

Update of Prior Probabilities by Minimal Divergence

Jan Naudts 

Departement Fysica, Universiteit Antwerpen, 2610 Antwerpen, Belgium; Jan.Naudts@uantwerpen.be

Abstract: The present paper investigates the update of an empirical probability distribution with the results of a new set of observations. The update reproduces the new observations and interpolates using prior information. The optimal update is obtained by minimizing either the Hellinger distance or the quadratic Bregman divergence. The results obtained by the two methods differ. Updates with information about conditional probabilities are considered as well.

Keywords: statistical update procedure; minimal divergence; Hellinger distance; Bregman divergence; Jeffrey conditioning

1. Introduction

The present work is inspired by the current practices in Information Geometry [1–3] where minimization of divergences is an important tool. In Statistical Physics a divergence is called a relative entropy. Its importance was noted rather late in the twentieth century, after the work of Jaynes on the maximal entropy principle [4]. Estimation in the presence of hidden variables by minimizing a divergence function is briefly discussed in Chapter 8 of [2].

Assume now that some observation or experiment yields new statistical data. The approach is then to look for a probability distribution that reproduces the newly observed probabilities and that interpolates the data with missing information coming from a prior.

No further model assumptions are imposed. Hence, the statistical model under consideration consists of all probability distributions that are consistent with the newly obtained empirical data. Internal consistency of the empirical data ensures that the model is not empty. The update is the model point that minimizes the chosen divergence function from the prior to the manifold of the model.

In the context of Maximum Likelihood Estimation (MLE) one usually adopts a parameterized model. The dimension of the model can be kept low and properties of the model can be used to ease the calculations. One assumes that the new data can lead to a more accurate estimation of the limited number of model parameters. It can then happen that the model is misspecified [5] and that the update is only a good approximation of the empirical data.

Here, the model is dictated by the newly acquired empirical data and the update is forced to reproduce the measured data. Finding the probability distribution is then an underdetermined problem. Minimization of the divergence from the prior probability distribution solves the underdetermination.

In Bayesian statistics, the update $q(B)$ of the probability $p(B)$ of an event B equals

$$q(B) = p^{\text{emp}}(A) p(B|A) + p^{\text{emp}}(A^c) p(B|A^c). \quad (1)$$

The quantities $p^{\text{emp}}(A)$ and $p^{\text{emp}}(A^c)$ are the empirical probabilities obtained after repeated measurement of event A and its complement A^c . Expression (1) has been called *Jeffrey conditioning* [6]. It implies the sufficiency conditions $q(B|A) = p(B|A)$ and $q(B|A^c) = p(B|A^c)$. It is an updating rule used in Radical Probabilism [7]. This expression is also obtained when minimizing the Hellinger distance between the prior and the model manifold. A proof of the latter follows later on in Section 4.



Citation: Naudts, J. Update of Prior Probabilities by Minimal Divergence. *Entropy* **2021**, *23*, 1668. <https://doi.org/10.3390/e23121668>

Academic Editors: Wolfgang von der Linden and Sascha Ranftl

Received: 18 November 2021

Accepted: 9 December 2021

Published: 11 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. 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 (<https://creativecommons.org/licenses/by/4.0/>).

The present approach is a special case of minimizing a divergence function in the presence of linear constraints. See the introduction of [8] for an overview of early applications of this technique. Two classes of generalized distance functions satisfy a natural set of axioms: the f -divergences of Csiszár and the generalized Bregman divergences. The squared Hellinger distance belongs to the former class. The other divergence function considered here is the square Bregman divergence. Both Hellinger and square Bregman have special properties that make it easy to work with them.

A broad class of generalized Bregman divergences satisfies the Pythagorean equality [8,9]. Pythagorean inequalities hold for an even larger class [10]. The Pythagorean relations derived in the present work make use of the specific properties of the Hellinger distance and of the quadratic Bregman divergence. It is unclear how to prove them for more general divergences.

One incentive for starting the present work is a paper of Banerjee, Guo, and Wang [11,12]. They consider the problem of predicting a random variable Z_1 given observations of a random variable Z_2 . It is well-known that the conditional expectation, as defined by Kolmogorov, is the optimal predictor. They show that this statement remains true when the metric distance is replaced by a Bregman divergence. It is shown in Theorem 2 below that a proof in a more general context yields a deviating result.

The next Section fixes notations. Section 3 collects some results about the squared Hellinger distance and the quadratic Bregman divergence. Section 4 discusses the optimal choice and contains the Theorems 1 and 2. The proof of the theorems can be adapted to cover the situation that a subsequent measurement also yields information on conditional probabilities. This is shown in Section 4.3. Section 5 treats a simple example. A final section summarizes the results of the paper.

