Stability and Sensitivity of the Capacity in Continuous Channels

Malcolm Egan

Univ. Lyon, INSA Lyon, INRIA

2019 European School of Information Theory

April 18, 2019

Capacity of Additive Noise Models

Consider the (memoryless, stationary, scalar) additive noise channel

$$Y = X + N,$$

where the noise N is a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with probability density function p_N.

The capacity is defined by

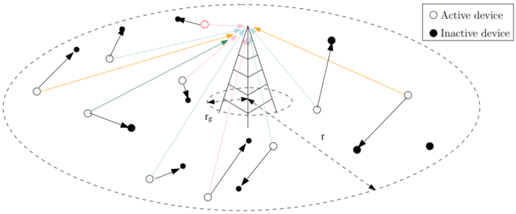
$$\begin{aligned} C = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\ \text{subject to} \quad & \mu_X \in \Lambda \end{aligned}$$

Key Question: What is the capacity for general constraints and non-Gaussian noise distributions?

Non-Gaussian Noise Models

In many applications, the noise is **non-Gaussian**.

Example 1: Poisson Spatial Fields of Interferers.



Noise in this model is the interference

$$Z = \sum_{i \in \Phi} r_i^{-\eta/2} h_i X_i.$$

Non-Gaussian Noise Models

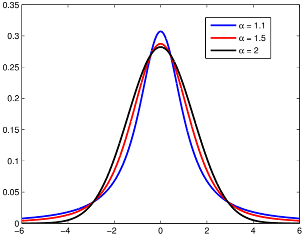
Suppose that

(i) Φ is a homogeneous Poisson point process;

(ii) (h_i) and (X_i) are processes with independent elements;

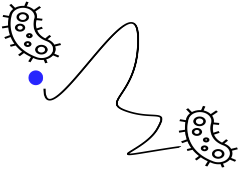
(ii) $\mathbb{E}[|h_i X_i|^{4/\eta}] < \infty$.

Then, the interference Z converges almost surely to a **symmetric α-stable random variable**.



Non-Gaussian Noise Models

Example 2: Molecular Timing Channel.



In the channel

$$Y = X + N,$$

the input X corresponds to time of release.

Non-Gaussian Noise Models

In the channel

$$Y = X + N,$$

the noise N corresponds to the diffusion time from the transmitter to the receiver.

Under Brownian motion models of diffusion, the noise distribution is **inverse Gaussian** or **Lévy stable**.

$$p_N(x) = \sqrt{\frac{\lambda}{2\pi x^2}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right).$$

Capacity of Non-Gaussian Noise Models

The capacity is defined by

$$\begin{aligned} C = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\ \text{subject to} \quad & \mu_X \in \Lambda \end{aligned}$$

The noise is in general non-Gaussian.

Question: What is the constraint set Λ?

Constraint Sets

A familiar constraint common in wireless communications is

$$\Lambda_P = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[X^2] \leq P\}$$

corresponding to an **average power constraint**.

Other constraints appear in applications. For example,

$$\Lambda_c = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[|X|^r] \leq c\}$$

where $0 < r < 2$. This corresponds to a **fractional moment constraint** (useful in the study of α-stable noise channels).

In the molecular timing channel,

$$\Lambda_T = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[X] \leq T,\ \mathbb{P}_{\mu_X}(X < 0) = 0\}$$

is the relevant constraint.

Capacity of Non-Gaussian Noise Channels

The capacity is defined by

$$\begin{aligned} C = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\ \text{subject to} \quad & \mu_X \in \Lambda \end{aligned}$$

Since the channel is additive,

$$I(\mu_X, P_{Y|X}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_N(y-x) \log \frac{p_N(y-x)}{p_Y(y)} dy d\mu_X(x).$$

There are two basic questions that can be asked:

(i) What is the value of the capacity C?

(ii) What is the optimal solution μ_X^*?

