To Surprise and Inform

Lav R. Varshney
IBM Thomas J. Watson Research Center

***Abstract*—In information overload regimes, it is necessary for messages to not only provide information but also to attract attention in the first place. Bayesian surprise is an information-theoretic functional that has been experimentally shown to measure the attraction of human attention. This paper studies the limits of reliable communication under a constraint on surprise so as to limit distraction: *surprise-constrained capacity*. It also considers *attention-seeking capacity*, where the goal is to maximize both information rate and surprise to attract attention. Properties of these functions are proven. There are no nontrivial tradeoffs for surprise-constrained capacity, but an interesting tradeoff arises for attention-seeking capacity; reversing the direction of constraint does not yield essentially equivalent problems.**

I. Introduction

Cognitive systems have the ability to automatically prioritize, which allows thinking and acting without being overwhelmed by either the external world or internal informational processes [1]. Attention placed on what is important while ignoring extraneous stimuli can enhance performance in challenging tasks or in the face of personal danger [2]. When the attention capacity of a system is exceeded, however, only a fraction of demands are completed [3]. Especially in the regime of *information overload* [4], the attention capacity of the receiver is a primary limiting factor for communication, beyond the physical limits of the communication channel.

In many communication settings the flood of messages in not only immense but also monotonously similar. Some have argued "it would be far more effective to send one very unusual message than a thousand typical ones" [5, p. 59]. Indeed, novel and surprising stimuli spontaneously attract attention [3], a fact well-known in marketing [6].

Surprising signals can sometimes draw the focus of cognitive systems and yet not provide them too much value [7]: such signals can be distractions that impair task performance [8]. Unlike in natural environments, ignoring stimuli in modern technological environments is often difficult for people [4]. Hence it may make sense to re-organize the environments themselves and the communication within them to be less extraneously surprising. To do so would require either full cooperation among senders and receivers or a regulatory authority like the FCC.

As a building block for this goal, the paper considers communication when a formal notion of signal surprise is restricted; in particular, the relationship between information rate that can be communicated and the amount of surprise that such communication causes. An explicit characterization for the relationship between the two quantities is determined and named *surprise-constrained capacity*.

When sender and receiver have aligned objectives or there is a regulatory regime in place, the level of surprise of a given signal can be restricted so as not to be distracting. On the other hand, the opposite problem is of a non-cooperative sender trying to be as surprising as possible to attract attention, while still furnishing information. By defining *attention-seeking capacity*, we further consider this alternate tradeoff between information rate and surprise.

Itti and Baldi have defined a measure of surprise termed *Bayesian surprise*, and have experimentally demonstrated its connection to attraction of human attention across different spatiotemporal scales, modalities, and levels of abstraction [9], [10]. The surprise of each location on a feature map is computed by comparing beliefs about what is likely to be in that location before and after seeing the information. This measure is also effective in judging perceived novelty within computational creativity systems [11]. Bayesian surprise is the measure of surprise used herein; its operational significance is due to extensive results from behavioral experiments.

The notion of information rate for reliable communication used herein is Shannon's; its operational significance will arise due to coding theorems.

Whereas we find no nontrivial tradeoffs between surprise and information rate for surprise-constrained capacity, there are rather intriguing tradeoffs for attention-seeking capacity under a maximum surprise criterion (as shown herein, an average surprise criterion is the same as mutual information). The difference between surprise-constrained and attention-seeking capacities is notable since reversing the direction of constraint in point-to-point communication problems usually yields largely equivalent problems [12].

Understanding the full strategic interaction between sender and receiver with misaligned objectives introduces game-theoretic considerations [13] and is beyond the current scope.

II. Bayesian Surprise

Bayesian surprise was first defined and endowed with psychological significance in [9], [10]. Let the prior probability on the received signal be $\{p(Y)\}_{Y \in \mathcal{Y}}$, for signals in the output alphabet, $\mathcal{Y}$. Given this prior probability on output signals, the effect of a transmitted signal, $X = x$, is to change the prior into the signaled posterior, $\{p(Y|X = x)\}_{Y \in \mathcal{Y}}$. The surprise $s(x)$ for a given transmitted signal $x \in \mathcal{X}$ is the relative entropy between the signaled posterior distribution and the prior channel output distribution:

$$s(x) \triangleq D\left(p(Y|X = x) || p(Y)\right). \tag{1}$$

Only channel inputs affecting the receiver's beliefs upon receiving the corresponding received signal are surprising.[1]

In thermodynamic formulations of Bayesian inference [14], an increase in Bayesian surprise is necessarily associated with a decrease in free-energy due to a reduction in prediction error. Bayes-optimal inference schemes, however, do not optimize for Bayesian surprise in itself [15].

