individuals can learn how to use communication to improve coordination even in the presence of noise.

Formally, we describe an individual by endowing them with both a communication device π out of the set of all measurable communication devices Σ and an interpretation map $\alpha \in T^{\#\mathrm{W}}$ for a fixed word space W. The strategy space $\Sigma \times T^{\#\mathrm{W}}$ is a complicated space, even more so when considering populations, i.e., probability distributions over strategies. The technical foundation has fortunately been established for certain dynamics and extends to our setting. These include the replicator, cf. Oechssler & Riedel (2001), Cressman et al. (2006), payoff monotone, cf. Heifetz et al. (2007), and Brown-von-Neumann-Nash dynamics, cf. Hofbauer et al. (2009).

We proceed along the lines of Jäger et al. (2011), considering a symmetric version of the cheap talk game studied so far. When two individuals meet, a fair coin toss decides about their roles of sender or receiver. An individual using (π, α) and meeting another one using (π', α') amounts to an expected loss of

$$\Lambda((\pi, \alpha), (\pi', \alpha')) = \frac{1}{2} L(\pi, \alpha') + \frac{1}{2} L(\pi', \alpha) \tag{23}$$

for both. Describing a population of individuals by a probability distribution P on $\Gamma := \Sigma \times T^{\#\mathrm{W}}$ the expected loss is generalized to

$$\Lambda(P, Q) := \int_\Gamma \int_\Gamma \Lambda((\pi, \alpha), (\pi', \alpha'))\, P(\mathrm{d}\pi, \mathrm{d}\alpha)\, Q(\mathrm{d}\pi', \mathrm{d}\alpha'). \tag{24}$$

Individuals are able to learn efficient communication under noise.

Proposition 7.1. *The following assertions hold.*

(i) The symmetrized loss function is a Lyapunov function for the replicator, regular and payoff monotone and the Brown-von-Neumann-Nash dynamics.

(ii) Locally optimal languages are Lyapunov stable w.r.t. the replicator, regular and payoff monotone and Brown-von-Neumann-Nash dynamics.

Figure 7 gives a tractable numerical illustration of evolution by means of the best reply dynamics in the setting of Section 6.2. In this case, a sender and a receiver meet every day and play our communication game. In the beginning, both had a random strategy, but after each encounter they learn the strategy of their peer and play a best response to that at the next time. As the figure shows, their

language quickly converges to a noise equilibrium. The depicted equilibrium is a local optimum that improves upon the pooling loss, but is not efficient, compare Section 6.2. Indeed, numerical simulations suggest that the languages 1a and 1b are stable and even attractors of the best reply dynamics, while languages 2a and 2b prove to be unstable.

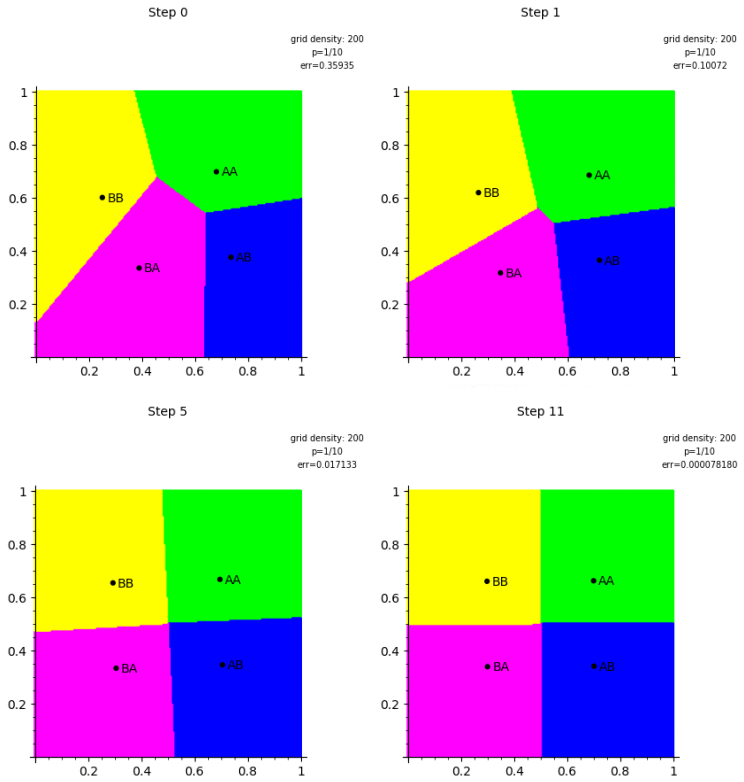

Figure 7: Numerical simulation of the best reply dynamics in the setting of Section 6.2 (shifted to the unit square). Starting with random interpretations, the agents take turn in playing a best reply to the previous strategy of their peer. Convergence to equilibrium is measured by the norm distance between two subsequent interpretation maps. Agents quickly learn one of the noise equilibria 1a or 1b from Figure 4 which, in contrast to 2a an 2b, appear to be stable and attractors.

8 Conclusion

Within our daily routine, errors in our communication are ubiquitous. Despite this inhibiting factor, humans have learned to communicate efficiently through the cause of their existence. This chapter formalizes noisy communication as a

cheap talk game of common interest and studies its structural properties, i.e., the grammar of communication. Efficient languages exist and can be learned by evolutionary dynamics. The sender optimally induces receiver actions, the spread of which should be as maximal as possible. As in non-noisy cases, the set of states which are referred to by a single word is convex and has sharp boundaries to reduce inefficient vagueness. If noise is present, examples illustrate that the sender can reduce the expected loss by using distant words to describe distant states. That way, they minimize the relatively high loss that results from confusing states that are very distinct. The more noisy communication is, the more important this becomes. If the dimension of the state space does not allow for errors to balance out, the communication device in place must adjust for higher errors by emphasizing extreme states at the cost of describing average states.

