The Information Bottleneck: A Unified Information Theoretic View

Shlomo Shamai

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


2020 Workshop on Machine Learning for Communication (MLCOM–2020)
Israel, September 7/8, 2020
Outline

- **Information Bottleneck:**
- **Information Bottleneck in Deep/Machine Learning in Wireless Networks:**
- **Connections:**
  - Remote Source Coding.
  - Common Reconstruction.
  - Information Combining.
  - Wyner-Ahlswede-Korner Problem.
  - Efficiency of Investment Information.
  - Hypothesis Testing.
  - Compound Wiretap Channel.
- **Distributed Information Bottleneck:**
  - CEO Source Coding Problem under Log-Loss.
  - Oblivious Relay Processing, CRAN.
  - Distributed Information Bottleneck for Learning.

- **Some Perspectives**
- **References**
Information Bottleneck

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

  **Rate (or Complexity):** $I(U; Y)$  
  **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].
Information Bottleneck in Deep/Machine Learning: Wireless Networks

- A theoretical tool to address unified strategies for communications:
  - Joint source channel coding; Channel state estimation; Universal Decoding; Modulation (de)/ Coding (de)/Equalization, Scheduling, Access, Resources (Frequency/Power/Bandwidth/Space), Muti-User Communications . . .

Relevant Machine Learning Aspects in Wireless Network: Some Overviews


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 $U = f(Y)$ measured by the pair

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

Example:

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

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

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


Solution also coined as the Information Bottleneck Method [Tishby’99]

\[ L_{IB}(\beta, P_{X,Y}) = \min_{p(u|y)} I(Y; U) - \beta I(X; U) \]
**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).

Other Connections (Cont.)

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

\[
I(U; X) = I(U; Y) - I(U; Y|X) \\
\leq R - I(U; Y|X)
\]


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

Other Connections (Cont.)

- **Information Combining**

\[ I(Y;U,X) = I(U;Y) + I(X;Y) - I(U;X) \quad \text{(since } X \perp Y \perp U) \]

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


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:

\[
\begin{align*}
H_0 : & \quad P_{X,Y} : (X,Y) \text{ Independent random variables.} \\
H_1 : & \quad P_{X,Y} : (X;Y) \text{ Dependent random variables.}
\end{align*}
\]

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

⇒ For \(n \to \infty\), the Stein error exponent (normalized by \(n\), of the Neuman-Pearson type II error: (the sequences were governed by \(H_0\), while the decision was \(H_1\)), 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 \(H_1\), while the decision was \(H_0\)) \(\leq \varepsilon\).

Compound Wiretap Channel

\[ X - Y - U, \text{ (} X \text{-input, } Y \text{-legitimate receiver, } U \text{-eavesdropper).} \]

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.
Elegant Proofs of Classical Bottleneck Results

- $X, Y$ binary symmetric connected through a Binary Symmetric Channel (error probability $e$): $U-Y$, also a BSC, $I(U;X) = \{1 - h(e^*v)\}$ where $e^*v = e(1 - v) + v(1 - e)$, $R = 1 - h(v)$.

Directly extends to $X - Y$ symmetric, where $Y$ is symmetric binary (one bit output quantization).

- $X$ standard Gaussian, and $Y = \sqrt{\text{snr}}X + N$ ($N$ standard Gaussian). Elegant proof via I-MMSE [Guo-Shamai-Verdu, FnT’13].

$$I(U;X) = \frac{1}{2} \log(1 + \text{snr}) - \frac{1}{2} \log\left(1 + \text{snr} \exp(-2R)\right)$$
Other Connections (Cont.)

**Proof:**

\[
\min I(Y; X, U) \text{ subject to: } I(Y; U) = R.
\]

Let

\[
\tilde{X} = \sqrt{1 + \text{snr}} X = \sqrt{\beta} Y + M, \quad M \sim N(0, 1)
\]

\[\beta = \text{snr}/(1 + \text{snr})\]

\[
I(Y; X, U) = I(Y; \tilde{X}, U) = I(Y; U) + I(Y; \tilde{X}|U)
\]

\[
I(Y; X|U) = I(Y; \tilde{X}|U) = \frac{1}{2} \int_0^\beta \text{mmse} (Y : \gamma, U) \, d\gamma
\]

\[
\text{mmse} (Y : \gamma, U) = E\left( Y - E(Y|\sqrt{\gamma} Y + M, U) \right)^2
\]
Other Connections (Cont.)

