Chebyshev Center Computation on Probability Simplex With α-Divergence Measure

Çağatay Candan, *Senior Member, IEEE*

***Abstract*—Chebyshev center computation problem, i.e. finding the point which is at minimum distance to a set of given points, on the probability simplex with α-divergence distance measure is studied. The proposed solution generalizes the Arimoto-Blahut (AB) algorithm utilizing Kullback-Leibler divergence to α-divergence, and reduces to the AB method as $\alpha \rightarrow 1$. Similar to the AB algorithm, the method is an ascent method with a guarantee on the objective value (α-mutual information or Chebyshev radius) improvement at every iteration. A practical application area for the method is the fusion of probability mass functions lacking a joint probability description. Another application area is the error exponent calculation.**

***Index Terms*—Arimoto-Blahut algorithm, minimax-redundancy, redundancy-capacity theorem, alpha-divergence, Rényi-divergence, information fusion, error exponent calculation.**

I. Introduction

To ILLUSTRATE the problem of interest, we consider a set of probability mass functions (pmf) $\mathbf{p}_j$, $j = \{1, \ldots, M\}$ each of which is said to represent a locally generated posterior distribution of a random variable of interest. The goal is to fuse the local posteriors to a set-representative pmf $\mathbf{q}$. In the absence of joint distribution information on $\mathbf{p}_j$, $j = \{1, \ldots, M\}$; the Bayesian approach can not be pursued. Instead, a minimax formulation $\widehat{\mathbf{q}} = \arg\min_{\mathbf{q}} \max_j D(\mathbf{p}_j || \mathbf{q})$ where $D(\mathbf{p}_j || \mathbf{q})$ is a distance measure between two distributions (possibly in a loose sense) can be suggested. The minimax solution $\widehat{\mathbf{q}}$ can be interpreted as the point on the probability simplex which minimizes the worst-case distance to the set members. This study presents an efficient method for the solution of the minimax problem with α-divergence distance measure.

The minimax problem $\widehat{\mathbf{q}} = \arg\min_{\mathbf{q}} \max_j D(\mathbf{p}_j || \mathbf{q})$ is also known as the Chebyshev center problem. Its optimizer $\widehat{\mathbf{q}}$ and optimal value $r = \min_{\mathbf{q}} \max_j D(\mathbf{p}_j || \mathbf{q})$ are called Chebyshev center and radius, respectively. The solution of minimax problem with Kullback-Leibler (KL) divergence is of major concern in information theory [1]. Source code design problem with minimax redundancy is identical to the Chebyshev center problem with KL-divergence, [1, Ch. 13]. Furthermore, Gallager and Ryabko [2], [3] have shown that the solution of the minimax problem coincides with the capacity calculation (mutual information maximization over all input distributions) for the special case of discrete memoryless channels with finite input/output alphabets [1, Theorem 13.1.1]. Capacity-redundancy theorem is considered as one of the cornerstone results of information theory and extended to infinite alphabets, probability density functions and measures under different divergences [4]–[7].

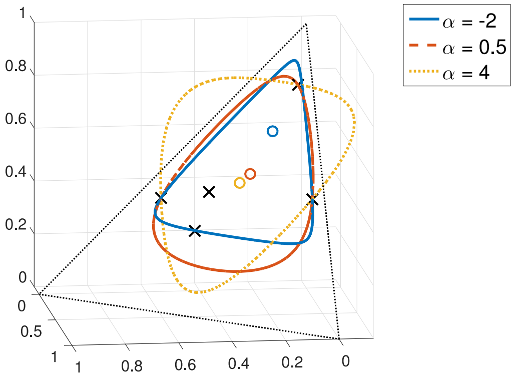

Fig. 1. 5 points (crosses) on probability simplex, Chebyshev centers (round markers) and Chebyshev circles for different orders of α-divergence.

Extending the capacity-redundancy discussions from KL divergence to α-divergence can be motivated by the desire of adjusting inclusiveness or exclusiveness of the solution [5], [8]. Fig. 1 shows five 3-dimensional probability vectors (crosses) and their Chebyshev centers (circles) according to different α-divergence orders. Loosely speaking, the divergence order affects how to measure the distance between points. Chebyshev circle for $\alpha = -2$ in Fig. 1 can be said to be risk-avoiding, that is showing a tendency not to include as many points as other circles. The circle for $\alpha = 0.5$ can be considered to have a balanced risk. The capability of adjusting the degree of inclusiveness/exclusiveness is especially important in approximate inference problems [8].

