Dissipation of information in channels with input constraints

Yury Polyanskiy

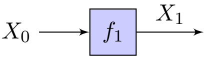
Department of EECS
MIT

yp@mit.edu

Joint work with Yihong Wu (UIUC)

Apr 28, 2014

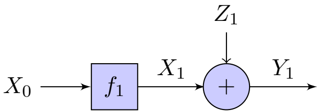
Relaying data across a chain of Gaussian channels



- Original message: X_0
- Encoders f_k satisfying power constraint

$$\mathbb{E}[X_k^2] \le P$$

Relaying data across a chain of Gaussian channels

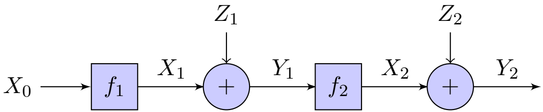


- Original message: X_0
- Encoders f_k satisfying power constraint

$$\mathbb{E}[X_k^2] \le P$$

- Noise: $Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

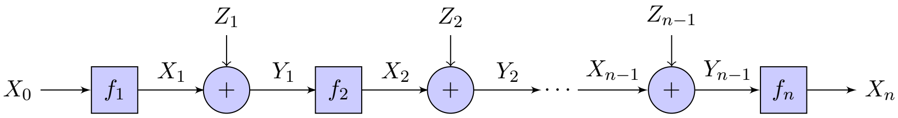
Relaying data across a chain of Gaussian channels



- Original message: X_0
- Encoders f_k satisfying power constraint

$$\mathbb{E}[X_k^2] \le P$$

- Noise: $Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

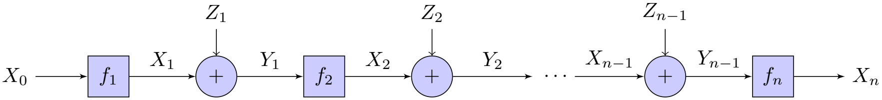


- Original message: X_0
- Encoders f_k satisfying power constraint

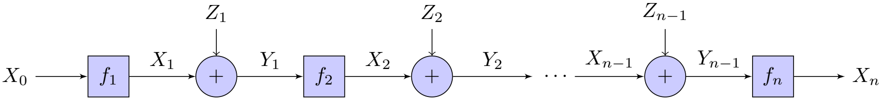
$$\mathbb{E}[X_k^2] \leq P$$

- Noise: $Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- Goal: Design f's so that $X_0 \approx X_n$

Can we preserve any information?



Can we preserve any information?



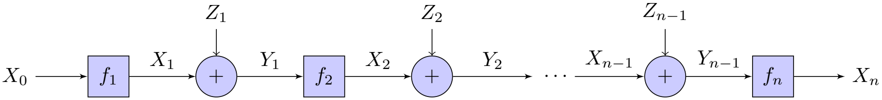
Intuition

each stage has finite energy budget

$\Rightarrow$ cannot denoise completely

$\Rightarrow$ noise accumulates and kills dependency

Can we preserve any information?



Intuition

each stage has finite energy budget

$\Rightarrow$ cannot denoise completely

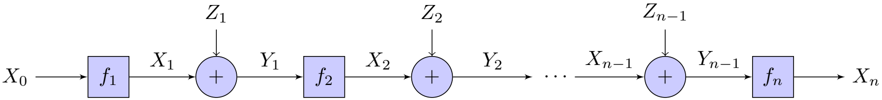
$\Rightarrow$ noise accumulates and kills dependency

Conjecture: asymptotic decoupling

Regardless of X_0 or $\{f_1, \ldots, f_n\}$

$$P_{X_0 X_n} \approx P_{X_0} P_{X_n} \qquad n \gg 1$$

How to gauge decoupling?

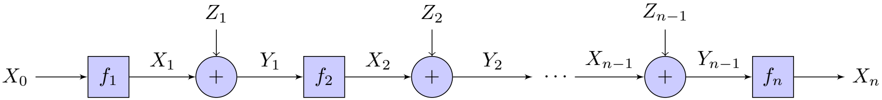


$$P_{X_0 X_n} \approx P_{X_0} P_{X_n}$$

Quantitatively:

- (TV) $\mathrm{TV}(P_{X_0 X_n}, P_{X_0} P_{X_n}) \to 0$
- (KL) $D(P_{X_0 X_n} \| P_{X_0} P_{X_n}) \to 0$
- (Correlation) $\rho(X_0, X_n) \to 0$

How to gauge decoupling?



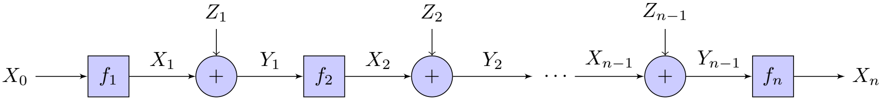
$$P_{X_0 X_n} \approx P_{X_0} P_{X_n}$$

Quantitatively:

- (TV) $\quad \mathrm{TV}(P_{X_0 X_n}, P_{X_0} P_{X_n}) \to 0$
- (KL) $\quad D(P_{X_0 X_n} \| P_{X_0} P_{X_n}) \to 0$

