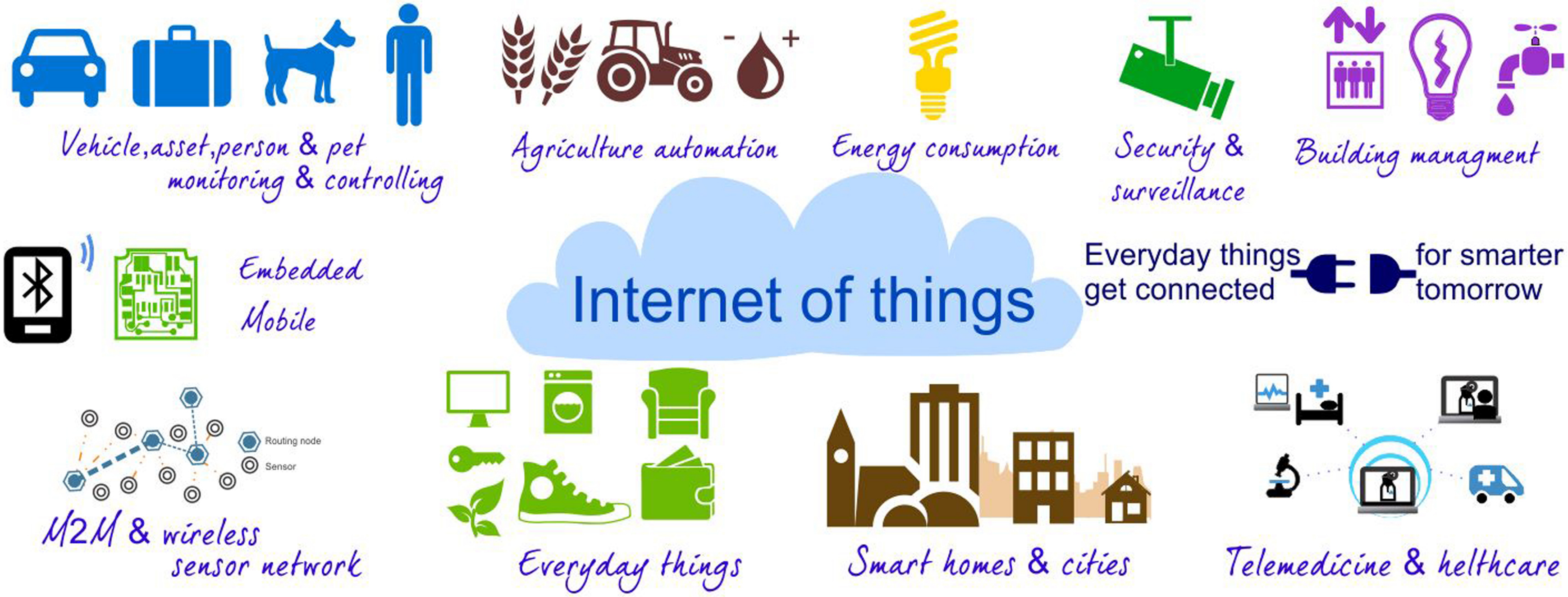
Fundamentals for Low Latency Communications

H. Vincent Poor

Princeton University

Supported in part by NSF Grants CCF-093970 and CCF-1513915.

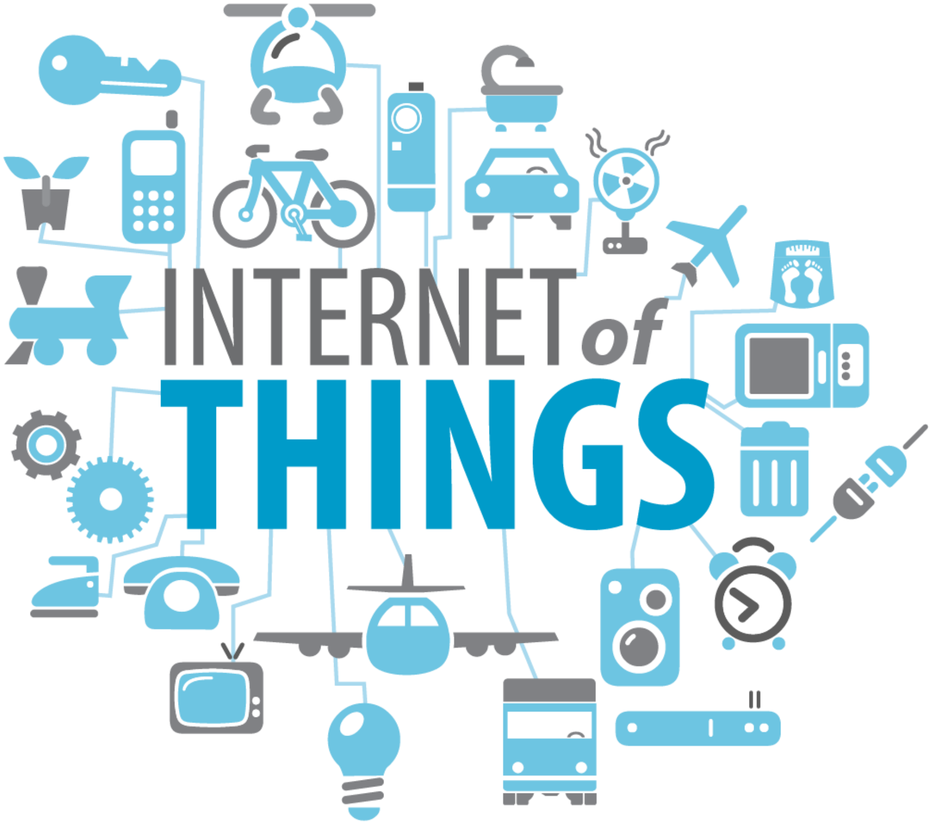
The Internet-of-Things (IoT) Vision



- Interconnecting perhaps 100s of billions of devices
- Key enabler: wireless communications

Salient Characteristics of IoT

- Massive connectivity
- High energy efficiency
- Low complexity
- High reliability
- Short packets
- Low latency



Requirements for URLLC in 5G [3GPP TS22.261]

Table 7.2.2-1 Performance requirements for low-latency and high-reliability scenarios.

Scenario	End-to-end latency (note 3)	Jitter	Survival time	Communication service availability (note 4)	Reliability (note 4)	User experienced data rate	Payload size (note 5)	Traffic density (note 6)	Connection density (note 7)	Service area dimension (note 8)
Discrete automation – motion control (note 1)	1 ms	1 µs	0 ms	99,9999%	99,9999%	1 Mbps up to 10 Mbps	Small	1 Tbps/km^2	100 000/km^2	100 x 100 x 30 m
Discrete automation	10 ms	100 µs	0 ms	99,99%	99,99%	10 Mbps	Small to big	1 Tbps/km^2	100 000/km^2	1000 x 1000 x 30 m
Process automation – remote control	50 ms	20 ms	100 ms	99,9999%	99,9999%	1 Mbps up to 100 Mbps	Small to big	100 Gbps/km^2	1 000/km^2	300 x 300 x 50 m
Process automation – monitoring	50 ms	20 ms	100 ms	99,9%	99,9%	1 Mbps	Small	10 Gbps/km^2	10 000/km^2	300 x 300 x 50
Electricity distribution – medium voltage	25 ms	25 ms	25 ms	99,9%	99,9%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	100 km along power line
Electricity distribution – high voltage (note 2)	5 ms	1 ms	10 ms	99,9999%	99,9999%	10 Mbps	Small	100 Gbps/km^2	1 000/km^2 (note 9)	200 km along power line
Intelligent transport systems – infrastructure backhaul	10 ms	20 ms	100 ms	99,9999%	99,9999%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	2 km along a road
Tactile interaction (note 1)	0,5 ms	TBC	TBC	[99,999%]	[99,999%]	[Low]	[Small]	[Low]	[Low]	TBC
Remote control	[5 ms]	TBC	TBC	[99,999%]	[99,999%]	[From low to 10 Mbps]	[Small to big]	[Low]	[Low]	TBC

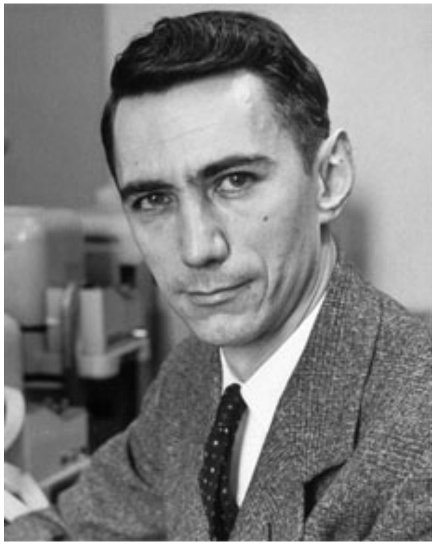
NOTE 1: Traffic prioritization and hosting services close to the end-user may be helpful in reaching the lowest latency values.
NOTE 2: Currently realised via wired communication lines.
NOTE 3: This is the end-to-end latency the service requires. The end-to-end latency is not completely allocated to the 5G system in case other networks are in the communication path.
NOTE 4: Communication service availability relates to the service interfaces, reliability relates to a given node. Reliability should be equal or higher than communication service availability.
NOTE 5: Small: payload typically ≤ 256 bytes
NOTE 6: Based on the assumption that all connected applications within the service volume require the user experienced data rate.
NOTE 7: Under the assumption of 100% 5G penetration.
NOTE 8 Estimates of maximum dimensions; the last figure is the vertical dimension.
NOTE 9: In dense urban areas.
NOTE 10: All the values in this table are targeted values and not strict requirements.

