A NOTE ON INFORMATION BOTTLENECK PRINCIPLE

YIMING XU

This note is written to assist my own understanding of the material, therefore may be imprecise or contain mistakes.

1. RATE DISTORTION THEORY

This part is based on the chapter 2 and 14 of the book [Wil13] as well as the notes by Tsachy Weissman, see http://web.stanford.edu/class/ee376a/outline.html.

Let \( x \in [M] \) be a discrete random variable with probability mass function (pmf) \( p \). A sequence of i.i.d. copies of \( x \), \( x^{(n)} = \{x_i\}_{i \in [n]} \), is called the source sequence in the language of information theory. For convenience, one may consider \( x^{(n)} \) as a random message of \( n \) words with each word independently generated from the source \( x \). By the law of large numbers, the empirical distribution of \( x^{(n)} \) converges to \( p \) with probability 1, which ensures that for sufficiently large \( n \), with overwhelming probability \( \#\{i : x_i = m\} \approx p_m n \) for \( m \in [M] \). The set of sequences satisfying these approximate identities are called typical sequences, which for convenience is denoted by \( T_x \). Note that \( T_x \) also depends on the proximity parameter \( \delta \) which can be treated as fixed thus omitted in our notation. Using standard asymptotics argument, one can check that

\[
|T_x| \approx \left( \frac{n!}{\prod_{m=1}^{M} (p_m n)!} \right) \approx \frac{1}{\prod_{m=1}^{M} (p_m)^{p_m n}} \cdot \sqrt{\frac{2np_n}{\prod_{m=1}^{M} 2\pi p_m n}},
\]

where

\[
H(p) = -\sum_{m=1}^{M} p_m \log p_m = \mathbb{E}[-\log p].
\]

In communication theory, supposing one wants to transmit all the elements in \( T_x \) using binary encoding, the above calculation shows that about \( nH(p) + o(n) \) bits are needed. Averaging over \( n \), the randomness per word in the element requires approximately \( H(p) \) bits for encoding. \( H(p) \) is referred to as the entropy of \( p \), and provides a lower bound for the compressibility of the source sequence under which the risk of information loss is negligible. This is the renowned source coding theorem which was discovered by C. Shannon in 1948.

In some situations, one may wish to know how far \( H(p) \) can be pushed down provided that a limited amount of information loss is allowed. An abstract approach for modeling the lossy communication is to assume that the binary code of \( x^{(n)} \) in \( T_x \) is decoded by some element \( c(x^{(n)}) \) in a codebook \( C \) of size \( 2^{nR} \), where \( R > 0 \) is the analogue of \( H(p) \) in the lossless case and called the transmission rate. The goal here is to make \( R \) as small as possible while maintaining the average distance between \( x^{(n)} \) and \( c(x^{(n)}) \) below some given \( \varepsilon > 0 \). The average distance here is measured by the average distortion function

\[
D = \mathbb{E}[d(x^{(n)}, c(x^{(n)}))],
\]

where \( d \) is some distance function defined on \([M]^n \times [M]^n\). The following picture summarizes our goal in a brief diagram:

- **lossless communication:** \( T_x \xrightarrow{encoder} \{0,1\}^{nH(p)} \xrightarrow{encoder^{-1}} T_x \)
- **lossy communication:** \( T_x \xrightarrow{encoder} \{0,1\}^{nR} \xrightarrow{decoder} C \).

For lossless communication, the encoder is injective on \( T_x \) therefore the decoder is simply the inverse of the encoder. For lossy communication, however, since \( R \leq H(p) \), the encoder may not be invertible on \( T_x \). In this case, a natural strategy is to map typical sequences in \( T_x \) that are close in the metric \( d \) to a fixed element in \( \{0,1\}^{nR} \), then decode this element by any typical sequence.
mapped to it. This process is most economically done by taking an \( \varepsilon \)-\( d \) covering of \( T_x \). The resulting compression scheme has average distortion \( D \leq \varepsilon \) with transmission rate

\[
R \geq \log \mathcal{N}(T_x, d, \varepsilon),
\]

where \( \mathcal{N}(T_x, d, \varepsilon) \) is minimal number of balls of radius \( \varepsilon \) that cover \( T_x \), and \( \log \mathcal{N}(T_x, d, \varepsilon) \) is called the metric entropy of \( T_x \).

Unfortunately, finding the covering number for an arbitrary set is extremely difficult (NP hard). An approximate way to do this is via constructing a randomized codebook such that with overwhelming probability the codebook is ‘good’. Precisely, the codebook \( \mathcal{C} \) consists of \( 2^{nR} \) independent sequences of form \( y(n) = \{ y_i \}_{i \in [n]} \), and each sequence is generated by independently sampling its components from some pmf. Since several pmfs will be used after this point, to keep notation simple, we will use the same \( p \) to denote pmfs with its arguments indicating which variable(s) it is associated with.