9 Appendix

9.1 Finite state space

While assuming a rich state space is compelling if we think of describing, say, colors, we can also think of scenarios in which states are finite, for instance telling somebody to go left or right. In fact, one can extend the presented framework by allowing for point masses of μ_0 on the state space. If μ_0 only has point masses, we can thus model a finite state space. In fact, many of the results we have seen carry over nicely, for instance the uniqueness of sender's best replies and the existence of efficient equilibria. The classical framework of Shannon (1948) indeed considered *finite sources* which is the natural assumption in information theory where data is discrete. One of its most important fields for all of our digital data transmission is *coding theory*, cf. Roth (2006). A code is a subset $C \subset \{0,1\}^n$ of binary sequences of a fixed length n. Before the transmission starts, agents declare C and agree that the sender only sends (code-)words $c \in C$. At first it might seem surprising that codes typically are a proper subset of all possible words $\{0,1\}^2$, but this is no coincidence. Including redundancy in communication enables the receiver to detect and probably even correct errors resulting from noise in the transmission channel. In the following, we illustrate the economic reasoning behind this using a finite state space with as many words as there are types.

Let $T = \{-2,-1,1,2\}$ be a discrete state space with uniform prior μ_0. The message space is $\mathrm{W} = \{A,B\}^2$. Communication is noisy and the error channel is given by a binary symmetric channel of length two with crossover probability p. Let $\mathrm{supp}(\pi) = \pi(T)$ be the support of a communication device π. For the sake of the argument, assume the receiver can still play 'hedging' actions, i.e., any point on the convex hull of T. The expected loss is quadratic. One can argue that,

up to isometry, there are unique best communication devices for each support of cardinality $2, 3, 4$ if the receiver plays their best response. We depict these in Table 2.

com.dev.	$\pi(-2)$	$\pi(-1)$	$\pi(1)$	$\pi(2)$
π_2	AA	AA	BB	BB
π_3	AA	AA	AB	BB
π_4	AA	AB	BA	BB

Table 2: The set of communication devices that are optimal given the cardinality of their support if the receiver plays a best response. Here $\mathrm{W} = \{A, B\}^2$ and $T = \{-2, -1, 1, 2\}$.

A straightforward calculation determines the expected losses given each of the communication devices from Table 2, see Figure 8 for a plot.

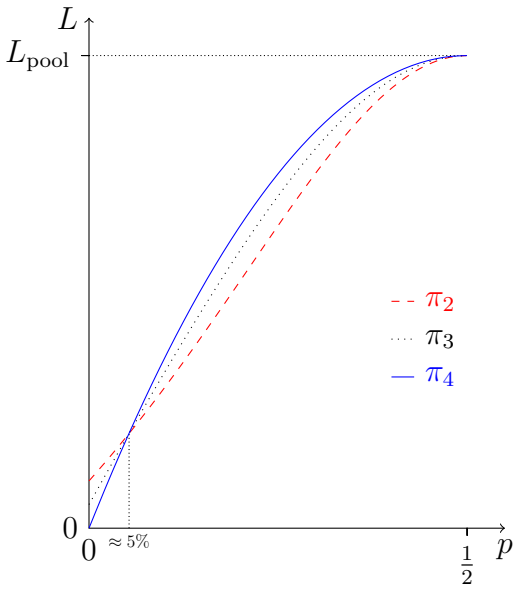

Figure 8: Expected loss of the communication devices in Table 2 under the receiver's best responses. The communication device using four words is the best for small values or p, while the one using only two words results in the lowest loss starting at $p \approx 5\%$. The agents are never better of using only three words.

The example provides an economic reasoning why codes in information theory feature redundancy. If there was no noise, the sender can only perfectly reveal the state to the receiver if they employ as many massages as there are states. However, if there is noise, all the receiver's actions get pulled towards the pooling action.

This effect is more pronounced the more words are used. As a result, there will be a threshold when using fewer words results in a lower expected loss, see Figure 8, justifying redundancy of a code.

9.2 Proofs & calculations

Proof of Lemma 3.1. The non-emptiness of the best reply set is given by continuity of the integrand and compactness of T. We will now prove that the function $T \to \mathbb{R}$, $s \mapsto \mathbb{E}_\mu[\ell(\|t-s\|)]$ is strictly convex, implying uniqueness of the minimizer. To this end, assume there are two distinct minimizers s_1, s_2 and let $\lambda \in (0,1)$. By the triangle inequality and convexity of ℓ we have

$$\begin{aligned}
&\mathbb{E}_\mu[\ell(\|t-(\lambda s_1+(1-\lambda)s_2)\|)] \\
\leq &\mathbb{E}_\mu[\ell(\lambda\,\|(t-s_1)\|+(1-\lambda)\,\|t-s_2\|)] && (25)\\
< &\mathbb{E}_\mu[\lambda\ell(\|t-s_1\|)]+\mathbb{E}_\mu[(1-\lambda)\ell(\|t-s_2\|)] && (26)\\
= &\mathbb{E}_\mu[\ell(\|t-\hat{s}\|)].
\end{aligned}$$

Since the strictness of the inequality in (26) is not too obvious, we give some more details. By strict convexity of ℓ, we have a strict inequality within the integrand for all t with $\|t-s_1\| \neq \|t-s_2\|$. Consequently, it suffices to prove that the set $M := \{t \mid \|t-s_1\| \neq \|t-s_2\|\}$ has positive μ-mass. By positive definiteness of the norm we have $s_1 \in M$. Since norms are continuous (even Lipschitz), M contains an open environment (in T) of s_1. Since μ is absolutely continuous w.r.t. the Lebesgue measure we find $\mu(M) > 0$ to conclude the proof. □

Proof of Proposition 3.2. We have

$$\begin{aligned}
L(\pi,\hat{\alpha}) &= \mathbb{E}_{\lambda^\pi}\left[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}\left[\ell(\|t-\hat{\alpha}(\mathrm{w})\|)\right]\right] && (27)\\
&\leq \mathbb{E}_{\lambda^\pi}\left[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}\left[\ell(\|t-\alpha_{\mathrm{pool}}\|)\right]\right] && (28)\\
&= \mathbb{E}_{\mu_0}[\ell(\|t-\alpha_{\mathrm{pool}}\|)] = L_{\mathrm{pool}}, && (29)
\end{aligned}$$

using the defining property for the inequality and Bayes-Plausibility to condense the expectations. The inequality is strict if and only if there is a word w with $\lambda^\pi(\mathrm{w}) > 0$ and $\hat{\alpha}(\mathrm{w}) \neq \alpha_{\mathrm{pool}}$ as the resp. minimizers are unique. □

Proof of Corollary 3.3. If π is constant, say $\pi \equiv \mathrm{v}$, then $\varepsilon(\mathrm{w}\,|\,\pi(t)) = \varepsilon(\mathrm{w}\,|\,\mathrm{v})$ is constant and equal to $\lambda^\pi(\mathrm{w})$ for any $\mathrm{w} \in \mathrm{W}$. This implies $f^\pi_{\mathrm{w}} = f_0$ and $\mu^\pi_{\mathrm{w}} = \mu_0$.