Talk Outline

- Traditional information theory - asymptotic performance
- Basics of finite blocklength information theory: point-to-point
- Multipoint - network models
- Age-of-Information (briefly)
- Conclusion

Traditional Information Theory

Traditional Information Theory

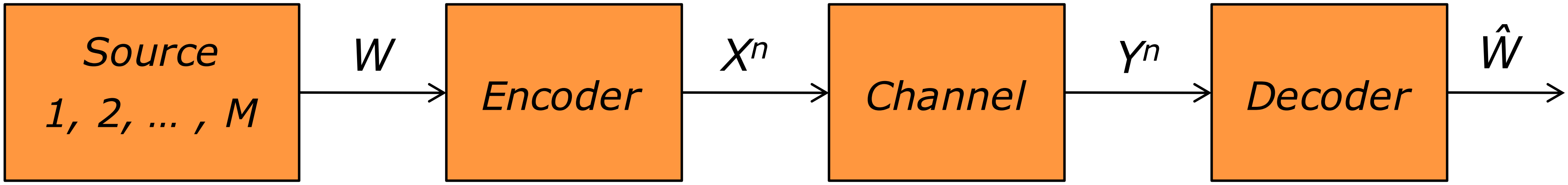


- Point-to-point: Shannon's pioneering work
  - Data transmission
  - Data compression
- Network information theory
  - Broadcast, multiple access, relay
  - Secure transmission (wiretap), secure compression
- Asymptotic: characterizes fundamental limits when delay is unimportant

- Benefit: characterizes operational, engineering problems in terms of elegant mathematical formulas

- Limitation: not suitable for low-latency applications as in IoT

Data Transmission

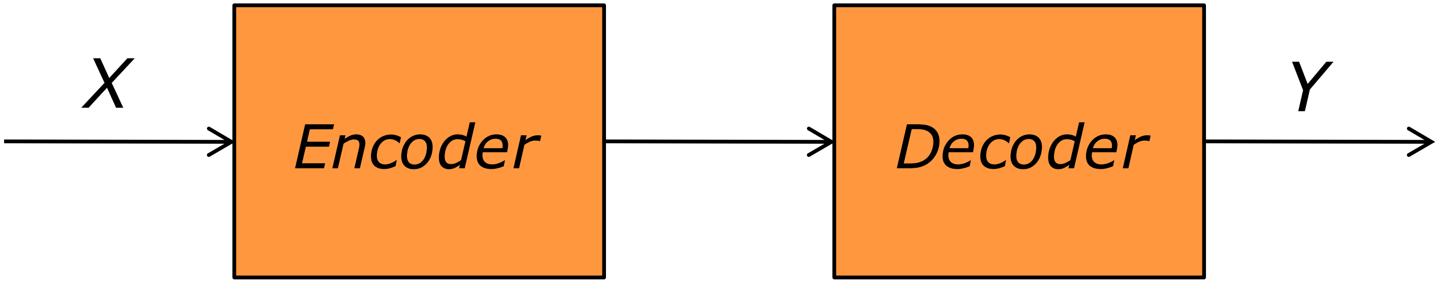


Capacity: largest rate in the asymptotic regime of

- Blocklength $n \to \infty$
- Probability of error $\mathbb{P}\left(W \neq \hat{W}\right) \to 0$

$$C = \max_{P_X} I(X;Y)$$

Data Compression



- Entropy: smallest asymptotic rate for lossless compression

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

- Rate-distortion function: smallest asymptotic rate for lossy compression

$$R(d) = \min_{P_{Y|X} \,:\, \mathbb{E}[d(X,Y)] \leq d} I(X;Y)$$

Basics of Finite-Blocklength Information Theory

How do we characterize non-asymptotic limits?

- Data transmission: the information density

$$\imath_{X;Y}(x;y) = \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

- Expectation of the information density is mutual information

$$I(X;Y) = \mathbb{E}\left[\imath_{X;Y}(X;Y)\right]$$

- Similarly, for lossless and lossy compression: information and d-tilted information, respectively

Example: Data Transmission

- Upper (achievability) bound:

smallest error for a code of length n and M codewords

information density

$$\epsilon^*(M,n) \leq \inf_{P_X} \mathbb{P}\left[\imath_{X^n;Y^n}(X^n;Y^n) \leq \log M + n\gamma\right] + \exp(-n\gamma)$$

"fudge" parameter, can be any positive number

- Lower (converse) bound:

$$\epsilon^*(M,n) \geq \inf_{P_X} \mathbb{P}\left[\imath_{X^n;Y^n}(X^n;Y^n) \leq \log M - n\gamma\right] - \exp(-n\gamma)$$

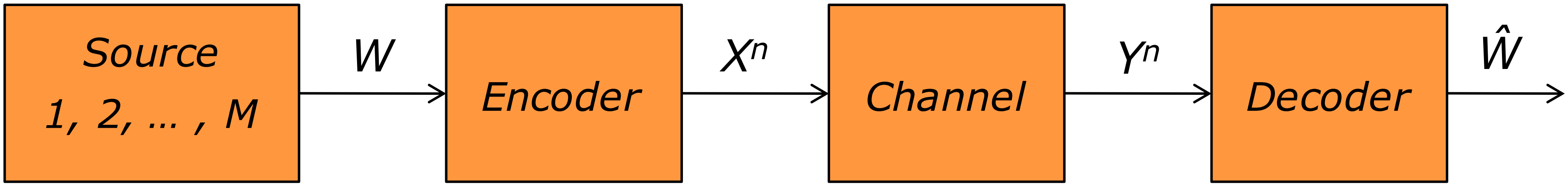
Non-asymptotic Information Theory

- Non-asymptotic fundamental limits are characterized by information density (data transmission), information (lossless compression), etc.

- Good upper (converse) and lower (achievability) bounds

- Refined asymptotic limits: better characterize fundamental limits when delay is important

Data Transmission Revisited



Source

1, 2, … , M

W

Encoder

X^n

Channel

Y^n

Decoder

$\hat{W}$

(n,M,ε) code: *P(W≠Ŵ)* ≤ ε

Fundamental limit: *M*(n,ε)* = max{*M:* ∃ *an (n,M,ε)* code}

$\mathbb{E}\left[\imath_{X;Y}(X;Y)\right]$ $\mathbb{V}ar\left[\imath_{X;Y}(X;Y)\right]$

$$\frac{1}{n}\log M^*(n,\epsilon) \approx C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon)$$

