Bottleneck Problems: Connections, Applications and Implications

Shlomo Shamai

Technion—Israel Institute of Technology
sshlomo@ee.technion.ac.il

Views reflected in joint studies with:
A. Zaidi, I.E. Auguerri, G. Caire, O. Simeone and S-H. Park

2020 Workshop on Coding, Cooperation, and Security in Modern Communication Networks (COCO 2020): July 16, 2020

Outline

* **Information Bottleneck:**
* **Connections:**
  - Remote Source Coding.
  - Common Reconstruction.
  - Information Combining.
  - Wyner-Ahlswede-Korner Problem.
  - Efficiency of Investment Information.
  - Hypothesis Testing.
  - Compound Wiretap Channel.
* **Some Perspectives**

Information Bottleneck

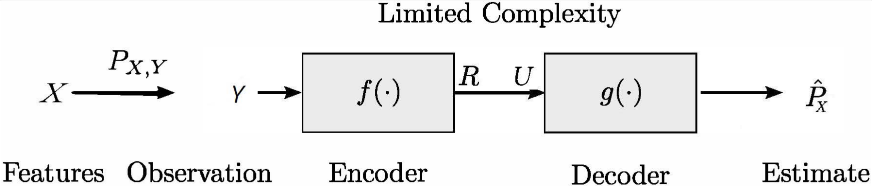
$$X \longrightarrow Y \longrightarrow U$$

- Efficiency of a given representation $U = f(Y)$ measured by the pair

  Rate (or *Complexity*): $I(U;Y)$ and Information (or *Relevance*): $I(U;X)$

- Information $I(X;U)$ can be achieved by OBLIVIOUS coding Y while with the logarithmic distortion with respect to X
- Single letter-wise, U is not necessarily a deterministic function of Y
- The non-oblivious bottleneck problem is immediate as the $\min(I(X;Y), R)$ is achievable by having the relay decoding the message transmitted by X
- The bottleneck problem connects to many timely aspects, such as 'deep learning' *[Tishby-Zaslavsky, ITW'15]*.

Digression: Learning via the Information Bottleneck Method



- Preserving all the information about X that is contained in Y, i.e., $I(X;Y)$, requires high *complexity* (in terms of *minimum description coding length*).
  - Other measures of complexity may be (Vapnik-Chervonenkis) VC-dimension, covering numbers, ..
- Efficiency of a given representation $\mathbf{U} = f(\mathbf{Y})$ measured by the pair

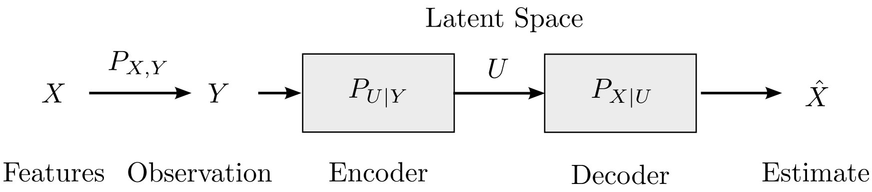
  Complexity: $I(U;Y)$ and Relevance: $I(U;X)$

- Example: $\mathbb{E}\Big(x - f^*(v)\Big)^2$, $f^*(v) = \mathbb{E}(x|v)$

$$\max_{p(u|x)} I(U;X) \quad \text{s.t.} \quad I(U;Y) \leq R, \quad \text{for} \quad 0 \leq R \leq H(Y)$$

$$\min_{p(u|x)} I(U;Y) \quad \text{s.t.} \quad I(U;X) \geq \Delta, \quad \text{for} \quad 0 \leq \Delta \leq I(X;Y)$$

Basically, a Remote Source Coding Problem !



- Reconstruction at decoder is under log-loss measure,

$$R(\Delta) = \min_{p(u|y)} I(U;Y)$$

where the minimization is over all conditional pmfs $p(u|y)$ such that

$$\mathbb{E}[\ell_{\log}(X,U)] \leq H(X) - H(X|U) = H(X) - \Delta$$

- R. L. Dobrushin and B. S. Tsybakov, *"Information transmission with additional noise"*, IRE Tran. Info. Theory, Vol. IT-8, pp. 293-304, 1962.

