Explaining Iterative Proportional Fitting (IPF) Convergence

Explaining IPF convergence

To see why IPF really does find the "maximum-cross-entropy" (i.e. minimum-KL) solution, it helps to think of each row- or column-scaling step as a special kind of projection in KL-space. In fact, IPF can be viewed as performing an alternating sequence of **"I-projections"** (Csiszár projections) onto two convex sets:

1. $\mathcal{R} = \{ P : \sum_j P_{ij} = r_i \ \forall i \}$, the set of all joint distributions whose row-marginals equal $\{r_i\}$.
2. $\mathcal{C} = \{ P : \sum_i P_{ij} = c_j \ \forall j \}$, the set of all joint distributions whose column-marginals equal $\{c_j\}$.

These two sets of distributions are both convex (in fact, each is an affine slice of the probability-simplex), and their intersection $\mathcal{R} \cap \mathcal{C}$ is exactly the set of all tables satisfying both row- and column-sums. We start from a prior $P^{(0)}$ (normalized so $\sum_{i,j} P^{(0)}_{ij} = 1$) and want

$$P^* \ = \ \operatorname*{arg\,min}_{P \in \mathcal{R} \cap \mathcal{C}} D_{\mathrm{KL}}\big(P \| P^{(0)}\big).$$

Because $D_{\mathrm{KL}}(\cdot \parallel P^{(0)})$ is strictly convex over the probability simplex, there is a unique minimizer in $\mathcal{R} \cap \mathcal{C}$. The magic of IPF is that by alternately projecting (in the KL-divergence sense) first onto $\mathcal{R}$ and then onto $\mathcal{C}$, one converges to exactly that unique minimizer. Here is why each step preserves or lowers the KL-value, and why the process converges.

1. "Projection onto a single marginal constraint" reduces KL

Suppose at some iteration t, we have a current estimate $P^{(t)}$. First we enforce the row-margins. That is, we look for

$$P^{(t+\frac{1}{2})} \ = \ \operatorname*{arg\,min}_{P \in \mathcal{R}} D_{\mathrm{KL}}\big(P \parallel P^{(t)}\big).$$

Because $\mathcal{R}$ is the set of distributions whose rows sum to $\{r_i\}$, we know from basic convex-analysis that this "I-projection" onto $\mathcal{R}$ can be done in closed form by simply re-scaling each row of $P^{(t)}$ so that it sums to r_i. Concretely, if

$$\tilde{r}_i^{(t)} \ = \ \sum_j P_{ij}^{(t)},$$

then for each cell (i, j) we set

$$P_{ij}^{(t+\frac{1}{2})} \ = \ P_{ij}^{(t)} \ \frac{r_i}{\tilde{r}_i^{(t)}}.$$

One can check (by taking the derivative of $\sum_{i,j} P_{ij} \ln(P_{ij}/P_{ij}^{(t)})$ subject to $\sum_j P_{ij} = r_i$) that this is exactly the minimizer. Because projecting onto a convex set in a strictly convex divergence always *lowers* (or at worst leaves unchanged) the divergence, we have

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})} \big\| P^{(t)}\big) \ \le \ 0,$$

and also

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})} \big\| P^{(0)}\big) \ \le \ D_{\mathrm{KL}}\big(P^{(t)} \big\| P^{(0)}\big),$$

because the KL-divergence is jointly convex and one can show (via the Pythagorean property of I-projections) that

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})} \| P^{(0)}\big) \ + \ D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})} \| P^{(t)}\big) \ = \ D_{\mathrm{KL}}\big(P^{(t)} \| P^{(0)}\big).$$

Since $D_{\mathrm{KL}}(P^{(t+\frac{1}{2})}\|P^{(t)}) \geq 0$, it follows that

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(0)}\big) \;\leq\; D_{\mathrm{KL}}\big(P^{(t)}\big\|P^{(0)}\big).$$

In other words, enforcing the row-margins "pushes us" closer (in KL-distance) to the prior $P^{(0)}$. By the same argument, the subsequent column-scaling step,

$$P^{(t+1)} \;=\; \arg\min_{P\in\mathcal{C}} D_{\mathrm{KL}}\big(P\|\, P^{(t+\frac{1}{2})}\big),$$

can be carried out by rescaling each column to sum to c_j. Again,

$$D_{\mathrm{KL}}\big(P^{(t+1)}\big\|P^{(0)}\big) \;\leq\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(0)}\big).$$

Hence each half-step (row-projection then column-projection) monotonically decreases the total divergence $D_{\mathrm{KL}}(\cdot\,\|\,P^{(0)})$. Because KL is bounded below by zero, these successive improvements force convergence (at least of the divergence values), and one can show the tables themselves converge entrywise to a unique limit $P^* \in \mathcal{R}\cap\mathcal{C}$.

