CLASSIFICATION OF DOUBLE ATTRIBUTES VIA MUTUAL SUGGESTION BETWEEN A PAIR OF CLASSIFIERS

Kazuyuki HIRAOKA and Taketoshi MISHIMA

Dept. of Information and Computer Sciences, Saitama University
255 Shimo-Okubo, Saitama-shi, 338-8570, Japan
email: hira@me.ics.saitama-u.ac.jp

ABSTRACT

Real-world objects often have two or more significant attributes. For example, face images have attributes of persons, expressions, and so on. Even if you are interested in only one of those attributes, additional informations on auxiliary attributes can help recognition of the main one. The authors have been proposed a method for classification with double attributes. Its main idea is mutual suggestion of hints between a pair of classifiers. In the present paper, we will reexamine the task based on information geometry, and propose a new method of EM-like iterations. We will also show experimentally that the heuristic method in our previous work can be used as a good approximation of the new method which has solid theoretical basis.

1. INTRODUCTION

Pattern recognition on one attribute has been studied widely [3] , while that on double attribute is not sufficiently studied in spite of its importance. Real-world problems often have two or more interesting attributes. A typical example is face images which have attributes of persons, expressions, and so on. Even if you are interested in only one of those attributes, additional informations on auxiliary attributes can help recognition of the main one.

From this point of view, the authors have been proposed a method for classification with double attributes [2]. Its main idea is mutual suggestion of hints between a pair of classifiers: each classifier correspond to one attribute (Fig. 1). Since decisions of classifiers are not consistent in general, a mediation mechanism is necessary. In [2], a heuristic mediation is applied. It has an advantage that the result is obtained without iterative procedures, while its theoretical meaning is not clear.

In the present paper, another mediation is proposed. It has clear meaning based on information geometry [4]. EM-like iterative algorithm can be applied for calculation of the estimated joint probability of two attributes.

Both methods give almost same results of classification in our experiments. Thus, the heuristic mediation in [2] can be used as a good approximation of theoretically solid mediation.

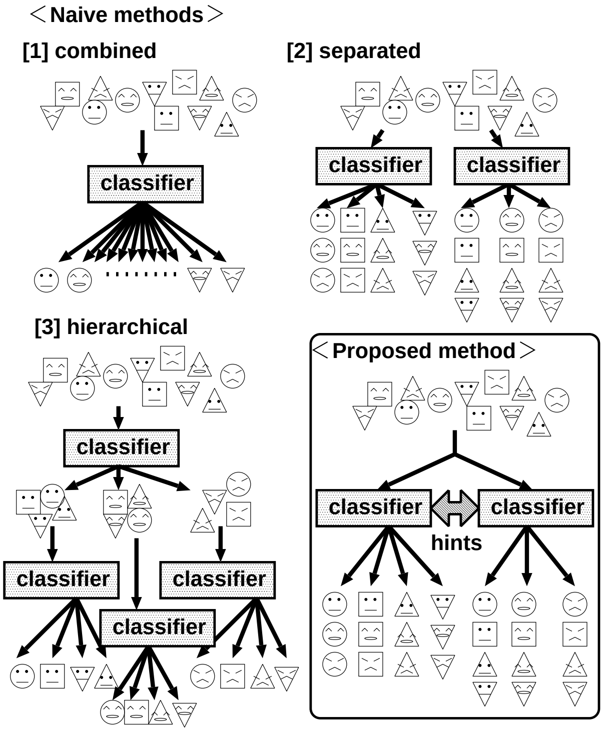

Figure 1: Comparison among naive methods and the proposed method

2. TASK

As training samples, n vector data $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n)$ are presented. In addition, double attributes (s, c) for each $\boldsymbol{x}$ are also presented:

$$\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T \in R^N, \quad (1)$$
$$s(t) \in \mathcal{S} = \{1, \cdots, S\}, \quad (2)$$
$$c(t) \in \mathcal{C} = \{1, \cdots, C\}, \quad (3)$$
$$(t = 1, \cdots, n), \quad (4)$$

where T denotes matrix transposition. Then, a new datum $\boldsymbol{x}$ is presented and estimation of its attributes (s, c) is required.

A solution for this task has been proposed in [1] for the case that the whole data can be approximated by the bilinear model. In our previous paper [2], different approach is discussed for general cases. We reexamine the task in the next section based on information geometry [4].

