Information Bottleneck

- Goal of the information bottleneck method
- Information bottleneck method
- General conditions for the solution
- Gaussian Information Bottleneck (GIB)
- GIB problem statement
- GIB optimal projection
- GIB Information curve

Information Bottleneck Method

- **Goal:** Extracting relevant aspects of a source using a compressed representation

- Source variable X with $f_X(x)$
- Relevance variable Y with $f_Y(y)$
- Compressed representation T with $f_T(t)$

- Y, X, T **form a Markov chain in that order**

$$Y - X - T$$

- Note, Y and T are conditionally independent given X

Information Bottleneck Method

- Given the joint distribution of source and relevance variables $f_{X,Y}(x,y)$, the Information Bottleneck (IB)
  - operates to compresses X
  - into the compressed representation T,
  - while preserving information about Y.

- Stated as a variational problem:

$$\inf_{f_{T|X}} I(X;T) - \beta I(T;Y) \quad \text{for} \quad \beta > 0$$

Information Bottleneck Method

- Data processing inequality for Y – X – T:

$$
\begin{aligned}
I(T;X) &\geq I(Y;T) \\
I(X;T) - \beta I(T;Y) &\geq (1-\beta) I(T;Y)
\end{aligned}
$$

- Case $0 < \beta < 1$: (degenerated problem)

$$
\begin{aligned}
\inf I(X;T) - \beta I(T;Y) &= \inf (1-\beta) I(T;Y) \\
&= \inf I(T;Y) = 0
\end{aligned}
$$

$$I(T;Y) = I(X;T) = 0$$

- Case $\beta > 1$:

$$
\begin{aligned}
\inf I(X;T) - \beta I(T;Y) &= \inf (1-\beta) I(T;Y) \\
&= (1-\beta) \sup I(T;Y)
\end{aligned}
$$

Information Bottleneck: Interpretation

- Shannon's perspective:
  - Minimization of mutual information corresponds to optimal compression in rate distortion theory
  - Maximization of mutual information corresponds to optimal information transmission in noisy channel coding

- Machine learning perspective:
  - Regularized generative modeling
  - Minimization of $I(X;T)$ penalizes complex models
  - Maximization of $I(T;Y)$ optimizes an empirical likelihood of a special mixture model

General Conditions for the IB Solution

- Solve unconstrained problem $\inf_{f_{T|X}} J(X,T)$ with $\beta > 1$ and

$$
\begin{aligned}
J(X,T) &= E_{Y,X,T}\left\{\ln\frac{f_{T|X}}{f_T} - \beta\ln\frac{f_{Y|T}}{f_Y}\right\} \\
&= E_{X,T}\left\{\ln\frac{f_{T|X}}{f_T} + \beta E_{Y|X}\{\ln\frac{f_Y}{f_{Y|T}}\}\right\} \\
&= E_{X,T}\left\{\ln\frac{f_{T|X}}{f_T} + \beta E_{Y|X}\{\ln\frac{f_{Y|X}}{f_{Y|T}} + \ln\frac{f_Y}{f_{Y|X}}\}\right\} \\
&= E_{X,T}\left\{-\ln\left[\frac{f_T}{f_{T|X}}e^{-\beta D_{KL}(f_{Y|X}||f_{Y|T})-\beta a(x)}\right]\right\} \\
&\geq -\ln E_{X,T}\left\{\frac{f_T}{f_{T|X}}e^{-\beta D_{KL}(f_{Y|X}||f_{Y|T})-\beta a(x)}\right\}
\end{aligned}
$$

General Conditions for the IB Solution

- Jensen: Lower bound is tight, iff

$$\frac{f_T}{f_{T|X}} e^{-\beta D_{KL}(f_{Y|X}||f_{Y|T}) - \beta a(x)} = \text{const.}$$

- We obtain the conditional PDF:

$$f_{T|X}(t|x) = \frac{f_T(t)}{\mu(\beta, x)} e^{-\beta D_{KL}(f_{Y|X}||f_{Y|T})}$$

- Conditions from Bayes' rule:

$$f_T = \int f_{T|X} f_X dx$$

$$f_{Y|T} = \frac{1}{f_T} \int f_{Y,X} f_{T|X} dx$$

General Conditions for the IB Solution

- Note, $f_X(x)$ and $f_{X,Y}(x,y)$ are given.
- Note, $Y - X - T$ form a Markov chain and we have

$$f_{Y,X,T} = f_{Y|X} f_{T|X} f_X$$

- Hence, above conditions can be iterated directly in a Blahut-Arimoto like algorithm.

Remark: Entropy Power Inequality

- Let X be an n-dimensional continuous-valued random variable with differential entropy h(X).

- The entropy power of X is defined to be:

$$N(X) = \frac{1}{2\pi e} e^{\frac{2}{n} h(X)}$$

- Let X and Y be independent random variables, then

$$N(X+Y) \geq N(X) + N(Y)$$

- Equality holds, iff X and Y are **multivariate normal** random variables with **proportional covariance matrices**.

Gaussian Information Bottleneck

- Let X and Y be two jointly multivariate Gaussian variables of dimensions n_x and n_y, let C_{xx} and C_{yy} be the covariance matrices, and let C_{xy} be the cross-covariance matrix.

- The **entropy power inequality** shows that the optimum is obtained by a variable T which is also jointly Gaussian with X.

