Center for Mathematical Economics
Working Papers

655

October 2021

Effects of Noise on the Grammar of Languages

Gerrit Bauch

Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
uni-bielefeld.de/zwe/imw/research/working-papers
ISSN: 0931-6558

Unless otherwise noted, this work is licensed under a Creative Commons Attribution 4.0 International (CC BY) license. Further information:

https://creativecommons.org/licenses/by/4.0/deed.en

https://creativecommons.org/licenses/by/4.0/legalcode.en

Gerrit Bauch October 5, 2021

Effects of Noise on the Grammar of Languages[1]

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 languages exist. In equilibrium, sender uses a tessellation consisting of convex cells while receiver converts posterior beliefs into Bayesian estimators serving as interpretations. Shannon entropy measures the noise level and describes to which extent communication is possible. A limit case of errors that respect the distance between words leads to concise interpretations in the decoding process. Comparative statics for different levels of noise reveal which grammatical structures are more robust towards noise. For increasing error separation between most distinct types becomes more important than precision about each single one. Furthermore, distinct words are saved for the description of opposite domains of the type space. Evolutionary modeling approaches converge to equilibria, 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 his tongue, background noise may make it harder to understand or the recipient may suffer from a hearing impairment. Thus, it is 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 that does not break down in presence of small errors. While stochastic error in communication can be analyzed generally, an intuitive approach to this problem is chosen in this paper. Precisely, in order to reflect properties of common day languages, such as English, a mistake in understanding is more likely to arise if words are used that are in a sense close to one another. The more two words differ in letters and pronunciation the less likely it becomes to confound one for the other. Led by this idea, it is natural to endow the set of words with a metric and make the noise depend on it. Having a common language in mind it stands to reason to pick an infinite, possibly high dimensional state space while on the other hand the set of messages or words is limited to a finite amount. This immediately prevents people to reveal perfect information, no matter if there's a common interest or not. It is thus interesting to see how rational individuals cope with this if in addition noise is added to the communication, further deteriorating the circumstances in a way that resembles to day-to-day routines. This situation is modeled by a cheap talk game in which a metric dependent noise confounds the signals in a publicly known way.

Cheap talk game are classically studied in situations of strategic conflict. The seminal paper [CS82] introduced a game of strategic information transmission between an informed sender and an uninformed receiver. In their model, sender learns the state of the world and can send a costless signal

[1]We gratefully acknowledge financial support by the DFG (Deutsche Forschungsgemeinschaft / German Research Foundation) via grant Ri 1128-9-1 (Open Research Area in the Social Sciences, Ambiguity in Dynamic Environments).

to receiver who in turn chooses an action. Both players utilities are depending on the state of the world and the chosen action. The strategic aspect in the equilibrium outcomes stems from a bias in the agents' utility functions. It turns out that in equilibrium, sender will choose a partition, revealing to receiver only to which element of the partition the state belongs. This partition is the coarser the higher the bias of the agents is, thus revealing less information. These partitions consist of convex subsets, even in settings of organizational codes ([Sob15]) or multidimensional applications ([FR11]), i.e. close types use the same word - a property that we will recover under noise and provide a reasoning for.
An introduction of stochastic noise to a game can indeed lead to welfare improvements as observed by [Mye91]. The same observation is made by [BBK07] in a strategic setting generalizing [CS82]. More precisely, they find that low and high levels of noise lead to equilibria that Pareto dominate the equilibria in the no-noise setting of [CS82]. Said differently, noise, although disturbing communication, can create incentives to reveal more information for sender.
In settings of common interest, sender would rather like to fully reveal her type to receiver. While there are instances not to do so in a communication, e.g. in elections ([BS07], [FV20]), the main purpose of a language is to exchange information. In the following are focusing on this aspect. In presence of noise, full revelation is not possible as receiver cannot distinguish between a message purposefully sent by sender or creates by error. In such a cheap talk game, sender and receiver must recalibrate their equilibrium strategies in order to deal with the distortion that pulls receiver to a pooling action, i.e. the equilibrium where information is ignored and communication thus has no effect. The model proposed in this paper follows the spirit of [BBK07] in the common interest setting while differing from it in two ways. Firstly, the message space will be taken finite whereas [BBK07] consider a continuum of messages. Secondly, the structure of the studied noise in this paper will depend on the sent signal to a significant degree, whereas in [BBK07] a fixed and independent noise distribution is mixed in the correct transmission (even if the convex combination can depend on the sent message). Concerning the message space, choosing one or the other setting per se does not pose any technical difficulties. However, having a finite set of messages resembles the proposed model to the one of Voronoi languages [JMR11], which also which will also serve us as the mathematical benchmark. The choice of noise proposed brings in line errors in common day language where it is more likely to mix up similar words.

The structure of the paper is as follows. In section 2 the model is introduced and its assumptions are briefly discussed. Section 3 deals with the informational environment that a language induces. Receiver's posterior beliefs are studied in general and a sharp upper bound on the expected loss of communication is given. Two conceptually different origins of non-beneficial communication are stated. The existence of efficient languages for arbitrary noise channels is discussed in section 4. The best reply correspondence of sender is stated. Furthermore, the restriction to pure strategies is explained and formalized. Section 5 introduces a natural class of noise frameworks that capture the idea that close words are more likely to be confounded. Shannon entropy serves as a measure for different levels or noise. Bayesian updates are possible even for the limiting case of no error, formalizing why slight stammers or spelling mistakes do not disturb a proper understanding. Section 6 analyzes the restriction to quadratic loss function on a Euclidean space. Receiver's best reply is a Bayesian estimator of the state space given the posterior belief. Sender's best communication strategy is built of convex tessellations and generically unique up to null sets. Sections 7 and 8 provide examples that best capture our main intuition that languages have error robust properties. The used word space is the simplest one not covered in the economic literature so far. Properties of languages as well as comparative statics w.r.t. to increasing error are discussed. A foundation of a dynamic evolutionary approach to the proposed model is given in section 9. Equilibria can be learned under many different dynamics, but their outcomes need not be stable. Section 10 concludes.

2 Model and notation

We adapt the setting of Voronoi languages, c.f. [JMR11]: A cheap talk game of common interest between a sender (she) and a receiver (he) is analyzed. Let $T \subsetneq \mathbb{R}^L$, $L \in \mathbb{N}\geq 1$, be a convex and compact set representing the sender's type which can also be thought of as the state of the world that she wants to inform receiver about. An atomless probability distribution μ_0 on T, absolutely continuous w.r.t. the Lebesgue measure on T with strictly positive and continuous density function f_0 is fixed and known to both players. Having any type $t \in T$, sender can choose a word v out of a finite set of messages/words W to be sent to receiver. The new feature of this model is the introduction of a commonly known stochastic error $\varepsilon\colon \mathrm{W} \to \Delta(\mathrm{W})$ that confounds the communication channel, making it possible that not the intended word v is being received, but w with a 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. Receiver interprets the received word w as some point $\alpha(\mathrm{w}) \in T$. Despite the inhibiting factor, we incentivize both players to want the type t and the interpretation $\alpha(\mathrm{w})$ to be as close as possible. To this end, we endow T with a norm $\|.\|$ and weigh the norm-difference in communication by a convex and strictly increasing function $\ell\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$. Thus, given (pure) strategies $\pi\colon T \to \mathrm{W}$ (measurable w.r.t. μ_0) and $\alpha\colon \mathrm{W} \to T$, players seek to minimize their expected loss in communication

$$\begin{aligned}\mathscr{L}(\pi,\alpha) :&= \mathbb{E}_{\mu_0}[\mathbb{E}_{\varepsilon(.\,|\,\pi(t))}[\ell(\|t-\alpha(\mathrm{w})\|)]] \\ &= \int_T \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\pi(t)) \cdot \ell(\|t-\alpha(\mathrm{w})\|)\,\mu_0(\mathrm{d}t).\end{aligned}$$

We call π a *communication device* and α an *interpretation (map)*. Note that null sets play no role for the integral. Furthermore, if there is no error, i.e. $\varepsilon(\mathrm{w}\,|\,\mathrm{v}) = 1$ if and only if $\mathrm{w} = \mathrm{v}$, the expected loss and hence the analysis collapses to the one in [JMR11], thus constituting a proper generalization.

We briefly discuss the reasons for choosing these requirements.The compactness of T together with the continuity of ℓ ensure integrability. Convexity of T guarantees that all optimal actions lie within T, i.e. can be played. Convexity along with monotonicity of ℓ prevents receiver from hedging by using randomized responses, as his optimal action becomes unique.

From a mathematical and game theoretic point of view, it is important to immediately clarify whether or not the minimization problem over all possible strategy profiles (π,α) is well-posed and if we need to allow for mixing. Any such solution is considered an *efficient language*. Fortunately, efficient languages always exist and can be assumed to be pure ones (see Theorem 3 and Lemma 4). As the concrete strucure of efficient outcomes is not easy to elicit, the game theoretic solution concept of perfect Bayesian Nash equilibria is used to give necessary conditions. To this end, the best reply correspondence is to be determined in the first place. For now, the general problem is studied from the point of view of posterior beliefs, pinning down the receiver's best reply.

3 Induced beliefs and receiver's best reply

A first step towards understanding the game is to pin down the best replies of both players in order to characterize Bayesian Nash equilibria. Since in games of common interest strategy profiles leading to an efficient outcome are necessarily Nash equilibria, we thus get a better understanding of efficient languages.

To this end, assume that receiver knows the sender's communication device $\pi\colon T \to \mathrm{W}$. Then, upon observing w, receiver's informational environment changes, leading to a new belief, called *posterior*

belief which is obtained by Bayes rule. To write it down we begin by noting that f_0 and π induce a distribution over received words which is given by

$$\lambda^\pi(\mathrm{w}) := \mathbb{E}_{\mu_0}[\varepsilon(\mathrm{w} \,|\, \pi(t))] = \int_T \varepsilon(\mathrm{w} \,|\, \pi(t))\, \mu_0(\mathrm{d}t). \tag{1}$$

The value $\lambda^\pi(\mathrm{w})$ represents the expected probability with which receiver will actually receive the word w if sender uses the communication device π.

If we think of ε as representing some possible error in communication, we may assume that there is always a small but positive probability of making any error. Following this, we often impose $\lambda^\pi(\mathrm{w}) > 0 \,\forall\, \mathrm{w} \in \mathrm{W}$ for convenience, discussing general cases within the proofs which can be found in the appendix. In the mentioned case, λ^π defines a fully supported probability measure over W. Thus, receiver can always use Bayes rule to update his prior belief to the posterior μ^π_{w} with density function

$$f^\pi_{\mathrm{w}}(t) := \lambda^\pi(\mathrm{w})^{-1} \cdot f_0(t) \cdot \varepsilon(\mathrm{w} \,|\, \pi(t)). \tag{2}$$

That is, knowing π and receiving signal w, receiver uses this information to re-evaluate the chances that sender is of some type t by applying μ^π_{w}.

