Universität Augsburg

Institut für Mathematik

Christoph Gietl, Fabian Reffel

Accumulation Points of the Iterative Proportional Fitting Procedure

Accumulation Points of the Iterative Proportional Fitting Procedure*

Christoph Gietl†   Fabian P. Reffel‡

6 July 2012

Abstract The asymptotic behavior of the iterative proportional fitting procedure (IPF procedure) is analyzed comprehensively. Given a nonnegative matrix as well as row and column marginals the IPF procedure generates a sequence of matrices, called the IPF sequence, by alternately fitting rows and columns to match their respective marginals. We prove that the IPF sequence has at most two accumulation points. They originate as the limits of the even-step subsequence, and of the odd-step subsequence. The well-known IPF convergence criteria are then retrieved easily. Our proof is based on Csiszár's and Tusnády's (1984) results on the interplay of the I-divergence geometry and alternating minimization procedures.

Keywords Iterative proportional fitting · Accumulation points · Alternating minimization · I-divergence · I-projection · Distributions with given marginals

Mathematics Subject Classification (2000) 68W40 · 62H17 · 62B10

---

†C. Gietl
  Institut für Mathematik
  Universität Augsburg
  86135 Augsburg, Germany
  Tel.: +49-821-5982160
  Fax: +49-821-5982280
  christoph.gietl@math.uni-augsburg.de
‡F. P. Reffel
  Institut für Mathematik
  Universität Augsburg
  86135 Augsburg, Germany
  Tel.: +49-821-5982202
  Fax: +49-821-5982280
  fabian.reffel@math.uni-augsburg.de

1. Introduction

Across different disciplines the following problem often occurs: Given a nonnegative $k \times \ell$ matrix A and given positive marginals, find a nonnegative $k \times \ell$ matrix B which fulfills the given marginals and is biproportional to the matrix A. The iterative proportional fitting procedure (IPF procedure) aims to solve this problem by iteratively fitting columns and rows to match their respective given marginals. It is in use for problems such as calculating maximum likelihood estimators in graphical log-affine models (Lauritzen 1996, Chapter 4.3.1), ranking webpages (Knight 2008) or determining passenger flows (McCord et al 2010).

Our motivation for this paper arose from the apparent lack of knowledge of applicability of Csiszár's and Tusnády's (1984) work to the analysis of IPF procedure. Having read the recent literature on the IPF procedure there is evidence to suggest that the cyclic behavior of the IPF procedure is still unknown. Whenever necessary and sufficient conditions for convergence of the whole sequence of the IPF procedure are not fulfilled, the procedure's even-step and odd-step subsequences still converge. Jiroušek and Vomlel (1994) stated this as a conjecture, Vomlel (2004) commented the case by saying the IPF procedure "tends to come in cycles" and Zhang et al (2008) similarly wrote that it "rather goes into cycles". Most recently, Pukelsheim (2012) proved the cyclic behavior under an "IPF conjecture", which remained open in the general case. The proof sought-after has already been given by Csiszár and Tusnády (1984). They investigated alternating minimization of I-divergence which is equivalent to the IPF procedure as we know due to Csiszár (1975) and Cramer (2000). We elaborate the equivalence between alternating minimization and iterative fitting, apply the proofs of Csiszár and Tusnády to the IPF procedure and simplify them dramatically.

The IPF procedure has been introduced by Kruithof (1937) for predicting telephone traffic between central stations and has become popular by the work of Deming and Stephan (1940). Thereafter, a vast amount of literature on the IPF procedure has been published. The first to prove necessary and sufficient conditions for convergence of the IPF procedure for nonnegative matrices is Bregman (1967) using an L_1-approach. Csiszár (1975), presumably not knowing the earlier results, gives an alternative proof by means of I-divergence. These two concepts, L_1 and I-divergence, both appear to be appropriate for the analysis of the IPF procedure. Recent results concerning the L_1-approach and a detailed overview of the literature can be found in Pukelsheim (2012). Also Brown et al (1993) give an exhaustive survey of the literature up to 1993.

Section 2 quotes essential properties of I-divergence and generalizes them to the case of weight matrices. Section 3 defines the IPF procedure in detail. The equivalence between iterative fitting and alternating minimization of I-divergence is shown in Section 4. From then on, the main focus lies on the procedure. Three multiple-points properties are proven in Section 5. Section 6 proves convergence of the even-step IPF subsequence and the odd-step IPF subsequence. From that, the aforementioned necessary and sufficient conditions for convergence of the whole sequence of the IPF procedure are derived.

In the sequel, all indices i belong to the set $\{1, \ldots, k\}$ whereas all indices j belong to the set $\{1, \ldots, l\}$. A $+$ as a subscript indicates the summation over the index that

would otherwise appear in its place. A set as a subscript denotes the summation over all entries belonging to that set, i.e. $r_I = \sum_{i \in I} r_i$. All matrices are real nonnegative $k \times \ell$ matrices. The symbol $\ll$ expresses dominance in the sense of measure theory. Thus, for two matrices S and T the expression $S \ll T$ is equivalent to $t_{ij} = 0$ implying $s_{ij} = 0$ for all entries (i, j). The symbol $\approx$ expresses equivalence of two measures. Thus, two matrices S and T are equivalent when they have the same zero entries. We distinguish between *weight matrices* and *probability matrices*. For both cases we demand positive row and column sums. This is not restrictive, since columns or rows of zeros do not change the IPF procedure. For probability matrices we further stipulate the total sum of all entries to sum up to one.

2. I-divergence between weight matrices

Information divergence (I-divergence), also known as Kullback-Leibler divergence, relative entropy or information gain, plays a key role in the analysis of the IPF procedure. Let us briefly recall its properties for probability matrices and generalize them to the case of weight matrices. In this section, the matrices $P = ((p_{ij}))$ and $Q = ((q_{ij}))$ denote two probability matrices, $p_{++} = q_{++} = 1$, whereas the matrices $S = ((s_{ij}))$ and $T = ((t_{ij}))$ are weight matrices. Thus, the matrices P and Q are seen as probability measures and the matrices S and T as finite measures.

We use the conventions $\log 0 := -\infty$, $\log x/0 := \infty$ for all $x \geq 0$ and $0 \cdot \pm\infty := 0$. The *I-divergence of the matrix S relative to the matrix T* is defined by

$$
\mathrm{I}\left(S \mid T\right) := \begin{cases} \sum_{i,j} s_{ij} \log \dfrac{s_{ij}}{t_{ij}} & \text{if } S \ll T, \\ \infty & \text{if } S \not\ll T. \end{cases} \tag{2.1}
$$

For two vectors, I-divergence is defined analogously. It follows from the definition that $\mathrm{I}\left(S \mid T\right)$ is finite if and only if $S \ll T$. In the pertinent literature, I-divergence is only defined for probability measures or a correction term for finite measures is added (Csiszár and Tusnády 1984). For simplicity, it deems appropriate to stick to the definition given in equation (2.1) and point out the minor restrictions when using weight matrices.

2.1. Probability matrices

We first cite the properties of I-divergence for probability matrices needed in the sequel before generalizing them to weight matrices. I-divergence is *nonnegative* and equals zero if and only if both probability matrices are equal (Kullback 1978, Theorem 3.1),

$$\mathrm{I}\left(P \mid Q\right) \geq 0 \text{ and} \tag{2.2}$$

$$\mathrm{I}\left(P \mid Q\right) = 0 \Leftrightarrow P = Q. \tag{2.3}$$

Hence, I-divergence serves as a distance between probability matrices. However, it is not a metric and cannot be transformed into a metric by any reasonable function as shown by Csiszár (1962).

