Learning from Both Experts and Data

Rémi Besson, Erwan Le Pennec, Stéphanie Allassonnière

▶ **To cite this version:**

Rémi Besson, Erwan Le Pennec, Stéphanie Allassonnière. Learning from Both Experts and Data. Entropy, 2019, 21 (12), pp.1208. 10.3390/e21121208 . hal-04432717

HAL Id: hal-04432717
https://hal.science/hal-04432717v1

Submitted on 6 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

entropy

MDPI

Article

Learning from Both Experts and Data

Rémi Besson [1,*], Erwan Le Pennec [1,2] and Stéphanie Allassonnière [3]

1 CMAP Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France; erwan.le-pennec@polytechnique.edu
2 XPop, Inria Saclay, 91120 Palaiseau, France
3 School of Medecine, Université Paris-Descartes, 75006 Paris, France; stephanie.allassonniere@polytechnique.edu
* Correspondence: remi.besson@polytechnique.edu

Received: 29 October 2019; Accepted: 6 December 2019; Published: 10 December 2019

Abstract: In this work, we study the problem of inferring a discrete probability distribution using both expert knowledge and empirical data. This is an important issue for many applications where the scarcity of data prevents a purely empirical approach. In this context, it is common to rely first on an a priori from initial domain knowledge before proceeding to an online data acquisition. We are particularly interested in the intermediate regime, where we do not have enough data to do without the initial a priori of the experts, but enough to correct it if necessary. We present here a novel way to tackle this issue, with a method providing an objective way to choose the weight to be given to experts compared to data. We show, both empirically and theoretically, that our proposed estimator is always more efficient than the best of the two models (expert or data) within a constant.

Keywords: maximum entropy; mixing expert and data; Kullback–Leibler centroid

1. Introduction

In this work, we present a novel way to estimate a discrete probability distribution, denoted $p^\star$, using both expert knowledge and data. This is a crucial aspect for many applications. Indeed, when deploying a decision support tool, we often rely entirely on expert/domain knowledge at the beginning; the data only comes with the use of the algorithm in real life. However, we need a good model of the environment to directly train the decision support tool with a planning algorithm. This model of the environment is to be refined and corrected as the data flow increases.

We assume here to have some expert knowledge in the form of an initial a priori on the marginals, the moments, and/or the support of $p^\star$ or any other relevant information. We also assume that we sequentially receive data. We denote $x^{(1)},\dots, x^{(n)}$ an independent and identically distributed (i.i.d) sample following a given unknown discrete probability distribution $p^\star$ in $\mathcal{P}$.

One example of application may come from the objective of building a symptom checker for rare diseases [1]. In this case, $p^\star$ represents the probabilities of the different possible combinations of symptoms, given the event that the disease of the patient is D. More precisely, we denote:

$$p^\star = (p^\star_1, ..., p^\star_K)^T = \begin{pmatrix} \mathbb{P}[\bar{B}_1, ..., \bar{B}_{J-1}, \bar{B}_J \mid D] \\ \mathbb{P}[\bar{B}_1, ..., \bar{B}_{J-1}, B_J \mid D] \\ \vdots \\ \mathbb{P}[B_1, ..., B_{J-1}, B_J \mid D] \end{pmatrix}. \tag{1}$$

We aim to estimate the distribution where D is the random variable disease. $B_1,\dots, B_J$ are the typical symptoms of the disease D; all are binary random variables, i.e., the symptom can be present

or absent. We aim to estimate the $2^J = K$ different combinations (as $\mathbb{P}[B_1, \ldots, B_L \mid D]$, for example) when we only have an expert a priori on the marginals $\mathbb{P}[B_i \mid D]$, for all $i \in [1, J]$.

Of course, a first idea would be to assume that the symptoms are conditionally independent given the disease. However, we expect complex correlations between the typical symptoms of a given disease. Indeed, we can imagine two symptoms that are very plausible individually, but which rarely occur together (or even never, in the case of incompatible symptoms like microcephaly and macrocephaly).

Note also that the assumption of conditional independence would make it possible to present a disease without having any of the symptoms related to this disease in the database (when there is no B_i such that $\mathbb{P}[B_i \mid D] = 1$), which should be impossible.

Generally speaking, if we had enough empirical data, we would no longer need the experts. Conversely, without empirical data, our model must be based entirely on experts. We detail here two different approaches to dealing with the intermediate regime where we do not have enough data to do without the a priori given by the experts, but where we have enough data to correct and specify this initial a priori. These approaches are meaningful as long as we do not know how much data have been used to build the initial a priori, and as long as we really try to combine two heterogeneous forms of information: Experts and empirical data.

In Section 2.1, we first recall the principle of maximum entropy, which is the basic brick we use to build an expert model. We then briefly introduce the proposed approach to mixing experts and data in Section 2.2. We underline the extent to which this approach is superior to the one we previously proposed in [1]. The Barycenter approach that we propose here provides an objective way to choose the weight to be given to experts compared to data. On the contrary, the maximum likelihood with entropic penalization approach of [1] was shown to be sensitive to the choice of the regularization parameter. In Section 3, we outline a review of the literature. Finally, in Section 4, we show both empirically and theoretically that our barycenter estimator is always more efficient than the best of the two models (expert or data) within a constant.

It should be noted that even though we will refer throughout the paper to our particular application in medicine, our framework is relevant for any inference problem involving an initial a priori with a particular form (marginals, moments, support, etc.) combined with data. Biology, ecology, and physics, to name a few, are areas where ideas of maximum entropy have been used for a long time and where the ideas developed in this work could be interesting. See [2] for an overview of the maximum entropy applications for inference in biology.

2. Mixing Expert and Empirical Data

2.1. Building an Expert Model: The Maximum Entropy Principle

Of course, the aim of benefiting simultaneously from expert data and empirical data has a very old history. This is the very essence of Bayesian statistics [3], which aims to integrate expert data in the form of an a priori, which is updated with empirical data using the Bayes' theorem to obtain what will be called the posterior.

Note that in our case, we do not have a classical a priori modeling the model parameters with probability distributions. We have an a priori on the marginals, such as a number of constraints on the distribution to be estimated. The absence of an obvious a priori to model the distribution of the parameters naturally leads us to the idea of maximum entropy, theorized by [4]. Indeed, if no model seems more plausible to us than another, then we will choose the least informative. This is a generalization of the principle of indifference often attributed to Laplace:

"We consider two events as equally probable, when we see no reason that makes one more probable than the other, because, even if there is an unequal possibility between them, since we don't know which is the biggest, this uncertainty makes us look at one as as likely as the other" [5].

This principle therefore takes the form of an axiom that allows us to construct a method to choose an a priori: The least informative possible a priori that is consistent with what we know.

We then define the distribution of maximum entropy as follows:

$$p^{\text{maxent}} = \underset{p/p\in\tilde{\mathcal{C}}}{\arg\max}\, H(p) \tag{2}$$

where $\tilde{\mathcal{C}} = \mathcal{C} \bigcap \mathcal{C}^{\text{expert}}$. $\mathcal{C}^{\text{expert}}$ is the set of constraints fixed by experts and $\mathcal{C}_K$ is the probability simplex of the discrete probability distributions of dimension K:

$$\mathcal{C}_K = \left\{ p = (p_1, \ldots, p_K) / \sum_{i=1}^{K} p_i = 1, p_i \geq 0 \right\}. \tag{3}$$

Note that p^{maxent} is well-defined; namely, it exists and is unique, as long as $\mathcal{C}^{\text{expert}}$ is a convex set. Indeed, the function $p \mapsto H(p)$ is strictly concave; it is a classic result that a strictly concave function under convex constraints admits a unique maximum.

It is well known that if $\mathcal{C}^{\text{expert}}$ only contains the constraints for the marginals, then p^{maxent} is nothing more than the independent distribution.

However, in our case, we can add some information about the structure of the desired distribution as constraints integrated into $\mathcal{C}^{\text{expert}}$. We judge that it is impossible to have a disease without having at least a certain number of its associated symptoms: One, two, or more depending on the disease. Indeed, the diseases we are interested in manifest themselves in combinations of symptoms. The combinations which allow the presence of two simultaneous but exclusive symptoms should also have constraints that are equal to 0. All combinations of constraints are conceivable, as long as $\tilde{\mathcal{C}}$ remains a convex closed space, in order to ensure the existence and uniqueness of p^{maxent}.