- H. Witsenhausen, A. Wyner, *"A conditional entropy bound for a pair of discrete random variables"*, IEEE Trans. on Info. Theory, Vol. 21, pp. 493-501, 1975.

- Solution also coined as the Information Bottleneck Method *[Tishby'99]*

$$L_{\mathrm{IB}}(\beta, P_{X,Y}) = \min_{p(u|y)} I(Y;U) - \beta I(X;U)$$

Other Connections

- **Efficiency of Investment Information**
  - X - Stock Market Data.

    Y - Correlated Information about X.

    $\Delta(R)$ the maximum increase in growth rate when Y is described to the investor at rate R (a logarithmic distortion that relates to the Wyner-Ahlswede-Korner Problem).
  - Solution of the bottleneck for: (X, Y) are binary and (X, Y) Gaussian (horse race examples).
  - E. Erkip and T. M. Cover, "The Efficiency of Investment Information", IEEE Trans. on Info. Theory, Vol. 44, May 1998.

Other Connections (Cont.)

- **Common Reconstruction**. Because $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- U$, we have

$$
\begin{aligned}
I(U;X) &= I(U;Y) - I(U;Y|X) \\
&\leq R - I(U;Y|X)
\end{aligned}
$$

- Y. Steinberg, *"Coding and common reconstruction"*, IEEE Trans. on Inform. Theory, vol. 55, no. 11, pp. 4995–5010, Nov. 2009 (X – side information is not used for the 'source' Y common reconstruction).

∗ Heegard-Berger Problem with Common Reconstruction: Y-source, to be commonly reconstructed (with logarithmic distortion), with and without side information (X), as to maximize $I(U;X)$.

- M. Benammar, A. Zaidi, "Rate-Distortion of a Heegard-Berger Problem with Common Reconstruction Constraint," IZS, March 2–4, 2016.

Other Connections (Cont.)

- **Information Combining**

  $$I(Y;U,X) = I(U;Y) + I(X;Y) - I(U;X) \qquad \text{(since } X - \circ - Y - \circ - U\text{)}$$

  Since $I(X;Y)$ is given and $I(Y;U) = R$, maximizing $I(U;X)$ is equivalent to minimizing $I(Y;U,X)$.

- I. Sutskover, S. Shamai and J. Ziv, "Extremes of Information Combining", IEEE Trans. Inform. Theory, vol. 51, no. 4, pp. 1313–1325, April 2005.
- I. Land and J. Huber, *"Information combining,"* Foundations and trends in Commun. and Inform. Theory, vol. 3, pp. 227–330, Nov. 2006.

Other Connections (Cont.)

- **Hypothesis Testing**

  Let $(X^n;Y^n)$ be an n length, iid sequence of pairs (X,Y). Assume that the sequences had been produced by two possible probability measures:

  $$\begin{array}{llll} H0: & P_X P_Y: & (X,Y) & \text{Independent random variables.} \\ H1: & P_{X,Y}: & (X;Y) & \text{Dependent random variables.} \end{array}$$

  X^n is available at the destination, and Y^n, is encoded at rate R.

  $\Rightarrow$ For $n \to \infty$, the Stein error exponent (normalized by n), of the Neuman-Pearson type II error: (the sequences were governed by $\mathrm{H0}$, while the decision was $\mathrm{H1}$), is lower bounded by:

  $$\max I(X;U), I(Y;U) \leq R : X - Y - U \,,$$

- (that is the information bottleneck result) for any type I decision error (the sequences were governed by $\mathrm{H1}$, while the decision was $\mathrm{H0}$) $\leq \varepsilon$.

  R. Ahlswede and I. Csiszár, "Hypothesis Testing with Communication Constraints," IEEE Trans. Inform. Theory, vol. IT–32, no. 4, pp. 533–542, July 1986.

Other Connections (Cont.)