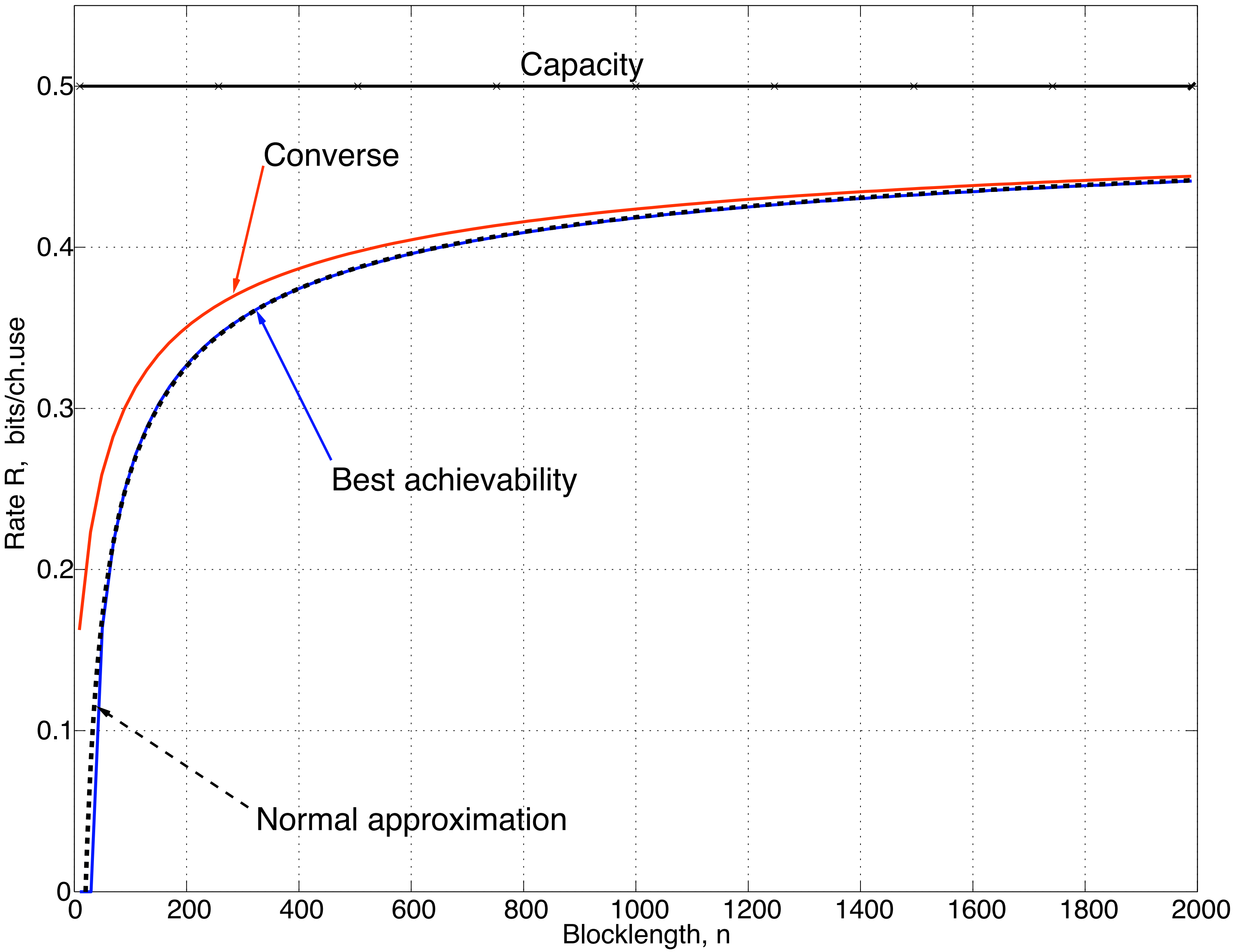
[Polyanskiy, et al. (2010)]

non-asymptotic fundamental limit

channel capacity

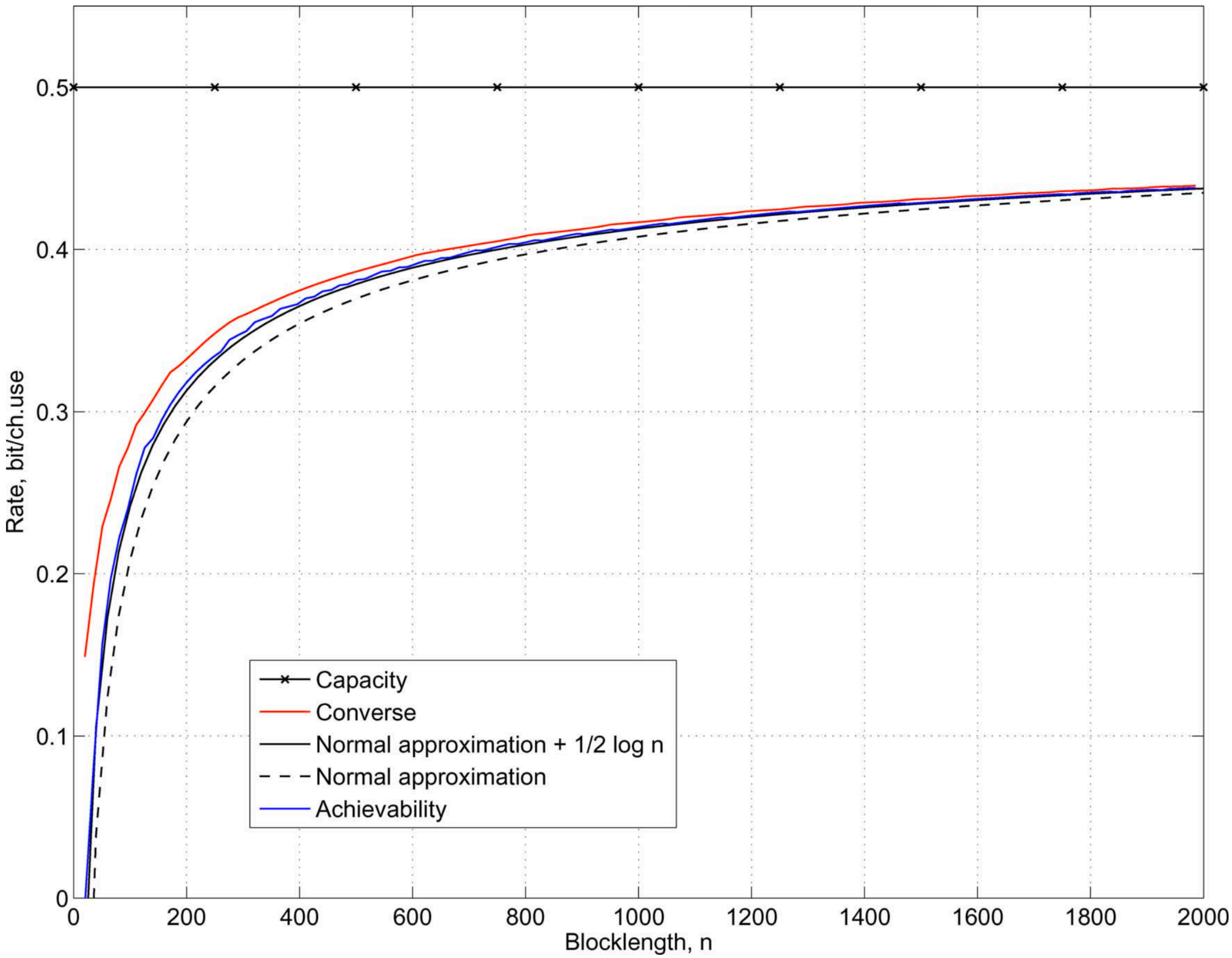
channel dispersion

Example: AWGN (SNR = 0 dB; ε = 10^{-3})



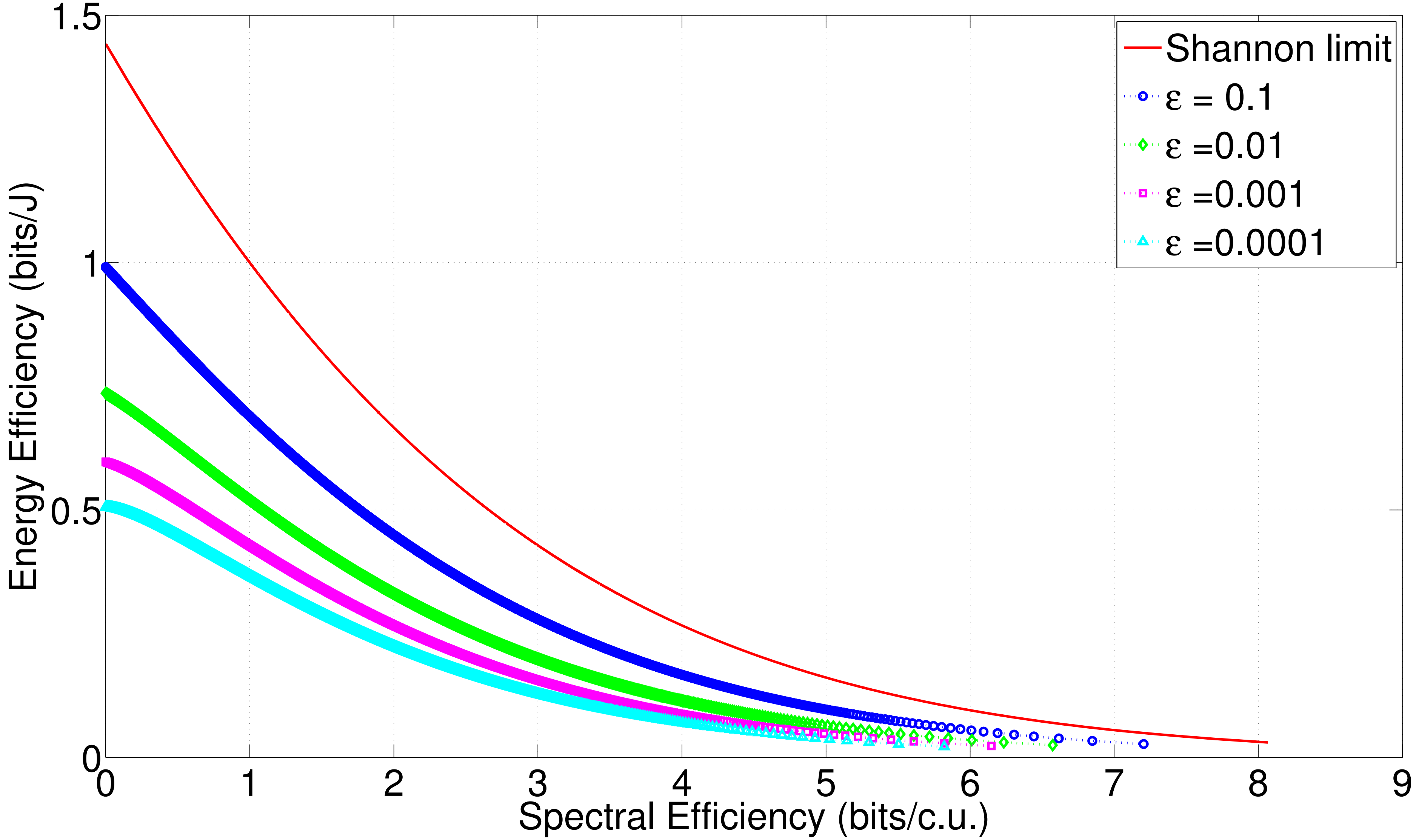
[Polyanskiy, et al. (2010)]

