THE UNIVERSITY OF
MELBOURNE

# Elements of a Nonstochastic Information Theory

Girish Nair
Dept. Electrical & Electronic Engineering
University of Melbourne

LCCC Workshop on Information and Control in Networks
Lund, Sweden
17 October 2012

# Random Variables in Communications

In communications, unknown quantities/signals are usually modelled as *random variables (rv's)* & *random processes*, for good reasons:

- Physical laws governing electronic/photonic circuit noise give rise to well-defined distributions & random models – e.g. Gaussian thermal electronic noise, binary symmetric channels, Rayleigh fading, etc.

- Telecomm. systems usually designed to be used many times, & each individual phone call/email/download may not be critically important...

  ➔ System designer need only seek good performance in an average or *expected* sense - e.g. bit error rate, signal-to-noise ratio, outage probability.

# Nonrandom Variables in Control

In contrast, unknowns in control are often treated as *non*stochastic variables or signals

- Dominant disturbances are not necessarily electronic/photonic circuit noise, & may not follow well-defined probability distributions.
- Safety- & mission-criticality

  ➔ Performance guarantees needed *every time* plant is used, not just on average.

# Networked Control

Networked control: combines both communications and control theories!

➔How may *nonstochastic* analogues of key probabilistic concepts like independence, Markovness and information be usefully defined?

# Another Motivation: Channel Capacity

The *ordinary capacity* $C$ of a channel is defined as

the highest block-code bit-rate that permits an

arbitrarily small probability of decoding error.

I.e. $C := \lim_{\varepsilon \to 0} \sup_{t \geq 0} \sup \frac{\log_2 |\boldsymbol{F}_t|}{t+1} \overset{\text{(subadditivity)}}{=} \lim_{\varepsilon \to 0} \lim_{t \to \infty} \sup \frac{\log_2 \left|\boldsymbol{F}_t\right|}{t+1},$

where $\boldsymbol{F}_t$:= a finite set of input words of length $t+1$,

& the inner supremums are over all $\boldsymbol{F}_t$ s.t. $\forall x(0:t) \in \boldsymbol{F}_t$,

the corresponding random channel output word $Y(0:t)$

can be mapped to an estimate $\hat{X}(0:t)$ with $\Pr\left[\hat{X}(0:t) \neq x(0:t)\right] \leq \varepsilon$.

# Information Capacity

Shannon's *Channel Coding Theorem* essentially gives an information-theoretic characterization of $C$ for *stationary memoryless stochastic channels*:

$$C = \sup_{t\geq 0} \sup \frac{\mathrm{I}\big[X(0:t);Y(0:t)\big]}{t+1} = \lim_{t\to\infty} \sup \frac{\mathrm{I}\big[X(0:t);Y(0:t)\big]}{t+1}$$
$$\big(= \sup \mathrm{I}[X(0);Y(0)]\big),$$

where I[·;·]:=Shannon's ***mutual information*** functional, and the inner supremums are over all random input sequences $X(0:t)$.

# Zero-Error Capacity

In 1956, Shannon also introduced the stricter notion of *zero-error capacity* $C_0$, the highest block-coded bit-rate that permits a probability of decoding error = 0 exactly.

I.e. $$C_0 := \sup_{t\geq 0} \sup \frac{\log_2 |\boldsymbol{F}_t|}{t+1} = \lim_{t\to\infty} \sup \frac{\log_2 \left|\boldsymbol{F}_t\right|}{t+1},$$

where $\boldsymbol{F}_t$ = a finite set of input words of length $t+1$,
& the inner supremums are over all $\boldsymbol{F}_t$ s.t. $\forall x(0:t) \in \boldsymbol{F}_t$,
the corresponding channel output word $Y(0:t)$
can be mapped to an estimate $\hat{X}(0:t)$ with $\Pr\left[\hat{X}(0:t) \neq x(0:t)\right] = 0.$

Clearly, $C_0$ is (usually strictly) smaller than $C$.

