MAXIMUM ENTROPY AND ITERATIVE SCALING

S. Della Pietra, V. Della Pietra, and J. Lafferty

(Excerpts from a paper submitted for publication)

1.1 Two Optimization Problems. Suppose that we are given an initial model $q_0 \in \Delta$, a reference distribution $\tilde{p}$, and a set of features $f = (f_0, f_1, \ldots, f_n)$. In practice, it is often the case that $\tilde{p}$ is the empirical distribution of a set of training samples $x^{(1)}, x^{(2)} \ldots x^{(N)}$, and is thus given by

$$\tilde{p}(x) = \frac{c(x)}{N} \tag{1.1}$$

where $c(x) = \sum_{1 \le i \le N} \delta(x, x^{(i)})$ is the number of times that configuration x appears among the training samples.

We wish to construct a probability distribution $q_\star \in \Delta$ that accounts for these data, in the sense that it approximates $\tilde{p}$ but does not deviate too far from q_0. We measure distance between probability distributions p and q in Δ using the Kullback-Leibler divergence

$$D(\tilde{p} \,\|\, p) = \sum_{x \in \Omega} \tilde{p}(x) \, \log \frac{\tilde{p}(x)}{p(x)} \,. \tag{1.2}$$

Throughout this paper we use the notation

$$p[g] = \sum_{x \in \Omega} g(x)\, p(x)$$

for the expectation of a function $g : \Omega \to \mathbf{R}$ with respect to the probability distribution p. For a function $h : \Omega \to \mathbf{R}$ and a distribution q, we use both the notation $h \circ q$ and q_h to denote the generalized Gibbs distribution given by

$$q_h(x) = (h \circ q)(x) = \frac{1}{Z_q(h)} e^{\,h(x)}\, q(x) \,.$$

Note that $Z_q(h)$ is not the usual partition function. It is a normalization constant determined by the requirement that $(h \circ q)(x)$ sums to 1 over x, and can be written as an expectation:

$$Z_q(h) = q[e^{\,h}] \,.$$

There are two natural sets of probability distributions determined by the data $\tilde{p}$, q_0, and f. The first is the set $\mathcal{P}(f, \tilde{p})$ of all distributions that agree with $\tilde{p}$ as to the expected value of the feature function f:

$$\mathcal{P}(f, \tilde{p}) = \{ p \in \Delta : \; p[f] = \tilde{p}[f] \; \} \,.$$

The second is the set $\mathcal{Q}(f, q_0)$ of generalized Gibbs distributions based on q_0 with feature function f:

$$\mathcal{Q}(f, q_0) = \{(\lambda \cdot f) \circ q_0 : \ \lambda \in \mathbf{R}^n \ \}.$$

We let $\bar{\mathcal{Q}}(f, q_0)$ denote the closure of $\mathcal{Q}(f, q_0)$ in Δ (with respect to the topology it inherits as a subset of Euclidean space).

There are two natural criteria for choosing $q_\star$:

- *Maximum Likelihood Gibbs Distribution.* Choose $q_\star$ to be a distribution in $\bar{\mathcal{Q}}(f, q_0)$ with maximum likelihood with respect to $\tilde{p}$:

$$q_\star = \operatorname*{arg\,min}_{q \in \bar{\mathcal{Q}}(f, q_0)} D(\tilde{p} \,\|\, q)$$

- *Maximum Entropy Constrained Distribution.* Choose $q_\star$ to be a distribution in $\mathcal{P}(f, \tilde{p})$ that has maximum entropy relative to q_0:

$$q_\star = \operatorname*{arg\,min}_{p \in \mathcal{P}(f, \tilde{p})} D(p \,\|\, q_0)$$

Although these criteria are different, they determine the same distribution. In fact, the following is true, as we prove in Section 1.3.

