JMLR: Workshop and Conference Proceedings 19 (2011) 397–419 24th Annual Conference on Learning Theory

Safe Learning: *bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity*

Peter Grünwald PDG@CWI.NL
CWI Amsterdam and Mathematical Institute, Leiden University, The Netherlands

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

We extend Bayesian MAP and Minimum Description Length (MDL) learning by testing whether the data can be substantially more compressed by a mixture of the MDL/MAP distribution with another element of the model, and adjusting the learning rate if this is the case. While standard Bayes and MDL can fail to converge if the model is wrong, the resulting "safe" estimator continues to achieve good rates with wrong models. Moreover, when applied to classification and regression models as considered in statistical learning theory, the approach achieves optimal rates under, e.g., Tsybakov's conditions, and reveals new situations in which we can penalize by $(-\log \textsc{prior})/n$ rather than $\sqrt{(-\log \textsc{prior})/n}$.

1. Introduction

1. Learning Theory; Predictor Models. In much of statistical learning and machine learning theory, the goal is to learn, based on a set of observed data $Z^n = (Z_1, \ldots, Z_n)$, a predictor $\breve{f}$ taken from some set of candidate prediction rules $\mathcal{F}$. Here each $Z_i = (X_i, Y_i)$, each X_i takes values in some set $\mathcal{X}$, each Y_i takes values in $\mathcal{Y}$, and $\mathcal{F}$ is a set of functions $f : \mathcal{X} \to \mathcal{Y}$. The Z_i are assumed to be sampled i.i.d. according to some distribution P^* on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The learned predictor $\breve{f}$ should have a small *generalization error* or *risk*, defined as $\textsc{risk}(f) := E^*[\textsc{loss}(Y, f(X))]$ where $\textsc{loss}$ is some given loss function and here, as elsewhere in this paper, $E^* = E_{(X,Y)\sim P^*}$ denotes joint expectation of (X, Y) over P^*. In a typical classification setting, $\mathcal{Y} = \{0, 1\}$ and $\textsc{loss}(y, \hat{y}) := |y - \hat{y}|$ is the 0/1-loss; in typical regression problems, $\mathcal{Y} = \mathbb{R}$ and $\textsc{loss}(y, \hat{y}) := (y - \hat{y})^2$ is the squared loss. Crucially, risk bounds are usually proved in worst-case settings, using only weak assumptions on P^*

2. Standard Statistics; Probability Models. Here one models uncertainty by a statistical model, i.e. a set of probability distributions $\mathcal{P}$, and the goal is to learn a distribution $\breve{p}$ that is a good representation of the underlying distribution P^* from which the data Z^n are sampled. Here we focus on the case that Z^n are i.i.d., $Z_i = (X_i, Y_i)$ as above, and $\mathcal{P}$ is a set of conditional distributions $p(y \mid x)$, identified by their mass functions (if $\mathcal{Y}$ is finite/countable) or densities, and extended to n outcomes by independence. Witness papers such as *The Two Cultures* (Breiman, 2001), the difference between statistical/machine learning theory and standard statistics based on probability models is often regarded as fundamental. Here, I propose a first, preliminary, attempt at an overarching, single theory of learning, as embodied by a new 'safe' estimator. It is called 'safe' because, when applied to probability models $\mathcal{P}$, then, unlike standard Bayes and MDL, it is guaranteed to perform well in the often inevitable situation that 'all models (elements of $\mathcal{P}$) are wrong, yet some are useful'.

© 2011 P. Grünwald.

Safe Estimation for Probability Models For probability models $\mathcal{P}$, the safe estimator behaves similarly to the Bayesian MAP or two-part MDL estimator. Following Barron and Cover (1991), we define the *κ-two part estimator*, written as $\ddot{p}_\kappa$, as a generalization of the MAP/MDL estimator, as follows: fix some prior distribution w and some $\kappa > 0$. For each $x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$, $\ddot{p}_\kappa$ is defined* as the $p \in \mathcal{P}$ achieving[1]

$$\min_{p \in \mathcal{P}} \{-\kappa \log w(p) - \log p(y^n \mid x^n)\}. \tag{1}$$

When $\kappa \geq 1$, then, via the Kraft inequality, (1) can be thought of as the number of bits needed to encode Y^n given X^n in a two-stage code; $-\kappa \log w(p)$ is the codelength needed to encode p, and acts as a complexity penalty. $-\log p(y^n \mid x^n)$ is the codelength of the data y^n when encoded with the help of p and x^n. To get good convergence rates, one needs to set $\kappa > 1$ (Zhang, 2006); while any fixed $\kappa > 1$ will do, for 'standard 2-part MDL' one takes $\kappa = 2$ which is mathematically convenient (Barron and Cover, 1991). In contrast, the safe estimator is defined (in Section 2, Eq. (9)) as $\breve{p}_{\text{SAFE}} = \ddot{p}_{2\breve{\kappa}_{\text{SAFE}}}$ where $\breve{\kappa}_{\text{SAFE}}$ is not fixed but determined by the data. $\breve{\kappa}_{\text{SAFE}}$ will be a small constant ≥ 2, *unless* the data indicate that the model is misspecified (wrong). Whereas ordinary Bayesian and MDL approaches can fail to converge in this case (Example 7 below), the safe estimator continues to perform well in the following sense: suppose that data Z^n are i.i.d. $\sim P^*$, as above, where for each x, $P^*(Y = \cdot \mid X = x)$ has conditional density $p^*(\cdot \mid x)$. Let q be the best approximation within $\mathcal{P}$ of p^* in terms of Kullback-Leibler (KL) divergence. Then the KL divergence between $\breve{p}_{\text{SAFE}}$ and p^* converges to the KL divergence between q and p^* at fast rates. To express this formally, for any two conditional densities p and p', we define* the *generalized KL divergence* (Grünwald, 2007) relative to P^* as

$$D^*(p' \| p) := E^*[-\log p(Y \mid X) + \log p'(Y \mid X)].$$

Then, for q satisfying $\inf_{p \in \mathcal{P}} D^*(p^* \| p) = D^*(p^* \| q)$ we prove, under suitable regularity conditions, in Theorem 1 in combination with Theorem 3 below that $D(p^* \| \breve{p}_{\text{SAFE}}) - D(p^* \| q) \to 0$, or equivalently, $D^*(q \| \breve{p}_{\text{SAFE}}) \to 0$, in probability at fast rates.

Safe Estimation for Predictor Models In our overarching approach, all models are formally defined as sets of probability distributions. Predictor models $\mathcal{F}$ are "transformed" into corresponding probability models $\mathcal{P}_{\mathcal{F}} := \{p_f \mid f \in \mathcal{F}\}$ by a standard transformation (called 'entropification' and extensively motivated from an MDL perspective by Grünwald (1999)): for each $f \in \mathcal{F}$,

$$p_f(y \mid x) := \tfrac{1}{\mathrm{Z}(\beta)} e^{-\beta \textsc{loss}(y, f(x))} \quad , \quad p_f(y^n \mid x^n) := \textstyle\prod_{i=1}^n p_f(y_i \mid x_i). \tag{2}$$

Here $\mathrm{Z}(\beta) = \int_{y \in \mathcal{Y}} \exp(\beta \textsc{loss}(y, f(x)))dy$ is a normalization factor (if $\mathcal{Y}$ is finite/countable, then here, as everywhere else in this paper, the integral should be replaced by a sum). In this preliminary study, we set β to some fixed value, say, 1 (but see Section 6). For the squared loss, $\mathrm{Z}(\beta)$ does not depend on $f(x)$; if we set $\beta = 1/2\sigma^2$, we see that (2) expresses that Y is Gaussian with mean $f(X)$ and variance σ^2. For the 0/1-loss, $\mathrm{Z}(\beta)$ does not

1. Distracting aspects of proofs (such as showing that the minimum of a function exists) have been omitted in this paper, but will be provided in the journal version. Such details are marked by a $*$, such as here*.

depend on $f(x)$ either; loss functions for which $\mathrm{Z}(\beta)$ depends on f are handled as described under Eq.(7) below. Taking logarithms in (2), we then get that the *excess risk* of any f as compared to any g is a linear function of the generalized KL divergence of the corresponding distributions:

$$\textsc{excess-risk}(g\|f) = \textsc{risk}(f) - \textsc{risk}(g) = \\ E^*[\textsc{loss}(Y, f(X)) - \textsc{loss}(Y, g(X))] = \frac{1}{\beta} D^*(p_g \| p_f). \quad (3)$$

Now let g be such that $\textsc{risk}(g) = \inf_{f \in \mathcal{F}} \textsc{risk}(f)$. Even if $\mathcal{F}$ is a good 'model', i.e. $\textsc{risk}(g)$ is small, the corresponding model $\mathcal{P}_{\mathcal{F}}$ will typically be misspecified (e.g. in the squared loss case, the 'true' noise may not be Gaussian at all). Since the safe estimator is immune to this problem, we can safely apply it to the model $\mathcal{P}_{\mathcal{F}}$. Then Theorems 1 and 3 show that the excess risk $\textsc{excess-risk}(g\|\breve{f}_{\textsc{safe}})$ converges to 0 at rates that are in many cases optimal; here $\breve{f}_{\textsc{safe}} := f$ for the f with $\breve{p}_{\textsc{safe}} = p_f$. Thus, by the construction (2), convergence in generalized KL-divergence becomes equivalent to convergence in the loss function of interest.

The Role of Convexity Our starting point is the known fact that 'standard MDL still works', i.e., (broadly speaking), $D^*(q\|\ddot{p}_2) \to 0$ at the appropriate rate if the closure $\langle \mathcal{P} \rangle$ (suitably defined as in (10) below) of the model $\mathcal{P}$ is *convex* (Li, 1999, Theorem 5.5); see also (Kleijn and van der Vaart, 2006). Our first observation is that, even if $\langle \mathcal{P} \rangle$ is not convex, then as long as we have the weaker condition

$$\inf_{p \in \mathcal{P}} D^*(p^*\|p) = \inf_{p \in \textsc{convex-hull}(\mathcal{P})} D^*(p^*\|p), \quad (4)$$

we still get that $D^*(q\|\ddot{p}_2) \to 0$ at the right rates. Now define, for $\eta \leq 1$, the model $\mathcal{P}^{(\eta)} := \{p^{(\eta)} \mid p \in \mathcal{P}\}$, where $p^{(\eta)}(y \mid x) \propto (p(y \mid x))^\eta$ (for predictor models $\mathcal{P}_{\mathcal{F}}$, this corresponds to replacing β in (2) by $\eta \cdot \beta$; a precise definition is beneath (7) below). Our second insight is that, even if (4) does not hold for $\mathcal{P}$, then still, for all η no greater than some critical value $\eta_{\textsc{crit}}$, (4) will actually hold with $\mathcal{P}$ replaced by $\mathcal{P}^{(\eta)}$ and thus 'standard MDL still works' for $\mathcal{P}^{(\eta)}$. The third insight is that the MDL estimator $\ddot{p}_{2\kappa}$ for model $\mathcal{P}$ with $\kappa = 1/\eta_{\textsc{crit}}$ is essentially equivalent to the standard MDL estimator $\ddot{p}_2^{(\eta)}$ for model $\mathcal{P}^{(\eta_{\textsc{crit}})}$; indeed, we will prove (implicitly in Theorem 3) that the MDL estimator $\ddot{p}_{2\kappa}$ with $\kappa = 1/\eta_{\textsc{crit}}$ leads to good results for the model $\mathcal{P}$. The fourth, and main, insight is that, for any given η, we can *test* whether $\eta \leq \eta_{\textsc{crit}}$, i.e. whether (4) is the case for $\mathcal{P}^{(\eta)}$, by looking at the observed data: *essentially, the likelihood of the data according to $\ddot{p}_2^{(\eta)}$ will be significantly smaller than the likelihood according to a 2-component mixture of $\ddot{p}_2^{(\eta)}$ and another, suitably chosen $p \in \mathcal{P}$, if and only if (4) does not hold.* The minus logarithm of this discrepancy is measured in terms of a function CONV-LACK, a key concept of this paper, defined formally in (8). The safe estimator is defined as $\breve{p}_{\textsc{safe}} = \ddot{p}_{2\breve{\kappa}_{\textsc{safe}}}$, i.e. it is the $2\breve{\kappa}_{\textsc{safe}}$-two-part estimator where $\breve{\kappa}_{\textsc{safe}} = 1/\eta$, and η is determined by the data: it is effectively set to the largest value for which CONV-LACK is small, i.e. for which we cannot fit the data better by a two-component mixture.

