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 $\widehat{W}$. This limit is the relative-entropy nearest solution to W among all matrices of prescribed row marginals $\mathbf{r}$ and column marginals $\mathbf{c}$. 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 $\mathbf{r} = \mathbf{c}$. 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 $\widehat{W}$ 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, Kruithof'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 $\mathbf{r}$ and $\mathbf{c}$ that specify new target marginals for the rows and columns of W, respectively. The traditional IPF-sequence $(W^{\langle k \rangle})$ is determined by $W^{\langle 0 \rangle} := W$, and for $k \geq 1$ in an alternating way by first re-scaling all rows to have marginals $\mathbf{r}$, then re-scaling all columns to have marginals $\mathbf{c}$, then re-scaling the rows again, and so forth. The IPF-sequence converges, under certain conditions, to a limit matrix $\widehat{W} = \lim_{k \to \infty} W^{\langle k \rangle}$ that simultaneously achieves the desired row marginals $\mathbf{r}$ and column marginals $\mathbf{c}$. In the case that all entries in W are positive, the IPF-sequence is well-understood: Sinkhorn (1967) proves that for *any* choice of $W \in \mathbb{R}^{m \times n}_{>0}$, $\mathbf{r} \in \mathbb{R}^m_{>0}$, $\mathbf{c} \in \mathbb{R}^n_{>0}$ with $\|\mathbf{r}\|_1 = \|\mathbf{c}\|_1$ the IPF-sequence converges to a unique limit matrix $\widehat{W} \in \mathbb{R}^{m \times n}_{>0}$ of row marginals $\mathbf{r}$ and column marginals $\mathbf{c}$. Further $\widehat{W}$ has the form $\widehat{W} = YWZ$ for positive diagonal matrices Y, Z that are unique up to a scaling factor. By the Lagrangian approach it is straightforward to prove that $\widehat{W}$ 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 $\mathbf{r} = \mathbf{c} = \mathbf{1}$, 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 $\mathbf{r}$, $\mathbf{c}$, 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: $\widehat{W}$ has *at least* the same zero-entries as W, and whenever $\widehat{W}$ 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.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

He represents $\widehat{W}$ and W as absolutely continuous measures, which allows to consider their Radon-Nikodym derivative. This extends the relative-entropy interpretation of $\widehat{W}$ 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 $\mathbf{f} := \mathbf{r} = \mathbf{c}$. 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 XWX for positive diagonal X.

In the symmetric setting, we derive in Section 5 necessary and sufficient conditions for the existence of $\widehat{W}$ in terms of expansion properties of the undirected weighted graph $\mathcal{G}(W)$. 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 Ω the set of non-negative $m \times n$ matrices that contain no zero row and no zero column, that is $\Omega := \{X \in \mathbb{R}^{m\times n}_{\geq 0} \mid X\mathbf{1} > 0, X^T\mathbf{1} > 0\}$ with all inequalities applied entry-wise and $\mathbf{1} := (1, \ldots, 1)^T$. The restriction of Ω to symmetric matrices is denoted by $\mathcal{S} := \{X \in \mathbb{R}^{n\times n}_{\geq 0} \mid X = X^T, X\mathbf{1} > 0\}$. 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 $\mathbf{f} \in \mathbb{R}^n_{>0}$ and a symmetric matrix $W \in \mathcal{S}$ we define the constrained subset $\mathcal{S}(\mathbf{f}, W) := \{X \in \mathbb{R}^{n\times n}_{\geq 0} \mid X = X^T, X\mathbf{1} = \mathbf{f}, E(X) \subseteq E(W)\} \subseteq \mathcal{S}$ of those matrices in $\mathcal{S}$ that have row (and column) marginals $\mathbf{f}$ while preserving the zeros of W. Similarly, for positive vectors $\mathbf{r}$ and $\mathbf{c}$ we define $\Omega(\mathbf{r}, \mathbf{c}, W) := \{X \in \mathbb{R}^{m\times n}_{\geq 0} \mid X\mathbf{1} = \mathbf{r}, X^T\mathbf{1} = \mathbf{c}, E(X) \subseteq E(W)\} \subseteq \Omega$. We drop individual constraints by a dot, for example $\Omega(\mathbf{r}, \cdot, \cdot) = \{X \in \mathbb{R}^{m\times n}_{\geq 0} \mid X\mathbf{1} = \mathbf{r}\}$ $(\not\subseteq \Omega)$.

