# Minimum Description Length Revisited

Peter Grünwald and Teemu Roos (12 March 2020)

International Journal of Mathematics for Industry Vol. 11, No. 1

**HVG Executive Summary (08/10/23)**

## 1. Introduction

- **The *MDL Principle* is a *Theory of Inductive Inference*** that can be applied to general problems in *Statistics, Machine Learning* and *Pattern Recognition*.
  - It starts by *assuming* (as an axiom) that ***modeling by data compression*** (or, equivalently, *sequential predictive log-loss minimization*) ***is the right thing to do***.
  - It states that the ***best explanation for a given set of data*** is provided by the ***shortest description of that data*** → obtained by ***choosing the model minimizing a code length*** (i.e., *minimizing prediction error in a stringent worst-case-over-all-data sense*).
  - *MDL*-type methods have a clear interpretation independent of whether any of the models under consideration is "*true*" (in the sense that it generates the data).
- ***Massive deployment of MDL has been hindered by two issues (addressed herein)*:**
  1) Applying MDL requires *basic knowledge of Statistics and Information Theory*.
  2) Many MDL procedures are *computationally highly intensive* or seem to require *arbitrary restrictions of parameter spaces*.
- ***Over the last 10 years, most of these issues have been resolved.***
- This paper is relevant to the ***Foundations of Statistics and Machine Learning***.

### Notational Preliminaries

- ***We are concerned with statistical models*** (families of probability distributions) $M = \{p_\theta : \theta \in \Theta\}$ parametrized by $\Theta$; and ***families of such models*** $\{M_\gamma : \gamma \in \Gamma\}$, where each $M_\gamma = \{p_\theta : \theta \in \Theta_\gamma\}$ is used to model ***data*** $z^n := (z_1, \dots z_n)$ with $z_i \in Z$, for some ***outcome space*** $Z$.
  - Each $p_\theta$ is defined on sequences of arbitrary length.
  - For i.i.d. data, we have $p_\theta(z^n) = \prod_{i=1}^n p_\theta(z_i)$
- The ML estimator given the model $M = \{p_\theta : \theta \in \Theta\}$ is denoted by $\hat{\theta}_{ML}$, and the ML estimator relative to model $M_\gamma$ is denoted by $\hat{\theta}_{ML|\gamma}$.
  - $\breve{\theta}$ denotes more general estimators
  - $\hat{\theta}_v$ denotes the *MDL estimator with luckiness function* $v$.

## 2. The Fundamental Concept: Universal Modeling

- **Let $M_\gamma;\ \gamma = 1,2,\dots,\gamma_{max}$ be a finite (or countably infinite) collection of statistical models, each of which is associated with a single distribution $\bar{p}_\gamma$, called a *Universal Distribution* relative to $M_\gamma$.**
  - The minus-log-likelihood $-log\left(\bar{p}_\gamma(Z^n)\right)$ is called the *code length of data* $Z^n$ *under universal code* $\bar{p}_\gamma$.
  - We equip the model indices $\Gamma := \{1,2,\dots,\gamma_{max}\}$ with a distribution $\pi(\gamma)$
  - Then, the model $M_\gamma$ giving the best explanation of data $z^n$ is obtained by maximizing $\bar{p}_\gamma(z^n)\pi(\gamma)$; or equivalently by minimizing: $-log\left(\bar{p}_\gamma(z^n)\right) - log\left(\pi(\gamma)\right)$
- ***Bayesian Universal Distribution***: In standard Bayesian model selection, for each $\gamma$, we set $\bar{p}_\gamma = p_{w_\gamma}^{BAYES}(z^n) = \int p_\theta(z^n) w_\gamma(\theta) d\theta$, for some prior density $w_\gamma(\theta)$ on the parameters in $\Theta_\gamma$, supplied by the user. However, defining $\bar{p}_\gamma$ this way is just *one particular way* to define an MDL universal distribution, and by no means the only one.

- ***Shtarkov Distribution*:** The Shtarkov distribution: $\bar{p}_\gamma = p_v^{NML}(z^n) \coloneqq \frac{\max_{\theta\in\Theta} p_\theta(z^n)v(\theta)}{\int \max_{\theta\in\Theta} p_\theta(z^n)v(\theta)dz^n}$ is perhaps the *most fundamental universal distribution*, where $v\colon \Theta \to R_0^+$ is a "*luckiness*" function (that does not have to be integrable, and is not necessarily a probability density).
  - ***Model Complexity:*** In the Shtarkov distribution above, the log of the integral in the denominator is called the *Model Complexity*: $COMP(M, v) \coloneqq \log \int \max_{\theta\in\Theta} p_\theta(z^n)v(\theta)\, dz^n$
  - ***MDL Estimators*:** Given a "*luckiness*" function $v$, the ***MDL Estimator Based On $v$*** defined as $\hat{\theta}_v \coloneqq \operatorname{argmax}_{\theta\in\Theta}\{p_\theta(z^n)v(\theta)\}$ can be considered a *penalized ML estimator* (coinciding with the *Bayes MAP estimator* whenever $v$ is a probability density).
    - By choosing $v$ sufficiently smooth, $\hat{\theta}_v$ will usually be almost indistinguishable from the ML estimator $\hat{\theta}_{ML}$ if the number of parameters is small relative to $n$.
    - Given (i) a set of models $M$ indexed by $\Gamma$, with (ii) luckiness functions $v$ specified on $\Theta_\Gamma$ for each $\gamma \in \Gamma$, and (iii) selecting a uniform distribution $\pi$ on $\Gamma$, the NML estimator is obtained by picking the model that minimizes: $-\log p_{\hat{\theta}_{v_\gamma(z^n)}} - \log v_\gamma\left(\hat{\theta}_{v_\gamma(z^n)}\right) + COMP\left(M_\gamma, v_\gamma\right)$ over $\gamma$, where $COMP\left(M_\gamma, v_\gamma\right) = \log \int p_{\hat{\theta}_{v_\gamma(z^n)}}(z^n) v_\gamma\left(\hat{\theta}_{v_\gamma}(z^n)\right) dz^n$ .
    - This estimator incorporates a trade-off between goodness of fit as measured by $-\log p_{\hat{\theta}_{v_\gamma(z^n)}}$ and model complexity as measured by $COMP$.
    - Although the *n-fold* integral inside $COMP$ looks daunting, Suzuki and Yamanishi show that in many cases it can be evaluated explicitly with appropriate choice of $v$.
- ***Normalized Maximum Likelihood (NML)*:** The NML distribution was originally defined by Shtarkov for special case with $v \equiv 1$, as $p^{NML}_{v=constant}(z^n) \coloneqq \frac{\max_{\theta\in\Theta} p_\theta(z^n)}{\int \max_{\theta\in\Theta} p_\theta(z^n)dz^n}$ so that $COMP(M, v) \coloneqq \log \int p_{\hat{\theta}_{ML(z^n)}}(z^n)dz^n$.
  - This is the version advocated by Rissanen as embodying the *purest form of the MDL Principle*.
  - For *finite outcome spaces*, $v \equiv 1$ usually "*works*", and the integral is well defined.
  - However, the integral is ill-defined for just about every parametric model defined on *unbounded outcome spaces* (such as $N;\ R\ or\ R^+$); using *nonuniform* $v$ allows one to deal with such cases.
- ***The Two-Part (Sub-)Distribution*:** In the two-part distribution $p_w^{2-P}$ (historically the *oldest* universal distribution), one *discretizes* $\Theta$ to some countable sub-set $\ddot{\Theta}$ and equips it with a probability mass function $w$ (that sums to 1) so that: $p_w^{NML}(z^n) \coloneqq \frac{\max_{\ddot{\theta}\in\ddot{\Theta}} p_{\ddot{\theta}}(z^n)w(\ddot{\theta})}{\int \max_{\ddot{\theta}\in\ddot{\Theta}} p_{\ddot{\theta}}(z^n)w(\ddot{\theta})dz^n}$
  - Since $\int \max_{\ddot{\theta}\in\ddot{\Theta}} p_{\ddot{\theta}}(z^n)w\left(\ddot{\theta}\right) dz^n \leq \sum_{\theta\in\ddot{\Theta}} w(\theta) \int p_{\ddot{\theta}}(z^n)dz^n = 1$, we can approximate $p_w^{NML}$ by sub-distribution $p_w^{2-P} \coloneqq \max_{\ddot{\theta}\in\ddot{\Theta}} p_{\ddot{\theta}}(z^n)w\left(\ddot{\theta}\right)$, which integrates to something smaller than 1.
  - We then imagine that $p_w^{2-P}$ puts its remaining mass on a special outcome, say "$\diamond$", which in reality will never occur.
- ***The Prequential Plug-In Distribution*:** Here, one takes any reasonable estimator $\breve{\theta}$ for the given model $M$, and define: $p_{\breve{\theta}}^{PREQ}(z^n) \coloneqq \prod_{i=1}^N p_{\breve{\theta}(z^{i-1})}\left(z_i|z^{i-1}\right)$ where for i.i.d. models, the probability inside the product simplifies to $p_{\breve{\theta}(z^{i-1})}(z_i)$. See ***Sec. 2.4***.
- ***The Switch Distribution*:** This universal distribution $\bar{p}^{SWITCH}$ behaves better in a particular type of nested model selection. See ***Sec. 3.1***.
- ***Universal Distributions $\bar{p}^{RIPR}$ based on the Reverse Information Projection*:** These lead to *improved error bounds and optional stopping behavior in hypothesis testing*; See ***Sec. 3.3***.

