# Comparison and Connection between the joint and the conditional Generalized Iterative Scaling Algorithm

**Carlotta Langer**
Martin-Luther-University
Halle-Wittenberg
carlotta.langer@student.uni-halle.de

**Nihat Ay**
MPI for Mathematics in the Sciences
Santa Fe Institute
University of Leipzig
nay@mis.mpg.de

**Abstract**

Iterative Scaling is a widely used method to solve maximum entropy problems. Depending on the application they are used for, there are many different versions of Iterative Scaling algorithms. This paper compares and reconnects two popular algorithms, which share a name, but are not equal and even converge to different limit points.

**Keywords:** iterative scaling, maximum entropy estimation

## 1 Introduction

Beginning with the Iterative Scaling algorithm described by Csiszár in [6], there now exist many different types of Iterative Scaling algorithms depending on the respective application. There are two frequently used algorithms, which converge to different limit points but are both called Generalized Iterative Scaling (GIS) algorithm. These algorithms are the one presented by Darroch and Ratcliff in [8] and the algorithm used for example by Goodman in [10] or Huang et al. in [11]. To prevent confusion we will call the last one Conditional Generalized Iterative Scaling (CGIS) algorithm.

In order to emphasize their similarities and differences we present both algorithms and categorize them according to two different properties. At first we take a look at the instance they are operating on. In the case of the GIS algorithm it is the full probability distribution, in contrast to the CGIS algorithm that operates on parameters $\lambda$. We introduce an intermediate algorithm, the *Joint GIS* algorithm to analyse this difference.

Secondly, the form of the probability distribution is different in both algorithms. The GIS and *Joint GIS* algorithm use joint distributions while the CGIS algorithm computes conditional distributions. This transformation is the reason for the different limit points. GIS and *Joint GIS* converge to the maximum entropy estimation, in contrast to CGIS which converges to the conditional maximum entropy estimation. Our tool to study this difference is the *conditional GIS* algorithm, a new algorithm which works on the whole probability distribution but with the conditional maximum entropy principle. The categorization of these four algorithms is shown in Figure 1.

| iteration on \ distribution | joint | conditional |
|---|---|---|
| full distribution | GIS [8], [5] Section 3 | *conditional GIS* Section 5 |
| parameter | *Joint GIS* Section 4 | CGIS [10], [11] Section 6 |

Figure 1: Categorization of the four different algorithms, namely GIS, *conditional GIS*, *Joint GIS* and CGIS according to the used distribution and the component over which they iterate.

## 2 Iterative Scaling

This section gives a brief introduction to Iterative Scaling. The probability distributions discussed here are discrete distributions on a finite set $X$ and the set of these distributions will be denoted by $\mathcal{P}$.

Iterative Scaling algorithms are a method to solve maximum entropy problems. In this setting, the algorithms determine a probability distribution on $X$ with predetermined properties. The properties are fixed by introducing constraints which describe the expected value of a feature and are defined for $i \in \{1, \dots, m\}$ as:

$$\sum_{x \in X} P(x) f_i(x) = k_i, \quad k_i \geqslant 0, \ \sum_i k_i = 1. \tag{1}$$

The constraints are called consistent, if the set of positive probability distributions on $X$, which fulfil these, is not empty. An important result for Iterative Scaling is the duality in Lemma 1. We will need the following sets:

$$L(f,k) = \{P \in \mathcal{P} \mid \sum_{x \in X} P(x) f_i(x) = k_i, \ i \ \in \{1, \dots, m\}\}$$

$$Q(f, P^{(0)}) = \left\{ P \in \mathcal{P} \mid P(x) = \frac{1}{Z_{P^{(0)}}(\lambda \cdot f)} e^{\sum_i \lambda_i f_i(x)} P^{(0)}(x), \ \lambda_i \in \mathbb{R}, \ x \in X \right\}.$$

