Iterative Proportional Fitting

Xuda Ye

Purdue University

November 13, 2025

1 Problem Statement

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sample spaces with $|\mathcal{X}| = M$ and $|\mathcal{Y}| = N$. Assume we are given the probability distributions $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, represented by the probability vectors

$$\mu = (\mu_i)_{i=1}^{M} \in \mathbb{R}_+^M, \qquad \nu = (\nu_j)_{j=1}^{N} \in \mathbb{R}_+^N.$$

Furthermore, let $Q = (Q_{ij})_{i,j=1}^{M,N} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a given reference probability distribution. We aim to find a new distribution $P = (P_{ij})_{i,j=1}^{M,N} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that *looks like* Q, but which has the specified marginal distributions μ and ν.

More precisely, we aim to solve the following optimization problem, which seeks the distribution P closest to Q in relative entropy while matching the marginals:

$$\min_{P \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})} D_{\mathrm{KL}}(P\|Q) \quad \text{subject to} \quad P_{\mathcal{X}} = \mu \text{ and } P_{\mathcal{Y}} = \nu. \tag{1}$$

Here, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ denote the marginal distributions of P on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the relative entropy (Kullback–Leibler divergence) $D_{\mathrm{KL}}(P\|Q)$ is defined via

$$D_{\mathrm{KL}}(P\|Q) = \sum_{i,j} P_{ij} \log \frac{P_{ij}}{Q_{ij}}.$$

Equivalently, (1) can be reformulated in terms of the components P_{ij}:

$$\min_{(P_{ij}) \in \mathbb{R}_+^{M\times N}} \sum_{i=1}^{M}\sum_{j=1}^{N} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \quad \text{subject to} \quad \sum_{j=1}^{N} P_{ij} = \mu_i \text{ and } \sum_{i=1}^{M} P_{ij} = \nu_j. \tag{2}$$

Clearly, (2) is a convex optimization problem with linear constraints.

In this paper, we introduce the **Iterative Proportional Fitting** (IPF) algorithm, a simple and efficient method for solving the optimization problem (1). Interestingly, the IPF algorithm was originally proposed by W. Edwards Deming and Frederick F. Stephan in 1940 to solve a practical problem involving the adjustment of contingency tables. Its application in information theory, particularly for solving the optimization problem (1), was discovered much later, after the 1970s.

2 Iterative Proportional Fitting

The Iterative Proportional Fitting algorithm is designed as follows:

Algorithm 1: Iterative Proportional Fitting (IPF)

Input: Marginal distributions $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$; Reference distribution $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$
Output: The distribution P solving the optimization problem (1)

Set $P^{(0)} = Q$;
for $k = 0, 1, 2, \ldots$ **do**
  `// Fit the` μ `marginal (row sums)`
  $P_{ij}^{(k+\frac{1}{2})} = P_{ij}^{(k)} \left(\frac{\mu_i}{\sum_{j'=1}^{N} P_{ij'}^{(k)}} \right)$, for all i, j;
  `// Fit the` ν `marginal (column sums)`
  $P_{ij}^{(k+1)} = P_{ij}^{(k+\frac{1}{2})} \left(\frac{\nu_j}{\sum_{i'=1}^{M} P_{i'j}^{(k+\frac{1}{2})}} \right)$, for all i, j;
Output $P = P^{(k)}$ until convergence.

This algorithm appears deceptively simple, as it merely adjusts row and column scaling factors alternately. A natural question is how this sequence of distributions $P^{(k)}$ can converge to the *solution* of the non-linear optimization problem (1).

To gain some insight, observe that the iterates $P^{(k)}$ retain a specific structure relative to the reference distribution Q. Starting from $P^{(0)} = Q$, each subsequent iterate generated by the algorithm can be expressed in the form:

$$P_{ij}^{(k)} = \alpha_i^{(k)} Q_{ij} \beta_j^{(k)}, \tag{3}$$

where $\alpha^{(k)} \in \mathbb{R}_+^M$ and $\beta^{(k)} \in \mathbb{R}_+^N$ are non-negative scaling vectors. In this sense, $P^{(k)}$ maintains the underlying structure of Q (i.e., it remains a biproportional scaling), and the algorithm iteratively updates these scaling vectors $\alpha^{(k)}$ and $\beta^{(k)}$.

