Iterative Algorithms with an Information Geometry Background

I. Csiszár, Rényi Institute, Budapest

Iterative projection algorithms that minimize Kullback information divergence (I-divergence) include iterative scaling, the EM algorithm, Cover's portfolio optimizing algorithm, etc. Such algorithms will be considered from an "information geometric" point of view; extensions to Bregman distances will also be addressed.

Overview

(i) Information geometry: I-divergence is an analogue of squared Euclidean distance, better suited than the latter for vectors with nonnegative components

(ii) I-projection to an intersection of affine families in $\mathbb{R}^k_+$ is the limit of iterated projections to these sets. Iterative scaling and generalized iterative scaling (or SMART) algorithms as special cases.

(iii) Minimum I-divergence between two convex sets via alternating projections. EM algorithm, portfolio optimization.

(iv) Iterative projection algorithms with Bregman distances. Dysktra's algorithm.

(v) Does the belief propagation algorithm admits an iterative projection interpretation?

Information Geometry

I-divergence (or relative entropy) for probability vectors $\underline{p} = (p_1, \ldots, p_k)$, $\underline{q} = (q_1, \ldots, q_k)$:

$$D(\underline{p}\|\underline{q}) = \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}$$

here $0 \log 0 = 0 \log \frac{0}{0} = 0, 0 \log \frac{t}{0} = +\infty (t > 0)$.

Key measure in information theory and statistics of the difference of probability distributions (though does not satisfy symmetry or triangle inequality).

Extension to arbitrary $\underline{p}, \underline{q}$ in $\mathbb{R}_+^k$

$$D(\underline{p}\|\underline{q}) = \sum_{j=1}^{k} \left[p_j \log \frac{p_j}{q_j} - p_j + q_j \right].$$

Nonnegative, equal to 0 if and only if $\underline{p} = \underline{q}$. Finite if and only if $\underline{p} \ll \underline{q}$ ($\underline{p}$ is dominated by $\underline{q}$), i.e,

$$S(\underline{p}) = \{i : p_i > 0\} \subseteq S(\underline{q}) = \{i : q_i > 0\}\,.$$

I-projection of $\underline{q} \in \mathbb{R}_+^k$ to a convex, closed set $C \subset \mathbb{R}_+^k$:

$$\Pi_C(\underline{q}) = \arg\min_{\underline{p} \in C} D\left(\underline{p} \| \underline{q}\right).$$

Well defined if C contains any $\underline{p} \ll \underline{q}$, thus always if $\underline{q}$ is strictly positive (and $C \neq \emptyset$). In many applications C consists of probability vectors, and $\underline{q}$ has equal components.
Then $\Pi_C(q)$ equals the maximizer of

$$H(\underline{p}) = -\sum_{j=1}^{k} p_j \log p_j$$

subject to $\underline{p} \in C$, the maximum entropy distribution in C.
For $\underline{q}$ strictly positive, $\log \underline{q} = (\log q_1, \ldots, \log q_k)$ is regarded as dual representation of $\underline{q}$.

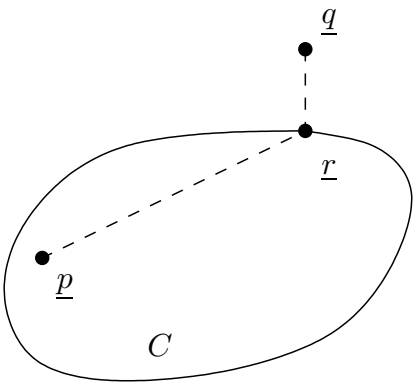
Lemma 1. If $\underline{q}$ is strictly positive and C contains some strictly positive $\underline{p}$, the I-projection $\Pi_C(\underline{q})$ is the unique strictly positive $\underline{r} \in C$ at which $\underline{r}$ the vector $\log \underline{q} - \log \underline{r}$ is normal to C,

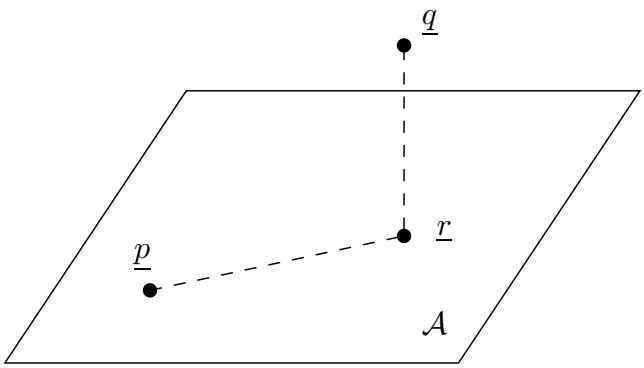
$$\left\langle \underline{p} - \underline{r}, \log \underline{q} - \log \underline{r} \right\rangle \leq 0 \qquad \forall \underline{p} \in C.$$

For C equal to an affine family $\mathcal{A}$, here the equality must hold, thus $\log \underline{q} - \log \underline{r} \perp \mathcal{A}$.

An affine family (also called mixture family) is the intersection with $\mathbb{R}^k_+$ of an affine subspace of $\mathbb{R}^k$.

Proof: $\log \underline{q} - \log \underline{r} = -\nabla_{\underline{p}} D\left(\underline{p} \| \underline{q}\right) /_{\underline{p}=\underline{r}}$.





Simple identity, analogue of cosine theorem:

$$D\left(\underline{p}\|\underline{r}\right)+D\left(\underline{r}\|\underline{q}\right)=D\left(\underline{p}\|\underline{q}\right)+\left\langle \underline{p}-\underline{r},\log\underline{q}-\log\underline{r}\right\rangle$$

for $\underline{p}\in\mathbb{R}_+^k$ arbitrary, $\underline{q}$, $\underline{r}$ strictly positive.

This and Lemma 1 imply the next assertion under the hypotheses of Lemma 1. Importantly, these hypotheses can be dispensed with.

Lemma 2. For any closed convex $C\subset\mathbb{R}_+^k$ and $\underline{q}\in\mathbb{R}_+^k$ such that $\Pi_C(\underline{q})$ exists, it equals the unique $\underline{r}\in C$ satisfying

$$D\left(\underline{p}\|\underline{q}\right)\geq D\left(\underline{p}\|\underline{r}\right)+D\left(\underline{r}\|\underline{q}\right)\quad\forall\underline{p}\in C.$$

For C equal to an affine family the equality holds: Pythagorean theorem.

