This is to certify that the

dissertation entitled

Asymptotically Optimal Bayes Compound and
Empirical Bayes Estimators in Exponential
Families with Compact Parameter Space

presented by

Somnath Datta

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

James Hannan
Major professor

Date July 25, 1988

ASYMPTOTICALLY OPTIMAL BAYES COMPOUND AND EMPIRICAL BAYES ESTIMATORS IN EXPONENTIAL FAMILIES WITH COMPACT PARAMETER SPACE

By

Somnath Datta

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1988

ABSTRACT

ASYMPTOTICALLY OPTIMAL BAYES COMPOUND AND EMPIRICAL BAYES ESTIMATORS IN EXPONENTIAL FAMILIES WITH COMPACT PARAMETER SPACE

By

Somnath Datta

The problem of finding admissible, asymptotically optimal compound and empirical Bayes rules is pursued in the infinite state case.

The component distributions considered in this work form a real exponential family of quite general nature with component parameter in a compact interval of the natural parameter space. The component problem estimates an arbitrary continuous transform of the natural parameter under squared error loss.

We consider the set, sequence compound and the empirical Bayes formulations of the above component and show that all Bayes estimators in the various formulations are admissible. Our main result is that any Bayes compound estimator versus a mixture of i.i.d. priors on the compound parameter is asymptotically optimal if the mixing hyperprior has full support. Analogously any Bayes empirical Bayes estimator is asymptotically optimal if the empirical Bayes prior has full support.

The exponential family structure has been used to treat the difference in risks of the Bayes estimators and the component Bayes versus the empiric for some special cases of the continuous transforms. The key to the proof of asymptotic optimality is an L_1 consistency of posterior mixtures, itself a major finding of the thesis and extendible far beyond the exponential context.

The thesis also derives an interesting uniform L_1 LLN for random continuous functions on a compact metric space which is applied in the proof of the last result.

The asymptotic optimality results are generalized to weighted squared error loss with continuous weight function and applications to some non–exponential situations are also considered.

Several examples of such hyperpriors/empirical Bayes priors are given and for some of them practically useful forms of the corresponding Bayes estimators are obtained.

To my parents and my wife

ACKNOWLEDGMENTS

I wish to express my deepest appreciation to my thesis adviser Professor James F. Hannan for everything he has done for me. His invaluable suggestions led the way to some results and continuous careful criticism greatly improved the overall presentation of this thesis.

I like to express my thanks to my other committee members Professors Hira L. Koul, Dennis C. Gilliland and Clifford Weil for their suggestions on an earlier draft.

Special thanks go to my wife Susmita for her total support and encouragement during the preparation of this thesis.

Finally, I would like to thank Ms. Loretta Ferguson for numerous helps during my typing of the manuscript.

TABLE OF CONTENTS

Chapter Page

CHAPTER 0

INTRODUCTION

We start with some notational conventions used throughout the body of the thesis. Let n be a positive integer. An n–vector $(x_1, \ldots, x_n)$ is denoted by $\underline{x}$ and for $1 \leq \alpha \leq n$, $(x_1, \ldots, x_\alpha)$ is denoted by $\underline{x}_\alpha$. For probabilities $P_1, \ldots, P_n$, $\overset{n}{\underset{\alpha=1}{\times}} P_\alpha$ denotes their measure theoretic product. Typically the letter P is used for probabilities and E for the corresponding expectations. For the sake of clarification, dummy variables are often displayed in integrals. Also mixed mode integral expressions like $\int X(\omega)dP$ are used. For a bounded function f, f_* and f^* denote its infimum and supremum respectively over its entire domain. For a measure m on the Borel σ–field of a topological space $\mathcal{T}$ the support of m is defined to be the set $\cap\{F \subset \mathcal{T}: F$ is closed and $m(F^c) = 0\}$. Note that the support of $m = \mathcal{T}$ iff, $\forall$ open $\phi \neq O \subset \mathcal{T}$, $m(O) > 0$. $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ stand for the set of reals, integers and non–negative integers respectively.

1. The component problem.

The component problem has the structure of the usual decision theory problem, i.e. we have a parameter space Θ, a family of probability measure $\{P_\theta : \theta \in \Theta\}$ on some common measurable space $\mathcal{X}$, an observable $\mathcal{X}$–valued random variable $X \sim P_\theta$ under θ, an action space $\mathcal{A}$, a loss function $L: \mathcal{A} \times \Theta \longrightarrow [0,\infty)$, decision rules t, $t: \mathcal{X} \longrightarrow \mathcal{A}$ such that $L(t,\theta)$ is measurable for each θ, with risk $R(t,\theta) = E_\theta \, L(t,\theta)$.

2. The set compound problem.

The set compound problem simultaneously considers a number, say n, of independent decision problems each of which is structurally identical to the above component problem, and allows the use of observations from all the problems in each of the decisions. The compound loss is taken to be the average of all the component losses.

Thus for each $n \geq 1$, the set compound problem can be formulated as a decision problem as follows. We have the parameter space Θ^n, the action space $\mathscr{A}^n$, observations $\underline{X} = (X_1, \ldots, X_n) \sim P_{\underline{\theta}} = \overset{n}{\underset{\alpha=1}{\times}} P_{\theta_\alpha}$, $\underline{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta^n$, compound rules $\underline{t} = (t_1, \ldots, t_n)$, where for each $1 \leq \alpha \leq n$ $t_\alpha : \mathscr{X}^n \longrightarrow \mathscr{A}$ such that $L(t_\alpha, \theta)$ is measurable for each θ, with loss $L_n(\underline{t}, \underline{\theta}) = n^{-1} \sum_1^n L(t_\alpha, \theta_\alpha)$ and risk

$$R_n(\underline{t}, \underline{\theta}) = E_{\underline{\theta}} L_n(\underline{t}, \underline{\theta}). \tag{2.1}$$

Let $\Omega = \{\omega : \omega \text{ is a probability on } \Theta\}$. For $\omega \in \Omega$, let $R(\omega)$ stand for the minimum Bayes risk versus ω in the component problem, i.e.

$$R(\omega) = \bigwedge_t \int R(t, \theta) d\omega(\theta).$$

For a traditional simple symmetric rule (i.e. $t_\alpha(\underline{x}) = t(x_\alpha)$ $\forall$ $1 \leq \alpha \leq n$ for some component rule t) the compound risk is easily seen to be at least $R(G_n)$, G_n being the empirical distribution of $\theta_1, \ldots, \theta_n$. In all non-trivial situations a component Bayes rule versus G_n is unavailable to the statistician because G_n is unknown and hence $R(G_n)$ cannot be achieved (for any n) via the use of a simple symmetric rule. Thus compound rules which attains risks asymptotically no more than $R(G_n)$ are of interest. Hannan (1957) used the term 'approximation to Bayes risk' to describe such effects.

For a compound rule $\underline{t}$, the difference $D_n(\underline{t},\underline{\theta}) = R_n(\underline{t},\underline{\theta}) - R(G_n)$ is called the modified regret of $\underline{t}$ at $\underline{\theta}$. We say that a rule $\underline{t}$ is asymptotically optimal (a.o.) if

$$(2.2) \qquad \bigvee_{\underline{\theta}} D_n(\underline{t},\underline{\theta})_+ \longrightarrow 0 \quad \text{as } n \longrightarrow \infty .$$

For the relation of this notion of optimality to that with more stringent envelopes in the finite Θ case, see Gilliland and Hannan (1986).

A set compound rule $\underline{t}$ is said to be admissible if for each $n \geq 1$, $R_n(\underline{t},\underline{\theta})$ is admissible in the usual decision theoretic sense as a function of $\underline{\theta}$ in the class of set compound rules.

3. The sequence compound problem.

The sequence compound problem also considers a number, say n, of independent repetitions of a component problem but allows only data up to stage α in making the α–th decision for $1 \leq \alpha \leq n$. Thus a sequence compound rule is $\underline{t} = (t_1, \ldots, t_n)$, $t_\alpha: \mathscr{X}^\alpha \longrightarrow \mathscr{A}$ such that $L(t_\alpha,\theta)$ is measurable for each θ, with the interpretation that the α–th decision made with the use of $\underline{t}$ is $t_\alpha(\underline{X}_\alpha)$ for each $1 \leq \alpha \leq n$.

In the weak sequence compound (w.s.c.) version n is known to the statistician apriori so that each t_α, $1 \leq \alpha \leq n$ may use n. We will use $\underline{t}_n$ for a w.s.c. rule to show its possible dependency on n. The other version is called the strong sequence compound (s.s.c.) problem and is more interesting. In both versions we are interested in the asymptotic risk behavior of compound rules as n tends to infinity. Hence a w.s.c. rule can be viewed as a triangular array $\underline{t}_n$, $n \geq 1$ and an s.s.c. rule as a sequence $\underline{t} = (t_1, t_2, \ldots \;)$. It should be noted that an s.s.c. rule is automatically a w.s.c. rule.

The risk (up to stage n) and the modified regret of a sequence compound rule $\underline{t}$ (or $\underline{t}_n$) is as given by (2.1) and (2.2) (with the understanding that $\underline{t}$ is viewed as a function on $\mathscr{X}^n$ as $t_\alpha(\underline{x}) = t_\alpha(\underline{x}_\alpha)$ for $1 \leq \alpha \leq n$). The notion of asymptotic optimality remains the same.

A sequence compound rule $\underline{t}$ (or $\underline{t}_n$) is said to be admissible if, for each $n \geq 1$, $R_n(\underline{t},\underline{\theta})$ (or $R_n(\underline{t}_n,\underline{\theta})$) is admissible in the usual decision theoretic sense as a function of $\underline{\theta}$ in the class of sequence compound rules. [This is the natural definition for w.s.c. rules and therefore more demanding of s.s.c. rules.]

4. The empirical Bayes problem.

The empirical Bayes problem considers a sequence of independent, identical Bayes decision problems.

In this case the component problem is the same as that in Section 1 with the additional notion that the component parameter θ is a random element having a (prior) distribution ω on Θ. Thus the risk of a component rule t is

$$R(t,\omega) = \int R(t,\theta)d\omega.$$

The prior ω is unknown to the statistician even though it is believed to exists.

The empirical Bayes problem has generic point $\underline{\theta} = (\theta_1, \theta_2, \ldots\ )$ representing the true states and data $\underline{X} = (X_1, X_2, \ldots\ )$ from the problems and the assumption is that (θ_1,X_1), (θ_2,X_2), ... are i.i.d. copies of (θ,X) having distribution ω on θ and, conditional on θ, P_θ on X. Let E denote the overall expectation. At stage n, a decision $t_n(\underline{X}_n)$ about θ_n is taken by t_n: $\mathscr{X}^n \longrightarrow \mathscr{A}$ with loss $L(t_n,\theta_n)$, which is jointly measurable, and the risk incurred is $R_n(t_n,\omega) = E\,L(t_n,\theta_n) = \int\int L(t_n,\theta_n)\, dP_{\theta_n} d\omega^n$. We call $<t_n: n\geq 1>$ an empirical Bayes (e.B.) rule.

An e.B. rule $<t_n>$ is called asymptotically optimal if

$$\lim_{n\to\infty} R_n(t_n,\omega) = R(\omega), \quad \forall\ \omega \in \Omega.$$

An e.B. rule $<t_n>$ is said to be admissible if for each n, $R_n(t_n,\omega)$ is admissible in the usual decision theoretic sense as a function of ω in the class of possible t_n.

