Predictive Hypothesis Identification

Marcus Hutter

RSISE @ ANU and SML @ NICTA

Canberra, ACT, 0200, Australia

marcus@hutter1.net www.hutter1.net

8 September 2008

Abstract

While statistics focusses on hypothesis testing and on estimating (properties of) the true sampling distribution, in machine learning the performance of learning algorithms on future data is the primary issue. In this paper we bridge the gap with a general principle (PHI) that identifies hypotheses with best predictive performance. This includes predictive point and interval estimation, simple and composite hypothesis testing, (mixture) model selection, and others as special cases. For concrete instantiations we will recover well-known methods, variations thereof, and new ones. PHI nicely justifies, reconciles, and blends (a reparametrization invariant variation of) MAP, ML, MDL, and moment estimation. One particular feature of PHI is that it can genuinely deal with nested hypotheses.

Contents

Keywords

parameter estimation; hypothesis testing; model selection; predictive inference; composite hypotheses; MAP versus ML; moment fitting; Bayesian statistics.

1 Introduction

Consider data D sampled from some distribution $p(D|\theta)$ with unknown $\theta \in \Omega$. The likelihood function or the posterior contain the complete statistical information of the sample. Often this information needs to be summarized or simplified for various reasons (comprehensibility, communication, storage, computational efficiency, mathematical tractability, etc.). Parameter estimation, hypothesis testing, and model (complexity) selection can all be regarded as ways of summarizing this information, albeit in different ways or context. The posterior might either be summarized by a single point $\Theta = \{\theta\}$ (e.g. ML or MAP or mean or stochastic model selection), or by a convex set $\Theta \subseteq \Omega$ (e.g. confidence or credible interval), or by a finite set of points $\Theta = \{\theta_1,...,\theta_l\}$ (mixture models) or a sample of points (particle filtering), or by the mean and covariance matrix (Gaussian approximation), or by more general density estimation, or in a few other ways [BM98, Bis06]. I have roughly sorted the methods in increasing order of complexity. This paper concentrates on set estimation, which includes (multiple) point estimation and hypothesis testing as special cases, henceforth jointly referred to as "*hypothesis identification*" (this nomenclature seems uncharged and naturally includes what we will do: estimation and testing of simple and complex hypotheses but not density estimation). We will briefly comment on generalizations beyond set estimation at the end.

Desirable properties. There are many desirable properties any hypothesis identification principle ideally should satisfy. It should

- lead to good predictions (that's what models are ultimately for),
- be broadly applicable,
- be analytically and computationally tractable,
- be defined and make sense also for non-i.i.d. and non-stationary data,
- be reparametrization and representation invariant,
- work for simple and composite hypotheses,
- work for classes containing nested and overlapping hypotheses,
- work in the estimation, testing, and model selection regime,
- reduce in special cases (approximately) to existing other methods.

Here we concentrate on the first item, and will show that the resulting principle nicely satisfies many of the other items.

The main idea. We address the problem of identifying hypotheses (parameters/models) with good *predictive performance* head on. If θ_0 is the true parameter, then $p(\boldsymbol{x}|\theta_0)$ is obviously the best prediction of the m future observations $\boldsymbol{x}$. If we don't know θ_0 but have prior belief $p(\theta)$ about its distribution, the predictive distribution $p(\boldsymbol{x}|D)$ based on the past n observations D (which averages the likelihood $p(\boldsymbol{x}|\theta)$ over θ with posterior weight $p(\theta|D)$) is by definition the best Bayesian predictor Often we cannot use full Bayes (for reasons discussed above) but predict with hypothesis $H = \{\theta \in \Theta\}$, i.e. use $p(\boldsymbol{x}|\Theta)$ as prediction. The closer $p(\boldsymbol{x}|\Theta)$ is

to $p(\boldsymbol{x}|D)$ or $p(\boldsymbol{x}|\theta_0,D)$[1] the better is H's prediction (by definition), where we can measure closeness with some distance function d. Since $\boldsymbol{x}$ and θ_0 are (assumed to be) unknown, we have to sum or average over them.

Definition 1 (Predictive Loss) *The predictive Loss/ $\widetilde{Loss}$ of Θ given D based on distance d for m future observations is*

$$\begin{aligned} Loss_d^m(\Theta,D) &:= \int d(p(\boldsymbol{x}|\Theta),p(\boldsymbol{x}|D))d\boldsymbol{x} \\ \widetilde{Loss}_d^m(\Theta,D) &:= \iint d(p(\boldsymbol{x}|\Theta),p(\boldsymbol{x}|\theta,D))\,p(\theta|D)d\boldsymbol{x}\,d\theta \end{aligned} \tag{1}$$

Predictive hypothesis identification (PHI) minimizes the losses w.r.t. some *hypothesis class* $\mathcal{H}$. Our formulation is general enough to cover point and interval estimation, simple and composite hypothesis testing, (mixture) model (complexity) selection, and others.

(Un)related work. The general idea of inference by maximizing predictive performance is not new [Gei93]. Indeed, in the context of model (complexity) selection it is prevalent in machine learning and implemented primarily by empirical cross validation procedures and variations thereof [Zuc00] or by minimizing test and/or train set (generalization) bounds; see [Lan02] and references therein. There are also a number of statistics papers on predictive inference; see [Gei93] for an overview and older references, and [BB04, MGB05] for newer references. Most of them deal with distribution free methods based on some form of cross-validation discrepancy measure, and often focus on model selection. A notable exception is MLPD [LF82], which maximizes the predictive likelihood including future observations. The full decision-theoretic setup in which a decision based on D leads to a loss depending on x, and minimizing the expected loss, has been studied extensively [BM98, Hut05], but scarcely in the context of hypothesis identification. On the natural progression of estimation→prediction→action, approximating the predictive distribution by minimizing (1) lies between traditional parameter estimation and optimal decision making. Formulation (1) is quite natural but I haven't seen it elsewhere. Indeed, besides ideological similarities the papers above bear no resemblance to this work.

Contents. The main purpose of this paper is to investigate the predictive losses above and in particular their minima, i.e. the best predictor in $\mathcal{H}$. Section 2 introduces notation, global assumptions, and illustrates PHI on a simple example. This also shows a shortcoming of MAP and ML esimtation. Section 3 formally states PHI, possible distance and loss functions, their minima, In Section 4, I study exact properties of PHI: invariances, sufficient statistics, and equivalences. Sections 5 investigates the limit $m\to\infty$ in which PHI can be related to MAP and ML. Section 6 derives large sample approximations $n\to\infty$ for which PHI reduces to sequential moment

[1] So far we tacitly assumed that given θ_0, $\boldsymbol{x}$ is independent D. For non-i.i.d. data this is generally not the case, hence the appearance of D.

fitting (SMF). The results are subsequently used for Offline PHI. Section 7 contains summary, outlook and conclusions. Throughout the paper, the Bernoulli example will illustrate the general results.

