Effects of noise on the category formation of languages

Gerrit Bauch[1*]

[1*]Center for Mathematical Economics, Bielefeld University, Universitätsstraße 25, Bielefeld, 33615, Germany.

Corresponding author(s). E-mail(s): gerrit.bauch@uni-bielefeld.de;

Abstract

I study an extension of the cheap-talk, common-interest game introduced by Jäger et al. (2011) to model the formation of convex categories in the presence noise. Given a finite set of messages, the sender partitions the state space into convex categories, extending a linguistic conjecture to communication under noise. If the message space is combinatorial, efficient communication selects robust partitions, assigning distant words to extreme ends of the state space and forming word clusters of similar states. Furthermore, the sender stresses extreme states under an increasing level of noise by reducing the size of cells describing average observations. Under evolutionary dynamics, I show that individuals can learn noise-buffering patterns and achieve efficient communication over time.

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

JEL Classification: C72 , C73 , D82 , D83

1 Introduction

Noise is ubiquitous in our daily communication. Words get swamped by background noises, people slip their tongue, stammer or suffer from a hearing impairment. It is natural to assume that communication is imperfect and prone to error. Consequently, natural languages have developed mechanisms to cope with environments of noise: The NATO phonetic alphabet is a famous example of communicating letters by clear-code words, using their redundancy and different sounds to reduce the likelihood of errors. Words marking extreme opposites often come with different sounds like yes↔no, left↔right or hot↔cold, indicated by the use of short vowels or diphthongs. Likewise, we can find words with similar meanings sharing similar pronunciation in phonesthemes, such as gleam, glow, glare, which in many cases could be mixed up without significant changes to the meaning of a sentence.[1]

This article proposes a game-theoretic approach to formalize the emergence of error-buffering patterns. To this end, I add a stochastic noise channel to a cheap-talk game of common interest where finitely many messages are used to describe a *conceptual space*, representing meanings and relationships of concepts in a structured way. This conceptual space is formally defined as a convex subset of a Euclidean space, thereby following Gärdenfors (2004)'s interpretation of concepts as regions in a geometric space which reflects similarity between concepts as geometric proximity. My contribution is two-fold: In the first part of the paper, I study general game-theoretic properties of communication under noise, characterizing best replies and providing an argument for convex category formation under noise. In the second part, I consider a combinatorial word space and a concrete noise respecting its structure. Together, they reveal three patterns of efficient noise-buffering that resonate with intuitive arguments: First, the geometric shape of the optimal concepts, bundling states together, is chosen in a way that limits the damage from miscommunication. Second, distant words are reserved for distant states in order to minimize the loss of miscommunication; likewise, similar states on each end of the state space are assigned similar words, forming phonesthemes. Third, the speaker stresses extreme states by enlarging the concepts on the outskirts, forsaking the use of words for average observations in the limit. Evolutionary dynamics allow individuals to learn these patterns and thereby efficient communication.

Formally, communication is modeled by a cheap-talk game of common interest, based on Jäger et al. (2011) which I have enriched by a stochastic noise channel. Since there are only finitely many messages to describe the compact convex state space, communication partitions observations into concepts. In contrast to the non-noisy model, the receiver cannot with certainty determine the concept from which the true observation stems. Instead, the posterior belief after observing a message is still fully supported on the whole state space, though with different weights. Likewise, the sender no longer faces a simple comparison of a single norm when contemplating which message to send, but rather a sum of norms, weighted by the transition probabilities of the noise channel. Despite these technical drawbacks, some results of the benchmark model

[1] While I argue that these features can be explained in a formal model, an anonymous referee has rightfully pointed out there are plenty counterexamples in natural languages such as contronyms and homophones that have the same or similar pronunciation, but different or even opposite meanings, e.g., *original* or *fast*.

generalize. The expected loss is still convex and allows for Pareto optimal, i.e., efficient, communication. The sender clusters observations into convex subsets, providing a foundation for the linguistic conjecture that simple words have convex categories even in a noisy environment. Communication improves welfare if and only if the sender induces the receiver to take at least two different actions. Under a quadratic loss, I show that the sender can indeed always improve welfare if the noise is informative. Efficient languages are as separating as possible, maximizing the spread of the induced actions. In contrast to the benchmark model of Jäger et al. (2011), equilibria no longer amount to Voronoi tessellations as the optimal receiver replies need not lie in the concepts proposed by the sender. As a result, the nice characterizations of equilibria of Jäger and co-authors do no longer hold. However, noise offers new insights in the ways natural languages deal with communication errors. For a combinatorial message space consisting of sequences of letters, I draw on a noise channel from information theory that respects the combinatorial message structure. I study two extensive examples to link intuitive linguistic properties of communication under noise to equilibrium outcomes of my model. Efficiency requires a robust choice of the concepts' geometry, making concepts in the shape of a square more efficient than triangular ones. The distribution of words to the concepts likewise influences the quality of communication and optimally assigns clearly distinguishable words to extreme ends of the state space. By restricting the usage of words, describing states close to the ex-ante expected state, harmful misunderstandings are buffered if noise levels increase. I show that efficient communication precludes vagueness and provide an evolutionary foundation, establishing that efficient communication under noise can be learned and produces the discussed patterns.

Related literature: This article builds upon the common-interest cheap-talk game of Jäger et al. (2011), in which a sender is restricted to a finite set of messages to communicate observations out of a continuum. Jäger and co-authors identify equilibria as Voronoi tessellations, i.e., a partition, the elements of which are convex sets, consisting of all states closest to a receiver's action. Interpreting the partition elements as categories in a linguistic context formalizes the idea that simple words have convex categories, a linguistic conjecture of Gärdenfors (2004). To this end, Jäger and co-authors build on the previous work of Jäger (2007) and give it a game-theoretic foundation that characterizes equilibria. By adding noise, I extend their model and replicate some of their properties, including convexity of categories. However, equilibria under noise need not be Voronoi tessellations as induced actions may not lie in their respective partition element. Studying economic models of communication under noise bears insights into natural languages. If communication is noisy, Nowak and Krakauer (1999) find that there is an upper bound for the number of signals that individuals can use to improve the welfare of communication. However, constructing sequences of signals can overcome this boundary, thus Nowak and Krakauer provide a theoretical framework for why natural languages use combinatorial messages as in section 4. Considering the workhorse of cheap-talk games (Crawford and Sobel, 1982), Blume et al. (2007) prove that small amounts of noise generically permit welfare improvements between a biased sender and an unbiased receiver. By studying a combinatorial message space as in section 4, Hernández and von Stengel (2014) link a communication game with encoding and decoding strategies under noise to coding theory when the state space is finite. While not all codes are part of an equilibrium, all receiver-optimal codes are. Surprisingly, for a binary alphabet and a weak monotonicity condition, all codes are part of an equilibrium. While I restrict the message space to be finite, Martel et al. (2019) argue that coarse communication arises naturally from a trade-off between precision and difficulty to interpret precise descriptions, in absence of conflict or bounded rationality, which is assumed in this article. Under coarse communication due to bounded rationality of the sender, Cremer et al. (2007) find that efficient communication reserves precise terms for frequent observations and vague ones for unusual events. The trade-off between complexity of messages and imprecision cost is also studied in Dilmé (2024). Dilmé again finds that simple messages are used for frequent pieces of information if the the set of messages is given, while only messages of the same precision are used if the word length is endogenously chosen by the sender. In contrast to the previous two sources, the example in section 4.5 reserves precision for states that are costly to be mixed up. While I focus on convex category formation under noise, Sobel (2016) explains why efficient organizational codes lump states together in intervals under weak assumptions on costs of differentiating (sets of) states by words. The seminal work of the linguist Gärdenfors (2004) brings together game theory and linguistics by introducing geometric objects, called conceptual spaces, which allow for modeling concept formation and the evolution of language. The linguistic work of Bergen and Goodman (2015) studies phenomena in pragmatics when communication is noisy. By using prosodic stress, a message gets less swamped by environmental sounds and therefore reduces the noise rate for the listener, a feature that resembles the insights from the example in section 4.5.

The rest of the paper is structured as follows. Section 2 introduces the formal model. Section 3 provides a game theoretic analysis of equilibria, identifying generalizations and differences to the benchmark model. Section 4 adds a combinatorial structure to the message space, allowing for the study of grammatical phenomena under noise. I show that efficient communication can be learned by means of evolution in Section 5. Section 6 concludes. The appendix contains proofs and calculations.

2 Model and notation

I follow the approach of Jäger et al. (2011), modeling the communication of an element out of a convex, compact state space $T \subset \mathbb{R}^L$ by means of finitely many messages from a set W as a common-interest cheap-talk game. An element $t \in T$ represents the sender's private information within a conversational topic and is interpreted as an observation the sender wants to inform the receiver about. The observation is drawn from a common prior distribution, given by a Borel measure μ_0 on T that is absolutely continuous w.r.t. the Lebesgue measure with density f_0. The sender communicates by sending a message $\mathrm{v} \in \mathrm{W}$. I generalize the framework by introducing a stochastic noise $\varepsilon\colon \mathrm{W} \to \Delta(\mathrm{W})$ that frustrates communication: It is possible that not the intended message v is being received, but instead an erroneous message w with probability $\varepsilon(\mathrm{w} \mid \mathrm{v})$. Having observed a message w, the receiver takes an action $\alpha(\mathrm{w}) \in T$, assigning w a representative state. The agents incur a loss in communication if the interpretation

$\alpha(\mathrm{w})$ and the state t do not coincide. The size of the loss is measured by a norm $\|.\|$ on T and weighted by a strictly convex and strictly increasing function $\ell\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, punishing large mismatches between state and interpretation.[2] A (pure) strategy of the sender is a (μ_0-measurable) function $\pi\colon T \to \mathrm{W}$, 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 of communication is given by

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

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

A strategy profile (π,α) is referred to as a *language* and 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 equilibrium, or equilibrium for short. Altering π on null sets does not change the expected loss. Note that if every message is flawlessly transmitted, i.e., $\varepsilon(\mathrm{w}\mid\mathrm{v}) = \mathbb{1}_{\mathrm{v}}(\mathrm{w})$ is the indicator function, the expected loss and hence the analysis reduces to the one in Jäger et al. (2011).