This study presents a method for the computation of Chebyshev center with α-divergence for orders $\alpha \in (0, 1)$ which is the interval in which risk balancing is possible [8]. The suggested method is an alternating minimization-maximization method with a Chebyshev-radius improvement guarantee at every iteration. The method approaches Arimoto-Blahut (AB) algorithm as $\alpha \rightarrow 1$ [9], [10] and can be interpreted as its generalization to α-divergence measure. Utilized proximal point based approach is an extension of earlier similar efforts by Chretien and Hero, in the context of Expectation-Maximization (EM) algorithm [11]; Matz and Duhamel, in the context of AB algorithm [12], to α-mutual information maximization problem. The method can

Manuscript received June 21, 2020; accepted August 17, 2020. Date of publication August 21, 2020; date of current version September 7, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qing Ling.

The author is with the Department of Electrical, and Electronics Engineering, Middle East Technical University (METU), Ankara 06800, Turkey (e-mail: ccandan@metu.edu.tr).

Digital Object Identifier 10.1109/LSP.2020.3018661

be utilized in information fusion [13], error-exponent calculation [14]–[16] and α-divergence applications [5], [17].

II. Preliminaries

The column vectors $\mathbf{p_j}$, $j = \{1, \ldots, M\}$ of dimension N with nonnegative entries, $p_j(i) \geq 0$, $i = \{1, \ldots, N\}$, denote probability mass functions (pmf) defined over the alphabet $\mathcal{Y} = \{1, 2, \ldots, N\}$. The set of all pmf's for alphabet $\mathcal{Y}$ is shown with $\mathcal{P}^{\mathcal{Y}}$. For $\mathbf{p} \in \mathcal{P}^{\mathcal{Y}}$ and $\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}$, f-divergence is defined as $D_f(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{N} q(i) f(p(i)/q(i))$, where $f(r)$ is a convex function with $f(1) = 0$, [18], [19]. Our focus is on α-divergence $D_\alpha(\mathbf{p}||\mathbf{q})$ which is a special case of f-divergence for $f_\alpha(r) = \frac{1-r^\alpha}{\alpha(1-\alpha)}$, $\alpha \in \mathbb{R} \setminus \{0, 1\}$:

$$D_\alpha(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{N} q(i) f_\alpha\left(\frac{p(i)}{q(i)}\right) = \frac{1 - \sum_{i=1}^{N} p^\alpha(i) q^{1-\alpha}(i)}{\alpha(1-\alpha)}. \tag{1}$$

It can be shown that as $\alpha \to 1$, $D_\alpha(\mathbf{p}||\mathbf{q})$ approaches KL-divergence $\mathrm{KL}(\mathbf{p}||\mathbf{q}) = \sum_i p(i) \log(p(i)/q(i))$ [8]. For additional properties of α-divergence and general properties of f-divergence, one can examine [8], [20].

The Chebyshev center problem can be expressed as

$$P_1 : \quad \min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} \max_{j} D_\alpha(\mathbf{p}_j||\mathbf{q}). \tag{2}$$

The outer minimization in (2) is a convex problem due to i. joint convexity of f-divergences over both arguments [20, Theorem 6.1], ii. convexity preservation by maximum function. Problem P_1 can be reparameterized as:

$$P_2 : \min_{z, \mathbf{q}} z \quad \text{s.t.} \begin{cases} D_\alpha(\mathbf{p}_j||\mathbf{q}) \leq z, \; j = \{1, \ldots, M\} \\ \mathbf{q} \in \mathcal{P}^{\mathcal{Y}} \end{cases}. \tag{3}$$

