Adaptive Alternating Minimization Algorithms

Urs Niesen, *Student Member, IEEE*, Devavrat Shah, and Gregory W. Wornell

***Abstract*—The classical alternating minimization (or projection) algorithm has been successful in the context of solving optimization problems over two variables. The iterative nature and simplicity of the algorithm has led to its application in many areas such as signal processing, information theory, control, and finance. A general set of sufficient conditions for the convergence and correctness of the algorithm are known when the underlying problem parameters are fixed. In many practical situations, however, the underlying problem parameters are changing over time, and the use of an adaptive algorithm is more appropriate. In this paper, we study such an adaptive version of the alternating minimization algorithm. More precisely, we consider the impact of having a slowly time-varying domain over which the minimization takes place. As a main result of this paper, we provide a general set of sufficient conditions for the convergence and correctness of the adaptive algorithm. Perhaps somewhat surprisingly, these conditions seem to be the minimal ones one would expect in such an adaptive setting. We present applications of our results to adaptive decomposition of mixtures, adaptive log-optimal portfolio selection, and adaptive filter design.**

***Index Terms*—Adaptive filters, adaptive signal processing, algorithms, Arimoto–Blahut algorithm, optimization methods.**

I. Introduction

A. Background

SOLVING an optimization problem over two variables in a product space is central to many applications in areas such as signal processing, information theory, statistics, control, and finance. The alternating minimization or projection algorithm has been extensively used in such applications due to its iterative nature and simplicity.

The alternating minimization algorithm attempts to solve a minimization problem of the following form: given $\mathcal{P}, \mathcal{Q}$ and a function $D : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$, minimize D over $\mathcal{P} \times \mathcal{Q}$. That is, find

$$\min_{(P,Q)\in\mathcal{P}\times\mathcal{Q}} D(P,Q).$$

Often minimizing over both variables simultaneously is not straightforward. However, minimizing with respect to one variable while keeping the other one fixed is often easy and sometimes possible analytically. In such a situation, the alternating minimization algorithm described next is well suited: start with an arbitrary initial point $Q_0 \in \mathcal{Q}$; for $n \geq 1$, iteratively compute

$$\begin{aligned} P_n &\in \arg\min_{P\in\mathcal{P}} D(P, Q_{n-1}) \\ Q_n &\in \arg\min_{Q\in\mathcal{Q}} D(P_n, Q). \end{aligned} \tag{1}$$

In other words, instead of solving the original minimization problem over two variables, the alternating minimization algorithm solves a sequence of minimization problems over only one variable. If the algorithm converges, the converged value is returned as the solution to the original problem. Conditions for the convergence and correctness of such an algorithm, that is, conditions under which

$$\lim_{n\to\infty} D(P_n, Q_n) = \min_{(P,Q)\in\mathcal{P}\times\mathcal{Q}} D(P,Q) \tag{2}$$

have been of interest since the early 1950s. A general set of conditions, stated in the paper by Csiszár and Tusnády [1, Theorem 2], is summarized in the next theorem.[1]

Theorem 1: Let $\mathcal{P}$ and $\mathcal{Q}$ be any two sets, and let $D : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ such that for all $\widetilde{P} \in \mathcal{P}, \widetilde{Q} \in \mathcal{Q}$

$$\arg\min_{P\in\mathcal{P}} D(P, \widetilde{Q}) \neq \emptyset$$

$$\arg\min_{Q\in\mathcal{Q}} D(\widetilde{P}, Q) \neq \emptyset.$$

Then the alternating minimization algorithm converges, i.e., (2) holds, if there exists a nonnegative function $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ such that the following two properties hold.

(a) Three-point property $(P, \widetilde{P}, \widetilde{Q})$: For all $P \in \mathcal{P}, \widetilde{Q} \in \mathcal{Q}$, $\widetilde{P} \in \arg\min_{P\in\mathcal{P}} D(P, \widetilde{Q})$

$$\delta(P, \widetilde{P}) + D(\widetilde{P}, \widetilde{Q}) \leq D(P, \widetilde{Q}).$$

(b) Four-point property $(P, Q, \widetilde{P}, \widetilde{Q})$: For all $P, \widetilde{P} \in \mathcal{P}$, $Q \in \mathcal{Q}$, $\widetilde{Q} \in \arg\min_{Q\in\mathcal{Q}} D(\widetilde{P}, Q)$

$$D(P, \widetilde{Q}) \leq D(P, Q) + \delta(P, \widetilde{P}).$$

B. Our Contribution

In this paper, we consider an adaptive version of the above minimization problem. As before, suppose we wish to find

$$\min_{(P,Q)\in\mathcal{P}\times\mathcal{Q}} D(P,Q)$$

by means of an alternating minimization algorithm. However, on the nth iteration of the algorithm, we are provided with sets $\mathcal{P}_n, \mathcal{Q}_n$ which are *time-varying* versions of the sets $\mathcal{P}$ and $\mathcal{Q}$, respectively. That is, we are given a sequence of optimization problems

$$\left\{ \min_{(P,Q)\in\mathcal{P}_n\times\mathcal{Q}_n} D(P,Q) \right\}_{n\geq 0}. \tag{3}$$

Manuscript received December 20, 2006; revised September 04, 2008. Current version published February 25, 2009. This work was supported in part by the National Science Foundation under Grant CCF-0515109, and by Hewlett-Packard through the MIT/HP Alliance. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Nice, France, June 2007.

The authors are with the Department of Electrical Engineering and Computer Science, the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: uniesen@mit.edu; devavrat@mit.edu; gww@mit.edu).

Communicated by G. Kramer, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2008.2011442

[1]The conditions in [1] are actually slightly more general than the ones shown here and allow for functions D that take the value $+\infty$, i.e., $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$.