3 Equilibrium Analysis

In this section, I derive the best replies of the agents and give first examples of equilibria. First, I formalize the change in belief of the receiver after observing a message and provide a necessary and sufficient condition for communication to improve welfare. Second, I show that the sender is optimally communicating by minimizing a weighted sum of norms and prove existence of efficient equilibria. Under a quadratic loss, efficient communication requires the use of convex concepts.

3.1 Induced beliefs and the receiver's best replies

Communication provides information to the receiver, thereby changing their belief about the sender's observation, eventually inducing a reply. Assume that the receiver knows the communication strategy $\pi\colon T \to \mathrm{W}$ of the sender. The 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 $\mathrm{v} \neq \mathrm{w}$). The expected likelihood of receiving w under π is

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

After receiving the message w, the receiver updates their informational environment. If $\lambda^\pi(\mathrm{w}) > 0$, the posterior μ^π_{w} is computed by Bayes rule and has the density

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

[2] I thank an anonymous referee for pointing out the importance of using a norm instead of a metric in combination with ℓ. In particular, Lemma 3.1 would not hold for a general metric.

If $\lambda^\pi(\mathrm{w}) = 0$, the posterior belief (and action) of the receiver does not influence the expected loss as receiving w is a null set. In that case we set $\mu^\pi_{\mathrm{w}} := \mu_0$ by convention, expressing that the receiver sticks to their prior belief if the Bayesian update is undefined.

The set of induced posterior beliefs $\{\mu^\pi_{\mathrm{w}}\}_{\mathrm{w}\in\mathrm{W}}$ is a decomposition of the prior belief μ_0, weighted by the distribution λ^π over W. Formally, for any bounded random variable $X : T \to \mathbb{R}$

$$\mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[X]] = \mathbb{E}_{\mu_0}[X]. \tag{5}$$

This property is referred to as *Bayes-Plausibility*, cf. Kamenica and Gentzkow (2011), and allows us to re-write the expected loss as follows.

$$L(\pi, \alpha) = \mathbb{E}_{\mu_0}[\mathbb{E}_{\varepsilon(.\,|\,\pi(t))}[\ell(\|t - \alpha(\mathrm{w})\|)]] \tag{6}$$

$$= \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\ell(\|t - \alpha(\mathrm{w})\|)]]. \tag{7}$$

Expression (6) describes the expected loss as a weighted sum of the deficits occurring due to the error for each realized state t. In contrast, expression (7) aggregates the expected losses under each posterior and weighs them according to the probability with which the posterior is induced. The latter expression aids in characterizing the receiver's loss minimizing interpretation for any (induced) posterior belief.

Lemma 3.1. *For any belief $\mu \in \Delta(T)$ with positive density f, the expected loss $\mathbb{E}_\mu[\ell(\|t - s\|)]$ has a unique minimizing action.*

Proof All proofs are delegated to the appendix. □

While simple, the statement does not resort to any differentiability requirements common in cheap-talk games and is more general than the Bayesian estimators in the benchmark model Jäger et al. (2011).

Henceforth, denote by $\hat\alpha(\mu)$ the unique minimizer of a receiver holding the belief μ. I 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 message. First, the receiver's reply is based on the received message and its implicit meaning and interpretation. Second, the received message changes the listener's belief about the sender's observation leading to their reply.

Consider now the situation in which the sender is sending the same message for every state, effectively pooling all states and not revealing any information. The receiver's posterior beliefs are then all equal to μ_0 and their best reply is the *pooling action* $\alpha_{\mathrm{pool}} := \hat\alpha(\mu_0)$, thus ultimately pooling their reply for every received message. Given the receiver's behavior, the sender has no incentive to deviate and their strategies form a so-called *babbling equilibrium*, resulting in the *pooling loss* $L_{\mathrm{pool}} := \mathbb{E}_{\mu_0}[\ell(\|t - \alpha_{\mathrm{pool}}\|)]$.

In a setting of common interest, communication serves the purpose of reducing misunderstandings, i.e., the expected loss. In particular, communication should foster understanding and result in a welfare improvement compared to the situation where individuals cannot or do not exchange information. The following result fully characterizes communication devices that improve upon the pooling loss if the receiver plays a best reply.

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

As long as the receiver knows the communication device π, the presence of signals cannot be detrimental to communication. For a strict welfare improvement, it is necessary and sufficient that the sender provides information in a way that makes the receiver deviate from the pooling action. The following corollary states two readily verifiable conditions under which communication does not improve upon 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,* $\hat{\alpha} \equiv \alpha_{\text{pool}}$ *and thus* $L(\pi, \hat{\alpha}) = L_{\text{pool}}$.

Corollary 3.3 demonstrates that non-beneficial communication can be due to bad communication or the perturbing noise by looking at two extreme cases: First, the communication device may not be meaningful, i.e., received messages never offer additional information. Second, if the error channel is *uninformative*, i.e., receiving a message w is equally likely under any sent message v and the channel can thus not convey any information. I call a channel *informative*, if it is not uninformative.
As in the non-noisy case, merely changing the belief is not enough to improve the outcome of communication upon the pooling loss, as the following example shows.

Example 3.4. Let $\mu_0 \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2}])$ with $\mathrm{W} = \{L, M, R\}$, $\ell \circ \|.\| = |.|^2$ and the noise fulfill $\varepsilon(M \mid M) = 1$ and $\varepsilon(\mathrm{w} \mid \mathrm{v}) = \frac{1}{2}$ for any $\mathrm{v}, \mathrm{w} \in \{L, R\}$. Consider $\pi([-\frac{1}{2}, -\frac{1}{6})) = L$, $\pi([-\frac{1}{6}, \frac{1}{6}]) = M$ and $\pi((\frac{1}{6}, \frac{1}{2}]) = R$. Then $\hat{\alpha}(\mathrm{w}) = \alpha_{\text{pool}} = 0$ for all $\mathrm{w} \in \mathrm{W}$ and $(\pi, \alpha_{\text{pool}})$ constitutes an equilibrium with pooling loss. However, the receiver's posterior belief is not equal to μ_0. For instance, the receiver can exclude the outer intervals if M has been received, resulting in the posterior $\mathcal{U}([-\frac{1}{6}, \frac{1}{6}])$. Figure 1 illustrates the example.

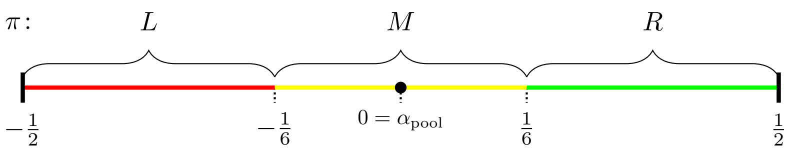


Fig. 1: Communication may not improve welfare even if it changes beliefs and involves convex partitions. Illustration of the equilibrium in Example 3.4, where $\mu_0 \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2}])$, $\mathrm{W} = \{L, M, R\}$, $\varepsilon(M \mid M) = 1$ and $\varepsilon(\mathrm{w} \mid \mathrm{v}) = \frac{1}{2}$ for any $\mathrm{v}, \mathrm{w} \in \{L, R\}$.

3.2 Efficient languages and the sender's best reply

In the previous section we have seen that communication cannot make individuals worse off in equilibrium despite an inhibiting noise. We now characterize and establish the existence of efficient communication by studying the sender's best reply.
A language (π, α) is called *efficient* if it minimizes the expected loss $L(\pi, \alpha)$ over all possible languages. As efficient communication is in both agents' interest, an efficient language is necessarily an equilibrium. Example 3.5 provides an example.

Example 3.5. Let $T = [-\frac{1}{2}, \frac{1}{2}]$, $\mu_0 \sim \mathcal{U}(T)$ and $\mathrm{W} = \{A, B\}$. The noisy channel confounds a sent message with probability $p \in [0, 1]$. Assume a quadratic loss, i.e., $\ell \circ \|.\| = |.|^2$. Figure 2 depicts an efficient language for every p. The sender cuts the interval into a left and a right one side to make the receiver move away from the pooling action 0. If p increases, the best replies come closer to the pooling action as long as $p \leq \frac{1}{2}$. At $p = \frac{1}{2}$ the noisy channel is uninformative.

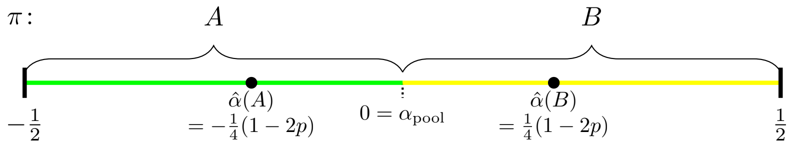


Fig. 2: An efficient language in the setting of Example 3.5. The sender uses different messages trying to reveal on which side of the interval the true state is situated. The closer the probability p of confusing the two words goes to $\frac{1}{2}$, the closer the optimal receiver replies are to the pooling action.

The above example marks a key difference to the benchmark model of Jäger et al. (2011): In their non-noisy model, an efficient language as well as a strict equilibrium both form a *Voronoi tessellation*, generated by the associated set of receiver actions (Theorem 1 and 2). However, for $p > \frac{1}{2}$, the induced actions in Example 2 do not lie in their corresponding cells anymore and thus do not form a Voronoi tessellation.

