Adaptive Alternating Minimization Algorithms

Urs Niesen, Devavrat Shah, Gregory Wornell

Abstract

The classical alternating minimization (or projection) algorithm has been successful in the context of solving optimization problems over two variables or equivalently of finding a point in the intersection of two sets. The iterative nature and simplicity of the algorithm has led to its application to 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 is quite well-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. 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 surprisingly, these conditions seem to be the minimal ones one would expect in such an adaptive setting. Our result is a generalization of the work by Csiszár and Tusnády [1] on alternating minimization procedures. We present applications of our results to adaptive decomposition of mixtures, adaptive log-optimal portfolio selection, and adaptive filter design.

I. Introduction

A. Background

The problem of finding a point in the intersection of two sets or equivalently of solving an optimization problem over two variables over 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

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

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 values are declared the solution to the original problem. Conditions for the convergence and correctness of such an algorithm, that is, conditions for

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

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

This work was supported in part by NSF under Grant No. CCF-0515109, and by HP through the MIT/HP Alliance.

The authors are with the Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA. Email: {uniesen,devavrat,gww}@mit.edu

Theorem 1. *Let $\mathcal{P}$ and $\mathcal{Q}$ be any two sets, and let $D : \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$. Then the alternating minimization algorithm converges, i.e. (2) holds, if there exists $P \in \mathcal{P}$ such that $D(P, Q_0) < \infty$, and if there exists a nonnegative function $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \cup \{+\infty\}$ such that for all $n \geq 1$ the following two properties hold:*

(a) *Three point property (P, P_n, Q_{n-1}):*

$$\delta(P, P_n) + D(P_n, Q_{n-1}) \leq D(P, Q_{n-1}), \quad \forall P \in \mathcal{P}.$$

(b) *Four point property (P, Q, P_n, Q_n):*

$$D(P, Q_n) \leq D(P, Q) + \delta(P, P_n), \quad \forall P \in \mathcal{P}, Q \in \mathcal{Q}.$$

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 n-th iteration, we are provided with sets $\mathcal{P}_n, \mathcal{Q}_n$ which are *noisy* versions of the sets $\mathcal{P}$ and $\mathcal{Q}$, respectively. That is, we are given a sequence of optimization problems

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

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: choose an arbitrary initial $Q_0 \in \mathcal{Q}_0$; for $n \geq 1$ compute (as in (1))

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

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 $D(P_n, Q_n)$ converges to

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

for large n. 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] described in Theorem 1. The precise results are stated in Theorem 5.

This work was motivated by several applications in which the need for an adaptive alternating minimization algorithm arises. We present three such applications from the areas of estimation, finance, and signal processing.

C. Organization

The remainder of this paper is organized as follows. In Section II, we describe the setup, 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. 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

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

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

$$\mathcal{G}(\mathcal{A}, \mathcal{B}) \triangleq \operatorname*{arg\,min}_{(A,B) \in \mathcal{A} \times \mathcal{B}} D(A, B).$$

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 (Lemma 1, [1])**.** *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[1]*,*

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

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 compact 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 $A_n \overset{d}{\to} A$. Then $A \in \mathcal{A}$.*

Proof: Consider the limit point A of the sequence $\{A_n\}_{n \geq 0}$. 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

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

and $d(\widehat{A}_n, A_n) \to 0$ as $n \to \infty$. Since $d(A_n, A) \to 0$, we have $d(\widehat{A}_n, A) \to 0$. Recall that the sequence $\{\widehat{A}_n\}_{n\geq 0}$ is entirely in $\mathcal{A}$. By compactness of $\mathcal{A}$, the limit points of $\{\widehat{A}_n\}_{n\geq 0}$ must belong to $\mathcal{A}$ (sequential compactness). That is, $A \in \mathcal{A}$. ■

Lemma 4 (Theorem 4.15, 4.19 [2])**.** *Let $(\mathcal{X}, d)$ be a compact metric space, and let $f : \mathcal{X} \to \mathbb{R}$ be continuous. Then, f is bounded and uniformly continuous. That is, for every $\delta > 0$ there exists $\varepsilon > 0$ such that*

$$|f(x) - f(x')| < \delta.$$

for all $x, x' \in \mathcal{X}$ for which $d(x, x') < \varepsilon$.