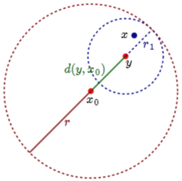
Topologies on Sets of Probability Measures

Point set topology plays an important role in optimization theory.

For example, it allows us to determine whether or not the optimum can be achieved (i.e., the sup becomes a max).

Topologies on Sets of Probability Measures

In applications, we usually optimize over $\mathbb{R}^n$, which has the standard topology induced by Euclidean metric balls.



In the capacity problem, we optimize over sets of probability measures in subsets of $\mathcal{P}$.

Question: What is a useful topology on the set of probability measures?

Topologies on Sets of Probability Measures

A useful choice is the **topology of weak convergence**.

Closed sets $\mathcal{S}$ are defined by sequences of probability measures $(\mu_i) \subset \mathcal{S}$ and a limiting probability measure $\mu \in \mathcal{S}$ such that

$$\lim_{i \to \infty} \int_{-\infty}^{\infty} f(x) d\mu_i(x) = \int_{-\infty}^{\infty} f(x) d\mu(x).$$

for all bounded and continuous functions f.

It turns out that the topology of weak convergence for probability measures is **metrizable**.

There exists a metric d on $\mathcal{P}$ such that d metric-balls generate the topology of weak convergence (known as the Lévy-Prokhorov metric).

Topologies on Sets of Probability Measures

In addition, Prokhorov's theorem gives a characterization of compactness.

Prokhorov's Theorem: If a subset $\Lambda \subset \mathcal{P}$ of probability measures is tight and closed, then Λ is compact in the topology of weak convergence.

A set of probability measures Λ is **tight** if for all $\epsilon > 0$, there exists a compact set $\mathcal{K}_\epsilon \subset \mathbb{R}$ such that

$$\mu(\mathcal{K}_\epsilon) \geq 1 - \epsilon, \ \forall \mu \in \Lambda.$$

Existence of the Optimal Input

The capacity is defined by

$$
\begin{aligned}
C = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\
\text{subject to} \quad & \mu_X \in \Lambda
\end{aligned}
$$

Question: Does the capacity-achieving input exist?

This is answered by the extreme value theorem.

Extreme Value Theorem: If Λ is weakly compact and $I(\mu_X, P_{Y|X})$ is weakly continuous on Λ, then μ_X^* exists.

Support of the Optimal Input

Question: When is the optimal input discrete and compactly supported?

The initial results on this question were due to Smith [Smith1971].

Theorem: For amplitude and average power constraints, the optimal input for the Gaussian noise channel is discrete and compactly supported.

Support of the Optimal Input

More generally, the support of the optimal input can be studied via the KKT conditions.

Let $\mathcal{M}$ be a convex and compact set of channel input distributions. Then, $\mu_X^* \in \mathcal{M}$ maximizes the capacity if and only if for all $\mu_X \in \mathcal{M}$

$$\mathbb{E}_{\mu_X}\left[\log\left(\frac{dP_{Y|X}(Y|X)}{dP_Y(Y)}\right)\right] \leq I(\mu_X^*, P_{Y|X}).$$

Equality holds at points of increase $\Rightarrow$ constraints on optimal inputs.

Significant progress recently; e.g., [Fahs2018,Dytso2019].

Characterizing the Capacity

In general, it is hard to compute the capacity in closed-form. Exceptions are Gaussian and Cauchy noise channels under various constraints.