2. Empirical Data

Consider a probability space Ω, μ . A measurable subset A of Ω is called an event. Its probability is denoted $p(A)$ and is given by

$$p(A) = \int_{\Omega} \mathbb{I}_A(x) d\mu(x),$$

where $\mathbb{I}_A(x)$ equals 1 when $x \in A$ and 0 otherwise. The conditional expectation of a random variable f given an event A with non-vanishing probability $p(A)$ is given by

$$\mathbb{E}_{\mu} f|A = \frac{1}{p(A)} \mathbb{E}_{\mu} f \mathbb{I}_A.$$

The probability space Ω, μ reflects the prior knowledge of the system at hand. When new data become available an update procedure is used to select the posterior probability space. The latter is denoted Ω, ν in what follows. The corresponding probability of an event A is denoted $q(A)$.

The outcome of repeated experiments is the empirical probability distribution of the events, denoted $p^{\text{emp}}(A)$. The question at hand is then to establish a criterion for finding the update ν of the probability distribution μ that is as close as possible to μ while reproducing the empirical results.

The event A defines a partition A, A^c of the probability space Ω, μ . As before A^c denotes the complement of A in Ω . In what follows a slightly more general situation is considered in which the event A is replaced by a partition $(O_i)_{i=1}^n$ of the measure space Ω, μ into subsets with non-vanishing probability. The notations p_i and μ_i are used, with

$$p_i = p(O_i) \quad \text{and} \quad d\mu_i(x) = \frac{1}{p_i} \mathbb{I}_{O_i}(x) d\mu(x). \quad (2)$$

Introduce the random variable g defined by $g(x) = i$ when $x \in O_i$. Repeated measurement of the random variable g yields the empirical probabilities

$$p_i^{\text{emp}} = \text{Emp Prob} \{x : g(x) = i\}.$$

They may deviate from the prior probabilities p_i . In some cases one also measures the conditional probabilities

$$p^{\text{emp}}(B|O_i) = \text{Emp Prob of } B \text{ given that } g(x) = i$$

of some other event B .

3. A Geometric Approach

In this section two divergences are reviewed, the squared Hellinger distance and the quadratic Bregman divergence.

3.1. Squared Hellinger Distance

For simplicity the present section is restricted to the case that the sample space Ω is the real line.

Given two probability measures μ and σ , both absolutely continuous w.r.t. the Lebesgue measure, the squared Hellinger distance is the divergence $D_2(\sigma||\mu)$ defined by

$$D_2(\sigma||\mu) = \frac{1}{2} \int_{\mathbb{R}} \left(\sqrt{\frac{d\sigma}{dx}} - \sqrt{\frac{d\mu}{dx}} \right)^2 dx.$$

It satisfies

$$D_2(\sigma||\mu) = 1 - \int_{\mathbb{R}} \sqrt{\frac{d\sigma}{dx} \frac{d\mu}{dx}} dx.$$

Let $(O_i)_i$ be a partition of Ω , μ and let $g(x) = i$ when x belongs to O_i , as before. Let p_i and μ_i be defined by (2). Consider the following functions of i , with i in $\{1, \dots, n\}$,

$$\begin{aligned} \tau^{(1)}(i) &= \mu, \quad \text{independent of } i, \\ \tau^{(2)}(i) &= \mu_i, \\ \tau^{(3)}(i) &= \sigma_i, \end{aligned}$$

where each of the σ_i is a probability distribution with support in O_i . The empirical expectation of a function $f(i)$ is given by $\mathbb{E}^{\text{emp}} f = \sum_i p_i^{\text{emp}} f(i)$.

Proposition 1. *If $p_i^{\text{emp}} > 0$ for all i and $\sum_i p_i^{\text{emp}} = 1$ then one has*

$$\mathbb{E}^{\text{emp}} D_2(\tau^{(1)}||\tau^{(3)}) \geq \mathbb{E}^{\text{emp}} D_2(\tau^{(1)}||\tau^{(2)})$$

with equality if and only if $\sigma_i = \mu_i$ for all i .

First prove the following two lemmas.

Lemma 1. *Assume that the probability measure ν_i is absolutely continuous w.r.t. the measure μ_i , with Radon-Nikodym derivative given by $d\nu_i(x) = f_i(x) d\mu_i$. Then one has*

$$D_2(\mu||\sigma_i) - D_2(\mu||\nu_i) = \sqrt{p_i} [D_2(\mu_i||\sigma_i) - D_2(\mu_i||\nu_i)]$$

and

$$D_2(\mu_i||\nu_i) = 1 - \int_{O_i} \sqrt{f_i(x)} d\mu_i(x).$$

Proof. One calculates

$$\begin{aligned}
 D_2(\mu||\sigma_i) - D_2(\mu||\nu_i) &= \int_{\mathbb{R}} \sqrt{\frac{d\mu}{dx}} \left[\sqrt{\frac{d\nu_i}{dx}} - \sqrt{\frac{d\sigma_i}{dx}} \right] dx \\
 &= \sqrt{p_i} \int_{O_i} \sqrt{\frac{d\mu_i}{dx}} \left[\sqrt{\frac{d\nu_i}{dx}} - \sqrt{\frac{d\sigma_i}{dx}} \right] dx \\
 &= \sqrt{p_i} \left[\int_{O_i} \sqrt{f_i(x)} d\mu_i(x) - \int_{O_i} \left[\frac{d\mu_i}{dx} \frac{d\sigma_i}{dx} \right]^{1/2} dx \right] \\
 &= \sqrt{p_i} \left[\int_{O_i} \sqrt{f_i(x)} d\mu_i(x) - 1 + D_2(\mu_i||\sigma_i) \right].
 \end{aligned}$$