A. The Surprise Functional in Channel Capacity Expressions

The Bayesian surprise $s(x) = D\left(p(\cdot|x)||p(\cdot)\right)$ appears in several different guises in prior work in statistics and communication theory. We review a few places where it arises in channel capacity investigations.

Consider an input-constrained memoryless communication channel, with the set of input distribution functions meeting the input constraint denoted by $\mathcal{F}$. Then the Shannon capacity of the channel C is the supremum of the input-output mutual information over the set of input distributions $\mathcal{F}$:

$$C = \sup_{F\in\mathcal{F}} \int\int p(y|x) \log \frac{p(y|x)}{p(y;F)} dy dF(x), \tag{2}$$

where $p(y;F)$ explicitly denotes the dependence of the output distribution on the input distribution F. One can often establish the KKT condition for achieving capacity [16].

Lemma 1: The input distribution F^* achieves Shannon capacity C under input cost function $b[x]$ and input cost constraint β if and only if there exists a Lagrange multiplier $\gamma \geq 0$ such that

$$\begin{aligned} 0 &\leq \gamma(b[x]-\beta) + C - \int p(y|x) \log \frac{p(y|x)}{p(y;F^*)} dy \\ &\leq \gamma(b[x]-\beta) + C - s(x). \end{aligned} \tag{3}$$

Without input constraint, Shannon capacity has a simple output-centered characterization from the KKT condition.

Lemma 2 (p. 142 of [17]): The Shannon capacity C is:

$$C = \min_{p_Y(y)} \max_{x\in\mathcal{X}} s(x). \tag{4}$$

Geometrically, the optimal *output* distribution will be the center of a "sphere" with radius measured by Bayesian surprise.

Indeed as its appearance in KKT conditions implies, the Bayesian surprise $s(x) = D\left(p(\cdot|x)||p(\cdot)\right)$ is related to the weak derivative of mutual information as:

$$\begin{aligned} I'_{F_0}(F) &= \lim_{\theta\downarrow 0} \frac{I\left((1-\theta)F_0 + \theta F\right) - I(F_0)}{\theta} \quad \text{for all } F\in\mathcal{F} \\ &= \int s(x) dF(x) - I(F_0). \end{aligned}$$

B. *Average Surprise* is *Information*

Having reviewed Bayesian surprise in some detail, consider the average of this functional. This would arise if considering formulations of average surprise-constrained capacity or average attention-seeking capacity.

As it turns out, expected surprise $E[s(X)]$ is actually equal to mutual information $I(X;Y)$.

Theorem 1: $E[s(X)] = I(X;Y)$.

Proof: Recall the definition of average Bayesian surprise.

$$\begin{aligned} E[s(X)] &= \int_{\mathcal{X}} p_X(x) D(p_{Y|X}(y|x) \| p_Y(y)) dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_X(x) p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{p_Y(y)} dx dy \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{Y|X}(y|x)}{p_Y(y)} dx dy \\ &= I(X;Y). \end{aligned}$$

■

Since the objective and constraint for the communication problems would coincide, there are no nontrivial tradeoffs.

Remark 1: Thm. 1 is instructive, since human experiments have shown people react not to the average surprise of symbols in a larger signal, but to the maximum surprise of symbols in signals [9]. That is to say, information (average surprise) is not what draws people's attention.

C. Shannon Capacity with Average Surprise Considerations

Rather than surprise-constrained or attention-seeking communication, one might consider the Shannon capacity-cost function. In [18], conditions on the channel input cost function are established that allow a given channel input distribution F to be optimal for a given channel transition probability assignment $p_{Y|X}$. A result of [18], which also appears as Prob. 2 in [17, p. 147], is the following.

Lemma 3: Input distribution F achieves Shannon capacity on the channel $p_{Y|X}$ with cost function $b[x]$ at input cost $\beta = E[b[X]]$, i.e. $I(X;Y) = C(\beta)$ if and only if the input cost function satisfies:

$$b[x] \begin{cases} = cs(x) + b_0, \text{ if } p(x) > 0 \\ \geq cs(x) + b_0, \text{ otherwise,} \end{cases}$$

where $c > 0$ and b_0 are arbitrary constants.

