Time-Sharing Vs. Source-Splitting in the Slepian-Wolf Problem: Error Exponents Analysis

Todd P. Coleman, Muriel Médard; Michelle Effros
{colemant,medard}@mit.edu; effros@caltech.edu
Massachusetts Institute of Technology; California Institute of Technology
Laboratory for Information and Decision Systems; Data Compression Laboratory
77 Massachusetts Ave, 32-D626; 1200 East California Boulevard, MS 136-93
Cambridge, MA 02139; Pasadena, CA 91125

Abstract

We discuss two approaches for decoding at arbitrary rates in the Slepian-Wolf problem - time sharing and source splitting - both of which rely on constituent vertex decoders. We consider the error exponents for both schemes and conclude that source-splitting is more robust at coding at arbitrary rates, as the error exponent for time-sharing degrades significantly at rates near vertices. As a by-product of our analysis, we exhibit an interesting connection between minimum mean-squared error estimation and error exponents.

1 Introduction

In this setting we discuss alternative approaches to attain achievable rates on the dominant face of the Slepian-Wolf region. Here we will focus on the two-user setting, but this can naturally be generalized. Consider two sources (U^1,U^2) with joint probability distribution $W(u^1,u^2)$. The two-user Slepian-Wolf region $\mathcal{R}[W]$ is given by

$$\mathcal{R}[W] = \left\{ \underline{R} \in \mathbb{R}_+^2 : \sum_{i \in S} R_i \geq H\left(U(S)|U(S^c)\right) \ \forall S \subseteq \{1,2\} \right\}$$

where $U(S) = \{U^j, j \in S\}$. Generally speaking, the decoder must find a pair of 'jointly typical' sequences [1, pp. 194-197] consistent with what is observed. This is in general a computationally difficult task. At *vertex* rate points, the joint search over both codebooks for a pair of 'jointly typical' sequences can be done successively. For instance, if users would like to communicate at the rate of $(R_1,R_2) = \left(H(U^1), H(U^2|U^1)\right)$, then we note that communicating at a rate of $H(U^1)$ can be done by simply entropy-encoding either a variable-rate lossless fashion or a near-lossless fixed-rate fashion. After successful decoding, U^1 can be passed as side information to help decode U^2 at a rate of $H(U^2|U^1)$. By exchanging the roles of U^1 and U^2, it follows that the same approach applies to encoding at the vertex rate $(R_1,R_2) = \left(H(U^1|U^2), H(U^2)\right)$. Recently, much attention has been paid to the construction of low-complexity decoders to achieve rates of R_2 very close to $H(U^2|U^1)$.

A more interesting question concerns communicating at *any* rate in the achievable rate region - not necessarily vertices. The most efficient communication schemes minimize sum rate and thus attain rates lying on the *dominant face*, $\mathcal{D}\{\mathcal{R}[W]\}$, given by $\left\{(R_1,R_2) \in \mathcal{R}[W] \ : \ R_1+R_2 = H\left(U^1,U^2\right)\right\}$. Two candidate approaches of using decoding strategies that rely upon vertex decoding are:

Proceedings of the Data Compression Conference (DCC'06)
0-7695-2545-8 /06 $20.00 © 2006 IEEE

- time-sharing, where coding for a non-vertex point is done by coding a certain fraction $\alpha \in [0,1]$ of the time at one vertex, and the remaining fraction $1-\alpha$ of the time at the other vertex
- source-splitting [2, 3], where coding for a non-vertex point in a two-source problem is done by splitting one of the sources and coding at a vertex rate in the corresponding three-source problem

We would like to understand here the performance of the two candidate approaches at rates near the joint entropy boundary, in terms of error probability. We illustrate below that the source-splitting approach is more robust for decoding at arbitrary rates on the dominant face as compared to time-sharing, which can have significant error exponent penalty at rates close to vertices. As a by-product of our analysis, we show an interesting connection between information theory and estimation theory: the error exponent of vertex decoding in an arbitrary instance of the Slepian-Wolf problem depends on the inverse of the Fisher information of Gallager's ρ-parametrized tilted distribution.

