Viewing All Models as 'Probabilistic'

Peter Grünwald[*]
Computer Science Department
Stanford University
Stanford CA 94305
grunwald@cs.stanford.edu

Abstract

In order to apply the Minimum Description Length Principle, one must associate each model in the model class under consideration with a corresponding *code*. For probabilistic model classes, there is a principled and generally agreed-upon method for doing this; for non-probabilistic model classes (i.e. classes of functions together with associated error functions) it is not so clear how to do this. Here, we present a new method for associating codes with models that works for probabilistic and non-probabilistic model classes alike. Our method can be re-interpreted as mapping arbitrary model classes to associated classes of probability distributions. The method can therefore also be applied in a Bayesian context. In contrast to earlier proposals by Barron, Yamanishi and Rissanen and to the ad-hoc solutions found in applications of MDL, our method involves *learning* the optimal scaling factor in the mapping from models to codes/probability distributions from the data at hand. We show that this method satisfies several optimality properties. We present several theorems that suggest that with the help of our mapping of models to codes, one can successfully learn using MDL and/or Bayesian methods when (1) almost arbitrary model classes and error functions are allowed, and (2) none of the models in the class under consideration are close to the 'truth' that generates the data.

1 INTRODUCTION

The Minimum Description Length (MDL) Principle [16, 11, 12, 1] is a powerful method for inductive inference. In order to apply it, one must associate each model in the model class under consideration with a corresponding *code*. For probabilistic model classes, there is a principled and generally agreed-upon method for doing this. For general, non-probabilistic model classes, i.e. classes of functions together with associated error functions, it is not so clear how to associate codes with models.

In this paper we present a general method, called 'entropification', for associating models with codes.

Because of the existence of a 1-1 correspondence between probability distributions and code length functions, 'entropification' can also be interpreted as a method for associating arbitrary models with corresponding probability distributions. For this reason, our results can also be interpreted from a Bayesian point of view. Moreover, the correspondence allows us to asses the performance of learning methods based on non-probabilistic model classes in probabilistic terms. Our theorems suggest that 'entropification' provides a well-behaved method for applying MDL and/or Bayesian methods to learning from data when (1) almost arbitrary model classes and error functions may be used, and (2) none of the models in the class under consideration are close to the 'truth' that generates the data.

The main idea behind entropification is to associate each model H and its associated error function ER with a code C such that, for all sequences of observations $(x, y)^n$, the code length of the encoding $C((x, y)^n)$ is a linear function of the error H makes on $(x, y)^n$. The main difference to earlier proposals by Rissanen, Barron and Yamanishi [12, 2, 18] is that the coefficients of this linear function are *learned* from the available data. Roughly speaking, our theorems state that, when enough data are available and under certain additional conditions, 'entropification' combined with MDL or Bayesian inference leads to the adoption of models that are *optimal* and *reliable*. Here 'optimal' means that, among all models in the class under consideration, they allow for the best predictions or decisions regarding future data, when the same error function ER is used as during the learning. 'reliable' means that, in a sense to be described below, the inferred models *give a faithful impression of their own performance over future data*, even if they are not at all close to the 'truth' that is supposed to generate data.

Below we first introduce notation. We then discuss our problem in more detail. In Section 4 we introduce 'entropification'. Section 5 discusses the framework of our theorems, which are presented in the remaining sections. In Section 8 we also summarize some further results we obtained. All

*The author is supported by a TALENT-grant awarded by the Netherlands Organization for Scientific Research (NWO).

proofs can be found in an appendix. Preliminary versions of some of the results reported here, along with much more detailed proofs as well as much more extensive discussion, can be found in the author's Ph.D. thesis [6].

2 PRELIMINARIES

We assume a setting with a sample space $\Omega = \Omega_x \times \Omega_y$. Here Ω_x can be written as $\Omega_x = \Omega_{x,1} \times \ldots \times \Omega_{x,m_x}$ for some $m_x \geq 0$. For simplicity, we will assume that each component $\Omega_{x,i}$ and Ω_y is either a finite subset of the natural numbers $\mathbf{N}$ or a compact subset of the reals $\mathbf{R}$ or identical to $\mathbf{R}$. In the first case we call $\Omega_{x,i}$ (Ω_y) *discrete*, in the later two cases we call it *real-valued*. We are given a set of data $(x,y)^n = (x_1, y_1), \ldots, (x_n, y_n)$ consisting of n observations. Here each $x_i \in \Omega_x$ and $y_i \in \Omega_y$. We are then presented the $(n+1)$-st partial observation x_{n+1} and we have to make a decision, prediction or classification considering the $(n+1)$-st y-outcome. We assume that there is some *decision space* Δ. The *error* or *loss* we incur at the $(n+1)$-st decision/prediction/classification (henceforth, all three possibilities are denoted simply by the word 'prediction') is given by an *error function*. This is a function $\text{ER}(\cdot;\cdot) : (\Omega_y; \Delta) \rightarrow \mathbf{R} \cup \{\infty\}$. ER is extended to several outcomes by $\text{ER}((y_1, \delta_1), \ldots, (y_n, \delta_n)) := \sum_{i=1}^{n} \text{ER}(y_i, \delta_i)$. A *decision rule* or *non-probabilistic model* is a function $H : \Omega_x \rightarrow \Delta$. The *error that non-probabilistic model H makes on data* $(x,y)^n$ is given by $\text{ER}((y_1; H(x_1)), \ldots, (y_n; H(x_n)))$, which we abbreviate to $\text{ER}((x,y)^n | H)$:

$$\begin{aligned} \text{ER}((x,y)^n|H) &:= \text{ER}((y_1; H(x_1)), \ldots, (y_n; H(x_n))) \\ &= \sum_{i=1}^{n} \text{ER}(y_i; H(x_i)). \end{aligned} \tag{1}$$

A *probabilistic model* is a probability distribution over Ω^∞. In this paper we restrict ourselves to distributions that render the data identically independently distributed (i.i.d.). We can therefore view them as probability distributions over Ω that, for each n, are extended to Ω^n by taking the product measure. If a component $\Omega_{x,i}$ (or Ω_y) of Ω is discrete, then the notation $P(x)$ for $x \in \Omega_{x,i}$ stands for the probability mass of x under distribution P. If the component $\Omega_{x,i}$ is real-valued, then the notation $P(x)$ for $x \in \Omega_{x,i}$ stands for the density of x under distribution P (we will only make use of this notation in the context of probability distributions for which such densities can be assumed to exist). We often consider classes of probabilistic models parameterized by some set $\Gamma \subset \mathbf{R}^k$. For each $\theta \in \Gamma$, $P(\cdot|\theta)$ indicates the probabilistic model associated with θ. We interchangeably write E_θ and $E_{P(\cdot|\theta)}$ to denote expectation under $P(\cdot|\theta)$. We write '$\theta \in \mathcal{M}$' as shorthand for $P(\cdot|\theta) \in \mathcal{M}$. If we integrate over a space Ω, the integration should be read as the corresponding sum if Ω is discrete.

For given sample space $\Omega = \Omega_x \times \Omega_y$, consider the decision space Δ consisting of all probability distributions over Ω_y. In the context of such a decision space, a conditional probabilistic model $P(\cdot|\cdot)$ (i.e. for each $x \in \Omega_x$, $P(\cdot|x)$ is a probabilistic model over Ω_y) can be viewed as a decision rule. The *error that probabilistic model H makes on data* $(x,y)^n$ is then given by $\text{ER}((y_1; P(\cdot|x_1)), \ldots, (y_n; P(\cdot|x_n)))$, which we abbreviate to $\text{ER}((x,y)^n|P)$:

$$\begin{aligned} \text{ER}((x,y)^n|P) &:= \text{ER}((y_1; P(\cdot|x_1)), \ldots, (y_n; P(\cdot|x_n))) \\ &= \sum_{i=1}^{n} \text{ER}(y_i; P(\cdot|x_i)). \end{aligned} \tag{2}$$

3 THE PROBLEM

In the present section we introduce the problem this paper is concerned with: how to associate non-probabilistic models with codes or equivalently, probability distributions.

Code lengths and Probabilities Let $L : \Omega \rightarrow \mathbf{R}$ be a code length function of a prefix code [4, 12] over Ω, where the code length is measured in nats. That is, there exists a prefix code $C : \Omega \rightarrow \{0,1\}^*$ such that, for each $x \in \Omega$, the length of $C(x)$ (the code word for x), when converted from bits to nats, is given by $L(x)$. Following Rissanen and Yamanishi [11, 18] we will use idealized prefix codes that can have non-integer code lengths. Let $\mathcal{L}$ be the set of all idealized code length functions for prefix codes over Ω. By the *Kraft Inequality* [12, 4], the function $L : \Omega \rightarrow \mathbf{R}$ is in $\mathcal{L}$ iff

$$\sum_{x \in \Omega} e^{-L(x)} \leq 1 \tag{3}$$

The Kraft Inequality shows that there exists a 1-1 correspondence between (subadditive) probability distributions and code length functions: for each function $P : \Omega \rightarrow \mathbf{R}$ such that $\sum_{x \in \Omega} P(x) \leq 1$, there exists a code length function $L_P : \Omega \rightarrow \mathbf{R}$ such that for all $x \in \Omega$,

$$L_P(x) = -\ln P(x) \tag{4}$$

and vice versa. For any given P, we call the code with lengths L_P achieving these lengths the *Shannon-Fano* code for P.

Two-Part Code MDL The original formulation of the MDL Principle [15, 10], also called '2-part code MDL' roughly works as follows: for data x^n, one chooses the model H in a class of models $\mathcal{M}$ that minimizes $L(H) + L(x^n|H)$. Here the first term stands for the description length of H when encoded using some fixed code C, and the second denotes the description length of data x^n *when it is encoded using the code corresponding to* H. For the second term to be well-defined, we must associate each $H \in \mathcal{M}$ with a code. In case $\mathcal{M}$ is a class of probabilistic models one usually picks, for each $P \in \mathcal{M}$, the code L_P with lengths (4). This code satisfies two properties which are necessary to make MDL 'work':

Fundamental Properties of Shannon-Fano Code:

1. The better $P \in \mathcal{M}$ fits the data (i.e. the higher $P(x)$), the shorter the code length $L_P(x)$.

2. If P is 'true', then among all possible codes L over Ω, the one with lengths (4) leads to the shortest expected code length $E_P[L(X)]$; i.e. $L_P = \arg\min E_P[L(X)]$ where the minimum is over all code length functions $L : \Omega \rightarrow \mathbf{R}$ (these functions L are just the functions satisfying the Kraft Inequality (3)).