If $\mathrm{v} \mapsto \varepsilon(.\,|\,\mathrm{v})$ is constant, we have $K_{\mathrm{w}} := \varepsilon(\mathrm{w}\,|\,\pi(t))$ is independent of t for any communication device π. Consequently, also $\lambda^\pi(\mathrm{w}) = K_{\mathrm{w}}$, implying $f^\pi_{\mathrm{w}}(t) = f_0(t)$ and thus $\mu^\pi_{\mathrm{w}} = \mu_0$. □

Proof of Theorem 4.2. We follow the proof of Lemma 1 in Jäger et al. (2011) which can be adjusted to incorporate any noisy channel.

Consider any pure strategy $\alpha\colon \mathrm{W} \to T$ of the receiver. Then, a type t-sender may optimally send any word v out of

$$\arg\min_{\mathrm{v}'\in\mathrm{W}} \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|). \tag{30}$$

Note that the set of minimizers is non-empty as W is finite. Now, fix any strict ordering $\leq_{\mathrm{W}}$ on W and define a partition of T by setting

$$C^{\alpha}_{\mathrm{v}} := \left\{ t \mid \ \mathrm{v} \text{ is smallest w.r.t. } \leq_{\mathrm{W}} \text{ in } \arg\min_{\mathrm{v}'\in\mathrm{W}} \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \tag{31}$$

for each $\mathrm{v} \in \mathrm{W}$. C^{α}_{v} is (Lebesgue-)measurable as all involved functions are continuous in t and it is the set difference of a closed set from a finite union of closed sets. To see this, start by collecting all t for which v is a minimizer, which is a closed set. Now, for all $\mathrm{v}' \leq_{\mathrm{W}} \mathrm{v}$ that are also minimizers take away the indifference sets, which are themselves closed, to obtain C^{α}_{v}.

We define a measurable function $\pi\colon T \to \mathrm{W}$ by $\pi(t) = \mathrm{v} \iff t \in C^{\alpha}_{\mathrm{v}}$ which represents one possible best reply of the sender.

Using any such choice we can internalize a sender's best reply and re-write the joint loss minimization as a function depending only on α, namely

$$\min_{\alpha} \int_{T} \min_{\mathrm{v}} \left\{ \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \mu_0(\mathrm{d}t). \tag{32}$$

We identify any strategy $\alpha\colon \mathrm{W} \to T$ with a point in T^N, $N := \#\mathrm{W}$. Thus, by Lebesgue's dominant convergence theorem, it suffices to prove continuity of the integrand in α for any fixed t. But this is obvious as the pointwise minimum of finitely many continuous functions is again continuous. □

Proof of Remark 5.1. The case $p = 0$ is clear. For $0 < \widetilde{p} < 1$ we have that

$$\varepsilon(\mathrm{w}\,|\,\mathrm{v}) > \varepsilon(\mathrm{w}'\,|\,\mathrm{v}) \iff \widetilde{p}^{d(\mathrm{w},\mathrm{v})} > \widetilde{p}^{d(\mathrm{w}',\mathrm{v})} \iff d(\mathrm{w},\mathrm{v}) < d(\mathrm{w}',\mathrm{v}). \tag{33}$$

The other cases follow similarly. □

Proof of Proposition 5.2. (i) Follows immediately from continuity of the integrand in $\widetilde{p}$ and Lebesgue's theorem.

(ii) We split the integral in three parts by disassembling the type space T into the three stets defined by $\{t' \mid d(\mathrm{w}, \pi(t')) \sim^* d(\mathrm{w}, \pi(t))\}$ for $\sim^* \in \{<, =, >\}$.

(a) The set $\{t' \mid d(\mathrm{w}, \pi(t')) < d(\mathrm{w}, \pi(t))\}$ has positive probability and the negative exponent $d(\mathrm{w}, \pi(t')) - d(\mathrm{w}, \pi(t))$ will let the integral go to infinity as $p \to 0$.

(b) The set $\{t' \mid d(\mathrm{w}, \pi(t')) < d(\mathrm{w}, \pi(t))\}$ has probability zero and can be neglected. For $p \to 0$, the integral over $\{t' \mid d(\mathrm{w}, \pi(t')) > d(\mathrm{w}, \pi(t))\}$ will vanish as the exponent of $\widetilde{p}$ is strictly positive. What is left of the overall integral is $\int_{\{t' \mid d(\mathrm{w}, \pi(t')) = d(\mathrm{w}, \pi(t))\}} \mu_0(\mathrm{d}t') = \mu_0(\{t' \mid d(\mathrm{w}, \pi(t')) = d(\mathrm{w}, \pi(t))\})$ which is strictly positive by assumption.

(c) Ignoring the integral over null sets, the limit $p \to 0$ makes the integral go to 0 making the limit meaningless.

□

Proof of Proposition 5.3. We start by calculating the entropy

$$
\begin{aligned}
& H(\varepsilon(. \mid \mathrm{v})) \\
= & - \sum_{w \in \mathrm{W}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \cdot \log\left(\varepsilon(\mathrm{w} \mid \mathrm{v})\right) \\
= & - \sum_{w \in \mathrm{W}} (1-p)^{n - d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})} \cdot \log\left((1-p)^{n-d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})}\right) && (34) \\
= & - \sum_{d=0}^{n} \binom{n}{d} \cdot m^d \cdot (1-p)^{n-d} \cdot \left(\frac{p}{m}\right)^{d} \cdot \log\left((1-p)^{n-d} \cdot \left(\frac{p}{m}\right)^{d}\right) && (35) \\
= & - \sum_{d=0}^{n} \binom{n}{d} \cdot (1-p)^{n-d} \cdot p^d \cdot \log\left((1-p)^{n-d} \cdot p^d\right) && (36) \\
& + \log(m) \cdot \sum_{d=0}^{n} \binom{n}{d} \cdot d \cdot (1-p)^{n-d} \cdot p^d && (37) \\
= & - \sum_{d=0}^{n} \binom{n}{d} \cdot (1-p)^{n-d} \cdot p^d \cdot \log\left((1-p)^{n-d} \cdot p^d\right) + np\log(m) && (38) \\
= & - np \cdot \log(p) - n(1-p) \cdot \log(1-p) + np\log(m) && (39) \\
= & \, n \cdot (H((p, 1-p)) + p\log(m)). && (40)
\end{aligned}
$$