- **Compound Wiretap Channel**
- $X - Y - U$, (X-input, Y-legitimate receiver, U-eavesdropper). The wiretap capacity is:

$$I(X;Y) - I(X;U)\,.$$

- The compound degraded wiretap channel: The wiretapper can have anything, satisfying $I(Y;U) \leq C$.
- Evidently, as known *[Liang-Kramer-Poor-Shamai, EURASIP 2009]*, *[Bjelakovic-Boche-Sommerfeld, Problems of Information Transmission, 2013]*:

\- The wiretap capacity is $\min : I(X;Y) - I(X;U)$, over the allowable set: $I(Y;U) \leq C$, which is the bottleneck solution.

Other Connections (Cont.)

- **Wyner-Ahlswede-Körner Problem**

  If X and Y are encoded at rates R_X and R_Y, respectively. For given $R_Y = R$, the minimum rate R_X that is needed to recover X losslessly is

$$R_X^\star(R) = \min_{p(u|y)\,:\,I(U;Y)\,\leq\,R} H(X|U)$$

  So, we get

$$\max_{p(u|y)\,:\,I(U;Y)\leq R} I(U;X) = H(X) - R_X^\star(R)$$

- R. F. Ahlswede and J. Korner, *"Source coding with side information and a converse for degraded broadcast channels"*, IEEE Trans. on Info. Theory, Vol. 21, pp. 629-637, 1975.

- A. D. Wyner, *"On source coding with side information at the decoder"*,
IEEE Trans. on Info. Theory, Vol. 21, pp. 294-300, 1975.

Vector Gaussian Information Bottleneck

- $(\mathbf{X}, \mathbf{Y})$ jointly Gaussian, $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$
- Optimal encoding $P_{U|Y}$ is a noisy linear projection to a subspace whose dimensionality is determined by the bottleneck Lagrangian multiplier β
  [Chechik-Globerson-Tushby-Weiss, '05]

$$\mathbf{U} = \mathbf{A}\mathbf{Y} + \mathbf{Z}, \qquad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

where

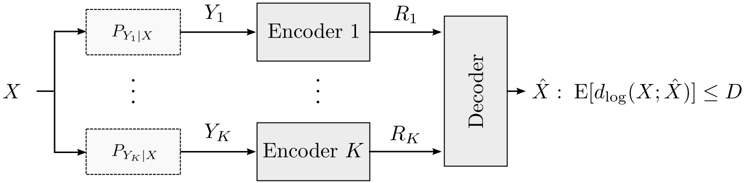
$$\mathbf{A} = \begin{cases} [\mathbf{0}^T; \dots; \mathbf{0}^T], & \text{if } 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 \mathbf{v}_1^T; \mathbf{0}^T; \dots; \mathbf{0}^T], & \text{if } \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; \mathbf{0}^T; \dots; \mathbf{0}^T], & \text{if } \beta_2^c \leq \beta \leq \beta_3^c \\ \qquad \vdots & \end{cases}$$

and $\{\mathbf{v}_1^T, \dots, \mathbf{v}_N^T\}$ are the left eigenvectors of $\mathbf{\Sigma}_{y|x}\mathbf{\Sigma}_y^{-1}$, sorted by their ascending eigenvalues $\{\lambda_1, \dots, \lambda_N\}$; $\beta_i^c = 1/(1-\lambda_i)$ are critical β values; $r_i = \mathbf{v}_i^T \mathbf{\Sigma}_y \mathbf{v}_i$ and

$$\alpha_i = \sqrt{\frac{\beta(1-\lambda_i) - 1}{\lambda_i r_i}}$$

Rate-Information Trade-off Gaussian Vector Channel [Winkelbauer-Matz, ISIT'14].

CEO Source Coding Problem under Log-Loss



- CEO source coding problem under log-loss distortion:

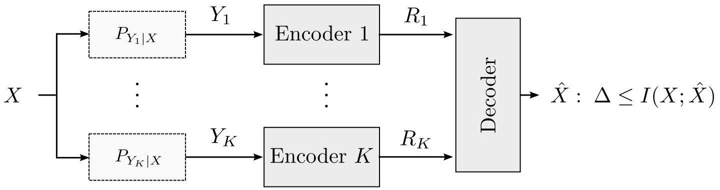
$$d_{\log}(x,\hat{x}) := \log\left(\frac{1}{\hat{x}(x)}\right)$$

  where $\hat{x} \in \mathcal{P}(\mathcal{X})$ is a probability distribution on $\mathcal{X}$.