One immediate observation is, 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 $\{\mu^\pi_{\mathrm{w}}\}_{\mathrm{w}\in\mathrm{W}}$. Then we have the following property which is referred to as *Bayes-Plausibility* in the setting of Bayesian Persuasion, c.f. [KG11]:

$$\sum_{\mathrm{w}\in\mathrm{W}} \lambda^\pi(\mu^\pi_{\mathrm{w}}) \cdot \mu^\pi_{\mathrm{w}} = \mu_0. \tag{3}$$

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

$$\mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[X]] = \mathbb{E}_{\mu_0}[X]. \tag{4}$$

Intuitively, by means of a communication device one can only induce such posterior beliefs that are in expectation (w.r.t. λ^π) the prior belief μ_0.

Before using the gathered knowledge about receiver's informational environment, we first state an additional way of writing down the loss functional by means of the introduced notations:

$$\begin{aligned}
&\mathscr{L}(\pi, \alpha) \\
&= \mathbb{E}_{\mu_0}[\mathbb{E}_{\varepsilon(.\,|\,\pi(t))}[\ell(\|t - \alpha(\mathrm{w})\|)]] = \int_T \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(w \,|\, \pi(t)) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\, \mu_0(\mathrm{d}t) &(5)\\
&= \sum_{\mathrm{w}\in\mathrm{W}} \lambda^\pi(\mathrm{w}) \cdot \int_T \lambda^\pi(\mathrm{w})^{-1} \cdot \varepsilon(w \,|\, \pi(t)) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\, \mu_0(\mathrm{d}t) \\
&= \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\ell(\|t - \alpha(\mathrm{w})\|)]]. &(6)
\end{aligned}$$

While expression (5) describes the expected loss by a weighted sum of the deficits that occur due to the error for each realized type, term (6) can be interpreted as counting the expected loss under each of the posteriors and assessing them under the accumulated probability with which the posterior is induced.

Especially, in terms of optimization, the last term just requires receiver to choose an optimal interpretation for any (induced) posterior belief: Having any (posterior) belief $\mu \in \Delta(T)$ with continuous and strictly positive density function f, receiver optimally responds by choosing the unique minimizer

$$\hat{s} \in \arg\min_{s\in T} \mathbb{E}_\mu[\ell(\|t - s\|)]. \tag{7}$$

The uniqueness arises due to the convexity assumption on ℓ. Henceforth, denote by $\hat{\alpha}(\mu)$ this unique minimizer. We can thus define $\hat{\alpha}(\mathrm{w}) := \hat{\alpha}(\mu_{\mathrm{w}}^{\pi})$ if π is understood and refer to $\hat{\alpha}$ as the unique best reply of receiver. This way, the profound connection between an action being taken as a response of either a word heard and interpreted by means of linguistics or an updated belief induced by the received word by means of game theory, is stressed. Furthermore, having a unique solution to the minimization problem, receiver will play a pure strategy in equilibrium.

If communication was not beneficial, the formation and use of languages would be questionable. It is thus of importance to compare the findings of the presented model to the situation without the possibility to send signals. In this default setting, it is only up to receiver to take an action and the only information at hand for him is the common prior μ_0. He thus optimally picks the *default action* or *pooling action* $\alpha_0 := \alpha(\mu_0) = \arg\min_{s\in T} \mathbb{E}_{\mu_0}[\ell(\|t-s\|)]$ and the *default loss* $L_0 := \mathbb{E}_{\mu_0}[\ell(\|t-\alpha_0\|)]$ is realized.

Proposition 1. *Given any communication device π, we have $\mathscr{L}(\pi,\hat{\alpha}) \leq L_0$. The inequality is strict if and only if there is a word $\mathrm{w}\in\mathrm{W}$ with $\lambda^{\pi}(\mathrm{w})>0$ and $\hat{\alpha}(\mathrm{w})\neq\alpha_0$.*

In words, as long as receiver knows the communication device π the presence of signals can not be detrimental.

We conclude this section by stating two readily verified conditions under which communication is not beneficial.

Corollary 2. *If π is constant or if ε is the constant uniform distribution on W, then $\mu_{\mathrm{w}}^{\pi}=\mu_0$ for all $\mathrm{w}\in\mathrm{W}$. Especially, $\hat{\alpha}\equiv\alpha_0$ and thus $\mathscr{L}(\pi,\hat{\alpha})=\mathscr{L}_0$.*

Although easy and intuitive, this result demonstrates that there are two sources that can lead to non-profitable communication. Firstly, if the communication device is not meaningful, i.e. does not provide receiver with additional information in any realization. Secondly, if the error channel does not convey any information. If everything is equally likely to be received, no matter what is sent, any kind of communication is useless. While the latter problem is to be considered an exogenous problem of the environment, the first one lies within the reach of strategic choices on the side of the sender. However, not any non-constant communication device leads to a profitable language, even if it changes the informational environment of receiver, see Example 10.

4 Noise equilibria and efficient languages

An ideal outcome of cooperation is now being analyzed. As mentioned in [JMR11] one can think of rational players coordinating their strategies before playing the game in a meta-language with the aim to minimize the loss functional $\mathscr{L}(\pi,\alpha)$. A *language* (π,α) is called *efficient* if it minimizes the expected loss $\mathscr{L}(\pi,\alpha)$ over all (π,α). In the presented setting, this necessarily requires both agents to minimize the occurring loss over their own action sets, i.e. to play a best reply each. Any such Nash equilibrium is called *noise equilibrium* ([BBK07]). The examples in section 8 show, that being a noise equilibrium is not sufficient to be efficient.

While the best reply of receiver is always uniquely determined, the sender's response can obviously always be arbitrarily perturbed on a null set. More importantly for an economic interpretation, in the interim stage, sender might be indifferent between sending two or more different words. This can only be because the corresponding actions taken by receiver in response amount to the same loss. To see this more precisely, fix any interpretation $\alpha\colon \mathrm{W}\to T$ of receiver. Then, by focusing on interim equilibria[(2)] a type t-Sender can pick any word $\mathrm{v}\in\mathrm{W}$ to be sent out of the non-empty set

$$\arg\min_{\mathrm{v}'} \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}')\cdot\ell(\|t-\alpha(\mathrm{w})\|). \tag{8}$$

[(2)] Which is w.l.o.g., as the marginal distributions have full support.

Applying suitable choices in the presence of multiple minimizers (see the proof of Theorem 3 for details) we can derive possible partitions $C^\alpha = \{C^\alpha_{\mathrm{v}}\}_{\mathrm{v}\in\mathrm{W}}$ of T, where each C^α_{v} is (Lebesgue-)measurable and consists only of types where v is a minimizing response of sender given α. Any of these partitions can be used to define a best reply communication device by setting $\pi^{C^\alpha}(t) = \mathrm{v}$ 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, two best replies constructed by means of such partitions may differ perceptibly. However, for the Euclidean norm such cases do not occur generically (see Proposition 13) as long as interpretations differ. This will be discussed in the subsection of section 8.

Employing any such a best reply for sender, one can prove the existence of efficient languages and thus of noise equilibria.

Theorem 3. *Efficient languages (π, α) exist.*

We conclude this section by arguing why there is no reason to include the possibility of mixing for languages. Before this detour is started, remember that a purely cooperative setting is being analyzed in which preferences are aligned and it would be most preferred to sender if she could reveal her true type to receiver. By introducing new randomness into the signaling or interpretation procedure coordination is harder to achieve while also the convexity of the loss functional ℓ prevents the agents from any kind of hedging.

To begin the formal analysis, one can restrict to pure strategies of receiver as he always favors the pure strategy that consists of Bayesian estimators for each induced belief. Focusing on sender now, denote by $\sigma\colon T \to \Delta(\mathrm{W})$ a mixed strategy, specifying a probability $\sigma(\mathrm{w}|t)$ of sending $\mathrm{w} \in \mathrm{W}$ if she is of type $t \in T$. Thus, expected payoff is given by

$$\mathscr{L}(\sigma, \alpha) = \int_T \sum_{\mathrm{v}\in\mathrm{W}} \sigma(\mathrm{v}|t) \cdot \sum_{\mathrm{w}\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|)\ \mu_0(\mathrm{d}t). \tag{9}$$

This equation immediately shows that sending any pure $\mathrm{v} \in \arg\min_{\mathrm{v}'} \sum_{\mathrm{w}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|)$ is weakly preferred by a type-t sender, even strictly if $\sigma(.|t)$ assigns a positive probability to any $\widetilde{\mathrm{v}}$ not being a minimizer. This result is summarized in Lemma 4.

Lemma 4. *For any mixed language (σ, τ) with strategies $\sigma\colon T \to \Delta(\mathrm{W})$ and $\tau\colon \mathrm{W} \to \Delta(T)$ there is a pure language (π, α) with weakly smaller loss $\mathscr{L}(\pi, \alpha) \le \mathscr{L}(\sigma, \tau)$.*
If τ is non-degenerated, the inequality can even be made strict by using receiver's best reply $\hat{\alpha}$.
Especially, the payoff of any efficient language is attained by a pure one.

5 The q-ary symmetric error channel of length n

In the following we will give an example of a class of error functions that is usually studied in information theory: The symmetric q-ary channel (c.f. [Rot06]). In order to apply this framework, the word space is resembled to actual words and expressions, making it accessible to both, settings in computer sciences and applications in linguistics. Another argument for this setup to be a canonical choice is made by the fact that it behaves well in comparative static analysis, e.g., if the error is approaching zero, see Proposition 6, giving at hand Bayesian updates even in cases when they are formally not defined.

To start with, it is reasonable to give the space of words W a measure of distance between words in order to distinguish how similar or different two words are. To this end, a metric d is bestowed upon W, making it a metric space. As to confound similar words is more likely, it is assumed that the error

[(3)] Figure 1 in [JMR11] gives an example for this using the maximum norm.

should depend only on the distance of the words and is the larger the further away two words are from one another.

Our proposed notion takes a finite alphabet $\mathcal{A}$ whose elements we interpret as letters (a,b,c,...) or expressions (tiny,tall; left, right; etc). The message space consists of a sequence of letters of length n, i.e. $\mathrm{W} = \mathcal{A}^n$. The *Hamming distance*

$$d\colon \mathrm{W}\times\mathrm{W}\to\mathbb{N}_0, ((\mathrm{w}_k)_k,(\mathrm{v}_k)_k)\mapsto \#\{k\in\{1,\dots,n\}\,|\,\mathrm{w}_k\neq\mathrm{v}_k\}, \tag{10}$$

first introduced and studied in [Ham50], is used as a measure of distance. Thus, words are to be considered further away from one another the more letters in the order of appearance differ. It is worth noting that the Hamming distance plays a crucial role in any applied fields related to information theory, especially coding theory (c.f. [Rot06]).

Turning to the definition of an error, we start by defining it on letters. To this end we set $\widetilde{\varepsilon}\colon \mathcal{A}\to\Delta(\mathcal{A})$ to be the function

$$a\mapsto\widetilde{\varepsilon}(.\,|\,a),\quad \widetilde{\varepsilon}(b\,|\,a):=\widetilde{\varepsilon}(a)(b)=\begin{cases}1-p & ,\, b=a,\\ \frac{p}{\#A-1} & ,\, b\neq a\end{cases}, \tag{11}$$

where the exogenous parameter $p\in[0,1]$ is the probability of wrongly transmitting the intended letter a and in case of an error, each of the other $\#\mathcal{A}-1$ symbols is equally likely received. For an alphabet of length q (in our case $q=\#\mathcal{A}$), this type of error structure is called symmetric q-ary symmetric channel (c.f. [Rot06]). It is a well-known noise channel that is used to model error transitions in telecommunication, data storage, but also finds application in DNA heritage of cell-divisions (c.f. [Mac02], [CT06]).