Proposition. *Suppose that $D(\tilde{p} \,\|\, q_0) < \infty$. Then there exists a unique $q_\star \in \Delta$ satisfying*

(1) $q_\star \in \mathcal{P}(f, \tilde{p}) \cap \bar{\mathcal{Q}}(f, q_0)$

(2) $D(p \,\|\, q) = D(p \,\|\, q_\star) + D(q_\star \,\|\, q)$ *for any* $p \in \mathcal{P}(f, \tilde{p})$ *and* $q \in \bar{\mathcal{Q}}(f, q_0)$

(3) $q_\star = \operatorname*{arg\,min}_{q \in \bar{\mathcal{Q}}(f, q_0)} D(\tilde{p} \,\|\, q)$

(4) $q_\star = \operatorname*{arg\,min}_{p \in \mathcal{P}(f, \tilde{p})} D(p \,\|\, q_0)$.

Moreover, any of these four properties determines $q_\star$ uniquely.

When $\tilde{p}$ is the empirical distribution of a set of training examples $x^{(1)}, x^{(2)} \ldots x^{(N)}$, minimizing $D(\tilde{p} \,\|\, p)$ is equivalent to maximizing the probability that the field p assigns to the training data, given by

$$\prod_{1 \le i \le N} p(x^{(i)}) = \prod_{x \in \Omega} p(x)^{\,c(x)} \ \propto \ e^{-N D(\tilde{p} \,\|\, p)}. \tag{1.3}$$

With sufficiently many parameters it is a simple matter to construct a field for which $D(\tilde{p} \,\|\, p)$ is arbitrarily small. In fact, we can construct a field with $N+1$ features and small Kullback-Leibler divergence with respect to $\tilde{p}$ by taking

$$f_i(x) = \delta(x, x^{(i)}), \quad \lambda_i = \log c(x^{(i)}) \tag{1.4}$$

for $1 \leq i \leq N$ and

$$f_{N+1}(x) = \prod_{1 \leq i \leq N} (1 - f_i(x)), \quad \lambda_{N+1} \ll -1. \tag{1.5}$$

While such a model has small divergence with respect to the empirical distribution of the samples $x^{(i)}$, it does not generalize to other, previously unseen configurations. This is the classic problem of *over-training*.

1.2 Duality, Auxiliary Functions, and Iterative Scaling. In this section we present an algorithm for selecting the parameters associated with the features of a random field. The algorithm is closely related to the Generalized Iterative Scaling algorithm of Darroch and Ratcliff [8]. Like the Darroch and Ratcliff procedure, the algorithm requires that the features f_i are non-negative: $f_i(x) \geq 0$ for all $x \in \Omega$. Unlike the Darroch and Ratcliff procedure, however, our method does not require the features to be normalized to sum to a constant.

Throughout this section we hold the set of features $f = (f_0, f_1, \ldots, f_n)$, the initial model q_0 and the reference distribution $\tilde{p}$ fixed, and we simplify the notation accordingly. In particular, we write $\gamma \circ q$ instead of $(\gamma \cdot f) \circ q$ for $\gamma \in \mathbf{R}^n$. We assume that $\tilde{p}(x) = 0$ whenever $q_0(x) = 0$. This condition is commonly written $\tilde{p} \ll q_0$, and it is equivalent to $D(\tilde{p} \,\|\, q_0) < \infty$.

A description of the algorithm requires an additional piece of notation. Let

$$M(x) = \sum_{i=0}^{n} f_i(x). \tag{1.6}$$

If the features are binary, then $M(x)$ is the total number of features that are "on" for the configuration x.

Improved Iterative Scaling.

Initial Data:

A reference distribution $\tilde{p}$ and an initial model q_0, with $\tilde{p} \ll q_0$, and non-negative features $f_0, f_1, \ldots, f_n$.

Output:

The distribution $q_\star = \operatorname{arg\,min}_{q \in \bar{\mathcal{Q}}(f, q_0)} D(\tilde{p} \,\|\, q)$

Algorithm:

(0) *Set* $q^{(0)} = q_0$.