The *Pinsker inequality* between total variation $|P - Q| := \frac{1}{2}\sum_{i,j}|p_{ij} - q_{ij}|$ and I-divergence (Pinsker 1964; Reid and Williamson 2009) provide a relation to the L_1-norm,

$$|P - Q| \leq 2\sqrt{\mathrm{I}(P \,|\, Q)}. \tag{2.4}$$

The monotonicity theorem of Liese and Vajda 1987, Corollary 1.29 is applied to the row sums of the probability matrices P and Q. Therefore, we obtain the *row sum inequality*

$$\mathrm{I}\left((p_{i+}) \,\middle|\, (q_{i+})\right) \leq \mathrm{I}(P \,|\, Q). \tag{2.5}$$

I-divergence is *lower semicontinuous* (Liese and Vajda 1987, Theorem 1.47). For two convergent sequences of probability matrices (P^n) and (Q^n), it holds

$$\mathrm{I}\left(\lim_{n \to \infty} P^n \,\middle|\, \lim_{n \to \infty} Q^n\right) \leq \liminf_{n \to \infty} \mathrm{I}(P^n \,|\, Q^n). \tag{2.6}$$

2.2. Weight matrices

We generalize the properties cited above to weight matrices. Setting the total sums $s := s_{++}$ and $t := t_{++}$, we get whenever the weight matrix S is dominated by the weight matrix T the conversion and *lower boundary*

$$\mathrm{I}(S \,|\, T) = \sum_{i,j} s\frac{s_{ij}}{s}\log\frac{s\frac{s_{ij}}{s}}{t\frac{t_{ij}}{t}} = s\,\mathrm{I}\left(\frac{1}{s}S \,\middle|\, \frac{1}{t}T\right) + s\log\frac{s}{t}$$
$$\geq s\log\frac{s}{t} > -\infty. \tag{2.7}$$

Consequently, for weight matrices S and T with the same total sum, $s = t$, I-divergence is nonnegative and equals zero if and only if both weight matrices are equal.

Whenever the weight matrices S and T have the same total sum, $s = t$, we get the *generalized Pinsker inequality*

$$|S - T| = s\left|\frac{1}{s}S - \frac{1}{s}T\right| \leq 2s\sqrt{\mathrm{I}\left(\frac{1}{s}S \,\middle|\, \frac{1}{s}T\right)} = 2\sqrt{s\,\mathrm{I}(S \,|\, T)}. \tag{2.8}$$

The row sum inequality holds for weight matrices as well, since

$$\mathrm{I}\left((s_{i+}) \,|\, (t_{i+})\right) = s\,\mathrm{I}\left(\left(\frac{s_{i+}}{s}\right) \,\middle|\, \left(\frac{t_{i+}}{t}\right)\right) + s\log\frac{s}{t}$$
$$\leq s\,\mathrm{I}\left(\frac{1}{s}S \,\middle|\, \frac{1}{t}T\right) + s\log\frac{s}{t} = \mathrm{I}(S \,|\, T). \tag{2.9}$$

I-divergence is also lower semicontinuous for sequences of weight matrices (S^n) and (T^n) with constant total sums $s^n_{++} = s$ and $t^n_{++} = t$ for all $n \in \mathbb{N}$, since

$$\mathrm{I}\left(\lim_{n\to\infty} S^n \,\middle|\, \lim_{n\to\infty} T^n\right) = s \sum_{i,j} \lim_{n\to\infty} \frac{s_{ij}^n}{s} \log \frac{\lim_{n\to\infty} \frac{s_{ij}^n}{s}}{\lim_{n\to\infty} \frac{t_{ij}^n}{t}} + s \sum_{i,j} \lim_{n\to\infty} \frac{s_{ij}^n}{s} \log \frac{s}{t}$$

$$\leq s \liminf_{n\to\infty} \mathrm{I}\left(\frac{1}{s} S^n \,\middle|\, \frac{1}{t} T^n\right) + \lim_{n\to\infty} s_{++}^n \log \frac{s}{t}$$

$$= \liminf_{n\to\infty} \left(\sum_{i,j} s_{ij}^n \log \frac{s_{ij}^n}{t_{ij}^n} + \sum_{i,j} s_{ij}^n \log \frac{t}{s}\right) + s \log \frac{s}{t}$$

$$= \liminf_{n\to\infty} \mathrm{I}\left(S^n \,\middle|\, T^n\right) + s \log \frac{t}{s} + s \log \frac{s}{t}$$

$$= \liminf_{n\to\infty} \mathrm{I}\left(S^n \,\middle|\, T^n\right). \tag{2.10}$$

3. IPF procedure

We specify the IPF procedure in full detail. Our notation sticks to Pukelsheim (2012) as close as possible. The IPF procedure takes as input an arbitrary weight matrix $A = ((a_{ij}))$ and two vectors with positive entries $r = (r_1, \ldots, r_k)$ and $c = (c_1, \ldots, c_\ell)$. The matrix A is referred to as the *input matrix*, whereas r is called the *row marginals* and c is called the *column marginals*.

The procedure is initialized by setting $A(0) := A$. Subsequently, the *IPF sequence* $(A(t))$ is calculated by alternately repeating the following two steps:

- Odd steps $t+1$ fit row sums to row marginals. To this end, all entries in the same row are multiplied by the same multiplier yielding

$$a_{ij}(t+1) := \frac{r_i}{a_{i+}(t)} a_{ij}(t) \text{ for all entries } (i,j). \tag{3.1}$$

- Even steps $t+2$ fit column sums to column marginals. To this end, all entries in the same column are multiplied by the same multiplier yielding

$$a_{ij}(t+2) := \frac{c_j}{a_{+j}(t+1)} a_{ij}(t+1) \text{ for all entries } (i,j). \tag{3.2}$$

By induction, for all steps $t \geq 0$ the inequality $a_{ij}(t) > 0$ holds if and only if $a_{ij} > 0$ holds. Consequently, all row sums $a_{i+}(t)$ and all column sums $a_{+j}(t)$ always stay positive. Thus, the IPF procedure is well defined. We say that the *IPF procedure converges*, when the IPF sequence $(A(t))$ converges.

4. Equivalence of iterative fitting and alternating minimization

In this section, we elaborate that each fitting step done by the IPF procedure is equivalent to a minimization with respect to I-divergence.

Let the input matrix A, the row marginals r and the column marginals c be given as in Section 3. We define the set $\mathcal{C}$ of matrices C that fulfill the column marginals and are

dominated by matrix A, and the set $\mathcal{R}$ of matrices R that fulfill the row marginals and are dominated by matrix A,

$$\mathcal{C} := \left\{ C = ((c_{ij})) \in \mathbb{R}^{k \times l}_{\geq 0} \,\middle|\, c_{+j} = c_j \text{ and } C \ll A \right\}, \tag{4.1}$$

$$\mathcal{R} := \left\{ R = ((r_{ij})) \in \mathbb{R}^{k \times l}_{\geq 0} \,\middle|\, r_{i+} = r_i \text{ and } R \ll A \right\}. \tag{4.2}$$

For the two sets $\mathcal{C}$ and $\mathcal{R}$ and an arbitrary weight matrix S we define

$$\mathrm{I}\,(\mathcal{C} \mid S) := \inf_{C \in \mathcal{C}} \mathrm{I}\,(C \mid S), \tag{4.3}$$

$$\mathrm{I}\,(S \mid \mathcal{R}) := \inf_{R \in \mathcal{R}} \mathrm{I}\,(S \mid R) \text{ and} \tag{4.4}$$

$$\mathrm{I}\,(\mathcal{C} \mid \mathcal{R}) := \inf_{C \in \mathcal{C}, R \in \mathcal{R}} \mathrm{I}\,(C \mid R). \tag{4.5}$$

In most parts of the sequel, the matrices $C \in \mathcal{C}$ appear in the I-divergence's first argument, and the matrices $R \in \mathcal{R}$ its second argument. This arrangement facilitates a comparison with the results in Pukelsheim (2012) by setting the focus on the even-step IPF subsequence. However, the reverse arrangement would be equally legitimate of course.