Overview of Results In Section 2 we formally define CONV-LACK and $\breve{p}_{\textsc{safe}}$. In Section 3 and 4, $\breve{p}_{\textsc{safe}}$ will be shown to converge at optimal rates up to log factors in a variety of

situations, as illustrated by examples in Section 4. Convergence of $\breve{p}_{\text{SAFE}}$ is shown in two stages: Theorem 1 bounds $D^*(q\|\breve{p})$ for arbitrary estimators $\breve{p}$ in terms of a 'redundancy' term RED and the CONV-LACK term. The redundancy term also shows up in classical MDL analyses and tends to 0 if we set $\breve{p}$ to a two-part estimator. Theorem 3 shows (essentially) that if $\breve{p}$ is set to $\ddot{p}_{2\kappa}$ with $\kappa \geq 1/\eta_{\text{CRIT}}$, then the CONV-LACK-term is small as well. Taken together, Theorem 1 and 3 imply that $\breve{p}_{\text{SAFE}}$ converges (a) at the right rates if the model is correct or convex; and (b) also if the model is "incorrect in the worst possible manner", and finally, (c) also if the model is a classification model, i.e. $\mathcal{P} = \mathcal{P}_{\mathcal{F}}$ for some set of classifiers $\mathcal{F}$, and a Tsybakov condition holds for $\mathcal{F}$.

In case (a), $\eta_{\text{CRIT}} = 1$. As shown in Example 4, if the model is in fact correct ($p^* = q$) we get the same bound on $D^*(p^*\|\breve{p}_{\text{SAFE}})$ as the bounds obtained on $D^*(p^*\|\ddot{p}_2)$ by Barron and Cover (1991), but with a larger constant factor — this is the price we have to pay for using a method that still works if the model is incorrect in a situation in which the model in fact, turns out to be correct. In the special case that $w(p^*) = w(q) > 0$, $D^*(q\|\breve{p}_{\text{SAFE}})$ will tend to 0 as $O((\log n)/n)$. In case (b), η_{CRIT} may be as small as $C/\sqrt{n}$ for some $C > 0$. Example 6 shows that $D^*(q\|\breve{p}_{\text{SAFE}})$ may then tend to 0 at rate as slow as $(\log n)/\sqrt{n}$, a worst-case bound familiar from the statistical learning literature. In case (c), $\eta_{\text{CRIT}} \asymp n^{-(1-\nu)/(2-\nu)}$ for some $0 \leq \nu \leq 1$ and the convergence rate depends on ν. In the special case that $w(q) > 0$, $D^*(q\|\breve{p}_{\text{SAFE}})$ will tend to 0 as $O((\log n)n^{-1/(2-\nu)})$; see above Example 7. The examples illustrate the generality of $\breve{p}_{\text{SAFE}}$, capturing both the common asymptotics for density estimation if the model is correct and for statistical classifier learning under the celebrated Tysbakov condition. $\breve{p}_{\text{SAFE}}$ also gives a new interpretation of the difference in complexity penalties prescribed by MDL/Bayes on the one hand and learning theory approaches (such as Structural Risk Minimization and PAC-Bayesian methods) on the other: since the $\mathcal{P}_{\mathcal{F}}$ constructed from a predictor model $\mathcal{F}$ is in general nonconvex, we may have $\eta_{\text{CRIT}} \ll 1$, and the standard MDL penalties become too small.

A third main result is Theorem 2, which gives a new PAC-Bayes style empirical generalization bound in which, if we are 'lucky' on the observed data, the codelength $(-\log w(\breve{p}_{\text{SAFE}}))$ only appears in an $O(1/n)$ rather than an $O(1/\sqrt{n})$ term, even if the empirical error of the learned classifier is not close to 0. As such it provides another step in the race to "root out the square root" that characterizes so much of the work on classification bounds in learning theory.

Related Work — Learning the Learning Rate The larger $\eta = 1/\kappa$, the more influence the data has on the chosen hypothesis $\ddot{p}_{2\kappa}$. For predictor models, the same holds for the β appearing in (2). Thus $\eta \cdot \beta$ may be viewed as the learning rate. A straightforward approach to learn it from data is to fix η and instead pick the β in (2) that minimizes overall description length of the data, as suggested by Grünwald (1999). Soon after publishing that paper, it became clear to me that this does not work (this was shown formally in (Grünwald and Langford, 2007)), and I started looking for an estimator that performs as well as if the optimal learning rate η_{CRIT} had been known in advance. The safe estimator does achieve this goal, thus ending a twelve-year long search. In a sense, $\breve{p}_{\text{SAFE}}$ learns the optimal learning rate. Note however that we cannot prove that the $\breve{\kappa}_{\text{SAFE}}$ selected by the safe estimator is equal or close to $1/\eta_{\text{CRIT}}$; we can only prove that it leads to the same asymptotic performance bounds.

A transformation similar to (2) is done in PAC-Bayesian methods (McAllester, 2003), where Bayesian averages of p_f are viewed as Gibbs distributions. Our approach is similar, yet closer in spirit to standard Bayes and MDL — There may be some relation with the advanced PAC-Bayesian analyses of Audibert (2004); Catoni (2007), who provide algorithms for predictor models that learn a learning rate (similar to $\eta = 1/\kappa$) determined by the amount of 'disagreement' on the input data $(X_1, \ldots, X_n)$ between the chosen predictor $\breve{f}$ and other predictors in $\mathcal{F}$. It would be interesting to study the connections further.

Finally, our approach can (broadly) be seen as equipping (a form of) Bayesian inference with a frequentist test, and adjusting the priors if the test indicates that the model is misspecified. Such an idea was already suggested in broad terms by other researchers, e.g. Dawid (1982). It can also be viewed as equipping (a form of) MDL with a randomness test (can we compress the data more by stepping outside the model?), an idea that goes back to the Kolmogorov complexity roots of MDL.

2. The Safe Estimator

Preliminaries A *probabilistic model* $\mathcal{P}$ is a *countable* set of conditional distributions on $\mathcal{Y}$ given $\mathcal{X}$, identified with their mass functions (in case $\mathcal{Y}$ is finite or countable) or otherwise their densities relative to Lebesgue measure, which we assume to exist; $\mathcal{X}$ can be arbitrary. We allow the $p \in \mathcal{P}$ to be defective (sum to less than one). That is, $p(y \mid x)$ can be any function such that, for all x, for all y, $p(y \mid x) \geq 0$ and $\sum_{y \in \mathcal{Y}} p(y \mid x) \leq 1$. We extend p to n outcomes by independence, i.e. $p(y^n \mid x^n) := \prod_{i=1}^n p(y_i \mid x_i)$. For given $z^n = (x_1, y_1), \ldots (x_n, y_n)$, $z_{\shortmid}^n$ is shorthand for $y^n \mid x^n$, i.e. $p(z_{\shortmid}^n) = p(y^n \mid x^n)$. All logarithms are to base e.

Crucially, the models $\mathcal{P}$ we are to consider, though countable, will usually represent very "large", "complex" sets of distributions, which may be thought of as dense (in the information closure sense, see Section 3) subsets of an even larger, "nonparametric" $\bar{\mathcal{P}}$ with $\mathcal{P} \subset \bar{\mathcal{P}}$: for example, we may consider the set $\bar{\mathcal{P}}$ of all Gaussian mixtures with an arbitrary number of components, and then define $\mathcal{P}$ as the subset of all $p \in \bar{\mathcal{P}}$ with rational-valued means and mixture coefficients.

We may fix a probability mass function w on $\mathcal{P}$, which we shall think of as the *prior distribution* on $\mathcal{P}$. An *estimator* at sample size n is a function $\breve{p} : \mathcal{Z}^n \to \mathcal{P}$ that maps each possible sequence of observations $z^n = (x_1, y_1), \ldots (x_n, y_n) \in \mathcal{Z}^n$ to some $p \in \mathcal{P}$, Following e.g. Barron and Cover (1991), the notation $\breve{p}(Z^n)$ denotes the density of the observed data Z^n according to the $\breve{p}$ that was selected (estimated) based on the same data Z^n.

Conditions We only consider combinations of $(\mathcal{P}, P^*)$ and prior w for which **(A)** for each $(x, y) \in \mathcal{Z}$, there is a $p \in \mathcal{P}$ such that $w(p)p(y \mid x) > 0$. We also assume **(B)** that for some finite integer $L_{\max} > 0$, for all n, all $z^n \in \mathcal{Z}^n$, $-\log w(\ddot{p}_2) \leq nL_{\max}$. Hence the codelength of the 2-MDL estimator is of order no larger than n. This assumption is innocuous, since it can always be satisfied by adding one or a few distributions to $\mathcal{P}$ (proof sketch in appendix). Finally, let

$$V = V(\mathcal{P}, P^*) := \operatorname{ess\,sup}_Z \sup_{p,p' \in \mathcal{P}} \frac{p(Z_{\shortmid})}{p'(Z_{\shortmid})}. \tag{5}$$

We assume **(C)** $1 < V$ and **(D)** $V < \infty$. We may think of $V(\mathcal{P}, P^*)$ as the maximum ratio between the density of $z_{\shortmid} = y \mid x$ assigned by different $p \in \mathcal{P}$, where the maximum is over

all (x, y) in the support of P^*. In case P^* has full support, the essential supremum can be replaced by the standard supremum. Assumptions (A), (B) and (C) are harmless; (D) is further discussed in Section 6.

Mixing The safe estimator is the $\breve{\kappa}_{\text{SAFE}}$-two part estimator, with $\breve{\kappa}_{\text{SAFE}}$ determined by the data. To find $\breve{\kappa}_{\text{SAFE}}$, we test, for each fixed κ, whether we can get a better fit of the data/additionally compress the data by a convex combination of $\ddot{p}_\kappa$ with a single other distribution $p' \in \mathcal{P}$. Of course, since $\mathcal{P}$ may be infinite and arbitrary, it may be the case that, no matter what data we observe, there is *always* some $p \in \mathcal{P}$ such that a convex combination of $\ddot{p}_\kappa$ and p gives a substantially better fit to the data. This problem, it turns out, can be addressed by only looking at distributions p with prior mass not much smaller than $w(\ddot{p}_\kappa)$: specifically, we require $-\log w(p) \leq \lceil -\log w(\ddot{p}_\kappa) \rceil$, where $\lceil \cdot \rceil$ denotes rounding up.

To formalize this, for any $p, p' \in \mathcal{P}$ and any $\lambda \in [0, 1]$, we define the mixture distribution $\text{MIX}(p, p', \lambda)$ as $(1 - \lambda)p + \lambda p'$, so that for a single outcome z, $\text{MIX}(p, p', \lambda)(z_{\dot{1}}) := (1 - \lambda)p(z_{\dot{1}}) + \lambda p'(z_{\dot{1}})$ (note the somewhat special notation). MIX is extended to n outcomes by independence:

$$\text{MIX}(p, p', \lambda)(z_{\dot{1}}^n) := \textstyle\prod_{i=1}^n ((1 - \lambda)p(z_{i\dot{1}}) + \lambda p'(z_{i\dot{1}})). \tag{6}$$

Our test is defined in terms of how much better fit can be achieved by the best-fitting convex combination of this form. To this end, for an arbitrary estimator $\breve{p}$, we let

$$\text{SUPMIX}(\breve{p})(Z_{\dot{1}}^n) := \sup_{p \in \mathcal{P}: -\log w(p) \leq \lceil -\log w(\breve{p}) \rceil, \lambda \in [0,1]} \text{MIX}(p, \breve{p}, \lambda)(Z_{\dot{1}}^n). \tag{7}$$

Let $\mathrm{Z}(\eta) := \sup_{x \in \mathcal{X}, p \in \mathcal{P}} \int_{y \in \mathcal{Y}} (p(y \mid x))^\eta dy$. For each $p \in \mathcal{P}$, and each $\eta \in \mathbb{R}$ with $\mathrm{Z}(\eta) < \infty$, we define $p^{(\eta)}(y|x) := (p(y|x))^\eta / \mathrm{Z}(\eta)$. Note that the $p^{(\eta)}$ all represent distributions which, in general, may be defective, even if p is not. We define $\mathcal{P}^{(\eta)} := \{p^{(\eta)} \mid p \in \mathcal{P}\}$. Since, in all our equations, every occurrence of a density $p^{(\eta)}$ will actually be as a *ratio* of two densities, i.e. $p^{(\eta)}(Z_{\dot{1}}^n)/q^{(\eta)}(Z_{\dot{1}}^n)$, we can safely write p^η instead of $p^{(\eta)}$ everywhere without affecting the results. This is what we do below. However, for interpreting our results it is useful to think of the $p^{(\eta)}$ as densities. For predictor models, η corresponds to β in (2); but see Section 6.