5. Literature review and a summary of the present work.

The pioneering paper of compound decision theory is by Robbins (1951). His featured example was decision between N(–1,1) and N(1,1). He exhibited an a.o. compound procedure and called it asymptotically subminimax by comparison with the simple symmetric minimax rule.

A.o. compound and empirical Bayes rules have been worked out for many choices of component problem. Typically they are bootstrap (or delete bootstrap) in nature, – rules whose components are Bayes versus some estimate of the unknown G_n (or ω in the e.B. case) or direct estimates of the Bayes rule versus G_n (or ω). In particular, when the component problem is an estimation problem under squared error loss, Gilliland (1968) and Singh (1974) obtained a.o.sequence compound rules with rates (we say $\underset{\sim}{t}$ is a.o. with rate α_n if $\bigvee_{\underset{\sim}{\theta}} D_n(\underset{\sim}{t},\underset{\sim}{\theta})_+ = o(\alpha_n)$) for discrete and Lebesgue exponential components respectively.

Though the above mentioned rules satisfy the criterion of asymptotic optimality, they are not very satisfactory as far as their finite n behaviors are concerned. In fact they turn out to be inadmissible where admissible is as defined in the previous sections.

Thus the problem of exhibiting compound and e.B. rules which are a.o. as well as admissible has been an interesting and challenging question ever since it was put forward by Robbins (1951) in the sense that he proposed the Bayes compound rule versus the symmetric prior uniform on proportions for his featured example and conjectured that it might have better risk behavior than his asymptotically subminimax rule. Inglis (1973) studied, i.a., the asymptotic optimality of a class of admissible Bayes rules for two state components under the finiteness of the expected log–densities and tacit (cf. Inglis (1977)) non–atomicity conditions for his "generalization" of the Hannan–Robbins theorem; cf. the addenda of the next two works. Gilliland and Hannan (1974, The finite state compound decision problem, equivariance and restricted risk components, RM 317, Department of Statistics and Probability, Michigan State University), which was later published in 1986, treated the more general problem of restricted risk components in the finite Θ case. They worked with a more stringent envelope and reduced the problem of asymptotic optimality to the problem of establishing the L_1 consistency of certain induced estimators. Gilliland, Hannan and Huang (1976) established that consistency, in two state components, for Bayes compound estimators versus certain symmetric priors including Robbins prior. This approach yielded for them admissible rules which are a.o. with rates as good as $O(n^{-1/2})$ in the general two state component case. Vardeman (1978) successfully exploited a result by the last authors to obtain admissible, a.o. sequence compound rules in the two state component case.

None of the results mentioned in the previous paragraph go beyond the finite Θ case. Meeden (1972) obtained admissible, a.o. empirical Bayes rule in two special infinite state examples, where the component problems are (i) squared error loss estimation of Geometric parameter and (ii) linear loss testing of Poisson mean. Inglis (1973) attempted to prove the admissibility and asymptotic optimality of a class of Bayes compound estimators versus mixtures of i.i.d. priors in example (i) above with compact parameter space. Unfortunately his proof of asymptotic optimality appears to contain certain serious gaps. For a discussion on this see the addendum of Gilliland and Hannan (1986).

The present work, which subsumes Inglis's example, seems to be the first successful attempt in the literature to accomplish compound admissibility and asymptotic optimality simultaneously in non–finite state case. Our component distributions form a one dimensional exponential family of quite general nature, whose examples include well known exponential families such as Normal, Exponential, Geometric, Poisson and Negative Binomial, where the parameter space is any compact interval of the natural parameter space on which the first moment is finite. The component problem is to estimate an arbitrary continuous transform of the natural parameter under squared error loss. We note that all compound Bayes estimators in our set and sequence compound problems are admissible. Our main result is that those Bayes versus a mixture of i.i.d. priors on the compound parameter are a.o. if the mixing hyperprior has full support. In the empirical Bayes version of our problem, our conclusion is that all Bayes e.B. estimators are admissible and those versus a prior with full support are a.o.

In the set compound situation, for a dense class of continuous functions, the question of asymptotic optimality is reduced to the question of the L_1 consistency of a posterior mixture which itself is of independent interest. We make use of an inequality suggested in the addendum of Gilliland, Hannan and Huang (1976), develop and use an uniform L_1– LLN and the full support property of Λ to treat the resulting terms and prove the required consistency. The results in the sequence compound and empirical Bayes problems, to a large extent, follow from the compound results.

The thesis is organized as follows.

Chapter 1 treats the set compound problem described above. Section 2 obtains the admissibility of all Bayes estimators from Lemma A.4, describes the Bayes estimator versus the above mentioned prior and establishes a bound on its modified regret. Section 3 establishes a bound, in terms of the L_1 distance between the corresponding mixtures, on the L_1 distance between two component Bayes estimators of $\phi(\theta) = e^{\theta k}$ for $k \in \mathbb{N}$ and θ the natural parameter. Section 4 proves the consistency result as detailed in the previous paragraph and adapts it to the delete versions. Section 5 combines the results of Sections 2–4 and establishes the desired asymptotic optimality.

Chapter 2 and 3 considers the sequence compound and empirical Bayes formulations respectively and obtains similar admissibility and optimality results.

Chapter 4 contains various examples of hyperpriors having full support. In some cases practically useful forms of the Bayes estimators are obtained. To this end some possible generalizations are recorded.

Finally some results of possible independent interest in more general contexts, which include the aforementioned uniform L_1–LLN for random continuous functions on a compact metric space, are derived in the appendix. They are used in the body of the thesis.

CHAPTER 1

THE SET COMPOUND ESTIMATION

1. Introduction.

1.1. Notations and conventions.

In this chapter we consider the set compound problem as described in Chapter 0 corresponding to the component problem to be introduced in the next section. We assume that the reader is familiar with the notations and definitions of the general set compound problem from Chapter 0. The following additional notations and conventions will be used throughout this chapter.

We will use **P** for $P_{\underline{\theta}} = \times_{\alpha=1}^{n} P_{\theta_\alpha}$ and **E** for the corresponding expectation. Given any vector $\underline{u} = (u_1, ..., u_n)$, for each $1 \leq \alpha \leq n$, $\check{\underline{u}}_\alpha$ will mean the vector $(v_1, ..., v_{n-1})$ with $v_j = u_j$ for $j < \alpha$ and $= u_{j+1}$ for $j \geq \alpha$. Since $\mathscr{X} = \mathbb{R}$ in our problem (see the next section), we will view $X_1, ..., X_n$ to be the co-ordinate functions on $\mathbb{R}^n$. On the other hand, any measurable function H on $\mathbb{R}^n$ will be viewed as the random variable $H(\underline{X})$ whenever convenient.

The next section describes our component problem and records a few elementary but useful results related to the component distributions. A summary of the rest of the chapter was given in Section 5 of Chapter 0.

1.2. Exponential family component.

Our component problem is the following. $\Theta = [c,d]$, $-\infty < c < d < \infty$, $\mathscr{A} = \mathbb{R}$, $L(a,\theta) = (a - \phi(\theta))^2$ where ϕ is any real continuous function on Θ and $\forall\ \theta \in \Theta$, P_θ admits a density p_θ wrt a common σ–finite μ on $\mathbb{R}$

given by

$$(1.1) \qquad p_\theta(x) = e^{\theta x} h(\theta), \quad x \in \mathbb{R}.$$

In addition μ is assumed to satisfy $\mu_1 \ll \mu$, where $\mu_k = \mu s_k^{-1}$, $s_k(x) = x + k$, $x \in \mathbb{R}$; $k \in \mathbb{Z}$.

Clearly $c, d \in \tilde{\Theta} = \{\theta : \int e^{\theta x} d\mu(x) < \infty\}$. Throughout we will assume

$$(1.2) \qquad \int x e^{cx} d\mu(x) < \infty, \quad \int x e^{dx} d\mu(x) < \infty.$$

It is well known (e.g. Lehmann (1959, 1986), Theorem 2.9) that(1.2) holds if $c, d \in$ interior of $\tilde{\Theta}$.

For $\omega \in \Omega$, let P_ω denote the ω-mix of P_θ's and p_ω denote its μ density, $p_\omega(x) = \int p_\theta(x) d\omega$, $x \in \mathbb{R}$.

The following consequences of our assumptions are worth noting and some will be used in the later sections.

C1. $h(\theta) = (\int e^{\theta x} d\mu(x))^{-1}$ and h is continuous and positive on compact Θ. Consequently, $0 < h_* \le h^* < \infty$.
[Indeed $h_* = h(c) \wedge h(d)$, since log h is concave by the Hölder inequality.]

C2. $\theta \rightsquigarrow \int x e^{\theta x} d\mu(x)$ is continuous on Θ.

[Since $e^{cx} \wedge e^{dx} \le e^{\theta x} \le e^{cx} \vee e^{dx}$ on Θ, the continuity of h and $\theta \rightsquigarrow \int x e^{\theta x} d\mu(x)$ including one sided continuity at the end points follow readily by the Dominated Convergence Theorem (D.C.T. hereafter).]

C3. For any $\theta \in \Theta$,

$$(1.3) \qquad p_\theta \leq (h^*/h_*)(p_c + p_d),$$

and hence any $f \in L_1(P_c) \cap L_1(P_d)$ is uniformly integrable wrt the family of probability measures $\{P_\theta , \theta \in \Theta\}$. In view of (1.2) the identity function is such.

C4. Since, for any $\theta \in \Theta$,

$$(1.4) \qquad h_*(e^{cx} \wedge e^{dx}) \leq p_\theta(x) \leq h^*(e^{cx} \vee e^{dx}),$$

and for any $\omega \in \Omega$, p_ω inherits the above bound (1.4) on p_θ, we have

$$(1.5) \qquad |\log \frac{p_\omega(x)}{p_{\omega'}(x)}| \leq |x|(d - c) + \log \frac{h^*}{h_*} , x \in \mathbf{R}$$

for any ω, $\omega' \in \Omega$.

2. Estimators induced by priors on Ω.

2.1. Bayes versus mixture of i.i.d. priors.

Consider Ω with the topology of weak convergence. Let $B(\Omega)$ denote the Borel σ-field of Ω. Let Λ be a probability on $(\Omega, B(\Omega))$. Define the prior (for each n) $\bar{\omega}_\Lambda^n$ on Θ^n as follows

$$(2.1) \qquad \bar{\omega}_\Lambda^n(B_1 \times \ldots \times B_n) = \int \prod_{i=1}^{n} \omega(B_i)\, d\Lambda,$$

for $B_1,\ldots,B_n$ Borels of Θ. [Note that the above integral makes sense because the integrand is non-negative and measurable. For a proof of measurability it suffices to take $n=1$ and B_1 open. But then it follows from a defining property of weak convergence (Billingsley (1968), Theorem 2.1.iv) that $\{\omega : \omega(B_1) > \epsilon\}$ is open and hence measurable $\forall\ \epsilon>0$.] Hereafter we will drop the superscript n in $\bar{\omega}^n_\Lambda$, as it will be clear from the context.