(1) *For each* i *let* $\gamma_i^{(k)} \in [-\infty, \infty)$ *be the unique solution of*

$$q^{(k)}[\, f_i \, e^{\,\gamma_i^{(k)}\, M}\,] = \tilde{p}[\, f_i\,]\,. \tag{1.7}$$

(2) *Set* $q^{(k+1)} = \gamma^{(k)} \circ q^{(k)}$ *and* $k \leftarrow k+1$.

(3) *If* $q^{(k)}$ *has converged, set* $q_\star = q^{(k)}$ *and terminate. Otherwise go to step (1).*

In other words, this algorithm constructs a distribution $q_\star = \lim_{n\to\infty} \gamma_n \circ q_0$ where $\gamma_n = \sum_{k=0}^{n} \gamma_i^{(k)}$ and $\gamma_i^{(k)}$ is determined as the solution to the equation

$$\sum_x q^{(k)}(x)\, f_i(x)\, e^{\,\gamma_i^{(k)}\, M(x)} = \sum_x \tilde{p}(x)\, f_i(x)\,. \tag{1.8}$$

When used in the n-th iteration of the field induction algorithm, where a candidate feature $g = f_n$ is added to the field $q = q_n$, we choose the initial distribution q_0 to be $q_0 = q_{\hat{\alpha} g}$, where $\hat{\alpha}$ is the parameter that maximizes the gain of g. In practice, this provides a good starting point from which to begin iterative scaling. In fact, we can view this distribution as the result of applying one iteration of an Iterative Proportional Fitting Procedure [1,4] to project $q_{\alpha g}$ onto the linear family of distributions with g-marginals constrained to $\tilde{p}[g]$.

Our main result in this section is

Proposition 1.1. *Suppose* $q^{(k)}$ *is the sequence in* Δ *determined by the Improved Iterative Scaling algorithm. Then* $D(\tilde{p} \,\|\, q^{(k)})$ *decreases monotonically to* $D(\tilde{p} \,\|\, q_\star)$ *and* $q^{(k)}$ *converges to* $q_\star = \arg\min_{q \in \bar{\mathcal{Q}}} D(\tilde{p} \,\|\, q) = \arg\min_{p \in \mathcal{P}} D(p \,\|\, q_0)$.

In the remainder of this section we present a self-contained proof of the convergence of the algorithm. The key idea of the proof is to express the incremental step of the algorithm in terms of an auxiliary function which bounds from below the likelihood objective function. This technique is the standard means of analyzing the EM algorithm [9], but it has not previously been applied to iterative scaling. Our analysis of iterative scaling is different and simpler than previous treatments. In particular, in contrast to Csiszár's proof of the Darroch-Ratcliff procedure [5], our proof does not rely upon the convergence of alternating I-projection [4].

We begin by proving the basic duality theorem which states that the maximum likelihood problem for a Gibbs distribution and the maximum entropy problem subject to linear constraints have the same solution. We then turn to the task of computing this

solution. After introducing auxiliary functions in a general setting, we apply this method to prove convergence of the Improved Iterative Scaling algorithm. We finish the section by discussing Monte Carlo methods for estimating the equations when the size of the configuration space prevents the explicit calculation of feature expectations.

1.3 Duality. In this section we prove

Proposition 1.2. *Suppose that $\tilde{p} \ll q_0$. Then there exists a unique $q_\star \in \Delta$ satisfying*

(1) $q_\star \in \mathcal{P} \cap \bar{\mathcal{Q}}$

(2) $D(p \,\|\, q) = D(p \,\|\, q_\star) + D(q_\star \,\|\, q)$ *for any* $p \in \mathcal{P}$ *and* $q \in \bar{\mathcal{Q}}$

(3) $q_\star = \arg\min_{q \in \bar{\mathcal{Q}}} D(\tilde{p} \,\|\, q)$

(4) $q_\star = \arg\min_{p \in \mathcal{P}} D(p \,\|\, q_0)$.

