A Geometric Interpretation of Darroch and Ratcliff's Generalized Iterative Scaling

Imre Csiszar

The Annals of Statistics, Vol. 17, No. 3. (Sep., 1989), pp. 1409-1413.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198909%2917%3A3%3C1409%3AAGIODA%3E2.0.CO%3B2-9

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1989, Vol. 17, No. 3, 1409–1413

A GEOMETRIC INTERPRETATION OF DARROCH AND RATCLIFF'S GENERALIZED ITERATIVE SCALING[1]

BY IMRE CSISZÁR

Hungarian Academy of Sciences

Darroch and Ratcliff's iterative algorithm for minimizing I-divergence subject to linear constraints is equivalent to a cyclic iteration of explicitly performable I-projection operations.

1. Introduction. The following problem often occurs in statistics: Given a probability distribution (PD) Q on a finite set $\mathscr{X}$ and a linear family

$$\mathscr{L} = \left\{P: \sum_x P(x) f_i(x) = a_i,\ i = 1, \dots, k\right\} \tag{1}$$

of PD's on $\mathscr{X}$, find the I-projection of Q on $\mathscr{L}$, i.e., that P^* which minimizes the (Kullback–Leibler) I-divergence

$$I(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{2}$$

subject to the linear constraints in (1).

In addition to being inherent to Kullback's "minimum discrimination information" approach [3, 4] and to maximum entropy methods (maximizing entropy is the same as minimizing divergence from the uniform distribution on $\mathscr{X}$), this numerical problem arises also in maximum likelihood estimation; cf. Section 3.

When $\mathscr{X}$ is a product space and $\mathscr{L}$ in the set of all PD's with given marginals of certain kinds, a very intuitive method known as iterative proportional fitting or iterative scaling is available for computing I-projection on $\mathscr{L}$. This method is extensively used in the analysis of contingency tables; cf., e.g., [3] and the references in [1] and [2].

I-projection on a general linear family (1) can be determined by "generalized iterative scaling" due to Darroch and Ratcliff [2]. The author and some of his colleagues have been wondering for some time whether this method also has an intuitive interpretation, within the framework of I-divergence geometry [1]. In this communication, using a suitable extension of the sample space $\mathscr{X}$, generalized iterative scaling is shown to be equivalent to a cyclic iteration of explicitly performable I-projection operations. Whereas this renders the convergence of Darroch and Ratcliff's algorithm a consequence of Theorem 3.2 of Csiszár [1], it should be noted that the proof of the latter—while more intuitive—is mathematically very similar to the original proof of the former [2].

Received June 1988.

[1]Research supported by Hungarian National Foundation for Scientific Research, Grant 1806.

AMS 1980 *subject classifications.* 62B10, 65K10.

Key words and phrases. Generalized iterative scaling, I-divergence geometry, minimum discrimination information, maximum entropy, maximum likelihood.

2. The result. Following Darroch and Ratcliff [2], we start with the observation that any linear family of PD's has a representation (1) with nonnegative functions f_i satisfying

$$\sum_{i=1}^{k} f_i(x) = 1 \quad \text{for every } x \in \mathscr{X}. \tag{3}$$

Henceforth we assume that $\mathscr{L}$ is so represented. Then, of course, the constants a_i in (1) are also nonnegative, and their sum must be 1. Unlike in [1], we do not require the a_i's to be positive. We do assume, without any loss of generality, that $Q(x) > 0$ for every $x \in \mathscr{X}$.

Introduce $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$ where $\mathscr{Y} = \{1, \dots, k\}$, let $\tilde{Q}$ be the PD on $\mathscr{Z}$ defined by

$$\tilde{Q}(x, i) = Q(x) f_i(x) \tag{4}$$

and let $\tilde{\mathscr{L}}$ be the linear family of those PD's $\tilde{P}$ on $\mathscr{Z}$ which are of form

$$\tilde{P}(x, i) = P(x) f_i(x) \tag{5}$$

and whose $\mathscr{Y}$-marginal equals $\mathbf{a} = (a_1, \dots, a_k)$.

Then there is a one-to-one correspondence between the I-projection P^* of Q on $\mathscr{L}$ and the I-projection $\tilde{P}^*$ of $\tilde{Q}$ on $\tilde{\mathscr{L}}$, namely

$$\tilde{P}^*(x, i) = P^*(x) f_i(x). \tag{6}$$

We recall Theorem 3.2 of Csiszár [1]: Let $\mathscr{E}$ be the intersection of linear families $\mathscr{E}_1, \dots, \mathscr{E}_k$ of PD's on a finite set and let Q be a PD to which there exists $P \in \mathscr{E}$ with $P \ll Q$. Then the sequence of PD's recursively defined by letting P_n be the I-projection of P_{n-1} on $\mathscr{E}_n$ (where $\mathscr{E}_n = \mathscr{E}_i$ if $n = mk + i$), with $P_0 = Q$, converges pointwise to the I-projection of Q on $\mathscr{E}$. We also recall, from the proof of this theorem, that $I(P_{n+1} \| P_n) \to 0$ as $n \to \infty$.