The equivalent formulation P_2 has linear objective with convex constraints. Lagrangian function for problem P_2 is

$$L(z, \mathbf{q}, \mathbf{p}_x) = z + \sum_{j=1}^{M} p_x(j)(D_\alpha(\mathbf{p}_j||\mathbf{q}) - z), \tag{4}$$

along with the constraint of $\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}$ not noted in (4) for the sake of expression clarity. In (4), $p_x(j)$ are nonnegative valued Lagrange multipliers which are the unknowns of dual problem [21]. Differentiating $L(z, \mathbf{q}, \mathbf{p}_x)$ with respect to z, immediately yields the stationarity condition of $\sum_{j=1}^{M} p_x(j) = 1$. Hence, we reach the important conclusion that elements of $M \times 1$ dimensional Lagrange multiplier vector $\mathbf{p}_x$ are nonnegative valued with a cumulative sum of 1. Therefore, we consider Lagrange multiplier vector $\mathbf{p}_x$ as a probability vector lying in M-dimensional probability simplex $\mathcal{P}^{\mathcal{X}}$, $\mathcal{X} = \{1, \ldots, M\}$.

The dual problem $\mathcal{D} : \max_{\mathbf{p_x}} \min_{z, \mathbf{q}} L(z, \mathbf{q}, \mathbf{p}_x)$ can be stated as

$$\mathcal{D} : \max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} \min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} \sum_{j=1}^{M} p_x(j) D_\alpha(\mathbf{p}_j||\mathbf{q}). \tag{5}$$

In this study, we use of the dual problem statement in (5) to formulate an alternating maximization-minimization solution.

To establish connections with other works in the literature, we introduce $N \times M$ dimensional probability transition matrix (channel) $\mathbf{P}_{Y|X} = [\mathbf{p}_1 \, \mathbf{p}_2 \, \ldots \, \mathbf{p}_M]$ which is formed by the juxtaposition of column vectors $\mathbf{p}_j$. Stated differently, the j'th column of $\mathbf{P}_{Y|X}$ is $\mathbf{P}_{Y|X=j} = \mathbf{p}_j$. With this definition, the objective of dual problem $\mathcal{D}$ coincides with conditional α-divergence definition in [6]:

$$D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x) \triangleq E_{\mathbf{x} \sim \mathbf{p}_x}\{D_\alpha(\mathbf{P}_{Y|X=j}||\mathbf{q}).\}. \tag{6}$$

With this definition, the dual problem (5) can be expressed as:

$$\mathcal{D} : \max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} \underbrace{\min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x)}_{I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x)}. \tag{7}$$

In [22], Sibson introduced a definition for α-mutual information, through information radius considerations, as follows:

$$I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x) \triangleq \min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x). \tag{8}$$

This definition generalizes the conventional mutual-information definition.[1] In fact, the conventional mutual information is the special case of Sibson's definition as $\alpha \to 1$. In this study, we recognize the fact that the dual problem of minimax redundancy for α-divergence measure, say for the sake of information fusion, given in (7) is the capacity maximization problem with Sibson's α-mutual information definition, that is $C_\alpha = \max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x)$.

As a final remark, we note that Slater's strong duality condition [21] is satisfied for the given problem. Hence, with the equality of primal (P_1) and dual optimal values, we have:

$$\begin{aligned} C_\alpha &\triangleq \max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} \min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x) \\ &= \min_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} \max_{j} D_\alpha(\mathbf{p}_j||\mathbf{q}). \end{aligned}$$

The optimizer $\mathbf{q}^* = \mathrm{argmin}_{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}} \max_j D_\alpha(\mathbf{p}_j||\mathbf{q})$ and its value $C_\alpha = \max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x)$ are denoted as the Chebyshev center and radius, respectively. An efficient method for the Chebyshev center computation is the topic of this study.

III. Proposed Method

Proposed method iteratively solves the optimization problem given in (5) by fixing either $\mathbf{p}_x$ or $\mathbf{q}$, alternatively and optimizing over the other variable.

