Electronic Journal of Statistics
Vol. 3 (2009) 1455–1472
ISSN: 1935-7524
DOI: 10.1214/09-EJS458

Asymptotic properties of predictive recursion: Robustness and rate of convergence

Ryan Martin

Department of Mathematical Sciences
Indiana University-Purdue University Indianapolis
e-mail: rgmartin@math.iupui.edu

and

Surya T. Tokdar

Department of Statistical Science
Duke University
e-mail: st118@stat.duke.edu

Abstract: Here we explore general asymptotic properties of Predictive Recursion (PR) for nonparametric estimation of mixing distributions. We prove that, when the mixture model is mis-specified, the estimated mixture converges almost surely in total variation to the mixture that minimizes the Kullback-Leibler divergence, and a bound on the (Hellinger contrast) rate of convergence is obtained. Simulations suggest that this rate is nearly sharp in a minimax sense. Moreover, when the model is identifiable, almost sure weak convergence of the mixing distribution estimate follows.

PR assumes that the support of the mixing distribution is known. To remove this requirement, we propose a generalization that incorporates a sequence of supports, increasing with the sample size, that combines the efficiency of PR with the flexibility of mixture sieves. Under mild conditions, we obtain a bound on the rate of convergence of these new estimates.

AMS 2000 subject classifications: Primary 62G20; secondary 62G05, 62G07, 62G35.
Keywords and phrases: Almost supermartingale, density estimation, empirical Bayes, Kullback-Leibler projection, mixture models.

Received August 2009.

1. Introduction

Despite a well-developed theory and numerous applications of mixture models, estimation of a mixing distribution remains a challenging statistical problem. However, some recent progress has been made through a computationally efficient nonparametric estimate due to Newton [17]; see also Newton, et al. [18] and Newton and Zhang [19]. A mixture model views data $(X_1, \ldots, X_n) \in \mathcal{X}^n$

as independent observations from a density $m(x)$ of the form

$$m_F(x) = \int_\Theta p(x|\theta)\,dF(\theta), \quad x \in \mathscr{X},\ F \in \mathbb{F} \tag{1.1}$$

where $\mathbb{F} = \mathbb{F}(\Theta, \mu)$ is the class of probability measures on a parameter space $(\Theta, \mathscr{B})$ dominated by a σ-finite measure μ, and $\{p(\cdot|\theta) : \theta \in \Theta\}$ is a parametric family of densities on $(\mathscr{X}, \mathscr{A})$, dominated by a σ-finite measure ν. More succinctly, the mixture model (1.1) assumes $m \in \mathbb{M} := \{m_F : F \in \mathbb{F}\}$. To estimate F from the data $X_1, \ldots, X_n$, Newton [17] proposed the following n-step recursive algorithm, called Predictive Recursion (PR):

Algorithm PR. Choose an initial measure $F_0 \in \mathbb{F}$ having μ-density f_0, and a sequence of weights $\{w_i : i \geq 1\} \subset (0, 1)$. For $i = 1, \ldots, n$, compute

$$f_i(\theta) = (1 - w_i) f_{i-1}(\theta) + w_i \, \frac{p(X_i|\theta) f_{i-1}(\theta)}{\int_\Theta p(X_i|\theta') f_{i-1}(\theta')\, d\mu(\theta')}, \quad \theta \in \Theta, \tag{1.2}$$

and produce F_n, the measure with μ-density f_n, as the final estimate of F.

Key features of PR include its speed and its unique flexibility to estimate a mixing distribution which has a density with respect to any user-defined dominating measure μ. The latter is a practically important property as a number of modern applications demand existence of a mixing density with respect to a specified dominating measure. For example, high-dimensional empirical Bayes analysis, spurred mainly by recent developments in DNA microarray technologies, starts with a Bayesian model whose prior/mixing distribution is itself a mixture of both discrete and continuous components; see Efron [5] and the references therein. Estimation of prior/mixing distributions in this “two-groups model” context is a promising application of PR; see Bogdan, et al. [3].

Until recently, very little was known about the large-sample behavior of PR. Ghosh and Tokdar [8] used a novel martingale argument to prove, when Θ is finite and $m = m_F \in \mathbb{M}$, that $F_n \to F$ a.s. Martin and Ghosh [16] proved a slightly stronger consistency theorem using tools from stochastic approximation theory. Most recently, Tokdar, Martin, and Ghosh [23] (henceforth, TMG) handled the case of a more general parameter space Θ by extending the martingale argument to the $\mathscr{X}$-space, proving that the mixture density estimate

$$m_n(x) := m_{F_n}(x) = \int p(x|\theta)\,dF_n(\theta), \quad x \in \mathscr{X}, \tag{1.3}$$

converges a.s. to $m_F(x)$ in the L_1 topology. From L_1 convergence of m_n, consistency of F_n in the weak topology on Θ is obtained.

In this paper, we extend the convergence results of TMG in two important directions. First, in the more general context, where the mixture model $\mathbb{M}$ need not contain the true density m, we show that the estimated mixture m_n in (1.3) is asymptotically robust in the sense that it converges almost surely in the total variation (or Hellinger) topology to the mixture m_F that minimizes the Kullback-Leibler (KL) divergence $K(m, m_\Phi) = \int \log(m/m_\Phi) m \, d\nu$. When the mixing distribution is identifiable, we also obtain weak convergence of F_n. Our second main result is a bound on the rate of convergence for m_n. For a unified treatment of the well- and mis-specified cases, we consider the *Hellinger contrast*

$$\rho(m_n, m_F) = \left[\frac{1}{2} \int \left(\sqrt{\frac{m_n}{m_F}} - 1 \right)^2 m \, d\nu \right]^{1/2}, \tag{1.4}$$

where m_F is a mixture that minimizes $K(m, m_\Phi)$. The Hellinger contrast has been previously used to study asymptotics of maximum likelihood and Bayes estimates under model mis-specification; see, for example, Patilea [20], and Kleijn and van der Vaart [10]. In Section 4, we show that if $\mathbb{F}$ is compact, then $\sqrt{a_n}\rho(m_n, m_F) \to 0$ almost surely, where $a_n = \sum_{i=1}^n w_i$. This establishes a direct connection between the choice of weights w_i and the performance of the resulting PR estimate. Moreover, this bound is derived without using any structural knowledge about the mixands $p(x|\theta)$ and applies to a wide range of such densities, including Normal, Gamma, and Poisson. The conditions on w_i required for this result are satisfied by $w_n \asymp n^{-\gamma}$, $\gamma \in (2/3, 1]$, leading to $a_n \asymp n^{1-\gamma}$. Consequently, the Hellinger contrast convergence rate of m_n is strictly faster than $n^{-(1-\gamma)/2}$. How this relates to the rate of convergence for F_n remains an important open problem.