  Equiv.: $\quad I(X_0; X_n) \to 0$
- (Correlation) $\quad \rho(X_0, X_n) \to 0$

  Equiv.: $\quad \mathrm{mmse}(X_0 | X_n) \to \mathrm{Var}[X_0]$



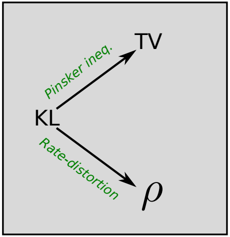
$$P_{X_0 X_n} \approx P_{X_0} P_{X_n}$$

Quantitatively:

- (TV) $\quad \mathsf{TV}(P_{X_0 X_n}, P_{X_0} P_{X_n}) \to 0$
- (KL) $\quad D(P_{X_0 X_n} \| P_{X_0} P_{X_n}) \to 0$

  Equiv.: $\quad I(X_0; X_n) \to 0$
- (Correlation) $\quad \rho(X_0, X_n) \to 0$

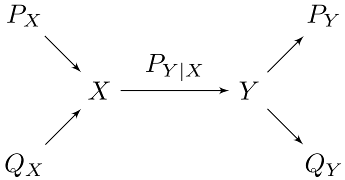
  Equiv.: $\quad \mathrm{mmse}(X_0 | X_n) \to \mathrm{Var}[X_0]$



Attempt 1: data processing

Data processing inequality

- KL divergence

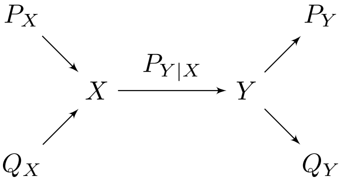


$$\Rightarrow D(P_Y||Q_Y) \leq D(P_X||Q_X)$$

(applies to any f-divergence, in particular TV)

Data processing inequality

- KL divergence



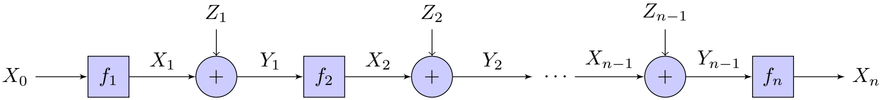
$$\Rightarrow D(P_Y||Q_Y) \le D(P_X||Q_X)$$

(applies to any f-divergence, in particular TV)

- mutual information

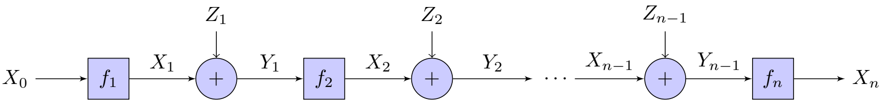
$$U \to X \to Y \Rightarrow I(U;Y) \le I(U;X)$$

Applying data processing



$$I(X_0;X_n) \le I(X_0;Y_{n-1}) \le I(X_0;X_{n-1}) \le \cdots \le I(X_0;X_1)$$

Applying data processing



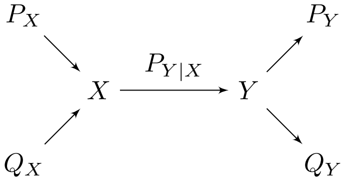
$$I(X_0; X_n) \le I(X_0; Y_{n-1}) \le I(X_0; X_{n-1}) \le \cdots \le I(X_0; X_1)$$

- OK: $I(X_0; X_n)$ is non-increasing
- Need a quantitative data processing inequality

Attempt 2: strong data processing

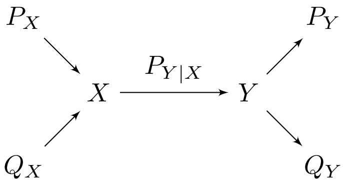
Strong data processing inequality

- KL divergence



$$\Rightarrow D(P_Y||Q_Y) \le \eta_{\mathrm{KL}} D(P_X||Q_X)$$

- KL divergence

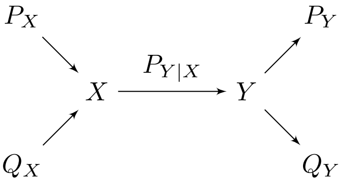


$$\Rightarrow D(P_Y||Q_Y) \le \eta_{\mathrm{KL}} D(P_X||Q_X)$$

- mutual information

$$U \to X \to Y \Rightarrow I(U;Y) \le \eta_{\mathrm{KL}} I(U;X)$$

Strong data processing inequality

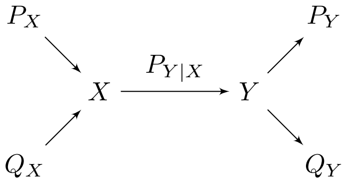


- For fixed $P_{Y|X}$, contraction ratio:

$$\eta_{\mathrm{KL}} = \sup_{P_X \neq Q_X} \frac{D(P_Y||Q_Y)}{D(P_X||Q_X)} = \sup_{U \to X \to Y} \frac{I(U;Y)}{I(U;X)}$$

- Ahlswede-Gács: for discrete indecomposable channels $\eta_{\mathrm{KL}} < 1$

Strong data processing inequality

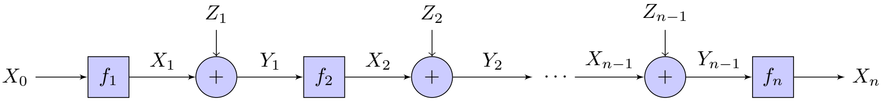


- For fixed $P_{Y|X}$, contraction ratio:

$$\eta_{\mathrm{KL}} = \sup_{P_X \neq Q_X} \frac{D(P_Y \| Q_Y)}{D(P_X \| Q_X)} = \sup_{U \to X \to Y} \frac{I(U;Y)}{I(U;X)}$$

- Ahlswede-Gács: for discrete indecomposable channels $\eta_{\mathrm{KL}} < 1$
- ... $\eta_{\mathrm{KL}} =$ hypercontractivity ratio
- ... related to LSI etc: [Witsenhausen], [Erkip-Cover], [Cohen-Kempermann-Zbăganu], [Del Moral-Ledoux-Miclo], [Anantharam-Gohari-Kamath-Nair], [Raginsky], ...