Such situations arise naturally in many applications. For example, in adaptive signal processing problems, the changing parameters could be caused by a slowly time-varying system, with the index n representing time. An obvious approach is to solve each of the problems in (3) independently (one at each time instance n). However, since the system varies only slowly with time, such an approach is likely to result in a lot of redundant computation. Indeed, it is likely that a solution to the problem at time instance $n-1$ will be very close to the one at time instance n. A different approach is to use an *adaptive* algorithm instead. Such an adaptive algorithm should be computationally efficient: given the tentative solution at time $n-1$, the tentative solution at time n should be easy to compute. Moreover, if the time-varying system eventually reaches steady state, the algorithm should converge to the optimal steady-state solution. In other words, instead of insisting that the adaptive algorithm solves (3) for every n, we only impose that it does so as $n \to \infty$.

Given these requirement, a natural candidate for such an algorithm is the following adaptation of the alternating minimization algorithm: start with an arbitrary initial $Q_0 \in \mathcal{Q}_0$; for $n \geq 1$ compute (cf. (1))

$$P_n \in \arg\min_{P \in \mathcal{P}_n} D(P, Q_{n-1})$$
$$Q_n \in \arg\min_{Q \in \mathcal{Q}_n} D(P_n, Q).$$

Suppose that the sequences of sets $\{\mathcal{P}_n\}_{n\geq 0}$ and $\{\mathcal{Q}_n\}_{n\geq 0}$ converge (in a sense to be made precise later) to sets $\mathcal{P}$ and $\mathcal{Q}$, respectively. We are interested in conditions under which

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

As a main result of this paper, we provide a general set of sufficient conditions under which this adaptive algorithm converges. These conditions are essentially the same as those of [1] summarized in Theorem 1. The precise results are stated in Theorem 4.

C. Organization

The remainder of this paper is organized as follows. In Section II, we introduce notation, and some preliminary results. Section III provides a convergence result for a fairly general class of adaptive alternating minimization algorithms. We specialize this result to adaptive minimization of divergences in Section IV, and to adaptive minimization procedures in Hilbert spaces (with respect to inner product induced norm) in Section V. This work was motivated by several applications in which the need for an adaptive alternating minimization algorithm arises. We present an application in the divergence minimization setting from statistics and finance in Section IV, and an application in the Hilbert space setting from adaptive signal processing in Section V. Section VI contains concluding remarks.

II. Notations and Technical Preliminaries

In this section, we setup notations and present technical preliminaries needed in the remainder of the paper. Let $(\mathcal{M}, d)$ be a compact metric space. Given two sets $\mathcal{A}, \mathcal{B} \subset \mathcal{M}$, define the *Hausdorff distance* between them as

$$d_H(\mathcal{A}, \mathcal{B}) \triangleq \max\left\{\sup_{A\in\mathcal{A}} \inf_{B\in\mathcal{B}} d(A,B), \sup_{B\in\mathcal{B}} \inf_{A\in\mathcal{A}} d(A,B)\right\}.$$

It can be shown the d_H is a metric, and in particular satisfies the triangle inequality.

Consider a continuous function $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. For compact sets $\mathcal{A}, \mathcal{B} \subset \mathcal{M}$, define the set

$$\mathcal{G}(\mathcal{A}, \mathcal{B}) \triangleq \arg\min_{(A,B)\in\mathcal{A}\times\mathcal{B}} D(A, B).$$

With slight abuse of notation, let

$$D(\mathcal{A}, \mathcal{B}) \triangleq \min_{(A,B)\in\mathcal{A}\times\mathcal{B}} D(A, B).$$

Due to compactness of the sets $\mathcal{A}, \mathcal{B}$ and continuity of D, we have $\mathcal{G}(\mathcal{A}, \mathcal{B}) \neq \emptyset$, and hence $D(\mathcal{A}, \mathcal{B})$ is well defined.

A. Some Lemmas

Here we state a few auxiliary lemmas used in the following.

Lemma 2: Let $\{a_n\}_{n\geq 0}, \{b_n\}_{n\geq 0}$ be sequences of real numbers, satisfying

$$a_n + b_n \leq b_{n-1} + c$$

for all $n \geq 1$ and some $c \in \mathbb{R}$. If $\limsup_{n\to\infty} b_n > -\infty$ then

$$\liminf_{n\to\infty} a_n \leq c.$$

If, in addition[2]

$$\sum_{n=0}^{\infty} (c - a_n)^+ < \infty$$

then

$$\lim_{n\to\infty} a_n = c.$$

Lemma 3: Let $\{\mathcal{A}_n\}_{n\geq 0}$ be a sequence of subsets of $\mathcal{M}$. Let $\mathcal{A}$ be a closed subset of $\mathcal{M}$ such that $\mathcal{A}_n \overset{d_H}{\to} \mathcal{A}$. Consider any sequence $\{A_n\}_{n\geq 0}$ such that $A_n \in \mathcal{A}_n$ for all $n \geq 0$, and such that $A_n \overset{d}{\to} A \in \mathcal{M}$. Then $A \in \mathcal{A}$.

Proof: Since $A_n \in \mathcal{A}_n$ and $\mathcal{A}_n \overset{d_H}{\to} \mathcal{A}$, the definition of Hausdorff distance implies that there exists a sequence $\{\widehat{A}_n\}_{n\geq 0}$ such that $\widehat{A}_n \in \mathcal{A}$ for all n and $d(\widehat{A}_n, A_n) \to 0$ as $n \to \infty$. Therefore

$$d(\widehat{A}_n, A) \leq d(\widehat{A}_n, A_n) + d(A_n, A) \to 0$$

as $n \to \infty$. Since the sequence $\{\widehat{A}_n\}_{n\geq 0}$ is entirely in $\mathcal{A}$, this implies that A is a limit point of $\mathcal{A}$. As $\mathcal{A}$ is closed, we therefore have $A \in \mathcal{A}$. □

Let $(\mathcal{X}, d)$ be a metric space and $f : \mathcal{X} \to \mathbb{R}$. Define the *modulus of continuity* $\omega_f : \mathbb{R}_+ \to \mathbb{R}_+$ of f as

$$\omega_f(t) \triangleq \sup_{\substack{x,x'\in\mathcal{X}:\\ d(x,x')\leq t}} |f(x) - f(x')|.$$

[2]We use $(x)^+ \triangleq \max\{0, x\}$.