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

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

Remark 1. Note that if f is uniformly continuous then $w(t) \to 0$ as $t \to 0$. In particular by Lemma 4, this holds if $(\mathcal{X}, d)$ is compact and f is continuous.

III. Adaptive Alternating Minimization Algorithms

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

A. Setup

Consider a compact metric space $(\mathcal{M}, d)$, compact sets $\mathcal{P}, \mathcal{Q} \subset \mathcal{M}$, and a continuous cost 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 $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0}$: $\mathcal{P}_n, \mathcal{Q}_n \subset \mathcal{M}$ 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 $(P_n, Q_n) \in \mathcal{P}_n \times \mathcal{Q}_n$ so that

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

B. Algorithm

The setup described in the last section suggests the following adaptive version of the alternating minimization algorithm for the above setup. Initially, we have $(P_0, Q_0) \in \mathcal{P}_0 \times \mathcal{Q}_0$. Define 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 AAM (Adaptive Alternating Minimization) algorithm in the following. Note that if $\mathcal{P}_n = \mathcal{P}$ and $\mathcal{Q}_n = \mathcal{Q}$ for all n, then the above algorithm is the same as 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 (equivalent to those in [1]) also in the adaptive setup. To this end, assume there exists a continuous function $\delta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ such that the following conditions are satisfied.

C1. *Three point property* (P, P_n, Q_{n-1}): for all $n \geq 1$,

$$\delta(P, P_n) + D(P_n, Q_{n-1}) \leq D(P, Q_{n-1}), \quad \forall P \in \mathcal{P}_n.$$

C2. *Four point property* (P, Q, P_n, Q_n): for all $n \geq 1$,

$$D(P, Q_n) \leq D(P, Q) + \delta(P, P_n), \quad \forall P \in \mathcal{P}_n, Q \in \mathcal{Q}_n.$$

We are now ready to show convergence and correctness of the AAM algorithm.

Theorem 5. *Let $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0}$ be compact subsets of $\mathcal{M}$ such that*

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

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 lim inf 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 ω 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], the only 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 5.

D. Proof of Theorem 5

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

$$d_2((A, B), (A', B')) = d(A, A') + d(B, B')$$

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 $d_2((A, B), (A', B')) \leq \varepsilon$, $|D(A, B) - D(A', B')| \leq \omega(\varepsilon)$. Moreover, continuity of D and compactness of $\mathcal{M} \times \mathcal{M}$ imply by Lemma 4 (and subsequent 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 5, 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 $n \to \infty$.

Now we are ready to embark on the proof of Theorem 5. From the hypothesis of Theorem 5, conditions C1 and C2 are satisfied. Add the inequalities provided by these two conditions to obtain the following: for all $n \geq 1$, $P \in \mathcal{P}_n, Q \in \mathcal{Q}_n$,

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

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

$$\left|D(P_n, \widehat{Q}_n) - D(P_n, Q_{n-1})\right| \leq \omega(\gamma_n), \tag{5}$$

since $d_2((P_n, \widehat{Q}_n), (P_n, Q_{n-1})) \leq \gamma_n$. From (5) and the AAM algorithm, we have

$$\begin{aligned} D(P_n, Q_n) &= \min_{Q \in \mathcal{Q}_n} D(P_n, Q) \\ &\leq D(P_n, \widehat{Q}_n), \quad (\text{since } \widehat{Q}_n \in \mathcal{Q}_n), \\ &\leq D(P_n, Q_{n-1}) + \omega(\gamma_n). \end{aligned} \tag{6}$$

Adding inequalities (4) and (6),

$$D(P_n, Q_n) + D(P, Q_n) \leq D(P, Q_{n-1}) + D(P, Q) + \omega(\gamma_n), \tag{7}$$

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

From the hypothesis of the theorem, 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 $n \geq 0$. Replacing (P, Q) in (7) by this (P_n^*, Q_n^*), we obtain

$$D(P_n, Q_n) + D(P_n^*, Q_n) \leq D(P_n^*, Q_{n-1}) + D(P_n^*, Q_n^*) + \omega(\gamma_n). \tag{8}$$