2. The Pythagorean property (why the minimizer is in $\mathcal{R}\cap\mathcal{C}$)

A key fact about I-projections in KL-space is the so-called **"Pythagorean theorem for KL"**. Suppose you have a prior distribution Q, and a convex set $\mathcal{S}$. Then the KL-projection of Q onto $\mathcal{S}$,

$$Q_{\mathcal{S}} \;=\; \arg\min_{P\in\mathcal{S}} D_{\mathrm{KL}}(P\|Q),$$

satisfies, for every $P \in \mathcal{S}$,

$$D_{\mathrm{KL}}(P\|Q) \;=\; D_{\mathrm{KL}}(P\|Q_{\mathcal{S}}) \;+\; D_{\mathrm{KL}}(Q_{\mathcal{S}}\|Q).$$

Because KL is nonnegative, this immediately implies

$$D_{\mathrm{KL}}\big(Q_{\mathcal{S}}\big\|Q\big) \;\leq\; D_{\mathrm{KL}}\big(P\big\|Q\big) \quad \text{for all } P\in\mathcal{S}.$$

Hence the projection $Q_{\mathcal{S}}$ is the unique *closest* point (in KL-sense) to Q within $\mathcal{S}$; and you see that "once you've projected, you can never go 'backward' in divergence," because every subsequent projection onto some (other) convex set must satisfy a similar Pythagorean decomposition.

In IPF, at iteration t we project from $P^{(t)}$ onto $\mathcal{R}$, obtaining $P^{(t+\frac{1}{2})}$. By Pythagoras,

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(0)}\big) \;+\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(t)}\big) \;=\; D_{\mathrm{KL}}\big(P^{(t)}\big\|P^{(0)}\big).$$

Because $P^{(t+\frac{1}{2})} \in \mathcal{R}$, we also know

$$D_{\mathrm{KL}}\big(P\big\|P^{(t)}\big) \;=\; D_{\mathrm{KL}}\big(P\big\|P^{(t+\frac{1}{2})}\big) \;+\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(t)}\big) \quad \text{for any } P\in\mathcal{R}.$$

That second Pythagorean decomposition shows that once we land in $\mathcal{R}$, any other point in $\mathcal{R}$ must have strictly larger KL to $P^{(t)}$ than $P^{(t+\frac{1}{2})}$ does. In particular, if P^* is the true minimum of $D_{\mathrm{KL}}(\cdot\,\|P^{(0)})$ over $\mathcal{R}\cap\mathcal{C}$, then

$$D_{\mathrm{KL}}\big(P^*\big\|P^{(t)}\big) \;=\; D_{\mathrm{KL}}\big(P^*\big\|P^{(t+\frac{1}{2})}\big) \;+\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(t)}\big) \;\geq\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(t)}\big),$$

so

$$D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(0)}\big) \;=\; D_{\mathrm{KL}}\big(P^{(t)}\big\|P^{(0)}\big) \;-\; D_{\mathrm{KL}}\big(P^{(t+\frac{1}{2})}\big\|P^{(t)}\big) \;\geq\; D_{\mathrm{KL}}\big(P^{(t)}\big\|P^{(0)}\big) \;-\; D_{\mathrm{KL}}\big(P^*\big\|P^{(t)}\big).$$

In the limit, as we alternate between projecting onto $\mathcal{R}$ and onto $\mathcal{C}$, both sequences of "errors" $\big\{D_{\mathrm{KL}}(P^{(t+\frac{1}{2})}\|P^{(t)})\big\}$ and $\big\{D_{\mathrm{KL}}(P^{(t+1)}\|P^{(t+\frac{1}{2})})\big\}$ shrink to zero, which forces $P^{(t)} \to P^*$. Equivalently, one can see that any limit point must lie in $\mathcal{R}\cap\mathcal{C}$, and—because each projection never increases divergence—its divergence to $P^{(0)}$ must be the minimal possible.