2 Error Exponents

Here we discuss the near-lossless fixed-to-fixed distributed data compression setting where n samples of the memoryless source $\left\{(U_i^1, U_i^2)\right\}_{i=1}^n$ are separately encoded. For each source j, the $\left\{U_i^j\right\}_{i=1}^n$ symbols will be mapped to 2^{nR_j} output symbols. The error exponent for a particular coding scheme k will be denoted by

$$E^k(R_1,R_2) \triangleq \liminf_{n\to\infty} -\frac{1}{n}\log P_e^k(R_1,R_2).$$

As illustrated in the appendix (13), for random variables X,Y with joint distribution W, the error exponent $E_{x|y}(R)$ for source coding X at rate R with side information Y has a flat slope at $R = H(X|Y)$:

$$\frac{d}{dR}\left\{E_{x|y}(R)\right\}_{R=H(X|Y)} = 0.$$

Thus to capture the behavior of the exponent at $R = H(X|Y)+\delta$, we must consider second order effects via a Taylor series expansion:

$$\begin{aligned} E_{x|y}(H(X|Y)+\delta) &= \frac{1}{2}\delta^2\frac{d^2}{dR^2}\left\{E_{x|y}(R)\right\}_{R=H(X|Y)} + o(\delta^3) \\ &= \frac{1}{2}\delta^2 E''_{x|y}(H(X|Y)) + o(\delta^3). \end{aligned}$$

We will denote the error exponent for time-sharing as $E^t(R_1,R_2)$ and that for source-splitting as $E^s(R_1,R_2)$. We are interested in the behavior of the error exponent at rates near the dominant face.

2.1 Time-Sharing

Time-sharing is one approach to attain any rate on the dominant face. For $\alpha \in [0,1]$, αn of the samples are encoded near the vertex $(R_1,R_2) = \left(H\left(U^1\right), H\left(U^2|U^1\right)\right)$ and the remaining $(1-\alpha)n$ samples are encoded near the other vertex $(R_1,R_2) = \left(H\left(U^1|U^2\right), H\left(U^2\right)\right)$. We will assume that decoding is done with the pipelined vertex decoding approach described above. Thus for the

decoding of the αn symbol pairs at the rate $\left(H\left(U^1\right)+\delta, H\left(U^2|U^1\right)+\delta\right)$, we have

$$P_e^{t,\alpha} \le P\left(\left[\underline{\hat{U}}^1\right]_1^{\alpha n} \neq \left[\underline{U}^1\right]_1^{\alpha n}\right) + P\left(\left[\underline{\hat{U}}^2\right]_1^{\alpha n} \neq \left[\underline{U}^2\right]_1^{\alpha n} \middle| \left[\underline{U}^1\right]_1^{\alpha n}\right)$$
$$= 2^{-n\alpha\left[E_{u^1}\left(H\left(U^1\right)+\delta\right)-o(n)\right]} + 2^{-n\alpha\left[E_{u^2|u^1}\left(H\left(U^2|U^1\right)+\delta\right)-o(n)\right]}$$

For the decoding of the $(1-\alpha)n$ symbol pairs at the rate $\left(H\left(U^1|U^2\right)+\delta, H\left(U^2\right)+\delta\right)$, we have

$$P_e^{t,1-\alpha} \le P\left(\left[\underline{\hat{U}}^2\right]_{\alpha n+1}^{n} \neq \left[\underline{U}^2\right]_{\alpha n+1}^{n}\right) + P\left(\left[\underline{\hat{U}}^1\right]_{\alpha n+1}^{n} \neq \left[\underline{U}^1\right]_{\alpha n+1}^{n} \middle| \left[\underline{U}^2\right]_{\alpha n+1}^{n}\right)$$
$$= 2^{-n(1-\alpha)\left[E_{u^2}\left(H\left(U^2\right)+\delta\right)-o(n)\right]} + 2^{-n(1-\alpha)\left[E_{u^1|u^2}\left(H\left(U^1|U^2\right)+\delta\right)-o(n)\right]}$$