During the calculation we used the functional equation of the logarithm and formulae occurring often when dealing with binomial distributions, e.g., its mean. This function is concave in p since the entropy $H((p, 1-p)) = -p\log(p) - (1-p)\log(1-p)$

over a binary source with probability p is. The maximum is attained for the uniform distribution, i.e., if $p = \frac{m}{m+1}$ by Remark 5.1 and yields $H(\mathcal{U}(\mathrm{W})) = \log(\#\mathrm{W})$. The other assertions follow readily from the calculated expression. □

Formulae for Example 5.4. For $p \in [0,1]$ the optimal response $\hat{\alpha}$ of the receiver is given by

$$\hat{\alpha}(L) = \frac{-2+3p}{8-4p}, \quad \hat{\alpha}(M) = 0, \quad \hat{\alpha}(R) = -\hat{\alpha}(L) \tag{41}$$

and the expected loss given p can be calculated to be

$$L(\pi, \hat{\alpha})(p) = \frac{1}{12} - 2^{-5} \cdot \frac{(-2+3p)^2}{2-p}. \tag{42}$$

□

Proof of Lemma 6.1. The receiver's minimization problem reads as

$$\min_{\alpha \in T} \mathbb{E}_\mu[\|t-\alpha\|_2^2] = \int_T \sum_{k=1}^{L} (t_k - \alpha_k)^2 \, \mu(\mathrm{d}t). \tag{43}$$

Using the Leibniz rule we check the first and second order conditions for each k and obtain the unique local and global minimum by choosing

$$\hat{\alpha}_k(\mu) = \int_T t_k \, \mu(\mathrm{d}t) = \mathbb{E}_\mu[t_k]. \tag{44}$$

Plugging $\hat{a}(\mu) = \mathbb{E}_\mu[t]$ back into the expected loss and using the scalar product $\langle \, . \, , \, . \, \rangle$ we get

$$\mathbb{E}_\mu[\|t - \mathbb{E}_\mu[t']\|_2^2] = \mathbb{E}_\mu[\|t\|_2^2] - \|\mathbb{E}_\mu[t]\|_2^2 \,, \tag{45}$$

which is the trace norm of the variance matrix, i.e., the sum of the variances over each dimension. If a communication device π is given the expected loss is

$$\begin{aligned} \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[\left\|t - \mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t']\right\|_2^2]] &= \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[\|t\|_2^2] - \left\|\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]\right\|_2^2] && (46) \\ &= \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\left\|\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]\right\|_2^2] && (47) \end{aligned}$$

where Bayes-Plausibility has been used. Finally, applying Bayes-Plausibility one more time we observe

$$\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]] = \mathbb{E}_{\mu_0}[t] = \hat{\alpha}(\mu_0) = \alpha_{\mathrm{pool}}. \tag{48}$$

□

Proof of Proposition 6.2. Without loss of generality, we assume $\alpha_{\text{pool}} = 0$ by translating the state space and the measure by $-\alpha_{\text{pool}}$. Since ε is not uninformative, there are two words v, v' with $\varepsilon(\mathrm{v} \,|\, \mathrm{v}) \neq \varepsilon(\mathrm{v} \,|\, \mathrm{v}')$. Assume that $\varepsilon(\mathrm{v} \,|\, \mathrm{v}) > 0$, applying similar arguments with swapped roles in the following otherwise.

Note that for any normal vector $\vec{n} \in \mathbb{R}^L$ the corresponding $L-1$ dimensional hyperplane $H := \{t \mid \langle \vec{n}, t \rangle = 0\}$ separated T into two disjoint convex cells $C_{\mathrm{v}} := T \cap H_+ = \{t \in T \mid \langle \vec{n}, t \rangle > 0\}$ and $C_{\mathrm{v}'} = T \cap H_- = \{t \in T \mid \langle \vec{n}, t \rangle \leq 0\}$. Since μ_0 is absolutely continuous w.r.t. to the Lebesgue measure, we can choose $\vec{n}$ in a way that both cells have positive μ_0-measure and $0 = \alpha_{\text{pool}} \in C_{\mathrm{v}'}$. Note that $0 \neq \mu_0(C_{\mathrm{v}}) \cdot \int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t) = \mathbb{E}_{C_{\mathrm{v}}, \mu_0}[t] \in C_{\mathrm{v}}$ by convexity of C_{v} and Lemma 6.1 and thus $0 \neq \int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t)$.

Now define $\pi(t) = \mathrm{v}$ if $t \in C_{\mathrm{v}}$ and $\pi(t) = \mathrm{v}'$ otherwise. Since we have $\lambda^\pi(\mathrm{v}) = \varepsilon(\mathrm{v} \,|\, \mathrm{v})\mu_0(C_{\mathrm{v}}) + \varepsilon(\mathrm{v} \,|\, \mathrm{v}')\mu_0(C_{\mathrm{v}'}) > 0$, it is sufficient to show $\hat{\alpha}(\mathrm{v}) \neq 0 = \alpha_{\text{pool}}$ to prove $L(\pi, \hat{\alpha}) > L_{\text{pool}}$ by Proposition 3.2. We use Lemma 6.1 again and evaluate

$$0 \neq \hat{\alpha}(\mathrm{v}) \cdot \lambda^\pi(\mathrm{v}) = \int_T \varepsilon(\mathrm{v} \,|\, \pi(t)) \cdot t \; \mu_0(\mathrm{d}t) \tag{49}$$

$$\iff 0 \neq \varepsilon(\mathrm{v} \,|\, \mathrm{v}) \int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t) + \varepsilon(\mathrm{v} \,|\, \mathrm{v}') \int_{C_{\mathrm{v}'}} t \; \mu_0(\mathrm{d}t). \tag{50}$$