Lem. 3 leads to the following rather surprising result.

Theorem 2: For a given channel $p_{Y|X}$ with expected Bayesian surprise constraint $E[s(x)] \leq S$, *any* input distribution F that meets the surprise constraint achieves the Shannon capacity-cost function $C(\beta = S)$ for the surprise cost function $b[x] = s(x)$ that is induced.

Proof: Directly from Lem. 3 by setting arbitrary constants $c = 1$ and $b_0 = 0$. ■

Notice this result deals with the Shannon capacity-cost function $C(\beta)$ rather than the surprise-constrained or attention-seeking communication problems of central interest herein. Optimizing the information rate among the several admissible input distributions that achieve the appropriate Shannon capacity-cost function would provide results for surprise-constrained or attention-seeking communication.

D. Maximum Surprise

As noted in Rem. 1, maximum symbol surprise is the key driver for drawing automatic attention to a message signal in cognitive systems [3], [9].

[1] When the logarithm is base 2, surprise is measured in *wows*.

Following experimental results in psychology, the Bayesian surprise for a sequence of random variables is defined on the level of individual symbols in the alphabet $x \in \mathcal{X}$ as the sequence-wide, position-by-position average of the Bayesian surprise for that symbol.

Definition 1: The Bayesian surprise functional, $s(\cdot)$, for n-letter inputs and outputs X_1^n and Y_1^n is a function of single input symbols $x \in \mathcal{X}$:

$$s(x) = \frac{1}{n} \sum_{i=1}^{n} D\left(p(Y_i | X_i = x) || p(Y_i)\right). \tag{5}$$

III. Surprise-Constrained Capacity

Let us consider the problem of reliable communication with maximum surprise constrained.

A. Definitions, Coding Theorem, and Properties

Definition 2: Given $0 \leq \epsilon < 1$, a non-negative number R is an ϵ-achievable rate for the channel $p_{Y|X}$ with surprise constraint Σ if for every $\delta > 0$ and every sufficiently large n there exists a block code with maximum error probability $\eta < \epsilon$ of rate exceeding $R - \delta$ for which $s(x) \leq \Sigma$ for each letter $x \in \mathcal{X}$. R is an achievable rate if it is ϵ-achievable for all $0 < \epsilon < 1$. The supremum of achievable rates is called the surprise-constrained capacity of the channel, denoted $\mathcal{C}_O^{(\mathrm{s})}(\Sigma)$.

Definition 3: For each n, the nth surprise-constrained capacity function $\mathcal{C}_n^{(\mathrm{s})}(\Sigma)$ of the channel is defined as

$$\mathcal{C}_n^{(\mathrm{s})}(\Sigma) = \sup_{X_1^n : s(x) \leq \Sigma, \forall x \in \mathcal{X}} I(X_1^n; Y_1^n). \tag{6}$$

The surprise-constrained capacity function of the channel is:

$$\mathcal{C}^{(\mathrm{s})}(\Sigma) = \sup_n \frac{1}{n} \mathcal{C}_n^{(\mathrm{s})}(\Sigma).$$

These definitions are used in the coding theorem.

Theorem 3: $\mathcal{C}_O^{(\mathrm{s})}(\Sigma) = \mathcal{C}^{(\mathrm{s})}(\Sigma)$.

Proof: Surprise-constrained communication can be cast as standard point-to-point communication with a constrained set of possible input distributions $\mathcal{F} = \{X_1^n : s(x) \leq \Sigma, \forall x \in \mathcal{X}\}$. Hence, achievability follows from random coding arguments and the converse follows from Fano's inequality. ■

It is immediate that $\mathcal{C}_n^{(\mathrm{s})}(\Sigma)$ is non-decreasing, since the feasible set in the optimization becomes larger as Σ increases. The function is also concave $\cap$.

Theorem 4: $\mathcal{C}_n^{(\mathrm{s})}(\Sigma)$ is a concave $\cap$ function of Σ.