# *C0* as an "Information" Capacity?

***Fact:*** *C0* does not depend on the nonzero transition probabilities of the channel,

and can be defined without any probability theory, in terms of the input-output graph that describes permitted channel transitions.

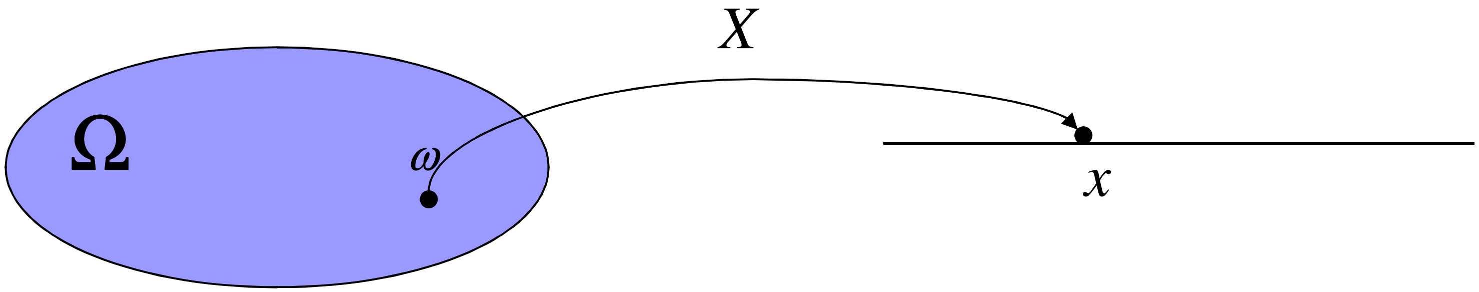
➔ ***Q:*** Can we express *C0* as the maximum rate of some *nonstochastic* information functional?

# Outline

- (Motivation)
- Uncertain Variables
- Taxicab Partitions & Maximin Information
- *C0* via Maximin Information
- Uniform LTI State Estimation over Erroneous Channels
- Conclusion
- Extension & Future Work

# The Uncertain Variable Framework

- Similar to probability theory, let an *uncertain variable (uv)* be a mapping *X* from some sample space **Ω** to a space **X**.
- E.g., each *ω* ϵ **Ω** may represent a particular combination of disturbances & inputs entering a system, & *X* may represent an output/state variable
- For any particular *ω*, the value *x=X(ω)* is *realised.*



Unlike prob. theory, assume *no* σ-algebra or measure on Ω.

# Ranges

As in prob. theory, the $\omega$-argument will often be omitted.

*Marginal range* $[\![X]\!] := \{X(\omega) : \omega \in \mathbf{\Omega}\} \subseteq \mathbf{X}.$

*Joint range* $[\![X, Y]\!] := \{(X(\omega), Y(\omega)) : \omega \in \mathbf{\Omega}\} \subseteq \mathbf{X} \times \mathbf{Y}.$

*Conditional range* $[\![X \mid y]\!] := \{X(\omega) : Y(\omega) = y, \omega \in \mathbf{\Omega}\} \subseteq \mathbf{X}.$

In the absence of statistical structure, the joint range completely characterises the relationship between uv's $X \,\&\, Y$.

As

$$[\![X, Y]\!] = \bigcup_{y \in [\![Y]\!]} [\![X \mid y]\!] \times \{y\},$$

the joint range can be determined from the conditional & marginal ranges, similar to the relationship between joint, conditional & marginal probability distributions.

