Entire Relaxation Path for Maximum Entropy Problems

Moshe Dubiner
Google
moshe@google.com

Yoram Singer
Google
singer@google.com

Abstract

We discuss and analyze the problem of finding a distribution that minimizes the relative entropy to a prior distribution while satisfying max-norm constraints with respect to an observed distribution. This setting generalizes the classical maximum entropy problems as it relaxes the standard constraints on the observed values. We tackle the problem by introducing a re-parametrization in which the unknown distribution is distilled to a single scalar. We then describe a homotopy between the relaxation parameter and the distribution characterizing parameter. The homotopy also reveals an aesthetic symmetry between the prior distribution and the observed distribution. We then use the reformulated problem to describe a space and time efficient algorithm for tracking the *entire* relaxation path. Our derivations are based on a compact geometric view of the relaxation path as a piecewise linear function in a *two* dimensional space of the relaxation-characterization parameters. We demonstrate the usability of our approach by applying the problem to Zipfian distributions over a large alphabet.

1 Introduction

Maximum entropy (max-ent) models and its dual counterpart, logistic regression, is a popular and effective tool in numerous natural language processing tasks. The principle of maximum entropy was spelled out explicitly by E.T. Jaynes (1968). Applications of maximum entropy approach to natural language processing are numerous. A notable example and probably one of the earliest usages and generalizations of the maximum entropy principle to language processing is the work of Berger, Della Pietra$\times 2$, and Lafferty (Berger et al., 1996, Della Pietra et al., 1997). The original formulation of max-ent cast the problem as the task of finding the distribution attaining the highest entropy subject to *equality* constraints. While this formalism is aesthetic and paves the way to a simple dual in the form of a unique Gibbs distribution (Della Pietra et al., 1997), it does not provide sufficient tools to deal with input noise and sparse representation of the target Gibbs distribution. To mitigate these issues, numerous relaxation schemes of the equality constraints have been proposed. A notable recent work by Dudik, Phillips, and Schapire (2007) provided a general constraint-relaxation framework. See also the references therein for an in depth overview of other approaches and generalizations of max-ent. The constraint relaxation surfaces a natural parameter, namely, a relaxation value. The dual form of this free parameter is the regularization value of penalized logistic regression problems. Typically this parameter is set by experimentation using cross validation technique. The relaxed maximum-entropy problem setting is the starting point of this paper.

In this paper we describe and analyze a framework for *efficiently* tracking the entire relaxation path of constrained max-ent problems. We start in Sec. 2 with a generalization in which we discuss the problem of finding a distribution that minimizes the relative entropy to a given prior distribution while satisfying max-norm constraints with respect to an observed distribution. In Sec. 3 we tackle the problem by introducing a re-parametrization in which the

unknown distribution is distilled to a single scalar. We next describe in Sec. 4 a homotopy between the relaxation parameter and the distribution characterizing parameter. This formulation also reveals an aesthetic symmetry between the prior distribution and the observed distribution. We use the reformulated problem to describe in Secs. 5-6 space and time efficient algorithms for tracking the *entire* relaxation path. Our derivations are based on a compact geometric view of the relaxation path as a piecewise linear function in a *two* dimensional space of the relaxation-characterization parameters. In contrast to common homotopy methods for the Lasso Osborne et al. (2000), our procedure for tracking the max-ent homotopy results in an uncharacteristically low complexity bounds thus renders the approach applicable for large alphabets. We provide preliminary experimental results with Zipf distributions in Sec. 8 that demonstrate the merits of our approach. Finally, we conclude in Sec. 9 with a brief discussion of future directions.