Minimization over $\mathbf{q}$ for a fixed $\mathbf{p}_x^{(k)}$: When $\mathbf{p}_x$ is fixed to $\mathbf{p}_x^{(k)}$, the dual problem in (5) reduces to an optimization problem involving a weighted average of α-divergences,

$$\begin{aligned} \mathbf{q}^{(k)} &= \underset{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}}{\mathrm{argmin}} \sum_{j=1}^{M} p_x^{(k)}(j) D_\alpha(\mathbf{p}_j||\mathbf{q}) \\ &\overset{(b)}{=} \underset{\mathbf{q} \in \mathcal{P}^{\mathcal{Y}}}{\mathrm{argmax}} \sum_{j=1}^{M} p_x^{(k)}(j) \sum_{i=1}^{N} p_j^\alpha(i) q^{1-\alpha}(i). \end{aligned} \tag{9}$$

Line-(b) is obtained based on (1). By calculating the gradient of the objective function in (9), the elements of $\mathbf{q}^{(k)}$ can be written as

$$q^{(k)}(i) = \frac{1}{c\left(\mathbf{p}_x^{(k)}\right)} \left(\sum_{j=1}^{M} p_x^{(k)}(j) P_{Y|X=j}^\alpha(i)\right)^{1/\alpha}. \tag{10}$$

Here $P_{Y|X=j}^\alpha(i)$ is the (i,j) entry of $\mathbf{P}_{Y|X}$ raised to the power α, that is $P_{Y|X=j}^\alpha(i) = ([\mathbf{P}_{Y|X}]_{i,j})^\alpha = p_j^\alpha(i)$ and

[1] Sibson's definition is given for Rényi divergence, which is not an f-divergence; but in one-to-one correspondence with α-divergence in (1). Also see [7, Section I] for more connections with existing works in the literature.

$c(\mathbf{p}_x^{(k)})$ is a normalization constant defined as $c(\mathbf{p}_x^{(k)}) = \sum_{i=1}^{N}(\sum_{j=1}^{M} p_x^{(k)}(j)P^{\alpha}_{Y|X=j}(i))^{1/\alpha}$. The normalization constant $c(\mathbf{p}_x^{(k)})$ can also be expressed in terms of Gallager function [23] as $c(\mathbf{p}_x) = \exp(-E_0(\frac{1-\alpha}{\alpha}, \mathbf{p}_x))$ where

$$E_0(\rho, \mathbf{p}_x) = -\log\left(\sum_{i=1}^{N}\left(\sum_{j=1}^{M} p_x(j)P_{Y|X=j}^{\frac{1}{1+\rho}}(i)\right)^{1+\rho}\right). \tag{11}$$

The optimal vector $\mathbf{q}^{(k)}$ in (10) is denoted as the α-response to $\mathbf{p}_x^{(k)}$ in the literature [6]. It is easy to see that as $\alpha \to 1$, $\mathbf{q}^{(k)} \to \mathbf{P}_{Y|X}\mathbf{p}_x^{(k)}$, which is the output distribution of channel $\mathbf{P}_{Y|X}$ for the input $\mathbf{p}_x^{(k)}$. Normalization constant $c(\mathbf{p}_x)$ is upper bounded by 1 for $\alpha \in (0, 1)$ and $c(\mathbf{p}_x) \to 1$ as $\alpha \to 1$.

Inserting α-response to $\mathbf{p}_x^{(k)}$ into (8) and simplifying, we get the α-mutual information induced by $\mathbf{p}_x^{(k)}$ as

$$I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)}) = \frac{1 - c^{\alpha}(\mathbf{p}_x^{(k)})}{\alpha(1-\alpha)} = \frac{1 - e^{-\alpha E_0\left(\frac{1-\alpha}{\alpha}, \mathbf{p}_x^{(k)}\right)}}{\alpha(1-\alpha)}. \tag{12}$$

Here $I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$ is the value of the dual problem objective, given in (7), at the k'th iteration.

Maximization over $\mathbf{p}_x$ for a fixed $\mathbf{q}^{(k)}$: Exact solution of the dual problem in (7) requires the solution of $\max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x)$ which is guaranteed to be a concave maximization problem by duality [21]. An iterative solution is given by Arimoto in the context of Gallager function maximization in [14]. We present a novel method based on proximal point iterations [24]. Approach is in principle similar to the ones given for EM and AB algorithms in [11], [12].