Strict contraction

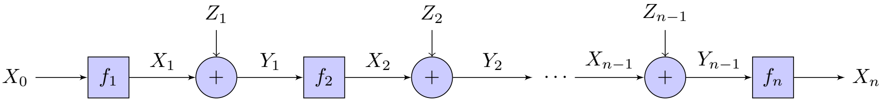


If $\eta_{\mathrm{KL}} < 1$, then

$$I(X_0; X_n) \leq I(X_0; Y_{n-1}) \leq \eta_{\mathrm{KL}} I(X_0; X_{n-1}) \leq \cdots \leq \eta_{\mathrm{KL}}^{n-1} I(X_0; X_1)$$

$$\to 0 \quad \text{exponentially}$$

Strict contraction

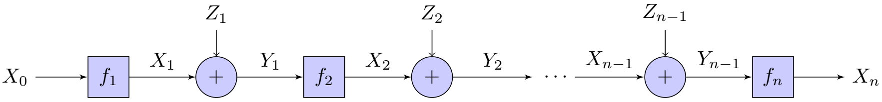


If $\eta_{\mathrm{KL}} < 1$, then

$$I(X_0; X_n) \le I(X_0; Y_{n-1}) \le \eta_{\mathrm{KL}} I(X_0; X_{n-1}) \le \cdots \le \eta_{\mathrm{KL}}^{n-1} I(X_0; X_1)$$
$$\to 0 \quad \text{exponentially}$$

- Sad: $\eta_{\mathrm{KL}} = 1$ for $Y = X + Z$, $\mathbb{E}[|X|^2] \le P$ (AWGN)

Strict contraction



If $\eta_{\mathrm{KL}} < 1$, then

$$I(X_0; X_n) \le I(X_0; Y_{n-1}) \le \eta_{\mathrm{KL}} I(X_0; X_{n-1}) \le \cdots \le \eta_{\mathrm{KL}}^{n-1} I(X_0; X_1)$$

$\to 0$ exponentially

- Sad: $\eta_{\mathrm{KL}} = 1$ for $Y = X + Z$, $\mathbb{E}[|X|^2] \le P$ (AWGN)
- Good: $\eta_{\mathrm{KL}} < 1$ for $Y = X + Z$, $|X| \le \sqrt{P}$ (amplitude constr.)

Dobrushin coefficient

- total variation

$$\mathsf{TV}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$$

- Dobrushin coefficient

$$\eta_{\mathrm{TV}} = \sup_{P_X \neq Q_X} \frac{\mathsf{TV}(P_Y, Q_Y)}{\mathsf{TV}(P_X, Q_X)} = \sup_{x,x'} \mathsf{TV}(P_{Y|X=x}, P_{Y|X=x'}).$$

- Markov process, stat physics, uniqueness of Gibbs measure
- [Cohen-Kempermann-Zbăganu'98]

$$\eta_{\mathrm{KL}} \leq \eta_{\mathrm{TV}}$$

Dobrushin coefficient

- If $|X| \leq A$ a.s.,

$$\eta_{\mathrm{KL}} \leq \eta_{\mathrm{TV}} = \sup_{x,x' \in [-A,A]} \mathsf{TV}(\mathcal{N}(0,x), \mathcal{N}(0,x')) \\ = \mathsf{TV}(\mathcal{N}(-A,1), \mathcal{N}(A,1)) < 1.$$

- Amplitude constraint $\Rightarrow$ exponential convergence of mutual information, etc.
- Short proof *and extension* of strongest results on Gaussian line-networks [Subramanian et al. '11, '12]

Dobrushin coefficient

- If $|X| \le A$ a.s.,

$$\eta_{\mathrm{KL}} \le \eta_{\mathrm{TV}} = \sup_{x,x' \in [-A,A]} \mathsf{TV}(\mathcal{N}(0,x), \mathcal{N}(0,x')) \\ = \mathsf{TV}(\mathcal{N}(-A,1), \mathcal{N}(A,1)) < 1.$$

- Amplitude constraint $\Rightarrow$ exponential convergence of mutual information, etc.
- Short proof *and extension* of strongest results on Gaussian line-networks [Subramanian et al. '11, '12]
- What about average power constraint?

Attempt 3: truncation arguments

FAIL.

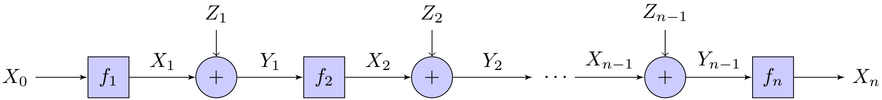
No contraction!