Assuming, that the error of transmitting a letter is independent of the symbols transmitted before, we can extend the error channel on W by gluing n independent copies of $\widetilde{\varepsilon}$ together. The result is called q-ary symmetric channel of length n Using the Hamming distance, this can explicitly be written as follows.

$$\begin{aligned}&\varepsilon\colon \mathrm{W}\to\Delta(\mathrm{W}),\\ &\qquad \mathrm{v}\mapsto\varepsilon(.\,|\,\mathrm{v}),\quad \varepsilon(\mathrm{w}\,|\,\mathrm{v})=(1-p)^{n-d(\mathrm{v},\mathrm{w})}\cdot\left(\frac{p}{\#\mathcal{A}-1}\right)^{d(\mathrm{w},\mathrm{w}')},\end{aligned} \tag{12}$$

In words, $\varepsilon(\mathrm{w}\,|\,\mathrm{v})$ refers to the probability that w is received if v is sent and it only depends on the Hamming distance $d(\mathrm{v},\mathrm{w})$ and the letter transmitting error p. Noting that 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, the probability distribution of the family $(\{\mathrm{w}\in\mathrm{W}\,|\,d(\mathrm{w},\mathrm{v})=d\})_{d=0}^n$ follows a binomial distribution. However, as we are interested in the transition probability from v to a particular w rather than to such a class of words, ε is itself the natural probability distribution to consider for us.

Using the notation $m:=\#\mathcal{A}-1$ and $\widetilde{p}:=\frac{p}{(1-p)m}$ we can rewrite $\varepsilon(\mathrm{w}\,|\,\mathrm{v})=(1-p)^n\cdot\widetilde{p}^{d(\mathrm{w},\mathrm{v})}$ which is often convenient.

Note that for fixed p, $\varepsilon(\mathrm{w}\,|\,\mathrm{v})$ only depends on $d(\mathrm{w},\mathrm{v})$. By symmetry of the metric, this in turn implies that $\varepsilon(\mathrm{w}\,|\,\mathrm{v})=\varepsilon(\mathrm{v}\,|\,\mathrm{w})$ for all $\mathrm{v},\mathrm{w}\in\mathrm{W}$. Thus, there are mathematically indistinguishable twin languages created by isometries, i.e. metric-preserving bijections of (W,d). When analyzing examples, this enables us to restrict our attention to a particular generic word v as for any word v' there is an isometry mapping v' to v. The downside of this is that we have to give up uniqueness of any equilibrium and, although we are able to pin down geometric structures, will not be able to explain the forming or evolution of one particular language. This problem can be tackled by resolving symmetries by dynamic learning as is done in [Blu04].

The following result about the introduced error channel is immediate, but captures the qualitative connection between the metric between words and the probability distribution with which one is confounded for the other for different levels of the single error probability p.

Lemma 5. *Let* $\mathrm{v} \in \mathrm{W}$ *be sent by sender.*

(i) If $p = 0$ *receiver will receive* v *with probability* 1*, i.e. there is no error and the model becomes the one for Voronoi languages.*

(ii) If $0 < p < \frac{m}{m+1}$ *the probability of receiving a word* $\mathrm{w} \neq \mathrm{v}$ *is decreasing in the distance* $d(\mathrm{w}, \mathrm{v})$*, i.e.* $d(\mathrm{w}, \mathrm{v}) > d(\mathrm{w}', \mathrm{v}) \implies \varepsilon(\mathrm{w} \,|\, \mathrm{v}) < \varepsilon(\mathrm{w}' \,|\, \mathrm{v})$.

(iii) If $p = \frac{m}{m+1}$*, there is no information to be gained from communication as* ε *is constant and equal to the uniform distribution.*

(iv) If $\frac{m}{m+1} < p < 1$ *the probability of receiving a word* $\mathrm{w} \neq \mathrm{v}$ *is increasing in the distance* $d(\mathrm{w}, \mathrm{v})$*, i.e.* $d(\mathrm{w}, \mathrm{v}) > d(\mathrm{w}', \mathrm{v}) \implies \varepsilon(\mathrm{w} \,|\, \mathrm{v}) > \varepsilon(\mathrm{w}' \,|\, \mathrm{v})$.

(v) If $p = 1$ *receiver will receive a word* w *with maximum distance* $n = d(\mathrm{w}, \mathrm{v})$ *to* v *and any such with equal probability.*

Starting from the classical case without error (i.e. $p = 0$), for values $0 < p < \frac{m}{m+1}$ words are more likely to be received (or sent from the receiver's point of view) that are closer to the word sent (received). Especially it is most likely to receive the unique word with distance 0 to v, i.e. $\mathrm{w} = \mathrm{v}$ itself. Reaching the threshold $p = \frac{m}{m+1}$, any communication using this channel is useless, see corollary 2).

Interestingly, for $\frac{m}{m+1} < p \leq 1$ informative communication can take place again, even though in general less efficiently then before. This is due to the fact that receiver now puts more weight on the event that the received word stems from one among those having the maximal distance to the it, extracting some information but not as much as if a single word would be the most likely one.[4] This feature is illustrated and further discussed in the subsection about entropy where we provide a quantitative measure for the information permitted by the channel.

Most often, we reckon the single error probability to fulfill $0 < p < \frac{m}{m+1}$ as there is always reason to assume some error, but not too much. The upcoming proposition to studies the limit cases towards both border. To start with, a short calculation reveals that the investigated error function induces posteriors beliefs with densities of the form

$$f_{\mathrm{w}}^{\pi}(t) = f_0(t) \cdot \left(\int_T \widetilde{p}^{\,d(\mathrm{w},\pi(t'))-d(\mathrm{w},\pi(t))} \; \mu_0(\mathrm{d}t') \right)^{-1}. \tag{13}$$

Note that the integrand is continuous in $\widetilde{p}$.

Proposition 6. *Let* π *be a communication device, which is known to receiver,* $p \in (0, \frac{m}{m+1})$ *and* $\mathrm{w} \in \mathrm{W}$ *be the observed word by receiver. Then the following properties hold.*

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

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

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

[4] For binary channels, i.e. $m = 1$, the roles of the two letters simply switch and there is no such loss in information, whereas for $m > 1$ there is.

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

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 error is getting infinitesimally small. In the presented non-discrete type space setting[(5)] it is important to keep track of null sets but can still be explained. To start with, after receiving w and considering t as a possible type, receiver determines the closest words to w that are sent by a set of sender types of positive probability w.r.t. μ_0. This distance is called d^*. Let p be close to 0 now so that the probability of wrongly transmitting a word is close to 0.

If $d^* < d(\mathrm{w}, \pi(t))$ then it is unreasonable for receiver to assume that sender is of type t as there is a set of sender types with positive probability mass that is sending words closer to w then type t does.

If $d^* = d(\mathrm{w}, \pi(t))$ then there is no set of sender types with positive probability mass that send words strictly closer to w than type t does. As there is an arbitrarily small error probability, receiver will not consider types that send a word even further away from w than type t. He only believes that types t' with $d^* = d(\mathrm{w}, t')$ are possible and forms the update over all those types. This even holds true if $\mathrm{w} \notin \pi(T)$ which is surprising as if compared to the classical setting (i.e. $p = 0$) there would be no possible Bayesian update: If an English receiver hears the word "orunge" while assuming an arbitrarily small error would correctly conclude that "orange" would be the word sent.

However, the case $d^* > d(\mathrm{w}, \pi(t))$ makes it impossible to use Bayes rule as receiver neither expects w to be sent with positive probability nor does he believe that type t or any other type t' that sends a word in distance $d(\mathrm{w}, \pi(t))$ can be the source of the received word. Although this is a drawback to the promising former point, for a finite type space this last case disappears.[(6)]

Entropy

In the following we provide a quantitative measure of the noise ε in terms of the single letter error probability p by using (Shannon) entropy [Sha48]. Intuitively, the more noise, the harder it gets for Receiver to properly decode an observed message. We will see that the noise is maximal for $p = \frac{m}{m+1}$ and strictly monotonically increasing for both $p \nearrow \frac{m}{m+1}$ and $p \searrow \frac{m}{m+1}$. While $p = 0$ always corresponds to a non-noisy channel, for $p = 1$ not all uncertainty in the channel can be resolved provided a non-binary alphabet, i.e. $m > 1$.

Let us start with a short introduction and discussion of entropy. Formally, for a discrete probability measure P on a finite set X, entropy is defined as follows.

$$H(P) = -\sum_{x \in X} P(x) \cdot \log(P(x)), \tag{14}$$

where the base choice of the logarithm is a question of normalization and usually chosen to be $\#X$ which is convenient for our setting later. It is the average of the *information content* $-\log(P(x))$ which describes how surprising the observation of an element x is given its probability $P(x)$. It thus associates a value of average surprisal or uncertainty to the distribution.

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.

(5) The analogous finite type set case is not that cumbersome in this regard.

(6) For fully supported μ_0 any type t occurs with probability > 0 thus we always have $d^* \leq d(\mathrm{w}, \pi(t))$.

In our setting, for any arbitrary but fixed sent word v and any error probability $p \in [0,1]$, $\varepsilon(.\,|\,\mathrm{v})$ defines a probability distributions on the set W. As the choice of v leads only to a permutation of the probabilities across W and the symmetry of H, $H(\varepsilon(.\,|\,\mathrm{v}))$ does not depend on v and can thus be defined as $H_\varepsilon(p) = H(\varepsilon)$ only in dependence on the single error probability p. We have the following proposition, characterizing properties of $H_\varepsilon(p)$.

Proposition 7. *The entropy of the error channel is*

$$H_\varepsilon(p) = -n \cdot (p\log(p) + (1-p)\log(1-p)) + n \cdot p\log(m). \tag{15}$$

It is a concave in $p \in [0,1]$ with the 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 quantifies the observations we made in Lemma 5. 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. This is interpreted as each message being equally likely received and thus the channel conveys no information. For $p = 0$ entropy is zero, showing that there is no noise and each piece of information is transmitted truthfully and can be correctly decoded. If $p = 1$ entropy is $\log(m/(m+1))$, telling us how much information is lost. For $m = 1$, i.e. a binary alphabet, this expression is again zero. This makes sense since the roles of the letters simply swap. For $m > 1$ however, we can see that although information can be recovered and thus reasonable communication takes place, there is still noise left and 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. I.e., for $p \nearrow$ communication gets hindered more and more on $[0, \frac{m}{m+1}]$, while afterwards communication gets facilitated again.

We conclude the discussion about the error channel by an example that illustrates the relation between an increasing noise in terms of entropy and the corresponding loss for a fixed communication device with a non-binary alphabet.

