The Information Bottleneck Method was proposed by Tishby, Pereira, and Bialek (Allerton, 1999). Lecture based on Goldfeld & Polyanskiy (2020) and Zaidi et al (2020).

An information-theoretic framework for learning.

Considers extracting information about a target signal $Y$ through a correlated observable $X$.

The extracted information is quantified by a variable $T = T(X)$ which is a possibly randomized function of $X$, thereby forming Markov chain

$Y - X - T$.

Objective is to find a $T$ that minimizes the mutual information $I(X;T)$ while keeping $I(Y;T)$ above a certain threshold.

Threshold determines how informative the representation $T$ is about $Y$.

$$\min \; I(X;T)$$
$$\text{s.t. } I(Y;T) \geq \alpha.$$

where minimization over all randomized mappings of $X$ to $T$.

This was originally motivated axiomatically rather than with a coding theorem, or a precise operational problem: want to keep as much relevant/informative info as possible.

$\rightarrow$ leads to a clustering algorithm.

IB framework concerned with finding $P_{T|X}$ that extracts information about $Y$, i.e. high $I(Y;T)$, while compressing $X$, quantified as keeping $I(X;T)$ small.

Data processing inequality in information theory for Markov chain $Y - X - T$ implies $I(Y;T) \leq I(X;Y)$, so compressed representation $T$ can't convey more information than original signal.

so, IB basically trying to find approximate version of minimal sufficient statistic

We pass the information $X$ contains about $Y$ through a bottleneck via the representation $T$.

$\Rightarrow$ claim is that this objective promotes minimality, sufficiency, and disentanglement of representations.

A slightly different form of the same ~~formulation~~ problem

$$\min_{P_{T|X} : H(X|T) \geq \alpha} H(Y|T)$$

was developed by Witsenhausen and Wyner (1975), in context of common information.

The original problem statement not convex, but can be made into a convex rate-distortion problem. Commonly solved by introducing Lagrange multiplier $\beta$ and considering the functional

$$\mathcal{L}_\beta(p_{T|X}) = I(X;T) - \beta I(T;Y).$$

Here $\beta$ controls amount of compression

→ small $\beta$ implies more compression

→ large $\beta$ leads to more informativeness / relevance.

The Lagrangian formulation then leads to a variational characterization of stationary point of $\mathcal{L}_\beta(p_{T|X})$

A stationary point $p^{(\beta)}_{T|X}$ must satisfy

① $$P_T^{(\beta)}(t) = \int_x p^{(\beta)}_{T|X}(t|x)\, dP_X(x)$$ (this is Radon-Nikodym derivative)

② $$P^{(\beta)}_{Y|T}(y|t) = \frac{1}{p^{(\beta)}_T(t)} \int_x P_{Y|X}(y|x)\, p^{(\beta)}_{T|X}(t|x)\, dP_X(x)$$

③ $$P^{(\beta)}_{T|X}(t|x) = \frac{P_T^{(\beta)}(t)}{Z_\beta(x)}\, e^{-\beta D_{KL}(P_{Y|X}(\cdot|x)\,\|\,P^{(\beta)}_{Y|T}(\cdot|t))}$$

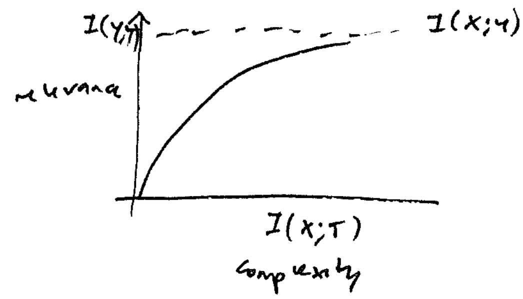
where $Z_\beta(x)$ is normalization constant (partition function).

If $X, Y, T$ take values in finite sets, and $P_{X,Y}$ known, then alternating iterations of $\begin{bmatrix} ① \\ ② \\ ③ \end{bmatrix}$ locally converge to a solution for any initial $P_{T|X}$.

Reminiscent of Blahut-Arimoto algorithm for rate distortion.

The IB curve is obtained by plotting

$(I_\beta(X;T),\ I_\beta(Y;T))$ for an optimal $P^{(\beta)}_{T|X}$ for each $\beta \in [0,\infty)$.




---

Come back to operational question of what problem is the information bottleneck the solution to? Looks a lot like a rate distortion problem with a particular information-based distortion function.

Consider the remote source coding

$$Y \longrightarrow \boxed{P_{X|Y}} \longrightarrow X \longrightarrow \boxed{enc} \xrightarrow{T} \boxed{dec} \longrightarrow \hat{Y}$$

with a distortion measure that is the log-loss. fidelity criterion.

The decoder generates a soft estimate $\hat{y}^n$ of $y^n$ in the form of a probability distribution over $\mathcal{Y}^n$, ie. $\hat{y} = \hat{P}_{Y^n|M}(\cdot)$.

The incurred discrepancy between $y^n$ and the estimation $\hat{y}^n$ under log-loss for the observation $x^n$ is then the per-letter log-loss distortion:

$\ell_{\log}(y,\hat{y}) = \log \frac{1}{\hat{y}(y)}$ for $y \in \mathcal{Y}$ and $\hat{y} \in \mathcal{P}(\mathcal{Y})$ designates a probability distribution on $\mathcal{Y}$ and $\hat{y}(y)$ is value of that distribution evaluated at outcome $y \in \mathcal{Y}$.

---

So want $E\left[\ell^{(n)}_{\log}(Y^n, \hat{Y}^n)\right] \leq D$

where incurred distortion between sequences $Y^n$ and $\hat{Y}^n$ is measured as

$$\ell^{(n)}_{\log}(y^n, \hat{y}^n) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\log}(y_i, \hat{y}_i).$$

It is known, rate-distortion region is $(R,D)$ that satisfy

$$R \geq I(U;X)$$
$$D \geq H(Y|U)$$

where ~~union~~ allowed is over all auxiliary random variables satisfying $U - X - Y$, and $|\mathcal{U}| \leq |\mathcal{X}|+1$.

Using substitution $\Delta = H(Y) - D$, get pairs $(R, H(Y)-D)$ that satisfy

$$R \geq I(U;X)$$
$$\Delta < I(U;Y)$$

for $U$ that satisfies $U - X - Y$ with $|\mathcal{U}| \leq |\mathcal{X}|+1$.

Also connected to WAK problem.

Sources $X$ and $Y$ encoded separately at rates $R_x$ and $R_y$

decoder recover $Y$ losslessly.

For given $R_x = R$, minimum rate $R_y$ to losslessly recover $Y$ is

$$R_Y^*(R) = \min_{P_{U|X} : I(U;X) \leq R} H(Y|U)$$

so we get

$$\max_{P_{U|X} : I(U;X) \leq R} I(U;Y) = H(Y) - R_Y^*(R).$$

So solving IB problem equivalent to solving WAK problem.