Moreover, any of these four properties determines $q_\star$ uniquely.

This result is well known, although perhaps not quite in this packaging. In the language of constrained optimization, it expresses the fact that the maximum likelihood problem for Gibbs distributions is the convex dual to the maximum entropy problem for linear constraints. We include a proof here to make this paper self-contained and also to carefully address the technical issues arising from the fact that $\mathcal{Q}$ is not closed. The proposition would not be true if we replaced $\bar{\mathcal{Q}}$ with $\mathcal{Q}$. In fact, $\mathcal{P} \cap \mathcal{Q}$ might be empty. Our proof is elementary and does not rely on the Kuhn-Tucker theorem or other machinery of constrained optimization.

Our proof of the proposition will use a few lemmas. The first two lemmas we state without proof.

Lemma 1.3.

(1) $D(p \,\|\, q)$ *is a non-negative, extended real-valued function on* $\Delta \times \Delta$.

(2) $D(p \,\|\, q) = 0$ *if and only if* $p = q$.

(3) $D(p \,\|\, q)$ *is strictly convex in* p *and* q *separately.*

(4) $D(p \,\|\, q)$ *is* C^1 *in* q.

Lemma 1.4.

(1) The map $(\gamma, p) \mapsto \gamma \circ p$ *is smooth in* $(\gamma, p) \in \mathbf{R}^n \times \Delta$.

(2) The derivative of $D(p \,\|\, \lambda \circ q)$ *with respect to* λ *is*

$$\frac{d}{dt}\Big|_{t=0} D(p \,\|\, (t\lambda) \circ q) = \lambda \cdot (p[f] - q[f]) \,.$$

Lemma 1.5. *If* $\tilde{p} \ll q_0$ *then* $\mathcal{P} \cap \bar{\mathcal{Q}}$ *is nonempty.*

Proof. Define $q_\star$ by property (3) of Proposition 1.2; that is, $q_\star = \arg\min_{q \in \bar{\mathcal{Q}}} D(\tilde{p} \,\|\, q)$. To see that this makes sense, note that since $\tilde{p} \ll q_0$, $D(\tilde{p}, q)$ is not identically ∞ on $\bar{\mathcal{Q}}$. Also, $D(p \,\|\, q)$ is continuous and strictly convex as a function of q. Thus, since $\bar{\mathcal{Q}}$ is closed, $D(\tilde{p} \,\|\, q)$ attains its minimum at a unique point $q_\star \in \bar{\mathcal{Q}}$. We will show that $q_\star$ is also in $\mathcal{P}$. Since $\bar{\mathcal{Q}}$ is closed under the action of $\mathbf{R}^n$, $\lambda \circ q_\star$ is in $\bar{\mathcal{Q}}$ for any λ. Thus by the definition of $q_\star$, $\lambda = 0$ is a minimum of the function $\lambda \to D(\tilde{p} \,\|\, \lambda \circ q_\star)$. Taking derivatives with respect to λ and using Lemma 1.4 we conclude $q_\star[f] = \tilde{p}[f]$. Thus $q_\star \in \mathcal{P}$. □

Lemma 1.6. *If $q_\star \in \mathcal{P} \cap \bar{\mathcal{Q}}$ then for any $p \in \mathcal{P}$ and $q \in \bar{\mathcal{Q}}$*

$$D(p \,\|\, q) = D(p \,\|\, q_\star) + D(q_\star \,\|\, q)\,.$$

This is called the *Pythagorean property* since it resembles the Pythagorean theorem if we imagine that $D(p \,\|\, q)$ is the square of Euclidean distance and $(p, q_\star, q)$ are the vertices of a right triangle.