2 Notations and Problem Setting

We denote vectors with bold face letters, e.g. $\boldsymbol{v}$. Sums are denoted by calligraphic letters, e.g. $\mathcal{M} = \sum_j m_j$. We use the shorthand $[n]$ to denote the set of integers $\{1, \ldots, n\}$. The n'th dimensional simplex, denoted Δ, consists of all vectors $\boldsymbol{p}$ such that, $\sum_{j=1}^n p_j = 1$ and for all $j \in [n]$, $p_j \geq 0$. We generalize this notion to multiplicity weighted vectors. Formally, we say that a vector $\boldsymbol{p}$ with multiplicity $\boldsymbol{m}$ is in the simplex, $(\boldsymbol{p}, \boldsymbol{m}) \in \Delta$, if $\sum_{j=1}^n m_j p_j = 1$, and for all $j \in [n]$, $p_j \geq 0$, and $m_j \geq 0$.

The generalized relaxed maximum-entropy problem is concerned with obtaining an estimate $\boldsymbol{p}$, given a prior distribution $\boldsymbol{u}$ and an observed distribution $\boldsymbol{q}$ such that the relative entropy between $\boldsymbol{p}$ and $\boldsymbol{u}$ is as small as possible while $\boldsymbol{p}$ and $\boldsymbol{q}$ are within a given max-norm tolerance. Formally, we cast the following constrained optimization problem,

$$\min_{\boldsymbol{p}} \sum_{j=1}^n m_j p_j \log\left(\frac{p_j}{u_j}\right) \quad , \tag{1}$$

such that $(\boldsymbol{p}, \boldsymbol{m}) \in \Delta$; $\|\boldsymbol{p} - \boldsymbol{q}\|_\infty \leq 1/\nu$. The vectors $\boldsymbol{u}$ and $\boldsymbol{q}$ are dimensionally compatible with $\boldsymbol{p}$, namely, $(\boldsymbol{q}, \boldsymbol{m}) \in \Delta$ and $(\boldsymbol{u}, \boldsymbol{m}) \in \Delta$. The scalar ν is a relaxation parameter. We use $1/\nu$ rather than ν itself for reasons that become clear in the sequel.

We next describe the dual form of (1). We derive the dual by introducing Lagrange-Legendre multipliers for each of the constraints appearing in (1). Let $\alpha_j^+ \geq 0$ denote the multiplier for the constraint $q_j - p_j \leq 1/\nu$ and $\alpha_j^- \geq 0$ the multiplier for the constraint $q_j - p_j \geq -1/\nu$. In addition, we use γ as the multiplier for the constraint $\sum_j m_j p_j = 1$. fter some routine algebraic manipulations we get that the Lagrangian is,

$$\begin{aligned} &\textstyle\sum_{j=1}^n m_i \left(p_j \log\left(\frac{p_j}{u_j}\right) + \alpha_j (q_j - p_j) + \frac{|\alpha_j|}{\nu} \right) \\ &\qquad + \gamma \left(\textstyle\sum_{j=1}^n m_j p_j - 1 \right) \quad . \end{aligned} \tag{2}$$

To find the dual form we take the partial derivative of the Lagrangian with respect to each p_j, equate to zero, and get that $\log\left(\frac{p_j}{u_j}\right) + 1 - \alpha_j + \gamma = 0$, which implies that $p_j \sim u_j e^{\alpha_j}$. We now employ the fact that $(\boldsymbol{p}, \boldsymbol{m}) \in \Delta$ to get that the exact form for p_j is

$$p_j = \frac{u_j \mathrm{e}^{\alpha_j}}{\sum_{i=1}^n m_i u_i \mathrm{e}^{\alpha_i}} \quad . \tag{3}$$

Using (3) in the compact form of the Lagrangian we obtain the following dual problem

$$\max_{\boldsymbol{\alpha}} - \left\{ \log(Z) - \sum_{j=1}^n m_j q_j \alpha_j + \sum_{j=1}^n \frac{m_j}{\nu} |\alpha_j| \right\} \quad , \tag{4}$$

where $Z = \sum_{j=1}^n m_j u_j \mathrm{e}^{\alpha_j}$. We make rather little use of the dual form of the problem. However, the complementary slackness conditions that are necessary for optimality to hold play an important role in the next section in which we present a reformulation of the relaxed maximum entropy problem.