### *2.1. Motivation*

- ***The objective is a high-level motivation that avoids direct use of data compression arguments*.**

- Consider models $M_\gamma$ and define "*the fit of the model to the data*" as: $F_\gamma(z^n) := p_{\hat{\theta}_{ML|\gamma}}(z^n)$, where $\hat{\theta}_{ML|\gamma}$ is the *Maximum Likelihood* estimator within $M_\gamma$
- If we enlarge $M_\gamma$ (by adding distributions to it) then $F_\gamma(z^n)$ can only increase, and for big enough $M_\gamma$ we can have $F_\gamma(z^n) = 1$ on all data, and so be prone to ***severe overfitting***.
- Instead, a central idea of MDL is to associate each model $M_\gamma$ with a *single* corresponding distribution $\bar{p}_\gamma$ (set $F_\gamma(z^n) := \bar{p}_\gamma(z^n)$) so that no matter the choice of $\bar{p}_\gamma$ we chose, we have $\sum_{z^n} F_\gamma(z^n) = 1$, and so ***inherently prevent overfitting***.
- To decide *which* $\bar{p}$ are best associated with a given $M$, we *postulate* that a good choice is one in which the ***Fitness Ratio*** $FR(\bar{p}, z^n) = \frac{\bar{p}(z^n)}{\max_{\theta\in\Theta} p_\theta(z^n)v(\theta)}$ (where $v: \Theta \to R_0^+$ is a non-negative function) ***tends to be as large as possible***, define as being ***as large as possible in the worst-case*** (pick the $\bar{p}$ achieving $\max_{\bar{p}} \min_{z^n} FR(\bar{p}, z^n)$ ).
- This *maximin* problem has a solution *iff* complexity $COMP(M, v)$ is finite, and the unique solution is given by setting $\bar{p} = \bar{\boldsymbol{p}}^{\boldsymbol{NML}} := \frac{\max_{\theta\in\Theta} p_\theta(z^n)v(\theta)}{\int \max_{\theta\in\Theta} p_\theta(z^n)v(\theta)dz^n}$, which *has a special status as the most robust choice of universal* $\bar{p}$.

- ***This choice/interpretation does not require us to assume the model $M$ is "true" in any sense***.
- The nicest sub-case is with $v(\theta) \equiv 1$, but for most popular models with infinite $Z$, the problem $\max_{\bar{p}} \min_{z^n} FR(\bar{p}, z^n)$ usually has no solution due to the complexity $COMP(M, v)$ being infinite.
  - For all sufficiently "*regular*" models (curved exponential families) this problem can be solved by restricting $\Theta$ to a bounded subset, but since it can be unclear where to put the boundaries it is more natural to introduce a nonuniform $v$, chosen so that the complexity $COMP(M, v)$ is finite.

### *2.2. Asymptotic Expansions*

### *2.3. Unifying Model Selection and Estimation*

### *2.4. Log-loss Prediction and Universal Distributions*

### *2.5. The Luckiness Function*

## 3. Novel Universal Distributions

### *3.1 The Switch Distribution and the AIC–BIC Dilemma*

### *3.2. Hybrids between NML Bayes and Prequential Plug-in*

### *3.3 Hypothesis Testing: Universal Distributions Based on the Reverse Information Projection*

## 4. Graphical Models

- Graphical models are a framework for representing multivariate probabilistic models.
- *Choosing the right level of parsimony in graphical models is an ideal problem for MDL model selection*.
- While in Bayesian network model selection, the prevailing Bayesian paradigm embodies a particular form/variation of MDL, *recent studies have proposed new model selection criteria that exploit the NML distribution.*
  - One approach involves a continuous relaxation of NML-type complexities in which model selection problem takes on a Lasso-type L1-minimization form.
  - In other approaches, *NML* (or approximations thereof) are used directly for encoding parts of the model, including the *factorized NML (fNML)* score, and the *quotient NML (qNML)* score.
  - *In all these papers, both simulated and real-world data experiments suggest that the MDL-based criteria are quite robust with respect to the parameters in the underlying data source*.

- *Asymptotic expansion forms exist for certain model classes, revealing systematic differences between the complexities of different models even if they have the same number of parameters*.

## 5. Latent Variable and Irregular Models

- Tractable approximations of *NML-type* distributions have been developed for some of the most important *irregular (i.e., non-exponential family)* models such as hierarchical latent variable models, and the related Gaussian mixture models.
- For *irregular* models, Watanabe has proposed the *Widely Applicable Information Criterion* (***WAIC***) and the *Widely Applicable Bayesian Information Criterion* (***WBIC***), where the latter coincides with *BIC* when applied to regular models but is applicable even for irregular models.
  - The asymptotic form of WBIC is: $WBIC(M) = -\log\left(p_{\theta_0}(z^n)\right) + \lambda \log(n) + O_p\left(\sqrt{\log(n)}\right)$ where $\theta_0$ is the parameter value minimizing the Kullback–Leibler divergence from the model to the true underlying distribution, and $\lambda > 0$ is a rational number called the real log-canonical threshold, which can be interpreted as the *effective number of parameters (times two)*.

## 6. Frequentist Convergence of MDL and Its Implications

- In general, *MDL* procedures behave desirably (consistency and rates of convergence) under *standard frequentist assumptions*.
- ***Sec. 6.1*** (*Frequentist Convergence of MDL Estimation*) shows that the link between data compression and consistent estimation is very strong.
- ***Sec. 6.2*** (*From MDL to Lasso*) discusses an *MDL* approach *(Grünwald and Mehta 2019)* ***that can fully handle Supervised Learning***, and *also be used with large classes of loss functions including squared error (without normality assumption) and zero/one-loss.*
  - This is achieved by associating predictors $f$ with densities $p_f(x, y) \propto \exp\left(-\ell(f(x), y)\right)$, so the log-loss relative to density $p_f$ on data $(x, y)$ becomes linearly related to the loss of $f$ on $(x, y)$.
- ***Sec. 6.3*** (*Misspecification*) examines what happens when the data comes from a distribution for which *all considered models are wrong* but some are useful (lead to good predictions).
  - It turns out that the reason most MDL approaches *cannot* show convergence under misspecification is related to the ***no-hypercompression property*** (discussed by *Grunwald (2018), Safe probability*) requiring that $P_0\left(\frac{p_0(z^n)}{p(z^n)} \leq \alpha\right) \leq \alpha$, which can be achieved by replacing the $p_0$ inside the brackets by $\tilde{p}$, the distribution/density in $M$ that is closest to $P_0$ in KL-divergence, one approach being to use the generalized likelihood $p_\theta^\eta(z^n)$ for some $1 > \eta > 0$;
- ***Sec. 6.4*** (*PAC-MDL Bounds and Deep Learning*) shows that MDL provides useful intuitions about *Deep Learning* (which can have many millions of parameters), which can be validated by frequentist results.
  - The ***PAC-Bayesian Bounds*** show that generalization performance of any classifier can be directly linked to a quantity that gets smaller (a) as soon as one needs *less bits to describe the parameter* and (b) as soon as one needs *less bits to describe the data given the parameters*;
  - *Dziugaite and Roy (2017)* and *Zhou et al. (2018)* show that one can predict nontrivial *generalization* using *Deep Neural Nets* by looking at the number of bits needed to describe the parameters and applying PAC-Bayesian bounds.
  - *Neural Networks* that generalize well tend to have parameters lying in very flat minima, and *Hinton and van Camp (1993)* and *Hochreiter and Schmidhuber (1997)* pointed out that *describing weights in flat minima requires substantially lower precision, thus connecting to the MDL idea*.

## 7. Concluding Remarks

- Some additional developments are:

- *Grunwald and Mehta (2019)* provided a major step towards understanding the relation of *MDL* to other complexity notions (*Vapnik–Chervonkis Dimension*, *Entropy Numbers, Rademacher Complexity*, etc.), by showing that the *NML* complexity for models of the form $p_\theta(z) \propto \exp\left(-\eta\ LOSS_\theta(z)\right)$ can be precisely bounded in terms of the Rademacher complexity defined relative to loss.
- *Rissanen (Information and Complexity in Statistical Modeling, 1989)* is developing a different (but compatible) direction by proposing foundations of statistics in which ***no underlying "true model" is ever assumed to exist***, and expanding *MDL* and *NML* ideas in the direction of the *Kolmogorov Structure Function*
- Since 2007, numerous *MDL* and *MDL-like* applications have appeared in the literature, particularly flourishing in the field of *Data Mining*.

# Minimum Description Length Revisited

Peter Grünwald and Teemu Roos (12 March 2020)

International Journal of Mathematics for Industry Vol. 11, No. 1

**HVG Point-wise Summary (07/31/23)**

## Abstract


- Provides an up-to-date *introduction to* and *overview of* the Minimum Description Length (MDL) Principle
  - MDL is a *theory of inductive inference* that can be applied to general problems in statistics, machine learning and pattern recognition.
- While MDL was *originally based on data compression ideas*, this introduction can be read without any knowledge thereof.
  - It discusses *all major developments since 2007*, the last time an extensive overview was written.
  - These include:
    - new methods for model selection and averaging and hypothesis testing, as well as the
    - first completely general definition of MDL estimators.
- Incorporating these developments, MDL can be seen as a powerful extension of both penalized likelihood and Bayesian approaches, in which
  - penalization functions and prior distributions are replaced by more general luckiness functions,
  - average-case methodology is replaced by a more robust worst-case approach, and in which
  - methods classically viewed as highly distinct, such as AIC versus BIC and cross-validation versus Bayes can, to a large extent, be viewed from a unified perspective.


## 1. Introduction

- The ***MDL Principle is a theory of inductive inference*** that can be applied to general problems in *statistics, machine learning* and *pattern recognition*.
- Broadly speaking, it states that ***the best explanation for a given set of data is provided by the shortest description of that data***.
- In 2007, one of us published the book *The Minimum Description Length Principle* (Ref. 4 referred to as G07 from now on), giving a detailed account of most works in the MDL area that had been done until then.
- During the last 10 years, several new practical MDL methods have been designed, and there have been exciting theoretical developments as well.
- It therefore seemed time to present an up-to-date combined introduction and review.

**Why Read this Overview?**

- While the MDL idea has been shown to be very powerful in theory, and there have been a fair number of successful practical implementations, ***massive deployment has been hindered by two issues***:
  - First, to apply MDL, *one needs basic knowledge of both statistics and information theory*.
    - To remedy this, we present, for the first time, the MDL Principle *without resorting to information theory*:
      - all the material can be understood without any knowledge of data compression
      - this, should make it easier for statisticians and ML researchers new to MDL.
  - Second is that many *classical MDL procedures are either computationally highly intensive* (*for example, MDL variable selection as in Example 4 below*) and hence less suited for our big data age, or they *seem to require somewhat arbitrary restrictions of parameter spaces* (*e.g., NML with* $v \equiv 1$ *as in Sec. 2*).
- Yet, over the last 10 years, there have been exciting developments — some of them very recent — which mostly resolve these issues.

- Incorporating these developments, MDL can be seen as a powerful extension of both penalized likelihood and Bayesian approaches, in which
  - penalization functions and prior distributions are replaced by more general luckiness functions,
  - average-case methodology is replaced by a more robust worst-case approach, and in which
  - methods classically viewed as highly distinct, such as AIC versus BIC and cross-validation versus Bayes can, to some extent, be viewed from a unified perspective;
- As such, this paper should also be of interest to researchers working on the foundations of statistics and machine learning.