Proof. A straightforward calculation shows that

$$D(p_1 \,\|\, q_1) - D(p_1 \,\|\, q_2) - D(p_2 \,\|\, q_1) + D(p_2 \,\|\, q_2) = \lambda \cdot (p_1[f] - p_2[f])$$

for any $p_1, p_2, q_1, q_2 \in \Delta$ with $q_2 = \lambda \circ q_1$. It follows from this identity and the continuity of D that

$$D(p_1 \,\|\, q_1) - D(p_1 \,\|\, q_2) - D(p_2 \,\|\, q_1) + D(p_2 \,\|\, q_2) = 0$$

if $p_1, p_2 \in \mathcal{P}$ and $q_1, q_2 \in \bar{\mathcal{Q}}$. The lemma follows by taking $p_1 = q_1 = q_\star$. □

Proof of Proposition 1.2. Choose $q_\star$ to be any point in $\mathcal{P} \cap \bar{\mathcal{Q}}$. Such a $q_\star$ exists by Lemma 1.5. It satisfies property (1) by definition, and it satisfies property (2) by Lemma 1.6. As a consequence of property (2), it also satisfies properties (3) and (4). To check property (3), for instance, note that if q is any point in $\bar{\mathcal{Q}}$, then $D(\tilde{p} \,\|\, q) = D(\tilde{p} \,\|\, q_\star) + D(q_\star \,\|\, q) \geq D(\tilde{p} \,\|\, q_\star)$.

It remains to prove that each of the four properties (1)–(4) determines $q_\star$ uniquely. In other words, we need to show that if m is any point in Δ satisfying any of the four properties (1)–(4), then $m = q_\star$. Suppose that m satisfies property (1). Then by property (2) for $q_\star$ with $p = q = m$, $D(m \,\|\, m) = D(m \,\|\, q_\star) + D(q_\star \,\|\, m)$. Since $D(m \,\|\, m) = 0$, it follows that $D(m, q_\star) = 0$ so $m = q_\star$. If m satisfies property (2), then the same argument with $q_\star$ and m reversed again proves that $m = q_\star$. Suppose that m satisfies property (3). Then

$$D(\tilde{p} \,\|\, q_\star) \geq D(\tilde{p} \,\|\, m) = D(\tilde{p} \,\|\, q_\star) + D(q_\star \,\|\, m)$$

where the second equality follows from property (2) for $q_\star$. Thus $D(q_\star \,\|\, m) \le 0$ so $m = q_\star$. If m satisfies property (4), then a similar proof shows that once again $m = q_\star$. □

1.4 Auxiliary functions. In the previous section we proved the existence of a unique probability distribution $q_\star$ that is both a maximum likelihood Gibbs distributions and a maximum entropy constrained distribution. We now turn to the task of computing $q_\star$.

Fix $\tilde{p}$ and let $L : \Delta \to \mathbf{R}$ be the log-likelihood objective function

$$L(q) = -D(\tilde{p} \,\|\, q)\,.$$

Definition 1.7. *A function $A : \mathbf{R}^n \times \Delta \to R$ is an auxiliary function for L if*

(1) For all $q \in \Delta$ and $\gamma \in \mathbf{R}^n$

$$L(\gamma \circ q) \ge L(q) + A(\gamma, q)$$

(2) $A(\gamma, q)$ is continuous in $q \in \Delta$ and C^1 in $\gamma \in \mathbf{R}^n$ with

$$A(0, q) = 0 \quad \textit{and} \quad \frac{d}{dt}\Big|_{t=0} A(t\gamma, q) = \frac{d}{dt}\Big|_{t=0} L((t\gamma) \circ q)\,.$$

We can use an auxiliary function A to construct an iterative algorithm for maximizing L. We start with $q^{(k)} = q_0$ and recursively define $q^{(k+1)}$ by

$$q^{(k+1)} = \gamma^{(k)} \circ q^{(k)} \quad \text{with} \quad \gamma^{(k)} = \arg\max_\gamma A(\gamma, q^{(k)})\,.$$

It is clear from property (1) of the definition that each step of this procedure increases L. The following proposition implies that in fact the sequence $q^{(k)}$ will reach the maximum of L.