This second property is a direct consequence of the Information Inequality [4]. It is the property that makes MDL 'self-consistent', see the remark after Theorem 1, Section 7.

Non-Probabilistic Models How should we associate codes with non-probabilistic models? We will assume that the quality of models is judged against a fixed error function ER (but see Section 8). We would like to have properties similar to 1. and 2. above in this case. In order to retain property 1. ("good fit implies short code length"), we could devise a code such that, for each $H \in \mathcal{M}$, for all $(x,y)^n$,

$$L((x,y)^n|H) = \beta \text{ER}((x,y)^n|H) + C. \quad (5)$$

where C depends on $(x,y)^n$ only through n; it does not depend on the specific values of x_i's and y_i's. As will be seen below, codes with such lengths exist - for most error functions, we can choose any $\beta > 0$ we like; the constant C, needed to make $L((x,y)^n|H)$ interpretable as a code length, is determined by n and the actual choice of β. Variations of codes (5) have been proposed by Rissanen [12], Barron [2] and Yamanishi [18]. In Rissanen's proposal β was given the ad-hoc value 1; in Barron's and Yamanishi's proposal β was determined by the *number* of available observations in a way that guaranteed certain rates of convergence. In practical applications of MDL the issue arises too; see for example [5] where an ad-hoc value $\beta = 2$ was chosen. Here we propose a new method: we learn β from the data, using the MDL Principle itself to obtain a good estimate for it. This allows us to prove (Section 7, Theorem 1) a version of Fundamental Property 2 above that does not refer to a 'true' distribution. As will be seen, this theorem provides a strong justification, both of (1) using codes of the form (5) (and not say, codes with lengths $L(x,y)$ that are some non-linear function of $\text{ER}(x,y)$) and of (2) learning β from the data at hand (rather than by some other method).

'Probabilistic' Interpretation By the Kraft Inequality (3), associating non-probabilistic models with codes as in (5) can also be interpreted as associating non-probabilistic models with *probability distributions*[1] satisfying $-\ln P((x,y)^n|H) = \beta\text{ER}((x,y)^n|H) + C$. However, the resulting functions will only be probability distributions in the purely mathematical sense that $\int P(x,y)dxdy = 1$. It is not clear a priori whether expectations taken over these distributions make sense; in the next sections we will show that, when we allow β to be determined from data, then expectations $E[\phi(X)]$ over such distributions do make sense for a limited class of functions ϕ. This 'probabilistic' interpretation makes our proposal also applicable to Bayesian Inference and, in some cases, maximum likelihood inference. In the next section, we shall formally define a mapping that takes as input an arbitrary model class $\mathcal{M}$ that can be either probabilistic[2] or non-probabilistic, together with an associated error function ER. The output of the mapping is an associated class of probabilistic model classes (or equivalently, class of codes) $\langle\mathcal{M}\rangle_{\text{ER}}$. We call the mapping 'entropification'.

[1] By a proper choice of the constant C in (5), the resulting distributions are 'true' probability distributions and not subadditive.

[2] The advantages of setting up the mapping so that it is also applicable to model classes $\mathcal{M}$ that are already probabilistic are discussed in the author's thesis.

4 ENTROPIFICATION

Definition 1 (Entropification) *Let Ω be a sample space, $\mathcal{M}$ be a class of models and* ER $: \Omega \times \mathcal{M} \rightarrow \mathbf{R} \cup \{\infty\}$ *be an error function as in Equation 1 (if $\mathcal{M}$ is non-probabilistic) or Equation 2 (if $\mathcal{M}$ is probabilistic). The entropification of model class $\mathcal{M}$ with respect to error function* ER *is the class of probabilistic models*

$$\langle\mathcal{M}\rangle_{\text{ER}} := \{P(\cdot|H,\beta) \mid H \in \mathcal{M}; \beta \in \Gamma_{nat}\} \quad (6)$$

with $P(\cdot|H,\beta)$ given by

$$P(x,y|H,\beta) = \frac{1}{Z_H(\beta)}\exp(-\beta\text{ER}(x,y|H)) \quad , \quad (7)$$

for each $(x,y) \in \Omega$, and

$$P((x,y)^n|H,\beta) = \prod_{i=1}^{n} P(x_i,y_i|H,\beta).$$

for all $(x,y)^n \in \Omega^n$. The normalizing factor $Z_H(\beta)$ is given by

$$Z_H(\beta) := \int_{x,y\in\Omega} \exp(-\beta\text{ER}(x,y|H))dxdy \quad (8)$$

The set of possible scaling factors Γ_{nat} ('nat' standing for 'natural parameter space') is given[3] *by*

$$\Gamma_{nat} = \{\beta \in \mathbf{R}\cup\{-\infty,\infty\}\} \mid \forall H \in \mathcal{M}: \ Z_H(\beta) < \infty\}.$$

We will sometimes be interested in the subclass of $\langle\mathcal{M}\rangle_{\text{ER}}$ involving only models $P(\cdot|H,\beta)$ in which β is constrained to some subset Γ of Γ_{nat}. We define *the entropification of $\mathcal{M}$ with respect to* ER *and* $\Gamma \subseteq \Gamma_{nat}$, denoted by $\langle\mathcal{M}\rangle_{\text{ER}}^{\Gamma}$, to be the class of models

$$\langle\mathcal{M}\rangle_{\text{ER}}^{\Gamma} = \{P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle \mid \beta \in \Gamma\}$$

Usually, we take $\Gamma = \{\beta \in \Gamma_{nat} | \beta \geq 0\}$. In that case, we write $\langle\mathcal{M}\rangle_{\text{ER}}^{\beta\geq 0}$ for $\langle\mathcal{M}\rangle_{\text{ER}}^{\Gamma}$.

Code lengths based on models in $\langle\mathcal{M}\rangle$; three interpretations of β The code corresponding to $P(\cdot|H,\beta)$ has the following code lengths:

$$\begin{aligned} L((x,y)^n|H,\beta) &= -\ln P((x,y)^n|H,\beta) \\ &= \beta\sum_{i=1}^{n}\text{ER}((x,y)^n|H) + n\ln Z_H(\beta) \end{aligned} \quad (9)$$

These code lengths consist of an error term and a 'uniform' term $n \ln Z_H(\beta)$ that grows linearly in n and is equal for all $(x,y)^n$. This shows that β can be interpreted as determining how strongly the error should be weighted in the code length corresponding to hypothesis (H,β). The extreme case $\beta = 0$ corresponds to the uniform distribution over all $(x,y)^n \in \Omega^n$. Two more interpretations of β follow from the fact that 'entropified model classes' are essentially *exponential families* or equivalently, 'maximum entropy model classes' [7, 6]: (1) The entropy of $P(\cdot|H,\beta)$ is determined

[3] The distribution $P(\cdot|H,\beta)$ for $\beta = \infty$ is defined by $P(x,y|H,\infty) = \lim_{\beta\rightarrow\infty} P(x,y|H,\beta)$.

by β (more precisely, it is a monotonically decreasing function of $|\beta|$; see Proposition 5 in the appendix) – hence the name 'entropification'. (2) β, when inferred from data, can be seen as a monotonic transform of an estimate of the error H makes on future data; we will focus on this property in the examples below.

Example 1 [Squared Error] Let Ω_x be a closed interval in $\mathbf{R}$, let $\Omega_y = \Delta = \mathbf{R}$ and let $\mathcal{M}_k$ be the class of k-th degree polynomials $H : \Omega_x \to \mathbf{R}$. Consider the expected squared error $\mathrm{ER}_{sq}((x,y)^n|H) = \sum_{i=1}^{n}(y_i - H(x_i))^2$. The models $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\mathrm{ER}_{sq}}$ can be written as $P((x,y)^n|H,\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{(y_i-H(x_i))^2}{2\sigma^2})$ with $\beta = 1/2\sigma^2$. Hence, according to $P(\cdot|H,\beta)$ the data are independently normally distributed; the scaling parameter β is a monotonic transform of σ^2, which in turn is equal to $E_{P(\cdot|H,\beta)}[\mathrm{ER}_{sq}(X,Y|H)]$, the squared error of H (this is straightforward to verify). Hence β is a monotonic transform of expected error.

Associating the squared error with the normal distribution has been common practice since Gauss [12]; 'entropification' can be seen as a generalization of this. However, there is a crucial difference to previous treatments. Usually, the distributions are constructed as *conditional* distributions, i.e. one writes $P(y|H,\beta,x) \propto \exp(\frac{(y-H(x))^2}{2\sigma^2})$ and the distribution $P(x|H,\beta)$ is left unspecified, as it is irrelevant to one's prediction task: one wants to make predictions concerning y-values for *given* values $x \in \Omega_x$.

In contrast, we define $P(\cdot|H,\beta)$ as a distribution over the full sample space. This turns out to be unproblematic[4] *as long as we do not use the models $P(\cdot|H,\beta)$ to infer anything about the distribution of the x-values.* Here, we will only use models $P(\cdot|H,\beta)$ to make predictions of Y given X, so this condition is automatically satisfied. In fact, one can develop a theory of 'safe statistics', identifying exactly what distributions that have been inferred as formal re-representations of non-probabilistic models can be used for and what not; see [6].

Example 2 [Classification] Let $\Omega_y = \Delta = \{0,1\}$ and let $\mathcal{M}$ be a class of *classifiers* $H : \Omega_x \to \{0,1\}$. Consider the 0/1-loss $\mathrm{ER}_{01}((x,y)^n|H) = \sum_{i=1}^{n}|y_i - H(x_i)|$. Basic calculus shows that the models $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\mathrm{ER}_{01}}$ can be written as

$$P((x,y)^n|H,\beta) = \prod_{i=1}^{n} P(y_i \neq H(x_i)|H,\beta)^{n_0} P(y_i = H(x_i)|H,\beta)^{(n-n_0)}.$$

Here $n_0 = \mathrm{ER}_{01}((x,y)^n|H)$ is the accumulated error H makes on $(x,y)^n$ and

$$P(y_i \neq H(x_i)|H,\beta) = E_{P(\cdot|H,\beta)}[\mathrm{ER}_{01}(X,Y|H)] = \eta$$

with $\beta = \ln(1-\eta) - \ln\eta$. Once again, β is a monotonic transform of the expected error of H. Note the possibility that $\beta < 0$; this is discussed further in Example 4.

[4]The reason we use the full sample space is that it makes notation and proofs much simpler. However, though we have no formal proof of this in all cases, it seems that all our theorems would still hold if we would redefine $\langle\mathcal{M}\rangle_{\mathrm{ER}}$ as a class of conditional distributions. The conditional road was taken in the author's thesis [6], Chapter 5.