Example 8. Let $T = [-\frac{1}{2}, \frac{1}{2}]$ with μ_0 being the uniform distribution. Let $\mathrm{W} = \{L, M, R\}$ and fix $\pi\colon T \to \mathrm{W}, \pi([-\frac{1}{2}, 0]) = L, \pi((0, \frac{1}{2}]) = R$. For $p \in [0,1]$ the optimal response $\hat{\alpha}$ of Receiver is given by

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

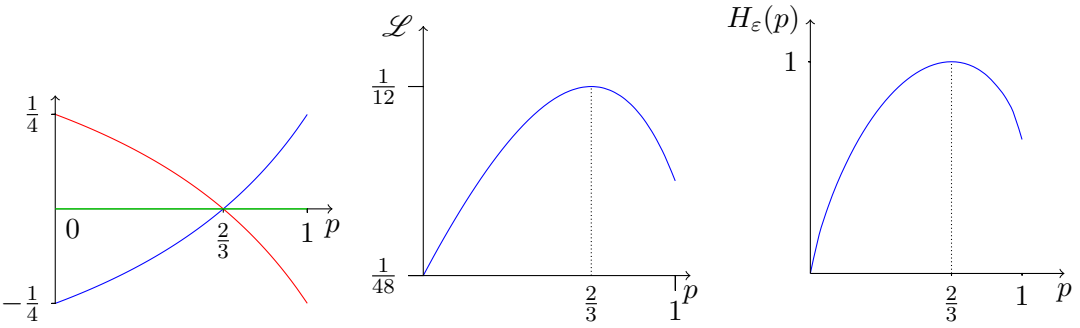
and the expected loss in dependence of p is given by

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

A graphical illustration can be found in figure 1.

It's worth noting that for $p \to 1$ Receiver can perfectly decode L and R as R and L respectively. However, he now also received message M with positive probability, but cannot determine from which originally sent message this observation stems. Thus, communication under $p = 1$ is worse than for $p = 0$. In numerical terms, we find that $\mathscr{L}(1)/\mathscr{L}(2/3) = 0.625 \approx 0.63 \approx H_\varepsilon(1)$, thus entropy roughly captures the expected loss.

Figure 1: Three Letters on a uniform Interval



6 Quadratic loss function

In this section, we analyze the structure of the loss aggregation by restricting our attention to the Euclidean norm $\|.\| = \|.\|_2$, induced by the scalar product $\langle .\,, .\rangle$ and a quadratic loss $\ell(x) = x^2$ which is common in cheap talk games (following [CS82]). In this case, for any posterior belief the optimal response of receiver can be written down explicitly. It is a Bayesian estimator for the whole type set T according to the induced posterior belief.

Lemma 9. *Having any (posterior) belief μ, with continuous density $f > 0$, receiver's unique best action is given by*

$$\hat{\alpha}(\mu) = \mathbb{E}_\mu[t]. \tag{16}$$

Thus, the expected loss of a publicly known communication device π is

$$\mathscr{L}(\pi, \hat{\alpha}) = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\left\|t - \mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t']\right\|_2^2]] = \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2]. \tag{17}$$

Furthermore, receiver plays on average the default action

$$\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \alpha_0.$$

Lemma 9 gives at hand a nice formula to calculate the expected loss of a communication device, separating the problem into several steps. On the one hand, there is a language independent threshold that stems from the probabilistic setting alone. On the other hand, the split of square norms of the optimal interpretations induced by π is the determinant factor of the quality of the communication device. This indicates that optimal languages are the ones maximizing this spread. In other words, efficient languages are "as separable as possible". Adding and subtracting the term $\|\alpha_0\|_2^2 = \|\mathbb{E}_{\mu_0}[t]\|_2^2$ to the expected loss we can write it as

$$\mathscr{L}(\pi, \hat{\alpha}) = L_0 - \left(\mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2] - \|\alpha_0\|_2^2\right). \tag{18}$$

This expression tells us how much better the communication device π is in comparison with the setting of no (reasonable) communication.

Recall Corollary 2 where two rather restrictive necessary conditions for a profitable communication device have been stated: Under both conditions the induced beliefs μ^π_{w} where shown to be the same and equal to the common belief μ_0. This in turn implies that the induced actions are the same and equal to the default action. One now might ask whether these conditions are sufficient and if any non-profitable communication device always induces the same beliefs. However, the following counterexample illustrates that neither one of the two statements is true.

Example 10. Let $T = [0,1], \mu_0 \sim \mathcal{U}[0,1], \mathrm{W} = \{A,B\}^2$ and take the symmetric binary channel with single error probability $0 < p < \frac{1}{2}$, i.e. $0 < \widetilde{p} < 1$. Consider the non-constant communication device

$$\pi \colon T \to \mathrm{W}, \omega \mapsto \begin{cases} BB & , t \in [\frac{1}{3}, \frac{2}{3}], \\ AA & , \text{otherwise.} \end{cases}$$

Some calculations reveal

$$f^{\pi}_{AA}(t) = \begin{cases} \frac{3\widetilde{p}^2}{2+\widetilde{p}^2} & , t \in [\frac{1}{3}, \frac{2}{3}], \\ \frac{3}{2+\widetilde{p}^2} & , \text{otherwise} \end{cases}, \quad f^{\pi}_{BB}(t) = \begin{cases} \frac{3}{2+\widetilde{p}^2} & , t \in [\frac{1}{3}, \frac{2}{3}], \\ \frac{3\widetilde{p}^2}{2+\widetilde{p}^2} & , \text{otherwise} \end{cases}$$

as well as $f^{\pi}_{AB} = f^{\pi}_{BA} = f_0$. Hence $\mu^{\pi}_{AA}, \mu^{\pi}_{BB} \neq \mu_0$ and $\mu^{\pi}_{AB} = \mu^{\pi}_{BA} = \mu_0$. However, $\hat{a}(\mu^{\pi}_{\mathrm{w}}) = \hat{a}(\mu_0) = \frac{1}{2}$ for all w, i.e. the induced actions and therefore the payoff (irrespectively of the error p), are always the same as in the setting without communication. Thus, conveying different information optimal interpretations may be the same. If, e.g., the received word is AA, receiver assigns a higher probability to the state lying in the complement of $[\frac{1}{3}, \frac{2}{3}]$ then in the mentioned interval itself. However, the suboptimal symmetric structure of the language does not allow him to get any benefit out of this updated belief. The expected state still lies at $\frac{1}{2}$, which will be his action taken, leading to the default loss. Note the action choice of receiver makes all communication devices a best reply for sender, including the proposed one. Thus, this language defines a noise equilibrium, but not a strict one.

Example 10 can thus be taken as a bad choice of a language. Note that the cell for AA is not convex. As we will see in the next results, this is an indication that we are either not having a noise equilibrium or (which was the case in the example) that sender is indifferent between at least two words on a set of positive probability mass, limiting a fruitful communication in these cases. More interestingly, we will rediscover a crucial property of Voronoi languages with regard to linguistics: If the Euclidean norm is used, the set of types for which a particular word v is the unique optimal choice by sender constitutes a convex set. This provides evidence for the linguistic conjecture that simple words have convex categories (c.f. [Jä07]) even in the presence of error.

Proposition 11. *Let α be an interpretation. For any word $\mathrm{v} \in \mathrm{W}$, the set of types for which v is a respectively the best reply by sender are convex sets. In the first case this set is a closed subset of T, in the latter an open one.*

The convexity result does not hold true for an arbitrary choice of the norm as the following example demonstrates.

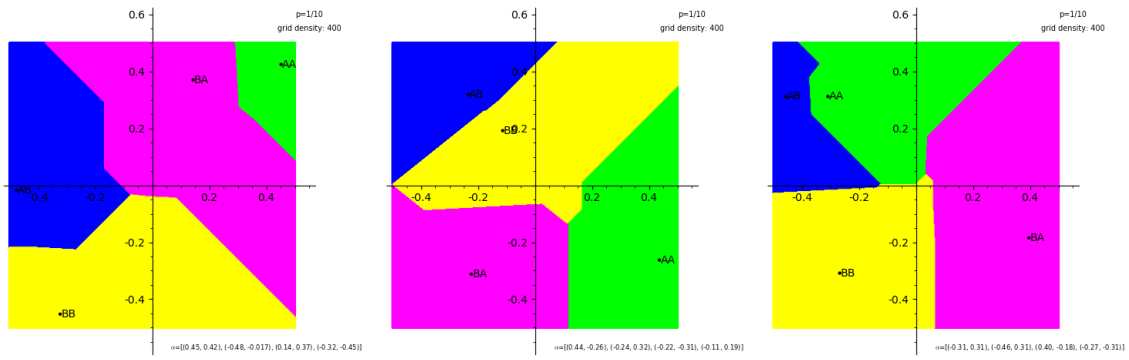
Example 12. Consider a unit square with uniform prior distribution, $\mathrm{W} = \{A,B\}$, $\ell \equiv \mathrm{id}$ and $\|.\| = \|.\|_{\infty}$. Then, the best tessellation corresponding to different receiver strategies are in general non convex as the numerical simulations in figure 2 show.

Having studied a special error function and the quadratic loss case with Euclidean norm separately, we will end this section by turn towards the interplay of the two.

Proposition 13. *Let α be an interpretation. For any pair of words $\mathrm{v}, \mathrm{v}' \in \mathrm{W}$, the set of sender types for which she is indifferent between sending v and v' is either a null set or T. The latter case can only occur for at most $n-2$ values of $p \in (0,1)$ if $\alpha(\mathrm{v}) \neq \alpha(\mathrm{v}')$.*

To conclude this section, we will give another, yet for the moment again not efficient, example of a noise equilibrium in the one dimensional case. The following language will constitute an almost surely strict noise equilibrium while not using a *full vocabulary*, i.e. not all of the available words.

Figure 2: Optimal cells for the maximum norm



Example 14. In the setting of Example 10 consider the communication device

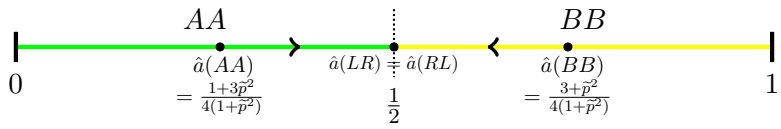
$$\pi\colon \Omega \to \mathrm{W}, \omega \mapsto \begin{cases} BB & ,\omega \in [\frac{1}{2}, 1], \\ AA & ,\text{otherwise.} \end{cases}$$

Optimal actions can be calculated to be

$$\hat{\alpha}(AA) = \frac{1+3\widetilde{p}^2}{4(1+\widetilde{p}^2)}, \quad \hat{\alpha}(BB) = 1 - \hat{\alpha}(BB), \quad \hat{\alpha}(AB) = \hat{\alpha}(BA) = \frac{1}{2}.$$

The tessellation and optimal actions are illustrated in Figure 3.

Figure 3: Strict Nash but not full vocabulary



Using this and equation (8), we find that any type-t sender with $t \in [0, \frac{1}{2})$ prefers to send AA over BB as well as over AB. By symmetry arguments we thus find that the optimal communication device, given the above interpretations, is uniquely π up to $t = \frac{1}{2}$ - a null set. The proposed language is thus a strict noise equilibrium under which the communication device does not use all possible words (with positive probability).

This indicates a violation of Theorem 2 in [JMR11] where a classification of strict Nash equilibria is given by Voronoi languages with full vocabulary. However, this finding is not that surprising as the presence of an error lets each word be received with positive probability and implies a unique optimal reply by receiver. Thus, the freedom of choice on words, that are not actually sent to (and for $p = 0$ also not received by) receiver must be given up.