# Unrelatedness

$X,Y$ called *unrelated* if

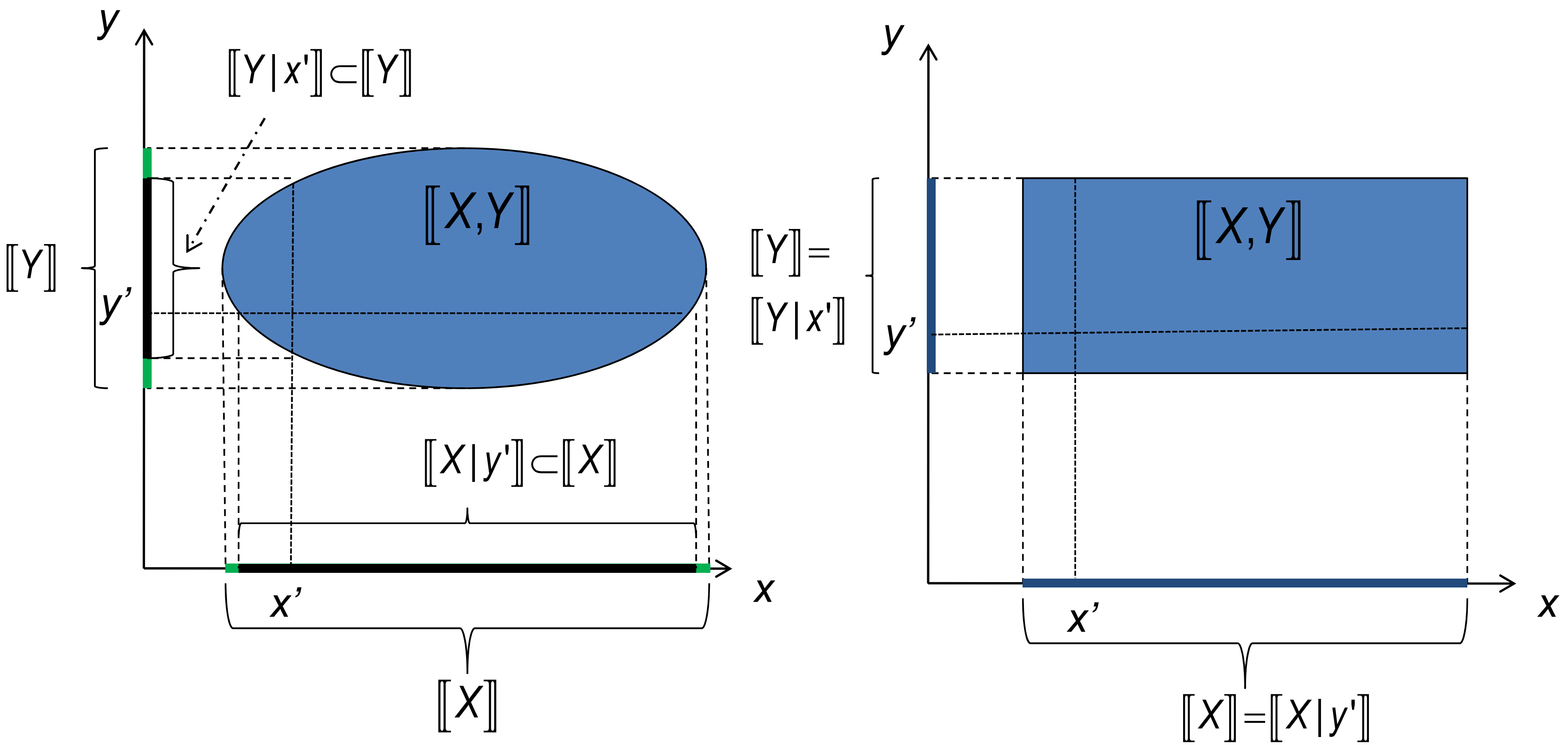
$$[\![X,Y]\!]=[\![X]\!]\times[\![Y]\!],$$

or equivalently if

$$[\![X\mid y]\!]=[\![X]\!],\quad \forall y\in[\![Y]\!].$$

Parallels the definition of mutual independence for rv's.