Proof: Let $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$. The inequality to be proven is that:

$$\mathcal{C}_n^{(\mathrm{s})}(\alpha_1 \Sigma_1 + \alpha_2 \Sigma_2) \geq \alpha_1 \mathcal{C}_n^{(\mathrm{s})}(\Sigma_1) + \alpha_2 \mathcal{C}_n^{(\mathrm{s})}(\Sigma_2). \tag{7}$$

Let X_1 and X_2 be n-dimensional test sources distributed according to p_1 and p_2 that achieve $\mathcal{C}_n^{(\mathrm{s})}(\Sigma_1)$ and $\mathcal{C}_n^{(\mathrm{s})}(\Sigma_2)$ respectively. By definition, $\max_{\mathcal{X}} s(X_i) \leq \Sigma_i$ and $I(X_i; Y_i) = \mathcal{C}_n^{(\mathrm{s})}(\Sigma_i)$ for $i = 1, 2$. Define another source X distributed according to $p = \alpha_1 p_1 + \alpha_2 p_2$. For arbitrary densities p, q_1, q_2, and $0 \leq \lambda \leq 1$, the relative entropy is convex: $D(p \| \lambda q_1 + (1 - \lambda) q_2) \leq \lambda D(p \| q_1) + (1 - \lambda) D(p \| q_2)$. Therefore it follows that as a function of p_Y, for fixed $p_{Y|X=x}$ Bayesian surprise is convex for each x. Further since p_Y is a linear function of p_X, Bayesian surprise is convex in p_X:

$$s(X) \leq \alpha_1 s(X_1) + \alpha_2 s(X_2) \text{ for all } x \in \mathcal{X}. \tag{8}$$

Since the pointwise maximum of convex functions is convex:

$$\max_{\mathcal{X}} s(X) \leq \alpha_1 \max_{\mathcal{X}} s(X_1) + \alpha_2 \max_{\mathcal{X}} s(X_2) \tag{9}$$

$$\leq \alpha_1 \Sigma_1 + \alpha_2 \Sigma_2. \tag{10}$$

Hence X is admissible under the surprise constraint.

Now, by definition of $\mathcal{C}_n^{(\mathrm{s})}$, it is known $I(X;Y) \leq \mathcal{C}_n^{(\mathrm{s})}(\alpha_1 \Sigma_1 + \alpha_2 \Sigma_2)$. However, since $I(X;Y)$ is a concave $\cap$ function of the input probability,

$$I(X;Y) \geq \alpha_1 I(X_1; Y_1) + \alpha_2 I(X_2; Y_2) \tag{11}$$

$$= \alpha_1 \mathcal{C}_n^{(\mathrm{s})}(\Sigma_1) + \alpha_2 \mathcal{C}_n^{(\mathrm{s})}(\Sigma_2). \tag{12}$$

Linking these two inequalities yields (7). ■

A single-letter expression may also be obtained.

Theorem 5:

$$\mathcal{C}^{(\mathrm{s})}(\Sigma) = \sup_{X : s(X) \leq \Sigma} I(X;Y). \tag{13}$$

Proof: Follows from Thm. 4 and channel memorylessness, nearly identically as the proof of [12, Thm. 3]. ■

In solving the information-theoretic optimization problem, one might think there would be a desire to reduce the alphabet to a subset that is allowable under a maximum surprise constraint, just as the signaling alphabet is restricted when considering maximum amplitude constraints [16]. But as noted previously, the Bayesian surprise depends on the signaling scheme itself, so things are not quite so straightforward.

Let us see some particular examples.

B. Binary and Ternary Channels

Consider a memoryless binary channel with alphabet $\mathcal{X} = \{0, 1\}$ and channel transition probability assignment

$$p_{Y|X}(y|x) = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}. \tag{14}$$

The first letter 0 is used with probability π and the second letter 1 is used with probability $1 - \pi$.