Matrix scaling. $B \in \Omega$ is a *biproportional scaling* of $W \in \Omega$ if it can be expressed as $B = \lim_{k\to\infty} Y^{(k)} W Z^{(k)}$ for two sequences of diagonal matrices $(Y^{(k)})$ and $(Z^{(k)})$ with positive diagonals $\mathbf{y}^{(k)}$ and $\mathbf{z}^{(k)}$. Such scalings of W often aim at fitting B to prescribed row marginals $\mathbf{r} \in \mathbb{R}^m_{>0}$ and column marginals $\mathbf{c} \in \mathbb{R}^n_{>0}$. For that reason we also denote a biproportional scaling B with $\mathbf{r} = B\mathbf{1}$ and $\mathbf{c} = B^T\mathbf{1}$ as a *biproportional fit* of W to row marginals $\mathbf{r}$ and column marginals $\mathbf{c}$. 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 $Y, Z \in \mathrm{diag}(\mathbb{R}^n_{>0})$ 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 $\Psi(W) := \{XWX \mid X \in \mathrm{diag}(\mathbb{R}^n_{>0})\}$.

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 $\mathbf{r} \in \mathbb{R}^m_{>0}$ and column marginals $\mathbf{c} \in \mathbb{R}^n_{>0}$, then it holds that $B_1 = B_2$.*

This justifies to talk about *the* biproportional fit of W to (row and column marginals) $\mathbf{r}$ and $\mathbf{c}$. 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 $D_h(X\|W) := h(X) - h(W) + \langle \nabla h(W), X - W\rangle$ for a continuously differentiable strictly convex function h, here $h(X) = \sum_{ij} x_{ij}\log(x_{ij}) - x_{ij}$. 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

$$\mathcal{P}^h_{\mathcal{M}}(Q) \;:=\; \arg\min_{M\in\mathcal{M}} \; D_h(M\|Q), \qquad (1)$$

which is unique whenever existing. For $D_h = RE$ we refer to (1) as the *RE-projection*, denoted by $\mathcal{P}_{\mathcal{M}}$. For

$\mathcal{R} := \Omega(\mathbf{r}, \cdot, W)$ and $\mathcal{C} := \Omega(\cdot, \mathbf{c}, W)$ it holds that

$$\begin{aligned} \mathcal{P}_{\mathcal{R}}(X) &= \operatorname{diag}(\mathbf{r}) \cdot \operatorname{diag}(X\mathbf{1})^{-1} \cdot X \quad , \\ \mathcal{P}_{\mathcal{C}}(X) &= X \cdot \operatorname{diag}(X^T\mathbf{1})^{-1} \cdot \operatorname{diag}(\mathbf{c}) \quad . \end{aligned} \tag{2}$$

Both *RE*-projections in (2) have the same zeros as X, that is $E(\mathcal{P}_{\mathcal{R}}(X)) = E(\mathcal{P}_{\mathcal{C}}(X)) = E(X) \subseteq E(W)$. The corresponding minimization problem (1) is non-smooth because there exists $Z \in \mathcal{R}$ (resp. $Z \in \mathcal{C}$) with $z_{ij} = 0 < x_{ij}$ for some ij, which implies the partial derivative $\lim_{\epsilon \to 0} \log(\epsilon / x_{ij}) = -\infty$. 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, $c\, m(x, y) = m(cx, cy)$ 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 $(W^{\langle k \rangle})$

PSIPF: Pseudo-Symmetric IPF-sequence $(W^{\langle\langle k \rangle\rangle})$

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 $\widehat{W} \in \mathcal{S}(\mathbf{f}, W)$ whenever $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$.

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 $W \in \mathcal{S}$ to row and column marginals $\mathbf{f} \in \mathbb{R}^n_{>0}$. 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 $W^{\langle k \rangle} \notin \Psi(W)$ in general. The PSIPF-sequence is derived from the IPF-sequence such that $W^{\langle\langle k \rangle\rangle} \in \Psi(W)$. 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 $W \in \Omega$, $\mathbf{r} \in \mathbb{R}^m_{>0}$, and $\mathbf{c} \in \mathbb{R}^n_{>0}$, the IPF-sequence is generated by alternately applying the two *RE*-projections in (2), that is with $W^{\langle 0 \rangle} := W$:

$$W^{\langle k+1 \rangle} := \begin{cases} \mathcal{P}_{\mathcal{R}}(W^{\langle k \rangle}) & \text{if } k \text{ even} \ , \\ \mathcal{P}_{\mathcal{C}}(W^{\langle k \rangle}) & \text{if } k \text{ odd} \ . \end{cases}$$

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

$$\begin{cases} \mathbf{y}^{\langle k \rangle} = \mathbf{r} / (W \mathbf{z}^{\langle k-1 \rangle}) \ , \ \mathbf{z}^{\langle k \rangle} = \mathbf{z}^{\langle k-1 \rangle} & \text{if } k \text{ odd} \ , \\ \mathbf{y}^{\langle k \rangle} = \mathbf{y}^{\langle k-1 \rangle} \ , \ \mathbf{z}^{\langle k \rangle} = \mathbf{c} / (W^T \mathbf{y}^{\langle k-1 \rangle}) & \text{if } k \text{ even} \ , \end{cases}$$

