Symmetric Iterative Proportional Fitting

Sven Kurras
Department of Computer Science, University of Hamburg, Germany
sven.kurras@uni-hamburg.de

Abstract

Iterative Proportional Fitting (IPF) generates from an input matrix $W$ a sequence of matrices that converges, under certain conditions, to a specific limit matrix $\hat{W}$. This limit is the relative-entropy nearest solution to $W$ among all matrices of prescribed row marginals $r$ and column marginals $c$. We prove this known fact by a novel strategy that contributes a pure algorithmic intuition. Then we focus on the symmetric setting: $W = W^T$ and $r = c$. Since IPF inherently generates non-symmetric matrices, we introduce two symmetrized variants of IPF. We prove convergence for both of them. Further, we give a novel characterization for the existence of $\hat{W}$ in terms of expansion properties of the undirected weighted graph represented by $W$. Finally, we show how our results contribute to recent work in machine learning.

1 INTRODUCTION

Iterative Proportional Fitting (IPF) refers to an iterative algorithm whose origins date back to research on traffic networks in the 1930s. It was rediscovered in other fields, in several variants, and in a large variety of different names (for example as Sheleikhovskii’s method, Kruthof’s algorithm, Furness method, Sinkhorn-Knopp algorithm, or RAS method, just to name a few). Nowadays, IPF is well-known in machine learning and many other disciplines like statistics, optimization, matrix factorization, economics, or network theory. In particular it serves as a bridge that allows to transfer results and interpretations between these disciplines. IPF takes as its input a non-negative matrix $W$ together with two positive vectors $r$ and $c$ that specify new target marginals for the rows and columns of $W$, respectively. The traditional IPF-sequence $(W^{(k)})$ is determined by $W^{(0)} := W$, and for $k \geq 1$ in an alternating way by first re-scaling all rows to have marginals $r$, then re-scaling all columns to have marginals $c$, then re-scaling the rows again, and so forth. The IPF-sequence converges, under certain conditions, to a limit matrix $\hat{W} = \lim_{k \to \infty} W^{(k)}$ that simultaneously achieves the desired row marginals $r$ and column marginals $c$. In the case that all entries in $W$ are positive, the IPF-sequence is well-understood: Sinkhorn (1967) proves that for any choice of $W \in \mathbb{R}^{m \times n}_{\geq 0}$, $r \in \mathbb{R}^m_{\geq 0}$, $c \in \mathbb{R}^n_{\geq 0}$ with $\|r\|_1 = \|c\|_1$ the IPF-sequence converges to a unique limit matrix $\hat{W} \in \mathbb{R}^{m \times n}_{\geq 0}$ of row marginals $r$ and column marginals $c$. Further $\hat{W}$ has the form $\hat{W} = Y W Z$ for positive diagonal matrices $Y, Z$ that are unique up to a scaling factor. By the Lagrangian approach it is straightforward to prove that $\hat{W}$ is the unique solution (among all matrices of the given marginals) that is closest to $W$ with respect to relative-entropy error. However, as soon as one allows for zero-entries in $W$, both the feasibility problem and the optimization problem become much harder. For the special case of $r = c = 1$, Sinkhorn and Knopp (1967) show that convergence and uniqueness only hold under specific structural constraints on the zero-pattern. These dependencies get even more complicated in the general case of arbitrary positive target marginals $r, c$, as handled in Section 5. Hence, assuming positivity is not just “simplifying the argument”, as stated by Ireland and Kullback (1968, p.182); indeed there is still active research on the non-negative case. One challenge for the Lagrangian approach is that the relative-entropy objective function becomes non-smooth: $\hat{W}$ has at least the same zero-entries as $W$, and whenever $\hat{W}$ has a zero at a position where $W$ has a non-zero, the optimal solution lies at a non-differentiable point. Several publications omit this detail, and simply apply the Lagrangian approach to the non-negative setting, although it is no longer valid. Csiszar (1975) provides a technically sound proof that avoids the Lagrangian.
He represents \( \hat{W} \) and \( W \) as absolutely continuous measures, which allows to consider their Radon-Nikodym derivative. This extends the relative-entropy interpretation of \( W \) to the non-negative setting. However, the proof lacks an algorithmic intuition. In this work we contribute a novel and intuitive proof for the convergence of the IPF-sequence to the relative-entropy optimal solution by identifying IPF as a special instance of a more general iterative projection algorithm.

We are particularly interested in the symmetric setting: \( W = W^T \) and \( f := r = c \). We study two different symmetrizations of IPF: Pseudo-Symmetric IPF, and Symmetric IPF, where we refer to either of both as “symmetrized IPF”. In Section 4 we prove that both symmetrizations converge to the same limit as the traditional IPF, however in contrast to IPF, every intermediate matrix in these sequences is symmetric and even more, of the form \( XWXX \) for positive diagonal \( X \).

In the symmetric setting, we derive in Section 5 necessary and sufficient conditions for the existence of \( \hat{W} \) in terms of expansion properties of the undirected weighted graph \( G(W) \). In Section 6 we show how our results contribute to recent work in machine learning.

## 2 Preliminaries

This section introduces the notation and some basic results from matrix scaling, Bregman projections, and mean functions. For \( m, n \geq 2 \) we denote by \( \Omega \) the set of non-negative \( m \times n \) matrices that contain no zero row and no zero column, that is \( \Omega := \{ X \in \mathbb{R}^{m \times n} \mid X1 > 0, X^T1 > 0 \} \) with all inequalities applied entry-wise and \( 1 := (1, \ldots, 1)^T \).

The restriction of \( \Omega \) to symmetric matrices is denoted by \( S := \{ X \in \mathbb{R}^{n \times n} \mid X = X^T, X1 > 0 \} \). For any matrix \( A = [a_{ij}] \) the set of index pairs of its non-zero entries is denoted by \( E(A) := \{ ij \mid a_{ij} \neq 0 \} \). We say that \( A \) preserves the zeros of \( B \) if \( E(A) \subseteq E(B) \) and that they have the same zeros if \( E(A) = E(B) \).