Dual counterpart of affine families in $\mathbb{R}_+^k$:
An exponential family consists of those strictly positive $\underline{r}\in\mathbb{R}_+^k$ whose dual representation belongs to a given affine subspace of $\mathbb{R}^k$.

Key concept in statistics, of course with attention restricted to probability vectors. Typically, one row of A below is $(1, \ldots, 1)$, then $\underline{r} \in \mathcal{E}$ implies $\alpha \underline{r} \in \mathcal{E}$ for all $\alpha > 0$, and restricting attention to the probability vectors in $\mathcal{E}$ means a simple normalization.

Mutually orthogonal affine and exponential families:

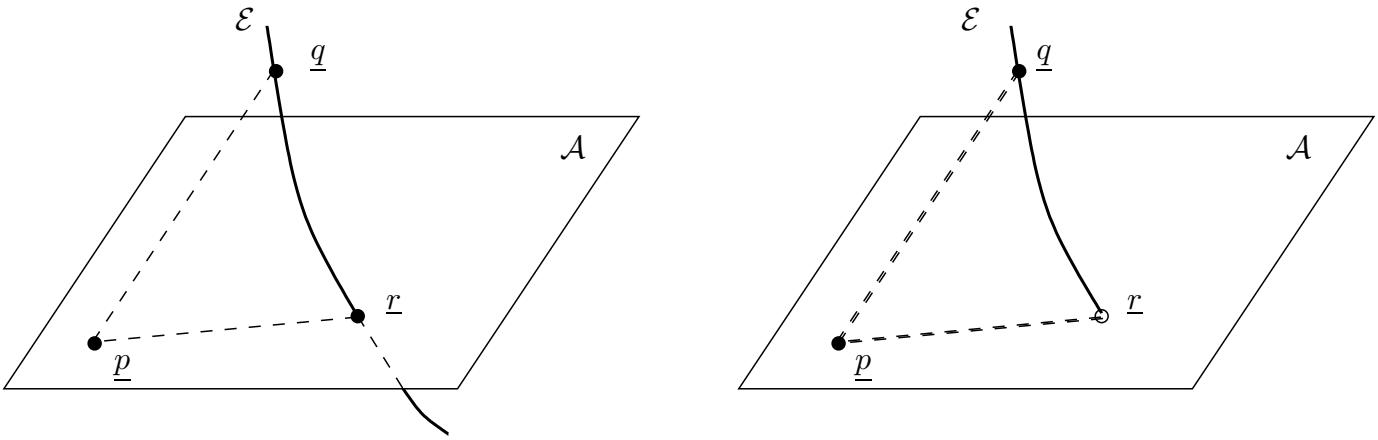
$$\mathcal{A} = \left\{ \underline{p} \in \mathbb{R}^k_+ : A\underline{p} = \underline{b} \right\} \qquad \left(A \; k \times \ell \text{ matrix}, \; \underline{b} \in \mathbb{R}^k \right)$$

$$\mathcal{E} = \left\{ \underline{r} : \log \underline{r} = \underline{\vartheta} A + \underline{c}, \underline{\vartheta} \in \mathbb{R}^\ell \right\} \qquad \left(\underline{c} \in \mathbb{R}^k \text{ fixed} \right)$$

Lemma 3. If $\mathcal{A}$ above contains some strictly positive $\underline{p}$ then $C \cap \mathcal{E} = \{\underline{r}\}$ with $\underline{r}$ satisfying the Pythagorean identity

$$D\left(\underline{p} \| \underline{q}\right) = D\left(\underline{p} \| \underline{r}\right) + D\left(\underline{r} \| \underline{q}\right) \qquad \forall \underline{p} \in \mathcal{A}, \underline{q} \in \mathcal{E}.$$

If no strictly positive $\underline{p} \in C$ exists (but $C \neq 0$) then $C \cap cl(\mathcal{E}) = \{\underline{r}\}$, again with $\underline{r}$ satisfying the above identity. Thus in both cases, $\underline{r} = \Pi_{\mathcal{A}}(\underline{q})$ for all $\underline{q} \in \mathcal{E}$.



When $\mathcal{A} \cap \mathcal{E} = \{\underline{r}\}$, the Pythagorean identity implies that $\underline{r}$ is both the I-projection of any $\underline{q} \in \mathcal{E}$ to $\mathcal{A}$ and the reverse I-projection

$$\arg\min_{\underline{q} \in \mathcal{E}} D\left(\underline{p} \| \underline{q}\right)$$

of any $\underline{p} \in \mathcal{A}$ to $\mathcal{E}$. In statistics, a maximum likelihood (ML) estimate equals the reverse I-projection of the empirical distribution to the model family. Thus, algorithms to compute I-projections to affine families also serve to compute ML estimates for exponential families.

Iterative I-projection algorithm

Let $\mathcal{A}_1, \ldots, \mathcal{A}_m$ be affine families in $\mathbb{R}_+^k$. Let Π_n denote I-projection to $\mathcal{A}_n$ or, if $n > m$, to $\mathcal{A}_i$ with $1 \leq i \leq m$, $i \equiv n \pmod{m}$.

Algorithm: $\underline{p}^0 = \underline{q}$, $\underline{p}^n = \Pi_n \left(\underline{p}^{n-1} \right)$, $n \geq 1$.

Theorem 1: (iterative I-projection).
If $\mathcal{A} = \bigcap_{i=1}^{m} \mathcal{A}_i$ contains some $\underline{p} \ll \underline{q}$ then the algorithm is well-defined, and $\underline{p}^n \to \Pi_{\mathcal{A}}(\underline{q})$ as $n \to \infty$.

Note that no strict positivity assumption is needed, though strict positivity of $\underline{q}$ always suffices if $\mathcal{A} \neq 0$.

