Computation of Csiszár's Mutual Information of Order α

Damianos Karakos*, Sanjeev Khudanpur* and Carey E. Priebe†
*Department of Electrical and Computer Engineering
and Center for Language and Speech Processing
† Department of Applied Mathematics and Statistics
Johns Hopkins University, Baltimore, MD 21218, USA
Email: {damianos, khudanpur, cep}@jhu.edu

***Abstract*— Csiszár introduced the mutual information of order α in [1] as a parameterized version of Shannon's mutual information. It involves a minimization over the probability space, and it cannot be computed in closed form. An alternating minimization algorithm for its computation is presented in this paper, along with a proof of its convergence. Furthermore, it is proved that the algorithm is an instance of the Csiszár-Tusnády iterative minimization procedure [2].**

I. Introduction

Csiszár [1] defined the mutual information of order α between two discrete random variables X, Y (with joint distribution P_{XY}) as

$$I_\alpha(X;Y) \stackrel{\triangle}{=} \min_Q \sum_y P_Y(y) D_\alpha(P_{X|Y}(\cdot|y)\|Q(\cdot)), \tag{1}$$

where $0 \leq \alpha \neq 1$, and

$$D_\alpha(P\|Q) \stackrel{\triangle}{=} \frac{1}{\alpha - 1} \log \sum_x P^\alpha(x) Q^{1-\alpha}(x), \tag{2}$$

is the *Rényi divergence of order* α [3] between distributions P, Q. The limit of (2), as $\alpha \to 1$, is the well-known Kullback-Leibler information divergence. Furthermore, we adopt the same convention as in [1]: if $\alpha > 1$ and $Q(x) = 0$ for some x, then $P^\alpha(x)Q^{1-\alpha}(x)$ is equal to 0 or $+\infty$ when $P(x) = 0$ or $P(x) > 0$, respectively.

It can be proved [1] that $I_\alpha(X;Y)$ retains most of the properties of $I(X;Y)$ (i.e., it is a non-negative, continuous, and concave function of P_X, and, when $\alpha < 1$, it is convex with respect to $P_{X|Y}$). Also, as $\alpha \to 1$, $I_\alpha(X;Y)$ converges to $I(X;Y)$.

Observe that (2) is computed by *exponentiating* the probability measures. If the exponents are strictly less than unity (i.e., $0 \leq \alpha \leq 1$), this results in "smoothed" values, lessening possible extreme variations in $P(x), Q(x)$. Smoothing is important in cases where the distributions are computed from limited amounts of data; sparsity typically leads to underestimates of the least likely events, and to overestimates of the most likely events. Such phenomena are prevalent when the underlying spaces are very large, compared to the sample size; see, for instance, [4]–[7] for various treatments of the sparsity problem in natural language processing. This fact partly explains the improvement in performance of the text categorization algorithms of [8], [9], when Shannon's mutual information was replaced with (1) as a criterion for clustering sparse empirical distributions.

There is no analytic form for the minimizer of the right-hand side of (1). The next section describes an alternating minimization algorithm which converges to the desired minimum.

Use of notation: Random variables (r.v.) are denoted by capital letters, while their realizations are denoted by the corresponding lowercase letters. All random variables are assumed to lie in discrete finite spaces, denoted by the corresponding calligraphic letter (e.g., $X \in \mathcal{X}$). The probability mass function (pmf) of a random quantity is denoted using the appropriate subscript, e.g., P_X for r.v. X, or P_{XY}

for the joint pmf of X, Y. The conditional pmf of X given Y is denoted by $P_{X|Y}$; letter W is also used to denote a conditional pmf. Finally, for the rest of the paper it will be assumed that $0 < \alpha < 1$.

II. The Algorithm

The algorithm for the computation of (1) is described in Figure 1, and it takes as input the joint distribution P_{XY} and $\gamma > 0$ (the latter is typically a small number, and is used in the termination condition). The algorithm is based on the following identity (proved in Section III): for any distributions P, Q, and $0 < \alpha < 1$,

$$D_\alpha(P\|Q) = \frac{\alpha}{1-\alpha} D(P^*\|P) + D(P^*\|Q), \quad (3)$$

where

$$P^*(x) = \frac{(P(x))^\alpha (Q(x))^{1-\alpha}}{\sum_x (P(x))^\alpha (Q(x))^{1-\alpha}}.$$