In the classification setting there is another natural code of associating a classifier H with a code with lengths $L(\cdot|H)$: for given H and data $(x,y)^n$, one codes the indices i of all the y_i for which $H(x_i) \neq y_i$. This code has been used in, for example, [9, 8]. Interestingly, as shown in the author's thesis [6, Chapter 5, Example 5.21], the code lengths obtained in this way coincide with the code lengths obtained by 'entropification'.

5 ON OPTIMALITY AND RELIABILITY

Our theorems will be about inferring models from data that are *optimal* and *reliable*.

Definition 2 *The* ER*-optimal model $\tilde{H}$ (relative to true distribution P^* and model class $\mathcal{M}$), if it exists, is the model*

$$\tilde{H} := \arg\min_{H \in \mathcal{M}} \{E_{P^*}[\mathrm{ER}(X,Y|H)]\} \qquad (10)$$

The definition of 'reliable' goes via a definition of probability distributions that partially agree with some other distribution P^* that is supposed to generate the data.

Definition 3 *Let ϕ be a function $\phi : \Omega \to \mathbf{R}$ and P and P^* be probability distributions over Ω. We say that P is ϕ-true relative to true distribution P^* if $E_P[\phi] = E_{P^*}[\phi]$.*

Definition 4 *Let $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\mathrm{ER}}$ where $\langle\mathcal{M}\rangle_{\mathrm{ER}}$ is defined as in Definition 1. $P(\cdot|H,\beta)$ is* reliable *relative to distribution P^* if it is $\mathrm{ER}(\cdot|H)$-true relative to P^*, i.e. if $E_{P(\cdot|H,\beta)}[\mathrm{ER}(X,Y|H)] = E_{P^*}[\mathrm{ER}(X,Y|H)]$.*

The idea behind optimality and reliability is the following: assume data are distributed i.i.d. according to some unknown distribution P^*. If, given a model class $\mathcal{M}$, we somehow manage to obtain a model $(\tilde{H},\tilde{\beta})$ such that $\tilde{H}$ is ER-optimal and $(\tilde{H},\tilde{\beta})$ is reliable with respect to the unknown P^*, then (1) we have the 'best' model in $\mathcal{M}$ that leads, in the long run, to the smallest average prediction error among all models in $\mathcal{M}$ (this follows from the law of large numbers); and (2) we have a correct impression of how good $\tilde{H}$ really is - $P(\cdot|\tilde{H},\tilde{\beta})$ gives a correct impression of the average error $\tilde{H}$ will make over future data; we can use the probability distribution $P(\cdot|\tilde{H},\tilde{\beta})$ *as if* it were true as long as we only use it in the context of error function ER. In this way, it is not such a disaster if we use a model class $\mathcal{M}$ that does not contain any really 'good' model H for the data at hand - if we can make sure that the inferred model $(\tilde{H},\tilde{\beta})$ is reliable, then this model *tells* us that it itself is not very good - we do not know much about the data generating machinery, but we know that we do not know much.

Let $\langle\mathcal{M}\rangle_{\mathrm{ER}}$ be the 'entropification' of $\mathcal{M}$. Whenever ER is clear from the context, we simply call an inference method *optimal* if, as more data becomes available, with high P^*-probability, the H-component of the model (H,β) we infer from data converges to the ER-optimal model $\tilde{H}$ (as given by (10)) within $\mathcal{M}$. We call it *reliable* if, with high P^*-probability, the model (H,β) we infer from data converges to a reliable model within $\langle\mathcal{M}\rangle_{\mathrm{ER}}$ (the exact mode of convergence to be discussed later). In Section 7 we state our theorems, which show that statistical inference methods used with entropified model classes are under certain conditions indeed optimal and reliable. Below we first state the regularity conditions for these theorems.

6 REGULARITY CONDITIONS

In all our theorems we will assume the following regularity conditions **C1** $-$ **C4** to hold:

C1 Ω_x can be written as $\Omega_x = \Omega_{x,1} \times \ldots \times \Omega_{x,m_x}$ for some $m_x \geq 0$. Each component $\Omega_{x,i}$ and Ω_y is either a finite subset of the natural numbers $\mathbf{N}$ or a compact subset of the reals $\mathbf{R}$ or identical to $\mathbf{R}$.

We will assume that data are i.i.d. according to an almost arbitrary distribution P^*. Nevertheless, in order for our theorems to go through, we need a form of *uniform convergence* (see Lemma 7 in the appendix). This can be obtained by assuming that P^* is zero outside of a compact subset of the sample space . More precisely, let $\Omega_x = \Omega_{x,1} \times \ldots \times \Omega_{x,m_x}$ for some $m_x \geq 0$ and consider P_2^*, the extension of P^* from $\Omega = \Omega_x \times \Omega_y$ to sample space $\mathbf{R}^{m_x+1}$. That is, P_2^* is a probability distribution over $\mathbf{R}^{m_x+1}$ satisfying $P_2^*(S) = P^*(S)$ for every set S in the intersection of Ω and the Borel σ-algebra on $\mathbf{R}^{m_x+1}$. We say that P^* has compact support iff there exists a compact subset C of $\mathbf{R}^{m_x+1}$ such that $P_2^*(C) = 1$.

C2 P^* has compact support.

The following condition ensures that $\langle \mathcal{M} \rangle_{\text{ER}}$ is well-defined:

C3 ER is a function $\text{ER} : \Omega \times \mathcal{M} \to \mathbf{R} \cup \{\infty\}$. $\mathcal{M}$ and ER (and hence $\langle \mathcal{M} \rangle_{\text{ER}}$) are such that there exists at least one $\beta > 0$ such that $E_{(H,\beta)}[\text{ER}(X, Y|H, \beta)] < \infty$ for all $H \in \mathcal{M}$.

Note that, for **C3** to be satisfied, it is sufficient if the following holds: Ω_x is compact and (1) Ω_y is finite or (2) Ω_y is compact, or (3) for all δ there exists a polynomial f_δ such that $\forall y : \text{ER}(y, \delta) \geq f_\delta(y)$. One of these three conditions is satisfied for most of the error functions found in the literature.

We say that a (general) model class $\mathcal{M}$ is parameterized by a set $\Gamma_{\mathcal{M}}$, where $\Gamma \subseteq \mathbf{R}^k$, if there exists a bijection $g : \Gamma_{\mathcal{M}} \to \mathcal{M}$. If $\mathcal{M}$ is probabilistic, then we say that $\Gamma_{\mathcal{M}}$ is a *continuous, twice differentiable parameterization of* $\mathcal{M}$ if both of the following hold: (1) if Ω is discrete, then for each $\omega \in \Omega$, $P(\omega|\theta)$ is a continuous, twice differentiable function of θ; (2) if Ω is real-valued, then for each $\omega \in \Omega$, $P(\omega|\theta)$ is a continuous, twice differentiable function of both ω and θ.

Similarly, if $\Omega = \Omega_{x,1} \times \ldots \times \Omega_{x,m} \times \Omega_y$ where some of the components are discrete and others are real-valued, $P(\omega|\theta)$ should, for each possible combination of values for the discrete subspaces, be a continuous function in θ and all the variables in the real subspaces. These definitions are needed in our last regularity condition:

C4 $\mathcal{M}$ can be parameterized by a compact set $\Gamma_{\mathcal{M}}$. $\langle \mathcal{M} \rangle_{\text{ER}}$ can be continuously, twice differentiably parameterized by the set $\{(\theta, \beta) | \theta \in \Gamma_{\mathcal{M}}, \beta \in \Gamma_{nat}\}$.

In some of our theorems, we will restrict ourselves to *simple* error functions. It is straightforward to verify that many of the usual error functions, and all those discussed here in Examples 1,2,3,4 are 'simple' with respect to all possible model classes. The logarithmic loss for probabilistic model classes however is *not* simple.

Definition 5 *If* ER *is such that for all* $H_1, H_2 \in \mathcal{M}$ *and all* β, $Z_{H_1}(\beta) = Z_{H_2}(\beta)$ *where* $Z_H(\beta)$ *is defined as in Equation 8, then we call* ER *a* simple *error function for* $\mathcal{M}$.

In some of our theorems, we assume that the error functions used are *symmetric*:

Definition 6 *We call* $\text{ER} : \Omega \times \mathcal{M} \to \mathbf{R}$ *a symmetric error function for* $\mathcal{M}$ *if the following holds: for all* $H \in \mathcal{M}$ *there exists an 'inverse' function* $\overline{H} : \Omega_x \to \Delta$ *satisfying* $\forall (x, y) \in \Omega : \text{ER}(x, y|H) = -\text{ER}(x, y|\overline{H})$.

Not many error functions are symmetric, but some interesting error functions can be *made* symmetric by defining $\text{ER}'(x, y|H) = \text{ER}(x, y|H) - C$ for some constant C. Clearly, obtaining the H that minimizes expected or empirical $\text{ER}(x, y|H)$ (which is what we are interested in) is equivalent to obtaining the H that minimizes expected or empirical $\text{ER}'(x, y|H)$. Note that the 0/1-error and the absolute error (defined by $\Omega_y = \Delta = [0, 1]$; $\text{ER}(y; \delta) = |y - \delta|$) can be made symmetric. The other error functions mentioned here cannot.

7 THEOREMS ON ENTROPIFICATION

We now present our main results. First, in Theorem 1, we show an analogue of the 'fundamental properties' of Section 3 that holds for entropified model classes. This theorem will be the cornerstone in the proofs of Theorems 2–4, which show that Maximum Likelihood (and hence MDL and Bayesian methods) in combination with entropified model classes with a bounded number of parameters leads to the inference of asymptotically optimal and reliable models.

Theorem 1 *Let sample space* Ω*, model class* $\mathcal{M}$*, error function* ER *and* P^* *satisfy conditions* **C1**-**C4**. *Assume* ER *is 'simple' (Definition 5). Then*

1. *The following minima exist:*
$$\widetilde{\text{ER}}(P^*) := \min_{H \in \mathcal{M}} E_{P^*}[\text{ER}(X, Y|H)] \qquad (11)$$
$$\widetilde{L}(P^*) := \min_{P(\cdot|H,\beta) \in \langle \mathcal{M} \rangle_{\text{ER}}^{\beta \geq 0}} E_{P^*}[-\ln P(X, Y|H, \beta)] \qquad (12)$$
2. *There exists a unique value* $\tilde{\beta}$ *such that for all models* (H, β) *minimizing (12),* $\beta = \tilde{\beta}$.
3. *If* $\tilde{\beta} > 0$*, then a model* $(H, \tilde{\beta})$ *minimizes (12) iff* H *minimizes (11)*
*(***optimal model minimizes expected code length***).*
4. *If* $\tilde{\beta} > 0$*, then for all models* $(H, \tilde{\beta})$ *minimizing (12), we have*
$$E_{P^*}[\text{ER}(X, Y|H)] = E_{(H,\tilde{\beta})}[\text{ER}(X, Y|H)].$$
(reliable model minimizes expected code length).