We therefore construct our a priori by taking the maximum entropy distribution, checking the constraints imposed by the experts. Thus, among the infinite distributions that verify the constraints imposed by the experts, we choose the least informative distribution p^{maxent}; in other words, the one closest to the conditional independence distribution.

We need to add information to move from the information provided by the experts to the final distribution, and we want to add as little as possible to what we do not know. This approach is referred to as maxent (maximum entropy) and has been widely studied in the literature [4,6,7].

2.2. Barycenters between Experts and Data

Our target probability distribution is denoted $p^\star = (p_1^\star, \ldots, p_K^\star)$ and is defined on the probability simplex $\mathcal{C}_K$ of Equation (3).

The i.i.d. sample of $p^\star$ is denoted $x^{(1)}, \ldots, x^{(n)}$, where $x^{(i)} \in \mathbb{R}^K$. The empirical distribution $p_n^{\text{emp}} = (p_{n,i}^{\text{emp}})_{i=1}^K$ is given by:

$$p_{n,i}^{\text{emp}} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{x^{(j)}=i\}}. \tag{4}$$

Following the ideas of Section 2.1, we define the expert distribution as the distribution which maximizes entropy while satisfying the constraints fixed by experts:

$$p^{\text{expert}} = \underset{p/p\in\tilde{\mathcal{C}}}{\arg\max}\, H(p) \tag{5}$$

where $\tilde{\mathcal{C}}$ is the intersection of the simplex probabilities with the set of constraints fixed by experts. In our medical context, the set of constraints is composed of a list of censured combinations and a list of marginals coming from the literature. Censured combinations are combinations of symptoms that are set to zero because they involve the simultaneous presence of two incompatible symptoms and/or combinations that do not involve enough of a presence of typical symptoms.

Note that it is possible to give more or less credit to the marginals given by experts by formulating the constraint as an interval (wider or narrower) rather than as a strict equality. The distribution of experts is then defined as the least informative distribution consistent with what we know.

Let $\mathcal{L}$ be any dissimilarity measured between two probability distributions. Our barycenter estimator mixing expert and empirical data is then defined as:

$$\hat{p}_{\epsilon_n}^{\mathcal{L}} = \underset{p \in \mathcal{C} / \mathcal{L}(p_n^{\text{emp}}, p) \leq \epsilon_n}{\arg\min} \mathcal{L}(p^{\text{expert}}, p) \tag{6}$$

where

$$\epsilon_n := \epsilon_n^{\delta} = \arg\min_{l} \mathbb{P}[\mathcal{L}(p_n^{\text{emp}}, p^{\star}) \leq l] \geq 1 - \delta \tag{7}$$

and $\mathbb{P}$ is the probability measure defined on the product space $\{(x^{(1)}, \ldots, x^{(n)}) \sim \otimes_{i=1}^{n} p^{\star}; n \geq 1\}$.

$\hat{p}_n^{\mathcal{L}}$ is then defined as the closest distribution from the experts, in the sense of the dissimilarity measure $\mathcal{L}$, which is consistent with the observed data.

For such a construction to be possible, we will therefore have to choose a measure of dissimilarity $\mathcal{L}$ so that we have a concentration of the empirical distribution around the true distribution for $\mathcal{L}$.

Such a formulation has several advantages over the maximum likelihood with entropic penalization approach previously proposed in [1]. First, we do not have to choose a regularization parameter, which seems to have a strong impact on the results of the estimator (see [1]). This parameter is replaced by the parameter δ, for which it is reasonable not to take more than 0.1 and which appears to have low impact on the result of $\hat{p}_n^{\mathcal{L}}$ (see Section 5). Secondly, the solution of (6) can be (depending on the choice of the dissimilarity measure $\mathcal{L}$) easier to compute than that of the optimization problem associated with the penalization approach, for which a closed form of the solution could not be derived [1]. Mathematically, $\hat{p}_n^{\mathcal{L}}$ is the projection of the experts on the confidence interval centered on the empirical distribution and radius ϵ_n. Figures 1 and 2 give a visual interpretation of such a construction. These representations should not be taken literally. The objects we work on live in the simplex of probabilities, and their geometry is very different from the Euclidean interpretation of Figures 1 and 2. These figures are mainly here to illustrate the two different cases we can have. In Figure 1, we do not have much data and the confidence interval is wide. In this case, the projection of the experts on the confidence interval centered on the empirical distribution is the expert distribution itself. We do not have enough elements to modify our initial a priori. This case can also occur when the initial a priori of the experts is very close to the true distribution. On the contrary, in Figure 2, we modify the initial a priori because the experts do not belong to the confidence interval anymore.

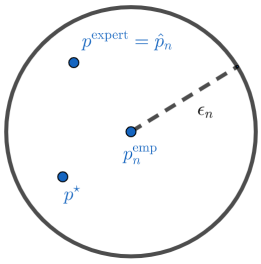

Figure 1. Barycenter between expert and data when the expert belongs to the confidence interval centered in the empirical distribution. In this case, there is no sufficient empirical evidence that the expert is wrong.

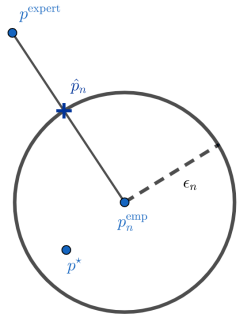

Figure 2. Barycenter between expert and data when the expert does not belong to the confidence interval centered in the empirical distribution. There is a high probability that the expert is outside the set where the target is located and therefore needs to be corrected.

2.3. A Baseline: Mixture Model with Bayesian Weights

We present here an alternative approach that will be our baseline in the numerical experiments of Section 5. We still aim to mix the empirical distribution, p_n^{emp}, built with an i.i.d. sample of $p^\star$: $x^{(1)},\dots,x^{(n)}$, with the distribution of the experts p^{expert}.

The idea is to make a linear combination of these two models:

$$\hat{p}_n^{\text{Bayes}} \propto \gamma_1 p_n^{\text{emp}} + \gamma_2 p^{\text{expert}}$$

where the mixture parameters are proportional to the log-likelihood of the data according to the model considered, namely:

$$\gamma_1 = \ell(x^{(1)},\dots,x^{(n)} \mid p_n^{\text{emp}})$$

and

$$\gamma_2 = \ell(x^{(1)},\dots,x^{(n)} \mid p^{\text{expert}})$$

where ℓ stands for the log-likelihood.

This is a parametric Bayesian approach, since we apply the Bayes theorem, stating that the posterior is proportional to the product of the prior with the likelihood function.

3. Related Works

3.1. Expert System with Probabilistic Reasoning

The creation of a decision support tool for medical diagnosis has been an objective since the beginning of the computer age. Most of the early work proposed a rule-based expert system, but in the 1980s, a significant part of the community studied the possibility of building an expert system using probabilistic reasoning [8]. Bayesian probabilities and methods were therefore considered as good ways to model the uncertainty inherent in medical diagnosis relatively early.

The assumption of conditional independence of symptoms given the disease has been intensively discussed, as it is of crucial importance for computational complexity. Some researchers considered this hypothesis harmless [9], while others already proposed a maximum entropy approach to face this issue [10–12].

However, it seems that none of the work of that time considered the expert vs. empirical data trade-off that we face. In the review article [13] presenting the state-of-the-art of the research of that time (1990) about this issue, it is clearly mentioned that these methods only deal with data of probabilistic forms. More precisely, they assume that they have an a priori on the marginal but also on some of the combinations of symptoms (in our case, we would assume that we have an a priori on $\mathbb{P}[B_1, B_2 \mid D]$, for example), and propose a maximum entropy approach where these expert data are

treated as constraints in the optimization process. Once again, this is not the case for us, since we have only an a priori on the marginal (and a certain number of constraints), as well as experimental data. This field of research was very active in the 1980s and then gradually disappeared, probably due to the computational intractability of the algorithms proposed for the computer resources of the time.