- The linear projection of X, which is also Gaussian, attains the maximum information.

$$T = AX + Z \quad \text{with} \quad Z \sim N(0, C_{ZZ})$$

GIB Problem Statement

- Optimize

$$\min_{A,C_{ZZ}} I(X;T) - \beta I(T;Y)$$

- Over the noisy linear transformations of A, C_{ZZ}

$$T = AX + Z \quad \text{with} \quad Z \sim N(0, C_{ZZ})$$

- T is normal distributed $T \sim N(0, C_{TT})$ with

$$C_{TT} = AC_{XX}A^T + C_{ZZ}$$

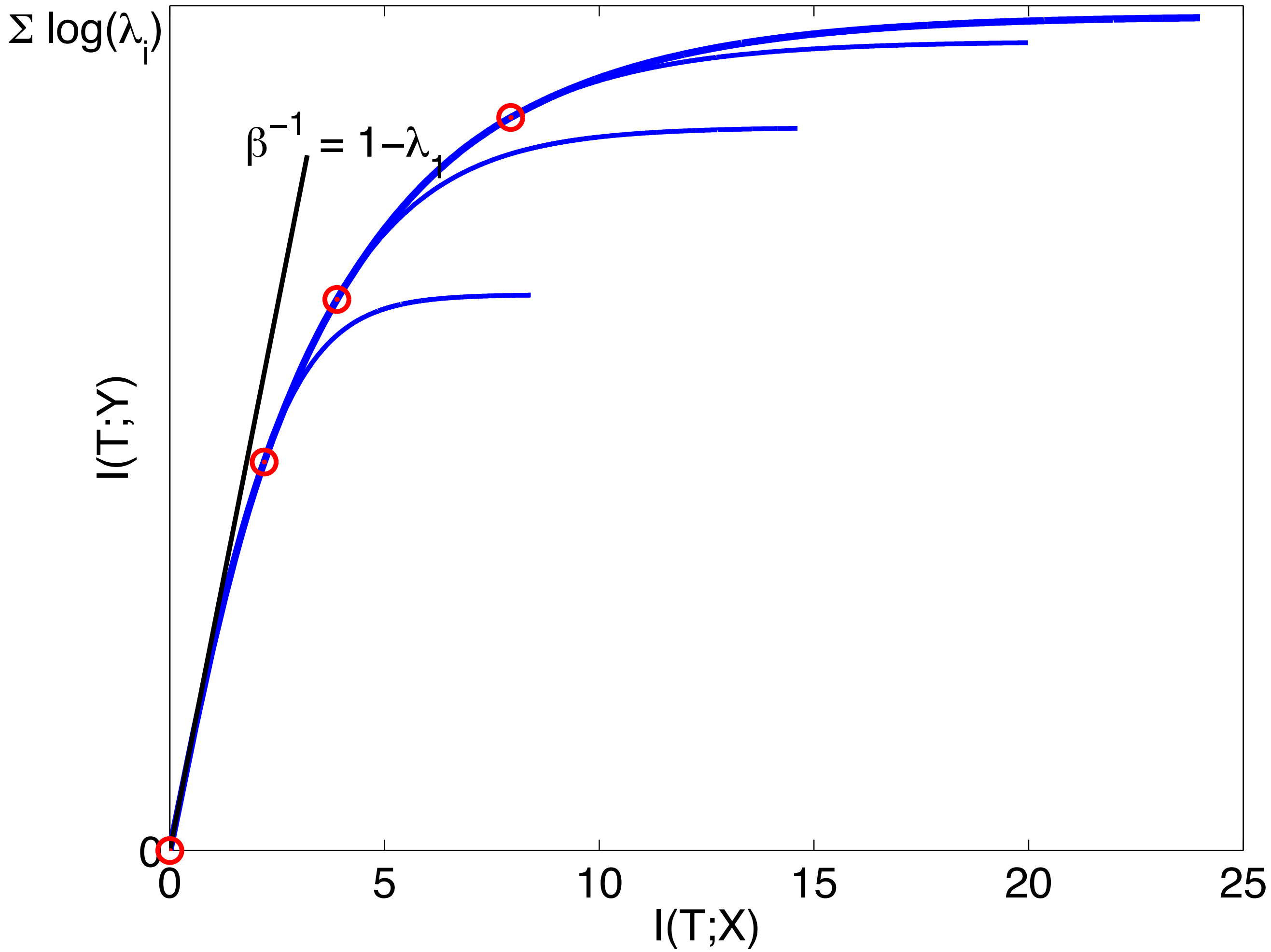
GIB Optimal Projection

- The optimal projection $T = AX + Z$ is given by $C_{ZZ} = I_x$ and

$$A = \begin{cases} \left[\mathbf{0}^T; \ldots; \mathbf{0}^T\right] & 0 \leq \beta \leq \beta^c{}_1 \\ \left[\alpha_1 \mathbf{v}_1^T, \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & \beta_1^c \leq \beta \leq \beta^c{}_2 \\ \left[\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & \beta^c{}_2 \leq \beta \leq \beta^c{}_3 \\ \vdots & \end{cases}$$

- v_i^T are left eigenvectors of $C_{X|Y} C_{XX}^{-1}$
- Sorted by ascending eigenvalues λ_I
- $\beta_i^c = \dfrac{1}{1 - \lambda_i}$ are critical β values
- $\alpha_i = \sqrt{\dfrac{\beta(1 - \lambda_i) - 1}{\lambda_i r_i}}$ with $r_i = v_i^T C_{XX} v_i$

GIB Information Curve



Deep Variational Information Bottleneck

- The original IB model as well as the Deep Variational Information Bottleneck (DVIB) assume the Markov chain

$$Y - X - T$$

- For the DVIB, the Markov chain $X - T - Y$ appears by construction.

Further Reading

- Chechik, Globerson, Tishby, and Weiss, Information Bottleneck for Gaussian Variables, Journal of Machine Learning Research, no. 6, pp. 165-188, 2005.
- Wieczorek, Roth, On the Difference between the Information Bottleneck and the Deep Information Bottleneck, Entropy, 22, 131, 2020.