Proposition 1.8. *Suppose $q^{(k)}$ is any sequence in Δ with*

$$q^{(0)} = q_0 \quad \textit{and} \quad q^{(k+1)} = \gamma^{(k)} \circ q^{(k)}$$

where $\gamma^{(k)} \in \mathbf{R}^n$ satisfies

$$A(\gamma^{(k)}, q^{(k)}) = \sup_\gamma A(\gamma, q^{(k)})\,. \tag{1.9}$$

Then $L(q^{(k)})$ increases monotonically to $\max_{q \in \bar{\mathcal{Q}}} L(q)$ and $q^{(k)}$ converges to $q_\star = \arg\max_{q \in \bar{\mathcal{Q}}} L(q)$.

Equation (1.9) assumes that the supremum $\sup_\gamma A(\gamma, q^{(k)})$ is achieved at finite γ. In the next section, under slightly stronger assumptions, we present a extension of Proposition 4.8 that allows some components of $\gamma^{(k)}$ to take the value $-\infty$.

To use the proposition to construct a practical algorithm we must determine an auxiliary function $A(\gamma, q)$ for which $\gamma^{(k)}$ satisfying the required condition can be determined efficiently. In Section 1.3 we present a choice of auxiliary function which yields the Improved Iterative Scaling updates.

To prove Proposition 1.8 we first prove three lemmas.

Lemma 1.9. *If m is a cluster point of $q^{(k)}$, then $A(\gamma, m) \leq 0$ for all $\gamma \in \mathbf{R}^n$.*

Proof. Let $q^{(k_l)}$ be a sub-sequence converging to m. Then for any γ

$$A(\gamma, q^{(k_l)}) \leq A(\gamma^{(k_l)}, q^{(k_l)}) \leq L(q^{(k_l+1)}) - L(q^{(k_l)}) \leq L(q^{(k_l+1)}) - L(q^{(k_l)}) \,.$$

The first inequality follows from property (1.9) of $\gamma^{(n_k)}$. The second and third inequalities are a consequence of the monotonicity of $L(q^{(k)})$. The lemma follows by taking limits and using the fact that L and A are continuous. □

Lemma 1.10. *If m is a cluster point of $q^{(k)}$, then $\frac{d}{dt}|_{t=0}\, L((t\gamma) \circ m) = 0$ for any $\gamma \in \mathbf{R}^n$.*

Proof. By the previous lemma, $A(\gamma, m) \leq 0$ for all γ. Since $A(0, m) = 0$, this means that $\gamma = 0$ is a maximum of $A(\gamma, m)$ so that

$$0 = \frac{d}{dt}|_{t=0}\, A(t\gamma, m) = \frac{d}{dt}|_{t=0}\, L((t\gamma) \circ m) \,.$$

□

Lemma 1.11. *Suppose $\{q^{(k)}\}$ is any sequence with only one cluster point q_*. Then $q^{(k)}$ converges to q_*.*

Proof. Suppose not. Then there exists an open set B containing q_* and a subsequence $q^{(n_k)} \notin B$. Since Δ is compact, $q^{(n_k)}$ has a cluster point $q'_* \notin B$. This contradicts the assumption that $\{q^{(k)}\}$ has a unique cluster point. □

Proof of Proposition 1.8. Suppose that m is a cluster point of $q^{(k)}$. It follows from Lemma 1.10 that $\frac{d}{dt}|_{t=0}\, L((t\gamma) \circ m) = 0$, and so $m \in \mathcal{P} \cap \bar{\mathcal{Q}}$ by Lemma 1.4. But $q_\star$ is the only point in $\mathcal{P} \cap \bar{\mathcal{Q}}$ by Proposition 1.2. It follows from Lemma 1.11 that $q^{(k)}$ converges to $q_\star$. □

