Projection Theorems for the Rényi Divergence on α-Convex Sets

M. Ashok Kumar   Igal Sason

Abstract

This paper studies forward and reverse projections for the Rényi divergence of order $\alpha \in (0, \infty)$ on α-convex sets. The forward projection on such a set is motivated by some works of Tsallis *et al.* in statistical physics, and the reverse projection is motivated by robust statistics. In a recent work, van Erven and Harremoës proved a Pythagorean inequality for Rényi divergences on α-convex sets under the assumption that the forward projection exists. Continuing this study, a sufficient condition for the existence of a forward projection is proved for probability measures on a general alphabet. For $\alpha \in (1, \infty)$, the proof relies on a new Apollonius theorem for the Hellinger divergence, and for $\alpha \in (0, 1)$, the proof relies on the Banach-Alaoglu theorem from functional analysis. Further projection results are then obtained in the finite alphabet setting. These include a projection theorem on a specific α-convex set, which is termed an *α-linear family*, generalizing a result by Csiszár to $\alpha \neq 1$. The solution to this problem yields a parametric family of probability measures which turns out to be an extension of the exponential family, and it is termed an *α-exponential family*. An orthogonality relationship between the α-exponential and α-linear families is established, and it is used to turn the reverse projection on an α-exponential family into a forward projection on an α-linear family. This paper also proves a convergence result of an iterative procedure used to calculate the forward projection on an intersection of a finite number of α-linear families.

Keywords: α-convex set, exponential and linear families, relative entropy, variational distance, forward projection, reverse projection, Hellinger divergence, Rényi divergence.

M. Ashok Kumar is with the Indian Institute of Technology, Indore 453552, India (e-mail: ashokm@iiti.ac.in). The work was partially done while he was a post-doctoral fellow at the Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

I. Sason is with the Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sason@ee.technion.ac.il).

This work has been supported by the Israeli Science Foundation (ISF) under Grant 12/12. It has been presented in part at the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain, July 10–15, 2016.

I. Introduction

Information projections of relative entropy have been extensively studied due to their various applications in large deviations theory (e.g., Sanov's theorem and the conditional limit theorem), maximum likelihood estimation (MLE), statistical physics, and so on. Some of the pioneering works studying information projections include Barron [2], Čencov [5], Chentsov [6], Csiszár [12], [13], Csiszár and Matúš [15], and Topsøe [40]. The broader subject areas using information projections as a major component are known as *Information Theory and Statistics* and *Information Geometry* (see, e.g., [7, Chapter 11], [16] and references therein).

Given a probability measure Q, and a set of probability measures $\mathcal{P}$ defined on an alphabet $\mathcal{A}$, a *forward projection* of Q on $\mathcal{P}$ is a $P^* \in \mathcal{P}$ which minimizes $D(P\|Q)$ subject to $P \in \mathcal{P}$. Forward projections appear predominantly in large deviations theory. By Sanov's theorem, the exponential decay rate of the probability of rare events is strongly related to forward projections (see [7, Theorem 11.4.1]); furthermore, in view of the conditional limit theorem, the forward projection of Q on $\mathcal{P}$ arises as the limiting conditional probability measure of a random variable with distribution $Q \notin \mathcal{P}$, given that the type of its i.i.d. samples belongs to $\mathcal{P}$ (see [7, Theorem 11.6.2]). The forward projection of a generalization of the relative entropy has been proposed by Sundaresan in [38] and [39] in the context of guessing under source uncertainty, and it was further studied in [24].

The Rényi divergence, introduced in [32] and further studied, e.g., in [18] and [37], has been investigated so far in various information-theoretic contexts. These include generalized cutoff rates and error exponents for hypothesis testing (e.g., [14]), guessing moments (e.g., [17]), source and channel coding error exponents (e.g., [20], [34], [36]), and other information-theoretic problems.

A motivation for the study of forward projections for the Rényi divergence on some generalized convex sets stems from the following maximum entropy problem which was proposed by Tsallis in statistical physics [41], [42]:

$$\arg\max_{(p_i)} \; S_\alpha(P) := \frac{1}{\alpha-1}\Bigl(1-\sum_{i=1}^{W} p_i^\alpha\Bigr) \tag{1}$$

$$\text{subject to} \quad \frac{\sum_{i=1}^{W} p_i^\alpha \epsilon_i}{\sum_{i=1}^{W} p_i^\alpha} = U^{(\alpha)}, \tag{2}$$

where $\alpha \in (0,1) \cup (1,\infty)$ is a free parameter, W is the number of microscopic states, $\{\epsilon_i\}$ are the eigenvalues of the Hamiltonian, and $U^{(\alpha)}$ is the total internal energy of the system. The functional $S_\alpha(P)$ in (1) is known as the *Tsallis entropy*. The constraint in (2) is on the *escort probability measure*

$$\begin{aligned} P^{(\alpha)} &:= \bigl(P_1^{(\alpha)}, \ldots, P_W^{(\alpha)}\bigr), \\ P_i^{(\alpha)} &:= \frac{p_i^\alpha}{\sum_{j=1}^{W} p_j^\alpha}, \quad i \in \{1, \ldots, W\} \end{aligned} \tag{3}$$

in contrast to the usual constraint in the Boltzmann-Gibbs statistical physics

$$\sum_{i=1}^{W} p_i \epsilon_i = U^{(1)}. \tag{4}$$

The constraint in (2) corresponds to an α-linear family (to be formally defined in Section IV), whereas (4) corresponds to a linear family [25, Definition 4]. If $Q = U$ is the equiprobable measure on the state space $\{1, \ldots, W\}$, then the Rényi divergence $D_\alpha(P\|U)$ is related to the objective function $S_\alpha(P)$ in (1) via the equation

$$D_\alpha(P\|U) = \log W + \frac{1}{\alpha - 1} \log\bigl(1 - (\alpha - 1) S_\alpha(P)\bigr) \tag{5}$$

which implies that the maximization of $S_\alpha(P)$ over the set which is defined in (2) is equivalent to the minimization of $D_\alpha(P\|U)$ on the same set of probability measures in (2) which corresponds to an α-convex set.

The other problem of interest in this paper is the *reverse projection* where the minimization is over the second argument of the divergence measure. This problem is intimately related to maximum-likelihood estimation and robust statistics. Suppose $X_1, \ldots, X_n$ are i.i.d. samples drawn according to a probability measure which is modelled by a parametric family of probability measures $\Pi = \{P_\theta \colon \theta \in \Theta\}$ where Θ is a parameter space, and all the members of Π are assumed to have a common finite support $\mathcal{A}$. The maximum-likelihood estimator of the given samples (if it exists) is the minimizer of $D(\hat{P}\|P_\theta)$ subject to $P_\theta \in \Pi$, where $\hat{P}$ is the empirical probability measure of the observed samples (see, e.g., [16, Lemma 3.1]). The minimizing probability measure (if it exists) is called the reverse projection of $\hat{P}$ on Π. Other divergences that have natural connection to statistical estimation problems include the Hellinger divergence of order $\frac{1}{2}$ (see, e.g., [4]), Pearson's χ^2-divergence [30], and so on. All of these information measures are f-divergences ([1], [9]) in the family of Hellinger divergences of order $\alpha \in (0, \infty)$ (note that, up to a positive scaling factor, Hellinger divergences are equal to the power divergences introduced by Cressie and Read [8]). The Hellinger divergences possess a very good robustness property when a significant fraction of the observed samples are outliers; the textbooks by Basu et al. [3] and Pardo [29] address the developments of studies on inference based on f-divergences. Since the Rényi divergence is a monotonically increasing function of the Hellinger divergence (as it follows from (14)), minimizing the Hellinger divergence of order $\alpha \in (0, \infty)$ is equivalent to minimizing the Rényi divergence of the same order. This motivates the study of reverse projections of the Rényi divergence in the context of robust statistics. In [27, Section 4], an iterative message-passing algorithm (a.k.a. belief propagation) was used to approximate reverse projections for the Rényi divergence.

In the following, we further motivate our study of forward and reverse projections for the Rényi divergence of order $\alpha \in (0, \infty)$ on α-convex sets (note that these terms are formally defined in Section II):

a) In view of existing projection theorems for the relative entropy (e.g., [6], [12], [13], [15]) and Sundaresan's relative α-entropy on convex sets [24], [25]), we study forward and reverse projections for the Rényi divergence of order $\alpha \in (0, \infty)$ on α-convex sets. Our problem reduces to the study of information projections for the relative entropy on convex sets when $\alpha = 1$. Note also that the Rényi divergence $D_\alpha(P\|Q)$ and Sundaresan's relative α-entropy $\mathscr{I}_\alpha(P, Q)$ are related according to the equality (see [24, Lemma 2c)])

$$\mathscr{I}_\alpha(P, Q) = D_{\frac{1}{\alpha}}(P^{(\alpha)}\|Q^{(\alpha)}) \tag{6}$$

where $P^{(\alpha)}$ and $Q^{(\alpha)}$ are, respectively, the associated escort probability measures of P and Q in (3).

b) In a recent work [18], van Erven and Harremoës proved a Pythagorean inequality for Rényi divergences of order $\alpha \in (0, \infty)$ on α-convex sets under the assumption that the forward projection exists.[1] Continuing this study, one of the main objectives of this work is to provide a sufficient condition for the existence of such a forward projection on an α-convex set of probability measures defined on a general alphabet (see Theorem 1). Our proof is inspired by the proof of the existence of the forward projection for the relative α-entropy on a convex set (see [24, Proposition 6] and [24, Theorem 8]).

c) Forward projections of the relative entropy on linear families and their orthogonality relationship to exponential families were studied in [12] and [16, Chapter 3]. We generalize these results by studying forward projection theorems for the Rényi divergence on α-linear families. The solution of this problem yields a parametric family of probability measures which turns out to be an extension of the exponential family, and it is termed an *α-exponential family*. An orthogonality relationship between the α-exponential and α-linear families is also established.

d) The orthogonality property of linear and exponential families was used to transform a reverse projection of relative entropy on an exponential family into a forward projection on a linear family [16, Theorem 3.3]. In this work, we make use of the generalized orthogonality relationship in Item c) to transform a reverse projection for the Rényi divergence of order α on an α-exponential family into a forward projection on an α-linear family.

e) In [12, Theorem 3.2], Csiszár proposed a convergent iterative process for finding the forward projection for the relative entropy on a finite intersection of linear families. This result is generalized in this work for the Rényi divergence of order $\alpha \in (0, \infty)$ on a finite intersection of α-linear families.