Convexity Lack We define the *convexity lack* of an arbitrary estimator $\breve{p}$ on data $Z_{\dot{1}}^n$ as

$$\text{CONV-LACK}(\eta, \breve{p}) = -\frac{c_\eta}{\eta} \log \frac{\breve{p}^\eta(Z_{\dot{1}}^n) w^2(\breve{p})}{\text{SUPMIX}(\breve{p}^\eta)(Z_{\dot{1}}^n)}, \tag{8}$$

where $c_\eta = 1 + C_\eta C'_\eta$, and $C_\eta = 2 + 2\eta \log V$ and $C'_\eta = 2V^{2\eta}$. The rationale behind these values will become clear in Theorem 1 and Lemma 9 below. CONV-LACK is a measure of how many more bits are needed to encode the data using a two-part code (with $\kappa = 2$) based on $\breve{p}^\eta$ (the numerator in the logarithm) as compared to the number of bits needed by the two-component mixture that provides the best fit (smallest codelength) with hindsight (the denominator). The larger this number, the more could have been gained by modelling the data with the convex hull of $\mathcal{P}$ rather than just $\mathcal{P}$. Our main insight (see Theorem 3 below) is that, if $\eta \leq \eta_{\text{CRIT}}$ (η_{CRIT}, introduced in Section 1, is formally defined below), then CONV-LACK is guaranteed to be small. This suggests to test various values of κ,

and define the safe estimator as the κ-two part estimator for the smallest value of κ for which CONV-LACK$(1/\kappa, \ddot{p}_\kappa)$ is below some fixed threshold. Here we opt for the essentially equivalent, but mathematically more convenient option to simply *add* CONV-LACK as an additional penalty to the codelength of the two-part estimator:

Safe Estimation Let $\kappa_{\max} = \lceil \sqrt{n}/(2 \log V) \rceil$ (this value is motivated below Lemma 5). The *safe estimator* $\breve{p}_{\text{SAFE}}$ is defined as the 2κ- two-part estimator for the $\kappa \in \{1, 2, \ldots, \kappa_{\max}\}$ that minimizes

$$-\log \ddot{p}_{2\kappa}(Z_1^n) - 2\kappa \log w(\ddot{p}_{2\kappa}) + \text{CONV-LACK}(\kappa, \ddot{p}_{2\kappa}). \tag{9}$$

This is just the formula for $\ddot{p}_{2\kappa}$, but with the term CONV-LACK$(\kappa, \ddot{p}_{2\kappa})$ added. Here we establish that $\breve{p}_{\text{SAFE}}$ has good theoretical properties; whether it is useful in practice is discussed in Section 6.

3. Generalization Bounds for the Safe Estimator

Preliminaries We define the (generalized) *information closure* of $\mathcal{P}$ (Barron and Cover, 1991) as

$$\langle \mathcal{P} \rangle := \{p' \mid \text{for some } P^*,\ \inf_{p \in \mathcal{P}} D^*(p^* \| p) = D^*(p^* \| p')\}, \tag{10}$$

where p' ranges over *all* conditional densities for Y given X (i.e. p' is not necessarily in $\mathcal{P}$), and P^* ranges over *all* distributions on $\mathcal{Z}$ that have some conditional density; we denote the density corresponding to P^* by $p^*(y \mid x)$. Henceforth we assume that data Z^n are sampled from a P^* that admits such a p^*. We also assume that P^* and $\mathcal{P}$ are such that there exists a *best-approximating* q, defined as a q such that:

$$D^*(p^* \| q) = \inf_{p \in \mathcal{P}} D^*(p^* \| p) \text{ and } \ V(\mathcal{P} \cup \{q\}, P^*) = V(\mathcal{P}, P^*). \tag{11}$$

From now on, for given $(\mathcal{P}, P^*)$, we fix a particular best-approximating density once and for all and keep denoting it by q. We must have $q \in \langle \mathcal{P} \rangle$, and $D^*(q \| p) \geq 0$ for all $p \in \mathcal{P}$. Our assumption that p^* and q exist simplifies the formulation of our theorems. Still, all our results continue to hold, with appropriately generalized definitions, if no such q or p^* exist*. In the well-specified case, with $q = p^* \in \langle \mathcal{P} \rangle$, we trivially have that, for $\eta = 1$,

$$\text{For all } p \in \mathcal{P}\text{: } \ E^*\left[\left(\frac{p(Z_1)}{q(Z_1)}\right)^\eta\right] \leq 1, \text{ or equivalently, } d^*_\eta(q \| p) := -\frac{1}{\eta} \log E^*\left[\left(\frac{p(Z_1)}{q(Z_1)}\right)^\eta\right] \geq 0, \tag{12}$$

as is seen by writing out the expectation in full and substituting q by p^*. Here d^*_η is the generalized Rényi divergence (Li, 1999); by a result of Li (1999), repeated as Proposition 15 in Section 5, if (12) holds then $d^*_{\eta/2}$ may be viewed as a proxy for the generalized KL divergence, since then, for all $p \in \mathcal{P}$, $D^*(q \| p) \leq C_\eta d^*_{\eta/2}(q \| p) \leq C_\eta D^*(q \| p)$, where $C_\eta = 2 + 2\eta \log V$ is a constant. This is a key idea for our proofs. Classical theorems on two-part MDL inference for the well-specified case (Barron and Cover, 1991; Zhang, 2006; Grünwald, 2007) invariably make use of (12) at some point in the proofs; so do classical results on Bayesian consistency (Doob, 1949), in which (12) is used to establish that $\{p(Z_1^n)/q(Z_1^n)\}_{n=1,2\ldots}$ forms a martingale. It can be shown (Li, 1999; Kleijn and van der Vaart, 2006) that (12) still holds for $\eta = 1$ if $\langle \mathcal{P} \rangle$ is convex, or, more generally, if (4) holds. This is the fundamental

reason why the MDL and Bayesian convergence bounds still hold in that setting. In fact, (4) with $\mathcal{P}^{(\eta)}$ in the role of $\mathcal{P}$ is equivalent to (12), as follows from Lemma 9 in Section 5 (proof sketch in Appendix). If (4) does not hold for $\eta = 1$, then (12) does not hold for $\eta = 1$ and MDL and Bayes may not converge (Example 7 below). Luckily, for many types of $\mathcal{P}$, one can still show that (12) holds for some $\eta < 1$. Thus it makes sense to define the *critical exponent* η_{CRIT} as the largest value of η such that, for all $p \in \mathcal{P}$, (12) holds. It is useful to generalize the idea slightly. We define, for $u \geq 0$, the *u-critical exponent* $\eta_{\text{CRIT}}(u)$ as

$$\eta_{\text{CRIT}}(u) := \sup \left\{ \eta \leq 1 \ : \ \text{for all } p \in \mathcal{P}, \quad \log E^* \left[\left(\frac{p(Z_1)}{q(Z_1)} \right)^{\eta} \right] \leq \frac{u}{n} \right\}. \tag{13}$$

$\eta_{\text{CRIT}}(0)$ is just the critical value as defined before. In Section 1 we cheated a little, writing η_{CRIT} for $\eta_{\text{CRIT}}(u)$ for the u which gives the best bounds; see below. Whenever we write "WHP" ('with high probability'), we really mean "for all $K \geq 0$, with P^*-probability at least $1 - e^{-K}$, Z^n satisfies...".

Theorem 1 (Oracle Bound) *Let $(\mathcal{P}, P^*)$ and w satisfy conditions (A)-(D) of Section 2 with V as in (5), q as in (11), η_{CRIT} as in (13) and $\kappa_{\max}$ as above (9). Let $\breve{p}$ be an arbitrary estimator. Let $Z^n \sim P^*$. WHP, uniformly for all $\eta \in \{1, 1/2, 1/3, 1/4, \ldots, 1/\kappa_{\max}\}$, all $u \in \{0, 1, \ldots, nL_{\max}\}$, we have*

$$D^*(q \| \breve{p}) \leq \frac{C_\eta}{n} \left(\text{RED}(2/\eta, \breve{p}) + \text{CONV-LACK}(\eta, \breve{p}) + \frac{u}{\min\{\eta, \eta_{\text{CRIT}}(u)\}} + R \right), \tag{14}$$

where $C_\eta = 2 + 2\eta \log V$. The term CONV-LACK *is given by (8). The term* RED *is given by*

$$\text{RED}(2/\eta, \breve{p}) = -\log \frac{w^{2/\eta}(\breve{p}) \breve{p}(Z_1^n)}{q(Z_1^n)} = -\frac{2}{\eta} \log w(\breve{p}) - \log \breve{p}(Z_1^n) + \log q(Z_1^n). \tag{15}$$

The remainder term is $R = O\left(\frac{K + \log n + \log(2 + \lceil -\log w(\breve{p}) \rceil)}{\min\{\eta, \eta_{\text{CRIT}}(u)\}} \right)$.*

RED stands for 'redundancy'. It can be interpreted as the extra number of nats needed to code the data using a two-part code based on $\breve{p}$ (with $\kappa = 2/\eta$), as compared to the code based on the best-approximating q, and under mild conditions RED$/n$ will tend to 0, WHP (Example 4). For predictor models as in (2), $\breve{p} = p_{\breve{f}}$ and $q = p_g$, and then RED$/n$ can be thought of as the difference in empirical risk between $\breve{f}$ and the optimal g, 'penalized' by $-(2/\eta n) \log w(\breve{p})$. The RED and CONV-LACK terms depend on the data; the third term in (14) becomes 0 if we set $u = 0$; for the possibility $u > 0$ see Example 6. The fourth term is a remainder term which does not depend on the data except for the term $\log(2 - \log w^{1/\eta}(\breve{p}))$ which, by assumption (B), if $\breve{p}$ is a 2-part or the safe estimator, is bounded by $O(\log n)$.

We can now motivate the definition $\breve{p}_{\text{SAFE}}$: among all two-part estimators, it is the one minimizing (14), ignoring the remainder term. To see this, note that the third term in (14) does not depend on the data, and the redundancy term can be written as the sum of two terms plus a term $\log q(Z_1^n)$ that does not depend on our choice of estimator $\breve{p}$. Thus, $\breve{p}_{\text{SAFE}}$ minimizes an upper bound on the KL divergence between $\breve{p}$ and the best-approximating q.

Let us compare (14) to the bounds on the standard two-part MDL estimator as derived by Barron and Cover (1991) under the assumption that $q = p^*$. One of their main results

(Theorem 4, as later strengthened by Zhang (2006) and (Grünwald, 2007, Section 15.3)) implies that, for all $\kappa \geq 1$,

$$E_{Z^n \sim P^*}\left[\, D^*(p^* \| \ddot{p}_{2\kappa}) \,\right] \leq \tfrac{C}{n} E_{Z^n \sim P^*}\left[\, \textsc{red}(2\kappa, \ddot{p}_{2\kappa}) \,\right], \tag{16}$$

where $C \leq 2 + 2 \log V$ as above. This provides a frequentist justification for the 2-part MDL estimator $\ddot{p}_{2\kappa}$ with $\kappa \geq 1$, since $\ddot{p}_{2\kappa}$ is in fact defined as the $p \in \mathcal{P}$ that minimizes $\textsc{red}(2\kappa, p)$. Apart from the 'in-expectation' rather than 'in-probability' formulation, the "only" relevant difference to our result is the CONV-LACK term, which appears in (14) because we do not require a correct model. In Section 4 we show that for many combinations of P^* and $\mathcal{P}$, the CONV-LACK term will be small, WHP, for $\breve{p} = \breve{p}_{\textsc{safe}}$. In such cases Theorem 1 states something similar to Barron and Cover's result ($D^*(q\|\breve{p}) \to 0$), but without the often unrealistic requirement that $p^* \in \langle \mathcal{P} \rangle$.

Theorem 2 (Empirical Bound) *Assume the notations and conditions of Theorem 1. WHP, uniformly for all $\eta \in \{1, 1/2, 1/3, \ldots, 1/\kappa_{\max}\}$, we have,*

$$E^*[-\log \breve{p}(Z_{\text{l}})] \leq \tfrac{1}{n}\left(-\log \breve{p}(Z^n_{\text{l}}) - \log w(\breve{p})^{2/\eta} + 2\textsc{conv-lack}(\eta, \breve{p}) + R\right), \tag{17}$$

with remainder term $R = O\left(\sqrt{n(K + \log n + \log(2 - \log w(\breve{p})))}\right)$.

The proof of this result is similar to the proof of Theorem 1 and will be provided in the full paper. Note that the weights of the main terms on the right side in Theorem 2 and 1 are different. In Theorem 2, the left term's weight is reduced from C_η to 1, and the weight of the right term (CONV-LACK) is reduced from C_η (which is always ≥ 2) to 2. Unlike the 'oracle' bound Theorem 1, the 'empirical' bound Theorem 2 gives useful information without knowledge of $E^*[-\log q(Z_{\text{l}})]$. If $\mathcal{P} = \mathcal{P}_{\mathcal{F}}$ for a classification model $\mathcal{F}$, then the first term on the right-hand side is the empirical risk $\beta^{-1} n^{-1} \sum_{i=1}^n \textsc{loss}(y_i, \breve{f}(x_i))$ and the bound becomes similar to the PAC-Bayesian and Occam's Razor (OR) bounds (McAllester, 2003; Blumer et al., 1987). Yet, by Condition (B), for $\breve{p} = \breve{p}_{\textsc{safe}}$, the remaining error term R/n is of order $O(\log n/\sqrt{n})$ rather than $O\left(\sqrt{-\log w(\breve{p})/n}\right)$ as it would be for PAC-Bayes and OR-bounds. In this sense, for data Z^n such that for some $\gamma > 0$ (say, $\gamma = 0.1$), for all $f \in \mathcal{F}$, the empirical loss of f on Z^n is larger than γ, the bound of Theorem 2 will be stronger than the best PAC-Bayesian or OR bounds. In other cases Theorem 2 gives weaker bounds than PAC-Bayes, since unlike PAC-Bayes it does not improve if $\breve{f}$ has empirical error ≈ 0; also it is not suitable for randomized classifiers. Theorem 2 is thus a first step, to be improved in future work.