To get an idea of what this theorem means, assume that there exists a single $\tilde{H}$ minimizing (11). In this case, one of the codes corresponding to this $\tilde{H}$ (namely, the one with $\beta = \tilde{\beta}$) minimizes the expected code length among all code length functions within the class (now viewed as a class of codes) $\langle \mathcal{M} \rangle_{\text{ER}}$. For this special $(\tilde{H}, \tilde{\beta})$, both *optimality* and *reliability* hold, at least if $\tilde{\beta} > 0$ (the possibility $\tilde{\beta} = 0$ will be dealt with in Theorem 3). Compare this with our second 'fundamental property of Shannon-Fano codes' (Section 3). There, we associated each model P with a code such that if the model P is true, its associated code leads to the shortest expected code lengths. Such a choice is needed to make MDL-based inference *consistent*. Here, we associate each model H with a class of 'codes' (with lengths $L(\cdot|H, \beta) = -\ln P(\cdot|H, \beta)$; one code for each value of β) such that if H is optimal, (a) one of the codes in this class leads to the shortest expected code lengths; (b) this particular code, when viewed as a probability distribution, is also reliable *no matter whether it is equal to the true P^* or not*. Such a choice is needed to make MDL-related inference *optimal and reliable* (rather than consistent).

Convergence Theorem The next theorem shows that, for model classes that are the result of 'entropification' of a class with a fixed number of parameters, maximum likelihood estimation converges, with probability one, to an optimal and reliable model. Theorem 4 will show that the error goes down fast as the number of examples increases.

The maximum likelihood estimator in $\langle \mathcal{M} \rangle_{\text{ER}}$ for data $(x, y)^n$, if it exists, is denoted by $(\hat{H}_n, \hat{\beta}_n)$. We will allow for more than one model maximizing the likelihood at the same time. Let us take a fixed parameterization of $\langle \mathcal{M} \rangle_{\text{ER}}$ and order the models in $\langle \mathcal{M} \rangle_{\text{ER}}$ lexicographically according to this parameterization. In case there is more than one pair $(H, \beta) \in \langle \mathcal{M} \rangle_{\text{ER}}$ maximizing the likelihood of $(x, y)^n$, the notation $(\hat{H}_n, \hat{\beta}_n)$ refers to the smallest (in the lexicographical sense) (H, β) within the set of all (H, β) maximizing the likelihood (our regularity condition **C4** implies that if an ML model exists at all, then there must also be a smallest one).

Theorem 2 *Let sample space Ω, model class $\mathcal{M}$, error function* ER *and P^* be as required in* **C1** – **C4**. *Assume* ER *is a 'simple' error function. Assume data are i.i.d. according to P^*. Assume an* ER*-optimal hypothesis $\tilde{H}$, as given by (10), exists. Then there exists a unique $\tilde{\beta}$ depending on $\tilde{H}$ such that either $\tilde{\beta} = 0$ or $\tilde{\beta} > 0$. If $\tilde{\beta} > 0$, then all of the following hold: (1) $P(\cdot|H, \tilde{\beta})$ is reliable relative to P^*:*

$$E_{(\tilde{H}, \tilde{\beta})}[\text{ER}(X, Y|\tilde{H})] = E_{P^*}[\text{ER}(X, Y|\tilde{H})]. \quad (13)$$

Furthermore, (2) for all n, $(\hat{H}_n, \hat{\beta}_n)$ exists, and (3) with P^*-probability 1:

$$\lim_{n \to \infty} E_{P^*}[\text{ER}(X, Y|\hat{H}_n)] = E_{P^*}[\text{ER}(X, Y|\tilde{H})] \quad \textbf{(optimality)} \quad (14)$$

and also (4)

$$\lim_{n \to \infty} E_{(\hat{H}_n, \hat{\beta}_n)}[\text{ER}(X, Y|\hat{H}_n)] = E_{P^*}[\text{ER}(X, Y|\tilde{H})] \quad \textbf{(reliability)} \quad (15)$$

Example 3 [Squared Errors] Assume the setting of Example 1. Theorem 2 shows that, if data are i.i.d. according to a distribution P^* such that, among *all* models in the class $\mathcal{M}_k$ of k-th degree polynomials, a certain polynomial $\tilde{H}$ is optimal in that it minimizes the true expected squared error $E_{P^*}[\text{ER}_{sq}(X, Y|H)]$, then the following holds: the true expected error of predictions based on the $\hat{H}_n$-component of $(\hat{H}_n, \hat{\beta}_n)$ will converge to the optimal expected error $E_{P^*}[\text{ER}_{sq}(X, Y|\tilde{H})]$. Also, $E_{(\hat{H}_n, \hat{\beta}_n)}[\text{ER}_{sq}(X, Y|\hat{H}_n)]$, the estimate of the expected error that $\hat{H}_n$ will make over future data, will converge to the true value $E_{P^*}[\text{ER}_{sq}(X, Y|\tilde{H})]$ (note that the possibility $\tilde{\beta} = 0$ cannot happen with the squared error since $0 \notin \Gamma_{nat}$). Even though $\langle \mathcal{M}_k \rangle_{\text{ER}_{sq}}$ is a class of Gaussians with conditional means given by a k-th degree polynomial, all this happens even if neither $E_{P^*}[Y|X]$ is a polynomial nor $P^*(Y|X = x)$ is a Gaussian.

Sophisticated Convergence Theorem We now move to error functions like the 0/1-loss for which β may become 0 or even negative. We now need some new notation: given sample space Ω, a class $\mathcal{M}$ of decision functions $H : \Omega_x \to \Delta$ and a distribution P over Ω, define

$$\text{ER}(x, y|P) := \text{ER}(x, y|\tilde{H}_P) \quad (16)$$

where

$$\tilde{H}_P = \arg\min E_P[\text{ER}(X, Y|H)] \quad (17)$$

and the minimum is taken over the space of all decision functions $H : \Omega_x \to \Delta$ (in the theorem below, the minimum can always be assumed to exist; if the minimum is attained for several H, we take an arbitrary one among them). We write $\text{ER}(x, y|H, \beta)$ as shorthand for $\text{ER}(x, y|P(\cdot|H, \beta))$. We will use the models $(\hat{H}_n, \hat{\beta}_n)$ that we infer from the data $(x, y)^n$ in a different way now: instead of predicting future y directly using $\hat{H}_n(x)$, we will predict using Equation 17 with $P(\cdot|\hat{H}_n, \hat{\beta}_n)$ substituted for the P appearing in the equation. Hence the error that model $(\hat{H}_n, \hat{\beta}_n)$ makes on data (x, y) is given by $\text{ER}(x, y|\hat{H}_n, \hat{\beta}_n)$ rather than $\text{ER}(x, y|\hat{H}_n)$. It turns out that if $\hat{\beta}_n > 0$, then this does not make any difference. This is shown by the following fact, proven in the appendix:

Fact 1 *If $\beta > 0$, then for all $H \in \mathcal{M}$, for all $(x, y) \in \Omega$, we have*

$$\text{ER}(x, y|H) = \text{ER}(x, y|H, \beta).$$

However, if $\beta < 0$, then $\text{ER}(x, y|H) \neq \text{ER}(x, y|H, \beta)$. which is indeed what we want; see Example 4 below.

Theorem 3 below is an analogue of Theorem 2 without restrictions on β; it works for model classes like the squared error for which $\Omega_y = \mathbf{R}$ and for the symmetric error functions which we defined earlier. In case $\Omega_y = \mathbf{R}$ Theorem 3 follows almost immediately from Theorem 2 because of Fact 1 above together with the following:

Fact 2 *If $\Omega_y = \mathbf{R}$, then Γ_{nat} contains only values $\beta > 0$.*

In order to state the theorem, we must first define the *KL-optimal model* $P(\cdot|\tilde{H}, \tilde{\beta}) \in \langle \mathcal{M} \rangle_{\text{ER}}$ as the model that minimizes expected code length (or equivalently, KL distance to the true model):

$$P(\cdot|\tilde{H},\tilde{\beta}) = \arg\min_{P(\cdot|H,\beta)\in\langle\mathcal{M}\rangle_{\text{ER}}} E_{P^*}[-\ln P(X,Y|H,\beta)] \tag{18}$$

To get a more tangible idea of what the following theorem says, we refer to Example 4.

Theorem 3 *Let sample space Ω, model class $\mathcal{M}$, error function* ER *and P^* be as required in* **C1** – **C4**. *Assume* ER *is 'simple' and EITHER $\Omega_y = \mathbf{R}$ OR* ER *is 'symmetric'. Assume data are i.i.d. according to P^*. Assume a KL-optimal model $P(\cdot|\tilde{H},\tilde{\beta}) \in \langle\mathcal{M}\rangle_{\text{ER}}$ as in (18) exists. Then (1)*

$$\begin{aligned} E_{(\tilde{H},\tilde{\beta})}[\text{ER}(X,Y|\tilde{H},\tilde{\beta})] &= E_{P^*}[\text{ER}(X,Y|\tilde{H},\tilde{\beta})] \\ &\leq \min_{H\in\mathcal{M}}\{E_{P^*}[\text{ER}(X,Y|H)]\} \end{aligned} \tag{19}$$

where the last inequality becomes an equality if $\tilde{\beta} > 0$. Furthermore, (2) with P^*-probability 1, for all large n, $(\hat{H}_n, \hat{\beta}_n)$ exists, and:

$$\lim_{n\to\infty} E_{P^*}[\text{ER}(X,Y|\hat{H}_n,\hat{\beta}_n)] = E_{P^*}[\text{ER}(X,Y|\tilde{H},\tilde{\beta})] \quad \textbf{(optimality)} \tag{20}$$

and also

$$\lim_{n\to\infty} E_{(\hat{H}_n,\hat{\beta}_n)}[\text{ER}(X,Y|\hat{H}_n,\hat{\beta}_n)] = E_{P^*}[\text{ER}(X,Y|\tilde{H},\tilde{\beta})] \quad \textbf{(reliability)} \tag{21}$$

Example 4 [Classification] Let Ω_x be a rectangle in $\mathbf{R}^l$ for some $l > 0$, $\Omega_y = \Delta = \{0,1\}$. Let $\mathcal{M}_k$ be the class of all back-propagation neural networks with l inputs, k hidden units and one output. Each $H \in \mathcal{M}_k$ is turned into a decision rule by setting $H(x) := 1$ iff the output of H on x is greater than $1/2$, and $H(x) := 0$ otherwise. What happens if $(\tilde{H},\tilde{\beta})$ is such that $\tilde{\beta} < 0$? For example, suppose that the ER_{01}-optimal $H_0 \in \mathcal{M}$ has expected error $E_{P^*}[\text{ER}_{01}(X,Y|H_0)] = 0.2$. However, suppose also that $\mathcal{M}$ contains another H_1 with $E_{P^*}[\text{ER}_{01}(X,Y|H_1)] = 0.95$. By *inverting* this H_1 (predicting a 0 when $H_1(x) = 1$ and vice versa) this H_1 can be modified into a hypothesis that has smaller expected error than H_0. Assume further that H_1 *maximizes* the expected error among all models in $\mathcal{M}$. Applying maximum likelihood to $\langle\mathcal{M}\rangle_{\text{ER}_{01}}$, $\hat{H}_n$ will converge to H_1 rather than H_0. Associating predictions with the inferred model $(\hat{H}_n,\hat{\beta}_n)$ using Equation 16 makes sure that the predictions obtained from $\hat{H}_n$ are actually inverted.