- No contraction even with power constraint

$$\begin{cases} P_X = (1-t)\delta_0 + t\delta_{\sqrt{P/t}} \\ Q_X = (1-t)\delta_0 + t\delta_{-\sqrt{P/t}} \end{cases} \Rightarrow \frac{\mathsf{TV}(P_Y, Q_Y)}{\mathsf{TV}(P_X, Q_X)} \to 1$$

- same for divergence and mutual information
- Maybe subexponential convergence?

Main results: Information nevertheless dissipates



Theorem

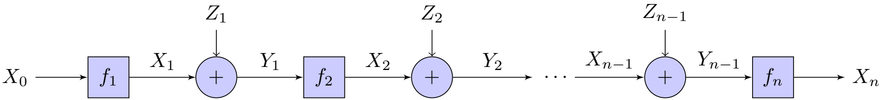
For any processors $\{f_n\}$,

$$\mathrm{TV}(P_{X_0 X_n}, P_{X_0} P_{X_n}) \le \frac{CP}{\log n}$$

$$\rho(X_0, X_n) \le \sqrt{\frac{CP \log\log n}{\log n}}$$

$$I(X_0; X_n) \le \frac{CP \log\log n}{\log n}$$

Main results: Information nevertheless dissipates



Theorem

For any processors $\{f_n\}$,

$$\mathrm{TV}(P_{X_0 X_n}, P_{X_0}P_{X_n}) \le \frac{CP}{\log n}$$

$$\rho(X_0, X_n) \le \sqrt{\frac{CP \log\log n}{\log n}}$$

$$I(X_0; X_n) \le \frac{CP \log\log n}{\log n}$$

Theorem

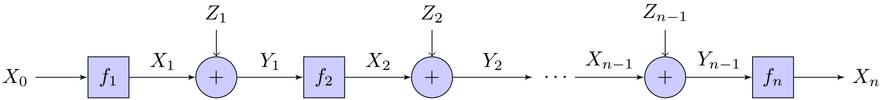
There exist $\{f_n\}$,

$$\mathrm{TV}(P_{X_0 X_n}, P_{X_0}P_{X_n}) \ge \frac{cP}{\log n}$$

$$\rho(X_0, X_n) \ge \sqrt{\frac{cP}{\log n}}$$

$$I(X_0; X_n) \ge \frac{cP}{\log n}$$

where c, C are universal constants.



Theorem

For any processors $\{f_n\}$,

$$\mathrm{TV}(P_{X_0X_n}, P_{X_0}P_{X_n}) \le \frac{CP}{\log n}$$

$$\rho(X_0, X_n) \le \sqrt{\frac{CP\log\log n}{\log n}}$$

$$I(X_0; X_n) \le \frac{CP\log\log n}{\log n}$$

Theorem

There exist $\{f_n\}$,

$$\mathrm{TV}(P_{X_0X_n}, P_{X_0}P_{X_n}) \ge \frac{cP}{\log n}$$

$$\rho(X_0, X_n) \ge \sqrt{\frac{cP}{\log n}}$$

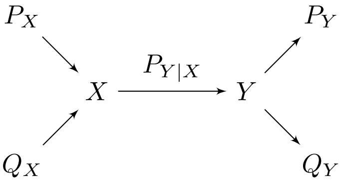
$$I(X_0; X_n) \ge \frac{cP}{\log n}$$

where c, C are universal constants.

Everything converges but über slowly!

Proof ideas

Revisit strong data-processing

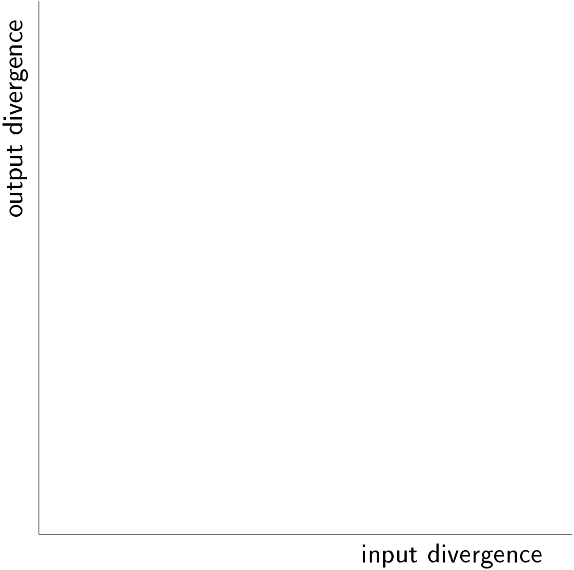


- Strong D.P.:

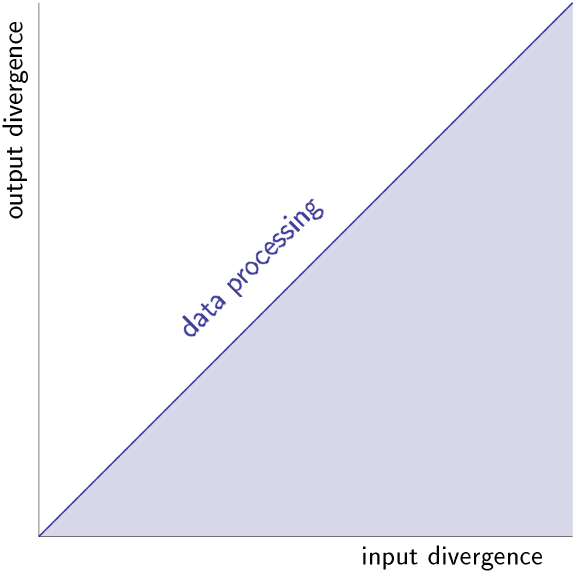
$$D(P_Y \| Q_Y) \leq \eta_{KL} D(P_X \| Q_X)$$