Our nearly $n^{-1/6}$ bound on the convergence rate for m_n closely matches the results in Genovese and Wasserman [6] derived for the special case of finite Gaussian mixtures. But it falls short of the nearly parametric rates obtained in Li and Barron [14] and Ghosal and van der Vaart [7]. However, the simulation results presented in Section 4.3 indicate that our rate is minimax in nature and, therefore, should not be directly compared to the rates in these two papers. In fact, empirical evidence suggests that, in some cases, the PR estimates can converge faster than what our bounds indicate.

A shortcoming of PR is that one needs to specify the compact mixing parameter space Θ *a priori*. To remove this requirement, we propose, in Section 5, a generalized PR (GPR) algorithm that features an increasing *sieve-like* sequence of supports $\Theta_i \subset \Theta_{i+1}$. We obtain a bound on the rate of convergence for GPR in the special case where $m = m_F$ and F has an unknown compact support.

2. Almost supermartingales

The primary tool used to prove the new results of this paper is an "almost supermartingale" convergence theorem of Robbins and Siegmund [22]. For convenience, we give the statement of this result here. Let $\{M_n : n \geq 1\}$ be a sequence of non-negative random variables adapted to a filtration $\{\mathscr{F}_n : n \geq 1\}$. Suppose there are non-negative random variables $\{\beta_n, \xi_n, \zeta_n\}$ such that

$$\mathsf{E}(M_n | \mathscr{F}_{n-1}) \leq (1 + \beta_{n-1}) M_{n-1} + \xi_{n-1} - \zeta_{n-1}, \quad n \geq 1. \tag{2.1}$$

If both $\beta_n \equiv 0$ and $\xi_n \equiv 0$, then M_n would be exactly a supermartingale. But, more generally, Robbins and Siegmund [22] call a sequence $\{M_n\}$ that satisfies (2.1) an *almost supermartingale*.

Theorem 2.1 (Robbins-Siegmund). *Suppose* (2.1) *holds, and that* $\sum_n \beta_n < \infty$ *and* $\sum_n \xi_n < \infty$ *a.s. Then* M_n *converges a.s. and* $\sum_n \zeta_n < \infty$ *a.s.*

That is, even if the sequence is not exactly a supermartingale, as long as the perturbations β_n and ξ_n vanish fast enough, then the conclusions of the usual martingale convergence theorem remain valid. For further discussion on Theorem 2.1 and its extensions, see Lai [11].

3. Kullback-Leibler projections

It is quite natural to use the KL divergence to study the large-sample properties of PR. Indeed, Martin and Ghosh [16] remark that, for Θ finite, $\ell(\Phi) = K(m, m_\Phi)$ is a Lyapunov function for the differential equation that characterizes the asymptotics of PR: roughly, the KL divergence controls the dynamics of PR, driving the estimates toward a stable equilibrium. But when $m \notin \mathbb{M}$, this equilibrium cannot be m. In such cases, we consider the mixture m_F that is closest to m in a KL sense; that is,

$$K(m, m_F) = K(m, \overline{\mathbb{M}}) := \inf\{K(m, m_\Phi) : \Phi \in \overline{\mathbb{F}}\}, \tag{3.1}$$

where $\overline{\mathbb{F}}$ is the weak closure of $\mathbb{F}$, and $\overline{\mathbb{M}} = \{m_F : F \in \overline{\mathbb{F}}\}$. We call m_F the *KL projection* of m onto $\mathbb{M}$. Similar ideas may be found in [12, 20, 10].

Existence of the KL projection is an important issue, and various results are available; see, for example, Liese and Vadja [13, Chap. 8]. Here we prove a simple result which gives sufficient conditions for the existence of a KL projection in our special case of mixtures. Assume the following:

A1. $\mathbb{F}$ is pre-compact with respect to the weak topology.
A2. $\theta \mapsto p(x|\theta)$ is bounded and continuous for ν-almost all x.

Lemma 3.1. *Under* A1–A2*, there exists* $F \in \overline{\mathbb{F}}$ *such that* $K(m, m_F) = K(m, \overline{\mathbb{M}})$.

Proof. Choose any $\Phi \in \overline{\mathbb{F}}$ and any $\{\Phi_s\} \subset \mathbb{F}$ such that $\Phi_s \to \Phi$ weakly. Then A2 and Scheffé's theorem imply $m_{\Phi_s} \to m_\Phi$ in the $L_1(\nu)$ and, hence, the weak topology. Further,

$$\kappa(\Phi) := K(m, m_\Phi) \leq \liminf_{s\to\infty} K(m, m_{\Phi_s}) = \liminf_{s\to\infty} \kappa(\Phi_s),$$

where the inequality follows from weak lower semi-continuity of $K(m, \cdot)$; see Liese and Vadja [13], Theorem 1.47. Consequently, $\kappa(\cdot)$ is weakly lower semi-continuous and, therefore, must attain its infimum on the compact $\overline{\mathbb{F}}$. A similar proof may be given based on Lemma 4 of Brown, et al. [4]. □

Remark 3.2. Convexity of the set $\overline{\mathbb{M}}$ and of the mapping $K(m, \cdot)$ together imply that the KL projection m_F in Lemma 3.1 is unique. However, in general, there could be many mixing distributions $\Phi \in \mathbb{F}$ whose mixture m_Φ corresponds to the KL projection m_F. Identifiability is needed to guarantee uniqueness of F.

Next we state one more important property of the KL projection which will be useful in what follows. Various proofs of this result can be found in the literature; see, e.g., Patilea [20] or Kleijn and van der Vaart [10].

Lemma 3.3. *If m_F is the KL projection of m onto the set $\overline{\mathbb{M}}$ of mixtures, then $\int (m_\Phi/m_F) m \, d\nu \leq 1$ for all $m_\Phi \in \overline{\mathbb{M}}$.*

4. Robustness and rate of convergence

4.1. Preliminaries

We begin with our assumptions and some preliminary lemmas; proofs can be found in the Appendix. Let $\{w_i : i \geq 1\}$ be the user-specified weight sequence in the PR algorithm, and define the sequence of partial sums $a_n = \sum_{i=1}^n w_i$. In addition to A1–A2 in Section 3, assume the following:

A3. $w_n > 0$, $w_n \downarrow 0$, $\sum_n w_n = \infty$, and $\sum_n a_n w_n^2 < \infty$.
A4. There exists $B < \infty$ such that

$$\sup_{\theta_1, \theta_2 \in \Theta} \int \left[\frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 m(x) \, d\nu(x) < B.$$

Condition A3 is satisfied if $w_n \asymp n^{-\gamma}$, for $\gamma \in (2/3, 1]$. The square-integrability condition A4 is the strongest assumption, but it does hold for Exponential family mixands (including Normal or Poisson) with sufficient statistic $S(x)$, provided that Θ is compact and mS^{-1} admits a moment-generating function on Θ. If one is willing to assume that $m \in \mathbb{M}$, then A4 can be replaced by a less restrictive condition, depending only on $p(x|\theta)$; cf. assumption A5 in TMG (p. 2505).