Suppose that the distance between \( x(n) \) and \( y(n) \) is additive (such as the Hamming distance). Particularly, \( d(x(n), y(n)) = \sum_{i=1}^{n} d(x_i, y_i) \), where \( d \) is usually taken as an explicit distance function on \([M] \times [M]\), but can be appropriately generalized. Under these assumptions, the average distortion condition (with the multiple constant \( n \) omitted) simplifies to

\[
D = \mathbb{E}[d(x, y)] \leq \varepsilon,
\]

where the expectation is taken over the joint pmf \( p(x, y) \). Almost equivalently, for a jointly typical sequence \( (x(n), y(n)) \) (with respect to \( p(x, y) \)),

\[
\frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i) \leq \varepsilon.
\]

But how large should the codebook be chosen so that every \( x(n) \) can find an element \( c \in \mathcal{C} \) such that \( d(x(n), c) \leq \varepsilon \)? To answer this question we will need two facts:

- \( p( \cdot | y(n) \in T_y) \) is almost uniform on \( T_y \);
- For fixed \( x(n) \) and any random sequence \( y(n) \) drawn from \( p(y) \), the probability (under \( p(y) \)) that \( y(n) \) resembles a typical sequence of \( p(y|x(n)) \) (therefore \( (x(n), y(n)) \) looks like a jointly typical sequence of \( p(x, y) = p(x)p(y|x) \) is approxiamtely \( 2^{-nI(x, y)} \), where \( I(x, y) \) is the mutual information between \( x \) and \( y \) defined by,

\[
I(x, y) = \mathbb{E} \left[ \log \frac{p(y)}{p(y|x)} \right],
\]

where the expectation is taken over the joint pmf \( p(x, y) \). \( I \) appears in other places as well, such as the KL-divergence (relative entropy) between \( p(x, y) \) and \( p(x) \otimes p(y) \), or the transportation cost incurred by changing \( p(x) \otimes p(y) \) to \( p(x, y) \).

Therefore, for any \( p(x, y) \) (or \( p(y|x) \) since \( p(x) \) is known) satisfying the average distortion constraint, \( |\mathcal{C}| \) needs to be at least \( 2^{nI(x, y)} \), implying that \( R \geq I(x, y) \). A main result of the rate distortion theory tells us that this lower bound is achievable (which needs to be defined separately). Hence, the smallest \( R \) can be obtained by solving the following minimization problem:

\[
\min_{p(y|x)} I(x, y) \text{ s.t. } \mathbb{E}[d(x, y)] \leq \varepsilon,
\]

which is equivalent to the unconstrained minimization problem

\[
\min_{p(y|x)} I(x, y) + \beta \mathbb{E}[d(x, y)],
\]

where \( \beta > 0 \) depends on \( \varepsilon \). For fixed \( x \) and \( y \), taking derivative with respect to \( p(y|x) \), one can verify that the optimal solution satisfies

\[
p(y|x) = p(y) \times \frac{e^{-\beta d(x, y)}}{\sum_y e^{-\beta d(x, y)}}
\]

\[
p(y) = \sum_x p(y|x)p(x),
\]

where the second equation follows from calculating the marginal distribution \( p(y) \). Treating \( p(y) \) as a separate variable, one may use an alternating minimization algorithm to solve for \( p(y|x) \) over
A NOTE ON INFORMATION BOTTLENECK PRINCIPLE

the convex sets of the normalized distributions. This is known as the Blahut-Arimoto algorithm. It can be shown that iterations in the Blahut-Arimoto algorithm converge to a unique minimum in the convex sets of two distributions.

2. The information bottleneck principle

One of the restrictions in applying the rate distortion theory in practice is the ambiguity in choosing the right distortion function \( d \). A new variational principle called the Information Bottleneck Principle (IB) was proposed in the seminal paper [ST00], in which the loss functional is constructed from a slightly different perspective. For the random pair \((x, y)\), let \( x \) be the total information and \( y \) be the relevant information to be preserved in compression. One may compare this to the parametric inference in statistics, where \( x \) is the complete dataset and \( y \) is the sufficient statistics of the parameters of interest. Here we assume that \( p(x, y) \) is given. The goal is to best squeeze the information of \( x \) through a bottleneck \( p(z|x) \) to a random codebook \( z \) while keeping the information of \( y \) above some threshold. This constrained optimization can be equivalently stated as the following unconstrained minimization problem:

\[
(2.1) \quad \min_{p(z|x)} I(x, z) - \beta I(y, z),
\]

where \( \beta > 0 \) is some tunable parameter. It is shown in [ST00] that by assuming the two-way Markov chain condition:

\[
z \leftrightarrow x \leftrightarrow y,
\]

the optimal solution of the above problem satisfies

\[
p(z|x) = p(z) \times \frac{e^{-\beta \sum_y p(y|x) \log \frac{p(x|y)}{p(y)}}}{\sum_y e^{-\beta \sum_y p(y|x) \log \frac{p(x|y)}{p(y)}}} = \frac{p(z|x)}{p(z)},
\]

\[
p(y|z) = \sum_x p(y|x)p(x|z) = \sum_x p(x, y) \frac{p(z|x)}{p(z)}.
\]

The last two equations are standard from marginal distribution calculations. The first equation, with a closer look, is similar to the one derived in the rate distortion theory by identifying \( d(x, z) = D_{KL}(p(y|x) || p(y|z)) \), where \( D_{KL} \) is the KL-divergence. However, this \( d \) cannot be computed explicitly since it requires knowledge on \( p(z|x) \), which is the variable optimized over. In practice, one can treat \( p(y|z) \) as an additional separate variable from \( p(z|x) \) and \( p(z) \) and minimize the three terms in an alternating manner, just as in the Blahut-Arimoto algorithm. Similar convergence guarantees hold as well in this case.

3. Deep variational information bottleneck principle

A nice application of IB is in deep neural networks [TZ15]. In a deep neural network, suppose that the transformation between layers functions as an autoencoder, through which the relevant information of the data is preserved while its complexity (dimension) is reduced. This makes (2.1) a natural candidate objective minimized by the encoder, where the admissible set is a parametric family of conditional distributions. We use the same notation here as in (2.1), but allow \( x, y, z \) to be continuous and \( p \) to be density functions accordingly. A major drawback of the information bottleneck principle is that for general continuous distributions, finding the mutual information is computationally infeasible. Yet this is quite often in deep neural networks. To overcome this, a relaxation scheme IB called deep variational information bottleneck was proposed in [Ale+16]. The idea is based on the variational autoencoder (VAE) [KW13] and the relaxed objective has two major advantages. It can be approximated by the Monte Carlo sampling, and as a consequence, the stochastic gradient descent applies.
Suppose that $q(y|z)$ is a variational approximation to $p(y|z)$ and $r(z)$ is a variational approximation to the marginal $p(z)$. Write the objective function in (2.1) as

$$I(x, z) - \beta I(y, z) = \int p(x, y)p(z|x) \log p(z|x) dz dx dy - \int p(z) \log p(z) dz$$

$$- \beta \left( \int p(x, y)p(z|x) \log p(y|z) dz dx dy - \int p(y) \log p(y) dy \right).$$

We analyze the computability of each term on the right-hand side respectively. (a) can be evaluated via Monte Carlo sampling since $p(x, y)$ is known. (b) can be lower bounded by $\int p(z) \log r(z)$, if $p(z) = \int p(z|x)p(x) dx$ is expensive to compute. (c) is intractable since it requires evaluation of the unknown decoder $p(y|z)$. A common practice is to substitute $p(y|z)$ by $q(y|z)$ in the original equation to obtain a lower bound. (d) is the entropy of $y$ which is independent of $z$, thus can be neglected. Putting these ingredients together, we can upper bound $I(x, z) - \beta I(y, z)$ by

$$I(x, z) - \beta I(y, z) - \beta(d)$$

$$\leq \int p(x, y)p(z|x) \log \frac{p(z|x)}{r(z)} dz dx dy - \beta \int p(x, y)p(z|x) \log \frac{q(y|z)}{r(z)} dz dx dy$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \left( \int p(z|x_i) \log \frac{p(z|x_i)}{r(z)} dz - \beta \int p(z|x_i) \log \frac{q(y_i|z)}{r(z)} dz \right),$$

where the last step follows from the Monte Carlo approximation, with $\{(x_i, y_i)\}_{i \in [n]} \overset{\text{i.i.d.}}{\sim} p(x, y)$. (3.1) is the objective function to be minimized under the deep variational information bottleneck framework. To see its computational convenience, note that in a deep neural network, $p(z|x)$ belongs to a parametric family whose parameters are given by the transformations (usually nonlinear) of $x$. A reparameterization trick in [KW13] (decouple the randomness from the parameters) could be used to make the differentiation of parameters commute with taking expectations. This makes the stochastic gradient operator an unbiased estimator for the true gradient.

**References**