4. What Actually Happens

Theorem 3 *Assume the notations and conditions of Theorem 1. Let c_η be as in (8). Fix $u \geq 0$ and let $\eta \leq \eta_{\textsc{crit}}(u)$. WHP, we have* $\textsc{conv-lack}(\eta, \breve{p}) \leq c_\eta\left(3\textsc{red}(2/\eta, \breve{p}) + \frac{u}{\eta}\right) + R$, *with remainder term $R = O((\log n + K + u)/\eta)$*

Applying Theorem 1 to the safe estimator $\breve{p}_{\textsc{safe}}$ with $\eta \leq \eta_{\textsc{crit}}(u)$, and using Theorem 3 to rewrite CONV-LACK, and using the fact that, if two inequalities hold with high probability,

the combined inequality also holds with high probability (see Proposition 12 in Section 5), we see that for all $\eta \in \{1, 1/2, 1/3, \ldots, \kappa_{\max}\}$ with $\eta \leq \eta_{\text{CRIT}}(u)$, the safe estimator achieves, WHP,

$$\begin{aligned} D^*(q\|\breve{p}_{\text{SAFE}}) &\leq \frac{C}{n}\left(\text{RED}(2/\eta, \breve{p}_{\text{SAFE}}) + \frac{u}{\eta} + c_\eta 3\text{RED}(2/\eta, \breve{p}_{\text{SAFE}}) + \frac{c_\eta}{\eta}u + R'\right) \\ &= \frac{C(1+3c_\eta)}{n}\left(\text{RED}(2/\eta, \ddot{p}_{2/\eta}) + R''\right) + \frac{C(1+c_\eta)}{n} \cdot \frac{u}{\eta} \\ &\leq \frac{C''}{n}\left(\text{RED}(2/\eta, \ddot{p}_{2/\eta}) + \frac{u}{\eta} + R''\right), \end{aligned} \tag{18}$$

with constant $C'' = C(1 + 3c_\eta)$ and new remainder term $R'' = O((\log n + K)/\eta)$. As long as we use (18) with $u = 0$ (directly below) or $u = 1$ (in Lemma 5) and $\eta \leq \eta_{\text{CRIT}}(u)$, then the terms u/η and R'' are at most of the same order as the first term $(-2/\eta) \log w(\ddot{p}_{2\eta})$ in RED, and hence do not affect the obtained convergence rates of $\breve{p}_{\text{SAFE}}$.

Example 4 [Best-Case: Model $\mathcal{P}$ correct or convex] Suppose that P^* is in the information closure of $\mathcal{P}$, i.e. $q = p^*$. Then $\eta_{\text{CRIT}}(0) = 1$, and, using (18) with $u = 0$ and $\eta = \eta_{\text{CRIT}}(0)$, WHP,

$$D^*(p^*\|\breve{p}_{\text{SAFE}}) = D^*(q\|\breve{p}_{\text{SAFE}}) \leq \frac{C''}{n}\left(\text{RED}(2, \ddot{p}_2) + R''\right), \tag{19}$$

where by Barron and Cover's original analysis, we would get (16). Except for the in-probability rather than in-expectation formulation, the only real difference is that the KL divergence is bounded in terms of a larger constant factor. This is the price we pay for not knowing in advance that our model was, in fact, correct, while using a procedure that still leads to good results if it is incorrect.

Barron and Cover (1991) show that, for a wide variety of probabilistic models $\mathcal{M}$, there exist countable discretizations $\mathcal{P} \subset \mathcal{M}$ and corresponding priors w on $\mathcal{P}$ such that $\frac{1}{n}E_{Z^n \sim P^*}\left[\text{RED}(2, \ddot{p}_2)\right]$ is equal to the minimax convergence rate in KL risk if $\mathcal{M}$ is "non-parametric", or equal to the minimax rate up to a $\log n$-factor if $\mathcal{M}$ is "parametric". Using Markov's inequality*, WHP $\text{RED}(2, \ddot{p}_2)$ on the data (as in (19)) is not larger than a constant factor times its expectation $E_{Z^n \sim P^*}\left[\text{RED}(2, \ddot{p}_2)\right]$. This implies that the standard MDL estimator also achieves the minimax rate in probability (up to a logarithmic factor in the parametric case). Hence, by (19), so do we. A similar story* can be told if $\langle\mathcal{P}\rangle$ is convex or if the weaker condition (4) holds; again, up to constant factors, the safe estimator performs as well as the two-part estimator, which converges at near-optimal rates.

When the Model is Wrong Define $D^*_{\text{SQ}}(q, p) = E^*[(\log p(Z_\text{I})/q(Z_\text{I}))^2]$. Such a variation of generalized KL divergence was earlier considered by, e.g., Kleijn and van der Vaart (2006). The lemma below shows that if the model is wrong, then the value of $\eta_{\text{CRIT}}(u)$ depends on the relation between D^*_{SQ} and D^*. The lemma is not really new, being a direct translation of existing results of e.g. Tsybakov (2004) from '$\mathcal{F}$-space with loss function LOSS' to '$\mathcal{P}$-space with log-loss'.

Lemma 5 *Assume the notations and conditions of Theorem 1. Suppose further* **(E)** *that for some $A > 0$ and some $0 \leq \nu \leq 1$, for all $p \in \mathcal{P}$, $D^*_{\text{SQ}}(q, p) \leq A(D^*(q\|p))^\nu$. Then, for all $u > 0$, we have*

$$\eta_{\text{CRIT}}(u) \geq \min\left\{\frac{1}{2\log V}, B\left(\frac{u}{n}\right)^{\frac{1-\nu}{2-\nu}}\right\} \quad \textit{where } B = (2/eA)^{\frac{1}{2-\nu}}. \tag{20}$$

When $\mathcal{P}$ represents a classification model containg the Bayes classifier, condition (E) above specializes to the celebrated condition of (Mammen and Tsybakov, 1999; Tsybakov, 2004); the κ in (Tsybakov, 2004) is equal to ν^{-1} in our notation. In particular, we automatically have $D^*_{\mathrm{SQ}}(q,p) \leq (\log V)^2$ so (E) always holds for $\nu = 0$ and $A = (\log V)^2$. Using (20) with these values, and using $1/2 < \sqrt{2/e}$, it follows that for all $u \geq 1$,

$$\eta_{\mathrm{CRIT}}(u) \geq \tfrac{1}{2\log V}\sqrt{\tfrac{u}{n}} \geq 1/\kappa_{\max}, \tag{21}$$

which explains why we could restrict η to $\eta > \kappa_{\max}$; see Example 6. If (E) holds for some $\nu > 0$ though, then $\eta_{\mathrm{CRIT}}(u)$ is of larger order than $\sqrt{u/n}$ and things get better; see below Example 6.

Example 6 [Worst-Case] Let $u \geq 1$ and let $\eta = \eta_{\mathrm{CRIT}}(u)$. Using (18), and (21), we see that WHP,

$$\begin{aligned} D^*(q\|\breve{p}_{\mathrm{SAFE}}) &\leq \tfrac{C''}{n}\left(\mathrm{RED}(2/\eta, \ddot{p}_{2/\eta}) + \tfrac{u}{\eta} + R''\right) \\ &= \tfrac{C''}{n}\left(-2c\sqrt{\tfrac{n}{u}}\log w(\ddot{p}_{2/\eta}) - \log \tfrac{\ddot{p}_{2/\eta}(Z_1^n)}{q(Z_1^n)} + c\sqrt{u\cdot n} + R''\right) \end{aligned}$$

for some constant $c = 2\log V$. Differentiating with respect to u shows that a minimum is achieved* for $u \approx -2\log w(\ddot{p}_{2/\eta})$. The resulting expression becomes

$$C''\cdot 2c\sqrt{\tfrac{-\log w(\ddot{p}_{2/\eta})}{n}} + C''\left(-\tfrac{1}{n}\log \tfrac{\ddot{p}_{2/\eta}(Z_1^n)}{q(Z_1^n)} + R''\right). \tag{22}$$

For classification models, this bound is familiar from the computational learning literature. Now suppose that (E) holds for some $\nu > 0$. Then we can achieve better bounds: by the reasoning below Theorem 3, $\breve{p}_{\mathrm{SAFE}}$ converges at the same rate as the κ-MDL estimator with $\kappa = \eta_{\mathrm{CRIT}}(1)^{-1} = O(n^{\frac{1-\nu}{2-\nu}})$. From (18) we see that in the special case that $q \in \mathcal{P}$, $w(q) > 0$, (18) gives a rate in probability of $(\log n)/n^{1/(2-\nu)}$, which, for classification models, is equal to the minimax optimal rate in expectation (Tsybakov, 2004) up to a log factor. The next example illustrates this for $\nu = 1$; in the full paper* we will also provide examples involving regression and classification with $0 < \nu < 1$.

Example 7 [Bayesian inconsistency and Tsybakov's Condition]
Grünwald and Langford (2007) showed that standard MDL and Bayesian inference can be inconsistent in various ways if $P^* \notin \langle\mathcal{P}\rangle$, for countable models $\mathcal{P} = \{p_0, p_1, \ldots\}$ that are really classification models, i.e. $\mathcal{P} = \mathcal{P}_{\mathcal{F}}$ with $\mathcal{F} = \{f_0, f_1, \ldots\}$ with $p_j = p_{f_j}$ as given by (2), where $\mathcal{Y} = \{0,1\}$ and LOSS is the 0/1-LOSS. In these examples, p_0 has positive prior $w(p_0) > 0$ independent of the sample size, and for some $\delta > 0$, for all $j > 0$, it holds $D^*(p_0\|p_j) > \delta$, i.e. $\mathrm{RISK}(p_j) > \mathrm{RISK}(p_0) + \beta^{-1}\delta$. In the examples Tsybakov's condition (E) holds with $\nu = 1$ but only for very large A. Since thus $q = p_0$ and, by Lemma 5, $\eta_{\mathrm{CRIT}}(1) > 1/C$ for some very large constant independent of n, it follows from (18) that the safe estimator converges WHP at rate $O((-\log w(p_0) + R'')/n) = O(\log n/n)$, much faster than the worst-case $O(1/\sqrt{n})$. However, explicit calculation of $\eta_{\mathrm{CRIT}}(1)$ shows that it is indeed very small, and since standard MDL and Bayesian MAP use an η equal to 1 or 2, it comes as no surprise that in this scenario they do not converge at all, i.e. with probability 1, for all large n, they select a distribution/classifier $p \neq p_0$, as was shown formally by Grünwald and Langford (2007).

5. The Proofs

Preliminary Results Our main tool is Proposition 8 below, a bound for ratios of probability densities, similar to earlier inequalities by Barron and Cover (1991); Li (1999); Zhang (2006). Below TR is a function mapping distributions to other distributions (we use notation as in (6)).