3.2. Bayesian Networks

Bayesian networks [14] were then quickly considered as a promising alternative for modeling probabilistic dependency relationships between symptoms and diseases [8]. These are now used in most expert systems, particularly in medicine [15].

A Bayesian network is generally defined as an acyclically oriented graph. The nodes in this graph correspond to the random variables: Symptoms or diseases in our case. The edges link two correlated random variables by integrating the information of the conditional law of the son node with respect to the father node. The main advantage of such a model is that it can factorize the joint distribution using the so-called global Markov property. The joint law can indeed be expressed as the product of the conditional distributions of each node given its direct parents in the graph [16].

First of all, the construction of a Bayesian network implies the inference of its structure, i.e., to determine the nodes that must be linked by an edge to those that can be considered conditionally independent of the rest of the graph (structure learning). Then, learning the network implies learning the parameters, i.e., the probabilities linking the nodes (parameter learning).

It is therefore natural to also find works that aimed at mixing expert and empirical data in this area of the literature. In [17], the experts' indications take a particular form since they indicate by hand correlations, positive or negative, between variables. The approach of [18] is also quite distant, because it is prefers being based on data. [18] only uses expert indications for additional variables for which there are no data, typically rare events never observed in the database. A work closer to ours is [19], where the authors assume that they have a first Bayesian network built entirely by the experts, to which they associate a degree of trust. The authors then use the available data to correct this expert network. We distinguish ourselves from this work in our effort to find an objective procedure for the weight to be given to experts in relation to the data (and for this weight not to be set by the experts themselves).

Note also that the main interest of Bayesian networks is to take advantage of conditional independence relationships known in advance, as they are pre-filled by experts or inferred from a sufficient amount of data. However, in our case, we do not have such an a priori knowledge about the dependency relationships between symptoms and or enough data to infer them.

3.3. From the Marginals to the Joint Distribution

Estimating the joint distribution from the marginal is an old problem, which is obviously not necessarily related to expert systems. This problem is sometimes referred to in the literature as the "cell probabilities estimation problem in contingency tables with fixed marginals". The book [20] gives a good overview of this field. We can trace back to the work of [21], which assumes knowing the marginal and having access to a sample of empirical data, and aims to estimate the joint distribution. In this article, they proposed the "iterative proportional fitting procedure" (IPFP) algorithm, which is still very popular for solving this problem.

An important assumption of [21] is that each cell of the contingency table receives data. In [22], the authors prove that the asymptotic estimator obtained by an IPFP algorithm is the distribution that minimizes the Kullback–Leibler divergence from the empirical distribution under the constraint to respect the marginal experts.

However, an IPFP algorithm is not suitable for our problem for two main reasons: First, we do not have absolute confidence in the marginals given by experts (we want to allow ourselves to modify them as we collect more data) and second, because we are interested in rare diseases, we do not expect to have a sufficient amount of data. In fact, many of the cells in the contingency table we are trying to

estimate will not receive data, but it would be disastrous in our application to assign a zero probability to the corresponding symptom combination.

In a sense, an IPFP algorithm does exactly the opposite of what we are aiming for: It modifies empirical data (as little as possible) to adapt them to experts, while we aim to modify experts (as little as possible) to make them consistent, in a less restrictive sense, with empirical data.

We should also mention the work related to our problem in applications of statistics in the social sciences, where researchers aim to construct a synthetic population from the marginal, coming from several inconsistent sources [23]. Their proposed approach also uses ideas of maximum entropy, but it is still different from our trade-off of expert vs. empirical data, since they built their model without samples.

3.4. The Kullback Centroid

Our optimization problem in Equation (6) in the particular case where the dissimilarity measure $\mathcal{L}$ is the Kullback–Leibler divergence is called moment-projection (M-projection) in the literature. The properties of these projections have been intensely studied [24].

Note that the Lagrangian associated with such an optimization problem is then nothing more than a Kullback–Leibler centroid. These objects or variations/generalizations of them (with Jeffrey's, Bregman's divergences, etc.) have been the subject of research since the paper of [25]. For example, articles [26,27] study cases where an exact formula can be obtained, and propose algorithms when this is not the case.

However, we have not found any use of these centroids to find a good trade-off of expert vs. empirical data as we propose in this paper. Bregman's divergence centroids have been used to mix several potentially contradictory experts; the interested reader may refer to the work of [28,29]. We could certainly consider that the empirical distribution p_n^{emp} is a second expert, and that our problem is the same as mixing two experts: Literature and data. However, the question of the weight to be given to each expert, which is the question that interests us here, will not be resolved. In [28], the aim is rather to synthesize contradictory opinions of different experts by fixing the weight to be given to each expert in advance. We propose, for our part, an objective procedure to determine the weight to be given to experts compared to empirical data.

4. Theoretical Properties of the Barycenter Estimator

4.1. Barycenter in L^p Space

In this section we work in the L^p space. Let us recall that the classic norm on the space L^p is given by: $\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$.

Following the ideas presented in Section 2.2, we define our estimator, $\forall i \geq 1$ as follows:

$$\hat{p}_n^i = \underset{p \in \mathcal{C} / \|p - p_n^{\text{emp}}\|_i \leq \epsilon_n}{\arg\min} \|p - p^{\text{expert}}\|_i \tag{8}$$

where

$$\epsilon_n := \epsilon_n^\delta = \underset{l}{\arg\min}\, \mathbb{P}[\|p_n^{\text{emp}} - p^\star\|_i \leq l] \geq 1 - \delta. \tag{9}$$

To control ϵ_n, we use the concentration inequality obtained in the recent work of [30]. In the literature, most of the concentration inequalities for the empirical distribution use the L^1 norm. This is why, even though we will present the following results by trying to generalize to spaces L^p for all p, in practice, only $\hat{p}_n^1$ interests us. The proofs for the different theoretical results of this section are relegated to the Appendix A.

Proposition 1 (Existence and uniqueness). *The estimator $\hat{p}_n^i$ defined by (8) exists for all $i \geq 1$.*
$\hat{p}_n^i$ is unique if and only if $i \neq 1$.
In the following, $\hat{p}_n^1$ therefore refers to a set of probability measures.

Proof. See Appendix A.1. □

The next proposition shows that one of the solutions of (8) can always be written as a barycenter between p_n^{emp} and p^{expert}. This property therefore provides us with an explicit expression of a solution of (8), which was not otherwise trivial to obtain by a direct calculation looking for the saddle points of the Lagrangian (for example, in the case $i = 1$).

Proposition 2. *Let $\hat{p}_n^i$ be defined by (8); then for all i, it exists $\tilde{p} \in \hat{p}_n^i$ such that $\exists \alpha_n \in [0,1]$:*

$$\tilde{p} = \alpha_n p^{expert} + (1 - \alpha_n) p_n^{emp} \tag{10}$$

where $\alpha_n = \dfrac{\epsilon_n}{\|p_n^{emp} - p^{expert}\|_i}$ if $\epsilon_n \leq \|p_n^{emp} - p^{expert}\|_i$ and $\alpha_n = 1$ otherwise.

Proof. See Appendix A.2. □

Therefore, one of the elements of $\hat{p}_n^1$ can be written under the form of a barycenter. For the sake of simplicity, in the following, we will designate $\hat{p}_n^1$ as the solution of (8) for $i = 1$, which can be written under the form of (10) rather than using the whole set of solutions.

It is now a question of deriving a result proving that mixing experts and data, as we do with $\hat{p}_n^1$, represents an interest rather than a binary choice of one of the two models. For this reason, we show in the following proposition that, with a high probability, our estimator $\hat{p}^{1,1}$ is always better than the best of the models within a constant.

Theorem 1. *Let $\hat{p}_n^1$ be defined by (8). Then, we have with probability of at least $1 - \delta$:*

$$\|p^\star - \hat{p}_n^1\|_1 \leq 2 \min\{\epsilon_n, \|p^\star - p^{expert}\|_1\}. \tag{11}$$

Proof. See Appendix A.3. □

4.2. Barycenter Using the Kullback–Leibler Divergence

In this section, we study the theoretical properties of the solution of Equation (6) in the particular case where the dissimilarity measure $\mathcal{L}$ is the Kullback–Leibler divergence. The proofs for the different theoretical results of this section are relegated to the Appendix B.