- **I-MMSE + Single Crossing Property**
  
  [Guo-Shamai-Verdú, FnT’13] \[\Rightarrow\]

\[
\frac{1}{2} \int_0^\beta \text{mmse}(Y : \gamma, U) \, d\gamma = \frac{1}{2} \int_0^\beta \frac{\rho \sigma_{Y|U}^2}{1 + \gamma \rho \sigma_{Y|U}^2} \, d\gamma \\
= \frac{1}{2} \log \left(1 + \beta \rho \sigma_{Y|U}^2\right)
\]

\[0 \leq \rho \leq 1\, , \quad \sigma_{Y|U}^2 = E \left( Y - E(Y|U) \right)^2 = \text{mmse}(Y : 0, U)\]
Other Connections (Cont.)

\[ R = I(Y; U) = h(Y) - h(Y|U) \]
\[ h(Y) = \frac{1}{2} \log \left( 2\pi \exp (\text{snr} + 1) \right) \]

\[ h(Y|U) = \frac{1}{2} \int_{0}^{\infty} \left( \text{mmse} \left( Y : \gamma, U \right) - \frac{1}{2\pi \rho + \gamma} \right) d\gamma \]
\[ \leq \text{single crossing point} \quad \frac{1}{2} \int_{0}^{\infty} \left( \frac{\rho \sigma_Y^2}{1 + \gamma \rho \sigma_Y^2} - \frac{1}{2\pi e + \gamma} \right) d\gamma \]
Other Connections (Cont.)

\[ \Rightarrow \quad \rho \sigma^2_{Y|U} \geq \exp(-2R)(1 + \text{snr}) \]

\[ \Rightarrow \quad \text{information combining} \quad I(Y; X, U) = I(Y; \tilde{X}, U) \geq R + \frac{1}{2} \log(1 + \text{snr} \exp(-2R)) \]

\[ \Rightarrow \quad \text{bottleneck} \quad I(X; U) \leq \frac{1}{2} \log(1 + \text{snr}) - \frac{1}{2} \log(1 + \text{snr} \exp(-2R)) \]

- Directly extends to the Gaussian vector case, where the vector version of the single crossing point [Bustin-Payaro-Palomar-Shamai, IT13] is used.
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

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

So, we get

$$\max_{p(u|y) : I(U;X) \leq R} I(U;X) = H(X) - R^*_X(R)$$

Vector Gaussian Information Bottleneck

- \((X, Y)\) jointly Gaussian, \(X \in \mathbb{R}^N\) and \(Y \in \mathbb{R}^M\)

- Optimal encoding \(P_{U|Y}\) is a noisy linear projection to a subspace whose dimensionality is determined by the bottleneck Lagrangian multiplier \(\beta\)

\[ U = AY + Z, \quad Z \sim \mathcal{N}(0, I) \]

where

\[
A = \begin{cases} 
[0^T; \ldots; 0^T], & \text{if } 0 \leq \beta \leq \beta_1^c \\
[\alpha_1 v_1^T; 0^T; \ldots; 0^T], & \text{if } \beta_1^c \leq \beta \leq \beta_2^c \\
[\alpha_1 v_1^T; \alpha_2 v_2^T; 0^T; \ldots; 0^T], & \text{if } \beta_2^c \leq \beta \leq \beta_3^c \\
& \vdots 
\end{cases}
\]

and \(\{v_1^T, \ldots, v_N^T\}\) are the left eigenvectors of \(\Sigma_{y|x} \Sigma_y^{-1}\), sorted by their ascending eigenvalues \(\{\lambda_1, \ldots, \lambda_N\}\); \(\beta_i^c = 1/(1 - \lambda_i)\) are critical \(\beta\) values; \(r_i = v_i^T \Sigma_y v_i\) 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 \([\text{Courtade-Weissman’14}]\)

- Key step: log-loss admits a lower bound in the form of conditional entropy of the source conditioned on the compression indices:

\[ nD \geq \mathbb{E}[d_{\log}(X^n; \hat{X}^n)] \geq H(X^n | J_K) = H(X^n) - I(X^n; J_K) \]
Converse of Theorem 1 for Oblivious CRAN leverages on this relation applied to multiple channel inputs, which can be designed. Multiple description CEO problem-logloss distortion [Pichler-Piantanida-Matz, ISIT’17].