The output distribution is then

$$p(y = 0) = \pi(1 - \alpha) + (1 - \pi)\beta \tag{15}$$

$$p(y = 1) = (1 - \beta)(1 - \pi) + \alpha\pi. \tag{16}$$

Bayesian surprise is computed to be:

$$s(0) = (1 - \alpha) \log_2 \left[\frac{1 - \alpha}{\pi(1 - \alpha) + (1 - \pi)\beta}\right] \tag{17}$$

$$+ \alpha \log_2 \left[\frac{\alpha}{(1 - \beta)(1 - \pi) + \alpha\pi}\right]$$

$$s(1) = \beta \log_2 \left[\frac{\beta}{\pi(1 - \alpha) + (1 - \pi)\beta}\right] \tag{18}$$

$$+ (1 - \beta) \log_2 \left[\frac{1 - \beta}{(1 - \beta)(1 - \pi) + \alpha\pi}\right].$$

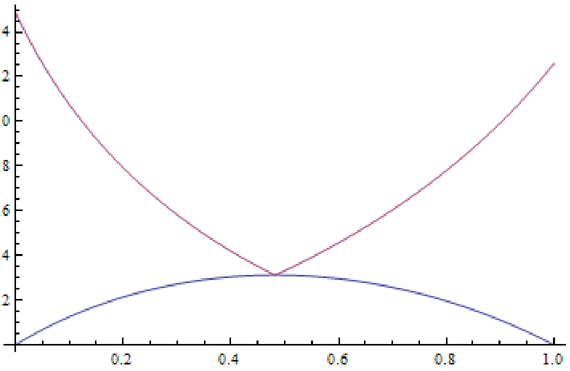

Fig. 1. Information rate (bits) in blue and maximum surprise (wows) in purple as a function of input probability distribution π for a binary channel with $\alpha = 1/4$ and $\beta = 1/8$. The point that maximizes information rate $\pi = 0.482103$ also minimizes the maximum surprise.

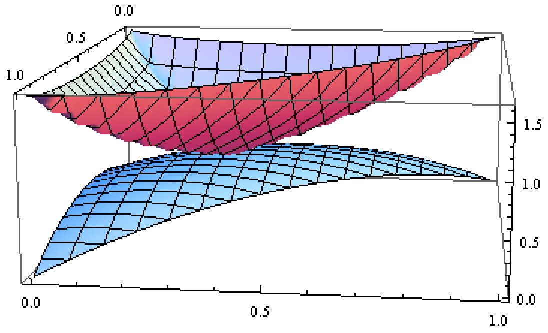

Fig. 2. Information rate (bits) in the bottom and maximum surprise (wows) in the top as a function of input probability distribution (p, q) on the unit simplex for a ternary symmetric channel with $\epsilon = 1/8$. The point that maximizes information rate $(p = 1/3, q = 1/3)$ is also the point that minimizes the maximum surprise.

Hence, the maximum surprise T incurred by a given signaling scheme π is

$$T(\pi) = \max\left(s(0;\pi), s(1;\pi)\right), \tag{19}$$

where dependence of $s(\cdot)$ on π has been explicitly noted.

The rate R achieved using signaling scheme π is just

$$R(\pi) = h_2\left(\pi(1-\alpha) + (1-\pi)\beta\right) - \pi h_2(\alpha) - (1-\pi)h_2(\beta), \tag{20}$$

where $h_2(\cdot)$ is the binary entropy function.

As can be observed by carrying out computations, the following result holds:

$$\arg\max_{\pi} R(\pi) = \arg\min_{\pi} T(\pi). \tag{21}$$

The values achieved (R [bits] & T [wows]) are also the same:

$$R(\pi^*) = T(\pi^*), \tag{22}$$

where π^* is the extremizing value in (21).

Hence there is no tradeoff between maximum information rate and minimax surprise: the same signaling scheme optimizes both the objective and the constraint. This is depicted for an example in Fig. 1 and formalized as follows.

Proposition 1: For binary memoryless channels,

$$\mathcal{C}^{(\mathrm{s})}(\Sigma) = \begin{cases} C(\beta = \infty), & \Sigma \geq C(\beta = \infty) \\ 0, & \text{otherwise}, \end{cases} \tag{23}$$

where $C(\cdot)$ is the Shannon capacity-cost function.

Fig. 2 shows the same phenomenon for ternary channels.

C. A General Result

Computations showed that the same extremal input distribution achieves both the minimax surprise and the unconstrained Shannon capacity for two specific examples. In fact, this is a general phenomenon, following from a mechanical spring interpretation of the KKT conditions for Shannon capacity (Lem. 2), which implies $s(x)$ for all letters $x \in \mathcal{X}$ should be equilibrated at the unconstrained capacity point.

Theorem 6: For any discrete memoryless channel,

$$\mathcal{C}^{(\mathrm{s})}(\Sigma) = \begin{cases} C(\beta = \infty), & \Sigma \geq C(\beta = \infty) \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where $C(\cdot)$ is the Shannon capacity-cost function.