3. INFORMATION-GEOMETRIC APPROACH

Suppose that we have a pair of classifiers

$$\boldsymbol{f}(\boldsymbol{x}, c) = (f_1(\boldsymbol{x}, c), \ldots, f_S(\boldsymbol{x}, c)), \quad (5)$$
$$\boldsymbol{g}(\boldsymbol{x}, s) = (g_1(\boldsymbol{x}, s), \ldots, g_C(\boldsymbol{x}, s)), \quad (6)$$

where $f_s(\boldsymbol{x}, c)$ and $g_c(\boldsymbol{x}, s)$ are estimations of conditional probabilities $q(s|\boldsymbol{x}, c)$ and $q(c|\boldsymbol{x}, s)$, respectively. The classifier $\boldsymbol{f}$ is trained for combined input $(\boldsymbol{x}(t), c(t))$ and simple output $s(t)$, while $\boldsymbol{g}(t)$ is for $(\boldsymbol{x}(t), s(t))$ and $c(t)$. These $\boldsymbol{f}$ and $\boldsymbol{g}$ are blackboxes throughout the proposed method: arbitrary classifiers can be used for $\boldsymbol{f}$ and $\boldsymbol{g}$ as far as they output conditional (posterior) probabilities of classes (Fig. 1).

We want to estimate the marginal probabilities $q(s|\boldsymbol{x})$, $q(c|\boldsymbol{x})$, and/or the joint probability $q(s, c|\boldsymbol{x})$ based on the guessed conditional probabilities $\boldsymbol{f}$ and $\boldsymbol{g}$. As we will show soon, $\boldsymbol{f}$ and $\boldsymbol{g}$ are not consistent generally in the sense that there is no joint probability $q(s, c|\boldsymbol{x})$ whose conditional probability $q(s|\boldsymbol{x}, c)$ and $q(c|\boldsymbol{x}, s)$ are equal to $f_s(\boldsymbol{x}, c)$ and $g_c(\boldsymbol{x}, s)$, respectively. Hence, a mediation mechanism is required. In our previous work [2], a heuristic mediation is applied (appendix A). Now we will reexamine the mediation problem based on information geometry [4].

From the view of information geometry [4], the present situation is illustrated as follows. Let $\mathcal{P}$ be the space of probability distributions on $\mathcal{S} \times \mathcal{C}$:

$$\mathcal{P} \equiv \left\{ q(\cdot, \cdot) \;\middle|\; q(s, c) \geq 0, \quad \sum_{s=1}^{S} \sum_{c=1}^{C} q(s, c) = 1 \right\}. \quad (7)$$

From now on, the variable $\boldsymbol{x}$ is fixed and it is omitted for simple notations. In addition, $f_s(c)$ and $g_c(s)$ are denoted as $f(s|c)$ and $g(c|s)$ respectively. The estimated conditional probability $\boldsymbol{f}$ determines a submanifold $\mathcal{F}$ in $\mathcal{P}$:

$$\mathcal{F} \equiv \left\{ q(s, c) \in \mathcal{P} \;\middle|\; q(s|c) \equiv q(s, c) / \sum_{s'} q(s', c) = f(s|c) \right\}. \quad (8)$$

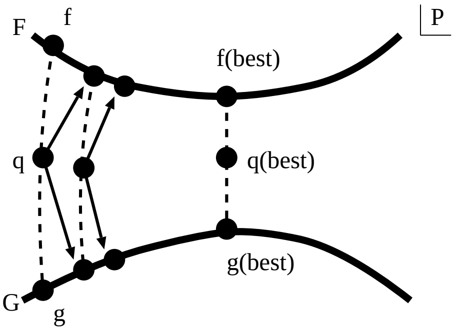

Figure 2: Information-geometric illustration of mediation

This $\mathcal{F}$ denotes the opinion of one classifier. The opinion $\mathcal{G}$ of the other classifier is also determined for $\boldsymbol{g}$ in the same way. Dimensions of $\mathcal{P}$, $\mathcal{F}$, $\mathcal{G}$ are $SC - 1$, $C - 1$, $S - 1$, respectively. In particular, $\dim \mathcal{P} > \dim \mathcal{F} + \dim \mathcal{G}$ for $S, C > 1$. This means $\mathcal{F} \cap \mathcal{G} = \emptyset$ in general. Hence, we want to find $q \in \mathcal{P}$ which is 'near' to both $\mathcal{F}$ and $\mathcal{G}$ (Fig. 2). As for measure of distance between two probability distributions $q(s, c)$ and $p(s, c)$, Kullback-Leibler divergence

$$D(p||q) = \sum_{s=1}^{S} \sum_{c=1}^{C} p(s, c) \log \frac{p(s, c)}{q(s, c)} \quad (9)$$

is appropriate since it is connected to asymptotic probability to judge q as p wrongly from samples.