On the other hand, we can prove that the unique solution $P^* = (P_{ij}^*)$ to the optimization problem (2) has exactly the same biproportional form:

$$P_{ij}^* = \alpha_i^* Q_{ij} \beta_j^*, \tag{4}$$

where $\alpha^* \in \mathbb{R}_+^M$ and $\beta^* \in \mathbb{R}_+^N$ are non-negative scaling vectors. This form can be derived formally using the method of Lagrangian multipliers. Let $f = (f_i)_{i=1}^M$ and $g = (g_j)_{j=1}^N$ be the Lagrange multipliers associated with the row-sum (μ_i) and column-sum (ν_j) constraints, respectively. Utilizing the matrix form (2), we define the Lagrangian $\mathcal{L}$ as

$$\mathcal{L} = \sum_{i=1}^{M} \sum_{j=1}^{N} P_{ij} \left(\log \frac{P_{ij}}{Q_{ij}} - 1 \right) - \sum_{i=1}^{M} f_i \left(\sum_{j=1}^{N} P_{ij} - \mu_i \right) - \sum_{j=1}^{N} g_j \left(\sum_{i=1}^{M} P_{ij} - \nu_j \right).$$

By setting the partial derivative of $\mathcal{L}$ with respect to P_{ij} to zero at the optimal point, we obtain the KKT condition satisfied by the optimal solution P^*:

$$\frac{\partial \mathcal{L}}{\partial P_{ij}} = \log \frac{P_{ij}^*}{Q_{ij}} - f_i - g_j = 0 \quad \Longrightarrow \quad P_{ij}^* = e^{f_i} Q_{ij} e^{g_j}.$$

This matches the form (4) by choosing the non-negative scalings $\alpha_i^* = e^{f_i}$ and $\beta_j^* = e^{g_j}$.

The consistency between the iterative form (3) and the optimal form (4) is a key insight. It demonstrates that the IPF algorithm implicitly searches for the correct scaling factors by producing a sequence $P^{(k)}$ that always preserves the required structure of the optimal solution.

3 Convergence Analysis

We now present a rigorous convergence analysis for the IPF algorithm. First, we define the concept of an information projection. Let C be a set of probability distributions. The **information projection** of a distribution P onto the set C, denoted $P^\perp$, is defined as the distribution in C that minimizes the relative entropy from P:

$$P^\perp := \arg\min_{R \in C} D_{\mathrm{KL}}(R \| P). \tag{5}$$

This projection satisfies the following generalized Pythagorean theorem for relative entropy.

Lemma 1 *Let C be a convex set of probability distributions, and let $P^\perp$ be the information projection of P onto C. Then, for any distribution $Q \in C$, we have*

$$D_{\mathrm{KL}}(Q \| P) \geqslant D_{\mathrm{KL}}(Q \| P^\perp) + D_{\mathrm{KL}}(P^\perp \| P). \tag{6}$$

Proof Expanding the terms in (6), the inequality is equivalent to:

$$\begin{aligned}
&\sum_i Q_i \log \frac{Q_i}{P_i} \geqslant \sum_i Q_i \log \frac{Q_i}{P_i^\perp} + \sum_i P_i^\perp \log \frac{P_i^\perp}{P_i} \\
\iff \quad &\sum_i Q_i \left(\log \frac{Q_i}{P_i} - \log \frac{Q_i}{P_i^\perp} \right) \geqslant \sum_i P_i^\perp \log \frac{P_i^\perp}{P_i} \\
\iff \quad &\sum_i Q_i \log \frac{P_i^\perp}{P_i} \geqslant \sum_i P_i^\perp \log \frac{P_i^\perp}{P_i} \\
\iff \quad &\sum_i (Q_i - P_i^\perp) \log \frac{P_i^\perp}{P_i} \geqslant 0.
\end{aligned} \tag{7}$$