1.5 Dealing with ∞. In order to prove the convergence of the Improved Iterative Scaling algorithm, we need an extension of Proposition 1.8 that allows the components of γ to

equal $-\infty$. For this extension, we assume that all the components of the feature function f are non-negative:

$$f_i(x) \geq 0 \quad \text{for all } i \text{ and all } x. \tag{1.10}$$

Let $R \cup -\infty$ denote the partially extended real numbers with the usual topology. The operations of addition and exponentiation extend continuously to $R \cup -\infty$. Let $\mathcal{S}$ be the open subset of $(R \cup -\infty)^n \times \Delta$ defined by

$$\mathcal{S} = \{\, (\gamma, q) \in (R \cup -\infty)^n \times \Delta : q(x)e^{\gamma \cdot f(x)} > 0 \text{ for some } x \,\}$$

Observe that $R^n \times \Delta$ is a dense subset of $\mathcal{S}$. The map $(\gamma, q) \mapsto \gamma \circ p$, which up to this point we defined only for finite γ, extends uniquely to a continuous map from all of $\mathcal{S}$ to Δ. (The condition on $(\gamma, q) \in \mathcal{S}$ ensures that the normalization in the definition of $\gamma \circ p$ is well defined, even if γ is not finite.)

Definition 1.12. *We call a function $A : \mathcal{S} \to R \cup -\infty$ an extended auxiliary function for L if when restricted to $R^n \times \Delta$ it is an ordinary auxiliary function in the sense of Definition 1.7, and if, in addition, it satisfies property (1) of Definition 1.7 for any $(q, \gamma) \in \mathcal{S}$, even if γ is not finite.*

Note that if an ordinary auxiliary function extends to a continuous function on $\mathcal{S}$, then the extension is an extended auxiliary function.

We have the following extension of Proposition 1.8:

Proposition 1.8′. *Suppose the feature function f satisfies the non-negativity condition (1.10) and suppose A is an extended auxiliary function for L. Then the conclusion of Proposition 1.8 continues to hold if the condition on $\gamma^{(k)}$ is replaced by:*

$$(\gamma^{(k)}, q^{(k)}) \in \mathcal{S} \quad \textit{and} \quad A(\gamma^{(k)}, q^{(k)}) \geq A(\gamma, q^{(k)}) \quad \textit{for any} \quad (\gamma, q^{(k)}) \in \mathcal{S}\,.$$

Proof. Lemma 1.9 is valid under the altered condition on $\gamma^{(k)}$ since $A(\gamma, q)$ satisfies property (1) of Definition 1.7 for all $(\gamma, q) \in \mathcal{S}$. As a consequence, Lemma 1.10 also is valid, and the proof of Proposition 1.8 goes through without change. □

1.6 Improved Iterative Scaling. We now prove the monotonicity and convergence of the Improved Iterative Scaling algorithm by applying Proposition 1.8 to a particular choice of auxiliary function. We continue to assume, as in the previous section, that each component of the feature function f is non-negative.

For $q \in \Delta$ and $\gamma \in \mathbf{R}^n$, define

$$A(\gamma, q) = 1 + \gamma \cdot \tilde{p}[f] - \sum_x q(x) \sum_i f(i \,|\, x)\, e^{\,\gamma_i\, M(x)}$$

where $f(i \,|\, x) = \frac{f_i(x)}{M(x)}$. It is easy to check that A extends to a continuous function on $(R \cup -\infty)^n \times \Delta$.

