Towards a Unified View of Communication and Control

Sanjoy K. Mitter
MIT

Talk on the occasion of the Sontagfest
May 24, 2011

Slides courtesy of Sekhar Tatikonda

Tatikonda, S. and Mitter, S.K.,
"The Capacity of Channels with Feedback,"
IEEE Trans. on Info. Theory, Vol. 55, January 2009

Agarwal, M., Sahai,A. and Mitter, S.K., “Coding into a source: a direct inverse rate-distortion theorem,” Allerton Conference Proceedings 2006, full journal version is in preparation, 2008.

Borkar, V.S., Mitter, S.K. and Venkatesh, S.R., “Variations on a theme by Neyman and Pearson,” *Sankhya*, Volume 66, Part 2, pp. 292-305, May 2004.

Borkar, V.S., Konda, V.R. and Mitter, S.K., “On De finetti Coherence and Kolmogorov Probability,” *Stat. Prob. Lett.* 66 (2004) pp. 417-421.

Borkar, V.S.,Tatikonda, S. and Mitter, S.K., “Markov Control Problems Under Communication Constraints,” *Comm. Inf. Sys.* Vol. 1., No. 1., pp. 15-32

Borkar, V.S., Mitter, S.K., Sahai, A. and Tatikonda, S., “Sequential Source Coding: An Optimization Viewpoint,” in IEEE Conference Proceedings, CDC-ECC 2005, Seville, Spain.

Mitter, S.K., “Control with Limited Information,” *Eur. Jrn. Control*, Vol. 7, pp. 122-131, December 2000.

Mitter, S.K. and Newton, N., “Variational Bayes and a Problem of Reliable Communication I: Finite Systems,” *Comm. in Inf. and Sys.*, Vol. 10, No. 3, pp. 155-182, 2010.

Mitter, S.K. and Newton, N., “Variational Bayes and a Problem of Reliable Communication II: Infinite Systems,” submitted to *Annals of Applied Probability*, 2010.

Mitter, S.K. and Newton, N., “A Variational Approach to Nonlinear Estimation,” *SIAM Jrn. on Control*, Volume 42, Number 5 (2004), pp. 1813-1833.

Mitter, S.K. and Newton, N.J., “Information and Entropy Flow in the Kalman-Bucy Filter,” *J. of Stat. Phys.,* Vol. 118, Nos. 1/2, January 2005.

Mitter, S.K., and Tatikonda, S., “Control over Noisy Channels,” *IEEE Trans. on Auto. Control*, Vol. 49, July 2004, pp. 1196-1201.

Mitter, S.K., and Tatikonda, S., “Control under Communication Constraints,” *IEEE Trans. on Auto. Control*, Vol. 49, July 2004, pp. 1056-1068.

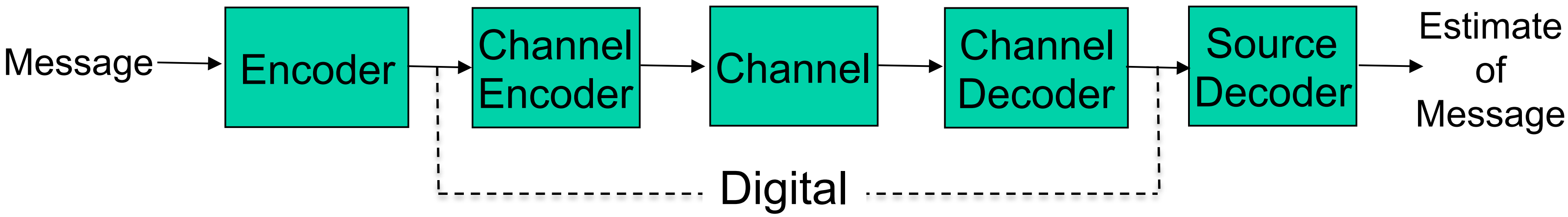
Sahai, A. and Mitter, S.K., “The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link, Part I: Scalar Systems,” *IEEE Trans. on Inf. Theory*, Vol. 52, No. 8, pp. 3369-3395, August 2006.

Sahai, A. and Mitter, S.K., “The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link, Part II: Vector Systems,” under revision, *IEEE Trans. on Inf. Theory*, April 2007.

Tatikonda, S. and Mitter, S.K. “The Capacity of Channels with Feedback,” *IEEE Trans. on Info. Theory*, Vol. 55, p. 323-349, January 2009.

Tatikonda, S., Sahai, A. and Mitter, S.K., “Stochastic Linear Control Over a Communication Channel,” *IEEE Trans. on Auto Control,* Special Issue on Networked Control Systems, Vol.49, Sept. 2004, pp. 1549-1561.

Shannon Picture



Source Encoder: Compression
Channel Encoder: Expansive Map

- Source Coding Theorem
- Noisy Channel Coding Theorem
- Source Channel Separation
- Can recover message with probability of error going to zero (block length $\longrightarrow \infty$), provided

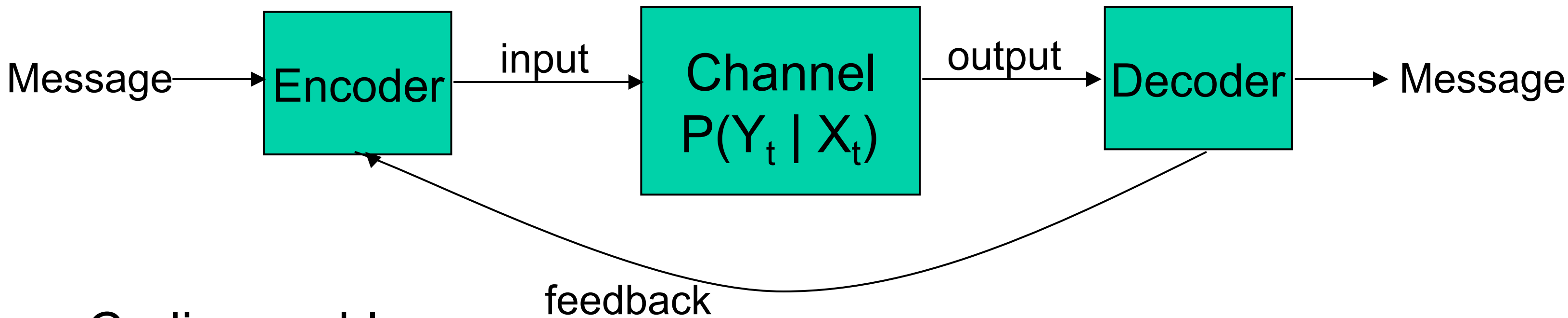
$$R < C, \quad C = \text{capacity}$$

- $R > C$, cannot recover message with probability of error $\longrightarrow 0$
- Separation architecture

Feedback Communication Problems

- Feedback from output of the channel to input of the channel encoder
- Causality
- Can we use the language and machinery of ergodic partially observed stochastic control to understand fundamental limitations of feedback communication problems?