We prove (7) by contradiction. Assume the inequality does not hold, i.e.,

$$\sum_i (Q_i - P_i^\perp) \log \frac{P_i^\perp}{P_i} < 0.$$

Since C is convex, for any $t \in [0,1]$, the distribution $P^t = (1-t)P^\perp + tQ$ is also in C. In particular, $P^0 = P^\perp$ is the projection. Next, we compute the derivative of $D_{\mathrm{KL}}(P^t \| P)$ with respect to t:

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(P^t \| P) = \frac{\mathrm{d}}{\mathrm{d}t} \sum_i P_i^t \left(\log \frac{P_i^t}{P_i} - 1 \right) = \sum_i (Q_i - P_i^\perp) \log \frac{P_i^t}{P_i}.$$

Evaluating this derivative at $t = 0$ (where $P_i^0 = P_i^\perp$), we have

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(P^t \| P) \bigg|_{t=0} = \sum_i (Q_i - P_i^\perp) \log \frac{P_i^\perp}{P_i} < 0.$$

This implies that for small $t > 0$, we have $D_{\mathrm{KL}}(P^t \| P) < D_{\mathrm{KL}}(P^0 \| P) = D_{\mathrm{KL}}(P^\perp \| P)$. However, $P^t \in C$, and this contradicts the definition of $P^\perp$ as the unique minimizer of $D_{\mathrm{KL}}(R\|P)$ over all $R \in C$. Therefore, the inequality (7) must hold. ■

Next, we define the two sets of distributions that satisfy one of the two marginal constraints:

$$C_\mu = \left\{ P \in \mathbb{R}_+^{M\times N} : \sum_{j=1}^{N} P_{ij} = \mu_i, \text{ for all } i \right\}, \quad C_\nu = \left\{ P \in \mathbb{R}_+^{M\times N} : \sum_{i=1}^{M} P_{ij} = \nu_j, \text{ for all } j \right\}. \tag{8}$$

It is clear that C_μ and C_ν are convex sets, as they are defined by linear equality constraints. The key insight is that the IPF algorithm is geometrically equivalent to performing alternating information projections onto these two sets.

Lemma 2 *The IPF algorithm (Algorithm 1) is equivalent to the alternating projection sequence:*

$$\boxed{P^{(k+\frac{1}{2})} = \arg\min_{R\in C_\mu} D_{\mathrm{KL}}(R\|P^{(k)}), \qquad P^{(k+1)} = \arg\min_{R\in C_\nu} D_{\mathrm{KL}}(R\|P^{(k+\frac{1}{2})}).} \tag{9}$$

Proof By symmetry, we only need to verify the first step: the projection of $P^{(k)}$ onto C_μ. By definition, this projection $P^{(k+\frac{1}{2})}$ is the solution R to the optimization problem:

$$\min_{R\in\mathbb{R}_+^{M\times N}} D_{\mathrm{KL}}(R\|P^{(k)}) \qquad \text{subject to} \qquad \sum_{j=1}^{N} R_{ij} = \mu_i, \text{ for all } i.$$

This problem is separable and can be solved independently for each row i:

$$\min_{(R_{ij})_{j=1}^N \in \mathbb{R}_+^N} \sum_{j=1}^{N} R_{ij} \log \frac{R_{ij}}{P_{ij}^{(k)}} \qquad \text{subject to} \qquad \sum_{j=1}^{N} R_{ij} = \mu_i.$$

Using the same Lagrangian multiplier approach as before (introducing f_i for the constraint on row i), the optimality condition is

$$\frac{\partial}{\partial R_{ij}} \left[R_{ij} \left(\log \frac{R_{ij}}{P_{ij}^{(k)}} - 1 \right) - f_i(R_{ij}) \right] = 0 \quad \Longrightarrow \quad \log \frac{R_{ij}}{P_{ij}^{(k)}} - f_i = 0.$$

This implies the solution $P^{(k+\frac{1}{2})}$ must have the form $P_{ij}^{(k+\frac{1}{2})} = e^{f_i} P_{ij}^{(k)}$, which is exactly the first update step in Algorithm 1. An identical argument holds for the projection onto C_ν (which is separable by columns j), thus verifying the equivalence (9). ■

Since the objective function in (2) is strictly convex over the constraint set, there exists a unique optimal solution, which we denote by P^*. The following theorem establishes that the IPF algorithm's iterates monotonically approach this solution in the sense of relative entropy.