Called *related* if $[\![X,Y]\!]\subset[\![X]\!]\times[\![Y]\!]$, without equality.


y
$[\![Y|x']\!] \subset [\![Y]\!]$
$[\![X,Y]\!]$
$[\![Y]\!]$
y'
$[\![X|y']\!] \subset [\![X]\!]$
x'
x
$[\![X]\!]$
y
$[\![X,Y]\!]$
$[\![Y]\!] =$
$[\![Y|x']\!]$
y'
x'
x
$[\![X]\!] = [\![X|y']\!]$

a) *X, Y* related

b) *X, Y* unrelated

# Nonstochastic Entropy

The *a priori* uncertainty associated with a uv $X$ is captured by

*Hartley entropy* $\mathrm{H}_0[X] := \log_2 \left| [\![ X ]\!] \right| \in [0,\infty].$

Continuous-valued uv's yield $\mathrm{H}_0[X] = \infty.$

$\Rightarrow$ For uv's with Lebesgue-measurable range in $\mathbb{R}^n$,

the 0-*th order Re′nyi differential entropy*

$$\mathrm{h}_0[X] := \log_2 \mu [\![ X ]\!] \in [-\infty,\infty]$$

is more useful.

# Nonstochastic Information – Previous Definitions

H. Shingin & Y. Ohta, NecSys09:

$$\mathrm{I}_0[X;Y] := \begin{cases} \inf\limits_{y \in [\![Y]\!]} \log_2\left(\dfrac{\left|[\![X]\!]\right|}{\left|[\![X|y]\!]\right|}\right), & X \text{ discrete-valued} \\ \inf\limits_{y \in [\![Y]\!]} \log_2\left(\dfrac{\mu[\![X]\!]}{\mu[\![X|y]\!]}\right), & X \text{ continuous-valued} \end{cases}.$$

(expressed in the uv framework here)

G. Klir, 2006:

$$\mathrm{T}[X;Y] := \begin{cases} \mathrm{H}_0[\![X]\!] + \mathrm{H}_0[\![Y]\!] - \mathrm{H}_0[\![X,Y]\!], & X,Y \text{ finite-valued} \\ \text{Something complex}, & (X,Y) \text{ cont.-valued w. convex range } \subset \mathbb{R}^n \end{cases}.$$

# Comments on Previous Definitions

- Each gives different treatments of continuous & discrete-valued variables.
- Klir’s information has natural properties, but is purely axiomatic. No demonstrated relevance to problems in communications or control.
- Shingin & Ohta’s information: inherently asymmetric, but shown to be useful for studying control over errorless digital channels.

# Taxicab Connectivity

A pair of points $(x,y)$, $(x',y') \in [\![X,Y]\!]$ is called *taxicab connected,*

denoted $(x,y) \leftrightarrow (x',y')$, if $\exists$ a finite sequence $\left((x_i,y_i)\right)_{i=1}^{n}$ in $[\![X,Y]\!]$

i) beginning from $(x_1,y_1) = (x,y)$,

ii) ending in $(x_n,y_n) = (x',y')$,

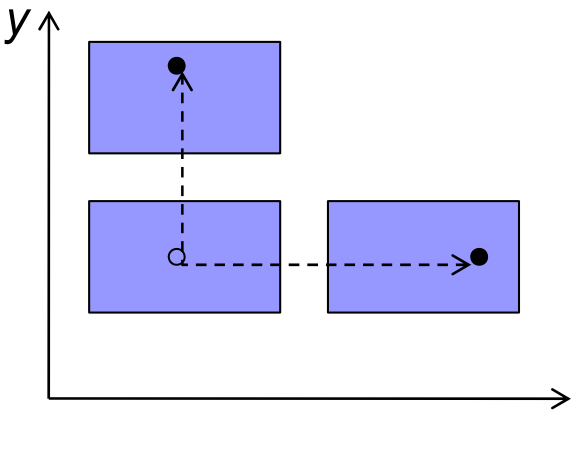
iii) and with each point in the sequence differing in at *most* one coordinate from its predecessor.

Every point in this sequence must yield the *same* $z$-value as its predecessor, since it has either the same $x$- or $y$-coordinate.