**Lemma 1.** Suppose that the distribution $\widehat{P}$ satisfies the constraints and that $D(\widehat{P} \parallel P^{(0)}) < \infty$ for a probability distribution $P^{(0)}$. Then any of the following properties determine $P^\star$ uniquely and the following statements are equivalent:

(1) $P^\star = \underset{\widehat{Q} \in \overline{Q}(f, P^{(0)})}{\arg\min} \; D(\widehat{P} \parallel \widehat{Q})$

(2) $P^\star = \underset{P \in L(f,k)}{\arg\min} D(P \parallel P^{(0)})$

(3) $P^\star \in L(f,k) \cap \overline{Q}(f, P^{(0)})$

*Proof.* The proof is given by Ay et al. in Section 2.8.3 Theorem 2.8 in [1]. □

# 3 The GIS algorithm

In order to prove the convergence of the algorithm Darroch and Ratcliff apply in [8] the following restrictions to the features:

$$\sum_x f_i(x) = 1 \quad \text{and} \quad f_i(x) \geqslant 0. \tag{2}$$

It is possible to define less strict restrictions as for example Curran and Clark point out in [7], but at this point it is sufficient to use the original ones.

**Theorem 3.1** (The GIS algorithm, [8])**.** Let $P^{(0)}$ be the uniform distribution, $f_i$ as in (2), $n \in \mathbb{N}$ and

$$P^{(n)}(x) = P^{(n-1)}(x) \prod_{i=1}^{m} \left( \frac{k_i}{\sum\limits_{x' \in X} P^{(n-1)}(x') f_i(x')} \right)^{f_i(x)}.$$

The $P^{(n)}$ converges to a positive and unique solution $P^\star \in \overline{Q}(f, P^{(0)})$ fulfilling the constraints (1) and the properties described in Lemma 1.

Additionally, it is possible to prove that $P^\star \in L(f,k) \cap Q(f, P^{(0)})$ under the condition that the constraints are consistent.

*Proof.* This was proven by Csiszár in [5]. □

# 4 The *Joint GIS* algorithm

The parameters $\lambda_i$ and the features $f_i$ determine the density of a Gibbs distribution uniquely. Therefore, we do not have to compute $P^{(n)}$ in each step of the algorithm in order to find $P^\star$. That is why it is sufficient to find an iteration for the parameters. In each step, the parameters $\lambda_i$ will be altered by $\delta_i$, so that $\lambda_i^{(n+1)} = \lambda_i^{(n)} + \delta_i$.

Although it is easy to check that the iteration below and the one described in Theorem 3.1 are related, we are not able to conclude immediately that this new algorithm converges. Darroch and Ratcliff state in [8] at the end of Section 2 without a proof that the parameters can be compiled in an easy way. In [2] Brown et al. are able to give this proof for some cases, but not for the general case.

In conclusion, we will prove the convergence of the algorithm based on the framework Pietra et al. use in [13] to prove the convergence of another form of iterative scaling algorithm, their Improved Iterative Scaling algorithm. Denote by

$$P_\delta(x) = \frac{1}{Z(\lambda+\delta, f)} \cdot e^{\sum_{i=1}^{m} (\lambda_i+\delta_i) f_i(x)}, \; Z(\lambda+\delta, f) = \sum_{x \in X} e^{\sum_{i=1}^{m} (\lambda_i+\delta_i) f_i(x)}. \quad (3)$$

The goal is to maximize $M(Q, P_\delta) := -D(Q \parallel P_\delta)$ with a fixed $Q \in \mathcal{P}$ that satisfies the constraints and to use Lemma 1. Therefore we need to find a lower bound of the steps of the iteration which is easy to maximize respecting $\delta$. The following Definition 1 and Theorem 4.1 are the ones Pietra et al. use in [13] in Section 4.B.

**Definition 1.** A function $B : \mathbb{R}^m \times \mathcal{P} \to \mathbb{R}$ is an auxiliary function for $M(Q, P_\delta)$ if it holds the following properties:

(1) For all $P \in \mathcal{P}$ and $\delta \in \mathbb{R}^m$ we have: $M(Q, P_\delta) \geqslant M(Q, P) + B(\delta, P)$.