Efficient languages can be explicitly computed by internalizing a best reply of the sender for any interpretation map. For any interpretation $\alpha \colon \mathrm{W} \to T$ of the receiver and state t, the sender best responds by sending any message out of the (non-empty) set

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

Notably and in contrast to the receiver's case, the sender's best reply is not unique. Despite changing the sender's strategy on a null set, the sender more importantly can be indifferent between sending two or more different messages. While the latter issue only occurs for a null set under the quadratic loss induced by the Euclidean norm, cf. Proposition 4.4, there can be thick indifference sets for other norms, such as the maximum norm, cf. Figure 1 in Jäger et al. (2011).

Fixing any strict ordering on W, we nevertheless can derive a measurable partition $C^\alpha = \{C^\alpha_{\mathrm{v}}\}_{\mathrm{v}\in \mathrm{W}}$ of T, where each *concept* or *cell* C^α_{v} consists only of states where v is a minimizing reply of the sender given α, i.e., an element of (8). Assigning $\pi(t) = \mathrm{v}$ if and only if $t \in C^\alpha_{\mathrm{v}}$ yields thus a best reply to α. Details are given in the proof of the upcoming proposition. As a straightforward generalization of Lemma 1 in Jäger et al. (2011), we establish the existence of efficient languages in the presence of noise.[3]

Proposition 3.6. *Efficient languages exist.*

Even if agents were unable to find or agree on an efficient equilibrium play, they should not use mixed strategies, i.e., randomize their behavior. Any randomization further blurs proper understanding under common interest, thereby increasing the expected loss, cf. Blackwell (1953). In further analogy to Jäger et al. (2011), we thus have:

Proposition 3.7. *For any mixed language (σ, τ) there is a pure language (π, α) with weakly smaller loss $L(\pi, \alpha) \le L(\sigma, \tau)$.*
The inequality can be made strict if τ is not a pure strategy or the support of $\sigma(. \mid t)$ contains a message not in the argmin-set (8) *for positive mass of states t.*

3.3 Convexity of meaning under quadratic loss

Ever since the seminal work of Crawford and Sobel (1982), the most prominent loss function in the study of communication games is the quadratic loss. In the following, we restrict our attention to the multidimensional quadratic loss functional, given by $\ell \circ \|.\| = \|.\|^2$, i.e., the square value of the Euclidean norm. For an arbitrary loss, I show that the sender strictly improves welfare by maximizing a certain spread of receiver interpretations, and communicates by using convex concepts that do not allow for vagueness.

Our first observation pins down immediate analytical properties of languages under noise and a quadratic loss.

Lemma 3.8. *Having any (posterior) belief μ with continuous density $f > 0$, the receiver's unique best interpretation is $\hat{\alpha}(\mu) = \mathbb{E}_\mu[t]$. Given a communication device π and the induced best reply $\hat{\alpha}$, the expected loss is*

$$L(\pi, \hat{\alpha}) = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu^\pi_{\mathrm{w}}}[\big\|t - \mathbb{E}_{\mu^\pi_{\mathrm{w}}}[t']\big\|_2^2]] = \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2]. \tag{9}$$

Furthermore, the receiver plays the pooling action on average, $\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \alpha_{\mathrm{pool}}$.

The following result implies that, under a quadratic loss, efficient communication can always strictly improve upon babbling if the noisy channel is informative. Proposition 3.9 is thus the counterpart of Corollary 3.3. In conjunction they assert that communication can improve upon the pooling loss if and only if the noise channel is informative.

[3]Under common interest, Pareto optima are equilibria. I thank an anonymous referee for pointing out much weaker conditions of continuity than the ones used here that guarantee the existence of Pareto optima.

Proposition 3.9. *If ε is informative, there is (π, α) with $L(\pi, \alpha) < L_{\text{pool}}$.*

While at first glance it seems convincing that Proposition 3.9 should for most general ℓ and $\|.\|$, there is a quaint observation about convex functions that has raised my doubt: The proper convex combination of strictly convex functions may not have a minimizer on (an open set around) the line segment of their respective minimizers.[4] Consequently, it is neither clear that the minimizer of $\mathbb{E}_{C,\mu_0}[\ell(\|t-s\|)]$ has a minimum on the interior of a convex cell C, nor that the solution the the receiver's problem $\arg\min_s \sum_{\mathrm{w}} \varepsilon(\mathrm{v} \mid \mathrm{w})\mu_0(\pi^{-1}(\mathrm{w}))\mathbb{E}_{\pi^{-1}(\mathrm{w}),\mu_0}[\ell(\|t-s\|)]$ is not equal to α_{pool}.

In the remainder of this section, I elicit some more geometric and structural properties of efficient communication. By expression (9), 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. Re-writing

$$L(\pi, \hat{\alpha}) = L_{\text{pool}} - \left(\mathbb{E}_{\lambda^\pi}[\|\hat{\alpha}(\mathrm{w})\|_2^2] - \|\alpha_{\text{pool}}\|_2^2\right). \tag{10}$$

and recalling $\alpha_{\text{pool}} = \mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})]$, an efficient language is any 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.*

Another necessary condition of efficiency that is shared with Jäger et al. (2011) is the partition into convex cells under a quadratic loss.

Proposition 3.10. *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.*

If the sender plays a best reply to the receiver's interpretation map α, an induced cell corresponds to the indicative meaning of its corresponding message. Each cell thus forms a category in the sense of (Lewis, 1969, p. 165ff), cf. also (Jäger et al., 2011, p. 5). Interpreting abstract messages as words in a natural language, our result thus generalizes the linguistic idea that *(simple) words have convex categories* to communication under noise, cf. (Gärdenfors, 2004; Jäger, 2007; Jäger et al., 2011).
Note Proposition 3.10 draws heavily on the assumption of a quadratic loss, as this induces linear indifference curves for types for sending any of two messages.

4 The geometry of efficient communication

In this section, I impose more structure on the message space and the error term, which allows to study the geometry and patterns of efficient communication. To this end, I introduce a combinatorial message space, the messages of which are supposed to resemble words in a natural languages. The noise function that I consider respects the combinatorial structure of words. I discuss how this noisy channel changes beliefs and

[4] Consider, e.g., $\frac{1}{2}(f_1 + f_2)$, with $f_1(x) = x_1^2 + x_2^2$, $f_2(x) = 10(x_1 - 1)^2 + (x - 1)^2$, found at https://math.stackexchange.com/questions/1866379/minimum-of-the-sum-of-strictly-convex-functions.

allows for equilibrium selection under limit cases for the noise. Adding a quadratic loss, I then prove that vagueness does not aid communication. The section concludes with the extensive study of two examples, indicating the following geometric properties: Noise is best dealt with by reserving distant words to states that result in a huge loss if confused. Some geometric shapes for concepts buffer errors better than others. 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.

4.1 Combinatorial message space and metric-dependent noise channel

Natural languages are highly combinatorial: Text is made up of sentences, sentences consist of words, words are sequences of letters, syllables, phonems. We formalize this observation by considering sequences of building blocks of a fixed length n, which is common in information theory and has proven a solid baseline for coding theory, cf. Roth (2006). The word space is $\mathrm{W} = \mathcal{A}^n$ where we refer to the finite, non-singleton set $\mathcal{A}$ as *alphabet* and to its elements as *letters*. From an intuitive point of view, mixing up two words should happen more frequently for words that are similar than for ones that are very different from one another. For instance, the English words "flank" and "plank" can be more easily misunderstood than "flank" and "igloo". A natural noisy channel should respect this observation.

For the remainder of this section, I restrict attention to a noise channel on $\mathrm{W} = \mathcal{A}^n$ that is motivated as follows. The letters of a word v are sent independently one after another. There is a fixed *crossover probability* $p \in [0,1]$ with which a single letter a is transmitted incorrectly. In that case, the received letter is drawn from the uniform distribution over all remaining $m := \#\mathcal{A} - 1$ letters in $\mathcal{A} \setminus \{a\}$. Mathematically, the noisy channel is defined by the transition probabilities

$$\varepsilon(\mathrm{w} \mid \mathrm{v}) := (1-p)^{n-d(\mathrm{w},\mathrm{v})} \cdot \left(\frac{p}{\#\mathcal{A}-1}\right)^{d(\mathrm{w},\mathrm{v})}, \tag{11}$$

where $d(\mathrm{w},\mathrm{v})$ counts the number of entries in which v and w differ

$$d\colon \mathrm{W} \times \mathrm{W} \to \mathbb{N}_0, ((\mathrm{w}_k)_k, (\mathrm{v}_k)_k) \mapsto \#\{k \in \{1,\dots,n\} \mid \mathrm{w}_k \neq \mathrm{v}_k\}, \tag{12}$$

and under the usual convention $0^0 := 1$. The metric $d(\mathrm{w},\mathrm{v})$ on W is called *Hamming distance* and the noise channel the $\#\mathcal{A}$-ary symmetric channel of length n, cf. Hamming (1950). They play a crucial role in many applied fields related to information theory, modeling error transitions in telecommunication, data storage, but also in DNA heritage of cell-divisions, cf. (Roth, 2006; MacKay, 2002; Cover and Thomas, 2006)).

Note that the probability $\varepsilon(\mathrm{w} \mid \mathrm{v})$ that w is received if v is sent depends only on the Hamming distance $d(\mathrm{w},\mathrm{v})$ and the crossover probability p. Especially, $\varepsilon(\mathrm{w} \mid \mathrm{v}) = \varepsilon(\mathrm{v} \mid \mathrm{w})$. The following remark captures the intuition about distant words being less likely received by error.

Remark 4.1. Suppose $\mathrm{v} \in \mathrm{W}$ is the word sent.