- Characterization of rate-distortion region in *[Courtade-Weissman'14]*
  - Key step: log-loss admits a lower bound in the form of conditional entropy of the source conditioned on the compression indices:

$$nD \geq \mathrm{E}[d_{\log}(X^n;\hat{X}^n)] \geq H(X^n|J_{\mathcal{K}}) = H(X^n) - I(X^n;J_{\mathcal{K}})$$

Distributed Information Bottleneck



- Information Bottleneck introduced by *[Tishby'99]* and *[Witsenhausen'80]* "Indirect Rate Distortion Problems", IT–26, no. 5, pp. 518–521, Sept. 1980.
- It is a CEO source-coding problem under log-loss!

Theorem (Distributed Information Bottleneck [Estella-Zaidi, IZS'18])

The D-IB region is the set of all tuples $(\Delta, R_1, \ldots, R_K)$ *which satisfy*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(Y_k; U_k | X, Q)] + I(X; U_{\mathcal{S}^c} | Q), \qquad \text{for all } \mathcal{S} \subseteq \mathcal{K}$$

for some joint pmf $p(q)p(x) \prod_{k=1}^{K} p(y_k|x) \prod_{k=1}^{K} p(u_k|y_k, q)$.

Vector Gaussian Distributed Information Bottleneck

- $(\mathbf{Y}_1, \cdots, \mathbf{Y}_K, \mathbf{X})$ jointly Gaussian, $\mathbf{Y}_k \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^M$,

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X} + \mathbf{N}_k, \qquad \mathbf{N}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}_k})$$

- Optimal encoding $P^*_{U_k|Y_k}$ is Gaussian and $Q = \emptyset$ *[Estella-Zaidi'17]*

Theorem ([Estella-Zaidi, IZS'18], [Ugur-Aguerri-Zaidi, arxiv:1811.03933])

If $(\mathbf{X}, \mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ are jointly Gaussian, the D-IB region is given by the set of all tuples $(\Delta, R_1, \ldots, R_L)$ satisfying that for all $\mathcal{S} \subseteq \mathcal{K}$

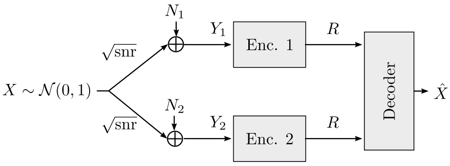
$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k + \log |\mathbf{I} - \mathbf{B}_k|] + \log \left| \sum_{k \in \mathcal{S}^c} \bar{\mathbf{H}}_k^H \mathbf{B}_k \bar{\mathbf{H}}_k + \mathbf{I} \right|$$

for some $\mathbf{0} \preceq \mathbf{B}_k \preceq \mathbf{I}$, where $\bar{\mathbf{H}}_k = \mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2} \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}}^{1/2}$, and achievable with

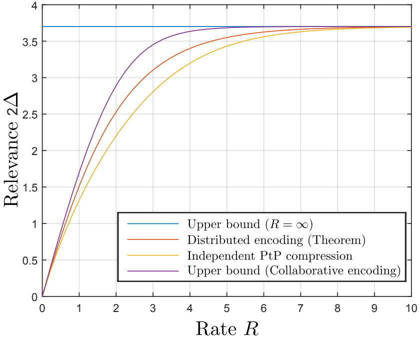
$$p^*(\mathbf{u}_k | \mathbf{y}_k, q) = \mathcal{CN}(\mathbf{y}_k, \Sigma_{\mathbf{n}_k}^{1/2} (\mathbf{B}_k - \mathbf{I}) \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2})$$

- Reminiscent of the sum-capacity in Gaussian Oblivious CRAN with Constant Gaussian Input constraint.