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

$$D(P_n^*, Q_n^*) \leq D(P^*, Q^*) + \omega(\varepsilon_n). \tag{9}$$

Moreover, by definition of d_2 and choice of the (P_n^*, Q_n^*), we have $d(P_{n-1}^*, P_n^*) \leq \gamma_n$. Therefore

$$D(P_n^*, Q_{n-1}) \leq D(P_{n-1}^*, Q_{n-1}) + \omega(\gamma_n). \tag{10}$$

Combining inequalities (9) and (10) with (8), we obtain

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

Define $a_n \triangleq D(P_n, Q_n) - 2\omega(\gamma_n) - \omega(\varepsilon_n)$, $b_n \triangleq D(P_n^*, Q_n)$ and $c \triangleq D(P^*, Q^*)$. Since D is a bounded function over $\mathcal{M} \times \mathcal{M}$, we have $\limsup_{n\to\infty} |b_n| < \infty$. Applying Lemma 2,

$$\liminf_{n\to\infty} D(P_n, Q_n) \leq D(P^*, Q^*) + \limsup_{n\to\infty} 2\omega(\gamma_n) + \omega(\varepsilon_n). \tag{11}$$

Since $\gamma_n \to 0$ and $\varepsilon_n \to 0$ imply $2\omega(\gamma_n) + \omega(\varepsilon_n) \to 0$, (11) yields

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

Finally, any limit point of $\{(P_n, Q_n)\}_{n\geq 0}$ belongs to $\mathcal{P} \times \mathcal{Q}$ by Lemma 3, which by continuity of D implies 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 lim inf belong to $\mathcal{G}(\mathcal{P}, \mathcal{Q})$. This completes the proof the first part of Theorem 5.

Suppose now that we have in addition

$$\sum_{n=0}^{\infty} \omega(2\varepsilon_n) < \infty. \tag{12}$$

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

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

Thus by (12),

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

and applying again Lemma 2 yields

$$\lim_{n \to \infty} D(P_n, Q_n) = D(P^*, Q^*). \tag{13}$$

As every limit point of $\{(P_n, Q_n)\}_{n \geq 0}$ belongs to $\mathcal{P} \times \mathcal{Q}$, (13) and continuity of D imply that if (12) 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 5.

IV. Divergence Minimization

In this section, we specialize the setup and algorithm from Section III to the special 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 [3], [4], selection of log-optimal portfolios [5], and maximum likelihood estimation from incomplete data [6]. These problems were shown to be divergence minimization problems in [1]. For further applications of alternating divergence minimization algorithms, see [7]. We describe applications to the problem of adaptive mixture decomposition and of adaptive log-optimal portfolio selection.

A. Setup

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

$$\sum_{\sigma \in \Sigma} P(\sigma) \leq \Theta, \text{ and } P(\sigma) \geq \theta, \ \forall \ \sigma \in \Sigma. \tag{14}$$

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

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

for any $P, Q \in \mathcal{M}$. Note that (14) 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 δ

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

for any $P, Q \in \mathcal{M}$.

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 as required in Theorem 1. The next corollary of Theorem 5 extends this result to the adaptive setup considered here.

Corollary 6. *Let $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0}$ be compact convex subsets of $\mathcal{M} = \mathcal{M}(\Sigma, \theta, \Theta)$ such that*

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

Then, with the AAM *algorithm applied to the divergence cost function D as defined above,*

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

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

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

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 $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

Proof: As stated above, the space $\mathcal{M} = \mathcal{M}(\Sigma, \theta, \Theta)$ with metric d is a compact metric space, and the function D is continuous. Conditions C1 and C2 follow from the "three point" and "four point" properties for divergences established in [1]. The corollary follows now from applying Theorem 5. ■

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.