The following is an outline of the paper. Section II provides preliminary material which is essential to this paper. In Section III, we study a sufficient condition for the existence of the forward projection for the Rényi divergence on generalized convex sets. In Section IV, we revisit the Pythagorean property for Rényi divergence and prove the iterated projections property as a consequence. In Section V, we establish the form of forward D_α-projection on an α-linear family and identify the α-exponential family as an extension of the exponential family. In Section VI, we establish an orthogonality relationship between the α-linear and α-exponential families, and in Section VII we use this orthogonality property to convert the reverse projection on an α-exponential family into a forward projection on an α-linear family. Finally, Section VIII briefly summarizes this paper and provides some concluding remarks.

II. Preliminaries

In this section, we set the notation and formally define terms which are used in this paper.

Let $(\mathcal{A}, \mathscr{F})$ be a measurable space, and let $\mathcal{M}$ denote the space of all probability measures defined on $\mathcal{A}$.

Definition 1: For $P, Q \in \mathcal{M}$, the *total variation distance* between P and Q is defined as

$$|P - Q| := 2 \sup_{\mathcal{F} \in \mathscr{F}} |P(\mathcal{F}) - Q(\mathcal{F})|. \tag{7}$$

If P and Q are absolutely continuous with respect to a common σ-finite measure μ (denoted by $P, Q \ll \mu$), let $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}, q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ denote their respective densities (Radon-Nikodym derivatives) with respect to μ (called μ-densities). Then,

$$|P - Q| := \|p - q\|_1 = \int |p - q| \, \mathrm{d}\mu, \tag{8}$$

[1]It should be noted that the Rényi divergence does not necessarily satisfy a Pythagorean inequality on convex sets. For a counter example, see [19, Appendix A on p. 19].

and $\mathcal{M}$ together with the total variation distance forms a metric space. Throughout the paper, the Lebesgue integrals are over the set $\mathcal{A}$.

Pinsker's inequality [31] states that

$$\tfrac{1}{2}\,|P-Q|^2 \log e \le D(P\|Q). \tag{9}$$

Eq. (9) was proved by Csiszár [10] and Kullback [23], with Kemperman [22] independently a bit later. From Pinsker's inequality (9), it follows that convergence in relative entropy also yields convergence in total variation distance (i.e., if $D(P_n\|P) \to 0$ as $n \to \infty$, then $|P_n - P| \to 0$).

Definition 2 (Rényi divergence): Let $\alpha \in (0,1) \cup (1,\infty)$. The *Rényi divergence* [32] of order α from P to Q is given by

$$D_\alpha(P\|Q) := \frac{1}{\alpha-1}\,\log\left(\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu\right). \tag{10}$$

If $\alpha = 1$, then

$$D_1(P\|Q) := D(P\|Q), \tag{11}$$

which is the continuous extension of $D_\alpha(P\|Q)$ at $\alpha = 1$.

Definition 3: The Hellinger divergence [26, Definition 2.10] of order $\alpha \in (0,1)\cup(1,\infty)$ from P to Q is given by

$$\mathscr{H}_\alpha(P\|Q) := \frac{1}{\alpha-1}\,\left(\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu - 1\right). \tag{12}$$

The continuous extension of $\mathscr{H}_\alpha(P\|Q)$ at $\alpha = 1$ yields

$$\mathscr{H}_1(P\|Q)\,\log e = D(P\|Q). \tag{13}$$

Note that $|P-Q|$, $D_\alpha(P\|Q)$ and $\mathscr{H}_\alpha(P\|Q)$ are non-negative, and are equal to zero if and only if $P = Q$. These measures can be expressed in terms of f-divergences [1], [9], [10], and they do not depend on the choice of the reference measure μ. Note that, from (10) and (12),

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\,\log\bigl(1 + (\alpha-1)\mathscr{H}_\alpha(P\|Q)\bigr), \tag{14}$$

showing that the Rényi divergence is monotonically increasing with the Hellinger divergence.

Definition 4 ((α,λ)-mixture [18]): Let $P_0, P_1 \ll \mu$, let $\alpha \in (0,\infty)$, and let $\lambda \in (0,1)$. The (α,λ)-*mixture* of (P_0, P_1) is the probability measure $S_{0,1}$ with μ-density

$$s_{0,1} := \frac{1}{Z}\Bigl[(1-\lambda)p_0^\alpha + \lambda p_1^\alpha\Bigr]^{\frac{1}{\alpha}}, \tag{15}$$

where Z is a normalizing constant such that $\int s_{0,1}\; d\mu = 1$, i.e.,

$$Z = \int \Bigl[(1-\lambda)p_0^\alpha + \lambda p_1^\alpha\Bigr]^{\frac{1}{\alpha}}\,\mathrm{d}\mu. \tag{16}$$

Here, for simplicity, we suppress the dependence of $S_{0,1}$ and Z on α, λ. Note that $s_{0,1}$ is well-defined as Z is always positive and finite. Indeed, for $\alpha \in (0,\infty)$ and $\lambda \in [0,1]$,

$$0 \le \Bigl[(1-\lambda)p_0^\alpha + \lambda p_1^\alpha\Bigr]^{\frac{1}{\alpha}} \le \max\{p_0, p_1\} \le p_0 + p_1 \tag{17}$$

which implies that $0 < Z \le 2$. From (15), for $\lambda \in [0,1]$, the (α,λ)-mixture of (P_0, P_1) is the same as the $(\alpha, 1-\lambda)$-mixture of (P_1, P_0).

Definition 5 (α-convex set): Let $\alpha \in (0,\infty)$. A set of probability measures $\mathcal{P}$ is said to be α-*convex* if, for every $P_0, P_1 \in \mathcal{P}$ and $\lambda \in (0,1)$, the (α,λ)-mixture $S_{0,1} \in \mathcal{P}$.

III. Existence of Forward D_α-Projections on α-Convex Sets

In this section, we define what we mean by a forward D_α-projection, and then provide a sufficient condition for the existence of forward D_α-projections on α-convex sets.

Definition 6 (Forward D_α-projection): Let $Q \in \mathcal{M}$, $\mathcal{P} \subseteq \mathcal{M}$, and $\alpha \in (0, \infty)$. If there exists $P^* \in \mathcal{P}$ which attains the global minimum of $D_\alpha(P\|Q)$ over all $P \in \mathcal{P}$ and $D_\alpha(P^*\|Q) < \infty$, then P^* is said to be a *forward D_α-projection* of Q on $\mathcal{P}$.

We next proceed to show the existence of a forward D_α-projection on an α-convex set. It has been shown in [18, Theorem 14] that if $\mathcal{P}$ is an α-convex set and P^* exists, then the *Pythagorean inequality* holds, i.e.,

$$D_\alpha(P\|Q) \geq D_\alpha(P\|P^*) + D_\alpha(P^*\|Q), \quad \forall\, P \in \mathcal{P}. \tag{18}$$

However, the existence of the forward D_α-projection was not addressed in [18]. We show that if the α-convex set $\mathcal{P}$ is closed with respect to the total variation distance, then the forward D_α-projection exists. The proof is inspired by the proof of the existence of a forward projection for the relative α-entropy on a convex set [24, Theorem 8]. Before getting to the main result of this section, we prove the following inequality for the Hellinger divergence.

Lemma 1 (Apollonius theorem for the Hellinger divergence): If $\alpha \in (1, \infty)$, $\lambda \in (0, 1)$, and P_0, P_1, Q are probability measures where $P_0, P_1, Q \ll \mu$, then

$$\begin{aligned}&(1-\lambda)\big(\mathscr{H}_\alpha(P_0\|Q) - \mathscr{H}_\alpha(P_0\|S_{0,1})\big)\\ &+ \lambda\big(\mathscr{H}_\alpha(P_1\|Q) - \mathscr{H}_\alpha(P_1\|S_{0,1})\big) \geq \mathscr{H}_\alpha(S_{0,1}\|Q),\end{aligned} \tag{19}$$

and the inequality in (19) is reversed for $\alpha \in (0, 1)$.

Proof: The left side of (19) simplifies to

$$\begin{aligned}&\big(1-\lambda)(\mathscr{H}_\alpha(P_0\|Q) - \mathscr{H}_\alpha(P_0\|S_{0,1})\big) + \lambda\big(\mathscr{H}_\alpha(P_1\|Q) - \mathscr{H}_\alpha(P_1\|S_{0,1})\big)\\
&= \frac{1-\lambda}{\alpha-1}\int p_0^\alpha\big(q^{1-\alpha} - s_{0,1}^{1-\alpha}\big)\,\mathrm{d}\mu + \frac{\lambda}{\alpha-1}\int p_1^\alpha\big(q^{1-\alpha} - s_{0,1}^{1-\alpha}\big)\,\mathrm{d}\mu\\
&= \frac{1}{\alpha-1}\int\big((1-\lambda)p_0^\alpha + \lambda p_1^\alpha\big)\big(q^{1-\alpha} - s_{0,1}^{1-\alpha}\big)\,\mathrm{d}\mu\\
&= \frac{1}{\alpha-1}\int Z^\alpha\, s_{0,1}^\alpha\,\big(q^{1-\alpha} - s_{0,1}^{1-\alpha}\big)\,\mathrm{d}\mu\\
&= \frac{Z^\alpha}{\alpha-1}\left(\int s_{0,1}^\alpha\, q^{1-\alpha}\,\mathrm{d}\mu - 1\right)\\
&= Z^\alpha\,\mathscr{H}_\alpha(S_{0,1}\|Q).\end{aligned} \tag{20}$$

The result follows since, by invoking Jensen's inequality to (16) (see [18, Lemma 3]), $Z \geq 1$ if $\alpha \in (1, \infty)$, and $0 < Z \leq 1$ if $\alpha \in (0, 1)$. ∎

Remark 1: Lemma 1 is analogous to the Apollonius theorem for the relative α-entropy [24, Proposition 6] where $S_{0,1}$ is replaced by a convex combination of P_0 and P_1. In view of (13) and since $Z = 1$ when $\alpha = 1$ (see (16)), it follows that (20) reduces to the *parallelogram law* for the relative entropy [12, (2.2)] when $\alpha = 1$ and $\lambda = \frac{1}{2}$.

We are now ready to state our first main result.

Theorem 1 (Existence of forward D_α-projection): Let $\alpha \in (0, \infty)$, and let Q be an arbitrary probability measure defined on a set $\mathcal{A}$. Let $\mathcal{P}$ be an α-convex set of probability measures defined on $\mathcal{A}$, and assume that $\mathcal{P}$ is closed with respect to the total variation distance. If there exists $P \in \mathcal{P}$ such that $D_\alpha(P\|Q) < \infty$, then there exists a forward D_α-projection of Q on $\mathcal{P}$.