If $\varepsilon(\mathrm{v} \,|\, \mathrm{v}') = 0$, we conclude $\hat{\alpha}(\mathrm{v}) \neq 0$ as $\varepsilon(\mathrm{v} \,|\, \mathrm{v}) > 0$ and $\int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t) \neq 0$. If $\varepsilon(\mathrm{v} \,|\, \mathrm{v}') > 0$ we divide by $\varepsilon(\mathrm{v} \,|\, \mathrm{v}')$ and find

$$0 = \alpha_{\text{pool}} = \mathbb{E}_{\mu_0}[t] = \int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t) + \int_{C_{\mathrm{v}'}} t \; \mu_0(\mathrm{d}t), \tag{51}$$

$$0 \neq \frac{\varepsilon(\mathrm{v} \,|\, \mathrm{v})}{\varepsilon(\mathrm{v} \,|\, \mathrm{v}')} \cdot \int_{C_{\mathrm{v}}} t \; \mu_0(\mathrm{d}t) + \int_{C_{\mathrm{v}'}} t \; \mu_0(\mathrm{d}t). \tag{52}$$

As $\varepsilon(\mathrm{v} \,|\, \mathrm{v}) \neq \varepsilon(\mathrm{v} \,|\, \mathrm{v}')$, we conclude $\hat{\alpha}(\mathrm{v}) \neq 0$. □

Proof of Proposition 6.3. Revisit the argmin-set of sender (11) and recall that $\|x - y\|_2^2 = \|x\|_2^2 - 2\,\langle x \,,\, y \rangle + \|y\|_2^2$. In state t, the sender strictly prefers to send v instead of v′ if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \cdot \|t - \alpha(\mathrm{w})\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}') \cdot \|t - \alpha(\mathrm{w})\|_2^2 \tag{53}$$

$$\iff \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \left(-2\,\langle t \,,\, \alpha(\mathrm{w}) \rangle + \|\alpha(\mathrm{w})\|_2^2\right) > 0 \tag{54}$$

By linearity of the scalar product, convexity and the topological properties become clear. For the weak preference substitute the proper inequality accordingly. □

Proof of Proposition 6.4. From (54) the sender is indifferent between sending v and v′ if the state is t if and only if

$$0 = \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \left(-2 \left\langle t \,,\, \alpha(\mathrm{w})\right\rangle + \|\alpha(\mathrm{w})\|_2^2\right) \tag{55}$$

$$\begin{aligned} \iff 0 = &- 2 \cdot \left\langle t \,,\, \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \alpha(\mathrm{w}) \right\rangle \\ &+ \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \|\alpha(\mathrm{w})\|_2^2 \,. \end{aligned} \tag{56}$$

Note that if $\vec{x} := \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \alpha(\mathrm{w}) \neq 0$ the solution set to (56) is the translation of the $L-1$ dimensional hyperplane perpendicular to the vector $\vec{x}$ by a particular solution (if it exists, otherwise it is the empty set) and thus a null set in $\mathbb{R}^L$ w.r.t. μ_0. If $\vec{x} = 0$ we can only have indifference if also $\sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \|\alpha(\mathrm{w})\|_2^2 = 0$. If this is the case, any t in T (even $\mathbb{R}^L$) solves (56). Otherwise, (56) is equivalent to

$$0 = \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \,|\, \mathrm{v}') - \varepsilon(\mathrm{w} \,|\, \mathrm{v})\right) \cdot \alpha(\mathrm{w}) \tag{57}$$

$$\iff 0 = \sum_{\mathrm{w}} \left(\widetilde{p}^{d(\mathrm{w},\mathrm{v}')} - \widetilde{p}^{d(\mathrm{w},\mathrm{v})}\right) \cdot \alpha(\mathrm{w}) =: Q(\widetilde{p}), \tag{58}$$

where Q is a (vector-valued) polynomial of degree at most n in $\widetilde{p}$. Evidently, Q has a zero in $\widetilde{p} = 1$,i.e., $p = \frac{m}{m+1}$, representing the uninformativeness bound. It does not vanish in $p = \widetilde{p} = 0$ as its constant coefficient is $\alpha(\mathrm{v}') - \alpha(\mathrm{v}) \neq 0$. Consequently, Q is not the zero polynomial and has at most $n-1$ further zeros on $p \in (0,1] \setminus \{\frac{m}{m+1}\}$. □

Proof of Proposition 7.1. We adapt the proof of Jäger et al. (2011) to account for noise. It suffices to show continuity and boundedness of L as this implies continuity of Λ in the weak topology. The rest of the assertions follow well-known lines (Heifetz et al. (2007), Hofbauer et al. (2009)) as well as Bhatia & Szegö (2002) for the last statement.

Let $(\pi_k)_k$ be a sequence of communication strategies converging uniformly to π, i.e., for all $\rho' > 0$ there is an M such that for all $t \in T$ we simultaneously find $d(\pi_k(t), \pi(t)) < \rho'$ for $k > M$. As d has only values in $\{0, \ldots, n\}$, this is equivalent to $\pi_k \equiv \pi$ for all $k > N_0$ for some N_0. Let $\rho > 0$ be arbitrary. As T is compact and $|.|$ as well as ℓ are continuous,there is $\delta > 0$ such that $|\ell(a) - \ell(b)| < \rho$ if $\|a - b\| < \delta$. Furthermore, let $(\alpha_k)_k$ be a sequence converging to α uniformly on $T^{\#\mathrm{W}} \subsetneq \mathbb{R}^{\#\mathrm{W}}$. Then there is $N_1 \geq N_0$ such that for all $t \in T$ and $\mathrm{w} \in \mathrm{W}$ we have

$\|t - \alpha_k(\mathrm{w}) - (t - \alpha(\mathrm{w}))\| = \|\alpha_k(\mathrm{w}) - \alpha(\mathrm{w})\| < \delta$ for all $k > N_1$. Hence, for all $k > N_1$ we find

$$|L(\pi_k, \alpha_k) - L(\pi, \alpha))| \tag{59}$$

$$\leq \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi_k(t)) \ell(\|t - \alpha_k(\mathrm{w})\|) - \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{60}$$

$$= \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left(\ell(\|t - \alpha_k(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|) \right) \right| \mu_0(\mathrm{d}t) \tag{61}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left| \ell(\|t - \alpha_k(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{62}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \rho \, \mu_0(\mathrm{d}t) = \rho. \tag{63}$$