Theorem [Lapidoth and Moser]: Let the input alphabet $\mathcal{X}$ and the output alphabet $\mathcal{Y}$ of a channel $W(\cdot|\cdot)$ be seperable metric spaces, and assume that for any Borel subset $\mathcal{B} \subset \mathcal{Y}$ the mapping $x \mapsto W(\mathcal{B}|x)$ from $\mathcal{X}$ to $[0,1]$ is Borel measurable. Let $Q(\cdot)$ be any probability measure on $\mathcal{X}$, and $R(\cdot)$ any probability measure on $\mathcal{Y}$. Then, the mutual information $I(Q;W)$ can be bounded by

$$I(Q;W) \leq \int D(W(\cdot|x)||R(\cdot))dQ(x)$$

A Change in Perspective

New Perspective: the capacity is a map

$$(p_N, \Lambda) \mapsto C.$$

Definition

Let $\mathcal{K} = (p_N, \Lambda)$ and $\hat{\mathcal{K}} = (\hat{p}_N, \hat{\Lambda})$ be two tuples of channel parameters. The *capacity sensitivity* due to a perturbation from channel $\mathcal{K}$ to the channel $\hat{\mathcal{K}}$ is defined as

$$C_{\mathcal{K} \to \hat{\mathcal{K}}} \triangleq |C(\mathcal{K}) - C(\hat{\mathcal{K}})|.$$

Egan, M., Perlaza, S.M. and Kungurtsev, V., "Capacity sensitivity in additive non-Gaussian noise channels," *Proc. IEEE International Symposium on Information Theory*, Aachen, Germany, Jun. 2017.

A Strategy

- Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, which admits a Taylor series representation

$$f(\mathbf{x} + \|\mathbf{e}\|\tilde{\mathbf{e}}) = f(\mathbf{x}) + \|\mathbf{e}\| D_{\tilde{\mathbf{e}}} f(\mathbf{x})^T \tilde{\mathbf{e}} + o(\|\mathbf{e}\|).$$

  ($\tilde{\mathbf{e}}$ is unit norm).

- This yields

$$|f(\mathbf{x} + \|\mathbf{e}\|\tilde{\mathbf{e}}) - f(\mathbf{x})| \leq \|D_{\tilde{\mathbf{e}}} f(\mathbf{x})\| \|\mathbf{e}\| + o(\|\mathbf{e}\|),$$

  i.e., the *sensitivity*.

Question: what is the directional derivative of the optimal value function of an optimization problem (e.g., the capacity)?

A Strategy

- In the case of vector, smooth optimization problems there is a good theory.
- E.g., envelope theorems.

Proposition

Let the real valued function $f(\mathbf{x}, y) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *be twice differentiable on a compact convex subset* $\mathcal{X}$ *of* $\mathbb{R}^{n+1}$*, strictly concave in* $\mathbf{x}$*. Let* $\mathbf{x}^*$ *be the optimal value of* f *on* $\mathcal{X}$ *and denote* $\psi(y) = f(\mathbf{x}^*, y)$*. Then, the derivative of* $\psi(y)$ *exists and is given by*

$$\psi'(y) = f_y(\mathbf{x}^*, y).$$

A Strategy

A sketch of the proof:

1. Use the implicit function theorem to write $\psi(y) = f(\mathbf{x}^*(y), y)$.

2. Observe that

$$\begin{aligned}\psi'(y) &= f_y(\mathbf{x}^*(y), y) + (\nabla_{\mathbf{x}} f(\mathbf{x}^*(y), y))^T \frac{d\mathbf{x}^*(y)}{dy} \\ &= f_y(\mathbf{x}^*(y), y).\end{aligned}$$

Generalizations of this result due to Danskin and Gol'shtein.

A Strategy

Recall:

$$C(\Lambda, p_N) = \sup_{\mu_X \in \Lambda} I(\mu_X, p_N)$$

Question: What is the effect of

- Constraint perturbations: $C(\Lambda)$ (fix p_N)?
- Noise distribution perturbations: $C(p_N)$ (fix Λ)?

Constraint Perturbations

Common Question: What is the effect of power on the capacity?

Another Formulation: What is the effect of changing the set of probability measures

$$\Lambda^2 = \{\mu_X : \mathbb{E}_{\mu_X}[X^2] \leq P\}.$$

Natural Generalization: What is the effect of changing Λ on

$$\begin{aligned} C(\Lambda) = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\ \text{subject to} \quad & \mu_X \in \Lambda. \end{aligned}$$

Constraint Perturbations

Question: Do small changes in the constraint set lead to small changes in the capacity?