In summary, two submanifolds $\mathcal{F}, \mathcal{G} \subset \mathcal{P}$ are presented, and we want to find $q \in \mathcal{P}$, $f \in \mathcal{F}$, $g \in \mathcal{G}$ which minimizes $d(q, f, g) = D(f||q) + D(g||q)$.

4. EM-LIKE ALGORITHM

The minimization problem in the previous section can be solved by alternative iteration of partial minimizations (Fig. 2).

e-step: For given q, find f, g which minimize $d(q, f, g)$.

m-step: For given f, g, find q which minimizes $d(q, f, g)$.

Calculation of partial minimizations are described in appendix B.

According to the above approach, we propose the following algorithm for mediation, i.e. estimation of joint probability $q(s, c)$ from guessed conditional probabilities $f(s|c)$ and $g(c|s)$.

1. Set the initial values $f(c) \leftarrow 1/C$ and $g(s) \leftarrow 1/S$.
2. Repeat the following two updates alternatively until convergence, and answer the final $q(s, c)$.

m-step:

$$q(s, c) \leftarrow \frac{1}{2} \Big\{ f(s|c) f(c) + g(c|s) g(s) \Big\} \quad (10)$$

e-step:

$$f(c) \leftarrow e^{-\zeta(c)}/Z, \quad (11)$$

$$g(s) \leftarrow e^{-\psi(s)}/\Psi, \quad (12)$$

where

$$\zeta(c) = \sum_{s=1}^{S} f(s|c) \log \frac{f(s|c)}{q(s,c)}, \quad (13)$$

$$\psi(s) = \sum_{c=1}^{C} g(c|s) \log \frac{g(c|s)}{q(s,c)}, \quad (14)$$

$$Z = \sum_{c=1}^{C} e^{-\zeta(c)}, \quad \Psi = \sum_{c=1}^{C} e^{-\psi(c)}. \quad (15)$$

5. EXPERIMENTS

The proposed method is experimentally compared with a heuristic method in [2] for basic artificial tasks. Parameters of the tasks are shown in Table 1, where I denotes the identity matrix.

Table 1: Parameters of experiments

number of classes	$(S,C) = (3,3)$
number of samples	$n = 50 \times S \times C = 450$
dimension of data $\boldsymbol{x}$	$N = 2$
within-class distribution of data $\boldsymbol{x}$ (within-class variance)	Gaussian ($V = 0.3^2 I$)
classifiers f_s, g_c	Fisher linear discriminant
dimension of projected discriminant space	$L = 1$
criterion of convergence	$\lvert\lvert q_{\text{new}} - q \rvert\rvert^2 < 10^{-24}$
max number of iterations	1000

Classifier $\boldsymbol{f}$ for these experiments consists of three "experts" $\boldsymbol{f}(\cdot,1), \boldsymbol{f}(\cdot,2), \boldsymbol{f}(\cdot,3)$ which correspond to $c = 1,2,3$, respectively. These experts are independently trained with only samples which have corresponding value of $c(t)$. Assuming that within-class distribution of $\boldsymbol{x}(t)$ in each (s,c) is Gaussian with a common unknown variance matrix V, we use Fisher linear discriminant[3] as each expert. Classifier $\boldsymbol{g}$ is also constructed similarly. Final decision of classification is obtained according to marginal probabilities $q(s) = \sum_c q(s,c)$ and $q(c) = \sum_c q(s,c)$ of estimated joint probability $q(s,c)$.

The results are shown in Fig. 3 and 4. Though the proposed method tends to answer a probability near to 0 or 1, both methods give almost same results of classification in our experiments. Thus, the heuristic method in [2] can be used as a good approximation of theoretically solid method.

6. CONCLUSION

In the present paper, A new method is proposed for classification of double attributes. Though it is based on the same idea as our previous method, it has clear meaning based on information geometry while the previous method is heuristic. It is experimentally shown that the previous method can be used as a good approximation of the new method which has solid theoretical basis.

There are many points which must be studied for establishment of the proposed method. A major one is comparison of merits and demerits with other approaches.