Strictly speaking, Theorem 3 does not apply to this neural network case, since the entropified model class $\langle\mathcal{M}_k\rangle_{\text{ER}_{01}}$ does not admit a continuous parameterization (this is due to the discontinuity in the decision rules). We can overcome this problem by using *randomized decision rules*, which means that, for each given $x \in \Omega_x$, using H we predict a '1' with probability $P(Y = 1) := H(x)$. Using randomized decision rules, a version of all our theorems can still be proved; we omit details here.

In order to express our rate of convergence results, we use a framework and notation due to Barron and Cover [3]. For sequences of positive random variables X_n, the notation $X_n \lesssim R_n$ stands for convergence in probability at the indicated rate. This means that the ratio X_n/R_n is bounded in probability, i.e. for every $\epsilon > 0$, there exists a $c > 0$ such that $P(X_n/R_n > c) \leq \epsilon$ for all large n.

Theorem 4 (Rate of convergence) *Let Ω, $\mathcal{M}$,* ER *and P^* satisfy conditions* **C1**-**C4**. *Assume* ER *is 'simple'. Assume further EITHER $\Omega_y = \mathbf{R}$ OR* ER *is 'symmetric'. Assume data are i.i.d. according to P^*. Assume a single* ER*-optimal hypothesis $\tilde{H}$, as given by (10), exists; assume that (the parameter standing for) $\tilde{H}$ lies in the interior of $\Gamma_{\mathcal{M}}$, the parameterization of $\mathcal{M}$. Then*

$$E_{P^*}[\text{ER}(X,Y|\hat{H}_n,\hat{\beta}_n)] - E_{P^*}[\text{ER}(X,Y|\tilde{H},\tilde{\beta})] \lesssim 1/n \tag{22}$$

Hence the convergence, at least asymptotically and in probability, is as fast as convergence of ordinary ML estimation of probabilistic models in the classical setting where one of the models in the class is supposed to generate the data (see for example [17]).

8 HOW GENERAL IS THIS? – ADDITIONAL RESULTS

What have we achieved? We give a brief discussion.

8.1 A MISSING RESULT

Obviously, one key result is missing: can an analogue of Theorems 2–4 be proved if model classes with an unbounded number of parameters are used and MDL and/or Bayesian methods are used in stead of Maximum Likelihood? We are still working on this question.

8.2 HOW GENERAL IS ENTROPIFICATION?

Beyond 'simple' error functions The proofs of all our theorems make heavy use of the fact that only 'simple' error functions are used in the entropification procedure. Many practically relevant error functions are, unfortunately, not simple. The most important are perhaps the *logarithmic loss* for probabilistic model classes and *asymmetric loss functions* for the classification setting (where, say, classifying x as 1 where it should be 0 may lead to a much higher loss than classifying x as 0 where it should be 1). Fortunately, both kinds of loss functions can be handled. It is straightforward to adapt all theorems presented here to the logarithmic loss function; a preliminary version of this result can be found in the author's thesis [6]. In the logarithmic loss case, the definition of entropification does not need to be changed. To deal with asymmetric loss functions, the definition needs to be generalized. This generalization turns out to be possible but by no means trivial (many of the proof techniques used here cannot be applied any more) and will be published in a future paper. Finally, it turns out to be possible to entropify with respect to several error functions at the same time (so that the inferred models become reliable with respect to all the error functions used).

Predicting against $\text{ER}' \neq \text{ER}$ In practical applications, one is often forced to *learn* data using one error function ER, while the inferred model is then *used* in predictions/decisions involving a different error function ER. Examples are neural networks and Naive Bayes classifiers which are usually trained with respect to the squared error and the logarithmic error respectively, and then used for classification, i.e. prediction with respect to $0/1$-error. In principle, one could 'entropify' both neural nets and Naive Bayes model classes with respect to the $0/1$-error but the resulting inference procedure would become computationally very expensive. Is there any chance of obtaining good results when using a model class that is entropified with respect to ER for predictions against ER$'$ where ER$'$ is not a linear function of ER? Interestingly, even though we cannot expect that the inferred model $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\text{ER}}$ for data $(x,y)^n$ is close to the 'true' P^*, it sometimes make sense to use $E_{(H,\beta)}[\text{ER}'(Y;\delta)]$ as a *guess* of $E_{P^*}[\text{ER}'(Y;\delta)]$ even though ER$'$ is an error function different (not a linear transformation of) ER. The reason is the following

Fact 3 *Assume the setting of Theorem 3. Assume data are i.i.d. according to P^*. Assume a KL-optimal model $P(\cdot|\tilde{H},\tilde{\beta}) \in \langle\mathcal{M}\rangle_{\text{ER}}$ as in (18) exists and let $t = E_{P^*}[\text{ER}(X,Y|\tilde{H})]$. The KL-optimal model $P(\cdot|\tilde{H},\tilde{\beta})$ is the* maximum entropy model *for the constraint*

$$E[\text{ER}(X,Y|\tilde{H})] = t$$

This means [7, 13] that there are several reasons for adopting $P(\cdot|\tilde{H},\tilde{\beta})$ as a 'best available guess' of P^*, even though it is unreasonable to assume that really $P(\cdot|\tilde{H},\tilde{\beta}) = P^*$.

Safe Statistics Related is the question of 'what can entropified model classes be safely used for and what not'; in general, one can make 'safe' inferences (using the inferred model $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\text{ER}}$ to predict against error function ER); 'risky' inferences (using the inferred model $P(\cdot|H,\beta)$ to predict against an error function ER$'$ that is not a linear function of ER, but that is still such that $E_{(H,\beta)}[\text{ER}'(Y;\delta)]$ provides a reasonable guess of $E_{P^*}[\text{ER}'(Y;\delta)]$) and finally 'dangerous' inferences (inferences that would only make sense if the inferred model $P(\cdot|H,\beta)$ is almost identical to P^* - something which we certainly cannot hope for). Our results in this direction will be left for another paper.

9 THE UNIVERSAL YARDSTICK

The title of this paper is inspired by a claim by Rissanen [12, page 20]: 'we then see that the unification obtained by interpreting all models as probabilistic has given us an immutable yardstick, the code length, which can never be reduced to zero by scaling or other devices...'. In the present paper we analyzed and partially substantiated this claim.

10 ACKNOWLEDGMENTS

Most of the work reported here was done while the author was a Ph.D. student at the CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. The author is currently supported by a TALENT-grant awarded by the Netherlands Organization for Scientific Research (NWO). He wishes to thank Phil Dawid, Jorma Rissanen, Paul Vitányi, Ronald de Wolf and Kenji Yamanishi for many helpful discussions and lots of encouragement in pursuing this research.

References

[1] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[2] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576, Dordrecht, 1990. Kluwer Academic Publishers.

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

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Interscience, New York, 1991.

[5] Q. Gao and M. Li. An application of minimum description length principle to online recognition of handprinted alphanumerals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pages 843–848, 1989.

[6] P.D. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, University of Amsterdam, The Netherlands, October 1998. Available as ILLC Dissertation Series 1998-03; see http://robotics.stanford.edu/~grunwald.

[7] E.T. Jaynes. Probability theory: the logic of science. Available at ftp://bayes.wustl.edu/Jaynes.book/, 1996. Jaynes' forthcoming monumental treatise on probability theory as extended logic. This preliminary version (1996) is available via the web.

[8] M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

[9] J. Quinlan and R. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, 1989.

[10] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[11] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society, series B*, 49:223–239, 1987. Discussion: pages 252–265.

[12] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, 1989.

[13] J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, 1980.

[14] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[15] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computing Journal*, 11:185–195, 1968.

[16] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, 49:240–251, 1987. Discussion: pages 252–265.

[17] S.S Wilks. *Mathematical statistics*. John Wiley, 1962.

[18] K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*,

44(4):1424–1439, 1998.

APPENDIX: PROOFS OF THEOREMS 1–4

All proofs make heavy use of the following Claim 1 and the (more interesting) Proposition 5 which both list a number of properties of entropified model classes. These properties are all directly related to the fact that entropified model classes are essentially 1-parameter exponential families. Claim 1 is restricted to models $P(\cdot|H,\beta)$ where H is fixed and β may vary; in this case, exponentiality is clear. The requirement of using 'simple' error functions makes sure that, for fixed β, the distributions $P(\cdot|H_1,\beta)$ and $P(\cdot|H_2,\beta)$ for $H_1 \neq H_2$ are still 'the same' in that their moments correspond: $E_{(H_1,\beta)}[\text{ER}(X,Y|H_1)] = E_{(H_2,\beta)}[\text{ER}(X,Y|H_2)]$, which is shown formally in Proposition 5.