We are given a sequence of 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, and $c_i \geq c_0 > 0$ 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 $\overline{P}_n$ 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., [8, Lemma 3.1])

$$\arg\min_{\{\tilde{c}_i\}} D\Big(\overline{P}_n \Big\| \textstyle\sum_{i=1}^{I} \tilde{c}_i \mu_i\Big),$$

Following [8, Example 5.1], we define

$$\begin{aligned}
&\Sigma \triangleq \{1, \ldots, I\} \times \mathcal{Y},\\
&\mathcal{Q}_n = \mathcal{Q} \triangleq \{Q : Q(i, y) = \tilde{c}_i \mu_i(y), \text{ for some } \{\tilde{c}_i\} \text{ with } \textstyle\sum_i \tilde{c}_i = 1, \tilde{c}_i \geq c_0 \forall i\},\\
&\mathcal{P}_n \triangleq \{P : \textstyle\sum_{i=1}^{I} P(i, y) = \overline{P}_n(y), P(i, y) \geq 0 \forall i, y\}.
\end{aligned} \tag{15}$$

Note that $\mathcal{P}_n$ and $\mathcal{Q}$ are convex and compact. From [8, Lemma 5.1], we have

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

and the minimizer of the left hand side can be recovered from the corresponding marginal of the optimal Q on the right hand side.

Fix a P, assuming without loss of generality that

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

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

$$\frac{1}{\eta(J^*)} \sum_{y\in\mathcal{Y}} P(J^*,y) > c_0,$$
$$\frac{1}{\eta(J^*)} \sum_{y\in\mathcal{Y}} P(J^*+1,y) \leq c_0,$$

where $P(I+1,y) \triangleq 0$. Then the optimal $\{\tilde{c}_i\}$ are given by

$$\tilde{c}_i = \frac{1}{\eta(J^*)} \sum_{y\in\mathcal{Y}} P(i,y)$$

for $i \leq J^*$ and $\tilde{c}_i = c_0$ for $J^* < i \leq I$. For fixed Q, the minimizing P is

$$P(i,y) = \frac{\tilde{c}_i \mu_i(y)}{\sum_i \tilde{c}_i \mu_i(y)} \overline{P}_n(y). \tag{16}$$

We now check that (14) is satisfied. As $\mathcal{P}_n$ and $\mathcal{Q}$ are sets of distributions, we can choose $\Theta = 1$. For all $Q \in \mathcal{Q}$, $i \in \{1,\ldots,I\}$, $y \in \mathcal{Y}$, we have $Q(i,y) \geq \mu_0 c_0 > 0$. However, for $P \in \mathcal{P}_n$, we have in general only $P(i,y) \geq 0$. In order to apply Corollary 6, 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

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

for all $y \in \mathcal{Y}$. As $\sum_i c_i \mu_i(y) \geq \mu_0$ this implies $\overline{P}_n(y) \geq \frac{\mu_0}{2}$ for all y. From (16), this implies that the projection P on $\mathcal{P}$ of any point in $\mathcal{Q}$ satisfies $P(i,y) \geq \frac{1}{2} c_0 \mu_0^2 \triangleq p_0$ for all $i \in \{1,\ldots,I\}, y \in \mathcal{Y}$. Hence $\mathcal{M}(\Sigma,\theta,\Theta)$ satisfies (14) with $\theta = \frac{1}{2} c_0 \mu_0^2$ and $\Theta = 1$.

It remains to argue that $\overline{P}_n$ is close to $\sum_i c_i \mu_i(y)$. Suppose instead of constructing the set $\mathcal{P}_n$ (see (15)) with respect to $\overline{P}_n$, we construct it with respect to the distribution $\overline{\overline{P}}_n$ defined as

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

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

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

Hence we have $\mathcal{P}_n \overset{d_H}{\to} \mathcal{P}$ almost surely, where $\mathcal{P}$ is constructed as in (15) with respect to the true distribution $\sum_i c_i \mu_i$.

Applying now Corollary 6 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 $\{(P_n, Q_n)\}_{n\geq 0}$ achieving this lim inf is an element of $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

Since by the law of the iterated logarithm, convergence of $\bar{P}_n$ to P is only $\Theta(\sqrt{\log\log n}/\sqrt{n})$ as $n \to \infty$ almost surely, and since $\omega(\varepsilon) = o(\varepsilon)$ as $\varepsilon \to 0$ only if D is a constant [9], we can in this scenario *not* conclude from Corollary 6 that $\lim_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$.

As noted in [8], 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 [5]. 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 setup and algorithm from Section III to the special 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 [10]. In the context of Hilbert spaces, the alternating minimization algorithm is often called POCS (Projection Onto Convex Sets).