Finally, boundedness of L follows from compactness of T and continuity of ℓ, since $\bar{\ell} := \sup_{t \in T} |\ell(\|t\|)| < \infty$. For any π, α

$$|L(\pi, \alpha)| \leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)\cdot) \left| \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{64}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \bar{\ell} \, \mu_0(\mathrm{d}t) = \bar{\ell} < \infty. \tag{65}$$

□

9.3 Calculations of Section 6.2

We prove the assertions and formulae of Section 6.2 by the following lengthy calculations which are structured as follows. We begin by calculating the minimizing interpretations and the expected loss. Afterwards, fixing any of the calculated interpretation maps we show that the given tessellation is indeed an optimal one, even uniquely up to null sets. Denote by $\mathbb{E}_C$ the expectation operator of the measure μ_0 restricted to the cell C. We begin by noting that for any considered cell $\mu_0(C_{\mathrm{v}}) = \frac{1}{4}$ and thus for any w

$$\lambda^{\pi}(\mathrm{w}) = \int_T \varepsilon(\mathrm{w} \,|\, \pi(t)) \, \mathrm{d}t = \sum_{\mathrm{v}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \cdot \mu_0(C_{\mathrm{v}}) = \frac{1}{4}. \tag{66}$$

(i) Interpretations and expected Loss

Denote by $\square(AA)$ and $\Delta(AA)$ the resp. square or triangular cell in Figure 4. The following calculations are down for the resp. languages.

1 a) We derive the center of gravity of each cell, say the one for AB.

$$\mathbb{E}_{\square(AB)}[t] := \mathbb{E}_{\mu_0,\square(AB)}[t] = \mu_0(\square(AB))^{-1} \cdot \int_{\square(AB)} t \,\mathrm{d}t \tag{67}$$

$$= 4 \cdot \int_{-\frac{1}{2}}^{0} \int_{-\frac{1}{2}}^{0} (t_1, t_2) \,\mathtt{d}t_1 \mathrm{d}t_2 = \left(-\frac{1}{4}, -\frac{1}{4}\right). \tag{68}$$

Similarly, or by using symmetry arguments, we obtain the expected values for AA, BA, BB which are, resp. $(-\frac{1}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{1}{4})$,$(\frac{1}{4}, -\frac{1}{4})$. Let us now calculate, e.g., the optimal action $\hat{\alpha}(AA)$.

$$\hat{\alpha}(AA) = \mathbb{E}_{\mu^{\pi}_{AA}}[t] = \lambda^{\pi}(AA)^{-1} \cdot \int_T \varepsilon(AA \,|\, \pi(t)) \cdot t \,\mu_0(\mathrm{d}t) \tag{69}$$

$$= 4 \cdot \sum_{\mathrm{w}} \varepsilon(AA \,|\, \mathrm{w}) \cdot \int_{\square(\mathrm{w})} t \,\mathrm{d}t \tag{70}$$

$$= \sum_{\mathrm{w}} \varepsilon(AA \,|\, \mathrm{w}) \cdot \mathbb{E}_{\square(\mathrm{w})} \tag{71}$$

$$= (1-p)^2 \cdot (-\tfrac{1}{4}, \tfrac{1}{4}) + p(1-p) \cdot \big((-\tfrac{1}{4}, -\tfrac{1}{4}) + (\tfrac{1}{4}, \tfrac{1}{4})\big) + p^2 \cdot (\tfrac{1}{4}, -\tfrac{1}{4}) \tag{72}$$

$$= \frac{1}{4} \cdot (-1 + 2p, 1 - 2p)\,. \tag{73}$$

Analogously, by symmetry arguments or using Lemma 6.1 we get $\hat{\alpha}(AB) = \frac{1}{4}\cdot(-1+2p, -1+2p)$, $\hat{\alpha}(BA) = \frac{1}{4}\cdot(1-2p, 1-2p)$, $\hat{\alpha}(BB) = \frac{1}{4}\cdot(1-2p, -1+2p)$. For each word w we see $\|\alpha(\mathrm{w}) - \alpha_{\mathrm{pool}}\|_2 \searrow 0$ for $p \to \frac{1}{2}$, where $\alpha_{\mathrm{pool}} = (0,0)$ is the center of the whole state space.

We are now set to calculate the expected loss and start by observing that each interpretation has the same norm:

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \cdot \left\| \frac{1}{4} \cdot (1-2p, 1-2p) \right\|_2^2 = \frac{1}{8} \cdot (1-2p)^2. \tag{74}$$

Having calculated $\mathbb{E}_T[\|t\|_2^2] = \frac{1}{6}$, we use (20) to obtain the expected loss

$$L(\pi_{1,a}, \hat{\alpha}) = \frac{1}{6} - \sum_{\mathrm{w}} \frac{1}{4} \cdot \frac{1}{8} \cdot (1-2p)^2 = \frac{1}{6} - \frac{1}{8} \cdot (1-2p)^2. \tag{75}$$

One clearly sees that the expected loss is monotonically increasing in $p \in [0, \frac{1}{2}]$

b) Using the calculations from (a) we can directly compute the optimal interpretations, only keeping in mind that the centers of gravity are switched for BA and BB. We obtain $\hat{\alpha}(AA) = \frac{1}{4} \cdot (-1 + 2p, 1 - 4p + 4p^2)$, $\hat{\alpha}(AB) = \frac{1}{4} \cdot (-1 + 2p, -1 + 4p - 4p^2)$, $\hat{\alpha}(BA) = \frac{1}{4} \cdot (1 - 2p, -1 + 4p - 4p^2)$, $\hat{\alpha}(BB) = \frac{1}{4} \cdot (1 - 2p, 1 - 4p + 4p^2)$. Thus, for any word w we have

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{16} \cdot ((1-2p)^2 + (1-2p)^4), \tag{76}$$

resulting in an expected loss of

$$L(\pi_{1,b}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{16} \cdot \left((1-2p)^2 + (1-2p)^4\right) \tag{77}$$

$$= \frac{1}{6} - \frac{1}{8} \cdot (1-2p)^2 \cdot (1 - 2p + 2p^2). \tag{78}$$

We observe for $0 < p < \frac{1}{2}$

$$L(\pi_{1,a}, \hat{\alpha}) < L(\pi_{1,b}, \hat{\alpha}), \tag{79}$$

thus, the language putting distant words farther away from one another achieves a lower expected loss.