Lemma 4.1 (Minimization over $\mathcal{C}$ and $\mathcal{R}$)**.**

(i) *The sets $\mathcal{C}$ and $\mathcal{R}$ are compact.*

(ii) *The infima of equations* (4.3), (4.4) *and* (4.5) *are attained.*

(iii) *If the infima of equations* (4.3) *and* (4.4) *are finite, the attaining matrices are unique.*

Proof. (i) It is straightforward to show that the two sets are closed and bounded.

(ii) Matrices in $\mathcal{C}$ and $\mathcal{R}$ have constant total sums. Therefore, I-divergence is lower semicontinuous in both arguments on the space $\mathcal{C} \times \mathcal{R}$ as shown in inequality (2.10). Then, I-divergence is also lower semicontinuous in each single argument when keeping the other argument constant. On compact sets, each lower semicontinuous function attains its minimum (Barbu and Precupanu 2010, Theorem 2.8).

(iii) For a matrix C attaining the finite infimum of equation (4.3), it holds $C \ll S$. Hence, we can rewrite the I-divergence of C relative to S as $\mathrm{I}\,(C \mid S) = \sum_{i,j} (c_{ij}/s_{ij}) \cdot \log(c_{ij}/s_{ij}) \cdot s_{ij}$. Since the mapping $x \mapsto x \log x$ is strictly convex, C is unique.

For a matrix R attaining the finite infimum of equation (4.4), it holds $S \ll R$. Hence, we can rewrite the I-divergence of S relative to R as $\mathrm{I}\,(S \mid R) = \sum_{i,j} s_{ij} \log s_{ij} - \sum_{i,j} s_{ij} \log r_{ij}$. In order to minimize $\mathrm{I}\,(S \mid R)$, the value of r_{ij} seems to be arbitrary if $s_{ij} = 0$. However, the constraint of fixed row marginals forces $r_{ij} = 0$ whenever $s_{ij} = 0$. Otherwise a transfer of the value r_{ij} within row i to an entry (i, j_*) with $s_{ij_*} > 0$ reduces the I-divergence, since the mapping $x \mapsto -\log x$ is strictly decreasing. From the strict convexity of $x \mapsto -\log x$, we conclude the uniqueness of R. $\qquad\square$

When minimizing the I-divergence's second argument only, we define for all matrices $C \in \mathcal{C}$ with $I(C \mid \mathcal{R}) < \infty$ the matrix $P_{\mathcal{R},2}(C) \in \mathcal{R}$ with $I(C|P_{\mathcal{R},2}(C)) = I(C \mid \mathcal{R})$ as the $I_{\mathcal{R},2}$-*projection of the matrix C on the set $\mathcal{R}$ in the second argument.* It is denoted by

$$P_{\mathcal{R},2}(C) := \operatorname{argmin}_{R \in \mathcal{R}} I(C \mid R) < \infty. \tag{4.6}$$

Similarly, we define for all matrices $R \in \mathcal{R}$ with $I(\mathcal{C} \mid R) < \infty$ the matrix $P_{\mathcal{C},1}(R) \in \mathcal{C}$ with $I(P_{\mathcal{C},1}(R)|R) = I(\mathcal{C} \mid R)$ as the $I_{\mathcal{C},1}$-*projection of the matrix R on the set $\mathcal{C}$ in the first argument.* It is denoted by

$$P_{\mathcal{C},1}(R) := \operatorname{argmin}_{C \in \mathcal{C}} I(C \mid R) < \infty. \tag{4.7}$$

The explicit form of the matrices $P_{\mathcal{R},2}(C)$ and $P_{\mathcal{C},1}(R)$ are given in Theorem 4.2. There, the matrix $A(t)$ in part (i) can be replaced by an arbitrary matrix in $\mathcal{C}$ and the matrix $A(t+1)$ in part (ii) can be replaced by an arbitrary matrix in $\mathcal{R}$.

Theorem 4.2 (Equivalence of fitting and I-projecting). *Let the input matrix A, the row marginals r and the column marginals c be given. Let $t \geq 0$ be an even step of the IPF sequence $(A(t))$.*

(i) *The matrix $A(t+1)$ is the fitting of the matrix $A(t)$ to the row marginals r if and only if the matrix $A(t+1)$ is the $I_{\mathcal{R},2}$-projection of the matrix $A(t)$ on the set $\mathcal{R}$, formally*

$$a_{ij}(t+1) = \frac{r_i}{a_{i+}(t)} a_{ij}(t) \text{ for all entries } (i,j) \Leftrightarrow A(t+1) = P_{\mathcal{R},2}(A(t)). \tag{4.8}$$

(ii) *The matrix $A(t+2)$ is the fitting of the matrix $A(t+1)$ to the column marginals c if and only if the matrix $A(t+2)$ is the $I_{\mathcal{C},1}$-projection of the matrix $A(t+1)$ on the set $\mathcal{C}$, formally*

$$a_{ij}(t+2) = \frac{c_j}{a_{+j}(t+1)} a_{ij}(t+1) \text{ for all entries } (i,j) \Leftrightarrow A(t+2) = P_{\mathcal{C},1}(A(t+1)).$$

$$\tag{4.9}$$

Proof. For both cases we assume the left-hand side and prove that it is the respective finite minimum. Uniqueness of the finite minimum (Lemma 4.1 (iii)) guarantees equivalence.

(i) Let $A(t+1)$ be defined by $a_{ij}(t+1) = (r_i/a_{i+}(t)) \cdot a_{ij}(t)$. It holds $a_{i+}(t+1) = \sum_j (r_i/a_{i+}(t)) \cdot a_{ij}(t) = r_i$. From Section 3, it follows that $A(t+1) \approx A$. Thus, $A(t+1) \in \mathcal{R}$ and $I(A(t) \mid A(t+1)) < \infty$.

We prove $I(A(t) \mid A(t+1)) \leq I(A(t) \mid R)$ for all $R \in \mathcal{R}$, which shows that for $A(t+1)$ the desired minimum is attained. Without loss of generality we assume $A(t) \ll R$, since otherwise the intended inequality is trivial. We make use of $r_{i+} = r_i$ for all $R \in \mathcal{R}$ and the row sum inequality (2.9) and obtain for all $R \in \mathcal{R}$ with $A(t) \ll R$ the inequality

$$I(A(t) \mid A(t+1)) = \sum_{i,j} a_{ij}(t) \log \frac{a_{ij}(t)}{\frac{r_i}{a_{i+}(t)} a_{ij}(t)} = \sum_{i,j} a_{ij}(t) \log \frac{a_{i+}(t)}{r_i}$$

$$= \sum_i a_{i+}(t) \log \frac{a_{i+}(t)}{r_{i+}} = I((a_{i+}(t)) \mid (r_{i+})) \leq I(A(t) \mid R). \tag{4.10}$$

(ii) Let $A(t+2)$ be defined by $a_{ij}(t+2) = (c_j/a_{+j}(t+1)) \cdot a_{ij}(t+1)$. It holds $a_{+j}(t+2) = \sum_i (c_j/a_{+j}(t+1)) \cdot a_{ij}(t+1) = c_j$. From Section 3, it follows that $A \approx A(t+1) \approx A(t+2)$. Thus, $A(t+2) \in \mathcal{C}$ and $\mathrm{I}(A(t+2)\,|\,A(t+1)) < \infty$.