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 $(m(W^{\langle k \rangle}, (W^{\langle k \rangle})^T))$. 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 $(W^{\langle\langle k \rangle\rangle})$ of W for $k \geq 0$ as

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

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

$$W^{\langle\langle k \rangle\rangle} = S^{\langle\langle k \rangle\rangle} \, W \, S^{\langle\langle k \rangle\rangle} = \left[\sqrt{y_i^{\langle k \rangle} z_i^{\langle k \rangle}} \; w_{ij} \; \sqrt{y_j^{\langle k \rangle} z_j^{\langle k \rangle}} \right] .$$

In particular it holds that $W^{\langle\langle k \rangle\rangle} \in \Psi(W)$.

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,

$$\begin{aligned} W^{(k+1)} &:= m(\mathcal{P}_{\mathcal{R}}(W^{(k)}), \mathcal{P}_{\mathcal{C}}(W^{(k)})) \\ &= \left[m\left(\tfrac{f_i}{d_i^{(k)}} \cdot w_{ij}^{(k)}, \tfrac{f_j}{d_j^{(k)}} \cdot w_{ij}^{(k)} \right) \right] \\ &= \left[w_{ij} \cdot \textstyle\prod_{\ell=0}^{k} m\left(\tfrac{f_i}{d_i^{(\ell)}}, \tfrac{f_j}{d_j^{(\ell)}} \right) \right], \end{aligned} \tag{3}$$

where $\mathbf{d}^{(k)} = [d_i^{(k)}] := W^{(k)}\mathbf{1}$ denotes the positive row (and column) marginals of symmetric $W^{(k)} = [w_{ij}^{(k)}]$. For example, the arithmetic mean gives the m_A-sequence $((\mathcal{P}_{\mathcal{R}}(W^{(k)}) + \mathcal{P}_{\mathcal{C}}(W^{(k)}))/2)$. The SIPF-sequence is defined as the m_G-sequence. It allows for the following recursive and factorized representations, where $F := \operatorname{diag}(\mathbf{f})$ and $D^{(k)} := \operatorname{diag}(\mathbf{d}^{(k)})$:

$$\begin{aligned} W^{(k+1)} &= \sqrt{\tfrac{F}{D^{(k)}}} \cdot W^{(k)} \cdot \sqrt{\tfrac{F}{D^{(k)}}} \\ &= S^{(k)} \cdot W \cdot S^{(k)} \quad \text{for} \quad S^{(k)} = \sqrt{\textstyle\prod_{\ell=0}^{k} \tfrac{F}{D^{(\ell)}}}\,. \end{aligned}$$

This particularly implies that $W^{(k)} \in \Psi(W)$.

SIPF has already been studied implicitly in the literature for the special case of $\mathbf{f} = \mathbf{1}$, that is for a doubly stochastic limit. One aspect that makes this case special is that it represents the projection to the Birkhoff polytope, which allows for a variety of specialized arguments. Zass and Shashua (2005) sketch a proof idea based on the monotony of the matrix permanent. Recently, Knight et al. (2014) provide a rigorous convergence proof based on the monotony of $\prod_i d_i^{(k)} (= \prod_i d_i^{(k)}/f_i)$. However, the general case $\mathbf{f} \neq \mathbf{1}$ loses several monotony properties, in particular these two stated here.

Recursive versus factorized. In case of non-direct $B = \lim_{k\to\infty} T^{(k)} W T^{(k)}$, some entries in $T^{(k)}$ tend to infinity. This makes any factorized approach numerically infeasible. All recursive approaches guarantee bounded values by avoiding to represent $T^{(k)}$. But whenever applicable, the factorized approach has the strong advantage of providing self-stabilization: any numerical errors $\xi \in \mathbb{R}^n$ affect the intermediate result only in the form of $(T^{(k)} + \operatorname{diag}(\xi)) \cdot W \cdot (T^{(k)} + \operatorname{diag}(\xi))$, which is always close to $\Psi(W)$ up to machine precision. Further, the limit remains the same for *any* starting point in $\Psi(W)$, even in presence of arbitrary large sporadic errors ξ. So the limit is ensured to equal B up to machine precision. In contrast to that, the limit of a recursive approach can drift away under numerical errors towards another limit $\tilde{B} \in \Psi(W + \Delta)$ for cumulated errors $\Delta \in \mathbb{R}^{n\times n}$. Although $\tilde{B}$ provides the specified marginals, it can differ from the intended biproportional fit $B \in \Psi(W)$ by more than machine precision. Thus one should prefer the factorized approach whenever it is feasible, that is whenever the limit B is direct: $B = TWT \in \Psi(W)$. We characterize in Section 5 all cases where directness holds true.