1. If $p = 0$, there is no noise, i.e., v is received with probability 1.
2. If $0 < p < \frac{m}{m+1}$, $\varepsilon(\mathrm{w} \mid \mathrm{v})$ is decreasing in $d(\mathrm{w}, \mathrm{v})$. Especially it is most likely to receive v.
3. If $p = \frac{m}{m+1}$, ε is uninformative.
4. If $\frac{m}{m+1} < p < 1$, i.e., $\varepsilon(\mathrm{w} \mid \mathrm{v})$ is increasing in $d(\mathrm{w}, \mathrm{v})$. Especially, words with maximum distance $d(\mathrm{w}, \mathrm{v}) = n$ are most likely received.
5. If $p = 1$, only words with the maximal 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 vanishes completely at the *uninformativeness bound* $p = \frac{m}{m+1}$.

The amount of information transmitted through the noisy channel is quantitatively measured by *Shannon entropy*. Recall that entropy for a probability distribution P on a finite set X, it is defined as $H(P) = -\sum_{x \in X} P(x) \cdot \log(P(x))$, with the convention $0 = P(x) \cdot \log(P(x))$ if $P(x) = 0$. Note that $\varepsilon(. \mid \mathrm{v})$ defines a probability distributions on W for any v. Choosing a different v only permutes the probabilities on W, thus $H_\varepsilon(p) := H(\varepsilon(. \mid \mathrm{v}))$ is well-defined and independent of v. Choosing the base of the logarithm as $\#\mathrm{W}$, $H_\varepsilon(p)$ can be interpreted as the percentage of how much information is lost during the transmission.

Proposition 4.2. *The entropy of the $\#\mathcal{A}$-ary symmetric error channel of length n with crossover probability p is*

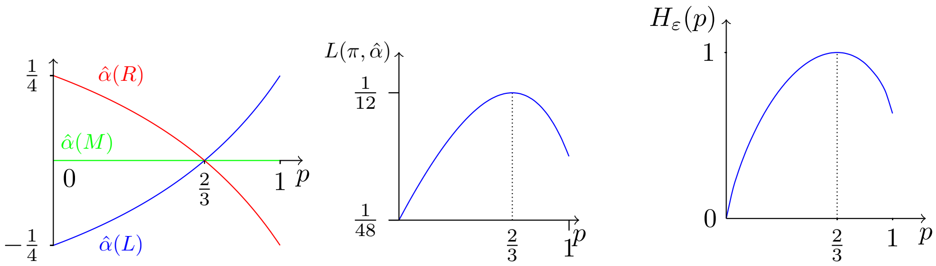
$$H_\varepsilon(p) = -n \cdot (p \log(p) + (1 - p) \log(1 - p)) + n \cdot p \log(m). \tag{13}$$

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

Interestingly, informative communication is possible beyond the uninformativeness bound, $\frac{m}{m+1} < p$, though 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. For binary channels ($m = 1$) the roles of the two letters (bits) simply switch and entropy is symmetric around the uninformativeness bound $p = \frac{1}{2}$. Figure 3 relates entropy and expected loss in an example.

4.2 Bayesian updates and limit cases

How does communication change beliefs when noise increases from a non-noisy level to an uninformative one? For instance, if you hear the non-existent word "orunge", you



(a) Actions (b) Expected loss (c) Entropy

Fig. 3: Optimal replies, expected loss and entropy of the communication device $\pi\colon T \to \mathrm{W}, \pi([-\frac{1}{2}, 0]) = L, \pi((0, \frac{1}{2}]) = R$ with $\mu_0 \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2}])$ and $\mathrm{W} = \{L, M, R\}$. The uninformativeness bound is $p = \frac{2}{3}$ where all actions coincide with the pooling action $\alpha_{\text{pool}} = 0$ and expected loss and entropy are maximal. For $p = 1$, optimal receiver replies to the messages L, R can be recovered. However, M will now be received with positive probability, leading to a loss in information in comparison to $p = 0$. Since $L(1)/L(2/3) = 0.625 \approx H_\varepsilon(1)$, entropy roughly captures expected loss.

probably conclude that the sender intended to say “orange”. This reasoning applies for arbitrary small amounts of noise, allowing thus for equilibrium selection in the case without noise. Similarly, if you believe that the noise produces completely random words that are not in correlation to the sender’s intended message, you ignore it. I formalize these insights by calculating the updated beliefs of a message under the $\#\mathcal{A}$-ary symmetric channel of length n, and taking limit cases for the crossover probabilities. For $0 < p < \frac{m}{m+1}$, Bayes rule can always be applied to compute the posterior beliefs, the density of which is given by

$$f_{\mathrm{w}}^{\pi}(t) = f_0(t) \cdot \left(\int_T \left(\frac{p}{(1-p)m} \right)^{d(\mathrm{w}, \pi(t')) - d(\mathrm{w}, \pi(t))} \mu_0(\mathrm{d}t') \right)^{-1} \tag{14}$$

for a received word w. The following proposition elicits the receiver’s posterior beliefs for the limit cases $p \to 0$ and $p \to \frac{m}{m+1}$.

Proposition 4.3. *Let π be known to the receiver who observes $\mathrm{w} \in \mathrm{W}$. Then the following properties hold.*

1. $\lim\limits_{p \to \frac{m}{m+1}} f_{\mathrm{w}}^{\pi}(t) = f_0(t)$.
2. *Let* $d^* := \min\{d \in \{0, \dots, n\} \mid \mu_0(\{t' \mid d(\mathrm{w}, \pi(t')) = d\}) > 0\}$.
   (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* $\lim\limits_{p \to 0} f_{\mathrm{w}}^{\pi}(t) = f_0(t) \cdot \mu_0(\{t' \mid d(\mathrm{w}, \pi(t')) = d(\mathrm{w}, \pi(t))\})^{-1}$.
   (c) *If* $d^* > d(\mathrm{w}, \pi(t))$ *the limit of the posterior belief for* $p \to 0$ *is not defined.*

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. First, the receiver determines the closest words to the received w that are sent by sender with positive probability. Let the according distance be d^* and let the receiver contemplate about the sender's observation 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 $d^* = d(\mathrm{w}, \pi(t'))$. 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. 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.

4.3 Non-vagueness of communication

Starting at which temperature does one 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. (Lipman, 2009; Sorensen, 2023). I generalize a result from Jäger et al. (2011), showing that the sender does not use vague concepts in equilibrium under a quadratic loss even when facing noise in the form of a $\#\mathcal{A}$-ary symmetric channel of length n. Efficient communication under noise thus does not profit from vagueness and switches from "hot" to "not hot" at a precise temperature.

Proposition 4.4. *Let ε by the $\#\mathcal{A}$-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.

The proposition states that concepts in equilibrium generically have sharp boundaries, if two words v, v' are not *synonyms*, i.e., $\alpha(\mathrm{v}) \neq \alpha(\mathrm{v}')$. In particular, there is no smooth transition between two concepts in efficient communication.

4.4 Example: The shape of cells and their labeling

The patterns in which we structure our communication determines how robust a language is against noise. On a two dimensional state space, I analyze four different language which differ in two aspects. The first aspect concerns the geometric shape of the cells the sender uses to pool observations. In the example, concepts take on either a quadratic or a triangular shape. The second aspect is about the choice of messages attached to these cells. We can either use similar words for close concepts or different ones. The studied equilibria suggest that that quadratic concepts buffer errors better than triangular ones and that distant states should be assigned distant words.

Let $T = [-\frac{1}{2}, \frac{1}{2}]^2$ be endowed with the uniform distribution $\mu_0 \sim \mathcal{U}(T)$ and consider a quadratic loss, in particular $\alpha_{\text{pool}} = (0, 0)$. The message space is $\mathrm{W} = \{A, B\}^2$ and the noise is a binary symmetric channel of length two and crossover probability p. The communication devices depicted in Figure 4 constitute equilibria together with the replies from Table 1 for every p. The precise derivations and proofs are given in the appendix.

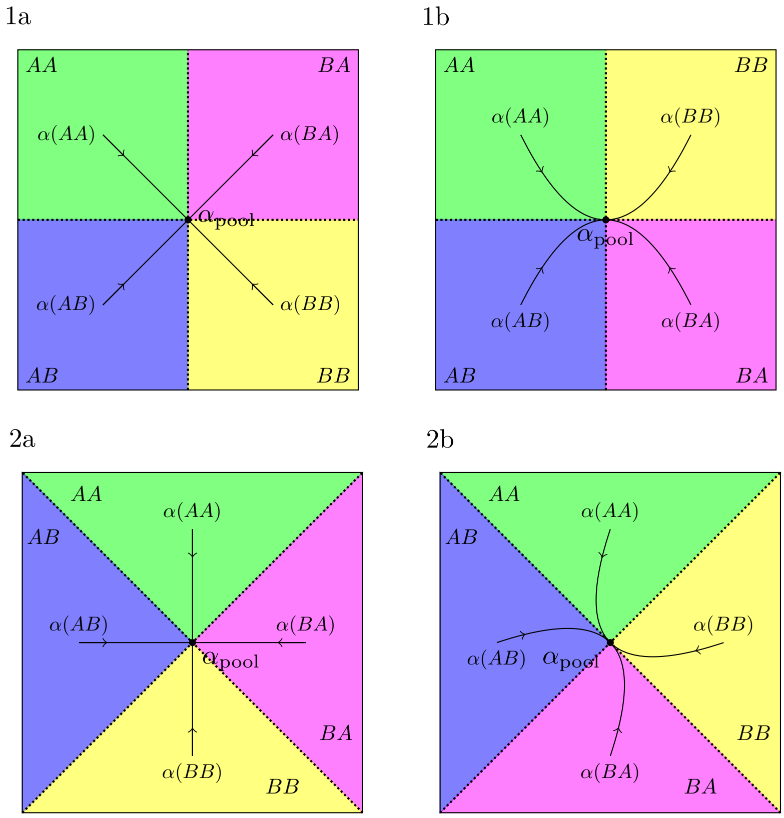


Fig. 4: Four communication devices on $[-\frac{1}{2}, \frac{1}{2}]^2$. The black lines depict the loci of the receiver's best replies for different crossover probabilities p, given in Table 1. The best replies move from the center of the resp. cells to the pooling action for $0 \leq \leq \frac{1}{2}$, establishing equilibrium. While the best replies in 1a and 2a move to α_{pool} in a straight line, the best replies in 1b and 2b are curved, being pulled closer to the cell of their neighbor with closer words.