In section 5, "expert" classifiers are trained independently with only data which have corresponding attributes. This makes number of available samples for each expert smaller. In order to utilize informations in samples more efficiently, we have to adopt a classifier $\boldsymbol{f}$ which can deal with c more properly as a "hint". One natural idea is the use of single classifier with input vector $(x_1, \ldots, x_n, \delta_{1c}, \ldots, \delta_{Cc})^T$.

This work has been partly supported by JSPS (Japan Society for the Promotion of Science), 14580405.

7. REFERENCES

[1] Joshua B. Tenenbaum, "Separating style and content with bilinear models", *Neural Computation*, Vol. 12, pp. 1247–1283, 2000.

[2] K. Hiraoka and T. Mishima, "Complementary Discriminant Analysis for Classification of Double Attributes", *Proc. of ITC-CSCC 2002*, CD-ROM, 2002.

[3] Richard O. Duda, E. Hart, and David G. Stork, *Pattern classification*, 2nd ed., John Wiley & Sons, Inc., New York, 2001.

[4] Shun-ichi Amari, *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, 28, Springer-Verlag, 2nd printing, 1990.

A. BRIEF DESCRIPTION OF THE METHOD IN [2]

Notations in section 3 are used. For given $f(s|c)$ and $g(c|s)$, we can always find a pair of probabilities $F(s)$ and $G(c)$ which satisfy

$$F(s) = \sum_c f(s|c)G(c), \quad G(c) = \sum_s g(c|s)F(s), \quad (16)$$

because f and g can be viewed as a transition matrix of Markov chain on bipartite graph. In addition, when $\mathcal{F} \cap \mathcal{G} = \{q\}$, marginal probabilities of q are equal to F and G. From these considerations, the authors have been proposed to use F and G as estimated marginal probabilities [2].