For a positive vector \( f \in \mathbb{R}^n_0 \) and a symmetric matrix \( W \in \mathcal{S} \) we define the constrained subset \( S(f,W) := \{ X \in \mathbb{R}^{m \times n} \mid X = X^T, X1 = f, E(X) \subseteq E(W) \} \subseteq \mathcal{S} \) of those matrices in \( \mathcal{S} \) that have row (and column) marginals \( f \) while preserving the zeros of \( W \). Similarly, for positive vectors \( r \) and \( c \) define \( \Omega(r,c,W) := \{ X \in \mathbb{R}^{m \times n} \mid X1 = r, X^T1 = c, E(X) \subseteq E(W) \} \). We drop individual constraints by a dot, for example \( \Omega(r,\ldots) := \{ X \in \mathbb{R}^{m \times n} \mid X1 = r \} \) (\( \not\subseteq \Omega \)).

**Matrix scaling.** \( B \in \Omega \) is a biproportional scaling of \( W \in \Omega \) if it can be expressed as \( B = \lim_{k \to +\infty} Y^{(k)}WZ^{(k)} \) for two sequences of diagonal matrices \( (Y^{(k)}) \) and \( (Z^{(k)}) \) with positive diagonals \( y^{(k)} \) and \( z^{(k)} \). Such scalings of \( W \) often aim at fitting \( B \) to prescribed row marginals \( r \in \mathbb{R}^m_0 \) and column marginals \( c \in \mathbb{R}^n_0 \). For that reason we also denote a biproportional scaling \( B \) with \( r = B1 \) and \( c = B^T1 \) as a biproportional fit of \( W \) to row marginals \( r \) and column marginals \( c \). The corresponding biproportional scaling can be seen as an iterative transformation of \( W \) into another matrix \( B \) that achieves the desired marginals.

A biproportional scaling is direct if the sequences can be chosen to be constant, that is \( Y = Y^{(k)} \) and \( Z = Z^{(k)} \) for some diagonal matrices \( Y, Z \in \text{diag}(\mathbb{R}^{m}_0) \) and all \( k \), hence \( B = YWZ \). In this case \( W \) factorizes as \( W = Y^{-1}BZ^{-1} \). For \( W \in \mathcal{S} \), we denote the set of all symmetric direct biproportional fits of \( W \) by \( \Psi(W) := \{ XWXX \mid X \in \text{diag}(\mathbb{R}^{n}_0) \} \).

The following lemma (Pukelsheim, 2014, Theorem 1) shows that any choice of row and column marginals determines a biproportional fit uniquely, if existing.

**Lemma 2.1 (Biproportional fits are unique).** If \( B_1, B_2 \) are two biproportional fits of \( W \in \Omega \) to row marginals \( r \in \mathbb{R}^m_0 \) and column marginals \( c \in \mathbb{R}^n_0 \), then it holds that \( B_1 = B_2 \).

This justifies to talk about the biproportional fit of \( W \) to (row and column marginals) \( r \) and \( c \). The next lemma can be derived from results by Menon (1968). It characterizes directness as exactly those biproportional scalings that satisfy \( E(B) = E(W) \).

**Lemma 2.2 (Directness).** Let \( B \) denote any biproportional scaling of \( W \). Then \( E(B) \subseteq E(W) \). Further \( E(B) = E(W) \) if and only if \( B \) is direct.

If a biproportional scaling \( B \) is not direct then \( E(W) \setminus E(B) \neq \emptyset \) implies that some diagonal entries in \( Y^{(k)} \) and \( Z^{(k)} \) must tend to zero, hence some others to infinity because of the positive marginals constraint.

**Bregman projections.** For matrices \( X, W \in \Omega \) with \( E(X) \subseteq E(W) \), the generalized relative-entropy error (generalized Kullback-Leibler divergence) is defined as

\[
RE(X\|W) := \sum_{i,j} x_{ij} \log(x_{ij}/w_{ij}) - x_{ij} + w_{ij}
\]

with the continuous extension \( 0 \cdot \log(0/w_{ij}) := 0 \). \( RE \) is a Bregman divergence, that is it can be represented as \( D_h(X\|W) := h(X) - h(W) + \langle h(W), X - W \rangle \) for a continuously differentiable strictly convex function \( h \), here \( h(X) = \sum_{i,j} x_{ij} \log(x_{ij}) - x_{ij} \). It holds that \( D_h(X\|W) \geq 0 \), with equality if and only if \( X = W \).

For any closed convex \( \mathcal{M} \subseteq \Omega \) the corresponding Bregman projection \( \mathcal{P}^h_{\mathcal{M}}(Q) \) of \( Q \in \Omega \) to \( \mathcal{M} \) is defined as

\[
\mathcal{P}^h_{\mathcal{M}}(Q) := \arg \min_{M \in \mathcal{M}} D_h(M\|Q),
\]

which is unique whenever existing. For \( \mathcal{D}_h = \mathcal{R}E \) we refer to (1) as the \( \mathcal{R}E \)-projection, denoted by \( \mathcal{P}_\mathcal{M} \). For
\[ R := \Omega(r, \cdot, W) \text{ and } C := \Omega(\cdot, c, W) \text{ it holds that} \]
\[ PR(X) = \text{diag}(r) \cdot \text{diag}(X^1)^{-1} \cdot X, \]
\[ PC(X) = X \cdot \text{diag}(X^T)^{-1} \cdot \text{diag}(c) \quad (2) \]

Both \( RE \)-projections in (2) have the same zeros as \( X \), that is \( E(PR(X)) = E(PC(X)) = E(X) \subseteq E(W) \). The corresponding minimization problem (1) is nonsmooth because there exists \( Z \in R \) (resp. \( Z \in C \)) with \( z_{ij} = 0 < x_{ij} \) for some \( ij \), which implies the partial derivative \( \lim_{t \to 0} \log(\epsilon/x_{ij}) = -\infty \). See the supplement for a proof of Equation (2) that adapts the standard Lagrangian approach to this setting.