Proof: (Csiszár 1975, using ideas of Ireland-Kullback 1968)

For $\underline{p} \in \mathcal{A}$ with $\underline{p} \ll \underline{q}$

$$D\left(\underline{p}\|\underline{p}^{n-1}\right) = D\left(\underline{p}\|\underline{p}^{n}\right) + \left(\underline{p}^{n}\|\underline{p}^{n-1}\right), n = 1, 2, \ldots$$

by Lemma 2 (Pythagorean theorem).
All terms are finite by induction. Summing for $n = 1, \ldots, N$ gives

$$(*) \quad D\left(\underline{p}\|\underline{q}\right) = D\left(\underline{p}\|\underline{p}^{0}\right) = \\ = D\left(\underline{p}\|\underline{p}^{N}\right) + \sum_{n=1}^{N} D\left(\underline{p}^{n}\|\underline{p}^{n-1}\right)$$

Hence $D\left(\underline{p}\|\underline{p}^{N}\right)$ is nonincreasing, $\underline{p}^{N}$ is bounded, $\underline{p}^{N_i} \to \underline{p}^{*}$ for a subsequence, clearly $\underline{p}^{*} \ll \underline{q}$. Also, $D\left(\underline{p}^{n}\|\underline{p}^{n-1}\right) \to 0$, thus the subsequences $\underline{p}^{N_i+1}, \ldots, \underline{p}^{N_i+m-1}$ also converge to $\underline{p}^{*}$, and therefore $\underline{p}^{*} \in \mathcal{A}$.

Then $D\left(\underline{p}^*\|\underline{p}^N\right)$ is nonincreasing,

$$\lim_{N\to\infty} D\left(p^*\|p^N\right) = \lim_{i\to\infty} D\left(p^*\|\underline{p}^{N_i}\right) = 0$$

thus $\underline{p}^N \to \underline{p}^*$, and $(*)$ with $\underline{p} = \underline{p}^*$ gives

$$D\left(\underline{p}^*\|\underline{q}\right) = \sum_{n=1}^{\infty} D\left(\underline{p}^n\|\underline{p}^{n-1}\right).$$

Hence $\underline{p}^* = \Pi_{\mathcal{A}}(\underline{q})$ follows.

Conceptually important: $D\left(\underline{p}\|\underline{p}^N\right)$ is nonincreasing for each $\underline{p} \in \mathcal{A}$. Fejér monotonicity of the sequence $\underline{p}^N$ relative to $\mathcal{A}$.

Special case: iterative scaling

The following widely used algorithm goes back to Kruithof (1937) and Deming-Stephan (1940).

Given a $k \times \ell$ matrix $Q = (q_{ij})$ with nonnegative elements, and vectors $\underline{b} = (b_1, \ldots, b_k) \in \mathbb{R}_+^k$ $\underline{c} = (c_1, \ldots, c_\ell) \in \mathbb{R}_+^\ell$ as required marginals of an adjustment of Q.

Algorithm: $P^0 = Q$.
For $n \geq 1$ define $P^n = \left(p_{ij}^n\right)$ by

$$p_{ij}^n = \begin{cases} p_{ij}^{n-1} \dfrac{b_i}{p_{i\bullet}^{n-1}} & n \text{ odd} \\ p_{ij}^{n-1} \dfrac{c_j}{p_{\bullet j}^{n-1}} & n \text{ even} \end{cases}$$

with notation

$$p_{i\bullet} = \sum_{j=1}^{\ell} p_{ij} \quad p_{\bullet j} = \sum_{i=1}^{k} p_{ij}$$

Ireland-Kullback (1968) pointed out that the iterative steps are I-projections, in step n the I-projection of P^{n-1} is taken to the affine family of matrices whose first marginal in $\underline{b}$, respectively second marginal is $\underline{c}$.

Corollary: P^n converges to the I-projection of Q to the family of matrices $P = (p_{ij})$ with marginals $\underline{b}$ and $\underline{c}$, provided some matrix in this family satisfies $p_{ij} = 0$ whenever $q_{ij} = 0$.

Multidimensional iterative scaling is also widely used. Given a d-dimensional array of nonnegative numbers

$$q_{i_1 \ldots i_d}, \quad 1 \leq i_1 \leq k_1, \ldots, 1 \leq i_d \leq k_d,$$

and m marginals of perhaps different dimensions of a requested adjustment of this array, the adjustment is performed by a cyclic iteration. In each step one of the m marginals is adjusted by scaling, which amounts to an I-projection as before. As noted previously, such algorithms are suitable also for computing ML estimates, specifically for log-linear models in the analysis of multidimensional contingency tables.

Generalized iterative scaling

The iterative I-projection theorem appears of limited value for computing I-projection to $\mathcal{A} = \bigcap_{i=1}^{m} \mathcal{A}_i$ when no explicit formulas are available for I-projection to the individual $\mathcal{A}_i$'s.

Actually, with a twist, the theorem admits to design an effective algorithm for computing I-projections to any affine family $\mathcal{A}$ that consist of vectors $\underline{p}$ with $\sum_{j=1}^{k} p_j =$ constant (as in most applications).

In that case, it may be assumed that in the representation

$$\mathcal{A} = \left\{ \underline{p} : A\underline{p} = \underline{b} \right\},$$

the columns of the $k \times \ell$ matrix $A = (a_{ij})$ are probability vectors.

Then, for any $\underline{p}, \underline{q}$ in $\mathbb{R}_+^k$ and matrices P, Q defined by $p_{ij} = p_j a_{ij}$, $q_{ij} = q_j a_{ij}$, it holds that $D\left(\underline{p} \| \underline{q}\right) = D\left(P \| Q\right)$. Hence, minimizing $D\left(\underline{p} \| \underline{q}\right)$ for $\underline{p} \in \mathcal{A}$ is equivalent to minimizing $D\left(P \| Q\right)$ for matrices $P \in \tilde{\mathcal{A}}_1 \cap \tilde{\mathcal{A}}_2$

$$\tilde{\mathcal{A}}_1 = \left\{ P = (p_{ij}) \in \mathbb{R}_+^{k\ell} : p_{\bullet j} = b_j, j = 1, \ldots, \ell \right\}$$
$$\tilde{\mathcal{A}}_2 = \Big\{ P \in \mathbb{R}_+^{k\ell} : p_{ij} = p_{\bullet j} a_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, \ell \Big\}.$$