Therefore,

$$I_\alpha(X;Y) = \min_Q \sum_y P_Y(y) \Big(D\big(P^*(\cdot|y)\|Q\big) + \frac{\alpha}{1-\alpha} D\big(P^*(\cdot|y)\|P(\cdot|y)\big) \Big),$$

which, by virtue of Lemma 2 of Section III, can be written as

$$I_\alpha(X;Y) = \min_Q \min_W \sum_y P_Y(y) \Big(D\big(W(\cdot|y)\|Q\big) + \frac{\alpha}{1-\alpha} D\big(W(\cdot|y)\|P(\cdot|y)\big) \Big). \quad (4)$$

The double minimization of (4) is achieved through the alternating minimization of the algorithm of Figure 1, as is proved in Theorem 1 of Section III.

III. Convergence of the Algorithm

First, we will prove identity (3). Given distributions P, Q and a scalar $0 < \alpha < 1$, we define:

$$P^*(x) = C^{-1} (P(x))^\alpha (Q(x))^{1-\alpha},$$

where $C = \sum_x (P(x))^\alpha (Q(x))^{1-\alpha}$ is just a normalizing constant. Then, we have the following chain of equalities

$$\begin{aligned} D(P^*\|Q) &= \sum_x P^*(x) \log \left(C^{-1} \left(\frac{P(x)}{Q(x)} \right)^\alpha \right) \\ &= -\log(C) + \alpha \sum_x P^*(x) \log \left(\frac{P(x)}{Q(x)} \right) \\ &= -\log(C) + \alpha (D(P^*\|Q) - D(P^*\|P)) \\ &= (1-\alpha) D_\alpha(P\|Q) + \\ &\quad \alpha (D(P^*\|Q) - D(P^*\|P)). \end{aligned} \quad (5)$$

By re-arranging terms in (5), we obtain (3).

Now, for fixed distributions $P_Y, P_{X|Y}$, let us define the functional

$$J(W,Q) \triangleq \sum_y P_Y(y) \Big(D\big(W(\cdot|y)\|Q\big) + \frac{\alpha}{1-\alpha} D\big(W(\cdot|y)\|P_{X|Y}(\cdot|y)\big) \Big), \quad (6)$$

which is a function of distribution Q and conditional distribution W. The algorithm of Figure 1 performs an alternating update of Q, W. As is proved in the

Algorithm for computing Csiszár's mutual information of order α.

Input: Joint distribution P_{XY}, $0 < \alpha < 1$, and threshold $\gamma > 0$.

Initialization:

$t = 0$
$Q_t^*(x) = \sum_y P_Y(y) P_{X|Y}(x|y)$
$I_t = -\infty$

Loop:

Repeat
$t = t + 1$
$P_t^*(x|y) = C_t^{-1}(y) (P_{X|Y}(x|y))^\alpha (Q_{t-1}^*(x))^{1-\alpha}$
$Q_t^*(x) = \sum_y P_Y(y) P_t^*(x|y)$
$I_t = \sum_y P_Y(y) \big(\frac{\alpha}{1-\alpha} D(P_t^*(\cdot|y)\|P_{X|Y}(\cdot|y)) + D(P_t^*(\cdot|y)\|Q_t^*) \big)$
Until $|I_t - I_{t-1}| < \gamma$

Output: I_t.

Fig. 1. Alternating minimization algorithm for computing $I_\alpha(X;Y)$. $C_t(y)$ is just a normalizing constant, computed for each conditioning y.

next two lemmas, this alternating update can only reduce $J(W,Q)$ (alternating minimization). Then, convergence to the global minimum is guaranteed by the fact that Q, W lie in convex (probability) spaces, and that J is a convex function of its arguments Q, W.

Lemma 1: For a fixed conditional distribution W, the functional $J(W,Q)$ is minimized by

$$Q^*(x) = \sum_y P_Y(y) W(x|y).$$

Proof: Only the first term in the right-hand side of (6) needs to be minimized. It can be easily checked that this term is equal to $D(P_Y \times W \| P_Y \times Q)$, and hence the proof follows immediately from Lemma 13.8.1 of [10]. ∎

Alternative proof: Let Q be any arbitrary pmf on $\mathcal{X}$, and Q^* as specified in the statement of the lemma. Then,

$$\begin{aligned}
&\sum_y P_Y(y) D(W(\cdot|y)||Q) \\
&\qquad - \sum_y P_Y(y) D(W(\cdot|y)||Q^*) \\
&= \sum_y P_Y(y) \Big(\sum_x W(x|y) \log \frac{W(x|y)}{Q(x)} \\
&\qquad - \sum_x W(x|y) \log \frac{W(x|y)}{Q^*(x)} \Big) \\
&= \sum_x Q^*(x) \log \frac{Q^*(x)}{Q(x)} \geq 0,
\end{aligned}$$