Proof: We first consider the case where $\alpha \in (1, \infty)$. Let $\{P_n\}$ be a sequence in $\mathcal{P}$ such that $D_\alpha(P_n \| Q) < \infty$ and $D_\alpha(P_n \| Q) \to \inf_{P \in \mathcal{P}} D_\alpha(P \| Q) =: D_\alpha(\mathcal{P} \| Q)$. Then, in view of (14), $\mathscr{H}_\alpha(P_n \| Q) < \infty$ and $\mathscr{H}_\alpha(P_n \| Q) \to \inf_{P \in \mathbb{E}} \mathscr{H}_\alpha(P \| Q) =: \mathscr{H}_\alpha(\mathcal{P} \| Q)$.

Let $m, n \in \mathbb{N}$, and let $S_{m,n}$ be the (α, λ)-mixture of (P_m, P_n), i.e., $S_{m,n}$ is the probability measure with μ-density

$$s_{m,n} = \frac{1}{Z_{m,n}} \left[(1-\lambda) p_m^\alpha + \lambda p_n^\alpha \right]^{1/\alpha}, \tag{21}$$

where $Z_{m,n}$ is the normalizing constant such that $\int s_{m,n} \, d\mu = 1$. Applying Lemma 1, we have

$$0 \le (1-\lambda) \mathscr{H}_\alpha(P_m \| S_{m,n}) + \lambda \mathscr{H}_\alpha(P_n \| S_{m,n}) \tag{22}$$

$$\le (1-\lambda) \mathscr{H}_\alpha(P_m \| Q) + \lambda \mathscr{H}_\alpha(P_n \| Q) - \mathscr{H}_\alpha(S_{m,n} \| Q). \tag{23}$$

Since $\mathscr{H}_\alpha(P_n \| Q) \to \mathscr{H}_\alpha(\mathcal{P} \| Q)$ as we let $n \to \infty$, and $\mathscr{H}_\alpha(S_{m,n} \| Q) \ge \mathscr{H}_\alpha(\mathcal{P} \| Q)$ (note that $S_{m,n} \in \mathcal{P}$ due to the α-convexity of the set $\mathcal{P}$), the limit supremum of the right side of (23) is non-positive as $n, m \to \infty$. From the left side of (22), the limit infimum of the right side of (23) is also non-negative. This implies that the limit of the right side of (23) is zero, which also implies that the right side of (22) converges to zero as we let $m, n \to \infty$; consequently, $\mathscr{H}_\alpha(P_n \| S_{m,n}) \to 0$ and $\mathscr{H}_\alpha(P_m \| S_{m,n}) \to 0$ as $m, n \to \infty$. Since the Hellinger divergence, $\mathscr{H}_\alpha(\cdot \| \cdot)$, is monotonically increasing in α [26, Proposition 2.7][2], it follows from (13) that $D(P_n \| S_{m,n}) \to 0$ and $D(P_m \| S_{m,n}) \to 0$ as $m, n \to \infty$, which, in turn implies (via Pinsker's inequality (9)) that $|P_n - S_{m,n}| \to 0$ and $|P_m - S_{m,n}| \to 0$ as $m, n \to \infty$. The triangle inequality for the total variation distance yields that $|P_n - P_m| \to 0$ as $m, n \to \infty$, i.e., $\{P_n\}$ is a Cauchy sequence in $\mathcal{P}$, which therefore converges to some $P^* \in \mathcal{P}$ due to the completeness of $\mathcal{P}$ with respect to the total variation distance. Subsequently, the corresponding sequence of μ-densities $\{p_n\}$ converges to the μ-density p^* in L^1; this implies that there exists a sub-sequence $\{p_{n_k}\}$ which converges μ-almost everywhere (a.e.) to p^*. By Fatou's lemma and (12), it follows that for $\alpha \in (1, \infty)$

$$\begin{aligned} \mathscr{H}_\alpha(\mathcal{P} \| Q) &= \lim_{n \to \infty} \mathscr{H}_\alpha(P_n \| Q) \\ &= \lim_{k \to \infty} \mathscr{H}_\alpha(P_{n_k} \| Q) \\ &\ge \mathscr{H}_\alpha(P^* \| Q) \end{aligned} \tag{24}$$

which implies that P^* is a forward $\mathscr{H}_\alpha$-projection of Q on $\mathcal{P}$. In view of (14), this is equivalent to saying that P^* is a forward D_α-projection of Q on $\mathcal{P}$.

We next consider the case where $\alpha \in (0, 1)$. Abusing notation a little, we use the same letter $\mathcal{P}$ to denote a set of probability measures as well as the set of their corresponding μ-densities. Since $\alpha < 1$,

$$\inf_{P \in \mathcal{P}} D_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \left(\sup_{p \in \mathcal{P}} \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu \right) \tag{25}$$

$$= \frac{1}{\alpha - 1} \log \left(\sup_{g \in \widehat{\mathcal{P}}} \int gh \, \mathrm{d}\mu \right), \tag{26}$$

where $g := sp^\alpha, h := q^{1-\alpha}$ and

$$\widehat{\mathcal{P}} := \{ sp^\alpha \colon p \in \mathcal{P}, \ 0 \le s \le 1 \}. \tag{27}$$

[2]A simple proof of the monotonicity of the Hellinger divergence in α appears in [35, Theorem 33].

Notice that the multiplication of p^α by the scalar $s \in [0,1]$ in the right side of (27) does not affect the supremum in (26). This supremum, if attained, is obtained by some $g = sp^\alpha$ with $s = 1$ and $p \in \mathcal{P}$. The purpose of introducing $s \in [0,1]$ is to make the optimization in (26) over a convex set (as it is shown in the sequel).

Let $\beta = \frac{1}{\alpha}$ and $\beta' := \frac{1}{1-\alpha}$; note that β and β' are Hölder conjugates (i.e., $\frac{1}{\beta} + \frac{1}{\beta'} = 1$). Then, $\int h^{\beta'} \, \mathrm{d}\mu = \int q \, \mathrm{d}\mu = 1$, so $h \in L^{\beta'}(\mu)$. By invoking Hölder's inequality, it follows that $F_h(g) := \int gh \, \mathrm{d}\mu$ is a continuous linear functional on $L^\beta(\mu)$. Thus, the supremum is of a continuous linear functional on the reflexive Banach space $L^\beta(\mu)$. We claim that $\widehat{\mathcal{P}}$ is closed and convex in $L^\beta(\mu)$. For the moment, we assume that the claim holds, and later prove it. A convex set which is closed with respect to the norm topology is also closed with respect to the weak topology [33, Ch. 10, Cor. 23]. Note that the weak topology on $L^\beta(\mu)$ is the smallest topology on $L^\beta(\mu)$ for which the continuity of the linear functionals on $L^\beta(\mu)$ is preserved. Moreover, for any $g = sp^\alpha \in \widehat{\mathcal{P}}$, $\|g\|_\beta = s \le 1$. Hence, $\widehat{\mathcal{P}}$ is a subset of the unit sphere of $L^\beta(\mu)$. By the Banach-Alaoglu theorem [33, Ch. 10, Th. 17] and the fact that $L^\beta(\mu)$ is a reflexive Banach space, it follows that the unit sphere $\{g \colon \|g\|_\beta \le 1\}$ is compact with respect to the weak topology of L^β. Hence, $\widehat{\mathcal{P}}$ is a closed subset of a compact set with respect to the weak topology of $L^\beta(\mu)$, so $\widehat{\mathcal{P}}$ is also compact in the weak topology. Thus, the supremum in (26) is of a continuous linear functional over a compact set in $L^\beta(\mu)$, which yields that this supremum is attained.

To complete the proof for $\alpha \in (0,1)$, we prove the claim that $\widehat{\mathcal{P}}$ is convex and closed. To verify that $\widehat{\mathcal{P}}$ is convex, let $s_1 p_1^\alpha, s_2 p_2^\alpha \in \widehat{\mathcal{P}}$ and $\lambda \in (0,1)$. We can write $\lambda s_1 p_1^\alpha + (1-\lambda) s_2 p_2^\alpha = sp^\alpha$ with

$$p = \frac{1}{Z} \left(\frac{\lambda s_1 p_1^\alpha + (1-\lambda) s_2 p_2^\alpha}{\lambda s_1 + (1-\lambda) s_2} \right)^{1/\alpha}, \tag{28}$$

where Z is the normalizing constant, and $s = \bigl(\lambda s_1 + (1-\lambda) s_2\bigr) Z^\alpha$. For $\alpha \in (0,1)$, $0 < Z \le 1$ by [18, Lemma 3] which implies that $s \in [0,1]$. This proves the convexity of $\widehat{\mathcal{P}}$.

Next, to prove that $\widehat{\mathcal{P}}$ is closed, let $g_n := s_n p_n^\alpha \in \widehat{\mathcal{P}}$ be such that $g_n \to g$ in $L^\beta(\mu)$. We need to show that $g \in \widehat{\mathcal{P}}$. Since $s_n = \|g_n\|_\beta \to \|g\|_\beta$, we have $\|g\|_\beta \le 1$. If $\|g\|_\beta = 0$, then $g = 0$ μ-a.e., and hence obviously $g \in \widehat{\mathcal{P}}$. Since $\beta = \frac{1}{\alpha} > 1$, it follows that if $\|g\|_\beta > 0$, then $p_n^\alpha = g_n / \|g_n\|_\beta \to g / \|g\|_\beta$ in $L^\beta(\mu)$, and therefore $p_n \to (g/\|g\|_\beta)^\beta$ in $L^1(\mu)$.[3] Since $\mathcal{P}$ is closed in $L^1(\mu)$, we have $g/\|g\|_\beta = p^* \in \mathcal{P}$, and $g = \|g\|_\beta \cdot p^* \in \widehat{\mathcal{P}}$.

■

Remark 2: The fact underlying the above proof is that the maximum or minimum of a continuous function over a compact set is always attained. Although the actual set $\mathcal{P}$ in (25), over which we wish to optimize the functional, is not compact, it was possible to modify it into the set $\widehat{\mathcal{P}}$ in (27) without affecting the optimal value in (26); the modified set $\widehat{\mathcal{P}}$ was compact in an appropriate topology where the functional also remains to be continuous.

[3] If $\beta > 1$ and $\{f_n\}$ converges to f in L^β, then an application of the mean-value theorem and Hölder's inequality yields $\bigl\| |f_n|^\beta - |f|^\beta \bigr\| \le \beta (\|f_n\|_\beta + \|f\|_\beta)^{\beta-1} \, \|f_n - f\|_1$; hence, $\{|f_n|^\beta\}$ converges to $|f|^\beta$ in L^1. Since non-negative functions are considered in our case, we can ignore the absolute values so $\{f_n^\beta\}$ converges to f^β in L^1.