(2) $B(\delta, P)$ is continuous in $P \in \mathcal{P}$ and $\mathrm{C}^1$ in $\delta \in \mathbb{R}^m$.

(3) Let $t \in \mathbb{R}$. Then $B(0, P) = 0$ and:

$$\left.\frac{d}{dt}\right|_{t=0} B(t \cdot \delta, P) = \left.\frac{d}{dt}\right|_{t=0} M(Q, P_{t \cdot \delta}).$$

It is possible to define the following sequence:

$$P^{(n+1)} = P^{(n)}_{\delta^{(n)}}, \text{ with } \delta^{(n)} = \arg\max_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}).$$

Property (1) of Definition 1 makes sure that $M(Q, P_{\delta^{(n)}})$ increases with every step. With this we can get to the next result:

**Theorem 4.1.** Let $P^{(n)} \in \mathcal{P}$ be a sequence where the support of $P^{(0)}$ is $X$ and the properties

$$P^{(n+1)} = P^{(n)}_{\delta^{(n)}}, \; \delta^{(n)} \in \mathbb{R}^{(m)}, \quad B(\delta^{(n)}, P^{(n)}) = \sup_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}).$$

Then $M(Q, P_{\delta^{(n)}})$ increases monotonically, it converges to

$$\max_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q}) \text{ and } \lim_{n \to \infty} P^{(n)} = P^\star = \arg\max_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q}).$$

*Proof.* This is proven in [13] Section 4.B. □

In order to resemble the CGIS algorithm, we will now make use of new restrictions towards the features:

$$f^c := \max_{x \in X} \sum_{i=1}^{m} f_i(x) \quad 0 \leqslant f_i(x) \leqslant 1, \quad i = 1, \ldots, m.$$

Additionally, we assume that $f^c \geqslant 1$. Note that we no longer need the $f_i$ to sum up to 1 or any fixed constant. In [7] Curran and Clark proved the convergence of the algorithm without a correction feature

$$f_{m+1} := 1 - \sum_{i=1}^{m} \frac{f_i(x)}{f^c}$$

by fixing its value with $\lambda_{m+1} \equiv 0$ to zero. In order to apply the Jensen's inequality we will use the same trick in the next lemma.

**Lemma 2.** Let $Q, P \in \mathcal{P}$ and $\delta \in \mathbb{R}^m$. Then the function

$$B(\delta, P) = 1 + \sum_{x \in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x \in X} P(x) \sum_{i=1}^{m+1} \frac{f_i(x)}{f^c} e^{\delta_i f^c}$$

is an auxiliary function for $M(Q, P)$ with $\delta_{m+1} = 0$ fixed.

*Proof of Lemma 2.* We will prove the properties listed in Definition 1. To prove the first property we use $log(x) \leqslant x - 1$, for all $x > 0$ and the Jensen's inequality.

$$\begin{aligned}(1)\ M(Q, P_\delta) - M(Q, P) &\geqslant \sum_{x \in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x \in X} Q(x) \left( \sum_{x' \in X} e^{\sum_{i=1}^{m} \delta_i f_i(x')} P(x') - 1 \right) \\ &\geqslant B(\delta, P).\end{aligned}$$

(2) The definition of $f^c$ assures that $f^c > 0$. As a sum of continuous functions $B(\delta, P)$ is continuous in $P$ and

$$\frac{d}{d\delta} B(\delta, P) = \left( \sum_{x \in X} Q(x) f_1(x) - P(x) f_1(x) e^{\delta_1 f^c} \ \ldots \ \sum_{x \in X} Q(x) f_m(x) - P(x) f_m(x) e^{\delta_m f^c} \right).$$

Every entry of the Jacobian matrix is continuous in $\delta_i$ and we gain property (2).