Applying the I-projection algorithm to the latter problem, I-projections to $\tilde{\mathcal{A}}_1$ are performed simply by scaling, and I-projections to $\tilde{\mathcal{A}}_2$ also admit an explicit formula. Simple calculation gives that the $2n$'th step (the n'th projection to $\tilde{\mathcal{A}}_2$) results in a matrix $\left(p_j^n a_{ij}\right)$ where

$$p_j^n = p_j^{n-1} \prod_{i=1}^{\ell} \left(\frac{b_i}{\left\langle \underline{a}_i, \underline{p}^{n-1} \right\rangle} \right)^{a_{ij}}, \ n \geq 1, \ p_j^0 = q_j.$$

Corollary: This iteration gives $\underline{p}^n \to \Pi_{\mathcal{A}}(\underline{q})$, provided $\mathcal{A}$ contains some $\underline{p}$ with $\underline{p} \ll \underline{q}$.

The last iteration is known as generalized iterative scaling (Darroch-Ratcliff 1972) or SMART algorithm. The above derivation is from Csiszár 1989.

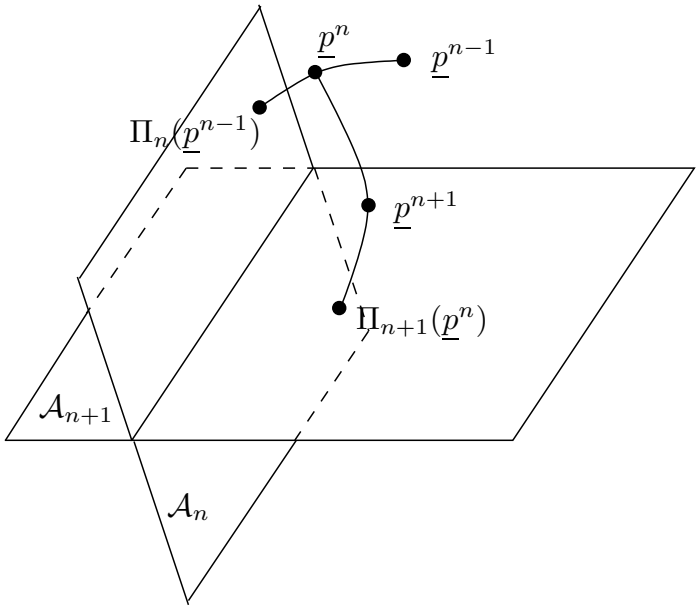
Remark: In cases when iterative scaling applies, generalized iterative scaling does not give the same algorithm. For example, to compute the I-projection of a matrix Q to the family of matrices with given marginals $\underline{b}, \underline{c}$, a generalized iterative scaling algorithm is $P^0 = Q$,

$$p_{ij}^n = p_{ij}^{n-1} \sqrt{\frac{b_i}{p_{i\bullet}^{n-1}} \cdot \frac{c_j}{p_{\bullet j}^{n-1}}}, \; n \geq 1.$$

Extensions of Theorem 1

(i) Non-cyclic iteration also works

(ii) Relaxation: $\underline{p}^n$ is a convex combination of $\underline{p}^{n-1}$ and $\Pi_n(\underline{p}^{n-1})$, in dual representation



(iii) Affine families are replaceded by halfspaces or any closed convex sets C_i. I-projection iteration still converges to some $\underline{p}^* \in \bigcap_{i=1}^m C_i$ (if nonempty), thus solves the convex feasibility problem. A modified algorithm is available (Dysktra 1985) that converges to I-projection to $\bigcap_{i=1}^m C_i$.

Alternating minimization algorithm

Given: Convex compact subsets B, C of $\mathbb{R}^k_+$ such that there exist $\underline{p} \in B$, $\underline{q} \in C$ with $\underline{p} \ll \underline{q}$. Denote $S(C) = \{i : q_i > 0 \text{ for some } \underline{q} \in C\}$

Algorithm: $\underline{q}_0 \in C$ arbitrary with $S(\underline{q}_0) = S(C)$.

$\underline{p}^n = \arg\min_{\underline{p} \in B} D(\underline{p} \| \underline{q}^{n-1})$ I-projection,

$\underline{p}^n = \arg\min_{\underline{q} \in C} D(\underline{p}^n \| \underline{q})$ reverse I-projection.

The latter minimizer need not be unique if $S(\underline{p}^n)$ is a proper subset of $S(C)$; then either minimizer may be taken as $\underline{q}^n$.

Theorem 2 (alternating minimization):

$$D(\underline{p}^1\|\underline{q}^0) \geq D(\underline{p}^1\|\underline{q}^1) \geq \\ \geq D(\underline{p}^2\|\underline{q}^1) \geq D(\underline{p}^2\|\underline{q}^2) \geq \dots$$

converges to $D_{\min} = \min_{\underline{p}\in B, \underline{q}\in C} D(\underline{p}\|\underline{q})$, and $\underline{p}^n$ converges to a limit $\underline{p}^* \in B$ such that for all accumulation points $\underline{q}^*$ of the sequence $\underline{q}^n$

$$D(\underline{p}^*\|\underline{q}^*) = D_{\min}$$

The proof shows that the sequence $\underline{p}^n$ is Fejér monotone relative to the set of those $\underline{p} \in B$ to which there exists $\underline{q} \in C$ with $D\left(\underline{p}\|\underline{q}\right) = D_{\min}$: $D\left(\underline{p}\|\underline{p}^n\right)$ is nonincreasing for each such $\underline{p}$.

Theorem 2 is relevant also for the convex feasibility problem: $B \cap C \neq \emptyset$ if and only if $D\left(\underline{p}^n\|\underline{q}\right) \to 0$, or when both $\underline{p}^n$ and $\underline{q}^n$ converge to the same limit.

Moreover, if $B \cap C \neq \emptyset$ then the sequence $\underline{p}^n$ is Fejér monotone relative to $B \cap C$.