The Kullback–Leibler divergence between two discrete probability measures p and q is defined as:

$$\mathbb{KL}(p||q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right).$$

Let us recall that the Kullback-Leibler divergence is not a distance, since it is not symmetric and does not satisfy the triangular inequality; however, it is positively defined [6].

We define our estimator as:

$$\hat{p}_n^L = \underset{p \in \mathcal{C} / \mathbb{KL}(p_n^{\text{emp}}||p) \leq \epsilon_n}{\arg\min} \mathbb{KL}(p^{\text{expert}}||p) \tag{12}$$

where

$$\epsilon_n := \epsilon_n^\delta = \arg\min_l \mathbb{P}[\mathbb{KL}(p_n^{\text{emp}}||p^\star) \leq l] \geq 1 - \delta. \tag{13}$$

To calibrate ϵ_n, we can use the concentration inequality obtained in [30]. More precisely, we have:

$$\epsilon_n = \frac{1}{n}\left(-\log(\delta) + \log \underbrace{\left(3 + 3\sum_{i=1}^{K-2} \left(\sqrt{\frac{e^3 n}{2\pi i}} \right)^i \right)}_{=:G_n} \right). \tag{14}$$

In the following proposition, we show the existence and uniqueness of our estimator $\hat{p}_n^L$ and the fact that our estimator is a barycenter. However, unlike in the case of $\hat{p}_n^1$ of Equation (8), it does not seem possible to obtain a closed form for $\hat{p}_n^L$ this time.

Proposition 3. *Let $\hat{p}_n^L$ be defined by (12); then, $\hat{p}_n^L$ exists and is unique. Moreover, $\hat{p}_n^L$ can be written under the following form:*

$$\hat{p}_n^L = \frac{1}{1+\tilde{\lambda}} p^{expert} + \frac{\tilde{\lambda}}{1+\tilde{\lambda}} p_n^{emp} \tag{15}$$

where $\tilde{\lambda}$ is a non-negative real such that:

$$\tilde{\lambda} \geq \frac{\mathbb{KL}(p_n^{emp} || p^{expert})}{\epsilon_n} - 1. \tag{16}$$

Proof. See Appendix B.1. □

The following proposition is intended to be the analog of the proposition 1 when $\mathcal{L}$ is the Kullback–Leibler divergence. We prove that the centroid $\hat{p}_n^L$ is better than the experts (with high probability). On the other hand, we obtain that when $\mathbb{KL}(p_n^{\text{emp}} || p^\star) > \mathbb{KL}(p^{\text{expert}} || p^\star)$, the $\hat{p}_n^L$ barycenter is better than the empirical distribution. To obtain guarantees when $\mathbb{KL}(p_n^{\text{emp}} || p^\star) \leq \mathbb{KL}(p^{\text{expert}} || p^\star)$ seems less obvious and requires control over the quantity $\mathbb{KL}(p_n^{\text{emp}} || p^{\text{expert}})$.

Theorem 2. *Let $\hat{p}_n^L$ be defined by (12); then, we have with probability at least $1-\delta$:*

$$\mathbb{KL}(\hat{p}_n^L || p^\star) \leq \min\left\{ \mathbb{KL}(p^{expert} || p^\star), \epsilon_n \left(L_n + 1 \right) \right\} \tag{17}$$

where

$$L_n = \frac{\mathbb{KL}(p^{expert} || p^\star) - \mathbb{KL}(p_n^{emp} || p^\star)}{\mathbb{KL}(p_n^{emp} || p^{expert})}.$$

Proof. See Appendix B.2. □

Remark 1. *Note that $\mathbb{KL}(\hat{p}_n^L || p^\star)$ is infinite if p^{expert} does not have the same support as that of $p^\star$. Nevertheless, obtaining a result for $\mathbb{KL}(p^\star || \hat{p}_n^L)$ would require us to have a concentration on $\mathbb{KL}(p^\star || \hat{p}_n^{emp})$, which we do not have. Note that $\mathbb{KL}(p^\star || \hat{p}_n^{emp})$ is infinite until we have sampled all the elements of the support of $p^\star$ at least one time.*

5. Numerical Results

For each experiment in this section, we generate a random distribution $p^\star$ that we try to estimate. To do this, we simulate some realizations of a uniform distribution and renormalize in order to sum up to 1.

We also generate four different distributions that will serve as a priori for the inference: $p^{\text{expert},i}, \forall i \in \{1,2,3,4\}$. The first three priors are obtained by a maximum entropy procedure under constraint to respect marginals of $p^\star$ having undergone a modification. We added to the marginals

of $p^\star$ a Gaussian noise of zero expectation and variance equal to $\sigma_1^2 = 0.1$, $\sigma_2^2 = 0.2$ and $\sigma_3^2 = 0.4$, respectively. The last prior $p^{\text{expert,4}}$ is chosen to be equal to the distribution $p^\star$ (the experts provided us with the right distribution).

We then sequentially sample data from $p^\star$, i.e., we generate patients, and update for each new datum and each different a priori, the left centroid $\hat{p}_n^L$ (using an Uzawa algorithm [31]), the barycenter $\hat{p}_n^{1,1}$, and the empirical distribution p_n^{emp}, as well as the divergences $\mathbb{KL}(\hat{p}_n^L||p^\star)$ and $\mathbb{KL}(p_n^{\text{emp}}||p^\star)$ and the norms $\|\hat{p}_n^{1,1} - p^\star\|_1$ and $\|p_n^{\text{emp}} - p^\star\|_1$.

The experiments of Figures 3–6 were conducted on the case of a disease with $J = 7$ typical symptoms and where there are therefore $K = 2^7 = 128$ possible combinations. The experiments of Figures 7–9 were conducted on the case of a disease with 9 typical symptoms and where there are therefore $K = 2^9 = 512$ possible combinations.

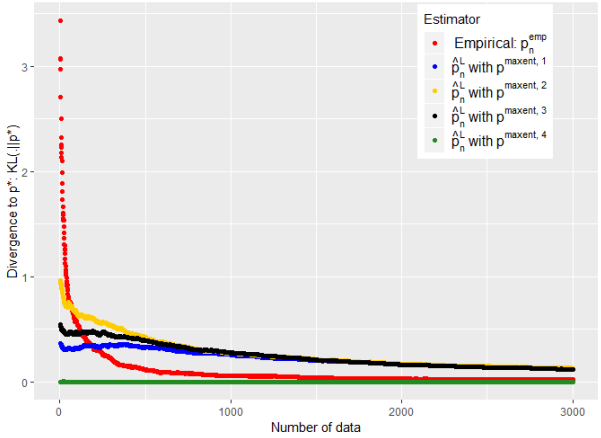

Figure 3. Evolution of the performance of $\hat{p}_n^L$ as a function of the available number of empirical data. ϵ_n is defined by Equation (18).

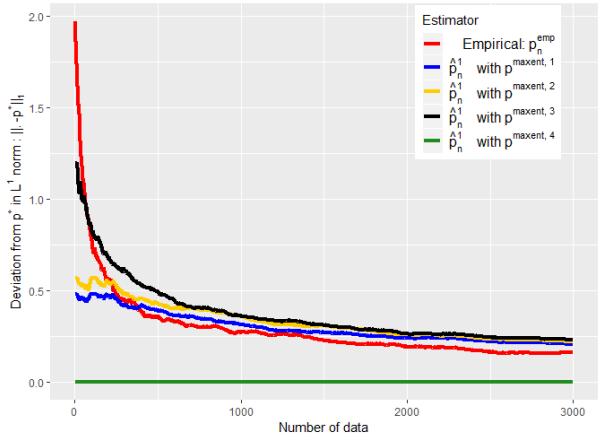

Figure 4. Evolution of the performance of $\hat{p}_n^1$ as a function of the available number of empirical data. ϵ_n is defined by Equation (20).