Vector Gaussian CEO Problem Under Logarithmic Loss and Applications [Ugur-Aguerri-Zaidi, arxiv:1811.03933]: Accounts also for Gaussian side information about the source at the decoder.

- Full characterization (not the case for MMSE Distortion, [Ekrem-Ulukos, IT0214]).

Implications [Ugur-Aguerri-Zaidi, arxiv:1811.03933] Solutions of:

- Vector Gaussian distributed hypothesis testing against conditional independence [Rahman-Wagner, IT2012].

- A quadratic vector Gaussian CEO problem with determinant constraint.

- Vector Gaussian distributed Information Bottleneck Problem.
Distributed Information Bottleneck

- 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

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

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

For every \((\Delta, R_{\text{sum}}) \in \mathbb{R}^2_+\) that lies on the boundary of the optimal relevance complexity region there exist \(s \geq 0\) such that \((\Delta, R_{\text{sum}}) = (\Delta_s, R_s)\), with

\[
\Delta_s = \frac{1}{(1 + s)} \left[ (1 + sK)H(X) + sR_s + \max_{P} \mathcal{L}_s(P) \right]
\]

\[
R_s = I(X; U^*_K) + \sum_{k=1}^{K} \left[ I(Y_k; U^*_k) - I(X; U^*_k) \right]
\]

and \(P^*\) is the set of conditional pmfs \(P\) that maximize the cost function

\[
\mathcal{L}_s(P) := -H(X | U_K) - s \sum_{k=1}^{K} \left[ H(X | U_k) + I(Y_k; U_k) \right].
\]
A Variational Bound

- Let $L^{\text{VB}}_s(P, Q)$ denote

$$\mathbb{E}[\log Q_{X|U_k}(X|U_k)] + s \sum_{k=1}^{K} \left( \mathbb{E}[\log Q_{X|U_k}(Y|U_k)] - D_{\text{KL}}(P_{U_k|Y_k} \parallel Q_{U_k}) \right).$$

- It is not difficult to see that

$$\max_P L_s(P) = \max_P \max_Q L^{\text{VB}}_s(P, Q)$$

and

$$Q^*_{U_k} = P_{U_k}, \quad Q^*_{Y|U_k} = P_{Y|U_k}, \quad Q^*_{Y|U_1,\ldots,U_k} = P_{Y|U_1,\ldots,U_K}$$
Let
\[ \mathcal{L}^{\text{NN}}_s(\theta, \phi, \varphi) := \mathbb{E}_{P_Y, X} \mathbb{E}\{P_{\theta_k}(U_k|X_k)\} \left[ \log Q_{\phi_k}(Y|U_k) 
+ s \sum_{k=1}^K \left( \log Q_{\phi_k}(Y|U_k) - D_{\text{KL}}(P_{\theta_k}(U_k|X_k) || Q_{\varphi_k}(U_k)) \right) \right]. \]

We have
\[ \max_P \max_Q \mathcal{L}^{\text{VB}}_s(P, Q) \geq \max_{\theta, \phi, \varphi} \mathcal{L}^{\text{NN}}_s(\theta, \phi, \varphi) \]

Optimization through Stochastic Gradient Descendent
Vector Gaussian Distributed Information Bottleneck

- \((Y_1, \cdots, Y_K, X)\) jointly Gaussian, \(Y_k \in \mathbb{R}^N\) and \(X \in \mathbb{R}^M\), 
  \[Y_k = H_k X + N_k, \quad N_k \sim \mathcal{N}(0, \Sigma_{n_k})\]
- Optimal encoding \(P^*_{U_k|Y_k}\) is Gaussian and \(Q = \emptyset\) [Estella-Zaidi’17]

**Theorem ([Estella-Zaidi, IZS’18], [Ugur-Aguerri-Zaidi, IT July 2020])**

If \((X, Y_1, \ldots, Y_K)\) 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 \(S \subseteq K\)