Now, let $\tilde{\mathscr{L}}_1$ be the family of those PD's on $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$ whose $\mathscr{Y}$-marginal is equal to $\mathbf{a}$ and let $\tilde{\mathscr{L}}_2$ be the family of the PD's on $\mathscr{Z}$ of form (5). Then $\tilde{\mathscr{L}} = \tilde{\mathscr{L}}_1 \cap \tilde{\mathscr{L}}_2$ and the above theorem applies whenever $\mathscr{L} \neq 0$. Thus the cyclic iteration of I-projections on $\tilde{\mathscr{L}}_1$ and $\tilde{\mathscr{L}}_2$ leads to a sequence of PD's $\tilde{P}_n \to \tilde{P}^*$. More exactly, let $\tilde{P}_0 = \tilde{Q}$ and, for $n = 0, 1, \dots,$ let $\tilde{P}_{2n+1}$ be the I-projection of $\tilde{P}_{2n}$ on $\tilde{\mathscr{L}}_1$ and $\tilde{P}_{2n+2}$ the I-projection of $\tilde{P}_{2n+1}$ on $\tilde{\mathscr{L}}_2$. Then

$$\lim_{n \to \infty} \tilde{P}_n = \tilde{P}^*. \tag{7}$$

The iteration yielding the PD's $\tilde{P}_n$ can be given explicitly. To this end, write $\tilde{P}_{2n}$ [which, by definition is of form (5)] as

$$\tilde{P}_{2n}(x, i) = P_n(x) f_i(x), \qquad n = 0, 1, \dots, \tag{8}$$

where $P_0 = Q$. It is well-known (and easy to check) that if a family of PD's is defined by a fixed marginal, I-projection on this family is obtained simply by scaling. Thus the I-projection of $\tilde{P}_{2n}$ [cf. (8)] on $\tilde{\mathscr{L}}_1$ is given by

$$\tilde{P}_{2n+1}(x, i) = P_n(x) f_i(x) \frac{a_i}{a_{i,n}}, \qquad a_{i,n} = \sum_x P_n(x) f_i(x). \tag{9}$$

Here we understand $\frac{0}{0} = 0$. Notice that $a_{i,n}$ is always positive if a_i is, provided that P_n is strictly positive on

$$\mathscr{X}^+ = \{x\colon f_i(x) > 0 \text{ for some } i \text{ with } a_i > 0\}. \tag{10}$$

The latter certainly holds for $n = 0$ and can be verified, by induction, for every n; cf. below.

Next, the I-projection $\tilde{P}_{2n+2}$ of $\tilde{P}_{2n+1}$ on $\tilde{\mathscr{L}}_2$ is obtained by minimizing the I-divergence of PD's of form (5) from $\tilde{P}_{2n+1}$, i.e., by minimizing

$$\sum_{x,i} P(x) f_i(x) \log \frac{P(x) f_i(x)}{P_n(x) f_i(x) a_i / a_{i,n}} = \sum_x P(x) \left[\log \frac{P(x)}{P_n(x)} + \sum_i f_i(x) \log \frac{a_{i,n}}{a_i} \right];$$

cf. (3). Write

$$R_{n+1}(x) = P_n(x) \prod_{i=1}^{k} \left(\frac{a_i}{a_{i,n}} \right)^{f_i(x)}, \tag{11}$$

where 0^0 is understood as 1. Then, denoting by c_{n+1} a constant that makes $c_{n+1} R_{n+1}$ a PD, the last sum equals $I(P \| c_{n+1} R_{n+1}) + \log c_{n+1}$. It follows that the minimizing P is $P_{n+1} = c_{n+1} R_{n+1}$ and the minimum, i.e., $I(\tilde{P}_{2n+2} \| \tilde{P}_{2n+1})$, equals $\log c_{n+1}$. Thus

$$\tilde{P}_{2n+2}(x, i) = P_{n+1}(x) f_i(x), \qquad P_{n+1}(x) = c_{n+1} P_n(x) \prod_{i=1}^{k} \left(\frac{a_i}{a_{i,n}} \right)^{f_i(x)}. \tag{12}$$

In particular, this completes the inductive proof of the positivity of P_n on $\mathscr{X}^+$.

By (6), (7) and (8) we have $P_n \to P^*$. Since $\log c_{n+1} = I(\tilde{P}_{2n+2} \| \tilde{P}_{2n+1}) \to 0$ and $P_n = c_n R_n$, this means that also $R_n \to P^*$. Finally, substituting $P_n = c_n R_n$ in (11) and (9), it follows that R_n satisfies the recurrence

$$R_{n+1}(x) = R_n(x) \prod_{i=1}^{k} \left(\frac{a_i}{b_{i,n}} \right)^{f_i(x)}, \qquad b_{i,n} = \sum_x R_n(x) f_i(x), \tag{13}$$

with $R_0 = P_0 = Q$.

But (13) in exactly the generalized iterative scaling algorithm of Darroch and Ratcliff [2], which, we believe, thereby has been given an intuitive understanding.