By the Fubini Theorem, the α-component Bayes risk against $\bar{\omega}_\Lambda$ of an estimator $\underline{t} = (t_1,\dots,t_n)$ in our compound problem is

$$(2.2) \qquad R(t_\alpha, \bar{\omega}_\Lambda) = \int_{\mathbb{R}^n}(\int_{\Theta^n}(t_\alpha - \phi(\theta_\alpha))^2 \prod_{i=1}^{n} p_{\theta_i}(x_i)d\bar{\omega}_\Lambda)d\mu^n$$

Let $e^{g_\alpha(\omega)} = \prod_{i \neq \alpha}^{n} p_\omega(x_i)$, Λ_α denote the probability measure on Ω with density proportional to e^{g_α} wrt Λ and $\omega_\alpha = \Lambda_\alpha \circ \omega$. By the Fubini Theorem (on the space $\Omega \times \Theta^n$), the inner integral in (2.2) is

$$\int\int(t_\alpha - \phi(\theta_\alpha))^2 p_{\theta_\alpha}(x_\alpha)d\omega \, e^{g_\alpha(\omega)} d\Lambda,$$

which then, by definitions of g_α and ω_α, equals

$$(2.3) \qquad (\int e^{g_\alpha(\omega)} d\Lambda) \int(t_\alpha - \phi(\theta_\alpha))^2 p_{\theta_\alpha}(x_\alpha)d\omega_\alpha(\theta_\alpha).$$

The compound risk being the average of the risks across the components, it follows from (2.2) and (2.3) that the estimator which plays component Bayes versus ω_α in the α-th estimation $\forall \ \alpha = 1, \dots, n$ is Bayes versus $\bar{\omega}_\Lambda$ in the compound problem. Let τ_ω denote the Bayes estimator of $\phi(\theta)$ versus a prior ω on Θ in the component problem,

$$\tau_\omega(x) = \int \phi(\theta) p_\theta(x) d\omega \, / \, p_\omega(x).$$

Thus the Bayes estimator $\hat{\underline{t}}$ versus $\bar{\omega}_\Lambda$ is then given by

$$(2.4) \qquad \hat{t}_\alpha(\underline{x}) = \tau_{\omega_\alpha}(x_\alpha) , \qquad \alpha = 1, ..., n.$$

2.2. Admissibility.

Since, for each $\underline{\theta} \in \Theta^n$, P and μ^n are mutually absolutely continuous, we have $P \ll \int P d\zeta$ for any prior ζ on Θ^n. Hence, by an immediate application of Lemma A.4, we get that for each $n \geq 1$, all Bayes estimators in our compound problem are admissible. In particular, $\hat{\underline{t}}$ is so.

2.3. A useful inequality on the absolute modified regret.

Recall that G_n stands for the empirical distribution of $\theta_1, ..., \theta_n$. Let $\tilde{\underline{t}}$ be the Bayes estimator versus G_n^n. Then $\tilde{t}_\alpha(\underline{x}) = \tau_{G_n}(x_\alpha)$, $1 \leq \alpha \leq n$ and the modified regret of $\hat{\underline{t}}$ at $\underline{\theta}$ is

$$(2.5) \qquad D_n(\hat{\underline{t}}, \underline{\theta}) = n^{-1} \sum_{\alpha=1}^{n} \{E(\hat{t}_\alpha - \phi(\theta_\alpha))^2 - E(\tilde{t}_\alpha - \phi(\theta_\alpha))^2\}.$$

Since $\hat{t}_\alpha, \tilde{t}_\alpha, \phi(\theta_\alpha) \in \phi[\Theta]$,

$$(2.6) \qquad |D_n(\hat{\underline{t}}, \underline{\theta})| \leq 2 \text{ diam } \phi[\Theta] \; n^{-1} \sum_{\alpha=1}^{n} E|\hat{t}_\alpha - \tilde{t}_\alpha|.$$

Note that it follows from the definitions of $\hat{t}_\alpha$ and $\tilde{t}_\alpha$ that $E|\hat{t}_\alpha - \tilde{t}_\alpha| = E_{\underline{\theta}_\alpha^{\vee}} E_{\theta_\alpha} |\tau_{\omega_\alpha} - \tau_{G_n}|$. Hence in order to investigate the bound on the modified regret given by (2.6), it is useful to consider $E_\theta|\tau_\omega - \tau_{\omega'}|$, for $\theta \in \Theta$ and $\omega, \omega' \in \Omega$.

3. A bound on the $L_1(E_\theta)$ distance between two component Bayes rules when $\phi(\theta) = e^{\theta k}$, $k \in \mathbb{N}$.

Throughout this section, let $\phi(\theta) = e^{\theta k}$ for some $k \in \mathbb{N}$. For this special case, we establish a bound, uniform in θ, essentially in terms of the L_1 distance between the corresponding mixtures.

For any two ω, $\omega' \in \Omega$, let

$$\|P_\omega - P_{\omega'}\| = \int |p_\omega - p_{\omega'}| d\mu,$$

and for a function f on $\mathbb{R}$ and any $k \in \mathbb{N}$, let $f^{(k)} = f o s_k$.

Note that $\mu_k \ll \mu$ follows from the assumption $\mu_1 \ll \mu$.

Lemma 3.1. For any ω and $\omega' \in \Omega$ and $m, m' \in (0,\infty)$,

$$(3.1) \qquad \frac{h_*}{h^*} E_\theta|\tau_\omega - \tau_{\omega'}| \le (e^{dk} - e^{ck})(P_c + P_d)[|\cdot|>m \text{ or } f^{(k)}>m']$$

$$+ e^{(d-c)m}(2e^{dk} - e^{ck} + m')\|P_\omega - P_{\omega'}\|,$$

with $f = \frac{d\mu_k}{d\mu}$.

Proof. Since τ_ω and $\tau_{\omega'} \in (e^{ck}, e^{dk})$, by C3 it follows that

$$\frac{h_*}{h^*} E_\theta|\tau_\omega - \tau_{\omega'}|[|\cdot| > m \text{ or } f^{(k)} > m']$$

is bounded by the first term in the RHS of (3.1). To bound the expectation over the other region first observe that

$$\tau_\omega(x) = \int e^{\theta k} p_\theta(x) d\omega \,/ p_\omega(x) = p_\omega^{(k)}(x)/p_\omega(x)$$

for any $\omega \in \Omega$ since $e^{\theta k}p_\theta(x) = p_\theta(x+k) = p_\theta^{(k)}(x)$. Lemma A.1 then applies to yield

$$(3.2) \qquad p_\omega|\tau_\omega - \tau_{\omega'}| \le (2e^{dk} - e^{ck})|p_\omega - p_{\omega'}| + |p_\omega^{(k)} - p_{\omega'}^{(k)}|.$$

Since ${}_{h_*}^{h^*}p_\theta \le e^{(d-c)|\cdot|}\ p_\omega$ follows from (1.5), (3.2) shows that $(h_*/h^*)E_\theta|\tau_\omega - \tau_{\omega'}|[|\cdot|\le m,\ f^{(k)}\le m']$ is bounded by

$$e^{(d-c)m}\ (2e^{dk} - e^{ck})\|P_\omega - P_{\omega'}\| \ + \int_{f^{(k)}\le m'} |p_\omega^{(k)} - p_{\omega'}^{(k)}|d\mu\ \}.$$

This completes the proof because, by the transformation theorem, the above integral wrt μ is

$$\int_{f\le m'} |p_\omega - p_{\omega'}|d\mu_k = \int_{f\le m'} |p_\omega - p_{\omega'}|fd\mu \le m'\|P_\omega - P_{\omega'}\|.$$

4. Consistency of the posterior mixtures.

In view of (3.1) and the paragraph following (2.6), the question of null convergence of the modified regret reduces to the question, loosely speaking, whether ω_α is L_1 consistent for P_{G_n}. In Theorem 4.1 we establish such a consistency result for the non-delete version. The result for the delete versions will follow as a corollary (i.e. Corollary 4.1). Note that the proof of this theorem can be extended far beyond the exponential families of Section 1.

Replace n by n+1 in Section 2.1. Denote g_{n+1}, Λ_{n+1}, ω_{n+1} (of the second paragraph of 2.1) by g, $\hat{\Lambda}$, $\hat{\omega}$ respectively. [$\hat{\omega}$ can shown to be the posterior distribution of θ_{n+1} given the data $\underline{X} = (X_1, ..., X_n)$ under the prior $\bar{\omega}_{\Lambda}^{n+1}$ on $(\theta_1, ..., \theta_n, \theta_{n+1})$.] We will use $\hat{\omega}_{(n)}$, if necessary, to exhibit the number of arguments.

We will first prove three lemmas which will be used for the proof of Theorem 4.1.

Let

$$\mathcal{V} = \bigvee_{\omega} \mid n^{-1} \sum_{1}^{n} \log p_{\omega}(x_{\alpha}) - \int \log p_{\omega} dP_{G_n} \mid$$

and for any $\omega, \omega' \in \Omega$,

$$\Delta_{\omega'}(\omega) = \int \log (p_{\omega'}/p_{\omega}) dP_{\omega'}.$$

Lemma 4.1. For each $\delta > 0$,

$$\frac{1}{2} E\| P_{\hat{\omega}} - P_{G_n} \| < \sqrt{2\delta} + P(\mathcal{V} > \frac{\delta}{4}) + \frac{e^{-\frac{1}{2} n\delta}}{\Lambda(\mathcal{U}_{\delta})},$$

where

$$\mathcal{U}_{\delta} = \{\Delta_{G_n} < \delta\} \subset \Omega. \tag{4.1}$$

Proof. By definition of $\hat{\omega}$,

$$\|P_{\hat{\omega}} - P_{G_n}\| = \int | \int p_{\theta} d(\hat{\Lambda} \circ \omega) - p_{G_n} | d\mu,$$

$$= \int | \int (\int p_{\theta} d\omega - p_{G_n}) d\hat{\Lambda}(\omega) | d\mu \text{ by Fubini on } \Omega \times \Theta, \tag{4.2}$$

$$\leq \int\int | p_{\omega} - p_{G_n} | d\hat{\Lambda}(\omega) d\mu = \int \|P_{\omega} - P_{G_n}\| d\hat{\Lambda}(\omega).$$

The last two steps follow by taking absolute under integral and Fubini on $\Omega \times \mathbb{R}$ respectively.

For any ω, by (3.6) of Hannan (1960),

$$\frac{1}{2}\|P_\omega - P_{G_n}\| \le \sqrt{\Delta_{G_n}(\omega)}.$$

Clearly LHS above ≤ 1 everywhere and, by the above, $< \sqrt{2\delta}$ on $\mathcal{U}_{2\delta}$. Combining this with (4.2), we get

$$(4.3) \qquad \frac{1}{2}\|P_{\hat{\omega}} - P_{G_n}\| < \sqrt{2\delta} + \hat{\Lambda}((\mathcal{U}_{2\delta})^c),$$

where the superscript c denotes complement.

Since $\hat{\Lambda}$ has density wrt Λ proportional to e^g and $\mathcal{V}$ is the sup norm of $n^{-1}g + \Delta_{G_n} - \int \log p_{G_n} dP_{G_n}$, one easily gets (equation (iii)′ of the addendum of Gilliland, Hannan and Huang (1976))

$$(4.4) \qquad \frac{\hat{\Lambda}((\mathcal{U}_{2\delta})^c)}{\hat{\Lambda}(\mathcal{U}_\delta)} \le \frac{e^{-2n\delta + n\mathcal{V}}}{\Lambda(\mathcal{U}_\delta)e^{-n\delta - n\mathcal{V}}}$$

by bounding g above on $(\mathcal{U}_{2\delta})^c$ and below on $\mathcal{U}_\delta$. Since $\hat{\Lambda}$ is a probability the LHS bounds $\hat{\Lambda}((\mathcal{U}_{2\delta})^c)$; while on the set $[\mathcal{V} \le \delta/4]$, the RHS is bounded by $(e^{-n\delta/2})/\Lambda(\mathcal{U}_\delta)$. Using these in (4.3) and taking expectation we get the asserted bound.