**History of the Field, Recent Advances and Overview of this Paper**

- ***MDL*** was introduced in 1978
  - *Jorma Rissanen* (*Modeling by the Shortest Data Description*). The paper:
    - *coined the term* ***MDL*** and
    - *introduced and analyzed the two-part code for parametric models*.
  - The *two-part code is the simplest instance of a universal code* (universal probability distribution), which is the cornerstone concept of MDL theory.
- ***MDL*** theory was greatly extended in the 1980s
  - Rissanen published a sequence of ground-breaking papers, several of which introduced new types of universal distributions.
- It came to full blossom in the 1990s
  - With further major contributions from, primarily, *Jorma Rissanen*, *Andrew Barron* and *Bin Yu*,
  - culminating in their overview paper[1] and the collection[2] with additional chapters by other essential contributors such as *Kenji Yamanish*i.
  - The book G07 provides a more exhaustive treatment of this early work, including discussion of *important precursors/alternatives to* ***MDL*** *such as* ***MML***[3], "*ideal*", Kolmogorov complexity-based MDL[4] and Solomonoff's theory of induction[5].
- Universal distributions are still central to ***MDL***.
  - ***Sec. 2*** introduces them in a concise yet self-contained way, including substantial underlying motivation, incorporating the extensions to and new insights into these basic building blocks that have been gathered over the last 10 years.
    - These include more general formulations of arguably *the most fundamental universal code*, the *Normalized Maximum Likelihood (**NML**) Distribution*, including faster ways to calculate it as well.
  - ***Sec. 3*** devotes a separate section to *new universal codes*, with quite pleasant properties for practical use,
    - most notably the *switch distribution* (Sec. 3.1), which can be used for model selection combining almost the best of ***AIC*** and ***BIC***; and the
    - *Reverse Information Projection* (***RIPr***)-universal code (Sec. 3.3) specially geared to hypothesis testing with composite null hypotheses, leading to several advantages over classical Neyman–Pearson tests.
- ***Sec. 4*** reviews recent developments on fast calculation of ***NML***-type distributions for model selection for graphical models (Bayesian networks and the like), leading to methods which appear to be more robust in practice than the standard Bayesian ones.

---

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

[2] P. D. Grünwald, I. J. Myung and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications* (The MIT Press, 2005).

[3] C. S. Wallace and D. M. Boulton, *An information measure for classification*, Comput. J. 11 (1968) 185–195.

[4] P. M. B. Vitanyi and M. Li, *Minimum description length induction, Bayesianism, and Kolmogorov complexity*, IEEE Trans. Inform. Theory IT-46(2) (2000) 446–464.

[5] T. F. Sterkenburg, *Universal prediction: A philosophical investigation*, Ph.D. thesis, University of Groningen (2018).

- ***Sec. 5*** treats recent extensions of ***MDL*** theory and *practical implementations to latent variable and irregular models* are treated.
- ***Sec. 6*** reviews developments relating to consistency and convergence properties of ***MDL*** methods.
  - First, while originally ***MDL*** estimation was formulated solely in terms of discretized estimators (reflecting the fact that coding always requires discretization),
    - it has gradually become clear that a much larger class of estimators (including maximum likelihood for "*simple*" models, and, in some circumstances, the Lasso — see Example 4) can be viewed from an ***MDL*** perspective, and
    - this becomes clearest if one investigates asymptotic convergence theorems relating to ***MDL***.
  - Second, it was found that ***MDL*** (and ***Bayes***), without modification, can behave sub-optimally under misspecification, *i.e., when all models under consideration are wrong, but some are useful* — see ***Sec. 6.3***.
  - Third, very recently, it was shown how *some of the surprising phenomena underlying the deep learning revolution in machine learning* can be explained from an ***MDL***-related perspective;
    - we briefly review these developments in ***Sec. 6.4***.
  - Finally, we note that G07 presented many explicit open problems, most of which have been resolved:
    - we mention throughout the text whenever a new development solved an old open problem, deferring some of the most technical issues to the Appendix.

**Notational preliminaries**

- ***We shall mainly be concerned with statistical models*** (families of probability distributions) of the form $M = \{p_\theta : \theta \in \Theta\}$ parametrized by some set $\Theta$ which is usually but not always a subset of Euclidean space; and ***families of models*** $\{M_\gamma : \gamma \in \Gamma\}$, where each $M_\gamma = \{p_\theta : \theta \in \Theta_\gamma\}$ is a statistical model, used to model the data $z^n := (z_1, \ldots z_n)$ with each $z_i \in Z$, for some outcome space $Z$.
  - Each $p_\theta$ represents a probability density function (pdf) or probability mass function, defined on sequences of arbitrary length.
    - With slight abuse of notation, we also denote the corresponding probability distribution by $p_\theta$ (rather than the more common $P_\theta$).
  - In the simple case that the data are i.i.d. according to each $p_\theta$ under consideration, we have $p_\theta(z^n) = \prod_{i=1}^n p_\theta(z_i)$
- We denote the maximum likelihood (ML) estimator given the model $M = \{p_\theta : \theta \in \Theta\}$ by $\hat{\theta}_{ML}$, whenever it exists and is unique; the ML estimator relative to model $M_\gamma$ is denoted by $\hat{\theta}_{ML|\gamma}$.
  - We shall, purely for simplicity, generally assume its existence and uniqueness, although nearly all results can be generalized to the case where it does not.
  - We use $\breve{\theta}$ to denote more general estimators, and $\hat{\theta}_v$ to denote what we call the MDL estimator with *luckiness function* $v$, see (5).

## 2. The Fundamental Concept: Universal Modeling

- ***MDL*** is best explained by starting with one of its prime applications, *model comparison*
  - we will generalize to prediction and estimation later, in Secs. 2.3 and 2.4.
- **Assume then that we are given a finite or countably infinite collection of statistical models** $M_1; M_2; \ldots$ ; each consisting of a set of probability distributions.
  - The fundamental idea of MDL is to associate each $M_\gamma$ with a single distribution $\bar{p}_\gamma$, often called a universal distribution relative to $M_\gamma$.
- We call the minus-log-likelihood $-log\left(\bar{p}_\gamma(Z^n)\right)$ the *code length of data* $Z^n$ *under universal code* $\bar{p}_\gamma$.
  - This terminology, and how MDL is related to coding (lossless compression of data), is briefly reviewed in ***Secs. 2.3 and 2.4***;
    - *but a crucial observation at this point is that the main **MDL** ideas can be understood abstractly, without resorting to the code length interpretation.*

- We also equip the model indices $\Gamma := \{1,2,\ldots,\gamma_{max}\}$ (where we allow $|\Gamma| = \gamma_{max} = \infty$) with a distribution, say $\pi$;
  - if the number of models to be compared is small (e.g., bounded independently of $n$ or at most a small polynomial in $n$), we can take $\pi$ to be uniform distribution
  - for large (exponential in $n$) and infinite $\Gamma$, see ***Sec. 2.3 and Example 4***.
- We then take, as our best explanation of the given data $z^n$, the model $M_\gamma$ minimizing:

$$-log(\pi(\gamma)) - log\left(\bar{p}_\gamma(z^n)\right) \qquad (1)$$

  - or, equivalently, we maximize $\pi(\gamma)\bar{p}_\gamma(z^n)$;
  - when $\pi$ is uniform this simply amounts to picking the $\gamma$ maximizing $\bar{p}_\gamma(z^n)$.
  - ***Eq. (1)*** will later be generalized to $\pi$ that are not distributions but rather more general "*luckiness functions*" — see ***Sec. 2.3***.

**(1) The Bayesian Universal Distribution**

- The reader may recognize this as being formally equivalent to the standard Bayesian way of model selection, the *Bayes factor method* as long as the $\gamma$ are defined as Bayesian marginal distributions, i.e., for each $\gamma$, we set $\bar{p}_\gamma = p_{w_\gamma}^{BAYES}$, where

$$p_{w_\gamma}^{BAYES}(z^n) = \int p_\theta(z^n) w_\gamma(\theta) d\theta \qquad (2)$$

  - for some prior probability density $w_\gamma$ on the parameters in $\Theta_\gamma$, which has to be supplied by the user.
  - When $w_\gamma$ is clear from the context, we shall write $p_\gamma^{BAYES}$ rather than $p_{w_\gamma}^{BAYES}$.
- Using Bayesian marginal distributions $\bar{p}^{BAYES}$ is indeed one possible way to instantiate ***MDL*** model selection, but it is not the only way:
  - ***MDL*** can also be based on other distributions such as $\bar{p}^{NML} = p_v^{NML}$ (depending on a function $v$), $\bar{p}^{PREQ} = p_{\breve{\theta}}^{PREQ}$ (depending on an estimator $\breve{\theta}$) and others;
    - in general, we add a bar to such distributions if the "*parameter*" $w, v$ *or* $\breve{\theta}$ is clear from the context.
- Before we continue with these other instantiations of $\bar{p}_\gamma$ we proceed with an example.

**Example 1 (Bernoulli).**

- Let $M = \{p_\theta : \theta \in [0,1]\}$ represent the *Bernoulli* model, extended to $n$ outcomes by independence.
- We then have for each $z^n \in \{0,1\}^n$ that $p_\theta(z^n) = \theta^{n_1}(1-\theta)^{n_0}$ where $n_1 = \sum_{i=1}^n z_i$ and $n_0 = n - n_1$.
- Most standard prior distributions one encounters in the literature are beta priors, for which $w(\theta) \propto (z^n) = \theta^\alpha(1-\theta)^\beta$, so that $p_w^{BAYES}(z^n) \propto \int \theta^{n_1+\alpha}(1-\theta)^{n_0+\beta} d\theta$.
- Note that $p_w^{BAYES}(z^n)$ is not itself an element of the Bernoulli model.
- One could use $p_w^{BAYES}$ to compare the Bernoulli model, via (1), to, for example, a first-order Markov model, with Bayesian marginal likelihoods defined analogously.
- We shall say a lot more about the choice of prior below.

**Example 2 (Gauss and General Improper Priors).**

- A second example is the Gaussian location family $M_{GAUSS}$ with fixed variance (say 1), in which $Z = R$ and $p_\theta(z^n) \propto \exp \sum_{i=1}^n (z_i - \theta)^2/2$.
- A standard prior for such a model is the uniform prior, $w(\theta) = 1$, which is improper (it does not integrate, hence does not define a probability distribution).
- Improper priors cannot be directly used in (2), and hence they cannot be directly used for model comparison as in (1) either.
- Still, we can use them in an indirect manner, as long as we are guaranteed that, for all $M_\gamma$ under consideration, after some initial number of $m$ observations, the Bayesian posterior $w_\gamma(\theta|z^m)$ is proper.