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 $W \in \Omega$, $\mathbf{r} \in \mathbb{R}^m_{>0}$, $\mathbf{c} \in \mathbb{R}^n_{>0}$ such that $\Omega(\mathbf{r}, \mathbf{c}, W) \neq \emptyset$. Then the IPF-sequence converges to $\mathcal{P}_{\Omega(\mathbf{r},\mathbf{c},W)}(W)$.*

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 $\mathcal{F} \subseteq \mathbb{R}^N$ can be written as the non-empty intersection of finitely many closed convex sets, that is $\mathcal{F} = \mathcal{C}_1 \cap \ldots \cap \mathcal{C}_\ell \neq \emptyset$, and let $\mathcal{P}_i(z)$ denote the *RE*-projection of z to $\mathcal{C}_i$. Our goal is to determine $\mathcal{P}_{\mathcal{F}}(x_0) \in \mathcal{F}$ for some given x_0. The easier goal of finding an *arbitrary* element from $\mathcal{F}$ can be solved by the *iterative projection method (with Bregman projections)*, that is to cycle again and again through all the $\mathcal{C}_i$'s while projecting the previous result to $\mathcal{C}_i$:

$$x_k := \mathcal{P}_{[k]}(x_{k-1}) \tag{4}$$

for $k \geq 1$, where $[k] := 1 + (k - 1 \bmod \ell)$. Observe that (4) equals the IPF-sequence for $\mathcal{C}_1 := \mathcal{R}$, $\mathcal{C}_2 := \mathcal{C}$, and $\mathcal{F} := \Omega(\mathbf{r}, \mathbf{c}, W) = \mathcal{R} \cap \mathcal{C}$, which aims at approximating $P_{\mathcal{F}}(W)$ for some given $x_0 := W \in \mathbb{R}^{m\cdot n}$. Bregman (1967) proves that (4) converges to *some* solution $x^* \in \mathcal{F}$, but in general with $x^* \neq \mathcal{P}_{\mathcal{F}}(x_0)$. For the special case of orthogonal projections $\mathcal{P}^{\perp}$, Boyle and Dykstra (1985) provide a strategy for ensuring that $x^* = \mathcal{P}^{\perp}_{\mathcal{F}}(x_0)$ 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:

$$\begin{aligned} x_k &:= (\mathcal{P}^h_{[k]} \circ \nabla h^*)(\nabla h(x_{k-1}) + r_{k-\ell}) \\ \text{and} \quad r_k &:= \nabla h(x_{k-1}) + r_{k-\ell} - \nabla h(x_k) \end{aligned} \tag{5}$$

for $k \geq 1$ with $r_i := 0$ whenever $i \leq 0$, and h^* the convex conjugate of h. Bauschke and Lewis (2000) prove convergence of (5) to the limit $x^* = \mathcal{P}^h_{\mathcal{F}}(x_0)$. Further, their Theorem 4.3 shows that if all $\mathcal{C}_i$'s are affine, 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 $\nabla h^* \circ \nabla h = \mathrm{id}$ that (5) coincides

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 $x^* = \mathcal{P}^h_{\mathcal{F}}(x_0)$ 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 $\{x \in \mathcal{A}_{[k]} \mid \|x - x_k\|_2 \leq \epsilon_k\} \subseteq \mathcal{C}_{[k]}$. 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: $\mathcal{C}_1 (= \mathcal{R})$ is a subset of the affine space $\mathcal{A}_1 := \{X \in \mathbb{R}^{m \times n} \mid X\mathbf{1} = \mathbf{r}\}$. Similarly for $\mathcal{C}_2$. From (2) we get that $E(W^{\langle k \rangle}) = E(W)$. Thus, for $k \geq 1$ and $\epsilon_k < \min\{w_{ij}^{\langle k \rangle} \mid w_{ij}^{\langle k \rangle} > 0\}$, every matrix from the ball of radius ϵ_k around $W^{\langle k \rangle}$ in $\mathcal{A}_{[k]}$ 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 $\mathcal{P}_{\Omega(\mathbf{f},\mathbf{f},W)}(W) = \mathcal{P}_{\mathcal{S}(\mathbf{f},W)}(W)$ 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 $(m(W^{\langle k \rangle}, (W^{\langle k \rangle})^T))$ converges to $\mathcal{P}_{\mathcal{S}(\mathbf{f},W)}(W)$ 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 $\mathbf{f} \in \mathbb{R}^n_{>0}$ such that $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$. Then the PSIPF-sequence $(W^{\langle\langle k \rangle\rangle})$ converges to $\mathcal{P}_{\mathcal{S}(\mathbf{f},W)}(W)$. Further, $W^{\langle\langle k \rangle\rangle} \in \Psi(W)$ 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 $\mathbf{f} \in \mathbb{R}^n_{>0}$ such that $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$. Then the SIPF-sequence $(W^{(k)})$ converges to $\mathcal{P}_{\mathcal{S}(\mathbf{f},W)}(W)$. Further, $W^{(k)} \in \Psi(W)$ for all $k \geq 0$, and $\|\mathbf{f} - W^{(k)}\mathbf{1}\|_1 \to 0$ monotonously decreasing.*