7 Separation over Precision

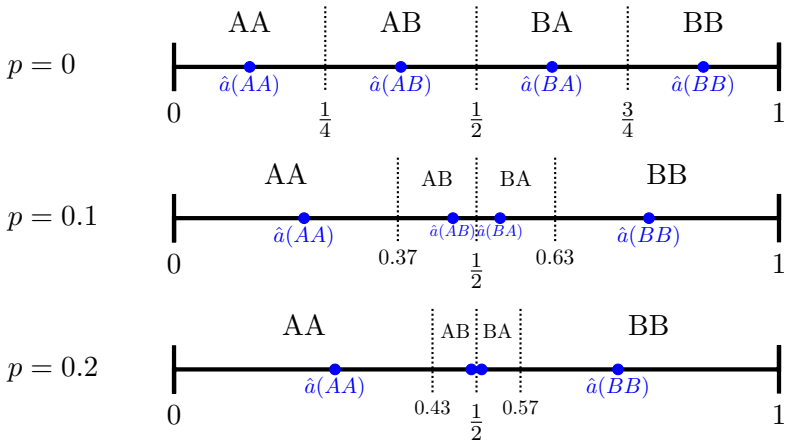
Only a few words are really necessary to capture the main idea of a sentence. The remaining words are either decorative, concerned with details or simply redundant. Emphasizing the important words thus becomes thus more important the more likely communication may be flawed. The extreme case is that of coding theory, where the sender deliberately refrains to use all accessible words so that receiver has a chance to spot and correct small errors. This technique is of crucial importance not

only in telecommunication but also in all kinds of storage devices. Obviously, this comes at the price of having some kind of redundancy in the communication as it is no longer theoretically able to convey as much information as possible.

This waiving of additional information that could be transmitted can also be found within our framework. In the following example, sender gives up precision over types that are close to the pooling action in order to prevent confusion about states that would come with a large loss if they are mistaken.

Example 15. Let $L = 1$ and $T = [0, 1]$ endowed with the uniform distribution. The word space is given by $\mathrm{W} = \{A, B\}^2$. For different values of p, we can categorize efficient languages. For some error levels, these are depicted in Figure 4.

Figure 4: Efficient languages



We observe that most opposite types (those close to 0 and 1) are separated by words that are less easily confused (here AA and BB). This lowers the probability of huge losses resulting from a confounding of very distinct types. While all messages are used, we see that for increasing noise the words that mark the boundaries of the interval absorb more and more space and that the words used to describe types close to the pooling equilibrium are less often used. That indicates that for higher levels of noise, it becomes more reasonable to separate clearly extreme types than keeping up a uniform labeling that could potentially be used to describe the state space with higher precision. This is achieved by putting more weight on clearly distinguishable words and decreasing the use of words that result in more likely spillovers from and to the former.
Imagine that sender wants to inform receiver about the height of a person, where 0 and 1 refer to the tiniest or tallest possible option. The used words can be thought of as "very tiny" (AA), "rather tiny" (AB) and so forth. Note that already the individual letters are endowed with an intrinsic meaning, e.g. A describing the property "small'. The distinction between AB and BA is non canonical and could be swapped by symmetry. Now, as can be seen for increasing values of $p \in [0, \frac{1}{2}]$, sender puts the more emphasize on the extreme cases than on the average sized persons, relinquishing on precision in order to prevent misunderstandings that would lead to a large loss.

8 A two-dimensional state space with four words

We turn towards illustrative examples in a two dimensional state space and analyze the robustness of different geometric language structures. The word space will again consist of two letters and words of length 2, i.e. $\mathrm{W} = \{A, B\}^2$.

Let $L = 2$ and $T = [-\frac{1}{2}, \frac{1}{2}]^2$ be endowed with the uniform distribution $\mu_0 \sim \mathcal{U}(T)$ and $M = (0, 0)$. Assume further that $p < \frac{1}{2}$, i.e. $\widetilde{p} < 1$, so that the error function is not uninformative and it is less likely to confound words that are farther away from one another.

We will now study the four special kinds of tessellations, depicted in Figure 5. The respective cells of the communication device are given by the following color schemes belonging to words: AA (green), AB (blue), BA (violet), BB (yellow). The black lines indicate the respective optimal interpretations for each $p \in [0, \frac{1}{2}]$, going to the center M if $p \in [0, \frac{1}{2}]$ increases. The considered languages differ in two different aspects. Firstly, the shape of cells is either a square or a triangle (classes 1 or 2). Secondly, the distribution of words to the cells in a way that leads to different Hamming distances of the used words, differs across class-a and class-b languages. Note that while in class-a languages the proper neighbors always use words with a Hamming distance of 1, while for class-b languages each word has a neighbor with Hamming distance 2 as well.

Figure 5: Four Languages

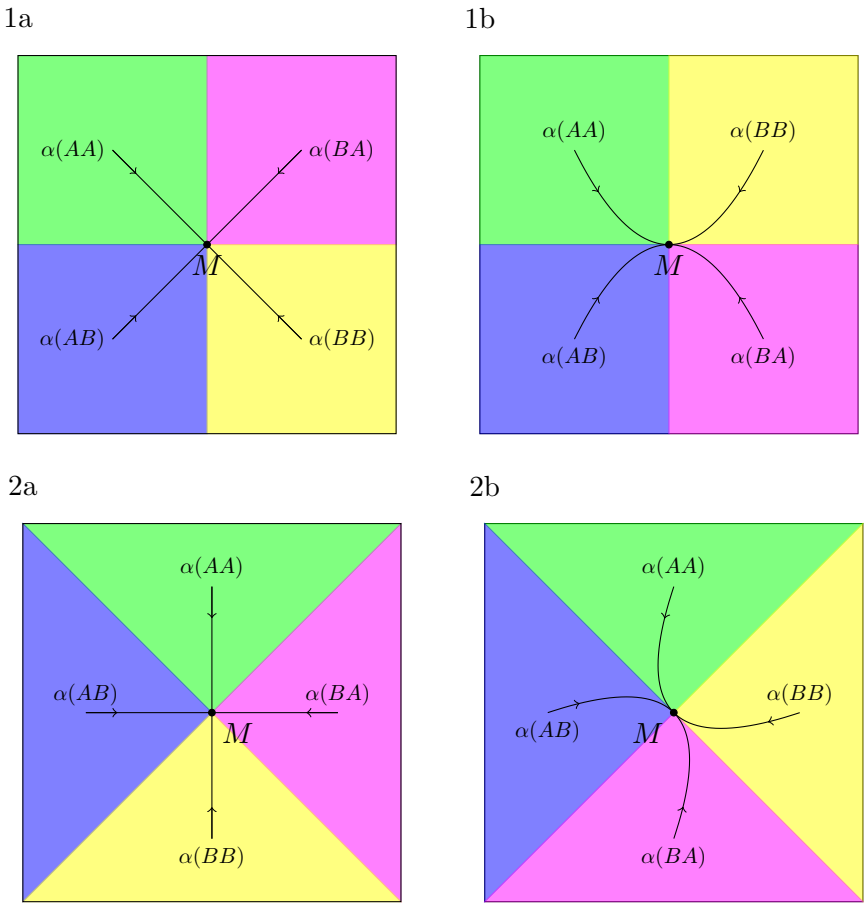


Table 1: Optimal Values

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

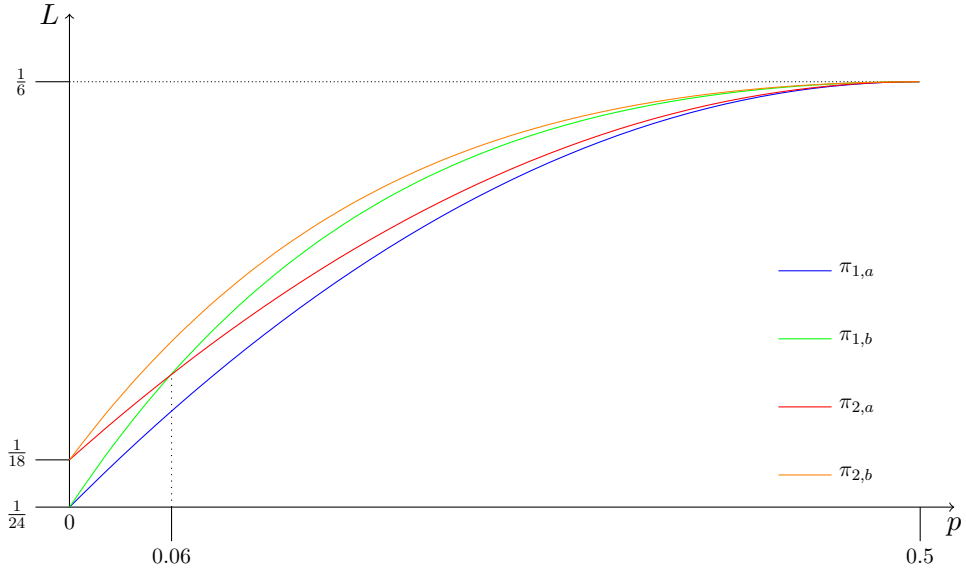
Lemma 16. *Let $p \in [0, \frac{1}{2}]$ be arbitrary. Then the following assertions hold about any of the four languages in Figure 5.*

(i) The optimal interpretations and the expected loss are given by Table 1.

(ii) Given the interpretations in Table 1, the indicated tessellation is an optimal one.

Thus, each of the considered languages constitutes a noise equilibrium.

Figure 6: Expected Loss



There are several interesting comparisons that can be made with these languages. First of, the class-a languages have interpretations that go in a straight line towards the default action $\alpha_0 = M$ while the class-b ones do so quadratically[(7)] with the bump towards the edge where words are closer w.r.t. the Hamming distance. The intuition for this is straightforward: Look w.l.o.g. at $\hat{\alpha}(AA)$: Within the class-a languages, the cells of the words w with length $d(AA, \mathrm{w}) = 1$ (AB and BA) absorb the same amount of mistakes from and towards AA, while the word BB with length $d(AA, BB) = 2$ does so quadratically but also point-symmetrically w.r.t. the default action. Increasing p thus shifts the interpretation on a straight line towards M. In contrast, for the class-b languages, the locus of $\hat{\alpha}(AA)$ is again pulled linearly by the cells of AB and BA while quadratically of BB. As the cell of BB now borders the one of AA and it is less likely to confound AA with BB, $\hat{\alpha}(AA)$ is quadratically less pulled by the cell of BB, leading to the parabola shaped locus.

Comparing the expected loss (see also Figure 6) we first note that in the classical case without error (i.e. $p = 0$), the respective class-a and class-b languages are equivalent while in the presence of error, this is no longer so. In general, increasing the error monotonically decreases the loss, and, reaching the uninformative bound $p = \frac{1}{2}$, any communication collapses and the default loss is reached.

In general, we find that the class-a languages serve as better communication structures for any single error probability $0 < p < \frac{1}{2}$. Intuitively and mathematically, receiving AA given a class-b language shifts our belief (and thus the interpretation) closer to the default one than a class-a language, as more of the mass of distance 1 words is concentrated opposite of AA's cell than in the class-a languages. Even though the communication of class-b languages is more precise in distinguishing types from AA's cell to the ones in BB's cell, this does not make up for the deterioration caused by the even stronger pull towards the default belief. Thus we find again that it is better to clearly distinguish states that are furthest away from each other by using words that are least likely to be confounded.

[(7)] Note that as $0 \leq p \leq 1$, a quadratical pull is less strong than a linear one.