Theorem 1 *The sequence $P^{(k)}$ generated by the IPF algorithm (Algorithm 1) satisfies*

$$D_{\mathrm{KL}}(P^*\|P^{(k+1)}) \leqslant D_{\mathrm{KL}}(P^*\|P^{(k+\frac{1}{2})}) \leqslant D_{\mathrm{KL}}(P^*\|P^{(k)}).$$

Moreover, $\lim_{k\to\infty} D_{\mathrm{KL}}(P^*\|P^{(k)}) = 0.$

Proof $P^{(k+\frac{1}{2})}$ is the information projection of $P^{(k)}$ onto the convex set C_μ. Since P^* is the optimal solution, it must satisfy the marginal constraints, which implies $P^* \in C_\mu$. We can now apply the Pythagorean theorem (Lemma 1) to obtain

$$D_{\mathrm{KL}}(P^* \| P^{(k)}) \geqslant D_{\mathrm{KL}}(P^* \| P^{(k+\frac{1}{2})}) + D_{\mathrm{KL}}(P^{(k+\frac{1}{2})} \| P^{(k)}). \tag{10}$$

Since $D_{\mathrm{KL}}(P^{(k+\frac{1}{2})} \| P^{(k)}) \geqslant 0$, equation (10) immediately implies the first inequality

$$D_{\mathrm{KL}}(P^* \| P^{(k+\frac{1}{2})}) \leqslant D_{\mathrm{KL}}(P^* \| P^{(k)}).$$

The second inequality, $D_{\mathrm{KL}}(P^* \| P^{(k+1)}) \leqslant D_{\mathrm{KL}}(P^* \| P^{(k+\frac{1}{2})})$, holds true for an identical reason by considering the projection onto C_ν (which also contains P^*).

Thus, the sequence $D_{\mathrm{KL}}(P^* \| P^{(k)})$ is monotonically decreasing in k and bounded below by 0. It must therefore converge to a limit, $a \geqslant 0$:

$$\lim_{k\to\infty} D_{\mathrm{KL}}(P^* \| P^{(k)}) = a. \tag{11}$$

Next, we prove $a = 0$. Note that the results (10) and (11) imply

$$\lim_{k\to\infty} D_{\mathrm{KL}}(P^{(k+\frac{1}{2})} \| P^{(k)}) = 0.$$

A similar argument shows $\lim_{k\to\infty} D_{\mathrm{KL}}(P^{(k+1)} \| P^{(k+\frac{1}{2})}) = 0$.

In Algorithm 1, the entries of $P_{ij}^{(k)}$ are bounded by 1, hence the sequence $\{P^{(k)}\}_{k=1}^{\infty}$ is bounded. By the Bolzano–Weierstrass theorem, there exists a convergent subsequence. Let R^* be the limit of such a subsequence:

$$R^* = \lim_{k\in I} P^{(k)},$$

where I is an infinite subset of $\mathbb{N}$. Denote the information projection onto C_μ by

$$\mathrm{Proj}_\mu(P) := \arg\min_{R\in C_\mu} D_{\mathrm{KL}}(R \| P).$$

Then $P^{(k+\frac{1}{2})} = \mathrm{Proj}_\mu(P^{(k)})$. By the continuity of the projection and the KL divergence, we take the limit along the subsequence I:

$$D_{\mathrm{KL}}\big(\mathrm{Proj}_\mu(R^*) \big\| R^*\big) = \lim_{k\in I, k\to\infty} D_{\mathrm{KL}}\big(\mathrm{Proj}_\mu(P^{(k)}) \big\| P^{(k)}\big) = 0,$$

which implies $\mathrm{Proj}_\mu(R^*) = R^*$ or $R^* \in C_\mu$. In other words, the subsequence limit R^* satisfies the marginal constraint for μ. By a similar argument using the second half-step, we also have $R^* \in C_\nu$.

