Exact Minimax Strategies for Predictive Density Estimation, Data Compression, and Model Selection

Feng Liang and Andrew Barron, *Senior Member, IEEE*

***Abstract*—For location and scale families of distributions and related settings of linear regression, we determine minimax procedures for predictive density estimation, for universal data compression, and for the minimum description length (MDL) criterion for model selection. The analysis gives the best invariant and indeed minimax procedure for predictive density estimation by directly verifying extended Bayes properties or, alternatively, by general aspects of decision theory on groups which are shown to simplify in the case of Kullback–Leibler loss. An exact minimax rule is generalized Bayes using a uniform (Lebesgue measure) prior on the location and log-scale parameters, which is made proper by conditioning on an initial set of observations.**

***Index Terms*—Haar measure, Hunt–Stein, invariance, Kullback–Leibler divergence, minimum description length (MDL), minimax risk, predictive density estimation, universal coding.**

I. Introduction

LET $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ be a random vector to which we wish to assign a distribution given observed data $Y = (Y_1, \ldots, Y_m)$. For each model it is assumed that there is a parametric family of distributions $P_{Y|\theta}$ and $P_{\tilde{Y}|Y,\theta}$ with densities $p(y|\theta)$ and $p(\tilde{y}|y,\theta)$ depending on a d-dimensional parameter vector θ which takes values in a parameter space Θ, possibly consisting of all of $\mathbb{R}^d$. To each choice of predictive distribution $Q_{\tilde{Y}|Y}$ with density $q(\tilde{y}|y)$ we incur a loss given by the Kullback–Leibler information divergence

$$D(P_{\tilde{Y}|Y,\theta}||Q_{\tilde{Y}|Y}) = \int p(\tilde{y}|y,\theta) \log \frac{p(\tilde{y}|y,\theta)}{q(\tilde{y}|y)} d\tilde{y} \qquad (1)$$

and a resulting risk $R(\theta, Q) = \mathbb{E}_{Y|\theta} D(P_{\tilde{Y}|Y,\theta}||Q_{\tilde{Y}|Y})$. Our interest is in the minimax risk

$$R = \min_{Q} \max_{\theta \in \Theta} R(\theta, Q) \qquad (2)$$

and in the determination of a predictive distribution $Q_{\tilde{Y}|Y}$ that achieves it for location and scale families of distributions. Also of interest is the maximin risk defined as the supremum over choices of proper prior distributions on Θ of the Bayes average risk.

We provide exact solution to this minimax problem for certain families of densities parameterized by location or scale. To show exact minimaxity, we use methods from statistical decision theory adapted to the information-theoretic choice of loss. In particular, we identify constant risk procedures and find among them the procedure which is extended Bayes and hence minimax. Auxiliary to this demonstration of the minimax procedure, it follows that the minimax value is equal to the maximin value for such location and scale problems conditioning on Y.

Implications are discussed for predictive density estimation, for universal data compression, and for the minimum description length (MDL) criterion.

A. Density Estimation

In density estimation, our aim is to estimate the density function for $\tilde{Y}$ using the data Y in the absence of knowledge of θ. Estimators $q(\tilde{y}|y)$ are required to be nonnegative and to integrate to one for each y, and as such can be interpreted as predictive densities for $\tilde{y}$ given y. The risk function is the expected Kullback–Leibler loss $R(\theta, q)$ defined before. It may be customary to use plug-in type estimators $q(\tilde{y}|y) = p(\tilde{y}|\hat{\theta}(y))$, however, one finds that optimal density estimators (from Bayes and minimax perspectives) take on the form of an average of members of the family with respect to a posterior distribution given y. We remind the readers of the Bayes optimality property: with prior w and Kullback–Leibler loss, the Bayes risk $R_w(q) = \int R(\theta, q) w(\theta) d\theta$ is minimized by choosing q to be the Bayes predictive density

$$p_w(\tilde{y}|y) = \int p(\tilde{y}|y,\theta) w(\theta|y) d\theta = \frac{\int_\Theta p(y, \tilde{y}|\theta) w(\theta) d\theta}{\int_\Theta p(y|\theta) w(\theta) d\theta} \qquad (3)$$

as shown in [18], [1], [13], [5].

A procedure is said to be generalized Bayes if it takes the same form as in (3), with a possibly improper prior (i.e., $\int w(\theta) d\theta$ might not be finite), but proper posterior (i.e., $\int p(y|\theta) w(\theta) d\theta$ is finite for each y) [21]. Such generalized Bayes procedures arise in our examination of minimax optimality.

Asymptotics of the Kullback–Leibler risk for large sizes m of the conditioning sample Y and smooth parametric families of densities have been explored by Hartigan in [26]. He identifies the second-order asymptotic risk as a function of the parameters and the choice of prior. His asymptotics provide a differential inequality which may be used to identify asymptotically minimax procedures, as further illuminated in [2]. A formulation of cumulative Kullback–Leibler risk was considered earlier in [14] and asymptotically minimax procedures were identified therein. Here we establish for particular types of parametric families

Manuscript received June 24, 2002; revised August 2, 2004.

F. Liang is with the Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 USA(e-mail: feng@stat.duke.edu).

A. Barron is with the Department of Statistics, Yale University, New Haven, CT 06520 USA (e-mail: andrew.barron@yale.edu).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2004.836922

exact minimax procedures, valid for all finite sample sizes, not only asymptotically.

For location families with Kullback–Leibler loss, the minimax procedure $q^*(\tilde{y}|y)$ is the generalized Bayes procedure using a uniform (Lebesgue) prior. A similar conclusion holds for a univariate scale parameter $\theta \neq 0$ with the minimax procedure using a uniform prior on $\log|\theta|$. Likewise, when one has both multivariate location ($\theta_1 \in \mathbb{R}^d$) and univariate scale ($\theta_2 \neq 0$) parameters, the minimax procedure uses a Lebesgue product measure for θ_1 and $\log|\theta_2|$. These procedures have constant risk. As we shall discuss, there is more than one way to proceed in the establishment of minimaxity. Here we emphasize simple information-theoretic identities and inequalities to show the procedures are extended Bayes.

Families defined by other groups of transformations including linear transformations and affine transformations, may also be addressed in part by the information-theoretic techniques. For theses families, the procedure that is identified is minimax among invariant procedures, but not necessarily minimax over all procedures. Additionally, in Appendix D we point to general group-theoretic abstraction which yields the prior of best invariant procedures.

In the location family case, as we have said, the minimax procedure is based on the uniform prior. This is so no matter whether one is doing parameter estimation or density estimation. In the case of parameter estimation with squared error loss, the use of this prior produces Pitman's procedure which is best invariant [42] and minimax [23]. We note a relationship between these estimators. If the family is such that given θ the random variable $\tilde{Y}$ (with $n = 1$) has mean θ, then (for every conditioning size m) our minimax density estimator has mean $\int \tilde{y} q^*(\tilde{y}|y)d\tilde{y}$ equal to the posterior mean of θ given Y. That is, our minimax density estimator is a density function centered (in mean) at Pitman's estimator of location. For scale problems, Pitman's [42] best invariant procedures (for estimation of various functions of scale with corresponding invariant loss functions) do naturally involve the same prior (which makes the log-scale be uniform), however, the scale of our minimax density estimates for Kullback–Leibler loss does not appear to correspond to any of the standard Pitman estimators of scale.

B. Data Compression and Information Capacity

An objective of data compression is to provide a uniquely decodable code for data $\tilde{Y}$, given the value of Y. Such codes correspond to probability measures for $\tilde{Y}$ given Y (or innocuously more generally subprobability measures summing to not more than 1) via the Kraft–McMillan theorem [15]. If θ is known, codes of optimal expected length are based on the true conditional distribution $P_{\tilde{Y}|Y,\theta}$, whereas in universal data compression [45], [34], [51], [18], [15], [5], without knowledge of the parameter θ, a choice of predictive distribution $Q_{\tilde{Y}|Y}$ is used to construct the code instead.

The expected Kullback–Leibler loss arises as the excess average code length (redundancy). Indeed, If $\tilde{Y}$ is discrete, $\log_2 1/q(\tilde{y}|y)$ provides the code length in bits and the redundancy

$$\mathbb{E}_{\tilde{Y}|Y,\theta}\left[\log_2 1/q(\tilde{Y}|Y) - \log_2 1/p(\tilde{Y}|Y,\theta)\right] \qquad (4)$$

is the Kullback–Leibler loss. If $\tilde{Y}$ is continuous valued with a density, such redundancy arises for each choice of discretization (that is, for each rule of quantization or partition of the space for $\tilde{Y}$). Now if the codes are constructed from a predictive distribution that has a density function $q(\tilde{y}|y)$, then the integral giving the Kullback–Leibler loss as in expression (1) arises as the least upper bound (supremum) of the redundancies for the collection of all discretizations of $\tilde{Y}$ (see, e.g., [41], [35], [24] or references cited therein). Furthermore, this Kullback–Leibler integral arises as the limit of the redundancies as the discretizations become infinitessimally fine [16], [17], [56]. Thus, it is now customary in universal data compression, following [45], to refer to the Kullback–Leibler loss as the redundancy for both discrete and continuous settings.

Of particular interest in universal data compression is the determination of a distribution Q that provides the minimax redundancy [18], [49], [19], [14]. The minimax redundancy in the unconditional case is $R = \min_Q \max_\theta D(P_{\tilde{Y}|\theta}\|Q_{\tilde{Y}})$ and in the conditional case it is given by (2).

Bayes optimal codes play a central role in universal data compression. For each proper prior w on the parameter space, the Bayes codes are based on the Bayes predictive density (as in (3)) and the Bayes risk is equal to Shannon's conditional mutual information between the parameter and the sample

$$R_w = I_w(\Theta;\tilde{Y}|Y) = E_{\theta,\tilde{Y},Y} \log p(\tilde{Y}|Y,\theta)/p_w(\tilde{Y}|Y).$$

The Bayes story for unconditional redundancy is comparable. In the absence of prior knowledge, to focus greater attention on some parts of the parameter space, one may study the maximin value $C = \sup_w R_w$, which is recognized as the information capacity of the family of distributions in the unconditional case [18], [14], [15]. In accordance with game theory and statistical decision theory [7], [21], the minimax and maximin values agree, $R = C$, as shown in the unconditional setting in [22], [49], [19], [27].

Minimaxity of the unconditional redundancy is a workable criterion for parametric families with parameter spaces that are not too extensive [34], [58], [14], [48]. In this literature, a key quantity, shown to determine the minimax unconditional redundancy in large sample cases, is the integral over the parameter space of the square root of Fisher information. Indeed, Jeffreys prior (proportional to the square root of the determinant of the Fisher information) provides asymptotically maximin and minimax procedures for the unconditional redundancy, if the parameter is restricted such that the Jeffreys integral is finite [14]. (This prior is historically important [30], [25] because of a local invariance property—small diameter Kullback–Leibler balls have approximately the same prior probability in different parts of the parameter space.) Unfortunately, for various unbounded parameter spaces, including location and scale families, the minimax unconditional redundancy is infinite. We shall see that conditioning on initial data Y not only provides a finite minimax expected redundancy, but also an *exact* minimax procedure (even in the case of finite samples) for location and scale families.

There is a simple relationship between unconditional and conditional redundancies. The total Kullback divergence (unconditional redundancy for the description of Y and $\tilde{Y}$) has a decomposition as a sum of a divergence for the description of

Y (this is the part that is unbounded as a function of location or scale parameters) plus an expected conditional divergence for the description of $\tilde{Y}$ given Y (this is the part made constant in our setting for each conditioning sample size m that is not too small). Procedures that are minimax for unconditional and conditional redundancy can be rather different. Indeed, a rule that produces constant total redundancy (if possible) does not necessarily make the terms in the decomposition constant as well. A consequence is that minimax conditional redundancy procedures (or their limits as studied in [26], [2]) do not have to match with the use of Jeffreys prior. It happens for location or for scale that our exact minimax rules for conditional redundancy do correspond to the use of Jeffreys improper prior. However, in joint location and scale cases the exact minimax procedure uses a prior that is not proportional to the square-root determinant of the information.

Universal data compression provides the foundation for the MDL criterion for model selection, as arise in particular in problems of linear regression, which provides location and scale type families to which our theory is adapted.

C. Minimum Description Length

Consideration of model selection in linear regression from the MDL perspective began in Rissanen [44]. Suppose we have a total of N observations Y_i predicted using given d-dimensional explanatory vectors x_i for $i = 1, 2, \ldots, N$ with normal errors $N(0, \sigma^2)$. If σ^2 is fixed and θ is estimated, these models lead to description length criteria of the form

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^t\hat{\theta})^2 + \frac{1}{2}\log|S_N| + c \tag{5}$$

where $S_N = \sum_{i=1}^{N} x_i x_i^t$ is the d by d information matrix. When several candidates are available for the explanatory variables x, the model selection criterion picks out the subset of the variables that leads to the shortest total description length achieving the best tradeoff between sum of squared errors and the complexity of the model.

In Rissanen's original two-stage code formulation, the parameter θ is estimated by least squares (maximum likelihood) and his complexity penalty term corresponds to the length of description of the coordinates of the maximum-likelihood estimate $\hat{\theta}$ to certain precision. Various values for c have arisen in the literature corresponding to different schemes of quantization of θ, or to the use of mixture or predictive coding strategies rather than two-stage codes [47], [5]. Asymptotics in N have also played a role in justifying the form of the criterion [5]. Rissanen has also developed Bayes and predictive formulations of the MDL criterion [47]

Previous work in information theory has identified the role of the information matrix in asymptotically optimal two-stage codes [3], in stochastic complexity (Bayes mixture codes) [3], [46], [13], and in asymptotically minimax codes [14], [48]. For the regression problem, Jeffreys prior is improper, as the information matrix is constant (not depending on θ), commensurate with infinite minimax redundancy on unbounded parameter spaces.

In this paper, we obtain an explicit exact minimax optimal MDL criterion for any such regression problem, when one conditions on m initial observations with m at least as large as the parameter dimension d.

The paper is arranged as follows. In Section II, we study the best invariant predictive densities, which have constant risk. An information inequality is shown to reveal the best invariant predictive density as the one which is Bayes with an uniform prior. In Section III, we prove that this procedure is minimax with Kullback–Leibler loss for location families, scale families, and multivariate location with univariate scale families, by directly demonstrating an extended Bayes property. Section IV extends the results to the linear regression problem. Section V considers general aspects of decision theory on groups which are shown to simplify in the case of Kullback–Leibler loss. Discussion and conclusion are given in Section VI followed by Appendices.

II. Best Invariant Strategies

Our goal here is to find the best invariant estimator or coding strategy $q^*(\tilde{y}|y)$.