3. Discussion. As the PD's P_n in Section 2 are positive on $\mathscr{X}^+$ [cf. (10)] they are everywhere positive if $\mathscr{X}^+ = \mathscr{X}$ (in particular, if each a_i is positive). In this case it also follows [by induction, using (12)] that each P_n belongs to the exponential family

$$\mathscr{E} = \left\{ Q_{t_1,\ldots,t_k} \colon Q_{t_1,\ldots,t_k}(x) = c_{t_1,\ldots,t_k} Q(x) \exp \sum_{i=1}^{k} t_i f_i(x) \right\}. \tag{14}$$

If the I-projection $P^* = \lim_{n\to\infty} P_n$ is everywhere positive, it can be concluded that also $P^* \in \mathscr{E}$, whereas otherwise P^* belongs only to the closure of $\mathscr{E}$. Recall that P^* is everywhere positive iff there exists at least one everywhere positive $P \in \mathscr{L}$ (cf., e.g., Csiszár [1], Remark to Theorem 2.2).

Clearly, the exponential family (14) does not depend on the actual representation of $\mathscr{L}$ in the form (1). Changing this representation amounts only to a reparametrization of $\mathscr{E}$. In particular, $\mathscr{L}$ can always be represented in terms of strictly positive functions f_i satisfying (3), making $\mathscr{X}^+$ in (10) equal to $\mathscr{X}$. Hence, by the previous paragraph, the I-projection of Q on $\mathscr{L}$ always belongs to the closure of $\mathscr{E}$.

Darroch and Ratcliff [2] proved the convergence of their algorithm under the hypothesis that an everywhere positive $P \in \mathscr{L}$ exists and also showed that $P^* \in \mathscr{E}$. We see that their hypothesis is needed only for the latter, whereas $P_n \to P^*$ (or $R_n \to P^*$) always holds whenever $\mathscr{L} \neq \varnothing$.

It is a simple fact (dating back at least to Kullback [4]) that if a $P^* \in \mathscr{L} \cap \mathscr{E}$ exists, it satisfies

$$I(P\|Q) = I(P\|P^*) + I(P^*\|Q) \quad \text{for every } P \in \mathscr{L}. \tag{15}$$

This "Pythagorean identity" implies, in particular, that if $\mathscr{L} \cap \mathscr{E}$ is nonvoid, it consists of a single PD and this equals the I-projection of Q on $\mathscr{L}$. The result of [2] cited above is of interest also because it establishes that $\mathscr{L} \cap \mathscr{E}$ is, indeed, nonvoid if an everywhere positive $P \in \mathscr{L}$ exists at all. For a direct proof that under the last hypothesis the I-projection of Q on $\mathscr{L}$ belongs to $\mathscr{E}$ and that the Pythagorean identity (15) holds for the I-projection P^* of Q on $\mathscr{L}$ even if $P^* \notin \mathscr{E}$, see Csiszár ([1], Corollary 3.1, where $\mathscr{X}$ is not required to be finite).

Finally, let us briefly discuss the significance of computing I-projections for maximum likelihood estimation (cf. also [2] and [4]). For an i.i.d. sample from an unknown distribution on $\mathscr{X}$, the log-likelihood as a function of the underlying distribution Q can be written as

$$n \sum_x \hat{P}(x) \log Q(x),$$

where $\hat{P}$ in the empirical distribution of the sample. Comparing this with (2) shows that maximizing the likelihood over a given family of PD's Q called the model family, is equivalent to minimizing $I(\hat{P}\|Q)$.

Suppose that the model family is an exponential family as in (14). Let $\mathscr{L}$ be the linear family (1) defined by the same functions f_i, with constants a_i equal to the sample averages of the f_i's, i.e.,

$$a_i = \sum_x \hat{P}(x) f_i(x).$$

It is easy to see (and well-known) that all PD's in the exponential family (14) have the same I-projection P^* on $\mathscr{L}$. Thus by (15), with $P = \hat{P}$, we have

$$I\left(\hat{P}\|Q_{t_1,\dots,t_k}\right) = I(\hat{P}\|P^*) + I\left(P^*\|Q_{t_1,\dots,t_k}\right) \tag{16}$$

for each $Q_{t_1,\dots,t_k} \in \mathscr{E}$. It follows that if $P^* \in \mathscr{E}$, the left-hand side of (16) is minimized by $Q_{t_1,\dots,t_k} = P^*$, i.e., the (unique) ML estimate of the unknown distribution equals the I-projection of Q on $\mathscr{L}$. On the other hand, if P^* is not in

$\mathscr{E}$ only in its closure, the left-hand side of (16) can be made arbitrarily close to but is always larger than $I(\hat{P} \| P^*)$; hence in this case the ML estimate does not exist.

REFERENCES

[1] CSISZÁR, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

[2] DARROCH, J. N. and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43** 1470–1480.

[3] GOKHALE, D. V. and KULLBACK, S. (1978). *The Information in Contingency Tables*. Dekker, New York.

[4] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES
P.O. BOX 127
BUDAPEST 1364
HUNGARY