We prove for all $C \in \mathcal{C}$ the so-called three-points property

$$\mathrm{I}(C\,|\,A(t+2)) + \mathrm{I}(A(t+2)\,|\,A(t+1)) = \mathrm{I}(C\,|\,A(t+1)). \tag{4.11}$$

Since $C \ll A$ by definition (4.1) and $A \approx A(t+1) \approx A(t+2)$, it holds $C \ll A(t+1)$ and $C \ll A(t+2)$. We make use of $c_{+j} = c_j$ for all $C \in \mathcal{C}$ and obtain for all $C \in \mathcal{C}$ the equation

$$\mathrm{I}(C\,|\,A(t+2)) + \mathrm{I}(A(t+2)\,|\,A(t+1))$$

$$= \sum_{i,j} c_{ij} \log \frac{c_{ij}}{\frac{c_j}{a_{+j}(t+1)} a_{ij}(t+1)} + \sum_{i,j} \frac{c_j}{a_{+j}(t+1)} a_{ij}(t+1) \log \frac{\frac{c_j}{a_{+j}(t+1)} a_{ij}(t+1)}{a_{ij}(t+1)}$$

$$= \sum_{i,j} c_{ij} \log \frac{c_{ij}}{a_{ij}(t+1)} + \sum_{i,j} c_{ij} \log \frac{a_{+j}(t+1)}{c_j} + \sum_{i,j} \frac{c_j a_{ij}(t+1)}{a_{+j}(t+1)} \log \frac{c_j}{a_{+j}(t+1)}$$

$$= \mathrm{I}(C\,|\,A(t+1)) + \sum_{j} c_{+j} \log \frac{a_{+j}(t+1)}{c_j} + \sum_{j} \frac{c_j a_{+j}(t+1)}{a_{+j}(t+1)} \log \frac{c_j}{a_{+j}(t+1)}$$

$$= \mathrm{I}(C\,|\,A(t+1)) + \sum_{j} c_j \log \frac{a_{+j}(t+1)}{c_j} - \sum_{j} c_j \log \frac{a_{+j}(t+1)}{c_j}$$

$$= \mathrm{I}(C\,|\,A(t+1)). \tag{4.12}$$

With $c_{++} = a_{++}(t+2)$, it follows $\mathrm{I}(C\,|\,A(t+2)) \geq 0$ and, consequently, $\mathrm{I}(A(t+2)\,|\,A(t+1)) \leq \mathrm{I}(C\,|\,A(t+1))$ for all $C \in \mathcal{C}$. This shows that for $A(t+2)$ the desired minimum is attained. $\square$

Given an arbitrary probability measure as well as an arbitrary marginal and assuming the existence of an I-projection of the probability measure on the set of probability measures fulfilling the given marginal, Cramer (2000) proved the result of the above Theorem 4.2 in his Theorems 3.4 and 3.8.

The *alternating minimization procedure* is defined by alternating $\mathrm{I}_{\mathcal{R},2}$- and $\mathrm{I}_{\mathcal{C},1}$-projections starting with the $\mathrm{I}_{\mathcal{R},2}$-projection of the matrix $A(0) = A$. Hence, in an even step t the matrix $A(t)$ of the alternating minimization procedure is calculated by $t/2$ of $P_{\mathcal{R},2}$-projections and $t/2$ of $P_{\mathcal{C},1}$-projections in the alternating form

$$P_{\mathcal{C},1} \circ P_{\mathcal{R},2} \circ P_{\mathcal{C},1} \dots P_{\mathcal{R},2} \circ P_{\mathcal{C},1} \circ P_{\mathcal{R},2}(A(0)). \tag{4.13}$$

Since $A(t+1) = P_{\mathcal{R},2}(A(t))$ and $A(t+2) = P_{\mathcal{C},1}(A(t+1))$ by Theorem 4.2, iterative fitting and alternating minimization are equivalent.

5. Multiple-points properties

The proofs in Section 6 rely on how a matrix $C \in \mathcal{C}$ and a matrix $R \in \mathcal{R}$ relate to progressing in the odd-step IPF subsequence from one of its members to the next,

that is, from $A(t+1)$ via $A(t+2)$ to $A(t+3)$. This five-points relation is detailed in Theorem 5.3. The preparatory Lemmas 5.1 and 5.2 start out with three-points and four-points relations. A geometrical interpretation can be found in Appendix A. The following lemma recalls equation (4.11) for further reference.

Lemma 5.1 (Three-points property)**.** *Let $t \geq 0$ be an even step of the IPF sequence $(A(t))$ and let $C \in \mathcal{C}$ be arbitrary. We then have*

$$\mathrm{I}\left(C \,|\, A(t+2)\right) + \mathrm{I}\left(A(t+2) \,|\, A(t+1)\right) = \mathrm{I}\left(C \,|\, A(t+1)\right). \tag{5.1}$$

Whereas the three-points property deals with the fitting of columns, the following four-points property deals with row fitting.

Lemma 5.2 (Four-points property)**.** *Let $t \geq 0$ be an even step of the IPF sequence $(A(t))$ and let $C \in \mathcal{C}$ and $R \in \mathcal{R}$ be arbitrary. We then have*

$$\mathrm{I}\left(C \,|\, A(t+3)\right) \leq \mathrm{I}\left(C \,|\, A(t+2)\right) + \mathrm{I}\left(C \,|\, R\right). \tag{5.2}$$

Proof. If $\mathrm{I}\left(C \,|\, A(t+2)\right) = \infty$ or $\mathrm{I}\left(C \,|\, R\right) = \infty$ holds, inequality (5.2) also holds. We therefore suppose $C \ll A(t+2)$ and $C \ll R$. From Section 3, we have

$$a_{ij}(t+3) = \frac{r_i}{a_{i+}(t+2)} a_{ij}(t+2) \text{ for all entries } (i,j). \tag{5.3}$$

Hence, we obtain $A(t+2) \approx A(t+3)$ and, consequently, $C \ll A(t+3)$. Applying equation (5.3) we get

$$
\begin{aligned}
\mathrm{I}\left(C \,|\, A(t+3)\right) &= \sum_{i,j} c_{ij} \log \frac{c_{ij}}{\frac{r_i}{a_{i+}(t+2)} a_{ij}(t+2)} \\
&= \sum_{i,j} c_{ij} \log \frac{c_{ij}}{a_{ij}(t+2)} + \sum_{i,j} c_{ij} \log \frac{a_{i+}(t+2)}{r_i} \\
&= \mathrm{I}\left(C \,|\, A(t+2)\right) + \sum_{i} c_{i+} \log \frac{a_{i+}(t+2)}{r_i}. \tag{5.4}
\end{aligned}
$$

We have $r_i = r_{i+}$ for all rows i and, thus,

$$
\begin{aligned}
\mathrm{I}&\left(C \,|\, A(t+2)\right) + \sum_{i} c_{i+} \log \frac{a_{i+}(t+2)}{r_i} \\
&= \mathrm{I}\left(C \,|\, A(t+2)\right) + \sum_{i:\, c_{i+}>0} c_{i+} \log \frac{c_{i+} a_{i+}(t+2)}{r_{i+} c_{i+}} \\
&= \mathrm{I}\left(C \,|\, A(t+2)\right) + \sum_{i:\, c_{i+}>0} c_{i+} \log \frac{c_{i+}}{r_{i+}} + \sum_{i:\, c_{i+}>0} c_{i+} \log \frac{a_{i+}(t+2)}{c_{i+}} \\
&= \mathrm{I}\left(C \,|\, A(t+2)\right) + \mathrm{I}\left((c_{i+}) \,|\, (r_{i+})\right) - \mathrm{I}\left((c_{i+}) \,|\, (a_{i+}(t+2))\right). \tag{5.5}
\end{aligned}
$$

Applying the row sum inequality inequality (2.9) and inequality (2.7) with $c_{++} = a_{++}(t+2)$ we get

$$
\begin{aligned}
&\mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left((c_{i+}) \mid (r_{i+})\right) - \mathrm{I}\left((c_{i+}) \mid (a_{i+}(t+2))\right) \\
&\leq \mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left(C \mid R\right).
\end{aligned}
\tag{5.6}
$$

Combining equations (5.4) and (5.5) and inequality (5.6) completes the proof. □

The three-points property and the four-points property together yield the five-points property.