Languages 1a and 1b assign quadratic concepts to observations, while 2a and 2b use triangularly shaped ones. 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 neighboring cell, the assigned word of which has distance 2.

The best replies are sketched in Figure 4, the precise loci for varying p are given in Table 1. As is clear from Proposition 4.3, the receiver's interpretations start at the center of each cell and continuously move to the pooling action. Notably, the loci in 1a and 2a are straight lines, whereas 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 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.

The expected loss of the four languages is depicted in Figure 5 and provides an assessment of their respective performance, summarized by two observations. First, each type-a language outperforms their type-b counterpart for each noise level. The assignment of messages to the cells thus is an important determinant of the efficiency of a language. In order to reduce the harm from miscommunication, the sender assigns similar words to describe similar states and uses distant words to describe distant states. Second, the languages with quadratic cells outperform the triangularly 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.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 with quadratic cells to be stable outcomes of evolution, whereas type-2 languages are unstable, cf. Section 5.

4.5 Example: Stressing important states

Only a few words are really necessary to get the main idea of a sentence. The remaining words are either concerned with details or redundant. Emphasizing key words becomes more important the noisier the communication. In the linguistic field of pragmatics, Bergen and Goodman (2015) study how noise drives two phenomena of pragmatic inference: First, prosodic stress of the speaker ensures a stronger belief on the listener's end, thereby countering noise. Second, fragments of a sentence can communicate full propositions. The upcoming example on an interval hints at these phenomena as follows. For an increasing level of noise, the sender gives more stress to words describing

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 or by symmetry.

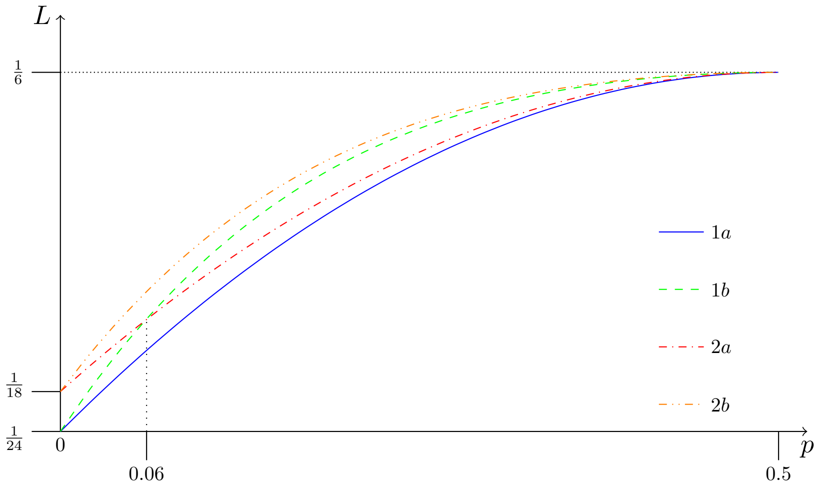


Fig. 5: Expected loss of the four languages from Figure 4 for different crossover probabilities p. For $p = 0$ the assignments of words to cells is irrelevant and 1a and 1b have a lower loss than 2a and 2b. At the uninformativeness bound $p = \frac{1}{2}$ the pooling loss realizes for all languages.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

states on the opposite ends of the interval by enlarging their cells. This allows for a clearer distinction between states which result in a high loss when confounded. At the same time, the cells close to the pooling action shrink with increasing noise, eventually forsaking a more nuanced division of the state space by letting only two words remain that constitute the proposition “left” and “right”.

Consider $T = [-\frac{1}{2}, \frac{1}{2}]$ with $\mu_0 \sim \mathcal{U}(T)$ and a quadratic loss. The word space is $\mathrm{W} = \{A, B\}^2$ and the noise is the binary symmetric channel of length two with crossover probability p. The efficient languages can be computed and are depicted in Figure 6. As in Subsection 4.4, opposite concepts are articulated by sending words that are not easily mistaken. In contrast to the example in Section 4.4, the low dimension of the current state space does not allow to sustain its structure for an increasing noise: The cells on the extreme ends get larger while the cells around the pooling action shrink.

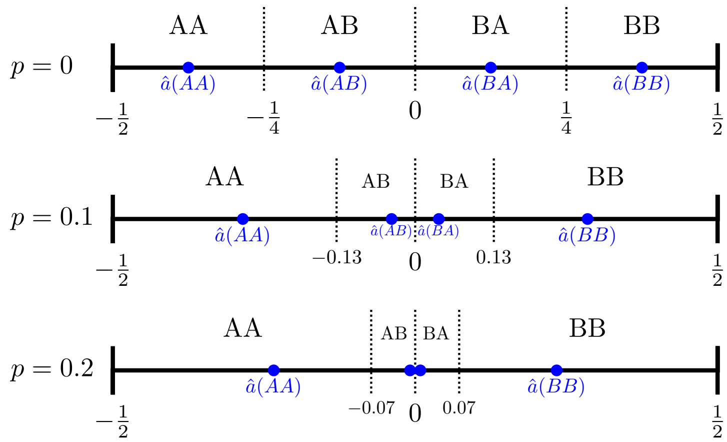


Fig. 6: Efficient languages in the setting of Section 4.5 for different crossover probabilities p. Words of maximal distance are used to articulate the boundaries and their cells enlarge with increasing error.

Think of a speaker describing 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 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. 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 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 simply becomes "tiny" and "tall".

5 Evolution

In this section, I address whether or not agents can learn (efficient) communication in face of noise over time. To this end, I straightforwardly extend the framework of evolutionary game theory developed in Jäger et al. (2011) to the more general setting with noise. I find that efficient communication is a stable outcome of many evolutionary dynamics under arbitrary noise channels. A numerical simulation for the best-reply dynamics shows however, that inefficient equilibria can also be stable.

To this end, interpret the population of a species at each point in time as 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 identify an individual by a language (π, α), consisting of 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. Hereby, we endow Σ with the topology of convergence in probability and identify communication devices that agree almost-surely. 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 for infinite strategy spaces has fortunately been established for certain dynamics and extends to our setting. These include the replicator, cf. Oechssler and 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

$$\Lambda((\pi, \alpha), (\pi', \alpha')) = \frac{1}{2} L(\pi, \alpha') + \frac{1}{2} L(\pi', \alpha) \tag{15}$$

for both. Describing populations of individuals by two probability distributions P, Q on the Borel space $\Gamma := \Sigma \times T^{\#\mathrm{W}}$, the expected loss is generalized to

$$\Lambda(P, Q) := \int_\Gamma \int_\Gamma \Lambda((\pi, \alpha), (\pi', \alpha'))\, P(\mathrm{d}\pi, \mathrm{d}\alpha)\, Q(\mathrm{d}\pi', \mathrm{d}\alpha'). \tag{16}$$

In the following, we explain in some more details the replicator dynamics, Oechssler and Riedel (2001). First, we measure the fitness of an individual using a pure strategy profile $\gamma := (\pi, \alpha)$ against a population Q by their average excess success which is given by the difference

$$\Psi(\gamma, Q) := \Lambda((\pi, \alpha), Q) - \Lambda(Q, Q). \tag{17}$$

Second, we assume that the fitness is proportional to the infinitesimal increase of the pure strategy of the next generation. This is formally described by a population Q_t for any point in time $t \geq 0$ and initial population P by means of the differential equations

$$\dot{Q}_t(A) = \int_A \Psi(\gamma, Q_t)\, Q_t(\mathrm{d}\gamma) \quad \text{and} \quad Q_0 = P, \tag{18}$$

for all measurable sets $A \subseteq \Gamma$. The following proposition shows that individuals are able to learn efficient communication in the presence of noise.

Proposition 5.1. *The following assertions hold.*

1. *The symmetrized loss function is a Lyapunov function for the replicator dynamics.*
2. *Locally optimal languages are Lyapunov stable w.r.t. the replicator dynamics.*

Note that the results also hold true for other evolutionary dynamics in a continuous setting, such as the regular and payoff monotone, cf. Heifetz et al. (2007), and the Brown-von-Neumann-Nash, cf. Hofbauer et al. (2009).

Figure 7 gives a tractable numerical illustration of evolution by means of the best-reply dynamics in the setting of Section 4.4. 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 reply to that at the next time. As the figure shows, their language quickly converges to an equilibrium. The depicted equilibrium is a local optimum that improves upon the pooling loss, but is not efficient, compare Subsection 4.4. 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.

6 Conclusion

Within our daily routine, errors in communication are ubiquitous. Despite this inhibiting factor, humans have found ways to communicate efficiently. This article formalizes noisy communication as a cheap-talk game of common interest and connects patterns of robust communication under a combinatorial message space to intuitive ones known in linguistics or day-to-day business, such as stress and the use of phonetic alphabets. Future research may characterize these observation in more general settings and answer the question whether efficient communication requires the sender to use all available words in the proposed noisy setting.