Thus it follows that for $(R_1,R_2) \in \mathcal{D}$,

$$E^t(R_1+\delta,R_2+\delta) = \min\Bigg[\alpha E_{u^1}\left(H\left(U^1\right)+\delta\right), \alpha E_{u^2|u^1}\left(H\left(U^2|U^1\right)+\delta\right),$$
$$(1-\alpha)E_{u^2}\left(H\left(U^2\right)+\delta\right), (1-\alpha)E_{u^1|u^2}\left(H\left(U^1|U^2\right)+\delta\right)\Bigg]$$
$$= \frac{1}{2}\delta^2 \min\Bigg[\alpha E''_{u^1}\left(H\left(U^1\right)\right), \alpha E''_{u^2|u^1}\left(H\left(U^2|U^1\right)\right),$$
$$(1-\alpha)E''_{u^2}\left(H\left(U^2\right)\right), (1-\alpha)E''_{u^1|u^2}\left(H\left(U^1|U^2\right)\right)\Bigg] + o(\delta^3)$$

where α satisfies

$$R_1 = \alpha H\left(U^1\right) + (1-\alpha)H\left(U^1|U^2\right). \tag{1}$$

2.2 Source-Splitting

Source-splitting transforms a point on the dominant face of the two-source problem to a vertex point in a three-source problem. This is done by

$$U_i^1 \quad \mapsto \quad \left(\begin{array}{l} U_i^{1a} = f_a\left(U_i^1\right) \\ U_i^{1b} = f_b\left(U_i^1\right) \end{array}\right) \mapsto U_i^1 = f\left(U_i^{1a}, U_i^{1b}\right) \tag{2}$$

where the functions $f_a : \mathcal{U}_1 \to \mathcal{U}_1, f_b : \mathcal{U}_1 \to \mathcal{U}_1$ and $f : \mathcal{U}_1 \to \mathcal{U}_1$ satisfy

$$f\left(f_a(u), f_b(u)\right) = u \;\forall\; u \in \mathcal{U}_1.$$

As an example [3], this can be done as follows:

$$U_i \quad \mapsto \quad \left(\begin{array}{l} U_i^a = \min(\pi(U_i), T) \\ U_i^b = \max(\pi(U_i), T) - T \end{array}\right) \mapsto U_i = \pi^{-1}\left(U_i^a + U_i^b\right), \tag{3}$$

where $T \in \mathcal{U}$ operates as a thresholder and $\pi \in \Pi(\mathcal{U})$ is a permutation operator.

If we have two discrete memoryless sources (U^1,U^2) drawn according to $P\left(u^1,u^2\right)$, then we can split U^1 to form (U^{1a},U^{1b}) as shown in (3). At this point, we have three sources, each of which can be encoded separately at rates R_{1a},R_{1b},R_2. We note that because $U \leftrightarrow (U^{1a},U^{1b})$, $H\left(U^1,U^2\right) = H\left(U^{1a},U^{1b},U^2\right)$. Through the chain rule for entropy, we consider the rates

$$\begin{aligned} R_{1a} &= H\left(U^{1a}\right) && \text{(4a)}\\ R_2 &= H\left(U^2|U^{1a}\right) && \text{(4b)}\\ R_{1b} &= H\left(U^{1b}|U^2,U^{1a}\right) && \text{(4c)}\\ R_1 &= R_{1a}+R_{1b}. && \text{(4d)} \end{aligned}$$

For any nontrivial split, (R_1,R_2) is not a vertex in $\mathcal{R}\left[P\left(u^1,u^2\right)\right]$, but (R_{1a},R_2,R_{1b}) *is* a vertex in $\mathcal{R}\left[P\left(u^{1a},u^2,u^{1b}\right)\right]$. This directly implies a parallelizable encoding strategy and pipelined single-user decoding strategy that operates with the complexity of a smaller-alphabet decoder.