Now take $\sigma_i = \nu_i$ to obtain the desired results. \square

Lemma 2. (Pythagorean relation) For any i is

$$D_2(\mu||\sigma_i) = D_2(\mu||\mu_i) + \sqrt{p_i}D_2(\mu_i||\sigma_i).$$

Proof. The proof follows by taking $\nu_i = \mu_i$ in the previous lemma. \square

Proof. (Proposition 1)

From the previous lemma it follows that $D_2(\tau^{(1)}||\tau^{(3)}) \geq D_2(\tau^{(1)}||\tau^{(2)})$. Note that $\sigma_i = \mu_i$ implies that $\tau^{(3)} = \tau^{(2)}$ and hence $D_2(\tau^{(1)}||\tau^{(3)}) = D_2(\tau^{(1)}||\tau^{(2)})$. Conversely, if

$$\mathbb{E}^{\text{emp}} D_2(\tau^{(1)}||\tau^{(3)}) = \mathbb{E}^{\text{emp}} D_2(\tau^{(1)}||\tau^{(2)})$$

then it follows from the previous lemma that $\mathbb{E}^{\text{emp}} D_2(\tau^{(2)}||\tau^{(3)}) = 0$. If in addition $p_i^{\text{emp}} > 0$ for all i then it follows that for all i

$$0 = D_2(\tau^{(2)}(i)||\tau^{(3)}(i)).$$

Because the squared Hellinger distance is a divergence, this implies that $\tau^{(2)}(i) = \tau^{(3)}(i)$, which is equivalent with $\mu_i = \sigma_i$. \square

3.2. Bregman Divergence

In the present section the squared Hellinger distance, which is an f-divergence, is replaced by a divergence of the Bregman type. In addition let Ω be a finite set equipped with the counting measure ρ . It assigns to each subset A of Ω the number of elements in A . This number is denoted $|A|$. The expectation value $\mathbb{E}_\mu f$ of a random variable f w.r.t. the probability measure μ is given by

$$\mathbb{E}_\mu f = \sum_{k \in \Omega} \mu(k) f(k).$$

Given a partition of Ω into sets O_i one can define conditional probability measures with probability mass function ρ_i given by

$$\begin{aligned}
 \rho_i(k) &= \frac{1}{|O_i|} \quad \text{if } k \in O_i, \\
 &= 0 \quad \text{otherwise.}
 \end{aligned} \tag{3}$$

Similarly, conditional probability measures with probability mass function μ_i are given by

$$\begin{aligned}
 \mu_i(k) &= \frac{\mu(k)}{\mu(O_i)} \quad \text{if } k \in O_i, \\
 &= 0 \quad \text{otherwise.}
 \end{aligned} \tag{4}$$

Fix a strictly convex function $\phi : \mathbb{R} \mapsto \mathbb{R}$. The Bregman divergence of the probability measures σ and μ is defined by

$$D_\phi(\sigma||\mu) = F(\sigma) - F(\mu) - \langle \nabla F, \sigma - \mu \rangle$$

with

$$F(\sigma) = \sum_k \phi(\sigma(k)) \quad \text{and} \quad \nabla_k F(\sigma) = \phi'(\sigma(k)).$$

In the case that $\phi(x) = x^2/2$, which is used below, it becomes

$$D_\phi(\sigma||\mu) = \frac{1}{2} \sum_k [\sigma(k) - \mu(k)]^2. \tag{5}$$

For convenience, this case is referred to as the *quadratic Bregman divergence*.

The following result, obtained with the quadratic Bregman divergence, is more elegant than the result of Lemma 2.