Example



$Y_k = \sqrt{\text{snr}} X + N_k$, with $N_k \sim \mathcal{N}(0,1)$, $k = 1,2$



- Optimal information (relevance):

$$\Delta^*(R, \text{snr}) = \frac{1}{2} \log \left(1 + 2\,\text{snr}\, \exp(-4R) \left(\exp(4R) + \text{snr} - \sqrt{\text{snr}^2 + (1 + 2\,\text{snr}) \exp(4R)}\right)\right)$$

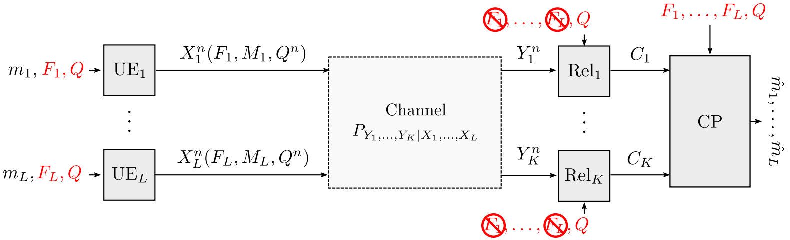
- Collaborative encoding upper bound: (Y_1, Y_2) encoded at rate $2R$

$$\Delta_{\text{ub}}(R, \text{sr}) = \frac{1}{2} \log(1 + 2\,\text{snr}) - \frac{1}{2} \log\left(1 + 2\,\text{snr}\, \exp(-4R)\right)$$

- Lower bound: Y_1 and Y_2 independently encoded

$$\Delta_{lb}(R, \text{snr}) = \frac{1}{2} \log\left(1 + 2\,\text{snr} - \text{snr}\, \exp(-2R)\right) - \frac{1}{2} \log\left(1 + \text{snr}\, \exp(-2R)\right)$$

Oblivious Relay Processing-CRAN System



- Resource-sharing random variable Q^n available at all terminals *[Simeone et al'11]*.
- Q^n way easier to share, (e.g., on/off activity).
- Memoryless Channel: $P_{Y_1,\ldots,Y_K|X_1,\ldots,X_1}$
- User $l \in \{1,\ldots,L\}$: $\phi_l^n : [1, |\mathcal{X}_l|^{n2^{nR_l}}] \times [1, 2^{nR_l}] \times \mathcal{Q}^n \to \mathcal{X}_l^n$
- Relay $k \in \{1,\ldots,K\}$: $g_k^n : \mathcal{Y}_k{}^n \times \mathcal{Q}^n \to [1, 2^{nC_k}]$
- Decoder:

$$\psi^n : [1, |\mathcal{X}_1|^{n2^{nR_1}}] \times \cdots \times [1, 2^{nC_K}] \times \mathcal{Q}^n \to [1, 2^{nR_1}] \times \ldots \times [1, 2^{nR_L}]$$

Capacity Region of a Class of CRAN Channels

Theorem (Aguerri-Zaidi-Caire-Shamai 'IT19)

For the class of discrete memoryless channels satisfying

$$Y_k -\!\!\ominus\!\!- X_{\mathcal{L}} -\!\!\ominus\!\!- Y_{\mathcal{K}\setminus k}$$

with oblivious relay processing and enabled resource-sharing, a rate tuple $(R_1, \ldots, R_L)$ *is achievable if and only if for all* $\mathcal{T} \subseteq \mathcal{L}$ *and for all* $\mathcal{S} \subseteq \mathcal{K}$*,*

$$\sum_{t \in \mathcal{T}} R_t \leq \sum_{s \in \mathcal{S}} [C_s - I(Y_s; U_s | X_{\mathcal{L}}, Q)] + I(X_{\mathcal{T}}; U_{\mathcal{S}^c} | X_{\mathcal{T}^c}, Q),$$

for some joint measure of the form

$$P_Q \prod_{l=1}^{L} P_{X_l|Q} \prod_{k=1}^{K} P_{Y_k|X_{\mathcal{L}}} \prod_{k=1}^{K} P_{U_k|Y_k,Q},$$

with the cardinality of Q *bounded as* $|\mathcal{Q}| \leq K + 2$*.*

$\Rightarrow$ *Equivalent to Noisy Network Coding [Lim-Kim-El Gamal-Chung, IT '11].*