Theorem 5.3 (Five-points property). *Let $t \geq 0$ be an even step of the IPF sequence $(A(t))$ and let $C \in \mathcal{C}$ and $R \in \mathcal{R}$ be arbitrary. We then have*

$$
\mathrm{I}\left(A(t+2) \mid A(t+3)\right) + \mathrm{I}\left(C \mid A(t+3)\right) \leq \mathrm{I}\left(C \mid A(t+1)\right) + \mathrm{I}\left(C \mid R\right).
\tag{5.7}
$$

Proof. Summation of the three-points property and the four-points property yields

$$
\begin{aligned}
&\mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left(A(t+2) \mid A(t+1)\right) + \mathrm{I}\left(C \mid A(t+3)\right) \\
&\leq \mathrm{I}\left(C \mid A(t+1)\right) + \mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left(C \mid R\right).
\end{aligned}
\tag{5.8}
$$

If $\mathrm{I}\left(C \mid A(t+1)\right) = \infty$ holds, inequality (5.7) also holds. We therefore suppose $C \ll A(t+1)$. From Section 3, we have $C \ll A(t+2)$ and, thus, $\mathrm{I}\left(C \mid A(t+2)\right) < \infty$. For this reason, we can subtract $\mathrm{I}\left(C \mid A(t+2)\right)$ from both sides of inequality (5.8) and get

$$
\mathrm{I}\left(A(t+2) \mid A(t+1)\right) + \mathrm{I}\left(C \mid A(t+3)\right) \leq \mathrm{I}\left(C \mid A(t+1)\right) + \mathrm{I}\left(C \mid R\right).
\tag{5.9}
$$

From Theorem 4.2 (i), we know

$$
\mathrm{I}\left(A(t+2) \mid A(t+3)\right) = \min_{R \in \mathcal{R}} \mathrm{I}\left(A(t+2) \mid R\right) \leq \mathrm{I}\left(A(t+2) \mid A(t+1)\right).
\tag{5.10}
$$

Combining inequalities (5.9) and (5.10) completes the proof. □

6. Accumulation points of the IPF sequence

We now present the central results of this paper. The role of the following monotonicity lemma is similar to the role of Lemma 1 in Pukelsheim (2012), where monotonicity is proved for an L_1-error function.

Lemma 6.1 (Monotonicity). *There exists a matrix $C^* \in \mathcal{C}$ such that $\mathrm{I}\left(C^* \mid \mathcal{R}\right) = \mathrm{I}\left(\mathcal{C} \mid \mathcal{R}\right)$ holds. For all even steps $t \geq 2$ of the IPF sequence $(A(t))$, this matrix C^* satisfies the inequality*

$$
\mathrm{I}\left(C^* \mid A(t+2)\right) \leq \mathrm{I}\left(C^* \mid A(t)\right).
\tag{6.1}
$$

Proof. Lemma 4.1 (ii) guarantees the existence of a matrix $C^* \in \mathcal{C}$ such that $I\left(C^* \mid \mathcal{R}\right) = I\left(\mathcal{C} \mid \mathcal{R}\right)$ holds. By applying the three-points property (Lemma 5.1) to $C^* \in \mathcal{C}$, we get

$$I\left(C^* \mid A(t+2)\right) + I\left(A(t+2) \mid A(t+1)\right) = I\left(C^* \mid A(t+1)\right). \qquad (6.2)$$

Since $t \geq 2$ holds, we can apply the four-points property (Lemma 5.2) to $C^* \in \mathcal{C}$ and step $t-2$ and obtain

$$I\left(C^* \mid A(t+1)\right) \leq I\left(C^* \mid A(t)\right) + I\left(C^* \mid R\right) \text{ for all } R \in \mathcal{R}. \qquad (6.3)$$

Combining equation (6.2) and inequality (6.3) yields

$$\begin{aligned}
I\left(C^* \mid A(t+2)\right) + I\left(A(t+2) \mid A(t+1)\right) &\leq I\left(C^* \mid A(t)\right) + I\left(C^* \mid \mathcal{R}\right) \\
&= I\left(C^* \mid A(t)\right) + I\left(\mathcal{C} \mid \mathcal{R}\right). \qquad (6.4)
\end{aligned}$$

Since $I\left(A(t+2) \mid A(t+1)\right) \geq I\left(\mathcal{C} \mid \mathcal{R}\right)$ holds by definition of $I\left(\mathcal{C} \mid \mathcal{R}\right)$, we get

$$I\left(C^* \mid A(t+2)\right) \leq I\left(C^* \mid A(t)\right). \qquad (6.5)$$

$\square$

Whereas Lemma 6.1 deals with the even-step IPF subsequence only, the following lemma gives a limit for the I-divergence of an even-step matrix $A(t+2)$ relative to its successor $A(t+3)$.

Lemma 6.2 (Convergence of I-divergence)**.** *We have*

$$\lim_{t=0,2,4,\dots} I\left(A(t+2) \mid A(t+3)\right) = I\left(\mathcal{C} \mid \mathcal{R}\right). \qquad (6.6)$$

Proof. Lemma 4.1 (ii) guarantees the existence of a matrix $C^* \in \mathcal{C}$ and a matrix $R^* \in \mathcal{R}$ such that $I\left(C^* \mid R^*\right) = I\left(\mathcal{C} \mid \mathcal{R}\right)$ holds. We now apply the five-points property (Theorem 5.3) to an arbitrary even step $t \geq 0$ and the matrices $C^* \in \mathcal{C}$ and $R^* \in \mathcal{R}$. This yields

$$I\left(A(t+2) \mid A(t+3)\right) + I\left(C^* \mid A(t+3)\right) \leq I\left(C^* \mid A(t+1)\right) + I\left(C^* \mid R^*\right). \qquad (6.7)$$

By definition of the sets $\mathcal{C}$ and $\mathcal{R}$, we have $C^* \ll A$ and $R^* \ll A$. Hence, all I-divergences appearing in inequality (6.7) are finite. Thus, for all even steps $t \geq 0$, we have

$$I\left(A(t+2) \mid A(t+3)\right) - I\left(C^* \mid R^*\right) \leq I\left(C^* \mid A(t+1)\right) - I\left(C^* \mid A(t+3)\right). \qquad (6.8)$$