Proposition 2. Consider the quadratic Bregman divergence D_ϕ as given by (5). Let $\nu_i = p_i\mu_i + (1 - p_i)\rho_i$. Let σ_i be any probability measure with support in O_i . Then the following Pythagorean relation holds.

$$D_\phi(\mu||\sigma_i) = D_\phi(\mu||\nu_i) + D_\phi(\nu_i||\sigma_i).$$

Proof. One calculates

$$\begin{aligned} D_\phi(\mu||\sigma_i) - D_\phi(\mu||\nu_i) &= D_\phi(\nu_i||\sigma_i) + \sum_x [\mu(x) - \nu_i(x)] [\phi'(\nu_i(x)) - \phi'(\sigma_i(x))] \\ &= D_\phi(\nu_i||\sigma_i) + \sum_{x \in O_i} [p_i\mu_i(x) - \nu_i(x)] [\phi'(\nu_i(x)) - \phi'(\sigma_i(x))] \\ &= D_\phi(\nu_i||\sigma_i) - (1 - p_i) \frac{1}{|O_i|} \sum_{x \in O_i} [\phi'(\nu_i(x)) - \phi'(\sigma_i(x))]. \end{aligned}$$

Use now that $\phi'(u) = u$ and the normalization of the probability measures ν_i and σ_i to find the desired result. \square

4. The Optimal Choice

4.1. Updated Probabilities

The following result proves that the standard Kolmogorovian definition of the conditional probability minimizes the Hellinger distance between the prior probability measure μ and the updated probability measure ν . The optimal choice of the updated probability measure ν is given by corresponding probabilities $q(B)$. They satisfy

$$q(B) = \sum_{i=1}^n p_i^{emp} p(B|O_i) \quad \text{for any event } B.$$

Theorem 1. Let be given a partition $(O_i)_{i=1}^n$ of the probability space Ω , μ with $\Omega = \mathbb{R}$. Let μ_i be given by (2). Let $p_i = p(O_i) > 0$ denote the probability of the event O_i and let be given strictly positive empirical probabilities p_i^{emp} , $i = 1, \dots, n$. The squared Hellinger distance $D_2(\sigma||\mu)$ as a function of σ is minimal if and only if $\sigma_i = \mu_i$ for all i . Here, σ is any probability measure on Ω satisfying

$$\sigma = \sum_{i=1}^n p_i^{emp} \sigma_i,$$

and each of the σ_i is a probability measure with support in O_i and absolutely continuous w.r.t. μ_i .

Note that the probability measure ν given by

$$\nu(x) = \sum_{i=1}^n p_i^{\text{emp}} \mu_i(x)$$

uses the Kolmogorovian conditional probability as the predictor because the probabilities determined by the μ_i are obtained from the prior probability distribution μ by $p_i(x) = p(x|O_i)$. By the above theorem this predictor is the optimal one w.r.t. the squared Hellinger distance.

Proof. With the notations of the previous section is

$$D_2(\sigma||\mu) = \mathbb{E}^{\text{emp}} D_2(\tau^{(1)}||\tau^{(3)}).$$

Proposition 1 shows that it is minimal if and only if $\sigma_i = \mu_i$ for all i . \square

Next, consider the use of the quadratic Bregman divergence in the context of a finite probability space.

Theorem 2. Let be given a partition $(O_i)_{i=1}^n$ of the finite probability space Ω, μ . Let ρ_i be the counting measure on O_i defined by (3). Let μ_i be given by (2). Let $p_i = p(O_i) > 0$ denote the probability of the event O_i and let be given strictly positive empirical probabilities $p_i^{\text{emp}}, i = 1, \dots, n$ summing up to 1. Assume that

$$p_i^{\text{emp}} \geq p_i [1 - |O_i| \mu_i(x)] \quad \text{for all } x \in O_i \text{ and for } i = 1, \dots, n. \tag{6}$$

Then the following hold.

(a) A probability distribution ν is defined by $\nu = \sum_i p_i^{\text{emp}} \nu_i$ with

$$\nu_i = \left(1 - \frac{p_i}{p_i^{\text{emp}}}\right) \rho_i + \frac{p_i}{p_i^{\text{emp}}} \mu_i. \tag{7}$$

(b) Let σ be any probability measure on Ω satisfying $\sigma = \sum_{i=1}^n p_i^{\text{emp}} \sigma_i$, where each of the σ_i is a probability distribution with support in O_i . Then the quadratic Bregman divergence satisfies the Pythagorean relation

$$D_\phi(\sigma||\mu) = D_\phi(\nu||\mu) + \sum_{i=1}^n (p_i^{\text{emp}})^2 D_\phi(\sigma_i||\nu_i). \tag{8}$$

(c) The quadratic Bregman divergence $D_\phi(\sigma||\mu)$ is minimal if and only if $\sigma = \nu$.

Proof.

(a)

The assumption (6) guarantees that the $\nu_i(x)$ are probabilities.