Remark 1: Note that if f is uniformly continuous then $w_f(t) \to 0$ as $t \to 0$. In particular, if $(\mathcal{X}, d)$ is compact and f is continuous then f is uniformly continuous, and hence $\lim_{t\to 0} w_f(t) = 0$.

III. Adaptive Alternating Minimization Algorithms

Here we present the precise problem formulation. We then present an adaptive algorithm and sufficient conditions for its convergence and correctness.

A. Problem Statement

Consider a compact metric space $(\mathcal{M}, d)$, compact sets $\mathcal{P}, \mathcal{Q} \subset \mathcal{M}$, and a continuous function $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. We want to find $D(\mathcal{P}, \mathcal{Q})$. However, we are not given the sets $\mathcal{P}, \mathcal{Q}$ directly. Instead, we are given a sequence of compact sets $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0} : \mathcal{P}_n, \mathcal{Q}_n \subset \mathcal{M}$ that are revealed at time n such that as $n \to \infty, \mathcal{P}_n \overset{d_H}{\to} \mathcal{P}$ and $\mathcal{Q}_n \overset{d_H}{\to} \mathcal{Q}$. Given an arbitrary initial $(P_0, Q_0) \in \mathcal{P}_0 \times \mathcal{Q}_0$, the goal is to find a sequence of points $(P_n, Q_n) \in \mathcal{P}_n \times \mathcal{Q}_n$ such that

$$\lim_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q}).$$

B. Algorithm

The problem formulation described in the last section suggests the following adaptive version of the alternating minimization algorithm. Initially, we have $(P_0, Q_0) \in \mathcal{P}_0 \times \mathcal{Q}_0$. Recursively for $n \geq 1$, pick any

$$P_n \in \arg\min_{P\in\mathcal{P}_n} D(P, Q_{n-1})$$
$$Q_n \in \arg\min_{Q\in\mathcal{Q}_n} D(P_n, Q).$$

We call this the Adaptive Alternating Minimization (AAM) algorithm in the sequel. Note that if $\mathcal{P}_n = \mathcal{P}$ and $\mathcal{Q}_n = \mathcal{Q}$ for all n, then the above algorithm specializes to the classical alternating minimization algorithm.

C. Sufficient Conditions for Convergence

In this section, we present a set of sufficient conditions under which the AAM algorithm converges to $D(\mathcal{P}, \mathcal{Q})$. As we shall see, we need "three-point" and "four-point" properties (generalizing those in [1]) also in the adaptive setup. To this end, assume there exists a function[3] $\delta : \mathcal{M}\times\mathcal{M} \to \mathbb{R}$ such that the following conditions are satisfied.

(C1) *Three-point property* $(P, \widetilde{P}, Q)$: for all $n \geq 1, P \in \mathcal{P}_n, Q \in \mathcal{Q}_{n-1}, \widetilde{P} \in \arg\min_{P\in\mathcal{P}_n} D(P, Q)$

$$\delta(P, \widetilde{P}) + D(\widetilde{P}, Q) \leq D(P, Q).$$

(C2) *Four-point property* $(P, Q, \widetilde{P}, \widetilde{Q})$: for all $n \geq 1, P, \widetilde{P} \in \mathcal{P}_n, Q \in \mathcal{Q}_n, \widetilde{Q} \in \arg\min_{Q\in\mathcal{Q}_n} D(\widetilde{P}, Q)$,

$$D(P, \widetilde{Q}) \leq D(P, Q) + \delta(P, \widetilde{P}).$$

Our main result is as follows.

[3]Note that unlike the condition in [1], we do not require δ to be nonnegative here.

Theorem 4: Let $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0}, \mathcal{P}, \mathcal{Q}$ be compact subsets of the compact metric space $(\mathcal{M}, d)$ such that

$$\mathcal{P}_n \overset{d_H}{\to} \mathcal{P}, \quad \mathcal{Q}_n \overset{d_H}{\to} \mathcal{Q}$$

and let $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be a continuous function. Let conditions C1 and C2 hold. Then, under the AAM algorithm

$$\liminf_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$$

and all limit points of subsequences of $\{(P_n, Q_n)\}_{n\geq 0}$ achieving this $\liminf$ belong to $\mathcal{G}(\mathcal{P}, \mathcal{Q})$. If, in addition

$$\sum_{n=0}^{\infty} \omega(2\varepsilon_n) < \infty$$

where $\varepsilon_n \triangleq d_H(\mathcal{P}_n, \mathcal{P}) + d_H(\mathcal{Q}_n, \mathcal{Q})$, and $\omega \triangleq \omega_D$ is the modulus of continuity of D, then

$$\lim_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$$

and all limit points of $\{(P_n, Q_n)\}_{n\geq 0}$ belong to $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

Remark 2: Compared to the conditions of [1, Theorem 2] summarized in Theorem 1, the main additional requirement here is in essence uniform continuity of the function D (which is implied by compactness of $\mathcal{M}$ and continuity of D), and summability of the $\omega(2\varepsilon_n)$. This is the least one would expect in this adaptive setup to obtain a conclusion as in Theorem 4.