To answer this question, we need to formalize what a small change means.

Key Idea: The constraint set is viewed as a point-to-set map.

Example: Consider the power constraint

$$\Lambda^2(P) = \{\mu_X : \mathbb{E}_{\mu_X}[X^2] \leq P\}$$

is a map from $\mathbb{R}$ to a compact set of probability measures $\Lambda^2 : \mathbb{R} \rightrightarrows \mathcal{P}$

Constraint Perturbations

When the power P (or more generally, any other parameter) is changed, $\Lambda^2(P)$ can expand or contract.

There are therefore two aspects to continuity of a point-to-set map.

Definition

A point-to-set map $\Lambda : \mathbb{R} \rightrightarrows \mathcal{P}$ is upper hemicontinuous at $P \in \mathbb{R}$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $d(\overline{P}, P) < \delta$ implies that $\Lambda(\overline{P}) \subseteq \eta_\epsilon(\Lambda(P))$.

Definition

A point-to-set map $\Lambda : \mathbb{R} \rightrightarrows \mathcal{P}$ is lower hemicontinuous at $P \in \mathbb{R}$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $d(\overline{P}, P) < \delta$ implies that $\Lambda(P) \subseteq \eta_\epsilon(\Lambda(\overline{P}))$.

Λ is **continuous** if it is both upper and lower hemicontinous.

Lemma 1: Berge's Maximum Theorem

Theorem (Berge's Maximum Theorem)

Let Θ and S be two metric spaces, $\Gamma : \Theta \rightrightarrows S$ a compact-valued point-to-set map, and $\varphi : S \times \Theta \to \mathbb{R}$ be a continuous function on $S \times \Theta$. Define

$$\sigma(\theta) = \arg\max\{\varphi(s,\theta) : s \in \Gamma(\theta)\}, \ \forall \theta \in \Theta$$
$$\varphi^*(\theta) = \max\{\varphi(s,\theta) : s \in \Gamma(\theta)\}, \ \forall \theta \in \Theta$$

and assume that Γ **is continuous** *at $\theta \in \Theta$. Then, $\varphi^* : \Theta \to \mathbb{R}$ is continuous at θ.*

Implication: continuity of the capacity in P if

1. $I(\mu_X, P_{Y|X})$ is weakly continuous on Λ
2. $\Lambda : \mathbb{R} \rightrightarrows \mathcal{P}$ is continuous.

Bounding the Capacity Sensitivity

We now have general conditions to ensure that the capacity sensitivity

$$|C(\Lambda(P)) - C(\Lambda(P'))| \to 0, \;\; P \to P'.$$

However, the capacity is in general a complicated function of the constraint parameters.

Question: Is there a general way of bounding the capacity sensitivity?

Bounding the Capacity Sensitivity

Key tool: Regular subgradients.

Definition

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $\bar{\mathbf{x}} \in \mathbb{R}^n$ with $f(\bar{\mathbf{x}})$ finite. For a vector, $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}$ is a regular subgradient of f at $\bar{\mathbf{x}}$, denoted by $\mathbf{v} \in \hat{\partial} f(\bar{\mathbf{x}})$, if there exists $\delta > 0$ such that for all $\mathbf{x} \in B_\delta(\bar{\mathbf{x}})$

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \mathbf{v}^T(\mathbf{x} - \bar{\mathbf{x}}) + o(|\mathbf{x} - \bar{\mathbf{x}}|).$$

Related to subgradients in convex optimization.

What are conditions for existence?