3. Putting it all together: IPF is "valid" because it's just alternating I-projections

1. **Convexity.** The function $P \mapsto D_{\mathrm{KL}}(P\|P^{(0)})$ is strictly convex over the probability simplex. The constraint sets
$$\mathcal{R} = \{P : \sum_j P_{ij} = r_i\}, \quad \mathcal{C} = \{P : \sum_i P_{ij} = c_j\}$$
are convex (actually affine) subsets. Hence there is a unique global minimizer
$$P^* \;=\; \arg\min_{P \in \mathcal{R}\cap\mathcal{C}} D_{\mathrm{KL}}\big(P\|P^{(0)}\big).$$

2. **I-projections coincide with simple scalings.** Minimizing $D_{\mathrm{KL}}(P\|Q)$ over the set of distributions that match prescribed row-margins forces a closed-form solution: you simply re-scale each row of Q to hit the target sums. Likewise for columns. Therefore "project onto $\mathcal{R}$" is exactly "multiply each row of your current table by the ratio of (desired row-sum)/(current row-sum)." And "project onto $\mathcal{C}$" is "multiply each column by (desired col-sum)/(current col-sum)."

3. **Monotonic decrease of KL and convergence.** Each I-projection onto a convex set cannot increase KL-divergence to the original prior; instead, it strictly decreases it (unless the current table already satisfies that constraint exactly). From the Pythagorean property, the error (in KL-terms) from the true minimizer is strictly eaten away by each projection. As you alternate row and column projections, both the divergence to $P^{(0)}$ and the total violation of the constraints shrink to zero. In the limit, you must land at exactly the unique point in $\mathcal{R}\cap\mathcal{C}$ that has lowest KL to $P^{(0)}$.

Hence **IPF is "valid"** because it exactly implements a rigorously-justified alternating projection onto two convex sets in a strictly convex divergence. The final table is guaranteed to:

1. satisfy both row- and column-marginals, and
2. minimize $D_{\mathrm{KL}}\big(P\|P^{(0)}\big)$ over all tables with those margins.

Because "maximizing cross-entropy relative to $P^{(0)}$" is equivalent to "minimizing KL-divergence to $P^{(0)}$, " the limit of IPF is *by construction* the maximum-cross-entropy solution.

A more intuitive summary:

- **Each scaling step is the "best possible fix" for one family of constraints:** You start with a prior table $P^{(0)}$. Suppose right now your rows don't match $\{r_i\}$. If you insist that rows equal $\{r_i\}$, then the distribution that is *closest* to your current guess $P^{(t)}$ (in the information-theoretic sense) is just "take each row of $P^{(t)}$ and multiply it by a constant so its sum becomes r_i." That's exactly one I-projection; it never moves you farther from $P^{(0)}$ than you already were, and it enforces one half of the constraint.

- **Alternate with columns:** Once the rows are correct, your columns may now be wrong; project onto the set of models with correct columns. This again cannot increase (and in fact typically decreases) the KL-distance to $P^{(0)}$. **At convergence, you cannot decrease KL any further without breaking one of the margin constraints:** Thus the limit must lie in $\mathcal{R}\cap\mathcal{C}$, and it is the unique point there that has the smallest KL to $P^{(0)}$.

Because of these two facts—*(a)* each step is an exact projection that lowers KL, and *(b)* the intersection $\mathcal{R}\cap\mathcal{C}$ is convex and the KL-objective is strictly convex—the algorithm is guaranteed to converge to the exact global minimizer.

Hence:

- **Why is the *cross-entropy* maximum?** If you view "cross-entropy" $H(P, P^{(0)}) = -\sum_{i,j} P_{ij} \log P^{(0)}_{ij}$, then minimizing $D_{\mathrm{KL}}(P\|P^{(0)})$ is equivalent (up to an additive constant) to maximizing that cross-entropy, subject to the same margins. IPF finds the unique minimizer of $D_{\mathrm{KL}}(\cdot\|P^{(0)})$. Therefore it also finds the unique maximizer of cross-entropy.

- **Why is IPF *valid*?** Because each half-step is provably the exact I-projection onto one convex family of distributions, and the theory of alternating I-projections (Csiszár's theorem) guarantees that repeating these two projections converges to the single point in $\mathcal{R} \cap \mathcal{C}$ that has lowest KL to $P^{(0)}$. In plain terms: "you can't do better—each scaling is the optimal way to satisfy one set of margins without increasing your distance to the prior."

Once you accept that enforcing one family of marginals (rows or columns) by simple rescaling is exactly a KL-projection, it follows immediately that alternating those two projections must land you at the unique joint distribution which satisfies *both* margins *and* is closest (in KL-divergence) to your prior. Equivalently, that joint distribution is the one of **maximum cross-entropy**.