Review of Basic Set-Up



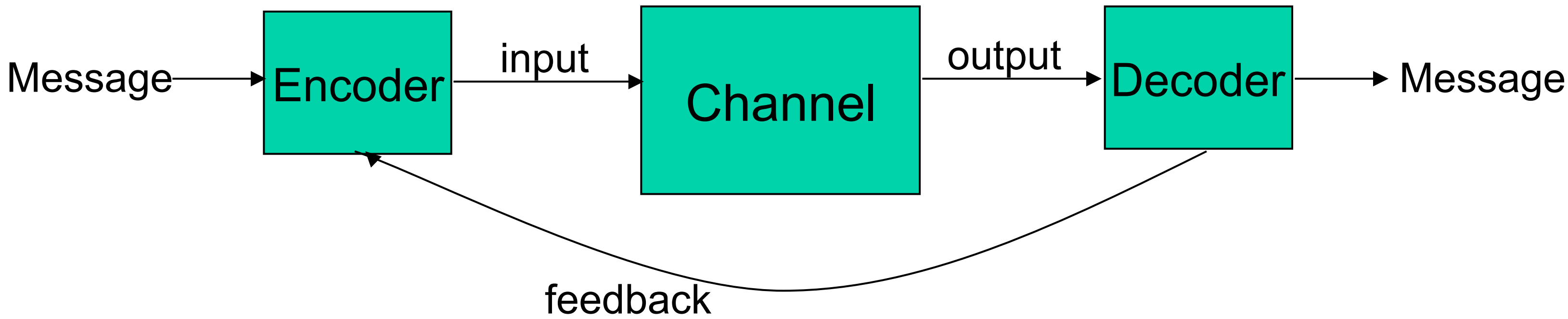
- Coding problem:
  - memoryless channel: $P(Y_t \mid X_t)$
  - messages: { m }
  - codewords: encoder maps $m \rightarrow X^T = (X_1,\ldots,X_T)$
  - decoder: $Y^T \rightarrow \hat{m}$

- Objective: maximize number of messages subject to small error probability: $P(m \neq \hat{m})$

- Shannon's theorem: $C = \max_{P(X)} I(X; Y)$
  - single letter characterization

- If no memory (or state) then feedback does not increase capacity

Time-Ordering of Events and Code-Functions



- With feedback one can adapt the channel input symbol. Use code-functions: $X_t = F_t(Y^{t-1})$

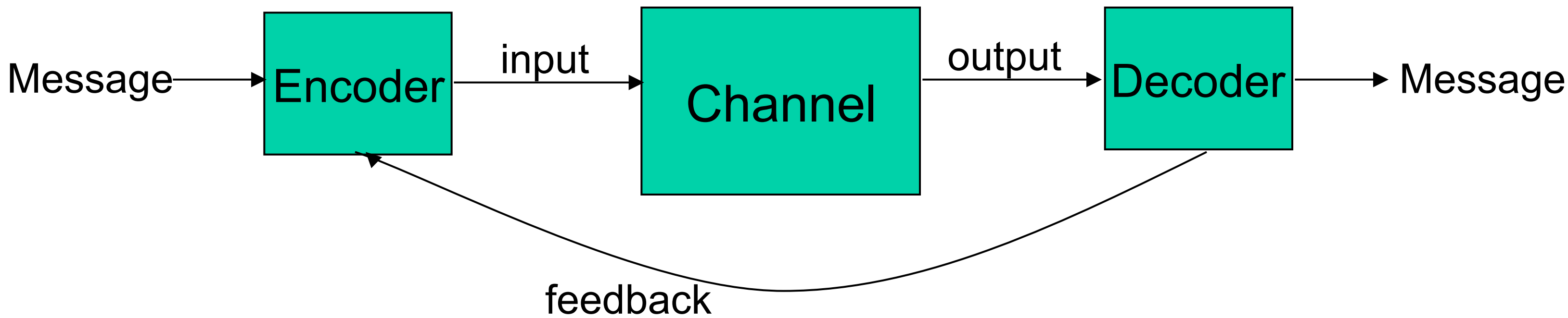
- Time order of event.

  non-causal: $M, X_1, X_2, \ldots, X_T, Y_1, Y_2, \ldots, Y_T, M$

  causal: $M, X_1, Y_1, X_2, Y_2, \ldots, X_T, Y_T, M$

- Every message is assigned a sequence of code-functions $F_1, F_2, \ldots, F_T$ (as opposed to a codeword.)

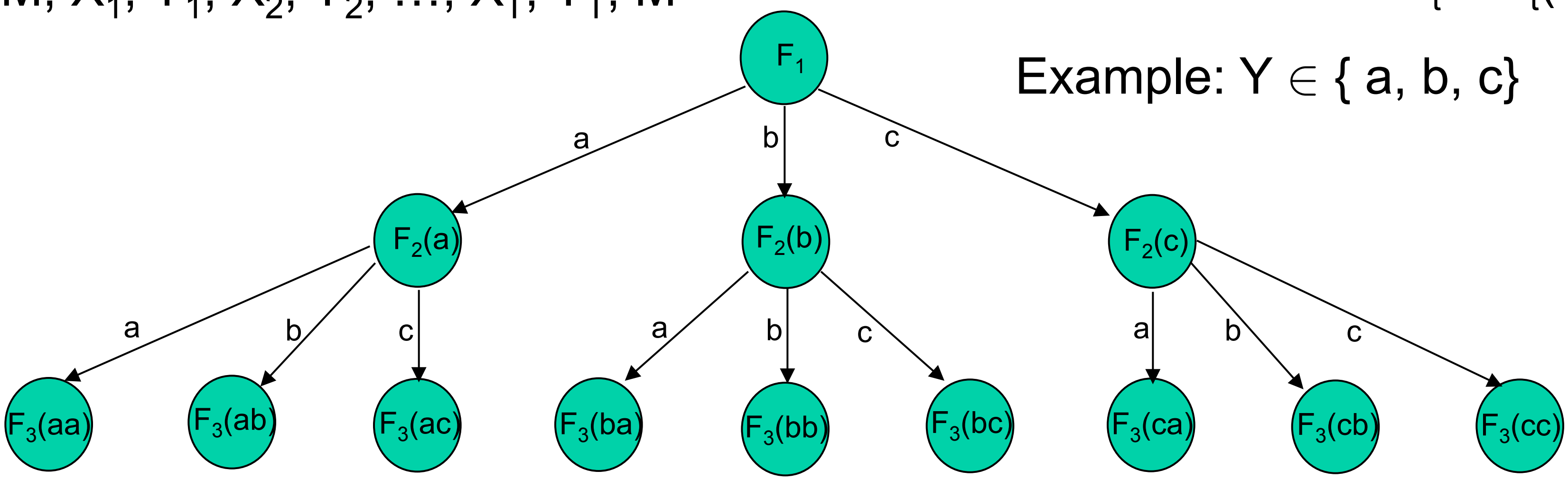
Code-Functions



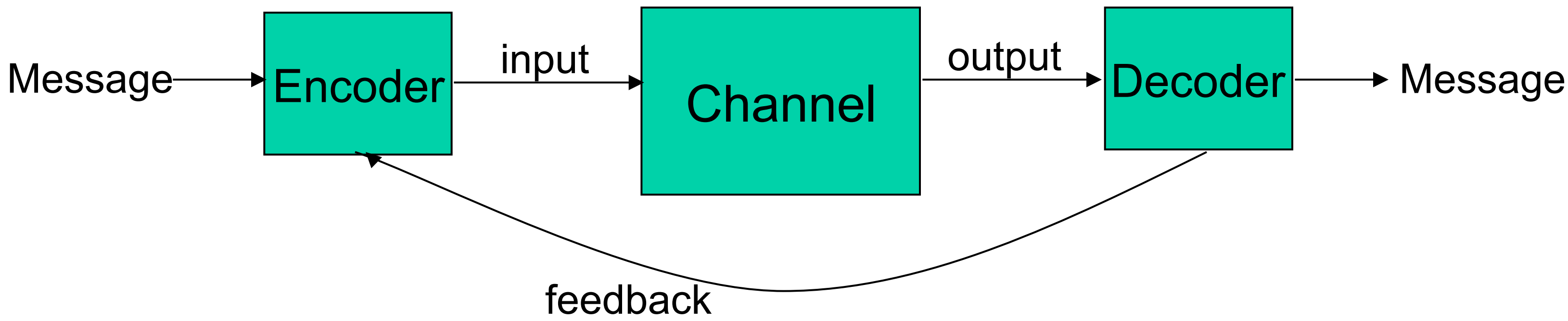
Time order:
$M, X_1, Y_1, X_2, Y_2, \ldots, X_T, Y_T, M$