As a naive attempt, we may try to fix $\mathbf{q}$ to $\mathbf{q}^{(k)}$ in (5), as in earlier sub-problem, and optimize over $\mathbf{p}_x$. This attempt leads to a linear program without any interior point solution and not particularly suitable for an iterative optimization. Instead, we suggest to modify the problem to

$$\mathbf{p}_x^{(k+1)} = \underset{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}}{\operatorname{argmax}}\, f(\mathbf{p}_x, \mathbf{p}_x^{(k)}), \tag{13}$$

and $f(\mathbf{p}_x, \mathbf{p}_x^{(k)}) \triangleq \sum_{j=1}^{M} p_x(j)D_\alpha(\mathbf{p}_j||\mathbf{q}^{(k)}) - \mu^{-1}\mathrm{KL}(\mathbf{p}_x||\mathbf{p}_x^{(k)})$. The modified problem aims to update the solution in the proximity of the previous primal variable estimate $\mathbf{p}_x^{(k)}$. The deviation from earlier iteration $\mathbf{p}_x^{(k)}$ is penalized with μ^{-1}. It is shown that if the step-size parameter μ is chosen properly, the algorithm monotonically converges to the optimum.

The problem in (13) is additively separable. By evaluating $\frac{\partial f(\mathbf{p}_x, \mathbf{p}_x^{(k)})}{\partial p_x(j)} = D_\alpha(\mathbf{p}_j||\mathbf{q}^{(k)}) - \frac{\log(p_x(j))+1-\log(p_x^{(k)}(j))}{\mu}$ and optimizing over the simplex, we get the update equation as:

$$p_x^{(k+1)}(j) = p_x^{(k)}(j)\frac{e^{\mu D_\alpha(\mathbf{p}_j||\mathbf{q}^{(k)})}}{\sum_j \underset{\hookrightarrow P_x^{[k]}(J)}{e}^{\mu D_\alpha(\mathbf{p}_j||\mathbf{q}^{(k)})}}. \tag{14}$$

Selection of step-size parameter: It is shown below that the step-size μ can be selected to guarantee the monotonic increase of dual problem objective, $I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$ in (7), at every iteration.

From (13), $f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x=\mathbf{p}_x^{(k+1)}} \geq f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x=\mathbf{p}_x^{(k)}}$; since $\mathbf{p}_x^{(k+1)}$ is the maximizer of the problem. The evaluation of right hand side, $f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)}) = f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x=\mathbf{p}_x^{(k)}}$, is immediate, $f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)}) = I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$; since $\mathrm{KL}(\mathbf{p}_x^{(k)}||\mathbf{p}_x^{(k)}) = 0$. For the evaluation of left hand side, $f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x=\mathbf{p}_x^{(k+1)}} = f(\mathbf{p}_x^{(k+1)}, \mathbf{p}_x^{(k)})$, we examine the summation in (13), $S = \sum_{j=1}^{M} p_x^{(k+1)}(j)D_\alpha(\mathbf{p}_j||\mathbf{q}^{(k)})$ first:

$$\begin{aligned}
S &\overset{(a)}{=} \sum_{j=1}^{M} p_x^{(k+1)}(j)\sum_{i=1}^{N} q^{(k)}(i)f_\alpha\left(\frac{p_j(i)}{q^{(k)}(i)}\right) \\
&\overset{(b)}{=} \frac{1 - \sum_{j=1}^{M} p_x^{(k+1)}(j)\sum_{i=1}^{N}(q^{(k)}(i))^{1-\alpha}p_j^{\alpha}(i)}{\alpha(1-\alpha)} \\
&\overset{(c)}{=} \frac{1 - c^{\alpha}(\mathbf{p}_x^{(k+1)})\sum_{i=1}^{N} q^{(k)}(i)\left(\frac{q^{(k+1)}(i)}{q^{(k)}(i)}\right)^{\alpha}}{\alpha(1-\alpha)}
\end{aligned} \tag{15}$$

Line-(a) follows the α-divergence definition in (1). In line-(b), $f_\alpha(r) = \frac{1-r^\alpha}{\alpha(1-\alpha)}$ is substituted. In line-(c), summation over j is recognized from (10) as $(q^{(k+1)}(i)c(\mathbf{p_x}^{(k+1)}))^{\alpha}$.