(b)

One calculates

$$\begin{aligned} D_\phi(\sigma||\mu) - D_\phi(\nu||\mu) &= \frac{1}{2} \sum_x [\sigma(x) - \nu(x)] [\sigma(x) + \nu(x) - 2\mu(x)] \\ &= \sum_{i=1}^n p_i^{\text{emp}} \frac{1}{2} \sum_{x \in O_i} [\sigma_i(x) - \nu_i(x)] \\ &\quad \times [p_i^{\text{emp}} \sigma_i(x) + p_i^{\text{emp}} \nu_i(x) - 2p_i \mu_i(x)] \\ &= \sum_{i=1}^n (p_i^{\text{emp}})^2 \frac{1}{2} \sum_{x \in O_i} [\sigma_i(x) - \nu_i(x)]^2 \end{aligned}$$

$$\begin{aligned}
 & + \sum_{i=1}^n p_i^{\text{emp}} \sum_{x \in O_i} [\sigma_i(x) - \nu_i(x)] (p_i^{\text{emp}} - p_i) \rho_i(x) \\
 & = \sum_{i=1}^n (p_i^{\text{emp}})^2 D_\phi(\sigma_i || \nu_i).
 \end{aligned}$$

In the above calculation the third line is obtained by eliminating $p_i \mu_i$ using the definition of ν_i . This gives

$$\begin{aligned}
 & p_i^{\text{emp}} \sigma_i(x) + p_i^{\text{emp}} \nu_i(x) - 2p_i \mu_i(x) \\
 & = p_i^{\text{emp}} \sigma_i(x) + p_i^{\text{emp}} \nu_i(x) - 2p_i^{\text{emp}} \left[\nu_i(x) - \left(1 - \frac{p_i}{p_i^{\text{emp}}} \right) \rho_i(x) \right] \\
 & = p_i^{\text{emp}} [\sigma_i(x) - \nu_i(x)] + 2(p_i^{\text{emp}} - p_i) \rho_i(x).
 \end{aligned}$$

The term

$$\sum_{i=1}^n p_i^{\text{emp}} \sum_{x \in O_i} [\sigma_i(x) - \nu_i(x)] (p_i^{\text{emp}} - p_i) \rho_i(x)$$

vanishes because $\rho_i(x)$ is constant on the set O_i and the probability measures ν_i and σ_i have support in O_i .

(c)

From (b) it follows that $D_\phi(\sigma || \mu) \geq D_\phi(\nu || \mu)$, with equality when $\sigma = \nu$. Conversely, when $D_\phi(\sigma || \mu) = D_\phi(\nu || \mu)$ then (8) implies that

$$\sum_{i=1}^n (p_i^{\text{emp}})^2 D_\phi(\sigma_i || \nu_i) = 0.$$

The empirical probabilities are strictly positive by assumption. Hence, it follows that $D_\phi(\mu || \sigma_i) = D_\phi(\mu || \nu_i)$ for all i and hence, that $\sigma_i = \nu_i$ for all i . The latter implies $\sigma = \nu$. \square

The optimal update ν can be written as

$$\nu = \sum_i [(p_i^{\text{emp}} - p_i) \rho_i + p_i \mu_i] = \mu + \sum_i (p_i^{\text{emp}} - p_i) \rho_i.$$

This result is in general quite different from the update proposed by Theorem 1, which is

$$\nu = \sum_i p_i^{\text{emp}} \mu_i.$$

The updates proposed by the two theorems coincide only in the special cases that either $p_i^{\text{emp}} = p_i$ for all i or that $\mu_i = \rho_i$ for all i . In the latter case the prior distribution $\mu = \sum_i p_i \rho_i$ is replaced by the update $\nu = \sum_i p_i^{\text{emp}} \rho_i$.

The entropy of the update when event O_i is observed, according to Theorem 1, equals $S(\nu_i) = S(\mu_i)$. According to Theorem 2 it equals

$$S(\nu_i) = S\left(\left[1 - \frac{p_i}{p_i^{\text{emp}}}\right] \rho_i + \frac{p_i}{p_i^{\text{emp}}} \mu_i\right).$$

If $p_i \leq p_i^{\text{emp}}$ then it follows that

$$\begin{aligned}
 S(\nu_i) & \geq \left[1 - \frac{p_i}{p_i^{\text{emp}}}\right] S(\rho_i) + \frac{p_i}{p_i^{\text{emp}}} S(\mu_i) \\
 & \geq S(\mu_i).
 \end{aligned}$$

The former inequality follows because the entropy is a concave function. The latter follows because entropy is maximal for the uniform distribution ρ_i . On the other hand, if $p_i > p_i^{\text{emp}}$ then one has

$$\begin{aligned} S(\mu_i) &= S\left(\left[1 - \frac{p_i^{\text{emp}}}{p_i}\right]\rho_i + \frac{p_i^{\text{emp}}}{p_i}v_i\right) \\ &\geq \left[1 - \frac{p_i^{\text{emp}}}{p_i}\right]S(\rho_i) + \frac{p_i^{\text{emp}}}{p_i}S(v_i) \\ &\geq S(v_i). \end{aligned}$$

In the latter case the decrease of the entropy is stronger than in the case of the update based on the squared Hellinger distance. In conclusion, the update relying on the quadratic Bregman divergence loses details of the prior distribution by making a convex combination with a uniform distribution weighed with the probabilities of the observation. It does this more so for the events with observed probability larger than predicted; this is when $p_i^{\text{emp}} > p_i$.