Example: BSC (crossover = 0.11; ε = 10^{-2})



[Polyanskiy, et al. (2010)]

Example: Energy/Spectral Efficiency Tradeoff



[Gorce, et al. (2016)]

Lossy Compression Revisited

- Refined asymptotic limit: stationary source with per-letter distortion

$$R(d) = \min_{P_{Y|X}\,:\,\mathbb{E}[d(X,Y)] \leq d} I(X;Y)$$

$$R(n,d,\epsilon) \approx R(d) + \sqrt{\frac{V(d)}{n}} Q^{-1}(\epsilon)$$

non-asymptotic fundamental limit

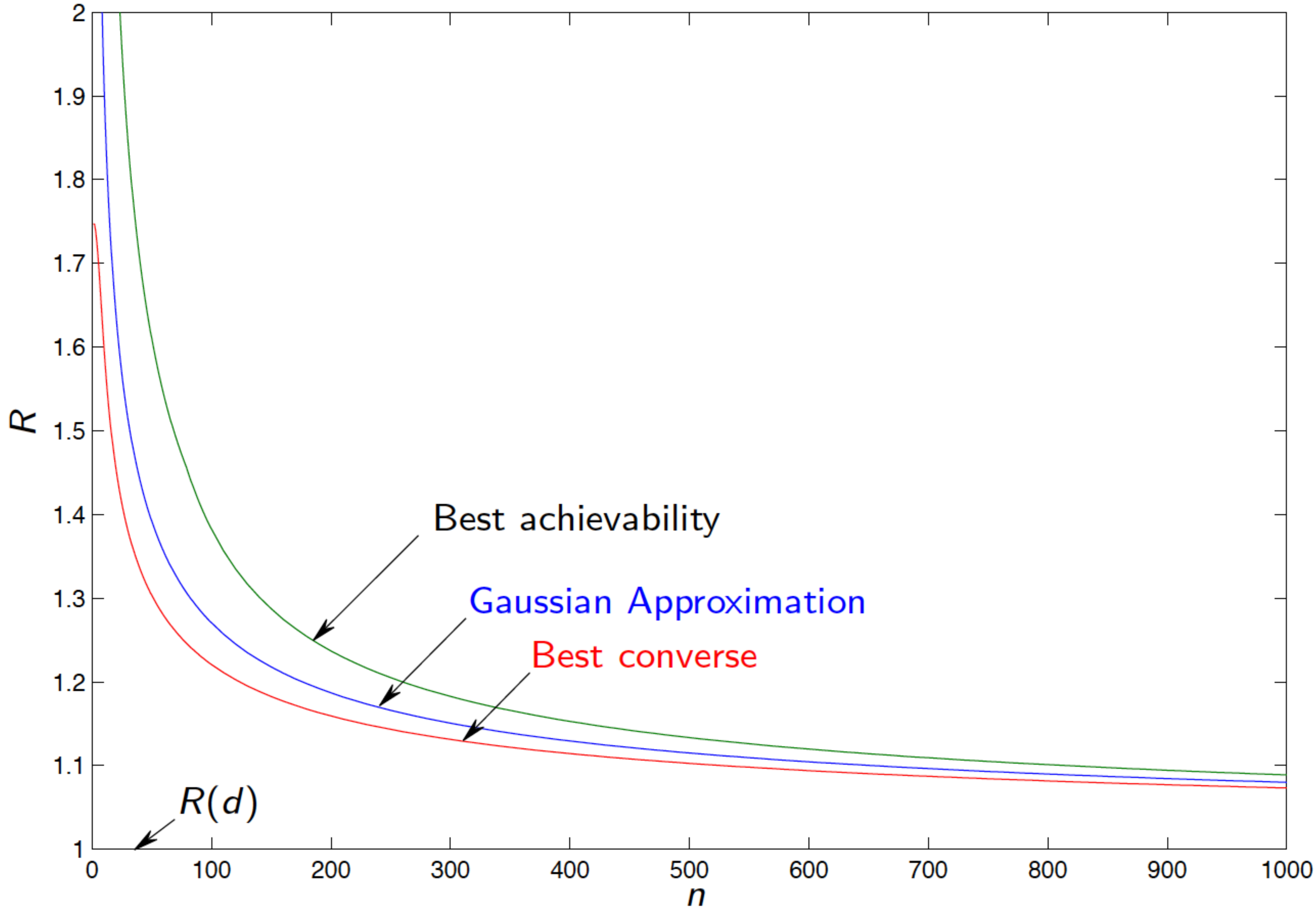
ϵ is the probability that the distortion incurred by the reproduction exceeds d

rate-dispersion function

Note: plus not minus

[Kostina, et al. (2012)]

Compression of Memoryless *N*(0,1) Source; d = 1/4 ; ε = 10^{-4}



[Kostina, et al. (2012)]

Lossless Compression Revisited

- Refined asymptotic limit: memoryless source with entropy H(X)

$$\sigma^2(X) = \mathsf{Var}\left(\imath_X(X)\right)$$

$$R^\star(n,\epsilon) \approx H(X) + \sqrt{\frac{\sigma^2(X)}{n}}Q^{-1}(\epsilon)$$