Consider first location families. We are to observe $Y = (Y_1, \ldots, Y_m)$ and want to encode or provide predictive distribution for the future observations $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$, where $Y_i = Z_i + \theta$, $\tilde{Y}_i = \tilde{Z}_i + \theta$ with unknown $\theta \in \mathbb{R}^d$. We assume that $Z = (Z_1, \ldots, Z_m)$ and $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)$ have a known joint density $p_{\tilde{Z},Z}$. Then the joint density for Y and $\tilde{Y}$ is given by

$$p(y, \tilde{y}|\theta) = p_{Z,\tilde{Z}}(y - \theta, \tilde{y} - \theta).$$

We use $y - \theta$ and $\tilde{y} - \theta$ as shorthand notations for $y_1 - \theta, \ldots, y_m - \theta$ and $\tilde{y}_1 - \theta, \ldots, \tilde{y}_n - \theta$, respectively. When the context is clear, we will write $p_{Z,\tilde{Z}}$ as p and write $\mathbb{E}_{Z,\tilde{Z}} f(Z, \tilde{Z})$ as $\mathbb{E} f(Z, \tilde{Z})$.

We allow our procedures $q(\tilde{y}|y)$ to be subprobability densities, that is, $\int q(\tilde{y}|y)d\tilde{y} \leq 1$. This provides a closer correspondence with Kraft's inequality for the coding story. Nonetheless, the Kullback–Leibler loss is always improved by renormalization of a strict subprobability density, though the renormalization might impact the invariance. Nonetheless, the best invariant rule (even within the class of subprobabilities) will be seen to be a particular probability density.

Definition 1: A procedure q is *invariant* under location shift, if for each $a \in \mathbb{R}^d$ and all $y, \tilde{y}$, $q(\tilde{y}|y + a) = q(\tilde{y} - a|y)$.

That is, adding a constant a to the observations $y = (y_1, \ldots, y_m)$ shifts the density estimator for $\tilde{y}$ by the same amount a. Consequently, if we shift both y and $\tilde{y}$ by the same amount, the value of $q(\tilde{y}|y)$ is unchanged

$$q(\tilde{y} + a|y + a) = q(\tilde{y}|y). \tag{6}$$

Proposition 1: Invariant procedures have constant risk.

Proof: By the invariance of q, the risk $R(\theta, q)$ is equal to

$$\mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(Y - \theta|\tilde{Y} - \theta)}{q(\tilde{Y} - \theta|Y - \theta)} = \mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z}|Z)}{q(\tilde{Z}|Z)}, \tag{7}$$

a quantity not depending on θ, therefore, q has constant risk.

Now we derive the best invariant procedure. The idea is to express the risk in terms of transformed variables that are invariant to the location shift. Applying the invariance property (6) with $a = -Z_1$ in (7), we obtain

$$R(\theta, q) = \mathbb{E} \log \frac{p(\tilde{Z}|Z)}{q(\tilde{U}|0, U)}$$

where $\tilde{U} = \tilde{Z} - Z_1$ and $U_i = Z_{i+1} - Z_1$ for $i = 1, \ldots, m-1$, which have a distribution not depending on θ. We will show that the conditional density of $\tilde{U}$ given $U = (U_1, \ldots, U_{m-1})$ provides the optimal q. Indeed, for any q, the risk satisfies

$$R(\theta, q) \geq \mathbb{E} \log \frac{p(\tilde{Z}|Z)}{p(\tilde{U}|U)} \tag{8}$$

because the difference

$$\mathbb{E} \log \frac{p(\tilde{U}|U)}{q(\tilde{U}|0, U)} = \mathbb{E}_U \left[\mathbb{E}_{\tilde{U}|U} \log \frac{p(\tilde{U}|U)}{q(\tilde{U}|0, U)} \right]$$

is an expected Kullback–Leibler divergence that is greater than or equal to zero, and it is equal to zero (i.e., achieves the smallest risk) if and only if $q(\tilde{u}|0, u) = p(\tilde{u}|u)$.

Next we solve for $p(\tilde{u}|u) = p(u, \tilde{u})/p(u)$. Note that the mapping from $(Z, \tilde{Z})$ to $(Z_1, U, \tilde{U})$ has unit Jacobian. So the joint density $p(u, \tilde{u})$ is given by $p_{\tilde{Z},Z}(z_1, u + z_1, \tilde{u} + z_1)$. Integrating out z_1, we obtain

$$p(u, \tilde{u}) = \int p_{Z,\tilde{Z}}(z_1, u + z_1, \tilde{u} + z_1) dz_1. \tag{9}$$

Use the fact that $u_i = y_{i+1} - y_1$ for $i = 1, \ldots, m-1$ and $\tilde{u} = \tilde{y} - y_1$, then (9) is equal to

$$\int p_{Z,\tilde{Z}}(z_1, y_2 - y_1 + z_1, \ldots, y_m - y_1 + z_1, \tilde{y} - y_1 + z_1) dz_1.$$

Changing the variable of integration z_1 to $\theta = y_1 - z_1$, we have (9) equal to $\int p(y, \tilde{y}|\theta) d\theta$. Similarly, $p(u) = \int p(y|\theta) d\theta$. Thus, the conditional density for $\tilde{u}$ given u (expressed as a function of y and $\tilde{y}$) is the ratio

$$p(\tilde{u}|u) = \frac{\int p(y, \tilde{y}|\theta) d\theta}{\int p(y|\theta) d\theta} \tag{10}$$

which we denote by $q^*(\tilde{y}|y)$. One can check that q^* is an invariant procedure under location shift. Our analysis at inequality (8) and the following show that this predictive density q^* has the smallest risk among all invariant estimators. It is also the unique best invariant one due to the strict convexity of the Kullback–Leibler loss.

Proposition 2: The unique best invariant predictive density for a location family is

$$q^*(\tilde{y}|y) = \frac{\int p(y, \tilde{y}|\theta) d\theta}{\int p(y|\theta) d\theta}. \tag{11}$$

The procedure q^* which we have showed to be the best invariant can be interpreted as a generalized Bayes procedure with uniform (improper) prior $w(\theta)$ on $\mathbb{R}^d$ (Lebesgue measure) for location families. Bayes prediction densities are not invariant in general, except for certain improper priors, identified in [25] as *relatively invariant* priors, for which $w(\theta + t) = c(t)w(\theta)$, e.g., $w(\theta) = ce^{a\theta}$. A corollary then of Proposition 2 is that the relatively invariant prior with the smallest constant risk is the uniform prior on $\mathbb{R}^d$ ($w(\theta) = c$).

Analysis of groups of transformations provides means to study invariant estimators in more general settings. For the location families we consider here, the group of transformations is the location shift operating on the sample space of observations Y and $\tilde{Y}$ and it induces group operations on the spaces of the parameter θ and of the procedure q. The uniform prior, which we identify to provide the best invariant procedure, is an invariant prior (unique up to a multiplicative constant) for the parameter space under location shift. Under general conditions, the best invariant estimator is the generalized Bayes estimator using the (right) invariant prior ("right" is specified here because some transformations, such as the affine groups, will yield different priors when being applied on the right or on the left). For such general treatment, one may see [6, pp. 410–412] which we specialize to Kullback–Leibler loss in our Appendix D.

The demonstration of best invariance we give here (in Section II) is based on the information inequality as in inequality (8), which directly quantifies the excess risk of any nonoptimal invariant rule as a Kullback–Leibler divergence, in a manner analogous to Pitman's analysis for squared-error loss [42] (cf. [21, pp. 186–187]). As in the squared-error case, the information inequality method does extend to other cases beyond that of location families, as subsequently given in Proposition 3.

Some particular cases we consider include.

1) Linear transformation family: $Y_i = \theta^{-1} Z_i$, $\tilde{Y}_i = \theta^{-1} \tilde{Z}_i$, where θ is a nonsingular $d \times d$ matrix and
$$p(y, \tilde{y}|\theta) = |\theta|^{m+n} p_{Z,\tilde{Z}}(\theta y, \theta \tilde{y}).$$
Specially, when $d = 1$, it is called a univariate *scale* family. A procedure q is invariant under linear transformation if $|b|^n q(b\tilde{y}|by) = q(\tilde{y}|y)$ for any nonsingular $d \times d$ matrix b and all y, $\tilde{y}$.
2) Affine family: $Y_i = \theta_2^{-1} Z_i + \theta_1$, $\tilde{Y}_i = \theta_2^{-1} \tilde{Z}_i + \theta_1$, $\theta_1 \in \mathbb{R}^d$, θ_2 nonsingular $d \times d$ matrix
$$p(y, \tilde{y}|\theta) = |\theta_2|^{m+n} p_{Z,\tilde{Z}}(\theta_2(y - \theta_1), \theta_2(\tilde{y} - \theta_1)).$$
A procedure q is invariant if
$$|b|^n q(b(\tilde{y} - a)|b(y - a)) = q(\tilde{y}|y)$$
for any $a \in \mathbb{R}^d$ and nonsingular $d \times d$ matrix b, and all y, $\tilde{y}$.
3) Multivariate location with univariate scale: same as in the case of affine families with $\theta_1 \in \mathbb{R}^d$, but with scalar $\theta_2 \in \mathbb{R}/\{0\}$
$$p(y, \tilde{y}|\theta) = |\theta_2|^{(m+n)d} p_{Z,\tilde{Z}}(\theta_2(y - \theta_1), \theta_2(\tilde{y} - \theta_1)).$$
A procedure q is invariant if
$$|b|^{nd} q(b(\tilde{y} - a)|b(y - a)) = q(\tilde{y}|y)$$
for any $a \in \mathbb{R}^d$ and nonzero scalar b, and all y, $\tilde{y}$.

Proposition 3: The unique best invariant predictive density is a generalized Bayes (taking the form (3)) with prior $w(\theta) = 1/|\theta|^d$ for a linear transformation family, with prior $w(\theta_1, \theta_2) = 1/|\theta_2|^d$ for an affine family, and with prior

$w(\theta_1, \theta_2) = 1/|\theta_2|$ for a multivariate location with univariate scale family.

For the proof, see Appendix C.

A. Examples

The best invariant estimator q^* is calculated for some examples in which we have m observations $Y = (Y_1, \ldots, Y_m)$ and want to estimate the density for the next observation $\tilde{Y}$. Let $Y_{(i)}$ be the ith-order statistic (the ith smallest value) among $Y_1, \ldots, Y_m$.

- *Shifted exponential family:* $p(\tilde{y}|\theta) = \exp(-(\tilde{y}-\theta))1_{\{\tilde{y}\geq\theta\}}$

$$q^*(\tilde{y}|Y) = \begin{cases} \frac{m}{m+1} e^{-(\tilde{y}-Y_{(1)})}, & \text{if } \tilde{y} \geq Y_{(1)} \\ \frac{m}{m+1} e^{-m(Y_{(1)}-\tilde{y})}, & \text{if } \tilde{y} < Y_{(1)}. \end{cases}$$

- *Uniform family* (with scale parameter):

$$p(\tilde{y}|\theta) = |\theta| 1_{\{0 \leq \theta\tilde{y} \leq 1\}}.$$

Even though θ can take any value in $\mathbb{R}$ except 0, we will know θ is positive or negative once one observation is given. Here suppose Y_1 is positive, then θ ranges from 0 to ∞

$$q^*(\tilde{y}|Y) = \begin{cases} \frac{m}{m+1} \frac{(Y_{(m)})^m}{\tilde{y}^{m+1}}, & \text{if } \tilde{y} > Y_{(m)} \\ \frac{m}{m+1} \frac{1}{Y_{(m)}}, & \text{if } \tilde{y} \leq Y_{(m)}. \end{cases}$$

In Fig. 1, we plot the true density (solid line) and the best invariant estimator q^* (dashed line) for the above two families for $\theta = 1$ and $m = 5$.

- *Normal Location:* $\text{Normal}(\theta, \sigma^2)$, θ unknown, and σ^2 fixed with $p(y|\theta) = \phi_{\sigma^2}(y - \theta)$

$$q^*(\tilde{y}|Y) = \phi_{\sigma^2(1+\frac{1}{m})}(\tilde{y} - \bar{Y}).$$

This is the normal density with mean

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i$$

and a slightly larger variance $\sigma^2(1 + \frac{1}{m})$.

- *Normal location and scale:* $\text{Normal}(\theta, \sigma^2)$, $\theta \in \mathbb{R}^d$, $\sigma^2 \geq 0$, both unknown

$$q^*(\tilde{y}|Y) \propto \left[\frac{\|\tilde{y} - \bar{Y}\|^2}{(1 + 1/m)\hat{s}^2} + (m-1)d\right]^{-md/2}$$

where

$$\hat{s}^2 = \sum_{i=1}^{m} \|Y_i - \bar{Y}\|^2 / ((m-1)d)$$

is the sample variance. Thus,

$$T = (\tilde{Y} - \bar{Y}) / [(1 + \frac{1}{m})\hat{s}^2]^{1/2}$$

is assigned a predictive distribution which is the multivariate t distribution with $(m-1)d$ degrees of freedom.

- *Uniform on Parallelograms:*

$$p(\tilde{y}|\theta_1, \theta_2) = |\theta_2| 1_{(0,1)\times(0,1)}(\theta_2(\tilde{y} + \theta_1))$$

where $\theta_1 \in \mathbb{R}^2$ and θ_2 is a 2×2 matrix with determinant not equal to 0. Conditioning on at least three observations, one can show that the best invariant density estimation q^* is constant in the convex hull spanned by the observations, and tapers down toward zero as one moves away from the convex hull.

III. Minimax and Extended Bayes Strategies

Since the risk is constant for invariant predictive density estimators, the best invariant estimator q^* is the minimax procedure among all invariant procedures. If a constant risk procedure is shown to have an extended Bayes property then it is, in fact, minimax over all procedures, and we shall demonstrate such extended Bayes properties in standard transformation families in this section. Alternatively, Hunt–Stein theory provides means by which to show under some conditions the best invariant estimator is in fact minimax over all rules, as we shall develop in Section V for application to our situation. But first, it is fruitful to see directly by information inequalities given here that in standard transformation families the best invariant rule is extended Bayes. In some respect this is more than a demonstration of minimaxity as it explicitly prescribes sequences of priors for which the Bayes risk is close to the minimax value.

Definition 2: A predictive procedure q is called *extended Bayes*, if there exists a sequence of Bayes procedures $\{p_{w_k}\}$ with proper priors w_k such that their Bayes risk differences go to zero, that is,

$$R_{w_k}(q) - R_{w_k}(p_{w_k}) \to 0, \qquad \text{as } k \to \infty.$$

For a procedure q that has constant risk C, the extended Bayes property is the existence of a sequence of proper priors with Bayes risk $R_{w_k}(p_{w_k})$ converging to C, which implies minimaxity.

Theorem 1: Assume for the location family that at least one of the $Z_1, \ldots, Z_m$ has finite second moment. Then, for any dimension d, under Kullback–Leibler loss, the best invariant predictive procedure $q^*(\tilde{y}|y)$ as in (11) is extended Bayes for sequences of priors that we exhibit. Hence, it is minimax.

Proof: We take a sequence of priors w_k to be the normal distributions with mean zero and variance k (broader tail priors which allow a relaxing of the moment condition are considered in Appendix B).