3 Problem Reformulation

First note that the primal problem is a strictly convex function over a compact convex domain. Thus, its optimum exists and is unique. Let us now characterize the form of the solution. We partition the set of indices in $[n]$ into three disjoint sets depending on whether the constraint $|p_j - q_j| \leq 1/\nu$ is active and its form. Concretely, we define

$$\begin{aligned} I_- &= \{1 \leq j \leq n \mid p_j = q_j - 1/\nu\} \\ I_0 &= \{1 \leq j \leq n \mid |p_j - q_j| < 1/\nu\} \\ I_+ &= \{1 \leq j \leq n \mid p_j = q_j + 1/\nu\} \quad . \end{aligned} \tag{5}$$

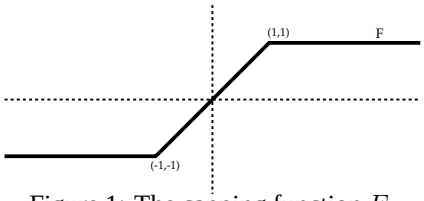

Figure 1: The capping function F.

Recall that $Z = \sum_{j=1}^{n} m_j u_j e^{\alpha_j}$. Thus, from (3) we can rewrite $p_j = u_j e^{\alpha_j}/Z$. We next use the complementary slackness conditions (see for instance (Boyd and Vandenberghe, 2004)) to further characterize the solution. For any $j \in I_-$ we must have $\alpha_j^- = 0$ and $\alpha_j^+ \geq 0$ therefore $\alpha_j \geq 0$, which immediately implies that $p_j \geq u_j/Z$. By definition we have that $p_j = q_j - 1/\nu$ for $j \in I_-$. Combining these two facts we get that $u_j/Z \leq q_j - 1/\nu$ for $j \in I_-$. Analogous derivation yields that $u_j/Z \geq q_j + 1/\nu$ for $j \in I_+$. Last, if the set I_0 is not empty then for each j in I_0 we must have $\alpha_j^+ = 0$ and $\alpha_j^- = 0$ thus $\alpha_j = 0$. Resorting again to the definition of $\boldsymbol{p}$ from (3) we get that $p_j = u_j/Z$ for $j \in I_0$. Since $|p_j - q_j| < 1/\nu$ for $j \in I_0$ we get that $|u_j/Z - q_j| < 1/\nu$. To recap, there exists $Z > 0$ such that the optimal solution takes the following form,

$$p_j = \begin{cases} q_j - 1/\nu & u_j/Z \leq q_j - 1/\nu \\ u_j/Z & |u_j/Z - q_j| < 1/\nu \\ q_j + 1/\nu & u_j/Z \geq q_j + 1/\nu \end{cases} \quad . \tag{6}$$

We next introduce an key re-parametrization, defining $\mu = \nu/Z$. We also denote by $F(\cdot)$ the capping function $F(x) = \max\{-1, \min\{1, x\}\}$. A simple illustration of the capping function is given in Fig. 1. Equipped with these definition we can rewrite (6) as follows,

$$p_j = q_j + \frac{1}{\nu} F(\mu u_j - \nu q_j) \ . \tag{7}$$

Given $\boldsymbol{u}$, $\boldsymbol{q}$, and ν, the value of μ can be found by using $\sum_j m_j p_j = \sum_j m_j q_j = 1$, which implies

$$G(\nu, \mu) \stackrel{\text{def}}{=} \sum_{j=1}^{n} m_j F(\mu u_j - \nu q_j) = 0 \ . \tag{8}$$

We defer the derivation of the actual algorithm for computing μ (and in turn $\boldsymbol{p}$) to the next section. In the meanwhile let us continue to explore the rich structure of the general solution. Note that $\mu, \boldsymbol{u}$ are interchangeable with $\nu, \boldsymbol{q}$. We can thus swap the roles of the prior distribution with the observed distribution and obtain an analogous characterization. In the next section we further explore the dependence of μ on ν. The structure we reveal shortly serves as our infrastructure for deriving efficient algorithms for following the regularization path.