**Mean functions.** A function \( m : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \) is a (homogeneous) mean function if it is symmetric, \( m(x, y) = m(y, x) \), minmax-bounded, \( m(x, y) \in [\min(x, y), \max(x, y)] \), and homogeneous, \( cm(x, y) = m(cx, cy) \) for all \( c > 0 \). As a rich example consider the family of Hölder \( p \)-means, defined as \( m_p(x, y) := \sqrt[p]{(x^p + y^p)/2} \) for parameter \( p \in \mathbb{R} \cup \{\pm \infty\} \). It contains the minimum function for \( p = -\infty \), the maximum function for \( p = \infty \), the arithmetic mean \( m_A(x, y) := (x + y)/2 \) for \( p = 1 \), and for \( p = 0 \) the geometric mean \( m_G(x, y) := \sqrt{xy} \). A mean function \( m \) is sub-arithmetic if \( m(x, y) \leq m_A(x, y) \) for all \( x, y \in \mathbb{R}_{\geq 0} \). It is strictly sub-arithmetic if \( x \neq y \) implies strict inequality. Similarly for (strictly) super-arithmetic.

**3 SYMMETRIZATION**

In this section we define three sequences of matrices:

**IPF:** traditional IPF-sequence (\( W^{(k)} \))

**PSIPF:** Pseudo-Symmetric IPF-sequence (\( W^{(lk)} \))

**SIPF:** Symmetric IPF-sequence (\( W^{(k)} \)).

In the symmetric setting, all three start with the same matrix \( W \) and then generate an individual sequence of matrices. We show that all sequences converge to the same limit \( W \in S(f, W) \) whenever \( S(f, W) \neq \emptyset \).

Our first lemma summarizes specific properties of biproportional fits in the symmetric setting. The proof can be found in the supplement.

**Lemma 3.1 (Symmetric biproportional fit).** Let \( B \) denote the biproportional fit of \( W \in S \) to row and column marginals \( f \in \mathbb{R}_{\geq 0}^n \). Then

(i) \( B = B^T \) is symmetric

(ii) \( B = \lim_{k \to \infty} W_k \) for a sequence of \( W_k \in \Psi(W) \)

(iii) \( B \in \Psi(W) \) if and only if \( B \) is direct

Lemma 3.1 shows that, in the symmetric setting, we can find a sequence of matrices in \( \Psi(W) \) that converges to the biproportional fit. The IPF-sequence is not of this type because its row-column-alternation inherently generates non-symmetric matrices \( W^{(k)} \notin \Psi(W) \) in general. The PSIPF-sequence is derived from the IPF-sequence such that \( W^{(k)} \in \Psi(W) \). The SIPF-sequence satisfies that \( W^{(k)} \in \Psi(W) \) without being based on the IPF-sequence. Further, SIPF arises naturally in some applications, see Section 6.

We define each of the three sequences in two ways: recursively by a first-order recursion, and factorized by referring inductively back to the initial matrix \( W \).

**IPF-sequence.** For \( W \in \Omega, r \in \mathbb{R}_{\geq 0}^n \) and \( c \in \mathbb{R}_{\geq 0}^n \), the IPF-sequence is generated by alternately applying the two \( RE \)-projections in (2), that is with \( W^{(0)} := W \):

\[ W^{(k+1)} := \begin{cases} PR(W^{(k)}) & \text{if } k \text{ even} \\ PC(W^{(k)}) & \text{if } k \text{ odd} \end{cases} \]

This implies the factorization \( W^{(k)} = Y^{(k)}WZ^{(k)} \) for some positive diagonal matrices \( Y^{(k)} = \text{diag}(y^{(k)}) \) and \( Z^{(k)} = \text{diag}(z^{(k)}) \) with \( y^{(0)} = z^{(0)} = 1 \). The factorization can be computed explicitly by the RAS-method (Bacharach, 1965), which directly updates the diagonals \( y^{(k)} \) and \( z^{(k)} \) for \( k \geq 1 \) in each step by

\[ \begin{align*}
  y^{(k)} &= r/(Wz^{(k-1)}) \\
  z^{(k)} &= c/(Wy^{(k-1)})
\end{align*} \]

where the division of vectors is meant to be entry-wise.

**PSIPF-sequence.** A straightforward strategy for symmetrizing the IPF-sequence is to choose any mean function \( m \) and apply it entry-wise along the sequence \( (m(W^{(k)}), (W^{(k)})^T) \). Every matrix in this sequence is symmetric. In order to get that it is further in \( \Psi(W) \), we choose the geometric mean function \( m_G \). This defines the PSIPF-sequence (\( W^{(lk)} \)) of \( W \) for \( k \geq 0 \) as

\[ W^{(lk)} := m_G(W^{(k)}), (W^{(k)})^T = \left[ \sqrt{w_{ij}^{(k)} w_{ji}^{(k)}} \right] \]

which implies that \( W^{(0)} = W \). The factorized representation is given with \( S^{(lk)} := \sqrt{Y^{(k)}Z^{(k)}} \) as

\[ W^{(lk)} = S^{(lk)} W S^{(lk)^T} = \left[ \sqrt{y_{ij}^{(k)} z_{ij}^{(k)}} w_{ij} \sqrt{y_{ji}^{(k)} z_{ji}^{(k)} w_{ji}} \right] \]

In particular it holds that \( W^{(lk)} \in \Psi(W) \).