- We can then replace $p_{w_\gamma}^{BAYES}(z^n)$ in (2) by $p_{w_\gamma}^{BAYES}(z_{m+1},\dots,z_n|z^m) := \int p_\theta(z_{m+1},\dots,z_n|z^n)w_\gamma(\theta|z^m)d\theta$.
- We extend all these *conditional universal distributions* to distributions on $Z^n$ by defining $p_{w_\gamma}^{BAYES}(z_1,\dots,z_n) := p_{w_\gamma}^{BAYES}(z_{m+1},\dots,z_n|z^m)p_0(z^m)$ for some distribution $p_0$ on $Z^m$ that is taken to be the same for all models $M_\gamma$ under consideration.
- We can now use (1) again for model selection based on $p_{w_\gamma}^{BAYES}(z_1,\dots,z_n)$, where we note that the choice of $p_0$ plays no role in the minimization, which is equivalent to minimizing $-log(\pi(\gamma)) - log\left(p_{w_\gamma}^{BAYES}(z_{m+1},\dots,z_n|z^m)\right)$.
- Now comes the crux of the story, which makes MDL, in the end, quite different from Bayes:
  - defining the $\bar{p}_\gamma$ as in (2) is just *one particular way* to define an MDL universal distribution — but it is by no means the only one.
  - There are several other ways, and *some of them are sometimes preferable to the Bayesian choice*.
- Here we list the most important ones.

**(2) NML or Shtarkov[6] Distribution, and MDL Estimators**

- This is perhaps the *most fundamental universal distribution*, leading also to the definition of an MDL estimator.
- In its general form, the NML distribution and "*MDL estimators*" depend on a function $v: \Theta \to R_0^+$.
- The definition is then given by:

$$p_v^{NML}(z^n) := \frac{\max_{\theta\in\Theta} p_\theta(z^n)v(\theta)}{\int \max_{\theta\in\Theta} p_\theta(z^n)v(\theta)dz^n} \underset{=}{(if\ v\ is\ constant)} \frac{\max_{\theta\in\Theta} p_\theta(z^n)}{\int \max_{\theta\in\Theta} p_\theta(z^n)dz^n} \qquad (3)$$

  which is defined whenever the normalizing integral is finite.

- The logarithm of this integral is called the ***Model Complexity*** and is thus given by:

$$COMP(M,v) := \log \int \max_{\theta\in\Theta} p_\theta(z^n)v(\theta)\,dz^n \underset{=}{(if\ v\ is\ constant)} \log \int p_{\hat{\theta}_{ML}(z^n)}(z^n)dz^n \quad (4)$$

  - Here the *integral* is replaced by a *sum* for discrete data, and *max* is replaced by *sup* if necessary.
- This means that any function $v: \Theta \to R_0^+$ such that (4) is finite is allowed;
  - we call any such $v$ a luckiness function, a terminology we explain later.
- Note that $v$ is not necessarily a probability density — it does not have to be integrable.
- For any luckiness function $v$, we define the *MDL Estimator Based On* $v$ as

$$\hat{\theta}_v := \underset{\theta\in\Theta}{\operatorname{argmax}}\{p_\theta(z^n)v(\theta)\} = \underset{\theta\in\Theta}{\operatorname{argmin}}\{-\log p_\theta - [-\log v(\theta)]\} \qquad (5)$$

- The $v-MDL$ estimator is a *penalized ML estimator*, which coincides with the *Bayes MAP estimator* based on prior $v$ whenever $v$ is a probability density.
- **Although this has only become clear gradually over the last 10 years, *estimators of form* (5) *are the prime way of using MDL for estimation*;**
  - there is, however, a second, "*improper*" way for estimating distributions within MDL though, see ***Sec.2.4***.
- In practice, we will choose $v$ that are sufficiently smooth so that, if the number of parameters is small relative to $n$, $\hat{\theta}_v$ will usually be almost indistinguishable from the ML estimator $\hat{\theta}_{ML}$.
- $COMP$ indeed measures something one could call a "*complexity*" — this is easiest to see if $v = 1$, for then, if $M$ contains just a single distribution, we must have $COMP(M,v) = 0$, and the more distributions we add to $M$, the larger $COMP(M,v)$ gets — this is explored further in ***Sec. 2.2***.
- Now suppose we have a collection of models $M$ indexed by finite $\Gamma$ and we have specified luckiness functions $v$ on $\Theta_\Gamma$ for each $\gamma \in \Gamma$, and we pick a uniform distribution $\pi$ on $\Gamma$.
- As can be seen from the above, if we base our model choice on NML, we pick the model minimizing

$$-\log p_{\hat{\theta}_{v_\gamma(z^n)}} - \log v_\gamma\left(\hat{\theta}_{v_\gamma(z^n)}\right) + COMP(M_\gamma, v_\gamma) \qquad (6)$$

[6] Yu. M. Shtarkov, *Universal sequential coding of single messages*, Probl. Inf. Transm. 23(3) (1987) 3–17.

- over $\gamma$, where $COMP(M_\gamma, v_\gamma)$ is given by:

$$COMP(M_\gamma, v_\gamma) = \log \int \max_{\theta \in \Theta_\gamma} \left( p_\theta(z^n) v_\gamma(\theta) \right) dz^n = \log \int p_{\hat{\theta}_{v_\gamma(z^n)}}(z^n) v_\gamma \left( \hat{\theta}_{v_\gamma}(z^n) \right) dz^n \qquad (7)$$

- Thus, by (6), MDL incorporates a trade-off between goodness of fit and model complexity as measured by $COMP$.
- Although the *n-fold* integral inside $COMP$ looks daunting, Suzuki and Yamanishi[7] show that in many cases (e.g., normal, Weibull–Laplace models) it can be evaluated explicitly with appropriate choice of $v$.
- Originally, the NML distribution was defined by Shtarkov for the special case with $v \equiv 1$, leading to the rightmost definition in (3), and hence the term NML (in the modern version, perhaps "*normalized penalized ML*" would be more apt).
- This is also the version that Rissanen[8] advocated as embodying the *purest form of the MDL Principle*.
- However, the integral in (3) is ill-defined for just about every parametric model defined on unbounded outcome spaces (such as $N;\ R\ or\ R^+$ þ ), including the simple normal location family.
- Using nonuniform $v$ allows one to deal with such cases in a principled manner after all, see ***Sec. 2.5***.
- For finite outcome spaces though, $v \equiv 1$ usually "*works*", and (3) is well defined, as we illustrate for the Bernoulli model (see ***Sec. 4*** for more examples).

**Example 3 (Continuation of Example 1).**

- For the Bernoulli model, $\hat{\theta}_{ml}(z^n) = n_1/n$ and $COMP(M, v)$ as in (7) with $v \equiv 1$ can be rewritten as $\log \sum_{n_1=0}^{n} \binom{n_1}{n} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}$ , which, as we shall see in ***Sec. 2.2***, is within a constant of $\left(\frac{1}{2}\right) \log n$.
- As reviewed in that sub-section, the resulting $p_v^{NML}$ is asymptotically (essentially) indistinguishable from $p_{wJ}^{BAYES}$ where the latter is equipped with Jeffreys' prior, defined as $wJ(\theta) \propto \sqrt{I(\theta)} = \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, with $I(\theta)$ being the Fisher information at $\theta$.

**(3) The two-part (sub-)distribution**

- Here one first discretizes $\Theta$ to some countable sub-set $\ddot{\Theta}$ which one equips with a probability mass function $w$;
  - in contrast to the $v$ above, this function must sum to 1.
- One then considers:

$$p_w^{NML}(z^n) := \frac{\max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta})}{\int \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta}) dz^n} \qquad (8)$$

  - `which is just a special case of (3).
- But since $\int \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta})\, dz^n \leq \int \sum_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta}) dz^n = \sum_{\theta \in \ddot{\Theta}} w(\theta) \int p_{\ddot{\theta}}(z^n) dz^n = 1$, we can approximate $p_w^{NML}$ by the sub-distribution $p_w^{2-P} := \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta})$.
- This "*distribution*" adds or integrates to something smaller than 1.
- This can be incorporated into the general story by imagining that $p_w^{2-P}$ puts its remaining mass on a special outcome, say "$\diamond$", which in reality will never occur (while sub-distributions are thus "*allowed*", measures that add up to something larger than 1 have no place in MDL).
- The two-part distribution $p_w^{2-P}$ is historically the oldest universal distribution.
- The fact that it can be considered a special case of NML has only become fully clear very recently[9];
  - in that same paper, an even more general formulation of (3) is given that has all Bayesian, two-part and NML distributions as special cases.
- Despite its age, the two-part code is still important in practice, as we explain in ***Sec. 2.3***.

---

[7] A. Suzuki and K. Yamanishi, *Exact calculation of normalized maximum likelihood code length using Fourier analysis*, arXiv:1801.03705 [math.ST].

[8] J. Rissanen, *Fisher information and stochastic complexity*, IEEE Trans. Inform. Theory 42(1) (1996) 40–47.

[9] P.Grunwald and N.Mehta, *A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity*, in Proc. Thirtieth Conf. Algorithmic Learning Theory (ALT 2019) (2019), arXiv:1720.07732 [stat.ME].