The Bayes risk difference $R_{w_k}(q^*) - R_{w_k}(p_{w_k})$ is equal to

$$\int \left[R(\theta, q^*) - R(\theta, p_{w_k})\right] w_k(\theta) d\theta = \mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(\tilde{Y}|Y_1, \ldots, Y_m)}{q^*(\tilde{Y}|Y_1, \ldots, Y_m)}$$

where in the expectation $\mathbb{E}_{\tilde{Y},Y}$, also denoted $\mathbb{E}^{w_k}_{Y,\tilde{Y}}$, the distribution of $(Y, \tilde{Y})$ is taken to be a mixture with respect to the prior w_k.

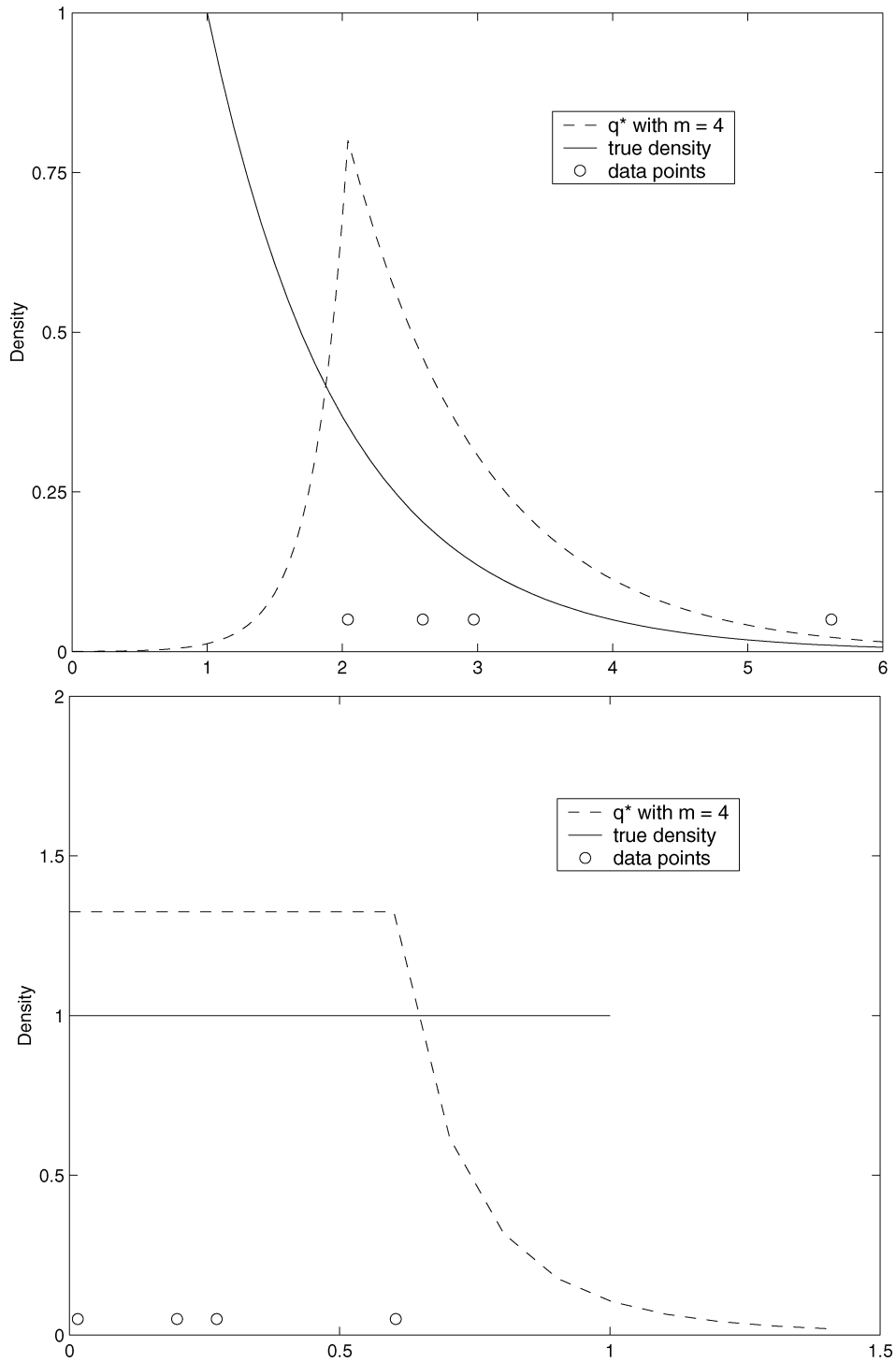

Fig. 1. Plot of true density versus q^*. Top: shifted exponential family. Bottom: uniform family with scale parameter.

By the chain rule of information theory, the Bayes risk difference is less than or equal to the following total Bayes risk difference (conditioning only on Y_1):

$$\mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(Y_2 \ldots, Y_m, \tilde{Y}|Y_1)}{q^*(Y_2, \ldots, Y_m, \tilde{Y}|Y_1)}$$
$$= \mathbb{E}_{Y,\tilde{Y}} \left[-\log \frac{\int p(Y, \tilde{Y}|\theta) w_k(\theta) \frac{1}{w_k(\theta)} d\theta}{\int p(Y, \tilde{Y}|\theta) w_k(\theta) d\theta} \right.$$
$$\left. - \log \frac{\int p(Y_1|\theta') w_k(\theta') d\theta'}{\int p(Y_1|\theta') d\theta'} \right]$$
$$= \mathbb{E}_{Y,\tilde{Y}} \left[-\log \mathbb{E}_{\theta|Y,\tilde{Y}} \frac{1}{w_k(\theta)} - \log \int p(Y_1|\theta') w_k(\theta') d\theta' \right]$$

where we use $\int p(Y_1|\theta') d\theta' = 1$ at the last equality. The variable on which to condition is chosen to be one for which the variance is finite (here Y_1, without loss of generality).

Invoking Jensen's inequality in both terms (using convexity of $-\log$), we get that the Bayes risk difference is less than or equal to

$$\begin{aligned}\mathbb{E}_\theta \log w_k(\theta) - \mathbb{E}_{Y_1} \int p(Y_1|\theta') \log w_k(\theta') d\theta' \\ = \int w_k(\theta) \log w_k(\theta) d\theta + \int\int\int w_k(\theta) p(y_1-\theta) \\ \times\, p(y_1-\theta') \log \frac{1}{w_k(\theta')} d\theta' dy_1 d\theta \end{aligned} \tag{12}$$

where $\int w_k(\theta)p(y_1-\theta)d\theta$ in the second term is the mixture giving the distribution of Y_1. Next we do a change of variables, where for each θ we replace y_1 and θ' with $z_1 = y_1 - \theta$ and $z_1' = y_1 - \theta'$. So (12) becomes

$$\mathbb{E}_{Z_1,Z_1',\theta} \log \frac{w_k(\theta)}{w_k(\theta + Z_1 - Z_1')} = \frac{\mathbb{E}||Z_1||^2}{k} \to 0, \quad k \to \infty. \tag{13}$$

So q^* is extended Bayes, and therefore minimax (as per Lemma 4 of Appendix A).

Remark: A similar but more involved argument using prior $w_k(\theta)$ with tails that decay at a polynomial rather than exponential rate (e.g., Cauchy priors) shows that a finite logarithmic moment (that is, $\mathbb{E}\log(1+|Z_i|)$ finite for some i) is sufficient for minimaxity of the best invariant rule (see Appendix B).

Next we consider extended Bayes and minimaxity for the cases of univariate scale (Theorem 2) and multivariate location with univariate scale (Theorem 3). The technique that we used in deriving the upper bounds for the Bayes risk differences in the proof for Theorem 1 turns out to be very useful for other cases too. So we summarize a key step in this technique as a more general lemma below.

Lemma 3: (Bayes Risk Difference Bound): Suppose there is a parametric family $\{p(y,\tilde{y}|\theta) : \theta \in \Theta\}$. Let v and w be two priors (v proper, w possibly improper) on θ and let $u = f(y)$ be a function of y with density $p_U(u|\theta)$ for which the posterior $w(\theta|u)$ is proper, that is, $\int p_U(u|\theta)w(\theta)d\theta$ is finite for all u. Then the Bayes risk difference satisfies the following inequality:

$$R_v(p_w) - R_v(p_v) \le \mathbb{E}_\theta^v \mathbb{E}_{U|\theta} \mathbb{E}_{\theta'|U}^w \log \frac{v(\theta)/w(\theta)}{v(\theta')/w(\theta')}$$

where $\mathbb{E}_{\theta'|U}^w$ denotes the expectation with respect to the posterior of θ' given U when θ' has prior w and $\mathbb{E}_\theta^v$ denotes the expectation with respect to the prior v on θ.

Proof: By definition, the risk difference $R_v(p_w) - R_v(p_v)$ is equal to

$$\mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(\tilde{Y}|Y)}{p_w(\tilde{Y}|Y)} = \mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(Y,\tilde{Y})}{p_w(Y,\tilde{Y})} - \mathbb{E}_Y^v \log \frac{p_v(Y)}{p_w(Y)}$$

which is upper-bounded by

$$\mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(Y,\tilde{Y})}{p_w(Y,\tilde{Y})} - \mathbb{E}_U^v \log \frac{p_v(U)}{p_w(U)} \tag{14}$$

due to the result from information theory (as can be verified by a chain rule) that for two density functions p_Y, q_Y, if u is a function of y with corresponding densities p_U and q_U, then $D(p_Y||q_Y) \ge D(p_U||q_U)$.

Similarly to the proof for Theorems 1, we express the first term of (14) as a conditional expectation and then apply Jensen's inequality using the convexity of $-\log$

$$\begin{aligned}\mathbb{E}_{Y,\tilde{Y}}^v \Big(-\log \frac{\int p(Y,\tilde{Y}|\theta)v(\theta)\frac{w(\theta)}{v(\theta)}d\theta}{\int p(Y,\tilde{Y}|\theta)v(\theta)d\theta} \Big) \\ \le \mathbb{E}_{Y,\tilde{Y}}^v \mathbb{E}_{\theta|Y,\tilde{Y}}^v \Big(-\log \frac{w(\theta)}{v(\theta)} \Big) = \mathbb{E}_\theta^v \log \frac{v(\theta)}{w(\theta)}. \end{aligned}$$

Apply the same steps on the second term of (14), we have

$$\begin{aligned}\mathbb{E}_U^v \log \frac{p_v(U)}{p_w(U)} &= \mathbb{E}_U^v \log \frac{\int p(U|\theta')w(\theta')\frac{v(\theta')}{w(\theta')}d\theta'}{\int p(U|\theta')w(\theta')d\theta'} \\ &\ge \mathbb{E}_U^v \mathbb{E}_{\theta'|U}^w \log \frac{v(\theta')}{w(\theta')}. \end{aligned}$$

So the Bayes risk difference is less than or equal to

$$\mathbb{E}_\theta^v \log \frac{v(\theta)}{w(\theta)} - \mathbb{E}_U^v \mathbb{E}_{\theta'|U}^w \log \frac{v(\theta')}{w(\theta')}$$

which completes the proof.

Theorem 2: Assume for the scale family (i.e., general linear transformation family with $d = 1$ and $\theta \neq 0$) that there exists $i \in \{1,\dots,m\}$ such that $\log(|Z_i|)$ is integrable. Then, under the Kullback–Leibler loss, the best invariant predictive procedure

$$q^*(\tilde{y}|y) = \frac{\int \frac{1}{|\theta|} p(y,\tilde{y}|\theta)d\theta}{\int \frac{1}{|\theta|} p(y|\theta)d\theta}$$

is extended Bayes and hence minimax.

Proof: We take a sequence of proper priors to be $w_k(\theta)$ proportional to $\min(|\theta|^{-1-\alpha_k}, |\theta|^{-1+\alpha_k})$, where $\alpha_k > 0$. For α_k small, these priors have behavior close to that of improper prior $w(\theta) = |\theta|^{-1}$.

By Lemma 3 with $v = w_k$ and $u = f(y) = y_1$, we have the risk difference less than or equal to

$$\mathbb{E}_\theta^{w_k} \log \frac{w_k(\theta)}{w(\theta)} - \mathbb{E}_{Y_1}^{w_k} \left[\mathbb{E}_{\theta'|Y_1}^w \log \frac{w_k(\theta')}{w(\theta')} \right]. \tag{15}$$

We change the integration variable θ' inside the brackets above to $z_1' = y_1\theta'$ for any given $Y_1 = y_1$. Calculation reveals that the posterior density (given y_1) for Z_1' is indeed $p_{Z_1}(z_1')$ independent of y_1. Thus,

$$\mathbb{E}_{\theta'|Y_1}^w \log \frac{w_k(\theta')}{w(\theta')} = \mathbb{E}_{Z_1'} \log \frac{w_k(Z_1'/Y_1)}{w(Z_1'/Y_1)}.$$

Apply another change of variable from y_1 to $z_1 = \theta y_1$, then we have (15) equal to

$$\begin{aligned}\mathbb{E}_\theta^{w_k} \mathbb{E}_{Z_1,Z_1'} \log \frac{|\theta| w_k(\theta)}{|\theta|\frac{|Z_1'|}{|Z_1|} w_k(\theta \frac{Z_1'}{Z_1})} \\ = \mathbb{E}^{w_k} \mathbb{E}_{Z_1,Z_1'} \min(-\alpha_k \log|\theta|,\ \alpha_k \log|\theta|) \\ - \min\left(-\alpha_k \log|\theta| \frac{|Z_1'|}{|Z_1|},\ \alpha_k \log|\theta| \frac{|Z_1'|}{|Z_1|} \right). \end{aligned} \tag{16}$$

By the fact that $\min(a,-a) - \min(-a-b, a+b) \le |b|$, (16) is less than or equal to $\alpha_k \mathbb{E}\,|\log \frac{|Z_1'|}{|Z_1|}|$, which goes to zero when α_k goes to zero by our assumption.

Theorem 3: For the multivariate location with univariate scale family, conditioning on at least two observations

$(m \geq 2)$, assume that there exist $i, j \in \{1, \ldots, m\}$ and $\ell \in \{1, \ldots, d\}$ such that $\log(|Z_{i\ell} - Z_{j\ell}|)$, $\log\left(1 + \left|\frac{Z_{i\ell} - Z_{j\ell}}{Z'_{i\ell} - Z'_{j\ell}}\right|\right)$, and $\log(1 + \|Z_i\|)$ are integrable, where Z'_i and Z'_j are independent copies of Z_i and Z_j, respectively, and $Z_{i\ell}$ denotes the ℓth coordinate of the d-dimensional vector Z_i. Then, under the Kullback–Leibler loss, the best invariant predictive procedure

$$q^*(\tilde{y}|y) = \frac{\int\int \frac{1}{|\theta_2|} p(y, \tilde{y}|\theta_1, \theta_2) d\theta_1 d\theta_2}{\int\int \frac{1}{|\theta_2|} p(y|\theta_1, \theta_2) d\theta_1 d\theta_2}$$

is extended Bayes and hence minimax.