This corresponds to a vertex in the U^{1a},U^{1b},U^2 problem by using the chain rule for entropy and encoding at rates given by (4). Also in this scheme, coding to attain a non-vertex is mapped to coding at a vertex. Here we also assume that decoding is done with the pipelined vertex decoding approach. Thus for the decoding, we have

$$\begin{aligned} P_e^s &\le P\left(\underline{\hat{U}}^{1a} \neq \underline{U}^{1a}\right) + P\left(\underline{\hat{U}}^{2} \neq \underline{U}^{2}|\underline{U}^{1a}\right) + P\left(\underline{\hat{U}}^{1b} \neq \underline{U}^{1b}|\underline{U}^{1a},\underline{U}^{2}\right)\\ &= 2^{-n\left[E_{u^{1a}}\left(H\left(U^{1a}\right)+\frac{1}{2}\delta\right)-o(n)\right]} + 2^{-n\left[E_{u^2|u^{1a}}\left(H\left(U^2|U^{1a}\right)+\delta\right)-o(n)\right]}\\ &+ 2^{-n\left[E_{u^{1b}|u^2,u^{1a}}\left(H\left(U^{1b}|U^2,U^{1a}\right)+\frac{1}{2}\delta\right)-o(n)\right]} \end{aligned}$$

Thus it follows that for $(R_1,R_2) \in \mathcal{D}$,

$$\begin{aligned} E^s(R_1+\delta,R_2+\delta) &= \min\left[E_{u^{1a}}\left(H\left(U^{1a}\right)+\frac{1}{2}\delta\right), E_{u^2|u^{1a}}\left(H\left(U^2|U^{1a}\right)+\delta\right),\right.\\ &\qquad \left. E_{u^{1b}|u^2,u^{1a}}\left(H\left(U^{1b}|U^2,U^{1a}\right)+\frac{1}{2}\delta\right)\right]\\ &= \frac{1}{2}\delta^2 \min\left[\frac{1}{4}E''_{u^{1a}}\left(H\left(U^{1a}\right)\right), E''_{u^2|u^{1a}}\left(H\left(U^2|U^{1a}\right)\right),\right.\\ &\qquad \left.\frac{1}{4}E''_{u^{1b}|u^2,u^{1a}}\left(H\left(U^{1b}|U^2,U^{1a}\right)\right)\right] + o(\delta^3). \end{aligned}$$

Note that to attain a rate of $(R_1+\delta,R_2+\delta)$ we have to allocate $\frac{1}{2}\delta$ extra rate to U^{1a} and $\frac{1}{2}\delta$ to U^{1b}, as compared to the usual δ to U^1 in the time-sharing case.

2.3 Comparison

It is the purpose of this discussion to observe how the error exponents behave for the two approaches when coding at rates $(R_1+\delta,R_2+\delta)$ where $(R_1,R_2) \in \mathcal{D}$. From the onset it is not clear which approach has better exponents - both cases exhibit error exponent degradation. In the case of time-sharing, error exponent degradation is caused by a reduction in the effective block length by

factors of α and $1-\alpha$. In the source-splitting scenario, error exponent degradation arises because of error propagation in decoding three sources rather than two, along with the reduction by a factor of $\frac{1}{4}$ due to the splitting operation. Furthermore, the comparison is not straightforward because the source-splitting operation creates a new joint distribution on the three sources, as compared to the original joint distribution on the two sources. Although these distributions are in some sense equivalent because (U_i^1, U_i^2) and $(U_i^{1a}, U_i^2, U_i^{1b})$ form a bijection, the behavior of the error exponent's second derivative involves more complicated functions of the distribution than just entropy:

Lemma 2.1 $E''_{x|y}(H(X|Y)) = \frac{1}{-H(X|Y)^2 + \sum_{x,y} Q(x,y)\log^2[Q(x|y)]}$.

Proof details are in the appendix. What is interesting about this denominator is that it can be characterized in terms of the Fisher information of Gallager's ρ-tilted distribution [4]. In particular, if we define $Q_\rho(x,y)$ as the product of $Q_\rho(y)$, given by (7), and $Q_\rho(x|y)$, given by (8), then we can calculate the Fisher information of this parametrized probability distribution:

$$F(\rho) = \sum_{x,y} \left(\frac{d \log Q_\rho(x,y)}{d\rho} \right)^2 Q_\rho(x,y) \tag{5}$$