Proposition 8 *Let Z^n be i.i.d. $\sim P^*$. Let $\mathcal{P}$ be a countable set of (possibly defective) conditional densities for $\mathcal{Z}^n$ and let $\breve{p}$ be an arbitrary estimator. Let $\mathcal{Q}$ be another set of (possibly defective) conditional densities for $\mathcal{Z}^n$. Let $\mathrm{TR} : \mathcal{P} \to \mathcal{Q}$ be a function mapping distributions in $\mathcal{P}$ to distributions in $\mathcal{Q}$. Let w be a (potentially defective) probability mass function on $\mathcal{P}$. Let $\eta > 0$. Then WHP,*

$$d^*_\eta(\mathrm{TR}(\breve{p})\|\breve{p}) \;\leq\; \tfrac{1}{n}\left(-\log \tfrac{w(\breve{p})\breve{p}(Z^n_{\text{I}})}{\mathrm{TR}(\breve{p})(Z^n_{\text{I}})} + \tfrac{K}{\eta} + \tfrac{1}{\eta}\log \textstyle\sum_{p\in\breve{\mathcal{P}}} w(p)^\eta\right). \tag{23}$$

Proof We bound the probability that (23) does *not* hold:

$$\begin{array}{l}
P^*\left(\eta d^*_\eta(\mathrm{TR}(\breve{p})\|\breve{p}) > \tfrac{1}{n}\left(-\eta\log\tfrac{w(\breve{p})\breve{p}(Z^n_{\text{I}})}{\mathrm{TR}(\breve{p})(Z^n_{\text{I}})} + K + \log\textstyle\sum_{p\in\breve{\mathcal{P}}} w(p)^\eta\right)\right) = \\
P^*\left(\ \log\left(\tfrac{\breve{p}(Z^n_{\text{I}})}{\mathrm{TR}(\breve{p})(Z^n_{\text{I}})}\right)^\eta > -\log\tfrac{w^\eta(\breve{p})}{\sum w^\eta(p)} + K + n\log E^*\left(\tfrac{\breve{p}(Z_{\text{I}})}{\mathrm{TR}(\breve{p})(Z_{\text{I}})}\right)^\eta\ \right) \leq \\
P^*\left(\text{There exists } p\in\mathcal{P} \text{ with } \left(\tfrac{p(Z^n_{\text{I}})}{\mathrm{TR}(p)(Z^n_{\text{I}})}\right)^\eta > e^K\left(\tfrac{\sum w(p)^\eta}{w^\eta(p)}\right)E^*\left(\tfrac{p(Z^n)}{\mathrm{TR}(p)(Z^n)}\right)^\eta\ \right) \leq \\
\textstyle\sum_{p\in\mathcal{P}} P^*\left(\ \left(\tfrac{p(Z^n_{\text{I}})}{\mathrm{TR}(p)(Z^n_{\text{I}})}\right)^\eta > e^K\left(\tfrac{\sum w(p)^\eta}{w^\eta(p)}\right)E^*\left(\tfrac{p(Z^n)}{\mathrm{TR}(p)(Z^n)}\right)^\eta\ \right) \leq e^{-K},
\end{array}$$

where the equality is basic rewriting, the first inequality follows from exponentiating both sides, absorbing n into the expectation E^* (which can be done since the Z_i are i.i.d.) and weakening, the second is the union bound, and the third is an instance of Markov's inequality. ∎

In some applications we set, for all $p \in \mathcal{P}$, $\mathrm{TR}(p)$ equal to the best approximating density q, and then the first term on the right in (23) is equal to $\mathrm{RED}(1/\eta, \breve{p})$; the inequality is then a weakening of Zhang's, who provides an expectation rather than an in-probability form. In other applications (e.g. below Eq. (37)), $\mathrm{TR}(p)$ actually varies with p, and in this form, the inequality is new. We will apply this proposition in two different ways. In the first type of application, the goal is to get a (high-probability) upper bound on the left-hand side of (23). In the second type of application, the goal is to upper bound $-\log \mathrm{TR}(\breve{p})(Z^n_{\text{I}})$. Thus, we rewrite (23) equivalently as:

$$-\frac{1}{n}\log \mathrm{TR}(\breve{p})(Z^n_{\text{I}}) \leq -\frac{1}{n}\log w(\breve{p})\breve{p}(Z^n_{\text{I}}) + R, \text{ with } R = -d^*_\eta(\mathrm{TR}(\breve{p})\|\breve{p}) + \frac{K + \log\sum_{p\in\breve{\mathcal{P}}} w(p)^\eta}{n\eta}. \tag{24}$$

In such applications, we take a value of η guaranteeing $d^*_\eta(\mathrm{TR}(\breve{p})\|\breve{p}) \geq 0$ or (as e.g. in Lemma 14), we allow $d^*_\eta(\mathrm{TR}(\breve{p})\|\breve{p})$ to be negative but not too negative, so that the bound remains useful.

Let $p, p' \in \mathcal{P}$ such that $D^*(p\|p') \geq 0$, and let* $\lambda^\circ := \arg\min_{\lambda\in[0,1]} D^*(p^*\|\mathrm{MIX}(p,p',\lambda)) = \arg\max_{\lambda\in[0,1]} D^*(\mathrm{MIX}(p,p',\lambda)\|p))$ (if more than one λ achieves the extremum, we take the smallest). Our second key result states that if $D^*(\mathrm{MIX}(p,p',\lambda^\circ)\|p)$ is small, then $E^*[\mathrm{MIX}(p,p',\lambda^\circ)(Z_{\text{I}})/p(Z_{\text{I}})]$ cannot be much larger than 1; equivalently $d^*_1(p\|\mathrm{MIX}(p,p',\lambda^\circ))$ cannot be much smaller than 0:

Lemma 9 *Let $(\mathcal{P}, P^*)$ be as on page 401, $V = V(\mathcal{P}, P^*)$ be as in (5), and let $p, p' \in \mathcal{P}$ be such that $D^*(p\|p') \geq 0$, and λ° be as above. (a) If $\lambda^\circ = 0$ (p is closer to p^* than any mixture of p and p') then for all $\lambda \in [0,1]$, $d_1^*(p\|\text{MIX}(p,p',\lambda)) \geq 0$; otherwise, (b), $-d_1^*(p\|\text{MIX}(p,p',\lambda^\circ)) \leq 2V^2 D^*(\text{MIX}(p,p',\lambda^\circ)\|p)$.*

Proof Let $g(\lambda) = D^*(\text{MIX}(p,p',\lambda)\|p)$. Then $g(0) = 0, g(1) \leq 0$. We first need the following (proof straightforward by differentiation, see appendix):

Proposition 10 *1. $g'(0) = E^*\left(\frac{p'(Z_\text{I})}{p(Z_\text{I})}\right) - 1$; if $\lambda^\circ = 0$ then (1a) $g'(0) \leq 0$; if $\lambda^\circ > 0$ then (1b) $g'(0) > 0$, $g'(\lambda^\circ) = 0$ and $g'(1) \leq 0$; 2. if $p(Z_\text{I}) = p'(Z_\text{I})$ P^*-almost surely, then (2a) $g'(\lambda) = g''(\lambda) = 0$ on $\lambda \in [0,1]$. Otherwise (2b) $g''(\lambda) < 0$ on $[0,1]$ and $\max_{\lambda\in[0,1]} |g''(\lambda)| \leq \min_{\lambda\in[0,1]} V^2 |g''(\lambda)|$.*

Abbreviate $d_1^*(p\|\text{MIX}(p,p',\lambda))$ to $d^*(\lambda)$ and $g(\lambda^\circ) = D^*(\text{MIX}(p,p',\lambda^\circ)\|p)$ to D^*, and note that

$$-d^*(\lambda) = \log E^*\left(\tfrac{(1-\lambda)p(Z_\text{I})+\lambda p'(Z_\text{I})}{p(Z_\text{I})}\right) \leq E^*\left(\tfrac{(1-\lambda)p(Z_\text{I})+\lambda p'(Z_\text{I})}{p(Z_\text{I})}\right) - 1 = \lambda g'(0). \tag{25}$$

In case (1a) and (2a) the result is now immediate, so assume (1b) and (2b). Then by a first-order Taylor approximation of g', for some $0 \leq \lambda_1 \leq \lambda^\circ$, $g'(0) = g'(\lambda^\circ) - \lambda^\circ g''(\lambda_1) = \lambda^\circ |g''(\lambda_1)|$, so that (25) gives $-d^*(\lambda^\circ) \leq (\lambda^\circ)^2 |g''(\lambda_1)|$. Also, by a 2nd-order Taylor expansion of g around λ° we find, for some $0 \leq \lambda_2 \leq \lambda^\circ$, that $0 = g(0) = (1/2)(\lambda^\circ)^2 g''(\lambda_2) + g(\lambda^\circ)$, so $D^* = (1/2)(\lambda^\circ)^2 |g''(\lambda_2)|$. Combining with our expression for $d^*(\lambda^\circ)$, we get $\frac{-d^*(\lambda^\circ)}{D^*} \leq 2\frac{|g''(\lambda_1)|}{|g''(\lambda_2)|}$. The result now follows by part (2b) of Proposition 10. ■

The next proposition is about varying exponents rather than mixture coefficients:

Proposition 11 *Let $\mathcal{P} = \{p, p'\}$ be such that $V(\mathcal{P}, P^*) < \infty$ and $D^*(p\|p') > 0$. Then (a): letting $g(\eta) = \log E^*(p'(Z_\text{I})/p(Z_\text{I}))^\eta = -\eta d_\eta^*(p\|p')$, we have $g(0) = 0$, $g(\eta)$ is decreasing at $\eta = 0$ and $\exp(g(\eta))$ is strictly convex, so that there exists at most one $\eta' > 0$ with $g(\eta') = 0$, and $g(\eta)$ is increasing for $\eta \geq \eta'$. (b) Define λ° as in Lemma 9. If $\lambda^\circ = 0$ (p is closer to p^* than any mixture of p and p') then $\forall \eta \in (0,1)$, $d_\eta^*(p\|p') > 0$.*

Proof (a) is just differentiation of $E^*(p'(Z_\text{I})/p(Z_\text{I}))^\eta$ (see proof of Lemma 5); details omitted. (b) is immediate from (a) because by Lemma 9, part (a), $d_1^*(p\|p') = d_1^*(p\|\text{MIX}(p,p',1)) = -g(1) \geq 0$. ■

The following proposition provides the glue that ties all our inequalities together:

Proposition 12 (log-Bonferroni) *Let $\mathcal{J}$ be a finite or countably infinite set. Let $\{Y_j\}_{j\in\mathcal{J}}$ be a collection of random variables, let $\{a_j\}_{j\in\mathcal{J}}$ be a collection of constants in $\mathbb{R}$ and let $\{f_j\}_{j\in\mathcal{J}}$ be a collection of increasing functions $\mathbb{R} \to \mathbb{R}$. Suppose that for all $j \in \mathcal{J}$, WHP, $Y_j \leq a_j + f_j(K)$. Then, for any collection of positive numbers $\{w_j\}_{j\in\mathcal{J}}$ such that $\sum_{j\in\mathcal{J}} w_j = 1$, we have, WHP,*

$$\textit{For all } j \in \mathcal{J},\ Y_j \leq a_j + f_j(K - \log w_j).$$

This result is a straightforward consequence of the union bound that appears in one form or other in many COLT papers; for convenience there is a proof in the appendix.

Notation common to the Proofs In all proofs below we make use of the following concepts: let w be a prior for a countable set of densities $\mathcal{P}$. Let $p \in \mathcal{P}$. Relative to w and P^*, we define* the *optimal density at p's description length* as

$$\textsc{opt}(p) := \mathop{\arg\min}_{p' \in \mathcal{P}: -\log w(p') \leq \lceil -\log w(p) \rceil} D^*(p^* \| p) \tag{26}$$

For an estimator $\breve{p}$, $\textsc{opt}(\breve{p})$ is itself a random variable, representing the best distribution (closest in KL divergence to p^*) with prior no smaller (up to rounding) (or "complexity" no larger, up to rounding) than the $\breve{p}$ selected for the given data Z_1^n. We further define

$$\textsc{opt}(\mathcal{P}) = \{p \in \mathcal{P} : p = \textsc{opt}(p') \text{ for some } p' \in \mathcal{P}\}. \tag{27}$$

Now define for $p \in \mathcal{P}$, $\textsc{optmix}(p) := \textsc{mix}(\textsc{opt}(p), p, \lambda^\circ)$, where $\lambda^\circ \in [0,1]$ minimizes*

$$E^*[-\log \textsc{mix}(\textsc{opt}(p), p, \lambda)(Z_1)] = E^*[-\log ((1-\lambda)\textsc{opt}(p)(Z_1) + \lambda p(Z_1))]. \tag{28}$$

$$\text{Note that } D^*(\textsc{optmix}(\breve{p}^\eta) \| \breve{p}^\eta) = \max_{\lambda \in [0,1]} D^*(\textsc{mix}(\textsc{opt}(\breve{p}^\eta), \breve{p}^\eta, \lambda) \| \breve{p}^\eta) \geq D^*(\textsc{opt}(\breve{p}^\eta) \| \breve{p}^\eta). \tag{29}$$

5.1. Proofs of Main Results

Proof of Theorem 1 We first consider an arbitrary fixed η and a fixed u. We have:

$$D^*(q\|\breve{p}) = D^*(q\|\textsc{opt}(\breve{p})) + \frac{1}{\eta} D^*(\textsc{opt}(\breve{p}^\eta)\|\breve{p}^\eta) \leq D^*(q\|\textsc{opt}(\breve{p})) + \frac{1}{\eta} D^*(\textsc{optmix}(\breve{p}^\eta)\|\breve{p}^\eta), \tag{30}$$

where the equality is straightforward from the definition of D^* and the inequality follows from (29). By Lemma 14, we can bound the term $D^*(q\|\textsc{opt}(\breve{p}))$ from above and rewrite (30) to get, WHP,

$$D^*(q\|\breve{p}) \leq \frac{C_\eta}{n} \left(-\frac{1}{\eta} \log \frac{w(\breve{p}) \textsc{optmix}(\breve{p}^\eta)(Z_1^n)}{q^\eta(Z_1^n)} + \frac{u}{\eta'} + R_1 \right) + T \tag{31}$$

with $C_\eta = 2 + 2\eta \log V$, $\eta' = \min\{\eta, \eta_{\text{CRIT}}(u)\}$, R_1 as in Lemma 14 and

$$T = -\frac{C_\eta}{\eta} d_1^*(\textsc{opt}(\breve{p}^\eta)\|\textsc{optmix}(\breve{p}^\eta)) + \frac{1}{\eta} D^*(\textsc{optmix}(\breve{p}^\eta)\|\breve{p}^\eta), \tag{32}$$