Claim 1 *Let sample space Ω, model class $\mathcal{M}$, error function* ER *and P^* satisfy conditions* **C1**-**C4**. *Let $\Gamma^{\mathbf{R}}_{nat} := \Gamma_{nat} \setminus \{-\infty,\infty\}$. Then $\Gamma^{\mathbf{R}}_{nat}$ is an open set. For all fixed $H \in \mathcal{M}$ and all $\beta \in \Gamma^{\mathbf{R}}_{nat}$ we have:*

1. *$Z_H(\beta)$ and $E_{P^*}[-\ln P(X,Y|H,\beta)]$ are continuous, twice differentiable functions at β.*
2. $(\partial/\partial\beta)\ln Z_H(\beta) = -E_{(H,\beta)}[\text{ER}(X,Y|H)]$.
3. $(\partial^2/\partial\beta^2)\ln Z_H(\beta) =$ $E_{(H,\beta)}[(\text{ER}(X,Y|H) - E_{(H,\beta)}[\text{ER}(X,Y|H)])^2] =$ $\text{var}_{(H,\beta)}(\text{ER}(X,Y|H)) \geq 0$.
4. $(\partial/\partial\beta)E_{(H,\beta)}[\text{ER}(X,Y|H)] =$ $-(\partial^2/\partial\beta^2)\ln Z_H(\beta) \leq 0$.
Hence $E_{(H,\beta)}[\text{ER}(X,Y|H)]$ is a monotonically decreasing function of β.
5. $(\partial/\partial\beta)E_{(H,\beta)}[-\ln P(X,Y|H,\beta)] =$ $\beta\cdot(\partial/\partial\beta)E_{(H,\beta)}[\text{ER}(X,Y|H)] =$ $-\beta\cdot(\partial^2/\partial\beta^2)\ln Z_H(\beta) = \begin{cases} \geq 0 & \text{if } \beta < 0 \\ \leq 0 & \text{if } \beta > 0 \end{cases}$
Hence $E_{(H,\beta)}[-\ln P(X,Y|H,\beta)]$ is a monotonically decreasing function of $|\beta|$.
6. $(\partial/\partial\beta)E_{P^*}[-\ln P(X,Y|H,\beta)] =$ $E_{P^*}[\text{ER}(X,Y|H)] - E_{(H,\beta)}[\text{ER}(X,Y|H)]$.
Hence $(\partial/\partial\beta)E_{P^*}[-\ln P(X,Y|H,\beta)]$ is a decreasing function of β.
7. $(\partial^2/\partial\beta^2)E_{P^*}[-\ln P(X,Y|H,\beta)] =$ $(\partial^2/\partial\beta^2)\ln Z_H(\beta) \geq 0$.
Hence $E_{P^*}[-\ln P(X,Y|H,\beta)]$ as a function of β is convex.

The proof of the statements in the claim is straightforward by differentiation and the fact that, by our regularity conditions, we are allowed to swap the order of differentiation and integration. Details are omitted.

In the remainder of this section, we will only prove propositions and theorems concerning 'simple' error functions, for which $Z_{H_1}(\beta) = Z_{H_2}(\beta)$ for all $H_1, H_2 \in \mathcal{M}$. Hence we can unambiguously write $Z(\beta)$.

Proposition 5 *Let sample space Ω, model class $\mathcal{M}$, error function* ER *and P^* satisfy conditions* **C1**-**C4**. *Assume* ER *is 'simple' (Definition 5). Then*

1. *For each fixed $H \in \mathcal{M}$,*
$$\min_{\beta\in\Gamma_{nat}} E_{P^*}[-\ln P(X,Y|H,\beta)] \tag{23}$$
exists and is attained at a single value $\beta = \tilde{\beta}$ satisfying
$$E_{(H,\tilde{\beta})}[\text{ER}(X,Y|H)] = E_{P^*}[\text{ER}(X,Y|H)] \tag{24}$$
and
$$E_{(H,\tilde{\beta})}[-\ln P(X,Y|H,\tilde{\beta})] = E_{P^*}[-\ln P(X,Y|H,\tilde{\beta})] \tag{25}$$
2. *For each fixed $H \in \mathcal{M}$ and each $(x,y)^n \in \Omega^n$,*
$$\min_{\beta\in\Gamma_{nat}}\{-\ln P((x,y)^n|H,\beta)\}$$
exists and is attained for a single ('maximum likelihood') value $\beta = \hat{\beta}_n$ satisfying
$$E_{(H,\hat{\beta}_n)}[\text{ER}(X,Y|H)] = \frac{1}{n}\sum_{i=1}^{n}\text{ER}(x_i,y_i|H) \tag{26}$$
3. *For all fixed $\beta \in \Gamma_{nat}$ and all $H_1, H_2 \in \mathcal{M}$ we have:*
$$E_{(H_1,\beta)}[\text{ER}(X,Y|H_1)] = E_{(H_2,\beta)}[\text{ER}(X,Y|H_2)] \tag{27}$$
and similarly
$$E_{(H_1,\beta)}[-\ln P(X,Y|H_1,\beta)] = E_{(H_2,\beta)}[-\ln P(X,Y|H_2,\beta)]. \tag{28}$$

Proof We start with item 1. By Claim 1, item 6., $(\partial/\partial\beta)$ $E_{P^*}[-\ln P(X,Y|H,\beta)]$ is a decreasing function for $\beta \in \Gamma^{\mathbf{R}}_{nat}$. Together with item 7. of the claim, this implies that $E_{P^*}[-\ln P(X,Y|H,\beta)]$ has at most one minimum for $\beta \in \Gamma^{\mathbf{R}}_{nat}$. Let us first suppose this minimum exists, and let it be attained for $\tilde{\beta}$. It now follows from item 1. and item 6. of the claim that (24) holds at this minimum. From the definition of $P(\cdot|H,\beta)$ for $\beta = (-)\infty$ we see that the minimum (23) is also the minimum over all Γ_{nat}. Now let us suppose the minimum (over $\beta \in \Gamma^{\mathbf{R}}_{nat}$) does not exist. From the fact that $Z_H(\beta)$ is continuous within $\Gamma^{\mathbf{R}}_{nat}$ and regularity condition **C3** it follows there are only two possibilities for $\Gamma^{\mathbf{R}}_{nat}$: EITHER $\Gamma^{\mathbf{R}}_{nat} = \{\beta \in \mathbf{R}|\beta > 0\}$ and $\lim_{\beta\downarrow 0} Z(\beta) = \infty$ OR $\Gamma^{\mathbf{R}}_{nat} = \mathbf{R}$ and $Z(0) < \infty$. We only consider the case $\Gamma^{\mathbf{R}}_{nat} = \{\beta \in \mathbf{R}|\beta > 0\}$; the other case can be handled in a similar, but simpler manner. By item 7 of Claim 1, $E_{P^*}[-\ln P(X,Y|H,\beta)]$ must have an infimum either at $\beta = 0$ or at $\beta = \infty$ (recall that we are assuming that no minimum for $\beta \in \Gamma^{\mathbf{R}}_{nat}$ exists). Since also $\lim_{\beta\downarrow 0} Z(\beta) = \infty$ we get $\lim_{\beta\downarrow 0} E_{P^*}[-\ln P(X,Y|H,\beta)]$ $= \lim_{\beta\downarrow 0} E_{P^*}[\beta\text{ER}(X,Y|H) + \ln Z(\beta)] = \infty$. On the other hand, for $\beta > 0$, $E_{P^*}[-\ln P(X,Y|H,\beta)] < \infty$ by our regularity conditions on ER and P^*. Hence the infimum cannot be at $\beta = 0$ and must be at $\beta = \infty$. From the definition of entropification we now see that $P(\cdot|H,\infty)$ is the uniform distribution on the set

$$S := \{(x,y) \in \Omega \mid \text{ER}(x,y|H) = \min_{(x,y)\in\Omega}\text{ER}(x,y|H)\}.$$

Hence EITHER $E_{P^*}[-\ln P(X,Y|H,\infty)] = \infty$ OR the support of P^* is the set S defined above. The first possibility is in contradiction with our assumption that $E_{P^*}[-\ln P(X, Y|H, \infty)]$ is an infimum and can be dismissed. Hence the latter case applies, the minimum in (23) exists, is reached only for $\beta = \infty$ and (24) is immediate. We have now proved item 1. except for Equation 25. Since, for fixed β, $-\ln P(x, y|H, \beta)$ is a linear function of $\text{ER}(x, y|H)$, (25) is immediate from (24). This completes the proof of item (1).

Item (2). can be handled analogously to item (1), using an analogue of Claim 1 (which we do not present here) in which statements about $E_{P^*}[-\ln P(X, Y|H, \beta)]$ are replaced by statements about $n^{-1}\sum_{i=1}^{n} -\ln P(x_i, y_i|H, \beta)$. We omit the details (see [6], Chapter 5).

Item (3). is straightforward from the definition of 'simple' error functions (Definition 5): for all $H_1 \neq H_2$, for all β: $Z_{H_1}(\beta) = Z_{H_2}(\beta)$ and hence also $(\partial/\partial\beta)\ln Z_{H_1}(\beta) = (\partial/\partial\beta)\ln Z_{H_2}(\beta)$ for all β for which the derivatives exist. By Claim 1, item 2., this proves (27) for $\beta \in \Gamma^{\mathbf{R}}_{nat}$. Since for fixed β, $-\ln P(x, y|H, \beta)$ is a linear function of $\text{ER}(x, y|H)$, (28) also follows. This leaves only the case $\beta = (-)\infty$, which can be handled by limit arguments; we omit details. □

PROOF OF THEOREM 1

The existence of the minimum $\widetilde{\text{ER}}(P^*)$ follows directly from our regularity conditions on P^* (**C2**, i.e. compactness) and $\mathcal{M}$ (**C4**, i.e. parameterizability); this proves the first part of Theorem 1, item (1). To prove the remainder of the theorem, let $\mathcal{M}_{\text{OPT}}$ be the set of all $H \in \mathcal{M}$ for which the minimum $\widetilde{\text{ER}}(P^*)$ is attained. Let $\tilde{H} \in \mathcal{M}$ be an (arbitrary) member of $\mathcal{M}_{\text{OPT}}$. From Proposition 5, item (1). together with Claim 1, item 7, we see that

$$\min_{\substack{\beta\in\Gamma_{nat}\\ \beta\geq 0}} \{E_{P^*}[-\ln P(X, Y|\tilde{H}, \beta)]\} = \min_{\substack{\beta\in\Gamma_{nat}\\ \beta\geq 0}} \{\beta \cdot \widetilde{\text{ER}}(P^*) + \ln Z_{\tilde{H}}(\beta)\} \tag{29}$$

exists and is reached for a single value of $\tilde{\beta}$. Since ER is 'simple' (Definition 5), for all $H \in \mathcal{M}$ it holds $Z_H(\beta) = Z_{\tilde{H}}(\beta)$. It now follows that the minimum in (29) is reached at the same value $\tilde{\beta}$ for *all* $H \in \mathcal{M}_{\text{OPT}}$. The remainder of the proof is split in two cases: $\tilde{\beta} = 0$ and $\tilde{\beta} > 0$.