Proof: We take the proper prior $w_k(\theta_1, \theta_2)$ to be the product of priors on θ_1 and θ_2 which we used in the proofs for location families (Appendix B, Theorem 1′) and scale families (Theorem 2). That is, $w_k(\theta_1, \theta_2) = w_k^{(1)}(\theta_1) w_k^{(2)}(\theta_2)$ and

$$\begin{aligned} w_k^{(1)}(\theta_1) &\sim \frac{1}{(1 + ||\theta_1||/k^4)^{d+1}} \\ w_k^{(2)}(\theta_2) &\sim \min(|\theta_2|^{-1-\alpha_k}, |\theta_2|^{-1+\alpha_k}). \end{aligned} \tag{17}$$

This provides our sequence of proper priors with behavior close to that of the improper prior $w(\theta_1, \theta_2) = 1/|\theta_2|$.

Without loss of generality, we assume the indices i, j, and ℓ in the assumption are equal to 1, 2, and 1. Apply Lemma 3 with $u = (y_1, y_{21})$ and $v = w_k$, where y_{21} is the first coordinate of y_2. Then the Bayes risk difference $R_{w_k}(p^*) - R_{w_k}(p_{w_k})$ is less than or equal to

$$\mathbb{E}_\theta^{w_k} \mathbb{E}_{Y_1, Y_{21}|\theta} \mathbb{E}^w_{\theta'|Y_1, Y_{21}} \log \frac{w_k(\theta)|\theta_2|}{w_k(\theta')|\theta'_2|}. \tag{18}$$

In a manner similar to the previous proofs, for given y_1 and y_{21}, we change variable (θ'_1, θ'_2) to (z'_1, z'_{21}) with

$$\begin{cases} z'_1 &= \theta'_2(y_1 - \theta'_1) \\ z'_{21} &= \theta'_2(y_{21} - \theta'_{11}) \end{cases} \Rightarrow \begin{cases} \theta'_1 &= y_1 - z'_1 \frac{y_{11} - y_{21}}{z'_{11} - z'_{21}} \\ \theta'_2 &= \frac{z'_{11} - z'_{21}}{y_{11} - y_{21}}. \end{cases}$$

The corresponding Jacobian is equal to $|\theta'_1|^{-d}|y_{21} - y_{11}|^{-1}$. We find that the joint density for (Z'_1, Z'_{21}) is independent of y_1, y_{21} and has the same distribution as (Z_1, Z_{21}). Replace (y_1, y_{21}) by $y_1 = z_1/\theta_2 + \theta_1$ and $y_{21} = z_{21}/\theta_2 + \theta_{11}$, then we have (18) equal to

$$\begin{aligned} \mathbb{E}_\theta^{w_k} \mathbb{E}_{Z_1, Z_{21}} \mathbb{E}_{Z'_1, Z'_{21}} \Bigg[&\log \frac{w_k^{(1)}(\theta_1)}{w_k^{(1)}(\frac{Z_1}{\theta_2} - \frac{Z'_1}{\theta_2}\frac{Z_{11} - Z_{21}}{Z'_{11} - Z'_{21}} + \theta_1)} \\ &+ \log \frac{w_k^{(2)}(\theta_2)|\theta_2|}{w_k^{(2)}(\frac{Z'_{11} - Z'_{21}}{Z_{11} - Z_{21}}\theta_2)\left|\theta_2 \frac{Z'_{11} - Z'_{21}}{Z_{11} - Z_{21}}\right|} \Bigg]. \end{aligned} \tag{19}$$

By the proof for Theorem 1′ (in Appendix B) and Theorem 2, we know that the quantities above go to zero if

$$\mathbb{E}^{w_k} \log\left(1 + \frac{1}{k}\left\|\frac{Z_1}{\theta_2} - \frac{Z'_1}{\theta_2}\frac{Z_{11} - Z_{21}}{Z'_{11} - Z'_{21}}\right\|\right)$$

and $\alpha_k \mathbb{E}|\log(\,|Z_{21} - Z_{11}|\,)|$ go to zero when k goes to infinity. Since $\mathbb{E}|\log(\,|Z_{21} - Z_{11}|\,)|$ is finite, $\alpha_k \mathbb{E}|\log(\,|Z_{21} - Z_{11}|\,)|$ goes to zero. Using the triangle inequality and the inequalities for positive a, b that $(1+a+b) \leq (1+a)(1+b)$ and $(1+ab) \leq (1+a)(1+b)$, we obtain

$$\begin{aligned} \mathbb{E}^{w_k} &\log\left(1 + \frac{1}{k^4}\left\|\frac{Z_1}{\theta_2} - \frac{Z'_1}{\theta_2}\frac{Z_{11} - Z_{21}}{Z'_{11} - Z'_{21}}\right\|\right) \\ &\leq 2\mathbb{E}^{w_k} \log\left(1 + \frac{1}{k|\theta_2|}\right) + 2\mathbb{E}\log\left(1 + \frac{\|Z_1\|}{k}\right) \\ &\quad + \mathbb{E}\log\left(1 + \frac{1}{k^2}\left|\frac{Z_{11} - Z_{21}}{Z'_{11} - Z'_{21}}\right|\right) \end{aligned}$$

where the last two terms in the final expression will go to zero since

$$\log\left(1 + \left|\frac{Z_{11} - Z_{21}}{Z'_{11} - Z'_{21}}\right|\right)$$

and $\log(1 + \|Z_1\|)$ are integrable by assumptions. For the first term $\mathbb{E}_\theta \log(1 + \frac{1}{k|\theta_2|})$, we consider the integration over $[0, 1]$ and $(1, \infty)$ separately. When $\theta_2 > 1$

$$\log\left(1 + \frac{1}{k\theta_2}\right) \leq \log\left(1 + \frac{1}{k}\right) \leq 1/k.$$

Thus,

$$\int_1^\infty w_k^{(2)}(\theta_2) \log\left(1 + \frac{1}{k\theta_2}\right) d\theta_2 \leq 1/k$$

which goes to zero when k goes to infinity. On the other hand, the integral over the range $[0, 1]$ is equal to

$$\frac{\alpha_k}{2} \int_0^1 \theta^{-1+\alpha_k} \log\left(1 + \frac{1}{k\theta_2}\right) d\theta_2. \tag{20}$$

Change the variable θ_2 to $\eta = 1/(k\theta_2)$, then (20) is equal to

$$(\alpha_k/2) k^{-\alpha_k} \int_{1/k}^\infty \eta^{-1-\alpha_k} \log(1 + \eta) d\eta.$$

Divide the integration range into two parts $(1, \infty)$ and $(k^{-1/4}, 1]$. Applying inequalities $\log(1+\eta) \leq \log 2 + \log \eta$ for $\eta > 1$ and $\log(1+\eta) \leq \eta$ for $\eta \leq 1$, we have the integral (20) is less than or equal to $\frac{1}{\alpha_k k^{\alpha_k}}(\frac{\alpha_k^2}{1-\alpha_k} + \alpha_k \log 2 + 1)$ which goes to zero if $\alpha_k k^{\alpha_k} \to \infty$, or equivalently, $\alpha_k \log k + \log \alpha_k \to \infty$, as is true if, for example, $\alpha_k = 1/\sqrt{\log k}$.

Since the minimax procedure q^* has constant risk and is extended Bayes, an immediate corollary of Theorems 1–3 is that the maximin value C is equal to the minimax value R for those location and scale problems conditioning on data Y.

Next we show that the minimax risk is infinite without conditioning on enough initial observations. Here the minimal number of initial observations required is one for location or scale families, and two for multivariate location with univariate scale families.

Proposition 4: For the location or scale families, the minimax risk (using Kullback–Leibler loss) is infinity if one does not condition on any observations. For multivariate location with univariate scale families, the minimax risk is infinity if conditioning on less than two observations.

Proof: When there is no conditioning, the conclusion is a special case of a result given by Haussler ([27, Lemma 4]) on the unconditional minimax risk. It is stated there that the minimax risk is infinity if the parametric family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is not uniformly tight, where *uniformly tight* means for every $\epsilon > 0$ there exists a compact set K of the sample space such that $P(K) > 1 - \epsilon$ for all $P \in \mathcal{P}$. It can be verified that location or scale families are not uniformly tight. For example, consider any location family. For any $0 < \epsilon < 1/2$ and any compact K with $P_{\tilde{Z}}(K) > 1 - \epsilon$. Let $B(0, r)$ be a ball centered at origin with radius r so large that $K \subset B$. Let $\theta_0 = 2r$, then $P_{\tilde{Y}|\theta_0}(K)$ is not more than

$$P_{\tilde{Y}|\theta_0}(B) = P_{\tilde{Z}}(B - \theta_0) \le \epsilon < 1 - \epsilon$$

since the shift of the ball $B(0, r) - 2r = B(-2r, r)$ is in B^c. Therefore, the location family is not uniformly tight and the uncondition risk is infinity. The scale case is similar.

Next we show that, for multivariate location with univariate scale families, the minimax risk is infinity when one conditions on only one observation. The risk of an estimator $q(\tilde{y}|y)$ is equal to

$$\mathbb{E}_{Z,\tilde{Z}} \log \frac{|\theta_2|^d p(\tilde{Z})}{q(\frac{\tilde{Z}}{\theta_2} + \theta_1 | \frac{Z}{\theta_2} + \theta_1)} = -H(\tilde{Z}) + d \log |\theta_2| - \mathbb{E}_{Z,T} \log q\left(\frac{Z}{\theta_2} + \theta_1 + \frac{T}{\theta_2} | \frac{Z}{\theta_2} + \theta_1\right)$$

where $H(\tilde{Z}) = -\mathbb{E}_{\tilde{Z}} \log p(\tilde{Z})$ is the entropy of $\tilde{Z}$ and T is a random variable equal to $\tilde{Z} - Z$ with distribution not depending on θ. Using Jensen's inequality, the risk is greater than or equal to

$$-H(\tilde{Z}) + d \log |\theta_2| - \mathbb{E}_T \log f(T/\theta_2) \tag{21}$$

where $f(T/\theta_2)$ denotes $\mathbb{E}_{Z|T}[q(\frac{Z}{\theta_2} + \theta_1 + \frac{T}{\theta_2} | \frac{Z}{\theta_2} + \theta_1)]$ (the expectation may also depends on θ_1, but such a dependence is irrelevant to this proof). Observe that the function $f(\cdot)$ is a probability density function, that is,

$$\int f(x) dx = \mathbb{E}_Z \int q\left(\frac{Z}{\theta_2} + \theta_1 + x | \frac{Z}{\theta_2} + \theta_1\right) dx = 1.$$

Let $X = T/\theta_2$, which has density $p_{X|\theta_2}(x|\theta_2) = |\theta_2|^d p_T(\theta_2 x)$, then our lower bound on the risk is

$$H(T) - H(\tilde{Z}) + \mathbb{E}_{X|\theta_2} \log \frac{p_{X|\theta_2}(X|\theta_2)}{f(X)}$$

where the last term is the Kullback risk of the estimator $f(X)$ (based on no data) of the scale family of the densities for X. So by the result for the unconditional risk for scale families, its supremum over θ_2 is infinity.

IV. Exact Minimax Rules for Regression

We consider a linear regression model

$$\tilde{y}_i = \tilde{x}_{i1}\theta_1 + \cdots + \tilde{x}_{id}\theta_d + \tilde{z}_i = \tilde{x}_i^t \theta + \tilde{z}_i$$

where $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{id})$ is a d-dimensional input vector, and $\tilde{z}_i$ is the random error. Our interest is in finding the exact minimax coding strategy (or predictive density estimator) for linear regression models. We use $Y = (Y_1, \ldots, Y_m)$ for the initial data, $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ for the data for which we want to predict the distribution, and $\tilde{Z}$, Z for the corresponding errors. Let $\tilde{x}$ denote the $d \times n$ matrix with $\tilde{x}_i$ as its ith column. Similarly, we have x denote the $d \times m$ matrix with x_i as its ith column.

Assume $(\tilde{Z}, Z)$ is modeled by a distribution P with density p. Then the density for $(\tilde{Y}, Y)$ is given by

$$p_{\tilde{Y},Y|\theta}(\tilde{y}, y|\theta) = p(\tilde{y} - \tilde{x}^t\theta, y - x^t\theta), \qquad \theta \in \mathbb{R}^d \tag{22}$$

which is different from the ordinary location families we studied before, but similar analysis can be applied and it reveals that the exact minimax strategy is the Bayes procedure with uniform prior over the parameter space $\mathbb{R}^d$, conditioning on at least $m \ge d$ observations.

Theorem 4: Assume that for the parametric family given in (22) with $m \ge d$ there exists a d-element subset from $(1, \ldots, m)$, denoted by $(i_1, \ldots, i_d)$, such that the d errors $(Z_{i_1}, \ldots, Z_{i_d})$ have finite second moments and that the $d \times d$ matrix composed by the d vectors $x_{i_1}, \ldots, x_{i_d}$ is nonsingular. Then the procedure

$$q^*(\tilde{y}|y) = \frac{\int p(\tilde{y} - \tilde{x}^t\theta, y - x^t\theta) d\theta}{\int p(y - x^t\theta) d\theta}$$

is extended Bayes with constant risk and hence minimax for Kullback–Leibler loss.

Proof: First observe that q^* is invariant to shift of y by $x^t\theta$ if $\tilde{y}$ is correspondingly shifted by $\tilde{x}^t\theta$, that is,

$$\begin{aligned} q^*(\tilde{y} - \tilde{x}^t\theta | y - x^t\theta) &= \frac{\int p(\tilde{y} - \tilde{x}^t\theta - \tilde{x}^t\alpha, y - x^t\theta - x^t\alpha) d\alpha}{\int p(y - x^t\theta - x^t\alpha) d\alpha} \\ &= \frac{\int p(\tilde{y} - \tilde{x}^t\theta', y - x^t\theta') d\theta'}{\int p(y - x^t\theta') d\theta'}, \qquad \theta' = \alpha + \theta. \end{aligned}$$

By invariance, it is easy to show that q^* has constant risk.