(3) For $\delta = 0^{(m)}$ we get: $B(0, P) = 1 - \sum_{x \in X} P(x) \sum_{i=1}^{m+1} \frac{f_i(x)}{f^c} = 1 - 1 = 0$. With $t \in \mathbb{R}$ the differentiation leads to:

$$\begin{aligned}\left. \frac{d}{dt} \right|_{t=0} M(Q, P_{t \cdot \delta}) &= \left. \frac{d}{dt} \right|_{t=0} \sum_{x \in X} ln \left( \frac{1}{Z_P(t \cdot \sigma, f)} e^{\sum_{i=1}^{m} t \cdot \delta_i f_i(x)} P(x) \right) - \sum_{x \in X} Q(x) ln(Q(x)) \\ &= \sum_{x \in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x \in X} P(x) \sum_{i=1}^{m} \delta_i f_i(x) = \left. \frac{d}{dt} \right|_{t=0} B(t \cdot \delta, P).\end{aligned}$$

□

It remains to show that $B(Q,P)$ and Theorem 4.1 result in the desired iteration:

**Lemma 3** (The *Joint GIS* algorithm)**.** Let $\lambda_i^{(0)} = 0$ and

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + \frac{1}{f^c} ln \left( \frac{k_i}{\sum\limits_{x \in X} P^{(n)}(x) f_i(x)} \right). \tag{4}$$

This converges to $\lambda_i^\star = \lim\limits_{n \to \infty} \lambda_i^{(n)}$. Additionally we have

$$P(x) = \frac{1}{Z(\lambda^\star, f)} e^{\sum\limits_{i=1}^{m} \lambda_i^\star f_i(x)} = P^\star(x)$$

for all $x$ and $P^\star$ is the same limit point as the one of GIS in Theorem 3.1.

*Proof of Lemma 3.* Theorem 4.1 provides us with an iteration that convergences to $\arg\max\limits_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q})$. Choosing $\lambda_i^{(0)}$ to be zero leads to $P^{(0)}$ as the uniform distribution. That means that the initial points of both algorithms are the same. Now we will take a look at $\delta^{(n)}$ defined in Theorem 4.1 and maximize $B(\delta, P^{(n)})$ in respect to $\delta$. For every $i \in \{1, \ldots, m\}$ we get:

$$\delta_i = \frac{1}{f^c} ln \left( \frac{k_i}{\sum\limits_{x \in X} P^{(n)}(x) f_i(x)} \right).$$

This is in fact a maximum of $B(\delta, P^{(n)})$ because of the negativity of the Hessian matrix. We are now able to iterate over the $\lambda_i$ separately, because of the following equality:

$$\sup_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}) = 1 + \sum_{i=1}^{m} \sup_{\delta_i \in \mathbb{R}} \left( \sum_{x \in X} Q(x) \delta_i f_i(x) - \sum_{x \in X} P(x) \frac{f_i(x)}{f^c} e^{\delta_i f^c} \right).$$

Together with (3) this leads to the iteration stated in this Lemma. This proves the convergence of the algorithm. Lemma 1 additionally yields the equality of the limit of this algorithm and the one of the algorithm described in Theorem 3.1. □

# 5 The *conditional GIS* algorithm

In this section we will perform the second step towards the implemented CGIS algorithm. This approach was also described in [14] in Section 4.5 and in [3] in Section 3. We will switch from joint distributions to conditional distributions. The GIS algorithm can be very expensive regarding the needed time for each step in the iteration. In each step the algorithm iterates over every $x \in X$. Consider an

experiment as an application for the algorithm with a huge space $X$ of possible outcomes. It is likely to assume that there are applications in which the space of actually occurring $x$ is a rather small subset $X' \subset X$. By substituting the joint probabilities with a special form of probability, we are able to iterate only over the $x$ that actually appear in the data. First, we will introduce this new form of probability distributions regarding a target distribution $P_t$ satisfying the constraints in general. This approach allows us to introduce the next step without assuming the existence of data.