Note that Theorem 2 cannot always be applied because it contains restrictions on the empirical probabilities. In particular, if the prior probability $\mu(x)$ of some point x in Ω vanishes then the condition (6) requires that the empirical probability p_i^{emp} of the partition O_i to which the point x belongs is larger than or equal to the prior probability p_i .

4.2. Update of Conditional Probabilities

The two previous theorems assume that no empirical information is available about conditional probabilities. If such information is present then an optimal choice should make use of it. In one case the solution of the problem is straightforward. If the probabilities p_i^{emp} are available together with all conditional probabilities $p^{\text{emp}}(B|O_i)$ and there exists an update ν which reproduces these results then it is unique. Two cases remain: (1) The information about the conditional probabilities is incomplete; (2) the information is internally inconsistent – no update exists which reproduces the data.

Let us tackle the problem by considering the case that the only information that is available besides the probabilities p_i^{emp} is the vector of conditional probabilities $p^{\text{emp}}(B|O_i)$ of a fixed event B , given the outcome of the measurement of the random variable g as introduced in Section 2.

The following result is independent of the choice of divergence function.

Proposition 3. Fix an event B in Ω . Assume that the conditional probabilities $p(B|O_i), i = 1, \dots, n$, are strictly positive and strictly less than 1. Assume in addition that $p_i^{\text{emp}} p^{\text{emp}}(B|O_i) \leq 1$ for all i . Then there exists an update ν with corresponding probabilities $q(\cdot)$ such that $q(O_i) = p_i^{\text{emp}}$ and $q(B|O_i) = p^{\text{emp}}(B|O_i), i = 1, \dots, n$.

Proof. An obvious choice is to take ν of the form $\nu = \sum_i p_i^{\text{emp}} \nu_i$ with ν_i of the form

$$d\nu_i(x) = [a_i \mathbb{1}_{B \cap O_i}(x) + b_i \mathbb{1}_{B^c \cap O_i}(x)] d\mu(x),$$

with $a_i \geq 0$ and $b_i \geq 0$. Normalization of the ν_i gives the conditions

$$1 = a_i p(B \cap O_i) + b_i p(B^c \cap O_i). \tag{9}$$

Reproduction of the conditional probabilities gives the conditions

$$p^{\text{emp}}(B|O_i) = \frac{q(B \cap O_i)}{q(O_i)} = a_i \frac{p(B \cap O_i)}{p_i^{\text{emp}}}.$$

The latter gives

$$a_i = \frac{p_i^{\text{emp}}}{p_i} \frac{p^{\text{emp}}(B|O_i)}{p(B|O_i)}.$$

The normalization condition (9) becomes

$$1 = p_i^{\text{emp}} p^{\text{emp}}(B|O_i) + b_i p(B^c \cap O_i).$$

It has a positive solution for b_i because $p_i^{\text{emp}} p^{\text{emp}}(B|O_i) \leq 1$ and $p(B^c \cap O_i) > 0$. \square

4.3. The Hellinger Case

The optimal updates can be derived easily from Theorem 1. Double the partition by introduction of the following sets

$$O_i^+ = B \cap O_i \quad \text{and} \quad O_i^- = B^c \cap O_i.$$

They have prior probabilities $p_i^\pm = p(O_i^\pm)$. Corresponding prior measures μ_i^\pm are defined by

$$d\mu_i^\pm(x) = \frac{1}{p_i^\pm} \mathbb{I}_{O_i^\pm}(x) d\mu(x).$$

The empirical probability of the set O_i^+ is taken equal to $p_i^{\text{emp}} p^{\text{emp}}(B|O_i)$, that of O_i^- equals $p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)]$. The optimal update ν follows from Theorem 1 and is given by

$$d\nu(x) = \sum_i p_i^{\text{emp}} p^{\text{emp}}(B|O_i) d\mu_i^+(x) + \sum_i p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)] d\mu_i^-(x). \quad (10)$$

By construction it is

$$q(O_i^+) = p_i^{\text{emp}} p^{\text{emp}}(B|O_i) \quad \text{and} \quad q(O_i^-) = p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)].$$

One now verifies that $q(O_i) = p_i^{\text{emp}}$ and $q(B|O_i) = p^{\text{emp}}(B|O_i)$, which is the intended result.