Proof: Directly from information geometry of Lem. 2. ■

So we see there is no nontrivial tradeoff for surprise-constrained capacity. As may be evident from Figs. 1–2, attention-seeking capacity will have a nontrivial tradeoff.

IV. Attention-Seeking Capacity

Now consider trying to attract attention to a message by maximizing surprise and information. Basic definitions and theorems are similar to the surprise-constrained setting, but the attention-seeking capacity function will be rather different. Due to similarity, theorem proofs are omitted for brevity.

A. Definitions, Coding Theorem, and Properties

First, the operational definition.

Definition 4: Given $0 \leq \epsilon < 1$, a non-negative number R is an ϵ-achievable rate for the channel $p_{Y|X}$ with minimum maximum surprise constraint Σ if for every $\delta > 0$ and every sufficiently large n there exists a block code with maximum error probability $\eta < \epsilon$ of rate exceeding $R - \delta$ for which $s(x) \geq \Sigma$ for at least one letter in the transmitted signal. R is an achievable rate if it is ϵ-achievable for all $0 < \epsilon < 1$. The supremum of achievable rates is called the attention-seeking capacity of the channel, denoted $\mathcal{C}_O^{(\mathrm{a})}(\Sigma)$.
Now the informational definition.

Definition 5: For each n, the nth attention-seeking capacity function $\mathcal{C}_n^{(\mathrm{a})}(\Sigma)$ of the channel is defined as

$$\mathcal{C}_n^{(\mathrm{a})}(\Sigma) = \sup_{X_1^n : \max_{\mathcal{X}} s(x) \geq \Sigma} I(X_1^n; Y_1^n), \tag{25}$$

where $I(X_1^n; Y_1^n)$ is the mutual information. The attention-seeking capacity function of the channel is defined as

$$\mathcal{C}^{(\mathrm{a})}(\Sigma) = \sup_n \frac{1}{n} \mathcal{C}_n^{(\mathrm{a})}(\Sigma).$$

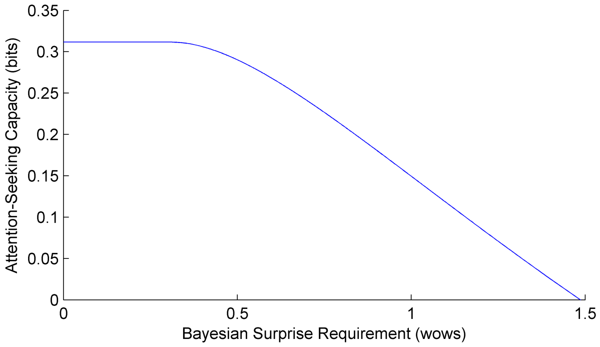

Fig. 3. The attention-seeking capacity function for a binary channel with $\alpha = 1/4$ and $\beta = 1/8$.

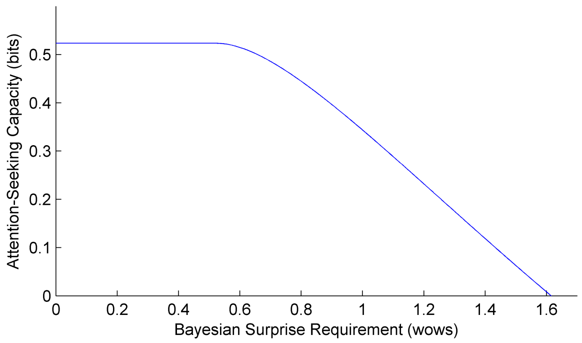

Fig. 4. The attention-seeking capacity function for a ternary symmetric channel with $\epsilon = 1/8$.

Next the coding theorem.

Theorem 7: $\mathcal{C}_O^{(\mathrm{a})}(\Sigma) = \mathcal{C}^{(\mathrm{a})}(\Sigma)$.

Concavity holds here too.

Theorem 8: $\mathcal{C}_n^{(\mathrm{a})}(\Sigma)$ is a non-increasing concave $\cap$ function of Σ, for $0 \le \Sigma \le \Sigma_{\max}$.

Finally, a single-letter expression.