The main aim of this paper is to introduce and motivate PHI, demonstrate how it can deal with the difficult problem of selecting composite and nested hypotheses, and show how PHI reduces to known principles in certain regimes. The latter provides additional justification and support of previous principles, and clarifies their range of applicability. In general, the treatment is exemplary, not exhaustive.

2 Preliminaries

Setup. Let $D \equiv D_n \equiv (x_1,...,x_n) \equiv x_{1:n} \in \mathcal{X}^n$ be the observed *sample* with *observations* $x_i \in \mathcal{X}$ from some measurable space $\mathcal{X}$, e.g. $\mathbb{R}^{d'}$ or $\mathbb{N}$ or a subset thereof. Similarly let $\boldsymbol{x} \equiv (x_{n+1},...,x_{n+m}) \equiv x_{n+1:n+m} \in \mathcal{X}^m$ be potential *future observations*. We assume that D and $\boldsymbol{x}$ are sampled from some *probability distribution* $\mathrm{P}[\cdot|\theta]$, where $\theta \in \Omega$ is some unknown parameter. We do *not* assume independence of the $x_{i \in \mathbb{N}}$ unless otherwise stated. For simplicity of exposition we assume that the *densities* $p(D|\theta)$ w.r.t. the default (Lebesgue or counting) measure ($\int d\lambda$, $\sum_x$, written both henceforth as $\int dx$) exist.

Bayes. Similarly, we assume a prior distribution $\mathrm{P}[\Theta]$ with density $p(\theta)$ over parameters. From *prior* $p(\theta)$ and *likelihood* $p(D|\theta)$ we can compute the *posterior* $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$, where normalizer $p(D) = \int p(D|\theta)p(\theta)d\theta$. The full Bayesian approach uses parameter averaging for prediction

$$p(\boldsymbol{x}|D) \;=\; \int p(\boldsymbol{x}|\theta, D)p(\theta|D)d\theta \;=\; \frac{p(D,\boldsymbol{x})}{p(D)}$$

the so-called *predictive distribution* (or more precisely predictive density), which can be regarded as the gold standard for prediction (and there are plenty of results justifying this [BCH93, Hut05]).

Composite likelihood. Let H_θ be the *simple hypothesis* that $\boldsymbol{x}$ is sampled from $p(\boldsymbol{x}|\theta)$ and H_Θ the *composite hypothesis* that $\boldsymbol{x}$ is "sampled" from $p(\boldsymbol{x}|\Theta)$, where $\Theta \subseteq \Omega$. In the Bayesian framework, the "*composite likelihood*" $p(\boldsymbol{x}|\Theta)$ is actually well defined (for measurable Θ with $\mathrm{P}[\Theta] > 0$) as an averaged likelihood

$$p(\boldsymbol{x}|\Theta) \;=\; \int p(\boldsymbol{x}|\theta)p(\theta|\Theta)d\theta, \qquad \text{where} \quad p(\theta|\Theta) = \frac{p(\theta)}{\mathrm{P}[\Theta]} \text{ for } \theta \in \Theta \text{ and 0 else.}$$

MAP and ML. Let $\mathcal{H}$ be the (finite, countable, continuous, complete, or else) class of hypotheses H_Θ (or Θ for short) from which the "best" one shall be selected. Each $\Theta \in \mathcal{H}$ is assumed to be a measurable subset of Ω. The *maximum a posteriori* (MAP) estimator is defined as $\theta^{\mathrm{MAP}} = \mathrm{argmax}_{\theta \in \mathcal{H}} p(\theta|D)$ if $\mathcal{H}$ contains only simple hypotheses

and $\Theta^{\rm MAP}=\text{argmax}_{\Theta\in\mathcal{H}}\text{P}[\Theta|D]$ in the general case. The composite *maximum likelihood* estimator is defined as $\Theta^{\rm ML}=\text{argmax}_{\Theta\in\mathcal{H}}p(D|\Theta)$, which reduces to ordinary ML for simple hypotheses.

In order not to further clutter up the text with too much mathematical gibberish, we make the following global assumptions during informal discussions:

Global Assumption 2 *Wherever necessary, we assume that sets, spaces, and functions are measurable, densities exist w.r.t. some (Lebesgue or counting) base measure, observed events have non-zero probability, or densities conditioned on probability zero events are appropriately defined, in which case statements might hold with probability 1 only. Functions and densities are sufficiently often (continuously) differentiable, and integrals exist and exchange.*

Bernoulli Example. Consider a binary $\mathcal{X}=\{0,1\}$ i.i.d. process $p(D|\theta)=\theta^{n_1}(1-\theta)^{n_0}$ with bias $\theta\in[0,1]=\Omega$, and $n_1=x_1+...+x_n=n-n_0$ the number of observed 1s. Let us assume a uniform uniform prior $p(\theta)=1$. Here but not generally in later continuations of the example we also assume $n_0=n_1$. Consider hypothesis class $\mathcal{H}=\{H_f,H_v\}$ containing simple hypothesis $H_f=\{\theta=\frac{1}{2}\}$ meaning "fair" and composite vacuous alternative $H_v=\Omega$ meaning "don't know". It is easy to see that

$$p(D|H_v)=p(D)=\frac{n_1!n_0!}{(n+1)!} \;<\; 2^{-n}=p(D|H_f) \quad \text{for } n>1 \qquad (\text{and} = \text{else})$$

hence $\Theta^{\rm ML}=H_f$, i.e. **ML** always suggests a fair coin however weak the evidence is. On the other hand, $\text{P}[H_f|D]=0<1=\text{P}[H_v|D]$, i.e. **MAP** never suggests a fair coin however strong the evidence is.

Now consider **PHI**. Let $m_1=x_{n+1}+...x_{n+m}=m-m_0$ be the number of future 1s. The probabilities of $\boldsymbol{x}$ given H_f, H_v, and D are, respectively

$$p(\boldsymbol{x}|H_f)=2^{-m}, \quad p(\boldsymbol{x}|H_v)=\frac{m_1!m_0!}{(m+1)!}, \quad p(\boldsymbol{x}|D)=\frac{(m_1+n_1)!(n_0+m_0)!}{(n+m+1)!}\frac{(n+1)!}{n_1!n_0!} \quad (2)$$

For $\boldsymbol{m=1}$ we get $p(1|H_f)=\frac{1}{2}=p(1|H_v)$, so when concerned with predicting only one bit, both hypotheses are equally good. More generally, for an interval $\Theta=[a,b]$, compare $p(1|\Theta)=\bar{\theta}:=\frac{1}{2}(a+b)$ to the full Bayesian prediction $p(1|D)=\frac{n_1+1}{n+2}$ (Laplace's rule). Hence if $\mathcal{H}$ is a class of interval hypotheses, then PHI chooses the $\Theta\in\mathcal{H}$ whose midpoint $\bar{\theta}$ is closest to Laplace's rule, which is reasonable. The size of the interval doesn't matter, since $p(x_{n+1}|\Theta)$ is independent of it.