**SIPF-sequence.** Both IPF and PSIPF carry out the \( RE \)-projections (2) alternately one after the other. The approach taken by SIPF is to aggregate both projections simultaneously at each step by taking their entry-wise geometric means. More general, for any mean function \( m \), we define the \( m \)-sequence (\( W^{(k)} \)) of \( W \) by \( W^{(0)} := W \) and, in the symmetric setting,
\[ W^{(k+1)} := m(\mathcal{P}_R(W^{(k)}), \mathcal{P}_C(W^{(k)})) \]
\[ = m \left( \frac{m}{d_{ij}^{(k)}}, \frac{m}{d_{ij}^{(k)}} \right) \]
\[ = \left[ w_{ij} \prod_{\ell=0} w_{ij} \right] \],

where \( d_{ij}^{(k)} = \left[ d_{ij}^{(k)} \right] := W^{(k)} \mathbf{1} \) denotes the positive row (and column) marginals of symmetric \( W^{(k)} = \left[ w_{ij}^{(k)} \right] \). For example, the arithmetic mean gives \( m \left( \frac{f_i}{d_{ij}}, \frac{f_j}{d_{ij}} \right) \).

In this case we get immediately from \( \nabla h^* \circ \nabla h = \text{id} \) that (5) coincides with (5) for all \( k \) without affecting the limit. In this case we get immediately from \( \nabla h^* \circ \nabla h = \text{id} \) that (5) coincides with (5) for all \( k \) without affecting the limit. In this case we get immediately from \( \nabla h^* \circ \nabla h = \text{id} \) that (5) coincides with (5) for all \( k \).

### 4 CONVERGENCE

In this section we study the convergence of IPF, PSIPF and SIPF, as defined in the previous section.

#### 4.1 Convergence of the IPF-sequence

The following theorem has been proved in the literature in various ways. See for example Pukelsheim (2014) and the references therein for an overview.

**Theorem 4.1 (Convergence of IPF).** Let \( W \in \Omega, r \in \mathbb{R}_{\geq 0}, c \in \mathbb{R}_{\geq 0} \) such that \( \Omega(r, c, W) \neq \emptyset \). Then the IPF-sequence converges to \( \mathcal{P}_{\Omega(r,c,W)}(W) \).

Here we present a novel proof that provides an intuitive understanding on why RE-optimality holds. We motivate this approach in more generality: assume that some set \( F \subseteq \mathbb{R}^N \) can be written as the non-empty intersection of finitely many closed convex sets, that is \( F = C_1 \cap \ldots \cap C_k \neq \emptyset \), and let \( \mathcal{P}_F(z) \) denote the RE-projection of \( z \) to \( C_i \). Our goal is to determine \( \mathcal{P}_F(x_0) \in F \) for some given \( x_0 \). The easier goal of finding an arbitrary element from \( F \) can be solved by the iterative projection method (with Bregman projections), that is to cycle again and again through all the \( C_i \)'s while projecting the previous result to \( C_i \):

\[ x_k := \mathcal{P}_{[k]}(x_{k-1}) \]

for \( k \geq 1 \), where \([k] := 1+(k-1 \mod \ell)\). Observe that (4) equals the IPF-sequence for \( C_i := R, C_2 := C, \) and \( F := \Omega(r, c, W) = R \cap C, \) which aims at approximating \( P_F(W) \) for some given \( x_0 := W \in \mathbb{R}^{m \times n} \). Bregman (1967) proves that (4) converges to some solution \( x^* \in F \), but in general with \( x^* \neq P_F(x_0) \). For the special case of orthogonal projections \( P^\perp \), Boyle and Dykstra (1985) provide a strategy for ensuring that \( x^* = P^\perp_F(x_0) \) by considering a modified sequence: they add a specific reflection term to each pre-image before applying the projection. Bauschke and Lewis (2000) generalize this idea to a large family of Bregman projections that particularly includes orthogonal projections and RE-projections. They introduce a similar reflection term that depends on the function \( h \) that induces the Bregman divergence. This defines Dykstra’s algorithm with Bregman projections by:

\[ x_k := [P^\perp_k] \circ \nabla h^*(\nabla h(x_{k-1}) + r_k) \]

and

\[ r_k := \nabla h(x_{k-1}) - \nabla h(x_k) \]

for \( k \geq 1 \) with \( r_1 := 0 \) whenever \( i \leq 0 \), and \( h^* \) the convex conjugate of \( h \). Bauschke and Lewis (2000) prove convergence of (5) to the limit \( x^* = P^\perp_F(x_0) \). Further, their Theorem 4.3 shows that if all \( C_i \)'s are affline, then one can drop all reflection terms (i.e., set \( r_k := 0 \) for all \( k \)) without affecting the limit. In this case we get immediately from \( \nabla h^* \circ \nabla h = \text{id} \) that (5) coincides with (5) for all \( k \).
with (4). In particular, this gives that IPF equals (5) with all reflection terms dropped. So it remains to show why dropping the reflection terms does not affect the (RE-optimal) limit \( x^* = \mathcal{P}_\mathcal{F}(x_0) \) in this case, although neither \( \mathcal{R} \) nor \( \mathcal{C} \) are affine.

Our key insight is that the limit \( x^* = \mathcal{P}_\mathcal{F}(x_0) \) is unaffected even under a weaker notion of affinity: it already suffices that each element \( x_k \) for \( k \geq 1 \) of the non-reflected sequence (4) is locally affine. That is, each \( \mathcal{C}_i \) is a subset of an affine space \( \mathcal{A}_i \), and there exists \( \epsilon_k > 0 \) such that \( \{ x \in \mathcal{A}_i \mid \| x - x_k \|_2 \leq \epsilon_k \} \subseteq \mathcal{C}_i \).