Acknowledgements. For inspiring discussions and comments I thank Andreas Blume, Gerhard Jäger, Christina Katt-Pawlowitsch, Frédéric Koessler, Salvador Mascarenhas, Frank Riedel, Joel Sobel, Benjamin Spector, three anonymous referees, the editorial board of the IJGT and participants of the 2nd Paris Workshop on Games, Decisions, and Language 2024 as well as the linguistic seminar at Tübingen University. Financial support by the DFG via grant Ri 1128-9-1 is gratefully acknowledged.

Appendix A Proofs & calculations

Proof of Lemma 3.1 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. 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

$$\begin{aligned}
&\mathbb{E}_\mu[\ell(\|t-(\lambda s_1+(1-\lambda)s_2)\|)] \\
\leq&\mathbb{E}_\mu[\ell(\lambda\|(t-s_1)\|+(1-\lambda)\|t-s_2\|)] \qquad \text{(A1)}\\
<&\mathbb{E}_\mu[\lambda\ell(\|t-s_1\|)]+\mathbb{E}_\mu[(1-\lambda)\ell(\|t-s_2\|)] \qquad \text{(A2)}
\end{aligned}$$

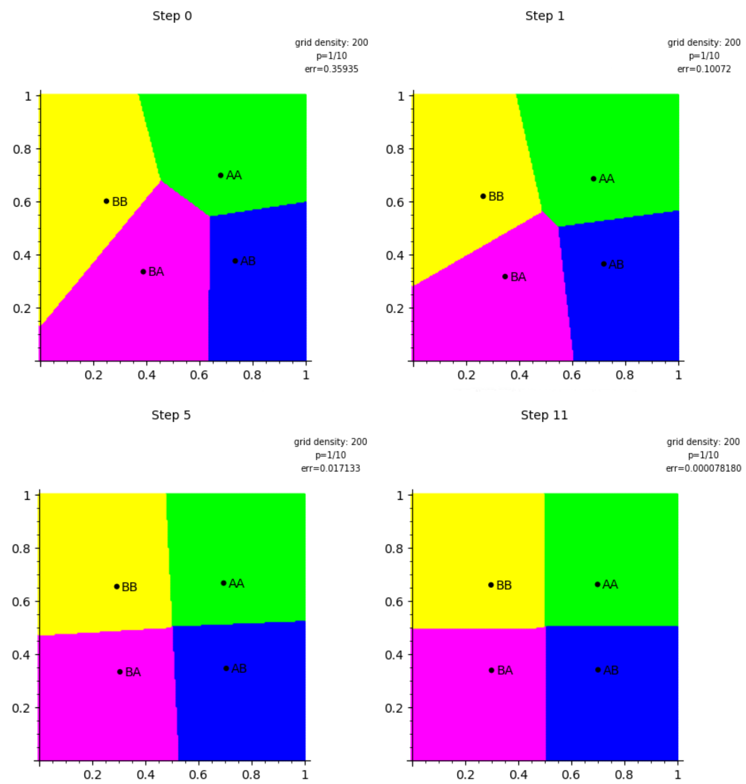


Fig. 7: Numerical simulation of the best-reply dynamics in the setting of Section 4.4 (shifted to the unit square). Starting with random interpretations, the agents take turn in playing a best reply to the previous strategy of their peer. Convergence to equilibrium is measured by the norm distance between two subsequent interpretation maps. Agents quickly learn one of the equilibria 1a or 1b from Figure 4 which, in contrast to 2a an 2b, appear to be stable and attractors. The sagemath-file to produce the above graphs is available on https://github.com/gbauch/noisy.

$$=\mathbb{E}_{\mu}[\ell(\|t-\hat{s}\|)].$$

Since the strictness of the inequality in (A2) is not obvious, I 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 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 Using Bayes-Plausibility, we have

$$L(\pi,\hat{\alpha}) = \mathbb{E}_{\lambda^{\pi}}\left[\mathbb{E}_{\mu^{\pi}_{\mathrm{w}}}\left[\ell(\|t-\hat{\alpha}(\mathrm{w})\|)\right]\right] \tag{A3}$$

$$\leq \mathbb{E}_{\lambda^{\pi}}\left[\mathbb{E}_{\mu^{\pi}_{\mathrm{w}}}\left[\ell(\left\|t-\alpha_{\mathrm{pool}}\right\|)\right]\right] \tag{A4}$$

$$= \mathbb{E}_{\mu_0}[\ell(\|t - \alpha_{\text{pool}}\|)] = L_{\text{pool}}. \tag{A5}$$

The inequality is strict if and only if there is a word w with $\lambda^\pi(\mathrm{w}) > 0$ and $\hat{\alpha}(\mathrm{w}) \neq \alpha_{\text{pool}}$ 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 Proposition 3.6 The proof is a simple extension to the the one of Lemma 1 in Jäger et al. (2011) to noise. Consider any pure strategy $\alpha\colon \mathrm{W} \to T$ of the receiver. Given t, the sender then optimally sends any word v out of the argmin-set in (8), which 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

$$C^\alpha_{\mathrm{v}} := \left\{ t \;|\; \text{ v is smallest w.r.t. } \leq_{\mathrm{W}} \text{ in } \operatorname*{arg\,min}_{\mathrm{v}' \in \mathrm{W}} \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}') \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \tag{A6}$$

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. Setting $\pi(t) = \mathrm{v} \iff t \in C^\alpha_{\mathrm{v}}$ defines a measurable best reply of the sender. We can thus re-write the joint loss minimization as a function depending only on α, namely

$$\min_\alpha \int_T \min_{\mathrm{v}} \left\{ \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|) \right\} \mu_0(\mathrm{d}t). \tag{A7}$$

Identify $\alpha\colon \mathrm{W} \to T$ with a point in T^N, $N := \#\mathrm{W}$. By Lebesgue's dominant convergence theorem and continuity of the integrand in α the minimum is attained. □

Proof of Proposition 3.7 First, the receiver always strictly favors a pure strategy $\alpha\colon \mathrm{W} \to T$ over a mixed one $\tau\colon \mathrm{W} \to \Delta(T)$ by Lemma 3.1. Second, 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 observe $t \in T$. The expected loss is thus given by

$$L(\sigma, \alpha) = \int_T \sum_{\mathrm{v} \in \mathrm{W}} \sigma(\mathrm{v} \mid t) \cdot \sum_{\mathrm{w} \in \mathrm{W}} \varepsilon(\mathrm{w} \,|\, \mathrm{v}) \cdot \ell(\|t - \alpha(\mathrm{w})\|) \; \mu_0(\mathrm{d}t). \tag{A8}$$

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

Proof of Lemma 3.8 The receiver's minimization problem reads as

$$\min_{\alpha \in T} \mathbb{E}_\mu[\|t - \alpha\|_2^2] = \int_T \sum_{k=1}^{L} (t_k - \alpha_k)^2 \, \mu(\mathrm{d}t). \tag{A9}$$

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

$$\hat{\alpha}_k(\mu) = \int_T t_k \, \mu(\mathrm{d}t) = \mathbb{E}_\mu[t_k]. \tag{A10}$$

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

$$\mathbb{E}_\mu[\left\|t - \mathbb{E}_\mu[t']\right\|_2^2] = \mathbb{E}_\mu[\|t\|_2^2] - \|\mathbb{E}_\mu[t]\|_2^2 \, , \tag{A11}$$

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

$$\mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[\left\|t - \mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t']\right\|_2^2]] = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[\|t\|_2^2] - \left\|\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]\right\|_2^2] = \mathbb{E}_{\mu_0}[\|t\|_2^2] - \mathbb{E}_{\lambda^\pi}[\left\|\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]\right\|_2^2]$$

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

$$\mathbb{E}_{\lambda^\pi}[\hat{\alpha}(\mathrm{w})] = \mathbb{E}_{\lambda^\pi}[\mathbb{E}_{\mu_{\mathrm{w}}^\pi}[t]] = \mathbb{E}_{\mu_0}[t] = \hat{\alpha}(\mu_0) = \alpha_{\mathrm{pool}}. \tag{A12}$$

□

Proof of Proposition 3.9 We can assume that $\alpha_{\mathrm{pool}} = 0$ by translating the state space and the measure. Since ε is informative, there are two words v, v' with $\varepsilon(\mathrm{v} \mid \mathrm{v}) \neq \varepsilon(\mathrm{v} \mid \mathrm{v}')$. Note that for any normal vector $\vec{n} \in \mathbb{R}^L$ the corresponding $L-1$ dimensional hyperplane $H := \{t \mid \langle \vec{n}, t\rangle = 0\}$ through $\alpha_{\mathrm{pool}} = 0$ separates T into two disjoint convex cells $C_1 := T \cap H_+ = \{t \in T \mid \langle \vec{n}, t\rangle > 0\}$ and $C_2 = T \cap H_- = \{t \in T \mid \langle \vec{n}, t\rangle \leq 0\}$. Since μ_0 is absolutely continuous w.r.t. to the Lebesgue measure, we can choose $\vec{n}$ in a way that the both cells have μ_0-measure $\frac{1}{2}$. Since μ_0 is absolutely continuous w.r.t. the Lebesgue measure, the minimizer $\mathbb{E}_{C_i,\mu_0}[t] = \arg\min_{s \in T} E_{C_i,\mu_0}[\|t - s\|_2^2]$ is an interior point of the respective C_i and unequal to α_{pool}. Now define $\pi(t) = \mathrm{v}$ if $t \in C_1$ and $\pi(t) = \mathrm{v}'$ otherwise. Since we have $\lambda^\pi(\mathrm{v}) = \varepsilon(\mathrm{v} \mid \mathrm{v})\mu_0(C_1) + \varepsilon(\mathrm{v} \mid \mathrm{v}')\mu_0(C_2) = \frac{1}{2}(\varepsilon(\mathrm{v} \mid \mathrm{v}) + \varepsilon(\mathrm{v} \mid \mathrm{v}')) > 0$, it is sufficient to show $\hat{\alpha}(\mathrm{v}) \neq \alpha_{\mathrm{pool}}$ in order to prove $L(\pi, \hat{\alpha}) > L_{\mathrm{pool}}$ by Proposition 3.2. To this end, consider the strictly convex function $\phi_\beta(s) := \beta\varphi_1 + (1-\beta)\varphi_2 \colon T \to \mathbb{R}$ for an arbitrary β. Since we consider a quadratic loss, a shot calculation reveals that the minimizer of ϕ_β is given by $\beta\mathbb{E}_{C_1,\mu_0}[t] + (1-\beta)\mathbb{E}_{C_2,\mu_0}[t]$. Since $\mathbb{E}_{C_1,\mu_0}[t] \neq \mathbb{E}_{C_2,\mu_0}[t]$, these minimizers are different for different β. Note that α_{pool} is the minimizer of $\phi_{\frac{1}{2}}$ and $\hat{\alpha}(\mathrm{v})$ is by definition the minimizer of ϕ_κ for $\kappa := \frac{\varepsilon(\mathrm{v} \mid \mathrm{v})\mu_0(C_1)}{\lambda^\pi(\mathrm{v})} = \frac{\varepsilon(\mathrm{v} \mid \mathrm{v})}{\varepsilon(\mathrm{v} \mid \mathrm{v}) + \varepsilon(\mathrm{v} \mid \mathrm{v}')} \neq \frac{1}{2}$. Consequently, $\alpha_{\mathrm{pool}} \neq \hat{\alpha}(\mathrm{v})$. □