Things start to change for $\boldsymbol{m=2}$. The following table lists $p(\boldsymbol{x}|D)$ for some D, together with $p(\boldsymbol{x}|H_f)$ and $p(\boldsymbol{x}|H_v)$, and their prediction error $\text{Err}(H):=\text{Loss}_1^2(H,D)$

for $d(p,q)=|p-q|$ in (1)

$p(\boldsymbol{x}\|D)$	$\boldsymbol{x}=00$	$\boldsymbol{x}=01\|10$	$\boldsymbol{x}=11$	$\mathrm{Err}(H_f) \gtrless \mathrm{Err}(H_v)$	Conclusion
$D=\{\}$	1/3	1/3	1/3	$1/3 > 0$	don't know
$D=01$	3/10	4/10	3/10	$1/5 > 2/15$	don't know
$D=0101$	2/7	3/7	2/7	$1/7 < 4/21$	fair
$D=(01)^\infty$	1/4	1/2	1/4	$0 < 1/3$	fair
$p(\boldsymbol{x}\|H_f)$	1/4	1/2	1/4		
$p(\boldsymbol{x}\|H_v)$	1/3	1/3	1/3		

The last column contains the identified best predictive hypothesis. For four or more observations, PHI says "fair", otherwise "don't know".

Using (2) or our later results, one can show more generally that PHI chooses "fair" for $\boldsymbol{n} \gg \boldsymbol{m}$ and "don't know" for $\boldsymbol{m} \gg \boldsymbol{n}$. ♢

MAP versus ML versus PHI. The conclusions of the example generalize: For $\Theta_1 \subseteq \Theta_2$, we have $\mathrm{P}[\Theta_1|D] \leq \mathrm{P}[\Theta_2|D]$, i.e. MAP always chooses the less specific hypothesis H_{Θ_2}. On the other hand, we have $p(D|\theta^{\mathrm{ML}}) \geq p(D|\Theta)$, since the maximum can never be smaller than an average, i.e. composite ML prefers the maximally specific hypothesis. So interestingly, although MAP and ML give identical answers for uniform prior on simple hypotheses, their naive extension to composite hypotheses is diametral. While MAP is risk averse finding a likely true model of low predictive power, composite ML risks an (over)precise prediction. Sure, there are ways to make MAP and ML work for nested hypotheses. The Bernoulli example has also shown that PHI's answer depends not only on the past data size n but also on the future data size m. Indeed, if we make only few predictions based on a lot of data ($m \ll n$), a point estimation (H_f) is typically sufficient, since there will not be enough future observations to detect any discrepancy. On the other hand, if $m \gg n$, selecting a vacuous model (H_v) that ignores past data is better than selecting a potentially wrong parameter, since there is plenty of future data to learn from. This is exactly the behavior PHI exhibited in the example.

3 Predictive Hypothesis Identification Principle

We already have defined the predictive loss functions in (1). We now formally state our predictive hypothesis identification (PHI) principle, discuss possible distances d, and major prediction scenarios related to the choice of m.

Distance functions. Throughout this work we assume that d is continuous and zero if and only if both arguments coincide. Some popular distances are: the (f) f-divergence $d(p,q)=f(p/q)q$ for convex f with $f(1)=0$, the (α) α-distance $f(t)=|t^\alpha-1|^{1/\alpha}$, the (1) absolute deviation $d(p,q)=|p-q|$ ($\alpha=1$), the (h) Hellinger distance $d(p,q)=(\sqrt{p}-\sqrt{q})^2$ ($\alpha=\frac{1}{2}$), the (c) chi-square distance $f(t)=(t-1)^2$, the (k) KL-divergence $f(t)=t\ln t$, and the (r) reverse KL-divergence $f(t)=-\ln t$. The only

distance considered here that is not an f divergence is the (2) squared distance $d(p,q)=(p-q)^2$. The f-divergence is particularly interesting, since it contains most of the standard distances and makes Loss representation invariant (RI).

Definition 3 (Predictive hypothesis identification (PHI)) *The best ($\widetilde{best}$) predictive hypothesis in $\mathcal{H}$ given D is defined as*

$$\hat{\Theta}_d^m := \arg\min_{\Theta\in\mathcal{H}} Loss_d^m(\Theta, D) \qquad (\tilde{\Theta}_d^m := \arg\min_{\Theta\in\mathcal{H}} \widetilde{Loss}_d^m(\Theta, D))$$

The PHI ($\widetilde{PHI}$) principle states to predict $\boldsymbol{x}$ with probability $p(\boldsymbol{x}|\hat{\Theta}_d^m)$ ($p(\boldsymbol{x}|\tilde{\Theta}_d^m)$), which we call PHI_d^m ($\widetilde{PHI}_d^m$) prediction.

Prediction modes. There exist a few distinct prediction scenarios and modes. Here are prototypes of the presumably most important ones: ***Infinite batch:*** Assume we summarize our data D by a model/hypothesis $\Theta\in\mathcal{H}$. The model is henceforth used as background knowledge for predicting and learning from further observations essentially indefinitely. This corresponds to $m\to\infty$. ***Finite batch:*** Assume the scenario above, but terminate after m predictions for whatever reason. This corresponds to a finite m (often large). ***Offline:*** The selected model Θ is used for predicting x_{k+1} for $k=n,...,n+m-1$ separately with $p(x_{k+1}|\Theta)$ without further learning from $x_{n+1}...x_k$ taking place. This corresponds to repeated $m=1$ with common Θ: $\text{Loss}_d^{1m}(\Theta,D):=\mathbf{E}[\sum_{k=n}^{n+m-1}\text{Loss}_d^1(\Theta,D_k)|D]$. ***Online:*** At every step $k=n,...,n+m-1$ we determine a (good) hypothesis Θ_k from $\mathcal{H}$ based on past data D_k, and use it only once for predicting x_{k+1}. Then for $k+1$ we select a new hypothesis etc. This corresponds to repeated $m=1$ with different Θ: $\text{Loss}=\sum_{k=n}^{n+m-1}\text{Loss}_d^1(\Theta_k,D_k)$.

The above list is not exhaustive. Other prediction scenarios are definitely possible. In all prediction scenarios above we can use $\widetilde{\text{Loss}}$ instead of Loss equally well. Since all time steps k in Online PHI are completely independent, online PHI reduces to 1-Batch PHI, hence will not be discussed any further.

4 Exact Properties of PHI

Reparametrization and representation invariance (RI). An important sanity check of any statistical procedure is its behavior under reparametrization $\theta\rightsquigarrow\vartheta=g(\theta)$ [KW96] and/or when changing the representation of observations $x_i\rightsquigarrow y_i=h(x_i)$ [Wal96], where g and h are bijections. If the parametrization/representation is judged irrelevant to the problem, any inference should also be independent of it. MAP and ML are both representation invariant, but (for point estimation) only ML is reparametrization invariant.