Code-functions: $X_t = F_t(Y^{t-1})$

Example: $Y \in \{ a, b, c\}$



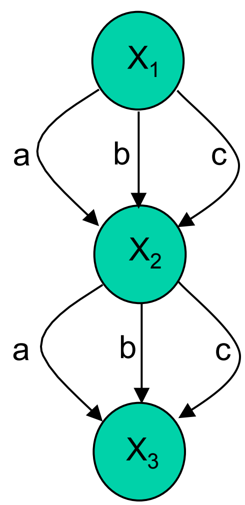
Codewords versus Code-Functions

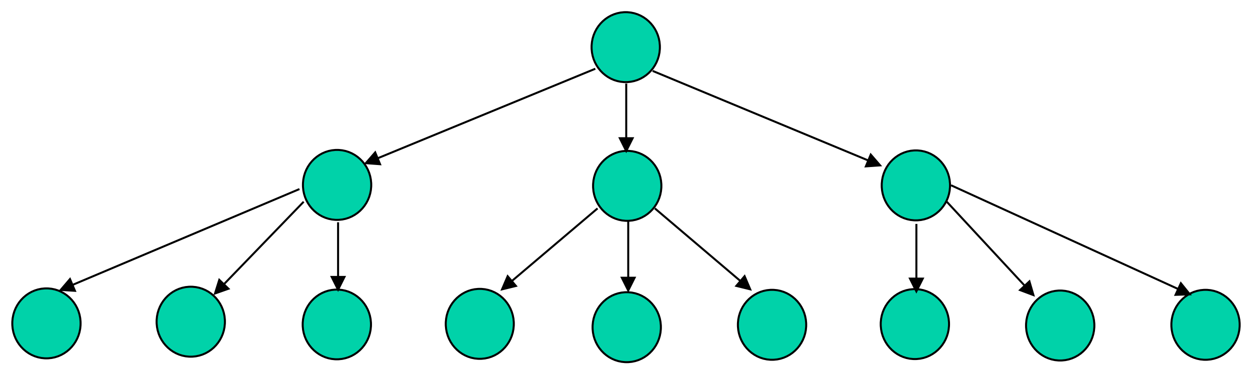


A codeword is a code-function that does not care about the feedback.