Loosely speaking, the border between AA and BB is less permeable than the one for AA and AB (or BA), meaning that errors diffuse easier between cells with harder border (of *degree* $d(\mathrm{v},\mathrm{w})$).

Furthermore, comparing type-1 languages to the type-2 ones we find that the former are superior to the latter (case a and b wise). Intuitively, the squares provide a more compact structure and less points near and on indifference level (in this case the respective bisectors between the interpretations[8]) than the triangle shapes. This also leads to the interpretations being farther away from the default action, decomposing the common belief less tightly than in case 2.

Concluding the example, we also find that language 2a is superior to 1b for a single error probability exceeding $p = \frac{1}{2} - \sqrt{2} \approx 6\%$. This indicates that the distribution of words to cells seems more important than the choice of stable cell structures.

Merging Words

So far, we did not encounter a non-null set of indifference points in the two dimensional case. Language 3 in Figure 7 provides such an example. Interestingly, the language can be condensed into one using fewer words, thus making the active use of the remaining words redundant.

Consider on $T = [-\frac{1}{2}, \frac{1}{2}]^2$ with $\mu_0 \sim \mathcal{U}(T)$ and $\mathrm{W} = \{A, B\}^2$ the languages depicted in Figure 7. The color schemes remain the same as before.

Figure 7: Merging words

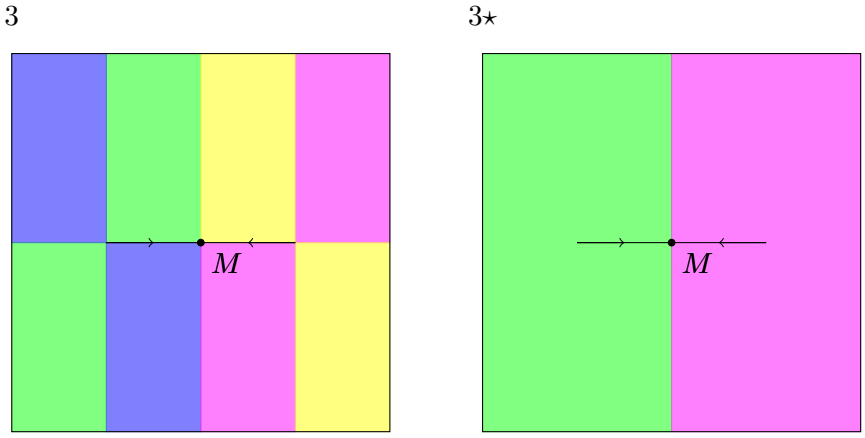


Table 2: Optimal Values - 3a,b

Case	$\alpha(AA)$	$\alpha(AB)$	$\alpha(BA)$	$\alpha(BB)$	$\mathscr{L}(\pi,\alpha)$
$3(\star)$	$\frac{1}{4}(-(1-2p),0)$	$\frac{1}{4}(-(1-2p),0)$	$\frac{1}{4}(1-2p,0)$	$\frac{1}{4}(1-2p,0)$	$\frac{1}{6} - \frac{1}{16}(1-2p)^2$

The optimal actions and the corresponding expected loss are exactly the same and stated in Table 2. Thus, both languages can be interpreted as (outcome) equivalent. The reason for this is that the confusing structure of language 3 amounts to the same information gain of receiver as the simpler language using only two words. Note that sender is indifferent between sending AA or AB on the left hand side of M (and likewise BB or BA on the right hand side).

[8] In case 1, theres a total border length of $1 + 1 + \sqrt{2}$, while in case 2 this is $\sqrt{2} + \sqrt{2} + 1$

Surprisingly, if sender switches from 3 to 3⋆, the optimal interpretations do not change. This stems from the fact that, for instance, the beliefs μ^{π}_{BB} and μ^{π}_{BA} are the same: The smaller likelihood of receiving BB (than BA) in language 3⋆ is exactly compensated by the increased confidence that sender actually sent BA (and not AA) if BB is received.

To see this formally, start by verifying $\lambda^{\pi^{3\star}}(AA) = \lambda^{\pi^{3\star}}(BA) = \frac{1}{2}(1-p)^2(1+\widetilde{p})$ and $\lambda^{\pi^{3\star}}(BB) = \lambda^{\pi^{3\star}}(AB) = \widetilde{p}\cdot\lambda^{\pi^{3\star}}(AA)$. Finally, the density function yields

$$f^{\pi^{3\star}}_{BB}(t) = \frac{\varepsilon(BB\,|\,\pi^{3\star}(t))\cdot f_0(t)}{\lambda^{\pi^{3\star}}(BB)} = \frac{\widetilde{p}\cdot\varepsilon(BA\,|\,\pi^{3\star}(t))\cdot f_0(t)}{\widetilde{p}\cdot\lambda^{\pi^{3\star}}(BA)} = f^{\pi^{3\star}}_{BA}(t).$$

Thus, there is the possibility to achieve the same expected loss as in 3, merging the words on the left resp. right hand side of the square in a way that leaves two words of distance 1 in W describing "left" and "right".

9 Evolution

As for usual languages we are interested in the formation and convergence process of communication. This leads to the question how evolutionary forces, modeled by dynamical systems, shape populations engaging in noisy communication.

As our strategy space consists not only of interpretations, amounting of points in $T^{\#\mathrm{W}}$, but also communication devices which are member of the set Σ of measurable functions $\pi\colon T\to \mathrm{W}$ we are facing rather complicated strategy spaces. Even more so when considering populations, i.e. probability distributions over pure strategies. The technical foundation for well-known dynamics is given. These include the replicator ([OR01], [CHR06]), payoff monotone ([HSS07]) and Brown-von-Neumann-Nash dynamics ([HOR09]) and extends to our setting.

We proceed along the lines of [JMR11], considering the symmetrized version of the cheap talk game. Agents are equally likely to become sender and receiver and thus needs to choose both, a communication device as well as an interpretation. Thus, an agent using (π,α) and meeting an individual using (π',α') occurs in expectation a loss of

$$\Lambda((\pi,\alpha),(\pi',\alpha')) = \frac{1}{2}\mathscr{L}(\pi,\alpha') + \frac{1}{2}\mathscr{L}(\pi',\alpha). \tag{19}$$

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

$$\Lambda(P,Q) := \int_\Gamma\int_\Gamma \Lambda((\pi,\alpha),(\pi',\alpha'))\,P(\mathrm{d}\pi,\mathrm{d}\alpha)\,Q(\mathrm{d}\pi',\mathrm{d}\alpha'). \tag{20}$$

Proposition 17. *The following assertions hold.*

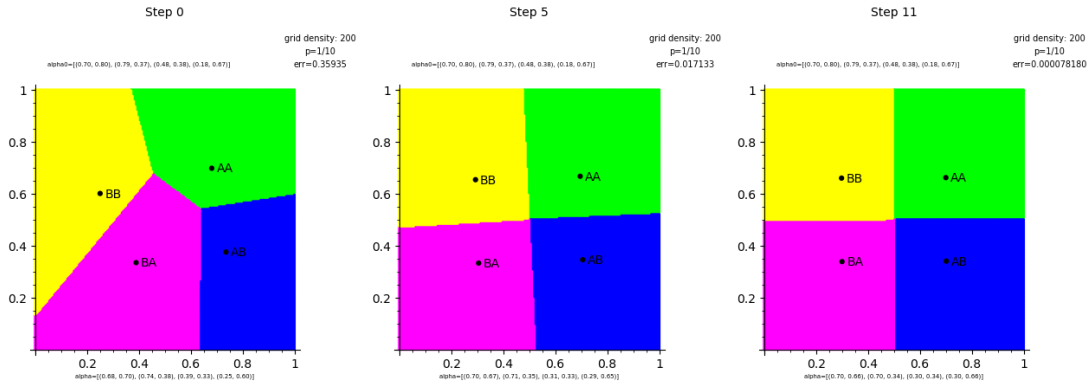
(i) The expected loss function is continuous with respect to the weak topology.

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

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

Approaching the question of language formation from a myopic point of view, one could ask whether or not the same sender and receiver may be able to learn from one another if they are playing the game many times. In fact, as Figure 8 demonstrates, a numerical analysis reveals that iteratively playing a best reply to the action taken by the other player before, converges to stable languages of the form 1, while indeed those of form 2 turn out not to be locally optimal.

Figure 8: Best Reply Dynamics



Depicted are the optimal communication device given the marked interpretations. The dispersion to equilibrium ('error') is measured as norm distance between two subsequent interpretations.

10 Conclusion

Within our daily routine, errors in our communication are ubiquitous and inevitable. However, this type of uncertainty is far from being arbitrary. In situations of communication, both, digitally or in person, confounding is more likely to arise between messages that are similar than between ones that are very distinct, thus giving noise a structure. Providing a mathematical foundation for such settings, this paper analyzes not only general properties of such noise channels, such as existence of equilibria, by incentivizing mutual understanding, but also provides normative and descriptive insights: On the one hand, we find geometric criteria for languages to be robust in presence of errors that can be used to develop message systems that minimize confusion. On the other hand, we find that such noise properties can be learned over time by means of evolution, indicating that our common day languages are shaped by forces of omnipresent small errors.

A Proofs

(best reply of receiver). The non-emptiness of the best reply set is given by continuity of the integrand and compactness of T. Assume there are two distinct minimizers s_1, s_2 and let $\lambda \in (0,1)$. Then by the triangle inequality and convexity of ℓ we find the contradiction

$$\begin{aligned}
\mathbb{E}_\mu[\ell(\|t-(\lambda s_1+(1-\lambda)s_2)\|)] &\leq \mathbb{E}_\mu[\ell(\lambda\,\|(t-s_1)\|+(1-\lambda)\,\|t-s_2\|)] && (21)\\
&< \mathbb{E}_\mu[\lambda\ell(\|t-s_1\|)]+\mathbb{E}_\mu[(1-\lambda)\ell(\|t-s_2\|)] && (22)\\
&= \mathbb{E}_\mu[\ell(\|t-\hat{s}\|)]. && (23)
\end{aligned}$$

The strict inequality holds as $f>0$ and ℓ is convex and strictly increasing. □

Proof. (Proposition 1) Consider the constant strategy α_0 of receiver. Then

$$\begin{aligned}
\mathscr{L}(\pi,\hat{\alpha}) &= \mathbb{E}_{\lambda^\pi}\left[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}\left[\ell(\|t-\hat{\alpha}(\mathrm{w})\|)\right]\right] && (24)\\
&\leq \mathbb{E}_{\lambda^\pi}\left[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}\left[\ell(\|t-\alpha_0\|)\right]\right] && (25)\\
&= \mathbb{E}_{\mu_0}[\ell(\|t-\alpha_0\|)] = L_0, && (26)
\end{aligned}$$

where we applied the minimizing 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 λ^π and $\hat{\alpha}(\mathrm{w}) \neq \alpha_0$. □

Proof. (Corollary 2) If π ist constant, say $\pi \equiv \mathrm{v}$, then $\varepsilon(\mathrm{w} \,|\, \pi(t)) = \varepsilon(\mathrm{w} \,|\, \mathrm{v})$ is constant and even equals $\lambda^\pi(\mathrm{w})$ for any $\mathrm{w} \in \mathrm{W}$. This implies $f^\pi_\mathrm{w} = f_0$ and thus $\mu^\pi_\mathrm{w} = \mu_0$.
If $\varepsilon(.\,|\,\mathrm{v})$ is the uniform distribution on W for all $\mathrm{v} \in \mathrm{W}$ it follows that $\varepsilon(\mathrm{w} \,|\, \mathrm{v}) = \#\mathrm{W}^{-1}$ for all $\mathrm{w}, \mathrm{v} \in \mathrm{W}$. Then $\lambda^\pi(\mathrm{w}) = \#\mathrm{W}^{-1}$ for all $\mathrm{w} \in \mathrm{W}$ and again $\mu^\pi_\mathrm{w} = \mu_0$. □

Proof. (Theorem 3) We follow the proof of Lemma 1 in [JMR11] which can be adjusted, incorporating the error function.
To this end, start by fixing a pure receiver strategy $\alpha\colon \mathrm{W} \to T$. Then, a type $t \in T$ sender optimally picks any word v out of

$$\arg\min_{\mathrm{v}' \in \mathrm{W}} \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|). \tag{27}$$

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

$$C^\alpha_\mathrm{v} := \left\{ t \in T \mid \mathrm{v} \text{ is smallest element w.r.t. } \leq_\mathrm{W} \text{ in } \arg\min_{\mathrm{v}' \in \mathrm{W}} \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \tag{28}$$

for each $\mathrm{v} \in \mathrm{W}$. Note that 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:
Start by collecting all the t where v is a minimizer, which is a closed set. Now, for all $\mathrm{v}' \leq_\mathrm{W} \mathrm{v}$ take away the indifference sets, which are closed, to obtain C^α_v.