$\Rightarrow$ By induction, $(x,y)$ & $(x',y')$ yield the same $z$-value.

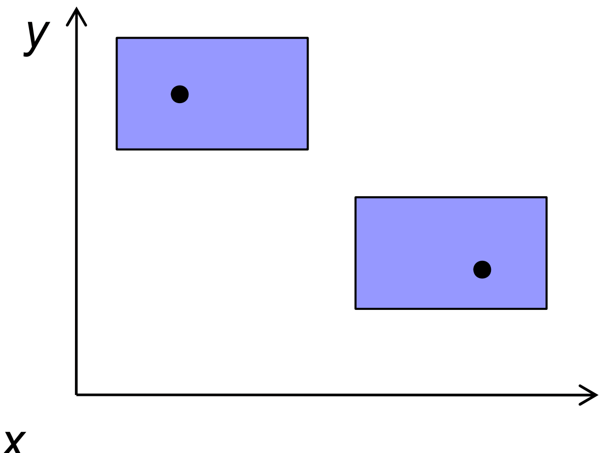
# Taxicab Connectedness Examples

([[$X, Y$]] = shaded area)



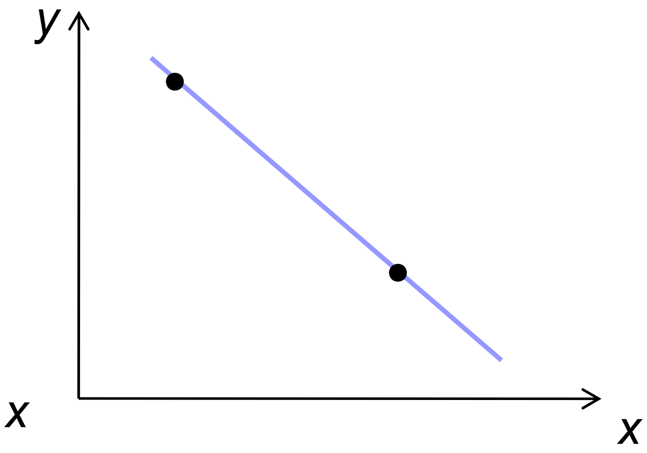
$(x,y) \leftrightarrow (x',y')$,

but disconnected in usual sense.



$(x,y) \not\leftrightarrow (x',y')$,

also disconnected in usual sense



$(x,y) \not\leftrightarrow (x',y')$,

but connected in usual sense.

# Taxicab Partition and Nonstochastic Information

***Thm :*** There is a unique partition $\mathcal{T}$ of $[\![X,Y]\!]$ in which

a) every pair of points in the same partition set is taxicab connected, but

b) *no* pair of points in different partition sets is taxicab connected.

Can be established that $\mathcal{T}$ defines the most refined shared data $Z$ that can be unambiguously determined from $X$ or $Y$ alone.

$\Rightarrow$ Define ***maximin information*** $\mathrm{I}^{*}[X;Y] := \log_2 |\mathcal{T}|$

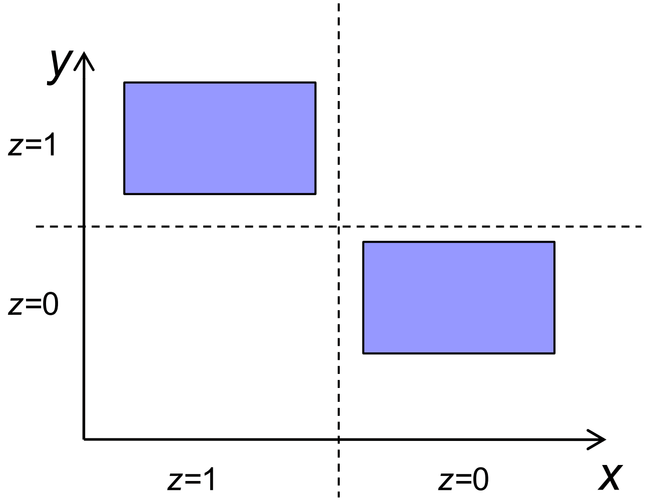
# Interpretation as a Common/Shared Variable