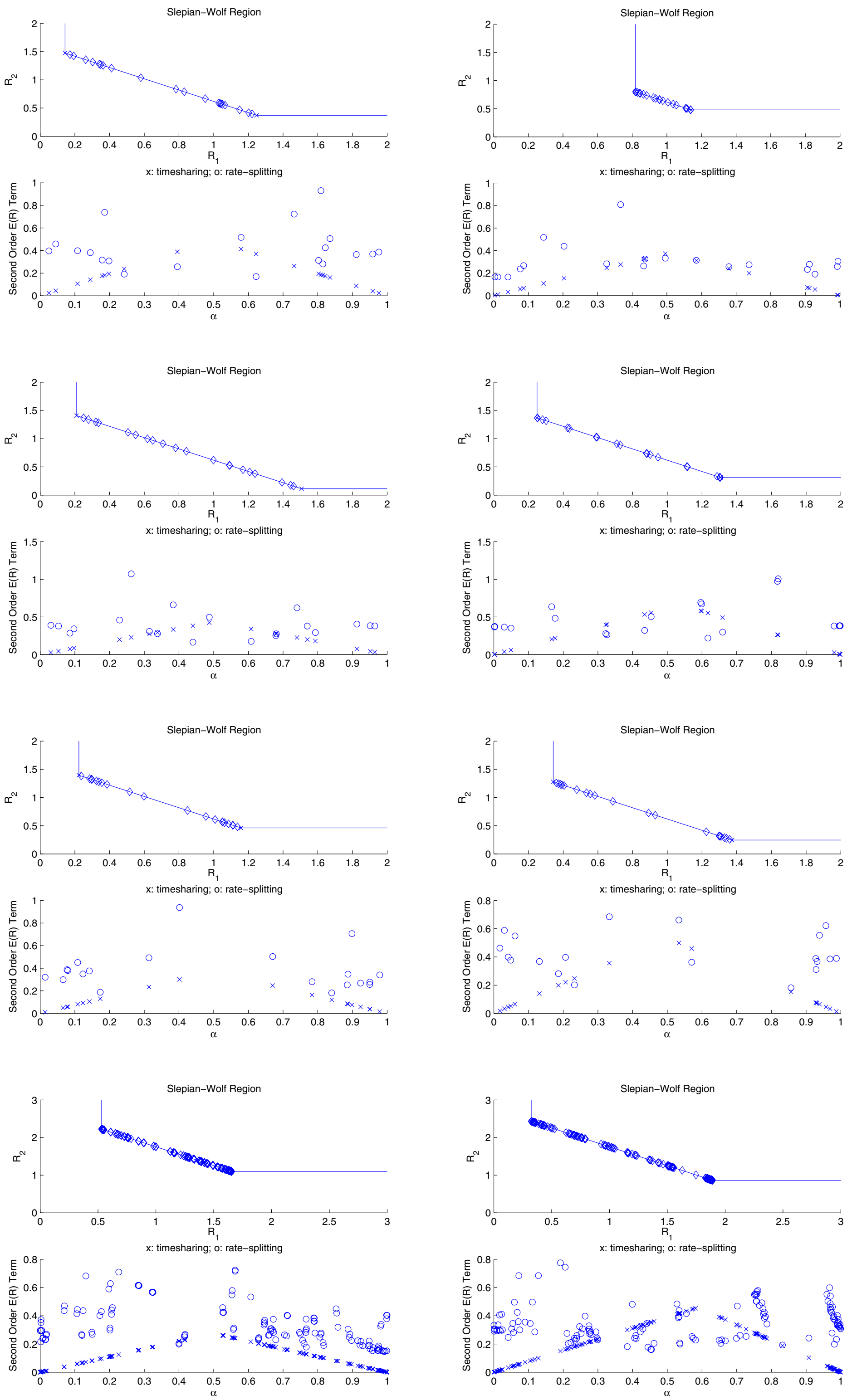
Then we can characterize the error exponent's second derivative in terms of the inverse of the above Fisher information quantity:

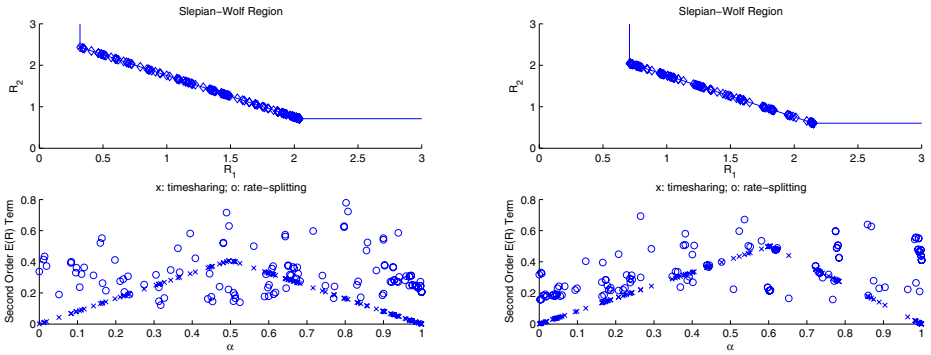
Lemma 2.2 $E''_{x|y}(H(X|Y)) = \frac{1}{F(\rho)}\Big|_{\rho=0}$.

Proof details are in the appendix. This leads to another interesting connection between information-theoretic quantities (error exponents) and estimation theoretic ones (MMSE and Fisher Information). In particular, the connection relates the error exponent's second derivative to the inverse of a Fisher information - which is a bound on minimum mean-squared error. The common thread appears to lie in the information geometry [5] interpretation of the Kullback-Leibler distance. However in our opinion, an in-depth understanding of this relation remains to be found.

3 Examples on Randomly Constructed Joint Probability Distributions

Here we randomly construct joint probability distributions W on (U^1, U^2) and compare $E^s(R_1+\delta, R_2+\delta)$ with $E^t(R_1+\delta, R_2+\delta)$. In the figure pairs below, the top figure in each pair shows the Slepian-Wolf achievable rate region and the target rate points on the dominant face. The bottom figure in each pair shows $E^s(R_1+\delta, R_2+\delta)$ and $E^t(R_1+\delta, R_2+\delta)$ as a function of α, where α satisfies (1). For the splitting case, splitting is done according to (3). The takeaway theme from all these examples is that the minimum $E^s(R_1+\delta, R_2+\delta)$ for points $(R_1, R_2) \in \mathcal{D}$ is bounded away from 0 whereas $E^t(R_1+\delta, R_2+\delta)$ decays to 0 linearly as α approaches 0 or 1. Consequently at rates close to vertices, the second order source-splitting exponent significantly dominates that of time-sharing. At rates halfway between vertices, in some cases source-splitting wins, and in other cases time-sharing does. We were not able to find many cases where the second-order time-sharing exponent significantly dominates that of source-splitting. Thus in terms of error exponents, source-splitting appears to be more robust across various rates than time-sharing.


Slepian–Wolf Region
R_2
R_1
x: timesharing; o: rate–splitting
Second Order E(R) Term
α



Acknowledgment

The authors would like to thank Ralf Koetter for his initial suggestion to consider this comparison.

A Proof of Lemma 2.1

Here we consider the ML decoding error exponent for source decoding x when side information y is known at the decoder. Denote $P_e(y)$ to be the error probability conditioned upon receiving y. Then from [4] we have that

$$\frac{-1}{n}\log P_e(y) \geq E_{x|y}(R,y) \triangleq \max_{0\leq\rho\leq 1} \rho R - E_{0,x|y}(\rho,y)$$

$$E_{0,x|y}(\rho,y) \triangleq (1+\rho)\log\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right] \tag{6}$$

For future reference, let us define the tilted distributions

$$Q_\rho(x|y) \triangleq \frac{Q(x|y)^{\frac{1}{1+\rho}}}{\sum_x Q(x|y)^{\frac{1}{1+\rho}}} \tag{7}$$

$$Q_\rho(y) \triangleq \frac{P(y)\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right]^{1+\rho}}{\sum_y P(y)\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right]^{1+\rho}}. \tag{8}$$