4 The function $\mu(\nu)$

In order to explore the dependency of μ on ν let us introduce the following sums

$$\begin{aligned} \mathcal{M} &= \sum_{j \in I_+} m_j - \sum_{j \in I_-} m_j \\ \mathcal{U} &= \sum_{j \in I_0} m_j \, u_j \\ \mathcal{Q} &= \sum_{j \in I_0} m_j \, q_j \ . \end{aligned} \tag{9}$$

Fixing ν and using (9), we can rewrite (8) as follows

$$\mu \mathcal{U} - \nu \mathcal{Q} + \mathcal{M} = 0 \ . \tag{10}$$

Clearly, so long as the partition of $[n]$ into the sets I_+, I_-, I_0 is intact, there is a simple linear relation between μ and ν. The number of possible subsets I_-, I_0, I_+ is finite. Thus, the range $0 < \nu < \infty$ decomposes into a finite number of intervals each of which corresponds to a fixed partition of $[n]$ into I_+, I_-, I_0. In each interval μ is a linear function of ν, unless I_0 is empty. Let ν_∞ be the smallest ν value for which I_0 is empty. Let μ_∞ be its corresponding μ value. If I_0 is never empty for any finite value of ν we define $\nu_\infty = \mu_\infty = \infty$. Clearly, replacing (ν, μ) with $(\kappa\nu, \kappa\mu)$ for any $\kappa \geq 1$ and $\nu \geq \nu_\infty$ yields the same feasible solution as $I_+(\kappa\nu) = I_+(\nu)$, $I_-(\alpha\nu) = I_-(\nu)$. Hence, as far as the original problem is concerned there is no reason to go past ν_∞ during the process of characterizing the solution. We recap our derivation so far in the following lemma.

Lemma 4.1 *For $0 \leq \nu \leq \nu_\infty$, the value of μ as defined by (7) is a unique. Further, the function $\mu(\nu)$ is a piecewise linear continuous function in ν. When $\nu \geq \nu_\infty$ letting $\mu = \mu_\infty \nu/\nu_\infty$ keeps (7) valid.*

We established the fact that $\mu(\nu)$ is a piecewise linear function. The lingering question is how many

linear sub-intervals the function can attain. To study this property, we take a geometric view of the plane defined by (ν, μ). Our combinatorial characterization of the number of sub-intervals makes use of the following definitions of lines in $\mathbb{R}^2$,

$$\ell_{+j} = \{(\nu,\mu) \mid u_j\mu - q_j\nu = +1\} \quad (11)$$
$$\ell_{-j} = \{(\nu,\mu) \mid u_j\mu - q_j\nu = -1\} \quad (12)$$
$$\ell_0 = \{(\nu,\mu) \mid \mu\mathcal{U} - \nu\mathcal{Q} + \mathcal{M} = 0\} \ , \quad (13)$$

where $-\infty < \nu < \infty$ and $j \in [n]$. The next theorem gives an upper bound on the number of linear segments the function $\mu()$ may attain. While the bound is quadratic in the dimension, for both artificial data and real data the bound is way too pessimistic.

Theorem 4.2 *The piecewise linear function $\mu(\nu)$ consists of at most n^2 linear segments for $\nu \in \mathbb{R}_+$.*

Proof Since we showed that that $\mu(\nu)$ is a piecewise linear function, it remains to show that it has at most n^2 linear segments. Consider the two dimensional function $G(\nu, \mu)$ from (8). The (ν, μ) plane is divided by the $2n$ straight lines $\ell_1, \ell_2, \ldots, \ell_n, \ell_{-1}, \ell_{-2}, \ldots, \ell_{-n}$ into at most $2n^2+1$ polygons. The latter property is proved by induction. It clearly holds for $n = 0$. Assume that it holds for $n-1$. Line ℓ_n intersects the previous $2n-2$ lines at no more than $2n-2$ points, thus splitting at most $2n-1$ polygons into two separate polygonal parts. Line ℓ_{-n} is parallel to ℓ_n, again adding at most $2n-1$ polygons. Recapping, we obtain at most $2(n-1)^2 + 1 + 2(2n-1) = 2n^2 + 1$ polygons, as required per induction. Recall that $\mu(\nu)$ is linear inside each polygon. The two extreme polygons where $G(\nu, \mu) = \pm\sum_{j=1}^n m_j$ clearly disallow $G(\nu, \mu) = 0$, hence $\mu(\nu)$ can have at most $2n^2 - 1$ segments for $-\infty < \nu < \infty$. Lastly, we use the symmetry $G(-\nu, -\mu) = -G(\nu, \mu)$ which implies that for $\nu \in \mathbb{R}_+$ there are at most n^2 segments. ■