Our new developments are partially based on calculations in TMG, and we begin by recording a few of these for future reference. First, let $R(x)$ be the remainder term of a first-order Taylor approximation of $\log(1+x)$ at $x = 0$; that is, $\log(1+x) = x - x^2 R(x)$, $x > -1$, where $R(x)$ satisfies

$$0 \leq R(x) \leq \max\{1, (1+x)^{-2}\}, \quad x > -1. \tag{4.1}$$

Next, write the PR estimate $m_n \in \mathbb{M}$ in (1.3) as

$$m_n(x) = (1 - w_n) m_{n-1}(x) + w_n h_{n,X_n}(x),$$

where

$$h_{n,y}(x) = \frac{\int p(x|\theta) p(y|\theta) dF_{n-1}(\theta)}{m_{n-1}(y)}, \quad x, y \in \mathcal{X}.$$

For notational convenience, also define the function

$$H_{n,y}(x) = \frac{h_{n,y}(x)}{m_{n-1}(x)} - 1, \quad x, y \in \mathcal{X}.$$

Then the KL divergence $K_n := K(m, m_n)$ satisfies

$$\begin{aligned} K_n - K_{n-1} &= \int m \log(m_{n-1}/m_n)\, d\nu \\ &= -\int m \log(1 + w_n H_{n,X_n})\, d\nu \\ &= -w_n \int m\, H_{n,X_n}\, d\nu + w_n^2 \int m\, H_{n,X_n}^2 R(w_n H_{n,X_n})\, d\nu \end{aligned} \tag{4.2}$$

where $R(x)$ satisfies (4.1). Let $\mathscr{A}_{n-1}$ be the σ-algebra generated by the data sequence $X_1, \ldots, X_{n-1}$. Since K_{n-1} is $\mathscr{A}_{n-1}$-measurable, upon taking conditional expectation with respect to $\mathscr{A}_{n-1}$ we get

$$\mathsf{E}(K_n | \mathscr{A}_{n-1}) = K_{n-1} - w_n T(F_{n-1}) + w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1}) \tag{4.3}$$

where $T(\cdot)$ and Z_n are defined as

$$T(\Phi) = \int_\Theta \left\{ \int_{\mathcal{X}} \frac{m(x)}{m_\Phi(x)} p(x|\theta)\, d\nu(x) \right\}^2 d\Phi(\theta) - 1, \quad \Phi \in \overline{\mathbb{F}} \tag{4.4}$$

$$Z_n = \int m\, H_{n,X_n}^2 R(w_n H_{n,X_n})\, d\nu \tag{4.5}$$

Note that $T(F_{n-1})$ is exactly M_n^* defined in TMG (p. 2508). The following properties of $T(\cdot)$ will be critical in the proof of our main result.

Lemma 4.1. (a) *$T(\Phi) \geq 0$ with equality iff $K(m, m_\Phi) = K(m, \overline{\mathbb{M}})$.*
(b) *Under* A1, A2 *and* A4*, T is continuous in the weak topology on $\mathbb{F}$.*

Remark 4.2. For some further insight into the relationship between $T(\Phi)$ and the KL divergence $K(m, m_\Phi)$, define

$$D(\theta; \Phi) = \int \frac{m(x)}{m_\Phi(x)} p(x|\theta)\, d\nu(x) - 1$$

and notice that $T(\Phi) = \int D^2(\theta; \Phi)\, d\Phi(\theta)$. Some analysis shows that $D(\theta; \Phi)$ is the negative Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\Phi$ in the direction of $p(\cdot|\theta)$. Now, if $T(\Phi) = 0$, then $D(\theta; \Phi) = 0$ for μ-almost all θ and, hence,

$$D(\Psi; \Phi) = \int D(\theta; \Phi)\, d\Psi(\theta),$$

the negative Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\Phi$ in the direction of m_Ψ, is zero for all $\Psi \in \mathbb{F}$. The fact that the Gâteaux derivative vanishes in all directions suggests that m_Φ is a point at which the infimum $K(m, \overline{\mathbb{M}})$ is attained, and this is exactly the conclusion of Lemma 4.1(a). A similar characterization of the NPMLE is given in Lindsay [15].

In light of Lemma 4.1(a) and (4.3), we see in K_n the makings of an almost supermartingale (2.1). The following bound on Z_n is needed to push through the argument based on Theorem 2.1.

Lemma 4.3. *Under condition* A4, $Z_n \leq 1 + B$ *a.s. for all* n.

Our last preliminary result of this section gets is needed to bound the convergence rate of $K(m, m_n)$. Define $K_n^* := K(m, m_n) - K(m, \overline{\mathbb{M}}) \geq 0$.

Lemma 4.4. *Suppose a KL minimizer* F *exists in the interior of* $\mathbb{F}$. *Then under conditions* A3–A4, $\sum_{n=1}^{\infty} w_n K_{n-1}^* < \infty$ *a.s.*

4.2. Main results

We are now ready to state and prove our main results. Those convergence properties advertised in Section 1 are corollaries to Theorems 4.5 and 4.8 that follow.

Theorem 4.5. *Under conditions* A1–A4, $K_n^* \to 0$ *a.s.*

Proof. From (4.3) we have

$$\mathsf{E}(K_n^* | \mathscr{A}_{n-1}) = K_{n-1}^* + w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1}) - w_n T(F_{n-1}). \tag{4.6}$$

This is of the form (2.1) with $\beta_n \equiv 0$, $\xi_{n-1} = w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1})$, and $\zeta_{n-1} = w_n T(F_{n-1})$. Therefore, from Theorem 2.1 we get $K_n^* \to K_\infty^* \geq 0$ a.s. and

$$\sum_n w_n T(F_{n-1}) < \infty \quad \text{a.s.} \tag{4.7}$$

It remains to show that $K_\infty^* = 0$ a.s. Suppose, on the contrary, that $K_\infty^* > 0$ with positive probability. Then there exists $\varepsilon > 0$ such that

$$K(m, m_n) > K(m, \overline{\mathbb{M}}) + \varepsilon$$

for all but perhaps finitely many n. Recall the proof of Lemma 3.1, which shows that the mapping $\kappa(\Phi) = K(m, m_\Phi)$ is lower semi-continuous with respect to the weak topology on $\mathbb{F}$. Consequently,

$$\mathbb{F}_\varepsilon := \{\Phi \in \mathbb{F} : \kappa(\Phi) > K(m, \overline{\mathbb{M}}) + \varepsilon\} \subset \mathbb{F}$$

is a weakly open set, and its closure $\overline{\mathbb{F}}_\varepsilon$ is compact by A1. Since $F_n \in \overline{\mathbb{F}}_\varepsilon$ for all but finitely many n, Lemma 4.1 implies that $T(F_{n-1})$ is bounded away from zero. But this and A3 together contradict (4.7). Therefore, $K_\infty^* = 0$ a.s. □

Next we show that $K_n^* \to 0$ implies $\|m_n - m_F\| \to 0$, where $\|\cdot\|$ denotes the $L_1(\nu)$ norm.