Case 1: $\tilde{\beta} = 0$ We only consider the case where $\mathcal{M}$ contains an H' such that $H' \notin \mathcal{M}_{\text{OPT}}$, i.e.

$$E_{P^*}[\text{ER}(X, Y|H')] > E_{P^*}[\text{ER}(X, Y|\tilde{H})] \tag{30}$$

(the proof is trivial if such an H' does not exist). We have for all $\beta > 0$,
$E_{P^*}[-\ln P(X, Y|H', \beta)] > E_{P^*}[-\ln P(X, Y|\tilde{H}, \beta)] >$
$E_{P^*}[-\ln P(X, Y|\tilde{H}, 0)] = \ln Z(0) =$
$= E_{P^*}[-\ln P(X, Y|H', 0)]$.
Here the first inequality follows from (30) and the fact that for fixed β, $-\ln P(x, y|H, \beta)$ is a linear function of $\text{ER}(x, y|H)$; the second inequality follows from the fact that $\tilde{\beta} = 0$ and the two equalities are immediate. The chain of inequalities shows that the minimum $\tilde{L}(P^*)$ in (12) exists and, for all $H \in \mathcal{M}$, is reached for $P(\cdot|H, \beta)$ iff $\beta = 0$. This completes the proof of the theorem for the case $\tilde{\beta} = 0$.

Case 2: $\tilde{\beta} > 0$ Again, we only consider the case where $\mathcal{M}$ contains an H' such that $H' \notin \mathcal{M}_{\text{OPT}}$, i.e. (30) holds. (the proof is trivial if such an H' does not exist). We have for all $\beta > 0$,
$E_{P^*}[-\ln P(X, Y|H', \beta)] > E_{P^*}[-\ln P(X, Y|\tilde{H}, \beta)] \geq$
$E_{P^*}[-\ln P(X, Y|\tilde{H}, \tilde{\beta})]$
and for $\beta = 0$ we have $E_{P^*}[-\ln P(X, Y|H', \beta)] = \ln Z(0)$
$= E_{P^*}[-\ln P(X, Y|\tilde{H}, 0)] > E_{P^*}[-\ln P(X, Y|\tilde{H}, \tilde{\beta})]$
(where the (in)equalities follow by similar reasoning as in Case 1 above). Combining both chains of (in)equalities, we find that

$$\min_{\substack{\beta\in\Gamma_{nat}\\ \beta\geq 0}} \{E_{P^*}[-\ln P(X, Y|H', \beta)]\} > E_{P^*}[-\ln P(X, Y|\tilde{H}, \tilde{\beta})] \tag{31}$$

From (31) together with the uniqueness of $\tilde{\beta}$ for all $H \in \mathcal{M}_{\text{OPT}}$, which we established above, we see that the minimum $\tilde{L}(P^*)$ in (12) exists and that a model (H, β) minimizes (12) iff $H \in \mathcal{M}_{\text{OPT}}$ and $\beta = \tilde{\beta}$. This proves items 1, 2 and 3 of the theorem for the case $\tilde{\beta} > 0$. Item 4. now follows directly from Proposition 5, item 1. □

PROOF OF THEOREM 2

We first need the following proposition and lemma:

Proposition 6 *Let Ω, $\mathcal{M}$ and* ER *satisfy conditions* **C1** *-***C4**. *Let* ER *be a simple error function. Then, if the ML estimator $(\hat{H}_n, \hat{\beta}_n)$ for data $(x, y)^n$ exists and $\hat{\beta}_n > 0$, $\hat{H}_n$ is given by*

$$\hat{H}_n = \arg\min_{H\in\mathcal{M}}\{\text{ER}((x, y)^n|H)\}.$$

In words, if only the H-component of a model in $\langle\mathcal{M}\rangle$ is taken into account and β may be assumed > 0, then ML estimation reduces to empirical error minimization.

Proof Straightforward from definition of entropification and of simple error functions; details omitted. □

Lemma 7 (Uniform Convergence) *Let Ω, $\mathcal{M}$,* ER *and P^* be as required in* **C1** *–* **C4**. *Assume data are i.i.d. according to P^*. Then with P^*-probability 1,*

$$\lim_{n\to\infty} \sup_{H\in\mathcal{M}} \{|\frac{1}{n}\sum_{i=1}^{n} \text{ER}(x_i, y_i|H) - E_{P^*}[\text{ER}(X, Y|H)]|\} = 0 \tag{32}$$

Proof Straightforward by the method of *ϵ-nets*, see Vapnik's book on uniform convergence theory [14]. Otherwise, see [6], Chapter 5, Lemma 5.14 and its proof. □

Proof (of Theorem 2)
Item 1. in the theorem is immediate from Theorem 1. Concerning item 2., we first prove existence of the ML estimator $(\hat{H}_n, \hat{\beta}_n)$ for large n and at the same time Equation 14.

Since $\mathcal{M}$ can be compactly parameterized, at least one H maximizing $n^{-1}\sum_{i=1}^{n} \text{ER}(x_i, y_i|H)$ exists for all n. Let $\mathcal{M}_{\text{ML}}$ be the set of all such H. It can now be established by repeatedly using Lemma 7 that uniformly for *all* $H \in \mathcal{M}_{\mathrm{ML}}$, with P^*-probability 1,

$$E_{P^*}[\mathrm{ER}(X,Y|H)] \rightarrow E_{P^*}[\mathrm{ER}(X,Y|\tilde{H})]. \quad (33)$$

Moreover we have by Proposition 5, item 2. that for all $H \in \mathcal{M}_{\mathrm{ML}}$,

$$\min_{\beta \in \Gamma_{nat}} \{-\ln P((x,y)^n | H, \beta)\}$$

exists and is attained for a single value $\hat{\beta}_n$, the same value being taken for all $H \in \mathcal{M}_{\mathrm{ML}}$. By Proposition 6 the set of models $P(\cdot|H,\hat{\beta}_n)$ with $H \in \mathcal{M}_{\mathrm{ML}}$ coincides exactly with the set S consisting of all $P(\cdot|H,\beta)$ maximizing the likelihood of $(x,y)^n$. By compactness (of $\mathcal{M}$) and continuity (of the parameterization of $\mathcal{M}$) arguments one can now establish that the set S of maximum likelihood estimators contains a minimal (in the lexicographical ordering induced by the parameterization) element, and hence the maximum likelihood estimator $(\hat{H}_n, \hat{\beta}_n)$ exists. Since $\hat{H}_n \in \mathcal{M}_{\mathrm{ML}}$, Equation 14 in the theorem now immediately follows from (33).

It only remains to prove Equation 15. By Proposition 5, item 2,

$$E_{(\hat{H}_n,\hat{\beta}_n)}[\mathrm{ER}(X,Y|\hat{H}_n)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{ER}(x_i, y_i|\hat{H}_n). \quad (34)$$

By uniform convergence (Lemma 7), this implies that with P^*-probability 1,

$$E_{(\hat{H}_n,\hat{\beta}_n)}[\mathrm{ER}(X,Y|\hat{H}_n)] \rightarrow E_{P^*}[\mathrm{ER}(X,Y|\hat{H}_n)] \quad (35)$$

which together with Equation 14 implies Equation 15. □

PROOF OF FACTS 1 AND 2 AND THEOREM 3

Proof (of Fact 1) Let $\beta > 0$. For any $H \in \mathcal{M}$, $\mathrm{ER}(x,y|H,\beta)$ is, by Equation 16, equal to $\mathrm{ER}(x,y|\tilde{H})$ where

$$\tilde{H} := \arg\min E_{(H,\beta)}[\mathrm{ER}(X,Y|H')]$$

and the minimum is over all decision functions H'. But

$$\begin{aligned}\arg\min_{H'} E_{(H,\beta)}[\mathrm{ER}(X,Y|H')] &= \\ \arg\min\{\beta E_{(H,\beta)}[\mathrm{ER}(X,Y|H')] + \ln Z(\beta)\} &= \\ \arg\min E_{(H,\beta)}[-\ln P(X,Y|H',\beta)] &= H \quad (36)\end{aligned}$$

where the last equality is an immediate consequence of the Information Inequality [4]. It follows that for all $(x,y) \in \Omega$, $\mathrm{ER}(x,y|H,\beta) = \mathrm{ER}(x,y|H)$. □

The proof of Fact 2 is straightforward and omitted.

Proof (of Theorem 3) (short sketch) We consider two cases.

Case 1: $\Omega_y = \mathbf{R}$ Using Facts 1 and 2 it is straightforward to show that Theorems 1 and 2 imply Theorem 3 for the case that $\Omega_y = \infty$. We omit details.

Case 2: ER **is symmetric** In this case, we still have (by Fact 1) that if $\beta > 0$, then $\mathrm{ER}(\cdot,\cdot|H,\beta) = \mathrm{ER}(\cdot,\cdot|H)$ for all $H \in \mathcal{M}$. In case $\beta < 0$, we can repeat the reasoning in the proof of Fact 1 with a little 'twist': for any $H \in \mathcal{M}$, $\mathrm{ER}(x,y|H,\beta)$ is equal to $\mathrm{ER}(x,y|\tilde{H})$ where $\tilde{H} := \arg\min E_{(H,\beta)}[\mathrm{ER}(X,Y|H')]$ and the minimum is over all decision functions H'. But, if $\beta < 0$,

$$\begin{aligned}\arg\min_{H'} E_{(H,\beta)}[\mathrm{ER}(X,Y|H')] &= \\ \arg\max\{\beta E_{(H,\beta)}[\mathrm{ER}(X,Y|H')] + \ln Z(\beta)\} &= \\ \arg\min\{\beta E_{(H,\beta)}[\mathrm{ER}(X,Y|\overline{H'})] + \ln Z(\beta)\} &= \\ \arg\min E_{(H,\beta)}[-\ln P(X,Y|\overline{H'},\beta)] &= \overline{H} \quad (37)\end{aligned}$$

where $\overline{H'}$ is the inverse of H' as defined in Definition 6. This implies that, if $\beta < 0$,

$$\mathrm{ER}(\cdot,\cdot|H,\beta) = \mathrm{ER}(\cdot,\cdot|\overline{H})$$

for all $H \in \mathcal{M}$. It is also easy to show that $Z(\beta) = Z(-\beta)$ for simple symmetric error functions. This means that, for any H and $\overline{H}$, we have $P(\cdot|H,\beta) = P(\cdot|\overline{H},-\beta)$ and $\mathrm{ER}(\cdot,\cdot|H,\beta) = \mathrm{ER}(\cdot,\cdot|\overline{H},-\beta)$. Armed with these facts, it is possible to make small adjustments to the proofs of Theorems 1 and 2 so that they become a proof of Theorem 3 for the case that ER is symmetric. We omit the details. □

PROOF OF THEOREM 4

The proof depends on Proposition 8 and Lemma 9 below.