Lemma 2: Existence of Regular Subgradients

Theorem (Rockafellar and Wets 1997)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ *is finite and lower semicontinuous at* $\bar{\mathbf{x}} \in \mathbb{R}^n$*. Then, there exists a sequence* $\mathbf{x}^k \underset{f}{\to} \bar{\mathbf{x}}$ *with* $\hat{\partial} f(\mathbf{x}^k) \neq \emptyset$ *for all* k*.*

Rockafellar, R. and Wets, R., *Variational Analysis*. Berlin Heidelbeg: Springer-Verlag, 1997

Implication:

1. Let $f(P) = C(\Lambda(P))$
2. Apply Berge's maximum theorem and regular subgradients.

This yields general estimates of the capacity sensitivity.

Example 1: RHS Constraint Perturbations

- Consider constraints

$$\Lambda(b) = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[f(|X|)] \leq b\}$$

where f is positive, non-decreasing and lower semicontinuous.

- The capacity is given by

$$\begin{aligned} \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\ \text{subject to} \quad & \mu_X \in \Lambda(b). \end{aligned}$$

- Need to establish continuity in b.

Example 1: RHS Constraint Perturbations

Theorem

Let $\overline{b} \in \mathbb{R}_+$ and suppose that the following conditions hold:

(i) $\Lambda(\overline{b})$ *is non-empty and compact.*

(ii) $I(\mu_X, P_{Y|X})$ *is weakly continuous on* $\Lambda(\overline{b})$*.*

Then, $C(b)$ *is continuous at* $\overline{b}$*.*

It is now possible to apply the Rockafellar-Wets regular subgradient existence theorem.

Suppose $\overline{b} > \tilde{b}$ with $C(\overline{b}) < \infty$ and $C(\tilde{b}) < \infty$. If $\overline{b} - \tilde{b}$ and $\epsilon > 0$ are sufficiently small

$$C(\overline{b}) - C(\tilde{b}) - \epsilon \leq |v||\overline{b} - \tilde{b}| + o(|\overline{b} - \tilde{b}|)$$

Example 2: Discrete Input Constraints

Consider the general constraint set Λ (allowing for continuous inputs).

$$
\begin{aligned}
C(\Lambda) = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\
\text{subject to} \quad & \mu_X \in \Lambda,
\end{aligned}
$$

E.g., $\Lambda = \Lambda^p = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[|X|^p] \leq b\}$.

Let $\mathcal{P}_\Delta$ be the set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that have mass points in the set $\cup_{\Delta' > \Delta} \Delta' \mathbb{Z}$. Let Λ be a compact subset of $\mathcal{P}$.

The **discrete approximation** of $C(\Lambda)$ is then defined as

$$
\begin{aligned}
C(\Lambda_\Delta) = \sup_{\mu_X \in \mathcal{P}} \quad & I(\mu_X, P_{Y|X}) \\
\text{subject to} \quad & \mu_X \in \Lambda_\Delta,
\end{aligned}
$$

where $\Lambda_\Delta = \mathcal{P}_\Delta \cap \Lambda$.

Example 2: Discrete Input Constraints

The **capacity sensitivity** in this case is:

$$C_{\Lambda \to \Lambda_\Delta} = |C(\Lambda) - C(\Lambda_\Delta)|,$$

I.e., the *cost* of discreteness.

Again, we need to establish continuity in order to apply the Rockafellar-Wets theorem.

Example 2: Discrete Input Constraints

Theorem (Egan, Perlaza 2018)

Let Λ be a non-empty compact subset of $\mathcal{P}$. If the mutual information $I(\cdot, P_{Y|X})$ is weakly continuous on Λ, then $C(\Lambda_\Delta) \to C(\Lambda)$ as $\Delta \to 0$.