Corollary 4.6. *Under* A1–A4, m_n *converges to* m_F *a.s. in* $L_1(\nu)$.

Proof. Suppose not; that is, there is an $\varepsilon > 0$ and a subsequence $\{n_s\}$ such that $\|m_{n_s} - m_F\| > \varepsilon$ for all s. Then, by A1, this sequence has a further subsequence $\{n_{s(t)}\}$ such that $F_{n_{s(t)}}$ converges weakly to some $F_\infty \in \overline{\mathbb{F}}$. By A2, $m_{n_{s(t)}}$ converges to $m_\infty := m_{F_\infty}$ pointwise and in $L_1(\nu)$ by Scheffé's theorem;

therefore, $m_\infty \neq m_F$. Define $u_t = m_{n_{s(t)}}/m_\infty - 1$. Then $u_t \to 0$ pointwise and $\{u_t\}$ is uniformly integrable with respect to $m\,d\nu$ by A4 and Jensen's inequality:

$$\sup_t \int u_t^2 m\,d\nu = \sup_t \int \left(\frac{m_{n_{s(t)}}}{m_\infty} - 1\right)^2 m\,d\nu < 1 + B.$$

Theorem 4.8, together with Vitali's theorem [2, Theorem 25.10], implies that

$$\begin{aligned} 0 &\leq K(m, m_\infty) - K(m, m_F) \\ &= K(m, m_\infty) - \lim_{t\to\infty} K(m, m_{n_{s(t)}}) \\ &= \lim_{t\to\infty} \int m \log(m_{n_{s(t)}}/m_\infty)\,d\nu \leq \lim_{t\to\infty} \int u_t m\,d\nu = 0 \end{aligned}$$

Therefore, $K(m, m_\infty) = K(m, m_F)$, and uniqueness of the KL projection implies $m_\infty = m_F$, which contradicts our supposition. □

Corollary 4.6 suggests that F_n converges to some $F \in \overline{\mathbb{F}}$ at which the infimum $K(m, \overline{\mathbb{M}})$ is attained. However, to conclude weak convergence of F_n from L_1 convergence of m_n, we need two additional conditions:

A5. Identifiability: $m_\Phi = m_\Psi$ ν-a.e. implies $\Phi = \Psi$; cf. Remark 3.2.

A6. For any $\varepsilon > 0$ and any compact $\mathcal{X}' \subset \mathcal{X}$, there exists a compact $\Theta' \subset \Theta$ such that $\int_{\mathcal{X}'} p(x|\theta)\,d\nu(x) < \varepsilon$ for all $\theta \notin \Theta'$.

With conditions A5–A6 and Theorem 3 of TMG, the next result follows immediately from Corollary 4.6.

Corollary 4.7. *Under* A1–A6, *$F_n \to F$ a.s. in the weak topology, where $F \in \overline{\mathbb{F}}$ is the unique mixing distribution that satisfies $K(m, m_F) = K(m, \overline{\mathbb{M}})$. In particular, if $m \in \mathbb{M}$, then F_n is a consistent estimate of the true mixing density, in the weak topology.*

A slight modification of Theorem 4.5 produces a bound on the rate of convergence. But one extra assumption is needed to push through the proof, namely, that the KL minimizer F is dominated by μ. The precise result is next.

Theorem 4.8. *In addition to* A1–A4, *assume a KL minimizer F exists in the interior of $\mathbb{F}$. Then $a_n K_n^* \to 0$ a.s.*

Proof. Multiply through (4.6) by a_n to get

$$\begin{aligned} \mathsf{E}(a_n K_n^* \mid \mathscr{A}_{n-1}) &= a_n \left[K_{n-1}^* + w_n^2 \mathsf{E}(Z_n|\mathscr{A}_{n-1}) - w_n T(F_{n-1})\right] \\ &= a_{n-1}K_{n-1}^* + w_n K_{n-1}^* + a_n w_n^2 \mathsf{E}(Z_n|\mathscr{A}_{n-1}) - a_n w_n T(F_{n-1}) \end{aligned}$$

This last line is also of the almost supermartingale form (2.1), with $\beta_n \equiv 0$, $\zeta_{n-1} = a_n w_n T(F_{n-1})$, and

$$\xi_{n-1} = a_n w_n^2 \mathsf{E}(Z_n|\mathscr{A}_{n-1}) + w_n K_{n-1}^*.$$

Since $\sum_n \xi_n < \infty$ a.s. by Lemmas 4.3–4.4, it follows from Theorem 2.1 that $a_n K_n^* \to K_\infty^*$ a.s., and $\sum_n a_n w_n T(F_{n-1}) < \infty$ a.s. To show that $K_\infty^* = 0$ a.s., proceed by contradiction as in the proof of Theorem 4.5. □

Remark 4.9. The extra condition that the KL minimizer F sits inside $\mathbb{F}$ can be viewed as an assumption about the quality of the model. That is, F should be inside $\mathbb{F}$ unless the mixture model $\mathbb{M}$ is "too bad." This notion of model quality is not yet fully understood, so sufficient conditions are currently not available. However, an example of a "bad" model is one where $m(x) = p(x|\theta)$ for some θ, which amounts to a mis-specification of the dominating measure μ. We suspect that the conclusion of Theorem 4.8 holds without this extra condition—see Section 4.3—but our proof hinges on Lemma 4.4, which we are unable to prove unless the KL minimizer F has a density f with respect to μ.

In the mis-specified case, even though $K_n^* \to 0$ implies $\|m_n - m_F\| \to 0$, the $L_1(\nu)$ rate does not easily follow without extra assumptions, such as a.e. boundedness of m_F/m. But a Hellinger contrast rate is a direct consequence of Theorem 4.8. In the well-specified case, when $m = m_F$, the Hellinger contrast reduces to the usual Hellinger distance, so our convergence rate results are comparable to those of, say, Genovese and Wasserman [6].

Corollary 4.10. *Choose* $w_n \asymp n^{-\gamma}$, $\gamma \in (2/3, 1]$, *and assume* A1–A4.