D. Proof of Theorem 4

We start with some preliminaries. Given that $(\mathcal{M}, d)$ is compact, the product space $(\mathcal{M} \times \mathcal{M}, d_2)$ with

$$d_2((A, B), (A', B')) \triangleq d(A, A') + d(B, B')$$

for all $(A, B), (A', B') \in \mathcal{M} \times \mathcal{M}$, is compact. Let $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ be the modulus of continuity of D with respect to the metric space $(\mathcal{M} \times \mathcal{M}, d_2)$. By definition of ω, for any $\varepsilon > 0$ and $(A, B), (A', B') \in \mathcal{M} \times \mathcal{M}$ such that

$$d_2((A, B), (A', B')) \leq \varepsilon$$

we have

$$|D(A, B) - D(A', B')| \leq \omega(\varepsilon).$$

Moreover, continuity of D and compactness of $\mathcal{M} \times \mathcal{M}$ imply (see Remark 1) that $\omega(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Recall the definition of

$$\varepsilon_n \triangleq d_H(\mathcal{P}_n, \mathcal{P}) + d_H(\mathcal{Q}_n, \mathcal{Q}).$$

By the hypothesis of Theorem 4, we have $\varepsilon_n \to 0$ as $n \to \infty$, and

$$d_H(\mathcal{P}_n, \mathcal{P}_{n-1}) + d_H(\mathcal{Q}_n, \mathcal{Q}_{n-1}) \leq \varepsilon_{n-1} + \varepsilon_n \triangleq \gamma_n$$

with $\gamma_n \to 0$ as $n \to \infty$.

We now proceed to the Proof of Theorem 4. Condition C1 implies that for all $n \geq 1, P \in \mathcal{P}_n, Q \in \mathcal{Q}_n$

$$\delta(P, P_n) + D(P_n, Q_{n-1}) \leq D(P, Q_{n-1}). \tag{4}$$

Condition C2 implies that for all $n \geq 1, P \in \mathcal{P}_n, Q \in \mathcal{Q}_n$

$$D(P, Q_n) \leq D(P, Q) + \delta(P, P_n). \tag{5}$$

Adding (4) and (5), we obtain that for all $n \geq 1, P \in \mathcal{P}_n$, $Q \in \mathcal{Q}_n$

$$D(P_n, Q_{n-1}) + D(P, Q_n) \leq D(P, Q_{n-1}) + D(P, Q). \tag{6}$$

Given that $d_H(\mathcal{Q}_{n-1}, \mathcal{Q}_n) \leq \gamma_n$, there exists $\widehat{Q}_n \in \mathcal{Q}_n$ such $d(Q_{n-1}, \widehat{Q}_n) \leq \gamma_n$. It follows that

$$d_2((P_n, \widehat{Q}_n), (P_n, Q_{n-1})) \leq \gamma_n$$

and hence

$$|D(P_n, \widehat{Q}_n) - D(P_n, Q_{n-1})| \leq \omega(\gamma_n). \tag{7}$$

From (7) and the AAM algorithm, we have

$$\begin{aligned} D(P_n, Q_n) &= \min_{Q \in \mathcal{Q}_n} D(P_n, Q) \\ &\leq D(P_n, \widehat{Q}_n) \quad (\text{since } \widehat{Q}_n \in \mathcal{Q}_n) \\ &\leq D(P_n, Q_{n-1}) + \omega(\gamma_n). \end{aligned} \tag{8}$$

Adding inequalities (6) and (8)

$$D(P_n, Q_n) + D(P, Q_n) \leq D(P, Q_{n-1}) + D(P, Q) + \omega(\gamma_n) \tag{9}$$

for all $P \in \mathcal{P}_n, Q \in \mathcal{Q}_n$.

Since $\mathcal{P}_n \overset{d_H}{\to} \mathcal{P}$ and $\mathcal{Q}_n \overset{d_H}{\to} \mathcal{Q}$, there exists a sequence $(P_n^*, Q_n^*) \in \mathcal{P}_n \times \mathcal{Q}_n$ such that $(P_n^*, Q_n^*) \to (P^*, Q^*) \in \mathcal{G}(\mathcal{P}, \mathcal{Q})$ and $d_2((P_n^*, Q_n^*), (P^*, Q^*)) \leq \varepsilon_n$ for all $n \geq 0$. Pick any such sequence $\{(P_n^*, Q_n^*)\}_{n \geq 0}$. Replacing (P, Q) in (9) by this (P_n^*, Q_n^*), we obtain

$$\begin{aligned} &D(P_n, Q_n) + D(P_n^*, Q_n) \\ &\quad \leq D(P_n^*, Q_{n-1}) + D(P_n^*, Q_n^*) + \omega(\gamma_n). \end{aligned} \tag{10}$$

By choice of the (P_n^*, Q_n^*)

$$D(P_n^*, Q_n^*) \leq D(P^*, Q^*) + \omega(\varepsilon_n). \tag{11}$$

Moreover

$$\begin{aligned} d(P_{n-1}^*, P_n^*) &\leq d(P_{n-1}^*, P^*) + d(P^*, P_n^*) \\ &\leq \varepsilon_{n-1} + \varepsilon_n \\ &= \gamma_n \end{aligned}$$

and therefore

$$D(P_n^*, Q_{n-1}) \leq D(P_{n-1}^*, Q_{n-1}) + \omega(\gamma_n). \tag{12}$$

Combining inequalities (11) and (12) with (13), we obtain

$$\begin{aligned} &D(P_n, Q_n) + D(P_n^*, Q_n) \\ &\leq D(P_{n-1}^*, Q_{n-1}) + D(P^*, Q^*) + 2\omega(\gamma_n) + \omega(\varepsilon_n). \end{aligned} \tag{13}$$

Define

$$\begin{aligned} a_n &\triangleq D(P_n, Q_n) - 2\omega(\gamma_n) - \omega(\varepsilon_n) \\ b_n &\triangleq D(P_n^*, Q_n) \\ c &\triangleq D(P^*, Q^*) \end{aligned}$$

and note that by (13)

$$a_n + b_n \leq b_{n-1} + c.$$

Since D is a continuous function over the compact set $\mathcal{M} \times \mathcal{M}$, it is also a bounded function. Hence, we have $\limsup_{n \to \infty} |b_n| < \infty$. Applying Lemma 2

$$\liminf_{n \to \infty} D(P_n, Q_n) \leq D(P^*, Q^*) + \limsup_{n \to \infty} (2\omega(\gamma_n) + \omega(\varepsilon_n)). \tag{14}$$