Proposition 4 (Invariance of Loss) *$Loss_d^m(\Theta,D)$ and $\widetilde{Loss}_d^m(\Theta,D)$ are invariant under reparametrization of Ω. If distance d is an f-divergence, then they are also independent of the representation of the observation space $\mathcal{X}$. For continuous $\mathcal{X}$, the transformations are assumed to be continuously differentiable.*

RI for Loss_f^m is obvious, but will see later some interesting consequences. Any exact inference or any specialized form of PHI_f will inherit RI. Similarly for approximations, as long as they do not break RI. For instance, PHI_h will lead to an interesting RI variation of MAP.

Sufficient statistic. For large m, the integral in Definition 1 is prohibitive. Many models (the whole exponential family) possess a sufficient statistic which allows us to reduce the integral over $\mathcal{X}^m$ to an integral over the sufficient statistic. Let

$$T : \mathcal{X}^m \to \mathbb{R}^{d'} \text{ be a sufficient statistic, i.e. } p(\boldsymbol{x}|T(\boldsymbol{x}),\theta) = p(\boldsymbol{x}|T(\boldsymbol{x})) \;\forall \boldsymbol{x},\theta \tag{3}$$

which implies that there exist functions g and h such that the likelihood factorizes into

$$p(\boldsymbol{x}|\theta) \;=\; h(\boldsymbol{x})g(T(\boldsymbol{x})|\theta) \tag{4}$$

The proof is trivial for discrete $\mathcal{X}$ (choose $h(\boldsymbol{x}) = p(\boldsymbol{x}|T(\boldsymbol{x}))$ and $g(t|\theta) = p(t|\theta) := \mathrm{P}[T(\boldsymbol{x}) = t|\theta]$) and follows from Fisher's factorization theorem for continuous $\mathcal{X}$. Let A be an event that is independent $\boldsymbol{x}$ given θ. Then multiplying (4) by $p(\theta|A)$ and integrating over θ yields

$$p(\boldsymbol{x}|A) \;=\; \int p(\boldsymbol{x}|\theta)p(\theta|A)d\theta \;=\; h(\boldsymbol{x})g(T(\boldsymbol{x})|A), \quad \text{where} \tag{5}$$

$$g(t|A) \;:=\; \int g(t|\theta)p(\theta|A)d\theta \tag{6}$$

For some $\beta \in \mathbb{R}$ let (non-probability) measure $\mu_\beta[B] := \int_{\{\boldsymbol{x}:T(\boldsymbol{x})\in B\}} h(\boldsymbol{x})^\beta d\boldsymbol{x}$ $(B \subseteq \mathbb{R}^{d'})$ have density $h_\beta(t)$ $(t \in \mathbb{R}^{d'})$ w.r.t. to (Lebesgue or counting) base measure dt ($\int dt = \sum_t$ in the discrete case). Informally,

$$h_\beta(t) \;:=\; \int h(\boldsymbol{x})^\beta \delta(T(\boldsymbol{x}) - t) d\boldsymbol{x} \tag{7}$$

where δ is the Dirac delta for continuous $\mathcal{X}$ (or the Kronecker delta for countable $\mathcal{X}$, i.e. $\int d\boldsymbol{x}\, \delta(T(\boldsymbol{x}) - t) = \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t}$).

Theorem 5 (PHI for sufficient statistic) *Let $T(\boldsymbol{x})$ be a sufficient statistic (3) for θ and assume $\boldsymbol{x}$ is independent D given θ, i.e. $p(\boldsymbol{x}|\theta,D) = p(\boldsymbol{x}|\theta)$. Then*

$$\begin{aligned} \mathit{Loss}_d^m(\Theta, D) &= \int d(g(t|\Theta), g(t|D))h_\beta(t)dt \\ \widetilde{\mathit{Loss}}_d^m(\Theta, D) &= \int d(g(t|\Theta), g(t|\theta))h_\beta(t)p(\theta|D)dtd\theta \end{aligned}$$

holds (where g and h_β have been defined in (4), (6), and (7)), provided one (or both) of the following conditions hold: (i) distance d scales with a power $\beta \in \mathbb{R}$, i.e. $d(\sigma p, \sigma q) = \sigma^\beta d(p,q)$ for $\sigma > 0$, or (ii) any distance d, but $h(\boldsymbol{x}) \equiv 1$ in (4). One can choose $g(t|\cdot) = p(t|\cdot)$, the probability density of t, in which case $h_1(t) \equiv 1$.

All distances defined in Section 3 satisfy (i), the f-divergences all with $\beta=1$ and the square loss with $\beta=2$. The independence assumption is rather strong. In practice, usually it only holds for some n if it holds for all n. Independence of $x_{n+1:n+m}$ from D_n given θ for all n can only be satisfied for independent (not necessarily identically distributed) $x_{i\in\mathbb{N}}$.

Theorem 6 (Equivalence of $\mathbf{PHI}^m_{2|r}$ and $\widetilde{\mathbf{PHI}}^m_{2|r}$) *For square distance ($d\widehat{=}2$) and RKL distance ($d\widehat{=}r$), $Loss^m_d(\Theta,D)$ differs from $\widetilde{Loss}^m_d(\Theta,D)$ only by an additive constant $c^m_d(D)$ independent of Θ, hence PHI and $\widetilde{PHI}$ select the same hypotheses $\hat\Theta^m_2=\tilde\Theta^m_2$ and $\hat\Theta^m_r=\tilde\Theta^m_r$.*

Bernoulli Example. Let us continue with our Bernoulli example with uniform prior. $T(\boldsymbol{x})=x_1+...+x_m=m_1=t\in\{0,...,m\}$ is a sufficient statistic. Since $\mathcal{X}=\{0,1\}$ is discrete, $\int dt=\sum_{t=0}^m$ and $\int dx=\sum_{\boldsymbol{x}\in\mathcal{X}^m}$. In (4) we can choose $g(t|\theta)=p(\boldsymbol{x}|\theta)=\theta^t(1-\theta)^{m-t}$ which implies $h(\boldsymbol{x})\equiv 1$ and $h_\beta(t)=\sum_{\boldsymbol{x}:T(\boldsymbol{x})=t}1=\binom{m}{t}$. From definition (5) we see that $g(t|D)=p(\boldsymbol{x}|D)$ whose expression can be found in (2). For RKL-distance, Theorem 5 now yields $\mathrm{Loss}^m_r(\Theta|D)=\sum_{t=1}^m h_\beta(t)g(t|D)\ln\frac{g(t|D)}{g(t|\Theta)}$. For a point hypothesis $\Theta=\{\theta\}$ this evaluates to a constant minus $m[\frac{n_1+1}{n+2}\ln\theta+\frac{n_1+1}{n+2}\ln(1-\theta)]$, which is minimized for $\theta=\frac{n_1+1}{n+2}$. Therefore the best predictive point $\hat\theta_r=\frac{n_1+1}{n+2}=\tilde\theta_r=$ Laplace rule, where we have used Theorem 6 in the third equality. $\diamondsuit$

5 PHI for ∞-Batch

In this section we will study PHI for large m, or more precisely, the $m\gg n$ regime. No assumption is made on the data size n, i.e. the results are exact for any n (small or large) in the limit $m\to\infty$. For simplicity and partly by necessity we assume that the $x_{i\in\mathbb{N}}$ are i.i.d. (lifting the "identical" is possible). Throughout this section we make the following assumptions.