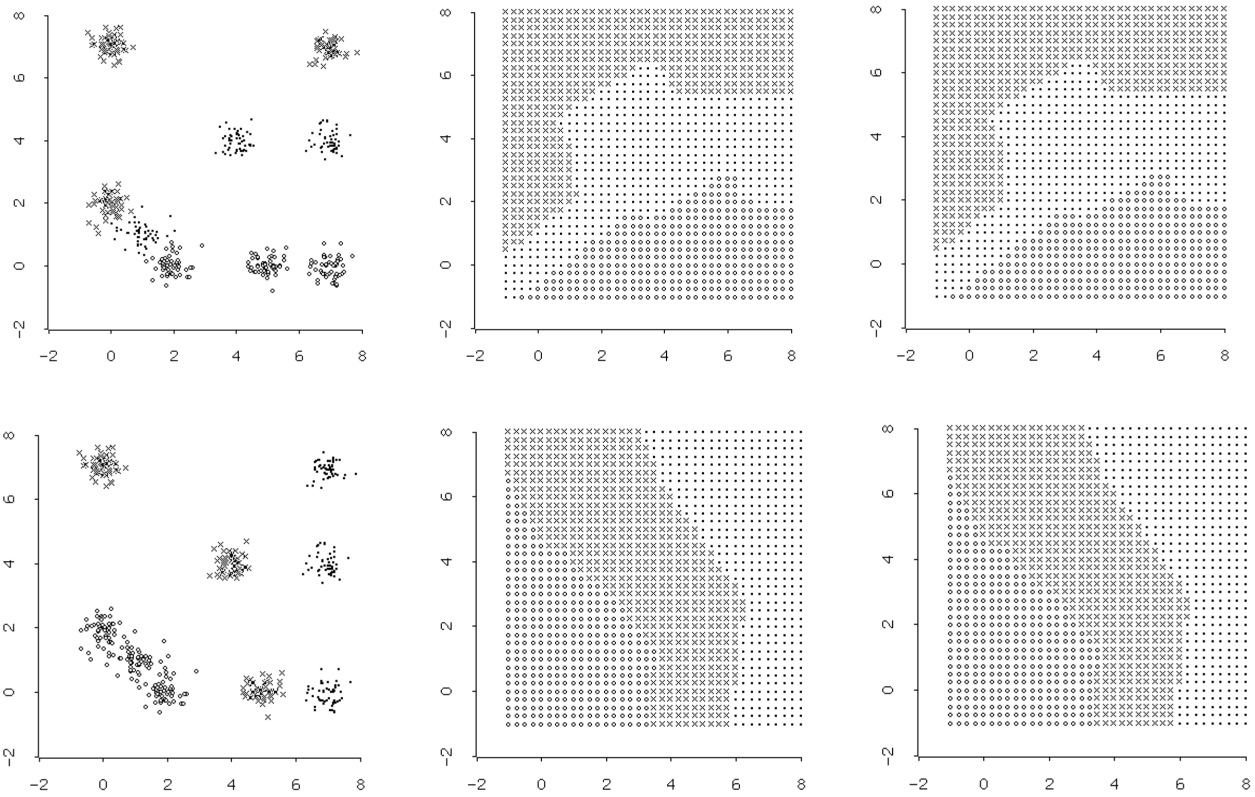

Figure 3: Experiment (matrix-type structure)
Upper: attribute s = 1(o), 2(x), 3(.). Lower: attribute c = 1(o), 2(x), 3(.).
Left: presented samples. Middle and Right: obtained boundaries of classification for proposed method and [2], respectively.

B. DERIVATION OF PROPOSED ALGORITHM

Before discussing partial minimizations in section 4, we show a lemma which suggests a good nature of the problem.

Lemma 1 *Both $\mathcal{F}$ and $\mathcal{G}$ are m-flat.*

Proof: We only show a proof on $\mathcal{F}$ because of symmetry. Suppose that $q(s|c) = q'(s|c) = \xi$ and $r(s,c) = \alpha q(s,c) + (1-\alpha)q'(s,c)$, where $0 \le \alpha \le 1$. They mean $q(s,c) = \xi q(c)$ and $q'(s,c) = \xi q'(c)$. Then

$$r(s|c) = \frac{\alpha q(s,c) + (1-\alpha)q'(s,c)}{\alpha q(c) + (1-\alpha)q'(c)} = \xi. \qquad (17)$$

Hence, $\alpha q + (1-\alpha)q' \in \mathcal{F}$ if $q, q' \in \mathcal{F}$ and $0 \le \alpha \le 1$. ■
Note that lemma 1 is trivial because we can rewrite $\mathcal{F}$ as

$$\mathcal{F} = \left\{ \sum_{c'=1}^{C} q_{c'} r_{c'}(\cdot,\cdot) \;\middle|\; q_c \ge 0, \quad \sum_{c'=1}^{C} q_{c'} = 1 \right\}, \qquad (18)$$

where $r_{c'}(s,c) = f(s|c')\delta_{cc'}$ and $\delta_{cc'} = 1(c = c'), 0(c \neq c')$.

E-step and M-step are calculated by next propositions.

Proposition 1 *Suppose that $q(s,c) \in \mathcal{P}$ and $\mathcal{F}$ with a conditional probability $f(s|c)$ are given. Then, the optimal $f \in \mathcal{F}$ which minimizes $D(f||q)$ is given by*

$$\begin{aligned} f(s,c) &= f(s|c)f(c), && (19)\\ f(c) &= e^{-\zeta(c)}/Z, && (20)\\ \zeta(c) &= \sum_{s=1}^{S} f(s|c)\log\frac{f(s|c)}{q(s,c)}, && (21)\\ Z &= \sum_{c=1}^{C} e^{-\zeta(c)}. && (22) \end{aligned}$$