- More precise characterization: joint range
$(P_X, Q_X) \mapsto (D(P_X \| Q_X), D(P_Y \| Q_Y)) \in \mathbb{R}_+^2$

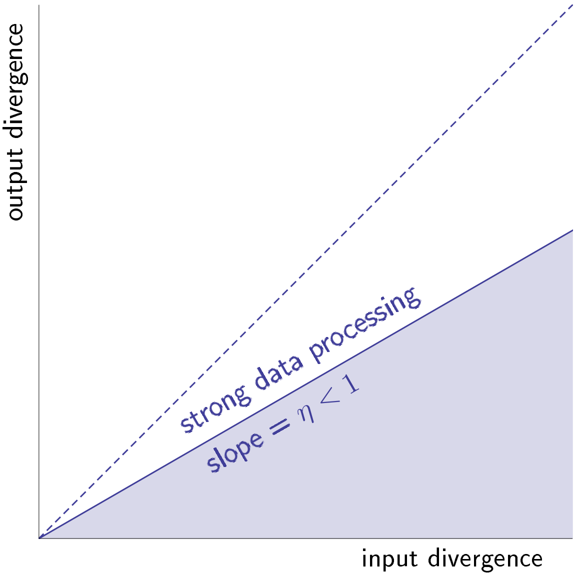
Joint range



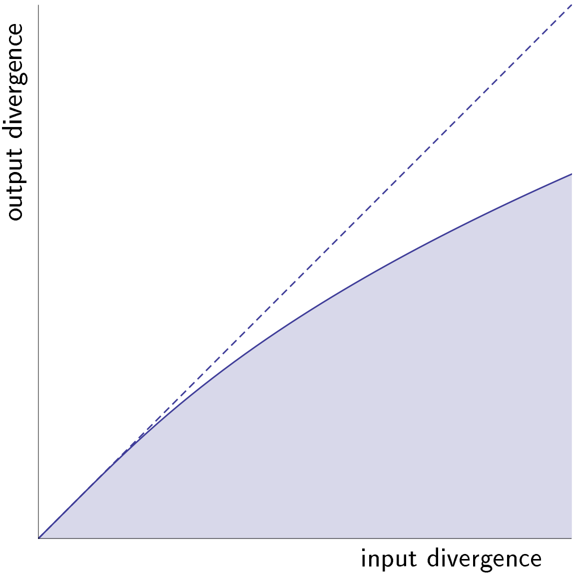
Joint range



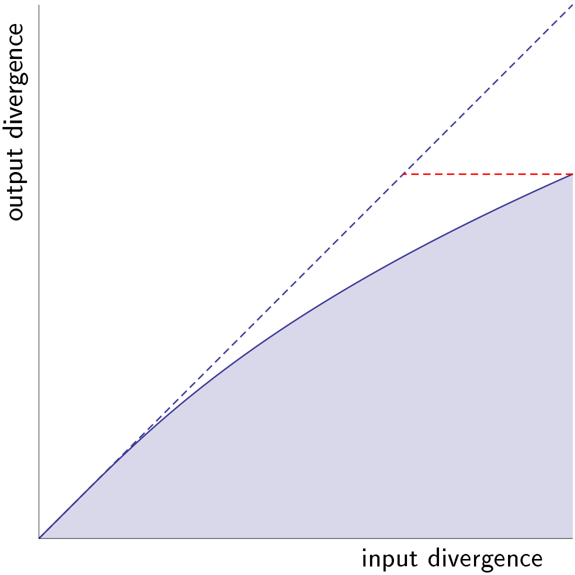
Joint range


output divergence
strong data processing
slope = $\eta < 1$
input divergence

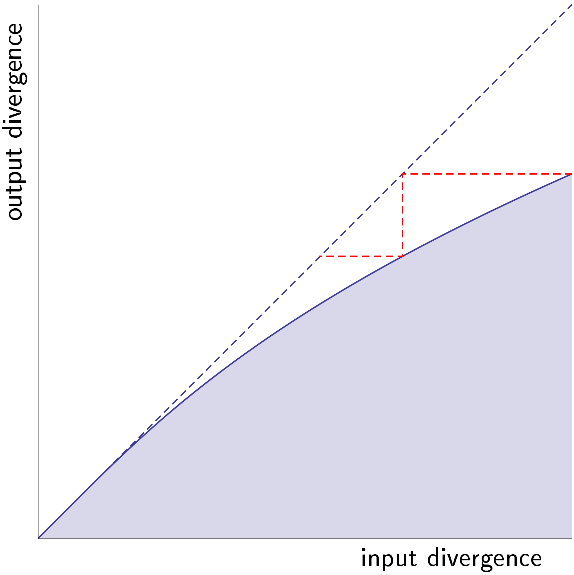
Joint range



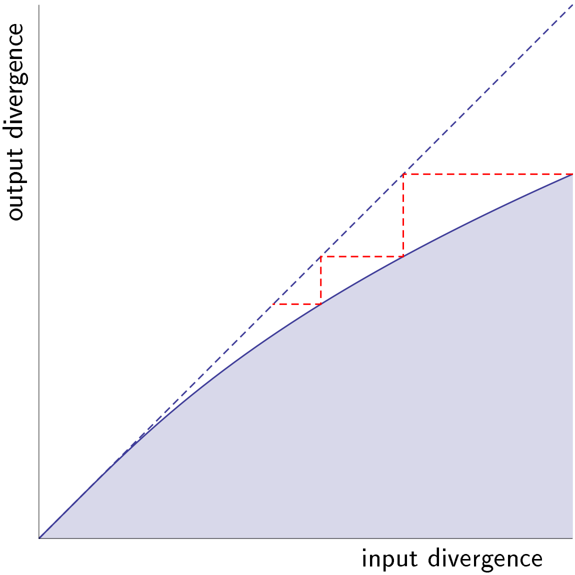
Joint range



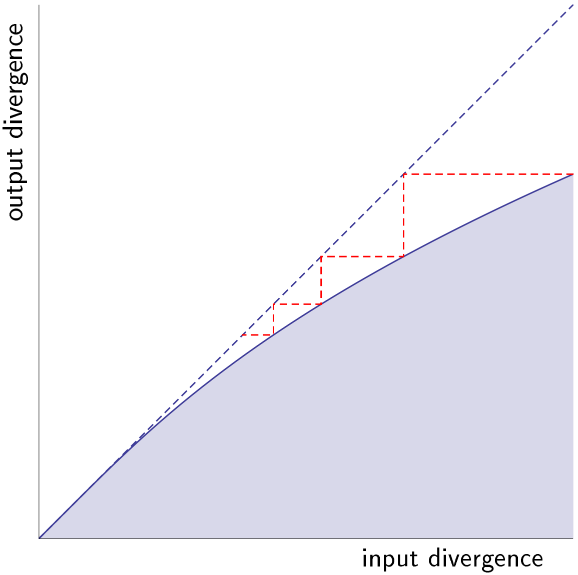
Joint range



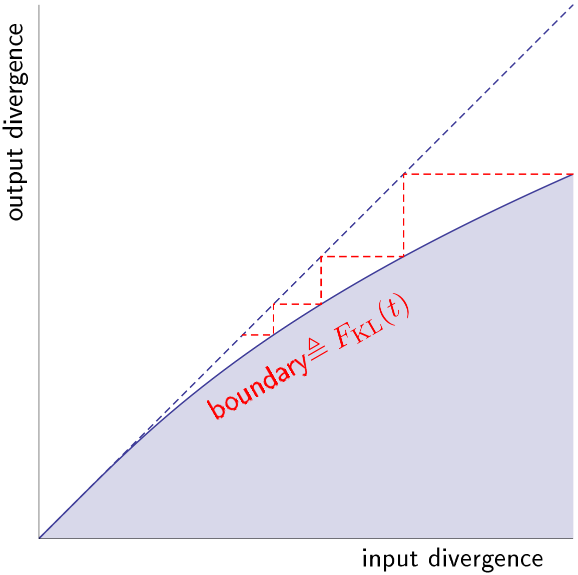
Joint range



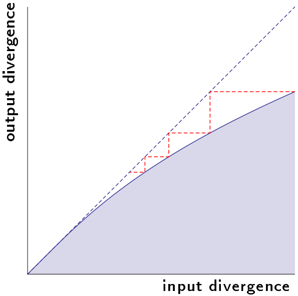
Joint range



Joint range

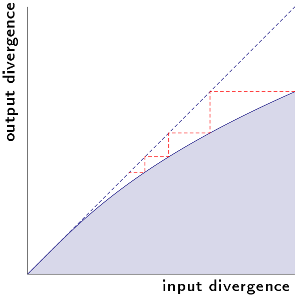

output divergence