(a) $\rho(m_n, m_F) = o(n^{-(1-\gamma)/2})$ *a.s.*
(b) *If* $m_F/m \in L_\infty(m\,d\nu)$, *then* $\|m_n - m_F\| = o(n^{-(1-\gamma)/2})$ *a.s.*

Proof. Set $\Gamma_n = \int (m_n/m_F) m\,d\nu$; then $\Gamma_n \leq 1$ by Lemma 3.3. Now part (a) follows from Lemma 2.4 of Patilea [20]. Indeed,

$$\begin{aligned}
2\rho^2(m_n, m_F) &= \int \left(\sqrt{\frac{m_n}{m_F}} - 1\right)^2 m\,d\nu \\
&= (\Gamma_n - 1) + 2\int \left(1 - \sqrt{\frac{m_n}{m_F}}\right) m\,d\nu \\
&\leq 2\int \log\sqrt{\frac{m_F}{m_n}}\, m\,d\nu \\
&= K_n^*
\end{aligned}$$

For part (b), let $q_n = m m_n/m_F\Gamma_n$, and notice that $K_n^* \geq K(m, q_n)$; see Barron [1, Theorem 3]. Then, Pinsker's inequality [9, Theorem 6.1] gives

$$\sqrt{2K_n^*} \geq \|m - q_n\| = \frac{1}{\Gamma_n}\left\|\frac{m_n}{m_F} - \Gamma_n\right\|_{L_1(m\,d\nu)} \geq \left\|\frac{m_n}{m_F} - \Gamma_n\right\|_{L_1(m\,d\nu)}.$$

The triangle inequality for $\|\cdot\|_{L_1(m\,d\nu)}$ implies

$$\left\|\frac{m_n}{m_F} - 1\right\|_{L_1(m\,d\nu)} \leq \left\|\frac{m_n}{m_F} - \Gamma_n\right\|_{L_1(m\,d\nu)} + (1 - \Gamma_n),$$

and hence

$$\sqrt{2K_n^*} + (1 - \Gamma_n) \geq \left\|\frac{m_n}{m_F} - 1\right\|_{L_1(m\,d\nu)}.$$

The right-hand side is related to the $L_1(\nu)$ error via Holder's inequality

$$\begin{aligned}\|m_n - m_F\| &= \int |m_n - m_F|\,d\nu \\ &= \int \frac{m_F}{m}\left|\frac{m_n}{m_F} - 1\right| m\,d\nu \le C\left\|\frac{m_n}{m_F} - 1\right\|_{L_1(m\,d\nu)}\end{aligned}$$

where $C := \|m_F/m\|_{L_\infty(m\,d\nu)}$ is finite by assumption. Therefore,

$$\|m_n - m_F\| \le C\left\{\sqrt{2K_n^*} + (1-\Gamma_n)\right\},$$

and part (b) follows from the fact that $(1-\Gamma_n) \le K_n^*$. □

4.3. Numerical illustrations

We present a brief simulation study to highlight an example where our bound on rates appears sharp. Let $p(\cdot|\theta)$ be a $\mathrm{N}(\theta, 0.1^2)$ density, with $\Theta = [0,1]$ and μ as Lebesgue measure on Θ. We simulate data $X_1, \ldots, X_n$ from $\mathrm{N}(0.5, 0.1^2)$, which equals the mixture m_F where $F \in \overline{\mathbb{F}}$ is the point mass at 0.5. We consider weight sequences of the form $w_i = (i+1)^{-\gamma}$ for $\gamma \in \{0.5, 0.6, 0.67, 0.7, 0.75, 0.8, 0.9, 1.0\}$. For each choice of γ, the KL distance $K_n = K(m, m_n)$ is computed for 500 simulated data sets of size n, for each $n \in \{10^3, 10^4, 10^5, 10^6\}$. In order to estimate the empirical rate of convergence, set $L_n = -\log_{10} K_n$ and consider the following linear models:

$$\text{Model 1:} \quad \mathsf{E}(L_n) = \beta_0 + \beta_1 \log_{10} n \tag{4.8}$$

$$\text{Model 2:} \quad \mathsf{E}(L_n) = \beta_0 + \beta_1 \log_{10} n + \beta_2 \log_{10}\log_{10} n \tag{4.9}$$

Figure 1 shows K_n, averaged over the 500 replications, against n for each choice of γ. Table 1 shows the estimated coefficients of the linear models. In either model we would expect β_1 to be close to $1-\gamma$ had our upper bound been sharp. This indeed appears to be the case, particularly for γ not too close to 1. Also, the two estimates of β_1 sandwich $1-\gamma$, perhaps indicating that the actual rate is $n^{-(1-\gamma)}$ modulo a factor of $\log n$.

We should point out that this example does not exactly satisfy the assumptions of Theorem 4.8. Indeed, the KL minimizer $F = \delta_{\{0.5\}}$ lies on the boundary of $\mathbb{F}$ and not in its interior. However, we interpret this as the limit of examples where the optimal F gets increasingly close to the boundary, and this limit-based argument points to a minimax-type sharpness of the derived bounds. On the other hand, in examples where F is in the interior of $\mathbb{F}$, simulation studies have shown convergence rates faster than $n^{1-\gamma}$. For example, when F is a Unif(Θ) distribution and $\gamma = 0.75$, a fit of Model 1 gives $\hat{\beta}_1 = 0.73$, and a fit of Model 2 gives $\hat{\beta}_1 = 0.53$ and $\hat{\beta}_2 = 1.98$. These empirical results suggest that a nearly parametric rate of convergence, like $(\log n)^k/\sqrt{n}$, may be attainable in some cases.

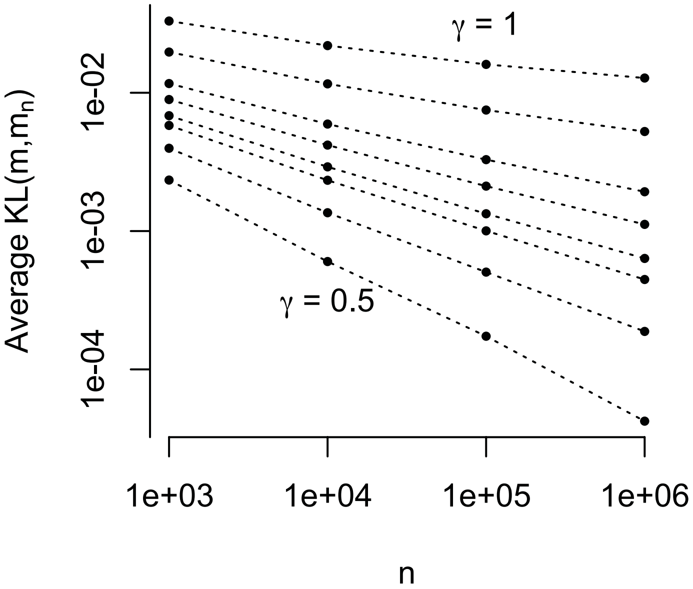

Fig 1. *KL divergence between m and m_n, averaged over 500 data sets of size n, against n for 8 choices of the weight sequence $w_i = (i+1)^{-\gamma}$, $\gamma \in \{0.5, 0.6, 0.67, 0.7, 0.75, 0.8, 0.9, 1\}$. Both axes are on the log scale. The average KL distances maintain the same ordering as of the γ values, the bottom curve is for $\gamma = 0.5$ and the top one is for $\gamma = 1$.*