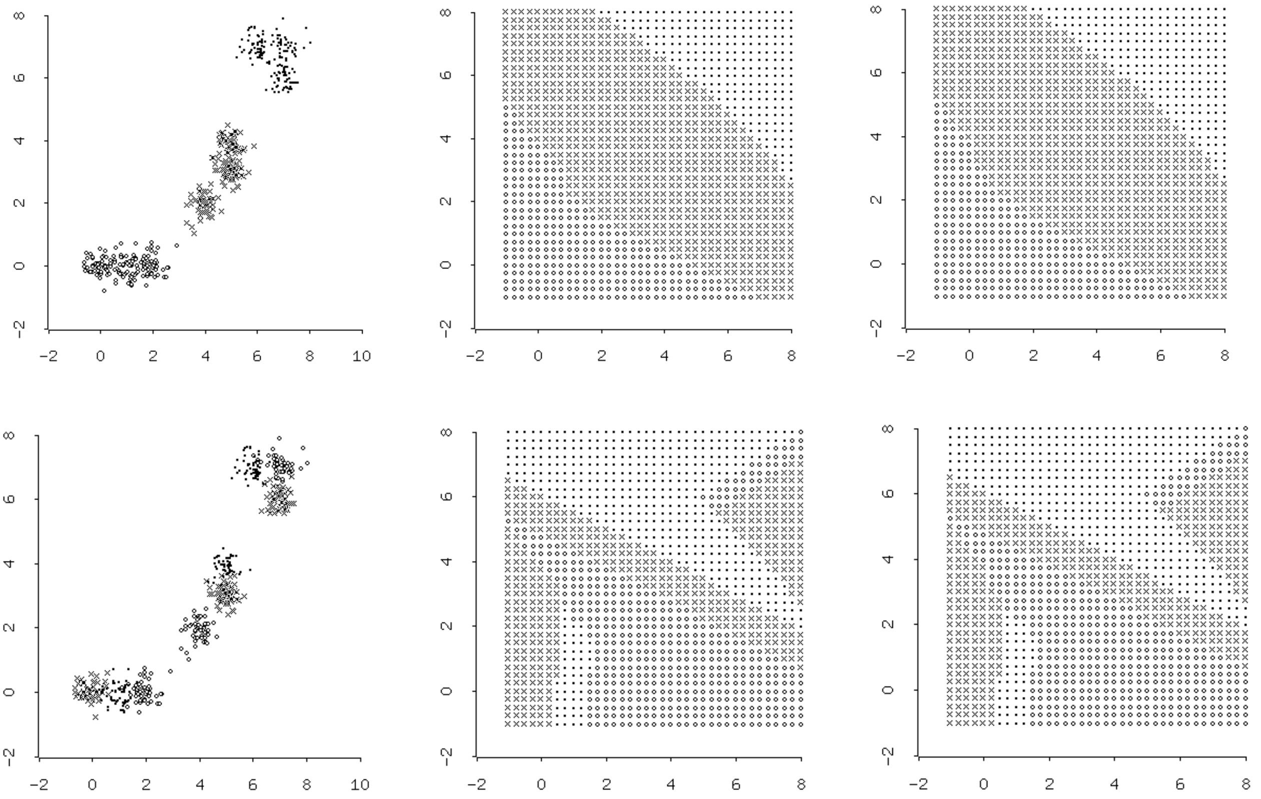

Figure 4: Experiment (cluster-type structure)
Upper: attribute s = 1(o), 2(x), 3(.). Lower: attribute c = 1(o), 2(x), 3(.).
Left: presented samples. Middle and Right: obtained boundaries of classification for proposed method and [2], respectively.

Proof: Since $f(s,c) = f(s|c)f(c)$ and $f(s|c)$ is given, we want to find the optimal probability $f(c)$ which minimizes

$$\begin{aligned} D(f||q) &= \sum_{s,c} f(s|c)f(c)\log\frac{f(s|c)f(c)}{q(s,c)} && (23)\\ &= \sum_c f(c)\sum_s f(s|c)\log\frac{f(s|c)}{q(s,c)} \\ &\quad + \sum_c \{f(c)\log f(c)\}\sum_s f(s|c), && (24)\\ &= \sum_c f(c)\zeta(c) + \sum_c f(c)\log f(c). && (25)\end{aligned}$$

In order to minimize $D(f||q)$ under the constraint $\sum_c f(c) = 1$, we define Lagrangian

$$L[f] \equiv \sum_c f(c)\zeta(c) + \sum_c f(c)\log f(c) - \lambda\left(\sum_c f(c) - 1\right). \quad (26)$$

From $\partial L/\partial\{f(c)\} = \zeta(c) + \log f(c) + 1 - \lambda = 0$, we obtain $f(c) \propto e^{-\zeta(c)}$ where proportional coefficient is determined by $\sum_c f(c) = 1$. This $f(s,c) = f(s|c)f(c)$ must be the minimum point since $\mathcal{F}$ is m-flat. ∎

Corresponding proposition for $g \in \mathcal{G}$ also holds, of course.

Proposition 2 *For arbitrary $f, g \in \mathcal{P}$,*

$$\arg\min_{q\in\mathcal{P}} d(q,f,g) = (f+g)/2. \quad (27)$$

Proof:

$$\begin{aligned} &d(q,f,g) && (28)\\ &= \sum f\log f + \sum g\log g - \sum (f+g)\log q && (29)\\ &= \sum (f+g)\log\frac{(f+g)/2}{q} \\ &\quad + \sum f\log\frac{f}{(f+g)/2} + \sum g\log\frac{g}{(f+g)/2} && (30)\\ &= 2D(h||q) + D(f||h) + D(g||h), && (31)\end{aligned}$$

where $h = (f+g)/2$. Since $D(f||h)$ and $D(g||h)$ are constants which are independent of q, $d(q,f,g)$ is minimized when $q = h$. ∎