IV. The Pythagorean Property and Iterated Projections

In this section we first revisit the Pythagorean property for a finite alphabet and use it to prove a convergence theorem for iterative projections. Throughout this section, we assume that the probability measures are defined on a *finite set* $\mathcal{A}$. For a probability measure P, let its support be given by $\mathrm{Supp}(P) := \{a \in \mathcal{A} \colon P(a) > 0\}$; for a set of probability measures $\mathcal{P}$, let

$$\mathrm{Supp}(\mathcal{P}) := \bigcup_{P \in \mathcal{P}} \mathrm{Supp}(P). \tag{29}$$

Let us first recall the Pythagorean property for a Rényi divergence on an α-convex set. As it is in the cases of relative entropy [16] and relative α-entropy [25], the Pythagorean property is crucial in establishing orthogonality properties. In the sequel, we assume that Q is a probability measure with $\mathrm{Supp}(Q) = \mathcal{A}$.

Proposition 1 (The Pythagorean property): Let $\alpha \in (0,1) \cup (1,\infty)$, let $\mathcal{P} \subseteq \mathcal{M}$ be an α-convex set, and $Q \in \mathcal{M}$.

a) If P^* is a forward D_α-projection of Q on $\mathcal{P}$, then

$$D_\alpha(P\|Q) \geq D_\alpha(P\|P^*) + D_\alpha(P^*\|Q), \quad \forall\, P \in \mathcal{P}. \tag{30}$$

Furthermore, if $\alpha > 1$, then $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathcal{P})$.

b) Conversely, if (30) is satisfied for some $P^* \in \mathcal{P}$, then P^* is a forward D_α-projection of Q on $\mathcal{P}$.

Proof: a) In view of the proof of [18, Theorem 14], for every $P \in \mathcal{P}$ and $t \in [0,1]$, let $P_t \in \mathcal{P}$ be the (α, t)-mixture of (P^*, P); since $D_\alpha(P_t\|Q)$ is minimized at $t = 0$, then (see [18, pp. 3806–3807] for detailed calculations)

$$\begin{aligned} 0 &\leq \frac{\mathrm{d}}{\mathrm{d}t}\, D_\alpha(P_t\|Q)\Big|_{t=0} \\ &= \frac{1}{\alpha - 1} \left(\frac{\sum_a P(a)^\alpha Q(a)^{1-\alpha}}{\sum_a P^*(a)^\alpha Q(a)^{1-\alpha}} - \sum_a P(a)^\alpha P^*(a)^{1-\alpha} \right) \end{aligned} \tag{31}$$

which is equivalent to (30). To show that $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathcal{P})$ for $\alpha > 1$, suppose that there exist $P \in \mathcal{P}$ and $a \in \mathcal{A}$ such that $P^*(a) = 0$ but $P(a) > 0$. Then (31) is violated since its right side is equal, in this case, to $-\infty$ (recall that by assumption $\mathrm{Supp}(Q) = \mathcal{A}$ so, if $\alpha > 1$, $\sum_a P(a)^\alpha Q(a)^{1-\alpha}$, $\sum_a P^*(a)^\alpha Q(a)^{1-\alpha} \in (0, \infty)$, and $\sum_a P(a)^\alpha P^*(a)^{1-\alpha} = +\infty$). This contradiction proves the last assertion in Proposition 1a).

b) From (30), we have

$$\begin{aligned} D_\alpha(P\|Q) &\geq D_\alpha(P\|P^*) + D_\alpha(P^*\|Q) \\ &\geq D_\alpha(P^*\|Q) \quad \forall P \in \mathcal{P}. \end{aligned} \tag{32}$$

■

Remark 3: The Pythagorean property (30) holds for probability measures defined on a general alphabet $\mathcal{A}$, as it is proved in [18, Theorem 14]. The novelty in Proposition 1 is in the last assertion of a), extending the result for the relative entropy in [16, Theorem 3.1], for which $\mathcal{A}$ needs to be a finite set.

Corollary 1: Let $\alpha \in (0, \infty)$. If a forward D_α-projection on an α-convex set exists, then it is unique.

Proof: For $\alpha = 1$, since an α-convex set is particularized to a convex set, this result is known in view of [16, p. 23]. Next, consider the case where $\alpha \in (0,1) \cup (1,\infty)$. Let P_1^* and P_2^* be forward D_α-projections of Q on an α-convex set $\mathcal{P}$. Applying Proposition 1, we have

$$D_\alpha(P_2^* \| Q) \geq D_\alpha(P_2^* \| P_1^*) + D_\alpha(P_1^* \| Q).$$

Since $D_\alpha(P_1^* \| Q) = D_\alpha(P_2^* \| Q)$, we must have $D_\alpha(P_2^* \| P_1^*) = 0$ which yields $P_1^* = P_2^*$. ■

The last assertion in Proposition 1a) shows that $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathcal{P})$ if $\alpha \in (1,\infty)$. The following counterexample illustrates that this equality does not necessarily hold for $\alpha \in (0,1)$.

Example 1: Let $\mathcal{A} = \{1,2,3,4\}$, $\alpha = \frac{1}{2}$, $f \colon \mathcal{A} \to \mathbb{R}$ be given by

$$f(1) = 1,\ f(2) = -3,\ f(3) = -5,\ f(4) = -6 \tag{33}$$

and let $Q(a) = \frac{1}{4}$ for all $a \in \mathcal{A}$. Consider the following α-convex set:[4]

$$\mathcal{P} := \left\{ P \in \mathcal{M} : \sum_{a \in \mathcal{A}} P(a)^\alpha f(a) = 0 \right\}. \tag{34}$$

Let

$$P^*(1) = \tfrac{9}{10},\ P^*(2) = \tfrac{1}{10},\ P^*(3) = 0,\ P^*(4) = 0. \tag{35}$$

It is easy to check that $P^* \in \mathcal{P}$. Furthermore, setting $\theta^* = \frac{1}{5}$ and $Z = \frac{2}{5}$ yields

$$P^*(a)^{1-\alpha} = Z^{\alpha-1} \Big[Q(a)^{1-\alpha} + (1-\alpha)\, f(a)\, \theta^* \Big], \tag{36}$$

for all $a \in \{1,2,3\}$, and

$$P^*(4)^{1-\alpha} > Z^{\alpha-1} \Big[Q(4)^{1-\alpha} + (1-\alpha)\, f(4)\, \theta^* \Big]. \tag{37}$$

From (34), (36) and (37), it follows that for every $P \in \mathcal{P}$

$$\sum_{a \in \mathcal{A}} P(a)^\alpha P^*(a)^{1-\alpha} \geq Z^{\alpha-1} \sum_{a \in \mathcal{A}} P(a)^\alpha Q(a)^{1-\alpha}. \tag{38}$$

Furthermore, it can be also verified that

$$Z^{\alpha-1} \sum_{a \in \mathcal{A}} P^*(a)^\alpha Q(a)^{1-\alpha} = 1. \tag{39}$$

Assembling (38) and (39) yields

$$\sum_{a \in \mathcal{A}} P(a)^\alpha P^*(a)^{1-\alpha} \geq \frac{\sum_{a \in \mathcal{A}} P(a)^\alpha Q(a)^{1-\alpha}}{\sum_{a \in \mathcal{A}} P^*(a)^\alpha Q(a)^{1-\alpha}}, \tag{40}$$

which is equivalent to (30). Hence, Proposition 1b) implies that P^* is the forward D_α-projection of Q on $\mathcal{P}$. Note, however, that $\mathrm{Supp}(P^*) \neq \mathrm{Supp}(\mathcal{P})$; to this end, from (34), it can be verified numerically that

$$P = (0.984688,\ 0.005683,\ 0.004180,\ 0.005449) \in \mathcal{P} \tag{41}$$

which implies that $\mathrm{Supp}(P^*) = \{1,2\}$ whereas $\mathrm{Supp}(\mathcal{P}) = \{1,2,3,4\}$.

[4]This set is characterized in (43) as an α-linear family.

Definition 7 (α-linear family): Let $\alpha \in (0,\infty)$, and $f_1,\ldots,f_k$ be real-valued functions defined on $\mathcal{A}$. The *α-linear family* determined by $f_1,\ldots,f_k$ is defined to be the following parametric family of probability measures defined on $\mathcal{A}$:

$$\mathscr{L}_\alpha := \left\{ P \in \mathcal{M} \colon P(a) = \left[\sum_{i=1}^{k} \theta_i f_i(a) \right]^{\frac{1}{\alpha}}, \quad (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k \right\}. \tag{42}$$

For typographical convenience, we have suppressed the dependence of $\mathscr{L}_\alpha$ in $f_1,\ldots,f_k$. It is easy to see that $\mathscr{L}_\alpha$ is an α-convex set. Without loss of generality, we shall assume that $f_1,\ldots,f_k$, as $|\mathcal{A}|$-dimensional vectors, are mutually orthogonal (otherwise, by the Gram-Schmidt procedure, these vectors can be orthogonalized without affecting the corresponding α-linear family in (42)). Let $\mathcal{F}$ be the subspace of $\mathbb{R}^{|\mathcal{A}|}$ spanned by $f_1,\ldots,f_k$, and let $\mathcal{F}^\perp$ denote the orthogonal complement of $\mathcal{F}$. Hence, there exist $f_{k+1},\ldots,f_{|\mathcal{A}|}$ such that $f_1,\ldots,f_{|\mathcal{A}|}$ are mutually orthogonal as $|\mathcal{A}|$-dimensional vectors, and $\mathcal{F}^\perp = \mathrm{Span}\{f_{k+1},\ldots,f_{|\mathcal{A}|}\}$. Consequently, from (42),

$$\mathscr{L}_\alpha = \left\{ P \in \mathcal{M} \colon \sum_a P(a)^\alpha f_i(a) = 0, \quad \forall\, i \in \{k+1,\ldots,|\mathcal{A}|\} \right\}. \tag{43}$$

From (43), the set $\mathscr{L}_\alpha$ is closed. We shall now focus our attention on forward D_α-projections on α-linear families.