**(4) The Prequential Plug-In Distribution[10],[11]**

- Here, one first takes any reasonable estimator $\breve{\theta}$ for the given model $M$.
- One then defines:

$$p_{\breve{\theta}}^{PREQ}(z^n) := \prod_{I=1}^{N} p_{\breve{\theta}(z^{i-1})}(z_i|z^{i-1}) \qquad (10)$$

  - where for i.i.d. models, the probability inside the product simplifies to $p_{\breve{\theta}(z^{i-1})}(z_i)$.
- For the normal location family, one could simply use the ML estimator: $\breve{\theta}(z^m) = \breve{\theta}_{ML}(z^m) = \sum_{J=1}^{m} z_j/m$.
- With discrete data though, the ML estimator should be avoided, since then one of the factors in (10) could easily become 0, making the product 0, so that the model for which $p_{\breve{\theta}}^{PREQ}$ is defined can never "win" the model selection contest even if most other factors in the product (10) are close to 1.
- Instead, one can use a slightly "*smoothed*" ML estimate (a natural choice for $\breve{\theta}$ is to take an MDL estimator for some $v$ as in (5), but this is not required).
- For example, in the Bernoulli model, one might take $\breve{\theta}(z^m) = \big(m_1 + (1/2)\big)/(m+1)$, where $m_1 = \sum_{J=1}^{m} z_j$.
- With this particular choice, $p_{\breve{\theta}}^{PREQ}$ turns out to coincide exactly with $p_{wJ}^{BAYES}$ with Jeffreys' prior $wJ$.
  - Such a precise correspondence between $\bar{p}^{PREQ}$ and $\bar{p}^{BAYES}$ is a special property of the Bernoulli and multinomial models though;
  - with other models, the two distributions can usually be made to behave similarly, but not identically.
- The rationale for using $\bar{p}^{PREQ}$ is described in ***Sec. 2.4***.
- In ***Sec. 3.2.1*** we will say a bit more about hybrids between prequential plug-in and Bayes (the *flattened leader distribution*) and between prequential and NML (*sequential NML*).
- Except for the just mentioned "*hybrids*", these first four universal distributions were all brought into MDL theory by Rissanen;
  - they are extensively treated by G07, in which one chapter is devoted to each, and to which we refer for details.
- The following two are much more recent.

**(5) The Switch Distribution $\bar{p}^{SWITCH}$ (Ref[12])**

- In a particular type of nested model selection, this universal distribution behaves arguably better than the other ones.
  - It will be treated in detail in ***Sec. 3.1***.

**(6) Universal Distributions $\bar{p}^{RIPR}$ based on the Reverse Information Projection**

- These universal distributions[13] lead to *improved error bounds and optional stopping behavior in hypothesis testing* and allow one to forge a connection with group-invariant Bayes factor methods; see ***Sec. 3.3***.

---

[10] J. Rissanen, *Universal coding, information, prediction and estimation*, IEEE Trans. Inform. Theory 30 (1984) 629–636.

[11] A. P. Dawid, *Present position and potential developments: Some personal views, statistical theory, the prequential approach*, J. R. Stat. Soc. A 147(2) (1984) 278–292.

[12] T. van Erven, P.D. Grunwald and S. de Rooij, *Catching up faster in Bayesian model selection and model averaging*, in Advances in Neural Information Processing Systems, Vol. 20 (Curran Associates, Inc. 2008), pp. 417–424.

[13] P. Grunwald, R. de Heide and W. Koolen, *Safe testing*, arXiv: 1906. 07801 [math.ST].

### *2.1. Motivation*

- *We first give a very high-level motivation* that avoids direct use of data compression arguments.
  - For readers interested in data compression, ***Sec. 2.3*** does make a high-level connection, but for more extensive material we refer to G07.
  - We do, in ***Sec. 2.4***, give a more detailed motivation in predictive terms, and, in ***Sec. 6***, we shall review mathematical results indicating that MDL methods are typically consistent and enjoy fast rates of convergence, providing an additional motivation in itself.
- Consider then models $M_\gamma$, where for simplicity we assume discrete data, and let $\hat{\theta}_{ML|\gamma}$ be the maximum likelihood estimator within $M_\gamma$.
- Define "*the fit of the model to the data*" in the standard way as the likelihood assigned to the data by the best fitting distribution within the model: $F_\gamma(z^n) := p_{\hat{\theta}_{ML|\gamma}(z^n)}(z^n)$.
  - Now if we enlarge the model $M_\gamma$, i.e., by adding several distributions to it, $F_\gamma(z^n)$ can only increase; and if we make $M_\gamma$ big enough such that for each $z^n$, it contains a distribution $p$ with $p(z^n) = 1$, we can even have $F_\gamma(z^n) = 1$ on all data.
  - If we simply picked the $\gamma$ maximizing $F_\gamma(z^n)$, we would be prone to severe overfitting.
    - For example, if models are nested, then, except for very special data, we would automatically pick the largest one.
- As we have seen, a central MDL idea is to instead associate each model $M_\gamma$ with a *single* corresponding distribution $\bar{p}_\gamma$, i.e., we set $F_\gamma(z^n) := \bar{p}_\gamma(z^n)$.
  - Then the total probability mass on all potential outcomes $z^n$ cannot be larger than 1, which makes it impossible to assign overly high fit $F_\gamma(z^n)$ to overly many data sequences:
    - no matter what distribution $\bar{p}_\gamma$ we chose, we must now have $\sum_{z^n} F_\gamma(z^n) = 1$, so a good fit on some $z^n$ necessarily implies a worse fit on others, and we will not select a model simply because it accidentally contained some distribution that fitted our data very well
    - thus, measuring fit by a distribution $\bar{p}_\gamma$ instead of $F_\gamma$ inherently prevents overfitting.
- This argument to measure fit relative to a model with a single $\bar{p}_\gamma$ is similar to *Bayesian Occam's Razor arguments*[14] used to motivate the Bayes factor;
  - the crucial difference is that we do not restrict ourselves to $\bar{p}_\gamma$ of the form $p_{w_\gamma}^{BAYES}(z^n) = \int p_\theta(z^n) w_\gamma(\theta) d\theta$;
  - inspecting the "*Bayesian*" Occam argument, there is, indeed, nothing in there which forces us to use distributions of Bayesian form.
- The next step is thus to decide *which* $\bar{p}$ are best associated with a given $M$.
  - To this end, we define the ***Fitness Ratio*** for data $z^n$ as

$$FR(\bar{p}, z^n) = \frac{\bar{p}(z^n)}{\max_{\theta \in \Theta} p_\theta(z^n) v(\theta)} \tag{11}$$

  - where $v: \Theta \to R_0^+$ is a non-negative function.
- To get a feeling for (11), it is best to first focus on the case with $v \equiv 1$; it then reduces to

$$FR(\bar{p}, z^n) = \frac{\bar{p}(z^n)}{\max_{\theta \in \Theta} p_\theta(z^n)} \tag{12}$$

  - We next postulate that a good choice for $\bar{p}$ relative to the given model is one in which $FR(\bar{p}, z^n)$ *tends to be as large as possible*.
    - The rationale is that, overfitting having already been taken care of by picking some $\bar{p}$ that is a probability measure (integrates to 1), it makes sense to take a $\bar{p}$ whose fit to data (as measured in terms of likelihood) is proportional to the fit to data of the best-fitting distribution in $M$ → whenever some distribution in the model $M$ fits the data $z^n$ well, the likelihood $\bar{p}(z^n)$ should be high as well.

[14] C. E. Rasmussen and Z. Ghahramani, *Occam's razor*, in Advances in Neural Information Proces- sing Systems, Vol. 13 (The MIT Press, 2000), pp. 294–300.

- One way to make "$FR$ *tends to be large*" precise is by *requiring it to be as large as possible in the worst-case*, i.e., we want to pick the $\bar{p}$ achieving:

$$\max_{\bar{p}} \min_{z^n} FR(\bar{p}, z^n) \qquad (13)$$

  - where the maximum is over all probability distributions over samples of length $n$.
- It turns out that this *maximin* problem has a solution if and only if the complexity $COMP(M, v)$ (4) is finite; and if it is fine, the unique solution is given by setting $\bar{p} = \bar{p}^{NML}$, with $\bar{p}^{NML}$ given by (3).
  - *The NML distribution thus has a special status as the most robust choice of universal* $\bar{p}$
    - Even though $\bar{p}$ is itself a probability distribution, it meaningfully assesses fit in the worst-case over all possible distributions, and its interpretation does not require one to assume that the model $M$ is "*true*" in any sense.
- The nicest sub-case is the one with $v(\theta) \equiv 1$, since then all distributions within the model $M$ are treated on exactly the same footing; no data or distribution is intrinsically preferred over any other one.

- Unfortunately, for most popular models with infinite $Z$, when taking $v(\theta) \equiv 1$, (13) usually has no solution since the integral $\int p_{\hat{\theta}_{ML(z^n)}}(z^n) dz^n$ diverges for such models, making the complexity $COMP(M, v)$ (4) infinite.
  - For all sufficiently "*regular*" models (curved exponential families, see below), this problem can invariably be solved by restricting $\Theta$ to a bounded subset of its own interior — one can show that the complexity (4) is finite with $v \equiv 1$, and thus (13) has a solution given by (3) if $\hat{\theta}_{ml}$ is restricted to a suitably bounded set.
  - Yet, restricting $\Theta$ to a bounded subset of itself is not satisfactory, since it is unclear where exactly to put the boundaries.
  - It is more natural to introduce a nonuniform $v$, which can invariably be chosen so that the complexity $COMP(M, v)$ (4) is finite and thus (13) has a solution — more on choosing $v$ at the end of Sec. 2.4.

- **A few remarks concerning this high-level motivation of MDL procedures are in order.**

(1) It is clear that, by requiring $FR$ to add to $1$, we will be less prone to overfitting than by setting it simply to $p_{\hat{\theta}_{ML(z^n)}}(z^n)$;
  - Whether the requirement to add (at most) to $1$, making $FR$ essentially a probability density function, is a clever way to avoid overfitting (leading to good results in practice) is not clear yet.
  - For this, we need additional arguments, which we very briefly review.
    - First, the sum-to-1 requirement *is the only choice for which the procedure can be interpreted as selecting the model which minimizes code length of the data* (the original interpretation of MDL)
    - Second, it is the only choice which has a predictive interpretation, which we review in ***Sec. 2.4*** below
    - Third, it is the only choice under which time-tested Bayesian methods fit into the picture
    - Fourth, with this choice we get desirable frequentist statistical properties such as consistency and convergence rates, see ***Sec. 6***.