Proof of Proposition 3.10 Revisit the argmin-set of sender (8) 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

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \cdot \|t - \alpha(\mathrm{w})\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \mid \mathrm{v}') \cdot \|t - \alpha(\mathrm{w})\|_2^2 \tag{A13}$$

$$\iff \sum_{\mathrm{w}} \left(\varepsilon(\mathrm{w} \mid \mathrm{v}') - \varepsilon(\mathrm{w} \mid \mathrm{v})\right) \cdot \left(-2\langle t \, , \, \alpha(\mathrm{w})\rangle + \|\alpha(\mathrm{w})\|_2^2\right) > 0 \tag{A14}$$

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

I introduce the notation $m := \#\mathcal{A} - 1$ and $\widetilde{p} := \frac{p}{(1-p)m}$ for the following proofs. We can thus compactly write $\varepsilon(\mathrm{w} \mid \mathrm{v}) = (1-p)^n \cdot \widetilde{p}^{d(\mathrm{w},\mathrm{v})}$.

Proof of Remark 4.1 The case $p = 0$ is clear. For $0 < p < \frac{m}{m+1}$, i.e., $0 < \widetilde{p} < 1$ we have that

$$\varepsilon(\mathrm{w} \mid \mathrm{v}) > \varepsilon(\mathrm{w}' \mid \mathrm{v}) \iff \widetilde{p}^{d(\mathrm{w},\mathrm{v})} > \widetilde{p}^{d(\mathrm{w}',\mathrm{v})} \iff d(\mathrm{w},\mathrm{v}) < d(\mathrm{w}',\mathrm{v}). \tag{A15}$$

The other cases follow similarly. □

Proof of Proposition 4.2 We start by calculating the entropy

$$
\begin{aligned}
&H(\varepsilon(.\,|\,\mathrm{v})) \\
&= -\sum_{w\in \mathrm{W}} (1-p)^{n-d(\mathrm{w},\mathrm{v})}\cdot\left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})}\cdot\log\left((1-p)^{n-d(\mathrm{w},\mathrm{v})}\cdot\left(\frac{p}{m}\right)^{d(\mathrm{w},\mathrm{v})}\right) && \text{(A16)}\\
&= -\sum_{d=0}^{n}\binom{n}{d}\cdot m^d\cdot(1-p)^{n-d}\cdot\left(\frac{p}{m}\right)^d\cdot\log\left((1-p)^{n-d}\cdot\left(\frac{p}{m}\right)^d\right) && \text{(A17)}\\
&= -\sum_{d=0}^{n}\binom{n}{d}\cdot(1-p)^{n-d}\cdot p^d\cdot\log\left((1-p)^{n-d}\cdot p^d\right) && \text{(A18)}\\
&\qquad +\log(m)\cdot\sum_{d=0}^{n}\binom{n}{d}\cdot d\cdot(1-p)^{n-d}\cdot p^d && \text{(A19)}\\
&= -\sum_{d=0}^{n}\binom{n}{d}\cdot(1-p)^{n-d}\cdot p^d\cdot\log\left((1-p)^{n-d}\cdot p^d\right)+np\log(m) && \text{(A20)}\\
&= -\,np\cdot\log(p)-n(1-p)\cdot\log(1-p)+np\log(m) && \text{(A21)}\\
&= n\cdot(H((p,1-p))+p\log(m)). && \text{(A22)}
\end{aligned}
$$

During the calculation we used the functional equation of the logarithm and formulae occurring often when dealing with binomial distributions, e.g., its mean. The 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 4.1 and yields $H(\mathcal{U}(\mathrm{W})) = \log(\#\mathrm{W})$. The other assertions follow readily from the calculated expression. □

Formulae for the example depicted in Figure 3. For $p\in[0,1]$ the optimal reply $\hat\alpha$ of the receiver is given by

$$\hat\alpha(L)=\frac{-2+3p}{8-4p},\quad \hat\alpha(M)=0,\quad \hat\alpha(R)=-\hat\alpha(L) \tag{A23}$$

and the expected loss given p can be calculated to be

$$L(\pi,\hat\alpha)(p)=\frac{1}{12}-2^{-5}\cdot\frac{(-2+3p)^2}{2-p}. \tag{A24}$$

□

Proof of Proposition 4.3 1. Follows immediately from continuity of the integrand in $\widetilde{p}$ and Lebesgue's theorem.

2. We split the integral in three parts by disassembling the state space T into the three stets defined by $\{t'\,|\,d(\mathrm{w},\pi(t'))\sim^* d(\mathrm{w},\pi(t))\}$ for $\sim^*\in\{<,=,>\}$.
   (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 of Proposition 4.4 From (A14) the sender is indifferent between sending v and v′ if the state is t if and only if

$$0 = \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \big(-2\,\langle t\,,\, \alpha(\mathrm{w})\rangle + \|\alpha(\mathrm{w})\|_2^2\big) \tag{A25}$$

$$\iff 0 = -\,2 \cdot \left\langle t\,,\, \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \alpha(\mathrm{w}) \right\rangle + \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \|\alpha(\mathrm{w})\|_2^2\,. \tag{A26}$$

Note that if $\vec{x} := \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \alpha(\mathrm{w}) \neq 0$ the solution set to (A26) 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 $\sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \|\alpha(\mathrm{w})\|_2^2 = 0$. If this is the case, any t in T (even $\mathbb{R}^L$) solves (A26). Otherwise, (A26) is equivalent to

$$0 = \sum_{\mathrm{w}} \big(\varepsilon(\mathrm{w}\,|\,\mathrm{v}') - \varepsilon(\mathrm{w}\,|\,\mathrm{v})\big) \cdot \alpha(\mathrm{w}) \tag{A27}$$

$$\iff 0 = \sum_{\mathrm{w}} \left(\widetilde{p}^{\,d(\mathrm{w},\mathrm{v}')} - \widetilde{p}^{\,d(\mathrm{w},\mathrm{v})}\right) \cdot \alpha(\mathrm{w}) =: Q(\widetilde{p}), \tag{A28}$$

where Q is a polynomial (with coefficients in the ring $\mathbb{R}^L$) 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 $p \in (0,1] \setminus \{\frac{m}{m+1}\}$. □

Calculations of Subsection 4.4 We prove the assertions and formulae of Subsection 4.4 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

$$\lambda^\pi(\mathrm{w}) = \int_T \varepsilon(\mathrm{w}\,|\,\pi(t))\,\mathrm{d}t = \sum_{\mathrm{v}} \varepsilon(\mathrm{w}\,|\,\mathrm{v}) \cdot \mu_0(C_{\mathrm{v}}) = \frac{1}{4}. \tag{A29}$$

(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.

$$\mathbb{E}_{\square(AB)}[t] := \mathbb{E}_{\mu_0,\square(AB)}[t] = \mu_0(\square(AB))^{-1} \cdot \int_{\square(AB)} t\,\mathrm{d}t \tag{A30}$$

$$= 4 \cdot \int_{-\frac{1}{2}}^{0} \int_{-\frac{1}{2}}^{0} (t_1, t_2)\, \mathtt{d}t_1 \mathrm{d}t_2 = \left(-\frac{1}{4}, -\frac{1}{4}\right). \quad \text{(A31)}$$

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

$$\hat{\alpha}(AA) = \mathbb{E}_{\mu^{\pi}_{AA}}[t] = \lambda^{\pi}(AA)^{-1} \cdot \int_T \varepsilon(AA \mid \pi(t)) \cdot t \, \mu_0(\mathrm{d}t) \quad \text{(A32)}$$

$$= 4 \cdot \sum_{\mathrm{w}} \varepsilon(AA \mid \mathrm{w}) \cdot \int_{\square(\mathrm{w})} t \, \mathrm{d}t \quad \text{(A33)}$$

$$= \sum_{\mathrm{w}} \varepsilon(AA \mid \mathrm{w}) \cdot \mathbb{E}_{\square(\mathrm{w})} \quad \text{(A34)}$$

$$= (1-p)^2 \cdot (-\tfrac{1}{4}, \tfrac{1}{4}) + p(1-p) \cdot \big((-\tfrac{1}{4}, -\tfrac{1}{4}) + (\tfrac{1}{4}, \tfrac{1}{4})\big) + p^2 \cdot (\tfrac{1}{4}, -\tfrac{1}{4}) \quad \text{(A35)}$$

$$= \frac{1}{4} \cdot (-1 + 2p, 1 - 2p). \quad \text{(A36)}$$

Analogously, by symmetry arguments or using Lemma 3.8 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 $\|\alpha(\mathrm{w}) - \alpha_{\text{pool}}\|_2 \searrow 0$ for $p \to \frac{1}{2}$, where $\alpha_{\text{pool}} = (0, 0)$ 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:

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \cdot \left\| \frac{1}{4} \cdot (1-2p, 1-2p) \right\|_2^2 = \frac{1}{8} \cdot (1-2p)^2. \quad \text{(A37)}$$

Having calculated $\mathbb{E}_T[\|t\|_2^2] = \frac{1}{6}$, we use (9) to obtain the expected loss

$$L(\pi_{1,a}, \hat{\alpha}) = \frac{1}{6} - \sum_{\mathrm{w}} \frac{1}{4} \cdot \frac{1}{8} \cdot (1-2p)^2 = \frac{1}{6} - \frac{1}{8} \cdot (1-2p)^2. \quad \text{(A38)}$$

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

$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{16} \cdot ((1-2p)^2 + (1-2p)^4), \quad \text{(A39)}$$

resulting in an expected loss of

$$L(\pi_{1,b},\hat{\alpha}) = \frac{1}{6} - \frac{1}{16}\cdot\left((1-2p)^2 + (1-2p)^4\right) \tag{A40}$$
$$= \frac{1}{6} - \frac{1}{8}\cdot(1-2p)^2\cdot(1-2p+2p^2). \tag{A41}$$

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

$$L(\pi_{1,a},\hat{\alpha}) < L(\pi_{1,b},\hat{\alpha}), \tag{A42}$$

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
$$\|\alpha(\mathrm{w})\|_2^2 = \left\|(0, \tfrac{1}{3} - \tfrac{2}{3}p)\right\|_2^2 = \frac{1}{9}\cdot(1-2p)^2. \tag{A43}$$
The resulting expected loss is
$$L(\pi_{2,a},\hat{\alpha}) = \frac{1}{6} - \frac{1}{9}\cdot(1-2p)^2, \tag{A44}$$
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
$$\|\hat{\alpha}(\mathrm{w})\|_2^2 = \frac{1}{9}(1-2p)^2(1-2p+2p^2) \tag{A45}$$
and hence
$$L(\pi_{2,b},\hat{\alpha}) = \frac{1}{6} - \frac{1}{9}(1-2p)^2(1-2p+2p^2), \tag{A46}$$
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 3.10 and Proposition 4.4 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.

1. In state t the sender prefers to send AA over BB if and only if

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w}\,|\,\mathrm{v})\,\|t - \alpha(AA)\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w}\,|\,\mathrm{v})\,\|t - \alpha(BB)\|_2^2 \tag{A47}$$
$$\iff \|t - \alpha(AA)\|_2 < \|t - \alpha(BB)\|_2 . \tag{A48}$$

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

$$\sum_{\mathrm{w}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \left\| t - \alpha(AA) \right\|_2^2 < \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \mid \mathrm{v}) \left\| t - \alpha(AB) \right\|_2^2 \tag{A49}$$

$$\begin{aligned} \iff & \left\| t - \alpha(AA) \right\|_2^2 - \left\| t - \alpha(AB) \right\|_2^2 \\ & < \widetilde{p} \left(\left\| t - \alpha(BB) \right\|_2^2 - \left\| t - \alpha(BA) \right\|_2^2 \right) \end{aligned} \tag{A50}$$

$$\begin{aligned} \iff & 2 \left\langle t \,,\, \alpha(AB) - \alpha(AA) + \widetilde{p}(\alpha(BB) - \alpha(BA)) \right\rangle \\ & + \left\| \alpha(AA) \right\|_2^2 - \left\| \alpha(AB) \right\|_2^2 + \widetilde{p}(\left\| \alpha(BA) \right\|_2^2 - \left\| \alpha(BB) \right\|_2^2) < 0. \end{aligned} \tag{A51}$$

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.

1. Observe that

$$\left\| t - AA \right\|_2 < \left\| t - AB \right\|_2 , \left\| t - BA \right\|_2 < \left\| t - BB \right\|_2 . \tag{A52}$$

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

$$\left\| t - AA \right\|_2^2 - \left\| t - AB \right\|_2^2 < 0 < \widetilde{p} \cdot \left(\left\| t - BB \right\|_2^2 - \left\| t - BA \right\|_2^2 \right) , \tag{A53}$$

   reveals that sending AA is preferred to AB (and analogously BA). Thus, AA is the unique best word to be send.
2. As before, preferring AA to BB is clear from $\left\| t - AA \right\|_2 < \left\| t - BB \right\|_2$. Since

$$0 \leq \left\| t - AA \right\|_2 < \left\| t - AB \right\|_2 , \left\| t - BB \right\|_2 < \left\| t - BA \right\|_2 , \tag{A54}$$

   we find

$$\left\| t - BA \right\|_2^2 - \left\| t - AA \right\|_2^2 > \left| \left\| t - AB \right\|_2^2 - \left\| t - BB \right\|_2^2 \right| \tag{A55}$$

$$> \widetilde{p} \cdot \left| \left\| t - AB \right\|_2^2 - \left\| t - BB \right\|_2^2 \right| \tag{A56}$$

$$\geq \widetilde{p} \cdot \left(\left\| t - AB \right\|_2^2 - \left\| t - BB \right\|_2^2 \right) , \tag{A57}$$

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

$$2 \cdot \left\langle t \,,\, -AA + AB + \widetilde{p}(BB - BA) \right\rangle$$

$$+\|AA\|_2^2 - \|AB\|_2^2 + \widetilde{p}(\|BB\|_2^2 - \|BA\|_2^2) \tag{A58}$$

$$= 4 \cdot \left\langle t\,,\, \underbrace{\tfrac{AB+BA}{2}}_{=:P} \right\rangle . \tag{A59}$$

The expression (A59) is smaller than zero in both cases for $t \in \Delta(AA)$:

(a) $t_1 < 0, t_2 > 0$ and $P_1 = 0, P_2 < 0$.

(b) $t = (y, z)$ 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. □

Proof of Proposition 5.1 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 and 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 $\{0, \dots, n\}$, 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

$$|L(\pi_k, \alpha_k) - L(\pi, \alpha))| \tag{A60}$$

$$\leq \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi_k(t))\ell(\|t - \alpha_k(\mathrm{w})\|) - \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t))\ell(\|t - \alpha(\mathrm{w})\|) \right| \mu_0(\mathrm{d}t) \tag{A61}$$

$$= \int_T \left| \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left(\ell(\|t - \alpha_k(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|)\right) \right| \mu_0(\mathrm{d}t) \tag{A62}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \left|\ell(\|t - \alpha_k(\mathrm{w})\|) - \ell(\|t - \alpha(\mathrm{w})\|)\right| \mu_0(\mathrm{d}t) \tag{A63}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \rho\, \mu_0(\mathrm{d}t) = \rho. \tag{A64}$$

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

$$|L(\pi, \alpha)| \leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t) \cdot) \left|\ell(\|t - \alpha(\mathrm{w})\|)\right| \mu_0(\mathrm{d}t) \tag{A65}$$

$$\leq \int_T \sum_{\mathrm{w}} \varepsilon(\mathrm{w} \,|\, \pi(t)) \cdot \bar{\ell}\, \mu_0(\mathrm{d}t) = \bar{\ell} < \infty. \tag{A66}$$

□

References

Bergen, L. and N.D. Goodman. 2015. The strategic use of noise in pragmatic reasoning. *Topics in cognitive science 7*(2): 336–350 .

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

Blackwell, D. 1953. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics 24*(2): 265–272 .

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

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

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

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

Cressman, R., J. Hofbauer, and F. Riedel. 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. https://doi.org/https://doi.org/10.1016/j.jtbi.2005.07.022 .

Dilmé, F. 2024. Precision vs complexity: Efficient communication and optimal codes. The article was formerly titled "Optimal languages", available at SSRN 3331077' (2018). The current version is available at the author's website https://www.francescdilme.com/ (accessed: November 8, 2024).

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. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x .

Heifetz, A., C. Shannon, and Y. Spiegel. 2007. What to maximize if you must. *Journal of Economic Theory 133*(1): 31–57 .

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

Hofbauer, J., J. Oechssler, and F. Riedel. 2009. Brown–von neumann–nash dynamics: The continuous strategy case. *Games and Economic Behavior 65*(2): 406–429. https://doi.org/https://doi.org/10.1016/j.geb.2008.03.006 .

Jäger, G. 2007, 10. The evolution of convex categories. *Linguistics and Philosophy* 30: 551–564. https://doi.org/10.1007/s10988-008-9024-3 .

Jäger, G., L.P. Metzger, and F. Riedel. 2011. Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals. *Games and Economic Behavior 73*(2): 517 – 537. https://doi.org/https://doi.org/10.1016/j.geb.2011.03.008 .

Kamenica, E. and M. Gentzkow. 2011, October. Bayesian persuasion. *American Economic Review 101*(6): 2590–2615. https://doi.org/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.* USA: Cambridge University Press.

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

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

Oechssler, J. and F. Riedel. 2001. Evolutionary dynamics on infinite strategy spaces. *Economic Theory 17*(1): 141–162 .

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

Sobel, J. 2016. 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 *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.)., eds. Zalta, E.N. and U. Nodelman. Metaphysics Research Lab, Stanford University.