By choosing \( \mathcal{A}_i \) as the affine hull of \( \mathcal{C}_i \), we get that this is equivalent to claiming that each \( x_k \) lies in the relative interior of \( \mathcal{C}[k] \), see Boyd and Vandenberghe (2004). Local affinity trivially holds if \( \mathcal{C}[k] \) is affine itself. Summarized, our contribution to this framework is that all reflection terms in Dykstra’s algorithm with Bregman projections can be dropped without affecting the limit whenever sequence (4) is locally affine. Please refer to the supplement for technical details.

It is straightforward to see that the IPF-sequence is locally affine: \( \mathcal{C}_i (= \mathcal{R}) \) is a subset of the affine space \( \mathcal{A}_i := \{ X \in \mathbb{R}^{m \times n} \mid X1 = r \} \). Similarly for \( \mathcal{C}_2 \).

From (2) we get that \( E(W^{(k)}) = E(W) \). Thus, for \( k \geq 1 \) and \( \epsilon_k = \min \{ w_i^{(k)} \mid w_i^{(k)} > 0 \} \), every matrix from the ball of radius \( \epsilon_k \) around \( W^{(k)} \) in \( \mathcal{A}_i \) lies in \( \mathcal{C}[k] \).

The fruit of the above is the following compact proof.

**Proof (of Theorem 4.1).** IPF equals Dykstra’s algorithm with Bregman projections and with all reflection terms dropped, which does not affect its RE-optimal limit because the IPF-sequence is locally affine. □

It further follows from uniqueness and Lemma 3.1 that \( \mathcal{P}_{S(\mathcal{F},W)}(W) = \mathcal{P}_{S(\mathcal{F},W)}(W) \) in the symmetric setting.

### 4.2 Convergence of the PSIPF-sequence

The convergence of PSIPF follows from the convergence of IPF. Indeed it is easy to see that \( (m(W^{(k)})^t, W^{(k)}) \) converges to \( \mathcal{P}_{S(\mathcal{F},W)}(W) \) for any mean function \( m \), along symmetric matrices. However, only for \( m = m_G \), that is for the PSIPF-sequence, all intermediate matrices lie in \( \Psi(W) \).

**Theorem 4.2 (Convergence of PSIPF).** Let \( W \in \mathcal{S} \) and \( f \in \mathbb{R}_{\geq 0}^n \) such that \( S(f, W) \neq 0 \). Then the PSIPF-sequence \( W^{(k)} \) converges to \( \mathcal{P}_{S(\mathcal{F},W)}(W) \). Further, \( W^{(k)} \in \Psi(W) \) for all \( k \geq 0 \).

The proof is by reduction to IPF, see the supplement.

### 4.3 Convergence of the SIPF-sequence

In this section we prove convergence of the SIPF-sequence, that is the \( m \)-sequence for \( m = m_G \).

**Theorem 4.3 (Convergence of SIPF).** Let \( W \in \mathcal{S} \) and \( f \in \mathbb{R}_{\geq 0}^n \) such that \( S(f, W) \neq 0 \). Then the SIPF-sequence \( (W^{(k)}) \) converges to \( \mathcal{P}_{S(\mathcal{F},W)}(W) \). Further, \( W^{(k)} \in \Psi(W) \) for all \( k \geq 0 \), and \( \| f - W^{(k)}1 \|_1 \to 0 \) monotonously decreasing.

In contrast to the PSIPF-sequence, the limit \( \hat{W}_m \) of an \( m \)-sequence can differ depending on the choice of \( m \). If \( m = \min \) or \( m = \max \), then the \( m \)-sequence can even converge to an infeasible limit \( \hat{W}_m \notin S(f, W) \neq 0 \). It is left open for future work whether \( \hat{W}_m \) can be related by an appropriate \( f \)-divergence to \( W \) for \( m \neq m_G \). The proof of Theorem 4.3 requires the following four lemmas, which are of their own interest because they even hold for \( m \neq m_G \). All proofs can be found in the supplement. Throughout this section, \( (W^{(k)}) \) denotes the \( m \)-sequence of \( W \) and \( d^{(k)} := W^{(k)}1 \). The first lemma guarantees \( L_1 \)-monotony for every \( m \)-sequence.

**Lemma 4.4 (\( L_1 \)-monotony).** For any \( W \in \mathcal{S} \) and any mean function \( m \), the \( m \)-sequence of \( W \) implies that \( \| f - d^{(k)} \|_1 \) is monotonously decreasing.

The second lemma bounds the “volume” \( \| d^{(k)} \|_1 \) from above or below by \( \| f \|_1 \) if the mean function is sub-arithmetic or super-arithmetic.

**Lemma 4.5 (Volume bounds).** For any \( W \in \mathcal{S} \) and any mean function \( m \), the \( m \)-sequence of \( W \) satisfies for all \( k \geq 1 \) that
\[
(i) \quad \| d^{(k)} \|_1 \leq \| f \|_1 \quad \text{if } m \text{ is sub-arithmetic}
\]
\[
(ii) \quad \| d^{(k)} \|_1 = \| f \|_1 \quad \text{if } m = m_A
\]
\[
(iii) \quad \| d^{(k)} \|_1 \geq \| f \|_1 \quad \text{if } m \text{ is super-arithmetic}
\]

If \( m \) is strict in (i) or (iii), then equality holds if and only if \( f_i/d_i^{(k)} = f_j/d_j^{(k)} \) for all \( w_{ij} \neq 0 \).

The third lemma characterizes all limit points of \( m \)-sequences. It shows that all that remains in order to prove Theorem 4.3 is to prove that \( \| f - d^{(k)} \|_1 \to 0 \).

**Lemma 4.6 (Limit points).** Every \( m \)-sequence is bounded and has at least one limit point \( W^* \). If \( \| f - d^{(k)} \|_1 \to 0 \), then every limit point \( W^* \) satisfies \( W^*1 = f \). If further \( m = m_G \), then \( W^* \) is the (unique) biproportional fit of \( W \) to row and column marginals \( f \), and it holds that \( W^{(k)} \to W^* \).