Next we show that q^* is extended Bayes. Take normal priors $w_k(\theta)$ as in the proof for Theorem 1. Let $w(\theta) = 1$ and take the reduced set of conditioning variables to be $u = (y_{i_1}, \ldots, y_{i_d})$. Then by Lemma 3

$$R_{w_k}(q^*) - R_{w_k}(p_{w_k}) \le \mathbb{E}_\theta^{w_k} \mathbb{E}_{U|\theta} \mathbb{E}_{\theta'|U}^{w} \log \frac{w_k(\theta)}{w_k(\theta')}. \tag{23}$$

For now (while working with the reduced set of conditioning variables), let x denote the $d \times d$ matrix $(x_{i_1}, \ldots, x_{i_d})$ which is nonsingular by our assumption. Change variables with $z' = u - x^t\theta'$ and $z = u - x^t\theta$. We find the posterior distribution of Z' given u is independent of u and has the same distribution as $Z = (Z_{i_1}, \ldots, Z_{i_d})$. So the right-hand side of inequality (23) is equal to

$$\begin{aligned} &\mathbb{E}_\theta^{w_k} \mathbb{E}_{Z,Z'} \log \frac{w_k(\theta)}{w_k(\theta')} \\ &\quad = \mathbb{E}_{Z,Z',\theta} \log \frac{w_k(\theta)}{w_k(\theta + (x^t)^{-1}(Z - Z'))} \\ &\quad = \frac{\mathrm{Trace}[(x^{-1})(x^{-1})^t \, \mathbb{E} Z Z^t]}{k} \end{aligned}$$

which goes to zero when k goes to infinity since x is nonsingular and Z has finite second moment by our assumptions. Thus, q^* is extended Bayes with constant risk, hence minimax.

In ordinary linear regression models, we often assume that the errors $\tilde{Z}_i$'s and Z_i's are distributed as independent $\text{Normal}(0,\sigma^2)$.

1) Known σ^2: The minimax predictive density q^* for future n observations $\tilde{Y} = (\tilde{Y}_1,\ldots,\tilde{Y}_n)$ based on the past observations $Y = (Y_1,\ldots,Y_m)$ is

$$q^*(\tilde{y}|y) = \frac{\int \phi_{\sigma^2}(\tilde{y}-\tilde{x}^t\theta)\phi_{\sigma^2}(y-x^t\theta)d\theta}{\int \phi_{\sigma^2}(y-x^t\theta)d\theta} \tag{24}$$

where ϕ_{σ^2} denotes the density function for $N(0,\sigma^2)$. Note that

$$\int \phi_{\sigma^2}(y-x^t\theta)d\theta = \frac{|S_m|^{-1/2}}{(\sqrt{2\pi\sigma^2})^{m-d}}\exp\left(-\frac{1}{2\sigma^2}\,\text{RSS}_m\right)$$

where $S_m = \sum_{i=1}^m x_i x_i^t$ is the information matrix, $\hat{\theta}_m = (x^tx)^{-1}x^ty$ is the least squares estimate of θ based on the m observations y, and $\text{RSS}_m = \|y - x^t\hat{\theta}_m\|^2$ is the residual sum of squares (RSS). Similarly simplifying the numerator of (24), we have the following expression for $(-log)$ predictive density and MDL code length:

$$\frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}(\text{RSS}_{m+n} - \text{RSS}_m) + \frac{1}{2}\log\frac{|S_{m+n}|}{|S_m|} \tag{25}$$

where

$$S_{m+n} = S_m + \sum_{i=1}^n \tilde{x}_i\tilde{x}_i^t$$

and

$$\text{RSS}_{m+n} = \|y - x^t\hat{\theta}_{m+n}\|^2 + \|\tilde{y} - \tilde{x}^t\hat{\theta}_{m+n}\|^2$$

respectively, are the information matrix and the residual sum of squares using all $N = m+n$ observations.

For regression model selection, we are looking for the optimal subset of $\tilde{x}$ to predict $\tilde{y}$. Here "optimal" means the resulting model has the shortest description length. We use γ to index the 2^d possible subsets (or *models*). The code length for the minimax coding strategy q^* given in (25) for each subset γ can be used as the criterion for model selection. Since the first term $(n/2)\log 2\pi\sigma^2$ is shared by all models, we omit it from the final MDL criterion

$$\frac{[\text{RSS}_{m+n}(\gamma) - \text{RSS}_m(\gamma)]}{2\sigma^2} + \frac{1}{2}\log\frac{|S_{m+n}(\gamma)|}{|S_m(\gamma)|} + L(\gamma) \tag{26}$$

where $L(\gamma)$ denotes the description length for the model index γ. When a uniform distribution is used to code the model index γ, the description length $L(\gamma)$ is the same for each subset and therefore it can be omitted from the final criterion.

If we set $R_N(\gamma) = \frac{1}{N}\sum_{i=1}^N x_{i\gamma}x_{i\gamma}^t$, then the main penalty terms in (26) are

$$\frac{d_\gamma}{2}\log\frac{N}{m} + \frac{1}{2}\log\frac{|R_N(\gamma)|}{|R_m(\gamma)|}$$

where d_γ is equal to the number of variables included in model γ. If $|R_N(\gamma)|$ is nearly constant (not depending strongly on N) this penalty is roughly in agreement with $\frac{d}{2}\log N$ penalty in simplified MDL or Bayesian information type criterion (BIC). Nonetheless, for some x_i's (e.g., those evolving according to some nonstationary time series models), the sum $S_N(\gamma)$ may grow at faster rates, e.g., of order N^2 rather than N, leading to $\frac{1}{2}\log|S_N(\gamma)|$ of order $d_\gamma\log N$ rather than $\frac{d_\gamma}{2}\log N$. In general, it is better to retain the $\frac{1}{2}\log|S_N(\gamma)|$ determinant form of the penalty rather than the sometimes inaccurate approximation $\frac{d_\gamma}{2}\log N$.

2) Unknown σ^2: In this case, we find that the minimax procedure q^* is the generalized Bayes procedure with a uniform prior on the location and log-scale parameters (Theorem 5)

$$\begin{aligned} q^*(\tilde{y}|y) &= \frac{\int\int \frac{1}{\sigma}\phi_{\sigma^2}(\tilde{y}-\tilde{x}^t\theta)\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma}{\int\int \frac{1}{\sigma}\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma} \\ &= \frac{\Gamma(\frac{m+n-d}{2})}{\Gamma(\frac{m-d}{2})}\frac{1}{(\pi)^{n/2}}\frac{|S_m|^{1/2}}{|S_{m+n}|^{1/2}}\frac{(\text{RSS}_m)^{(m-d)/2}}{(\text{RSS}_{m+n})^{(m+n-d)/2}} \end{aligned}$$

which leads to the following MDL criterion:

$$\begin{aligned} &\frac{m+n-d_\gamma}{2}\log\text{RSS}_{m+n}(\gamma) - \frac{m-d_\gamma}{2}\log\text{RSS}_m(\gamma) \\ &\qquad + \frac{1}{2}\log\frac{|S_{m+n}(\gamma)|}{|S_m(\gamma)|} - \log\frac{\Gamma(\frac{m+n-d_\gamma}{2})}{\Gamma(\frac{m-d_\gamma}{2})} + L(\gamma). \end{aligned}$$

Theorem 5: For the regression model with $m \geq d+1$, assume $(\tilde{Y},Y)$ is modeled by normal with mean $(\tilde{x}^t\theta, x^t\theta)$ and unknown variance σ^2. Then

$$q^*(\tilde{y}|y) = \frac{\int\int \frac{1}{\sigma}\phi_{\sigma^2}(\tilde{y}-\tilde{x}^t\theta)\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma}{\int\int \frac{1}{\sigma}\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma}$$

is minimax under the Kullback–Leibler loss.

We omit the proof for Theorem 5 here, which is similar to the proof for Theorem 3 (for detail, see [39]). It also can be regarded as a special case of a general proof we will give in the next section.

V. Hunt–Stein Analysis

There is a fairly general strategy, with several different particularizations [32], [33], [57], [11], [8], [54], [36], [38], by which to demonstrate under certain conditions that the best invariant procedure is minimax among all procedures, in group-theoretic settings. In addition to topological conditions on the parameter and action spaces, the key condition, used in essentially all versions of this strategy, is the so-called amenability of the transformation group G. One way of stating amenability is that there exists a sequence of probability measures λ_j on G that are asymptotically invariant, that is, $\int(\phi(ag)-\phi(a))\lambda_j(da)$ tends to zero as $j\to\infty$ for each g in G for bounded continuous functions ϕ. The groups we have studied here are amenable.

In brief, the strategy is as follows. One considers an arbitrary given procedure q (not necessarily invariant) and shows

that there is an invariant procedure q^* for which its constant risk is not more than the maximum risk of the given procedure. The idea is that one can consider transformations of the given procedure, here denoted q^a for a in G, and create rules that are closer to invariant by averaging with respect to λ_j to produce a sequence of procedures $q_j = \int q^a \lambda_j(da)$. Each of these will have risk functions less than or equal to the maximum of the risks of the q^a by the convexity of the loss function (in nonconvex loss cases randomized decision rules are used which provide a linearization of the loss). If one can establish that along some subsequence there is a limiting procedure q^* that is invariant and that has risk not more than the risks of the q_j then one is done.

We note that the estimators correspond to probability distributions (either by direct setup of the problem or by introduction of randomized rules). So extraction of the limit requires consideration of limits of probability measures. Difficulties can arise, e.g., in extracting a weak limit, from a mass leaking toward infinity.

The device championed in Strasser [54] and LeCam [36] is to consider averaged risks as linear both with respect to priors and with respect to randomization rules, so as to embed the problem in a setting of complete bilinear forms for which limiting bilinear forms can be extracted (in their case by appealing to a Markov–Kakatani fixed point theorem which also provides the invariance of the limit). Under further conditions which one must check in any problem at hand, they then show that there is an ordinary randomized statistical procedure whose risk is bounded by what is given by the extracted limiting bilinear form. The conditions entail compactness of level sets of the loss function in certain topologies.

We provide a simplification that covers cases of interest here. The idea is to consider problems for which the action space can be extended from collections of probability measures to the class of subprobability measures on given spaces, permitting consideration of vague convergence, and loss functions such as Kullback–Leibler which are convex and lower semicontinuous with respect to vague convergence. Then the extraction of the limiting measure and the verification that it provides a procedure with risk not larger than the Q_j are both automatic.

Vague convergence is for measures on sample spaces of suitable structure, such as the Euclidean space for the random variables $(Y, \tilde{Y})$ in our setting (or more generally locally compact topological spaces possessing a countable base). Here and in what follows we will use the notational convention that $Q(f) = \int f dQ$ for nonnegative finite measures Q and integrable f. A sequence of measures Q^j is said to *converge vaguely* to a limiting measure Q^* if for each function f that is bounded, continuous, and becomes zero outside some compact set (depending on f) one has $\lim_j Q^j(f) = Q^*(f)$ (see, e.g., [12], [31]). It is analogous to weak convergence except that in vague convergence, mass is permitted to leak away. For general bounded continuous f (not necessarily compactly supported) one has $\liminf_j Q^j(f) \geq Q^*(f)$. For example, even if every $Q^j(1) = 1$, the limit $Q^*(1)$ can be less than 1, where $Q(1) = \int dQ$. The fundamental usefulness of the notion of vague convergence is that for every sequence Q^j with bounded $Q^j(1)$ there is a vaguely convergence subsequence (see, for instance, [43], [31], [12]).

Our other main tool is lower semicontinuity of the relative entropy $D(P||Q)$. It is well known in the case of weak convergence. In Appendix E, we show it also holds true for vaguely convergent sequences Q^j.

General Group Setup: For the predictive density estimation problem, we have random variables $(Y, \tilde{Y})$ and a probability density function $p(y, \tilde{y}|\theta)$ with respect to Lebesgue measure, where θ takes values in a parameter space Θ. We denote the corresponding joint distribution as $P_{Y,\tilde{Y}|\theta}$. Given Y, and with θ unknown, we want to predict the distribution of $\tilde{Y}$. The action space A contains all subprobability measures Q on $\tilde{Y}$, that is, the Q measure of the whole $\tilde{Y}$ space is not more than 1. Decision procedures correspond to predictive distributions (also known as transition measures) $Q_{\tilde{Y}|Y=y}(\cdot) = Q(\cdot; y)$ that are subprobability measures on $\tilde{Y}$ for each y and measurable functions of y for each measurable set of $\tilde{Y}$.

Let G be a group which acts on the left of the three spaces $(Y, \tilde{Y})$, Θ, and A. The actions on the sample space and the parameter space are denoted by $(gy, g\tilde{y})$ and $g\theta$. The group action on any $Q \in A$ is defined such that for any bounded measurable function $f(\tilde{Y})$ we have $gQ(f(\tilde{Y})) = Q(f(g\tilde{Y}))$.

The decision problem we consider is invariant because the model $P_\theta = P_{Y,\tilde{Y}|\theta}$ is invariant, that is,

$$(Y, \tilde{Y}) \sim P_\theta \Rightarrow (gY, g\tilde{Y}) \sim P_{g\theta}$$

and the loss function is invariant, that is, for any probability distribution P on $\tilde{Y}$ and any action Q

$$D(gP_\theta || gQ) = D(P_\theta, Q).$$

For any procedure $Q_{\tilde{Y}|y}$ and $g \in G$ we use $Q^g_{\tilde{Y}|y}$ to denote a transformed procedure such that for any measurable set B

$$Q^g(B; y) = Q(g^{-1}B; g^{-1}y).$$

For bounded measurable functions $f(\tilde{y})$, the transformed procedure gives integral $Q^g_{\tilde{Y}|y}(f(Y)) = Q_{\tilde{Y}|g^{-1}y}(f(g\tilde{Y}))$. A procedure is said to be *invariant* if for every g in G we have $Q^g_{\tilde{Y}|y} = Q_{\tilde{Y}|y}$ for every y except on a Lebesgue null set of Y. It is said to be *almost invariant* if the null set depends of g.

A group G is *amenable* if there is a sequence of probability measures λ_j on G that is asymptotically invariant in the sense that for every $g \in G$ and every bounded measurable function ϕ on G

$$\lim_{j\to\infty} \int (\phi(ag) - \phi(a))\lambda_j(da) = 0. \tag{27}$$

Theorem 6: We have a family of densities $p(y, \tilde{y}|\theta)$ possessing a group structure as indicated above and we are estimating the conditional density $p(\tilde{y}|y, \theta)$ with Kullback–Leibler loss. Suppose that the marginal densities $p(y|\theta)$ are continuous and (at at least one parameter point) strictly positive in y. If G is amenable, then for any procedure $Q_{\tilde{Y}|y}$, there exists an invariant procedure $Q^*_{\tilde{Y}|y}$, such that

$$\max_\theta R(\theta, Q) \geq \max_\theta R(\theta, Q^*). \tag{28}$$

Proof: For any procedure $Q_{\tilde{Y}|y}$, the result is trivial if $\max_\theta R(\theta, Q) = \infty$, so assume $\max_\theta R(\theta, Q) < \infty$. Let

$q(\tilde{y}|y)$ be the density of the absolutely continuous part of $Q_{\tilde{Y}|y}$ with respect to Lebesgue measure. Since $P_{\tilde{Y}|y,\theta}$ is assumed absolutely continuous, the singular part of $Q_{\tilde{Y}|y}$ does not contribute to the Kullback–Leibler loss. So it is enough to show the result for procedures $Q_{\tilde{Y}|y}$ with no singular part. Likewise, if the procedure is a strict subprobability measure, then we may improve the Kullback–Leibler loss for every such y and every θ, by normalizing the procedure. So it is enough to show the result with $Q_{\tilde{Y}|y}$ restricted to be a proper predictive distribution with a predictive density that integrates to 1. (Subprobability measures may reappear when we extract Q^*.)