boundary $\triangleq F_{\mathrm{KL}}(t)$
input divergence

Convergence without contraction



Punchline:

If $D(P_X \| Q_X)$ vs $D(P_X * \mathcal{N} \| Q_X * \mathcal{N})$ curved $\Rightarrow$ done (KL$\to 0$).



Punchline:

If $D(P_X \| Q_X)$ vs $D(P_X * \mathcal{N} \| Q_X * \mathcal{N})$ curved $\Rightarrow$ done (KL$\to 0$).

- Sad news: For KL the boundary $F_{\mathrm{KL}}(t) = t$
- Good news: For TV the boundary $F_{\mathsf{TV}}(t) < t$ (!)

$F_{\mathsf{TV}}(t), t \in [0,1]$ – Dobrushin curve of the channel

Strategy

power constraint

$\Downarrow$

cannot transmit 1 bit

$\Downarrow$

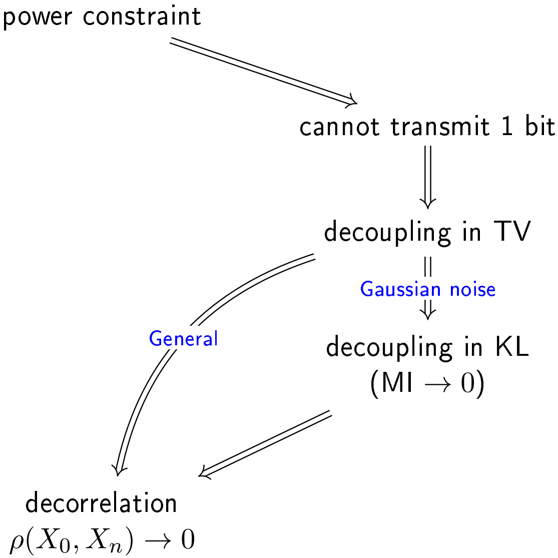
decoupling in TV

$\Downarrow$

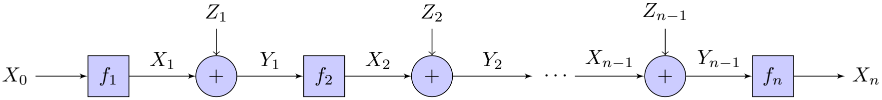
decoupling in KL
(MI $\to 0$)