Since $\gamma_n \to 0$ and $\varepsilon_n \to 0$ imply $2\omega(\gamma_n) + \omega(\varepsilon_n) \to 0$, (14) yields

$$\liminf_{n \to \infty} D(P_n, Q_n) \leq D(\mathcal{P}, \mathcal{Q}). \tag{15}$$

Now, let $\{n_k\}_{k \geq 0}$ be a subsequence such that

$$\liminf_{n \to \infty} D(P_n, Q_n) = \lim_{k \to \infty} D(P_{n_k}, Q_{n_k}).$$

By compactness of $\mathcal{M} \times \mathcal{M}$, we can assume without loss of generality that $P_{n_k} \overset{d}{\to} P, Q_{n_k} \overset{d}{\to} Q$ for some $P, Q \in \mathcal{M}$. Since $\mathcal{P}$ and $\mathcal{Q}$ are compact, Lemma 3 shows that $P \in \mathcal{P}, Q \in \mathcal{Q}$. By continuity of D this implies that

$$\begin{aligned} \liminf_{n \to \infty} D(P_n, Q_n) &= \lim_{k \to \infty} D(P_{n_k}, Q_{n_k}) \\ &= D(P, Q) \\ &\geq D(\mathcal{P}, \mathcal{Q}). \end{aligned}$$

Together with (15), this shows that

$$\liminf_{n \to \infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$$

and that all limit points of subsequences of $\{(P_n, Q_n)\}_{n \geq 0}$ achieving this $\liminf$ belong to $\mathcal{G}(\mathcal{P}, \mathcal{Q})$. This completes the proof the first part of Theorem 4.

Suppose now that we have in addition

$$\sum_{n=0}^{\infty} \omega(2\varepsilon_n) < \infty. \tag{16}$$

Since

$$\begin{aligned} D(P_n, Q_n) &\geq \min_{P \in \mathcal{P}_n, Q \in \mathcal{Q}_n} D(P, Q) \\ &\geq \min_{P \in \mathcal{P}, Q \in \mathcal{Q}} D(P, Q) - \omega(\varepsilon_n) \\ &= D(P^*, Q^*) - \omega(\varepsilon_n) \end{aligned}$$

we have

$$\begin{aligned} (c - a_n)^+ &= (D(P^*, Q^*) - D(P_n, Q_n) + 2\omega(\gamma_n) + \omega(\varepsilon_n))^+ \\ &\leq 2(\omega(\gamma_n) + \omega(\varepsilon_n)) \\ &\leq 2(\omega(2\varepsilon_n) + \omega(2\varepsilon_{n-1}) + \omega(\varepsilon_n)) \\ &\leq 2(2\omega(2\varepsilon_n) + \omega(2\varepsilon_{n-1})). \end{aligned}$$

Thus, by (16)

$$\sum_{n=0}^{\infty} (c - a_n)^+ < \infty$$

and applying again Lemma 2 yields

$$\lim_{n \to \infty} D(P_n, Q_n) = D(P^*, Q^*). \tag{17}$$

As every limit point of $\{(P_n, Q_n)\}_{n\geq 0}$ belongs to $\mathcal{P}\times\mathcal{Q}$ by Lemma 3, (17) and continuity of D imply that if (16) holds, then every limit point of $\{(P_n, Q_n)\}_{n\geq 0}$ must also belong to $\mathcal{G}(\mathcal{P},\mathcal{Q})$. This concludes the Proof of Theorem 4.

IV. Divergence Minimization

In this section, we specialize the algorithm from Section III to the case of alternating divergence minimization. A large class of problems can be formulated as a minimization of divergences. For example, computation of channel capacity and rate distortion function [2], [3], selection of log-optimal portfolios [4], and maximum-likelihood estimation from incomplete data [5]. These problems were shown to be divergence minimization problems in [1]. For further applications of alternating divergence minimization algorithms, see [6]. We describe applications to the problem of adaptive mixture decomposition and of adaptive log-optimal portfolio selection.

A. Setting

Given a finite set Σ and some constant $0 < b < B$, let $\mathcal{M} = \mathcal{M}(\Sigma, b, B)$ be the set of all measures P on Σ such that

$$\sum_{\sigma\in\Sigma} P(\sigma) \leq B, \text{and} P(\sigma) \geq b, \quad \forall\sigma\in\Sigma. \tag{18}$$

Endow $\mathcal{M}$ with the topology induced by the metric $d : \mathcal{M}\times\mathcal{M}\to\mathbb{R}_+$ defined as

$$d(P,Q) \triangleq \max_{\sigma\in\Sigma} |P(\sigma) - Q(\sigma)|.$$

It is easy to check that the metric space $(\mathcal{M}, d)$ is compact. The cost function D of interest is divergence[4]

$$D(P,Q) \triangleq D(P\|Q) \triangleq \sum_{\sigma\in\Sigma} P(\sigma)\log\frac{P(\sigma)}{Q(\sigma)}$$

for any $P, Q\in\mathcal{M}$. Note that (18) ensures that D is well defined (i.e., does not take the value ∞). It is well known (and easy to check) that the function D is continuous and convex in both arguments. Finally, define the function $\delta : \mathcal{M}\times\mathcal{M}\to\mathbb{R}$

$$\delta(P,\widetilde{P}) \triangleq D(P\|\widetilde{P}) - \sum_{\sigma\in\Sigma}(P(\sigma) - \widetilde{P}(\sigma)).$$

In [1], it has been established that for convex $\mathcal{P}$ and $\mathcal{Q}$, the pair of functions D, δ satisfy the "three-point" and "four-point" properties C1 and C2. As stated above, the space $\mathcal{M} = \mathcal{M}(\Sigma, b, B)$ with metric d is a compact metric space, and the function D is continuous. Hence, Theorem 4 applies in this setting.

B. Application: Decomposition of Mixtures and Log-Optimal Portfolio Selection

We consider an application of our adaptive divergence minimization algorithm to the problem of decomposing a mixture. A special case of this setting yields the problem of log-optimal portfolio selection.