(i) *Gaussian model*
- $p_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-x^2/(2\sigma^2)\right), \ \sigma > 0.$
- $\Lambda = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[X^2] \leq b\}, \ b > 0.$

(ii) *Cauchy model*
- $p_N(x) = \frac{1}{\pi\gamma\left(1+\left(\frac{x}{\gamma}\right)^2\right)}, \ \gamma > 0.$
- $\Lambda = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[|X|^r] \leq b\}, \ b > 0.$

(iii) *Inverse Gaussian model*
- $p_N(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\gamma)^2}{2\gamma^2 x}\right), \ x > 0, \ \lambda, \gamma > 0.$
- $\Lambda = \{\mu_X \in \mathcal{P} : \mathbb{E}_{\mu_X}[X] \leq b\}, \ b > 0.$

Example 2: Discrete Input Constraints

Theorem (Egan, Perlaza 2018)

Suppose that Λ *is a non-empty compact subset of* $\mathcal{P}$ *and the mutual information* $I : \mathcal{P} \to \mathbb{R}$ *is weakly continuous on* Λ *. If* $C = \sup_{\mu_X \in \Lambda} I(\mu_X, P_{Y|X}) < \infty$*, then for all* $\epsilon > 0$ *there exists* $v \in \mathbb{R}$ *such that for* Δ *sufficiently small,*

$$C(\Lambda) - C(\Lambda_\Delta) - \epsilon \leq |v|\Delta + o(\Delta)$$

holds.

Recipe Review

If the model parameter is **finite dimensional**:

1. Establish continuity via Berge's maximum theorem.
2. Apply regular subgradient existence theorem.

Remark: The method applies to more general channels; e.g., vector channels

Egan, M., "On Capacity Sensitivity in Additive Vector Symmetric -Stable Noise Channels", *Proc. IEEE WCNC (Invited Paper MoTION Workshop)*, 2019.

What if the model parameter is **not** finite dimensional?

E.g., the noise distribution?

Noise Distribution Perturbations

In the case of noise pdf perturbations, the relevant capacity sensitivity is

$$C_{p_N^0 \to p_N^1} = |C(p_N^0) - C(p_N^1)|.$$

Let $(p_N^i)_i$ be a sequence of pdfs converging to p_N^0 (in e.g., TV, weakly, KL divergence...).

Question: Does

$$C(p_N^i) \to C(p_N^0) \text{ as } i \to \infty$$

hold?

Noise Distribution Perturbations

Theorem (Egan, Perlaza 2017)

Let $\{p_N^i\}_{i=1}^{\infty}$ *be a pointwise convergent sequence with limit* p_N^0 *and let* Λ *be a compact set of probability measures not dependent on* p_N*. Suppose the following conditions hold:*

(i) *The mutual information* $I(\mu_X, p_N^i)$ *is weakly continuous on* Λ*.*

(ii) *For the convergent sequence* $\{p_N^i\}_{i=1}^{\infty}$ *and all weakly convergent sequences* $\{\mu_i\}_{i=1}^{\infty}$ *in* Λ*,*

$$\lim_{i \to \infty} I(\mu_i, p_N^i) = I(\mu_0, p_N^0).$$

(iii) *There exists an optimal input probability measure* μ_i^* *for each noise probability density* p_N^i*.*

Then, $\lim_{i \to \infty} C(p_N^i) = C(p_N^0)$*.*

Lemma 3: Mutual Information Bound

Lemma (Egan, Perlaza, Kungurtsev 2017)

Let p_N^0, p_N^1 be two noise probability density functions and Λ be a compact subset of $\mathcal{P}$ such that $C(p_N^0) < \infty$ and $C(p_N^1) < \infty$. Then, the capacity sensitivity satisfies

$$
\begin{aligned}
&|C(p_N^0) - C(p_N^1)| \\
&\leq \max\{|I(\mu_0^*, p_N^0) - I(\mu_0^*, p_N^1)|, |I(\mu_1^*, p_N^0) - I(\mu_1^*, p_N^1)|\}.
\end{aligned}
$$

Observation: To compute the estimate, we need to characterize the optimal input distribution.

I.e. is the support discrete, continuous, compact?

$\Rightarrow$

Connects to questions about the optimal input structure.

Conclusions

Key Question: How sensitive are information measures to model assumptions?

Many noise models and constraints are highly idealized.

The capacity sensitivity framework provides a means of investigating what happens when idealizations are relaxed.