We proceed to rewrite T. $-d_1^*(\textsc{opt}(\breve{p}^\eta)\|\textsc{optmix}(\breve{p}^\eta))$ may very well be *positive*, but by Lemma 9, applied with $(\mathcal{P}, p, p') \leftarrow (\mathcal{P}^{(\eta)}, \textsc{opt}(\breve{p})^\eta, \breve{p}^\eta)$ (the notation indicates that e.g. $\mathcal{P}$ in Lemma 9 is instantiated to $\mathcal{P}^{(\eta)}$ as used above), we can bound (32) to get: $T \leq \eta^{-1}(C_\eta C'_\eta + 1) D^*(\textsc{optmix}(\breve{p}^\eta)\|\breve{p}^\eta)$, where we used
$D^*(\textsc{optmix}(\breve{p}^\eta)\|\textsc{opt}(\breve{p})^\eta) \leq D^*(\textsc{optmix}(\breve{p}^\eta)\|\breve{p}^\eta)$, and $C'_\eta = 2V^{2\eta}$. Letting $w'(p^\eta) := w(p)$, this can be further bounded using Lemma 13 below, with $(\breve{p}, \mathcal{P}, w) \leftarrow (\breve{p}^\eta, \mathcal{P}^{(\eta)}, w')$ (notation as explained above). This gives, WHP:

$$T \leq \frac{C_\eta}{n} \left(-\frac{c_\eta}{\eta} \log \frac{\breve{p}^\eta(Z_1^n) w(\breve{p})^2}{\textsc{supmix}(\breve{p}^\eta)(Z_1^n)} + R'_1 \right) = \frac{C_\eta}{n} \left(\textsc{conv-lack}(\eta, \breve{p}) + R'_1 \right), \tag{33}$$

where $R_1' = \eta^{-1} c_\eta 2K$ and $c_\eta = C_\eta C_\eta' + 1$ and we used the definition of CONV-LACK. Now apply Proposition 8 as in (24), with $(\mathcal{P}, \breve{p}, \mathcal{Q}, \mathrm{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}^{(\eta)}, \breve{p}^\eta, \mathcal{Q}, \mathrm{OPTMIX}(\cdot), w', 1)$, where $\mathcal{Q} = \{\mathrm{OPTMIX}(p^\eta) \mid p \in \mathcal{P}\}$ and w' as above. Since, by Lemma 9, part (b), $d_1^*(\mathrm{OPTMIX}(p^\eta) \| \mathrm{OPT}(p)^\eta) \geq 0$, we find that WHP, $-\log \mathrm{OPTMIX}(\breve{p}^\eta)(Z_{\mathrm{I}}^n) \leq -\log w(\breve{p}) \breve{p}^\eta(Z_{\mathrm{I}}^n) + K$. Substituting this and (33) into (31), we get with Proposition 12 (with $|\mathcal{J}| = 3, w_j = 1/3$) that WHP,

$$D^*(q \| \breve{p}) \leq \frac{C_\eta}{n} \left(-\frac{1}{\eta} \log \frac{w(\breve{p})^2 \breve{p}^\eta(Z_{\mathrm{I}}^n)}{q^\eta(Z_{\mathrm{I}}^n)} + \frac{u}{\eta'} + \text{CONV-LACK}(\eta, \breve{p}) + R_2 + R_2' \right), \tag{34}$$

where $R_2 = (4(K + \log 3) + 4 \log(2 + \lceil -\log w(\breve{p}) \rceil))/\eta'$ and $R_2' = c_\eta (2(K + \log 3))/\eta$. The result now follows for a fixed value of u and η. To prove that it holds uniformly for $u \in \{0, 1, 2, \ldots, nL_{\max}\}$, $\eta \in \{1, 1/2, 1/3, \ldots, 1/\kappa_{\max}\}$, we use Proposition 12; details omitted*.

Lemma 13 *Let $(\mathcal{P}, P^*)$ and w be as on page 401. We have WHP,*

$$D^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) \leq \tfrac{C_1}{n} \left(-\log \tfrac{\breve{p}(Z_{\mathrm{I}}^n) w^2(\breve{p})}{\mathrm{SUPMIX}(\breve{p})(Z_{\mathrm{I}}^n)} + 2K \right)$$

where $C_1 = 2 + 2 \log V$ is a constant and SUPMIX *is defined as in (7) above.*

Proof By Proposition 15, applied with $\eta = 1$, we get

$$D^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) \leq C_1 d_{1/2}^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) - (C_1 - 1) d_1^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) \leq C_1 d_{1/2}^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}),$$

where $C_1 = 2 + 2 \log V$ and the final inequality follows because by Proposition 11, $d_1^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) \geq 0$. We now let $\mathcal{Q} = \{\mathrm{MIX}(p_0, p_1, \lambda) : p_0, p_1 \in \mathcal{P}, \lambda \in [0, 1]\}$ be the set of two-component mixtures of elements of $\mathcal{P}$, and apply Proposition 8, with $(\mathcal{P}, \breve{p}, \mathcal{Q}, \mathrm{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}, \breve{p}, \mathcal{Q}, \mathrm{OPTMIX}(\cdot), w^2, 1/2)$ (notation as below (32)). We get that, WHP, $d_{1/2}^*(\mathrm{OPTMIX}(\breve{p}) \| \breve{p}) \leq -\frac{1}{n} \log \frac{w^2(\breve{p}) \breve{p}(Z_{\mathrm{I}}^n)}{\mathrm{OPTMIX}(\breve{p})(Z_{\mathrm{I}}^n)} + \frac{2K}{n} \leq -\frac{1}{n} \log \frac{w^2(\breve{p}) \breve{p}(Z_{\mathrm{I}}^n)}{\mathrm{SUPMIX}(\breve{p})(Z_{\mathrm{I}}^n)} + \frac{2K}{n}$. ■

Lemma 14 *Assume the conditions and notation of Theorem 1. For all $0 < \eta \leq 1$, WHP,*

$$D^*(q \| \mathrm{OPT}(\breve{p})) \leq \frac{C_\eta}{n} \left(-\frac{1}{\eta} \log \frac{w(\breve{p}) \mathrm{OPTMIX}(\breve{p}^\eta)(Z_{\mathrm{I}}^n)}{q^\eta(Z_{\mathrm{I}}^n)} + \frac{u}{\eta'} + R_1 \right) - \frac{C_\eta}{\eta} d_1^*(\mathrm{OPT}(\breve{p}^\eta) \| \mathrm{OPTMIX}(\breve{p}^\eta)),$$

where $C_\eta = 2 + 2\eta \log V$, $\eta' = \min\{\eta, \eta_{\mathrm{CRIT}}(u)\}$ and remainder $R_1 = (3K + 4 \log(2 + \lceil -\log w(\breve{p}) \rceil))/\eta'$. (Note that $d_1^*(\mathrm{OPT}(\breve{p}^\eta) \| \breve{p}^\eta)$ may be negative).

Proof We apply Proposition 15 with η set to η'. This gives $D^*(q \| \mathrm{OPT}(\breve{p})) \leq C_{\eta'} d_{\eta'/2}^*(q \| \mathrm{OPT}(\breve{p})) + (C_{\eta'} - 1)R$, where $R = -d_{\eta'}^*(q \| \mathrm{OPT}(\breve{p}))$. By Proposition 11, part(a), if $\eta' R \geq 0$ then, since $\eta' \leq \eta_{\mathrm{CRIT}}(u)$, we have $\eta' R \leq -\eta_{\mathrm{CRIT}}(u) d_{\eta_{\mathrm{CRIT}}(u)}^*(q \| \mathrm{OPT}(\breve{p})) \leq u/n$. It follows that $R \leq u/(\eta' n)$, so we have

$$D^*(q \| \mathrm{OPT}(\breve{p})) \leq C_{\eta'} d_{\eta'/2}^*(q \| \mathrm{OPT}(\breve{p})) + \tfrac{1}{\eta'}(C_{\eta'} - 1)\tfrac{u}{n}. \tag{35}$$

Now fix some $p_0 \in \mathcal{P}$. Set $w'(p_0) = 1$, $\mathrm{TR}(p_0) = q$ and apply Proposition 8 with $(\mathcal{P}, \breve{p}, \mathcal{Q}, \mathrm{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}, p_0, \{q\}, \mathrm{TR}(\cdot), \eta'/2)$. The $\breve{p}$ in the proposition is a degenerate

estimator that is always equal to the fixed p_0 and does not depend on the data, and $\mathrm{TR}(\breve{p})$ is always equal to q. We get that WHP,

$$d^*_{\eta'/2}(q\|p_0) \leq \tfrac{1}{n}\left(-\log \tfrac{p_0(Z_1^n)}{q(Z_1^n)} + \tfrac{2K}{\eta'}\right). \tag{36}$$

Note that we can enumerate the elements of $\mathrm{OPT}(\mathcal{P})$ as given by (27) as $\{p_1, p_2, \ldots\}$ where, for all j, it holds $j-1 \leq \lceil -\log w(p_j)\rceil \leq j$. Using the log-Bonferroni Proposition 12 with $|\mathcal{J}| = \mathbb{N}$, $w_j = 1/j(j+1)$ and $f_j(K) = 2K/\eta'(u)$, we get that, WHP, uniformly for all $p_j \in \mathrm{OPT}(P)$,

$$\begin{array}{rcl} d^*_{\eta'/2}(q\|p_j) & \leq & \tfrac{1}{n}\left(-\log \tfrac{p_j(Z_1^n)}{q(Z_1^n)} + \tfrac{2}{\eta'}(K + 2\log(j+1))\right) \\ & \leq & \tfrac{1}{n}\left(-\log \tfrac{p_j(Z_1^n)}{q(Z_1^n)} + \tfrac{2}{\eta'}(K + 2\log(2 - \lceil -\log w(p_j)\rceil))\right). \end{array} \tag{37}$$

We now let $\mathcal{R} = \{\mathrm{OPTMIX}(p^\eta) \mid p \in \mathcal{P}\}$, and define a prior w' on $\mathcal{R}$ with, for $r \in \mathcal{R}$, $w'(r) := w(p)$ for the p such that $r = \mathrm{OPTMIX}(p^\eta)$. We set $\mathrm{TR}(p') := \mathrm{OPT}(\breve{p}^\eta)$. We now use Proposition 8 again (in the form (24)) with $(\mathcal{P}, \breve{p}, \mathcal{Q}, \mathrm{TR}(\cdot), w, \eta) \leftarrow (\mathcal{R}, \mathrm{OPTMIX}(\breve{p}^\eta), \mathrm{OPT}(\mathcal{P})^\eta, \mathrm{TR}(\cdot), w', 1)$. (notation as below (32); effectively we use $\mathrm{OPTMIX}(\breve{p}^\eta)$ as an estimator here.). We get, WHP,

$$-\frac{1}{n}\log \mathrm{OPT}(\breve{p}^\eta)(Z_1^n) \leq -\frac{1}{n}\log w(\breve{p})\mathrm{OPTMIX}(\breve{p}^\eta)(Z_1^n) - d^*_1(\mathrm{OPT}(\breve{p}^\eta)\|\mathrm{OPTMIX}(\breve{p}^\eta)) + \frac{K}{n}. \tag{38}$$

Dividing (38) by η, using $\eta^{-1}\log \mathrm{OPT}(\breve{p}^\eta) = \log \mathrm{OPT}(\breve{p})$, and then combining with (35) and (37), with p_j set to $\mathrm{OPT}(\breve{p})$, we get, WHP,

$$\begin{array}{l} D^*(q\|\mathrm{OPT}(\breve{p})) \leq C_{\eta'} \cdot \\ \quad \left(\frac{1}{n}\left(-\frac{1}{\eta}\log \frac{w(\breve{p})\mathrm{OPTMIX}(\breve{p}^\eta)(Z_1^n)}{q^\eta(Z_1^n)} + R\right) - \frac{1}{\eta}d^*_1(\mathrm{OPT}(\breve{p}^\eta)\|\mathrm{OPTMIX}(\breve{p}^\eta)) + \frac{1}{\eta'}\frac{u}{n}\right), \end{array} \tag{39}$$

where $R = (2/\eta')(K + 2\log(2 - \lceil -\log w(p_j)\rceil)) + (1/\eta)K$, which is no greater than R_1 in the statement of the lemma. This proves the result for $\eta \leq \eta_{\mathrm{CRIT}}(u)$ (for then $\eta' = \eta$). For the case that $\eta > \eta_{\mathrm{CRIT}}(u)$, note that then $C_\eta > C_{\eta'}$. Because by definition of q the left-hand side of (39) must be nonnegative, we have WHP that both (39) holds and its right-hand side is nonnegative, so that with the same probability, (39) holds with $C_{\eta'}$ replaced by C_η. The result follows. ■

Proposition 15 *Let P^* be a distribution on $\mathcal{Z}$, let p an q be conditional distributions for Y given X, and let $V = V(\{p, q\}, P^*)$ defined as in (5). For all $\eta \leq 1$ and all $C_\eta \geq 2 + 2\eta \log V$, we have*

$$D^*(q\|p) \leq C_\eta \cdot d^*_{\eta/2}(q\|p) - (C_\eta - 1)d^*_\eta(q\|p).$$

In particular, if $d^*_\eta(q\|p) \geq 0$, then $D^*(q\|p) \leq C_\eta d^*_{\eta/2}(q\|p)$, i.e. the generalized KL divergence is upper bounded by a constant times the generalized Rényi divergence of order $1/2\eta$.