In contrast to the PSIPF-sequence, the limit $\widehat{W}_m$ 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 $\widehat{W}_m \notin \mathcal{S}(\mathbf{f}, W) \neq \emptyset$. It is left open for future work whether $\widehat{W}_m$ 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 $\mathbf{d}^{(k)} := W^{(k)}\mathbf{1}$. 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 $\|\mathbf{f} - \mathbf{d}^{(k)}\|_1$ is monotonously decreasing.*

The second lemma bounds the "volume" $\|\mathbf{d}^{(k)}\|_1$ from above or below by $\|\mathbf{f}\|_1$ 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*

$$\begin{array}{lrcll} \textit{(i)} & \|\mathbf{d}^{(k)}\|_1 & \leq & \|\mathbf{f}\|_1 & \textit{if } m \textit{ is sub-arithmetic} \\ \textit{(ii)} & \|\mathbf{d}^{(k)}\|_1 & = & \|\mathbf{f}\|_1 & \textit{if } m = m_A \\ \textit{(iii)} & \|\mathbf{d}^{(k)}\|_1 & \geq & \|\mathbf{f}\|_1 & \textit{if } m \textit{ is super-arithmetic} \end{array}$$

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 $\|\mathbf{f} - \mathbf{d}^{(k)}\|_1 \to 0$.

Lemma 4.6 (Limit points). *Every m-sequence is bounded and has at least one limit point W^*. If $\|\mathbf{f} - \mathbf{d}^{(k)}\|_1 \to 0$, then every limit point W^* satisfies $W^*\mathbf{1} = \mathbf{f}$. If further $m = m_G$, then W^* is the (unique) biproportional fit of W to row and column marginals $\mathbf{f}$, 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 $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{R}^n_{>0}$, $\mathbf{a} := (a_1, \ldots, a_n) \in \mathbb{R}^n_{>0}$ with $\sum_i x_i \leq \sum_i a_i$ let $f(\mathbf{x}) := \sum_i a_i \log \frac{a_i}{x_i}$. Then*

$$f(\mathbf{x}) \geq 0 \textit{ with equality iff } \mathbf{x} = \mathbf{a}. \tag{6}$$

Further, for any sequence $(\mathbf{x}^{(k)})_{k \geq 0}$ in $\mathbb{R}^n_{>0}$ with $\sum_i x_i^{(k)} \leq \sum_i a_i$ it holds that

$$\lim_{k \to \infty} f(\mathbf{x}^{(k)}) = 0 \quad \Leftrightarrow \quad \lim_{k \to \infty} \mathbf{x}^{(k)} = \mathbf{a}. \tag{7}$$

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

Proof (of Theorem 4.3). By assumption there exists some $A = [a_{ij}] \in \mathcal{S}(\mathbf{f}, W)$. Equation (3) gives that $w_{ij}^{(k+1)} = w_{ij} \cdot u_{ij}^{(k)}$ with $u_{ij}^{(k)} = \prod_{\ell=0}^{k} m\left(s_i^{(\ell)}, s_j^{(\ell)}\right) \neq 0$ and $s_i^{(\ell)} := f_i / d_i^{(\ell)}$ for all $i, j \in \{1, \dots, n\}$ and $k \geq 0$. From $w_{ij} = 0 \Leftrightarrow w_{ij}^{(k)} = 0 \Rightarrow a_{ij} = 0$ we get that

$$\begin{aligned} v^{(k+1)} &:= \textstyle\sum_{i,j} a_{ij} \log(w_{ij}^{(k+1)} / w_{ij}) = \sum_{i,j} a_{ij} \log u_{ij}^{(k)} \\ &= \textstyle\sum_{i,j} a_{ij} \sum_{\ell=0}^{k} \log m\left(s_i^{(\ell)}, s_j^{(\ell)}\right) \quad . \end{aligned}$$