This way, a measurable function $\pi\colon T \to \mathrm{W}$ is defined by $\pi(t) = \mathrm{v} \iff t \in C^\alpha_\mathrm{v}$, which represents one possible best reply of sender.
Using any such choice, the general loss minimization can be rewritten depending only on α as

$$\min_\alpha \int_T \min_\mathrm{v} \left\{ \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \mu_0(\mathrm{d}t). \tag{29}$$

Now, note that we can identify any strategy $\alpha\colon \mathrm{W} \to T$ as 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. (Lemma 5) The case $p = 0$ is clear. For $0 < \widetilde{p} < 1$ we have that

$$\varepsilon(\mathrm{w} \,|\, \mathrm{v}) > \varepsilon(\mathrm{w}' \,|\, \mathrm{v}) \iff \widetilde{p}^{d(\mathrm{w},\mathrm{v})} > \widetilde{p}^{d(\mathrm{w}',\mathrm{v})} \iff d(\mathrm{w},\mathrm{v}) < d(\mathrm{w}',\mathrm{v}). \tag{30}$$

The other cases follow similarly. □

Proof. (Proposition 6)

(i) Clear by continuity of the integrand in $\widetilde{p}$ and Lebesgue's theorem.

(ii) To start with, always split the integral in three parts by disassembling the type space T into $\{t' \,|\, d(\mathrm{w}, \pi(t')) \lessgtr d(\mathrm{w}, \pi(t))\}$.

   (a) The set $\{t' \,|\, d(\mathrm{w}, \pi(t')) < d(\mathrm{w}, \pi(t))\}$ 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 $\{t' \,|\, d(\mathrm{w}, \pi(t')) < d(\mathrm{w}, \pi(t))\}$ has probability zero and can be neglected. For $p \to 0$, the integral over $\{t' \,|\, d(\mathrm{w}, \pi(t')) > d(\mathrm{w}, \pi(t))\}$ will vanish as the exponent of $\widetilde{p}$ is strictly positive. What is left of the overall integral is $\int_{\{t' \,|\, d(\mathrm{w},\pi(t')) = d(\mathrm{w},\pi(t))\}} \mu_0(\mathrm{d}t') = \mu_0(\{t' \,|\, d(\mathrm{w}, \pi(t')) = d(\mathrm{w}, \pi(t))\})$ 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. (Proposition 7) We start of by calculating the entropy

$$H(\varepsilon(.\,|\,\mathrm{v})) = -\sum_{w\in\mathrm{W}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}) \cdot \log\left(\varepsilon(\mathrm{w}\,|\,\mathrm{v})\right) \tag{31}$$

$$= -\sum_{w\in\mathrm{W}} (1-p)^{n-d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})} \cdot \log\left((1-p)^{n-d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})}\right) \tag{32}$$

$$= -\sum_{d=0}^{n} \binom{n}{d} \cdot m^d \cdot (1-p)^{n-d} \cdot \left(\frac{p}{m}\right)^{d} \cdot \log\left((1-p)^{n-d} \cdot \left(\frac{p}{m}\right)^{d}\right) \tag{33}$$

$$= -\sum_{d=0}^{n} \binom{n}{d} \cdot (1-p)^{n-d} \cdot p^d \cdot \log\left((1-p)^{n-d} \cdot p^{d}\right) \tag{34}$$

$$+ \log(m) \cdot \sum_{d=0}^{n} \binom{n}{d} \cdot d \cdot (1-p)^{n-d} \cdot p^d \tag{35}$$

$$= -\sum_{d=0}^{n} \binom{n}{d} \cdot (1-p)^{n-d} \cdot p^d \cdot \log\left((1-p)^{n-d} \cdot p^{d}\right) + np\log(m) \tag{36}$$

$$= -np \cdot \log(p) - n(1-p) \cdot \log(1-p) + np\log(m) \tag{37}$$

$$= n \cdot (H((p, 1-p)) + p\log(m)). \tag{38}$$

This function is concave in p since $H((p,1-p)) = -p\log(p) - (1-p)\log(1-p)$ is. The maximum is attained for the uniform distribution which is attained for $p = \frac{m}{m+1}$ by Lemma 5 and yields $H(\mathcal{U}(\mathrm{W})) = \log(\#\mathrm{W})$. The other assertions follow readily from the calculated expression. □

Proof. (Lemma 9) To begin with recall receiver's minimization problem:

$$\min_{\alpha\in T} \mathbb{E}_\mu[\|t-\alpha\|_2^2] = \int_T \sum_{k=1}^{L} (t_k - \alpha_k)^2 \,\mu(\mathrm{d}t). \tag{39}$$

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

$$\hat{\alpha}_k(\mu) = \int_T t_k \,\mu(\mathrm{d}t) = \mathbb{E}_\mu[t_k]. \tag{40}$$

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

$$\mathbb{E}_\mu[\left\|t - \mathbb{E}_\mu[t']\right\|_2^2] = \mathbb{E}_\mu[\|t\|_2^2] - \|\mathbb{E}_\mu[t]\|_2^2\,, \tag{41}$$

which is the sum of the variances over the t_k's.

Thus, if a communication device π is given, an expected loss of

$$\mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\left\|t - \mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t']\right\|_2^2]] = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\|t\|_2^2] - \left\|\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t]\right\|_2^2] \tag{42}$$

$$= \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\left\|\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t]\right\|_2^2] \tag{43}$$

is faced, where Bayes-Plausibility was used. Finally, using again Bayes-Plausibility we observe

$$\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \mathbb{E}_{\lambda^\pi}\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t] = \mathbb{E}_{\mu_0}[t] = \hat{\alpha}(\mu_0) = \alpha_0. \tag{44}$$

□

Proof. (Proposition 11) Recall (8) and that $\|x-y\|_2^2 = \|x\|_2^2 - 2\langle x\,,\,y\rangle + \|y\|_2^2$. We start by reformulating the condition that a type-t sender strictly prefers to send v instead of v$'$:

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w}\,|\,\mathrm{v})\,\|t-\alpha(\mathrm{w})\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}')\,\|t-\alpha(\mathrm{w})\|_2^2 \tag{45}$$

$$\iff \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big)\cdot\Big(-2\,\langle t\,,\,\alpha(\mathrm{w})\rangle + \|\alpha(\mathrm{w})\|_2^2\Big) > 0 \tag{46}$$

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

Proof. (Proposition 13) Indifference for type $t \in T$ means

$$\sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big)\cdot\Big(-2\,\langle t\,,\,\alpha(\mathrm{w})\rangle + \|\alpha(\mathrm{w})\|_2^2\Big) = 0 \tag{47}$$

$$\iff -2\cdot\left\langle t\,,\,\sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big)\cdot\alpha(\mathrm{w})\right\rangle + \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big)\cdot\|\alpha(\mathrm{w})\|_2^2 = 0. \tag{48}$$

Now if $\sum_{\mathrm{w}} (\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v}))\cdot\alpha(\mathrm{w}) \neq 0$ the solution set is the translation of the $L-1$ dimensional hyperplane perpendicular to the vector $\sum_{\mathrm{w}}(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v}))\}\alpha(\mathrm{w})$ by a particular solution (if it exists, otherwise it's the empty set) and thus a null set in $\mathbb{R}^L$. If it is 0, then necessarily $\sum_{\mathrm{w}} (\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v}))\cdot\|\alpha(\mathrm{w})\|_2^2 = 0$ for indifference. But then any t in T (even $\mathbb{R}^L$) fits the equation.

In the latter case, it holds that

$$\sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big)\cdot\alpha(\mathrm{w}) = 0 \tag{49}$$

$$\iff \sum_{\mathrm{w}} \Big(\widetilde{p}^{\,d(\mathrm{w},\mathrm{v}')} - \widetilde{p}^{\,d(\mathrm{w},\mathrm{v})}\Big)\cdot\alpha(\mathrm{w}) = 0, \tag{50}$$

which is a polynomial of degree at most n in $\widetilde{p}$ with zeros in $\widetilde{p} \in \{0,1\}$. It is not the zero polynomial as the coefficient for $\widetilde{p}^{\,0}$ is $\alpha(\mathrm{v}') - \alpha(\mathrm{v}) \neq 0$. On $(0,1)$ it thus has at most $n-2$ other zeros. □

Proof. (Lemma 16) As the proof is mostly about calculations to various objects, we structure it as follows: We begin by calculating the optimal interpretations and the expected loss. Then we characterize conditions on optimal cells for the considered word space. Eventually, fixing any of the calculated interpretation we will show that the indicated tessellation is indeed an optimal one, even uniquely up to null sets.

To start with, note, that for any considered cell $\mu_0(C_{\mathrm{v}}) = \frac{1}{4}$ and thus for any w

$$\lambda^{\pi}(\mathrm{w}) = \int_T \varepsilon(\mathrm{w}\,|\,\pi(t))\,\mathrm{d}t = \sum_{\mathrm{v}} \varepsilon(\mathrm{w}\,|\,\mathrm{v})\cdot\mu_0(C_{\mathrm{v}}) = \frac{1}{4}. \tag{51}$$

in all cases.

(i) Interpretations and Expected Loss

Denote by $\square(AA)$ and $\Delta(AA)$ the respective area indicated in Figure 5.