Lemma 4.2. $E\,\mathcal{V} \to 0$ uniformly in $\underline{\theta}$ as $n \to \infty$.

Proof. The conclusion readily follows by an application of Theorem A.3 with $S = \Omega$, $d =$ the Levy metric, $\mathscr{N} = \Theta^\infty$, $P_{\underline{\theta}} = \times_\alpha P_{\theta_\alpha} = P$ for $\underline{\theta} \in \Theta^\infty$, $H_\alpha(\omega) = \log p_\omega(X_\alpha)$, $\omega \in \Omega$, $\alpha \geq 1$.

(S,d) is a compact metric space by Helly's theorem.

Continuity of $\omega \rightsquigarrow p_\omega(x)$ follows from the continuity and the boundedness of $\theta \rightsquigarrow p_\theta(x)$. Thus H_α's satisfy (i). We verify (ii+) and (iii+) of Remark A.3 in the present situation.

For (ii+), use (1.4) to observe that

$$\|H_\alpha\| \leq |X_\alpha|(-c \vee d) + \log (h_*^{-1} \vee h^*),$$

and

$$* \ (|X_\alpha| - M)_+ \leq \limsup_n \bigvee_{\underline{\theta}} \bigvee_{\alpha=1}^{n} P(|X_\alpha| - M)_+ = \bigvee_\theta P_\theta(|\cdot| - M)_+ \downarrow 0$$

as $M \uparrow \infty$ in view of (1.3) and (1.2).

Next observe that p_ω is convex (because it is log-convex by Hölder) and continuous (by continuity of p_θ and D.C.T.). Using also the previously noted continuity of $\omega \rightsquigarrow p_\omega(x)$ $\forall$ x, it follows by Lemma A.2 that for any $m < \infty$,

$$(4.5) \qquad \bigvee_{|x| \leq m} |p_\omega(x) - p_{\omega_0}(x)| \longrightarrow 0 \quad \text{as } \omega \longrightarrow \omega_0 \ .$$

Let $V_{\omega_0 \rho \alpha} = \vee\{|H_\alpha]_{\omega_0}^{\omega}| : d(\omega,\omega_0) < \rho\}$. Since $(a \wedge b)|\log \frac{b}{a}| \leq |b-a|$ for any $a, b \in (0,\infty)$, (4.5) implies that for each $\epsilon > 0$ and $m < \infty$, $\exists\ \rho_0 > 0$

such that

$$(4.6) \qquad [V_{\omega_0\rho\alpha} > \epsilon,\ |X_\alpha| \le m] = \phi\ ,\ \forall\ \alpha \ \text{ if } \ \rho < \rho_0.$$

Hence

$$(4.7) \quad \limsup_{\rho\downarrow 0} \ \bigvee_{\alpha,\underline{\theta}} P[V_{\omega_0\rho\alpha} > \epsilon] \ \le \ \bigvee_{\alpha,\underline{\theta}} P[|X_\alpha|>m] \ = \ \bigvee_{\theta} P_\theta(m,\infty).$$

Now let $m\uparrow\infty$ to conclude, LHS of (4.7) $= 0$. This establishes (iii+).

$$\text{Also } \|\overline{H}^{(\underline{\theta})}\| = \bigvee_{\omega} \Big| \ n^{-1}\sum_{\alpha}\log\ p_\omega(X_\alpha) - n^{-1}\sum_{\alpha} P\log\ p_\omega(X_\alpha)\Big| \ = \ \mathcal{V}.$$

Hence by Theorem A.3, $* \ \mathcal{V} = \ \limsup_n \bigvee_{\underline{\theta}} E\ \mathcal{V} = 0$.

Lemma 4.3. If support of $\Lambda = \Omega$ then for any $\delta > 0$,

$$\bigwedge_{\omega_0\in\Omega} \Lambda(\{\Delta_{\omega_0} < \delta\}) \ > 0.$$

Proof. Fix $\delta > 0$. For an $\omega_0 \in \Omega$, the continuity of the function $\omega \rightsquigarrow \int\log\ (p_{\omega_0}/p_\omega)dP_{\omega_0}$ follows by D.C.T since $\log\ p_{\omega_n} \to \log\ p_\omega$ as $\omega_n \to \omega$ and (1.5). So the set $\{\Delta_{\omega_0} < \delta\}$ is open, non-empty (because it contains ω_0) and hence

$$(4.8) \qquad \Lambda(\{\Delta_{\omega_0} < \delta\}) > 0,$$

since Λ has full support.

Next observe that, since the functions Δ_{ω_n} converge to Δ_{ω_0} pointwise and hence in Λ–distribution if $\omega_n \longrightarrow \omega_0$, by the defining property of the latter convergence (cf. our Section 2.1 usage)

$$\liminf_n \Lambda(\{\Delta_{\omega_n} < \delta\}) \geq \Lambda(\{\Delta_{\omega_0} < \delta\}) \quad \text{if} \quad \omega_n \longrightarrow \omega_0 .$$

This shows that the function $\omega_0 \rightsquigarrow \Lambda(\{\Delta_{\omega_0} < \delta\})$ is lower semi–continuous. Hence it attains its infimum because Ω is compact. The proof now ends by (4.8).

Theorem 4.1. Let the support of Λ be Ω. Then

$$E\| P_{\hat{\omega}} - P_{G_n} \| \longrightarrow 0 \text{ uniformly in } \underline{\theta} \text{ as } n \longrightarrow \infty.$$

Proof. Fix a $\delta > 0$. Consider the bound in Lemma 4.1. Now, as $n \longrightarrow \infty$, the second term of this bound goes to zero by Lemma 4.2 and so does the third term by Lemma 4.3, the convergences being uniform in $\underline{\theta}$. Thus

$$\limsup_n \bigvee_{\underline{\theta}} E\| P_{\hat{\omega}} - P_{G_n} \| \leq 2\sqrt{2\delta} .$$

The proof ends, $\delta > 0$ being arbitrary.

Corollary 4.1. If the support of $\Lambda = \Omega$ then

$$\bigvee_{\alpha \leq n} E \| P_{\omega_\alpha} - P_{G_n} \| \longrightarrow 0 \text{ uniformly in } \underline{\theta} \text{ as } n \longrightarrow \infty.$$

Proof. Fix $\underline{\theta} \in \Theta^n$ and $1 \leq \alpha \leq n$. Let $G_{n\alpha}$ denote the empirical based on $\check{\underline{\theta}}_\alpha$. Then we have the following.

(i) $G_{n\alpha}$ is G_{n-1} corresponding to $\check{\underline{\theta}}_\alpha \in \Theta^{n-1}$.

(ii) Letting $F: \mathbb{R}^{n-1} \longrightarrow \Omega$ be the function such that $F(\underline{X}_{n-1}) = \hat{\omega}_{(n-1)}$, it follows from the definition of ω_α that $\omega_\alpha = F(\check{\underline{X}}_\alpha)$.

(iii) Clearly $P_{\underline{\theta}} \check{\underline{X}}_\alpha^{-1} = P_{\check{\underline{\theta}}_\alpha} \underline{X}_{n-1}^{-1}$.

From (i), (ii) and (iii), we get that

$$P_{\underline{\theta}} \| P_{\omega_\alpha} - P_{G_{n\alpha}} \|^{-1} = P_{\check{\underline{\theta}}_\alpha} \| P_{\hat{\omega}_{(n-1)}} - P_{G_{n-1}} \|^{-1}.$$

Since $\underline{\theta}$ and α are arbitrary

$$\bigvee_{\underline{\theta} \in \Theta^n} \bigvee_{\alpha \leq n} E \| P_{\omega_\alpha} - P_{G_{n\alpha}} \| \leq \bigvee_{\underline{\theta} \in \Theta^{n-1}} E \| P_{\hat{\omega}_{(n-1)}} - P_{G_{n-1}} \| \longrightarrow 0$$

as $n \longrightarrow \infty$, by Theorem 4.1.

Next, because $G_n - G_{n\alpha} = n^{-1}(\delta_{\theta_\alpha} - G_{n\alpha})$ with δ_{θ_α} the distribution degenerate at θ_α, the variation norm of $G_n - G_{n\alpha}$ is no more than $2n^{-1}$. Thus, by definition of p_ω and by (1.3),

$$|p_{G_n} - p_{G_{n\alpha}}| \leq 2n^{-1} \bigvee_{\theta} p_\theta \leq 2n^{-1} \frac{h^*}{h_*}(p_c + p_d).$$

Consequently

$$\|P_{G_n} - P_{G_{n\alpha}}\| \;\le\; 2n^{-1} \int(\vee_{\theta}\, p_{\theta}(x))d\mu \;\le\; (4h^*/h_*)n^{-1}.$$

The proof now ends by the triangle inequality.

(Empirical Bayes, Estimation of mixtures.)

Consider the situation where $X_1, \ldots, X_n$ are i.i.d. observations from the mixed distribution P_{ω}, $\omega \in \Omega$ being unknown to the statistician. The problem is to estimate P_{ω}. This model obtains in the usual empirical Bayes context where $\underline{E}$ is an expectation under which $\theta_1, \ldots, \theta_n$ are i.i.d. $\sim \omega$ and, given $\underline{\theta}$, $\underline{X} \sim \times_{\alpha=1}^{n} P_{\theta_{\alpha}}$. It turns out that $P_{\hat{\omega}}$ is L_1 consistent for P_{ω} whenever Λ has full support.

Corollary 4.2. If the support of $\Lambda = \Omega$ then, for any ω,

$$\underline{E} \,\| P_{\hat{\omega}} - P_{\omega}\| \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty.$$

Proof. It follows from the continuity, noted in Lemma 4.2 's proof, of $\omega \leadsto p_{\omega}(x)$ $\forall$ x that $\omega \leadsto P_{\omega}$ is continuous in $\|.\|$ by the Scheffe′ theorem. Thus as $n \longrightarrow \infty$, $P_{G_n} \longrightarrow P_{\omega}$ a.s. (since $G_n \longrightarrow \omega$ a.s. by Glivenko–Cantelli) and hence $\underline{E}\| P_{G_n} - P_{\omega}\| \longrightarrow 0$ by D.C.T. The conclusion now follows by Theorem 4.1 and the triangle inequality.

5. Asymptotic optimality. Now we are in a position to prove our main result.

Theorem 5.1 (Main result). If the support of $\Lambda = \Omega$ and $\underset{\sim}{\hat{t}}$ is the Bayes estimator given by (2.4), then

$$(5.1) \qquad \bigvee_{\alpha\leq n} E|\hat{t}_\alpha - \tilde{t}_\alpha| \longrightarrow 0 \text{ uniformly in } \underset{\sim}{\theta} \text{ as } n \longrightarrow \infty.$$

Consequently $\underset{\sim}{\hat{t}}$ is a.o.

Proof. The second part of the assertion follows from (2.6).

For the first part, first consider the case $\phi(\theta) = e^{\theta k}$, $k\in\mathbb{N}$. By the representation noted after (2.6),

$$\begin{aligned} E|\hat{t}_\alpha - \tilde{t}_\alpha| &= E_{\underset{\sim}{\theta}_\alpha^{\smile}} E_{\theta_\alpha} |\tau_{\omega_\alpha} - \tau_{G_n}|, \\ &\leq \frac{h^*}{h_*} \{(e^{dk} - e^{ck})(P_c + P_d)[|\cdot|>m \text{ or } f^{(k)}>m'] \\ &\qquad + e^{(d-c)m}(2e^{dk} - e^{ck} + m')\, E\| P_{\omega_\alpha} - P_{G_n} \|\} \end{aligned}$$

by Lemma 3.1, where $m, m' < \infty$ are arbitrary. For each m and m', the second term of the above bound is o(1) uniformly in α and $\underset{\sim}{\theta}$ as $n \longrightarrow \infty$ by Corollary 4.1. The first term is independent of α and $\underset{\sim}{\theta}$ and can be made arbitrarily small by choosing m and m' large enough. This concludes the proof in the present case.

