Nonasymptotic Universal Channel Coding

Pierre Moulin

University of Illinois at Urbana-Champaign

Electrical and Computer Engineering

DARPA ITMANET meeting
Austin, TX
May 24, 2010

Outline

- Capacity limits for finite blocklength:
  Strassen (1962), Polyanskiy *et al.* (2008), Hayashi (2009)
- Universal coding (unknown channel):
  Csiszar and Körner (1981)
- Let's consummate the union

Capacity for finite blocklength

- Discrete memoryless channel $\{W(y|x),\, x \in \mathcal{X},\, y \in \mathcal{Y}\}$
- Codewords $\mathbf{x}(m) \in \mathcal{X}^n$ for $m = 1, 2, \cdots, M_n$
- Code rate $R_n \triangleq \frac{1}{n} \log M_n$
- Decoding rule $\hat{m} = \phi(\mathbf{y})$
- Average error probability

$$P_e(W) = \frac{1}{M} \sum_{m=1}^{M_n} \sum_{\mathbf{y} \in \mathcal{Y}^n} W^n(\mathbf{y}|\mathbf{x}(m))\, \mathbb{1}\{\phi(\mathbf{y}) \neq m\}$$

- ϵ-capacity for blocklength n:

$$C_n(W, \epsilon) = \sup\{R_n \;:\; P_e(W) \leq \epsilon\}$$

- Shannon (1948):

$$C(W) = \lim_{\epsilon \downarrow 0} \lim_{n \to \infty} C_n(W, \epsilon) = \max_{P_X} I(P_X; W)$$

- Strassen (1962) and Polyanskiy (2008):

$$C_n(W, \epsilon) = C(W) - \frac{\sigma(W)}{\sqrt{n}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right)$$

- They showed that the first two terms giving $C_n(W, \epsilon)$ are achieved by standard random codes and ML decoding:

$$\max_{1 \le m \le M_n} W^n(\mathbf{y}|\mathbf{x}(m)) \quad \Leftrightarrow \quad \max_{1 \le m \le M_n} i(\mathbf{x}(m); \mathbf{y})$$

where $i(\mathbf{x}(m); \mathbf{y})$ is the information density defined by

$$i(\mathbf{x}; \mathbf{y}) \triangleq \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} = \log \frac{W^n(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}$$

- For iid random codes, $p(\mathbf{y})$ factors and thus

$$\frac{1}{n} i(\mathbf{x};\mathbf{y}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \underbrace{\log\frac{W(y_i|x_i)}{P_Y(y_i)}}_{l_i = l(x_i, y_i) = \text{iid rv's}}$$

- Mutual information $I(P_X, W) = \mathbb{E}\left[l(X,Y)\right]$
- Channel dispersion $\sigma(P_X, W) = \text{s.d.}\left[l(X,Y)\right]$
- Can thus write

$$\frac{1}{n} i(\mathbf{x}(m);\mathbf{y}) = I(P_X, W) + \frac{\sigma(P,W)}{\sqrt{n}}\, Z_n$$

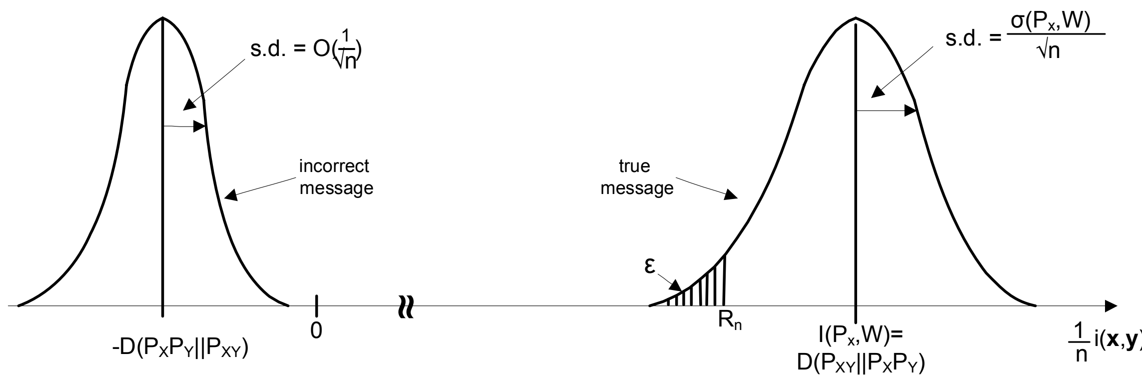
where $Z_n \xrightarrow{d} \mathcal{N}(0,1)$ by the Central Limit Theorem. Hence

$$\lim_{n\to\infty} Pr[\frac{1}{n} i(\mathbf{x}(m);\mathbf{y}) < \underbrace{I(P_X, W) - \frac{\sigma(P,W)}{\sqrt{n}} Q^{-1}(\epsilon)]}_{=R_n} = \epsilon$$