In contrast to the parameter estimation this step actually changes the limit of the convergence. By using a different form of probability distributions the maximum entropy principle turns into the conditional maximum entropy principle.

To do so, we have to be able to write $X$ as $X = X_1 \times X_2$ by defining two disjoint subsets $A_1, A_2$ of $\{1, \dots, r\}$ with $A_1 \cup A_2 = \{1, \dots, r\}$ and the alphabets $\mathcal{A}_\beta,\ \beta \in \{1, \dots, r\}$ of $x_\beta \in \mathcal{A}_\beta$:

$$X_i = \{x_i = (x_\beta)_{\beta \in A_i} \mid x \in \bigtimes_{\beta \in A_i} \mathcal{A}_\beta\},\ i \in \{1, 2\}.$$

The probability of $P \in \mathcal{P}$ on $X_1$ is defined by:

$$P(x_{A_1}) = \sum_{x \in X(x_{A_i})} P(x),\ x_{A_1} \in X_1 \text{ with } X(x_{A_i}) = \{y \in \bigtimes_{\beta=1}^{r} \mathcal{A}_\beta \mid y_{A_i} = x_{A_i}\}.$$

With the additional restriction that the marginal possibility of the $x_1 \in X_1$ equals the empirical distribution $\widehat{P}(x_1),\ x_1 \in X_1$ derived from a fixed set of data, we are able to define an algorithm iterating only over the $x_1 \in X_1$ occurring in the tests. Suppose that $P_t \in \mathcal{P}$ satisfies the constraints (1). Now we are able to change $P(x_1)$ to $P_t(x_1)$ to create the new constraints:

$$\sum_{x_1 \in X_1} P_t(x_1) \sum_{x_2 \in X_2} P(x_2 \mid x_1) f_i(x_1, x_2) = k_i. \tag{5}$$

This leads to the definition of a new probability distribution on $X$:

$$P^c(x) = P_t(x_1) \cdot P(x_2 \mid x_1), \quad \text{for all } x_1 \in X_1,\ x_2 \in X_2.$$

Now we take a closer look at the new probability distribution $P^c$. While $P_t$ is fixed to the target distribution, $P(x_2 \mid x_1)$ is a conditional Gibbs-distribution:

$$P(x_2 \mid x_1) = \frac{1}{Z_{x_2}(x_1)} e^{\sum_{i=1}^{m} \lambda_i f_i(x_1, x_2)},\ Z_{x_2}(x_1) = \sum_{x_2 \in X_2} e^{\sum_{i=1}^{m} \lambda_i f_i(x_1, x_2)}. \tag{6}$$

It is possible to define similar sets to the ones in Section 2:

$$L^c(f, k) = \left\{ P \in \mathcal{P} \mid P(x_1) = P_t(x_1), \text{ for all } x_1 \in X_1 \text{ and } \sum_{x \in X} P(x) f_i(x) = k_i \right\}$$

$$Q^c(f, P^{(0)}) = \left\{ P \in \mathcal{P} \mid P(x) = P_t(x_1) \frac{1}{Z_{x_2}(x_1)} e^{\sum_{i=1}^{m} \lambda_i f_i(x)} P^{(0)}(x),\ \lambda_i \in \mathbb{R},\ x \in X \right\}.$$

Now we are able to define a conditional equivalent to Lemma 1:

**Lemma 4.** Suppose that the distribution $\widehat{P} \in L^c(f,k)$ satisfies the constraints and that $D(\widehat{P},U) < \infty$. If $P^\star \in L^c(f,k) \cap Q^c(f,U)$ exists, it is unique and holds the properties:

(1) $P^\star = \underset{\widehat{Q} \in Q^c(f,U)}{\arg\min}\ D(\widehat{P}(X_2 \mid X_1) \parallel \widehat{Q}(X_2 \mid X_1))$

(2) $P^\star = \underset{P \in L(f,k)}{\arg\min}\ D(P(X_2 \mid X_1) \parallel U(X_2 \mid X_1))$

*Proof.* With the chain rule (9) the proof can be easily derived from the one presented by Pietra et Al. in Proposition 4 in [13]. □

All things considered, we are able to maximize the conditional entropy by adjusting the parameters $\lambda_i$ of (6). Now we define the *conditional GIS* and it converges to the distribution of the form of $P^c$ that maximizes the conditional entropy.