Next let $\phi(\theta) = \sum_k a_k e^{\theta k}$ be a polynomial in e^{θ}. By definition and the linearity property of conditional expectation (or integral) it follows that

$$\hat{t}_\alpha = \sum_k a_k \hat{t}^{[k]}_\alpha , \quad \tilde{t}_\alpha = \sum_k a_k \tilde{t}^{[k]}_\alpha$$

where $\hat{t}^{[k]}_\alpha$ and $\tilde{t}^{[k]}_\alpha$ are the corresponding Bayes estimators of $e^{\theta_\alpha k}$ for each k. Hence (5.1) holds since it holds with $\underline{\hat{t}}^{[k]}$ and $\underline{\tilde{t}}^{[k]}$ for each k by the previous case.

Finally for general continuous ϕ, given $\epsilon > 0$, choose a polynomial p such that $\forall_\theta \; |\phi(\theta) - p(e^{\theta})| < \epsilon$. Then, using definitions and taking absolute values under integrals,

$$|\hat{t}_\alpha - \hat{t}^{[p]}_\alpha| < \epsilon , \quad |\tilde{t}_\alpha - \tilde{t}^{[p]}_\alpha| < \epsilon$$

where $\hat{t}^{[p]}_\alpha$ amd $\tilde{t}^{[p]}_\alpha$ are the corresponding Bayes estimators of $p(e^{\theta_\alpha k})$ and so

$$|\hat{t}_\alpha - \tilde{t}_\alpha| \leq |\hat{t}^{[p]}_\alpha - \tilde{t}^{[p]}_\alpha| + 2\epsilon.$$

The proof is now complete by the previous case, ϵ being arbitrary.

CHAPTER 2

THE SEQUENCE COMPOUND ESTIMATION

1. Introduction.

Here we consider the sequence compound version of the problem treated in Chapter 1. In this formulation, at each stage α we estimate $\phi(\theta_\alpha)$ by estimators based on the data $\underline{X}_\alpha = (X_1, .., X_\alpha)$ then available. The sequence compound estimator, which for each n plays Bayes versus $\overline{\omega}_\Lambda^n$ with the compound loss $L_n(\underline{t},\underline{\theta}) = n^{-1}\sum_{\alpha=1}^{n}(t_\alpha - \phi(\theta_\alpha))^2$, turns out to be of s.s.c. type described in Chapter 0 and is a.o. if Λ has full support. The proof reduces to a corollary to the set compound result via an inequality due to Hannan (1957).

2. Bayes versus $\overline{\omega}_\Lambda^n$.

2.1. The Bayes sequence compound estimator.

Fix an $n \geq 1$. For $1 \leq \alpha \leq n$, a stage α sequence compound rule t_α has $R(t_\alpha, \overline{\omega}_\Lambda^n) = R(t_\alpha, \overline{\omega}_\Lambda^\alpha)$ which is its α component Bayes risk in the set problem with α components. But t_α which minimizes $R(t_\alpha, \overline{\omega}_\Lambda^\alpha)$ is the α-th component of the Bayes rule versus $\overline{\omega}_\Lambda^\alpha$ in the set problem with α components. Hence we get from Section 2 of Chapter 1 that in the sequence problem the estimator which minimizes $R_n(\underline{t}, \overline{\omega}_\Lambda^n) = n^{-1}\sum_{\alpha=1}^{n} R(t_\alpha, \overline{\omega}_\Lambda^n)$ is given by

$$\hat{t}_\alpha(\underline{x}_\alpha) = \tau_{\omega_\alpha}(x_\alpha), \quad 1 \leq \alpha \leq n, \tag{2.1}$$

where ω_α is as in (1.2.4) with $n=\alpha$. Note that the components do not depend on n and let $\hat{\underline{t}} = <\hat{t}_\alpha : \alpha \geq 1>$.

2.2. Admissibility.

It has been noted in Chapter 1 that the condition of Lemma A.4 holds in our case. Hence all Bayes sequence compound estimators are admissible. In particular $\hat{\underline{t}}$ is so.

Remark 2.1. Note that for each n, $\hat{t}_n$ is Bayes versus $\bar{\omega}_\Lambda^n$ in the class of all stage n sequence compound estimators t_n wrt the n–th estimation loss $L(t_n, \underline{\theta}_n) = (t_n - \phi(\theta_n))^2$ and, as recorded in the proof of Lemma A.4, has unique risk.

3. Asymptotic optimality.

Theorem 3.1. If Λ has full support then $\hat{\underline{t}}$ is a.o.

Proof. Let $\tilde{t}_{\alpha n}(\underline{x}_n) = \tau_{G_n}(x_\alpha)$ and $\tilde{\underline{t}}_n = (\tilde{t}_{1n}, \ldots, \tilde{t}_{nn})$ for $1 \leq \alpha \leq n < \infty$ and $\tilde{\underline{t}} = <\tilde{t}_{nn}>$. Now $R_n(\tilde{\underline{t}}, .) \leq R_n(\tilde{\underline{t}}_n, .)$, since its generalization (cf. inequality (8.8) of Hannan (1957)) holds without restriction and hence $D_n(\hat{\underline{t}}, \underline{\theta})_+$ is bounded by RHS (1.2.6) with $\hat{t}_\alpha$ as in this chapter and $\tilde{t}_\alpha$ replaced by $\tilde{t}_{\alpha\alpha}$. But since Λ has full support this bound is o(1) uniformly in $\underline{\theta}$, because $\bigvee_{\underline{\theta}} E|\hat{t}_n - \tilde{t}_{nn}|$ is o(1) by Theorem 1.5.1 and convergence implies Cesaro convergence.

Remark 3.1. In fact, in the above, $\bigvee_{\underline{\theta}} |D_n(\hat{\underline{t}}, \underline{\theta})| = o(1)$. To prove this note that a slight extension of (2.5) of Gilliland (1968) gives

$$|D_n(\hat{\underset{\sim}{t}},\underset{\sim}{\theta})| \leq 2 \text{ diam } \phi[\Theta] \; n^{-1} \sum_{\alpha=1}^{n} E|\hat{t}_\alpha - \tilde{t}_{\alpha\alpha}| + O(n^{-1}\log n),$$

where $O(n^{-1}\log n)$ is uniform in $\underset{\sim}{\theta}$. But the above proof has shown that first term is $o(1)$ uniformly in $\underset{\sim}{\theta}$.

CHAPTER 3

THE EMPIRICAL BAYES ESTIMATION

1. Introduction.

In this chapter we look at the empirical Bayes approach to our estimation problem. An empirical Bayes estimator $\underset{\sim}{t} = <t_n : n \geq 1>$ is such that t_n is a function of $\underline{X}_n$ with n–th estimation risk function

$$(1.1) \qquad R_n(t_n,\omega) = \int\int(t_n - \phi(\theta_n))^2 dP_{\underline{\theta}} d\omega^n , \quad \omega \in \Omega.$$

See Chapter 0 for further details.

We will prove that the empirical Bayes estimator $\hat{\underset{\sim}{t}} = <\hat{t}_n>$, where $\hat{t}_n$ is the Bayes estimator versus a prior Λ on Ω, is a.o. if Λ has full support.

2. Bayes versus Λ.

For any given n, any estimator t_n based on $\underline{X}_n$ has stage n Bayes risk wrt Λ in the empirical Bayes problem equal to its stage n Bayes risk versus $\bar{\omega}^n_\Lambda$ in the compound problem since iterated integral $d\omega^n d\Lambda$ and integral $d\bar{\omega}^n_\Lambda$ have same meaning. Hence the Bayes empirical Bayes estimator of $\phi(\theta_n)$ versus Λ is $\hat{t}_n$ of (2.2.1).

Admissibility. For any n, the stage n Bayes e.B. estimators wrt Λ are the stage n Bayes sequence compound estimators wrt $\bar{\omega}^n_\Lambda$ and, since the integral $d\omega^n$ maps the risks of the latter to those of the former, inherit the uniqueness of risk from that of the later (cf. Remark 2.2.1) and hence are admissible.

3. Asymptotic optimality.

Theorem 3.1. If Λ has full support then the Bayes empirical Bayes estimator $<\hat{t}_n>$ is a.o.,

$$\text{i.e.} \quad \forall\ \omega \in \Omega, \quad R_n(\hat{t}_n,\omega) \longrightarrow R(\omega) \quad \text{as } n \longrightarrow \infty.$$

Proof. First method: Since for each $n \geq 1$, $\hat{t}_n$ is the n–th component of the compound rule $\hat{\underset{\sim}{t}}$ of (1.2.4) whose equivariance follows from the definition of $\hat{\underset{\sim}{t}}$ and asymptotic optimality in the compound problem follows from Theorem 1.5.1, the asymptotic optimality of $\hat{t}_n$ in the empirical Bayes problem follows by Remark 1 of Gilliland and Hannan (1986).

Second method: A direct proof of the asymptotic optimality of $\hat{t}_n$ can be given in the present case along the same lines as that of Theorem 1.5.1. Interpret E as P^n_ω and $\tilde{t}_n$ as the component Bayes estimator versus ω based on X_n. Then as before, it is sufficient to prove that $E|\hat{t}_n - \tilde{t}_n| \longrightarrow 0$ as $n \longrightarrow \infty$. The proof goes through by the same steps with the use of Corollary 1.4.2 instead of 1.4.1.

Remark 3.1. For the asymptotic optimality of the Bayes rules in the general finite state empirical Bayes problem, Gilliland, Boyer and Taso (1982) came up with the same sufficient condition on Λ as the present work.

CHAPTER 4

EXAMPLES OF Λ AND CONCLUDING REMARKS

1. Introduction.

In Chapters 1–3 we have shown that the Bayes estimators under consideration in the set, sequence compound and the empirical Bayes versions of our problem are all related and are asymptotically optimal if Λ (of 1.2.1) has full support. In Bayesian contexts, we can say that such a prior is nonparametric in nature, a desirable property as indicated by Ferguson (1973) and others.

Priors on the set of probability distributions, or random probabilities or random distribution functions have been considered by many authors in the contexts of nonparametric Bayes and empirical Bayes estimation, estimation of mixing distributions, etc. In Section 2 we present brief descriptions of four examples of full support Λ from their works (with some modifications in the Rolph case).

In Section 3, some practically useful forms of the Bayes estimators corresponding to some of them are obtained. It is pointed out in the beginning of the section that the Bayes estimators can be expressed as a ratio of two multidimentional integrals involving the posterior means of ω. This form is useful if the posteriors of ω are analytically calculable (e.g. A and B).

Section 4 contains the concluding remarks which include some possible generalizations of the present work and its application to some non–exponential families. Also some related open problems are indicated.

2. Examples of Λ.

We list below five examples of Λ with support Ω.

A. (Dirichlet process) An important class of priors on the probabilities on $\mathbb{R}$ with manageable posteriors has been introduced by Ferguson (1973). Among many equivalent definitions, here we state Definition 1 of Ferguson (1973,1974). Other equivalent representations can be found in Blackwell and McQueen (1973), Ferguson (1974) and Sethuraman and Tiwari (1982).