- Suppose *X* & *Y* are separately observed by two agents.
- Let the agents have functions *f* & *g* respectively s.t. $f(X)=g(Y)=:Z$

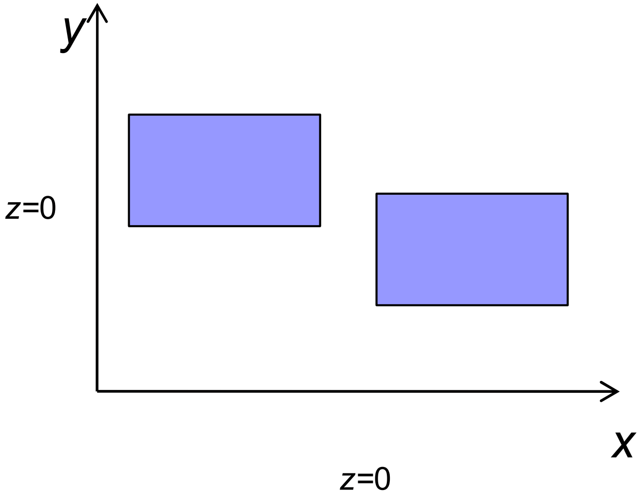
  ⇔ The agents can *unambiguously* agree on the value of the *common* variable *Z*.
- The more distinct values *Z* can take, the more refined is this shared knowledge.
- The values of *Z* induce a partition of the joint range [[*X,Y*]].
- Taxicab partition = the [[*X,Y*]]-partition induced by the *most refined common variable Z.*

# Examples

([[$X,Y$]] = shaded area)



$|\mathcal{T}|=2=\max.\#$ distinct values that can always be agreed on from separate observations of $X$ & $Y$.



$|\mathcal{T}|=1=\max.\#$ distinct values that can always be agreed on from separate observations of $X$ & $Y$.

# Some Key Properties of I*

***Symmetry :***

$$\mathrm{I}^*[X;Y] = \mathrm{I}^*[Y;X].$$

***More Data Can't Hurt :***

$$\mathrm{I}^*[X;Y] \leq \mathrm{I}^*[X;Y,W].$$

***"Data Processing" :***

If $W \leftrightarrow X \leftrightarrow Y$ is a Markov uncertainty chain, then

$$\mathrm{I}^*[W;Y] \leq \mathrm{I}^*[W;X].$$

# Uncertain Signals & Stationary Memoryless Channels

***Def :*** An *uncertain signal* $X$ is a mapping from $\boldsymbol{\Omega}$ to the space $\boldsymbol{X}^{\infty}$ of discrete-time signals $x:\mathbb{Z}_{\geq 0}\to \boldsymbol{X}$.

***Def :*** A *stationary memoryless uncertain channel* consists of a set-valued *transition function* $\boldsymbol{T}:\boldsymbol{X}\to \boldsymbol{2}^{\boldsymbol{Y}}$, and the family of all uncertain input-output signal pairs $(X,Y)$ s.t.

$$[\![ Y(k)\,|\,x(0:k),y(0:k-1)]\!] = [\![ Y(k)\,|\,x(k)]\!] = \boldsymbol{T}\big(x(k)\big)\subseteq \boldsymbol{Y},$$
$$\forall (x,y)\in [\![ X,Y]\!],\ k\in \mathbb{Z}_{\geq 0}.$$

# Channel Coding Theorem for Zero-Error Communication

***Thm*** : The zero-error capacity $C_0$ of a stationary memoryless uncertain channel coincides with the highest average rate of maximin information possible across it, i.e.

$$C_0 = \sup_{t \geq 0,\, X:\square X \square \subseteq \boldsymbol{X}^{\infty}} \frac{\mathrm{I}^*\left[X(0:t);Y(0:t)\right]}{t+1} = \lim_{t\to\infty} \sup_{X(0:t):\square X(0:t)\square \subseteq \boldsymbol{X}^{t+1}} \frac{\mathrm{I}^*\left[X(0:t);Y(0:t)\right]}{t+1}.$$

*Note* : $C_0$ is defined *operationally*, as the largest rate over all block codes that permit unambiguous recovery of the input sequence.