Differentiating with respect to ρ to find a stationary point, we can relate $E_{x|y}(R,y)$ and R parametrically in terms of ρ:

$$R = \frac{\partial E_{0,x|y}(\rho,y)}{\partial\rho} = H\left(X_\rho|y\right)$$

where the second equality above can be verified with calculation, as mentioned in [4]. Now consider averaging over a memoryless y. Again, from [4], we have:

$$\frac{-\log P_e}{n} \geq E_{x|y}(R) \triangleq \max_{0\leq\rho\leq 1} \rho R - E_0(\rho),$$

$$E_0(\rho) \triangleq \log\left(\sum_y P(y)\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right]^{1+\rho}\right) \tag{9}$$

We now relate $E_{x|y}(R)$ and R in terms of ρ through differentiation:

$$\begin{aligned} R &= \frac{\partial E_0(\rho)}{\partial \rho} = H\left(X_\rho|Y_\rho\right) && (10)\\ \Rightarrow E_{x|y}(R) &= \rho H\left(X_\rho|Y_\rho\right) - E_0(\rho) \\ \Rightarrow \frac{\partial}{\partial \rho}E_{x|y}(R) &= \rho\frac{\partial}{\partial \rho}H\left(X_\rho|Y_\rho\right) \\ \Rightarrow E'_{x|y}(R) &\triangleq \frac{dE_{x|y}(R)}{dR} = \frac{\partial E_{x|y}(R)}{\partial \rho} = \rho && (11)\\ \Rightarrow E''_{x|y}(R) &= \frac{\frac{\partial E'_{x|y}(R)}{\partial \rho}}{\frac{\partial R}{\partial \rho}} = \frac{1}{\frac{\partial}{\partial \rho}H\left(X_\rho|Y_\rho\right)} && (12) \end{aligned}$$

where the second equality in (10) can be verified with tedious calculations, as mentioned in [4]. Note from (10),(7) (8), and (11) that

$$E'_{x|y}(H(X|Y)) = 0. \quad (13)$$

$$\begin{aligned} \Rightarrow \frac{\partial}{\partial \rho}\left\{H\left(X_\rho|Y_\rho\right)\right\} &= \sum_y \frac{\partial}{\partial \rho}\left\{Q_\rho(y)H\left(X_\rho|y\right)\right\} \\ &= \sum_y H\left(X_\rho|y\right)\frac{\partial Q_\rho(y)}{\partial \rho} + \sum_y Q_\rho(y)\frac{\partial}{\partial \rho}\left\{H\left(X_\rho|y\right)\right\}. \quad (14) \end{aligned}$$

To address $\frac{\partial Q_\rho(y)}{\partial \rho}$, note that for any differentiable function g, $g'(\rho) = g(\rho)\frac{d}{d\rho}\{\log g(\rho)\}$.

$$\begin{aligned} \Rightarrow \log Q_\rho(y) &= \log P(y) + (1+\rho)\log\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right] \\ &\quad - \log\left[\sum_y P(y)\left[\sum_x Q(x|y)^{\frac{1}{1+\rho}}\right]^{1+\rho}\right] \\ &= \log P(y) + E_{0,x|y}(\rho, y) - E_0(\rho) \quad (15) \\ \Rightarrow \frac{\partial \log Q_\rho(y)}{\partial \rho} &= H\left(X_\rho|y\right) - H\left(X_\rho|Y_\rho\right) \quad (16) \\ \Rightarrow \frac{\partial Q_\rho(y)}{\partial \rho}\Big|_{\rho=0} &= Q(y)\left[H\left(X|y\right) - H\left(X|Y\right)\right] \quad (17) \end{aligned}$$

where (15) is due to (6), (9). As for $\frac{\partial H\left(X_\rho|y\right)}{\partial \rho}$, note that

$$\begin{aligned} H\left(X_\rho|y\right) &= -D\left(Q_\rho(x|y)\|\mathbb{U}\right) + \log|\mathcal{X}| \quad (18) \\ \Rightarrow \frac{\partial}{\partial \rho}\left\{H\left(X_\rho|y\right)\right\} &= -\frac{\partial}{\partial \rho}\left\{D\left(Q_\rho(x|y)\|\mathbb{U}\right)\right\} \quad (19) \end{aligned}$$

