Data Processing Theorems and the Second Law of Thermodynamics

Neri Merhav

Department of Electrical Engineering

Technion – Israel Institute of Technology

Technion City, Haifa, Israel

ITA Workshop, San Diego, CA, USA, February 6–11, 2011

Outline

Background:

- Generalized data processing theorems (DPT's).
- "Second law" (H–theorem) of Markov processes + extensions.

Results:

- A generalized principle in a unified framework.
- New perspectives on generalized DPT's.
- Example.

Introduction – Generalized DPT's

Csiszár (1972) defined a generalized divergence (f–divergence):

$$D_Q(P_1\|P_2) = \int \mathrm{d}x P_1(x) Q\left(\frac{P_2(x)}{P_1(x)}\right),$$

where $Q =$ general convex function. For $P_1 =$joint distribution and $P_2 =$product-of-marginals,

$$I_Q(X;Y) = \int \mathrm{d}x\mathrm{d}y P(x,y) Q\left(\frac{P(x)P(y)}{P(x,y)}\right)$$

this yields a generalized mutual information, which satisfies a DPT.

Ziv & Zakai (1973) – same idea independently with emphasis on improved lower bounds on distortion

$$R_Q(d) \le C_Q \quad \Longleftrightarrow \quad d \ge R_Q^{-1}(C_Q).$$

Introduction – Generalized DPT's (Cont'd)

Zakai & Ziv (1975) have further generalized their mutual information measure to be

$$I^Q(X;Y) = \int \mathrm{d}x\mathrm{d}y \cdot P(x,y) \cdot Q\left(\frac{\mu_1(x,y)}{P(x,y)}, \ldots, \frac{\mu_k(x,y)}{P(x,y)}\right),$$

where μ_i are arbitrary measures.

This class of info measures is rich enough to provide tight bounds: $\forall$ source and channel, $\exists\ Q$ and $\{\mu_i\}$ such that

$$[\text{lower bound on } d] = [d \text{ of optimum communication system}].$$

H–Theorem & Other Monotonicity Thms

The (microscopic) state of a physical system – normally modeled as a Markov process, $\{X_t\}$. In the discrete–state, continuous–time case define the state–transition rates according to:

$$\mathsf{Pr}\{X_{t+\delta} = x' | X_t = x\} = W_{xx'}\delta + o(\delta) \qquad x' \neq x$$

and

$$P_t(x) = \mathsf{Pr}\{X_t = x\}.$$

We then have

$$P_{t+dt}(x) = \sum_{x' \neq x} P_t(x') W_{x'x} \mathsf{d}t + P_t(x) \left(1 - \sum_{x' \neq x} W_{xx'} \mathsf{d}t \right),$$

which yields the Master equations:

$$\frac{\mathsf{d}P_t(x)}{\mathsf{d}t} = \sum_{x' \in \mathcal{X}} [P_t(x') W_{x'x} - P_t(x) W_{xx'}] \qquad x \in \mathcal{X}$$

Markov Processes, H–Theorem, ... (Cont'd)

In steady–state, $P_t(x) = P(x)$ are all time–invariant:

$$\sum_{x' \in \mathcal{X}} [P(x')W_{x'x} - P(x)W_{xx'}] = 0, \quad \forall\, x \in \mathcal{X}.$$

The net "probability flux" from/to each state vanishes (incoming flux = outgoing flux). In steady–state, there can be cyclic currents. For example,

$$1 \to 2 \to 3 \to 1 \to 2 \to 3 \to \quad \cdots$$

Stronger notion of time–invariance: detailed balance (DB):

$$P(x')W_{x'x} - P(x)W_{xx'} = 0, \quad \forall\, x, x' \in \mathcal{X}.$$

- DB occurs iff $\{X_t\}$ is time–reversible: $\mathcal{L}\{X_t\} = \mathcal{L}\{X_{-t}\}$.
- In physics, this corresponds to equilibrium (time–reversal symmetry).
- In DB, there are no cyclic probability currents. Ex: M/M/1 queue.
- In an isolated system, $P(x) = 1/|\mathcal{X}|$, and then DB means $W_{xx'} = W_{x'x}$.