(2) The motivation above only applies to the NML universal distributions. How about the other five types?
  - Originally, in the pure MDL approach mainly due to Rissanen, the NML was viewed as the optimal choice per se; other $\bar{p}$ should be used only for pragmatic reasons, such as them being easier to calculate.
    - One would then design them so as to be as close as possible to the NML distributions in terms of the fitness ratio they achieve.
  - In the following sub-section, we show that all six of them satisfy the same MDL/BIC asymptotics, meaning that their fitness ratio is never smaller than a constant factor of the NML one, either again in the worst-case over all $z^n$ or in some weaker expectation sense.
    - Thus, they are all "*kind of ok*" in a rather weak sense, and in practice one would simply revert to the one that is closest to NML and still usable in practice;

        - with the Bayesian $\bar{p}^{BAYES}$, as we shall see, one can even get arbitrarily close to NML as $n$ gets larger.
    - This classical story notwithstanding, it has become more and more apparent that, in practice, one sometimes wants or needs properties of model selection methods that are not guaranteed by NML — such as near-optimal predictions of future data or strong frequentist Type-I error guarantees.
        - This translates itself into universal codes $\bar{p}^{SWITCH}$ "p switch and $\bar{p}^{RIPR}$ that, for some special sequences, achieve much higher fitness ratio than $\bar{p}^{NML}$, while for all sequences having only very slightly smaller fitness ratio.
    - This more recent and pragmatic way of MDL is briefly reviewed in ***Secs. 3.1 and 3.3***.
  - This raises the question how we should define a universal distribution: what choices for $\bar{p}_\gamma$ are still "*universal*" (and define an MDL method) and what choices are not?
    - Informally, every distribution $\bar{p}_\gamma$ that for no $z^n \in Z^n$ has $\bar{p}_\gamma(z^n) \ll \bar{p}_\gamma^{NML}(z^n)$ is "*universal*" relative to $M_\gamma$.
    - For parametric models such as exponential families, the $\ll$ is partially formalized by requiring that at the very least, they should satisfy (14) below (G07 is much more precise on this).

(3) Third, we have not yet said how one should choose the "*luckiness function*" $v$ — and one needs to make a choice to apply MDL in practice.
  - The interpretation of $v$ is closely tied to the predictive interpretation of MDL, and hence we postpone this issue to the end of ***Sec. 2.4***.

(4) Fourth, the motivation so far is incomplete — we still need to explain why and how to incorporate the distribution $\pi$ on model index $\Gamma$.
  - This is done in ***Sec. 2.3*** below.

***2.2. Asymptotic Expansions***

***2.3. Unifying Model Selection and Estimation***

***2.4. Log-loss Prediction and Universal Distributions***

***2.5. The Luckiness Function***

## 3. Novel Universal Distributions

***3.1 The Switch Distribution and the AIC–BIC Dilemma***

***3.2. Hybrids between NML Bayes and Prequential Plug-in***

***3.3 Hypothesis Testing: Universal Distributions Based on the Reverse Information Projection***

## 4. Graphical Models

- Graphical models are a framework for representing multivariate probabilistic models in a way that encompasses a wide range of well-known model families, such as Markov chains, Markov random fields and Bayesian networks;
  - for a comprehensive overview, see Ref. 46[15].
- A key property of a graphical model is *parsimony*,
  - which can mean, for instance, a low-order Markov chain or more generally a sparse dependency graph that encodes conditional independence assumptions.
- *Choosing the right level of parsimony in graphical models is an ideal problem for MDL model selection*.
- In Bayesian network model selection, the prevailing paradigm is, unsurprisingly, the Bayesian one.
  - Especially the works of Geiger and Heckerman 47 and Heckerman et al. 48 have been extremely influential.

[15] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (The MIT Press, 2009).

- The main workhorse of this approach is the so-called Bayesian Dirichlet (BD) family of scores which is applicable in the discrete case where the variables being modeled are categorical.
  - Given a data sample, such scores assign a goodness value to each model structure.
  - Exhaustive search for the highest scoring structure is possible when the problem instance (characterized by the number of random variables) is of limited size, but heuristic search techniques such as variants of local search or "*hill-climbing*" can be used for larger problem.
- Different forms of the BD score imply different Dirichlet priors (different hyper-parameters) for the local multinomial distributions that comprise the joint distribution.
  - For example, in the commonly used BDeu score, the priors are determined by a single hyper-parameter, $\alpha$.
    - For a variable $X_i$ with $r$ distinct values and parents $P_{a_i}$ that can take $q$ possible combinations of values (configurations), the BDeu prior is $Dir(\alpha/rq, \ldots, \alpha/rq)$.
    - One of the main motivations for adopting this prior is that it leads to likelihood equivalence, i.e., *it assigns equal scores to all network structures that encode the same conditional independence assumptions*.
- In light of the fact that Bayesian model selection embodies a particular form/variation of MDL, these methods fit, at least to some extent, in the MDL framework as well.

- However, there also exist more "*pure*", non-Bayesian MDL methods for model selection in Bayesian networks; we mention Refs. 49 and 50 as early representative examples.
  - These early methods are almost invariably based on the two-part coding framework.

- More recently, several studies have proposed new model selection criteria that exploit the NML distribution.
  - One approach is a continuous relaxation of NML-type complexities proposed by Miyaguchi et al.[16] in which the model selection problem takes on a tractable Lasso-type L1-minimization form (see also Example 4).
  - *In other approaches, NML [or usually, approximations (but not relaxations) thereof] are used directly for encoding parts of the model; we now describe these latter approaches in a bit more detail*.

### *4.1. Factorized NML and variants*

- Silander et al. 52 propose the factorized NML (fNML) score for Bayesian network model selection which was designed to be decomposable, meaning that it can be expressed as a sum that includes a term for each variable in the network.
  - This property facilitates efficient search among the super-exponential number of possible model structures; see, e.g., Ref. 48.
  - The fNML score factors the joint likelihood not only in terms of the variables but also in terms of distinct configurations of the parent configurations.
  - Each factor in the product is given by a multinomial NML probability, for which a linear-time algorithm by Kontkanen and Myllymäki 53 can be used.

- A similar idea where a Bayesian network model selection criterion is constructed by piecing together multiple NML models under the multinomial model was proposed recently by Silander et al. 54
  - In the proposed quotient NML (qNML) score, the local scores corresponding to each variable in the network are defined as log-quotients of the form $\log \frac{NML_{FULL}(X_i \cup P_{a_i})}{NML_{FULL}(P_{a_i})}$, where $NML_{FULL}$ refers to an NML distribution defined by using a fully connected network to model the variable $X_i$ and its parents $P_{a_i}$ in the numerator and the same thing for the parent set $P_{a_i}$ in the denominator.
    - Technically, this amounts to collapsing the configurations of the variables into distinct values of a single categorical variable.

[16] K. Miyaguchi, S. Matsushima and K. Yamanishi, *Sparse graphical modeling via stochastic complexity*, in Proc. 2017 SIAM Int. Conf. Data Mining (SDM2017) (2017), pp. 723–731.

- Even though the resulting categorical variable may have a huge number of possible values, the linear time algorithm 53 or efficient approximations (see the next sub-section) can be used to implement the computations.
- A notable property of the $qNML$ score is that, unlike the $fNML$ score, it is likelihood- equivalent (see above).

- Eggeling et al.[17] apply similar ideas to a different model class, namely ***parsimonious Markov chains***.
  - There too, the likelihood is decomposed into factors depending on the configurations of other variables, and each part in the partitioning is modeled independently using the multinomial NML formula.
  - The authors demonstrate that the $fNML$-style criterion they propose leads to parsimonious models with good predictive accuracy for a wide range of different scenarios, whereas the corresponding Bayesian scores are sensitive to the choice of the prior hyper-parameters, which is important in the application where parsimonious Markov chains are used to model DNA binding sites.

- *In all these papers, both simulated and real-world data experiments suggest that the MDL-based criteria are quite robust with respect to the parameters in the underlying data source*.
  - In particular, the commonly used Bayesian methods (such as the BDeu criterion) that are being used as benchmarks are much more sensitive and fail when the assumed prior is a poor match to the data-generating model, whereas the *MDL methods are invariably very close to the Bayesian methods with the prior adapted to fit the data*.
  - This poses interesting questions concerning the proper choice of priors in the Bayesian paradigm.

- In fact, the prevalence of the Bayesian paradigm and the commonly used BD scores is challenged by two recent observations:
  - First, Silander et al. 52 show that the Dirichlet prior with hyper-parameters $(1/2, \ldots, 1/2)$, which is the invariant Jeffereys' prior for the multinomial model, but not likelihood-equivalent when used in the BD score, is very close to the *fNML* model and consequently, enjoys better robustness properties than the BDeu score which is the likelihood-equivalent BD score variant.
  - Second, Suzuki 57 shows that the BDeu criterion is irregular, i.e., prone to extreme overfitting behavior in situations where a deterministic relationship between one variable and a set of other variables holds in the data sample.
  - The MDL scores discussed above are regular in this respect and their robustness properties seem to be better than those of the BD scores, see Ref. 54.

### *4.2. Asymptotic expansions for graphical models*

- Asymptotic results concerning MDL-based criteria in graphical models are interesting for several reasons.
  1) For one, they lead to efficient scores that can be evaluated for thousands of different model structures.
  2) Second, asymptotic expansions can lead to insights about the relative complexity of different model structures.

- Various asymptotic forms exist for the point-wise and the expected regret depending on the model class in question.
  - For convenience we repeat the classical expansion of the *NML* (as well as the Bayesian marginal likelihood with Jeffreys' prior) regret/model complexity that applies for regular model classes $M = \{p_\theta : \theta \in \Theta\}$ for which $COMP(M, v)$ is finite with uniform $v$ (***see Sec. 2.2 above***)

$$COMP(M, v) = \frac{k}{2} \log \frac{n}{2\pi} + \int_\Theta \sqrt{|I(\theta)|}\, d\theta + O(1) \qquad (29)$$

  - where $k$ is the dimension of the model, $|I(\theta)|$ is the determinant of the Fisher information matrix at parameter $\theta$, the integral is over the parameter space, and the remainder term $O(1)$ vanishes as the sample size tends to infinity.

---

[17] R. Eggeling, T. Roos, P. Myllymäki and I. Grosse, Robust learning of inhomogeneous PMMs, in Proc. Seventeenth Int. Conf. Artificial Intelligence and Statistics (2014), pp. 229–237.