The fourth lemma proves strong convergence under relative-entropy error if the volumes bound each other.

**Lemma 4.7 (Strong convergence).** For any \( x := (x_1, \ldots, x_n) \in \mathbb{R}_{\geq 0}^n \), \( a := (a_1, \ldots, a_n) \in \mathbb{R}_{\geq 0}^n \) with \( \sum x_i \leq \sum a_i \), let \( f(x) := \sum a_i \log \frac{a_i}{x_i} \). Then
\[
f(x) \geq 0 \text{ with equality iff } x = a.
\]

Further, for any sequence \( (x^{(k)})_{k \geq 0} \) in \( \mathbb{R}_{\geq 0}^n \) with \( \sum x_i^{(k)} \leq \sum a_i \), it holds that
\[
\lim_{k \to \infty} f(x^{(k)}) = 0 \iff \lim_{k \to \infty} x^{(k)} = a.
\]

530
Now we are ready to prove that $\|f - d(k)\|_1 \to 0$. The proof is partly inspired by ideas of Pretzel (1980).

**Proof (of Theorem 4.3).** By assumption there exists some $A = \{a_{ij} \in S(f, W)\}$. Equation (3) gives that $w_{ij} = w_{ij} \cdot u_{ij}^{-1} \cdot (k)$ with $u_{ij} = \prod_{\ell=0}^k m \left( \frac{v_{ij}, s_{ij}}{s_{ij}} \right) \neq 0$ and $s_{ij}^{(k)} := f_i / d_i^{(k)}$ for all $i, j \in \{1, \ldots, n\}$ and $k \geq 0$.

From $w_{ij} = 0 \Rightarrow u_{ij}^{(0)} = 0 \Rightarrow a_{ij} = 0$ we get that

$$v(k+1) := \sum_i a_{ij} \log \left( \frac{w_{ij}^{(k+1)}}{w_{ij}} \right) = \sum_i a_{ij} \log \left( \frac{u_{ij}^{(k)}}{u_{ij}} \right) = \sum_i a_{ij} \sum_{\ell=0}^k m \left( \frac{v_{ij}, s_{ij}}{s_{ij}} \right).$$

For $k \geq 1$ this gives for $m = m_G$ that

$$v(k+1) - v(k) = \sum_i a_{ij} \log m \left( \frac{s_{ij}}{s_{ij}} \right) = \frac{1}{2} \sum_i a_{ij} \log \left( s_{ij} + \log s_{ij} \right) = \frac{1}{2} \sum_i f_i \log s_{ij} + \frac{1}{2} \sum_j f_j \log s_{ij} = \sum_i f_i \log s_{ij}.$$  

Since $m_G$ is sub-arithmetic, we get from Lemma 4.5 that $\sum_i d_i^{(k)} \leq \sum_i f_i$. This allows to apply Lemma 4.7 which gives that $v(k+1) \geq v(k)$, thus, $(v(k))_{k \geq 0}$ is monotonically increasing. Further $v(k)$ is bounded from above because $w_{ij} \leq \|d(k)\|_1 \leq \|f\|_1 = \sum_i a_{ij}$ implies together with $w_{ij} := \min\{w_{ij} | w_{ij} > 0\}$ that

$$v(k) = \sum_i a_{ij} \log \left( \frac{w_{ij}}{w_{ij}} \right) \leq \sum_i a_{ij} \log \left( \frac{\|f\|_1/w_{ij}}{w_{ij}} \right) = \|f\|_1 \cdot \log \left( \frac{\|f\|_1}{w_{ij}} \right) < \infty.$$

It follows that $\lim_{k \to \infty} v(k)$ exists, which implies that $\sum_i f_i \log f_i/d_i^{(k)} = v(k+1) - v(k) \to 0$ and hence with Lemma 4.7 that $d_i^{(k)} \to f_i$ for all $i \in \{1, \ldots, n\}$. Thus $\|f - d(k)\|_1 \to 0$ with monotonically given by Lemma 4.4. This proves that $S(f, W) \neq \emptyset$ is sufficient to let the $m_G$-sequence converge to some $\bar{W} \in S(f, W)$.

By Lemma 4.6 we get that $\bar{W}$ is the unique biproportional fit of $W$ to $f$, hence the same limit as for the IPF-sequence. In particular this implies that $\bar{W}$ is $RE$-optimal, thus $\bar{W} = \mathcal{P}_{S(f, W)}(W)$.  

**Corollary 4.8 (Maximality).** For all $A \in S(f, W)$ it holds that $E(A) \subseteq E(\mathcal{P}_{S(f, W)}(W))$.

**Proof.** For all $A \in S(f, W)$, equation (8) implies that $a_{ij} \log \left( \frac{w_{ij}^{(k)}}{w_{ij}} \right) + (\|f\|_1 - a_{ij}) \log \left( \frac{\|f\|_1}{w_{ij}} \right) \geq v(k) \geq v(1) > -\infty$, thus $a_{ij} \neq 0 \Rightarrow \lim_{k \to \infty} w_{ij}^{(k)} \neq 0$.  

## 5 Fitting Graph Matrices

In this section we study the applicability of the three convergence theorems (4.1, 4.2 and 4.3) to undirected weighted graphs. For any $d \in \mathbb{R}_{\geq 0}$ and $W \in S(d,)$, we denote by $\mathcal{G}(W) := (V, E, W)$ the graph on vertex set $V = \{1, \ldots, n\}$ with an undirected edge $ij = ji$ of weight $w_{ij} = w_{ji}$ for every $ij \in E(W)$. The edge weights sum up to the degree vector $d = W1 = W^T 1$.

All three theorems base on the non-emptiness assumption $\mathcal{G}(f, f, W) \neq \emptyset$ and $S(f, W) \neq \emptyset$, respectively, which are easily seen to be equivalent. Moreover, there exists a strictly positive solution $M \in \mathcal{G}(f, f, W)$, that is some $M$ with $E(M) = E(W)$, if and only if $S(f, W)$ contains a strictly positive solution. Thus, it suffices to focus only on $S(f, W)$ in the following.