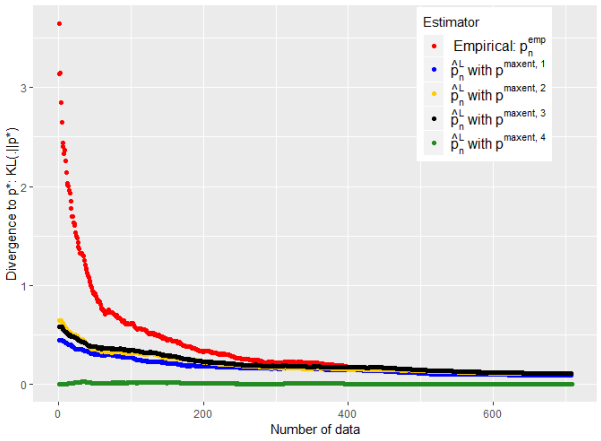

Figure 5. Evolution of the performance of $\hat{p}_n^L$ as a function of the available number of empirical data. ϵ_n is defined by Equation (19). Number of symptoms: 7.

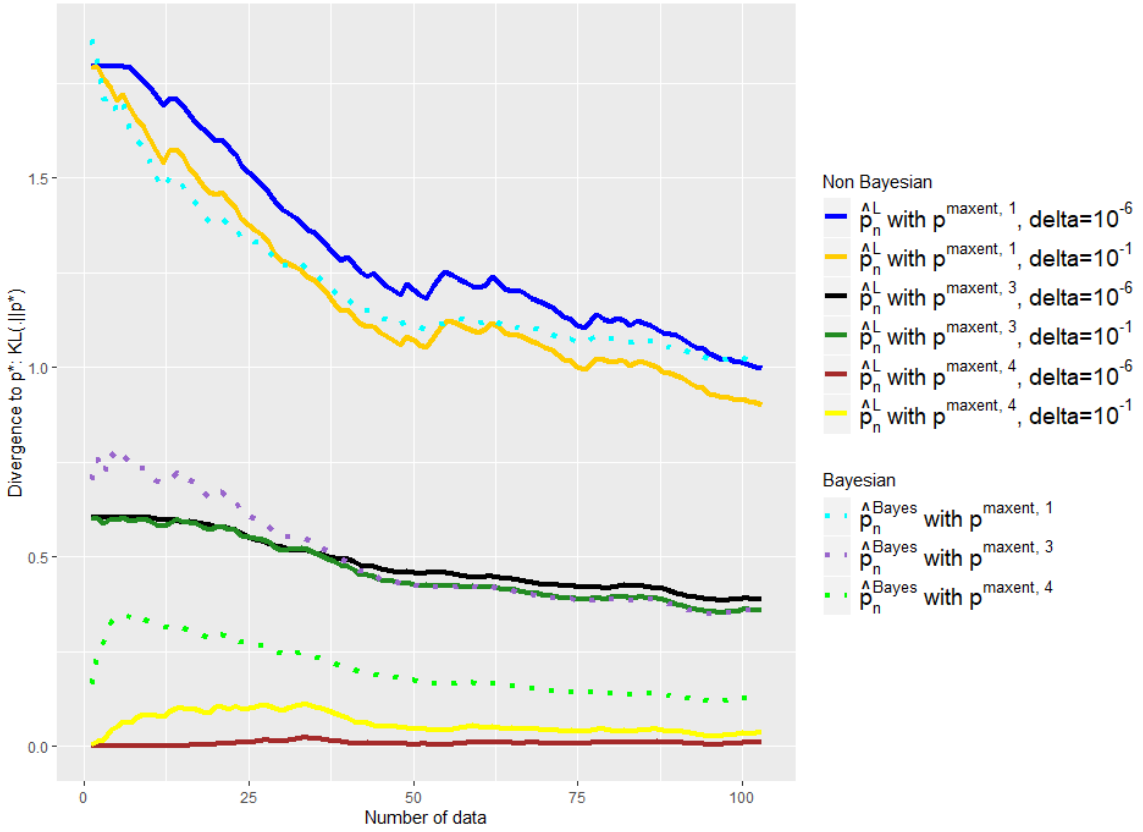

Figure 6. Comparison of the performances of $\hat{p}_n^L$ and $\hat{p}_n^{\text{Bayes}}$ as a function of the available number of empirical data with different initial a priori and δ. ϵ_n is defined by Equation (18). Number of symptoms: 7.

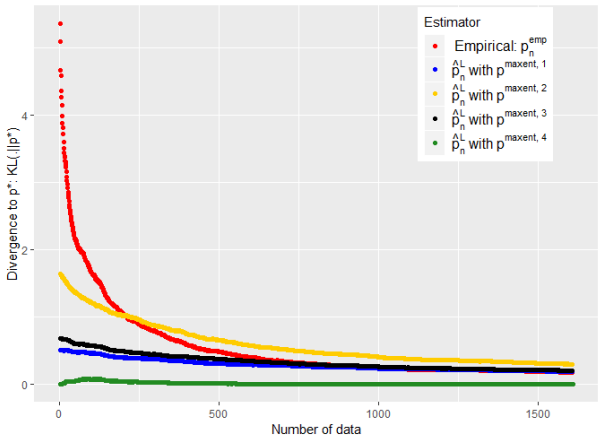

Figure 7. Evolution of the performance of $\hat{p}_n^L$ as a function of the available number of empirical data. ϵ_n is defined by Equation (19). Number of symptoms: 9.

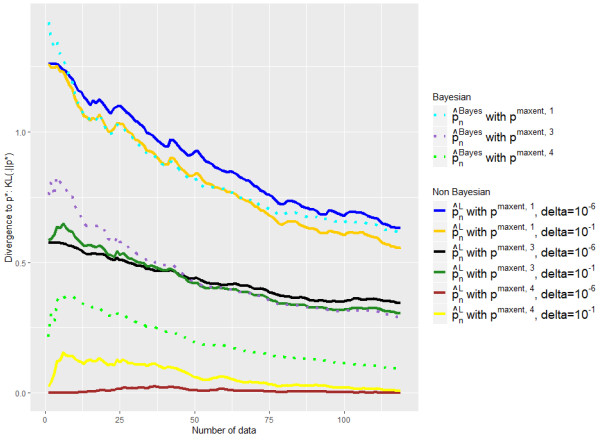

Figure 8. Comparison of the performances of $\hat{p}_n^L$ and $\hat{p}_n^{\text{Bayes}}$ as a function of the available number of empirical data with different initial a priori and δ. ϵ_n is defined by Equation (18). Number of symptoms: 9.

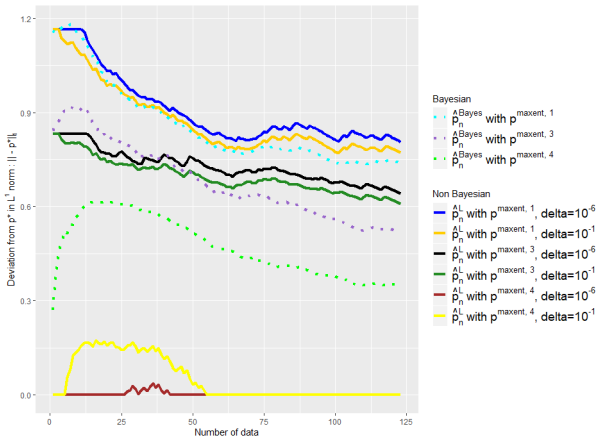

Figure 9. Comparison of the performances of $\hat{p}_n^1$ and $\hat{p}_n^{\text{Bayes}}$ as a function of the available number of empirical data with different initial a priori and δ. ϵ_n is defined by Equation (20). Number of symptoms: 9.

5.1. General Analysis of the Barycenter Performance and Choice of ϵ_n

The only parameter that we can control is the δ used to construct the confidence interval of the concentration of the empirical distribution around the true distribution. Let us recall that for the case of the Kullback centroid of Equation (12), we set:

$$\epsilon_n = \frac{1}{n}\left(-\log(\delta) + \log(G_n)\right) \tag{18}$$

where G_n is defined in Equation (14).

However, our first numerical experiments show that the choice of ϵ_n defined by Equation (18) is a little too conservative (see Figure 3). We need to converge ϵ_n faster towards 0 without abandoning our a priori when it is good.