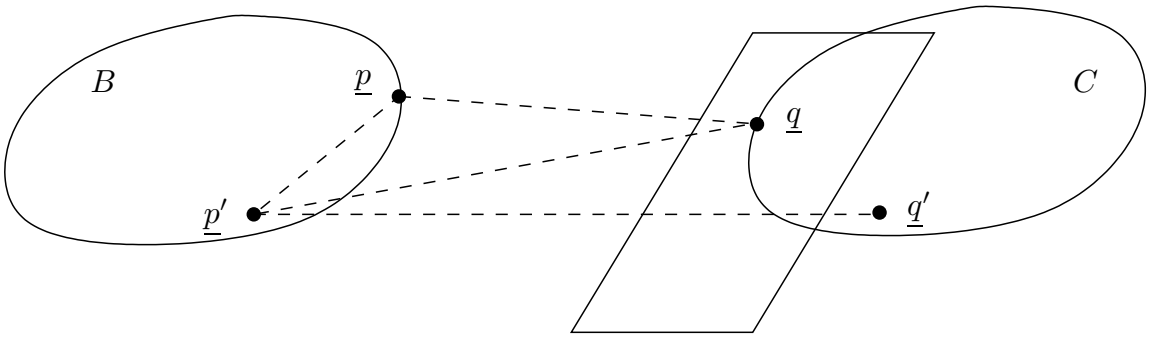
Key geometric ingredients of the proof

(i) Three points property:
for $\underline{p} \in B$, $\underline{q} \in C$, $\underline{r} = \Pi_B(\underline{q})$

$$D\left(\underline{p}\|\underline{q}\right) \geq D\left(\underline{p}\|\underline{r}\right) + D\left(\underline{r}\|\underline{q}\right) \qquad \text{(Lemma 2)}$$

(ii) Four points property:
for $\underline{p} \in B$ with $\min_{\underline{p} \in B} D\left(\underline{p}\|\underline{q}\right)$ finite, $\underline{q}$ attaining that minimum, and any $\underline{p'} \in B$, $\underline{q'} \in C$

$$D\left(\underline{p'}\|\underline{q'}\right) + D\left(\underline{p'}\|\underline{p}\right) \geq D\left(\underline{p'}\|\underline{q}\right) \qquad \text{(new)}$$



Theorem 2 is a special case of a more general result concerning alternating minimization of any function $d(x, y)$ of abstract variables (x, y) (Csiszár - Tusnády 1984).

There, the three and four point properties were postulates on $d(x, y)$ that involved an auxiliary nonnegative function $\delta(x, y)$.
In the present case $d = D$, also $\delta = D$. The same properties hold also for $d = \tilde{D}$ defined by

$$\tilde{D}\left(\underline{p}\|\underline{q}\right) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i} \qquad \left(\underline{p} \in \mathbb{R}_+^k, \underline{q} \in \mathbb{R}_+^k\right)$$

which $\tilde{D}$ can take also negative values (again with $\delta = D$).

Alternating minimization

and generalized iterative scaling

Generalized iterative scaling (GIS) is based on iterating I-projections to the affine families $\tilde{\mathcal{A}}_1$ and $\tilde{\mathcal{A}}_2$ defined there. Due to the special form of $\tilde{\mathcal{A}}_1$, I-projection and reverse I-projection to $\tilde{\mathcal{A}}_1$ are identical operations, given by scaling.

Hence GIS is equivalent to alternating minimization for $B = \tilde{\mathcal{A}}_2$, $C = \tilde{\mathcal{A}}_1$.

In the inconsistent case when

$$\mathcal{A} = \left\{\underline{p} \in \mathbb{R}_+^k : A\underline{p} = \underline{b}\right\} = \emptyset$$

the original GIS convergence result does not apply, but Theorem 2 does: when $\tilde{\mathcal{A}}_1 \cap \tilde{\mathcal{A}}_2 = \emptyset$, alternating minimization achieves, in the limit, the minimum I-divergence between $B = \tilde{\mathcal{A}}_2$ and $C = \tilde{\mathcal{A}}_1$.

Equivalently, the sequence $\underline{p}^n$ defined by GIS converges to a limit $\underline{p}^*$ that minimizes $D\left(A\underline{p}||\underline{b}\right)$.

EM algorithm

Goal: estimate an unknown probability distribution from partially observable drawings when ML estimation from fully observable drawings would be "easy". Specifically, ML estimate of the marginal distribution governing the partial observations is required.

Model: Let C be a family of probability distributions on pairs $(i,j), 1 \le i \le k, 1 \le j \le \ell$, given by probability matrices $Q = (q_{ij})$. For N independent drawings $(i_1,j_1),\ldots,(i_N,j_N)$ from an unknown $Q \in C$, only the second components are observed.
Denote $(i_1,\ldots,i_N) = \underline{x}$, $(j_1,\ldots,j_N) = \underline{y}$, let $\widehat{P}$ and $\underline{\widehat{p}}$ be the empirical distributions of the (unobserved) full sample $(\underline{x},\underline{y})$ and of the (observed) partial sample $\underline{y}$:

$$\widehat{p}_{ij} = \frac{1}{n}\left|\{t : i_t = i, j_t = j\}\right|,$$

$$\widehat{p}_j = \frac{1}{n}\left|\{t : j_t = j\}\right| = \widehat{p}_{\bullet j}.$$

Algorithm: $Q_0 \in C$ arbitrary, for $n \geq 1$

E-step: $P^n = E_{Q^{n-1}}\left(\widehat{P}|\underline{y}\right)$,

conditional expectation

M-step: $Q^n = \arg\min_{Q \in C} D\left(P^n \| Q\right)$,

ML estimate pretending that P^n is the empirical distribution of $(\underline{x}, \underline{y})$.

By simple algebra, $p_{ij}^n = q_{ij}^n \frac{\widehat{p}_j}{q_{\bullet j}^{n-1}}$.

Hence

$$D\left(P^n \| Q^{n-1}\right) = D\left(\widehat{p} \| \underline{q}^{n-1}\right)$$

(where $\underline{q}^{n-1} = \left(q_{\bullet 1}^{n-1}, \ldots, q_{\bullet \ell}^{n-1}\right)$) and P^n equals the I-projection of Q^{n-1} to the family

$$B = \left\{P = (p_{ij}) : p_{\bullet j} = \widehat{p}_j, j = 1, \ldots, \ell\right\}.$$

By definition, Q^n is the reverse I-projection of P^n to C; by assumption, it is "easy" to compute.