Definition: Let γ be a non–null, finite Borel measure on $\mathbb{R}$. Then Λ is called the Dirichlet process prior with parameter γ (hereafter we write $\Lambda = \mathcal{D}(\gamma)$) if for every finite measurable partition $\{B_1, ..., B_m\}$ of $\mathbb{R}$ the distribution of $(\omega(B_1), ..., \omega(B_m))$ under Λ (ω is the identity function on the space of probabilities on $\mathbb{R}$) is Dirichlet with parameters $(\gamma(B_1), ..., \gamma(B_m))$.

It is well known (e.g. Ferguson (1974)) that the support of $\mathcal{D}(\gamma)$ is the set of probability distributions on $\mathbb{R}$ whose support is contained in the support of γ. So if we choose γ with support of $\gamma = \Theta = [c,d]$ then $\Lambda = \mathcal{D}(\gamma)$ has support Ω.

B. (Processes neutral to the right) A more general class of priors than Dirichlet process has been introduced by Doksum (1974).

Definition: A random distribution function $F(t)$ on real line is said to be neutral to the right if, for every m and $t_1<t_2< .. <t_m$, $\exists$ independent random variables $V_1, V_2, .., V_m$ such that $(\overline{F}(t_1), \overline{F}(t_2), .., \overline{F}(t_m))$ has the same distribution as $(V_1, V_1V_2, .., \prod_{i=1}^{m} V_i)$, where $\overline{F} = 1-F$.

Doksum (1974, Theorem 3.1) gave a nice characterization of such processes in terms of independent increment processes. In his Example 3.1, he showed how such a process can constructed starting from any non–negative, infinitely divisible random variable Y and any distribution function β_0 on $\mathbb{R}$. If β_0 is absolutely continuous wrt the Lebesgue measure on [c,d] with positive density and the distribution function of Y is strictly increasing, then the resulting process has support Ω.

C. (Distributions on the moment space) Let
$D = \{(\mu_1, \mu_2, \dots) : \mu_i = \int \theta^i d\omega, \forall\ i \geq 1, \text{ for some } \omega \in \Omega\} \subset \mathbb{R}^\infty$ be the space of moment sequences of probabilities on Θ.

Since any $\omega \in \Omega$ is determined by its moment sequence $\{\mu_i\} \in D$, a prior on D induces a prior on Ω in the obvious way. To make the ideas precise consider Ω with the weak convergence topology and D with the product topology. Let μ be the mapping $\omega \rightsquigarrow (\mu_1(\omega), \mu_2(\omega), \dots)$, $\mu_i(\omega) = \int \theta^i d\omega$, $i \geq 1$. Then μ is 1–1, continuous, onto D and hence is a homeomorphism since Ω is compact and D is Hausdorff. Thus a prior Λ on $(D, \mathcal{B}(D))$ induces the prior $\Lambda_\Omega = \Lambda\mu$ on $(\Omega, \mathcal{B}(\Omega))$. Since μ is a homeomorphism, support of $\Lambda_\Omega = \Omega$ iff support of $\Lambda = D$. Hereafter we will write Λ for Λ_Ω too.

The structure of D, for the case $\Theta = [0,1]$, has been studied by many authors. Rolph (1968) exploited this structure to define his prior sequentially on the co–ordinates. His priors can be adapted to the case $\Theta = [c,d]$ by the reparametrization $\theta \rightsquigarrow (\theta - c)/(d - c)$.

Another way of putting priors on D would be to follow Rolph's approach directly for D = D[c,d]. It is easy to see that D[c,d] has the same structure as that of D[0,1]. Let us elaborate this more extensively. We use Rolph's notation of lower and upper bars to denote corresponding bounds on the range of moments given their predecessors – despite conflict with our n–tuple lower bar notation.

Let π_n be the projection of $\mathbb{R}^\infty$ onto its first n co–ordinates, $n \geq 1$ and $D_n = \pi_n(D)$. For $(\mu_1, ..., \mu_n) \in D_n[c,d]$ let

$$\underline{\mu}_{n+1}(\mu_1, ..., \mu_n) = (d-c)^{n+1}\underline{m}_{n+1}(m_1, ..., m_n) - \sum_{r=0}^{n} \binom{n+1}{r}(-c)^{n-r}\mu_r,$$

(2.1)

$$\overline{\mu}_{n+1}(\mu_1, ..., \mu_n) = (d-c)^{n+1}\overline{m}_{n+1}(m_1, ..., m_n) - \sum_{r=0}^{n} \binom{n+1}{r}(-c)^{n-r}\mu_r,$$

where $m_0 = 1$, $m_r = (d-c)^{-r}\sum_{i=0}^{r}\binom{r}{i}(-c)^{r-i}\mu_i$, $1 \leq r \leq n$ and $\underline{m}_{n+1}$, $\overline{m}_{n+1}$ are as in (5) of Rolph. Then for $(\mu_1, ..., \mu_n) \in D_n[c,d]$, $(\mu_1, ..., \mu_n, \mu_{n+1}) \in D_{n+1}[c,d]$ iff

(2.2) $$\underline{\mu}_{n+1}(\mu_1, ..., \mu_n) \leq \mu_{n+1} \leq \overline{\mu}_{n+1}(\mu_1, ..., \mu_n).$$

Thus starting from a sequence of measurable functions $\{h_n\}$ positive on [c,d] with $\int_{[c_n,d_n]} h_n dx < \infty$ (the integral is wrt Lebesgue), c_n and d_n being the minimum and the maximum of $\theta \rightsquigarrow \theta^n$ on [c,d], we construct the prior on D[c,d] exactly the same way given in Rolph with only changes of m_n, $\underline{m}_n$, $\overline{m}_n$ to μ_n, $\underline{\mu}_n$, $\overline{\mu}_n$ respectively.

Under this Λ, the distribution of any finite moment sequence has full support, h_n's being positive. It then follows from the definition of the product topology that Λ has full support. We will refer to this prior as Rolph's prior for [c,d].

D. (Random distribution functions of Dubins and Freedman) Starting from a probability (base probability in their terms) on the Borels of unit square assigning measure 0 to the corners (0,0) and (1,1), Dubin and Freedman (1966) constructed a random distribution on [0,1]. They gave (3.6,Theorem) precise conditions on the base probability so that the resulting prior has full support. Then an obvious transformation carries it over to a prior on Ω with full support. However for these priors we do not know any form of the Bayes estimator which can be computed in practice.

E. (Discrete priors) Since Ω is compact and metrizable, it is separable. Let $\{\omega_n\}$ be a dense sequence in Ω and $0<c_n<\infty$, $\Sigma c_n = 1$. Then $\Lambda = \Sigma c_n \delta_{\omega_n}$, δ_{ω_n} being the probability degenerate at ω_n, has support Ω.

This example shows that, contrary to as assumed in Inglis (1973), non–atomicity of Λ is not necessary for the purpose of asymptotic optimality.

3. The Bayes estimators.

It has been pointed out in the previous chapters that the Bayes estimators in the different formulations considered are all related. In this section we find out some practically useful expressions for $\hat{t}_\alpha$, the α-th component of the Bayes estimator in the set problem (with n components), $1 \leq \alpha \leq n$, for some of the examples considered above.

From (1.2.4), its preceding equation and the definition of ω_α it follows that

$$(3.1) \qquad \hat{t}_\alpha(\underline{x}) = \frac{\int \phi(\theta_\alpha) \prod_{i=1}^{n} p_{\theta_i}(x_i)\, d\bar{\omega}_\Lambda(\underline{\theta})}{\int \prod_{i=1}^{n} p_{\theta_i}(x_i) d\bar{\omega}_\Lambda(\underline{\theta})} .$$

Let ω denote a random probability having distribution Λ and, given ω, let $\underline{\theta}_n = (\theta_1, ..., \theta_n)$ be a random sample from ω ; equivalently consider P such that

$$(3.2) \qquad P(\omega\epsilon\Omega_0, \theta_1\epsilon B_1, ..., \theta_n\epsilon B_n) = \int I_{\Omega_0}(\omega) \prod_{i=1}^{n} \omega(B_i)\, d\Lambda(\omega)$$

$\forall\ \Omega_0 \in \mathscr{B}(\Omega)$, $B_1, ..., B_n \in \mathscr{B}(\Theta)$.

Note that the marginal distribution of $\underline{\theta}_n$ is $\overline{\omega}_\Lambda^n$, i.e.

$$(3.3) \qquad P(\theta_1\epsilon B_1, ..., \theta_n\epsilon B_n) = \int \prod_{i=1}^{n} \omega(B_i) d\Lambda.$$

Let $\Lambda^{\underline{\theta}_n}$ be a regular posterior distribution of ω given $\underline{\theta}_n$, so that

$$(3.4) \qquad \int_{B_1\times\ldots\times B_n} \Lambda^{\underline{\theta}_n}\ (\Omega_0)\ d\overline{\omega}_\Lambda(\underline{\theta}_n) = \text{LHS of (3.2)}$$

and let $\omega^{\underline{\theta}_n}$ be the posterior mean of ω given $\underline{\theta}_n$, i.e.

$$(3.5) \qquad \omega^{\underline{\theta}_n} = \int \omega d\Lambda^{\underline{\theta}_n} .$$

Then by repeated conditioning it follows that

$$\text{LHS of (3.3)} = \int_{B_1} .. \int_{B_{n-1}} \int_{B_n} d\omega^{\underline{\theta}_{n-1}}(\theta_n)\ d\omega^{\underline{\theta}_{n-2}}(\theta_{n-1})\ ..\ d\omega^{\underline{\theta}_0}(\theta_1),$$

where $\omega^{\underline{\theta}_0} = \int \omega d\Lambda$.

Using this via (3.3) in (3.1) we get

$$(3.6) \qquad \hat{t}_\alpha(\underline{x}) \quad = \quad \frac{\int ... \int \phi(\theta_\alpha) \prod_{i=1}^{n} p_{\theta_i}(x_i) \prod_{i=1}^{n} d\omega^{\underline{\theta}_{i-1}}\ (\theta_i)}{\int ... \int \prod_{i=1}^{n} p_{\theta_i}(x_i) \prod_{i=1}^{n} d\omega^{\underline{\theta}_{i-1}}\ (\theta_i)} .$$

For examples A and B (3.6) can be used to calculate $\hat{t}_\alpha$ from the data.

A. Let $\Lambda = \mathscr{D}(\gamma)$, with support $\gamma = \Theta$. Then by Proposition 1 and Theorem 1 of Ferguson (1973),

$$(3.7) \qquad \omega^{\underline{\theta}_n} = \frac{(\gamma + \sum_{i=1}^{n} \delta_{\theta_i})}{\gamma(\Theta) + n}, \quad n \geq 0.$$

B. In this case expression of the posterior means is known but complicated. It is given in Doksum, Example 4.1.

Remark 3.1. A trivial generalization of the (3.7) specialization of (3.6) is given in Kuo (1986). Also Kuo has described and exemplified a Monte Carlo method for its calculation.

When $\mu(\mathbb{N}^c) = 0$ consider the reparametrization $\eta = e^{\theta}$. Then we can use the geometric form of our $\tilde{p}_\eta(x) = \eta^x \tilde{h}(\eta)$ to obtain a manageable expression for $\hat{t}_\alpha$ in Example C when $\phi(\eta) = \eta^k$, $k \in \mathbb{N}$. For a general ϕ one can then use a suitable polynomial approximation.