As a consequence, for all even steps $\tau \geq 0$ we get

$$\begin{aligned}
&\sum_{t=0,2,4,\dots,\tau} \left[I\left(A(t+2) \mid A(t+3)\right) - I\left(C^* \mid R^*\right) \right] \\
&\leq \sum_{t=0,2,4,\dots,\tau} \left[I\left(C^* \mid A(t+1)\right) - I\left(C^* \mid A(t+3)\right) \right] \\
&= I\left(C^* \mid A(1)\right) - I\left(C^* \mid A(\tau+3)\right) \leq I\left(C^* \mid A(1)\right) - I\left(\mathcal{C} \mid \mathcal{R}\right) < \infty. \qquad (6.9)
\end{aligned}$$

Since $\mathrm{I}\left(A(t+2)\,|\,A(t+3)\right) \geq \mathrm{I}\left(C^*\,|\,R^*\right)$ holds for all even steps $t \geq 0$, the series is nondecreasing in $\tau = 0, 2, 4, \ldots$. Hence, it converges for $\tau \to \infty$ and its terms form a null sequence. As a result, we get

$$\lim_{t=0,2,4,\ldots} \mathrm{I}\left(A(t+2)\,|\,A(t+3)\right) = \mathrm{I}\left(C^*\,|\,R^*\right) = \mathrm{I}\left(\mathcal{C}\,|\,\mathcal{R}\right). \tag{6.10}$$

$\square$

We now make use of Lemma 6.1 and Lemma 6.2 to prove convergence of the even-step IPF subsequence as well as convergence of the odd-step IPF subsequence.

Theorem 6.3 (Accumulation points of the IPF sequence)**.** *Let $(A(t))$ be the IPF sequence as specified in Section 3. Then the following two statements hold:*

(i) The even-step IPF subsequence $(A(t))_{t=0,2,4,\ldots}$ converges.

(ii) The odd-step IPF subsequence $(A(t+1))_{t=0,2,4,\ldots}$ converges.

As a result, the IPF sequence $(A(t))$ has at most two accumulation points.

Proof. (i) According to Lemma 4.1 (i), the sets $\mathcal{C}$ and $\mathcal{R}$ are compact. Hence, there exists a sequence $(\tilde{t}_n)$ of even steps such that $(A(\tilde{t}_n + 2)) \subseteq \mathcal{C}$ converges. Moreover, there exists a subsequence (t_n) of $(\tilde{t}_n)$ such that $(A(t_n + 3)) \subseteq \mathcal{R}$ also converges. Set $C^* := \lim_{n\to\infty} A(t_n + 2)$ and $R^* := \lim_{n\to\infty} A(t_n + 3)$. Applying lower semicontinuity of I-divergence (inequality (2.10)) and Lemma 6.2 we get

$$\mathrm{I}\left(C^*\,|\,R^*\right) = \mathrm{I}\left(\lim_{n\to\infty} A(t_n + 2)\,\Big|\,\lim_{n\to\infty} A(t_n + 3)\right)$$
$$\leq \liminf_{n\to\infty} \mathrm{I}\left(A(t_n + 2)\,|\,A(t_n + 3)\right) = \mathrm{I}\left(\mathcal{C}\,|\,\mathcal{R}\right). \tag{6.11}$$

Since $\mathrm{I}\left(C^*\,|\,R^*\right) \geq \mathrm{I}\left(\mathcal{C}\,|\,\mathcal{R}\right)$ holds by definition of $\mathrm{I}\left(\mathcal{C}\,|\,\mathcal{R}\right)$, we have $\mathrm{I}\left(C^*\,|\,R^*\right) = \mathrm{I}\left(\mathcal{C}\,|\,\mathcal{R}\right)$.

We can now apply Lemma 6.1 to the matrix C^*. As a result, the sequence $(\mathrm{I}\left(C^*\,|\,A(t+2)\right))_{t=0,2,4,\ldots}$ is nonincreasing. Since $c_{++}^* = a_{++}(t+2)$ holds for all even steps $t \geq 0$, the sequence is also bounded from below by 0 (inequality (2.7)) and, thus, convergent. Consequently, we get

$$\lim_{t=0,2,4,\ldots} \mathrm{I}\left(C^*\,|\,A(t+2)\right) = \lim_{n\to\infty} \mathrm{I}\left(C^*\,|\,A(t_n + 2)\right) = \mathrm{I}\left(C^*\,|\,C^*\right) = 0. \tag{6.12}$$

Here, the penultimate equality follows from the continuity of I-divergence in the second argument. This holds, since $x \mapsto \log(1/x)$ is a continuous mapping and the I-divergence $\mathrm{I}\left(C^*\,|\,A(t_n + 2)\right)$ is defined as

$$\mathrm{I}\left(C^*\,|\,A(t_n + 2)\right) = \sum_{i,j} c_{ij}^* \log \frac{c_{ij}^*}{a_{ij}(t_n + 2)}. \tag{6.13}$$

Only the case $a_{ij}(t_n + 2) \to 0$ could cause problems. For this reason, we have to examine the two cases $c_{ij}^* = 0$ and $c_{ij}^* > 0$. Since $0 \cdot \pm\infty = 0$ holds by convention,

the first case turns out to be unproblematic. The second case cannot occur due to $c_{ij}^* = \lim_{n \to \infty} a_{ij}(t_n + 2)$.

By combining the generalized Pinsker inequality (2.8) with equation (6.12) we obtain

$$\lim_{t=0,2,4,\dots} |C^* - A(t+2)| = 0. \tag{6.14}$$

(ii) Since $A(t+1)$ continuously depends on $A(t)$ for all even steps $t \geq 0$ by definition of the IPF procedure (equation (3.1)), convergence of the even-step IPF subsequence $(A(t))_{t=0,2,4,\dots}$ implies convergence of the odd-step IPF subsequence $(A(t+1))_{t=0,2,4,\dots}$. $\square$

Theorem 6.3 proves Pukelsheim's conjecture on the accumulation points of the IPF sequence (Pukelsheim 2012, Lemma 4). Moreover, it permits us to derive the well-known criteria for convergence of the IPF procedure (Bregman 1967; Csiszár 1975).

Theorem 6.4 (Convergence of the IPF procedure)**.** *The IPF procedure as specified in Section 3 converges if and only if there exists a weight matrix D dominated by the input matrix A and fulfilling the row marginals r and the column marginals c.*

Proof. If, on the one hand, the entire IPF sequence $(A(t))$ converges to a limit B, we have $B \in \mathcal{R}$ and $B \in \mathcal{C}$. This yields $B \ll A$ as well as $b_{i+} = r_i$ for all rows i and $b_{+j} = c_j$ for all columns j.

If, on the other hand, there exists a weight matrix $D \ll A$ fulfilling the row marginals r and the column marginals c, we have $D \in \mathcal{R} \cap \mathcal{C}$ and $r_+ = c_+$. Hence, $\mathrm{I}(C \,|\, R)$ is bounded from below by 0 and this bound is attained for $C = R = D$. Set $C^* := \lim_{t=0,2,4,\dots} A(t+2)$ and $R^* := \lim_{t=0,2,4,\dots} A(t+3)$. Applying lower semicontinuity of I-divergence (inequality 2.10) and Lemma 6.2 we get

$$0 \leq \mathrm{I}(C^* \,|\, R^*) = \mathrm{I}\left(\lim_{t=0,2,4,\dots} A(t+2) \,\Big|\, \lim_{t=0,2,4,\dots} A(t+3)\right)$$
$$\leq \liminf_{t=0,2,4,\dots} \mathrm{I}(A(t+2) \,|\, A(t+3)) = \mathrm{I}(\mathcal{C} \,|\, \mathcal{R}) = 0. \tag{6.15}$$

Together with $r_+ = c_+$, this yields $C^* = R^*$ completing the proof. $\square$

The proof of Theorem 6.4 shows that convergence of the IPF procedure is equivalent with the two conditions $\mathrm{I}(\mathcal{C} \,|\, \mathcal{R}) = 0$ and $r_+ = c_+$. It is well-known that the second condition alone does not imply convergence. Neither does the first condition alone, as is illustrated in the following example.