Lemma 1.13. *$A(\gamma, q)$ is an extended auxiliary function for $L(q)$.*

The key ingredient in the proof of the lemma is the $\cap$-convexity of the logarithm and the $\cup$-convexity of the exponential, as expressed in the inequalities

$$e^{\sum_i t_i \alpha_i} \leq \sum_i t_i\, e^{\alpha_i} \qquad \text{if } t_i \geq 0 \text{ with } \sum_i t_i = 1 \tag{1.11}$$

$$\log x \leq x - 1 \qquad \text{for all } x > 0\,. \tag{1.12}$$

Proof of Lemma 1.13. Because A extends to a continuous function on $(R \cup -\infty)^n \times \Delta$, it suffices to prove that it satisfies properties (1) and (2) of Definition 1.7. To prove property (1) note that

$$\begin{aligned} L(\gamma \circ q) - L(q) &= \gamma \cdot \tilde{p}[f] - \log \sum_x q(x)\, e^{\gamma \cdot f(x)} && (1.13)\\ &\geq \gamma \cdot \tilde{p}[f] + 1 - \sum_x q(x)\, e^{\gamma \cdot f(x)} && (1.14)\\ &\geq \gamma \cdot \tilde{p}[f] + 1 - \sum_x q(x) \sum_i f(i \,|\, x)\, e^{\gamma_i\, M(x)} && (1.15)\\ &= A(\gamma, q)\,. \end{aligned}$$

Equality (1.13) is a simple calculation. Inequality (1.14) follows from inequality (1.12). Inequality (1.15) follows from the definition of M and Jensen's inequality (1.11). Property (2) of Definition 1.7 is straightforward to verify. □

Proposition 1.1 follows immediately from the above lemma and the extended Proposition 1.8. Indeed, it is easy to check that $\gamma^{(k)}$ defined in Proposition 1.1 achieves the maximum of $A(\gamma, q^{(k)})$, so that it satisfies the condition of Proposition 1.8′.

1.7 Monte Carlo methods. The Improved Iterative Scaling algorithm described above is well-suited to numerical techniques since all of the features take non-negative values. In

each iteration of this algorithm it is necessary to solve a polynomial equation for each feature f_i. That is, we can express equation (1.7) in the form

$$\sum_{m=0}^{M} a_{m,i}^{(k)} \, \beta_i^m = 0$$

where M is the largest value of $M(x) = \sum_i f_i(x)$ and

$$a_{m,i}^{(k)} = \begin{cases} \sum_x q^{(k)}(x) \, f_i(x) \, \delta(m, M(x)) & m > 0 \\ -\tilde{p}[\, f_i \,] & m = 0 \end{cases} \tag{1.16}$$

where $q^{(k)}$ is the field for the k-th iteration and $\beta_i = e^{\gamma_i^{(k)}}$. This equation has no solution precisely when $a_{m,i}^{(k)} = 0$ for $m > 0$. Otherwise, it can be efficiently solved using Newton's method since all of the coefficients $a_{m,i}^{(k)}$, $m > 0$, are non-negative. When Monte Carlo methods are to be used because the configuration space Ω is large, the coefficients $a_{m,i}^{(k)}$ can be simultaneously estimated for all i and m by generating a single set of samples from the distribution $q^{(k)}$.

References

[1] D. Brown, "A note on approximations to discrete probability distributions," Information and Control **2**, 386–392 (1959).

[2] L. Brown, *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics Lecture Notes–Monograph Series, Volume 9, Hayward, California, 1986.

[3] W. Byrne. "Alternating minimization and Boltzmann machine learning," IEEE Trans. on Neural Networks, **4** No. 4, 612–620, July 1992.

[4] I. Csiszár, "I-Divergence geometry of probability distributions and minimization problems," The Annals of Probability, **3**, No. 1, 146–158, 1975.

[5] I. Csiszár, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," The Annals of Statistics, **17**, No. 3, 1409–1413, 1989.

[6] _____, "An extended maximum entropy principle and a Bayesian justification," in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M Smith (Eds.), Elsevier Science Publishers, 1985, 83–98.

[7] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," Statistics & Decisions, Supplement Issue, **1**, 205–237, 1984.

[8] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," Ann. Math. Statist. **43**, 1470–1480, 1972.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society **39**, no. B, 1–38, 1977.

[10] E. T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, R. Rosenkrantz, ed., D. Reidel Publishing Co., Dordrecht–Holland, 1983.