The summation over i on the numerator of line-(c) of (15) is in the form of $\sum_{i=1}^{N} q^{(k)}(i)g_\alpha(\frac{q^{(k+1)}(i)}{q^{(k)}(i)})$, where $g_\alpha(r) = r^\alpha$. In order to write this summation in terms α-divergence with $f_\alpha(r) = \frac{1-r^\alpha}{\alpha(1-\alpha)}$, the function $g_\alpha(r)$ is expressed as $g_\alpha(r) = 1 - f(r)\alpha(1-\alpha)$:

$$\begin{aligned}
S &= \frac{1 - c^{\alpha}(\mathbf{p}_x^{(k+1)})\sum_{i=1}^{N} q^{(k)}(i)g_\alpha\left(\frac{q^{(k+1)}(i)}{q^{(k)}(i)}\right)}{\alpha(1-\alpha)} \\
&= \frac{1 - c^{\alpha}(\mathbf{p}_x^{(k+1)})}{\alpha(1-\alpha)} + c^{\alpha}(\mathbf{p}_x^{(k+1)})D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)}) \\
&\overset{(d)}{=} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k+1)}) + c^{\alpha}(\mathbf{p}_x^{(k+1)})D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)}).
\end{aligned} \tag{16}$$

In line-(d), α-mutual information definition from (12) is recognized. With these results, the inequality of $f(\mathbf{p}_x^{(k+1)}, \mathbf{p}_x^{(k)}) \geq f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)})$ is equivalent to

$$I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k+1)}) - I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)}) \geq \gamma \tag{17}$$

with $\gamma \triangleq \mu^{-1}\mathrm{KL}(\mathbf{p}_x^{(k+1)}||\mathbf{p}_x^{(k)}) - c^{\alpha}(\mathbf{p}_x^{(k+1)})D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)})$. We consider the value of γ as the margin. If margin is positive at an iteration, the objective value (α-mutual information) is improved at that iteration. In fact, setting

$$\mu \leq \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)}||\mathbf{p}_x^{(k)})}{c^{\alpha}(\mathbf{p}_x^{(k+1)})D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)})} \triangleq \bar{\mu}^{(k)} \tag{18}$$

guarantees the positivity of the margin at that iteration. Yet, this step-size selection rule is inadmissible; since the step-size depends on the divergence values after the update. A rather pessimistic approach can be the calculation of a lower bound for the right side of (18):

$$\begin{aligned}
\frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)}||\mathbf{p}_x^{(k)})}{c^{\alpha}(\mathbf{p}_x^{(k+1)})D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)})} &\overset{(a)}{\geq} \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)}||\mathbf{p}_x^{(k)})}{D_\alpha(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)})} \\
&\overset{(b)}{\geq} \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)}||\mathbf{p}_x^{(k)})}{\mathrm{KL}(\mathbf{q}^{(k+1)}||\mathbf{q}^{(k)})} \\
&\overset{(c)}{\geq} 1.
\end{aligned} \tag{19}$$

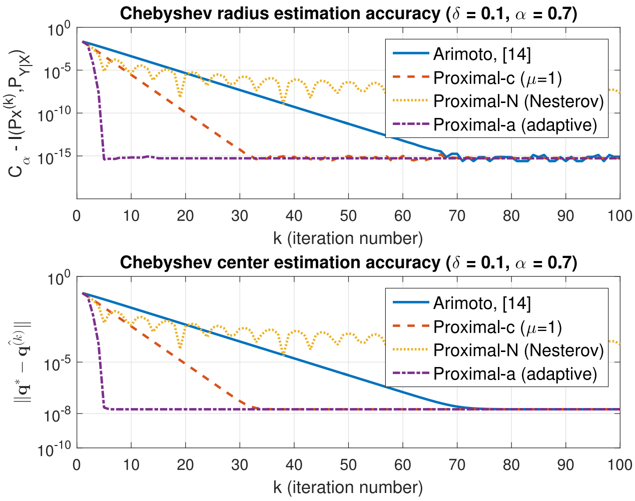

Fig. 2. Example 1 - Convergence of methods to Chebyshev circle parameters.

Line-(a) is due to the fact that $c(\mathbf{p}_x^{(k+1)}) = \exp(-E_0(\frac{1-\alpha}{\alpha}, \mathbf{p}_x)) \leq 1$ for $\alpha \in (0,1)$. Line-(b) follows from the monotonic increase of α-divergence with divergence order α, [6, Property 9]. Line-(c) is the data-processing inequality for KL-divergence [20], [25]. Hence, the choice of $\mu = 1$ guarantees the monotonic convergence. (Also, for $\mu = 1$, the method reduces to AB algorithm as $\alpha \to 1$.)