Our experiments suggest taking an ϵ_n consistent with the concentration proposed in a conjecture of [30] for Kullback–Leibler divergence:

$$\epsilon_n = \frac{-\log(\delta) + \frac{n}{2}\log\left(1 + \frac{K-1}{n}\right)}{n}. \tag{19}$$

Note that we added a constant $\frac{1}{2}$ to the conjecture of [30]. As for the choice of δ, this appears important mainly when n is small; taking it when sufficiently low avoids an overfitting situation when the number of data is still low, without being harmful when n is high. We took it equal to 10^{-6} in the experiments of Figures 3–5 and 7, and tried different values in Figure 6.

The Figures 5 and 7 show that such a choice for ϵ_n makes a good trade-off between expert and empirical data, because we are able to take advantage of these two sources of information when the number of data is small (typically when $n < K$), but also to quickly abandon our a priori when it is bad (see the black curves) or to keep it when it is good (the green curves). Eventually, the Figures 5 and 7 were performed on problems of 128 and 512, respectively, and this choice of ϵ_n therefore appears relatively robust with respect to changes in size.

Concerning $\hat{p}_n^{1,1}$, we took, still following the conjectures of [30]:

$$\epsilon_n = \sqrt{\frac{-\log(\delta) + \frac{n}{2}\log\left(1 + \frac{K-1}{n}\right)}{n}}. \tag{20}$$

The Figure 4 shows the error made by our barycenter in norm L^1: $\hat{p}_n^1$ using such an ϵ_n. We are again able to get rid of a bad a priori relatively quickly to follow the empirical (green curve) while keeping it if it is good (blue curve).

Moreover, we show with these experiments that there is an intermediate regime when we do not have much data, where our estimator is *strictly* better than the two individual models (experts and data alone). This is particularly visible when we used the ϵ_n of the conjecture of [30] (see Figure 7 and 5). It is then empirically evident that mixing these two heterogeneous sources of information, experts and empirical data, can be useful for statistical inference.

One might nevertheless wonder by looking at the experiments of Figures 3–5 and 7 why we propose a mixture of expert and data rather than just a binary choice of the best model. Indeed, both our theoretical and experimental results show that we can lose a constant when making a barycenter between expert and data instead of just a binary choice of the best of the two models. This is particularly true when the number of data tends to grow and when the initial expert a priori was misleading. Nevertheless, this constant is a price that we are willing to pay in order to avoid the undesirable consequences of a binary choice of model.

First, when making a binary choice of model, it is not that easy to determine when we should jump from the expert model to the empirical model. Note also that it would produce an undesirable discontinuity in the function of the data flow. Most importantly, it is crucial in our application that our estimator has the same support as the real distribution. It would be disastrous indeed to consider that

a disease is impossible because we never observed a particular combination of symptoms. This remark is somewhat linked to the well-known coupon collector's problem: How many samples do we need on average to observe all the modalities of the support of a given distribution at least one time? In the equal case (the target distribution is uniform), the average number of samples needed is of the order of $K\log(K)$, but it might be much more in the unequal case [32]. Nevertheless, let us emphasize here once again that we are particularly interested in the moment where we have little data. Then, the empirical distribution alone will never be a really good alternative. We could, of course, consider a Laplace smoothing in order to avoid this difficulty, but this would be nothing more than a less sophisticated maximum entropy approach.

5.2. Comparison with the Baseline and Choice of δ

In Figures 6, 8 and 9 we compare our approach with the Bayesian mixture of Section 2.3. We removed the empirical distribution curve for visual reasons, because it is always above the curves presented and thus distorts the representation by stretching the y-axis.

We tried two different values for our only parameter δ: $\delta = 10^{-1}$ and $\delta = 10^{-6}$. Note that the advantage of our method is that the parameter that we have to choose, δ, has an intelligible meaning: It refers to the probability that $p^\star$ is outside the confidence interval. That is why we do not consider higher values of δ.

First of all, one can note the influence of the δ parameter on Figures 6, 8 and 9. The light yellow curve is a clear example of where the δ has been chosen too high, 10^{-1}, giving too much credit to data in comparison with the expert. There is of course a trade-off; to choose a smaller δ, 10^{-6} has a cost, as we can see with the black and the dark blue curves, which are a bit too conservative in comparison with the dark green and the dark yellow ones.

Nevertheless, despite the variability of the observed performance of our method as a function of δ, it leads in any case to a better mixture than the baseline in all our experiments. Our barycenter then outperforms the baseline in this task of finding the right weight to give to data in relation to the expert. This is particularly true when $\delta = 10^{-6}$ and to a lesser extent when $\delta = 10^{-1}$.

Indeed, we aim at finding a mixture that would keep the expert a priori when it is good and quickly move away when it is bad. This is not what the baseline exhibits in our experiments, contrary to our estimator. The light green curve shows clearly that the weight placed on the data in relation to the expert is too high; the beginning of the purple curve also exhibits this behavior.

Once again, the Figure 8 shows that such observations are robust with respect to changes of dimension, and the Figure 8 has $2^9 = 512$ symptom combinations; meanwhile, the Figure 6 has $2^7 = 128$.

6. Conclusion and Perspectives

In this work, we have presented a way to combine expert knowledge—in the form of marginal probabilities and rules—together with empirical data in order to estimate a given discrete probability distribution. This problem emerged from our application, in which we aimed to learn the probability distribution of the different combinations of symptoms of a given disease. For this objective, we have an initial a priori consisting of the marginal distributions coming from the medical literature; clinical data collected is used as the decision support tool.

The particular form of the prior does not allow us to simply adopt a maximum a posteriori (MAP) approach. The absence of an obvious a priori to model the parameter's distribution naturally leads us to the idea of maximum entropy: If no model seems more plausible to us than another, then we will choose the least informative.

This idea of maximum entropy brings us back to the works of the 1980s and 1990s, where researchers also aimed to build symptom checkers using marginals. In our work, we go further by gradually integrating empirical data as the algorithm is used.

We are interested in the intermediate regime in which we do not have enough empirical data to do without experts, but have enough to correct them if necessary. Our proposal is to construct our estimator as the distribution closest to the experts' initial a priori, in the sense of a given dissimilarity measure, that is consistent with the empirical data collected.

We prove, both theoretically and empirically, that our barycenter estimator mixing the two sources of information is always more efficient than the best of the two models (clinical data or experts alone) within a constant.

We have empirically illustrated the effectiveness of the proposed approach by giving an a priori of different quality and incrementally adding empirical data. We have shown that our estimator allows a bad a priori to be abandoned relatively quickly when the inconsistency of the data collected with the initial a priori is observed. At the same time, this same mixture makes it possible to keep the initial a priori if it is good. Moreover, we show with this experiment that, in the intermediate regime, our estimator can be *strictly* better than the best of the two models (experts and data alone). It empirically confirms the idea that mixing these two heterogeneous sources of information can be profitable in statistical inference.

Future work will concentrate on several refinements, such as the addition of a kernel structure for the construction of the empirical distribution. Indeed, it is possible that there are omissions of some symptoms in the data collected. Then, a kernel approach that would consider closer states that only differ by some presences would capture such a difficulty and make a better use of empirical data. Other dissimilarity measures could also be investigated. Finally, having a true non-parametric Bayesian approach would be very interesting. However, closing the gap between classical Dirichlet priors on the marginal to a single prior on the joint distribution seems to be a real challenge.

Author Contributions: Conceptualization, R.B., E.L.P., and S.A.; Investigation, R.B.; Supervision, E.L.P. and S.A.; Writing—original draft, R.B.; Writing—review & editing, E.L.P. and S.A.

Funding: This work is supported by a public grant overseen by the French National research Agency (ANR) as part of the «Investissement d'Avenir» program, through the "IDI 2017" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02 and by the projet PR[AI]RIE.

Acknowledgments: The authors thanks Frédéric Logé-Munerel for fruitful discussions about this work.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Proof of the Theoretical Results of Our Barycenter Estimator in The L^p Spaces

Appendix A.1. Existence and Uniqueness

Proof of Proposition 1. The existence of a solution of (8) for all $i \geq 1$ is a consequence of the fact that the projection onto a finite dimension set always exists.

The uniqueness of a solution of (8) for all $i \neq 1$ is due to the fact that we aim to minimize a strictly convex function under convex constraints. When $i = 1$, the function that we aim to minimize is no longer strictly convex and some counterexamples can be exhibited.