2 a) The expected values of each colored area can be determined to be $\mathbb{E}_{\Delta(AA)}[t] = (0, \frac{1}{3})$, $\mathbb{E}_{\Delta(AB)}[t] = (-\frac{1}{3}, 0)$, $\mathbb{E}_{\Delta(BA)}[t] = (\frac{1}{3}, 0)$, $\mathbb{E}_{\Delta(BB)}[t] = (0, -\frac{1}{3})$.
Optimal actions can be computed to be $\hat{\alpha}(AA) = (0, -\frac{1}{3} + \frac{2}{3}p)$, $\hat{\alpha}(AB) = (-\frac{1}{3} + \frac{2}{3}p, 0)$, $\hat{\alpha}(BA) = (\frac{1}{3} - \frac{2}{3}p, 0)$, $\hat{\alpha}(BB) = (0, \frac{1}{3} + \frac{2}{3}p)$.
We thus get

$$\|\alpha(\mathrm{w})\|_2^2 = \left\|(0, \tfrac{1}{3} - \tfrac{2}{3}p)\right\|_2^2 = \frac{1}{9} \cdot (1-2p)^2. \tag{80}$$

The resulting expected loss is

$$L(\pi_{2,a}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{9} \cdot (1-2p)^2, \tag{81}$$

which is strictly higher than $L(\pi_{1,a}, \hat{\alpha})$ for any $p \in [0, \frac{1}{2})$.

b) Optimal actions can be calculated to be $\hat{\alpha}(AA) = \frac{1}{3} \cdot (-p + 2p^2, 1 - 3p + 2p^2)$, $\hat{\alpha}(AB) = \frac{1}{3} \cdot (-1 + 3p - 2p^2, p - 2p^2)$, $\hat{\alpha}(BA) = \frac{1}{3} \cdot (p - 2p^2, -1 + 3p - 2p^2)$, $\hat{\alpha}(BB) = \frac{1}{3} \cdot (1 - 3p + 2p^2, -p + 2p^2)$.
We thus get for any word w

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{9}(1-2p)^2(1-2p+2p^2) \tag{82}$$

and hence

$$L(\pi_{2,b}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{9}(1-2p)^2(1-2p+2p^2), \tag{83}$$

which is worse than $L(\pi_{2,a}, \hat{\alpha})$ for $0 < p < \frac{1}{2}$.

(ii) Optimal cell structure

To start with, we simplify the expressions from Proposition 6.3 and Proposition 6.4 for $\mathrm{W} = \{A, B\}^2$. To this end, fix w.l.o.g. the word AA and derive conditions on a fixed $t \in T$ for AA to be the optimal word.

(i) In state t the sender prefers to send AA over BB if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AA)\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(BB)\|_2^2 \tag{84}$$

$$\iff \|t - \alpha(AA)\|_2 < \|t - \alpha(BB)\|_2 \,. \tag{85}$$

(ii) In state t the sender prefers AA over AB (the case BA is analogous) if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AA)\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AB)\|_2^2 \tag{86}$$

$$\begin{aligned} \iff & \|t - \alpha(AA)\|_2^2 - \|t - \alpha(AB)\|_2^2 \\ & < \widetilde{p} \left(\|t - \alpha(BB)\|_2^2 - \|t - \alpha(BA)\|_2^2 \right) \end{aligned} \tag{87}$$

$$\begin{aligned} \iff & 2 \left\langle t \,,\, \alpha(AB) - \alpha(AA) + \widetilde{p}(\alpha(BB) - \alpha(BA)) \right\rangle \\ & + \|\alpha(AA)\|_2^2 - \|\alpha(AB)\|_2^2 + \widetilde{p}(\|\alpha(BA)\|_2^2 - \|\alpha(BB)\|_2^2) < 0. \end{aligned} \tag{88}$$

Whereas in (i) we clearly see that the set of states for which the sender is indifferent between sending AA and BB lie on the perpendicular bisector of $\alpha(AA)$ and $\alpha(BB)$ if the interpretations do not agree, it is not so obvious in case (ii). What we can say for sure is, that, as long as $\alpha(AB) - \alpha(AA) + \widetilde{p}(\alpha(BB) - \alpha(BA))$ is not the zero vector, the set of indifferent states is again a null set as it is the intersection of a line and T.

To drop some notation, we just write AA instead of $\alpha(AA)$ from Table 1 when talking about points in T. Consider the variants (a) and (b) respectively and let $t \in \square(AA)$ (resp. $t \in \Delta(AA)$) be in the interior.

(a) Observe that

$$\|t - AA\|_2 < \|t - AB\|_2 \,, \|t - BA\|_2 < \|t - BB\|_2 \,. \tag{89}$$

Obviously, sending AA is preferred to BB as $\|t - AA\|_2 < \|t - BB\|_2$.

Realizing that

$$\|t - AA\|_2^2 - \|t - AB\|_2^2 < 0 < \widetilde{p} \cdot \left(\|t - BB\|_2^2 - \|t - BA\|_2^2 \right), \tag{90}$$

reveals that sending AA is preferred to AB (and analogously BA). Thus, AA is the unique best word to be send.

(b) As before, preferring AA to BB is clear from $\|t - AA\|_2 < \|t - BB\|_2$. Since

$$0 \leq \|t - AA\|_2 < \|t - AB\|_2 \,, \|t - BB\|_2 < \|t - BA\|_2 \,, \tag{91}$$

we find

$$\begin{aligned} \|t - BA\|_2^2 - \|t - AA\|_2^2 &> \left| \|t - AB\|_2^2 - \|t - BB\|_2^2 \right| && (92)\\ &> \widetilde{p} \cdot \left| \|t - AB\|_2^2 - \|t - BB\|_2^2 \right| && (93)\\ &\geq \widetilde{p} \cdot \left(\|t - AB\|_2^2 - \|t - BB\|_2^2 \right), && (94) \end{aligned}$$

proving that AA is preferred to BA. Eventually, using $AA = -BA$, $AB = -BB$ and that $\|\alpha(\mathrm{w})\|$ is constant, we find

$$\begin{aligned} &2 \cdot \langle t \,, -AA + AB + \widetilde{p}(BB - BA) \rangle \\ &\quad + \|AA\|_2^2 - \|AB\|_2^2 + \widetilde{p}(\|BB\|_2^2 - \|BA\|_2^2) && (95)\\ =& 4 \cdot \left\langle t \,, \underbrace{\tfrac{AB+BA}{2}}_{=:P} \right\rangle . && (96) \end{aligned}$$

The expression (96) is smaller than zero in both cases for $t \in \Delta(AA)$:

1. $t_1 < 0, t_2 > 0$ and $P_1 = 0, P_2 < 0$.
2. $t = (y, z)$ with $z > 0, |y| < z$ and $P = (-x, x), x > 0$.