Theorem 2 (Pythagorean equality): Let $\alpha > 1$, and let P^* be the forward D_α-projection of Q on $\mathscr{L}_\alpha$. Then, P^* satisfies (30) with equality, i.e.,

$$D_\alpha(P\|Q) = D_\alpha(P\|P^*) + D_\alpha(P^*\|Q), \quad \forall\, P \in \mathscr{L}_\alpha. \tag{44}$$

Proof: For $t \in [0,1]$ and $P \in \mathscr{L}_\alpha$, let P_t be the (α,t)-mixture of (P,P^*), i.e.,

$$P_t(a) = \frac{1}{Z_t} \left[(1-t)P^*(a)^\alpha + tP(a)^\alpha \right]^{\frac{1}{\alpha}}, \tag{45}$$

where

$$Z_t := \sum_a \left[(1-t)P^*(a)^\alpha + tP(a)^\alpha \right]^{\frac{1}{\alpha}}. \tag{46}$$

Since $P_t \in \mathcal{P}$,

$$D_\alpha(P_t\|Q) \geq D_\alpha(P^*\|Q) = D_\alpha(P_0\|Q), \tag{47}$$

which yields

$$\lim_{t \downarrow 0} \frac{D_\alpha(P_t\|Q) - D_\alpha(P_0\|Q)}{t} \geq 0. \tag{48}$$

By Proposition 1a), if $\alpha \in (1,\infty)$, $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathscr{L}_\alpha)$. Hence, if $\alpha > 1$, for every $P \in \mathscr{L}_\alpha$ there exists $t' < 0$ such that

$$(1-t)P^*(a)^\alpha + tP(a)^\alpha > 0$$

for all $a \in \mathrm{Supp}(\mathscr{L}_\alpha)$ and $t \in (t',0)$. Since $\mathcal{A}$ is finite, the derivative of $D_\alpha(P_t\|Q)$ exists at $t = 0$. In view of (42) and since $P, P^* \in \mathscr{L}_\alpha$, for every $t \in (t',0)$, there exist $\theta_1^{(t)},\ldots,\theta_k^{(t)} \in \mathbb{R}$ such that

$$(1-t)P^*(a)^\alpha + tP(a)^\alpha = \sum_{i=1}^{k} \theta_i^{(t)} f_i(a)$$

which yields that $P_t \in \mathscr{L}_\alpha$ for $t \in (t', 0)$ (see (45)). Consequently, since (47) also holds for all $t \in (t', 0)$, then

$$\lim_{t \uparrow 0} \frac{D_\alpha(P_t \| Q) - D_\alpha(P_0 \| Q)}{t} \leq 0. \tag{49}$$

From (48), (49), and the existence of the derivative of $D_\alpha(P_t \| Q)$ at $t = 0$, it follows that this derivative should be equal to zero; since this derivative is equal to the right side of (31), it follows that (31) holds with equality. Hence, for every $P \in \mathcal{P}$,

$$\frac{\sum_a P(a)^\alpha Q(a)^{1-\alpha}}{\sum_a P^*(a)^\alpha Q(a)^{1-\alpha}} = \sum_a P(a)^\alpha P^*(a)^{1-\alpha}. \tag{50}$$

Taking logarithms on both sides of (50), and dividing by $\alpha - 1$, yields (44). ■

The following theorem suggests an iterative algorithm to find the forward D_α-projection when the underlying α-convex set is an intersection of a finite number of α-linear families.

Theorem 3 (Iterative projections): Let $\alpha \in (1, \infty)$. Suppose that $\mathscr{L}_\alpha^{(1)}, \ldots, \mathscr{L}_\alpha^{(m)}$ are α-linear families, and let

$$\mathcal{P} := \bigcap_{n=1}^{m} \mathscr{L}_\alpha^{(n)} \tag{51}$$

where $\mathcal{P}$ is assumed to be a non-empty set. Let $P_0 = Q$, and let P_n be the forward D_α-projection of P_{n-1} on $\mathscr{L}_\alpha^{(i_n)}$ with $i_n = n \bmod (m)$ for $n \in \mathbb{N}$. Then, $P_n \to P^*$ (a pointwise convergence by letting $n \to \infty$).

Proof: Since (by definition) P_n is a forward D_α-projection of P_{n-1} on an α-linear set which includes $\mathcal{P}$ (see (51)), it follows from Theorem 2 that for every $P \in \mathcal{P}$ and $N \in \mathbb{N}$

$$D_\alpha(P \| P_{n-1}) = D_\alpha(P \| P_n) + D_\alpha(P_n \| P_{n-1}), \quad \forall\, n \in \{1, \ldots, N\}. \tag{52}$$

Hence, since $P_0 = Q$, (52) yields

$$\begin{aligned} D_\alpha(P \| Q) &= D_\alpha(P \| P_N) + \sum_{n=1}^{N} \Big(D_\alpha(P \| P_{n-1}) - D_\alpha(P \| P_n) \Big) \\ &= D_\alpha(P \| P_N) + \sum_{n=1}^{N} D_\alpha(P_n \| P_{n-1}). \end{aligned} \tag{53}$$

Note that $\mathcal{P}$ in (51), being a non-empty intersection of a finite number of compact sets, is a compact set. Let $\{P_{N_k}\}$ be a subsequence of $\{P_n\}$ in $\mathcal{P}$ which pointwise converges to some P' on the finite set $\mathcal{A}$ (hence, $P' \in \mathcal{P}$). Letting $N_k \to \infty$ in (53) implies that, for every $P \in \mathcal{P}$,

$$D_\alpha(P \| Q) = D_\alpha(P \| P') + \sum_{n=1}^{\infty} D_\alpha(P_n \| P_{n-1}) \tag{54}$$

where, to obtain (54), $D_\alpha(P \| P_{N_k}) \to D_\alpha(P \| P')$ since $\mathcal{A}$ is finite and $P_{N_k} \to P'$. Since (54) yields $\sum_{n=1}^{\infty} D_\alpha(P_n \| P_{n-1}) < \infty$ then $D_\alpha(P_n \| P_{n-1}) \to 0$ as $n \to \infty$; consequently, since $D_\alpha(\cdot \| \cdot)$ is monotonically non-decreasing in α (see, e.g., [18, Theorem 3]) and $\alpha > 1$ then $D(P_n \| P_{n-1}) \to 0$, and by Pinsker's inequality $|P_n - P_{n-1}| \to 0$ as $n \to \infty$. From the periodic construction of $\{i_n\}$ (with period m), the subsequences $\{P_{N_k}\}, \{P_{N_k+1}\}, \ldots, \{P_{N_k+m-1}\}$ have their limits in $\mathscr{L}_\alpha^{(1)}, \ldots, \mathscr{L}_\alpha^{(m)}$, respectively. Since $|P_n - P_{n-1}| \to 0$ as $n \to \infty$, all these subsequences have the same limit P', which therefore implies that $P' \in \mathcal{P}$. Substituting $P = P'$ in (54) yields

$$D_\alpha(P'\|Q) = \sum_{n=1}^{\infty} D_\alpha(P_n\|P_{n-1}) \tag{55}$$

and assembling (54) and (55) yields

$$D_\alpha(P\|Q) = D_\alpha(P\|P') + D_\alpha(P'\|Q), \quad \forall\, P \in \mathcal{P}. \tag{56}$$

Hence, (56) implies that P' is the forward D_α-projection of Q on $\mathcal{P}$. Since $\{P_{N_k}\}$ is an arbitrary convergent subsequence of $\{P_n\}$, and the forward D_α-projection is unique, every convergent subsequence of $\{P_n\}$ has the same limit P^*. This proves that $P_n \to P^*$ as $n \to \infty$. ■

V. Forward Projection on an α-Linear Family

We identify in this section a parametric form of the forward D_α-projection on an α-linear family, which turns out to be a generalization of the well-known exponential family.

Theorem 4 (Forward projection on an α-linear family): Let $\alpha \in (0,1) \cup (1,\infty)$, and let P^* be the forward D_α-projection of Q on an α-linear family $\mathscr{L}_\alpha$ (as defined in (42) where $f_1, \ldots, f_k$, as $|\mathcal{A}|$-dimensional vectors, are mutually orthogonal). The following hold:

a) If $\text{Supp}(P^*) = \text{Supp}(\mathscr{L}_\alpha)$, then P^* satisfies (44).

b) If

$$\text{Supp}(P^*) = \text{Supp}(\mathscr{L}_\alpha) = \mathcal{A}, \tag{57}$$

then there exist $f_{k+1}, \ldots, f_{|\mathcal{A}|}$ such that $f_1, \ldots, f_{|\mathcal{A}|}$ are mutually orthogonal as $|\mathcal{A}|$-dimensional vectors, and $\theta^* = (\theta^*_{k+1}, \ldots, \theta^*_{|\mathcal{A}|}) \in \mathbb{R}^{|\mathcal{A}|-k}$ such that for all $a \in \mathcal{A}$

$$P^*(a) = Z(\theta^*)^{-1} \left[Q(a)^{1-\alpha} + (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta^*_i f_i(a) \right]^{\frac{1}{1-\alpha}} \tag{58}$$

where $Z(\theta^*)$ is a normalizing constant in (58).

Proof: The proof of Item a) follows from the proof of Theorem 2 which yields that P^* satisfies the Pythagorean equality (44).

We next prove Item b). Eq. (44) is equivalent to (50), which can be re-written as

$$\sum_a P(a)^\alpha \left[cP^*(a)^{1-\alpha} - Q(a)^{1-\alpha} \right] = 0, \quad \forall\, P \in \mathscr{L}_\alpha \tag{59}$$

with $c = \sum_a P^*(a)^\alpha Q(a)^{1-\alpha}$. Recall that if a subspace of the Euclidean space $\mathbb{R}^{|\mathcal{A}|}$ contains a vector whose all components are strictly positive, then this subspace is spanned by the vectors whose all components are nonnegative. In view of (42), the subspace $\mathcal{F}$ which is spanned by $f_1, \ldots, f_k$ (recall that these functions are regarded as $|\mathcal{A}|$-dimensional vectors) contains $(P^*)^\alpha$ whose support is $\mathcal{A}$ (see (57)). Consequently, it follows from (42) that $\{P^\alpha \colon P \in \mathscr{L}_\alpha\}$ spans the subspace $\mathcal{F}$ of $\mathbb{R}^{|\mathcal{A}|}$. From (59), it also follows that $c\,(P^*)^{1-\alpha} - Q^{1-\alpha} \in \mathcal{F}^\perp$, which yields the existence of $\theta^*_i \in \mathbb{R}$ for $i \in \{k+1, \ldots, |\mathcal{A}|\}$ such that for all $a \in \mathcal{A}$

$$cP^*(a)^{1-\alpha} - Q(a)^{1-\alpha} = (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta^*_i f_i(a) \tag{60}$$

with a scaling of $\{\theta^*_i\}$ by $1-\alpha \neq 0$ in the right side of (60). Hence, P^* satisfies (58) where c in the left side of (60) is the normalizing constant $Z(\theta^*)$ in (58). ■

Remark 4: In view of Example 1, the condition $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathscr{L}_\alpha)$ is not necessarily satisfied for $\alpha \in (0,1)$. However, due to Proposition 1 a), this condition is necessarily satisfied for all $\alpha \in (1,\infty)$.