[4]All logarithms are with respect to base e.

We are given a sequence of independent and identically distributed (i.i.d.) random variables $\{Y_l\}_{l\geq 0}$, each taking values in the finite set $\mathcal{Y}$. Y_l is distributed according to the mixture $\sum_{i=1}^I c_i\mu_i$, where the $\{c_i\}_{i=1}^I$ sum to one, $c_i \geq c_0 > 0$ for all $i\in\{1,\ldots,I\}$, and where $\{\mu_i\}_{i=1}^I$ are distributions on $\mathcal{Y}$. We assume that $\mu_i(y)\geq\mu_0 > 0$ for all $y\in\mathcal{Y}, i\in\{1,\ldots,I\}$. The goal is to compute an estimate of $\{c_i\}_{i=1}^I$ from $\{Y_l\}_{l=1}^n$ and knowing $\{\mu_i\}_{i=1}^I$.

Let $\overline{P}_n : \mathcal{Y}\to[0,1]$

$$\overline{P}_n(y) \triangleq \frac{1}{n}\sum_{\ell=1}^n \mathbf{1}_{\{Y_\ell = y\}}$$

be the empirical distribution of $\{Y_l\}_{l=1}^n$. The maximum-likelihood estimator of $\{c_i\}_{i=1}^I$ is given by (see, e.g., [7, Lemma 3.1])

$$\arg\min_{\{\tilde{c}_i\}} D\left(\overline{P}_n \middle\| \sum_{i=1}^I \tilde{c}_i\mu_i\right). \tag{19}$$

Following [7, Example 5.1], we define

$$\begin{aligned}
\Sigma &\triangleq \{1,\ldots,I\}\times\mathcal{Y}\\
\mathcal{Q}_n = \mathcal{Q} &\triangleq \Big\{Q : Q(i,y) = \tilde{c}_i\mu_i(y),\\
&\qquad \text{for some}\{\tilde{c}_i\} \text{ with } \textstyle\sum_i \tilde{c}_i = 1, \tilde{c}_i \geq c_0 \forall i\Big\}\\
\mathcal{P}_n &\triangleq \left\{P : \sum_{i=1}^I P(i,y) = \overline{P}_n(y), P(i,y)\geq 0\forall i,y\right\}.
\end{aligned} \tag{20}$$

Note that $\mathcal{P}_n$ and $\mathcal{Q}$ are convex and compact. From [7, Lemma 5.1], we have

$$\min_{\{\tilde{c}_i\}} D\left(\overline{P}_n \middle\| \sum_{i=1}^I \tilde{c}_i\mu_i\right) = \min_{P\in\mathcal{P}_n}\min_{Q\in\mathcal{Q}} D(P\|Q)$$

and the minimizer of the left-hand side (and hence (19)) is recovered from the corresponding marginal of the optimal Q on the right-hand side.

We now show how the projections on the sets $\mathcal{P}_n$ and $\mathcal{Q}$ can be computed. Fix a P, assuming without loss of generality that

$$\sum_{y\in\mathcal{Y}} P(1,y) \geq \sum_{y\in\mathcal{Y}} P(2,y) \geq \cdots \geq \sum_{y\in\mathcal{Y}} P(I,y).$$

We want to minimize $D(P\|Q)$ over all $Q\in\mathcal{Q}$, or, equivalently, over all valid $\{\tilde{c}_i\}$. The $\{\tilde{c}_i\}$ minimizing $D(P\|Q)$ can be shown to be of the form $\tilde{c}_i > c_0$ for all $i\leq J^*$ and $\tilde{c}_i = c_0$ for all $i > J^*$. More precisely, define

$$\eta(J) \triangleq \frac{1}{1-(I-J)c_0}\sum_{i=1}^J\sum_{y\in\mathcal{Y}} P(i,y)$$

and choose $J^*\in\{1,\ldots,I\}$ such that

$$\begin{aligned}
\frac{1}{\eta(J^*)}\sum_{y\in\mathcal{Y}} P(i,y) &> c_0, \quad \text{for } 1\leq i\leq J^*\\
\frac{1}{\eta(J^*)}\sum_{y\in\mathcal{Y}} P(i,y) &\leq c_0, \quad \text{for } J^* < i\leq I.
\end{aligned}$$

Then the optimal $\{\tilde{c}_i\}$ are given by

$$\tilde{c}_i = \frac{1}{\eta(J^*)} \sum_{y \in \mathcal{Y}} P(i,y), \quad \text{for } 1 \le i \le J^*$$
$$\tilde{c}_i = c_0, \quad \text{for } J^* < i \le I.$$

For fixed $Q(i,y) = \tilde{c}_i \mu_i(y)$, the minimizing P is

$$P(i,y) = \frac{\tilde{c}_i \mu_i(y)}{\sum_j \tilde{c}_j \mu_j(y)} \overline{P}_n(y). \tag{21}$$

We now check that (18) is satisfied for some values of b and B. As $\mathcal{P}_n$ and $\mathcal{Q}$ are sets of distributions, we can choose $B = 1$. For all $Q \in \mathcal{Q}, i \in \{1,\dots,I\}, y \in \mathcal{Y}$, we have $Q(i,y) \ge \mu_0 c_0 > 0$. However, for $P \in \mathcal{P}_n$, we have in general only $P(i,y) \ge 0$. In order to apply the results from Section IV-A, we need to show that we can, without loss of optimality, restrict the sets $\mathcal{P}_n$ to contain only distributions P that are bounded below by some $p_0 > 0$. In other words, we need to show that the projections on $\mathcal{P}_n$ are bounded below by p_0.