- For discrete data scenarios, by far the most interesting case is the multinomial model (extension of the Bernoulli distribution to an i.i.d. sequence of *r*-valued categorical random variables) since it is a building block of a number of MDL-criteria such as *fNML* and *qNML* (see above).
- There are many asymptotic expansions for the NML regret under the multinomial model. Probably the most useful is the one proposed by Szpankowski and Weinberger 58:
  - xxx
  - where *n* is the sample size, $\alpha = \frac{r}{n}$ and xxx
  - This simple formula is remarkably accurate over a wide range of finite values of *n* and *r* (see Ref. 54).
  - Note that the leading term is proportional to *n* (rather than *log(n)* as usual) because the formula is derived for the regime $r = \Theta(n)$ where the alphabet size grows proportionally to the sample size.
    - If *r* grows slower than *n* or not at all, the leading term tends to the classical form (29), where the leading term is $\frac{k}{2} log(n)$.
    - In practice, the approximation (30) is applicable for a wide range of $r/n$ ratios.
- *Roos, 59 and Zou and Roos 60 studied the second term in the expansion (29), namely the Fisher information integral, under Markov chains and Bayesian networks using Monte Carlo sampling techniques.*
  - ***This approach reveals systematic differences between the complexities of models even if they have the same number of parameters.***

## 5. Latent Variable and Irregular Models

- Although thus far we have highlighted exponential family and regression applications, *NML and other universal distributions can of course be used for model selection and estimation in complete generality* — and *many practical applications are in fact based on highly irregular models*.
  - Often, "*classical*" two-part distributions (based on discretized models) are used, since NML distributions often pose computational difficulties.
  - However, Yamanishi and collaborators have managed to come up with tractable approximations of NML-type distributions for some of the most important *irregular (i.e., non- exponential family)* models such as hierarchical latent variable models 61, and the related Gaussian mixture models 62,63.
- Suzuki et al. 64 provide an NML approach to nonnegative matrix factorization.
  - Two-part codes (and corresponding MDL estimators) for mixture families that come close to achieving the minimax regret were considered very recently by Miyamoto et al. 65
- When it comes to asymptotic approximations for code lengths/log-likelihoods based on NML and other universal distributions — all approximations so far (in ***Sec. 2.2***) were derived essentially assuming that the model under consideration is an exponential family.
  - Extensions to curved exponential families and generalized linear models are relatively straightforward (see G07 for details).
- For more irregular models, Watanabe has proposed the *Widely Applicable Information Criterion* (***WAIC***) and the *Widely Applicable Bayesian Information Criterion* (***WBIC***), see Refs. 66[18] and 67[19],
  - where the latter can be viewed as an asymptotic expansion of the log-likelihood based on a Bayesian universal distribution.
    - It coincides with *BIC* when applied to regular models but is applicable even for singular (irregular) models.
  - The asymptotic form of WBIC is: $WBIC(M) = -\log\left(p_{\theta_0}(z^n)\right) + \lambda \log(n) + O_p\left(\sqrt{\log(n)}\right)$

[18] S. Watanabe, *Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory*, J. Mach. Learn. Res. 11 (2010) 3571–3594.
[19] 67. S. Watanabe, *A widely applicable Bayesian information criterion*, J. Mach. Learn. Res. 14 (2013) 867–897.

- where $\theta_0$ is the parameter value minimizing the Kullback–Leibler divergence from the model to the true underlying distribution, and $\lambda > 0$ is a rational number called the real log-canonical threshold (see Ref. 67), which can be interpreted as the *effective number of parameters (times two)*.

## 6. Frequentist Convergence of MDL and Its Implications

- Rissanen first formulated the MDL Principle as — indeed — a Principle:
  - One can simply start by *assuming*, as an axiom, that modeling by data compression (or, equivalently, sequential predictive log-loss minimization) is the right thing to do.
- One can also take a more conventional, frequentist approach, and check whether MDL procedures behave desirably under standard frequentist assumptions.
  - We now review the results that show that, in general, they do — thus providing a frequentist justification of MDL ideas:
    - with some interesting caveats,
      - MDL model selection is typically consistent (the smallest model containing the true distribution is eventually chosen, with probability one) and
      - MDL prediction and estimation achieves good rates of convergence (the Hellinger distance between the estimated and the true density goes to zero, with high probability, quite fast).
  - In this section we review the most important convergence results.
    - ***Sec. 6.1*** shows that the link between data compression and consistent estimation is in fact very strong;
    - ***Sec. 6.4*** shows that, by taking MDL as a principle, one can get useful intuitions about deep questions concerning deep learning; and the intuitions can then, as a second step, be once again validated by frequentist results.
- ***Thus, let us assume, as is standard in frequentist statistics, that data are drawn from a distribution in one of the models under $M_\gamma$ under consideration.***
  - We consider *consistency* and *convergence* properties of the main MDL procedures in their main applications: *Model Selection, Prediction and Estimation*.
  - **Model selection:**
    - For model selection between a ***finite*** number of models, all universal codes mentioned here are consistent in wide generality;
      - for example, this has been explicitly proven if the data are i.i.d. and all models on the list are exponential families, but results for more complex models with dependent data have also been known for a long time; see G07 for an overview of results.
    - If the collection of models is ***countably infinite***, then results based on associating each $M_\gamma$ with $\bar{p}^{BAYES}$ have also been known for a long time;
      - such results typically hold for "*almost all*" (suitable defined) distributions in all $M_\gamma$; again, see G07 for a discussion of the (nontrivial) "*almost all*" requirement.
      - These countable-$\Gamma$ consistency results were extended to the switch distribution by van Erven et al. 29
  - **Prediction and "Improper" Estimation:**
    - As to sequential prediction (***Sec. 2.4***), the rate of convergence results are very easy to show (see Chap. 15 of G07), but these typically only demonstrate that the cumulative-log-loss prediction error of sequentially predicting with a universal distribution $\bar{p}$ behaves well as $n$ increases.
      - Thus, since the sum of prediction errors is small, say (for parametric models) of order $log(n)$, for *most* $t$ the individual cumulative prediction/estimation error at the $t^{\text{th}}$ sample point must be of order $1/t$, since $\sum_{t=1}^{n} \frac{1}{t} - log(t) = O(1)$.
    - Still, it remains an open question how to prove for individual $t$ what exactly the expected prediction error is at that specific $n$.

- Since one can view each prediction as an "*improper*" estimate (***end of Sec. 2.4***), the convergence rates of the resulting estimators, which estimate the underlying distribution based on a sample of size $t$ as $\bar{p}(Z^{t+1}|z^t)$, usually also behave well in a cumulative sense, but again it is very hard to say anything about individual $t$.

- The asymptotic expansions (15) and (16) imply that, for ***fixed parametric models*** $M_\gamma$, $\bar{p}^{BAYES}$ and $\bar{p}^{NML}$ achieve optimal cumulative prediction and estimation errors.
  - If, however, they are defined relative to a full model class $M = \bigcup_{\gamma\in\Gamma} M_\gamma$ consisting of at least two nested models, then they may fail to achieve optimal rates by a $log(n)$ factor.
  - van Erven et al. 29 show that sequential prediction/estimation based on the switch distribution achieves the minimax-optimal rates even in such cases.
  - van der Pas and Grunwald 30 show that, if only two models are compared, then the optimal obtainable rate for individual $n$ for any consistent procedure is achieved as well.

### *6.1. Frequentist convergence of MDL estimation*

- Very strong results exist concerning the convergence of MDL estimation based on an MDL estimator $\hat{\theta}_v$ as given by (5): $\hat{\theta}_v := \underset{\theta\in\Theta}{\operatorname{argmax}}\{p_\theta(z^n)v(\theta)\} = \underset{\theta\in\Theta}{\operatorname{argmin}}\{-\log p_\theta - [-\log v(\theta)]\}$
- A first, classical result was already stated by the ground-breaking work 68, establishing that consistency and good convergence rates can be obtained for the special case of a two- ……
- ……

### *6.2. From MDL to Lasso*

……

**Supervised Machine Learning**

- Importantly, all the works mentioned here except Ref. 13 cannot show convergence under mis-specification — for example, when applied to the Lasso, they would require an assumption of normal noise (corresponding to the squared error used in the Lasso fit, which is equivalent to the log-loss under a normal distribution for the noise).
  - In practice though, the Lasso (with the squared error) is often used in cases in which one cannot assume normally distributed errors.
  - Reference 13 contains results that can still be used in such cases [although the formula for $COMP_\eta(M, v)$ changes], based on ideas which we sketch in the following sub-section.
- More generally, one of the major areas within ***Machine Learning*** is ***Supervised Learning*** in which one assumes that data $(X_1; Y_1);\ (X_2; Y_2);\ \ldots$ are i.i.d. $\sim P_0$,with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$,and one aims to use the data to learn a predictor function $f: \mathcal{X} \to \mathcal{Y}'$ that has small expected loss or risk, defined as $E_{(X;Y)\sim P}[\ell(f(X), Y)]$, where $\ell: \mathcal{Y} \times \mathcal{Y}' \to R$ is some loss function of interest and $f$ is a member of some "*predictor model*" $F$.
- For example, the statistical notion of "*regression with random design*" corresponds, in *Machine Learning*, to a *Supervised Learning Problem* with $\mathcal{Y} = \mathcal{Y}' = R$ and $\ell(y', y) = (y' - y)^2$.
  - *Early MDL convergence results do not cover this "supervised" situation: they are not equipped to handle either random design or loss functions beyond the log-loss.*
  - *Some of the more recent works mentioned above are able to handle random design but not general loss functions (for example, for Lasso-type applications they require the noise to be normally distributed).*
  - ***Reference 13*** *seems to be the first that can fully handle supervised learning scenarios*:
    - *the convergence results can be used with random design,*
    - *they can also be used with large classes of loss functions including squared error (without normality assumption) and zero/one-loss.*
  - This is achieved by associating predictors $f$ with densities $p_f(x, y) \propto \exp\big(-\ell(f(x), y)\big)$, so that the log-loss relative to density $p_f$ on data $(x, y)$ becomes linearly related to the loss of $f$ on

$(x, y)$; the analysis then proceeds via analyzing convergence of *MDL* for the densities $p_f : f \in F$ as a mis-specified probability model.