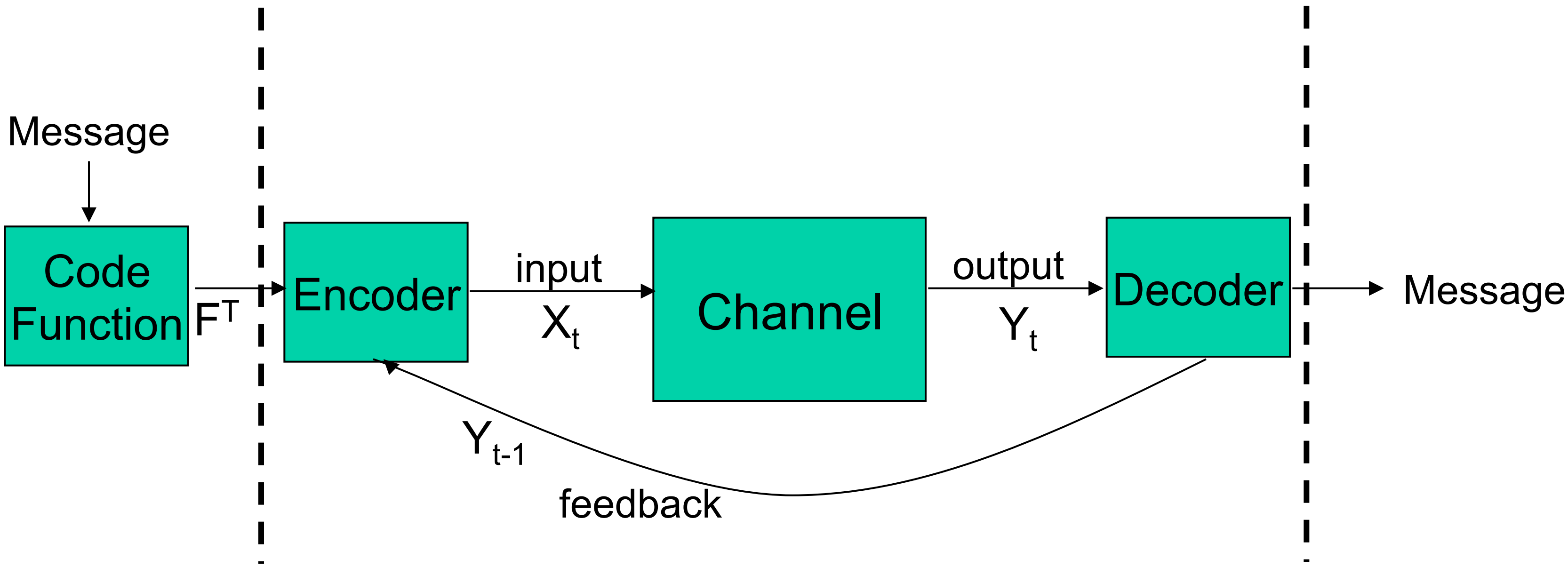
Code-functions: $X_t = F_t(Y^{t-1})$

Example: $Y \in \{ a, b, c\}$



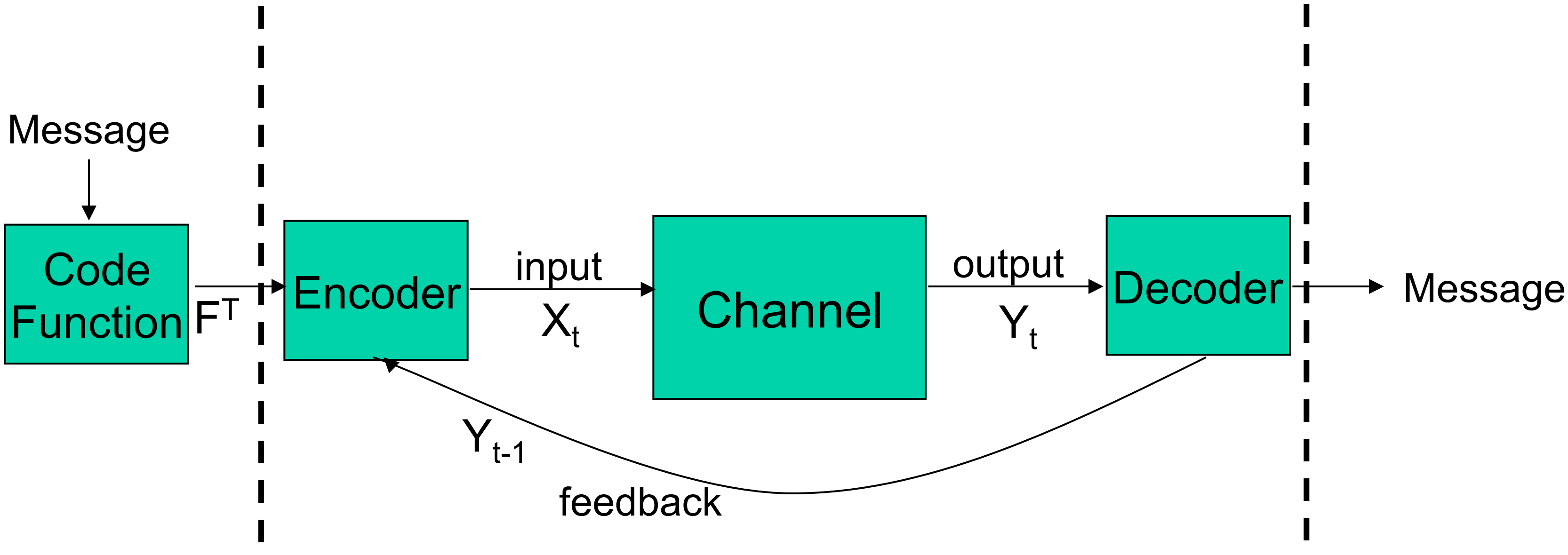


Converting a Feedback Channel to a Non-Feedback Channel



- A general *causal* channel has the form: $P(Y_t \mid X^t, Y^{t-1})$
- $P(Y_t \mid F^t, Y^{t-1}) = P(Y_t \mid X_1 = F_1, X_2 = F_2(Y_1), \ldots, X_t = F_t(Y^{t-1}), Y^{t-1})$
- The "F – Y" channel has no feedback

Converting a Feedback Channel to a Non-Feedback Channel – Part 2

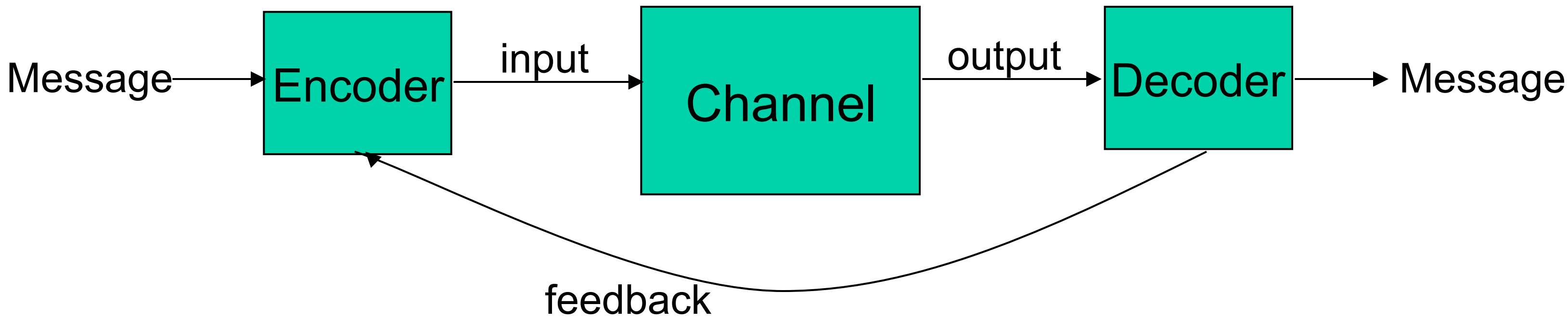


- Without feedback: $C = \lim_{T \to \infty} \max_{P(X^T)} 1/T \; I(X^T; Y^T)$

  With feedback: $C = \lim_{T \to \infty} \max_{P(F^T)} 1/T \; I(F^T; Y^T)$
  (both modulo information stability issues)

- Difficult optimization which does not give us much insight

Directed Information



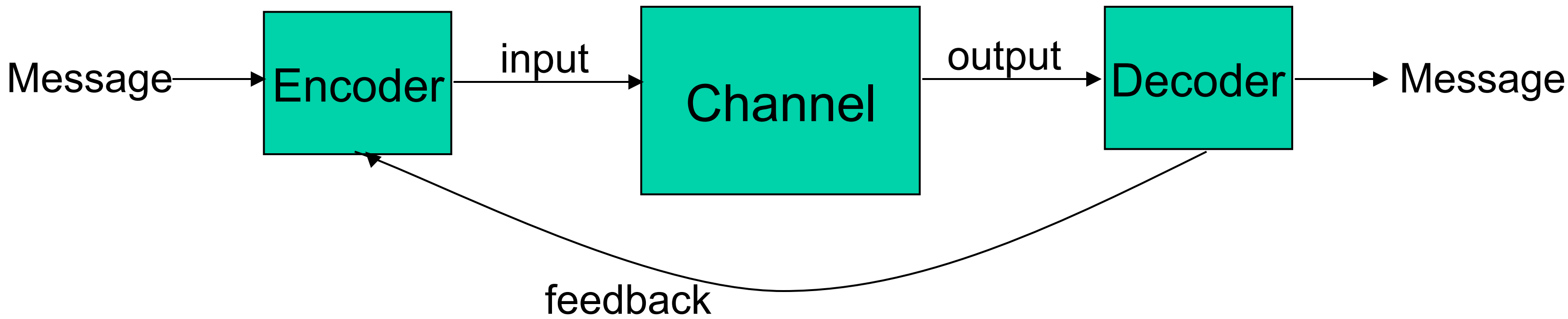
- $I(X^T; Y^T) = H(Y^T) - \sum_t H(Y_t \mid X^T, Y^{t-1}) = \sum_t I(X^T; Y_t \mid Y^{t-1})$

  - note dependence on future X's
  - Massey: "statistical dependence, unlike causality, has no inherent directivity."

- Massey's directed information, further developed by Kramer:

$$I(X^T \rightarrow Y^T) = H(Y^T) - \sum_t H(Y_t \mid X^t, Y^{t-1}) = \sum_t I(X^t; Y_t \mid Y^{t-1})$$