In the language of graphs, non-emptiness $S(f, W) \neq \emptyset$ means to assume that there exists an assignment of new weights $\bar{w}_{ij} \in [0, \infty)$ to all existing edges in $\mathcal{G}(W)$ such that the new vertex degrees $\bar{d}_i := \sum_{j \in V} \bar{w}_{ij}$ equal $f_i$ for all $i \in V$. Strict positivity further restricts the assignment to take only positive values $\bar{w}_{ij} \in (0, \infty)$.

We now reformulate the non-emptiness assumption and the existence of a strictly positive solution in terms of weighted vertex expansion properties in $\mathcal{G}(W)$. Let $N(S) := \{j \in V \mid \exists i \in S : ij \in E\}$ denote the vertex neighborhood of $S \subseteq V$. A graph is a weak $f$-expander for a positive vector $f$ if it holds for all $S \subseteq V$ that

$$\sum_{i \in N(S)} f_i \geq \sum_{i \in S} f_i.$$  

A weak $f$-expander is strict for $S$ if inequality (8) is strict for $S$. This is a weaker notion of “expansion”, which typically refers to the stronger assumption that (8) holds for the boundary $N(S) \setminus S$ instead of the neighborhood $N(S)$. In particular, if $W$ has a positive diagonal, then $\mathcal{G}(W)$ has self-loops at all vertices, and is a weak $f$-expander for every choice of $f \in \mathbb{R}_{\geq 0}$.

**Proposition 5.1 (Feasibility).** Let $W \in S$ and $f \in \mathbb{R}_{\geq 0}$. Then $S(f, W) \neq \emptyset$ if and only if $\mathcal{G}(W)$ is a weak $f$-expander.

Proposition 5.1 is already known in other formulations. For example in network theory for flows in directed graphs as the Gale-Hoffman theorem, which is a weighted variant of Hall’s Marriage Theorem. Further it appears as $(c) \Leftrightarrow (d)$ in Theorem 3 of Pukelsheim (2014), and it has also been proved for multi-graphs by Behrend (2013, Theorem 6). Proposition 5.1 implies convergence of IPF/PSIPF/SIPF to the biproportional fit $B \in S(f, W)$ if and only if $\mathcal{G}(W)$ is weakly $f$-expanding. However, it does not guarantee directeness of $\bar{W}$, so factorized approaches can be numerically infeasible. Corollary 4.8 gives that $W$ is direct if and only if there exists any strictly positive $A \in S(f, W)$. The key to the existence of $A$ is to claim that the weak $f$-expansion of $\mathcal{G}(W)$ is strict for all sets $S$ whenever “possible in principle”, that is up to “trivial” cases that enforce equality. Let us make this precise: the
vertex set of any graph can uniquely be partitioned into \( V = D_1 \cup ... \cup D_K \), where each \( D_\ell \) is either a non-bipartite connected component, or it forms for an \( \ell' \in \{ \ell - 1, \ell + 1 \} \) the unique bipartition \( D_\ell \cup D_{\ell'} \) of some bipartite connected component. We refer to \( V = \bigcup_{\ell=1}^K D_\ell \) as the \textit{non-bipartite decomposition} of the graph. \( K \) equals the number of non-bipartite connected components plus twice the number of bipartite connected components. Observe that whenever \( S \) equals the union of any of the \( D_\ell \)'s, its \( \mathbf{f} \)-weighted expansion (8) is forced to hold with equality. The following proposition shows that a strictly positive solution exists if and only if these are the only exceptions, so if every other \( S \) implies strict inequality in (8).

**Proposition 5.2 (Strictly positive feasibility).** Let \( G(W) = (V, E, W) \) denote the graph corresponding to \( W \in \mathcal{S} \), and \( V = \bigcup_{\ell=1}^K D_\ell \) its non-bipartite decomposition. For any \( \mathbf{f} \in \mathbb{R}^n_{\geq 0} \) there exists a strictly positive solution in \( \mathcal{S}(\mathbf{f}, W) \) if and only if \( G(W) \) is a weak \( \mathbf{f} \)-expander that is strict for all \( S \notin \{ \bigcup_{\ell=1}^K D_\ell \mid I \subseteq \{1, \ldots, K\} \} \).

Proposition 5.2 is implicitly proved by Brualdi (1968) in terms of sub-matrices of \( W \) and by Behrend (2013, Theorem 7) in terms of tri-partitions of multi-graphs. The special case of connected non-bipartite graphs, that is \( K = 1 \), is equivalent to considering symmetric “connected matrices” as Pukelsheim (2014, Theorem 2). However, it takes a considerable effort to transform any of these results to the plain formulation presented here. See the supplement for details.

Note that Zass and Shashua (2005) misleadingly state their Proposition 1 (convergence to a doubly stochastic limit) to hold for every non-negative symmetric matrix, which omitted the necessary feasibility conditions. A counterexample is the simple path graph on 3 vertices, for which the iteration does not converge to a doubly stochastic limit. We emphasize again that convergence is provided only for a specific family of matrices: \( G(W) \) must be a weak \( \mathbf{f} \)-expander, and whenever a factorized approach is used, \( G(W) \) must additionally satisfy the strictness assumptions in Proposition 5.2. In particular the factorization stated in (Zass and Shashua, 2006, Proposition 2) does not apply in general. The tempting Lagrangian approach is invalid whenever \( E(\bar{W}) \neq E(W) \).