- Using union bound and large deviations, can show that

$$
\begin{aligned}
& Pr[\exists m' \neq m \ : \ \frac{1}{n} i(\mathbf{x}(m'); \mathbf{y}) \geq R_n] \\
& \leq \quad (2^{nR_n} - 1) Pr[\frac{1}{n} i(\mathbf{x}(m'); \mathbf{y}) \geq R_n] \downarrow 0 \quad \text{as } n \to \infty
\end{aligned}
$$

[Polyanskiy (2010) used a different approach]



Unknown Channel / Informed Decoder

- Channel W belong to a compound class $\mathscr{W}$ of DMCs. Encoder does not know W but decoder does.
- Denote by P_X^* the *best input distribution* and by W^* the *worst channel* achieving compound capacity

$$C(\mathscr{W}) = \max_{P_X \in \mathscr{P}(\mathcal{X})} \min_{W \in \mathscr{W}} I(P_X, W)$$

  and compound dispersion

$$\sigma^2(\mathscr{W}) = \min_{P_X \in \mathscr{P}^*} \max_{W \in \mathscr{W}^*} \sigma(P_X, W)$$

- It is straightforward to extend the previous results to show that compound ϵ-capacity for blocklength n is given by

$$C_n(\mathscr{W}, \epsilon) = C(\mathscr{W}) - \frac{\sigma(\mathscr{W})}{\sqrt{n}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right)$$

Universal Coding

- Channel $W(y|x)$ is unknown, can't do ML decoding
- Universal code (if exists) achieves Shannon capacity $C(W)$
- For finite $\mathcal{X}, \mathcal{Y}$, define the joint type

$$\hat{P}_{\mathbf{xy}}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{N} \mathbb{1}\{x_i = x,\, y_i = y\}, \quad x \in \mathcal{X},\ y \in \mathcal{Y}$$

associated with length-n sequences $\mathbf{x}$ and $\mathbf{y}$

- Empirical mutual information

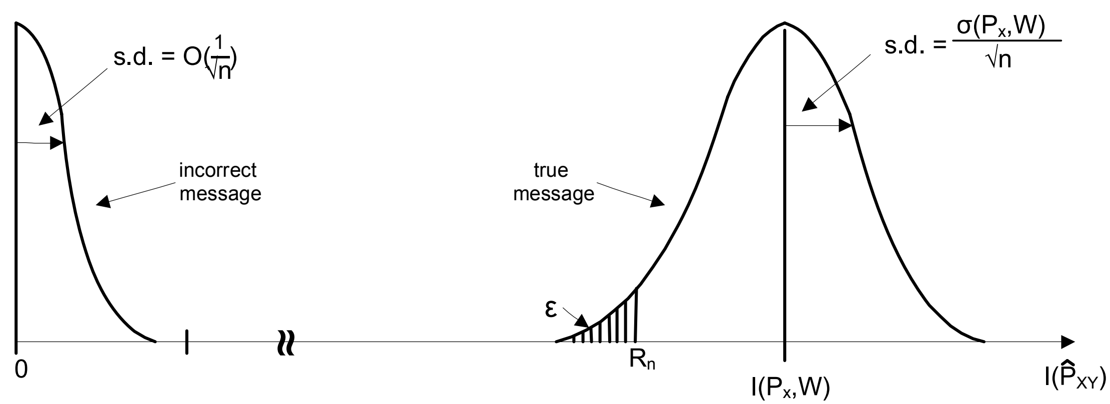
$$I(\hat{P}_{\mathbf{xy}}) = \sum_{x,y} \hat{P}_{\mathbf{xy}}(x,y) \frac{\hat{P}_{\mathbf{xy}}(x,y)}{\hat{P}_{\mathbf{x}}(x)\hat{P}_{\mathbf{y}}(y)}$$

- Use random constant-composition codes: $\mathbf{x}(m)$ have the same type P for all $1 \le m \le M_n$

- Maximum Mutual Information (MMI) decoder:

$$\max_{1 \le m \le M_n} I(\hat{P}_{\mathbf{x}(m)\mathbf{y}})$$

- Probability of error analysis:



Note that $I(\hat{P})$ is not a sum of iid rv's

- Csiszár and Körner (1981) have shown that this code achieves $C(\mathscr{W})$ as well as optimal error exponents at high rates.

Connection to $C_n(W,\epsilon)$?