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

$$\begin{aligned} v^{(k+1)} - v^{(k)} &= \textstyle\sum_{i,j} a_{ij} \log m\left(s_i^{(k)}, s_j^{(k)}\right) \\ &= \textstyle\frac{1}{2} \sum_{i,j} a_{ij} (\log s_i^{(k)} + \log s_j^{(k)}) \\ &= \textstyle\frac{1}{2} \sum_i f_i \log s_i^{(k)} + \frac{1}{2} \sum_j f_j \log s_j^{(k)} \\ &= \textstyle\sum_i f_i \log f_i / d_i^{(k)} \quad . \end{aligned}$$

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 monotonously increasing. Further $v^{(k)}$ is bounded from above because $w_{ij}^{(k)} \leq \|\mathbf{d}^{(k)}\|_1 \leq \|\mathbf{f}\|_1 = \sum_{i,j} a_{ij}$ implies together with $w_{min} := \min\{w_{ij} \mid w_{ij} > 0\}$ that

$$\begin{aligned} v^{(k)} &= \textstyle\sum_{i,j} a_{ij} \log(w_{ij}^{(k)} / w_{ij}) \leq \sum_{i,j} a_{ij} \log(\|\mathbf{f}\|_1 / w_{min}) \\ &= \|\mathbf{f}\|_1 \cdot \log(\|\mathbf{f}\|_1 / w_{min}) \ < \ \infty. \quad (\divideontimes) \end{aligned}$$

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 $i \in \{1, \dots, n\}$. Thus $\|\mathbf{f} - \mathbf{d}^{(k)}\|_1 \to 0$ with monotony given by Lemma 4.4. This proves that $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$ is sufficient to let the m_G-sequence converge to some $\widehat{W} \in \mathcal{S}(\mathbf{f}, W)$. By Lemma 4.6 we get that $\widehat{W}$ is the unique biproportional fit of W to $\mathbf{f}$, hence the same limit as for the IPF-sequence. In particular this implies that $\widehat{W}$ is *RE*-optimal, thus $\widehat{W} = \mathcal{P}_{\mathcal{S}(\mathbf{f}, W)}(W)$. □

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

Proof. For all $A \in \mathcal{S}(\mathbf{f}, W)$, equation ($\divideontimes$) implies that $a_{ij} \log(w_{ij}^{(k)} / w_{ij}) + (\|\mathbf{f}\|_1 - a_{ij}) \log(\|\mathbf{f}\|_1 / w_{min}) \geq v^{(k)} \geq v^{(1)} > -\infty$, 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 $\mathbf{d} \in \mathbb{R}^n_{>0}$ and $W \in \mathcal{S}(\mathbf{d}, \cdot)$, we denote by $\mathcal{G}(W) := (V, E, W)$ the graph on vertex set $V = \{1, \dots, n\}$ 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 $\mathbf{d} = W\mathbf{1} = W^T\mathbf{1}$. All three theorems base on the *non-emptiness assumption* $\Omega(\mathbf{f}, \mathbf{f}, W) \neq \emptyset$ and $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$, respectively, which are easily seen to be equivalent. Moreover, there exists a *strictly positive solution* $M \in \Omega(\mathbf{f}, \mathbf{f}, W)$, that is some M with $E(M) = E(W)$, if and only if $\mathcal{S}(\mathbf{f}, W)$ contains a strictly positive solution. Thus, it suffices to focus only on $\mathcal{S}(\mathbf{f}, W)$ in the following.

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

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* $\mathbf{f}$*-expander* for a positive vector $\mathbf{f}$ if it holds for all $S \subseteq V$ that

$$\textstyle\sum_{i \in N(S)} f_i \quad \geq \quad \sum_{i \in S} f_i \quad . \tag{8}$$

A weak $\mathbf{f}$-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 $\mathbf{f}$-expander for every choice of $\mathbf{f} \in \mathbb{R}^n_{>0}$.

Proposition 5.1 (Feasibility). *Let $W \in \mathcal{S}$ and $\mathbf{f} \in \mathbb{R}^n_{>0}$. Then $\mathcal{S}(\mathbf{f}, W) \neq \emptyset$ if and only if $\mathcal{G}(W)$ is a weak $\mathbf{f}$-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 $B \in \mathcal{S}(\mathbf{f}, W)$ if and only if $\mathcal{G}(W)$ is weakly $\mathbf{f}$-expanding. However, it does not guarantee directness of $\widehat{W}$, so factorized approaches can be numerically infeasible. Corollary 4.8 gives that $\widehat{W}$ is direct if and only if there exists *any* strictly positive $A \in \mathcal{S}(\mathbf{f}, W)$. The key to the existence of A is to claim that the weak $\mathbf{f}$-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 $V = D_1 \dot\cup \ldots \dot\cup D_K$, where each D_ℓ is either a non-bipartite connected component, or it forms for an $\ell' \in \{\ell-1, \ell+1\}$ the unique bipartition $D_\ell \dot\cup D_{\ell'}$ of some bipartite connected component. We refer to $V = \bigcup_{\ell=1}^{K} D_\ell$ as the *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_i'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 $\mathcal{G}(W) = (V, E, W)$ denote the graph corresponding to $W \in \mathcal{S}$, and $V = \bigcup_{i=\ell}^{K} D_\ell$ its non-bipartite decomposition. For any $\mathbf{f} \in \mathbb{R}^n_{>0}$ there exists a strictly positive solution in $\mathcal{S}(\mathbf{f}, W)$ if and only if $\mathcal{G}(W)$ is a weak $\mathbf{f}$-expander that is strict for all $S \notin \{\bigcup_{\ell\in I} D_\ell \mid I \subseteq \{1,\ldots,K\}\}$.*