For $\alpha \in (0,\infty)$, the forward D_α-projection on an α-linear family $\mathscr{L}_\alpha$ motivates the definition of the following parametric family of probability measures. Let $Q \in \mathcal{M}$, and let

$$\mathscr{E}_\alpha := \Biggl\{ P \in \mathcal{M} \colon P(a) = Z(\theta)^{-1} \Biggl[Q(a)^{1-\alpha} + (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i f_i(a) \Biggr]^{\frac{1}{1-\alpha}}, \quad \theta = (\theta_{k+1}, \ldots, \theta_{|\mathcal{A}|}) \in \mathbb{R}^{|\mathcal{A}|-k} \Biggr\}. \tag{61}$$

We shall call the family $\mathscr{E}_\alpha$ an *α-exponential family*,[5] which can be verified to be a $(1-\alpha)$-convex set. We next show that $\mathscr{E}_\alpha$ generalizes the *exponential family* $\mathscr{E}$ defined in [16, p. 24]:

$$\mathscr{E} = \Biggl\{ P \in \mathcal{M} \colon P(a) = Z(\theta)^{-1} Q(a) \exp\Biggl(\sum_{i=k+1}^{|\mathcal{A}|} \theta_i f_i(a) \Biggr), \quad \theta = (\theta_{k+1}, \ldots, \theta_{|\mathcal{A}|}) \in \mathbb{R}^{|\mathcal{A}|-k} \Biggr\}. \tag{62}$$

To this end, let the *α-exponential* and *α-logarithm* functions be, respectively, defined by

$$e_\alpha(x) := \begin{cases} \exp(x) & \text{if } \alpha = 1, \\ \Bigl(\max\bigl\{1 + (1-\alpha)x, \, 0\bigr\} \Bigr)^{\frac{1}{1-\alpha}} & \text{if } \alpha \in (0,1) \cup (1,\infty), \end{cases} \tag{63}$$

$$\ln_\alpha(x) := \begin{cases} \ln(x) & \text{if } \alpha = 1, \\ \frac{x^{1-\alpha}-1}{1-\alpha} & \text{if } \alpha \in (0,1) \cup (1,\infty). \end{cases} \tag{64}$$

In view of (61), (63) and (64), the α-exponential family $\mathscr{E}_\alpha$ includes all the probability measures P defined on $\mathcal{A}$ such that for all $a \in \mathcal{A}$

$$P(a) = Z(\theta)^{-1} \, e_\alpha\Biggl(\ln_\alpha(Q(a)) + \sum_{i=k+1}^{|\mathcal{A}|} \theta_i f_i(a) \Biggr), \tag{65}$$

whereas any $P \in \mathscr{E}$ can be written as

$$P(a) = Z(\theta)^{-1} \exp\Biggl(\ln(Q(a)) + \sum_{i=k+1}^{|\mathcal{A}|} \theta_i f_i(a) \Biggr). \tag{66}$$

This is an alternative way to notice that the family $\mathscr{E}_\alpha$ can be regarded as a continuous extension of the exponential family $\mathscr{E}$ when $\alpha \in (0,1) \cup (1,\infty)$. It is easy to see that the reference measure Q in the definition of $\mathscr{E}_\alpha$ is always a member of $\mathscr{E}_\alpha$. As in the case of the exponential family, the α-exponential family $\mathscr{E}_\alpha$ also depends on the reference measure Q only in a loose manner. In view of (61), any other member of $\mathscr{E}_\alpha$ can play the role of Q in defining this family. The proof is very similar to the one for the α-power-law family in [25, Proposition 22]. It should be also noted that, for $\alpha \in (1,\infty)$, all members of $\mathscr{E}_\alpha$ have the same support (i.e., the support of Q).

[5]Note that the *α-power-law family* in [25, Definition 8] is a different extension of the exponential family $\mathscr{E}$.

VI. Orthogonality of α-Linear and α-Exponential Families

In this section, we first prove an "orthogonality" relationship between an α-exponential family and its associated α-linear family. We then use it to transform the reverse D_α-projection on an α-exponential family into a forward D_α-projection on an α-linear family.

Let us begin by making precise the notion of orthogonality between two sets of probability measures with respect to D_α ($\alpha > 0$).

Definition 8 (Orthogonality of sets of probability measures): Let $\alpha \in (0,1) \cup (1,\infty)$, and let $\mathcal{P}$ and $\mathcal{Q}$ be sets of probability measures. We say that $\mathcal{P}$ *is α-orthogonal to $\mathcal{Q}$ at P^** if the following hold:

1) $\mathcal{P} \cap \mathcal{Q} = \{P^*\}$
2) $D_\alpha(P\|Q) = D_\alpha(P\|P^*) + D_\alpha(P^*\|Q)$ for every $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Note that, when $\alpha = 1$, this refers to the orthogonality between the linear and exponential families in [16, Corollary 3.1].

We are now ready to state our second main result namely, the orthogonality between $\mathscr{L}_\alpha$ and $\mathscr{E}_\alpha$.

Theorem 5 (Orthogonality of $\mathscr{L}_\alpha$ and $\mathscr{E}_\alpha$): Let $\alpha \in (1,\infty)$, let $\mathscr{L}_\alpha$ and $\mathscr{E}_\alpha$ be given in (42) and (61), respectively, and let P^* be the forward D_α-projection of Q on $\mathscr{L}_\alpha$. The following hold:

a) $\mathscr{L}_\alpha$ is α-orthogonal to $\mathrm{cl}(\mathscr{E}_\alpha)$ at P^*.
b) If $\mathrm{Supp}(\mathscr{L}_\alpha) = \mathcal{A}$, then $\mathscr{L}_\alpha$ is α-orthogonal to $\mathscr{E}_\alpha$ at P^*.

Proof: In view of Proposition 1 a), for $\alpha \in (1,\infty)$, the condition $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathscr{L}_\alpha)$ is satisfied. Consequently, for $\alpha \in (1,\infty)$, Theorem 4a) implies that P^* satisfies (44). We next prove the following:

i) Every $\tilde{P} \in \mathscr{L}_\alpha \cap \mathrm{cl}(\mathscr{E}_\alpha)$ satisfies (44) with $\tilde{P}$ in place of P^*.
ii) $\mathscr{L}_\alpha \cap \mathrm{cl}(\mathscr{E}_\alpha)$ is non-empty.

To prove Item i), since $\tilde{P} \in \mathrm{cl}(\mathscr{E}_\alpha)$, there exists a sequence $\{P_n\}$ in $\mathscr{E}_\alpha$ such that $P_n \to \tilde{P}$. Since $P_n \in \mathscr{E}_\alpha$, from (61), there exists $\theta^{(n)} := (\theta^{(n)}_{k+1}, \ldots, \theta^{(n)}_{|\mathcal{A}|}) \in \mathbb{R}^{|\mathcal{A}|-k}$ such that for all $a \in \mathcal{A}$

$$P_n(a)^{1-\alpha} = Z(\theta^{(n)})^{\alpha-1}\Big[Q(a)^{1-\alpha} + (1-\alpha)\sum_{i=k+1}^{|\mathcal{A}|}\theta^{(n)}_i f_i(a)\Big]. \tag{67}$$

Since $P, \tilde{P} \in \mathscr{L}_\alpha$, from (43), for all $i \in \{k+1, \ldots, |\mathcal{A}|\}$

$$\sum_a P(a)^\alpha f_i(a) = 0, \tag{68}$$

$$\sum_a \tilde{P}(a)^\alpha f_i(a) = 0. \tag{69}$$

Since $\mathcal{A}$ is finite, assembling (67)–(69) yields (after switching the order of summations over $a \in \mathcal{A}$ and $i \in \{k+1, \ldots, |\mathcal{A}|\}$)

$$\sum_a P(a)^\alpha P_n(a)^{1-\alpha} = Z(\theta^{(n)})^{\alpha-1}\sum_a P(a)^\alpha Q(a)^{1-\alpha}, \tag{70}$$

$$\sum_a \tilde{P}(a)^\alpha P_n(a)^{1-\alpha} = Z(\theta^{(n)})^{\alpha-1}\sum_a \tilde{P}(a)^\alpha Q(a)^{1-\alpha}, \tag{71}$$

and, from (70) and (71),

$$\sum_a \tilde{P}(a)^\alpha P_n(a)^{1-\alpha} = \frac{\sum_a P(a)^\alpha P_n(a)^{1-\alpha}}{\sum_a P(a)^\alpha Q(a)^{1-\alpha}} \cdot \sum_a \tilde{P}(a)^\alpha Q(a)^{1-\alpha}. \tag{72}$$

Since $P_n \to \tilde{P}$, letting $n \to \infty$ in (72) yields

$$1 = \frac{\sum_a P(a)^\alpha \tilde{P}(a)^{1-\alpha}}{\sum_a P(a)^\alpha Q(a)^{1-\alpha}} \cdot \sum_a \tilde{P}(a)^\alpha Q(a)^{1-\alpha}, \tag{73}$$

which is equivalent to (44) when P^* is replaced by $\tilde{P}$.

To prove Item ii), note that if $\text{Supp}(\mathscr{L}_\alpha) = \mathcal{A}$, then Theorem 4b) yields that $P^* \in \mathscr{L}_\alpha \cap \mathscr{E}_\alpha$, and we are done. So suppose that $\text{Supp}(\mathscr{L}_\alpha) \neq \mathcal{A}$, and consider the following sequence of α-linear families:

$$\mathscr{L}_\alpha^{(n)} := \Big\{ P \in \mathcal{M} \colon \sum_a P(a)^\alpha \tilde{f}_i(a) = 0, \quad i \in \{k+1, \ldots, |\mathcal{A}|\} \Big\}, \tag{74}$$

where

$$\tilde{f}_i(a) := f_i(a) - \eta_i^{(n)} Q(a)^{1-\alpha}, \quad \forall\, a \in \mathcal{A} \tag{75}$$

with

$$\eta_i^{(n)} := \frac{\frac{1}{n} \sum_a Q(a)^\alpha f_i(a)}{(1-\frac{1}{n}) \sum_a P^*(a)^\alpha Q(a)^{1-\alpha} + \frac{1}{n}}, \quad i \in \{k+1, \ldots, |\mathcal{A}|\}. \tag{76}$$

The $\tilde{f}_i$'s and $\eta_i^{(n)}$'s in (75) and (76) are selected such that the $\big(\alpha, \frac{1}{n}\big)$-mixture of (P^*, Q) is a member of $\mathscr{L}_\alpha^{(n)}$. This implies that $\text{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$ (recall that we assume that $\text{Supp}(Q) = \mathcal{A}$). Notice also that $\eta_i^{(n)} \to 0$ as $n \to \infty$. Hence, $\mathscr{L}_\alpha^{(n)}$ asymptotically coincides with $\mathscr{L}_\alpha$ as $n \to \infty$. Now, let P_n be the forward D_α-projection of Q on $\mathscr{L}_\alpha^{(n)}$. Then by Proposition 1, $\text{Supp}(P_n) = \mathcal{A}$, and hence by Theorem 4, there exists $\theta^{(n)} := (\theta_{k+1}^{(n)}, \ldots, \theta_{|\mathcal{A}|}^{(n)}) \in \mathbb{R}^{|\mathcal{A}|-k}$ such that for all $a \in \mathcal{A}$

$$\begin{aligned}
P_n(a)^{1-\alpha} &= Z(\theta^{(n)})^{\alpha-1} \Big[Q(a)^{1-\alpha} + (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} \tilde{f}_i(a) \Big] && (77)\\
&= Z(\theta^{(n)})^{\alpha-1} \Big[Q(a)^{1-\alpha} + (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} \big(f_i(a) - \eta_i^{(n)} Q(a)^{1-\alpha} \big) \Big] && (78)\\
&= Z(\theta^{(n)})^{\alpha-1} \Big[\Big(1 - (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} \eta_i^{(n)} \Big) Q(a)^{1-\alpha} \\
&\qquad + (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} f_i(a) \Big]. && (79)
\end{aligned}$$