Directed Information – Part 2

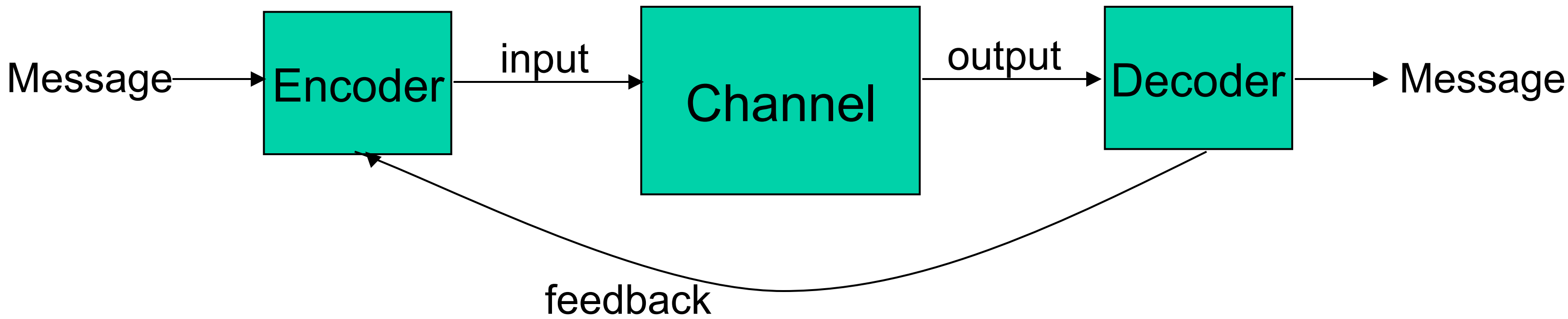


- $I(X^T \rightarrow Y^T) = H(Y^T) - \sum_t H(Y_t \mid X^t, Y^{t-1})$

- Consider example from before: $P(Y_t \mid X_t) = P(Y_t)$
  with feedback coding $X_t = Y_{t-1}$
  then $I(X^T \rightarrow Y^T) = H(Y^T) - \sum_t H(Y_t \mid X^t, Y^{t-1})$
  $= H(Y^T) - \sum_t H(Y_t)$
  $= 0$

- By DPI: $I(M; Y^T) = I(F^T; Y^T) \leq I(X^T; Y^T)$
  Message is not in one-to-one relation with X^T

Directed Information – Part 3



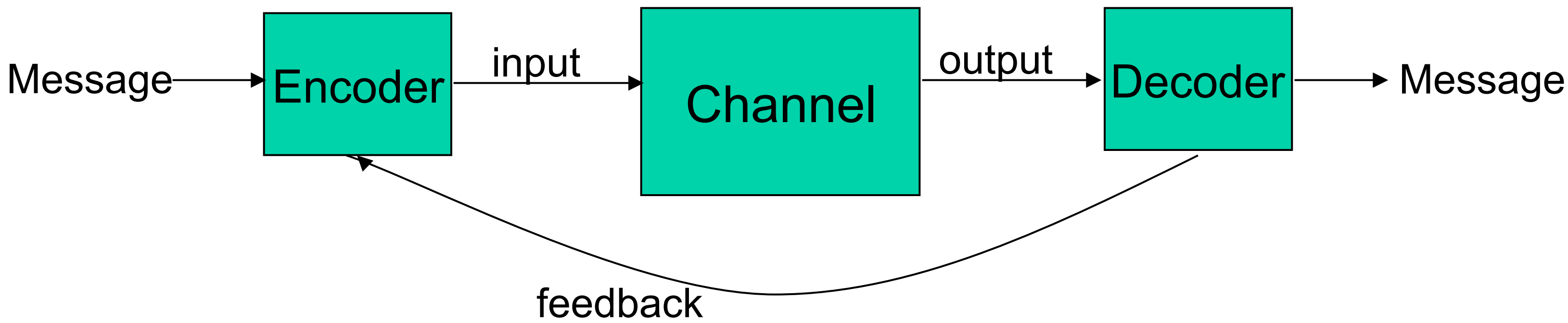
Massey showed:

- $I(X^T; Y^T) = I(X^T \to Y^T) + I(Y^T \to X^T)$
  conservation of information

- $I(X^T; Y^T) \geq I(X^T \to T^T)$
  with equality if and only if there is no feedback

- Directed information preserves causality, it is not symmetric

Information Spectrum Approach



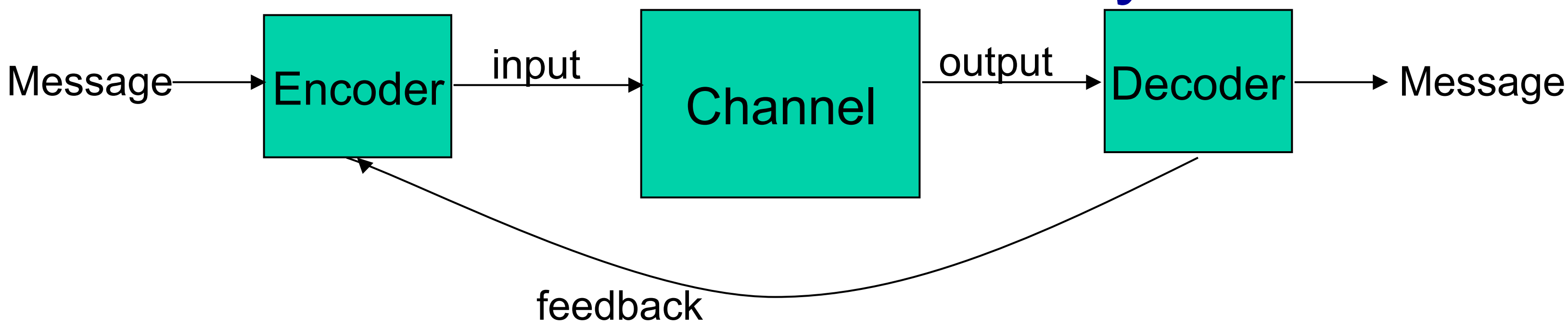
- $I(F^T; Y^T) = I(F^T \to Y^T)$
  $= H(Y^T) - \sum_t H(Y_t \mid F^t, Y^{t-1})$
  $= H(Y^T) - \sum_t H(Y_t \mid X^t, Y^{t-1})$
  $= I(X^T \to Y^T)$

- Coding theorem [Tat00] (based on Verdu/Han):

$$C = \sup_{\{P(X_t \mid X^{t-1}, Y^{t-1})\}} \text{ liminf in prob } 1/T\; i(X^T \to Y^T)$$

- Full-fledged coding theorem: both direct and converse; explicit construction of code-function distribution; and error exponent analysis. Can allow for arbitrary causal, deterministic feedback.

Information Stability

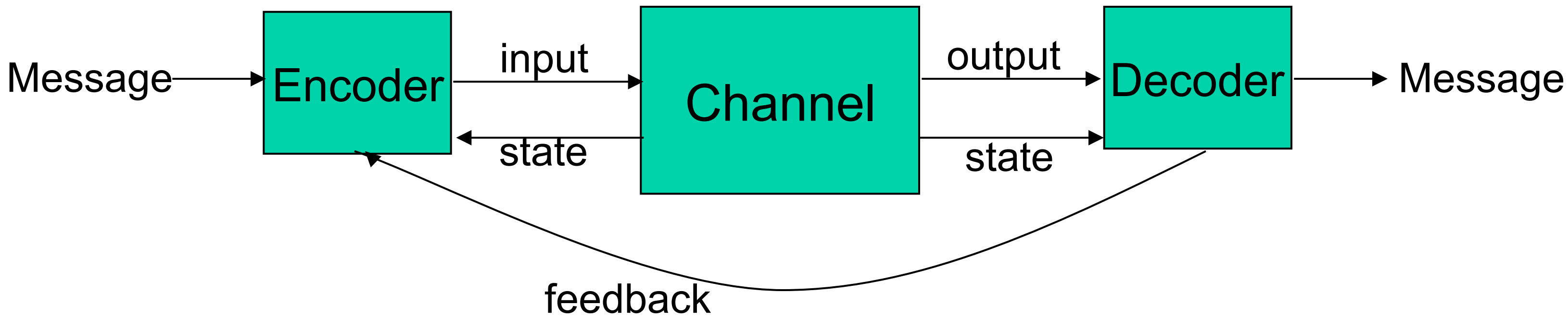


- How to compute!
- Two problems:
  - not information stable
  - not a single-letter characterization

- [Tat00] proposed an approach that *simultaneously* insured information stability and a single-letter characterization.