- In unpublished notes entitled "Behavior Near Channel Capacity", Shannon showed that the reliability function may be approximated as $E(W,R) = \frac{(C(W)-R)^2}{2\sigma^2(W)}$ for $R \approx C(W)$
- Try the approximation $\epsilon \overset{w.t.}{\approx} e^{-nE(W,R_n)+o(n)} \overset{w.t.}{\approx} e^{-nE(W,R_n)}$ where w.t. denotes *wishful thinking*

$$
\begin{aligned}
\overset{w.t.}{\Longrightarrow} R_n \quad &\approx \quad C(W) - \frac{\sigma(W)}{\sqrt{n}}\sqrt{\ln(2/\epsilon)} \quad \text{for "very large}'' n \\
&\sim \quad C(W) - \frac{\sigma(W)}{\sqrt{n}} Q^{-1}(\epsilon) \quad \text{for } \epsilon \downarrow 0
\end{aligned}
$$

- same as $C_n(W,\epsilon)$ given previously
- too much w.t. to be convincing

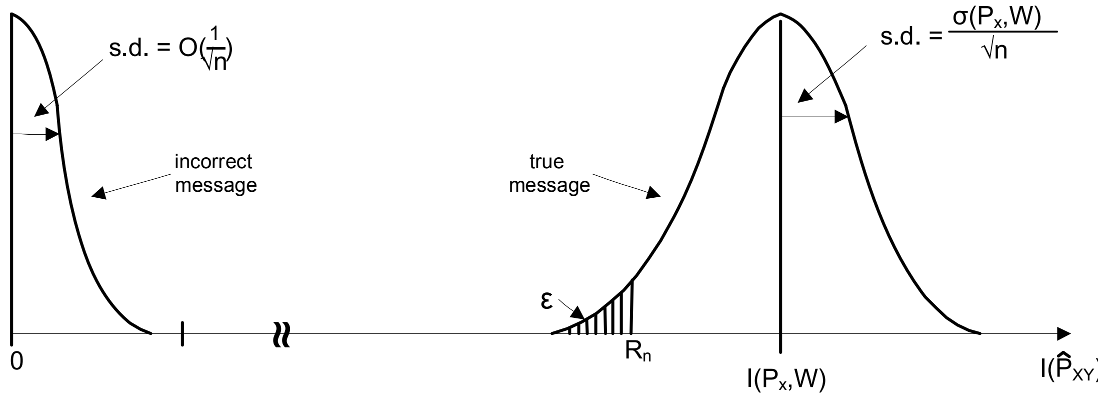
Towards a legal union

- Methodology: choose a random code and decoder and conduct precise analysis of error probability
- Use Shannon's iid random codes with $P = P^*$ and rate

$$R_n = I(P, W^*) - \frac{\sigma(P, W^*)}{\sqrt{n}} Q^{-1}(\epsilon)$$

  and Csiszár-Körner's MMI decoder
- **Main result:** The compound ϵ-capacity for blocklength n is the same as in the informed decoder case
- In other words, there is no penalty (neither in the first nor even in the second order) for the decoder not knowing channel W

- Error analysis: see distributions for e.m.i. statistic $I(\hat{P}_{\mathbf{x}(m)\mathbf{y}})$ for true message m and for incorrect message



- Decision rule: variable-size list decoders outputs list of all m such that $I(\hat{P}_{\mathbf{x}(m)\mathbf{y}}) \geq R_n$
- Erasure probability: $P_\emptyset(W) = Pr[I(\hat{P}_{\mathbf{x}(m)\mathbf{y}}) < R_n]$
Expected # of incorrect messages on list:

$$\mathbb{E}_W[N_i] = (2^{nR_n} - 1)Pr[I(\hat{P}_{\mathbf{x}(m')\mathbf{y}}) \geq R_n]$$

- True message m: the joint type $\hat{P}_{\mathbf{x}(m)\mathbf{y}}$ follows multinomial distribution with probabilities $P_X(x)W(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- Asymptotics are given by

$$\hat{P}(x,y) = P(x)W(y|x) + \frac{1}{\sqrt{n}} G_n(x,y), \quad x \in \mathcal{X},\ y \in \mathcal{Y}$$

where

$$\begin{aligned} \mathbb{E}[G_n] &= 0 \\ \mathrm{Cov}[G_n] &= \Sigma \quad \text{(rank deficient)} \\ G_n &\xrightarrow{d} \mathcal{N}(0, \Sigma) \end{aligned}$$

- Since $\hat{P} = P + \frac{1}{\sqrt{n}} G_n$ and $G_n \xrightarrow{d} \mathcal{N}(0, \Sigma)$, we have

$$I(\hat{P}) \xrightarrow{d} I(P,W) + \frac{1}{\sqrt{n}} \underbrace{G_n \cdot \nabla I(P,W)}_{\xrightarrow{d} N(0,\sigma^2(P,W))} + \frac{1}{n} \underbrace{\xi_n}_{\xrightarrow{d} N(0,\alpha(P,W))}$$