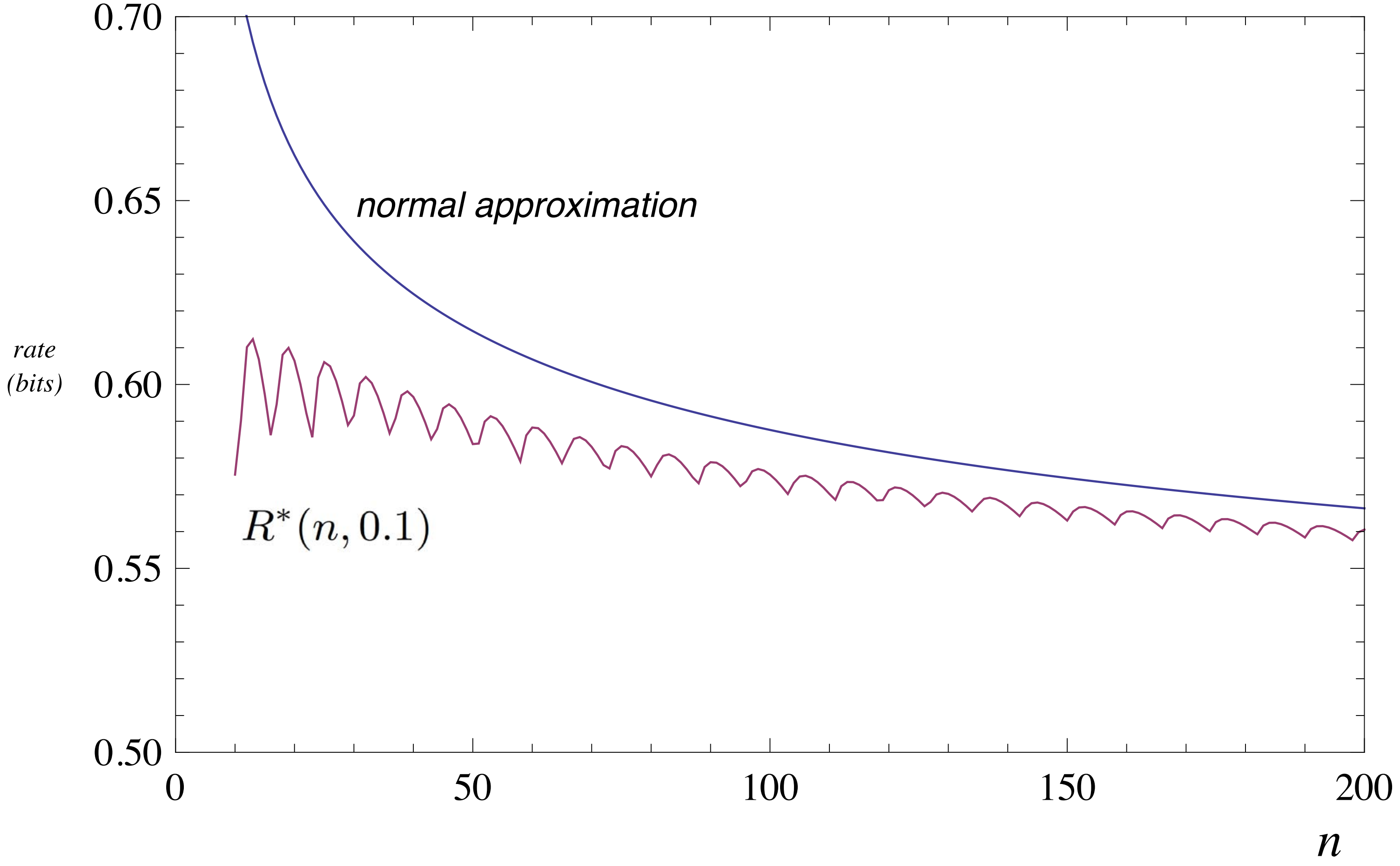
non-asymptotic best achievable rate

encoding failure probability

Called varentropy

[Kontoiannis, et al. (2014)]

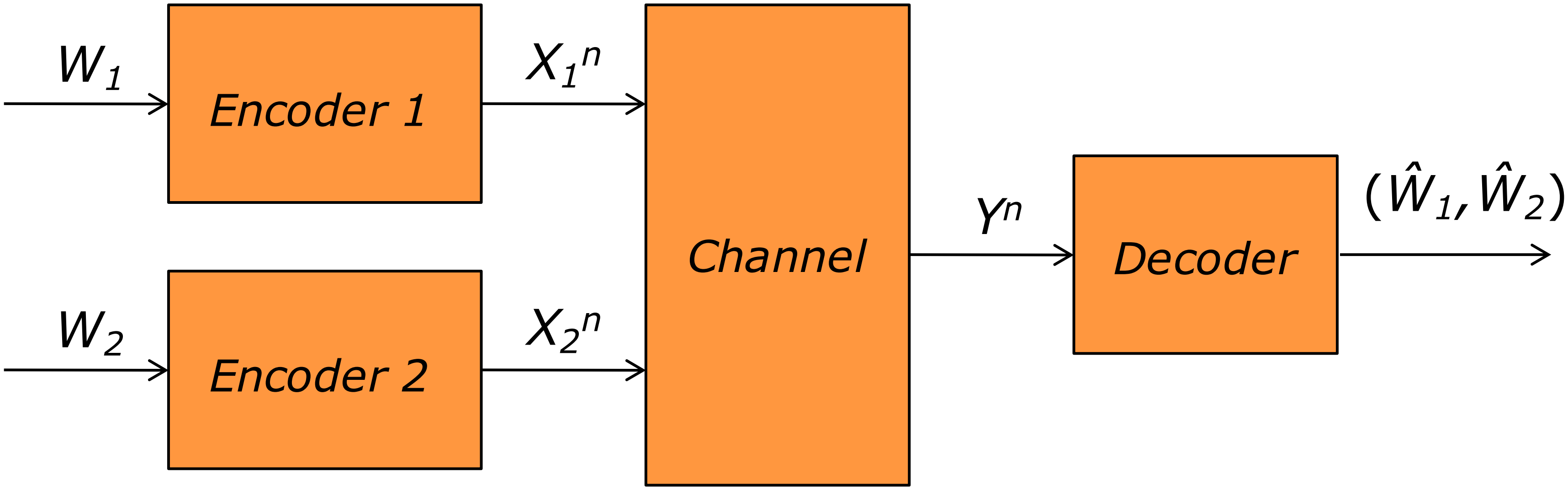
Example: Bernoulli-0.11 source ε = 10^{-1}



[Kontoiannis, et al. (2014)]

Extensions to Networks

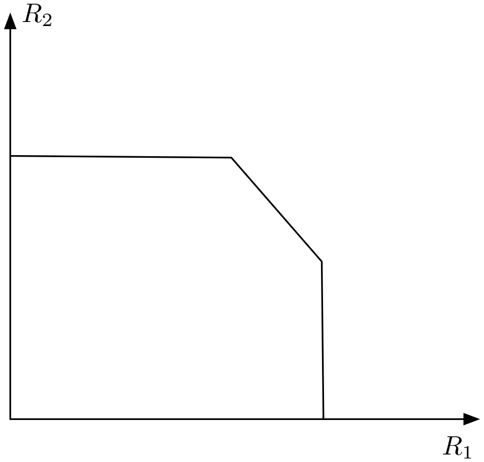
Network Information Theory (MAC - "Uplink")



- Capacity: largest rate region in the asymptotic regime of
  - Blocklength $n \to \infty$
  - Probability of error $\mathbb{P}\left((W_1, W_2) \neq (\hat{W}_1, \hat{W}_2)\right) \to 0$

$$C = \text{co}\left\{ \begin{array}{ll} R_1 & \leq I(X_1; Y | X_2) \\ R_2 & \leq I(X_2; Y | X_1) \\ R_1 + R_2 & \leq I(X_1, X_2; Y) \end{array} \right\}$$

for some $p(x_1)p(x_2)$



Non Asymptotic Version: Gaussian MAC

- An Achievability Result:

$$\left\{ \left(\frac{\log(M_1)}{n}, \frac{\log(M_2)}{n}\right) : \begin{array}{r} \frac{\log(M_1)}{n} \leq C(P_1) - \sqrt{\frac{V(P_1)}{n}}Q^{-1}(\lambda_1\epsilon) + O\left(\frac{1}{n}\right) \\ \frac{\log(M_2)}{n} \leq C(P_2) - \sqrt{\frac{V(P_2)}{n}}Q^{-1}(\lambda_2\epsilon) + O\left(\frac{1}{n}\right) \\ \frac{\log(M_1)}{n} + \frac{\log(M_2)}{n} \leq C(P_1 + P_2) - \sqrt{\frac{V(P_1+P_2)+V(P_1,P_2)}{n}}Q^{-1}(\lambda_3\epsilon) + O\left(\frac{1}{n}\right) \end{array} \right\}$$