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

for some \(0 \preceq B_k \preceq I\), where \(\bar{H}_k = \Sigma_{n_k}^{-1/2} H_k \Sigma_x^{1/2}\), and achievable with

\[
P^*(u_k|y_k, q) = \mathcal{CN}(y_k, \Sigma_{n_k}^{1/2}(B_k - I) \Sigma_{n_k}^{1/2})
\]

- Reminiscent of the sum-capacity in Gaussian Oblivious CRAN with Constant Gaussian Input constraint.
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 \(2R\)

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

Lower bound: \(Y_1\) and \(Y_2\) independently encoded

\[ \Delta_{lb}(R, \text{snr}) = \frac{1}{2} \log \left( 1 + 2 \text{snr} - \text{snr} \exp(-2R) \right) - \frac{1}{2} \log \left( 1 + \text{snr} \exp(-2R) \right) \]
The Cost of Oblivious Processing: an Example
Cut-Set Bound

\[ \sum (R, snr) = \min \left\{ 2R, \frac{1}{2} \log (1 + 2snr), \ R + \frac{1}{2} \log (1 + snr) \right\} \]

- **Improved Upper Bound:** geometric analysis of typical sets (equivalent in this case to the “information constrained transportation inequality”)
  
  [Wu-Ozgur-Peleg-Shamai, ITW’19 ]

There exists: \( \theta \in E[\arcsin(2^{-R}), \pi/2] \) such that:

\[ \sum (R, snr) \leq \frac{1}{2} \log (1 + snr) + R + \log \sin \theta, \]

\[ \sum (R, snr) \leq \frac{1}{2} \log (1 + snr) + \min_{\omega \in \left[\frac{\pi}{2} - \theta, \frac{\pi}{2}\right]} h(\omega; \theta) \]

\[ \sum (R, snr) \leq 2R + 2 \log \sin \theta \]

where

\[ h(\omega; \theta) = \frac{1}{2} \log \left( \frac{[2snr + \sin^2 \omega - 2snr \cos \omega] \sin^2 \theta}{(snr + 1)(\sin^2 \theta - \cos^2 \theta)} \right). \]
The Cost of Oblivious Processing: an Example
Cut-Set Bound (Cont).

**Achievable Scheme**

- Optimization (optimized time sharing)
  
  → Fully decode & forward (both relays decode) & rate splitting over the fronthaul links.

  → Optimal oblivious processing (distributed source coding under logarithmic loss).

  → Capacity achieving for: \( 2R \leq \frac{1}{2} \log (1 + \text{snr}) \).
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:

$$P_{Y_1, \ldots, Y_K | X_1, \ldots, X_L}$$

User $l \in \{1, \ldots, L\}$:

$$\phi^n_l : [1, |X_l|^{n2^{nR_l}}] \times [1, 2^{nR_l}] \times Q^n \rightarrow X_l^n$$

Relay $k \in \{1, \ldots, K\}$:

$$g^n_k : y_k^n \times Q^n \rightarrow [1, 2^{nC_k}]$$

Decoder:

$$\psi^n : [1, |X_1|^{n2^{nR_1}}] \times \cdots \times [1, 2^{nC_K}] \times Q^n \rightarrow [1, 2^{nR_1}] \times \cdots \times [1, 2^{nR_L}]$$
Theorem (Aguerri-Zaidi-Caire-Shamai 'IT19)

For the class of discrete memoryless channels satisfying

\[ Y_k \rightarrow X_\mathcal{L} \rightarrow Y_{\mathcal{K}\setminus k} \]

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 \(\mathcal{S} \subseteq \mathcal{K}\),

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

for some joint measure of the form

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

with the cardinality of \(Q\) bounded as \(|Q| \leq K + 2\).

⇒ Equivalent to Noisy Network Coding [Lim-Kim-El Gamal-Chung, IT '11].
⇒ Directly related to quantize-map-forward (QMF) [Avestimehr-Diggavi-Tian-Tse, FnT'15, and references therein].
Numerical Example

- Three-cell SISO circular Wyner model

- Each cell contains a single-antenna and a single-antenna RU.
- Inter-cell interference takes place only between adjacent cells.
- The intra-cell and inter-cell channel gains are given by 1 and $\alpha$, respectively.
- All RUs have a fronthaul capacity of $C$. 
Numerical Example (Cont.)