IV. Numerical Results

Example 1: Probability vectors $\mathbf{p}_j$, $j = \{1,2,3\}$ of dimension 4×1 are concatenated to form the columns of channel $\mathbf{P}_{Y|X}$, given below, which is also studied in [6, Ex. 6]:

$$\mathbf{P}_{Y|X} = \begin{bmatrix} \frac{1}{2} - \delta & \delta & \frac{1}{2} - \delta \\ \delta & \frac{1}{2} - \delta & \delta \\ \delta & \frac{1}{2} - \delta & \delta \\ \frac{1}{2} - \delta & \delta & \frac{1}{2} - \delta \end{bmatrix}, \quad \delta \in \left[0, \frac{1}{2}\right] \qquad (20)$$

Due to the symmetry in the problem, the Chebyshev center, or α-capacity achieving output distribution, is $[\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}]$. The Chebyshev radius, or α-capacity expression, is $C_\alpha = [1 - 2^{2\alpha-1}(\delta^\alpha + (\frac{1}{2} - \delta)^\alpha)]/[\alpha(1-\alpha)]$.

Fig. 2 shows the accuracy improvement for Chebyshev center and radius estimates versus iterations for $\delta = 0.1$ and α-divergence order $\alpha = 0.7$. Arimoto's error exponent calculation algorithm [14] and different versions of proximal point algorithm are compared. The curve with "Proximal-c" label shows the case with $\mu = 1$ at every iteration. "Proximal-N" shows the version with Nesterov's acceleration as given in [26]. "Proximal-a" is the adaptive version of the suggested scheme where μ changes at every iteration.

In "Proximal-a," the step-size of current iteration is taken as the bound $\bar{\mu}$ in (18) calculated from the data of the previous iteration, $\mu^{(k)} = \bar{\mu}^{(k-1)}$. It is observed from Fig. 2 that "Proximal-a" converges rapidly to the accuracy of the computing platform with this policy. Table I shows the number of iterations required to get 10 digit accuracy in the capacity estimate. Main conclusions of this experiment are i.) "Proximal-c" with a constant step-size of $\mu = 1$ has a monotonic performance as expected and outperforms Arimoto's method, ii.) Nesterov's acceleration brings some improvements over "Proximal-c"; but requires more effort for its adaptation to the probability simplex, iii.) "Proximal-a" with suggested step-size policy yields very rapid convergence.

TABLE I
Number of Iterations Required for 10 Digit Accuracy at α-Capacity in Example 1 Setting ($\delta = 0.1$)

	Arimoto [14]	Proximal-N	Proximal-c	Proximal-a
$\alpha = 0.1$	2034	119	17	5
$\alpha = 0.3$	228	72	18	5
$\alpha = 0.5$	84	79	19	5
$\alpha = 0.7$	44	101	21	5
$\alpha = 0.9$	28	10	22	5

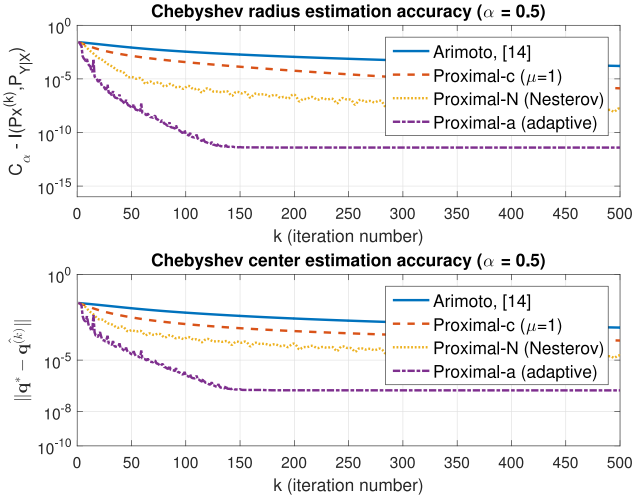

Fig. 3. Example 2 - Convergence of methods to Chebyshev circle parameters.