**Lemma 5** (The conditional GIS algorithm)**.** Let $k_i$ be defined as in (5) and $P^{(0)}$ be the uniform distribution. If a probability distribution of the form (6) satisfying the constraints exists, then

$$P^{(n)}(x) = P^{(n-1)}(x) \prod_{i=1}^{m} \left( \frac{k_i}{\sum\limits_{x_1 \in X_1} P_t(x_1) \sum\limits_{x_2 \in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x_1, x_2)} \right)^{\frac{f_i(x)}{f^c}}$$

converges to $P^\star$ of the form (6), satisfying the constraints and maximizing the conditional entropy.

*Proof.* The following proof is derived from the one Darroch and Ratcliff presented in [8] of Theorem 1. At first we will use the inequality between the generalized arithmetic and geometric means:

$$\prod_{i=1}^{m} \left( \frac{k_i}{k_i^{(n-1)}} \right)^{\frac{f_i(x)}{f^c}} \leqslant \sum_{i=1}^{m} \frac{f_i(x)}{f^c} \cdot \left( \frac{k_i}{k_i^{(n-1)}} \right) \tag{7}$$

with $k_i^{(n-1)} = \sum\limits_{x_1 \in X_1} P_t(x_1) \sum\limits_{x_2 \in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x_1, x_2)$. Applying this leads to

$$\sum_{x \in X} P^{(n)}(x) \leqslant \sum_{x \in X} P^{(n-1)}(x) \sum_{i=1}^{m} \frac{f_i(x)}{f^c} \cdot \left( \frac{k_i}{\sum\limits_{x \in X} f_i(x) P^{(n-1)}(x)} \right) = \frac{1}{f^c} \sum_{i=1}^{m} k_i = \frac{1}{f^c}$$

and we have $\sum_{x \in X} P^{(n)}(x) \leqslant \frac{1}{f^c} \leqslant 1$, therefore $\sum_{i=1}^{m} \sum_{x \in X} P^{(n)}(x) f_i(x) \leqslant \frac{1}{f^c} \leqslant 1$. For all $n$, $P^{(n)}(x) > 0$ and $k_i^{(n)} = \sum_{x \in X} P^{(n)}(x) f_i(x) > 0$. The positivity of the KL-divergence leads to

$$D(k_i \parallel k_i^{(n)}) = \sum_{i=1}^{m} k_i log_2 \left( \frac{k_i}{k_i^{(n)}} \right) \geqslant 0. \tag{8}$$

Let $Q$ be an arbitrary probability distribution satisfying (1). Then

$$D(Q \parallel P^{(n+1)}) = \sum_{x \in X} Q(x) log_2 \left( \frac{Q(x)}{P^{(n+1)}(x)} \right) = D(Q \parallel P^{(n)}) - \frac{1}{f^c} D(k \parallel k^{(n)}).$$

Now $\{D(Q \parallel P^{(n+1)}), n \in \mathbb{N}\}$ is a decreasing bounded sequence. Therefore it has a limit point and $D(k \parallel k^{(n)}) \to 0$ as $n \to \infty$. The properties of $k_i$ and Pinsker's inequality

$$D(k \parallel k^{(n)}) \geqslant \frac{1}{2} \sum_{i=1}^{m} \mid k_i - k_i^{(n)} \mid^2$$

lead to $k_i^{(n)} \to k_i$ as $n \to \infty$. Suppose $P_1$, $P_2$ are different limit points of the bounded sequence $\{P^{(n)}\}$. Because of (5), both satisfy the constraints (1) and are of the form (6). Additionally $P_1$, $P_2$ are positive probability distributions. Now we are able to apply Lemma 4 and this yields that $P_1 = P_2$. □