$\Rightarrow$ *Directly related to quantize-map-forward (QMF) [Avestimehr-Diggavi-Tian-Tse, FnT'15, and references therein].*

Some Perspectives

- Optimal input distributions for the input power constrained Gaussian bottleneck. Discrete signaling is already known to sometimes outperform Gaussian signaling for single-user Gaussian CRAN *[Sanderovich-Shamai-Steinberg-Kramer '08]*.

- It is conjectured that the optimal input distribution is discrete.

- Universal Distortion: $X \in \mathcal{X}$ – features, V – observation, $\ell\Big(X, f(V)\Big)$ distortion, $f(V) \in \mathcal{X}$ – estimate:

  $f^*(\cdot)$ optimal estimate: $L^*(X|V) = \inf\limits_{f(\cdot)} \mathbb{E}\,\ell\Big(X, f(V)\Big)$

  Example: MMSE – $\mathbb{E}\Big(X - f^*(V)\Big)^2$, $f^*(V) = \mathbb{E}(X|V)$

  $\| \ell \|_\infty = \sup \ell(\cdot,\cdot)$, $L^*(X|Y)$ – σ subGaussian or $\ell(\cdot,\cdot)$ uniformly bounded.

$$\Rightarrow L^*(X|U) - L^*(X|Y) \le \frac{\| \ell_\infty \|}{\sqrt{2}} I(Y;X|U)$$

$$= \frac{\| \ell_\infty \|}{\sqrt{2}} \{I(X;Y) - I(X;U)\}\,,\ X - Y - U\,,\ \text{[Linder, 20]}$$

  $IB \Rightarrow \max I(X;U)$, $I(Y;U) \le R$ relevant to any distortion measure.

Some Perspectives cont.'

- Two sided Information Bottleneck: For: $V - X - Y - U$, find:

$$\max\ I(U;V) \text{ subjected to: } I(V;X) \leq R_1,\ I(U;Y) \leq R_2\,.$$

- Entropy constaint bottleneck: $X - Y - U$ $\max\ I(X;U)$ under the constraint $H(U) \leq R$ practical applications: LZ distortionless compression.
  $\Rightarrow U = f(Y)$ is a deterministic function *[Homri-Peleg-Shamai, TCOM, Nov.'18]*.

- The The deterministic bottleneck: advantages in complexity as compared to a classical bottleneck: *[Strouse-Schwab, Neural Comp.'17]*.

Some Perspectives cont.'

- Privacy Funnel, dual of bottleneck: $X - Y - U$, minimize: $I(X;U)$, under the constraint: $I(Y;U) = R$. *[Calmon-Makhdoumi-Medard-Varia-Christiansen-Duffy IT2017]*.
  - Direct connection to Information combining, maximize: $I(Y;U,X) = I(X;Y) + I(U;Y) - I(U;X)$, under the constraint: $I(U;Y) = R$.
  - Example: (X,Y) binary symmetric connected via a BSC, $X - Y$. The channel $Y - U$ is an Erasure Channel.
  - Example (Ordentlich-Shamai): For the Gaussian model: $Y = \sqrt{(\mathrm{snr})}\,X + N$, where (X,N) are unit norm independent Gaussians: Take U to be a deterministic function of Y, say describes the m last digits of a b long $(b \to \infty)$ binary description of Y, such that $I(U;Y) = H(U) = R$ (m is R dependent). Evidently $I(U;X) \to 0$, as $I(Y;U,X) \to R + I(X;Y)$.
  - Helper problem *[Bross-Lapidoth, ITW2019]*: $Y = X + N, X, N$ independent finite differential entropy. Noise helper: $I(N;U) = R$. Direct solution via information combining (Ordentlich-Shamai): We have: $Y - N - U$, and (example above): $I(N;Y,U) = I(N;Y) + R \Rightarrow I(X;Y,U) = I(X;Y) + R$.

Thank you!