It follows that the sequence of divergences

$$D\left(P^n \| Q^{n-1}\right) = D\left(\widehat{p} \| \underline{q}^{n-1}\right)$$

is always nonincreasing.

In ideal case, it converges to

$$\min_{P \in B, Q \in C} D\left(P \| Q\right) = \min\left\{D\left(\widehat{p} \| \underline{q}\right) : \underline{q} \right.$$
$$\left. \text{marginal of some } Q \in C\right\}.$$

If the second minimum is attained by a unique $\underline{q}^*$ (which is then the ML estimate of the marginal distribution governing the observable part of the sample), the convergence of $D\left(\widehat{p} \| \underline{q}^n\right)$ to this minimum implies $\underline{q}^n \to \underline{q}^*$.

By Theorem 2, this ideal situation always obtains if the set C of feasible distributions is convex and closed, provided the initial $Q^0 \in C$ has maximal support. Moreover, in this case the sequence P^n is always convergent.

In practice, the EM algorithm is widely used even though the set of feasible distributions is seldom convex. Then the iteration may get stuck at a local minimum, but running the algorithm several times with different initial Q^0 typically leads to satisfactory results.

Example (decomposition of mixtures)

Suppose a sample $\underline{y} = (j_1, \ldots, j_N)$ with empirical distribution $\underline{\widehat{p}}$ has been drawn from a mixture distribution $\underline{q} = \sum_{i=1}^{k} p_i \underline{r}_i$ where

$$\underline{r}_i = (r_{i1}, \ldots, r_{i\ell}),\ i = 1, \ldots, k$$

are known probability vectors, and the weight vector $\underline{p} = (p_1, \ldots, p_k)$ is unknown.

To compute the ML estimate of $\underline{p}$ via the EM algorithm, pretend $\underline{y}$ consists of the second components of drawings $(i_1, j_1), \ldots, (i_N, j_N)$ from an unknown member of the family C of distributions $Q = (q_{ij})$ with $q_{ij} = p_i r_{ij}$.

The reverse I-projection of any $P = (p_{ij})$ to this family C is given by $q_{ij} = p_{i\bullet} r_{ij}$, hence the M-step of the EM algorithm is $q_{ij}^n = p_{i\bullet}^{n-1} r_{ij}$.

Combining this with the explicit form of the E-steps given before, we obtain that

$$q_{ij}^n = p_i^n r_{ij} \text{ with } p_i^n = p_i^{n-1} \sum_{j=1}^{\ell} r_{ij} \frac{\widehat{p}_j}{\sum_{h=1}^{k} p_h^{n-1} r_{hj}}$$

As the above set C is convex and closed, now the "ideal case" of the EM algorithm obtains. It follows that $\underline{p}^n$ defined by the last iteration converges to a limit $\underline{p}^*$ such that $\underline{q}^* = \sum p_i^* \underline{r}_i$ minimizes $D\left(\underline{\widehat{p}} \| \underline{q}\right)$; this holds even in the non-identifiable case when different weight vectors $\underline{p}$ may yield the same mixture $\underline{q}$.

Portfolio optimization

A portfolio is a probability vector $\underline{p} = (p_1, \ldots, p_k)$, where p_i represents the fraction of the total invested capital invested in stock $i \in \{1, \ldots, k\}$.

If one dollar invested in stock i returns X_i dollars by the end of the investment period, the portfolio with $E\left(\log \sum_{i=1}^{k} p_i X_i\right) = \text{maximum}$ achieves the highest long-term gain (subject to certain simplifying hypotheses).

Algorithm to compute the log-optimal portfolio when the joint distribution of the random variables X_i is known (Cover 1984):

$\underline{p}^0$: any strictly positive distribution,

$$p_i^n = p_i^{n-1} E\left(\frac{X_i}{\sum_{j=1}^{k} p_j^{n-1} X_j}\right), \quad n \geq 1$$

Effectively the same iteration as that for decomposition of mixtures, with the role of $\widehat{p}_j$ and r_{ij} now played by $X_i(\omega)$ and $p(\omega)$ (assuming the random variables X_i are defined on a finite sample space (Ω, p)).

The difference is the absence of an analogue of the assumption that the vectors $\underline{r}_i$ are probability vectors. Still, a convergence proof can be given along the same lines, via alternating minimization. In particular, $\underline{p}^n$ converges to a log-optimal $\underline{p}^*$ even in the case when uniqueness of the log-optimal portfolio does not hold.

Bregman projections

Key geometric properties of I-divergence are shared by Bregman distances.
Let f be a convex, lower semicontinuous function on $\mathbb{R}^k$, strictly convex and differentiable in the interior of

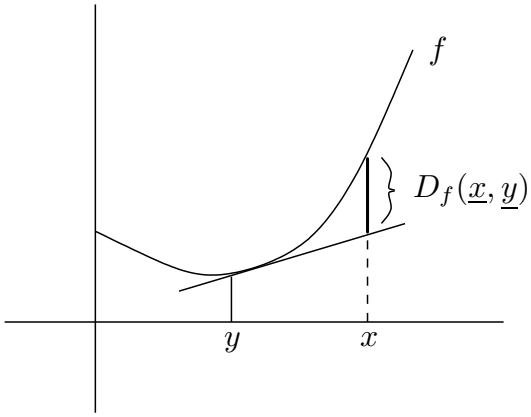
$$\mathrm{dom} f = \{x : f(\underline{x}) < +\infty\}\,.$$

This interior $S = \mathrm{int}(\mathrm{dom} f)$, called the zone of f, is assumed nonempty.

The Bregman distance (1967) of an $\underline{x} \in \mathrm{dom} f$ from an $\underline{y} \in S$ is

$$D_f\left(\underline{x}, \underline{y}\right) = f(\underline{x}) - f(\underline{y}) - \left\langle (\nabla f)(y),\, \underline{x} - \underline{y} \right\rangle$$

Finite, nonnegative, equals 0 only if $\underline{x} = \underline{y}$.