The H–Theorem

Defining

$$H(X_t) = -\sum_{x \in \mathcal{X}} P_t(x) \log P_t(x),$$

the H–theorem asserts that if:

- $\{X_t\}$ obeys DB, and
- $P(x) = 1/|\mathcal{X}|$ for all x,

then:

$$\frac{\mathrm{d}H(X_t)}{\mathrm{d}t} \geq 0.$$

Comments:

- Discrete–time analogue – holds too, and even without DB.
- Similar to the 2nd law of thermo, but not precisely equivalent.
- Arrow of time: how does this settle with time–reversal symmetry?

Extension to Non–Isolated Systems

What if $P(x)$ is not uniform? In [Cover & Thomas '06], it is shown that

$$D(P_t \| P) = \sum_{x \in \mathcal{X}} P_t(x) \log \frac{P_t(x)}{P(x)} \quad \searrow$$

Indeed, for P uniform

$$D(P_t \| P) = \log |\mathcal{X}| - H(X_t).$$

- Detailed balance is not needed.
- Maximum entropy $\rightarrow$ minimum free energy.
- Characterizes monotonic convergence $P_t \rightarrow P$ in the divergence sense.

More generally, for P_t and P'_t, two time–varying state distributions pertaining a given Markov process, $D(P_t \| P'_t) \quad \searrow$ [Cover & Thomas '06].

Monotonicity of the f-Divergence

In [Kelly '79]: If P is a steady–state distribution

$$D_Q(P\|P_t) = \sum_{x\in\mathcal{X}} P(x)Q\left(\frac{P_t(x)}{P(x)}\right) \quad \searrow$$

for whatever P_t that evolves according to the Markov process. This allows a general Q, and it covers both $D(P\|P_t)$ and $D(P_t\|P)$, but not $D(P_t\|P_t')$. To be handled soon...

Define $P_t(x,x') = P(X_0 = x, X_t = x')$ and $P_t'(x,x') = P(X_0 = x)P(X_t = x')$ then

$$D_Q(P_t\|P_t') = \sum_{x,x'} P_t(x,x')Q\left(\frac{P_t'(x,x')}{P_t(x,x')}\right) \quad \searrow$$

because here $D_Q(P_t\|P_t') = I_Q(X_0; X_t)$, and the above is the Ziv–Zakai–Csiszár DPT for the Markov chain $X_0 \to X_t \to X_{t+1}$.

A Unified Framework

This monotonicity thm does not cover the entire picture. Can we put everything under one umbrella?

Yes, we can! including the 1975 Ziv–Zakai information measure.
Two observations:

1. The above thm extends trivially to

$$U_t = \sum_{x \in \mathcal{X}} P(x) Q\left(\frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right)$$

where $\{\mu_t(x)\}$ all obey the Markov recursion $\mu_{t+1}^i(x) = \sum_{x'} \mu_t^i(x') P(x|x')$ and Q is jointly convex.

2. If $Q(u_1, \ldots, u_k)$ is convex, then so is its perspective

$$\tilde{Q}(v, u_1, \ldots, u_k) = v \cdot Q\left(\frac{u_1}{v}, \ldots, \frac{u_k}{v}\right) \quad v > 0.$$

A Unified Framework (Cont'd)

Thm: Let $\mu_t^0, \mu_t^1, \ldots, \mu_t^k$ be arbitrary measures that obey the Markov recursion and assume $P \gg \mu_t^0$ for all t. Then,

$$V_t \triangleq \sum_x \mu_t^0(x) Q\left(\frac{\mu_t^1(x)}{\mu_t^0(x)}, \ldots, \frac{\mu_t^k(x)}{\mu_t^0(x)}\right) \quad \searrow$$

Proof:

$$V_t = \sum_x P(x) \cdot \frac{\mu_t^0(x)}{P(x)} Q\left(\frac{\mu_t^1(x)/P(x)}{\mu_t^0(x)/P(x)}, \ldots, \frac{\mu_t^k(x)/P(x)}{\mu_t^0(x)/P(x)}\right)$$

$$= \sum_x P(x) \tilde{Q}\left(\frac{\mu_t^0(x)}{P(x)}, \frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right).$$

The assumption $P \gg \mu_t^0$ can be relaxed.