C. Let $\Sigma a_j \eta^j$ be a polynomial approximating $\tilde{h}$ on $[e^c, e^d]$. Then for $\tilde{\Lambda}$ = Rolph's prior for $[e^c, e^d]$ and $\phi(\eta) = \eta^k$, $\hat{t}_\alpha$ will be approximated by

$$(3.8) \qquad \int \Sigma a_j \mu_{X_\alpha + j + k} \prod_{i \neq \alpha}^{n} (\Sigma a_j \mu_{X_i + j})\, d\tilde{\Lambda} \Big/ \int \prod_{i=1}^{n} (\Sigma a_j \mu_{X_i + j})\, d\tilde{\Lambda}.$$

It is important to note that $\hat{t}_\alpha$ above corresponds to the $\tilde{\Lambda}$ mix of i.i.d. priors on $\underline{\eta}$ or, in terms of the original parametrization, the $\tilde{\Lambda}$ mix of i.i.d. priors on $\underline{e}^{\theta}$. The integrals in (3.8) reduce to Lebesgue integrals on some Euclidean spaces ($\mathbb{R}^{X^* + d + k}$ and $\mathbb{R}^{X^* + d}$ respectively to be precise where

$X^* = \vee <X_1, ..., X_n>$ and d is the degree of the polynomial) because any finite μ sequence has Lebesgue density under Λ.

To see (3.8), first use Fubini to rewrite (3.1) (with $\underline{\theta}$ replaced by $\underline{\eta}$, p_θ by $\tilde{p}_\eta$ and $\phi(\eta) = \eta^k$) as

$$(3.9) \qquad \hat{t}_\alpha(\underline{X}) = \frac{\int(\int \eta_\alpha^k \tilde{p}(X_\alpha) d\omega(\eta) \prod_{i \neq \alpha}^n \int \tilde{p}_\eta(X_i) d\omega(\eta))\ d\Lambda}{\int(\prod_{i=1}^n \int \tilde{p}_\eta(X_i) d\omega(\eta))\ d\Lambda} .$$

Now use $\tilde{p}_\eta(x) = \eta^x \tilde{h}(\eta)$, $\tilde{h}(\eta) \approx \Sigma a_j \eta^j$ in (3.9) to get (3.8).

Remark 3.2. For conditionally uniform prior, i.e. $h_i \equiv 1 \ \forall\ i$, (3.8) takes a simpler form. In cases like Geometric and Negative Binomial $\tilde{h}$ itself is a polynomial. For the Poisson case we can choose the polynomials $\Sigma \frac{(-1)^j}{j!} \eta^j$ for approximation. In general, since $\tilde{h}$ is continuous, a sequence of approximating polynomials always exists. Moreover such a sequence can be found numerically.

4. Remarks.

1. It should be obvious that we can also treat the cases where the components are 1–1 transforms of some exponential families we have been considering. Suppose that the component distributions P_θ , $\theta \in \Theta$ are such that $\{Q_\eta: \eta \in H\}$ form one such exponential family where $Q_\eta = P_{\psi^{-1}(\eta)} T^{-1}$, T and ψ are 1–1 transformations on $\mathbb{R}$ and Θ respectively and ψ^{-1} is continuous. Let $\underline{X} \sim P_{\underline{\theta}}$. Then $\underline{Y} \sim Q_{\underline{\eta}}$ where $Y_\alpha = T(X_\alpha)$, $\eta_\alpha = \psi(\theta_\alpha)$, $1 \leq \alpha \leq n$. Since T is 1–1, estimators (based on $\underline{Y}$) in the transformed problem are related in a 1–1 fashion to the estimators (based

on $\underline{X}$) in the original problem. Any such two estimators have identical risk function under a common parametrization. Moreover since ψ^{-1} is continuous, ϕ remains continuous in the reparametrization η of the transformed problem. Hence the conclusion of Chapter 1, Section 2 and Theorem 1.5.1 for the transformed problem implies that the set compound estimator

$$\hat{t}_{\alpha}(\underline{X}) = \frac{\int \phi(\psi^{-1}(\eta)) q_{\eta}(T(X_{\alpha}))\, d\omega_{\alpha}}{\int q_{\eta}(T(X_{\alpha}))\, d\omega_{\alpha}}, \quad \alpha = 1, \ldots, n,$$

is admissible and a.o. for estimating $\underline{\phi}$ in the original problem, where ω_{α} is as in Section 1.2.1 with $g_{\alpha}(\omega) = \sum_{i \neq \alpha}^{n} \log q_{\omega}(T(X_i))$. Analogous conclusions hold for the sequence compound and the empirical Bayes versions.

2. We can generalize the component loss to weighted squared error loss, where the weight function is positive and continuous. If $L(a,\theta) = w(\theta)(a-\phi(\theta))^2$ then a component Bayes estimator of $\phi(\theta)$ is the ratio of the corresponding Bayes estimators of $w(\theta).\phi(\theta)$ and $w(\theta)$ wrt the squared error loss. Since $w\phi$ is continuous and $w_* > 0$, the $L_1(E)$ case of Lemma A.1 with $L = 2w^* |\phi|^* / w_*$ and two applications of Theorem 1.5.1 imply that $E|\hat{t}_{\alpha} - \tilde{t}_{\alpha}| \longrightarrow 0$ uniformly in α and $\underline{\theta}$ where $\hat{t}_{\alpha}$ and $\tilde{t}_{\alpha}$ refer here to the weighted loss. As before this is sufficient to conclude the asymptotic optimality of $\hat{\underline{t}}$ since (1.2.6) holds with w^* multiple of the RHS. The same conclusion holds in the sequence compound and the empirical Bayes versions.

3. An interesting question seems to be how far we can relax the compactness assumption on the component parameter space. It is known that we can not always go up to the natural parameter space. An example where

no a.o. compound estimator exists is the Poisson family with unbounded parameter set. See Gilliland (1968, Section 3.3) for a proof.

4. Under the assumption $\mu_{-1} << \mu$ instead of $\mu_1 << \mu$ we can use the transformation $T(x) = -x$ in Remark 1 to obtain admissible, a.o. rules. An example where none of these holds is provided by the Binomial family and it is well known that in this case even the empirical Bayes problem has no a.o. solution.

5. A possible open question is whether the condition Λ has full support is necessary. Another interesting problem is to find examples of Λ for which a good lower bound (as a function of δ) to the quantity in Lemma 1.4.3. can be obtained so that a rate of convergence in the asymptotic optimality of $\hat{\underline{t}}$ can be established.

APPENDIX

APPENDIX

Here we present a few results of possible independent interest. They are used as technical tools in the body of the thesis.

1. On bounding the difference of two ratios.

Lemma A.1. For $\langle y,z,Y,Z,L\rangle \in \mathbb{R}^5$ and $z \neq 0 \leq L$,

$$(1.1)\qquad |z|\,\{|\tfrac{y}{z} - \tfrac{Y}{Z}| \wedge L\} \leq |z|^{-1}\,|yZ - zY| + L(|z| - |Z|)_+$$
$$\leq |y - Y| + (\,|\tfrac{y}{z}| + L)\,|z - Z|.$$

Proof. The first inequality holds because the RHS, less $|Y|$ if $Z = 0$, is the $|\frac{Z}{z}|$, $(1 - |\frac{Z}{z}|)_+$ weighted average of quantities whose minimum is the LHS. The second inequality follows by triangle inequality weakenings,

$$|yZ - zY| \leq |z|\,|y - Y| + |y|\,|z - Z|$$

and

$$(|z| - |Z|)_+ \leq |z - Z|\ ,$$

in the two LHS terms.

Remark A.1. Division by $|z|$ in (1.1) yields a pointwise improvement on a lemma of Singh (1974, Lemma A.2). When $\langle y,z,Y,Z\rangle$ are measurable functions on a space with an integral J, his lemma itself (and its extension, his Remark A.2) is further improved by the corollary to ours resulting from the subadditivity of the norm or metric distance from 0 in $L_\gamma(J)$ according as $\gamma \in [1,\infty]$ or $(0,1)$.

2. On uniform convergence of convex functions.

Lemma A.2. If $\{f_n\}$ is a sequence of convex functions on an interval $I \subset \mathbb{R}$ converging to a continuous real function f pointwise on I, then the convergence is uniform on compact subsets of I.

Proof. Let K be a compact subset of I, an interval w/o.l.g. Partition K into intervals of equal length.

Let ϵ denote the maximum of the oscillations of f within the subintervals. Let η_n denote the maximum of $|f_n - f|$ on the endpoints of the subintervals.

On a given subinterval with endpoints a and b, bound f_n above by its chord to obtain

$$f_n \leq f_n(a) \vee f_n(b) ;$$

bound f_n below by the line extending the chord from an adjacent interval to obtain

$$f_n \geq f_n(a) - | f_n]_{2a-b}^{a} | .$$

Thus

$$\bigvee_K |f_n - f| \leq 3\eta_n + 2\epsilon \longrightarrow 2\epsilon \text{ as } n \longrightarrow \infty .$$

The proof is now over since the uniform continuity of f over K permits arbitrarily small $\epsilon > 0$ based on the number of subintervals.

3. A uniform L_1– LLN for independent random continuous functions.

Let (S,d) be a compact metric space. Let $\| \ \|$ denote the sup norm on $C(S)$.

Let $\{P_\nu : \nu \in \mathcal{N}\}$ be an arbitrary family of probability measures. (We use the measure to denote the corresponding expectation too and use the superscript $^{(\nu)}$ to denote deviations of random elements from the values of their P_ν expectations.)

Let A_n denote the uniform expectation on $\{1,\dots,n\}$. (If $\{f_k\}$ is a sequence of elements of a linear space, $A_n f_.$ will be denoted by $\bar{f}$ when convenient.) Note that A_n commutes with $^{(\nu)}$. Let $*$ denote the (iterated) operation $\limsup_n \bigvee_\nu A_n \times P_\nu$ and note that it is subadditive.

Theorem A.3. If

(i) Under each P_ν, H_1, H_2, ... are independent $C(S)$ valued random elements with expectations $\in \mathbb{R}^S$ ($(P_\nu H_k)(s) = P_\nu H_k(s)$ $\forall$ k and s),

(ii) $$* \ (\|H^{(\nu)}_{.}\| - M)_+ \ \downarrow 0 \ \text{ as } \ M \uparrow \infty \ ,$$

(iii) $\forall$ $\epsilon > 0$ and $s \in S$, with $V_{s\rho\nu k} = \vee\{|H^{(\nu)}_k]^t_s| : d(s,t) < \rho\}$,

$$* \ [V_{s\rho\nu.} > \epsilon] \ \downarrow 0 \ \text{ as } \ \rho \downarrow 0 \ ,$$

then

$$* \ \|\overline{H}^{(\nu)}\| \ = \ 0 \ .$$

Proof. If card(S) = 1, $\| \ \|$ reduces to $| \ |$ and (iii) is vacuously satisfied. The H_k are then real valued random variables and will be denoted by X_k.