Hence, the first term in the right-hand side of (6) is minimized by Q^*, as required. ∎

Lemma 2: For a fixed distribution Q, the functional $J(W,Q)$ is minimized by

$$W^*(x|y) = C^{-1}(y) (P_{X|Y}(x|y))^\alpha (Q(x))^{1-\alpha},$$

where $C(y) = \sum_x (P_{X|Y}(x|y))^\alpha (Q(x))^{1-\alpha}$ is a normalizing constant.

Proof: It suffices to find the minimizer of the expression

$$\begin{aligned}
K\big(W(\cdot|y), Q\big) &= D\big(W(\cdot|y) \| Q\big) + \\
&\quad \frac{\alpha}{1-\alpha} D\big(W(\cdot|y) \| P_{X|Y}(\cdot|y)\big),
\end{aligned}$$

for any given y. This will imply minimization of $J(W,Q)$ as well. Notice that $K\big(W(\cdot|y), Q\big)$ is a convex function of $W(\cdot|y)$, for all y, by virtue of the convexity of the KL-divergence [10]. Hence, it has a unique minimizer, which can be found using Lagrange multipliers, by solving the equations

$$\frac{\partial (K\big(W(\cdot|y), Q\big) + \lambda_y (\sum_{x'} W(x'|y) - 1))}{\partial W(x|y)} = 0, \tag{7}$$

for all x, y, under the usual constraint $\sum_x W(x|y) = 1$. Note that there are $|\mathcal{X}| \times |\mathcal{Y}|$ equalities implied in Equation (7).

The left-hand-side of (7) is equal to

$$\begin{aligned}
&K'(W,Q) + \lambda_y \\
&= 1 + \log(W(x|y)) - \log(Q(x)) \\
&\quad + \frac{\alpha}{1-\alpha} (1 + \log(W(x|y)) \\
&\quad - \log(P_{X|Y}(x|y))) + \lambda_y \\
&= \frac{1}{1-\alpha} (1 + \log(W(x|y)) - (1-\alpha) \log(Q(x)) \\
&\quad - \alpha \log(P_{X|Y}(x|y)) + \lambda_y \\
&= \frac{1}{1-\alpha} (1 + \log(W(x|y)) \\
&\quad - \log((P_{X|Y}(x|y))^\alpha (Q(x))^{1-\alpha})) + \lambda_y. \qquad (8)
\end{aligned}$$

By setting (8) equal to zero and solving for $W(x|y)$, we obtain

$$W(x|y) = \exp(-(1-\alpha)\lambda_y - 1) (P_{X|Y}(x|y))^\alpha (Q(x))^{1-\alpha},$$

which, together with the constraint $\sum_x W(x|y) = 1$, gives us the required result. ∎

Alternative proof: Minimizing $K(W,Q)$ is equivalent to minimizing $(1-\alpha)K(W,Q) = (1-\alpha)D(W\|Q) + \alpha D(W\|P)$ with respect to W, for fixed Q, P. Then,

$$\begin{aligned}
&(1-\alpha) D(W\|Q) + \alpha D(W\|P) \qquad (9) \\
&\quad = \sum_x W(x) \log \frac{CW(x)}{CQ^{1-\alpha}(x) P^\alpha(x)} \\
&\quad = \sum_x W(x) \log \frac{CW(x)}{Q^{1-\alpha}(x) P^\alpha(x)} - \log C, \qquad (10)
\end{aligned}$$

where $C = \sum_x Q^{1-\alpha}(x) P^\alpha(x)$ and the first expression in (10) is the non-negative Kullback-Leibler distance of the distribution W and the

distribution $W^*(x) = C^{-1}Q^{1-\alpha}(x)P^{\alpha}(x)$. Hence (10) is minimized when $W = W^*$, as required. ■

Theorem 1: The alternating minimization algorithm of Figure 1 converges to $I_\alpha(X;Y)$, i.e., it computes the expression

$$\min_Q \min_W J(W,Q),$$

where $J(W,Q)$ was defined in (6).