Let $q(\tilde{y}|y)$ be the predictive densities of such a procedure. Fix a particular value of θ, say θ_0, for which the density $p_0(y) = p(y|\theta_0)$ is strictly positive for all y. Let $Q = Q_{Y,\tilde{Y}}$ be the joint measure constructed to have marginal density $q(y) = p_0(y)$ and conditional density $q(\tilde{y}|y)$. Now the risk of the procedure $Q_{\tilde{Y}|Y}$ at θ_0, usually expressed as $E_{Y|\theta_0} D(P_{\tilde{Y}|Y,\theta_0}\|Q_{\tilde{Y}|Y})$, may be also expressed as the total relative entropy between these joint distributions $D(P\|Q)$.

Define a sequence of measures Q^j on $Y \times \tilde{Y}$ with marginal density equal to $p_0(y)$ and transition measure given by the following:

$$Q^j_{\tilde{Y}|y}(\cdot) = \int Q^b_{\tilde{Y}|y}(\cdot)\lambda_j(db).$$

By convexity of the Kullback–Leibler divergence, we have

$$R(\theta_0, Q^j_{\tilde{Y}|Y}) \le \max_\theta R(\theta, Q_{\tilde{Y}|Y}).$$

Let Q^* be a vague subsequence limit of the sequence Q^j, that is, there is a subsequence J such that $\lim_{j\in J} Q^j(f) = Q^*(f)$ for every f that is bounded, continuous, and compactly supported. By lower semicontinuity we have

$$D(P\|Q^*) \le \liminf_{j\in J} D(P\|Q^j) \le \max_\theta R(\theta, Q_{\tilde{Y}|Y}). \quad (29)$$

Since this D is finite, Q^* has a nonzero absolutely continuous component Q^*_a with density $q^*(y,\tilde{y})$. Though all the marginal densities of Q^j are equal to $p_0(y)$, the marginal density $q^*(y) = \int q^*(y,\tilde{y})d\tilde{y}$ is less than or equal to $p_0(y)$ for almost every y, which can be seen from properties of vague convergence. Indeed, for bounded continuous functions $f(y,\tilde{y}) = f_Y(y)$ of the first argument,

$$Q^*(f_Y) \le \lim Q^j(f_Y) = P_{Y|\theta_0}(f_Y)$$

and since this is true for all such functions, we have $q^*(y) \le p_0(y)$ for almost every y.

Thus, considering the predictive density $q^{\rm sub}(\tilde{y}|y) = q^*(y,\tilde{y})/p_0(y)$, we see that it is a subprobability density (integrating to not more than one) for almost every y. It gives a representation of Q^*_a via the product of the density $p_0(y)$ for y and the subprobability density $q^{\rm sub}(\tilde{y}|y)$. Thus, we have risk

$$R(\theta_0, q^{\rm sub}) = D(P\|Q^*_a) = D(P\|Q^*) \le \max_\theta R(\theta, Q_{\tilde{Y}|Y}).$$

Next we consider invariance of a procedure with this risk. Toward that end we work with the full measure Q^* and not just its absolutely continuous part. A standard transition measure construction shows that there exists a family of regular transition measures $Q^*_{\tilde{Y}|y}$ that permit the representation

$$Q^*(f(Y,\tilde{Y})) = Q^*_Y(Q^*_{\tilde{Y}|Y}(f(Y,\tilde{Y}))).$$

Taking note again that the Y marginal of Q^* is absolutely continuous with respect to $P_{Y|\theta_0}$, we have in the same way that there is a family of subprobability transition measures $Q^{\rm sub}_{\tilde{Y}|y}(\cdot)$ that permit the representation $Q^*(\cdot) = P_{Y|\theta_0}(Q^{\rm sub}_{\tilde{Y}|Y}(\cdot))$ for bounded measurable f. The procedure $Q^{\rm sub}_{\tilde{Y}|y}$ is the right one to work with. Its risk $R(\theta_0, Q^{\rm sub})$, which depends only on its absolutely continuous component, is what we have bounded above.

It may be tempting to renormalize $Q^{\rm sub}_{\tilde{Y}|y}$ now to integrate to 1 and improve the risk further. However, such renormalization may disrupt the invariance property that we will next establish.

Next we show that $Q^{\rm sub}_{\tilde{Y}|y}$ is almost invariant. Let $Q^{{\rm sub},b}_{\tilde{Y}|y}$ denote its transformation by b. By the positivity of $p(y|\theta_0)$, it suffices to show for every bounded continuous and compactly supported f that

$$P_{Y|\theta_0}\, Q^{{\rm sub},b}_{\tilde{Y}|Y}\,(f(Y,\tilde{Y})) = P_{Y|\theta_0}\, Q^{\rm sub}_{\tilde{Y}|Y}(f(Y,\tilde{Y})). \quad (30)$$

The right-hand side of (30) is equal to $Q^*(f(Y,\tilde{Y}))$, which is the limit of $Q^j(f(Y,\tilde{Y}))$ for $j \in J$. Recall that Q^j is built from averaging transformed procedures. Let $\phi(a) = P_{Y|\theta_0}(Q^a_{\tilde{Y}|Y}(f(Y,\tilde{Y})))$ for a in G, then $Q^j(f(Y,\tilde{Y})) = \int \phi(a)d\lambda_j(a)$. By definition of the transformed procedure, followed by a change of variables, the left-hand side of (30) is

$$\begin{aligned} P_{Y|\theta_0}\, Q^{\rm sub}_{\tilde{Y}|b^{-1}Y}(f(Y,b\tilde{Y})) &= P_{Y|b\theta_0}\, Q^{\rm sub}_{\tilde{Y}|Y}(f(bY,b\tilde{Y})) \\ &= Q^*\left[f(bY,b\tilde{Y})\frac{p(Y|b\theta_0)}{p(Y|\theta_0)}\right]. \end{aligned}$$

Denote the function inside the brackets in the equation above by f_b. This function is compactly supported (since f is), so when $p(y|b\theta_0)/p(y|\theta_0)$ is continuous, we recognize $Q^*[f_b(Y,\tilde{Y})]$ as the limit for $j \in J$ of $Q^j[f_b(Y,\tilde{Y})]$, which in the same way is the value $P_{Y|\theta_0}Q^{j,b}_{\tilde{Y}|Y}(f(Y,\tilde{Y}))$ obtained when the procedure Q^j is transformed by b, namely, $\int \phi(ab)d\lambda_j(a)$.

By amenability, the limit of $\int(\phi(ab) - \phi(a))\lambda_j(da)$ is zero, so this shows that the almost invariance (30) holds.

Finally, there must be an invariant procedure with risk that is as good as almost invariant $Q^{\rm sub}$. The reasoning here is in accordance with [37, Sec. 6.5]. Indeed, by a standard Fubini trick, since $V_b = \{y : Q^{{\rm sub},b}_{\tilde{Y}|y} \ne Q^{\rm sub}_{\tilde{Y}|y}\}$ has measure zero for every b, so also, the set $V = \{(y,b): y \in V_b\}$ where invariance fails has measure zero, using the product of Lebesgue measure for y and a sigma-finite measure λ on the group G. Likewise, for almost every y, the slices $V_y = \{b : Q^{{\rm sub},b}_{\tilde{Y}|y} \ne Q^{\rm sub}_{\tilde{Y}|y}\}$ has λ measure 0. The measure λ is chosen, such as a Haar measure, to be such that it and its transformed measures are mutually absolutely continuous. One takes $\rho(a)$ to be an everywhere positive probability density with respect to λ, and let

$$Q^{\rm ave}(\cdot;y) = \int Q^{{\rm sub},a}(\cdot;y)\rho(a)\lambda(da).$$

Then $Q^{\rm ave}$ is almost everywhere equal to $Q^{\rm sub}$, it shares the same risk, and for almost every y, $Q^{{\rm ave},b} = Q^{\rm ave}$ simultaneously for all b in G. Since the null set does not depend on b, that satisfies the definition of invariance. If desired, one may also replace $Q^{\rm ave}$ by any fixed invariant procedure when y is in that null set, so as to arrive at a final procedure that is everywhere invariant, for all y and all b in G, and has the same risk as $Q^{\rm sub}$.

We have demonstrated a valid procedure (as a subprobability) that is invariant and that has risk at θ_0 bounded by the maximal risk of the original procedure. Finally, by the invariance, the risk at all other points θ is the same. Hence, we have that the maximal risk of the original procedure is never less than the risk of some invariant procedure. This completes the proof.

Comment: With these results, minimaxity in the class of all procedures is shown to coincide with the best invariant rule. This is because the maximal risk of any noninvariant procedure is shown to be never better than the risk of some invariant rule.

In each case studied, and in accordance with general theory as in Appendix D, the best invariant rule among all subprobabilities is indeed a proper probability distribution integrating to 1 (posterior Bayes with respect to right Haar measure). Thus, the side excursion into consideration of subprobabilities should be regarded as merely a convenient technical matter to ensure that we would be able to demonstrate the existence of the limiting Q^*. In the end, proper invariant procedures are found that are at least as good.

An advantage of the result in this section is its degree of generality. For instance, no moment condition is required, and it does not require calculations on a case-by-case basis other than demonstration of the group structure. While it does provide minimaxity of best invariant rules in some desirable generality that covers our contexts of interest, compared to what we developed in Section III, the result here, like other Hunt–Stein strategies, are perhaps less revealing in that it is an existence proof depending on comparatively more involved measure-theoretic matters. In contrast, in Section III we provided information inequalities which provide a concrete strategy to exhibit directly the extended Bayes conclusion for the best invariant procedure for particular sequences of proper priors.

VI. Discussion and Conclusion

In this paper, we considered the problem of finding exact minimax universal coding strategies conditioning on some initial observations, for ordinary location and scale families and for linear regression models. The minimax predictive density estimator (under Kullback–Leibler loss) is a Bayes estimator with uniform prior over the location and log-scale parameters. It provides an exact minimax optimal strategy for density estimation and for the MDL criterion for model selection in linear regression.

Here we mention some additional related topics.

The technique used in our proof for the minimaxity for location families (Theorem 1 and 1′) also provides the admissibility of q^* in dimension 1, but not in higher dimensions. This is similar to what arises in parameter estimation, for it is known that under certain moment conditions the best invariant estimator for a one-dimensional location parameter is admissible [53], [9]. However, in some cases, such as normal location families, the sample mean is not admissible for dimension three or higher, as shown by Stein [52], [29] (with extension to inadmissibity of best invariant estimators for various families and loss functions in dimension at least three in [9]). Furthermore, for the multivariate normal location problem of dimension at least five, Strawderman [55] showed that certain proper Bayes priors produce improved risk. It is intriguing to ask whether an analogous conclusion holds for the predictive density estimation using Kullback–Leibler risk. A solution is given in one of the authors' dissertation [39], in which a proper Bayes estimator is shown to be minimax and produce smaller risk everywhere than the constant risk minimax density estimator for normal location families, provided that the dimension is bigger than four.

Issues also arise as to whether it is possible to improve on the density estimator in combined location and scale families. Results of Brown [10] may be relevant here. He shows that the best invariant scale estimators are biased and inadmissible for all loss functions for scale estimation that possess suitable invariance properties, with the sole exception being a loss function for scale estimation due to Stein, associated with Kullback–Leibler loss of plug-in estimators of scale in Normal families. We note that best invariant predictive densities provide improved Kullback–Leibler risk compared to estimators which plug-in invariant estimators of scale. However, it has not been addressed whether the best invariant (and minimax) estimators of the density in combined location and scale families is admissible.

Concerning model selection settings, when multiple models (indexed by γ) are available for prediction and/or compression, instead of picking just one model, one may create in some cases a superior adaptive procedure by Bayes model averaging [28]. When such a data compression procedure provides small redundancy simultaneously for all models, it is called twice-universality [50], [40], which may be revealed by an Oracle inequality via the index of resolvability [13], [4]. Bayesian model averaging requires choices of predictive distributions for $\tilde{y}$ given y and γ as well as posterior probability weights for γ given y. For location and scale models (as in regression) the problem of choice of predictive distributions may be addressed by using minimax distribution for each model. However, at the level of detail of exact minimaxity there is no obvious choice of best posterior weight for γ given y. A possible formulation would be to find the adaptive procedure that minimizes the (maximal) additional expected Kullback–Leibler divergence beyond the minimax value R_γ for each family

$$\min_{Q} \max_{\gamma,\theta_\gamma} \{\mathbb{E}_{Y|\gamma,\theta_\gamma} D(P_{\tilde{Y}|Y,\gamma,\theta_\gamma} || Q_{\tilde{Y}|Y}) - R_\gamma\}.$$

Such a procedure could be said to be exact minimax for the problem of twice-universal coding.

Appendix A

First for completeness we give a standard fact from statistical decision theory (cf. [21, p. 91, Theorem 3])

Lemma 4: If procedure q is extended Bayes and has constant finite risk, then q is minimax.

Proof: Let C denote the constant risk of q, then

$$C \geq \min_{q'} \max_{\theta} R(\theta, q') \geq R_w(p_w)$$

where $R_w(p_w)$ is any Bayes risk. The extended Bayes property implies the existence of a sequence of proper priors with Bayes risk $R_{w_k}(p_{w_k})$ converging to C, hence q is minimax.

Appendix B

Here we relax the moment assumption in Theorem 1.

Theorem 1′: Assume for the location family that at least one of the $Z_1, \ldots, Z_m$ has finite expectation of $\log(1+|Z_i|)$. Then, under Kullback–Leibler loss, the best invariant predictive procedure $q^*(\tilde{y}|y)$ as in (11) is minimax for any dimension d.

Proof: Choose priors with polynomial tails $w_k(\theta) \sim (1+\|\theta\|/k)^{-d-1}$. Continue the calculation from (13)

$$\mathbb{E}_{Z,Z',\theta}(d+1)\log\frac{1+\|\theta+Z_1-Z_1'\|/k}{1+\|\theta\|/k} \leq \mathbb{E}_Z 2(d+1)\log\left(1+\frac{\|Z_1\|}{k}\right)$$

where we use the triangle inequalities and the fact that for a, b positive, $1+a+b \leq (1+a)(1+b)$. Since $\log(1+\|Z_1\|/k)$ is monotone decreasing with k and it is integrable when $k=1$ by our assumption, the risk difference goes to zero when k goes to infinity, as a result of the Monotone Convergence Theorem.

Appendix C

We give the proof for Proposition 3 using Pitman's technique which he developed from mean-squared error and we are here adapting to the Kullback–Leibler risk. The ideas from the location case are carried over to other transformations.