For example, if $p^{\text{expert}} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $p_n^{\text{emp}} = (\frac{1}{2}, 0, \frac{1}{2}, 0)$ and $\epsilon_n = \frac{9}{10}$.

Note that $\|p^{\text{expert}} - p_n^{\text{emp}}\|_1 = 1 > \frac{9}{10}$. Then, using Proposition 2, we know that

$$\begin{aligned} &\frac{\epsilon_n}{\|p_n^{\text{emp}} - p^{\text{expert}}\|_1} p^{\text{expert}} + \left(1 - \frac{\epsilon_n}{\|p_n^{\text{emp}} - p^{\text{expert}}\|_1}\right) p_n^{\text{emp}} \\ &= \left(\frac{11}{40}, \frac{9}{40}, \frac{11}{40}, \frac{9}{40}\right) =: \hat{p}_n^1 \end{aligned}$$

is a solution. However, $\tilde{p} = (\frac{10}{40}, \frac{9}{40}, \frac{12}{40}, \frac{9}{40})$ is a solution too. Indeed:

$$\|\tilde{p} - p^{\text{expert}}\|_1 = \frac{1}{10} = \|\hat{p}_n^{1,1} - p^{\text{expert}}\|_1$$

and:

$$\|\tilde{p} - p_n^{\text{emp}}\|_1 = \frac{36}{40} = \frac{9}{10}.$$

□

Appendix A.2. Linear Combination of Expert and Data: A Barycenter

Proof of Proposition 2. Let $\tilde{p} \in \mathcal{C}$ be such that there exists $\alpha \in [0,1]$ where $\tilde{p} = \alpha p^{\text{expert}} + (1-\alpha)p_n^{\text{emp}}$ and such that $\|\tilde{p} - p_n^{\text{emp}}\|_i = \epsilon_n$. We then have:

$$\|\tilde{p} - p_n^{\text{emp}}\|_i = \alpha \|p_n^{\text{emp}} - p^{\text{expert}}\|_i = \epsilon_n$$

and then

$$\alpha = \frac{\epsilon_n}{\|p_n^{\text{emp}} - p^{\text{expert}}\|_i}.$$

Moreover, note that we have the following equality, since $\tilde{p}$ can be written in the form of a barycenter:

$$\|\tilde{p} - p^{\text{expert}}\|_i + \underbrace{\|\tilde{p} - p_n^{\text{emp}}\|_i}_{=\epsilon_n} = \|p_n^{\text{emp}} - p^{\text{expert}}\|_i.$$

Let us make a reasoning by reductio ad absurdum. Let us assume that there exist $p' \in \mathcal{C}$ such that $\|p_n^{\text{emp}} - p'\|_i \leq \epsilon_n$ and $\|p^{\text{expert}} - p'\|_i < \|p^{\text{expert}} - \tilde{p}\|_i$.

We would then have:

$$\begin{aligned}
\|p_n^{\text{emp}} - p^{\text{expert}}\|_i &\leq \|p' - p^{\text{expert}}\|_i + \|p' - p_n^{\text{emp}}\|_i \\
&< \|\tilde{p} - p^{\text{expert}}\|_i + \|p' - p_n^{\text{emp}}\|_i \\
&= \|p_n^{\text{emp}} - p^{\text{expert}}\|_i - \epsilon_n + \|p' - p_n^{\text{emp}}\|_i \\
&\leq \|p_n^{\text{emp}} - p^{\text{expert}}\|_i
\end{aligned}$$

which leads to the desired contradiction. □

Appendix A.3. Our Barycenter Estimator Is More Efficient than the Best of the Two Models, Expert or Data, Within a Constant

Proof of Theorem 1. A simple application of the triangular inequality gives us:

$$\|p^\star - \hat{p}_n^1\|_1 \leq \|p^\star - p_n^{\text{emp}}\|_1 + \|p_n^{\text{emp}} - \hat{p}_n^1\|_1.$$

However, $\|p_n^{\text{emp}} - \hat{p}_n^1\|_1 \leq \epsilon_n$ by construction, and we have $\|p^\star - p_n^{\text{emp}}\|_1 \leq \epsilon_n$ with probability of at least $1 - \delta$.

In addition to that:

$$\|p^\star - \hat{p}_n^1\|_1 \leq \|p^\star - p^{\text{expert}}\|_1 + \|p^{\text{expert}} - \hat{p}_n^1\|_1.$$

However, using the definition of $\hat{p}_n^1$ and assuming $\|p^\star - p_n^{\text{emp}}\|_1 \leq \epsilon_n$, then:

$$\|p^{\text{expert}} - \hat{p}_n^1\|_1 \leq \|p^{\text{expert}} - p^\star\|_1.$$

We can conclude that if $\|p^\star - p_n^{\text{emp}}\|_1 \leq \epsilon_n$, which happens with probability of at least $1-\delta$, then:

$$\|p^\star - \hat{p}_n^1\|_1 \leq 2\min\{\epsilon_n, \|p^\star - p^{\text{expert}}\|_1\}. \tag{A1}$$

□

Appendix B. Proof of the Theoretical Results of Our Barycenter Estimator with the $\mathbb{KL}$ Divergence

Appendix B.1. Existence and Uniqueness. Formula as a Linear Combination of Experts and Empirical Data

Proof of Proposition 3. The existence and uniqueness of $\hat{p}_n^L$ is a consequence of the fact that $\mathcal{T} = \{p/p \in \mathcal{C} \text{ and } \mathbb{KL}(p_n^{\text{emp}}||p) \leq \epsilon_n\}$ is a convex set. Indeed, let $p, q \in \mathcal{T}$, $\alpha \in [0,1]$; then, using the classical log-sum inequality, we have:

$$\mathbb{KL}(p_n^{\text{emp}}||\alpha p + (1-\alpha)q) \leq \alpha\mathbb{KL}(p_n^{\text{emp}}||p) + (1-\alpha)\mathbb{KL}(p_n^{\text{emp}}||q) \leq \epsilon_n.$$

The Lagrangian associated to the optimization problem (12) can be written as:

$$\begin{aligned} L(p,\lambda,\mu) = &\sum_i p_i^{\text{expert}} \log\left(\frac{p_i^{\text{expert}}}{p_i}\right) \\ &+ \lambda\left(\sum_i p_i^{\text{emp}} \log\left(\frac{p_i^{\text{emp}}}{p_i}\right) - \epsilon_n\right) \\ &+ \mu\left(\sum_i p_i - 1\right). \end{aligned}$$

Deriving for all $i \in [1, K]$:

$$\frac{\partial L(p,\lambda,\mu)}{\partial p_i} = -\frac{p_i^{\text{expert}}}{p_i} - \lambda\frac{p_i^{\text{emp}}}{p_i} + \mu.$$

Equating this last expression to 0 and using the fact that the probability measure sums to 1, we find: $\mu = \lambda + 1$. Then, we have for all $i \in [1, K]$:

$$p_i = \frac{1}{1+\lambda} p_i^{\text{expert}} + \frac{\lambda}{1+\lambda} p_i^{\text{emp}}.$$

We know that $\hat{p}_n^L$ exists and is unique, and using the Kuhn–Tucker theorem (whose assumptions we satisfy since we minimize a convex function under convex inequality constraints), we know that the minimum of the optimization problem (12) is reached for the saddle-point of the Lagrangian: $(\tilde{\lambda}, \tilde{p}) = (\tilde{\lambda}, \hat{p}_n^L)$. We can then write:

$$\hat{p}_n^L = \frac{1}{1+\tilde{\lambda}} p^{\text{expert}} + \frac{\tilde{\lambda}}{1+\tilde{\lambda}} p_n^{\text{emp}}. \tag{A2}$$

We could not obtain a closed form for $\tilde{\lambda}$, unlike in the case of $\hat{p}^{i,i}$. However, we know that by construction, $\mathbb{KL}(p_n^{\text{emp}}||\hat{p}_n) \leq \epsilon_n$.