- Compare the following schemes [Park-Simeone-Sahin-Shamai ’14]
  - Single-cell processing
    - Each RU decodes the signal of the in-cell MS by treating all other MSs’ signals as noise.
  - Point-to-point fronthaul compression
    - Each RU compresses the received baseband signal and the quantized signals are decompressed in parallel at the control unit.
  - Distributed fronthaul compression [dCoso-Simoens ’09]
    - Each RU performs Wyner-Ziv coding on the received baseband signal and the quantized signals are successively recovered at the control unit.
    - Joint Decompression and Decoding (noisy network coding [Sanderovich-Shamai-Steinberg-Kramer’08])
  - Compute-and-forward [Hong-Caire ’11]
    - Each RU performs structured coding.
  - Oblivious processing upper bound
    - RUs cooperate and optimal compression is done over $3C$ fronthaul link.
  - Cutset upper bound [Simeone-Levy-Sanderovich-Somekh-Zaidel-Poor-Shamai ’12]
\[ \alpha = \frac{1}{\sqrt{2}} \text{ and } C = 3 \text{ bit/s/Hz} \]

- The performance advantage of distributed compression over point-to-point compression increases as SNR grows larger.
  - At high SNR, the correlation of the received signals at RUs becomes more pronounced.

- Compute-and-Forward
  - At low SNR, its performance coincides with single-cell processing.
  - RUs tend to decode trivial combinations.
  - At high SNR, the fronthaul capacity is the main performance bottleneck, so CoF shows the best performance.
Numerical Example (Cont.)

\[ \alpha = \frac{1}{\sqrt{2}} \text{ and } C = 3 \text{ bit/s/Hz} \]

- Distributed compression
- Joint decompression and decoding does not provide much gain compared to separate decompression and decoding.
- Optimality of joint decompression and decoding in symmetric case [Zaidi-Aguerri-Caire-Shamai’19].
Numerical Example (Cont.)

\[ \alpha = \frac{1}{\sqrt{2}} \text{ and } C = 5 \log_{10} P \text{ bit/s/Hz} \]

- When \( C \) increases as \( \log (\text{snr}) \), CoF is not the best for high SNR.
- i.e., if \( C \) does not limit the performance, the oblivious compression technique will be advantageous than CoF.
The Distributed Information Bottleneck for Learning

- For simplicity, we look at the D-IB under sum-rate [Aguerri-Zaidi'18]

\[
P^*_k|Y_k = \arg \min_{P_U|Y_k} I(X; U_K) + \beta \sum_{k=1}^{K} \left[ I(Y_k; U_k) - I(X; U_k) \right]
\]

- The optimal encoders-decoder of the D-IB under sum-rate constraint satisfy the following self consistent equations,

\[
p(u_k|y_k) = \frac{p(u_k)}{Z(\beta, u_k)} \exp \left(-\psi_s(u_k, y_k)\right),
\]

\[
p(x|u_k) = \sum_{y_k \in Y_k} p(y_k|u_k) p(x|y_k)
\]

\[
p(x|u_1, \ldots, u_K) = \sum_{y_K \in Y_K} p(y_K) p(u_K|y_K) p(x|y_K)/p(u_K)
\]

where

\[
\psi_s(u_k, y_k) := D_{KL}(P_X|y_k \parallel Q_X|u_k) + \frac{1}{s} E_{U_{\mathcal{K} \setminus k}|y_k} [D_{KL}(P_X|U_{\mathcal{K} \setminus k}, y_k \parallel Q_X|U_{\mathcal{K} \setminus k}, u_k)].
\]

- Alternating iterations of these equations converge to a solution for any initial \( p(u_k|x_k) \), similarly to a Blahut-Arimoto algorithm.
D-IB for Vector Gaussian Sources: Iterative Optimization

- \((Y_1, \cdots, Y_K, X)\) jointly Gaussian, \(Y_k \in \mathbb{R}^N\) and \(X \in \mathbb{R}^M\),

\[
Y_k = H_k X + N_k, \quad N_k \sim \mathcal{N}(0, I)
\]