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 omits 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: $\mathcal{G}(W)$ must be a weak $\mathbf{f}$-expander, and whenever a factorized approach is used, $\mathcal{G}(W)$ 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 $E(\widehat{W}) \neq E(W)$.

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: similar to Knight et al. (2014) for the doubly stochastic case, SIPF can serve as a preconditioner even for $\mathbf{f} \neq \mathbf{1}$. In Quadratic Non-Negative Matrix Factorization, SIPF is the canonical candidate for determining the factorization of $W \in \mathcal{S}$ into the form $W = YXY$ for a positive diagonal matrix Y and with X constrained to prescribed marginals $\mathbf{f}$. Further, Corollary 4.8 corresponds to the combinatorial problem of identifying the unique minimum set of edges in a graph that must be removed (set to weight 0) in order to be able to achieve degree vector $\mathbf{f}$ on the remaining edges. We now present three other related applications in more detail.

Earth Mover's Distance (EMD). For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d_{\geq 0}$ the EMD equals the minimum cost of a transportation plan for moving masses (initially distributed according to $\mathbf{a}$) between the dimensions such that the result is $\mathbf{b}$. 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 $\mathbf{a}$ and column marginals $\mathbf{b}$. Convergence is guaranteed for all $\mathbf{a}$, $\mathbf{b}$ 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 ∞-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 250×250 images already imply that $d = 62500$, which requires 32 GB memory for storing a dense K with double-precision. As long as $\mathbf{a}$ and $\mathbf{b}$ are not "fully incompatible" (i.e., $\Omega(\mathbf{a}, \mathbf{b}, K) = \emptyset$) 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 computation time. The same generalization from positive to non-negative IPF applies to the work of Cuturi and

Doucet (2014) on computing the Barycenter of multiple points $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n \in \mathbb{R}^d_{\geq 0}$ 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 $\mathcal{G}(W)$ with the locations as vertices and with the empirical amounts of pairwise interaction as edge weights. Degrees in this graph refer to the observed "visibility" of that location. These visibilities are non-uniform due to several biases in the experiment. Deeper domain knowledge suggests that the "true" visibilities should be uniform. This motivates to find a degree-balanced matrix that "best approximates" the empirical data W; such as the relative-entropy nearest doubly stochastic matrix $\widehat{W} := \mathcal{P}_{\mathcal{S}(\mathbf{1},W)}(W)$. It can be approximated by IPF or its symmetrized variants. The authors suggest another iterative algorithm, denoted as Iterative Correction (IC), without proving convergence. Indeed the convergence behavior is different, since IC can diverge for unconnected graphs. So **we suggest to replace IC by SIPF** whenever one-sided reads are skipped. This modification further enables to deal with non-uniform visibilities $\mathbf{f} \neq \mathbf{1}$. This allows for applications where uniform visibility does not hold, or where some biases are known and can be corrected separately, so we only need to correct for the remaining (non-uniform) bias. Another application is the comparison of two different genome matrices W_1 and W_2. The authors suggest to $\mathbf{1}$-balance W_1 and W_2 individually before comparing them. SIPF alternatively allows to directly compare the $\mathbf{d}_2$-fitted W_1 to W_2 and vice versa. The last step in the ICE method is an analysis of the largest eigenvectors of the "corrected" graph matrix $\widehat{W}$. For doubly stochastic $\widehat{W}$ this is equal to classical multi-way spectral analysis, that is to explore structural properties of $\widehat{W}$ by the smallest eigenvectors of its normalized Laplacian matrix. Hence, in this case the ICE method can compactly be summarized as classical spectral analysis of the normalized Laplacian of the "$\mathbf{1}$-fitted" graph. This motivates the following application, which generalizes this approach to $\mathbf{f} \neq \mathbf{1}$.