4.4. The Bregman Case

Next consider the optimization with the quadratic Bregman divergence. Probability distributions ρ_i^\pm are defined by

$$\rho_i^\pm(x) = \frac{1}{|O_i^\pm|} \mathbb{I}_{O_i^\pm}(x).$$

Introduce the notations

$$\begin{aligned} r_i^+ &= \frac{p_i^+}{p_i^{\text{emp}} p^{\text{emp}}(B|O_i)}, \\ r_i^- &= \frac{p_i^-}{p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)]}, \\ \nu_i^\pm(x) &= (1 - r_i^\pm) \rho_i^\pm + r_i^\pm \mu_i^\pm(x). \end{aligned}$$

Then the condition for Theorem 2 to hold is that $\nu_i^\pm(x) \geq 0$ for all x, i . The optimal probability distribution ν is given by

$$\begin{aligned} \nu(x) &= \sum_i p_i^{\text{emp}} p^{\text{emp}}(B|O_i) \nu_i^+(x) + \sum_i p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)] \nu_i^-(x) \\ &= \sum_i [p_i^{\text{emp}} p^{\text{emp}}(B|O_i) - p_i^+] \rho_i^+ + \sum_i p_i^+ \mu_i^+ \\ &\quad + \sum_i [p_i^{\text{emp}} [1 - p^{\text{emp}}(B|O_i)] - p_i^-] \rho_i^- + \sum_i p_i^- \mu_i^- \\ &= \sum_i p_i^{\text{emp}} p^{\text{emp}}(B|O_i) [\rho_i^+ - \rho_i^-] \\ &\quad - \sum_i p_i^+ \rho_i^+ + \sum_i [p_i^{\text{emp}} - p_i^-] \rho_i^- \\ &\quad + \mu. \end{aligned}$$

5. Example

Assume that the prior probability distribution is binomial with parameters n, λ , where n is known with certainty. The probability mass function is given by

$$\mu(k) = \text{Prob}(X = k) = \binom{n}{k} \lambda^k (1 - \lambda)^{n-k} \quad k = 0, 1, 2, \dots, n.$$

The probability distribution and the value of the parameter λ are for instance the result of theoretical modeling of the experiment. Or they are obtained from a different kind of experiment.

The experiment under consideration yields accurate values for the probability p^{emp} of the two events $X = 1$ and $X = 2$. The problem at hand is to predict by extrapolation the probability of the event $X = k$ for other values of k . A fit of the data with a binomial distribution is likely to fail because two accurate data points are given to determine a single parameter λ . The binomial model can be misspecified.

The geometric approach followed in the present paper yields an update from the binomial distribution to another distribution, one which is reproducing the data. The update is conducted in an unbiased manner. Quite often one is tempted to replace the model, in the case of the binomial model, by a model with one extra free parameter.

Let us see what are the results of minimizing divergence functions. The probability space Ω is the set of integers $0, 1, 2, \dots, n$ equipped with the uniform measure. Choose events

$$O_1 = \{1\}, \quad O_2 = \{2\}, \quad O_3 = \Omega \setminus (O_1 \cup O_2).$$

This gives for $p_i := \text{Prob}(X \in O_i)$

$$p_1 = \mu(1) = n \lambda (1 - \lambda)^{n-1}, \quad p_2 = \mu(2) = \frac{1}{2} n (n - 1) \lambda^2 (1 - \lambda)^{n-2}, \quad p_3 = 1 - p_1 - p_2.$$

The optimal update according to Theorem 1, minimizing the Hellinger distance, is given by the probabilities

$$\nu(B) = \sum_i p_i^{\text{emp}} \mu(B|O_i).$$

In particular, the probability mass function $\nu(k) := \nu(\{k\})$ becomes

$$\begin{aligned} \nu(1) &= p_1^{\text{emp}}, \\ \nu(2) &= p_2^{\text{emp}}, \\ \nu(k) &= \frac{p_3^{\text{emp}}}{p_3} \mu(k) \quad \text{otherwise.} \end{aligned}$$

The optimal update according to Theorem 2, minimizing the quadratic Bregman divergence, is given by (7). The auxiliary measures μ_i, ρ_i , and ν_i have probability mass functions given by

$$\mu_i(k) = \rho_i(k) = \nu_i = \delta_{k,i} \quad \text{for } i = 1, 2,$$

and

$$\begin{aligned} \mu_3(k) &= (1 - \delta_{k,1})(1 - \delta_{k,2}) \frac{\mu(k)}{p_3}, \\ \rho_3(k) &= (1 - \delta_{k,1})(1 - \delta_{k,2}) \frac{1}{n - 2} \\ \nu_3(k) &= (1 - \delta_{k,1})(1 - \delta_{k,2}) \left[\left(1 - \frac{p_3}{p_3^{\text{emp}}}\right) \frac{1}{n - 2} + \frac{\mu(k)}{p_3^{\text{emp}}} \right]. \end{aligned}$$

The probability mass function $\nu(k) := \nu(\{k\})$ becomes

$$\begin{aligned} \nu(k) &= p_1^{\text{emp}} \nu_1(k) + p_2^{\text{emp}} \nu_2(k) + p_3^{\text{emp}} \nu_3(k) \\ &= p_1^{\text{emp}} \quad \text{if } k = 1, \\ &= p_2^{\text{emp}} \quad \text{if } k = 2, \\ &= \frac{p_3^{\text{emp}} - p_3}{n - 2} + \mu(k) \quad \text{otherwise.} \end{aligned}$$

The condition (6) is the requirement that all $\nu(k)$ are non-negative. Because the probabilities $\mu(k)$ can become very small this essentially means that p_3^{emp} should be larger than p_3 . The amount of probability missing in the empirical probabilities p_1^{emp} and p_2^{emp} is equally distributed over the remaining $n - 1$ points of Ω . On the other hand, when minimizing the Hellinger distance the excess or shortage of probability is compensated by multiplying all remaining probabilities by a constant factor.