Assume for the moment that the empirical distribution $\overline{P}_n$ is close to the true one in the sense that

$$\left| \overline{P}_n(y) - \sum_i c_i \mu_i(y) \right| \le \frac{\mu_0}{2}$$

for all $y \in \mathcal{Y}$. As $\sum_i c_i \mu_i(y) \ge \mu_0$, this implies $\overline{P}_n(y) \ge \frac{\mu_0}{2}$ for all y. From (18), this implies that the projection P in $\mathcal{P}_n$ of any point in $\mathcal{Q}$ satisfies $P(i,y) \ge \frac{1}{2} c_0 \mu_0^2 \triangleq p_0$ for all $i \in \{1,\dots,I\}, y \in \mathcal{Y}$. Hence, in this case $\mathcal{M}(\Sigma, b, B)$ satisfies (21) with $b = \frac{1}{2} c_0 \mu_0^2$ and $B = 1$.

It remains to argue that $\overline{P}_n$ is close to $\sum_i c_i \mu_i(y)$. Suppose that instead of constructing the set $\mathcal{P}_n$ (see (20)) with respect to $\overline{P}_n$, we construct it with respect to the distribution $\overline{P}_n$ defined as

$$\overline{P}_n(y) \triangleq \frac{\mu_0}{2} + \lambda \left(\overline{P}_n(y) - \frac{\mu_0}{2} \right)^+$$

where λ is chosen such that $\sum_y \overline{P}_n(y) = 1$. $\overline{P}_n$ is bounded below by $\frac{\mu_0}{2}$ by construction. Moreover, by the strong law of large numbers

$$\mathbb{P}(\overline{P}_n \ne \overline{P}_n \text{i.o.}) = 0.$$

Hence, we have $\mathcal{P}_n \overset{d_H}{\to} \mathcal{P}$ almost surely, where $\mathcal{P}$ is constructed as in (20) with respect to the true distribution $\sum_i c_i \mu_i$.

Applying now the results from Section IV-A and Theorem 4 yields that under the AAM algorithm

$$\liminf_{n \to \infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$$

almost surely, and that every limit point of $\{(P_n, Q_n)\}_{n \ge 0}$ achieving this $\liminf$ is an element of $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

Since by the law of the iterated logarithm, convergence of $\overline{P}_n$ to P is only $\Theta(\sqrt{\log\log n}/\sqrt{n})$ as $n \to \infty$ almost surely, and since $\lim_{\varepsilon \to 0} \omega(\varepsilon)/\varepsilon = 0$ only if D is a constant [8], we can in this scenario *not* conclude from Theorem 4 that $\lim_{n \to \infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$.

As noted in [7], a special case of the decomposition of mixture problem is that of maximizing the expected value of $\log \sum_i c_i W_i$, where $\{W_i\}_{i=1}^I$ is distributed according to $\overline{P}_n$. The standard alternating divergence minimization algorithm is then the same as Cover's portfolio optimization algorithm [4]. Thus, the AAM algorithm applied as before yields also an adaptive version of this portfolio optimization algorithm.

V. Projections in Hilbert Space

In this section, we specialize the algorithm from Section III to the case of minimization in a Hilbert space. A large class of problems can be formulated as alternating projections in Hilbert spaces. For example, problems in filter design, signal recovery, and spectral estimation. For an extensive overview, see [9]. In the context of Hilbert spaces, the alternating minimization algorithm is often called POCS (projection onto convex sets).

A. Setting

Let $\mathcal{M}$ be a compact subset of a Hilbert space with the usual norm $d(A,B)^2 \triangleq \langle A - B, A - B \rangle$. Then $(\mathcal{M}, d)$ is a compact metric space. The cost function D of interest is

$$D(A,B) \triangleq d(A,B)^2.$$

The function D is continuous and convex. Define the function δ (as part of conditions C1 and C2), as

$$\delta(A, \widetilde{A}) \triangleq d(A, \widetilde{A})^2.$$

In [1], it is established that for convex $\mathcal{P}$ and $\mathcal{Q}$ the pair of functions D, δ satisfies the "three-point" and "four-point" properties C1 and C2. Hence, Theorem 4 applies in this setting.

B. Application: Set-Theoretic Signal Processing and Adaptive Filter Design

In this subsection, we consider a problem in the Hilbert space setting as defined in Section V-A. Let $\{\mathcal{S}_i\}_{i=1}^I$ be a collection of convex compact subsets of the Hilbert space $\mathbb{R}^k$ with the usual inner product, and let $\{c_i\}_{i=1}^I$ be positive weights summing to one. In set-theoretic signal processing, the objective is to find a point A minimizing

$$\sum_{i=1}^I c_i d(A, \mathcal{S}_i) \tag{22}$$

where $d(A, \mathcal{S}_i) \triangleq \min_{S \in \mathcal{S}_i} d(A,S)$. Many problems in signal processing can be formulated in this way. Applications can be found for example in control, filter design, and estimation. For an overview and extensive list of references, see [9]. As an example, in a filter design problem, the $\mathcal{S}_i$ could be constraints on the impulse and frequency responses of a filter [10], [11].

Following [12], this problem can be formulated in our framework by defining the Hilbert space $\mathcal{H} = \mathbb{R}^{Ik}$ with inner product

$$\langle A, B \rangle \triangleq \sum_{i=1}^I c_i \langle A_i, B_i \rangle,$$

where $A_i, B_i \in \mathbb{R}^k$ for $i \in \{1,\dots,I\}$ are the components of A and B. Let

$$\mathcal{S} \triangleq \operatorname{conv} \left\{ \cup_{i=1}^I \mathcal{S}_i \right\} \subset \mathbb{R}^k$$

be the convex hull of the union of the constraint sets $\{\mathcal{S}_i\}_{i=1}^I$, and let

$$\mathcal{M} \triangleq \mathcal{S}^I \subset \mathcal{H}$$

be its I-fold product. Since each of the sets $\mathcal{S}_i$ is compact, $\mathcal{M}$ is compact and by definition also convex. We define the set $\mathcal{P} \subset \mathcal{M}$ as

$$\mathcal{P} \triangleq \{(\widetilde{P}, \ldots, \widetilde{P}) \in \mathcal{H} : \widetilde{P} \in \mathcal{S}\}$$

and the set $\mathcal{Q} \subset \mathcal{M}$ as

$$\mathcal{Q} \triangleq \mathcal{S}_1 \times \cdots \times \mathcal{S}_I. \tag{23}$$