Bregman projection of $\underline{y} \in S$ to a closed convex set C intersecting dom$f$:

$$\Pi_{C,f}(\underline{y}) = \arg\min_{\underline{x} \in C} D_f(\underline{x}, \underline{y})$$

In practice, mostly separable Bregman distances are used, thus with

$$f(\underline{x}) = \sum_{i=1}^{k} f_i(x_i), \quad \text{typically } f_1 = \cdots = f_k$$

Examples

$f(\underline{x})$	$\mathrm{dom} f$	$D_f(\underline{x},\underline{y})$
$\frac{1}{2}\|\underline{x}\|^2$	$\mathbb{R}^k$	$\|\underline{x}-\underline{y}\|^2$
$\frac{1}{2}\|\underline{x}\|^2$	$\mathbb{R}^k_+$	$\|\underline{x}-\underline{y}\|^2$
$\sum_{i=1}^k (x_i \log x_i - x_i)$	$\mathbb{R}^k_+$	$D\left(\underline{x}\|\underline{y}\right)$
$-\sum_{i=1}^k \log x_i$	$\mathrm{int}\left(\mathbb{R}^k_+\right)$	$\sum_{i=1}^k \left[\log \frac{y_i}{x_i} + \frac{x_i}{y_i} - 1\right]$
$\sum_{i=1}^k [x_i \log x_i + (1-x_i)\log(1-x_i)]$		
	$[0,1]^k$	$\sum_{i=1}^k \left[x_i \log \frac{x_i}{y_i} + (1-x_i)\log \frac{1-x_i}{1-y_i}\right]$

For separable f, the definition of Bregman distances and projections could be extended to $\underline{y}$ on the boundary of $S = \mathrm{int}(\mathrm{dom} f)$ (as for I-divergences) but no natural extension appears possible for general f. Below, the second variable of $D_f(\underline{x},\underline{y})$ is always restricted to $\underline{y} \in S$.

Dual representation of $\underline{y} \in S$: the gradient vector $(\nabla f)(\underline{y})$.

With this, the analogues of Lemma 1-2 hold for Bregman distances, provided the convex set C (intersecting S) satisfies zone consistency: $\Pi_{C,f}(\underline{y}) \in S$ for each $\underline{y} \in S$.

May fail, for example, if $f(\underline{x}) = \frac{1}{2}\|\underline{x}\|^2$ for $\underline{x} \in \mathbb{R}^k_+$ and $+\infty$ otherwise. The (Euclidean) projection $\Pi_{C,f}(\underline{y})$ of a strictly positive $\underline{y} \in \mathbb{R}^k$ to a convex or even affine set $C \subset \mathbb{R}^k_+$ need not be strictly positive even though C contains strictly positive vectors.

One reason why I-divergence is preferable to Euclidean distance for vectors in $\mathbb{R}^k_+$.

Assume henceforth that f is steep: for $\underline{y}_n$ in S approaching a boundary point of S, the dual representations $(\nabla f)(\underline{y}_n)$ can not be bounded.

Then each C intersecting S is zone consistent, and the analogues of Lemmas 1-2 hold.

In particular, for an affine family $\mathcal{A}$ (intersection with dom$f$ of an affine subspace of $\mathbb{R}^k$) with $\mathcal{A} \cap \mathcal{S} \neq \emptyset$, the Bregman projection $\Pi_{\mathcal{A},f}(\underline{y})$ of an $\underline{y} \in S$ equals the unique $\underline{z} \in \mathcal{A} \cap S$ satisfying $(\nabla f)(\underline{y}) - (\nabla f)(z) \perp \mathcal{A}$, or equivalently

$$D_f(\underline{x}, \underline{y}) = D_f(\underline{x}, \underline{z}) + D_f(\underline{z}, \underline{y}) \quad \forall \underline{x} \in \mathcal{A}$$

Convex functions f with the postulated properties are called of Legendre type. The convex conjugate

$$f^*(\underline{a}) = \sup_{\underline{x}} [\langle \underline{a}, \underline{x} \rangle - f(\underline{x})]$$

of such f is also of Legendre type, and the map $\underline{y} \to (\nabla f)(\underline{y})$ is one-to-one from S onto the interior of the effective domain of f^*. The inverse map is $\underline{a} \to (\nabla f^*)(\underline{a})$.

Analogue of an exponential family: those $\underline{y} \in S$ whose dual representation $(\nabla f)(\underline{y})$ belongs to a given affine subspace of $\mathbb{R}^k$.

An f-exponential family orthogonal to the affine family

$$\mathcal{A} = \{\underline{x} \in \text{dom} f : A\underline{x} = \underline{b}\}$$

is

$$\begin{aligned}\mathcal{E}_f &= \left\{\underline{y} : (\nabla f)(\underline{y}) = \underline{\vartheta} A + \underline{c} \text{ for some } \underline{\vartheta} \in \mathbb{R}^\ell\right\} \\ &= \left\{\underline{y} = (\nabla f^*)(\underline{\vartheta} A + \underline{c}) \text{ for some } \underline{\vartheta} \in \mathbb{R}^\ell\right\} \\ &\qquad (\underline{c} \in \mathbb{R}^k, \text{ fixed})\end{aligned}$$

Such orthogonal families $\mathcal{A}$ and $\mathcal{E}_f$ intersect in a singleton $\{\underline{z}\}$, whenever $\mathcal{A} \cap S \neq \emptyset$, and the Pythagorean identity

$$D_f(\underline{x}, \underline{y}) = D_f(\underline{x}, \underline{z}) + D_f(\underline{z}, \underline{y}) \quad \forall \underline{x} \in \mathcal{A}, \underline{y} \in \mathcal{E}_f$$

holds (partial analogue of Lemma 3).

Iterative Bregman projection algorithm

Analogue of iterative I-projection algorithm. Let $\mathcal{A}_1, \ldots, \mathcal{A}_m$ be affine families in $\mathrm{dom} f$, each intersecting $S = \mathrm{int}(\mathrm{dom} f)$. Denote Π_n Bregman projection to $\mathcal{A}_n$, with cyclic convention. Let $\underline{y} \in S$.

Algorithm: $\underline{x}^0 = \underline{y}$, $\underline{x}^n = \Pi_n(\underline{x}^{n-1})$, $n \geq 1$.