Assumption 7 *Let $x_{i\in\mathbb{N}}$ be independent and identically distributed, $\Omega\subseteq\mathbb{R}^d$, the likelihood density $p(x_i|\theta)$ twice continuously differentiable w.r.t. θ, and the boundary of Θ has zero prior probability.*

We further define $x:=x_i$ (any i) and the partial derivative $\partial:=\partial/\partial\theta=(\partial/\partial\theta_1,...,\partial/\partial\theta_d)^\top=(\partial_1,...,\partial_d)^\top$. The (two representations of the) *Fisher information matrix* of $p(x|\theta)$

$$I_1(\theta)\;=\;\mathbf{E}[(\partial\ln p(x|\theta))(\partial\ln p(x|\theta))^\top|\theta]\;=\;-\int(\partial\partial^\top\ln p(x|\theta))p(x|\theta)dx \qquad (8)$$

will play a crucial role in this Section. It also occurs in *Jeffrey's prior*,

$$p_J(\theta)\;:=\;\sqrt{\det I_1(\theta)}/J,\qquad J\;:=\;\int\sqrt{\det I_1(\theta)}d\theta \qquad (9)$$

a popular reparametrization invariant (objective) reference prior (when it exists) [KW96]. We call the determinant (det) of $I_1(\theta)$, *Fisher information*. J can be interpreted as the intrinsic size of Ω [Grü07]. Although not essential to this work, it will be instructive to occasionally plug it into our expressions. As distance we choose the Hellinger distance.

Theorem 8 ($\widetilde{\text{Loss}}_h^m(\theta,D)$ for large m) *Under Assumption 7, for point estimation, the predictive Hellinger loss for large m is*

$$\begin{aligned} \widetilde{Loss}_h^m(\theta,D) &= 2-2\left(\frac{8\pi}{m}\right)^{d/2}\frac{p(\theta|D)}{\sqrt{\det I_1(\theta)}}[1+O(m^{-1/2})] \\ &\stackrel{J}{=} 2-2\left(\frac{8\pi}{m}\right)^{d/2}\frac{p(D|\theta)}{Jp(D)}[1+O(m^{-1/2})] \end{aligned}$$

where the first expression holds for any continuous prior density and the second expression ($\stackrel{J}{=}$) holds for Jeffrey's prior.

IMAP. The asymptotic expression shows that minimizing $\widetilde{\text{Loss}}_h^m$ is equivalent to the following maximization

$$\text{IMAP}: \qquad \tilde\theta_h^\infty \;=\; \theta^{\text{IMAP}} \;:=\; \arg\max_\theta \frac{p(\theta|D)}{\sqrt{\det I_1(\theta)}} \tag{10}$$

Without the denominator, this would just be MAP estimation. We have discussed that MAP is not reparametrization invariant, hence can be corrupted by a bad choice of parametrization. Since the square root of the Fisher information transforms like the posterior, their ratio is invariant. So PHI led us to a nice reparametrization invariant variation of MAP, immune to this problem. Invariance of the expressions in Theorem 8 is not a coincidence. It has to hold due to Proposition 4. For Jeffrey's prior (second expression in Theorem 8), minimizing $\widetilde{\text{Loss}}_h^m$ is equivalent to maximizing the likelihood, i.e. $\tilde\theta_h^\infty=\theta^{\text{ML}}$. Remember that the expressions are exact even and especially for small samples D_n. No large n approximation has been made. For small n, MAP, ML, and IMAP can lead to significantly different results. For Jeffrey's prior, IMAP and ML coincide. This is a nice reconciliation of MAP and ML: An "improved" MAP leads for Jeffrey's prior back to "simple" ML.

MDL. We can also relate PHI to MDL by taking the logarithm of the second expression in Theorem 8:

$$\tilde\theta_h^\infty \;\stackrel{J}{=}\; \arg\min_\theta\{-\log p(D|\theta)+\tfrac{d}{2}\log\tfrac{m}{8\pi}+J\} \tag{11}$$

For $m=4n$ this is the classical (large n approximation of) MDL [Grü07]. So presuming that (11) is a reasonable approximation of PHI even for $m=4n$, MDL approximately minimizes the predictive Hellinger loss *iff* used for $O(n)$ predictions. We will not expand on this, since the alluded relation to MDL stands on shaky grounds (for several reasons).

Corollary 9 ($\tilde\theta_h^\infty = \theta^{\mathbf{IMAP}} \stackrel{J}{=} \theta^{\mathbf{ML}}$) *The predictive estimator* $\tilde\theta_h^\infty = \lim_{m\to\infty}\operatorname{argmin}_\theta \widetilde{Loss}_h^m(\theta,D)$ *coincides with* θ^{IMAP}*, a representation invariant variation of MAP. In the special case of Jeffrey's prior, it also coincides with the maximum likelihood estimator* θ^{ML}*.*

Theorem 10 ($\widetilde{\mathbf{Loss}}_h^m(\Theta,D)$ for large m) *Under Assumption 7, for composite* Θ*, the predictive Hellinger loss for large* m *is*

$$\begin{aligned}\widetilde{Loss}_h^m(\Theta,D) &= 2-2\left(\frac{8\pi}{m}\right)^{d/4}\frac{1}{\sqrt{\mathrm{P}[\Theta]}}\int_\Theta p(\theta|D)\sqrt{\frac{p(\theta)}{\sqrt{\det I_1(\theta)}}}d\theta + o(m^{-d/4})\\ &\stackrel{J}{=} 2-2\left(\frac{8\pi}{m}\right)^{d/4}\sqrt{\frac{p(D|\Theta)\mathrm{P}[\Theta|D]}{J\mathrm{P}[D]}} + o(m^{-d/4})\end{aligned}$$

where the first expression holds for any continuous prior density and the second expression ($\stackrel{J}{=}$) *holds for Jeffrey's prior.*

MAP meets ML half way. The second expression in Theorem 10 is proportional to the geometric average of the posterior and the composite likelihood. For large Θ the likelihood gets small, since the average involves many wrong models. For small Θ, the posterior is proportional to the volume of Θ hence tends to zero. The product is maximal for some Θ in-between:

$$\mathrm{ML}\times\mathrm{MAP}:\ \sqrt{\tfrac{p(D|\Theta)\mathrm{P}[\Theta|D]}{\mathrm{P}[D]}} = \tfrac{\mathrm{P}[\Theta|D]}{\sqrt{P[\Theta]}} = \tfrac{p(D|\Theta)\sqrt{P[\Theta]}}{P[D]} \to \left\{\begin{array}{ccc} 1 & \text{for} & \Theta\to\Omega \\ 0 & \text{for} & \Theta\to\{\theta\} \\ O(n^{d/4}) & \text{for} & |\Theta|\sim n^{-d/2}\end{array}\right. \tag{12}$$