Table 1

Estimated coefficients for Models 1 and 2 (4.8–4.9). Standard errors for $\hat{\beta}_1$ are in $(1,2) \times 10^{-3}$ and $(1,2) \times 10^{-2}$ for Models 1 and 2, respectively, and standard errors for $\hat{\beta}_2$ are in $(1,2) \times 10^{-1}$. Our upper bound suggests β_1 should be close to $1-\gamma$. It is larger than $1-\gamma$ for Model 1 and smaller for Model 2 with the addition of a $\log\log n$ term. This indicates that the actual rate is unlikely to be faster than $n^{-(1-\gamma)}$ by a power of n

	Model 1		Model 2		
γ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.5	0.97	0.56	1.14	0.66	−0.91
0.6	1.12	0.43	1.01	0.38	0.59
0.67	1.16	0.37	1.00	0.29	0.81
0.7	1.17	0.34	1.00	0.25	0.88
0.75	1.18	0.30	0.98	0.19	1.03
0.8	1.19	0.26	0.95	0.14	1.22
0.9	1.17	0.19	0.88	0.03	1.55
1.0	1.11	0.13	0.79	−0.03	1.68

5. A generalized PR algorithm

To satisfy the conditions of Theorem 4.8, one typically needs the mixing parameter space Θ to be compact. Also, this Θ must be *known* in practice since computation requires integration over Θ in (1.2). These requirements can be somewhat restrictive, particularly when there is no natural choice of Θ. A potential solution is to use a *mixture sieve*, which allows the support of the estimated mixing distribution to grow with the sample size. A motivation for this dynamic choice of support would be that, eventually, the support will be large enough so that the class of all mixtures over that support will be sufficiently rich. Borrowing on this idea, we propose a sieve-like extension of the PR algorithm which, instead of requiring Θ to be fixed and known, incorporates a sequence of compact mixing parameter spaces that increases with n.

Let Θ denote the mixing parameter space, which may or may not be compact. For example, if the model is a Gaussian location-scale mixture, then $\Theta = \mathbb{R}\times\mathbb{R}_+$. The generalized PR algorithm, in terms of densities, is as follows.

Algorithm GPR. Choose an increasing sequence of compact sets Θ_n such that $\Theta_n \uparrow \Theta$, a bounded, strictly positive, μ-measurable function $g(\theta)$, and a sequence $c_n \geq 0$ that satisfies $\sum_n \log(1+c_n) < \infty$. Define

$$g_n(\theta) = g(\theta) I_{\Theta_n \setminus \Theta_{n-1}}(\theta)/d_n,$$

where $d_n = \int_{\Theta_n \setminus \Theta_{n-1}} g \, d\mu$ is the normalizing constant. Start with an initial estimate f_0 on Θ_0 and, for $n \geq 1$, define

$$f_n^*(\theta) = (1-w_n) f_{n-1}(\theta) + w_n \frac{p(X_n|\theta) f_{n-1}(\theta)}{m_{n-1}(X_n)}, \quad \theta \in \Theta_{n-1}$$

and then

$$f_n(\theta) = \begin{cases} \frac{1}{1+c_n} f_n^*(\theta), & \theta \in \Theta_{n-1} \\ \frac{c_n}{1+c_n} g_n(\theta), & \theta \in \Theta_n \setminus \Theta_{n-1} \\ 0, & \theta \in \Theta_n^c \end{cases} \tag{5.1}$$

As in (1.3), define $m_n := m_{f_n}$ as the final estimate of m.

Our motivation for using g_n is that if Θ_n is small compared to the unknown Θ, the estimate should be padded near the boundary of Θ_n to compensate for the possibly heavy tails of f assigning non-negligible mass to $\Theta \setminus \Theta_n$. The GPR algorithm requires specification of two parameters, namely, the weight sequence c_n and the function g. Simple choices are $c_n = w_n^2$ and $g(\theta) \equiv 1$, but the practical performance of GPR for these or other choices has yet to be studied.

We now consider convergence of the GPR estimate m_n. The primary obstacle in extending the results in Section 4 is that now the support of the mixing distributions is changing with n—this makes the comparisons of the mixing densities in the proof of Lemma 4.4 more difficult. Here we will consider only the case where $m = m_f$ for some mixing density f, but both f and its support $\Theta_f \subset \Theta$ are unknown. Theorem 5.1 below establishes a bound on the rate of convergence in the case where Θ_f is a compact subset of Θ. Note that, while we are restricting ourselves to the compact case, *we do not assume* Θ_f *is known.*

Recall condition A4 in Section 4 requiring that the likelihood ratio be square integrable uniformly over Θ. In this case, we have a sequence Θ_n, and we require a bound similar to that of A4 for each Θ_n. To this end, define the sequence

$$B_n = \sup_{\theta_1,\theta_2,\theta_3 \in \Theta_n} \int \left[\frac{p(x|\theta_1)}{p(x|\theta_2)}\right]^2 p(x|\theta_3)\, d\nu(x), \quad n \geq 0. \tag{5.2}$$

Since $\Theta_n \subset \Theta_{n+1}$, the sequence B_n is clearly increasing; if we are to push the proof of Theorem 4.8 through in this more general situation, we will need to control how fast B_n increases.

Theorem 5.1. *Assume that* $\Theta_f \subset \Theta$ *is compact and that conditions* A2–A3 *hold. Furthermore, assume that* $\sum_n a_n w_n^2 B_n < \infty$. *Then the* GPR *estimates* m_n *satisfy* $a_n K(m_f, m_n) \to 0$ *a.s.*

The proof is essentially the same as that of Theorem 4.8—we simply need to check that Lemmas 4.3 and 4.4 continue to hold in this new context. The details are provided in the Appendix.

Remark 5.2. The growth rate condition $\sum_n a_{n-1} w_n^2 B_n < \infty$ in Theorem 5.1 clearly holds provided that $\{w_n\}$ and $\{B_n\}$ satisfy $w_n \asymp (n^\alpha \log n)^{-1}$ for some $\alpha \in [2/3, 1]$, and $B_n = O(\log n)$.

Example 5.3. Let $p(x|\theta) = e^{-(x-\theta)^2/2}/\sqrt{2\pi}$ be a $N(\theta, 1)$ density (with respect to Lebesgue measure). Then

$$\int \left[\frac{p(x|\theta_1)}{p(x|\theta_2)}\right]^2 p(x|\theta_3)\, dx = \exp\{\theta_2^2 - \theta_1^2 + 2\theta_3(\theta_1 - \theta_2) + 2(\theta_1 - \theta_2)^2\}.$$

If we let $\Theta_n = [-t_n, t_n]$, then

$$B_n = \sup_{\theta_1, \theta_2, \theta_3 \in \Theta_n} \exp\{\theta_2^2 - \theta_1^2 + 2\theta_3(\theta_1 - \theta_2) + 2(\theta_1 - \theta_2)^2\} = \exp\{12 t_n^2\}.$$

If we choose $t_n \asymp (c + \frac{1}{12} \log\log n)^{1/2}$, for some constant $c > 0$, then $B_n = O(\log n)$. Taking w_n as in Remark 5.2 satisfies the conditions of the theorem.