Multiplying the left side of (77) and the right side of (79) by $P^*(a)^\alpha$, summing over all $a \in \mathcal{A}$, and using the fact that $\sum_a P^*(a)^\alpha f_i(a) = 0$ for all $i \in \{k+1, \ldots, |\mathcal{A}|\}$ yields

$$\sum_a P^*(a)^\alpha P_n(a)^{1-\alpha} = Z(\theta^{(n)})^{\alpha-1} \Big(1 - (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} \eta_i^{(n)} \Big) \sum_a P^*(a)^\alpha Q(a)^{1-\alpha}. \tag{80}$$

This implies that the term $1 - (1-\alpha) \sum_{i=k+1}^{|\mathcal{A}|} \theta_i^{(n)} \eta_i^{(n)}$ is positive for all n; hence, dividing the left side of (77) and the right side of (79) by this positive term yields that $P_n \in \mathscr{E}_\alpha$. This implies that the limit of every convergent subsequence of $\{P_n\}$ is a member of $\text{cl}(\mathscr{E}_\alpha)$, as well as of $\mathscr{L}_\alpha$.

In view of Items i) and ii), as listed at the beginning of this proof, it now follows from Proposition 1 b) and Corollary 1 that $\mathscr{L}_\alpha \cap \mathrm{cl}(\mathscr{E}_\alpha) = \{P^*\}$. Recall that, for $\alpha \in (1, \infty)$, Theorem 4a) implies that P^* satisfies (44); furthermore, since Q in (61) can be replaced by any other member of $\mathscr{E}_\alpha$, the satisfiability of (44) for $Q \in \mathscr{E}_\alpha$ yields its satisfiability with any other member of $\mathscr{E}_\alpha$ replacing Q. Since $\mathcal{A}$ is finite, (44) is also satisfied with any member of $\mathrm{cl}(\mathscr{E}_\alpha)$ replacing Q; this can be justified for any $\widetilde{Q} \in \mathrm{cl}(\mathscr{E}_\alpha)$ by selecting a sequence $\{\widetilde{Q}_n\}$ in $\mathscr{E}_\alpha$ which pointwise converges to $\widetilde{Q}$, and by letting $n \to \infty$. This proves Item a).

We next prove Item b). Since by our assumption $\mathrm{Supp}(\mathscr{L}_\alpha) = \mathcal{A}$ and $\alpha \in (1, \infty)$ then Proposition 1 a) implies that condition (57) holds. From Proposition 1 b), Corollary 1 and Theorem 4, it follows that the forward D_α-projection P^* is the unique member of $\mathscr{L}_\alpha \cap \mathscr{E}_\alpha$ satisfying (44). Similarly to the previous paragraph, (44) is satisfied not only for for $Q \in \mathscr{E}_\alpha$, but also for any other member of $\mathscr{E}_\alpha$ replacing Q. This proves Item b). ■

Remark 5: In view of Example 1, if $\alpha \in (0, 1)$, $\mathrm{Supp}(P^*)$ is not necessarily equal to $\mathrm{Supp}(\mathscr{L}_\alpha)$; this is consistent with Theorem 5 which is stated only for $\alpha \in (1, \infty)$. Nevertheless, in view of the proof of Theorem 2, the following holds for $\alpha \in (0, 1)$: if the condition $\mathrm{Supp}(P^*) = \mathrm{Supp}(\mathscr{L}_\alpha) = \mathcal{A}$ is satisfied, then $\mathscr{L}_\alpha$ is α-orthogonal to $\mathscr{E}_\alpha$ at P^*.

VII. Reverse Projection on an α-Exponential Family

In this section, we define reverse D_α-projections, and we rely on the orthogonality property in Theorem 5 (and the note in Remark 5) to convert the reverse D_α-projection on an α-exponential family into a forward projection on an α-linear family.

Definition 9 (Reverse D_α-projection): Let $P \in \mathcal{M}$, $\mathcal{Q} \subseteq \mathcal{M}$, and $\alpha \in (0, \infty)$. If there exists $Q^* \in \mathcal{Q}$ which attains the global minimum of $D_\alpha(P\|Q)$ over all $Q \in \mathcal{Q}$ and $D_\alpha(P\|Q^*) < \infty$, then Q^* is said to be a *reverse D_α-projection* of P on $\mathcal{Q}$.

Theorem 6: Let $\alpha \in (0, 1) \cup (1, \infty)$, and let $\mathscr{E}_\alpha$ be an α-exponential family determined by $Q, f_{k+1}, \ldots, f_{|\mathcal{A}|}$. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn at random according to a probability measure in $\mathscr{E}_\alpha$. Let $\hat{P}_n$ be the empirical probability measure of $X_1, \ldots, X_n$, and let P_n^* be the forward D_α-projection of Q on the α-linear family

$$\mathscr{L}_\alpha^{(n)} := \Big\{ P \in \mathcal{M} \colon \sum_{a \in \mathcal{A}} P(a)^\alpha \hat{f}_i(a) = 0, \quad i \in \{k+1, \ldots, |\mathcal{A}|\} \Big\}, \tag{81}$$

where

$$\hat{f}_i(a) := f_i(a) - \hat{\eta}_i^{(n)} Q(a)^{1-\alpha}, \quad \forall\, a \in \mathcal{A} \tag{82}$$

with

$$\hat{\eta}_i^{(n)} := \frac{\sum_a \hat{P}_n(a)^\alpha f_i(a)}{\sum_a \hat{P}_n(a)^\alpha Q(a)^{1-\alpha}}, \quad i \in \{k+1, \ldots, |\mathcal{A}|\}. \tag{83}$$

The following hold:

a) If $\mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$ for $\alpha \in (1, \infty)$ or $\mathrm{Supp}(P_n^*) = \mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$ for $\alpha \in (0, 1)$, then P_n^* is the reverse D_α-projection of $\hat{P}_n$ on $\mathscr{E}_\alpha$.

b) For $\alpha \in (1, \infty)$, if $\mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) \neq \mathcal{A}$, then the reverse D_α-projection of $\hat{P}_n$ on $\mathscr{E}_\alpha$ does not exist. Nevertheless, P_n^* is the reverse D_α-projection of $\hat{P}_n$ on $\mathrm{cl}(\mathscr{E}_\alpha)$.

Proof: To prove Item a), note that $\mathscr{L}_\alpha^{(n)}$ in (81)–(83) is constructed in such a way that

$$\hat{P}_n \in \mathscr{L}_\alpha^{(n)}. \tag{84}$$

Following (61), let $\mathscr{E}_\alpha = \mathscr{E}_\alpha(f_{k+1}, \ldots, f_{|\mathcal{A}|}; Q)$ denote the α-exponential family determined by $f_{k+1}, \ldots, f_{|\mathcal{A}|}$ and Q. We claim that

$$\mathscr{E}_\alpha(f_{k+1}, \ldots, f_{|\mathcal{A}|}; Q) = \mathscr{E}_\alpha(\hat{f}_{k+1}, \ldots, \hat{f}_{|\mathcal{A}|}; Q). \tag{85}$$

Indeed, if $P \in \mathscr{E}_\alpha(f_{k+1}, \ldots, f_{|\mathcal{A}|}; Q)$, then there exist $\theta = (\theta_{k+1}, \ldots, \theta_{|\mathcal{A}|}) \in \mathbb{R}^{|\mathcal{A}|-k}$ and a normalizing constant $Z = Z(\theta)$ such that for all $a \in \mathcal{A}$

$$P(a)^{1-\alpha} = Z^{\alpha-1}\Big[Q(a)^{1-\alpha} + (1-\alpha)\sum_{i=k+1}^{|\mathcal{A}|} \theta_i f_i(a)\Big] \tag{86}$$

$$= Z^{\alpha-1}\Big[\Big(1 + (1-\alpha)\sum_{i=k+1}^{|\mathcal{A}|} \theta_i \hat{\eta}_i^{(n)}\Big) Q(a)^{1-\alpha} + (1-\alpha)\sum_{i=k+1}^{|\mathcal{A}|} \theta_i \hat{f}_i(a)\Big] \tag{87}$$

where (86) and (87) follow, respectively, from (61) and (82). Multiplying the left side of (86) and the right side of (87) by $\hat{P}_n(a)^\alpha$, summing over all $a \in \mathcal{A}$, and using (84) yields

$$\sum_a \hat{P}_n(a)^\alpha P(a)^{1-\alpha} = Z^{\alpha-1}\left(1 + (1-\alpha)\sum_{i=k+1}^{|\mathcal{A}|} \theta_i \hat{\eta}_i^{(n)}\right) \sum_a \hat{P}_n(a)^\alpha Q(a)^{1-\alpha}. \tag{88}$$

Eq. (88) yields $1 + (1-\alpha)\sum\limits_{i=k+1}^{|\mathcal{A}|} \theta_i \hat{\eta}_i^{(n)} > 0$. Consequently, by rescaling (87) appropriately, it follows that $P \in \mathscr{E}_\alpha(\hat{f}_{k+1}, \ldots, \hat{f}_{|\mathcal{A}|}; Q)$ which therefore implies that

$$\mathscr{E}_\alpha(f_{k+1}, \ldots, f_{|\mathcal{A}|}; Q) \subseteq \mathscr{E}_\alpha(\hat{f}_{k+1}, \ldots, \hat{f}_{|\mathcal{A}|}; Q). \tag{89}$$

Similarly, one can show that the reverse relation of (89) also holds, which yields (85). The proof of a) is completed by considering the following two cases:

- If $\alpha \in (1, \infty)$ and $\mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$, in view of Theorem 5b), $\mathscr{L}_\alpha^{(n)}$ is α-orthogonal to $\mathscr{E}_\alpha = \mathscr{E}_\alpha(\hat{f}_{k+1}, \ldots, \hat{f}_{|\mathcal{A}|}; Q)$ at P_n^*; hence, due to (84),

$$D_\alpha(\hat{P}_n \| Q) = D_\alpha(\hat{P}_n \| P_n^*) + D_\alpha(P_n^* \| Q), \quad \forall\, Q \in \mathscr{E}_\alpha. \tag{90}$$

  Since $P_n^* \in \mathscr{E}_\alpha$, the minimum of $D_\alpha(\hat{P}_n \| Q)$ subject to $Q \in \mathscr{E}_\alpha$ is uniquely attained at $Q = P_n^*$.
- If $\alpha \in (0, 1)$ and $\mathrm{Supp}(P_n^*) = \mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$, then (90) holds in view of Remark 5 and (84). The minimum of $D_\alpha(\hat{P}_n \| Q)$ subject to $Q \in \mathscr{E}_\alpha$ is thus uniquely attained at $Q = P_n^*$.