where $\mathbb{U}(x) \triangleq \frac{1}{|\mathcal{X}|}$. By [6], for any two distributions W_0 and W_1, the distribution W_t

$$W_t(x) \triangleq \frac{W_0(x)^{1-t}W_1(x)^t}{\sum_a W_0(a)^{1-t}W_1(a)^t} \quad (20)$$

relates the Kullback-Leibler divergence to the Fisher information $F(t)$ according to:

$$\frac{d}{dt}\left\{D\left(W_t\|W_0\right)\right\} = tF(t), \tag{21}$$

$$F(t) \triangleq \sum_x \left(\frac{d\log W_t(x)}{dt}\right)^2 W_t(x) \tag{22}$$

We would like to characterize $Q_\rho(x|y)$ in terms of a W_t of the form (20):

$$\boxed{\begin{array}{c} 0 \le \rho \le \infty \\ \rho = 0 : Q_\rho(x|y) = Q(x|y) \\ \rho = \infty : Q_\rho(x|y) = \dfrac{1}{|\mathcal{X}|} \end{array}} \quad \overleftrightarrow{t = \tfrac{1}{1+\rho}} \quad \boxed{\begin{array}{c} 0 \le t \le 1 \\ t = 1 : W_t(x) = Q(x|y) \\ t = 0 : W_t(x) = \dfrac{1}{|\mathcal{X}|} \end{array}}$$

$$\begin{aligned} \Rightarrow \frac{\partial H\left(X_\rho|y\right)}{\partial\rho} &= -\frac{\partial}{\partial\rho}\left\{D\left(Q_\rho(x|y)\|\mathbb{U}\right)\right\} \\ &= -\frac{dt}{d\rho}\frac{d}{dt}\left\{D\left(W_t\|W_0\right)\right\}\Big|_{t=\frac{1}{1+\rho}} \\ &= \frac{1}{(1+\rho)^3}F\left(\frac{1}{1+\rho}\right) \\ \Rightarrow \frac{\partial H\left(X_\rho|y\right)}{\partial\rho}\Big|_{\rho=0} &= F(1) = -H(X|y)^2 + \sum_x Q(x|y)\log^2[Q(x|y)] \end{aligned} \tag{23}$$

Thus it follows from (12),(14), (17), and (23) that

$$\boxed{E''_{x|y}\left(H(X|Y)\right) = \frac{1}{-H(X|Y)^2 + \sum_{x,y} Q(x,y)\log^2[Q(x|y)]}}$$

B Proof of Lemma 2.2

$$\begin{aligned} \log Q_\rho(x,y) &= \log Q_\rho(x|y) + \log Q_\rho(y) \\ &= \frac{1}{1+\rho}\left[\log Q(x|y) - E_{0,x|y}(\rho,y)\right] + \log P(y) + E_{0,x|y}(\rho,y) - E_0(\rho) \\ \Rightarrow \frac{\partial\log Q_\rho(x,y)}{\partial\rho} &= \frac{-1}{1+\rho}\frac{\partial E_{0,x|y}(\rho,y)}{\partial\rho} + \frac{-1}{(1+\rho)^2}\left[\log Q(x|y) - E_{0,x|y}(\rho,y)\right] \\ &\quad + \frac{\partial E_{0,x|y}(\rho,y)}{\partial\rho} - \frac{\partial E_0(\rho)}{\partial\rho} \\ &= \frac{-1}{1+\rho}\left[H(X_\rho|y) + \log Q_\rho(x|y)\right] + H(X_\rho|y) - H(X_\rho|Y_\rho) \end{aligned} \tag{24}$$

Thus

$$\left(\frac{\partial \log Q_\rho(x,y)}{\partial \rho}\right)^2\Big|_{\rho=0} = (\log Q(x|y) + H(X|Y))^2$$
$$= H(X|Y)^2 + 2\log Q(x|y)H(X|Y) + \log^2[Q(x|y)]$$
$$\Rightarrow F(\rho)\Big|_{\rho=0} = -H(X|Y)^2 + \sum_{x,y} Q(x,y)\log^2[Q(x|y)]$$