Example 5.4. Suppose that $p(x|\theta) = e^{-\theta}\theta^x/x!$ is a Poisson density (with respect to counting measure). Then

$$\sum_x \left[\frac{p(x|\theta_1)}{p(x|\theta_2)}\right]^2 p(x|\theta_3) = \exp\{2(\theta_2 - \theta_1) - \theta_3 + \theta_1^2\theta_3/\theta_2^2\}.$$

Take $\Theta_n = [\alpha_n, \beta_n]$ where α_n and β_n are to be determined. Then,

$$B_n = \sup_{\theta_1, \theta_2, \theta_3 \in \Theta_n} \exp\{2(\theta_2 - \theta_1) - \theta_3 + \theta_1^2\theta_3/\theta_2^2\} \le \exp\{\beta_n^3/\alpha_n^2\}.$$

If $\beta_n \asymp (c + \log\log n)^{1/5}$, for some constant $c > 0$ such that Θ_0 suitably large, and $\alpha_n = \beta_n^{-1}$, then it is easy to check that $B_n = O(\log n)$. Therefore, the conditions of Theorem 5.1 are satisfied with w_n as in Remark 5.2.

6. Discussion

PR is an exciting stochastic algorithm for mixture models that is quite different from EM or MCMC in its focus and structure. While MCMC and EM focus exclusively on the mixture distribution, PR brings the mixing density to the fore.

Structurally, PR is not a hill-clmbing algorithm like EM or MCMC; rather, it draws inspiration from the recursive aggregation idea of stochastic approximation (SA). Martin and Ghosh [16] have established a precise connection between PR and SA for finite-dimensional $\mathbb{F}$, and this connection is likely to extend to the infinite-dimensional case as well. Interestingly, very little is known about convergence properties of general infinite-dimensional SA algorithms, even though their finite-dimensional counterparts are well understood. In this regard, our convergence results here can make a major contribution to the study of SA in general.

In this paper we have extended our recent work on asymptotic analysis of PR in two key directions. First, we have shown that PR is robust to mixture model mis-specification, in the sense that the PR estimate m_n converges to the mixture m_F which is closest in KL divergence to the true density m. This property is important since, typically, m itself is not a mixture, but is closely approximated by one. Second, we have established a bound on the PR rate of convergence, which makes a direct connection between the choice of weight sequence w_n and the performance of the PR estimates. We suspect that the extra condition needed for the rate in Theorem 4.8—namely, that a KL minimizer F sit inside $\mathbb{F}$—can be relaxed, but more work is needed. Simulation results presented in Section 4.3 reveal two interesting observations: (i) the bound on the rate is of a minimax nature, and (ii) the best (minimax) rates, when $w_n \asymp n^{-\gamma}$, are achieved for γ near 0.5. Further investigations are needed to better understand in what sense the rate is minimax, to extend Theorem 4.8 to the case when $\gamma \approx 0.5$, and to characterize the rate in a typical "non-minimax" problem.

We have also proposed a practical extension of the PR algorithm which does not require that the mixing distribution have a known compact support. Instead, GPR uses an increasing sieve-like sequence of increasing compact supports. Sufficient conditions are given—which essentially control the growth rate of the sieves—that guarantee consistency of the estimated mixture and bound the rate of convergence. Here, however, the growth of the sieve space is rather slow (see Examples 5.3–5.4), so the advantages of a dynamic support over a large fixed support may be difficult to see in finite samples. We suspect that the extension of Theorem 4.8 to handle γ near 0.5 can be used here to relax the growth rate conditions and allow the sieve space to grow more rapidly.

Acknowledgments

The authors thank the Associate Editor and referee for their helpful comments and suggestions. Also, special thanks go to Professor J. K. Ghosh for his support and guidance throughout this and other projects.

Appendix A: Proofs

Proof of Lemma 4.1. For part (a), treat θ as a random element in Θ, with distribution $\Phi \in \mathbb{F}$, and define the random variable

$$g_\Phi(\theta) = \int \frac{m(x)}{m_\Phi(x)} p(x|\theta)\, d\nu(x). \tag{A.1}$$

Then $\mathsf{E}_\Phi\{g_\Phi(\theta)\} = \int g_\Phi \, d\Phi = 1$ and $T(\Phi) = \mathsf{V}_\Phi\{g_\Phi(\theta)\} \geq 0$, with equality if and only if $g_\Phi = 1$ μ-a.e. Next define

$$G(\Phi) = \log\left\{\int_\Theta g_\Phi(\theta)\, dF(\theta)\right\} = \log\left\{\int_\mathcal{X} \frac{m_F}{m_\Phi} m\, d\nu\right\},$$

where $F \in \overline{\mathbb{F}}$ is such that $K(m, m_F) = K(m, \overline{\mathbb{M}})$. Note that $T(\Phi) = 0$ implies $G(\Phi) = 0$. By Jensen's inequality

$$G(\Phi) \geq \int_\mathcal{X} \log\left(\frac{m_F}{m_\Phi}\right) m\, d\nu = K(m, m_\Phi) - K(m, m_F) \geq 0,$$

so that $G(\Phi) = 0$ implies $K(m, m_\Phi) = K(m, \overline{\mathbb{M}})$.

For part (b), take a sequence $\{\Phi_s\} \subset \mathbb{F}$ such that Φ_s converges weakly to some $\Phi \in \overline{\mathbb{F}}$, and let g_{Φ_s} and g_Φ be as in (A.1). Let $r_s(x, \theta) = p(x|\theta)/m_{\Phi_s}(x)$ so that $g_{\Phi_s}(\theta) = \int r_s(x, \theta) m(x)\, d\nu(x)$. By A4 and Jensen's inequality,

$$\sup_s \int r_s^2(x, \theta) m(x)\, d\nu(x) \leq B,$$

which implies $\{r_s(\cdot, \theta)\}$ is uniformly integrable with respect to $m\, d\nu$; therefore, $g_{\Phi_s} \to g_\Phi$ μ-a.e. Since the weak topology on $\mathbb{F}$ is metrizable, to prove continuity of T it suffices to show that $T(\Phi_s) \to T(\Phi)$. We have

$$\begin{aligned} |T(\Phi_s) - T(\Phi)| &= \left|\int g_{\Phi_s}^2\, d\Phi_s - \int g_\Phi^2\, d\Phi\right| \\ &\leq \left|\int (g_{\Phi_s}^2 - g_\Phi^2)\, d\Phi_s\right| + \left|\int g_\Phi^2\, d(\Phi_s - \Phi)\right| \end{aligned} \tag{A.2}$$

The second term on the right-hand side of (A.2) goes to zero by definition of weak convergence, since $g_\Phi^2 \leq B$ (by A4) and g_Φ^2 is continuous (by A1). We also know the following:

- $|g_{\Phi_s}^2 - g_\Phi^2| \to 0$ μ-a.e.,
- $|g_{\Phi_s}^2 - g_\Phi^2| \leq 2B$, and
- $\lim_{s\to\infty} \int d\Phi_s = 1 = \int d\Phi$.