A. Setup

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, convex and nonnegative. Define the function δ (as part of conditions C1 and C2), as

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

In [1], it has been established that for convex $\mathcal{P}$ and $\mathcal{Q}$ the pair of functions D, δ satisfies the "three point" and "four point" properties as required in Theorem 1. The next corollary of Theorem 5 extends this result to the adaptive setup.

Corollary 7. *Let $\{(\mathcal{P}_n, \mathcal{Q}_n)\}_{n\geq 0}$ be convex compact subsets of $\mathcal{M}$ as defined above such that*

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

Then, with the AAM algorithm applied to the cost function D as defined above,

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

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

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

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 $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

Proof: As stated above, the space $(\mathcal{M}, d)$ is a compact metric space, and the function D is continuous. Conditions C1 and C2 follow from the "three point" and "four point" properties in [1]. Applying Theorem 5 yields the corollary. ∎

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

In this section, 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

$$\sum_{i=1}^{I} c_i d(A, \mathcal{S}_i), \tag{17}$$

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 [10]. As an example, in a filter design problem, the $\mathcal{S}_i$ could be constraints on the impulse and frequency responses of a filter [11], [12].

Following [13], 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 $i \in \{1, \ldots, I\}$ are the components of A and B. Let

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

and let

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

be the I-fold product of the convex hull of the constraint sets $\{\mathcal{S}_i\}_{i=1}^I$. Since each of the sets $\mathcal{S}_i$ is bounded, $\mathcal{M}$ is bounded and by definition also convex and closed. 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

$$\mathcal{Q} \triangleq \mathcal{S}_1 \times \cdots \times \mathcal{S}_I. \tag{18}$$

For a fixed $P \in \mathcal{P}$, the $Q \in \mathcal{Q}$ minimizing $D(P, Q)$ has the form

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

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

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

Moreover, a solution to (17) can be found from the standard alternating minimization algorithm for Hilbert spaces on $\mathcal{P}$ and $\mathcal{Q}$.

Up 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. Coming back 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., [14]).

The time-varying sets $\{\mathcal{S}_{i,n}\}_{i=1}^I$ give rise to sets $\mathcal{Q}_n$, defined in analogy to (18). 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 Corollary 7, we obtain that under the AAM algorithm $\liminf_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$, every limit point of subsequences of $\{(P_n, Q_n)\}_{n\geq 0}$ achieving this lim inf is in $\mathcal{G}(\mathcal{P}, \mathcal{Q})$, and if also $\sum_{n=0}^{\infty} \omega(2\varepsilon_n) < \infty$ then $\lim_{n\to\infty} D(P_n, Q_n) = D(\mathcal{P}, \mathcal{Q})$, and every limit point of $\{(P_n, Q_n)\}_{n\geq 0}$ is in $\mathcal{G}(\mathcal{P}, \mathcal{Q})$.

VI. Conclusions

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).

References

[1] I. Csiszár and G. Tusnády. Information geormetry and alternating minimization procedures. *Statistics & Decisions, Supplement Issue*, (1):205–237, 1984.

[2] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976.

[3] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, January 1972.

[4] R. E. Blahut. Computation of channel capacity and rate distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, July 1972.

[5] T. M. Cover. An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373, March 1984.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, November 1977.

[7] J. A. O'Sullivan. Alternating minimization algorithms: From Blahut-Arimoto to expecation-maximization. In A. Vardy, editor, *Codes, Curves, and Signals: Common Threads in Communications*, pages 173–192. Kluwer Academic, 1998.

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

[9] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer, 1993.

[10] P. L. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, February 1993.

[11] A. E. Çetin, Ö. N. Gerek, and Y. Yardimci. Equiripple FIR filter design by the FFT algorithm. *IEEE Signal Processing Magazine*, pages 60–64, March 1997.

[12] R. A. Nobakht and M. R. Civanlar. Optimal pulse shape design for digital communication systems by projections onto convex sets. *IEEE Transactions on Communications*, 43(12):2874–2877, December 1995.

[13] P. L. Combettes. Inconsistent signal feasibility problems: Least square solutions in a product space. *IEEE Transactions on Signal Processing*, 42(11):2955–2966, November 1994.

[14] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, 1996.