1. (a) Let us begin by calculating the expected values/center of gravity of each colored area, say the one for AB:

$$\mathbb{E}_{\square(AB)}[t] := \mathbb{E}_{\mu_0,\square(AB)}[t] = \mu_0(\square(AB))^{-1} \cdot \int_{\square(AB)} t \,\mathrm{d}t \tag{52}$$

$$= 4 \cdot \int_{-\frac{1}{2}}^{0} \int_{-\frac{1}{2}}^{0} (t_1, t_2) \,\mathrm{d}t_1 \mathrm{d}t_2 = \left(-\frac{1}{4}, -\frac{1}{4}\right). \tag{53}$$

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 $\hat{\alpha}(AA)$.

$$\hat{\alpha}(AA) = \mathbb{E}_{\mu^{\pi}_{AA}}[t] = \lambda^{\pi}(AA)^{-1} \cdot \int_T \varepsilon(AA \,|\, \pi(t)) \cdot t \,\mu_0(\mathrm{d}t) \tag{54}$$

$$= 4 \cdot \sum_{\mathrm{w}} \varepsilon(AA \,|\, \mathrm{w}) \cdot \int_{\square(\mathrm{w})} t \,\mathrm{d}t \tag{55}$$

$$= \sum_{\mathrm{w}} \varepsilon(AA \,|\, \mathrm{w}) \cdot \mathbb{E}_{\square(\mathrm{w})} \tag{56}$$

$$= (1-p)^2 \cdot (-\tfrac{1}{4}, \tfrac{1}{4}) + p(1-p) \cdot \big((-\tfrac{1}{4}, -\tfrac{1}{4}) + (\tfrac{1}{4}, \tfrac{1}{4})\big) + p^2 \cdot (\tfrac{1}{4}, -\tfrac{1}{4}) \tag{57}$$

$$= \frac{1}{4} \cdot (-1 + 2p, 1 - 2p). \tag{58}$$

Analogously, by symmetry arguments or using Lemma 9 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 $\|\alpha(\mathrm{w}) - \alpha_0\|_2 \searrow 0$ for $\widetilde{p} \to 1$, where $M = \alpha_0 = (0, 0)$ is the expected value over the whole state space.

We are now ready to calculate the expected loss. We start be observing that each interpretation has the same norm:

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \cdot \left\| \frac{1}{4} \cdot (1 - 2p, 1 - 2p) \right\|_2^2 = \frac{1}{8} \cdot (1 - 2p)^2. \tag{59}$$

Having calculated $\mathbb{E}_T[\|t\|_2^2] = \frac{1}{6}$, we use equation (17) to obtain the expected loss

$$\mathscr{L}(\pi_{1,a}, \hat{\alpha}) = \frac{1}{6} - \sum_{\mathrm{w}} \frac{1}{4} \cdot \frac{1}{8} \cdot (1 - 2p)^2 = \frac{1}{6} - \frac{1}{8} \cdot (1 - 2p)^2. \tag{60}$$

One clearly sees that the expected loss is increasing monotonically 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

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{16} \cdot ((1 - 2p)^2 + (1 - 2p)^4), \tag{61}$$

resulting in an expected loss of

$$\mathscr{L}(\pi_{1,b}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{16} \cdot \big((1 - 2p)^2 + (1 - 2p)^4\big) \tag{62}$$

$$= \frac{1}{6} - \frac{1}{8} \cdot (1 - 2p)^2 \cdot (1 - 2p + 2p^2). \tag{63}$$

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

$$\mathscr{L}(\pi_{1,a}, \hat{\alpha}) < \mathscr{L}(\pi_{1,b}, \hat{\alpha}), \tag{64}$$

thus, the language, putting far 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

$$\|\alpha(\mathrm{w})\|_2^2 = \left\|(0, \tfrac{1}{3} - \tfrac{2}{3}p)\right\|_2^2 = \frac{1}{9} \cdot (1-2p)^2. \tag{65}$$

The resulting expected loss is thus

$$\mathscr{L}(\pi_{2,a}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{9} \cdot (1-2p)^2, \tag{66}$$

which is strictly higher than $\mathscr{L}(\pi_{1,a}, \hat{\alpha})$ 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

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{9}(1-2p)^2(1-2p+2p^2) \tag{67}$$

and hence

$$\mathscr{L}(\pi_{2,b}, \hat{\alpha}) = \frac{1}{6} - \frac{1}{9}(1-2p)^2(1-2p+2p^2), \tag{68}$$

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

(ii) Optimal Cell Structure

To start with we simplify the expressions from Lemma 11 and Proposition 13 for $\mathrm{W} = \{A, B\}^2$. To this end, fix w.l.o.g. (by the subsection on language isometry) the word AA and derive conditions on a fixed $t \in T$ for AA to be the optimal word.

- Type t prefers to send AA over BB if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AA)\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(BB)\|_2^2 \tag{69}$$

$$\iff \|t - \alpha(AA)\|_2 < \|t - \alpha(BB)\|_2 \,. \tag{70}$$

- Type t prefers AA over AB (the case BA is analogous) if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AA)\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \, \|t - \alpha(AB)\|_2^2 \tag{71}$$

$$\iff \|t - \alpha(AA)\|_2^2 - \|t - \alpha(AB)\|_2^2 < \widetilde{p}\left(\|t - \alpha(BB)\|_2^2 - \|t - \alpha(BA)\|_2^2\right) \tag{72}$$

$$\iff 2\,\langle t \,,\, \alpha(AB) - \alpha(AA) + \widetilde{p}(\alpha(BB) - \alpha(BA))\rangle \tag{73}$$

$$+ \|\alpha(AA)\|_2^2 - \|\alpha(AB)\|_2^2 + \widetilde{p}(\|\alpha(BA)\|_2^2 - \|\alpha(BB)\|_2^2) < 0. \tag{74}$$

Whereas in (i) we clearly see that the set of sender types that are indifferent between AA and BB lie on the perpendicular bisector (and thus a null set in T) 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 types is again a null set as the intersection of a line and T.

To drop some notation and 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

$$\|t - AA\|_2 < \|t - AB\|_2 \,, \|t - BA\|_2 < \|t - BB\|_2 \,. \tag{75}$$

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

Realizing that

$$\|t - AA\|_2^2 - \|t - AB\|_2^2 < 0 < \widetilde{p} \cdot \left(\|t - BB\|_2^2 - \|t - BA\|_2^2 \right), \tag{76}$$

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

$$0 \leq \|t - AA\|_2 < \|t - AB\|_2 \,, \|t - BB\|_2 < \|t - BA\|_2 \,, \tag{77}$$

we observe

$$\begin{aligned} \|t - BA\|_2^2 - \|t - AA\|_2^2 &> \left| \|t - AB\|_2^2 - \|t - BB\|_2^2 \right| && (78) \\ &> \widetilde{p} \cdot \left| \|t - AB\|_2^2 - \|t - BB\|_2^2 \right| && (79) \\ &\geq \widetilde{p} \cdot \left(\|t - AB\|_2^2 - \|t - BB\|_2^2 \right), && (80) \end{aligned}$$

showing that AA is preferred toBA.

Finally, using $AA = -BA$, $AB = -BB$ and that $\|\alpha(\mathrm{w})\|$ is constant, we realize

$$\begin{aligned} & 2 \cdot \langle t \,, -AA + AB + \widetilde{p}(BB - BA) \rangle + \|AA\|_2^2 - \|AB\|_2^2 + \widetilde{p}(\|BB\|_2^2 - \|BA\|_2^2) && (81) \\ = & 4 \cdot \left\langle t \,, \underbrace{\tfrac{AB+BA}{2}}_{=:P} \right\rangle . && (82) \end{aligned}$$

This term is smaller than zero in both cases 1 and 2:

1. As $t_1 < 0, t_2 > 0$ 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.

From the calculations it is clear that the borders of the cells consist precisely of the indifference points for senders. □

Proof. (Proposition 17) We only show continuity and boundedness of L as this implies continuity of Λ. The rest of the assertions follow well-known lines ([HSS07], [HOR09]) as well as [BS02] for the last statement.

Let $(\pi_n)_n$ 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 $d(\pi_n(t), \pi(t)) < \rho'$ for $n > M$. As d has only values in $\{0, \dots, n\}$, this is equivalent to $\pi_n \equiv \pi$ for all $n > N_0$ 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 $(\alpha_n)_n$ 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 $\|t - \alpha_n(\mathrm{w}) - (t - \alpha(\mathrm{w}))\| = \|\alpha_n(\mathrm{w}) - \alpha(\mathrm{w})\| < \delta$ for all $n > N_1$. Thus, for all $n > N_1$ we find

$$|\mathscr{L}(\pi_n, \alpha_n) - \mathscr{L}(\pi, \alpha))| \leq \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi_n(t)) \ell(\|t - \alpha_n(\mathrm{w})\|) - \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{83}$$

$$= \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left(\ell(\|t - \alpha_n(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|) \right) \right| \mu_0(\mathrm{d}t) \tag{84}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left| \ell(\|t - \alpha_n(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{85}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \rho \, \mu_0(\mathrm{d}t) \tag{86}$$

$$= \rho. \tag{87}$$

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

$$|\mathscr{L}(\pi, \alpha)| \leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)\cdot) \left| \ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \bar{\ell} \, \mu_0(\mathrm{d}t) = \bar{\ell} < \infty. \tag{88}$$

□

References

[BBK07] Andreas Blume, Oliver Board, and Kohei Kawamura. Noisy Talk. Edinburgh School of Economics Discussion Paper Series 167, Edinburgh School of Economics, University of Edinburgh, April 2007.

[Blu04] Andreas Blume. A learning-efficiency explanation of structure in language. *Theory and Decision*, 57(3):265–285, 2004.

[BS02] N.P. Bhatia and G.P. Szegö. *Stability Theory of Dynamical Systems*. Classics in Mathematics. Springer Berlin Heidelberg, 2002.

[BS07] Timothy Besley and Michael Smart. Fiscal restraints and voter welfare. *Journal of public Economics*, 91(3-4):755–773, 2007.

[CHR06] Ross Cressman, Josef Hofbauer, and Frank Riedel. Stability of the replicator equation for a single species with a multi-dimensional continuous trait space. *Journal of Theoretical Biology*, 239(2):273–288, 2006. Special Issue in Memory of John Maynard Smith.

[CS82] Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[FR11] Manuel Förster and Frank Riedel. Distorted voronoi languages. Technical report, Working Papers, 2011.

[FV20] Manuel Foerster and Achim Voss. Believe me, i am ignorant, but not biased. *Available at SSRN 3301146*, 2020.

[Ham50] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950.

[HOR09] Josef Hofbauer, Jörg Oechssler, and Frank Riedel. Brown–von neumann–nash dynamics: The continuous strategy case. *Games and Economic Behavior*, 65(2):406–429, 2009.

[HSS07] Aviad Heifetz, Chris Shannon, and Yossi Spiegel. What to maximize if you must. *Journal of Economic Theory*, pages 31–57, 2007.

[JMR11] Gerhard Jäger, Lars P. Metzger, and Frank Riedel. Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals. *Games and Economic Behavior*, 73(2):517 – 537, 2011.

[Jä07] Gerhard Jäger. The evolution of convex categories. *Linguistics and Philosophy*, 30:551–564, 10 2007.

[KG11] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011.

[Mac02] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA, 2002.

[Mye91] Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[OR01] Jörg Oechssler and Frank Riedel. Evolutionary dynamics on infinite strategy spaces. *Economic Theory*, 17(1):141–162, 2001.

[Rot06] Ron Roth. *Introduction to Coding Theory*. Cambridge University Press, USA, 2006.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[Sob15] Joel Sobel. Broad terms and organizational codes. *Unpublished paper, Department of Economics, University of California, San Diego.[1138]*, 2015.