For $M \in (0,\infty)$ represent (as in the proof of Theorem 2.3.9 of Fabian and Hannan (1985), which the present real case greatly strengthens)

$$(3.1) \qquad X_k^{(\nu)} = U_k^{(\nu)} + P_\nu U_k + W_k$$

with U_k the projection of LHS into $[-M,M]$, so that

$$P_\nu|\overline{X}^{(\nu)}| \le P_\nu|\overline{U}^{(\nu)}| + |P_\nu\overline{U}| + P_\nu|\overline{W}|$$

and

$$|P_\nu\overline{U}| = |P_\nu\overline{W}|.$$

Thus

$$(3.2) \qquad P_\nu|\overline{X}^{(\nu)}| \le M/\sqrt{n} + 2\,P_\nu|\overline{W}|$$

(since $(P_\nu|\overline{U}^{(\nu)}|)^{\,2} \le P_\nu(\overline{U}^{(\nu)})^2 \le M^2/n$). Since $|W_k| = (|X_k^{(\nu)}| - M)_+$, $*|\overline{W}| \le *|W_{\cdot}| \downarrow 0$ as $M \uparrow \infty$ by (ii) and thus

$$(3.3) \qquad *\,|\overline{X}^{(\nu)}| = 0$$

follows from (3.2).

Let $\epsilon > 0$, $s \in S$. Since

$$(3.4) \qquad V_{s\rho\nu k} \le 2\|H_k^{(\nu)}\|,$$

$$(3.5) \qquad V_{s\rho\nu k} \leq \epsilon + M\,[V_{s\rho\nu k} > \epsilon] + (2\,\|H_k^{(\nu)}\| - M\,)_+$$

and by subadditivity

$$(3.6) \qquad \begin{aligned} {}^*\,V_{s\rho\nu.} &\leq \epsilon + M\ {}^*\,[V_{s\rho\nu.} > \epsilon] + {}^*\,(2\,\|H_k^{(\nu)}\| - M\,)_+ \\ &\leq 2\epsilon + M(\epsilon)\ {}^*\,[V_{s\rho\nu.} > \epsilon] \end{aligned}$$

by choice of $M = M(\epsilon)$ in (ii). And so

$$(3.7) \qquad {}^*\,V_{s\rho(s,\epsilon)\nu.} \leq 3\epsilon \quad \text{by choice of } \rho = \rho(s,\epsilon) \text{ in (iii).}$$

By compactness of S, ∃ a finite cover of S by spheres indexed by $o_i = (s_i, \rho_i)$ with $\rho_i = \rho(s_i,\epsilon)$, for $i = 1, \ldots, g$. Then

$$(3.8) \qquad \begin{aligned} \|\overline{H}^{(\nu)}\| &\leq \bigvee_1^g \{\ |\overline{H}^{(\nu)}(s_i)| + |\overline{V}^{(\nu)}_{o_i\nu}| + P_\nu \overline{V}_{o_i\nu}\ \}, \\ &\leq \sum_1^g \{\ |\overline{H}^{(\nu)}(s_i)| + |\overline{V}^{(\nu)}_{o_i\nu}|\ \} + \bigvee_1^g P_\nu \overline{V}_{o_i\nu} \ . \end{aligned}$$

From (3.4) it follows that $|V^{(\nu)}_{o_i\nu k}| \leq 2\,(\ \|H_k^{(\nu)}\| + P_\nu\|H_k^{(\nu)}\|\ )$ and hence

$$(3.9) \qquad P_\nu(\ |V^{(\nu)}_{o_i\nu k}| - M\ )_+ \leq 2\ P_\nu(\ 2\ \|H_k^{(\nu)}\| - M/2\ )_+ \ .$$

Thus the $V^{(\nu)}_{o_i\nu k}$, as well as the $H_k^{(\nu)}(s_i)$, inherit (ii). Since (i) obviously holds for both sequences, so does (3.3). So it follows from (3.8) and subadditivity of $*$ that

$$(3.10) \qquad * \|\overline{H}^{(\nu)}\| \leq * \bigvee_1^g P_\nu \overline{V}_{o_i \nu} = \limsup_n \bigvee_1^g \gamma_\nu P_\nu \overline{V}_{o_i \nu} .$$

Since $\limsup_n$ commutes with $\bigvee_1^g$ for finite g, we get RHS (3.10) $\leq 3\epsilon$. The proof ends since ϵ is arbitrary.

Remark A.3. Let (ii+) and (iii+) denote (ii) and (iii) respectively without the centerings $^{(\nu)}$.

Let (ii+) hold. Then, since $\|P_\nu H_k\| \leq P_\nu \|H_k\|$ and hence $(\|P_\nu H_k\| - M)_+ \leq P_\nu (\|H_k\| - M)_+$, (ii+) holds with H_k replaced by $P_\nu H_k$. This along with (ii+) then gives (ii) (since $(a + b - 2M)_+ \leq (a - M)_+ + (b - M)_+$).

Let (ii+) and (iii+) hold. Then, since $\vee\{ \, |P_\nu H_k]_s^t \, | : d(t,s) < \rho\} \leq P_\nu V_{s\rho k}$, $A_n[P_\nu V_{s\rho.} > \epsilon] \leq \epsilon^{-1} A_n P_\nu V_{s\rho.}$. So via (3.6+) and (ii+) $0 = \lim_{\rho \downarrow 0} * V_{s\rho.}$. Thus (iii+) holds with H_k replaced by $P_\nu H_k$. This along with (iii+) then gives (iii) (since $[\, \vee|H+G| > \epsilon] \subset [\, \vee|H| > \epsilon/2] + [\, \vee|G| > \epsilon/2]$).

4. Admissibility of Bayes estimators in the compound problem under squared error loss.

Let $\{P_\theta\}$, $\theta\in\Theta$ be the component distributions. Consider the compound problem of estimating $\underline{\phi}$ under squared error loss $L(\underline{t},\underline{\theta}) = n^{-1}\sum_{\alpha=1}^{n}(t_\alpha - \phi(\theta_\alpha))^2$ for any function ϕ on Θ. Let $P_{\underline{\theta}} = \times_{\alpha=1}^{n} P_{\theta_\alpha}$ for $\underline{\theta}\in\Theta^n$. Let ζ be a prior on $\underline{\theta}$. Denote the joint distribution $\zeta\circ P_{\underline{\theta}}$ on $<\underline{x},\underline{\theta}>$ by Q. Then the marginal of $\underline{x}$ is $Q\underline{x}^{-1} = \int P_{\underline{\theta}}d\zeta$. For a function f on Θ^n let $Q_{\underline{x}}f(\underline{\theta})$ denote the class of conditional expectations of $f(\underline{\theta})$ given $\underline{x}$.

Lemma A.4. If ζ is such that $P_{\underline{\theta}} \ll Q\underline{x}^{-1}$ $\forall\ \underline{\theta}\in\Theta^n$ then every Bayes estimator versus ζ is admissible.

Proof. First consider the set compound case.

Fix an $\alpha\in\{1,...,n\}$. Then $Q(t_\alpha - \theta_\alpha)^2$ is minimal iff $t_\alpha(\underline{x}) \in Q_{\underline{x}}\phi(\theta_\alpha)$. Hence t_α is determined up to $Q\underline{x}^{-1}$ null sets and so by the assumption of the lemma has unique risk $\underline{\theta} \rightsquigarrow \int(Q_{\underline{x}}\phi(\theta_\alpha) - \phi(\theta_\alpha))^2 dP_{\underline{\theta}}$. Thus, since $\alpha\in\{1,...n\}$ is arbitrary, the compound Bayes estimators have the unique compound risk $\underline{\theta} \rightsquigarrow n^{-1}\sum_{\alpha=1}^{n}\int(Q_{\underline{x}}\phi(\theta_\alpha) - \phi(\theta_\alpha))^2 dP_{\underline{\theta}}$ and hence are admissible.

For the sequence compound case, the given condition implies that for each $\alpha\in\{1, ..., n\}$, $P_{\underline{\theta}_\alpha} \ll Q\underline{x}_\alpha^{-1}$ $\forall\ \underline{\theta}_\alpha \in \Theta^\alpha$. Hence, by combining the intermediate results in set case with $n = \alpha$ for each α, we get that the sequence compound Bayes estimators have the unique compound risk $\underline{\theta} \rightsquigarrow n^{-1} \sum_{\alpha=1}^{n} \int (Q_{\underline{x}_\alpha} \phi(\theta_\alpha) - \phi(\theta_\alpha))^2 dP_{\underline{\theta}_\alpha}$ and hence are admissible.

BIBLIOGRAPHY

BIBLIOGRAPHY

Basu, D. and Tiwari, R. C. (1982). A note on the Dirichlet process. *Statist. Prob.: Essays in Honor of C. R. Rao.* North–Holland Publishing Comp, 89–103.

Billingsley, Patrick (1968). *Convergence of Probability Measures.* John Wiley & Sons.

Blackwell, David and McQueen, James B. (1973). Ferguson distribution via Polya urn schemes. *Ann. Statist.* **1**, 353–355.

Doksum, Kjell (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Statist.* **2**, 183–201.

Dubins, L. E. and Freedman, D. A. (1966). Random distribution functions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **II.1,** Univ. of California Press, 183–214.

Fabian, Václav and Hannan, James (1985). *Introduction to Probability and Mathematical Statistics.* John Wiley & Sons.

Ferguson, Thomas S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Ferguson, Thomas S. (1974). Prior distributions on space of probability measures. *Ann. Statist.* **2**, 615–629.

Gilliland, Dennis C. (1968). Sequential compound estimation. *Ann. Math. Statist.* **39**, 1890–1904.

Gilliland, Dennis C. and Hannan, James (1986). The finite state compound decision problem, equivariance and restricted risk components. *Adaptive Statistical Procedures and Related Topics*, IMS Lecture Notes – Monograph Series **8**, 129–145.

Gilliland, Dennis C., Hannan, James and Huang, J. S. (1976). Asymptotic solutions to the two state compound decision problem, Bayes versus diffuse priors on proportions. *Ann. Statist.* **4**, 1101–1112.

Gilliland, Dennis C., Boyer, John E. and Tsao, How Jan (1982). Bayes empirical Bayes : finite parameter case. *Ann. Statist.* **10**, 1277–1282.

Hannan, James F. (1957). Approximation to Bayes risk in repeated play. *Contribution to the Theory of Games* **3**, *Ann. Math. Studies*, No. 39, Princeton University Press, 97–139.

Hannan, J. (1960). Consistency of maximum likelihood estimation of discrete distributions. *Contributions to Prob. Statist.*, Stanford Univ. Press, 249–257.

Inglis, James (1973). Admissible decision rules for the compound problem. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Inglis, James (1977). Admissible decision rules for the compound decision problem : the two action two state case. *Ann. Statist.* **7**, 1127–1135.

Kuo, Lynn (1986). A note on Bayes empirical Bayes estimation by means of Dirichlet process. *Stat. Prob. Let.* **4**, 145–150.

Lehmann, E. L. (1959, 1986). *Testing Statistical Hypotheses.* John Wiley & Sons.

Meeden, Glen (1972). Some admissible empirical Bayes procedures. *Ann. Math. Statist.* **43**, 96–101.

Robbins, Herbert (1951). Asymptotically sub–minimax solutions of compound problems. *Proc. Second Berkeley Symp. Math. Statist. Prob.*, Univ. of California Press, 131–148

Robbins, Herbert (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, Univ. of California Press, 157–163.

Rolph, John E. (1968). Bayesian estimation of mixing distributions. *Ann. Math. Statist.* **39**, 1289–1302.

Sethuraman, Jayaram and Tiwari, Ram C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. *Statistical Decision Theory and RelatedTopics,* **III.2**, Academic Press, 305–315.

Singh, Radhey Shyam (1974). Estimation of derivatives of average μ–densities and sequence–compound estimation in exponential families. Ph.D. Thesis, Dept. of Statistics and Probability, Michigan State University

Vardeman, Stephen B. (1978). Admissible solutions of finite state sequence compound decision problems. *Ann. Statist.* **6**, 673–679.