Now, for the first term on the right-hand side of (A.2),

$$\left|\int (g_{\Phi_s}^2 - g_\Phi^2)\, d\Phi_s\right| \leq \int |g_{\Phi_s}^2 - g_\Phi^2|\, d\Phi_s \to 0,$$

where convergence follows from the above three properties and Theorem 1 of Pratt [21]. Therefore, $T(\Phi_s) \to T(\Phi)$ and since the sequence $\{\Phi_s\}$ was arbitrary, T must be continuous. □

Proof of Lemma 4.3. Note that for $a > 0$ and $b \in (0,1)$, we have

$$(a-1)^2 \max\{1, (1+b(a-1))^{-2}\} \leq \max\{(a-1)^2, (1/a-1)^2\}. \tag{A.3}$$

Combining inequalities (4.1) and (A.3) we see that

$$H_{n,X_n}^2 R(w_n H_{n,X_n}) \leq \max\left\{\left(\frac{h_{n,X_n}}{m_{n-1}} - 1\right)^2, \left(\frac{m_{n-1}}{h_{n,X_n}} - 1\right)^2\right\}$$

and, since both h_{n,X_n} and m_{n-1} belong to $\mathbb{M}$ for each n, we conclude from A4 and Jensen's inequality that $Z_n \leq 1 + B$ a.s. □

Proof of Lemma 4.4. Since $F \in \mathbb{F}$, the KL divergence of F_n from F is well-defined. Following the argument of TMG (p. 2505), it is possible to write a recursion for $K(F, F_n)$ as we did in (4.2). In particular,

$$\mathsf{E}\{K(F, F_n) \mid \mathscr{A}_{n-1}\} = K(F, F_{n-1}) - w_n D(F_{n-1}) + w_n^2 \mathsf{E}(Y_n | \mathscr{A}_{n-1}),$$

where Y_n, like Z_n, is uniformly bounded by $B+1$, and the functional $D(\cdot)$ is given by

$$D(\Phi) = \int \left(\frac{m_F}{m_\Phi} - 1\right) m \, d\nu.$$

It follows from Jensen's inequality that

$$D(\Phi) \geq \int \log \frac{m_F}{m_\Phi} m \, d\nu = K(m, m_\Phi) - K(m, m_F) \geq 0, \tag{A.4}$$

with equality iff $m_\Phi = m_F$. This is an almost supermartingale, so we conclude from Theorem 2.1 that $K(F, F_n)$ converges a.s. and $\sum_n w_n D(F_{n-1}) < \infty$ a.s. But in light of (A.4), we have $\sum_n w_n K_{n-1}^* < \infty$, the desired result. □

Proof of Theorem 5.1. Since B_n is increasing, the condition $\sum_n a_n w_n^2 B_n < \infty$ implies that the conclusion of Lemma 4.3 holds, and the remainder term Z_n in (4.5) satisfies $\sum_n w_n^2 \mathsf{E}(Z_n) < \infty$. The key observation is that, since Θ_f is compact, there is a number $N = N(f)$ such that $\Theta_f \subset \Theta_n$ for all $n \geq N$. Consequently, there are only N iterates $f_0, \ldots, f_{N-1}$ such that $K(f, f_i) = \infty$. Without loss of generality, we can shift the "time scale" so that $N = 0$. Now apply the same expansion as in the proof of Lemma 4.4 to get

$$\begin{aligned}\mathsf{E}\{K(f, f_n) \mid \mathscr{A}_{n-1}\} - K(f, f_{n-1}) &= -\int f \log(f_n/f_{n-1}) \, d\mu \\ &= -w_n D(f_{n-1}) + w_n^2 \mathsf{E}(Y_n | \mathscr{A}_{n-1}) + \log(1 + c_n)\end{aligned}$$

This is of the almost supermartingale form (2.1), and both $\sum_n w_n^2 \mathsf{E}(Y_n | \mathscr{A}_{n-1})$ and $\sum_n \log(1 + c_n)$ are finite. Therefore, Theorem 2.1 implies $K(f, f_n)$ converges, and $\sum_n w_n D(f_{n-1}) < \infty$. From this and (A.4) we can conclude that $\sum_n w_n K(m_f, m_n) < \infty$ a.s. To finish the proof, simply throw away the first N iterations and apply the argument used to prove Theorem 4.8. □

References

[1] Barron, A. R. (2000). Limits of information, Markov chains, and projection. In *IEEE International Symposium on Information Theory* 25.

[2] Billingsley, P. (1995). *Probability and measure*, Third ed. John Wiley & Sons Inc., New York. MR1324786

[3] Bogdan, M., Ghosh, J. K. and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg prodecure with some Bayesian Rules for Multiple Testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (N. Balakrishnan, E. Peña and M. Silvapulle, eds.) 211–230. IMS, Beachwood, OH. MR2462208

[4] Brown, L. D., George, E. I. and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. MR2418653

[5] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

[6] Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921

[7] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

[8] Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton's algorithm for estimating mixing distribution. In *Frontiers in statistics* 429–443. Imp. Coll. Press, London. MR2326012

[9] Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *Ann. Math. Statist.* **40** 2156–2177. MR0252112

[10] Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395

[11] Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31** 391–406. MR1983535

[12] Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360. MR1186253

[13] Liese, F. and Vajda, I. (1987). *Convex statistical distances.* Teubner, Leipzig.

[14] Li, J. Q. and Barron, A. R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems* (S. Solla, T. Leen and K.-R. Mueller, eds.) 279–285. MIT Press, Cambridge, Massachusetts.

[15] Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–94. MR684866

[16] Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statist. Sci.* **23** 365–382. MR2483909

[17] Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhyā Ser. A* **64** 306–322. MR1981761

[18] Newton, M. A., Quintana, F. A. and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical nonparametric*

and semiparametric Bayesian statistics, **133** 45–61. Springer, New York. MR1630075

[19] Newton, M. A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26. MR1688068

[20] Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123. MR1833960

[21] Pratt, J. W. (1960). On interchanging limits and integrals. *Ann. Math. Statist.* **31** 74–77. MR0123673

[22] Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)* 233–257. Academic Press, New York. MR0343355

[23] Tokdar, S. T., Martin, R. and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.* **37** 2502–2522. MR2543700