for some $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Where

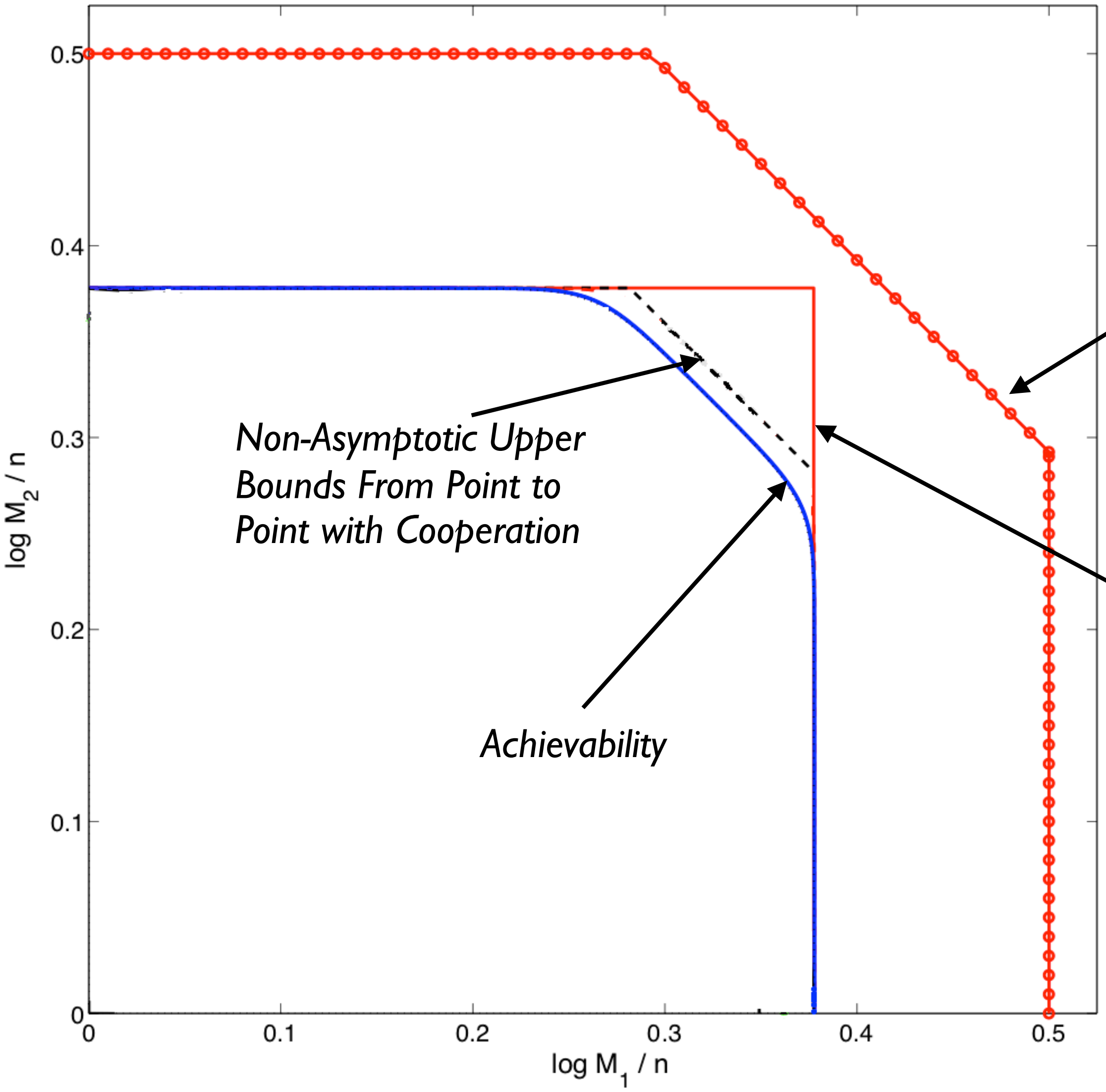
$$C(P) = \frac{1}{2}\log(1+P)$$

$$V(P) = \frac{(\log(\mathrm{e}))^2}{2}\frac{P(1+P)}{(1+P)^2}$$

$$V(P_1, P_2) = (\log(\mathrm{e}))^2\frac{P_1P_2}{(1+P_1+P_2)^2}$$

[MolavianJazi, et al. (2013)]

MAC Rate Region: *n* = 500; equal powers of 0dB; ε = 10^{-3}



Asymptotic Capacity

Non-Asymptotic Upper Bounds From Point to Point with Cooperation

Non-Asymptotic Upper Bounds From Point to Point

Achievability

[MolavianJazi, et al. (2013)]

Non Asymptotic Version: MAC with Degraded Message Sets

- Asymmetric MAC: encoder 1 knows both messages; encoder 2 only knows its own message.

second-order coding rate for j-th user

number of messages for j-th user as a function of blocklength

$$\frac{1}{n}\log M_{n,j} \approx R_j + \frac{1}{\sqrt{n}}L_j, \quad j = 1, 2$$

fraction characterizing boundary of capacity region

2-dimensional generalization of inverse Gaussian CDF

$$\begin{bmatrix} L_1 \\ L_1 + L_2 \end{bmatrix} \in \bigcup_{\beta \in \mathbb{R}} \left\{ \beta \mathbf{D}(\rho) + \Psi^{-1}\left(\mathbf{V}(\rho), \epsilon\right) \right\}$$

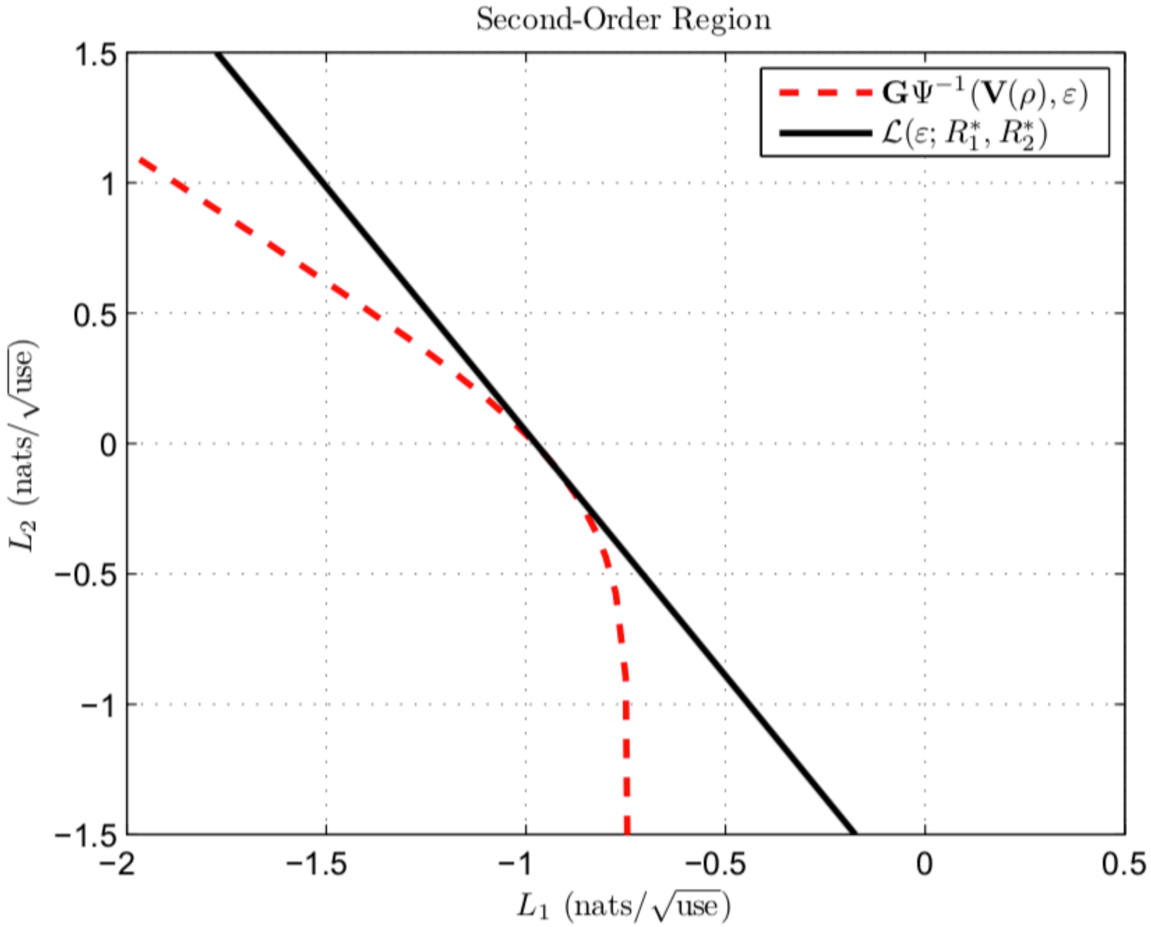
dispersion matrix

derivatives of asymptotic capacity region

[Scarlett, et al. (2015)]

Non Asymptotic Version: MAC with Degraded Message Sets

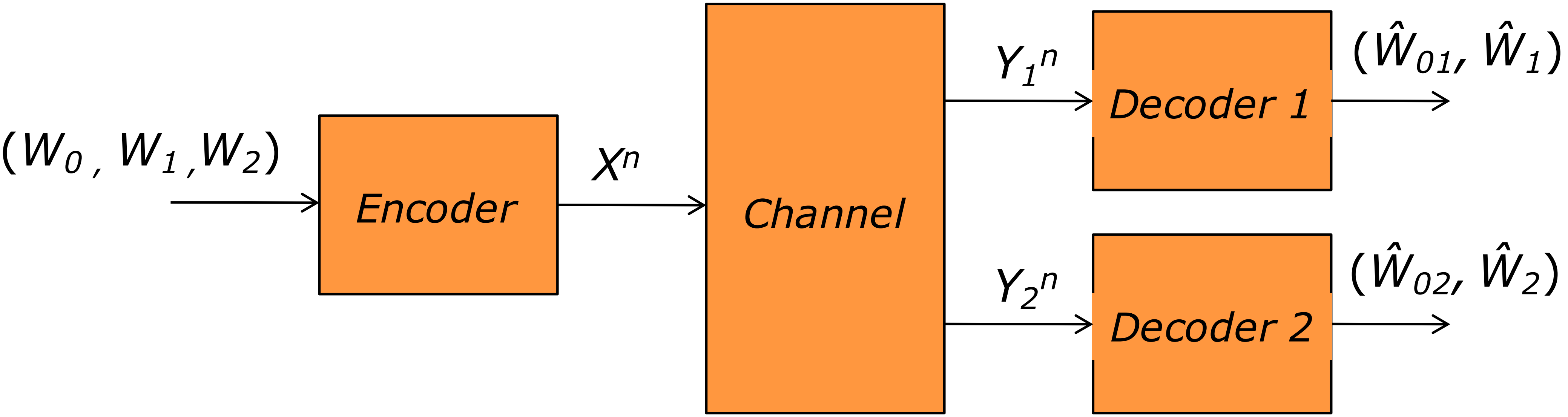
- Second order region: $\begin{bmatrix} L_1 \\ L_1 + L_2 \end{bmatrix} \in \bigcup_{\beta \in \mathbb{R}} \left\{ \beta \mathbf{D}(\rho) + \Psi^{-1}(\mathbf{V}(\rho), \epsilon) \right\}$