Proof This result is a straightforward extension of a result due to Andrew Barron and Jonathan Li, published in Li's (1999) thesis. See Lemma 5.11, page 67 and Lemma 5.12, page 73 of (Li, 1999). d_η^* corresponds to $\log c = \log \int fg/f^*$ in Li's notation; p^* corresponds to f in Li's notation, $p(\cdot \mid x)^\eta$ corresponds to g in Li, and $q(\cdot \mid x)^\eta$ corresponds to f^*. Our argument is slightly more involved than Li's since we allow conditioning on x; this also accounts for the extra $\eta \log V$ term in the constant C_η. For convience, we provide a full proof in the appendix. ■

Proof of Theorem 3 We fix a $\hat{p}$ and $\hat{\lambda}$ be such that $\textsc{supmix}(\breve{p}^\eta) = \textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \hat{\lambda})$, i.e. $\hat{p}$ and $\hat{\lambda}$ achieve* the supremum in (7) applied with estimator $\breve{p}^\eta$.
Let $\dot{\lambda} = \arg\min_{\lambda \in \{0, 1/n, 2/n, \ldots, 1\}} - \log \textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \lambda)(Z_{\shortmid}^n)$ be a discretized version of $\hat{\lambda}$. We have:

$$\begin{aligned} -\log \frac{\breve{p}^\eta(Z_{\shortmid}^n) w^2(\breve{p})}{\textsc{supmix}(\breve{p}^\eta)(Z_{\shortmid}^n)} &\leq -\log \frac{q^\eta(Z_{\shortmid}^n)}{\textsc{supmix}(\breve{p}^\eta)(Z_{\shortmid}^n)} - \log \frac{\breve{p}^\eta(Z_{\shortmid}^n) w^2(\breve{p})}{q^\eta(Z_{\shortmid}^n)} \\ &\leq -\log \frac{q^\eta(Z_{\shortmid}^n)}{\textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \dot{\lambda})(Z_{\shortmid}^n)} + V^\eta + \eta \textsc{red}(2/\eta, \breve{p}), \end{aligned} \tag{40}$$

where in the second inequality we used a simple first-order Taylor approximation, showing that $-\log \textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \dot{\lambda})(Z_{\shortmid}^n) \leq -\log \textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \hat{\lambda})(Z_{\shortmid}^n) + V^\eta$ (details omitted). We now come to the crucial step: we will prove that WHP, we have

$$-\log \frac{q^\eta(Z_{\shortmid}^n)}{\textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \dot{\lambda})(Z_{\shortmid}^n)} \leq -\log w^2(\breve{p}) + \log(n+1) + K + u. \tag{41}$$

This result follows by applying Proposition 8 to an extended model $\mathcal{P}'$ with a prior w' defined, at sample size n, as follows: $\mathcal{P}' = \{\textsc{mix}(p_0^\eta, p_1^\eta, \lambda) : p_0^\eta, p_1^\eta \in \mathcal{P}, \lambda \in [0,1]\}$; and, for $\lambda \in \Lambda := \{0, 1/n, \ldots, 1\}$, $w'(\textsc{mix}(p_0^\eta, p_1^\eta, \lambda)) := w(p_0^\eta) \cdot w(p_1^\eta)(n+1)^{-1}$. Thus, $\mathcal{P}'$ is the set of all two-component mixtures of $\mathcal{P}^{(\eta)}$; and w' has its support on all two-component mixtures with $\lambda \in \Lambda$, and puts mass 0 on all other mixtures. Note that w' is indeed a prior, i.e. $\sum_{p_0, p_1 \in \mathcal{P}, \lambda \in \Lambda} w'(\textsc{mix}(p_0^\eta, p_1^\eta, \lambda)) \leq 1$. We now apply Proposition 8 in the form (24) with, for all $p' \in \mathcal{P}'$, $\textsc{tr}(p') := q^\eta$, and η in the proposition set to 1, and with the estimator that, for data z^n, chooses $\textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \dot{\lambda})$. That is, we set
$(\mathcal{P}, \breve{p}, \mathcal{Q}, \textsc{tr}(\cdot), w, \eta) \leftarrow (\mathcal{P}', \textsc{mix}(\hat{p}^\eta, \breve{p}^\eta, \dot{\lambda}), \{q^\eta\}, \textsc{tr}(\cdot), w', 1)$. This gives (41), where we also used (a), by definition, $-\log w'(\textsc{mix}(\hat{p}_0^\eta, \breve{p}_1^\eta, \dot{\lambda})) \leq -\log w(\breve{p})^2 + \log(n+1)$; and (b): since $\eta \leq \eta_{\textsc{crit}}(u)$, we have that $d_1^*(q^\eta \| p^\eta) \geq -u/n$ for both $p = \breve{p}$ and $p = \hat{p}$, which implies, from the definition of d_1^*, that $d_1^*(q^\eta \| (1-\lambda)\hat{p}^\eta + \lambda \breve{p}^\eta) \geq -u/n$ for all $\lambda \in [0,1]$, in particular for $\dot{\lambda}$.
Combining (40) and (41), using the definition of CONV-LACK, it follows that, WHP,

$$\begin{aligned} \textsc{conv-lack} &\leq c_\eta \textsc{red}(2/\eta, \breve{p}) + \tfrac{c_\eta}{\eta} \left(-\log w^2(\breve{p}) + \log(n+1) + K + u + V^\eta \right) \\ &= c_\eta \textsc{red}(4/\eta, \breve{p}) + \tfrac{c_\eta}{\eta} \left(\log(n+1) + K + u + V^\eta \right). \end{aligned}$$

Now with some relatively straightforward manipulations* we get that we get,WHP, $\textsc{red}(4/\eta, \breve{p}) \leq 3\textsc{red}(2/\eta, \breve{p}) + \frac{2K + 2u + 2\log 2}{\eta}$, so that, using the log-Bonferroni Proposition 12 with $\mathcal{J} = 2$, the above becomes $\textsc{conv-lack} \leq 3c_\eta \textsc{red}(2/\eta, \breve{p}) + R$, and the result follows.

Proof of Lemma 5 A second-order Taylor expansion of $E^*(p(Z_i)/q(Z_i))^\eta = E^*(e^{\eta \log p(Z_i)/q(Z_i)})$ around $\eta = 0$ shows that, for all $\eta > 0$, for some $0 \leq \eta' \leq \eta$, we have:

$$E^*\left(\frac{p(Z_i)}{q(Z_i)}\right)^\eta = 1 - \eta E^*[-\log p(Z_i)/q(Z_i)] + \frac{1}{2}\eta^2 E^*\left(\frac{p(Z_i)}{q(Z_i)}\right)^{2\eta'}\left(\log \frac{p(Z_i)}{q(Z_i)}\right)^2 \leq 1 - \eta D^* + \frac{1}{2}\eta^2 V^{2\eta} D^*_{\mathrm{SQ}},$$

where we abbreviate $D^*(q\|p)$ to D^* and $D^*_{\mathrm{SQ}}(q,p)$ to D^*_{SQ}, and we replaced all factors in the expectation in the second order by their maximum. From now on we repeatedly use $D^*(q\|p) \geq 0$ which holds because q is best-approximating, It is sufficient to show that the right-hand side of this expression is bounded by $1+u/n$ if we plug in $\eta \leq \eta_{\mathrm{CRIT}}(u)$ as defined above. Dividing the inequality by η and using assumption (E), it is thus sufficient if we can show that

$$-D^* + \eta (D^*)^\nu \cdot b \leq \eta^{-1}(u/n) \tag{42}$$

where we set $b = \frac{A}{2}V^{2\eta}$. We may further assume $$(D^*)^{1-\nu} \leq \eta b, \tag{43}$$

for if this does not hold, then $-D^* = -(D^*)^\nu (D^*)^{1-\nu} \leq -(D^*)^\nu \eta b$ and then (42) holds trivially. Now first consider the case $0 < \nu < 1$. From (43) it follows that $D^* \leq (\eta b)^{1/(1-\nu)}$. By (42), it is thus sufficient if we can show that $\eta \cdot (\eta b)^{\nu/(1-\nu)} b \leq \eta^{-1} u/n$. Solving for η gives $\eta^{2+\frac{\nu}{1-\nu}} \leq \frac{u}{n} b^{-1/(1-\nu)}$, which can be rewritten to $\eta \leq C$, where $C = \left(\frac{u}{n}\right)^{\frac{1-\nu}{2-\nu}} b^{-1/(2-\nu)}$. Thus, weakening the requirement, it is sufficient if $\eta \leq \min\{1/(2\log V), C\}$. But if $\eta \leq 1/(2\log V)$, then $b^{-1} \geq 2/(eA)$, so it is also sufficient if $\eta \leq \min\left\{\frac{1}{2\log V}, B\left(\frac{u}{n}\right)^{\frac{1-\nu}{2-\nu}}\right\}$. (20) now follows for the case $0 < \nu < 1$. The limiting cases $\nu = 0$ and $\nu = 1$ can be handled similarly; we omit details.

6. Discussion and Future Work

The great advances we made were already summarized on page 3; but currently, our work also has at least two major restrictions : (a) $V = V(\mathcal{P}, P^*)$ as in (5) must be bounded; and (b) V occurs in the definition of CONV-LACK, so that it must be known in order to apply the safe estimator. Neither restriction is problematic for classification models, as long we used a fixed β in (2); both are problematic for e.g. standard regression models though. As to (a), currently our results only hold for such models if P^* has bounded support. In future work, we hope to replace the strong $V < \infty$ condition with a weaker condition on moments of P^*. As to (b), we do have a version of all our results in which V is replaced by its empirical counterpart $\bar{V} = \sup_{i\in\{1,\ldots,n\}} \sup_{p,p'\in\mathcal{P}} p(Z_{i\mathrm{i}})/p'(Z_{i\mathrm{i}})$, but with worse constants. We hope to refine this in future work.

Another issue is that, even if known, V or $\bar{V}$, appearing in CONV-LACK, may be so large as to make the approach useless in practice (even aside from computational issues, which, in this preliminary study, we decided not to deal with at all). We should note though that our current results hold for arbitrary priors w, in particular, priors with very heavy tails. Most priors used in practice have lighter tails, i.e. $\sum_{p\in\mathcal{P}} w^\rho(p) < \infty$ for some $\rho < 1$. For such priors, the theorems still hold for the prior w' defined as $w'(p) \propto w^\rho(p)$ rather

than the original w. As a result, the safe estimator $\breve{p}_{\text{SAFE}}$ defined relative to w' rather than w will effectively choose simpler distributions (with higher $w(p)$) for the same data, but all occurrences of V in our theorems can be replaced by V^ρ, which can lead to a serious improvement in the size of CONV-LACK. A related idea is to consider the β in predictor models $\mathcal{F}$ as in (2) as an additional parameter, to be equipped with a prior and fitted to the data. Since q and $\breve{p}$ in Theorem 1 may then refer to different predictors with different β's, β will act as a 'local' learning rate whereas η, shared by all distributions, is a 'global' learning rate. Preliminary investigations suggest that this leads to better bounds in some cases.

Acknowledgments

Supported in part by the IST Programme of the EU, under the PASCAL NoE, IST-2002-506778. I would like to thank Andrew Barron, Tim van Erven and Rui Castro for some very useful discussions.

References

J.Y. Audibert. *PAC-Bayesian statistical learning theory.* PhD thesis, Université Paris VI, 2004.

A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

L. Breiman. Statistical modeling: the two cultures (with discussion). *Statistical Science*, 16(3):199 –215, 2001.

O. Catoni. *PAC-Bayesian Supervised Classification.* Lecture Notes-Monograph Series. IMS, 2007.

A.P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–611, 1982. Discussion: pages 611–613.