Proof

Using the invariance property, we can show that invariant procedures q have constant risk which is equal to

$$\mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z}|Z)}{q(\tilde{Z}|Z)}. \tag{31}$$

For linear transformation families, let Z_1^d denote $(Z_1, \ldots, Z_d)$, the $d \times d$ matrix with Z_i in the ith column for $i = 1, \ldots, d$. Define

$$\tilde{U} = (Z_1^d)^{-1}\tilde{Z},\ U_i = (Z_1^d)^{-1}Z_i, \qquad i = d+1, \ldots, m. \tag{32}$$

Note that those variables are invariant to linear transformation of the Z_i and $\tilde{Z}$, so that

$$\tilde{U} = (Y_1^d)^{-1}\tilde{Y},\ U_i = (Y_1^d)^{-1}Y_i, \qquad i = d+1, \ldots, m \tag{33}$$

where Y_1^d is the $d \times d$ matrix formed from the initial portion of Y.

Apply the invariance property, then in a manner similar to the proof for location families (Proposition 2), we find that the best invariant estimator q^* satisfies

$$q^*(\tilde{u}|e_1, \ldots, e_d, u) = \frac{p(u,\tilde{u})}{p(u)} \tag{34}$$

where e_i is the ith column of the $d \times d$ identity matrix and $u = (u_{d+1}, \ldots, u_m)$.

Next we derive the expression (in terms of y and $\tilde{y}$) for both sides of (34). By the mapping between $U, \tilde{U}$, and $Z, \tilde{Z}$ given in (32), the joint density for $Z_1, \ldots, Z_d, U$, and $\tilde{U}$ is given by

$$|z_1^d|^{m+n-d} p_{Z,\tilde{Z}}(z_1, \ldots, z_d, z_1^d u, z_1^d \tilde{u})$$

where $|z_1^d|$ denotes the absolute value of the determinant of the matrix z_1^d and $|z_1^d|^{m+n-d}$ comes out as the Jacobian. Rewriting u and $\tilde{u}$ using (33) and changing the variables of integration $z_1^d = (z_1, \ldots, z_d)$ to $\theta = z_1^d (y_1^d)^{-1}$, a $d \times d$ matrix, we obtain

$$p(u,\tilde{u}) = |y_1^d|^{n+m} \int \frac{1}{|\theta|^d} p(y, \tilde{y}|\theta)\, d\theta.$$

Then the conditional distribution is equal to

$$|y_1^d|^n \frac{\int \frac{1}{|\theta|^d} p(y,\tilde{y}|\theta)\, d\theta}{\int \frac{1}{|\theta|^d} p(y|\theta)\, d\theta}.$$

On the other hand, using the equalities in (33) and the invariance property of q^*, we have the left-hand side of (34) equal to $|y_1^d|^n q^*(\tilde{y}|y)$. So

$$q^*(\tilde{y}|y) = \frac{\int \frac{1}{|\theta|^d} p(y,\tilde{y}|\theta)\, d\theta}{\int \frac{1}{|\theta|^d} p(y|\theta)\, d\theta}.$$

For the other two transformation families, once we define U and $\tilde{U}$, the remaining proofs are the same as the one given above.

For the affine families, we define

$$\begin{aligned} \tilde{U} &= [Z_2^{d+1} - Z_1\mathbf{1}]^{-1}(\tilde{Z} - Z_1) \\ U_i &= [Z_2^{d+1} - Z_1\mathbf{1}]^{-1}(Z_i - Z_1), \qquad i = d+2, \ldots, m \end{aligned}$$

where $\mathbf{1} = (1, \ldots, 1)$ is the row vector of all ones and, thus, $Z_1\mathbf{1}$ is the matrix with d identical columns Z_1.

For multivariate location with univariate scale families, define a scalar random variable W which is the first coordinate of the vector $Z_2 - Z_1$. The remaining $d-1$ coordinates divided by W is defined to be V (thus, $(1, V) = (Z_2 - Z_1)/W$). Then we define

$$\tilde{U} = \frac{\tilde{Z} - Z_1}{W},\ U_i = \frac{Z_i - Z_1}{W}, \qquad i = 3, \ldots, m. \tag{35}$$

Appendix D

Here we give a general group-theoretic framework that encompasses the best invariant calculations of Section II. Let G be a group of transformations acted on the sample space, parameter space, and action space in the same manner as described in Section V.

A couple of assumptions will be needed and it is easy to check that they are satisfied by all the cases we considered in Section II. Suppose that $G = \{g_\theta : \theta \in \Theta\}$ and Θ are isomorphic, such that

$$g_{\theta_1} g_{\theta_2} = g_{\theta_1 \theta_2}, \quad g_{\theta_1}\theta_2 = \theta_1\theta_2$$

where $g_{\theta_1} g_{\theta_2}$ denotes the composition of two group elements, $g_{\theta_1}\theta_2$ denotes a group action on a parameter, and $\theta_1\theta_2$ is the result of group operation in Θ. Assume that each transformation

$$g_\theta \in G : (y, \tilde{y}) \to g_\theta(y, \tilde{y})$$

has differential with Jacobian denoted by $D_\theta(y, \tilde{y})$. We regard $(Y, \tilde{Y})$ as being generated as $(Y, \tilde{Y}) = g_\theta(Z, \tilde{Z})$ with $(Z, \tilde{Z}) \sim p_{Z,\tilde{Z}}$, that is,

$$p(y, \tilde{y}|\theta) = p_{Z,\tilde{Z}}(g_{\theta^{-1}}(y, \tilde{y}))D_{\theta^{-1}}(y, \tilde{y}). \tag{36}$$

Recall that a predictive distribution $Q(\cdot; y)$ is invariant if for any measurable set B, $Q(B; y) = Q(gB; gy)$, that is,

$$q(\tilde{y}; y) = q(g_a\tilde{y}; g_a y)D_a(\tilde{y}). \tag{37}$$

We further assume that $Y = (Y_{(1)}, Y_{(2)})$ where $Y_{(1)}$ has the same space as Θ. The transformation $g_\theta \in G$ is assumed to be a triple of transformations on $Y_{(1)}$, $Y_{(2)}$, and $\tilde{Y}$ independently, that is,

$$g_\theta(y_{(1)}, y_{(2)}, \tilde{y}) = (g_\theta^{(1)} y_{(1)}, g_\theta^{(2)} y_{(2)}, g_\theta^{(3)} \tilde{y}).$$

When the context is clear and we do not need to make distinctions between $g^{(1)}$, $g^{(2)}$, and $g^{(3)}$, we simply denote them all by g. Due to the independence between the three transformations, we have $D_g(y, \tilde{y})$ equal to $D_g(y_{(1)})D_g(y_{(2)})D_g(\tilde{y})$, where we use the same notation $D_g(\cdot)$ to denote the Jacobians for each of the transformations of $y_{(1)}$, $y_{(2)}$, and $\tilde{y}$.

Recall that a main technique we used in Section II is an application of a transformation based on a portion of Y (i.e., $Y_{(1)}$) to yield variables not depending on θ. Likewise, here we assume that

$$(g^{(2)}_{Y_{(1)}^{-1}} Y_{(2)}, g^{(3)}_{Y_{(1)}^{-1}} \tilde{Y}) = (U, \tilde{U})$$

has a distribution not depending on θ. In particular, $(U, \tilde{U})$ is also equal to

$$(g^{(2)}_{Z_{(1)}^{-1}} Z_{(2)}, g^{(3)}_{Z_{(1)}^{-1}} \tilde{Z}).$$

It should be noted that for affine families, $Y_{(1)}$ is not equal to the first $d+1$ variables because they are not the same space as Θ, but rather $Y_{(1)}$ corresponds to their sample mean and sample standard deviation. Similarly for multivariate location with univariate scale families.

Our aim is to use the information-theoretic tools to confirm that the generalized Bayes rule with right Haar measure is the best invariant rule with Kullback–Leibler loss. First, similarly to what we showed for location families, we can show that the risk $R(\theta, q)$ is constant and equal to

$$\mathbb{E}_{Z,\tilde{Z}} \log \frac{p_{\tilde{Z}}(\tilde{Z})}{q(\tilde{Z}; Z_{(1)}, Z_{(2)})}. \tag{38}$$

To derive the best invariant estimator, one may apply invariance property (37) with $g_a = g_{Z_{(1)}^{-1}}$ in (38), define $U = g_{Z_{(1)}^{-1}} Z_{(2)}$ and $\tilde{U} = g_{Z_{(1)}^{-1}} \tilde{Z}$, and then obtain

$$R(\theta, q) = \mathbb{E}_{Z,\tilde{Z}} \log \frac{p_{\tilde{Z}}(\tilde{Z})}{q(\tilde{U}; e, U) D_{Z_{(1)}^{-1}}(\tilde{Z})}$$

where $e = g_{Z_{(1)}^{-1}} Z_{(1)} = Z_{(1)}^{-1} Z_{(1)}$ is the identity of the group G. Let $p(\tilde{u}|u)$ denote the conditional density for $\tilde{U}$ given U deduced from $p_{Z,\tilde{Z}}$, then the risk above is equal to

$$\mathbb{E}_{Z,\tilde{Z}} \log \frac{p_{\tilde{Z}}(\tilde{Z})}{p(\tilde{U}|U) D_{Z_{(1)}^{-1}}(\tilde{Z})} + \mathbb{E}_U \left[\mathbb{E}_{\tilde{U}|U} \log \frac{p(\tilde{U}|U)}{q(\tilde{U}; e, U)} \right]$$

where the first term is a constant not depending on q and the second one is an expected Kullback–Leibler distance which is nonnegative. So q achieves the smallest risk if and only if $q(\tilde{u}; e, u) = p(\tilde{u}|u)$. Denote such an estimator by q^*. Recalling that $(U, \tilde{U})$ is also equal to $(g_{Y_{(1)}^{-1}} Y_{(2)}, g_{Y_{(1)}^{-1}} \tilde{Y})$ and using the invariance property, we have that $q^*(\tilde{u}; e, u) = q^*(\tilde{y}; y) D_{y_{(1)}}(\tilde{u})$. So

$$q^*(\tilde{y}; y) = \frac{1}{D_{y_{(1)}}(\tilde{u})} \frac{p(u, \tilde{u})}{p(u)}. \tag{39}$$

To get the final expression for $q^*(\tilde{y}; y)$, we need to write the right-hand side of (39) in terms of y and $\tilde{y}$. We first work on $p(u)$. Since $u = g_{z_{(1)}^{-1}} z_{(2)} = g_{y_{(1)}^{-1}} y_{(2)}$

$$\begin{aligned} p(u) &= \int p_Z(z_{(1)}, g_{z_{(1)}} u) D_{z_{(1)}}(u)\, dz_{(1)} \\ &= \int p_Z(z_{(1)}, g_{z_{(1)} y_{(1)}^{-1}} y_{(2)}) D_{z_{(1)}}(u)\, dz_{(1)}. \end{aligned}$$

In the integration above, $z_{(1)}$ is just a dummy variable of integration. Make the change of variables $z_{(1)} = \theta^{-1} y_{(1)}$ where θ is the new variable of integration and $y_{(1)}$ is the true observed value of $Y_{(1)}$. Then $p(u)$ becomes

$$\begin{aligned} &\int p_Z(g_{\theta^{-1}} y_{(1)}, g_{\theta^{-1}} y_{(2)}) D_{\theta^{-1} y_{(1)}}(u) \left| \frac{\partial \theta^{-1} y_{(1)}}{\partial \theta} \right| d\theta \\ &\quad = D_{y_{(1)}}(u) \int p(y|\theta) \left| \frac{\partial \theta^{-1} y_{(1)}}{\partial \theta} \right| \Big/ D_{\theta^{-1}}(y_{(1)}) d\theta \end{aligned} \tag{40}$$

where we use (36) and the chain rule of Jacobians

$$D_{\theta^{-1} y_{(1)}}(u) = D_{\theta^{-1}}(y_{(2)}) D_{y_{(1)}}(u).$$

Before we show that the ratio of the two Jacobians within the integral in (40) is a right invariant measure on Θ, we recall some standard results on right and left invariant measures. In

the following, the right and left invariant densities are denoted by h^r and h^l, respectively.

- *Result 1* (R1): Connection between Jacobians and invariant densities (see [6, pp. 408–409])

$$\left|\frac{\partial gx}{\partial x}\right| = \frac{h^l(x)}{h^l(gx)}, \qquad \left|\frac{\partial xg}{\partial x}\right| = \frac{h^r(x)}{h^r(xg)}.$$

  It follows that $h^l(g) = 1/\left|\frac{\partial gx}{\partial x}\right|_{x=e}$ where e is the identity of the group G. For example, under our assumptions on Θ, we have $h^l(g) = D_g(e)$ where $D_g(e)$ is the Jacobian for the transformation $x \to gx$ evaluated at e.

- *Result 2* (R2): $h^r(x) = \Delta(x^{-1})h^l(x)$ where Δ is called the *modulus* and satisfies $\Delta(x_1x_2) = \Delta(x_1)\Delta(x_2)$ (see [20, p. 8]).

- *Result 3* (R3): $\left|\frac{\partial\theta^{-1}}{\partial\theta}\right| h^l(\theta^{-1}) = ch^r(\theta)$ where c is a constant (see [6, p. 411]).

Now return to (40). The ratio between two Jacobians within the integral is equal to

$$\begin{aligned}
&\left|\frac{\partial\theta^{-1}}{\partial\theta}\right| \left|\frac{\partial\theta^{-1}y_{(1)}}{\partial\theta^{-1}}\right| \Big/ \left|\frac{\partial g_{\theta^{-1}}y_{(1)}}{\partial y_{(1)}}\right| \\
&= \left|\frac{\partial\theta^{-1}}{\partial\theta}\right| \frac{h^r(\theta^{-1})}{h^r(\theta^{-1}y_{(1)})} \frac{h^l(\theta^{-1}y_{(1)})}{h^l(y_{(1)})} \quad \text{by (R1)} \\
&= \left|\frac{\partial\theta^{-1}}{\partial\theta}\right| \frac{\Delta(\theta)h^l(\theta^{-1})}{\Delta(\theta)\Delta(y_{(1)}^{-1})h^l(\theta^{-1}y_{(1)})} \frac{h^l(\theta^{-1}y_{(1)})}{h^l(y_{(1)})} \quad \text{by (R2)} \\
&= C(y_{(1)})h^r(\theta) \quad \text{by (R 3)}
\end{aligned}$$

where $C(y_{(1)})$ is a constant that may depend on $y_{(1)}$. So we have

$$p(u) = D_{y_{(1)}}(u)C(y_{(1)}) \int p(y|\theta)h^r(\theta)d\theta. \tag{41}$$

Similarly, we have

$$p(u,\tilde{u}) = D_{y_{(1)}}(u)D_{y_{(1)}}(\tilde{u})C(y_{(1)}) \int p(y,\tilde{y}|\theta)h^r(\theta)d\theta. \tag{42}$$

Combining (39), (41), and (42), we obtain

$$q^*(\tilde{y};y) = \frac{\int p(y,\tilde{y}|\theta)h^r(\theta)d\theta}{\int p(y|\theta)h^r(\theta)d\theta}$$

as the best invariant estimator as claimed.

Appendix E

Here we give two basic results for relative entropy concerning an extremal characterization and lower semicontinuity with respect to vague convergence.