Unit transmit powers, $\rho = 0.5$, and $\epsilon = 0.1$; red curve is with $\beta = 0$

[Scarlett, et al. (2015)]

Network Information Theory (BC - “Downlink”)

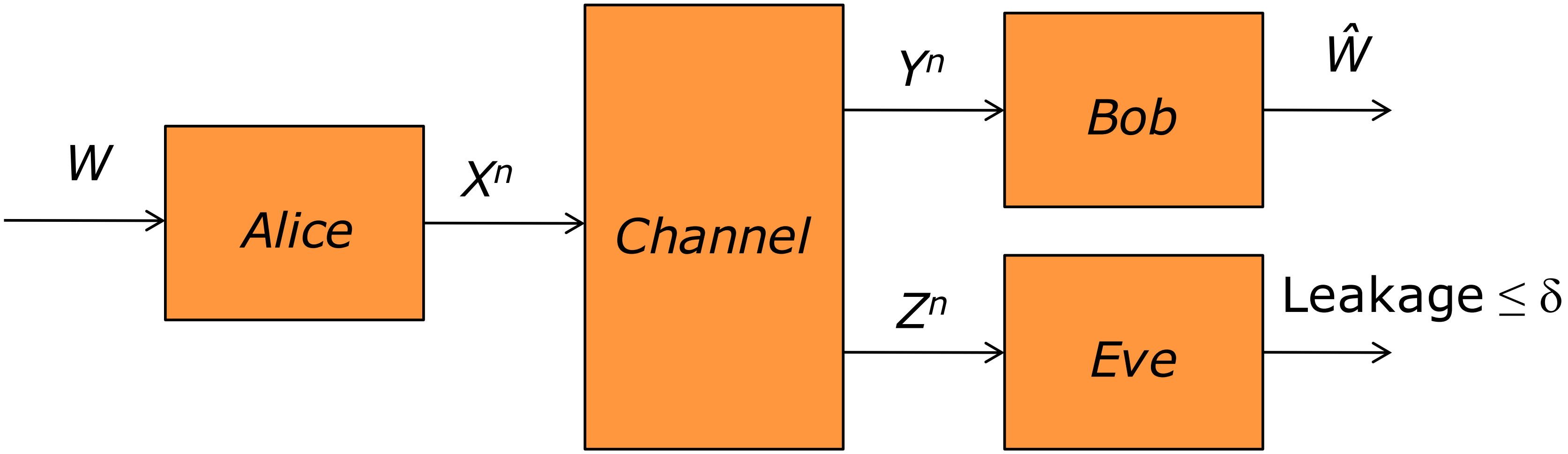


- Capacity: largest rate region in the asymptotic regime of
  - Blocklength $n \to \infty$
  - Probability of error $\mathbb{P}\Big((W_0, W_1) \neq (\hat{W}_{01}, \hat{W}_1) \text{ or } (W_0, W_2) \neq (\hat{W}_{02}, \hat{W}_2)\Big) \to 0$
  - Capacity is know only in special cases

- Non asymptotic results sparser here, but include a version of Marton’s inner bound with a common message.

[Liu, et al. (2015)]

Wiretap Channel and Secrecy Capacity

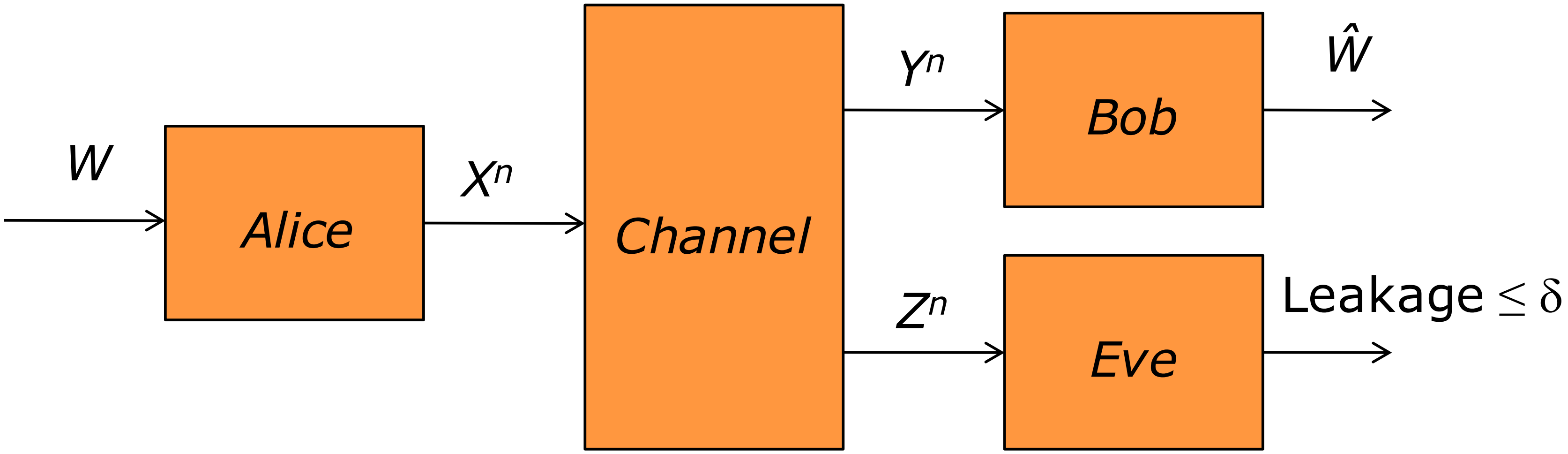


- Secrecy capacity: largest rate in the asymptotic regime of
  - Blocklength $n \to \infty$
  - Probability of error $\mathbb{P}\left(W \neq \hat{W}\right) \to 0$
  - Information leakage $\delta \to 0$

$$C_s = \max_{P_X} \{I(X;Y) - I(X;Z)\}$$

- Limitation: not suitable for low-latency applications as in IoT.

Wiretap Channel: Finite Blocklength



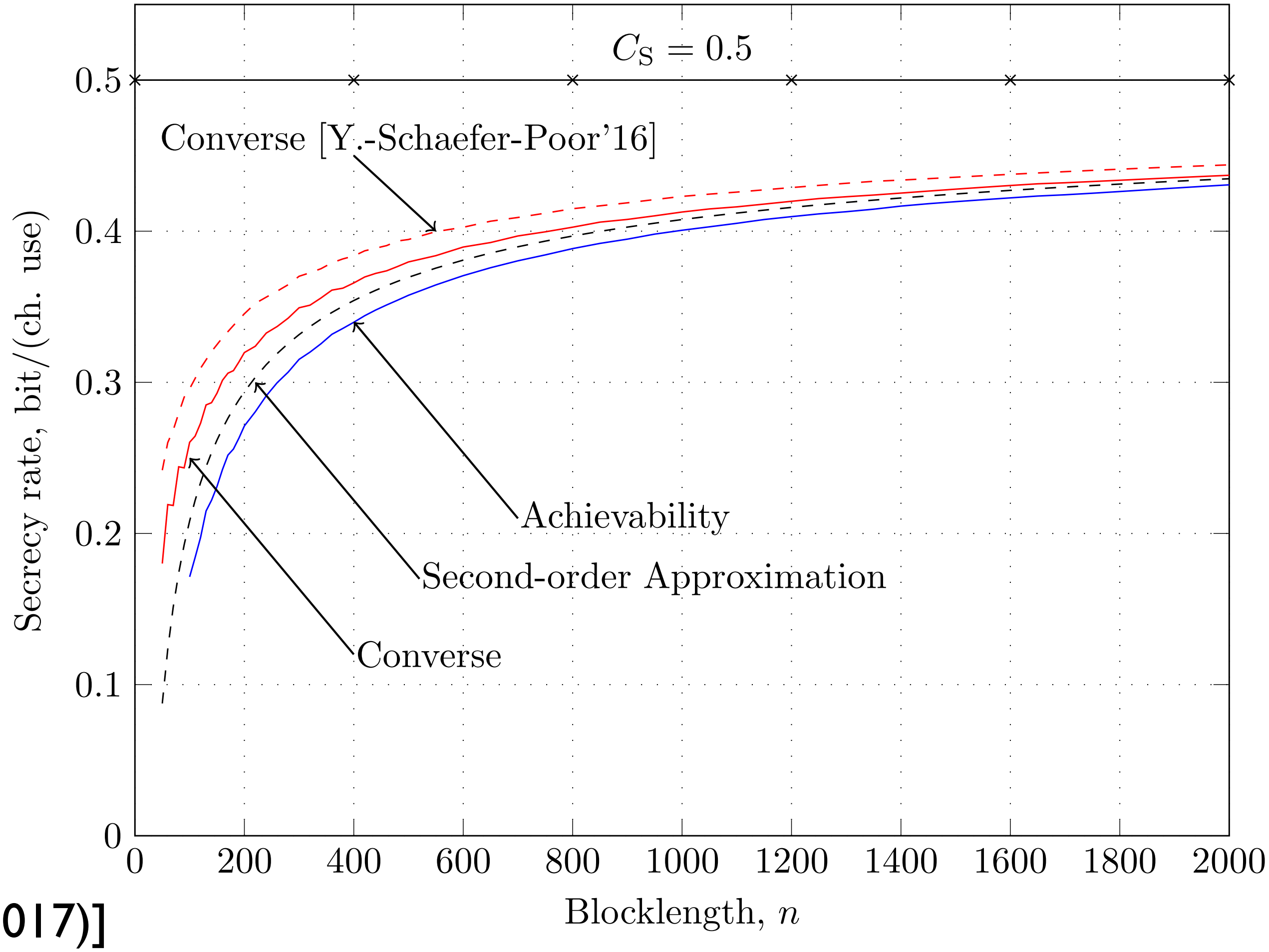
- (M, ϵ, δ) secrecy code:
  - Message $W \in \{1, \ldots, M\}$
  - Encoder $P_{X|W} : \{1, \ldots, M\} \to \mathcal{A}$; decoder $g : \mathcal{B} \to \{1, \ldots, M\}$
  - Average error probability: $\mathbb{P}\left(W \neq \hat{W}\right) \leq \epsilon$
  - Secrecy constraint: information leakage $\leq \delta$