Proof: It is easy to establish that $J(W,Q)$ is a convex function of the pair (Q,W), as it is a linear combination of KL-divergences having Q, W as their arguments. Also, Q and W belong to convex probability sets. Thus, the sufficient condition of [11] is satisfied, which implies that the alternating minimization algorithm of Figure 1 converges to the (unique) global minimum of $J(W,Q)$. ■

IV. Connection to the Csiszár-Tusnády Alternating Minimization Algorithm

In their paper [2], Csiszár and Tusnády show that an iterative algorithm that successively computes projections between two convex sets eventually converges, under certain conditions, to the minimum "distance" between the two sets. In this section, we prove that these conditions are satisfied for functional J of (6), thus establishing that the alternating minimization algorithm of Section II is an instance of the algorithm of [2]. This provides an alternative (albeit more complicated) proof of the convergence of the algorithm.

Let $\mathcal{P}, \mathcal{Q}$ represent two abstract convex and closed sets, and let $d : \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be an extended real-valued function. The "projection" of a member of one set to the other set is defined as follows:

$$P \xrightarrow{1} Q^* \quad \text{iff} \quad d(P,Q^*) = \min_{Q\in\mathcal{Q}} d(P,Q)$$

$$Q \xrightarrow{2} P^* \quad \text{iff} \quad d(P^*,Q) = \min_{P\in\mathcal{P}} d(P,Q)$$

where $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$. The alternating minimization algorithm of [2] computes a sequence of members of $\mathcal{P}$ and $\mathcal{Q}$ satisfying $Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \ldots$.

For a given function δ, we have the following definitions (from [2]):

Definition 1: The *three points property* holds for $P \in \mathcal{P}$ if

$$Q_0 \xrightarrow{2} P_1 \quad \Rightarrow \quad d(P,Q_0) \geq \delta(P,P_1) + d(P_1,Q_1).$$

Definition 2: The *four points property* holds for $P \in \mathcal{P}$ if

$$\forall\ Q \in \mathcal{Q}, \quad P_1 \xrightarrow{1} Q_1 \quad \Rightarrow \\ d(P,Q) + \delta(P,P_1) \geq d(P,Q_1).$$

The following theorem of [2] establishes convergence of the iterative minimization procedure to the minimum between the two sets:

Theorem 2: If the three points property and the four points property hold for all $P \in \mathcal{P}$, then

$$\lim_{n\to\infty} d(P_n,Q_n) = \min_{P\in\mathcal{P},Q\in\mathcal{Q}} d(P,Q) \triangleq d(\mathcal{P},\mathcal{Q}).$$

We will now show that the sequence of updates $Q_0 \xrightarrow{2} W_1 \xrightarrow{1} Q_1 \cdots$, as expressed in the algorithm of Figure 1, leads to convergence to the minimum of $J(W,Q)$. Note that $W \in \mathcal{P}_{\mathcal{X}}^{\mathcal{Y}}$ and $Q \in \mathcal{P}_{\mathcal{X}}$, where $\mathcal{P}_{\mathcal{X}}$ is the set of measures on $\mathcal{X}$. Furthermore, δ is the KL-divergence function.

Theorem 3: J satisfies the four points property for all $W \in \mathcal{P}_{\mathcal{X}}^{\mathcal{Y}}$ and all $Q \in \mathcal{P}_{\mathcal{X}}$.

Proof: We have

$$\begin{aligned}
&J(W,Q_1) - J(W,Q) \\
&= \sum_{y\in\mathcal{Y}} P_Y(y)(D(W(\cdot|y)\|Q_1) - D(W(\cdot|y)\|Q)) \\
&= \sum_{x\in\mathcal{X}} (P_Y W)(x) \log \frac{Q(x)}{Q_1(x)} \qquad (11) \\
&= D(P_Y W\|Q_1) - D(P_Y W\|Q) \\
&= D(P_Y W\|P_Y W_1) - D(P_Y W\|Q) \qquad (12) \\
&\leq \delta(P_Y W, P_Y W_1), \qquad (13)
\end{aligned}$$

where, in (11), $P_Y W$ is the X marginal of the joint distribution $W \times P_Y$, and in (12), Q_1 was replaced by $P_Y W_1$ by virtue of Lemma 1. ■

Theorem 4: J satisfies the three points property for all $W \in \mathcal{P}_{\mathcal{X}}^{\mathcal{Y}}$.