### *6.3. Misspecification*

- As beautifully explained by Rissanen[20], one of the main original motivations for *MDL*-type methods is that *they have a clear interpretation independent of whether any of the models under consideration is "true" in the sense that it generates the data*:
  - ***one chooses a model minimizing a code length** (i.e., a prediction error on unseen data), **which is meaningful and presumably might give something useful irrespective of whether the model is true (Rissanen even argues that the whole notion of a "true model" is misguided).***
  - *This model-free paradigm also leads one to define the NML distribution as minimizing prediction error in a stringent worst-case-over-all data sense [Eq. (13)] rather than a stochastic sense*.
- Nevertheless, it is of interest to see ***what happens if one samples data from a distribution for which all models under consideration are wrong***, but some are quite useful in the sense that they lead to pretty good predictions.
  - Doing this leads to rather unpleasant surprises: as first noted by Grunwald and Langford 77, *MDL* (and Bayesian inference) ***can become inconsistent***:
    - one can give examples of $\{M_\gamma : \gamma \in \Gamma\}$ with countably infinite $\Gamma$ and a "*true*" data generating distribution $P_0$ such that, when data are sampled i.i.d. from $P_0$, *MDL* will tend to select a sub-optimal model for all large $n$ →
      - while all sub-models $M_\gamma$ are wrong, one of them, $M_{\tilde{\gamma}}$ is optimal in several intuitive respects (closest in KL divergence to $P_0$, leading to best predictions under a number of loss functions), yet it will not be selected for large $n$.
  - While the models considered by Grunwald and Langford 77 were quite artificial, Grunwald and van Ommen 78 showed that the same can happen in a more natural linear regression setting;
    - moreover, they also showed that even if $\Gamma$ is finite, although then eventually *MDL* will select the best sub-model, for even relatively large $n$ it may select arbitrarily bad sub-models.
    - De Heide 79 shows that the problem also occurs with MDL and Bayesian regression with some real-world datasets.
- It turns out that the root of the problem is related to the ***no-hypercompression property*** (27)[21].
  - ***If the collection of models** $\{M_\gamma : \gamma \in \Gamma\}$ **contains the density** $p_0$* of the "*true*" distribution $P_0$, then any distribution $p \in \bigcup_{\gamma \in \Gamma} M_\gamma$ will satisfy no-hypercompression relative to the true $p_0$: $P_0\left(\frac{p_0(z^n)}{p(z^n)} \leq \alpha\right) \leq \alpha$.
    - This property underlies the proof of all MDL consistency and rate-of-convergence results, such as those by Barron and Cover 68, Zhang 69, and Grunwald and Mehta 13.
  - However, ***if the model class** $M$* does not contain the true $p_0$, then, in order to prove consistency, one needs $P_0\left(\frac{p_0(z^n)}{p(z^n)} \leq \alpha\right) \leq \alpha$ to hold with the $P_0$ outside the brackets unchanged, but the $p_0$ inside the brackets replaced by $\tilde{p}$, the distribution/density in $M$ that is closest to $P_0$ in KL-divergence (why it should be KL is explained at length by Grunwald and van Ommen 78).
    - Unfortunately, though, (33) does not necessarily hold with the $p_0$ replaced by $\tilde{p}$.
    - If it does not, *MDL* (and Bayesian methods, whose consistency relies on similar properties) may become inconsistent.
- Grunwald and van Ommen 78, based on earlier ideas in Refs. 80 and 81, propose a solution that works for *Bayesian universal distributions*:
  - it replaces the likelihoods $p_\theta(z^n)$ for every $p = p_\theta$ with $p \in M$ by the generalized likeli- hood $p_\theta^\eta(z^n)$ for some $\eta > 0$;
    - usually $\eta < 1$ — this $\eta$ has the same mathematical function as the $\eta$ appearing in (32).

---

[20] J. Rissanen, *Complexity of models*, in Complexity, Entropy and the Physics of Information, ed. W. H. Zurek (Addison-Wesley, 1991), pp. 117–125.

[21] P. Grunwald, *Safe probability*, J. Stat. Plan. Inference 195 (2018) 47–63.

- It turns out that with such a modification, if $\eta$ is chosen small enough, a version of the no-hypercompression inequality (33) holds after all.
  - References 78 and 81 also provide a method for learning $\eta$ from the data, the "*Safe Bayesian*" algorithm (note that $\eta$ cannot be learned from the data by standard *MDL* or Bayesian methods).
- The recent work of Grunwald and Mehta 13 suggests that the modification of likelihoods by exponentiating with $\eta$ should work for general *MDL* methods as well.

### *6.4. PAC-MDL Bounds and Deep Learning*

- One of the great mysteries of modern deep learning methods in machine learning is the following[22]:
  - *Deep Learning* is based on *Neural Network* models which can have many millions of parameters.
    - Although typically run on very large training samples $z^n$, $n$ is usually still so small that the data can be fit perfectly, with zero error on the training set.
    - Still, the trained models often perform very well on future test sets of data.
- How is this possible?
    - At first sight this contradicts the tenet, shared by *MDL* and just about any other method of statistics, that good generalization requires the models to be "*small*" or "*simple*" relative to the sample size
      - [small $COMP(M)$ in *MDL* analyses, small *VC* dimension or small entropy numbers in statistical learning analyses].
    - One of several explanations (which presumably all form a piece of the puzzle) is that the local minimum of the error function found by the training method is often very broad — if one moves around in parameter space near the minimum, the fit hardly changes.
  - Hochreiter and Schmidhuber 83 already observed that describing weights in sharp minima requires high precision in order to not incur non-trivial excess error on the data, whereas flat minima can be described with substantially lower precision, thus forging a connection to the *MDL* idea;
    - in fact, related ideas already appear in Ref. 84[23].
  - In these papers, the MDL Principle is used in a manner that is less direct than what was done thus far in this paper:
    - we (and, usually, Barron and Rissanen) directly hunt for the shortest description of the data.
- In contrast, the aforementioned authors simply note that, no matter how a vector of parameters for a model was obtained, if, with the obtained vector of parameters, the data can be compressed substantially, for example by coding first the parameters and then the data with the help of the parameters, then, if we believe the *MDL* Principle, with these parameters the model (network) should generalize well to future data.
- In modern practice, *Neural Networks* are often trained with *Stochastic Gradient Descent* (SGD), and it has been empirically found that networks that generalize well do tend to have parameters lying in very flat minima.
- While this use of the *MDL Principle* seems less precise than what we reviewed earlier in this paper, it can once again be given a frequentist justification, and this justification is mathematically precise after all:
  - the so-called ***PAC-Bayesian generalization bounds***[24] show that the generalization performance of any classifier can be directly linked to a quantity that gets smaller
    - (a) as soon as one needs *less bits to describe the parameter* and

---

[22] W. Zhou, V. Veitch, M. Austern, R. Adams and P. Orbanz, ***Compressibility and generalization in large-scale deep learning***, arXiv:1804.05862 [Stat.ML].

[23] G. E. Hinton and D. van Camp, ***Keeping the neural networks simple by minimizing the description length of the weights***, in Proc. Sixth Annu. Conf. Computational Learning Theory (ACM, 1993), pp. 5–13.

[24] D. McAllester, *PAC-Bayesian stochastic model selection*, Mach. Learn. 51(1) (2003) 5–21.

- (b) as soon as one needs *less bits to describe the data given the parameters*;
- Both the results and their proofs are very similar to the *MDL* convergence results by Barron and Cover 68, Zhang 69,70 and Grunwald and Mehta 13.
  - Although in general, the formulation is not as straightforward as a simple sum of the two description lengths (a) and (b), the connections between both the two-part code length and the Bayesian code length are quite strong, as was already noticed by Blum and Langford 86.
  - In particular, for discrete $\Theta$, such PAC-Bayes bounds contain a term $-log(\pi(\theta))$ which can be interpreted as the number of bits needed to encode $\theta$ using the codes based on some distribution $\pi$;
    - for general, uncountable $\Theta$, this term gets replaced by a KL-divergence term that can still be related to a code length via a so-called "*bits back argument*" pioneered by Hinton and van Camp 84.
    - Dziugaite and Roy 87 and Zhou et al. 82, inspired by earlier work by Langford and Caruana 88, indeed show that, for some real-world datasets, one can predict nontrivial generalization using deep neural nets by looking at the number of bits needed to describe the parameters and applying PAC-Bayesian bounds.

## 7. Concluding Remarks

- We have given a self-contained introduction to *MDL*, incorporating and highlighting recent developments.
- Of necessity, we had to make a choice as to what to cover in detail, and there are many things we omitted.
- We would like to end with briefly mentioning three additional developments.
  1) ***First, there has always been the question about how MDL relates to other complexity notions such as those considered in the statistical learning theory literature*** 26: Vapnik–Chervonkis dimension, entropy numbers, Rademacher complexity and so on.
     - A major step towards understanding the relation was made by Grunwald and Mehta 13 who show that for probability models with members of the form $p_\theta(z) \propto \exp(-\eta\ LOSS_\theta(z))$, where loss is an arbitrary bounded loss function, the *NML* complexity can be precisely bounded in terms of the Rademacher complexity defined relative to loss.
  2) ***Second, we should note that Rissanen's own views and research agenda have steered in a direction somewhat different from the developments we describe:***
     - Rissanen89 published *Information and Complexity in Statistical Modeling*, which proposes foundations of statistics in which no underlying "*true model*" is ever assumed to exist.
     - As Rissanen writes, "*even such a well-meaning statement as 'all models are wrong, but some are useful', is meaningless unless some model is `true'*"'.
     - Rissanen expands *MDL* and *NML* ideas in the direction of the *Kolmogorov Structure Function*, taking the idea of distinguishable distributions underlying Ref. 19[25] as the fundamental; while presumably compatible with the developments we describe here, the emphasis of this work is quite different.
- We end with a word about applications:
  - since 2007, numerous applications of *MDL* and *MDL-like* techniques have been described in the literature;
    - as discussed in ***Sec. 6.2***, highly popular methods such as Lasso and Bayes factor methods can often be seen as "*MDL-like*".

[25] I. J. Myung, V. Balasubramanian and M. A. Pitt, *Counting probability distributions: Differential geometry and model selection*, Proc. Natl. Acad. Sci. USA 97 (2000) 11170–11175.

- Even as to specific "*pure*" *MDL* applications (such as based on NML and two-part codes), the number and scope of applications are simply too large to give a succinct representative overview.
- However, there is one particular area which we would like to mention specifically, since that area had hardly seen any *MDL* applications before 2007 whereas nowadays such applications are flourishing:
  - this is the field of *Data Mining*.
    - Some representative publications are Refs. 90–92[26],[27],[28].
    - Most of this work centers on the use of two-part codes, but sometimes NML and other sophisticated universal distributions/codes are used as well 93[29].

[26] J. Vreeken, M. van Leeuwen and A. Siebes, Krimp: *Mining item sets that compress*, Data Min. Knowl. Disc. 23(1) (2011) 169–214.
[27] D. Koutra, U. Kang, J. Vreeken and C. Faloutsos, *Summarizing and understanding large graphs*, Stat. Anal. Data Min., ASA Data Sci. J. 8(3) (2015) 183–202.
[28] K. Budhathoki, J. Vreeken and J. Origo, *Causal inference by compression*, Knowl. Inf. Syst. 56(2) (2018) 285–307.
[29] N. Tatti and J. Vreeken, *Finding good item sets by packing data*, in Eighth IEEE Int. Conf. Data Mining (IEEE, 2008), pp. 588–597.