### 6 APPLICATIONS

IPF is widely used in many different fields, so there is a potential impact on a variety of applications. Let us point out some examples:

- **Earth Mover’s Distance (EMD).** For \( a, b \in \mathbb{R}^d_{\geq 0} \), the EMD equals the minimum cost of a transportation plan for moving masses (initially distributed according to \( a \)) between the dimensions such that the result is \( b \). The costs for moving masses are determined by some additional information \( M \in \mathbb{R}^{d \times d}_{\geq 0} \) on pairwise distances between dimensions, denoted as the ground metric. In image analysis the canonical candidate for the ground metric is the Euclidean distance between the pixel positions. Known algorithms for computing EMD take at least time super-cubic in \( d \).

- **Cuturi (2013) introduces the Sinkhorn distance by adding an entropic regularization to EMD that avoids over-complex transportation plans in a precise sense.** The dual of this modified distance can be approximated efficiently by performing IPF of the matrix \( K := \exp(-M) \) towards row marginals \( a \) and column marginals \( b \). Convergence is guaranteed for all \( a, b \) because all entries in \( K \) are positive. **We suggest to consider non-negative IPF here**, that is to allow for 0-entries in \( K \), which correspond to \( \infty \)-entries in \( M \). These entries refer to pairs of dimensions between which no direct mass transport is possible. Restricting the mass transport to, for example, nearby dimensions is an additional sense of regularization. This allows for sparse \( K \), which is a crucial improvement, since for example 250 \( \times \) 250 images already imply that \( d = 62500 \), which requires 32 GB memory for storing a dense \( K \) with double-precision. As long as \( a \) and \( b \) are not “fully incompatible” (i.e., \( \Omega(a, b, K) = 0 \)) their “sparse EMD” stays finite. This can be exploited as follows: in classification, data can often be normalized (e.g., images by centering, scaling, and rotation). After normalization, the mass transport can likely be restricted to smaller radii without leading to fully incompatible pairs of elements from the same or similar classes. Further, instead of a single sparse \( K \), we might consider a sequence \( K_1, K_2, ..., K_R \) with decreasing sparsity (e.g., by doubling the radii of possible mass transport when going from \( K_i \) to \( K_{i+1} \)). Then one may first compute very sparse EMD on \( K_1 \) very efficiently, and only in case of infinite distance continue with \( K_2, K_3, ..., \) until a finite distance is attained. This can drastically speed up the average computing time. The same generalization from positive to non-negative IPF applies to the work of Cuturi and...
Symmetric Iterative Proportional Fitting

Dooucet (2014) on computing the Barycenter of multiple points \(a_1, a_2, \ldots, a_n \in \mathbb{R}^2_0\) with respect to a regularized 2-Wasserstein distance.

**ICE method.** Imakaev et al. (2012) introduce a method called “Iterative Correction and Eigenvector decomposition” (ICE) for studying genomes by observing the interactions between different locations on it. The measurements give a histogram of pairwise interactions between the locations (double-sided reads). Optionally, additional one-directed actions (single-sided reads) are considered, which we leave out here. The double-sided reads define a graph \(G(W)\) with the locations as vertices and with the empirical amounts of pairwise interaction as edge weights.

Values and eigenvectors of \(L(W)\) refer to a first step approximation of scaling \(W\) towards degree vector \(f\) by SIPF. Indeed experiments show that \(W^{(1)} f \approx 1\), whenever the structure of the non-zero entries is not too restrictive; however, \(W^{(1)} f \approx 1\) is still correlated to the original degree vector \(d\). This relation between the Laplacian and SIPF generalizes to \(f \neq 1\). Kurra et al. (2014) introduce the \(f\)-adjusted Laplacian matrix \(L_f(W)\) for parameter \(f \in \mathbb{R}^n_0\), which generalizes the normalized Laplacian. \(L_f(W)\) refers to a modified graph \(\tilde{W}_f\) that is constructed from \(W\) by applying SIPF with \(f\) for a single step, followed by adding the residuals \(f - d^{(1)}\) along the main diagonal. Thus \(L_f(W)\) refers to a first step approximation of scaling \(W\) to fit degree vector \(f\) by SIPF. This interpretation motivates a new variant of a graph Laplacian matrix: for \(f \in \mathbb{R}^n_0\) we define the \(f\)-fitted Laplacian of \(W \in S\) by \(\hat{L}_f(W) := \mathcal{P}_{S_f(W)}\), where SIPF is the natural candidate in order to approximate \(\mathcal{P}_{S_f(W)}\). Whenever \(W^{(1)} \approx \tilde{W}\), the geometric interpretation of \(\tilde{W}_f\) as a density shift for geometric graphs also applies to \(W\). It is an interesting question for future work to determine the differences between \(L_f(W)\) and \(\hat{L}_f(W)\) in the case that \(W^{(1)} \approx \tilde{W}\). Summarized, spectral analysis of the \(f\)-fitted Laplacian \(\hat{L}_f(W)\) infers about \(W\) after “correcting” its degrees to \(f\) by replacing \(W\) with its relative-entropy nearest re-weighting that provides degrees \(f\). This approach already has a prominent application, as it captures the idea of the ICE method.

**7 CONCLUSION.** It is folklore that the iterative projection method converges to a feasible solution (if existing). We prove that local affinity is sufficient for this limit to be the Bregman projection of the initial element. Our result contributes a novel and purely algorithmic intuition for the fact that IPF converges to the relative-entropy optimum. However, the IPF-sequence does not fit well to the symmetric setting. We introduce two symmetric alternatives to IPF, and prove convergence. Both variants allow for a factorized approach that is preferable over the recursive approach whenever applicable. We characterize all feasible symmetric settings, in particular those in which the factorized approach is applicable, in a way that is far more intuitive than previous results. Finally, we point out open questions and various applications in order to motivate future work.

**ACKNOWLEDGEMENTS.** The author would like to thank Ulrike von Luxburg and Gilles Blanchard for helpful discussions. This research is supported by the German Research Foundation via the Research Unit 1735 “Structural Inference in Statistics: Adaptation and Efficiency”.

533
References


