Interpretation of the Erkip-Cover Strong Data Processing Inequality

Sudeep Kamath, Venkat Anantharam
e-mail: {sudeep, ananth}@eecs.berkeley.edu

February 18, 2013

Abstract

We provide a simple interpretation of the Erkip-Cover Strong Data Processing Inequality.

Consider any discrete random variable pair (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$ with probability distribution $P_{X,Y}(x, y)$, and marginals $P_X(x), P_Y(y)$. The following beautiful Strong Data Processing Inequality was proved in [1]:

Theorem 1

$$\sup_{W:W-X-Y} \frac{I(Y;W)}{I(X;W)} = \rho(X;Y)^2,$$

where $\rho(X;Y)$ *defined by* $\rho(X;Y) := \sup \mathbb{E} f(X)g(Y)$ *with the supremum taken over all real valued functions,* f, g *satisfying* $\mathbb{E}f(X) = \mathbb{E}g(Y) = 0$ *and* $\mathbb{E}f(X)^2 = \mathbb{E}g(Y)^2 = 1$, *is the Hirschfeld-Gebelein-Rényi maximal correlation of* X *and* Y.

Suppose we define T^* as a map from probability distributions on $\mathcal{X}$ to probability distributions on $\mathcal{Y}$ given by $T^*R = Q$ where $Q(y) = \sum_x \frac{P(x,y)}{P_X(x)} R(x)$. Then, the following theorem is shown in [2] with connections to hypercontractivity of a Markov operator.

Theorem 2

$$s^*(X;Y) := \sup_{R \neq P} \frac{D(T^*R||P_Y)}{D(R||P)} \geq \rho(X;Y)^2.$$

with strict inequality for some choices of $P_{X,Y}$.

The inequality in Theorem 2 can be shown by local approximation of the KL divergence in a neighborhood of P_X. We provide below an interpretation of Theorem 1 in light of Theorem 2 here. This authors have mentioned this in [1] but the purpose of this note is to state this connection explicitly.

Suppose we were to choose weights $\{w_i\}_{i=1}^k$ with $w_i \geq 0, i = 1, 2, \ldots, k, \sum_{i=1}^k w_i = 1$ and choose distributions on $\mathcal{X}$, $\{R_i\}_{i=1}^k$ with $R_i \neq P_X, i = 1, 2, \ldots, k$ then clearly,

$$\sup_{k,\{w_i\}_{i=1}^k,\{R_i\}_{i=1}^k,\sum_{i=1}^k w_i=1} \frac{\sum_{i=1}^k w_i D(T^*R_i||P_Y)}{\sum_{i=1}^k w_i D(R_i||P)} = s^*(X;Y) \geq \rho(X;Y)^2.$$

But now, suppose we additionally place the constraint $\sum_{i=1}^k w_i R_i = P_X$. Define a random variable W satisfying $W - X - Y$ taking values in $\{1, 2, \ldots, k\}$ with $P(W = i, X = x) = w_i r_i(x)$. Note that $\sum_{i=1}^k w_i D(R_i||P) = I(X : W)$ and $\sum_{i=1}^k w_i D(T^*R_i||P_Y) = I(Y;W)$. From Theorem 1, this means

$$\sup_{k,\{w_i\}_{i=1}^k,\{R_i\}_{i=1}^k,\sum_{i=1}^k w_i=1,\sum_{i=1}^k w_i R_i=P_X} \frac{\sum_{i=1}^k w_i D(T^*R_i||P_Y)}{\sum_{i=1}^k w_i D(R_i||P)} = \rho(X;Y)^2.$$

References

[1] E. Erkip and T. Cover, "The Efficiency of Investment Information," *IEEE Transactions On Information Theory*, vol. 44, pp. 1026–1040, May 1998.

[2] R. Ahlswede and P. Gács, "Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator," *Annals of Probability*, vol. 4, pp. 925–939, 1976.