Example 2: This example presents a higher dimensional comparison with $M = 25$ vectors on $N = 100$ dimensional probability simplex. Fig. 3 shows the results for randomly sampled probability vectors on the simplex. Analytical expressions for the Chebyshev center and radius are not available and calculated by the general purpose numerical optimization routines. We see that Arimoto's method and "Proximal-c" present monotonic performance improvements; while "Proximal-a" suffers intermittent performance losses. In this example, we limit the adaptive step size range to $\mu \in [1, 50]$ by setting $\mu^{(k)} = \min(50, \bar{\mu}^{(k-1)})$. Here, the factor 50 indicates that acceleration over "Proximal-c" can be up to 50 fold and observed sporadic performance losses are due to over-acceleration.

Computational Complexity Discussion: The sum for the update in (10) corresponds to a matrix and vector product, with a complexity $\mathcal{O}(NM)$ multiplications, assuming that $P^\alpha_{Y|X=j}(i)$ is pre-computed and stored. The update in (14) requires M fold α-divergence calculation, an operation of complexity $\mathcal{O}(NM)$ multiplications, and an additional $\mathcal{O}(M)$ multiplications. Hence, the overall complexity of suggested scheme is $\mathcal{O}(NM)$ multiplications per iteration.

V. Conclusion

An efficient method for the computation of Chebyshev center and radius (maximum α-mutual information) with α-divergence measure for finite samples spaces is given. The scheme generalizes the celebrated Arimoto-Blahut algorithm to α-divergence measure. An important open problem is the extension of the suggested method to continuous random variables. Ready-to-use Matlab codes reproducing the results in paper (and more) are available in [27].

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[2] R. G. Gallager, "Source coding with side information and universal coding," Laboratory for Information Decision Systems, MIT, Cambridge, MA, USA, Tech. Rep. LIDS-P-937, 1979.

[3] B. Y. Ryabko, "Coding a source with unknown but ordered probabilities," *Probl. Inf. Transmission*, vol. 15, no. 2, pp. 134–138, Oct. 1979.

[4] J. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae (Proc.)*, vol. 77, no. 2, pp. 101–115, 1974.

[5] S. Yagli, Y. Altug, and S. Verdú, "Minimax Rényi redundancy," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3715–3733, May 2018.

[6] C. Cai and S. Verdú, "Conditional Rényi divergence saddlepoint and the maximization of α-mutual information," *Entropy*, vol. 21, no. 10, 2019, Art. no. 969. [Online]. Available: https://www.mdpi.com/1099-4300/21/10/969

[7] B. Nakiboglu, "The Rényi capacity and center," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 841–860, Feb. 2019.

[8] T. Minka, "Divergence measures and message passing," Tech. Rep. MSR-TR-2005-173, Jan. 2005. [Online]. Available: https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/

[9] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[10] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[11] S. Chretien and A. O. Hero, "Kullback proximal algorithms for maximum-likelihood estimation," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1800–1810, Aug. 2000.

[12] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut–Arimoto-type algorithms," in *Proc. Inf. Theory Workshop*, 2004, pp. 66–70.

[13] S. J. Julier, "An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models," in *Proc. Int. Conf. Inf. Fusion*, 2006, pp. 1–8.

[14] S. Arimoto, "Computation of random coding exponent functions," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 6, pp. 665–671, Nov. 1976.

[15] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 1327–1333.

[16] Y. Jitsumatsu and Y. Oohama, "A new iterative algorithm for computing the correct decoding probability exponent of discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1585–1606, Mar. 2020.

[17] I. Sason and S. Verdú, "Arimoto–Rényi conditional entropy and Bayesian M-Ary hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 4–25, Jan. 2018.

[18] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Statist. Soc. B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[19] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[20] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," MIT, Cambridge, MA, USA, 2019.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[22] R. Sibson, "Information radius," *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[23] R. G. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ, USA: Wiley, 1968.

[24] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, 1976.

[25] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[26] R. Tibshirani, "Lecture notes on convex optimization," CMU, Pittsburgh, PA, USA, 2019.

[27] C. Candan. (2020). "Chebyshev center computation on probability simplex with α-divergence measure [source code]." [Online]. Available: https://doi.org/10.24433/CO.3654345.v1