This result gives an *intrinsic* characterization.

# Remarks

- The idea of a *common (random) variable Z* comes from cryptography [Wolf & Wullschleger, ITW2004]

- There, *Z* is formally defined by the *connected components* of the discrete bipartite graph describing *(x,y)* pairs having joint prob.> 0.

- Taxicab connectedness generalises this to continuous-valued and mixed pairs of variables, not representable by discrete graphs.

- *C0* was shown by Wolf & Wullschleger to coincide with the maximum Shannon entropy rate over all common rv's *Z*. However, this is still a probabilistic characterisation.

- Maximin information coincides with the *Hartley* entropy of the maximal common rv *Z.*

# State Estimation of Disturbance-Free LTI Systems

$$X(t+1) = AX(t), \quad Y(t) = GX(t), \quad X(0) \text{ a uv.}$$

**Coder :** $Y(0:t) \mapsto S(t) \in \mathbf{S}$. No channel feedback.

**Erroneous Channel :** $\mathbf{S} \to \mathbf{2^Q}$

**Estimator :** $Q(0:t) \mapsto \hat{X}(t+1)$

Given parameters $\rho, l > 0$, the objectives are

**I)** $\rho$ **- exponential uniformly bounded estimation errors** :

For any uv $X(0)$ s.t. $\| X(0) \| \le l$, $\quad \sup_{t \ge 0, \omega \in \Omega} \rho^{-t} \left\| X(t) - \hat{X}(t) \right\| < \infty.$

**II)** $\rho$ **- exponential uniform convergence :**

For any uv $X(0)$ s.t. $\| X(0) \| \le l$, $\quad \lim_{t \to \infty} \sup_{\omega \in \Omega} \rho^{-t} \left\| X(t) - \hat{X}(t) \right\| = 0.$

# Assumptions

**DF1** : $(G, A_\rho)$ is observable, where $A_\rho := A$ restricted to invariant subspace governed by $|\text{eigenvalue}|\text{'s} \geq \rho$.

**DF2** : The channel does not depend on the initial plant state, i.e. the output sequence $Q(0:t)$ is conditionally unrelated to $X(0)$, given channel input sequence $S(0:t)$,

$$X(0) \leftrightarrow S(0:t) \leftrightarrow Q(0:t)$$

**DF3** : $A$ has one or more $|\text{eigenvalue}|\text{'s} > \rho$

# Criterion without Disturbances

If $\rho$ - exponential uniformly bounded estimation errors are achieved for some $l > 0$, then

$$C_0 \geq \sum_{|\lambda_i| \geq \rho} \log_2 \left| \frac{\lambda_i}{\rho} \right| =: H_\rho \qquad (*)$$

Conversely, if $(*)$ holds strictly, then for any $l > 0$, a coder - estimator that achieves $\rho$ - exponential uniform convergence can be constructed.

$$\begin{pmatrix} \text{Proof of second part : constructive.} \\ \text{Proof of first part : maximin information theory} \end{pmatrix}$$

# LTI State Estimation With Plant Disturbances

$$X(t+1) = AX(t) + V(t), \quad Y(t) = GX(t) + W(t),$$

**Assumptions :**

**D0 :** $(G, A)$ is detectable.

**D1 :** $A$ has one or more $|\text{eigenvalue}|$'s $> 1$.

**D2 :** Realisations of $V$ & $W$ are uniformly bounded in $\ell_\infty$.

**D3 :** The null signals $v, w = 0$ are valid disturbance realisations.

**D4 :** $X(0), V$ & $W$ are mutually unrelated.

**D5 :** The channel does not depend on the plant states and disturbances, i.e. the channel output $Q(0:t)$ is conditionally unrelated with $\big(X(0), V(0:t-1), W(0:t)\big)$, given the channel input $S(0:t)$,

$$\big(X(0), V(0:t-1), W(0:t)\big) \leftrightarrow S(0:t) \leftrightarrow Q(0:t)$$