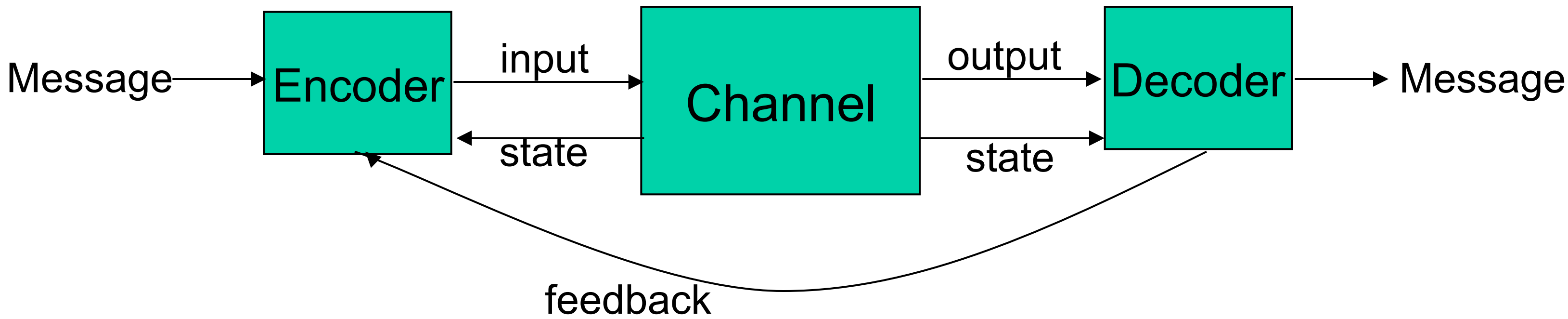
- Idea: Formulate optimization as an infinite horizon average cost Markov decision control problem.

Channels with Memory



- How to model memory in channels?
- General causal channel: $P(Y_t \mid X^t, Y^{t-1})$
- Need a compact representation
- Need stationarity (if we hope to calculate anything)

Markov Channel



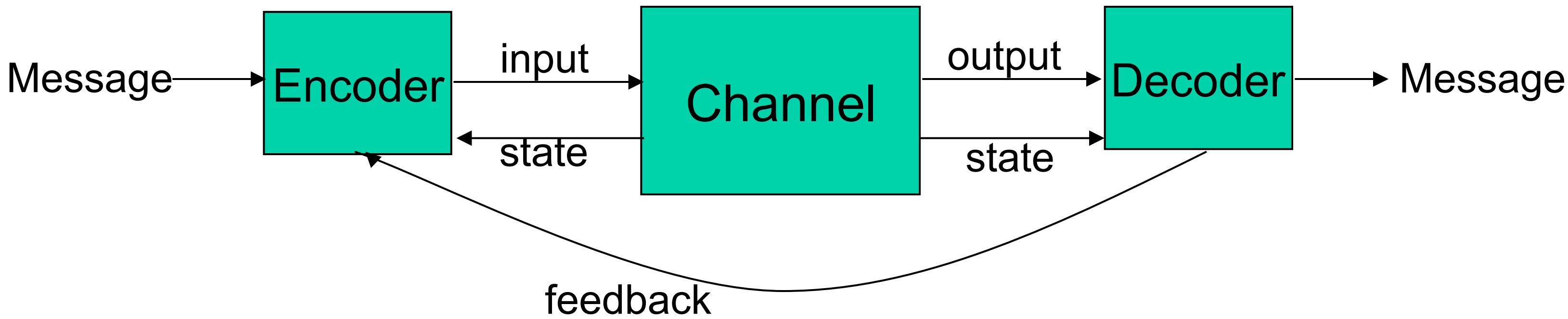
- A Markov Channel consists of: $P(S_1)$

  (1) state transition: $P(S_{t+1} \mid S_t, X_t)$

  (2) channel output: $P(Y_t \mid S_t, X_t)$

- If $P(S_{t+1} \mid S_t, X_t)$ is independent of X_t then non-ISI channel
- Assume time invariant (stationary)
- Time order: $M, S_1, X_1, Y_1, S_2, X_2, Y_2, \ldots, S_T, X_T, Y_T, M$

Side-Information versus Feedback



- Time order: M, S_1, X_1, Y_1, S_2, X_2, Y_2, …, S_T, X_T, Y_T, M

- Information pattern:
  - Tx: subset of (S^t, X^{t-1}, Y^{t-1})
  - Rx: subset of (S^t, Y^{t-1})
  - Nested information patterns

- Code-functions: $X_t = f_t(S^t, Y^{t-1})$
  (don't need X^{t-1})

- What is the difference between side-information and feedback?

A Quick Review of Dynamic Programming

- Sequential optimization.
  - State: S_t
  - Action: U_t
  - Dynamics: $P(S_{t+1} \mid S_t, U_t)$
  - Action:
  - Policy: $S_t \mapsto U_t$
  - Running cost: $c(S_t, U_t)$

- Infinite horizon average cost problem:

$$\sup \liminf 1/T \, E[\textstyle\sum_t c(S_t, U_t)]$$

- ACOE: Find J and w(s) that solve:

$$J + w(s) = \max_u c(s, u) + \textstyle\sum_{s+} P(s_+ \mid s,u) \, w(s_+)$$

then J is optimal cost and u*(s) is optimal policy.

Dynamic Programming Formulation

- ISI Markov Channel: $P(S_1)$, $P(S_{t+1} \mid S_t, X_t)$, $P(Y_t \mid S_t, X_t)$
- $I(X^T \rightarrow Y^T, S^T) = \sum_t I(X_t; Y_t, S_{t+1} \mid S_t)$
- $I(X_t; Y_t, S_{t+1} \mid S_t) = I(X_t; Y_t \mid S_t) + I(X_t; S_{t+1} \mid Y_t, S_t)$

- Dynamic programming framework:
  - state: S_t
  - action: $P(X_t)$
  - policy: $S_t \mapsto P(X_t)$, i.e. $P(X_t \mid S_t)$
  - running cost: $I(X_t; Y_t, S_{t+1} \mid S_t = s_t) = c(\, P(X_t), s_t)$

- Infinite horizon average cost problem:

$$\sup \liminf 1/T \, E[\, \textstyle\sum_t c(\, P(X_t), S_t \,)]$$

DP Formulation – Part 2

- Markov Channel: $P(S_1)$, $P(S_{t+1} \mid S_t, X_t)$, $P(Y_t \mid S_t, X_t)$
- Running cost: $c(P(X_t), s_t) = I(X_t; Y_t, S_{t+1} \mid S_t = s_t)$
- Infinite horizon average cost problem:

$$\sup \liminf 1/T\, E[\textstyle\sum_t c(P(X_t), S_t)]$$

- ACOE [Tat00]: If there exists a C and a w(S) such that $\forall$ s:

$$C + w(s) = \max_{P(X)}\{ I(X; Y, S_+ \mid s) + \textstyle\sum_{x,s_+} w(s_+)\, P(s_+ \mid x, s)\, P(x) \}$$

then C is the capacity of the ISI Markov channel. P(X|S) optimal input distribution.