f-fitted Laplacian matrix. The normalized graph Laplacian matrix of $W \in \mathcal{S}$ is defined by $\mathcal{L}(W) := I - D^{-1/2} W D^{-1/2}$. It has the same eigenvectors (with eigenvalue λ_i mapped to $1 - \lambda_i$) as the matrix $D^{-1/2} W D^{-1/2} = W^{(1)}$, the first element of the SIPF-sequence for $\mathbf{f} = \mathbf{1}$. This allows for a novel interpretation of the type of normalization in $\mathcal{L}(W)$: eigenvalues and eigenvectors of $\mathcal{L}(W)$ refer to a first step approximation of scaling W towards degree vector $\mathbf{1}$ by SIPF. Indeed experiments show that $W^{(1)}\mathbf{1} \approx \mathbf{1}$, whenever the structure of the non-zero entries is not too restrictive; however, $W^{(1)}\mathbf{1}$ is still correlated to the original degree vector $\mathbf{d}$. This relation between the Laplacian and SIPF generalizes to $\mathbf{f} \neq \mathbf{1}$: Kurras et al. (2014) introduce the $\mathbf{f}$-adjusted Laplacian matrix $\mathcal{L}_{\mathbf{f}}(W)$ for parameter $\mathbf{f} \in \mathbb{R}^n_{>0}$, which generalizes the normalized Laplacian. $\mathcal{L}_{\mathbf{f}}(W)$ refers to a modified graph $\overline{W}_{\mathbf{f}}$ that is constructed from W by applying SIPF with $\mathbf{f}$ for a single step, followed by adding the residuals $\mathbf{f} - \mathbf{d}^{(1)}$ along the main diagonal. Thus $\mathcal{L}_{\mathbf{f}}(W)$ refers to a first step approximation of scaling W to fit degree vector $\mathbf{f}$ by SIPF. This interpretation motivates a new variant of a graph Laplacian matrix: for $\mathbf{f} \in \mathbb{R}^n_{>0}$ **we define the f-fitted Laplacian** of $W \in \mathcal{S}$ by $\widehat{\mathcal{L}}_{\mathbf{f}}(W) := \mathcal{L}(\mathcal{P}_{\mathcal{S}(\mathbf{f},W)})$, where SIPF is the natural candidate in order to approximate $\mathcal{P}_{\mathcal{S}(\mathbf{f},W)}$. Whenever $W^{(1)} \approx \widehat{W}$, the geometric interpretation of $\overline{W}_{\mathbf{f}}$ as a density shift for geometric graphs also applies to $\widehat{W}$. It is an interesting question for future work to determine the differences between $\mathcal{L}_{\mathbf{f}}(W)$ and $\widehat{\mathcal{L}}_{\mathbf{f}}(W)$ in the case that $W^{(1)} \not\approx \widehat{W}$. Summarized, spectral analysis of the $\mathbf{f}$-fitted Laplacian $\widehat{\mathcal{L}}_{\mathbf{f}}(W)$ infers about W after "correcting" its degrees to $\mathbf{f}$ by replacing W with its relative-entropy nearest re-weighting that provides degrees $\mathbf{f}$. 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*".

References

M. Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6 (3):294–310, 1965.

H. H. Bauschke and A. S. Lewis. Dykstra's algorithm with Bregman projections: a convergence proof. *Optimization*, 48:409–427, 2000.

R. E. Behrend. Fractional perfect b-matching polytopes. I: General theory. *Linear Algebra and its Applications*, 439(12):3822–3858, 2013.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Lecture Notes in Statistics*, 37:28–47, 1985.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

R. A. Brualdi. Convex sets of nonnegative matrices. *Canadian Journal of Mathematics*, 20:144–157, 1968.

I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems (NIPS)*, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning (ICML)*, 2014.

M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012.

C. T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

P. Knight, D. Ruiz, and B. Uçar. A symmetry preserving algorithm for matrix scaling. *SIAM Journal on Matrix Analysis and Applications*, 35(3):931–955, 2014.

S. Kurras, U. von Luxburg, and G. Blanchard. The f-adjusted graph Laplacian: a diagonal modification with a geometric interpretation. In *International Conference on Machine Learning (ICML)*, 2014.

M. V. Menon. Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with predescribed row and column sums. *Canadian Journal of Mathematics*, 20:225–232, 1968.

O. Pretzel. Convergence of the iterative scaling procedure for non-negative matrices. *Journal of the London Mathematical Society*, 21(2):379–384, 1980.

F. Pukelsheim. Biproportional scaling of matrices and the iterative proportional fitting procedure. *Annals of Operations Research*, 215(1):269–283, 2014.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *International Conference on Computer Vision (ICCV)*, 2005.

R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. In *Neural Information Processing Systems (NIPS)*, 2006.