# 6 The CGIS algorithm

Applying both changes, the parameter estimation and the conditional distribution, leads to the desired CGIS algorithm:

**Lemma 6** (The CGIS algorithm, [10], [11], [7])**.** Let $\lambda_i^{(0)} = 0$, $k_i$ be consistent and as in (5) and $P^{(n)}$ of the form (6) with parameters $\lambda^{(n)}$. The iteration

$$\lambda_i^{(n)} = \lambda_i^{(n-1)} + \frac{1}{f^c} ln \left( \frac{k_i}{\sum_{x_1 \in X_1} P_t(x_1) \sum_{x_2 \in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x)} \right)$$

converges to the limit $\lambda_i^\star$ with $P^\star(x_2 \mid x_1) = \frac{1}{Z_{x_2}} e^{\sum_{i=1}^{m} \lambda_i^\star f_i(x)}$. Additionally, $P^\star(x) = P_t(x_1) P^\star(x_2 \mid x_1)$ is conditional maximum entropy estimation.

*Proof.* Curran and Clark provide a proof in the Appendix of [7]. □

# 7 Comparison

In order to understand the relationship between GIS and CGIS, we will take a look at the following chain rule for entropy:

$$H_P(X_1, X_2) = H_P(X_1) + H_P(X_2 \mid X_1) \tag{9}$$

with the conditional entropy defined as

$$H_P(X_2 \mid X_1) = - \sum_{x_1 \in X_1} P(x_1) \sum_{x_2 \in X_2} P(x_2 \mid x_1) log_2(P(x_2 \mid x_1)).$$

A proof for this rule is given by [4] in Theorem 2.2.1. Notice that the entropy equals the conditional entropy in the case of a fixed marginal distribution on $X_1$. That means that the algorithms GIS and *Joint GIS* iterate towards the same limit as the conditional ones under the restriction that the marginal distribution on $X_1$ is fixed in both cases to the same values. This can be easily done by introducing an additional feature for the desired distribution. However, in the general case the limits are not equal, as we can observe in Figure 2 (a). This example for the different limit points of maximum entropy and conditional maximum entropy estimation was given by Yuret in [15]. We gain the values calculated by Yuret with an implementation of *Joint GIS* and CGIS in C++ available at [9].

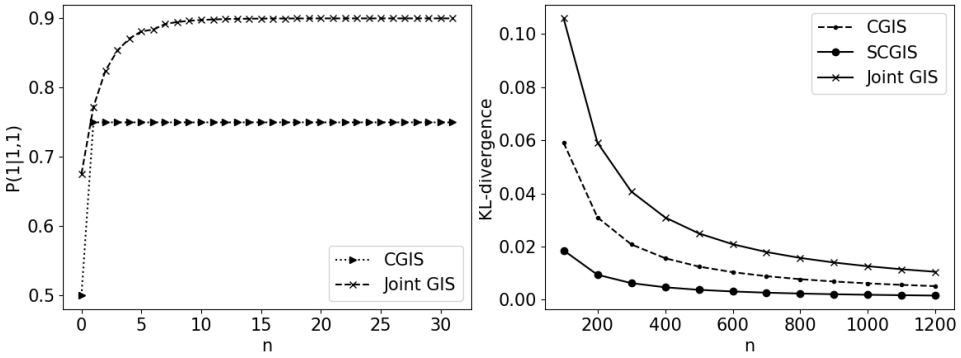


Figure 2: (a) CGIS vs. *Joint GIS* (b) AND: *Joint GIS*, CGIS and SCGIS

Additionally, we are able to compare the performances of the algorithms with an easy example. Consider $X$ as $X = X_1 \times X_2 \times X_3$, with $(X_1, X_2)$ as input and $X_3$ as output. Now, we are trying to predict the value of $X_3$ while only knowing the input. Our set of data is the logical AND gate listed in Figure 3(a).