Proof: Let $\mu \in [0,1]$ and $W^{(\mu)} = \mu W + (1-\mu)W_1$. It suffices to show that

$$J(W^{(\mu)}, Q_0) - J(W_1, Q_0) \geq D(P_Y W^{(\mu)} \| P_Y W_1) \tag{14}$$

for all $\mu \in [0,1]$ (and, hence, for $\mu = 1$).

First, note that, when $\mu = 0$, the left-hand side and the right-hand side of (14) are both equal to zero. Hence, it suffices to prove that

$$\frac{\partial J(W^{(\mu)}, Q_0)}{\partial \mu} \geq \frac{\partial D(P_Y W^{(\mu)} \| P_Y W_1)}{\partial \mu}, \quad \forall\ \mu.$$

We have

$$\begin{aligned}
&\frac{\partial D(P_Y W^{(\mu)} \| P_Y W_1)}{\partial \mu} \\
&= \sum_{x \in \mathcal{X}} (P_Y W(x) - P_Y W_1(x)) \log \frac{(P_Y W^{(\mu)})(x)}{(P_Y W_1)(x)} \\
&= \frac{1}{\mu} (D(P_Y W^{(\mu)} \| P_Y W_1) + D(P_Y W_1 \| P_Y W^{(\mu)}))
\end{aligned} \tag{15}$$

where μ is assumed non-zero in (15). Note that $\partial D(W^{(\mu)} P_Y \| P_Y W_1)/\partial \mu = 0$ when $\mu = 0$ (the same is true for $\partial J(W^{(\mu)}, Q_0)/\partial \mu$, as can be seen below).

Next, we have the following

$$\begin{aligned}
&\frac{\partial J(W^{(\mu)}, Q_0)}{\partial \mu} \\
&= \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x \in \mathcal{X}} (W(x|y) - W_1(x|y)) \times \\
&\quad \log \frac{(W^{(\mu)}(x|y))^{1/(1-\alpha)}}{Q_0(x)(P_{X|Y}(x|y))^{\alpha/(1-\alpha)}} \\
&= \frac{1}{1-\alpha} \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x \in \mathcal{X}} (W(x|y) - W_1(x|y)) \times \\
&\quad \log \frac{W^{(\mu)}(x|y)}{W_1(x|y)} \qquad (16) \\
&= \frac{1}{\mu(1-\alpha)} \sum_{y \in \mathcal{Y}} P_Y(y) (D(W^{(\mu)}(\cdot|y) \| W_1(\cdot|y)) \\
&\quad + D(W_1(\cdot|y) \| W^{(\mu)}(\cdot|y))) \\
&\geq \frac{1}{\mu} (D(P_Y W^{(\mu)} \| P_Y W_1) + D(P_Y W_1 \| P_Y W^{(\mu)})), \qquad (17)
\end{aligned}$$

where the update equation of Lemma 2 was used in (16) and $\mu > 0$ was assumed in (17). By combining (15) and (17) we obtain the required result. ∎

Finally, Theorems 3 and 4 establish that the iterative algorithm of Figure 1 results in the minimization of J, as required.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their comments. The alternative proofs of Lemmas 1 and 2 were supplied by the first reviewer. This work was partially supported by National Science Foundation grant No CCF-0728931.

REFERENCES

[1] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. on Information Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

[2] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions,* Supplement Issue 1, pp. 205–237, 1984.

[3] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symposium on Math. Statist. Probability*, vol. 1, 1961, pp. 547–561.

[4] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1997.

[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting of the ACL*, 1996, pp. 310–318.

[6] D. Karakos and S. Khudanpur, "Language modeling with the maximum likelihood set: Complexity issues and the back-off formula," in *Proc. 2006 IEEE Intern. Symposium on Information Theory (ISIT-06)*, Seattle, WA, July 2006.

[7] ——, "Error bounds and improved probability estimation using the maximum likelihood set," in *Proc. 2007 IEEE Intern. Symposium on Information Theory (ISIT-06)*, Nice, France, July 2007.

[8] D. Karakos, J. Eisner, S. Khudanpur, and C. E. Priebe, "Cross-instance tuning of unsupervised document clustering algorithms," in *Proc. 2007 Conference of the North American Chapter of the Assoc. for Computational Linguistics (NAACL-HLT 2007)*, April 2007.

[9] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proc. SIGIR'02, 25th ACM Int. Conf. on Research and Development of Inform. Retrieval*, 2002.

[10] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

[11] R. W. Yeung and T. Berger, "Multi-way alternating minimization," in *Proc. IEEE Int. Symposium Information Theory (ISIT-1995)*, September 1995, p. 192.