This result stands in contrast to the Lasso homotopy tracking procedure (Osborne et al., 2000), where the worst case number of segments seems to be exponential in n. Moreover, when the prior $\boldsymbol{u}$ is uniform, $u_j = 1/\sum_{j=1}^n m_j$ for all $j \in [n]$, the number of segments is at most $n+1$. We defer the analysis of the uniform case to a later section as the proof stems from the algorithm we describe in the sequel.

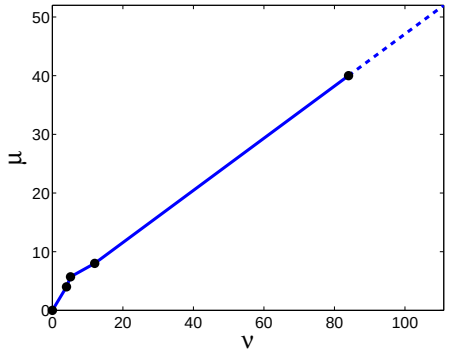

Figure 2: An illustration of the function $\mu(\nu)$ for a synthetic 3 dimensional example.

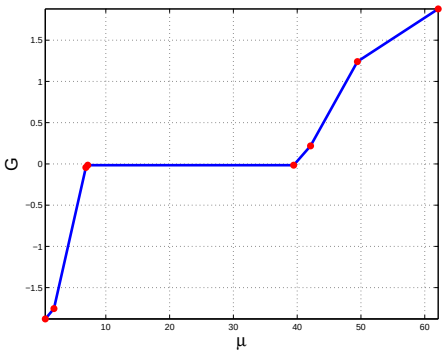

Figure 3: An illustration of the function $G(\mu)$ for a synthetic 4 dimensional example and a $\nu = 17$.

5 Algorithm for a Single Relaxation Value

Suppose we are given $\boldsymbol{u}, \boldsymbol{q}, \boldsymbol{m}$ and a specific relaxation value $\tilde{\nu}$. How can we find $\boldsymbol{p}$? The obvious approach is to solve the one dimensional monotonically nondecreasing equation $G(\mu) \stackrel{\text{def}}{=} G(\tilde{\nu}, \mu) = 0$ by bisection. In this section we present a more efficient and direct procedure that is guaranteed to find the optimal solution $\boldsymbol{p}$ in a finite number of steps. Clearly $G(\mu)$ is a piecewise linear function with at most $2n$ easily computable change points of the slope. See also Fig. (5) for an illustration of $G(\cdot)$. In order to find the slope change points we need to calculate the point (ν, μ_j) for all the lines $\ell_{\pm j}$ where $1 \le j \le n$. Concretely, these values are

$$\mu_j = \frac{\nu q_{|j|} + \text{sign}(j)}{u_{|j|}} \ . \quad (14)$$

We next sort the above values of μ_j and denote the resulting sorted list as $\mu_{\pi_1} \le \mu_{\pi_2} \le \cdots \le \mu_{\pi_{2n}}$. For any $0 \le j \le 2n$ let $\mathcal{M}_j, \mathcal{U}_j, \mathcal{Q}_j$ be the sums, defined