Assume P^* satisfies regularity condition **C2**. Let $\mathcal{M}$ be a probabilistic model class, parameterized by a compact set $\Gamma_{\mathcal{M}} \subset \mathbf{R}^k$ for some $k > 0$, such that the parameterization is continuous and twice differentiable. Assume

$$\tilde{\theta} = \arg\min_{\theta \in \Gamma_{\mathcal{M}}} E_{P^*}[-\ln P(X,Y|\theta)] \quad (38)$$

exists and lies in the interior of $\Gamma_{\mathcal{M}}$. Define $D_{P^*}(\theta||\tilde{\theta}) := E_{P^*}[-\ln P(X,Y|\theta)] - E_{P^*}[-\ln P(X,Y|\tilde{\theta})]$ for P^* and $\tilde{\theta}$ as above, as a function of θ. $D_{P^*}(\theta||\tilde{\theta})$ can be seen as a generalization of the Kullback-Leibler Divergence.

Proposition 8 *Assume $\mathcal{M}$, $\tilde{\theta}$ and $D_{P^*}(\theta||\tilde{\theta})$ are as above. Then $D_{P^*}(\tilde{\theta}||\tilde{\theta}) = 0$, and for $1 \leq i \leq k$,*

$$\{\frac{\partial}{\partial \theta_i} D_{P^*}(\theta||\tilde{\theta})\}_{\theta=\tilde{\theta}} = 0.$$

Proof Straightforward from the conditions imposed on P^* and $\mathcal{M}$; details are omitted. □

The following lemma shows that, if $\mathcal{M}$ is continuously parameterized and the true P^* is i.i.d., then the maximum likelihood estimator is asymptotically normally distributed around $\tilde{\theta}$, the optimal model in $\mathcal{M}$, even if $\tilde{\theta}$ is very different from P^*. In the following lemma we use $\hat{\theta}_n$ to denote the maximum likelihood estimator within model $\mathcal{M}$ for data $(x,y)^n$. If the maximum likelihood estimator is not unique, we again use the one with the smallest value in the lexicographical ordering induced by the parameterization (see the main part of the paper, Section 7, just above Theorem 2). Such an estimator must always exist due to the compactness of $\mathcal{M}$ and the continuity of the parameterization.

Lemma 9 (Asymptotic Normality) *Let sample space Ω and distribution P^* be as required in* **C1** *and* **C2**. *Let $\mathcal{M}$ be a probabilistic model class parameterized by some $\Gamma \subseteq \mathbf{R}^k$ such that the parameterization is continuous and twice differentiable. Assume a (KL-distance minimizing) model $\tilde{\theta} \in \Gamma$ as given by (38) exists and lies in the interior of Γ. Then $\sqrt{n}(\hat{\theta}_n - \tilde{\theta})$ converges in distribution to a normal distribution with mean 0 and some covariance matrix Σ.*

Proof (sketch): Write $\theta = (\theta_1, \ldots, \theta_k)$ and consider the k-dimensional vector

$$L'((x,y)^n|\tilde{\theta}) := \{((\partial/\partial\theta_1)[-\ln P((x,y)^n|\theta)], \ldots,$$

$$(\partial/\partial\theta_k)[-\ln P((x,y)^n|\theta)])\}_{\theta=\tilde{\theta}} = \sum_{i=1}^{n} L'(x_i, y_i|\tilde{\theta}). \quad (39)$$

Claim 2 *Under distribution P^*, the $L'(X_i, Y_i|\tilde{\theta})$ are i.i.d. random variables with mean 0 and some common covariance matrix Σ'.*

To see that this claim holds, note that i.i.d.-ness of $L'(X_i, Y_i | \tilde{\theta})$ is trivial. From Proposition 8 we have that

$$\frac{\partial}{\partial\theta_i} E_{P^*}[L(X_i, Y_i|\tilde{\theta})] = 0.$$

From the regularity conditions on P^* and $\mathcal{M}$, it follows that we can exchange the order of taking expectations and differentiating, hence $L'(X_i, Y_i|\tilde{\theta})$ has mean 0. Existence of the covariance matrix follows once more from our conditions on P^* and $\mathcal{M}$.

The claim implies that $(1/\sqrt{n}) \cdot L'((X,Y)^n|\tilde{\theta})$ converges in distribution to a normally distributed random variable with mean 0 and some covariance matrix Σ''. Once this has been established, a standard proof of asymptotic normality of ML estimators (by Taylor expansion of $L'((X,Y)^n|\tilde{\theta})$ around $\hat{\theta}_n$) applies, see, for example, Wilks [17, page 358]. (In that proof, it is assumed that $P^* \in \mathcal{M}$; yet in the proof that fact is only used to establish asymptotic normality of $(1/\sqrt{n}) \cdot L'((X,Y)^n|\tilde{\theta})$ which in our case we have already established; the remainder of Wilks' proof does not depend on P^* being in $\mathcal{M}$). □

Proof (of Theorem 4)
We will only prove the theorem for the case $\Omega_y = \mathbf{R}$; it can be easily extended to the alternative case mentioned in the statement of the theorem (i.e. the case where ER is symmetric). By Fact 2 and Fact 1 we know that if $\Omega_y = \mathbf{R}$, then Γ_{nat} contains only values $\beta > 0$ and also for all $P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\text{ER}}$ we have $\forall(x,y) \in \Omega$: $\text{ER}(x,y|H,\beta) = \text{ER}(x,y|H)$. Hence, since we assume $\Omega_y = \mathbf{R}$ we only need to show that

$$E_{P^*}[\text{ER}(X,Y|\hat{H}_n)] - E_{P^*}[\text{ER}(X,Y|\tilde{H})] \lesssim 1/n. \quad (40)$$

(we already know from Theorem 2 that $\hat{H}_n$ exists for all n). By Theorem 1, there exists $(\tilde{H}, \tilde{\beta})$ minimizing

$$E_{P^*}[-\ln P(X,Y|H,\beta)].$$

with $\tilde{H}$ being the ER-optimal hypothesis as in the statement of Theorem 4. By Fact 2, $\tilde{\beta} > 0$. Consider now the restricted model class

$$\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}} := \{P(\cdot|H,\beta) \in \langle\mathcal{M}\rangle_{\text{ER}} \mid \beta = \tilde{\beta}\}.$$

By rewriting $-\ln P((x,y)^n|H,\tilde{\beta})$ as in (9) we see that $(\hat{H}_n, \tilde{\beta})$ is the ML estimator of $(x,y)^n$ within $\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}}$ iff $\hat{H}_n$ minimizes the empirical error $\text{ER}((x,y)^n|H)$ among all $H \in \mathcal{M}$. Hence, by Proposition 6, this component of the ML estimator within $\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}}$ agrees with the $\hat{H}_n$-component of the ML estimator $(\hat{H}_n, \hat{\beta}_n)$ within the unrestricted $\langle\mathcal{M}\rangle_{\text{ER}}$. Together with the fact that Equation 40 only depends on $\hat{H}_n$ and not on $\hat{\beta}_n$ this implies the following: in order to prove Equation 40 for class $\langle\mathcal{M}\rangle_{\text{ER}}$, it is enough to prove Equation 40 for the restricted class $\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}}$. By conditions $\mathbf{C1} - \mathbf{C4}$, the following fact holds:

Fact 4 *$\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}}$ can be parameterized by a compact set $\Gamma_0 \subset \mathbf{R}^k$ for some $k > 0$, where the parameterization is continuous and twice differentiable.*

Note that (40) is equivalent to

$$[\tilde{\beta} E_{P^*}[\text{ER}(X,Y|\hat{H}_n)] + \ln Z(\tilde{\beta})] -$$
$$[\tilde{\beta} E_{P^*}[\text{ER}(X,Y|\tilde{H})] + \ln Z(\tilde{\beta})] \lesssim 1/n. \quad (41)$$

which in turn is equivalent to

$$E_{P^*}[-\ln P(X,Y|\hat{H}_n, \tilde{\beta})] -$$
$$E_{P^*}[-\ln P(X,Y|\tilde{H}, \tilde{\beta})] \lesssim 1/n. \quad (42)$$

Writing $\hat{\theta}_n$ for the parameter value in Γ_0 that indexes $P(\cdot|\hat{H}_n, \tilde{\beta})$ and $\tilde{\theta}$ for the parameter value indexing $P(\cdot|\tilde{H}, \tilde{\beta})$, we see that (42) is equivalent to

$$D_{P^*}(\hat{\theta}_n||\tilde{\theta}) \lesssim 1/n. \quad (43)$$

So, if we can show (43) we are done. For any $\theta, \tilde{\theta} \in \Gamma_0$ we now write, for simplicity, $(\theta - \tilde{\theta})^2$ to denote the dot product of the vector $(\theta_1 - \tilde{\theta}_1, \ldots, \theta_k - \tilde{\theta}_k)$ with itself. By Proposition 8 and a straightforward Taylor expansion (valid because of Fact 4 above), we see that there exists an $\epsilon > 0$ such that if $(\theta - \tilde{\theta})^2 < \epsilon$, then

$$c_1(\theta - \tilde{\theta})^2 \leq D_{P^*}(\theta||\tilde{\theta}) \leq c_2(\theta - \tilde{\theta})^2$$

for some constants $c_1, c_2 > 0$. Inspecting the definition of $\lesssim$, we find that in order to prove (43) it is sufficient to show that for all $\epsilon > 0$ there exists $c > 0$ such that for all large n,

$$P^*(n(\hat{\theta}_n - \tilde{\theta})^2 > c) \leq \epsilon$$

which implies it is also enough to show that for all $\epsilon > 0$, there exists a vector $c' = (c'_1, \ldots, c'_k)^T$ where c'_i are all > 0, such that for all large n,

$$P^*(-c' \leq \sqrt{n}(\hat{\theta}_n - \tilde{\theta}) \leq c) \leq \epsilon. \quad (44)$$

But Lemma 9, which by Fact 4 must hold for $\langle\mathcal{M}\rangle_{\text{ER},\tilde{\beta}}$, implies (44) as a straightforward consequence. The theorem has been proved. □

PROOF OF FACT 3

Fact 3 can immediately be verified using standard results on Maximum Entropy; for example, see [4].