The 1975 ZZ DPT for the Markov chain $X_0 \to X_t \to X_{t+1}$ is obtained for $\mu_t^0(x, x') = P(X_0 = x, X_t = x')$.

A New Perspective on the 1973 Ziv–Zakai DPT

While the 1973 ZZ info measure is

$$I_Q(X;Y) = \sum_{x,y} P(x,y) Q\left(\frac{P(x)P(y)}{P(x,y)}\right),$$

one can use any μ_0 and μ_1 (satisfying the Markov relations) and define

$$I_Q(X;Y) = \sum_{x,y} \mu_0(x,y) Q\left(\frac{\mu_1(x,y)}{\mu_0(x,y)}\right),$$

because

$$\begin{aligned} I_Q(X;Y) &= \sum_{x,y} P(x,y) \cdot \frac{\mu_0(x,y)}{P(x,y)} Q\left(\frac{\mu_1(x,y)/P(x,y)}{\mu_0(x,y)/P(x,y)}\right) \\ &= \sum_{x,y} P(x,y) \tilde{Q}\left(\frac{\mu_0(x,y)}{P(x,y)}, \frac{\mu_1(x,y)}{P(x,y)}\right) = \text{1975 ZZ info measure} \end{aligned}$$

A New Perspective ... (Cont'd)

Both μ's can be of the form

$$\mu(x,y) = s_0 P(x,y) + \sum_{x_i \in \mathcal{X}} s_i P(x) P(y|x = x_i)$$

with arbitrary positive coefficients s_0 and $\{s_i\}$.

For example,

$$I_Q(X;Y) = \sum_{x,y} [P(x,y) + sP(x)P(y)] \cdot Q\left(\frac{P(x)P(y)}{P(x,y) + sP(x)P(y)}\right)$$

satisfies a DPT for every $s \geq 0$.
$s = 0 \quad \rightarrow$ 1973 ZZ information measure.

Even for 1973 ZZ DPT (univariate Q), we have added a degree of freedom.
Important since only few functions Q, are easy to work with.

Example

Source U and the reconstuction V are uniform over $\{0, 1, \ldots, K-1\}$.

$$d(u,v) = \begin{cases} 0 & v = u \\ 1 & v = (u+1) \text{ mod } K \\ \infty & \text{elsewhere} \end{cases}$$

Channel: clean L–ary channel.

For $Q(z) = -\sqrt{z}$, we obtain

$$I_Q(U;V) = -\sum_{u,v} P(u)P(v)\sqrt{s + \frac{P(v|u)}{P(v)}}.$$

Example (Cont'd)

Applying the DPT $R_Q(d) \leq C_Q$ (for a given s), we obtain the lower bound

$$d \geq d_s.$$

For $s = 0$ (ZZ '73), we have:

$$d_0 = \frac{1}{2} - \frac{1}{2}\sqrt{2\theta - \theta^2},$$

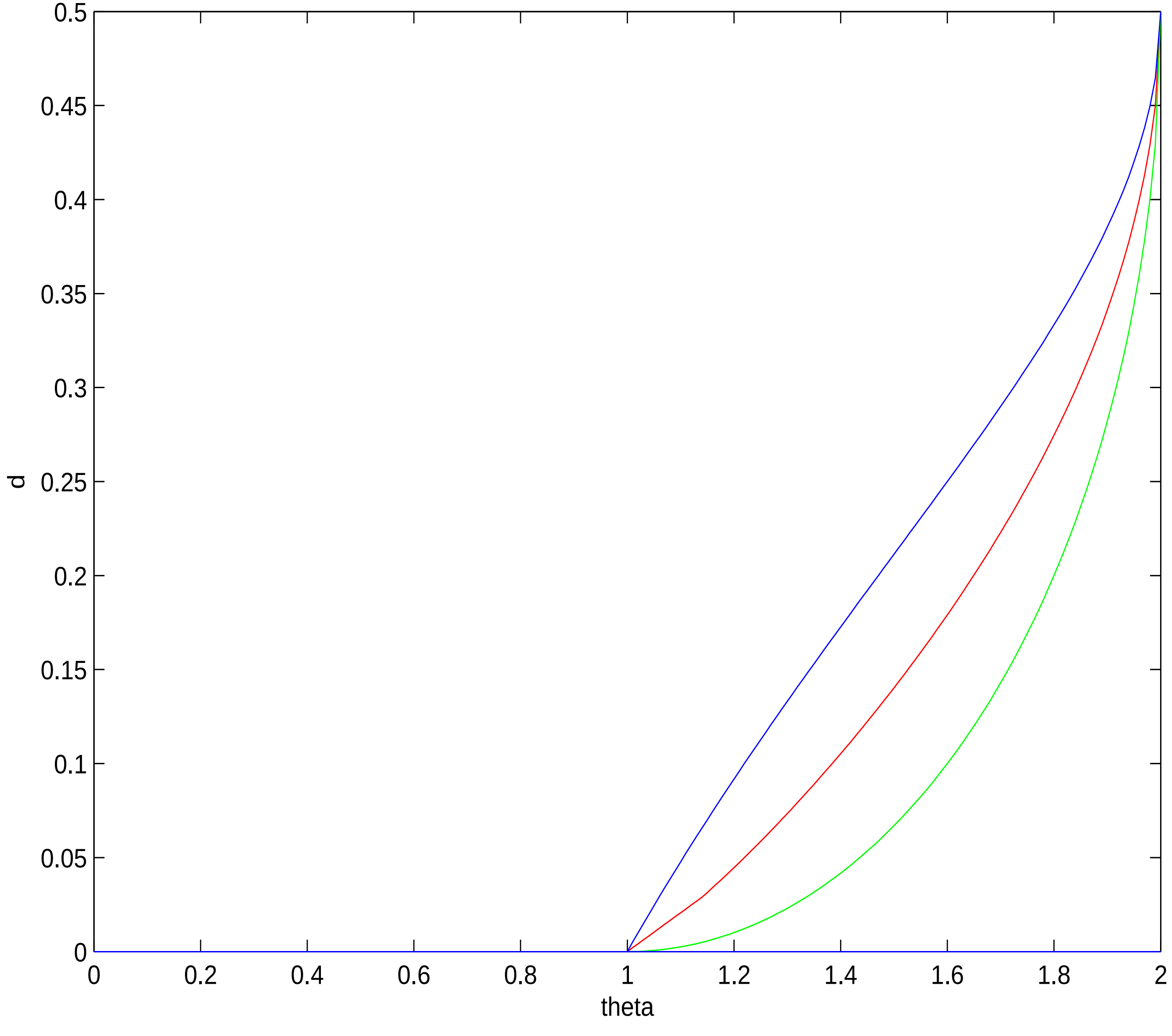
where $\theta \triangleq K/L$.

For $s \to \infty$,

$$d_\infty = \frac{1}{2} - \frac{1}{2\theta}\sqrt{2\theta - \theta^2},$$

which is larger than d_0 for all $1 < \theta < 2$.

The Shannon bound: $d_{Shannon} = h^{-1}(\log \theta)$ is in between.


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
d
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
theta

Conclusion

- Unified framework relating monotonicity theorems and generalized DPT's.
- The H–theorem was substantially generalized.
- A new perspective on the ZZ DPT that gives useful bounds.