- $R^*(n, \epsilon, \delta)$: maximum secret rate at a given blocklength.

Semi-deterministic Wiretap Channel (BSC): $\delta = \epsilon = 10^{-3}$

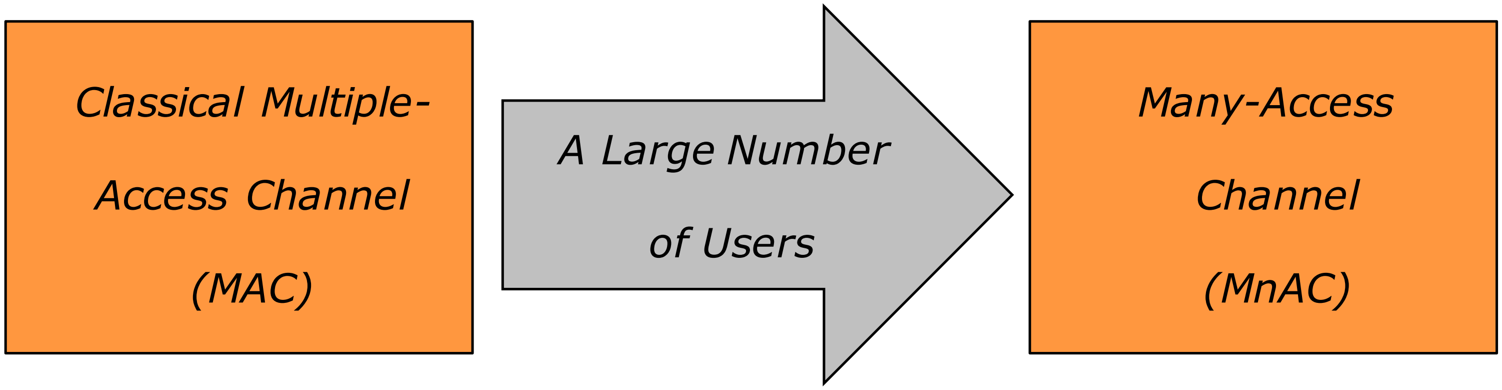
- Legitimate channel is deterministic, eavesdropper channel is BSC:

$$R^*(n, \epsilon, \delta) = C_s - \sqrt{\frac{V}{n}} Q^{-1}\left(\frac{\delta}{1-\epsilon}\right) + \mathcal{O}\left(\frac{\log n}{n}\right)$$



[Yang, et al. (2017)]

Latency in Large Networks



Classical Multiple-Access Channel (MAC)

A Large Number of Users

Many-Access Channel (MnAC)

The number of users *K(n)* is fixed as the blocklength *n* goes to infinity.

The number of users *K(n)* increases with the blocklength *n*.

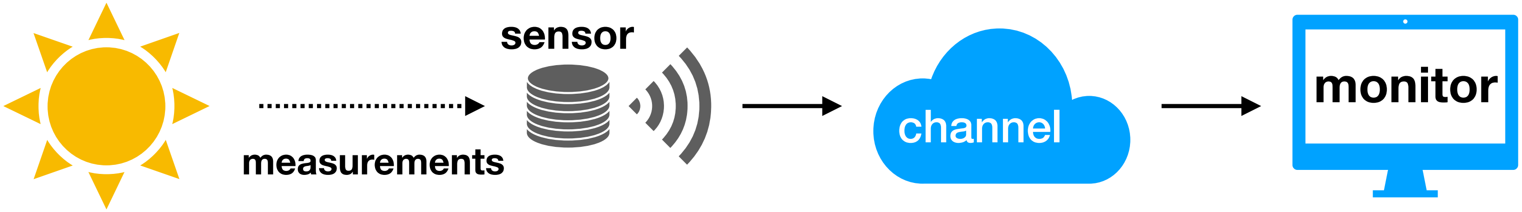
- Main Ideas:
  - Blocklength is proportional to latency
  - System latency per user $\ell = \frac{n}{K(n)}$
  - When is positive rate possible?

$$C = \begin{cases} \text{system rate is same} & K(n) = O(n) \\ \text{system rate decreases but is positive} & K(n) = O(n^p) \\ \text{system rate is zero} & K(n) = O(\mathrm{e}^{c \cdot n}) \end{cases}$$

- Message: We pay a rate penalty for low latency.

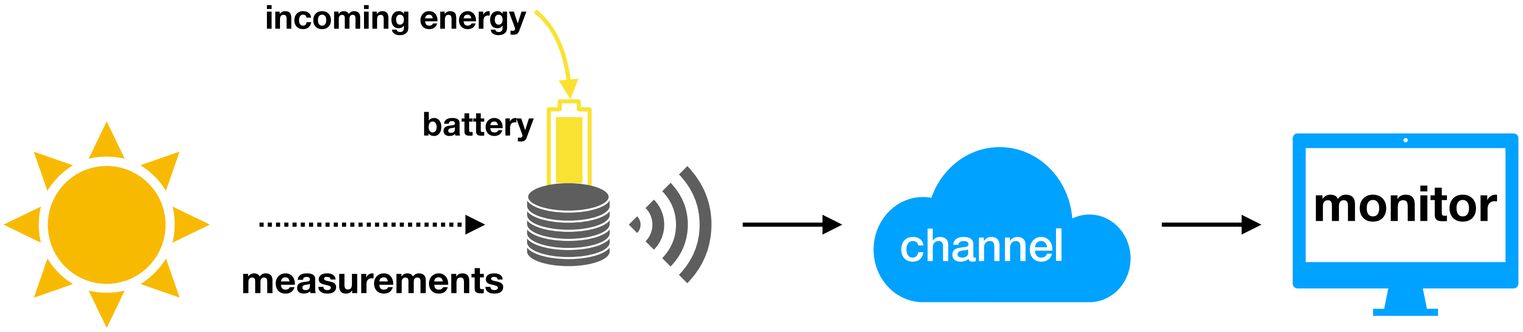
[Shahi, et al. (2016)]
[Chen, et al. (2017)]
[Cao, et al. (2018)]

Another Approach to Latency: Age of Information (AoI)



- AoI: time since latest measurement has reached destination
- Measures latency from destination’s perspective
- Assesses the freshness of data, in addition to distortion/error
- Suitable metric for real-time sensing applications in IoT
- Introduces queueing into the analysis

Example: AoI for Energy Harvesting Sensors



- Energy harvesting sensors cannot send data all the time
- Incoming energy needs to be optimally managed to minimize AoI
- Online threshold policies are age-minimal: send a new update only if AoI grows above a certain threshold

[Arafa, et al. (2018)]
[Bacinoglu, et al. (2018)

Conclusions

- In next-gen communications, latency tolerances will be much lower than in current generations because of time-critical machine-to-machine type applications

- Finite blocklength information theory is well-suited to assess latency in IoT applications, where the physical layer may predominate

- We examined:
  - point-to-point channels
  - multi-user channels
  - secrecy
  - large scale networks

- Age-of-Information: another approach is to assess latency via a different metric

- A rich area with much work left to do!

Thank you!