Let P and Q be positive measures on a measurable space X. The relative entropy is defined by $D(P\|Q) = P\log dP/dQ$ when $P \ll Q$ and $D(P\|Q) = \infty$, otherwise. It is nonnegative when $Q(1) \le P(1)$ and in such case it is equal to zero only when $Q = P$.

The following is a familiar characterization of relative entropy, expressed here for unnormalized measures.

Lemma 5: (Extremal Characterization): Consider functions f for which $\log f$ is P integrable and f is Q integrable. Then for every such f, the relative entropy satisfies the inequalities

$$\begin{aligned}
D(P\|Q) &\ge P(\log f) + P(1)\log P(1)/Q(f) \\
&\ge P(\log f) + P(1) - Q(f)
\end{aligned}$$

with equality in the first case when f is proportional to dP/dQ and equality in both cases when $f = dP/dQ$. Consequently

$$\begin{aligned}
D(P\|Q) &= \sup_f\{P(\log f) + P(1)\log P(1)/Q(f)\} \\
&= \sup_f\{P(\log f) + P(1) - Q(f)\}
\end{aligned}$$

where the supremum is over any class of nonnegative functions f such that the class of functions is dense for all positive functions in $L_1(Q)$ and the class of functions $\log f$ is dense in $L_1(P)$. (When $P \ll Q$, the supremum may be restricted to any class of functions f for which dP/dQ is a point of density in $L_1(Q)$ and $\log dP/dQ$ is a point of density of the $\log f$ in $L_1(P)$.)

The second expression represents $D(P\|Q)$ as a supremum of functionals linear in P and Q.

Proof: Consider $P \ll Q$ and f in $\mathcal{F}$. Define a positive measure R to have density $f(x)P(1)/Q(f)$ with respect to Q. Then $R(1) = P(1)$, so $D(P\|R)$ is nonnegative. The chain rule of densities gives

$$P\log dP/dQ = P\log dP/dR + P\log dR/dQ.$$

Throwing away the first term on the right since $D(P\|R) \ge 0$, we have

$$\begin{aligned}
D(P\|Q) &\ge P\log f(X)P(1)/Q(f) \\
&= P\log f + P(1)\log P(1)/Q(f)
\end{aligned}$$

with equality only when $R = P$, that is when f is proportional to dP/dQ. Equality is also reached in the limit if a supremum is taken over a class of functions which includes dP/dQ as a cluster point as indicated. For the other set of expressions use $-\log Q(f)/P(1) \ge (P(1) - Q(f))/P(1)$. Now equality holds for $f = dP/dQ$, and it is reached in limit by suprema of the lower bounds in the same way. This completes the proof.

Our key result of this appendix is the lower semicontinuity of relative entropy where the convergence for the second measure may be taken in the vague sense rather than the weak sense. As needed to support these notions of convergence, X is assumed to be a locally compact topological space with a countable base (such as a Euclidean space). Vague convergence of Q_j to some Q^* means that $Q_j(f)$ tends to $Q^*(f)$ for every bounded continuous and compactly supported f, and allows a limit with $Q^*(1) \le \liminf Q_j(1)$. Weak convergence of P_j to some P^* means not only that vague convergence holds but also that $P^*(1) = \lim_j P(1)$.

Lemma 6: (Lowers Semicontinuity): Suppose that P_j and Q_j are sequences of positive finite measures which converge vaguely to P^* and Q^*, respectively, and suppose the sequence

satisfies $P^*(1) = \lim_j P_j(1)$. Suppose also that there is an upper bound $Q_j(1) \leq a$. Then

$$\liminf_j D(P_j||Q_j) \geq D(P^*||Q^*).$$

Proof: First we refine the bounds from the previous lemma. Let P and Q be finite nonnegative measures with $Q(1) \leq a$, and let p and q be their densities with respect to some measure μ which dominates both, so that

$$D(P||Q) = P(\log p/q).$$

Under the indicated condition on the space X, it is sigma-compact. That is, as for Euclidean spaces, there is an increasing sequence of compact sets whose limit is X, and for each compact B in this sequence we can arrange to have a continuous function w with values satisfying $0 \leq w(x) \leq 1$ equaling 1 in B and tapering so that it is zero for any point in X outside the next compact set. Split

$$D(P||Q) = P(w \log p/q) + P((1-w)\log p/q).$$

For the second part on the right side we use the lower bound $P(1-w)\log P(1-w)/Q(1-w)$ which is at least

$$P(1-w)\log P(1-w)/a \geq g(P(1-w))$$

where $g(\alpha)$ is the continuous and decreasing function on the nonnegative reals which is equal to 0 for $\alpha = 0$, equal to $\alpha \log \alpha/a$ for $\alpha \leq a/e$, and equal to $-\alpha$ for $\alpha \geq a/e$.

For the first part write it as

$$P(w \log p/q) = \int (wp)\log(wp)/(wq),$$

the relative entropy between the measures with densities wp and wq. So by Lemma 5, we have that the preceding expression is

$$\sup_f [P(w \log f) + P(w) - Q(wf)]$$

where the supremum over f is taken over all functions for which $\log f$ is bounded and continuous. So adding the lower bounds on the first and second parts of the decomposition of D, we have the inequality for any such B

$$D(P||Q) \geq \sup_f [P(w\log f) + P(w) - Q(wf)] + g(P(1-w)).$$

Note that $w \log f$, wf, and w are bounded continuous and compactly supported and on the right side we have a supremum of integrals of such. Thus, applying the inequality for vaguely convergent P_j and Q_j we have that

$$\liminf_j D(P_j||Q_j) \geq \sup_f [P^*(w\log f) + P^*(w) - Q^*(wf)] + g(\limsup_j P_j(1-w)).$$

The next reasoning is similar to above steps but now applied to P^* and Q^*. Pick a dominating measure for P^* and Q^* (it may also dominate all the P_j and Q_j if we like) and let p^* and q^* be the associated densities. Using the preceding lemma once more (now for the measures with densities p^*w and q^*w) plus the hypothesis regarding $\lim_j P_j(1) = P^*(1)$ and $\lim_j P_j(w) = P^*(w)$, this bound becomes

$$\int wp^* \log p^*/q^* + g(P^*(1-w))$$

which we may write as

$$\int w[p^* \log p^*/q^* + q^* - p^*] + P^*(w) - Q^*(w) + g(P^*(1-w)).$$

Now take the limit for a sequence of such functions w increasing to 1. Noting the positivity of the integrand, we may employ monotone convergence to obtain

$$\int [p^* \log p^*/q^* + q^* - p^*] + P^*(1) - Q^*(1) + g(0)$$

which, since $g(0) = 0$, is equal to $D(P^*||Q^*)$. Thus, we conclude that

$$\liminf_j D(P_j||Q_j) \geq D(P^*||Q^*)$$

as desired.

Remark: In Section V, we apply the lower semicontinuity in the case of a fixed probability measure P and measures Q in the set of finite measures with bound $a = 1$.

Corollary: In the space of positive measures Q with bound $Q(1) \leq 1$, for any positive constant r, and any fixed finite probability measure P, the level sets $C = \{Q : D(P||Q) \leq r\}$ are compact in the vague topology.

Proof: Let Q_j be any sequence of such positive measures with bound 1 in the information ball $\{Q : D(P||Q) \leq r\}$. Since the measures have the sequence $Q_j(1)$ bounded, there will exist a vague subsequence limit Q^*. Let J be such a subsequence. Since the measures are in the information ball, by lower semicontinuity we have

$$D(P||Q^*) \leq \liminf_{j \in J} D(P||Q_j) \leq r.$$

So for any such sequence of Q_j in the information ball, there is a subsequence limit in the same ball. Thus, the information ball is compact.

Remarks: This implies for a fixed target distribution $P_{\tilde{Y}|\theta_0}$, the compactness of that part of the action space of predictive subprobability measures for which the loss is bounded by a constant.

One may also examine, for a fixed probability Q, the compactness of the set of all probabilities P for which the loss $D(P||Q) \leq r$. In this case, the set of such P is tight, because, taking $p(x) = dP/dQ(x)$, the boundedness of the Q integral of $p(X)\log_+ p(X)$ (by $r + e^{-1}$) implies the uniform integrability of the random variables $p(X)$ which is the same as tightness of the probabilities P. Thus, any vague limit of a sequence of such P_j is also a weak limit, and hence by lower semicontinuity is also in the ball.

Thus, both information balls $\{P : D(P\|Q) \leq r\}$ (for probabilities P with fixed Q) and $\{Q : D(P\|Q) \leq r\}$ (for subprobabilities Q with fixed P) are compact in the vague topology.

The link with the invariance characterization of minimax procedures is that the lower semicontinuity of the loss function or (what amounts to the same) the compactness of information balls, is at the heart of what is required for application of either the abstract Hunt–Stein results of [54], [36], or our simplified Hunt–Stein derivation for relative entropy loss.

Nonetheless, explicit demonstration of the extended Bayes property (and hence minimaxity) of the best invariant procedures is often possible by our information bound on the Bayes risk difference (Lemma 3), without any measure-theoretic or topological abstraction.

ACKNOWLEDGMENT

We are grateful for generous discussions with John Hartigan, Mihaela Aslan, Joseph Eaton, Larry Brown, and Ed George. We also want to thank the referees for their valuable suggestions.

REFERENCES

[1] J. Aitchison, "Goodness of prediction fit," *Biometrika*, vol. 62, pp. 547–554, 1975.

[2] M. Aslan, "Asymptotically minimax Bayes predictive densities," Ph.D. dissertation, Yale Univ., New Haven, CT, 2002.

[3] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1985.

[4] ——, "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems," in *Bayesian Statistics*, A. P. Dawid, J. M. Bernardo, J. O. Berger, and A. F. M. Smith, Eds. Oxford, U.K.: Oxford Univ. Press, 1998, vol. 6, pp. 27–52.

[5] A. R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[6] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.

[7] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. New York: Wieley, 1954.

[8] J. V. Bondar and P. Milnes, "Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups," *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete [Became: @J(ProbTher)]*, vol. 57, pp. 103–128, 1981.

[9] L. D. Brown, "On the admissibility of invariant estimators of one or more location parameters," *Ann. Math. Statist.*, vol. 37, pp. 1087–1136, 1966.

[10] ——, "Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters," *Ann. Math. Statist.*, vol. 39, pp. 29–48, 1968.

[11] ——, "Commentary on paper [19]," in *Jack Carl Kiefer. Collected Papers, Supplementary Volume*, L. D. Brown, I. Olkin, J. Sacks, and H. P. Wynn, Eds. New York: Springer-Verlag, 1986, pp. 20–27.

[12] K. L. Chung, *A Course in Probability Theory*. New York: Academic, 1974.

[13] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.

[14] ——, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*, vol. 41, pp. 37–60, 1994.

[15] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[16] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[17] ——, "Generalized entropy and quantization problems," in *Trans. 6th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes*, 1973, pp. 159–174.

[18] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[19] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.

[20] M. L. Eaton, *Group Invariance Applications in Statistics*. Beachwood, OH: Institute of Mathematical Statistics, 1989, IMS Lecture Notes—Monograph Series.

[21] T. S. Ferguson, *Mathematical Statistics, A Decision Theoretic Approach*. New York: Academic, 1967.

[22] R. Gallager, "Source coding with side information and universal coding," MIT Lab. Inform. Decision Syst., Cambridge, MA, Tch. Rep. LIDS-P-937, 1979.

[23] M. A. Girshick and L. J. Savage, "Bayes and minimax estimates for quadratic loss functions," in *Proc. 2nd Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1951, pp. 53–73.

[24] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[25] J. Hartigan, *Bayes Theory*. New York: Springer-Verlag, 1983.

[26] ——, "The maximum likelihood prior," *Ann. Statist.*, vol. 26, no. 6, pp. 2083–2103, 1998.

[27] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, July 1997.

[28] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Statist. Sci.*, vol. 14, no. 4, pp. 382–401, 1999.

[29] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, Berkeley, CA, 1960, pp. 361–379.

[30] H. Jeffreys, *Theory of Probability*. New York: Oxford Univ. Press, 1961.

[31] O. Kallenberg, *Foundations of Modern Probability*. Berlin, Germany: Springer-Verlag, 1997.

[32] J. Kiefer, "Invariance, minimax sequential estimation, and continuous time processes," *Ann. Math. Statist.*, vol. 28, pp. 537–601, 1957.

[33] ——, "Multivariate optimality results," *Multivariate Anal.*, pp. 255–274, 1966.

[34] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[35] S. Kullback, J. C. Keegel, and J. H. Kullback, *Topics in Statistical Information Theory*. Berlin, Germany: Springer-Verlag, 1980.

[36] L. M. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.

[37] E. L. Lehmann, *Testing Statistical Hypotheses*. New York: Springer-Verlag, 1986.

[38] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer-Verlag, 1998.

[39] F. Liang, "Exact minimax procedures for predictive density estimation and data compression," Ph.D. dissertation, Yale Univ., New Haven, CT, 2002.

[40] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714–722, May 1995.

[41] M. S. Pinsker, *Information and Information Stability of Random Variables*. San Francisco, CA: Holden Day, 1964. Translated by A. Feinstein.

[42] E. J. G. Pitman, "The estimation of location and scale parameters of a continuous population of any given form," *Biometrika*, vol. 30, pp. 391–421, 1939.

[43] D. Pollard, *A User's Guide to Measure Theoretic Probability*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[44] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[45] ——, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.

[46] ——, "Stochastic complexity," *J. Roy. Statist. Soc.*, vol. 49, no. 3, pp. 223–239, 1987.

[47] ——, *Stochastic Complexity and Statistical Inquiry*. Singapore: World Scientific, 1989.

[48] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.

[49] B. Y. Ryabko, "The encoding of a source with unknown but ordered probabilities," *Probl. Inform. Transm.*, vol. 14, pp. 71–77, 1979.

[50] ——, "Twice-universal coding," *Probl. Inform. Transm.*, vol. 3, pp. 173–177, 1984.

[51] ——, "Prediction of random sequences and universal coding," *Probl. Inform. Transm.*, vol. 24, pp. 87–96, 1988.

[52] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, Berkeley, CA, 1956, pp. 197–206.

[53] ——, "The sdmissibility of Pitman's estimator of a single location parameter," *Ann. Math. Statist.*, vol. 30, pp. 970–979, 1959.

[54] H. Strasser, *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Theory*. Berlin, Germany: De Gruyter, 1985, De Gruyter Studies in Mathematics.

[55] W. E. Strawderman, "Proper Bayes minimax estimators of the multivariate normal mean," *Ann. Math. Statist.*, vol. 42, pp. 385–388, 1971.

[56] I. Vajda, "On convergence of information contained in quantized observations," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2163–2172, 2002.

[57] O. Wesler, "Invariance theory and a modified minimax principle," *Ann. Math. Statist.*, vol. 30, pp. 1–20, 1959.

[58] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inform. Theory*, vol. 43, pp. 646–657, Mar. 1997.