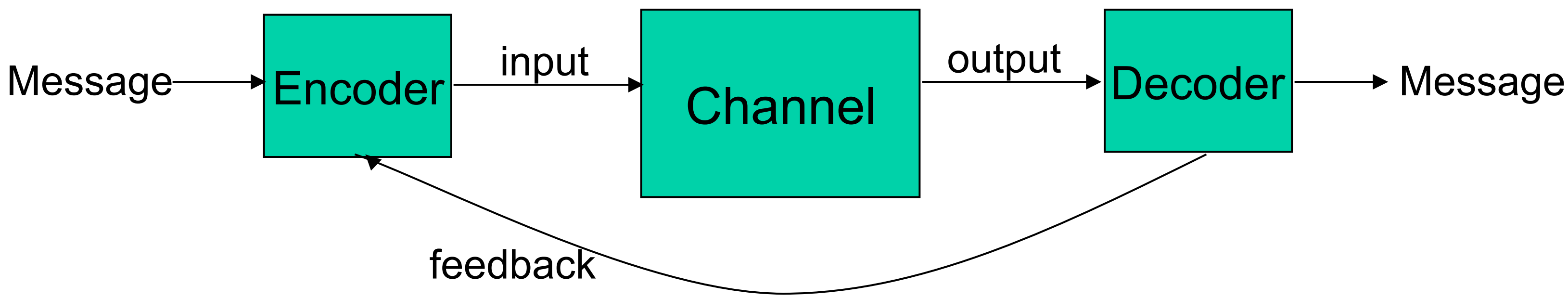
- Remarks:
  - Implicit single-letter characterization
  - If no ISI, ACOE becomes trivial
  - Multiplex between different codebooks indexed by S

DP Formulation – Part 3

- $C + w(s) = \max_{P(X)}\{ I(X; Y, S_+ \mid s) + \sum_{x,s_+} w(s_+) P(s_+ \mid x, s) P(x \mid s) \}$
- When does a solution to the ACOE exist? We need to insure ergodicity under the optimal policy.
- One sufficient condition is:
  $\| P(S_+ \mid S=s_1, X=x_1) - P(S_+ \mid S=s_2, X=x_2) \|_{TV} < 1 \quad \forall\, s_1, s_2, x_1, x_2$
- example: $P(s_+ \mid s, x) > 0 \quad \forall\, s, x, s_+$
- Related to Gallager's indecomposability:
  $\exists\, t$ s.t. $\| P(S_{t+1} \mid X^t=x^t, S_1=a) - P(S_{t+1} \mid X^t=x^t, S_1 = b) \|_{TV} < 1 \;\forall\, x^t, a, b$
  (in our setting this depends on the policy)
- Our sufficient condition insures that under any $P(X|S)$ the closed loop dynamic has a unique ergodic measure.

$$P(S_{t+1} \mid S_t) = \sum_x P(S_{t+1} \mid S_t, x) P(x \mid S_t)$$

Markov Channel with Output Feedback



- Markov channel: $P(S_1)$, $P(S_{t+1} \mid S_t, X_t)$, $P(Y_t \mid S_t, X_t)$.

- Now assume the state is not observed by either Tx or Rx. There is only output feedback. (Recall if state is known output feedback will not increase capacity.)

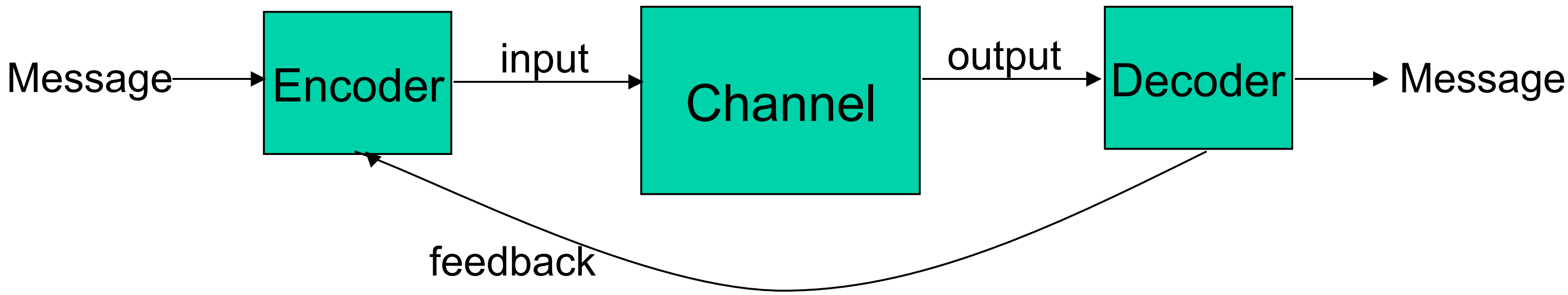
- At the beginning of the *t*-th epoch the

  Tx knows (X^{t-1}, Y^{t-1}) and Rx knows Y^{t-1}

  Note that the Rx's information pattern is nested in the Tx's information pattern. Find sufficient statistics (before it was S_t.)

- Input distribution has the form: $P(X_t \mid X^{t-1}, Y^{t-1})$

Encoder's Estimate: Π



- Use output feedback to estimate state at the encoder:

$$\Pi_t[X^{t-1}, Y^{t-1}] = P(S_t \mid X^{t-1}, Y^{t-1})$$

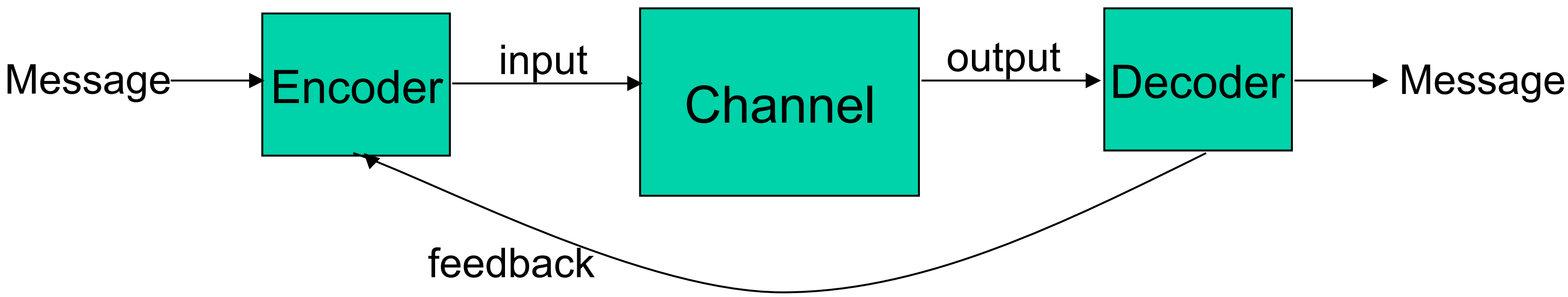
- There exists a policy independent function Φ_Π such that

$$\Pi_{t+1} = \Phi_\Pi(\Pi_t, X_t, Y_t).$$

This can be computed recursively at the Tx.