The regions where the posterior density $p(\theta|D)$ and where the (point) likelihood $p(D|\theta)$ are large are quite similar, as long as the prior is not extreme. Let Θ_0 be this region. It typically has diameter $O(n^{-1/2})$. Increasing $\Theta\supset\Theta_0$ cannot significantly increase $\mathrm{P}[\Theta|D]\le 1$, but significantly decreases the likelihood, hence the product gets smaller. Vice versa, decreasing $\Theta\subset\Theta_0$ cannot significantly increase $p(D|\Theta)\le p(D|\theta^{\mathrm{ML}})$, but significantly decreases the posterior. The value at Θ_0 follows from $\mathrm{P}[\Theta_0]\approx\mathrm{Volume}(\Theta_0)\approx O(n^{-d/2})$. Together this shows that Θ_0 approximately maximizes the product of likelihood and posterior. So the best predictive $\Theta_0=\tilde\Theta_h^\infty$ has diameter $O(n^{-1/2})$, which is a very reasonable answer. It covers well but not excessively the high posterior and high likelihood regions (provided $\mathcal{H}$ is sufficiently rich of course). By multiplying the likelihood or dividing the posterior with only the square root of the prior, they meet half way!

Bernoulli Example. A Bernoulli process with uniform prior and $n_0=n_1$ has posterior variance $\sigma_n^2=\frac{1}{4n}$. Hence any reasonable symmetric interval estimate $\Theta=[\frac12-z;\frac12+z]$ of θ will have size $2z=O(n^{-1/2})$. For PHI we get

$$\frac{\mathrm{P}[\Theta|D]}{\sqrt{\mathrm{P}[\Theta]}} = \frac{1}{\sqrt{2z}}\frac{(n+1)!}{n_1!n_0!}\int_\Theta\theta^{n_1}(1-\theta)^{n_0}d\theta \simeq \frac{1}{\sqrt{2z}}\mathrm{erf}\Big(\frac{z}{\sigma_n\sqrt{2}}\Big)$$

where equality $\simeq$ is a large n approximation, and $\mathrm{erf}(\cdot)$ is the error function [AS74]. $\mathrm{erf}(x)/\sqrt{x}$ has a global maximum at $x \doteq 1$ within 1% precision. Hence PHI selects an interval of half-width $z \doteq \sqrt{2}\sigma_n$.

If faced with a binary decision between point estimate $\Theta_f = \{\frac{1}{2}\}$ and vacuous estimate $\Theta_v = [0;1]$, comparing the losses in Theorems 8 and 10, we see that for large m, Θ_v is selected, despite σ_n being close to zero for large n. In Section 2 we have explained that this makes from a predictive point of view. $\diamondsuit$

Finally note that (12) does not converge to (any monotone function of) (10) for $\Theta \to \{\theta\}$, i.e. and $\tilde{\Theta}_h^\infty \not\to \tilde{\theta}_h^\infty$, since the limits $m \to \infty$ and $\Theta \to \{\theta\}$ do not exchange.

Finding $\tilde{\Theta}_h^\infty$. Contrary to MAP and ML, an unrestricted maximization of (12) over *all* measurable $\Theta \subseteq \Omega$ makes sense. The following result reduces the optimization problem to finding the level sets of the likelihood function and to a one-dimensional maximization problem.

Theorem 11 (Finding $\tilde{\Theta}_h^\infty = \Theta^{\mathbf{ML\times MAP}}$) *Let $\Theta_\gamma := \{\theta : p(D|\theta) \geq \gamma\}$ be the γ-level set of $p(D|\theta)$. If $\mathrm{P}[\Theta_\gamma]$ is continuous in γ, then*

$$\Theta^{ML\times MAP} \;:=\; \arg\max_\Theta \frac{\mathrm{P}[\Theta|D]}{\sqrt{\mathrm{P}[\Theta]}} \;=\; \arg\max_{\Theta_\gamma:\gamma\geq 0} \frac{\mathrm{P}[\Theta_\gamma|D]}{\sqrt{\mathrm{P}[\Theta_\gamma]}}$$

More precisely, every global maximum of (12) differs from the maximizer Θ_γ at most on a set of measure zero.

Using posterior level sets, i.e. shortest α-credible sets/intervals instead of likelihood level sets would not work (an indirect proof is that they are not RI). For a general prior, $p(D|\theta)\sqrt{p(\theta)/I_1(\theta)}$ level sets need to be considered. The continuity assumption on $\mathrm{P}[\Theta_\gamma]$ excludes likelihoods with plateaus, which is restrictive if considering non-analytic likelihoods. The assumption can be lifted by considering all Θ_γ in-between $\Theta_\gamma^o := \{\theta : p(D|\theta) > \gamma\}$ and $\bar{\Theta}_\gamma := \{\theta : p(D|\theta) \geq \gamma\}$. Exploiting the special form of (12) one can show that the maximum is attained for either Θ_γ^o or $\bar{\Theta}_\gamma$ with γ obtained as in the theorem.

Large n. For large n ($m \gg n \gg 1$), the likelihood usually tends to an (un-normalized) Gaussian with mean=mode $\bar{\theta} = \theta^{\mathrm{ML}}$ and covariance matrix $[nI_1(\bar{\theta})]^{-1}$. Therefore the levels sets are ellipsoids

$$\Theta_r \;=\; \{\theta : (\theta-\bar{\theta})^\top I_1(\bar{\theta})(\theta-\bar{\theta}) \leq r^2\}$$

We know that the size r of the maximizing ellipsoid scales with $O(n^{-1/2})$. For such tiny ellipsoids, (12) is asymptotically proportional to

$$\frac{\mathrm{P}[\Theta_r|D]}{\sqrt{\mathrm{P}[\Theta_r]}} \;\propto\; \frac{\int_{\Theta_r} p(D|\theta)d\theta}{\sqrt{\mathrm{Volume}[\Theta_r]}} \;\propto\; \frac{\int_{||z||\leq\rho} \mathrm{e}^{-||z||^2/2}dz}{\sqrt{\int_{||z||\leq\rho} 1dz}} \;\propto\; \frac{\int_0^{\rho^2/2} t^{d/2-1}\mathrm{e}^{-t}dt}{\rho^{d/2}} \;=\; \frac{\gamma(d/2,\rho^2/2)}{\rho^{d/2}}$$

where $z := \sqrt{nI_1(\bar\theta)}(\theta-\bar\theta) \in \mathbb{R}^d$, and $\rho := r\sqrt{n}$, and $t := \frac{1}{2}\rho^2$, and $\gamma(\cdot,\cdot)$ is the incomplete Gamma function [AS74], and we dropped all factors that are independent of r. The expressions also holds for general prior in Theorem 8, since asymptotically the prior has no influence. They are maximized for the following $\tilde r$:

d	1	2	3	4	5	10	100	$\cdots$	∞
$\tilde r\sqrt{n/d}$	1.400	1.121	1.009	0.947	0.907	0.819	0.721	$\cdots$	$1/\sqrt{2}$

i.e. for $m \gg n \gg 1$, unrestricted PHI selects ellipsoid $\tilde\Theta_h^\infty = \Theta_{\tilde r}$ of (linear) size $O(\sqrt{d/n})$.