To prove Item b), for $\alpha \in (1, \infty)$, note that $P_n^* \in \mathscr{E}_\alpha$ if and only if $\mathrm{Supp}(\mathscr{L}_\alpha^{(n)}) = \mathcal{A}$. Indeed, the 'if' part follows from Item a). The 'only if' part follows from the fact that all members of $\mathscr{E}_\alpha$ have the same support, Q is a member of $\mathscr{E}_\alpha$ which by assumption has full support, and P_n^* is in both $\mathscr{E}_\alpha$ (by assumption) and $\mathscr{L}_\alpha^{(n)}$ (by definition).

To prove the first assertion in Item b), note that by Theorem 5a), $P_n^* \in \mathrm{cl}(\mathscr{E}_\alpha)$ and (90) holds for every $Q \in \mathrm{cl}(\mathscr{E}_\alpha)$. Hence,

$$\min_{Q \in \mathrm{cl}(\mathscr{E}_\alpha)} D_\alpha(\hat{P}_n \| Q) = D_\alpha(\hat{P}_n \| P_n^*). \tag{91}$$

Due to the continuity of $D_\alpha(\hat{P}_n \| Q)$ for Q which is defined on the finite set $\mathcal{A}$, it follows from (91) that

$$\inf_{Q \in \mathscr{E}_\alpha} D_\alpha(\hat{P}_n \| Q) = D_\alpha(\hat{P}_n \| P_n^*). \tag{92}$$

In view of (90), the minimum of $D_\alpha(\hat{P}_n \| Q)$ over $Q \in \mathscr{E}_\alpha$ is not attained. Finally, the last assertion in b) is due to (90) which, in view of Theorem 5a), holds for all $Q \in \mathrm{cl}(\mathscr{E}_\alpha)$. ■

VIII. Summary and Concluding Remarks

In [18, Theorem 14], van Erven and Harremoës proved a Pythagorean inequality for Rényi divergences on α-convex sets under the assumption that the forward projection exists. Motivated by their result, we study forward and reverse projections for the Rényi divergence of order α on α-convex sets. The results obtained in this paper, for $\alpha \in (0, \infty)$, generalize the known results for $\alpha = 1$; this special case corresponds to projections of the relative entropy on convex sets, as studied by Csiszár *et al.* in [12], [13], [15], [16]. The main contributions of this paper are as follows:

1) we prove a sufficient condition for the existence of a forward projection in the general alphabet setting.
2) we prove a projection theorem on an α-linear family in the finite alphabet setting, and the parametric form of this projection gives rise to an α-exponential family.
3) we prove an orthogonality property between α-linear and α-exponential families; it yields a duality between forward and reverse projections, respectively, on these families.
4) we prove a convergence result of an iterative algorithm for calculating the forward projection on an intersection of a finite number of α-linear families.

For $\alpha = 0$, the notion of an α-convex set is continuously extended to a $\log$-convex set. Since $D_0(P\|Q) = -\log Q\big(\text{Supp}(P)\big)$ (see, e.g., [18, Theorem 4]), if there exists $P \in \mathcal{P}$ such that $\text{Supp}(P) = \text{Supp}(Q)$ then any such probability measure is a forward D_0-projection of Q on $\mathcal{P}$ for which $D_0(P\|Q) = 0$. Note that, in this case, a forward D_0-projection of Q on $\mathcal{P}$ is not necessarily unique.

For $\alpha = 0$ and a finite set $\mathcal{A}$, the notion of an α-linear family is the whole simplex of probability measures (with the convention that $0^0 = 1$ in (43)), provided that $\sum_a f_i(a) = 0$ for all $i \in \{k+1, \ldots, |\mathcal{A}|\}$; otherwise, the 0-linear family is an empty set. In the former case, the forward D_0-projection of Q on $\mathcal{P}$ is any probability measure P with a full support since in this case $D_0(P\|Q) = 0$; the forward D_0-projection is, however, meaningless in the latter case where $\mathcal{P}$ is an empty set.

The Rényi divergence of order ∞ is well defined (see, e.g., [18, Theorem 6]); furthermore, a set is defined to be ∞-convex if for all $P_0, P_1 \in \mathcal{P}$, the probability measure $S_{0,1}$ whose μ-density $s_{0,1}$ is equal to the normalized version of $\max\{p_0, p_1\}$, is also included in $\mathcal{P}$ (this definition follows from (15) by letting $\alpha \to \infty$). In this case, Theorems 1 and 2 continue to hold for $\alpha = \infty$ (recall that Theorem 2 refers to the setting where $\mathcal{A}$ is finite).

Consider the case where $\alpha = \infty$ and $\mathcal{A}$ is a finite set. By continuous extension, the ∞-linear family necessarily includes all the probability measures that are not a point mass (see (43)), and the ∞-exponential family only includes the reference measure Q (see (61)). Consequently, the results in Theorems 3–6 become trivial for $\alpha = \infty$.

Acknowledgment

This work has been supported by the Israeli Science Foundation (ISF) under Grant 12/12. We thank the Associate Editor, the anonymous reviewers of this journal paper and the ISIT '16 conference paper for their valuable feedback which helped to improve the presentation.

References

[1] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, series B, vol. 28, no. 1, pp. 131–142, 1966.

[2] A. Barron, "Limits of information, Markov chains, and projections," *Proceedings of the 2000 IEEE International Symposium on Information Theory*, p. 25, Sorrento, Italy, June 2000.

[3] A. Basu, H. Shioya and C. Park, "Statistical Inference: The Minimum Distance Approach," *Chapman & Hall/ CRC Monographs on Statistics and Applied Probability*, vol. 120, CRC Press, Boca Raton, Florida, USA, June 2011.

[4] R. Beran, "Minimum Hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5, no. 3, pp. 445–463, 1977.

[5] N. N. C̄encov, *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs, vol. 53, American Mathematical Society, 1982. Original Publication: Moscow, U.S.S.R.: Nauka, 1972.

[6] N. N. Chentsov, "Nonsymmetrical distance between probability distributions, entropy and the Phytagorean theorem" (in Russian), *Mat. Zametki*, vol. 4, no. 3, pp. 323–332, 1968.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, second edition, 2006.

[8] N. Cressie and T. R. C. Read, "Multinomial goodness of fit test," *Journal of the Royal Statistical Society*, vol. 463, no. 3, pp. 440–464, 1984.

[9] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, January 1963.

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

[11] I. Csiszár, "On topological properties of f-divergences," *Studia Scientiarum Mathematicarum Hungarica*, no. 2, pp. 329–339, 1967.

[12] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, pp. 146–158, 1975.

[13] I. Csiszár, "Sanov property, Generalized I-projection and a conditional limit theorem," *Annals of Probability*, vol. 12, no. 3, pp. 768–793, 1984.

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

[15] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. on Information Theory*, vol. 49, no. 6, pp. 1474–1490, June 2003.

[16] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[17] T. van Erven and P. Harremoës, "Rényi divergence and majorization," *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pp. 1335–1339, Austin, Texas, USA, June 2010.

[18] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[19] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," June 2012. [Online]. Available at `http://arxiv.org/abs/1206.2459v1`.

[20] R. G. Gallager, *Information Theory and Reliable Communications*, John Wiley, 1968.

[21] P. Harremoës and T. van Erven, personal communication, July 2016.

[22] J. H. B. Kemperman, "On the optimal rate of transmitting information," *Annals Mathematical Statistics*, vol. 40, pp. 2156–2177, December 1969.

[23] S. Kullback, "A lower bound for discrimination information in terms of variation," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 126–127, January 1967.

[24] M. A. Kumar and R. Sundaresan, "Minimization problems based on α-relative entropy I: Forward Projection," *IEEE Trans. on Information Theory*, vol. 61, no. 9, pp. 5063–5080, September 2015.

[25] M. A. Kumar and R. Sundaresan, "Minimization problems based on α-relative entropy II: Reverse Projection," *IEEE Trans. on Information Theory*, vol. 61, no. 9, pp. 5081–5095, September 2015.

[26] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner-Texte Zur Mathematik, vol. 95, Germany, 1987.

[27] T. Minka, "Divergence measures and message passing," *Microsoft Research Technical Report*, December 2005. [Online]. Available at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2005-173.pdf.

[28] T. Morimoto, "Markov processes and the H-theorem," *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, 1963.

[29] L. Pardo, "Statistical Inference based on Divergences," *CRC/Chapman–Hall*, 2006.

[30] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[31] M. S. Pinsker, *Information and Information Stability of Random Variables and Random Processes*, San-Fransisco: Holden-Day, 1964, originally published in Russian in 1960.

[32] A. Rényi, "On measures of entropy and information," *Proceedings of the 4th Berekely Symposium on Probability Theory and Mathematical Statistics*, pp. 547–561, Berekeley, California, USA, 1961.

[33] H. L. Royden, *Real Analysis*, third edition, Prentice-Hall, New Jersey, 1988.

[34] I. Sason, "On the Rényi divergence, joint range of relative entropies, and a channel coding theorem," *IEEE Trans. on Information Theory*, vol. 62, no. 1, pp. 23–34, January 2016.

[35] I. Sason and S. Verdú, "f-divergence inequalities," *preprint*, August 2015. [Online]. Available at http://arxiv.org/abs/1508.00335.

[36] C. Shannon, R. Gallager and E. Berlekamp, "Lower bounds to error probability for decoding on discrete memoryless channels - part I," *Information and Control*, vol. 10, pp. 65–103, February 1967.

[37] O. Shayevitz, "On Rényi measures and hypothesis testing," *Proceedings of the 2011 IEEE International Symposium on Information Theory*, pp. 800–804, Saint Petersburg, Russia, August 2011.

[38] R. Sundaresan, "A measure of discrimination and its geometric properties," *Proccedings of the IEEE 2002 International Symposium on Information Theory*, p. 264, Laussane, Switzerland, June 2002.

[39] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. on Information Theory*, vol. 53, no. 1, pp. 269–287, January 2007.

[40] F. Topsøe, "Information-theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. 7–17, 1979.

[41] C. Tsallis, "Possible generalization of Bolzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1-2, 1988.

[42] C. Tsallis, R. S. Mendes and A. R. Plastino, "The role of constraints within generalized nonextensive statistics," *Physica A*, vol. 261, no. 3–4, pp. 534–554, December 1998.