Example 6.5. Choose

$$A := \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}, \, r := \begin{pmatrix} \frac{3}{4}, \frac{1}{3} \end{pmatrix} \text{ and } c := \begin{pmatrix} \frac{1}{2}, \frac{1}{2} \end{pmatrix}. \tag{6.16}$$

This yields

$$\mathcal{C} = \left\{ \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \right\} \text{ and } \mathcal{R} = \left\{ \begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \right\}. \tag{6.17}$$

13

As a result, we get $\mathcal{C} \cap \mathcal{R} = \emptyset$ and the IPF procedure has two accumulation points. However, $I(\mathcal{C} \mid \mathcal{R}) = I(C^* \mid R^*) = 0$ holds.

Remark 6.6. For practical applications we need a criterion for when to stop the IPF procedure, especially when the procedure does not converge.

Pukelsheim (2012, Lemma 5) shows that the L_1-error of the IPF sequence converges to a limit which can be computed directly from the input matrix A and the marginals r and c. Thus, the IPF procedure can be stopped as soon as the L_1-error of the matrix $A(t)$ is sufficiently close to this limit.

Csiszár and Tusnády (1984, p. 219) suggest another approach, which is more closely related to I-divergence. They prove for all even steps $t \geq 0$ the upper boundary

$$I(A(t+2) \mid A(t+3)) - I(\mathcal{C} \mid \mathcal{R}) \leq c_+ \ln \max_{i,j} \frac{a_{ij}(t+3)}{a_{ij}(t+1)}. \tag{6.18}$$

Due to convergence of the odd-step IPF subsequence (Theorem 6.3 (ii)) the right-hand side of inequality (6.18) converges to 0 for $t \to \infty$ along $t = 0, 2, 4, \ldots$. Thus, the IPF procedure can be stopped as soon as this term is smaller than a given $\epsilon > 0$.

Inequality (6.18) can be seen as follows. Lemma 4.1 (ii) guarantees the existence of a matrix $C^* \in \mathcal{C}$ such that $I(C^* \mid \mathcal{R}) = I(\mathcal{C} \mid \mathcal{R})$ holds. Applying the five-points property (Theorem 5.3) we get

$$I(A(t+2) \mid A(t+3)) + I(C^* \mid A(t+3)) \leq I(C^* \mid A(t+1)) + I(C^* \mid \mathcal{R}). \tag{6.19}$$

This yields

$$I(A(t+2) \mid A(t+3)) - I(\mathcal{C} \mid \mathcal{R}) \leq I(C^* \mid A(t+1)) - I(C^* \mid A(t+3))$$
$$= \sum_{i,j} c_{ij}^* \ln \frac{a_{ij}(t+3)}{a_{ij}(t+1)} \leq c_+ \ln \max_{i,j} \frac{a_{ij}(t+3)}{a_{ij}(t+1)}. \tag{6.20}$$

An even sharper bound can be found by separately calculating the maximum of $a_{ij}(t+3)/a_{ij}(t+1)$ for each column j yielding

$$\sum_{i,j} c_{ij}^* \ln \frac{a_{ij}(t+3)}{a_{ij}(t+1)} \leq \sum_{j} c_j \ln \max_{i} \frac{a_{ij}(t+3)}{a_{ij}(t+1)}. \tag{6.21}$$

7. Outlook

The IPF procedure can also be applied to a multivariate input array with more than two sets of given multidimensional marginals. When fitting cyclically, necessary and sufficient conditions for convergence of the whole IPF sequence are well known (Csiszár 1975). However, convergence of the respective IPF subsequences is still an open problem.

Kullback (1968) introduced the IPF procedure on a continuous space, but failed to prove convergence of the whole IPF sequence. A complete proof was given by Rüschendorf (1995) under certain regularity constraints. Finding necessary and sufficient conditions for convergence of the whole IPF sequence in analogy with Theorem 6.4 is another open issue. Also, the convergence of the even-step IPF subsequence and the odd-step IPF subsequence on a continuous space remains a conjecture.

A. Geometric interpretation of the multiple-points properties

The multiple-points properties presented in Section 5 are interpreted geometrically. To present intelligible graphics, we stick to the two-dimensional Euclidean space. Here, the squared Euclidean distance d^2 plays the role of I-divergence. Thus, the asymmetry of I-divergence is neglected. The I-projections are seen as orthogonal projections. All matrices are handled as points. Csiszár (1975) as well as Csiszár and Tusnády (1984) adumbrate this interpretation.

The three-points property

$$\mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left(A(t+2) \mid A(t+1)\right) = \mathrm{I}\left(C \mid A(t+1)\right) \tag{A.1}$$

is displayed in Figure 1. The set $\mathcal{C}$ is a lower dimensional simplex in the space $\mathbb{R}_{\geq 0}^{k \times \ell}$
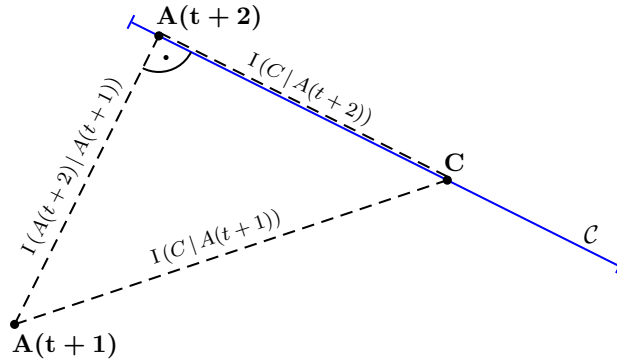
Figure 1: Geometric interpretation of the three-points property (Lemma 5.1) in the two-dimensional Euclidean space. The three-points property can be interpreted as the Pythagorean theorem with I-divergence as the squared Euclidean distance.

because of its linear restrictions. Hence, in two-dimensional space we depict $\mathcal{C}$ as a line segment. Now we fix an arbitrary point as $A(t+1)$ and another point on the line segment as C. The orthogonal projection of $A(t+1)$ on the line segment $\mathcal{C}$ is labeled $A(t+2)$. We assume the line segment $\mathcal{C}$ to be sufficiently long to secure a right angle between the line segment $\mathcal{C}$ and the line segment $A(t+1)A(t+2)$. The Euclidean distances between the three points $A(t+1), A(t+2)$ and C are displayed as dashed lines. Now, the three-points property can be interpreted as the Pythagorean theorem with I-divergence as the squared Euclidean distance.

The interpretation of the four-points property

$$\mathrm{I}\left(C \mid A(t+3)\right) \leq \mathrm{I}\left(C \mid A(t+2)\right) + \mathrm{I}\left(C \mid R\right) \tag{A.2}$$

is again based on the Pythagorean theorem. In analogy to the display of the set $\mathcal{C}$, the set $\mathcal{R}$ is depicted as a line segment, too. We fix an arbitrary point $A(t+2)$. The orthogonal projection of the point $A(t+2)$ on the line segment $\mathcal{C}$ is labeled as $A(t+3)$. Again, we assume the line segment $\mathcal{R}$ to be sufficiently long to secure a right angle between

the line segment $\mathcal{R}$ and the line segment $A(t+2)A(t+3)$. The origin of the Cartesian coordinate system is set in the point $A(t+3)$. The figure is rotated such that the x-axis of the coordinate system overlaps with the line segment $\mathcal{R}$ and the point $A(t+2)$ is on the positive y-axis. Now, we have to distinguish in which quadrant the point C lies. There, point C lying in the first quadrant is symmetrical to point C lying in the second quadrant and the same symmetry holds for the third and fourth quadrant.