We now show how the projections on the sets $\mathcal{P}$ and $\mathcal{Q}$ can be computed. For a fixed $P = (\widetilde{P}, \ldots, \widetilde{P}) \in \mathcal{P}$, the $Q \in \mathcal{Q}$ minimizing $D(P, Q)$ has the form

$$(S_1(\widetilde{P}), \ldots, S_I(\widetilde{P}))$$

where $S_i(\widetilde{P})$ is the $Q_i \in \mathcal{S}_i$ minimizing $\|\widetilde{P} - \widetilde{Q}_i\|^2$. For a fixed $Q = (Q_1, \ldots, Q_I) \in \mathcal{Q}$, the $P \in \mathcal{P}$ minimizing $D(P, Q)$ is given by

$$\left(\sum_{i=1}^I c_i Q_i, \ldots, \sum_{i=1}^I c_i Q_i\right).$$

Moreover, a solution to (22) can be found from the standard alternating minimization algorithm for Hilbert spaces on $\mathcal{P}$ and $\mathcal{Q}$.

To this point, we have assumed that the constraint sets $\{\mathcal{S}_i\}_{i=1}^I$ are constant. The results from Section III enable us to look at situations in which the constraint sets $\{\mathcal{S}_{i,n}\}_{i=1}^I$ are time-varying. Returning to the filter design example mentioned above, we are now interested in an adaptive filter. The need for such filters arises in many different situations (see, e.g., [13]).

The time-varying sets $\{\mathcal{S}_{i,n}\}_{i=1}^I$ give rise to sets $\mathcal{Q}_n$, defined in analogy to (23). We assume again that $\mathcal{S}_{i,n} \overset{d_H}{\to} \mathcal{S}_i$ for all $i \in \{1, \ldots, I\}$, and let $\mathcal{Q}$ be defined with respect to the limiting $\{S_i\}_{i=1}^I$ as before. Applying the results from Section V-A and Theorem 4, we obtain convergence and correctness of the AAM algorithm.

VI. Conclusion

We considered a fairly general adaptive alternating minimization algorithm, and found sufficient conditions for its convergence and correctness. This adaptive algorithm has applications in a variety of settings. We discussed in detail how to apply it to three different problems (from statistics, finance, and signal processing).

Acknowledgment

The authors would like to thank the anonymous reviewer as well as the Associate Editor Gerhard Kramer whose comments helped improving the final version of this manuscript.

References

[1] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions, Supplement Issue*, no. l, pp. 205–237, 1984.

[2] 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.

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

[4] T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 2, pp. 369–373, Mar. 1984.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, ser. B, vol. 39, no. 1, pp. 1–38, Nov. 1977.

[6] J. A. O'Sullivan, , A. Vardy, Ed., "Alternating minimization algorithms: From Blahut-Arimoto to expecation-maximization," in *Codes, Curves, and Signals: Common Threads in Communications*. Norwell, MA: Kluwer Academic, 1998, pp. 173–192.

[7] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*. Delft, The Netherlands: Now Publishers, 2004.

[8] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. New York: Springer-Verlag, 1993.

[9] P. L. Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, no. 2, pp. 182–208, Feb. 1993.

[10] A. E. Çetin, Ö. N. Gerek, and Y. Yardimci, "Equiripple FIR filter design by the FFT algorithm," *IEEE Signal Process. Mag.*, vol. 14, no. 2, pp. 60–64, Mar. 1997.

[11] R. A. Nobakht and M. R. Civanlar, "Optimal pulse shape design for digital communication systems by projections onto convex sets," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2874–2877, Dec. 1995.

[12] P. L. Combettes, "Inconsistent signal feasibility problems: Least square solutions in a product space," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 2955–2966, Nov. 1994.

[13] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 1996.

Urs Niesen (S'02) received the M.S. degree from the School of Computer and Communication Sciences at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2005.

He is currently at the Massachusetts Institute of Technology (MIT), Cambridge, in the Department of Electrical Engineering and Computer Science, where he is working toward the Ph.D. degree. His research interests are in the area of communication and information theory.

Devavrat Shah is currently a Jamieson career development Assistant Professor with the Department of Electrical Engineering and Computer Science, the Massachusetts Institute of Technology, Cambridge. His research focus is on theory of large complex networks which includes network algorithms, stochastic networks, network information theory, and large-scale statistical inference.

Prof. Shah was a corecipient of the IEEE INFOCOM best paper award in 2004 and ACM SIGMETRICS/Performance best paper awarded in 2006. He received 2005 George B. Dantzig best dissertation award from the INFORMS. He received an NSF CAREER award in 2006. He is the recipient of the first ACM SIGMETRICS Rising Star Award 2008 for his work on network scheduling algorithms.

Gregory W. Wornell (S'83–M'88–SM'01–F'04) received the B.A.Sc. degree from the University of British Columbia, Vancouver, BC, Canada, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, all in electrical engineering and computer science, in 1985, 1987, and 1991, respectively.

Since 1991, he has been on the faculty at MIT, where he is Professor of Electrical Engineering and Computer Science, and Co-Director of the Center for Wireless Networking. He has held visiting appointments at the former AT&T Bell Laboratories, Murray Hill, NJ, the University of California, Berkeley, and Hewlett-Packard Laboratories, Palo Alto, CA. His research interests and publications span the areas of signal processing, digital communication, and information theory, and include algorithms and architectures for wireless and sensor networks, broadband systems, and multimedia environments.

Dr. Wornell has been involved in the Information Theory and Signal Processing Societies of the IEEE in a variety of capacities, and maintains a number of close industrial relationships and activities. He has won a number of awards for both his research and teaching.