A numerical comparison with $n = 20$ and $\lambda = 1/8$ is found in Figure 1. The empirical values are $p_1^{\text{emp}} = 0.15$ and $p_2^{\text{emp}} = 0.25$. The difference with the prior values $p_1 \simeq 0.19774$ and $p_2 \simeq 0.26836$ is made large enough to amplify the effects of the update.

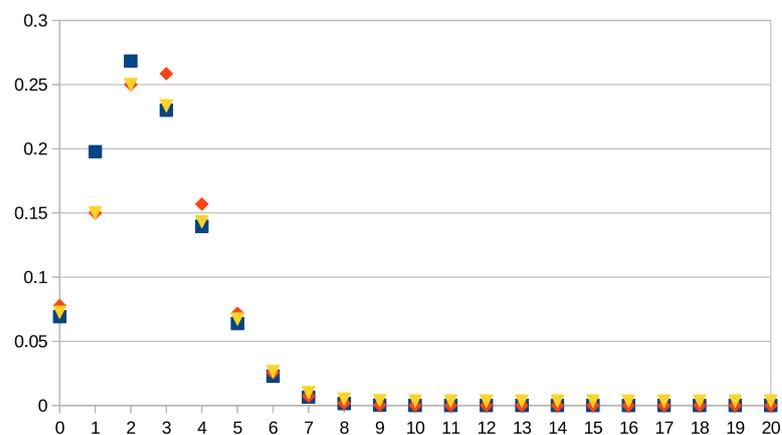


Figure 1. Probability as a function of the integer k running from 0 to 20, showing different updates of the binomial distribution with parameters $n = 20$ and $\lambda = 1/8$. The squares represent the binomial, the diamonds the update with the Hellinger distance, and the triangles the update with the square Bregman divergence. The empirical values are $p_1^{\text{emp}} = 0.15$ and $p_2^{\text{emp}} = 0.25$.

6. Summary

It is well known that the use of unmodified prior conditional probabilities is the optimal way for updating a probability distribution after new data become available. The update procedure minimizes the Hellinger distance between prior and posterior probability distributions. For the sake of completeness a proof is given in Theorem 1.

Alternatively, one can minimize the quadratic Bregman divergence instead of the Hellinger distance. The result is given in Theorem 2. The conservation of probability is handled in a different way in the two cases, either by multiplying prior probabilities with a suitable factor or by adding an appropriate term.

The example of Section 5 shows that the two update procedures have different effects and that neither of them may be satisfactory. This raises the question whether the present approach should be improved by choosing divergences other than Hellinger or Bregman.

In the present research, the work of Banerjee, Guo, and Wang [11] was considered as well. They prove that minimization of the Hellinger distance can be replaced by minimization of a Bregman divergence, without modifying the outcome. It is shown in Theorem 2 that, in a different context, the use of the Bregman divergence yields results quite distinct from those obtained by minimizing the Hellinger distance.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Originally published in Japanese by Iwanami Shoten, Tokyo, Japan, 1993; Oxford University Press: Oxford, UK, 2000.
2. Amari, S. *Information Geometry and Its Applications*; Springer Nature: Tokyo, Japan, 2016.
3. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Springer Nature: Basel, Switzerland, 2017.
4. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [[CrossRef](#)]
5. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [[CrossRef](#)]
6. Jeffrey, R. Alias Smith and Jones: The Testimony of the Senses. *Erkenntnis* **1987**, *26*, 391–399. [[CrossRef](#)]
7. Skyrms, B. The structure of Radical Probabilism. *Erkenntnis* **1997**, *45*, 285–297.
8. Csiszár, I. Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems. *Ann. Stat.* **1991**, *19*, 2032–2066. [[CrossRef](#)]
9. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158. [[CrossRef](#)]
10. Grünwald, P.D.; Dawid, A.P. Game Theory, Maximum Entropy, Minimum Discrepancy and robust Bayesian Decision Theory. *Ann. Stat.* **2004**, *32*, 1367–1433. [[CrossRef](#)]
11. Banerjee, A.; Guo, X.; Wang, H. On the Optimality of Conditional Expectation as a Bregman Predictor. *IEEE Trans. Inf. Theory* **2005**, *51*, 2664–2669. [[CrossRef](#)]
12. Frigyik, B.A.; Srivastava, S.; Gupta, M.R. Functional Bregman Divergences and Bayesian Estimation of Distributions. *IEEE Trans. Inf. Theory* **2008**, *54*, 5130–5139. [[CrossRef](#)]