- Optimal encoding \(P^*_U | Y_k\) is Gaussian [Aguerri-Zaidi'17] and given by

\[
U_k = A_k Y_k + Z_k, \quad Z_k \sim \mathcal{N}(0, \Sigma_{Z,k})
\]

- For this class of distributions, the updates in the Blahut-Arimoto type algorithm simplify to:

\[
\Sigma_{z_k}^{t+1} = \left( \left( 1 + \frac{1}{\beta} \right) \Sigma_{u_k}^{-1} | x - \frac{1}{s} \Sigma_{u_k}^{-1} | u_k \right)^{-1}
\]

\[
A_k^{t+1} = \Sigma_{z_k}^{-1} \left( \left( 1 + \frac{1}{\beta} \right) \Sigma_{u_k}^{-1} | x A_k (I - \Sigma_{y_k} | x \Sigma_{y_k}^{-1})
\]

\[
- \frac{1}{\beta} \Sigma_{u_k}^{-1} | u_k \right), A_k^t (I - \Sigma_{y_k} | u_k \Sigma_{y_k}^{-1}) \right).
\]
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.
  - Improved upper bounds (over cut-set) for non-oblivious relay based schemes, to better evaluate the cost of oblivious processing (á la: Vu-Barnes-Ozgur, arXiv:1701.02043 (IT’19) Gaussian primitive relay, [Wu-Ozgur-Peleg-Shamai, ITW’19]).
Some Perspectives cont.’

- Connections between classical bottleneck problems and Common Information [Wyner’75]: For given \((X, U)\) find \(Y : X \rightarrow Y \rightarrow U\) minimizing \(I(Y; X, U)\), and Gacs-Korner-Witsenhausen Common Information [Gacs-Korner ’73].

  - Lossy common information [Viswanatha-Akyol-Rose, IT2014].

  - Network source-coding [Gray-Wyner’74], viewed as a general common information characterization [El Gamal-Kim, Cambridge’15].

  - Gray-Wyner models with side information [Bennamar-Zaidi, Entropy’17].

Some Perspectives cont.’

- **Robust Information Bottleneck**
- Robustness is an important feature in bottleneck applications to deep learning.
- \((Y, U)\) joint random variables \(\Phi(U|Y)\)-statistical Fisher information
  \[
  = E_{Y,U} \left| \frac{\partial}{\partial Y} \log P(U|Y) \right|^2.
  \]

- **Robust bottleneck** \(X \rightarrow Y \rightarrow U\) (given \(P(X,Y)\))
  \[
  \max_{P(U|Y)} \left\{ I(X;U) - \beta I(Y;U) - \gamma \Phi(U|Y) \right\}
  \]
  - direct extensions to vector \((X,Y,U)\) spaces.
  - \((X,Y)\) jointly Gaussians \(\Rightarrow Y \rightarrow U\) Gaussian.
  - General \((X,Y)\) – stochastic gradient based algorithms.
  - MMSE based features: minimizing MMSE \((X|U)\) replaces maximizing \(I(X;U)\).
    \(\Rightarrow\) Connections: I-MMSE, De Bruijnis indentity, Cramer-Rao Inequality, Fano Inequality.
  - Strong Data Processing Inequalities for Input Constrained Additive Noise [Calmon-Polyansky-Wu, IT’18].
Some Perspectives cont.’


- A variety of related C-RAN & Distributed bottleneck problems:
  - Impact of block length $n$ [$R$ may not scale linearly with $n \Rightarrow$ Courtade conjecture ($R = 1$)] relates to [Courtade-Kumar, IT’14], [Yang-Wesel, arXiv:1807.11289, July’19], [Ordentlich-Shayevitz-Weinstein, ISIT’16].
    The $R = n - 1$ relates to [Huleihel-Ordentlich, arXiv:1701.03119v2, ISIT ’17].
  - Bandlimited time-continuous models [Homri-Peleg-Shamai, TCOM, Nov.’18], [Katz-Peleg-Shamai, COMCAS’19].
  - Broadcast Approach (oblivious and general) for the Information Bottleneck Channel [Steiner-Shamai, COMCAS’19], [Steiner-Shamai, ITA’20].
    - Channel State Information (CSI) availability and cost (fronthaul usage).
  - Gaussian version $\Rightarrow$ half space indicator [Kindler-O’Donnell-Witmer, arXiv July 2016].
Some Perspectives cont.’