Thus, sending AA is preferred to AB as well.

The calculations above show that the borders of the cells consist precisely of the points for which the sender is indifferent between sending the resp. messages.

References

Bhatia, N. & Szegö, G. (2002), *Stability Theory of Dynamical Systems*, Classics in Mathematics, Springer Berlin Heidelberg. https://books.google.de/books?id=wP5dwTS6jg0C.

Blume, A. & Board, O. (2013), 'Language barriers', *Econometrica* **81**(2), 781–812.

Blume, A., Board, O. J. & Kawamura, K. (2007), 'Noisy talk', *Theoretical Economics* **2**(4), 395–440.

Cover, T. M. & Thomas, J. A. (2006), *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience.

Crawford, V. P. & Sobel, J. (1982), 'Strategic information transmission', *Econometrica* **50**(6), 1431–1451. http://www.jstor.org/stable/1913390.

Cremer, J., Garicano, L. & Prat, A. (2007), 'Language and the theory of the firm', *The Quarterly Journal of Economics* **122**(1), 373–407.

Cressman, R., Hofbauer, J. & Riedel, F. (2006), 'Stability of the replicator equation for a single species with a multi-dimensional continuous trait space', *Journal of Theoretical Biology* **239**(2), 273–288. Special Issue in Memory of John Maynard Smith, Available at https://www.sciencedirect.com/science/article/pii/S0022519305003887.

Frege, G. (1892), 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und philosophische Kritik* **100**, 25–50.

Grice, H. P. (1975), Logic and conversation, *in* 'Speech acts', Brill, pp. 41–58.

Gärdenfors, P. (2004), *Conceptual spaces: The geometry of thought*, MIT press.

Hamming, R. W. (1950), 'Error detecting and error correcting codes', *The Bell System Technical Journal* **29**(2), 147–160.

Heifetz, A., Shannon, C. & Spiegel, Y. (2007), 'What to maximize if you must', *Journal of Economic Theory* pp. 31–57.

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

Hofbauer, J., Oechssler, J. & Riedel, F. (2009), 'Brown–von neumann–nash dynamics: The continuous strategy case', *Games and Economic Behavior* **65**(2), 406–429. Available at https://www.sciencedirect.com/science/article/pii/S0899825608000651.

Hofbauer, J. & Sigmund, K. (1998), *Evolutionary games and population dynamics*, Cambridge university press.

Jeitschko, T. D. & Normann, H.-T. (2012), 'Signaling in deterministic and stochastic settings', *Journal of Economic Behavior & Organization* **82**(1), 39–55.

Jäger, G. (2007), 'The evolution of convex categories', *Linguistics and Philosophy* **30**, 551–564.

Jäger, G., Metzger, L. P. & Riedel, F. (2011), 'Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals', *Games and Economic Behavior* **73**(2), 517 – 537.

Kamenica, E. & Gentzkow, M. (2011), 'Bayesian persuasion', *American Economic Review* **101**(6), 2590–2615. Available at http://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Lewis, D. (1969), *Convention: A philosophical study*, John Wiley & Sons.

Lipman, B. L. (2009), Why is language vague. Available at https://sites.bu.edu/blipman/files/2021/10/vague5.pdf.

MacKay, D. J. C. (2002), *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, USA.

Martel, J., Van Wesep, E. D. & Van Wesep, R. (2019), On ratings: A theory of non-strategic information transmission. Working paper, available at .

Miller, A. (2007), *Philosophy of language*, Taylor & Francis.

Myerson, R. B. (1991), *Game Theory: Analysis of Conflict*, Harvard University Press. Available at http://www.jstor.org/stable/j.ctvjsf522.

Nowak, M. A. & Krakauer, D. C. (1999), 'The evolution of language', *Proceedings of the National Academy of Sciences* **96**(14), 8028–8033.

Oechssler, J. & Riedel, F. (2001), 'Evolutionary dynamics on infinite strategy spaces', *Economic Theory* **17**(1), 141–162. https://ideas.repec.org/a/spr/joecth/v17y2001i1p141-162.html.

Ortoleva, P. (2012), 'Modeling the change of paradigm: Non-Bayesian reactions to unexpected news', *American Economic Review* **102**(6), 2410–2436.

Pelletier, F. J. (2001), 'Did Frege believe Frege's principle?', *Journal of Logic, Language and information* **10**, 87–114.

Roth, R. (2006), *Introduction to Coding Theory*, Cambridge University Press, USA.

Rubinstein, A. (1989), 'The electronic mail game: Strategic behavior under" almost common knowledge"', *The American Economic Review* **79**, 385–391.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell system technical journal* **27**(3), 379–423.

Smith, J. M. & Price, G. R. (1973), 'The logic of animal conflict', *Nature* **246**(5427), 15–18.

Sobel, J. (2015), Broad terms and organizational codes. Unpublished paper, Department of Economics, University of California, San Diego.[1138], Available at https://ipl.econ.duke.edu/seminars/system/files/seminars/1400.pdf.

Sorensen, R. (2023), Vagueness, *in* E. N. Zalta & U. Nodelman, eds, 'The Stanford Encyclopedia of Philosophy', Winter 2023 edn, Metaphysics Research Lab, Stanford University.

Spence, M. (1978), Job market signaling, *in* 'Uncertainty in economics', Elsevier, pp. 281–306.

Szabó, Z. G. (2022), Compositionality, *in* E. N. Zalta & U. Nodelman, eds, 'The Stanford Encyclopedia of Philosophy', Fall 2022 edn, Metaphysics Research Lab, Stanford University.