Analogue of Theorem 1:

$$\underline{x}^n \to \Pi_{\mathcal{A},f}(\underline{y}), \text{ if } \bigcap_{i=1}^{m} \mathcal{A}_i = \mathcal{A} \neq \emptyset.$$

Identical proof, if some technical conditions (obvious for I-divergence) are satisfied.

Conditions, for $\underline{x}_n, \underline{y}_n$ in S

(i) $D_f(\underline{x}, \underline{y}_n)$ bounded for some $\underline{x} \in \text{dom} f \Rightarrow \underline{y}_n$ bounded

(ii) $\underline{y}_n \to \underline{y}$, hypothesis of $(i) \Rightarrow \underline{y} \in \text{dom} f$, $D_f(\underline{y}, \underline{y}_n) \to 0$

(iii) $\underline{x}_n \to \underline{x} \in \text{dom} f$, $\underline{y}_n \to \underline{y} \in \text{dom} f$, $D_f(\underline{x}_n, \underline{y}_n) \to 0 \Rightarrow \underline{x} = \underline{y}$

Extensive literature of Bregman projection algorithms, goes back to Bregman 1967.

Censor-Lent 1981 formalized "standard" technical conditions for convergence proofs. Function satisfying them: Bregman functions.

Bauschke-Borwein 1997 proposed a slightly different class, called Bregman/Legendre functions: Legendre functions f satisfying the conditions on the previous page.

When $\underline{x} \in S$ or $\underline{y} \in S$ then (ii), (iii) hold for all Legendre functions; for separable Legendre function f, (ii), (iii) hold is full. Moreover, (i) implies that $\mathrm{dom} f^*$ is open; in turn, if $\mathrm{dom} f^*$ is open then (i) holds for $\underline{x} \in S$, and for all $\underline{x} \in \mathrm{dom} f$ if f is separable.

The results of Bauschke-Borwein 1997 imply convergence of the iterated Bregman projections $\underline{x}^n = \Pi_n(\underline{x}^{n-1})$ to the Bregman projection $\Pi_{\mathcal{A},f}(\underline{y})$ if f is Bregman/Legendre, in particular if f is separable Legendre with $\mathrm{dom} f^*$ open; or if $\mathcal{A} = \bigcap_{i=1}^{m} \mathcal{A}_i$ intersects S and f is any Legendre function with $\mathrm{dom} f^*$ open. It appears unknown whether the last condition can be dispensed with.

Many extensions are available, similar to those for iterative I-projections.

Byrne-Censor 2000 studied alternating minimization of Bregman distances. Established three and four points properties of $D_f(\underline{x}, \underline{y})$ if it is jointly convex in $(\underline{x}, \underline{y})$, and proved analogue of Theorem 2 for jointly convex Bregman distances.

Dysktra algorithm for Bregman projections

Let $C_1, \ldots, C_m$ be closed convex sets in $\mathbb{R}^k$, each intersecting S, let $\underline{y} \in S$. Denote Bregman projection to C_n by Π_n (cyclic convention). Iteration of these projection converges to a point in $\bigcap_{i=1}^{m} C_i = C$ (subject to technical conditions) but in general not to $\Pi_{C,f}(\underline{y})$.

Modified algorithm, "Bregman version" of the algorithms of Dysktra 1983, 1985 designed for Euclidean and I-projections: $\underline{x}^n$ is the projection to C_n of a "deflected version" of $\underline{x}^{n-1}$ obtained in step $n-1$; $\underline{x}^{n-1}$ is deflected by adding to its dual representation a vector $\underline{z}^{n-m}$ determined in step $n-m$, after the previous projection to C_n.

Algorithm: $\underline{x}^0 = \underline{y}$, the initial values of the deflecting vector are $\underline{0}$. For $n \geq 1$,

$$\begin{aligned}\underline{x}^n =& \Pi_n \left((\nabla f^*) \left((\nabla f)(\underline{x}^{n-1}) + \underline{z}^{n-m} \right) \right) \\ \underline{z}^n =& \underline{z}^{n-m} - (\nabla f)(\underline{x}^n) + (\nabla f)(\underline{x}^{n-1})\end{aligned}$$

The convergence $\underline{x}^n \to \Pi_{C,f}(\underline{y})$ was proved by

- Censor-Reich 1998 for the case when the sets C_i are halfspaces
- Bregman-Censor-Reich 1999, and Bauschke-Lewis 2000 for general convex, closed sets C_i, under different technical conditions. Both assumed f Legendre; the hypotheses of Bregman et al. included $\mathrm{dom} f$ closed, those of Bauschke-Lewis $\mathrm{dom} f^* = \mathbb{R}^k$.

The last hypothesis was needed to make sure that the iteration is well defined. Bregman et al. were able to dispense with it by a slight modification of the algorithm.

Iterative projections and belief propagation

Belief propagation or sum-product algorithm admits efficient decoding at transmission rates close to the Shannon limit. Originally an exact algorithm with finite number of steps (Pearl 1988), more powerful current versions are iterative, more heuristically than mathematically justified, remarkably successful in practice.

Given: a function of form $g(\underline{x}) = \prod_{k=1}^{K} g_k(\underline{x})$ of $\underline{x} = (x_1, \ldots, x_N)$, where N is very large but each factor $g_k(\underline{x})$ depends only on a few coordinates, say x_i with $i \in G_k$.

Goal: compute the one-dimensional marginals of g, thus sum $g(\underline{x})$ for all $\underline{x}$ with one components x_i fixed; often, the components x_i are binary.

The belief propagation (BP) algorithm involves iterated "message exchanges" between nodes of the factor graph a bipartite graph whose nodes represent the variables x_i, $1 \leq i \leq N$ and the factors g_k, $1 \leq k \leq K$.

Each message sent by a node is computed from previous received messages by product and sum operations.

Recent development:
Regalia-Walsh 2008 gave an equivalent description of the BP algorithm as a cyclic iteration resembling Dyskra's algorithm, but more complex than the latter, involving both I-projections and reverse I-projections.

It remains to be seen whether this relationship of BP to information projections algorithms is merely formal or will provide essential insights contributing to a better understanding of the mathematical intricacies of BP.