# Criterion with Disturbances

If uniformly bounded estimation errors are achieved for some $l > 0$, then

$$C_0 \geq \sum_{|\lambda_i| \geq 1} \log_2 |\lambda_i| =: H \,. \qquad\qquad (**)$$

Conversely, if $(**)$ holds strictly, then for any $l > 0$, a coder - estimator that achieves uniformly bounded estimation errors can be constructed.

# Remarks

- In a stochastic setting (i.e. random channel and *X(0))* with no plant noise, it is known that almost-sure asymptotic convergence is possible iff ordinary capacity $C > H$ (Matveev & Savkin 2007)*.*

  The criterion here is stricter because a law of large numbers cannot be used to average out decoding errors.

- If bounded, nonstochastic disturbances are present, they showed that a.s. uniformly bounded errors are possible iff $C0 > H$. Proof used no info theory

# Conclusion

- Formulated a framework for modelling unknown variables without assuming the existence of distributions
- Defined nonprobabilistic analogues of independence & Markovness
- Proposed maximin information as a nonstochastic index of the most refined knowledge that can be agreed on from separate observations of two variables
- Showed that zero-error capacity coincides with the highest maximin info rate possible across the channel
- Used maximin info theory to derive tight conditions for uniform state estimation of LTI plants

# Future Work

- Channels with input or memory constraints
- Network maximin information theory
- Systems with feedback – preliminary results to appear in CDC 2012

# Extension
# - Zero Error Feedback Capacity

Theorem (GN, to appear in *CDC*12) :

The operational zero - error feedback capacity of a stationary memoryless uncertain channel can be expressed in terms of *directed* maximin information :

$$C_{0\mathrm{F}} = \lim_{t\to\infty} \sup_{X(0:t),Y(0:t)} \frac{1}{t+1}\sum_{k=0}^{t} I^{*}\left[X(k);Y(k)\mid Y(0:k-1)\right] =: I^{*}[X \to Y],$$

where

$$I^{*}[X;Y\mid Z] := \min_{z\in[[Z]]} \log_2 \left|\mathcal{T}[X;Y\mid z]\right|$$

is *conditional* maximin information.

# Thank You!

References

- GN, “A nonstochastic information theory for communication and state estimation”, http://arxiv.org/abs/1112.3471. (Provisionally accepted by *IEEE Trans Auto. Contr*; short version in *Proc. 9th IEEE Int. Conf. Control & Automation*, Santiago, Chile, Dec. 2011.)
- --, “ A nonstochastic information theory for feedback”, to appear in *Proc. IEEE CDC*, Dec. 2012.
- G. Klir, *Uncertainty and Information Foundations of Generalized Information Theory*, Wiley, 2006, ch. 2.
- H. Shingin and Y. Ohta, “Disturbance rejection with information constraints: Performance limitations of a scalar system for bounded and Gaussian disturbances," *Automatica*, 2012.
- S. Wolf and J. Wullschleger, “Zero-error information and applications in cryptography”, *Info. Theory Workshop*, San Antonio, USA ,2004.
- C.E. Shannon, “The zero-error capacity of a noisy channel”, *IRE Trans. Info. Theory*, vol. 2, 1956.
- S. Tatikonda and S. Mitter, “Control under communication constraints,” *IEEE TAC*., 2004.
- A.S. Matveev and A.V. Savkin, ``Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels,“ *Int. Jour. Contr*., 2007.
- - , ``An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels”, *SIAM J. Contr. Optim*., 2007.
- J. Massey, ``Causality, feedback and directed information," in *Int. Symp. Inf. Theory App*., 1990.