Furthermore, as a consequence of the biproportional form (3), the subsequence limit R^* must also have the form

$$R^*_{ij} = \alpha^*_i Q_{ij} \beta^*_j$$

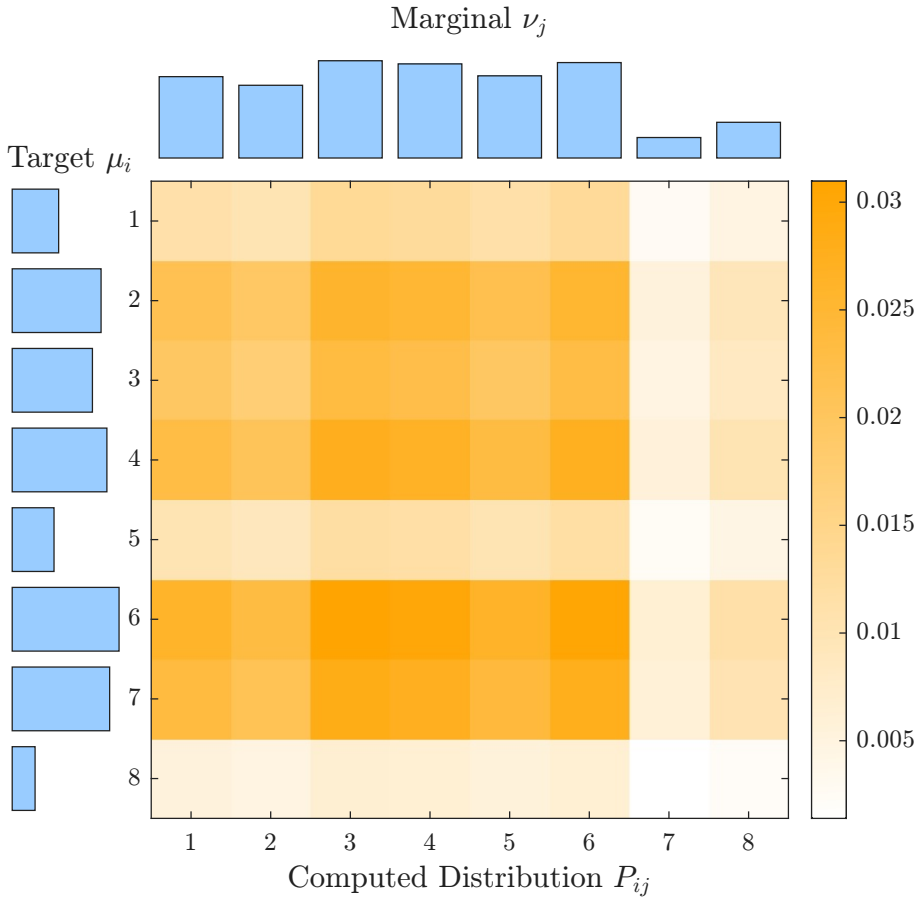
for some scaling vectors α^* and β^*. Since R^* has the biproportional form and also has marginal distributions μ and ν ($R^* \in C_\mu \cap C_\nu$), it satisfies the KKT conditions for the optimization problem (2). For this convex optimization problem with a strictly convex objective, the solution satisfying the KKT conditions is unique. Therefore, R^* is nothing but the unique optimal solution P^*. By the continuity of the KL divergence, we obtain

$$a = \lim_{k\in I, k\to\infty} D_{\mathrm{KL}}(P^* \| P^{(k)}) = D_{\mathrm{KL}}(P^* \| \lim_{k\to\infty} P^{(k)}) = D_{\mathrm{KL}}(P^* \| P^*) = 0,$$

which confirms $a = 0$, thus completing the proof. ■

4 Numerical Test

We implement a simple numerical test of the IPF algorithm. We set the marginals μ and ν to be 8-dimensional probability vectors (i.e., $M = N = 8$) and compute the optimal 8×8 probability matrix. The resulting distribution, along with the target marginals, is visualized as follows.



The source codes can be found at the following link: https://xuda-ye.wordpress.com/wp-content/uploads/2025/11/ipf.zip