- Indeed $\sigma^2(P,W) = [\nabla I(P,W)]^T \, \Sigma \, [\nabla I(P,W)]$. Thus $I(\hat{P})$ has the same asymptotic distribution as normalized info density!

$$\Rightarrow \lim_{n\to\infty} P_\emptyset(W) \;=\; \lim_{n\to\infty} Pr[I(\hat{P}_{\mathbf{x}(m)\mathbf{y}}) < \underbrace{I(P,W^*) - \frac{\sigma(P,W^*)}{\sqrt{n}} Q^{-1}(\epsilon)]}_{=R_n}$$

$$\leq \; \epsilon$$

with equality if $W = W^*$

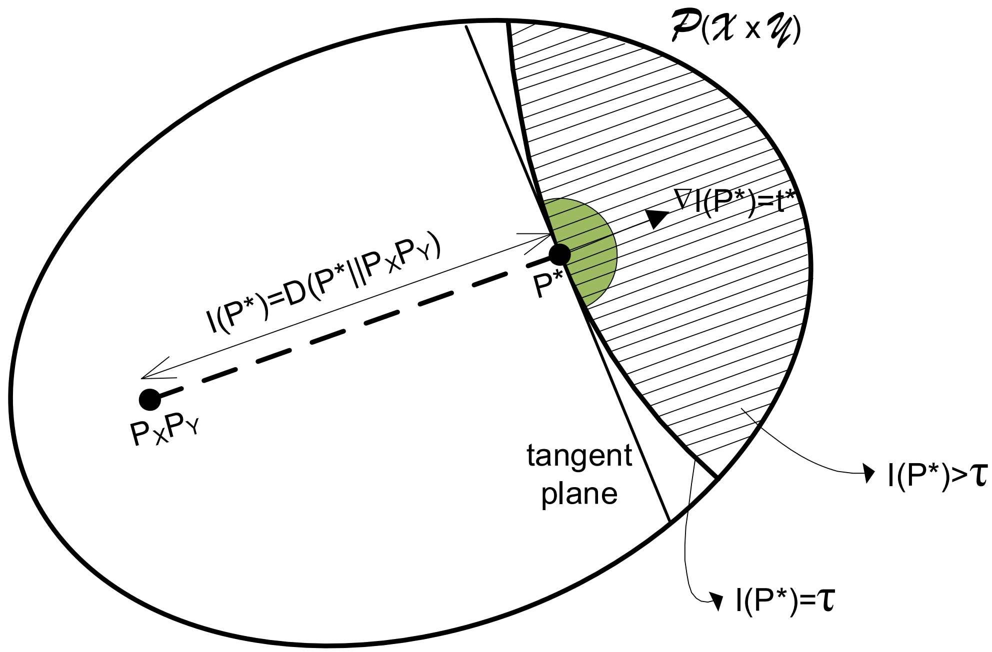
- Incorrect message: for $m' \neq m$, the joint type $\hat{P}_{\mathbf{x}(m')\mathbf{y}}$ follows a multinomial distribution with probabilities $P_X(x)P_Y(y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$

- Using large deviations (tilted distributions), can show that the expected # of incorrect messages

$$
\begin{aligned}
\mathbb{E}_W[N_i] &\leq (2^{nR_n} - 1) Pr[\frac{1}{n} I(\hat{P}_{\mathbf{x}(m')\mathbf{y}}) \geq R_n] \\
&\sim (2^{nR_n} - 1) \frac{2^{-nI(P_X, W)}}{\sqrt{2\pi n\, \zeta^2(P_X, W)}} \\
&\leq \frac{2^{-\sqrt{n}\, \sigma^* Q^{-1}(\epsilon)}}{\sqrt{2\pi n\, \zeta^2(P_X, W)}}
\end{aligned}
$$

(with equality for $W = W^*$) vanishes as $n \to \infty$.

- Geometric interpretation (where P^* is the tilted distribution and $\tau = R_n$):



Conclusion

- Combining Shannon's iid random codes with Csiszár and Körner's MMI decoder yields the same first <u>and second</u> order coding rate as in the case of an informed decoder.
- Hence there is no penalty for not knowing the channel!
- The bounds are independent of alphabet size. There is no large subexponential term of the form $(n+1)^{|\mathcal{X}|\,|\mathcal{Y}|}$ as in the error probability derivations of Csiszár and Körner
- The Gaussian approximation and the large-deviations approach based on tilted distributions and precise asymptotics are applicable to arbitrary large as well as continuous alphabets (by application of empirical process theory).
- Future work will explore more complicated channels (with memory) and multiterminal extensions