$\Downarrow$

decorrelation
$\rho(X_0, X_n) \to 0$

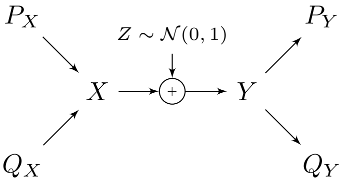

power constraint
cannot transmit 1 bit
decoupling in TV
Gaussian noise
General
decoupling in KL
(MI $\to 0$)
decorrelation
$\rho(X_0, X_n) \to 0$

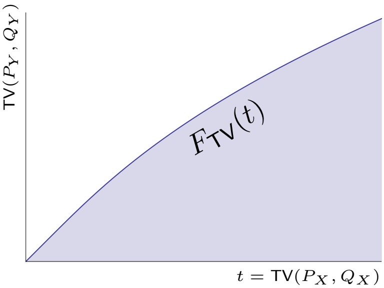
Transmit one bit



- $X_0 = \pm 1$ equiprobably
- Conditional law: $\mathbb{P}(\cdot|X_0 = 1) = P$ and $\mathbb{P}(\cdot|X_0 = -1) = Q$
- Reduce to testing P_{X_n} vs. Q_{X_n}

Dobrushin curve





- Upper boundary: $F_{\mathsf{TV}} : [0,1] \to [0,1]$

$$F_{\mathsf{TV}}(t) = \sup \Big\{ \mathsf{TV}(P_Y, Q_Y) : \mathsf{TV}(P_X, Q_X) \le t, \\ \mathbb{E}_{P_X}|X|^2 + \mathbb{E}_{Q_X}|X|^2 \le 2P \Big\}.$$

- Dobrushin coefficient η_{TV} = maximal slope of Dobrushin curve F_{TV}

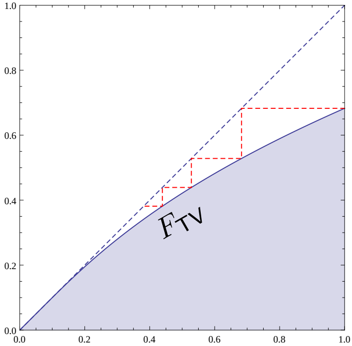
Dobrushin curve of AWGN

Theorem

Under power constraint $\mathbb{E}|X|^2 \le P$,

$$F_{\mathsf{TV}}(t) = t\left(1 - 2\mathsf{Q}\left(\sqrt{\frac{P}{t}}\right)\right)$$

where Q = complementary normal CDF.



Note:

- F_{TV} smooth but not analytic: $F_{\mathsf{TV}}'(0) = 1$, $F_{\mathsf{TV}}^{(k)}(0) = 0$
- iterative mapping:

$$\mathsf{TV}(P_{X_n}, Q_{X_n}) \le F_{\mathsf{TV}} \circ F_{\mathsf{TV}} \cdots \circ F_{\mathsf{TV}}(1) = O\left(\frac{1}{\log n}\right) \to 0$$

Proof ideas

$$F_{\mathsf{TV}}(t) = t\left(1 - 2\mathsf{Q}\left(\sqrt{\frac{P}{t}}\right)\right)$$

Proof.

- Given P_X, Q_X with $\mathsf{TV}(P_X, Q_X) = t$
- Need: upper-bound $\mathsf{TV}(P_X * \mathcal{N}, Q_X * \mathcal{N})$
- First couple P_X to Q_X:

$$\pi[X \neq X'] = \mathsf{TV}(P_X, Q_X) = t$$

- Idea: when $|X - X'| \ll 1 \quad \Rightarrow \quad Y \approx Y'$
  when $|X - X'| \gg 1 \quad \Rightarrow \quad$??? but:
  since $\mathbb{E}|X|^2 \leq P$ this is rare!

□

Details.

- Interplay between Euclidean distance and TV:

$$\theta(x) \triangleq \mathsf{TV}(\mathcal{N}(0,1), \mathcal{N}(x,1))$$

- Next notice:

$$\begin{aligned} \mathsf{TV}(P_X * \mathcal{N}, P_Y * \mathcal{N}) &\le \mathbb{E}[\theta(X - X')] && \text{[TV – Wasserstein distance!]} \\ &= \mathbb{E}[\theta(X - X') | X \neq X'] \cdot t \\ &\le t\theta\left(2\sqrt{\frac{P}{t}}\right) && \text{[θ is concave!]} \end{aligned}$$

- Lucky twice: choice of TV, unimodal noise.
- For the lower bound: take $(1-t)\delta_0 + t\delta_{\pm\sqrt{P/T}}$.