Moreover, using the log-sum inequality and our interpolation formula (A2), we have:

$$\begin{aligned} \mathbb{KL}(p_n^{\text{emp}}||\hat{p}_n) &= \mathbb{KL}\left(p_n^{\text{emp}}||\frac{1}{1+\tilde{\lambda}} p^{\text{expert}} + \frac{\tilde{\lambda}}{1+\tilde{\lambda}} p_n^{\text{emp}}\right) \\ &\leq \frac{1}{1+\tilde{\lambda}}\mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}}). \end{aligned}$$

We then have the following condition under $\tilde{\lambda}$:

$$\frac{1}{1+\tilde{\lambda}}\mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}}) \leq \epsilon_n \Leftrightarrow \tilde{\lambda} \geq \frac{\mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}})}{\epsilon_n} - 1.$$

□

Appendix B.2. Our Barycenter Is More Efficient Than the Best of the Two Models, Expert or Data, within a Constant

Proof of Theorem 2. Using the Proposition 3, we have:

$$\begin{aligned}
\mathbb{KL}(\hat{p}_n^L||p^\star) &= \mathbb{KL}\left(\frac{1}{1+\tilde{\lambda}}p^{\text{expert}} + \frac{\tilde{\lambda}}{1+\tilde{\lambda}}p_n^{\text{emp}}||p^\star\right) \\
&\leq \frac{1}{1+\tilde{\lambda}}\mathbb{KL}(p^{\text{expert}}||p^\star) + \frac{\tilde{\lambda}}{1+\tilde{\lambda}}\mathbb{KL}(p_n^{\text{emp}}||p^\star) \\
&= \frac{1}{1+\tilde{\lambda}}\left(\mathbb{KL}(p^{\text{expert}}||p^\star) - \mathbb{KL}(p_n^{\text{emp}}||p^\star)\right) \\
&+ \mathbb{KL}(p_n^{\text{emp}}||p^\star) \\
&\leq \epsilon_n \left(\frac{\mathbb{KL}(p^{\text{expert}}||p^\star) - \mathbb{KL}(p_n^{\text{emp}}||p^\star)}{\mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}})}\right) \\
&+ \mathbb{KL}(p_n^{\text{emp}}||p^\star)
\end{aligned}$$

where we used the available inequality to $\tilde{\lambda}$ (Proposition 3) in the last inequality, and the desired result is obtained by assuming that $\mathbb{KL}(p_n^{\text{emp}}||p^\star) \leq \epsilon_n$, which happens with probability of at least $1-\delta$.

In addition, note that:

$$\begin{aligned}
\epsilon_n \left(\frac{\mathbb{KL}(p^{\text{expert}}||p^\star) - \mathbb{KL}(p_n^{\text{emp}}||p^\star)}{\mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}})}\right) + \mathbb{KL}(p_n^{\text{emp}}||p^\star) \\
\leq \mathbb{KL}(p^{\text{expert}}||p^\star) \\
\Leftrightarrow \epsilon_n \leq \mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}}).
\end{aligned}$$

However, if $\epsilon_n \geq \mathbb{KL}(p_n^{\text{emp}}||p^{\text{expert}})$, we have by construction that $\hat{p}_n^L = p^{\text{expert}}$, and therefore $\mathbb{KL}(\hat{p}_n^L||p^\star) = \mathbb{KL}(p^{\text{expert}}||p^\star)$.

We can conclude from all of this that:

$$\mathbb{KL}(\hat{p}_n^L||p^\star) \leq \mathbb{KL}(p^{\text{expert}}||p^\star).$$

□

References

1. Besson, R.; Le Pennec, E.; Allassonnière, S.; Stirnemann, J.; Spaggiari, E.; Neuraz, A. A Model-Based Reinforcement Learning Approach for a Rare Disease Diagnostic Task. *arXiv* **2018**, arXiv:1811.10112.
2. De Martino, A.; De Martino, D. An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon* **2018**, *4*, e00596. [CrossRef] [PubMed]
3. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2004.
4. Jaynes, E. Information Theory and Statistical Mechanics. *Phys. Rev. (Ser. I)* **1957**, *106*, 620–630. [CrossRef]
5. Laplace, P.-S.D. Mémoire sur la Probabilité des Causes par les évènements. Available online: https://gallica.bnf.fr/ark:/12148/bpt6k77596b/f32 (accessed on 5 December 2019). (In French)

6. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
7. Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A Maximum Entropy Approach to Natural Language Processing. *Comput. Linguist.* **1996**, *22*, 39–71.
8. Pearl, J. Probabilistic reasoning in intelligent systems: networks of plausible inference. *J. Philos.***1988**, *88*, 434–437.
9. Charniak, E. *The Bayesian Basis of Common Sense Medical Diagnosis*; Brown University: Providence, RI, USA, 1983; pp. 70–73.
10. Hunter, D. Uncertain Reasoning Using Maximum Entropy Inference. *Mach. Intell. Pattern Recognit.* **1986**, *4*, 203–209.
11. Shore, J.E. Relative Entropy, Probabilistic Inference and AI. *arXiv* **2013**, arXiv:1304.3423.
12. Miller, J.W.; Goodman, R.M. A Polynomial Time Algorithm for Finding Bayesian Probabilities from Marginal Constraints. *arXiv* **2013**, arXiv:1304.1104.
13. Jirousek, R. A survey of methods used in probabilistic expert systems for knowledge integration. *Knowl.-Based Syst.* **1990**, *3*, 7–12. [CrossRef]
14. Jensen F.V. *An Introduction to Bayesian Networks;* Editions UCL Press: London, UK, 1996.
15. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning*; MIT Press: Cambridge, MA, USA, 2009.
16. Spiegelhalter, D.; Dawid, P.; Lauritzen, S.; Cowell, R. Bayesian Analysis in Expert Systems. *Stat. Sci.* **1993**, *8*, 219–247. [CrossRef]
17. Zhou, Y.; Fenton, N.; Zhu, C. An empirical study of Bayesian network parameter learning with monotonic influence constraints. *Decis. Support Syst.* **2016**, *87*, 69–79. [CrossRef]
18. Constantinou, A.C.; Fenton, N.; Neil, M. Integrating expert knowledge with data in Bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved. *Expert Syst. Appl.* **2016**, *56*, 197–208. [CrossRef] [PubMed]
19. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **1995**, *20*, 197–243. [CrossRef]
20. Beniger, J.R. Discrete Multivariate Analysis: Theory and Practice. *Am. Sociol. Assoc.* **1975**, *4*, 507–509. [CrossRef]
21. Deming, W. Edwards and Stephan, Frederick F., On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Ann. Math. Stat.* **1940**, *11*, 427–444. [CrossRef]
22. Ireland, C.T.; Kullback, S. Contingency tables with given marginals. *Biometrika* **1968**, *55*, 179–188. [CrossRef]
23. Barthelemy, J.; Toint, P.L. Synthetic Population Generation without a Sample. *Transp. Sci.* **2013**, *47*, 266–279. [CrossRef]
24. Csiszar, I.; Matus, F. Information projections revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1474–1490. [CrossRef]
25. Veldhuis, R. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Process Lett.* **2002**, *9*, 96–99. [CrossRef]
26. Nielsen, F.; Nock, R. Sided and Symmetrized Bregman Centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
27. Nielsen, F. The centroid of the Jeffreys Centroids: A Closed-Form Expression for Positive Histograms and a Guaranteed Tight Approximation for Frequency Histograms. *IEEE Signal Process Lett.* **2013**, *20*, 657–660. [CrossRef]
28. Adamcik, M. Collective Reasoning under Uncertainty and Inconsistency. Ph.D. Thesis, The University of Manchester, Manchester, UK, March 2014.
29. Adamk, M. The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning. *Entropy* **2014**, *16*, 6338–6381. [CrossRef]
30. Mardia, J.; Jiao, J.; Tánczos, E.; Nowak, R.D.; Weissman, T. Concentration Inequalities for the Empirical Distribution. *arXiv* **2018**, arXiv:1809.06522.
31. Uzawa, H. *Iterative Methods for Concave Programming*; Stanford University Press: Stanford, CA, USA, 1958.
32. Ferrante, M.; Saltalamacchia, M. The Coupon Collector's Problem. *Math. Mater.* **2014**, 1–35.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).