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

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

\[
\begin{align*}
\hat{f}(\cdot) \text{ optimal estimate: } L^*(X|V) &= \inf_{f(\cdot)} \mathbb{E} \ell(X, f(V)) \\
\| \ell \|_{\infty} &= \sup \ell(\cdot, \cdot), \quad L^*(X|Y) - \sigma \text{ subGaussian or } \ell(\cdot, \cdot) \text{ uniformly bounded.}
\end{align*}
\]

Example: MMSE – \( \mathbb{E} \left( X - \hat{f}(V) \right)^2 \), \( \hat{f}(V) = \mathbb{E}(X|V) \)

\[
\| \ell \|_{\infty} = \sup \ell(\cdot, \cdot), \quad L^*(X|Y) - \sigma \text{ subGaussian or } \ell(\cdot, \cdot) \text{ uniformly bounded.}
\]

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

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

\( IB \Rightarrow \max I(X; U), \, I(Y; U) \leq R \) relevant to any distortion measure.
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 constraint 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 deterministic bottleneck: advantages in complexity as compared to a classical bottleneck: [Strouse-Schwab, Neural Comp.’17].
Privacy Funnel, dual of bottleneck: $X - Y - U$, minimize: $I(X; U)$, under the constraint: $I(Y; U) = R$. \cite{Calmon:2017aa}.

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: $Y = \sqrt{(\text{snr})} X + N$, 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 $I(Y; U, X) \to R + I(X; Y)$.


$$I(N; Y, U) = I(N; Y) + R \Rightarrow I(X; Y, U) = I(X; Y) + R.$$


M. Benammar and A. Zaidi, “Rate-Distortion of a Heegard-Berger Problem with Common Reconstruction Constraint,” Zürich Seminar on Communications and Information, Zürich, Switzerland, March 2–4, 2016.

M. Benammar and A. Zaidi, “Rate-Distortion Region of a Gray-Wyner Model with Side Information,” Entropy, Special Issue Rate-Distortion Theory and Information Theory, December 2017.


References (cont.)


References (cont.)


References (cont.)


S. Shamai (Shitz) and A. Steiner, “Broadcast Approach under Information Bottleneck Capacity Uncertainty”, Information Theory & Applications (ITA2020), San Diego, USA, February 2–7, 2020.


Abstract:
This talk focuses on connections between relatively recent notions and variants of the Information Bottleneck and classical information theoretic frameworks, such as: Remote Source-Coding; Information Combining; Common Reconstruction; The Wyner-Ahlsvede-Korner Problem; The Efficiency of Investment Information; CEO Source Coding under Log-Loss, Hypothesis Testing Error Exponent and others.
We overview the upink Cloud Radio Access Networks (CRAN) with oblivious processing, which is an attractive model for future wireless systems and highlight the basic connections to distributed Gaussian information bottleneck framework. For this setting, the optimal trade-offs between rates (i.e. complexity) and information (i.e. accuracy) in the discrete and vector Gaussian schemes is determined, taking an information-estimation viewpoint. Further, the performance cost of the simple 'oblivious' universal processing in CRAN systems is exemplified via novel bounding techniques.
The concluding overview and outlook addresses in a unified way the dual problem of the privacy funnel and recent observations on the additive noise channels with a helper. Connections to the finite block length bottleneck features (related to the Courtade-Kumar conjecture) and entropy complexity measures (rather than mutual-information) are shortly discussed. Some challenging problems are mentioned such as the characterization of the optimal power limited inputs ('features') maximizing the 'relevance' for the Gaussian information bottleneck, under 'complexity' constraints.

The talk is based mainly on joint work with A. Zaidi, I.E. Auguerri, G. Caire, O. Simeone and S-H. Park.
The research of S. Shamai is supported by the European Union's Horizon 2020 Research and Innovation Programme: No. 694630.
Thank you!