J.L. Doob. Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique*, pages 23–27, Paris, 1949.

P. Grünwald. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA, 2007.

P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007. DOI 10.1007/s10994-007-0716-7.

P. D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 171–182, 1999.

B. Kleijn and A. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2), 2006.

J.Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27: 1808–1829, 1999.

D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

Tong Zhang. From ϵ-entropy to KL entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*, 34(5):2180–2210, 2006.

Appendix A. Additional Proofs

1. Condition (B) can be made to hold by adding one or a few distributions to $\mathcal{P}$ For example, in the classification case, it suffices to include the trivial distribution p_0 into $\mathcal{P}$, where, for all $x \in \mathcal{X}$, $p_0(Y = 1 \mid X = x) = p_0(Y = 0 \mid X = x) = 1/2$, and assign it some prior mass $w_0(p_0) > 0$. Then for all sequences z^n, for all $\kappa \geq 1$,

$$\begin{array}{rcl} -\log w(\ddot{p}_\kappa) & \leq & -\log w(\ddot{p}_1) \leq -\log w(\ddot{p}_1) - \log \ddot{p}_1(z_1^n) \\ & \leq & -\log w_0 - \log p_0(z_1^n) = -\log w_0 + n \log 2 \leq n L_{\max}, \end{array} \tag{44}$$

for suitably chosen $L_{\max}$. Clearly, this approach extends to all $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with finite or compact $\mathcal{Y}$. If $\mathcal{Y}$ is not compact, then, by our assumption (**D**) that $V(\mathcal{P}, P^*) < \infty$, the interval $[a, b]$ with $a = \operatorname{ess\,inf}_{Z \in \mathcal{Z}, p \in \mathcal{P}} p(Z_1)$ and $b = \operatorname{ess\,sup}_{p \in \mathcal{P}, Z \in \mathcal{Z}} p(Z_1)$, is bounded. It then suffices to include a density $p_{a,b}$ with prior w_0 such that, for all $x \in \mathcal{X}$, $p_{a,b}(Y = \cdot \mid X = x)$ is uniform on $[a, b]$. If the end points on the interval are not known, we can discretize candidate end points to integers and put a prior v on these end points satisfying, for both end points $c \in \{a, b\}$, $-\log v(c) \approx 2 \log \min\{|c|, 1\}$ (Grünwald, 2007). We can define the defective distribution $p_0(y \mid x) := \max_{a,b} p_{a,b}(y \mid x) v(a) v(b)$ and repeat the reasoning in (44).

2. Equivalence of (4) and (12) We will only show that equivalence holds in the idealized case in which the best-approximating q is actually a member of $\mathcal{P}$. This should be sufficient, since the goal of establishing equivalence is merely to give some intuition about the meaning of (12) (that's why we only put it in the appendix); the equivalence is not needed in any of our results, whose proofs invariably rely on (12) rather than (4). Assume then that $D(p^* \| q) = \inf_{p \in \mathcal{P}} D(p^* \| p)$ and that $q \in \mathcal{P}$. We only show equivalence for $\eta = 1$; extension to other η is immediate. We first prove (4) $\Rightarrow$ (12). If (4) holds, then for all $p \in \mathcal{P}$, $D^*(q \| \textsc{mix}(q, p, \lambda))$ has its minimum λ° (as defined above Lemma 9) at $\lambda^\circ = 0$. It then follows by Lemma 9 that $d_1^*(q \| p) = d_1^*(q \| \textsc{mix}(q, p, 1)) \geq 0$, and we see that (12) holds.

We next prove (12)$\Rightarrow$ (4). Suppose that (12) holds for all $p \in \mathcal{P}$. Without loss of generality let $\mathcal{P} = \{p_1, p_2, \ldots\}$. Then for any p' in the convex hull of $\mathcal{P}$, say $p' = \sum_{j=1}^{\infty} \alpha_j p_j$

with all $\alpha_j \geq 0$ and $\sum \alpha_j = 1$, we have $E^*(p'(Z_\text{l})/q(Z_\text{l})) = \sum_{j=1}^{\infty} \alpha_j E^*(p_j(Z_\text{l})/q(Z_\text{l})) \leq 1$. Thus $E^*(p'(Z_\text{l})/q(Z_\text{l})) - 1 \leq 0$ and hence, by Proposition 10, part (1), the derivative of the concave function $D^*(\text{MIX}(q, p', \lambda)\|q)$ is ≤ 0 at $\lambda = 0$. This implies that $D(p^*\|q) \leq D(p^*\|\text{MIX}(q, p', \lambda))$ for all $\lambda > 0$, in particular $D(p^*\|q) \leq D(p^*\|p')$; this shows that (4) holds.

3. Lemma 9 – proof of Proposition 10 Differentiation gives:

$$g'(\gamma) = \frac{d}{d\gamma} g(\gamma) = -E^* \left(\frac{p(Z) - p'(Z)}{(1-\gamma)p(Z) + \gamma p'(Z)} \right), \text{in particular } g'(0) = E^* \left(\frac{p'(Z)}{p(Z)} \right) - 1, \tag{45}$$

which shows the first part of 1. We now first show part 2(a) and (b). Note that

$$g''(\gamma) = -E^* \left(\frac{p(Z) - p'(Z)}{(1-\gamma)p(Z) + \gamma p'(Z)} \right)^2.$$

For all $\gamma \in [0,1]$ and all Z, the denominator inside the expectation must be bounded from below by $\underline{p} := \operatorname{ess\,inf}_{Z, p \in \mathcal{P}} p(Z)$ and from above by $\overline{p} := \operatorname{ess\,sup}_{Z, p \in \mathcal{P}} p(Z)$. We thus have, for all $\gamma \in [0,1]$, $1/\overline{p}^2 E^*(p' - p)^2 \leq |g''(\gamma)| \leq 1/\underline{p}^2 E^*(p' - p)^2$. Now suppose first that $E^*(p'(Z_\text{l}) - p(Z_\text{l}))^2 = 0$. Then $p'(Z_\text{l}) = p(Z_\text{l})$ almost surely, and $g'(\lambda) = g''(\lambda) = 0$ on $[0,1]$, almost surely, and part (2a) follows. If $E^*(p'(Z_\text{l}) - p(Z_\text{l}))^2 > 0$, then $p'(Z_\text{l}) \neq p(Z_\text{l})$ with positive probability, and part (2b) immediately follows. Having now established that $g''(\lambda) \leq 0$ on $[0,1]$, it follows by definition of λ° that $g'(0) > 0$ iff $\lambda^\circ > 0$. And since we assume $D^*(p\|p') \geq 0$, we have $g(1) \leq g(0)$, which implies that if $\lambda^\circ > 0$, then $g'(\lambda^\circ) = 0$ and $g'(1) \leq 0$.

4. Proof of Proposition 12 (log-Bonferroni) Let $X_j := e^{-Y_j}$ and $b_j = e^{-a_j}$. The assumption implies that, for any collection $\{K_j\}_{j \in \mathcal{J}}$ of positive real numbers, for all $j \in \mathcal{J}$,

$$P^* \left(X_j \geq b_j e^{-f_j(K_j)} \right) \geq 1 - e^{-K_j},$$

or equivalently,

$$P^* \left(X_j < b_j e^{-f_j(K_j)} \right) < e^{-K_j}.$$

Now, for fixed $K \geq 0$, define $K_j = K - \log w_j$. By the union bound, we have

$$P^* \left(\mathcal{A} \right) < \sum_{j \in \mathcal{J}} e^{-K + \log w_j} = \sum_{j \in \mathcal{J}} w_j e^{-K}.$$

where $\mathcal{A}$ is the event that for some $j \in \mathcal{J}$, $X_j < b_j e^{-f_j(K - \log w_j)}$. This implies that for $\bar{\mathcal{A}}$, the complement of $\mathcal{A}$, we have

$$P^* \left(\bar{\mathcal{A}} \right) \geq 1 - \sum_{j \in \mathcal{J}} w_j e^{-K}.$$

The result now follows by noting that the event whose probability is bounded in the statement of the proposition is just $\bar{\mathcal{A}}$.

5. Proof of Proposition 15 (Barron and Li's (1999) result) Define, for given η, p^*, p and q, the *affinity relative to* x as $A_x = \int_{y\in\mathcal{Y}} p^*(y \mid x) \cdot \left(\frac{p(y|x)}{q(y|x)}\right)^\eta$ and let

$$p^{\text{new}}(y \mid x) = \frac{1}{A(x)} p^*(y \mid x) \cdot \left(\frac{p(y \mid x)}{q(y \mid x)}\right)^\eta.$$

Next, recall that the *squared Hellinger distance*, between densities p and q on $\mathcal{Y}$, denoted by us as $H^2(q,p)$, is defined as

$$H^2(q,p) := \int_y (\sqrt{q(y)} - \sqrt{p(y)})^2 = 2\left(1 - \int_y \sqrt{q(y)p(y)}\right).$$

Also recall that the ordinary (nongeneralized) Rényi divergence of order $1/2$ is given by $d_{1/2}(q,p) = -2\log\int_y \sqrt{q(y)p(y)}dy$. Now, for $u \geq 0$, we have $1 - u \leq -\log u$ (this follows from $\log(1+z) \geq z$ and substituting $z = u - 1$). This implies the following well-known general relation between squared Hellinger distance and Rényi divergence:

$$H^2(q,p) \leq d_{1/2}(q\|p). \tag{46}$$

Moreover (Barron and Cover, 1991), when the ratio between p and q is bounded, then the standard (nongeneralized) KL divergence is upper-bounded by a multiple of the squared Hellinger distance. Yang and Barron (1999) proved the following precise relation:

$$D(q\|p) \leq (2 + \log V) H^2(q,p). \tag{47}$$

We will now use (46) and (47) to prove our result. We first need to clarify notation: for given x, the *generalized Rényi divergence between p and q, given x* is denoted as $d_\eta^{*|x}(q(\cdot \mid x)\|p(\cdot \mid x))$ and defined as

$$d_\eta^{*|x}(q(\cdot \mid x)\|p(\cdot \mid x)) = -\frac{1}{\eta}\log E^*\left[\left(\frac{p(Y \mid x)}{q(Y \mid x)}\right)^\eta \mid X = x\right].$$

We have for all $C_\eta \geq 2 + 2\eta\log V$, for each $x \in \mathcal{X}$, each $\eta \leq 1$,

$$\begin{aligned}
E^*\left[-\log\tfrac{p(Y|x)}{q(Y|x)} \mid X = x\right] &= \tfrac{1}{\eta}\cdot E^*\left[\log p^*(Y \mid x) - \log\left(p^*(Y \mid x)\left(\tfrac{q(Y|x)}{p(Y|x)}\right)^\eta\right) \mid X = x\right] \\
&\quad + \tfrac{1}{\eta}(\log A_x - \log A_x) \\
&= \tfrac{1}{\eta} D(p^*(\cdot|x)\|p^{\text{new}}(\cdot|x)) - \tfrac{1}{\eta}\log A_x \\
&\leq \tfrac{1}{\eta} C_\eta H^2(p^*(\cdot|x), p^{\text{new}}(\cdot|x)) - \tfrac{1}{\eta}\log A_x \\
&\leq \tfrac{1}{\eta} C_\eta d_{1/2}(p^*(\cdot|x), p^{\text{new}}(\cdot|x)) - \tfrac{1}{\eta}\log A_x \\
&= C_\eta\left(d_{\eta/2}^{*|x}(q(\cdot|x)\|p(\cdot|x)) + \tfrac{1}{\eta}\log A_x\right) - \tfrac{1}{\eta}\log A_x \\
&= C_\eta d_{\eta/2}^{*|x}(q(\cdot|x)\|p(\cdot|x)) + \tfrac{1}{\eta}(C_\eta - 1)\log A_x.
\end{aligned}$$

Here the first two equalities are just rewriting. In the first inequality we used (47), the fact that P^*-almost surely, $\sup_{X,Y} p^{\text{new}}(Y|X)/p^*(Y|X) \leq V^{2\eta}$, and the fact that $D(\cdot\|\cdot) \geq 0$, and the second inequality is just (46). In the fifth line we used some basic rewriting. Using

the notation E^*_X to denote expectation of X under P^*_X, the marginal distribution of X, we thus get:

$$\begin{aligned} D^*(q\|p) &\leq C_\eta E^*_X[d^{*|X}_{\eta/2}(q(\cdot \mid X)\|p(\cdot \mid X))] + (C_\eta - 1)\frac{1}{\eta}E^*_X[\log A_X] \\ &\leq C_\eta d^*_{2/\eta}(q\|p) + (C_\eta - 1)\frac{1}{\eta}\log E^*_X[A_X] \\ &= C_\eta d^*_{\eta/2}(q\|p) - (C_\eta - 1)d^*_\eta(q\|p). \end{aligned}$$

where the second inequality is Jensen's and the final equality is just the definition of Rényi divergence.