So far we have considered $\widetilde{\text{Loss}}_h^m$. Analogous asymptotic expressions can be derived for Loss_h^m: While Loss_h^m differs from $\widetilde{\text{Loss}}_h^m$, for point estimation their minima $\hat{\boldsymbol\theta}_{\boldsymbol d}^{\boldsymbol\infty} = \tilde{\boldsymbol\theta}_{\boldsymbol d}^{\boldsymbol\infty} = \boldsymbol\theta^{\text{IMAP}}$ coincide. For composite Θ, the answer is qualitatively similar but differs quantitatively.

6 Large Sample Approximations

In this section we will study PHI for large sample sizes n, more precisely the $n \gg m$ regime. For simplicity we concentrate on the univariate $d=1$ case only. Data may be non-i.i.d.

Sequential moment fitting (SMF). A classical approximation of the posterior density $p(\theta|D)$ is by a Gaussian with same mean and variance. In case the class of available distributions is further restricted, it is still reasonable to approximate the posterior by the distribution whose mean and variance are closest to that of $p(\theta|D)$. There might be a tradeoff between taking a distribution with good mean (low bias) or one with good variance. Often low bias is of primary importance, and variance comes second. This suggests to first fit the mean, then the variance, and possibly continue with higher order moments.

PHI is concerned with predictive performance, not with density estimation, but of course they are related. Good density estimation in general and sequential moment fitting (SMF) in particular lead to good predictions, but the converse is not necessarily true. We will indeed see that PHI for $n \to \infty$ (under certain conditions) reduces to an SMF procedure.

The SMF algorithm. In our case, the set of available distributions is given by $\{p(\theta|\Theta) : \Theta \in \mathcal{H}\}$. For some event A, let

$$\bar\theta^A := \mathbf{E}[\theta|A] = \int \theta\, p(\theta|A) d\theta \quad \text{and} \quad \mu_k^A := \mathbf{E}[(\theta - \bar\theta^A)^2|A] \quad (k \geq 2) \tag{13}$$

be the mean and central moments of $p(\theta|A)$. The posterior moments μ_k^D are known and can in principle be computed. SMF sequentially "fits" μ_k^Θ to μ_k^D: Starting with $\mathcal{H}_0 := \mathcal{H}$, let $\mathcal{H}_k \subseteq \mathcal{H}_{k-1}$ be the set of $\Theta \in \mathcal{H}_{k-1}$ that minimize $|\mu_k^\Theta - \mu_k^D|$:

$$\mathcal{H}_k := \{\arg\min_{\Theta \in \mathcal{H}_{k-1}} |\mu_k^\Theta - \mu_k^D|\}, \qquad \mathcal{H}_0 := \mathcal{H}, \qquad \mu_1^A := \bar\theta^A$$

Let $k^* := \min\{k : \mu_k^\Theta \neq \mu_k^D, \Theta \in \mathcal{H}_k\}$ be the smallest k for which there is no perfect fit anymore (or ∞ otherwise). Under some quite general conditions, in a certain sense, all and only the $\Theta \in \mathcal{H}_{k^*}$ minimize $\text{Loss}_d^m(\Theta, D_n)$ for large n.

Theorem 12 (PHI for large n by SMF) *For some $k \leq k^*$, assume $p(\boldsymbol{x}|\theta)$ is k times continuously differentiable w.r.t. θ at the posterior mean $\bar{\theta}^D$. Let $\beta > 0$ and assume $\sup_\theta \int |p^{(k)}(\boldsymbol{x}|\theta)|^\beta d\theta < \infty$, $\mu_k^D = O(n^{-k/2})$, $\mu_k^\Theta = O(n^{-k/2})$, and $d(p,q)/|p-q|^\beta$ is a bounded function. Then*

$$Loss_d^m(\Theta, D) \;=\; O(n^{-k\beta/2}) \quad \forall \Theta \in \mathcal{H}_k \quad (k \leq k^*)$$

For the $\alpha \leq 1$ distances we have $\beta = 1$, for the square distance we have $\beta = 2$ (see Section 3). For i.i.d. distributions with finite moments, the assumption $\mu_k^D = O(n^{-k/2})$ is virtually nil. Normally, no $\Theta \in \mathcal{H}$ has better loss order than $O(n^{-k^*\beta/2})$, i.e. $\mathcal{H}_{k^*}$ can be regarded as the set of all asymptotically optimal predictors. In many cases, $\mathcal{H}_{k^*}$ contains only a single element. Note that $\mathcal{H}_{k^*}$ does neither depend on m, nor on the chosen distance d, i.e. the best predictive hypothesis $\hat{\Theta} = \hat{\Theta}_d^m$ is essentially the same for all m and d if n is large.

Bernoulli Example. In the Bernoulli Example in Section 2 we considered a binary decision between point estimate $\Theta_f = \{\frac{1}{2}\}$ and vacuous estimate $\Theta_v = [0;1]$, i.e. $\mathcal{H}_0 = \{\Theta_f, \Theta_v\}$. For $n_0 = n_1$ we have $\bar{\theta}^{[0;1]} = \bar{\theta}^{1/2} = \frac{1}{2} = \bar{\theta}^D$, i.e. both fit the first moment exactly, hence $\mathcal{H}_1 = \mathcal{H}_0$. For the second moments we have $\mu_2^D = \frac{1}{4n}$, but $\mu_2^{[0;1]} = \frac{1}{12}$ and $\mu_2^{1/2} = 0$, hence for large n the point estimate matches the posterior variance better, so $\hat{\Theta} = \{\frac{1}{2}\} \in \mathcal{H}_2 = \{\Theta_f\}$, which makes sense. $\diamondsuit$