- Note that the statistic Π depends on information from both the Tx and the Rx

The Π Process and ISI



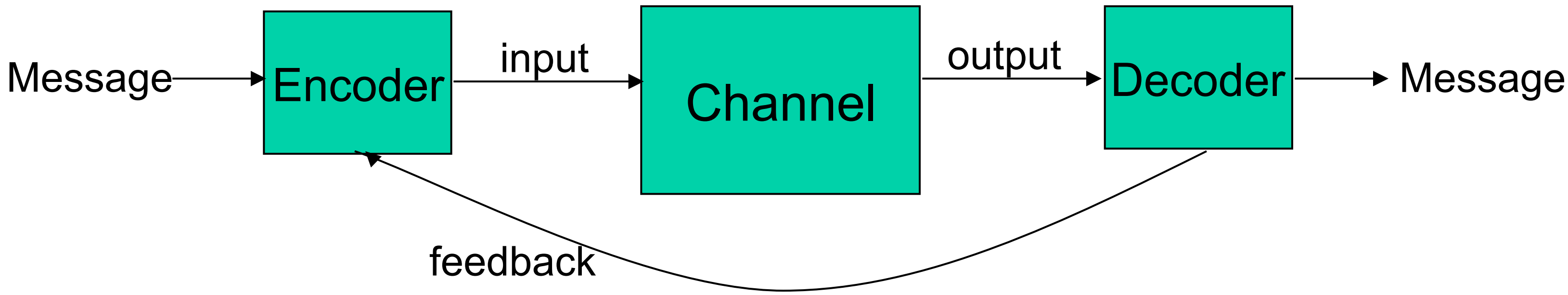
- $I(X^T \to Y^T) = H(Y^T) - \sum_t H(Y_t \mid X^t, Y^{t-1})$
  $= H(Y^T) - \sum_t H(Y_t \mid X_t, \Pi_t[X^{t-1}, Y^{t-1}])$
  $= \sum_t I(X_t, \Pi_t; Y_t \mid Y^{t-1})$

- Think of the pair (X_t, Π_t) as the input. The Rx does not know Π_t.

- Issue of dual effect. Even if underlying channel, $P(S_{t+1} \mid S_t)$, does not have ISI it is generically the case that the corresponding Π_t process *does* depend on the inputs:

$$P(\Pi_{t+1} \mid \Pi_t, X_t) = \sum_{s,y} \{\Pi_{t+1} = \Phi_\Pi(\Pi_t, X_t, y)\} P(y \mid s, X_t) \Pi_t(s)$$

Decoder's Estimate: Γ



- Tx: $(X^{t-1}, Y^{t-1}, \Pi_t)$, Rx: Y^{t-1}
- Goal: supremize $1/T \sum_t I(X_t, \Pi_t ; Y_t \mid Y^{t-1})$
- Rx needs estimate of Tx's estimate of the state:

$$\Gamma_t[Y^{t-1}] = P(\Pi_t \mid Y^{t-1})$$

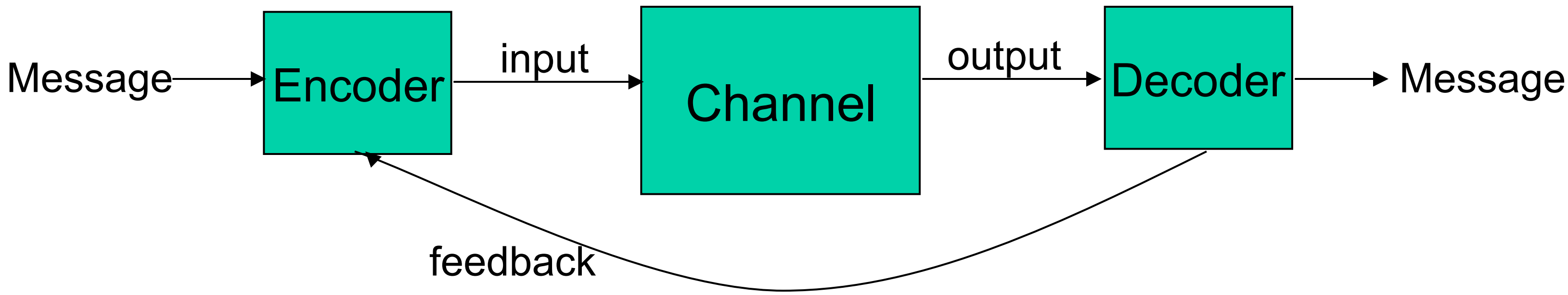
- note: not $P(S_t \mid Y^{t-1})$

- There exists a policy independent function Φ_Γ such that

$$\Gamma_{t+1} = \Phi_\Gamma(\Gamma_t, Y_t).$$

This can be computed at both the Tx and Rx.

Sufficient Statistics at Tx and Rx



- Before: Tx: (X^{t-1}, Y^{t-1}), Rx: Y^{t-1}

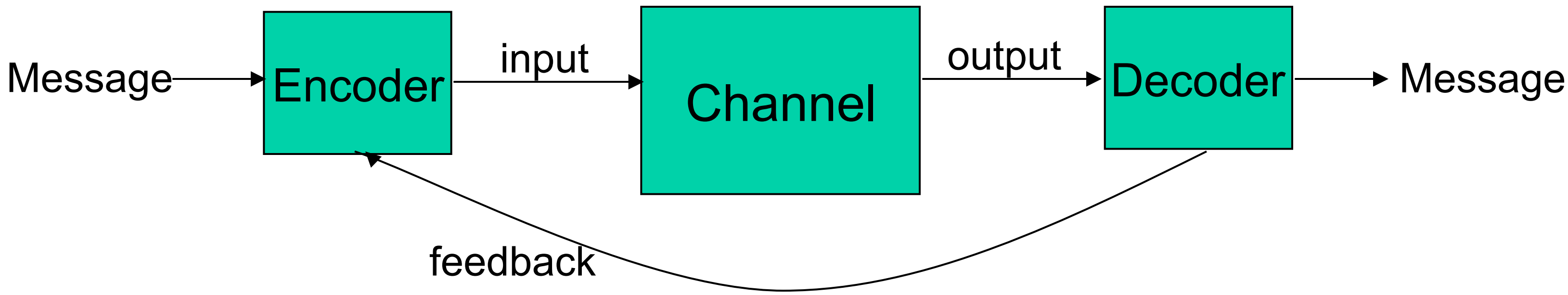
  Now: Tx: (Π_t, Γ_t), Rx: Γ_t

- Before: supremize $1/T \sum_t I(X^t ; Y_t \mid Y^{t-1})$

  Now: supremize $1/T \sum_t I(X_t, \Pi_t; Y_t \mid \Gamma_t)$

- Separation structure between estimation and coding.
Great simplification (though still complicated….)

ACOE



• Theorem [Tat05]: If there exists a bounded number C, a bounded function w: $\Gamma \mapsto R$, and a policy achieving the supremum for each $\Gamma = \gamma$ in the following ACOE:

$$C + w(\gamma) = \sup_{P(X,\,\Pi)} \left(I(X, \Pi; Y \mid \gamma) + \int w(\Gamma_+)\, P(d\Gamma_+ \mid \gamma, P(X, \Pi)) \right)$$

Then C is the capacity.

• Verification Theorem

In control problems with extensive sensing (vision sensor in feedback loop), control needs to act on “information” instead of signals.

Natural role for coding and decoding.