In Figure 2 we reuse the Figure 1 and display the case of point C lying in the first quadrant. When setting the point R on the negative x-axis, the distance $d(C, R)$ is obviously larger than $d(C, A(t+3))$ and the four-points property follows. Thus, we set the point R on the right-hand side of the point $A(t+3)$. The dotted lines depict the orthogonal projection a on the line segment $A(t+3)A(t+2)$ and the orthogonal projection r on the line segment $\mathcal{R}$. By the Pythagorean theorem it holds $d(C, A(t+3))^2 = d(C, a)^2 + d(C, r)^2$. With $d(C, a) \leq d(C, A(t+2))$ and $d(C, r) \leq d(C, R)$ the four-points property follows. In the case of point C lying in the third or fourth quadrant, we focus
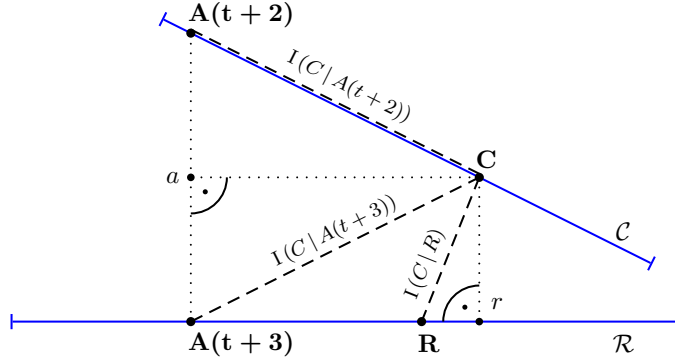

Figure 2: Geometric interpretation of the four-points property (Lemma 5.2) in the two-dimensional Euclidean space. This is again an application of the Pythagorean theorem.

on the triangle $A(t+3)A(t+2)C$. The angle α at point $A(t+3)$ is always obtuse, hence $\cos(\alpha) \leq 0$. The law of cosines

$$
\begin{aligned}
d(C, A(t+2))^2 =& d(C, A(t+3))^2 + d(A(t+2), A(t+3))^2 \\
&- 2d(C, A(t+3))d(A(t+2), A(t+3))\cos(\alpha)
\end{aligned} \tag{A.3}
$$

yields $d(C, A(t+3))^2 \leq d(C, A(t+2))^2$ and consequently $d(C, A(t+3))^2 \leq d(C, A(t+2))^2 + d(C, R)^2$.

The five-points property

$$
\mathrm{I}\left(A(t+2) \mid A(t+3)\right) + \mathrm{I}\left(C \mid A(t+3)\right) \leq \mathrm{I}\left(C \mid A(t+1)\right) + \mathrm{I}\left(C \mid R\right) \tag{A.4}
$$

is the summation of the three-points property and the four-points property with subsequent subtraction of $\mathrm{I}\left(C \mid A(t+2)\right)$ on both sides and the inequality
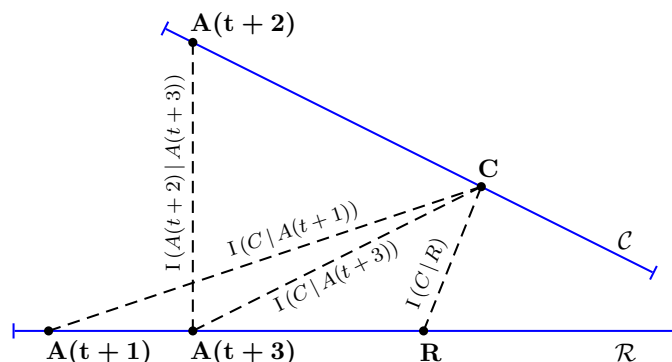
Figure 3: An elementary geometric interpretation of the five-points property (Theorem 5.3) in the two-dimensional Euclidean space fails. However, the geometry of the IPF procedure is revealed.

$I\left(A(t+2)\,|\,A(t+3)\right) \leq I\left(A(t+2)\,|\,A(t+1)\right)$. Figure 3 displays the outcome, but we are unable to give an elementary geometrical interpretation. Yet, the figure gives a geometric interpretation of the whole IPF sequence. The sequence can be seen as alternating orthogonal projections on the line segments $\mathcal{C}$ and $\mathcal{R}$. In the situation displayed here, the convergence of the even-step IPF subsequence to the right end of the upper line segment $\mathcal{C}$ is obvious.

References

Barbu V, Precupanu T (2010) Convexity and optimization in Banach spaces. Springer, Heidelberg

Bregman L (1967) Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints. USSR Computational Mathematics and Mathematical Physics 7:191–204

Brown J, Chase P, Pittenger A (1993) Order independence and factor convergence in iterative scaling. Linear Algebra and its Applications 190:1–38

Cramer E (2000) Probability measures with given marginals and conditionals: I–projections and conditional iterative proportional fitting. Statistics & Decisions 18:311–329

Csiszár I (1962) Über topologische und metrische Eigenschaften der relativen Information der Ordnung α. In: Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes (Prague), Publishing House of the Czechoslovak Academy of Sciences, pp 63–73

Csiszár I (1975) I–divergence geometry of probability distributions and minimization problems. Annals of Probability 3:146–158

Csiszár I, Tusnády G (1984) Information geometry and alternating minimization procedures. Statistics & Decisions, Supplement Issue 1:205–237

Deming W, Stephan F (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics 11:427–444

Jiroušek R, Vomlel J (1994) Inconsistent knowledge integration in a probabilistic model. In: Proceedings of the Workshop Mathematical Models for handling partial knowledge in A.I., Plenum Publ Corp, Erice, Sicily, pp 263–270

Knight P (2008) The Sinkhorn–Knopp algorithm: convergence and applications. SIAM Journal on Matrix Analysis and Applications 30:261–275

Kruithof J (1937) Telefoonverkeersrekening. De Ingenieur 52:E15–E25

Kullback S (1968) Probability densities with given marginals. Annals of Mathematical Statistics 39(4):1236–1243

Kullback S (1978) Information theory and statistics. Smith, Gloucester (Massachusetts)

Lauritzen S (1996) Graphical models. Clarendon Press, Oxford

Liese F, Vajda I (1987) Convex statistical distances, Teubner-Texte zur Mathematik, vol 95. Teubner, Leipzig

McCord M, Mishalani R, Goel P, Strohl B (2010) Iterative proportional fitting procedure to determine bus route passenger origin–destination flows. Transportation Research Record: Journal of the Transportation Research Board 2145:59–65

Pinsker M (1964) Information and information stability of random variables and processes. Holden-Day, San Francisco

Pukelsheim F (2012) Biproportional matrix scaling and the iterative proportional fitting procedure, submitted to Annals of Operations Research

Reid M, Williamson R (2009) Generalised Pinsker inequalities. arXiv:09061244v1

Rüschendorf L (1995) Convergence of the iterative proportional fitting procedure. Annals of Statistics 23:1160–1174

Vomlel J (2004) Integrating inconsistent data in a probabilistic model. Journal of Applied Non–Classical Logics 14:365–386

Zhang S, Peng Y, Wang X (2008) An efficient method for probabilistic knowledge integration. In: 2008 20th IEEE International Conference on Tools with Artificial Intelligence, IEEE, vol 2, pp 179–182