For unrestricted (single) point estimation, i.e. $\mathcal{H} = \{\{\theta\}, \theta \in \mathbb{R}\}$, one can typically estimate the mean exactly but no higher moments. More generally, finite mixture models $\Theta = \{\theta_1, ..., \theta_l\}$ with l components (degree of freedoms) can fit at most l moments. For large l, the number of $\theta_i \in \hat{\Theta}$ that lie in a small neighborhood of some θ (i.e. the "density" of points in $\hat{\Theta}$ at θ) will be proportional to the likelihood $p(D|\theta)$. Countably infinite and even more so continuous models if otherwise unrestricted are sufficient to get all moments right. If the parameter range is restricted, anything can happen ($k^* = \infty$ or $k^* < \infty$). For interval estimation $\mathcal{H} = \{[a;b] : a, b \in \mathbb{R}, a \leq b\}$ and uniform prior, we have $\bar{\theta}^{[a;b]} = \frac{1}{2}(a+b)$ and $\mu_2^{[a;b]} = \frac{1}{12}(b-a)^2$, hence the first two moments can be fitted exactly and the SMF algorithm yields the unique asymptotic solution $\hat{\Theta} = [\bar{\theta}^D - \sqrt{3}\mu_2^D \,; \bar{\theta}^D + \sqrt{3}\mu_2^D]$. In higher dimensions, common choices of $\mathcal{H}$ are convex sets, ellipsoids, and hypercubes. For ellipsoids, the mean and covariance matrix can be fitted exactly and uniquely similarly to 1d interval estimation. While SMF can be continued beyond k^*, $\mathcal{H}_k$ typically does *not* contain $\hat{\Theta}$ for $k > k^*$ anymore. The correct continuation beyond k^* is either $\mathcal{H}_{k+1} = \{\text{argmin}_{\Theta \in \mathcal{H}_k} \mu_k^\Theta\}$ or $\mathcal{H}_{k+1} = \{\text{argmax}_{\Theta \in \mathcal{H}_k} \mu_k^\Theta\}$ (there is some criterion for the choice), but apart from exotic situations this does not improve the order $O(n^{-k^*\beta/2})$ of the loss, and usually $|\mathcal{H}_{k^*}| = 1$ anyway.

Exploiting Theorem 6, we see that SMF is also applicable for $\widetilde{\text{Loss}}_2^m$ and $\widetilde{\text{Loss}}_r^m$. Luckily, Offline $\widetilde{\text{PHI}}$ can also be reduced to 1-Batch $\widetilde{\text{PHI}}$:

Proposition 13 (Offline = 1-Batch) *If $x_{i\in\mathbb{N}}$ are i.i.d., the Offline $\widetilde{Loss}$ is proportional to the 1-Batch $\widetilde{Loss}$:*

$$\widetilde{Loss}_d^{1m}(\Theta, D) \;:=\; \sum_{k=n}^{n+m-1} \int \widetilde{Loss}_d^{1}(\Theta, D_k)\, p(x_{n+1:k}|D) dx_{n+1:k} \;=\; m\, \widetilde{Loss}_d^{1}(\Theta, D)$$

In particular, Offline $\widetilde{PHI}$ equals 1-Batch $\widetilde{PHI}$: $\tilde{\Theta}_d^{1m} = \tilde{\Theta}_d^1$.

Exploiting Theorem 6, we see that also $\text{Loss}_{\mathbf{2|r}}^{\mathbf{1m}} = \boldsymbol{m}\text{Loss}_{\mathbf{2|r}}^{\boldsymbol{m}}$+constant. Hence we can apply SMF also for Offline $\text{PHI}_{2|r}$ and $\widetilde{\text{PHI}}_{2|r}$. For square loss, i.i.d. is not essential, independence is sufficient.

7 Discussion

Summary. If prediction is the goal, but full Bayes not feasible, one should *identify* (estimate/test/select) the *hypothesis* (parameter/model/interval) that *predicts* best. What best is can depend on the problem setup: What our benchmark is (Loss, $\widetilde{\text{Loss}}$), the distance function we use for comparison (d), how long we use the model (m) compared to how much data we have at hand (n), and whether we continue to learn or not (Batch,Offline). We have defined some reparametrization and representation invariant losses that cover many practical scenarios. Predictive hypothesis identification (PHI) aims at minimizing this loss. For $m\to\infty$, PHI overcomes some problems of and even reconciles (a variation of) MAP and (composite) ML. Asymptotically, for $n\to\infty$, PHI reduces to a sequential moment fitting (SMF) procedure, which is independent of m and d. The primary purpose of the asymptotic approximations was to gain understanding (e.g. consistency of PHI follows from it), without supposing that they are the most relevant in practice. A case where PHI can be evaluated efficiently exactly is when a sufficient statistic is available.

Outlook. There are many open ends and possible extensions that deserve further study. Some results have only been proven for specific distance functions. For instance, we conjecture that PHI reduces to IMAP for other d (seems true for α-distances). Definitely the behavior of PHI should next be studied for semi-parametric models and compared to existing model (complexity) selectors like AIC, LoRP [Hut07], BIC, and MDL [Grü07], and cross validation in the supervised case. Another important generalization to be done is to supervised learning (classification and regression), which (likely) requires a stochastic model of the input variables. PHI could also be generalized to predictive density estimation proper by replacing $p(\boldsymbol{x}|\Theta)$ with a (parametric) class of densities $q_\vartheta(\boldsymbol{x})$. Finally, we could also go the full way to a decision-theoretic setup and loss. Note that Theorems 8 and 12 combined with (asymptotic) frequentist properties like consistency of MAP/ML/SMF easily yields analogous results for PHI.

Conclusion. We have shown that predictive hypothesis identification scores well on all desirable properties listed in Section 3. In particular, PHI can properly deal with nested hypotheses, and nicely justifies, reconciles, and blends MAP and ML for $m \gg n$, MDL for $m \approx n$, and SMF for $n \gg m$.

Acknowledgements. Many thanks to Jan Poland for his help improving the clarity of the presentation.

References

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover publications, 1974.

[BB04] M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

[BCH93] A. R. Barron, B. S. Clarke, and D. Haussler. Information bounds for the risk of Bayesian predictions and the redundancy of universal codes. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 54–54, 1993.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BM98] A. A. Borovkov and A. Moullagaliev. *Mathematical Statistics*. Gordon & Breach, 1998.

[Gei93] S. Geisser. *Predictive Inference*. Chapman & Hall/CRC, 1993.

[Grü07] P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, Cambridge, 2007.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. 300 pages, http://www.hutter1.net/ai/uaibook.htm.

[Hut07] M. Hutter. The loss rank principle for model selection. In *Proc. 20th Annual Conf. on Learning Theory (COLT'07)*, volume 4539 of *LNAI*, pages 589–603, San Diego, 2007. Springer, Berlin.

[KW96] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[Lan02] J. Langford. Combining train set and test set bounds. In *Proc. 19th International Conf. on Machine Learning (ICML-2002)*, pages 331–338. Elsevier, 2002.

[LF82] M. Lejeune and G. D. Faulkenberry. A simple predictive density function. *Journal of the American Statistical Association*, 77(379):654–657, 1982.

[MGB05] N. Mukhopadhyaya, J. K. Ghosh, and J. O. Berger. Some Bayesian predictive approaches to model selection. *Statistics & Probability Letters*, 73(4):2005, 2005.

[Wal96] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society B*, 58(1):3–57, 1996.

[Zuc00] W. Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44(1):41–61, 2000.