At first we assume that $X_3$ depends on $X_1$ and $X_2$, but not their interactions as illustrated in Figure 3 (a) 2. Computing the data under this assumption leads to an probability distribution, which we compare to a second probability distribution by calculating the KL-divergence between them. The second probability distribution is gained by expecting that $X_3$ depends on $X_1$ and $X_2$ simultaneously and their interaction as visualized in Figure 3 (b) 2.

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

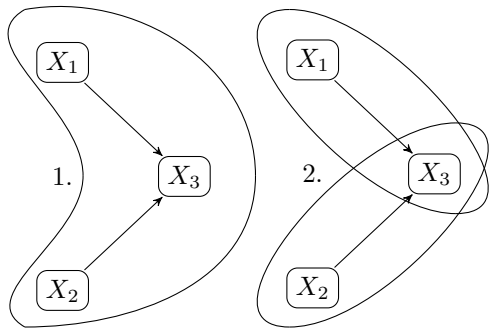


Figure 3: (a) AND gate (b) different systems of dependences

As indicated above, we used a third feature in case of the *Joint GIS* algorithm in order to gain the same limit point as the CGIS algorithm. Figure 2 (b) shows the results of this test. We observe that both algorithms now share the limit point 0 and that CGIS converges considerably faster than *Joint GIS*. The reason for this difference is in this case not the choice of the set $X$, but the additional feature we introduced for the *Joint GIS* algorithm.

A downside of iterative scaling algorithms is their poor performance compared to gradient methods shown for example by Huang et al. in [11] or by Minka in [12]. That is why we display a third algorithm in Figure 2 (b). This algorithm is a faster version of CGIS, the Sequential Conditional Generalized Iterative Scaling (SCGIS) algorithm presented by Goodman in [10]. Although SCGIS is considerably faster than the other algorithms presented here, it is still not as good as the gradient methods it was compared to by Huang et al. in [11]. This leads to the result that an iterative scaling method may not be the fastest way to calculate a maximum entropy model, but a reliable one.

In conclusion, we were able to fully explain the connection and highlight the differences between the considered Generalized Iterative Scaling algorithms.

# References


[1] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information Geometry*. Springer, 2017.

[2] J. B. Brown, P. J. Chase, and A. O. Pittenger. Order independence and factor convergence in iterative scaling. *Linear Algebra and its Applications*, 1993.

[3] S. Chen and R. Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, School of Computer Science, Carnegie Mellon University, 1999.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[5] I. Csiszár. A geometric interpretation of darroch and ratcliff's generalized iterative scaling. *The Annals of Statistics*.

[6] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 1975.

[7] J. R. Curran and S. Clark. Investigating gis and smoothing for maximum entropy taggers. *Proceedings of EACL'03*, 2003. URL: `http://www.aclweb.org/anthology/E/E03/E03-1071.pdf`.

[8] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 1972.

[9] K. Ghazi-Zahedi. Entropy++ github repository. 2017. URL: `http://github.com/kzahedi/entropy`.

[10] J. Goodman. Sequential conditional generalized iterative scaling. *In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2002.

[11] F.-L. Huang, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Iterative scaling and coordinate descent methods for maximum entropy models. *Journal of Machine Learning Research*, 2010.

[12] T. P. Minka. Algorithms for maximum-likelihood logistic regression. *Statistics Tech Report*, 2001.

[13] S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[14] R. Rosenfeld. *Adaptive Statistical Language Modeling: a Maximum Entropy Approach*. PhD thesis, Carnegie Mellon University, 1992.

[15] D. Yuret. Naive bayes is a joint maximum entropy model. 2010. URL: `http://www.denizyuret.com/2010/11/naive-bayes-is-joint-maximum-entropy.html`.