Theorem 9:

$$\mathcal{C}^{(\mathrm{a})}(\Sigma) = \sup_{X: \max_{\mathcal{X}} s(x) \ge \Sigma} I(X;Y). \tag{26}$$

Consider the same example channels as Sec. III.

B. Binary and Ternary Channels

Recall the binary channel (14), and its achievable information rate $R(\pi)$ and achievable Bayesian surprise $T(\pi)$ from (20) and (19), respectively, with signaling strategy π. From Fig. 1, we see there will be a nontrivial tradeoff; Fig. 3 depicts this. As seen, the attention-seeking capacity function has an initial flat portion, which achieves the unconstrained capacity, and then decreases as the constraint tightens.

In particular, the attention-seeking capacity function is:

$$\mathcal{C}^{(\mathrm{a})}(\Sigma) = \begin{cases} C(\beta = \infty), & 0 \le \Sigma \le R(\pi^*) \\ R(\pi), 0 \le \pi \le \pi^*, & \Sigma > R(\pi^*) \text{ and } \pi^* \le 1/2 \\ R(\pi), \pi^* \le \pi \le 1, & \Sigma > R(\pi^*) \text{ and } \pi^* > 1/2. \end{cases}$$

where $R(\cdot)$ is the expression (20) and π^* is the extremizing value in (21).

The attention-seeking capacity function for the ternary symmetric channel in Fig. 2 is shown in Fig. 4.

V. Bits and Wows

When designing communication strategies for the information overload regime, it is not enough to just transmit messages reliably. The receiver's attention must also be engaged. Transmitted messages must be both informative and surprising.

Any signaling scheme over any channel achieves Shannon capacity-cost when Bayesian surprise is the cost, so surprise is the natural cost function for communication. Hence there are no nontrivial tradeoffs in the surprise-limited setting, but under the important setting of attention-seeking, increasing the surprise requirement leads to a reduction in information rate. Network settings such as broadcast may have more intricate interplay between attention and information.

References

[1] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, "A taxonomy of external and internal attention," *Annu. Rev. Psychol.*, vol. 62, pp. 73–101, Jan. 2011.

[2] M. Mather and M. R. Sutherland, "Arousal-biased competition in perception and memory," *Perspect. Psychol. Sci.*, vol. 6, pp. 114–133, Mar. 2011.

[3] D. Kahneman, *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[4] J. B. Spira, *Overload!: How Too Much Information is Hazardous to your Organization*. Hoboken, NJ: John Wiley & Sons, 2011.

[5] T. H. Davenport and J. C. Beck, *The Attention Economy*. Boston: Harvard Business School Press, 2001.

[6] J. P. L. Schoormans and H. S. J. Robben, "The effect of new package design on product attention, categorization and evaluation," *J. Econ. Psychol.*, vol. 18, pp. 271–287, Apr. 1997.

[7] J. Wainer, L. Dabbish, and R. Kraut, "Should I open this email?: Inbox-level cues, curiosity and attention to email," in *Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst.*, May 2011, pp. 3439–3448.

[8] M. I. Garrido, R. J. Dolan, and M. Sahani, "Surprise leads to noisier perceptual decisions," *i-Perception*, vol. 2, pp. 112–120, 2011.

[9] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 547–554.

[10] P. Baldi and L. Itti, "Of bits and wows: A Bayesian theory of surprise with applications to attention," *Neural Netw.*, vol. 23, pp. 649–666, Jun. 2010.

[11] L. R. Varshney, F. Pinel, K. R. Varshney, A. Schörgendorfer, and Y.-M. Chee, "Cognition as a part of computational creativity," in *Proc. 12th IEEE Int. Conf. Cogn. Inform. Cogn. Comput.*, Jul. 2013.

[12] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. 2008 IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1612–1616.

[13] T. Van Zandt, "Information overload in a network of targeted communication," *Rand J. Econ.*, vol. 35, pp. 542–560, Autumn 2004.

[14] K. Friston, "The free-energy principle: a rough guide to the brain?" *Trends Cogn. Sci.*, vol. 13, pp. 293–301, Jul. 2009.

[15] H. Feldman and K. J. Friston, "Attention, uncertainty, and free-energy," *Front. Human Neurosci.*, vol. 4, 215, 2010.

[16] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inf. Control*, vol. 18, pp. 203–219, Apr. 1971.

[17] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Budapest: Akadémiai Kiadó, 1997.

[18] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1147–1158, May 2003.