□

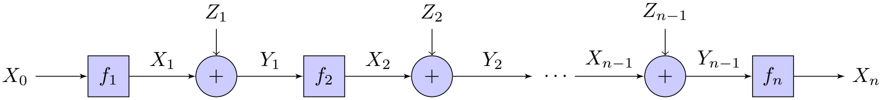
Generalizations

- cost: $\mathbb{E}|X|^2 \leq P$ $\rightarrow$ any convex cost
- noise: Gaussian $\rightarrow$ any unimodal density and more
- channels: scalar-input $\rightarrow$ vector-input (even ∞-dim!)

More than you want to know here:

P. & Yihong Wu (2014). *Dissipation of information in channels with input constraints*. Preprint.

One such generalization



Theorem

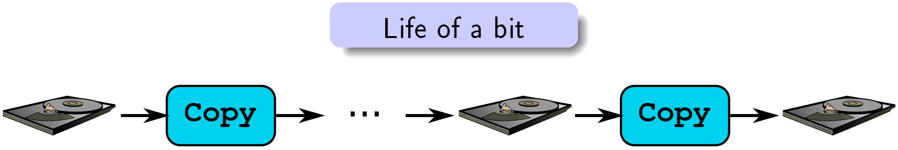
Let X_j, Z_j *be* d*-dimensional,* $d \in \mathbb{N} \cup \{+\infty\}$. *If* $Z_j \sim \mathcal{N}(0, I_d)$ *and*

$$\mathbb{E}\|X_j\|_2^2 \leq E < \infty$$

then

$$I(X_0; X_n) \leq \text{const} \cdot \frac{E \log \log n}{\log n}$$

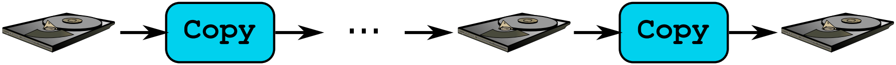
Is digital information immortal?



Copier has finite energy ⇒ information dissipates (slowly!)

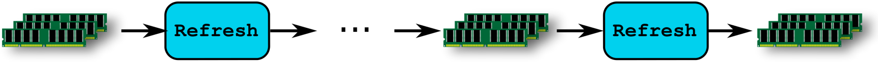
Is digital information immortal?

Life of a bit



Copier has finite energy ⇒ information dissipates (slowly!)

DRAM controller



Does refresh every 64ms!

Application: Uniqueness of Gibbs measures

Game:

$$\cdots - X_{-2} - X_{-1} - X_0 - X_1 - X_2 - \cdots$$

- Given: conditionals $P_{X_j|X_{j-1},X_{j+1}}$
- Find: joint distribution $P_{X_{-\infty}^{+\infty}}$

Some background:

- Shows how local interactions lead to funny global effects.
- Example: multiple solutions correspond to phase-transition (e.g. 2D-Ising)
- Rule of thumb: links are weak (high temp.) $\Rightarrow$ no phase tr.

Application: Uniqueness of Gibbs measures

Dobrushin's method: How to show uniqueness?

- Contrapositive:

$$\cdots - X_{-2} - X_{-1} - X_0 - X_{+1} - X_{+2} - \cdots$$

$$\cdots - \tilde{X}_{-2} - \tilde{X}_{-1} - \tilde{X}_0 - \tilde{X}_{+1} - \tilde{X}_{+2} - \cdots$$

- Couple $X_{\pm 1}$ to $\tilde{X}_{\pm 1}$
- If channel $(X_{-1}, X_{+1}) \mapsto X_0$ is TV-contractive ($\eta_{\mathrm{TV}} < 1$) then improve coupling of X_0 to $\tilde{X}_0$. Repeat.

Application: Uniqueness of Gibbs measures

Dobrushin's method: How to show uniqueness?

- Contrapositive:

$$\cdots - X_{-2} - X_{-1} - X_0 - X_{+1} - X_{+2} - \cdots$$

$$\cdots - \tilde{X}_{-2} - \tilde{X}_{-1} - \tilde{X}_0 - \tilde{X}_{+1} - \tilde{X}_{+2} - \cdots$$

- Couple $X_{\pm 1}$ to $\tilde{X}_{\pm 1}$
- If channel $(X_{-1}, X_{+1}) \mapsto X_0$ is TV-contractive ($\eta_{\mathrm{TV}} < 1$) then improve coupling of X_0 to $\tilde{X}_0$. Repeat.

Our contribution:

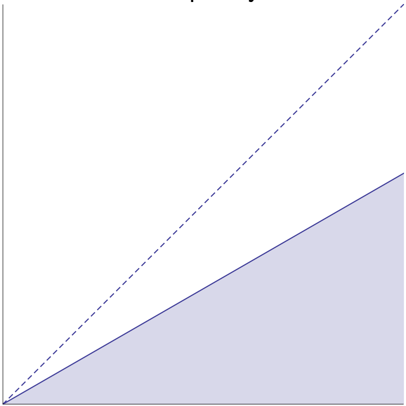
Theorem

Let pairwise potentials $\Phi_j(x_j, x_{j+1})$ be uniformly "lower-bounded". Then there exists at most one Gibbs measure s.t.

$$\sup_k \mathbb{E}|X_k|^2 < \infty$$

Take-away message

linear strong data processing inequality



nonlinear strong data processing inequality