in (9), for the line segment $\mu_{\pi_{j-1}} < \mu < \mu_{\pi_j}$ (denoting $\mu_{\pi_0} = -\infty$, $\mu_{\pi_{2n+1}} = \infty$). We compute the sums $\mathcal{M}_j, \mathcal{U}_j, \mathcal{Q}_j$ incrementally, starting from $\mathcal{M}_0 = -\sum_{i=1}^n m_i$, $\mathcal{U}_0 = \mathcal{Q}_0 = 0$. Once the values of $j-1$'th sums are known, we can compute the next sums in the sequence as follows,

$$\begin{aligned} \mathcal{M}_j &= \mathcal{M}_{j-1} + m_{|\pi_j|} \\ \mathcal{U}_j &= \mathcal{U}_{j-1} - \mathrm{sign}(\pi_j)\, m_{|\pi_j|}\, u_{|\pi_j|} \\ \mathcal{Q}_j &= \mathcal{Q}_{j-1} - \mathrm{sign}(\pi_j)\, m_{|\pi_j|}\, q_{|\pi_j|} \ . \end{aligned}$$

From the above sums we can compute the value of the function $G(\nu, \mu)$ at the end point of the line segment $(\mu_{\pi_{j-1}}, \mu_{\pi_j})$, which is the same as the start point of the line segment $(\mu_{\pi_j}, \mu_{\pi_{j+1}})$,

$$\begin{aligned} G_j &= \mathcal{M}_{j-1} + \mathcal{U}_{j-1}\,\mu_j - \mathcal{Q}_{j-1}\,\nu \\ &= \mathcal{M}_j + \mathcal{U}_j\,\mu_j - \mathcal{Q}_j\,\nu \ . \end{aligned}$$

The optimal value of μ resides in the line segment for which $G(\cdot)$ attains 0. Such a segment must exist since $G_0 = \mathcal{M}_0 = -\sum_{i=1}^n m_i < 0$ and $G_{2n} = -\mathcal{M}_0 > 0$. Therefore, there exists an index $1 \le j < 2n$, where $G_j \le 0 \le G_{j+1}$. Once we bracketed the feasible segment for μ, the optimal value of μ is found by solving the linear equation (10),

$$\mu = \left(\mathcal{Q}_j\,\nu - \mathcal{M}_j\right)/\mathcal{U}_j \quad . \tag{15}$$

From the optimal value of μ it is straightforward to construct $\boldsymbol{p}$ using (7). Due to the sorting step, the algorithm's run time is $O(n\log(n))$ and it takes linear space. The number of operations can be reduced to $O(n)$ using a randomized search procedure.

6 Homotopy Tracking

We now shift gears and focus on the main thrust of this paper, namely, an efficient characterization of the *entire* regularization path for the maximum entropy problem. Since we have shown that the optimal solution $\boldsymbol{p}$ can be straightforwardly obtained from the variable μ, it suffices to efficiently track the function $\mu(\nu)$ as we traverse the plane (ν, μ) from $\nu = 0$ through the last change point which we denoted as (ν_∞, μ_∞). In this section we give an algorithm that traverses $\mu(\nu)$ by locating the intersections of ℓ_0 with the fixed lines $\ell_{-n}, \ell_{-n+1}, \ldots, \ell_{-1}, \ell_1, \ldots, \ell_n$ and updating ℓ_0 after each intersection.

More formally, the local homotopy tracking follows the piecewise linear function $\mu(\nu)$, segment by segment. Each segment corresponds to a subset of the line ℓ_0 for a *given* triplet $(\mathcal{M}, \mathcal{U}, \mathcal{Q})$. It is simple to show that $\mu(0) = 0$, hence we start with

$$\nu = 0,\ \mathcal{M} = 0,\ \mathcal{U} = \mathcal{Q} = 1 \ . \tag{16}$$

We now track the value of μ as ν increases, and the relaxation parameter $1/\nu$ decreases. The characterization of ℓ_0 remains intact until ℓ_0 hits one of the lines ℓ_j for $1 \le |j| \le n$. To find the line intersecting ℓ_0 we need to compute the potential intersection points $(\nu_j, \mu_j) = \ell_0 \cap \ell_j$ which amounts to calculating $\nu_{-n}, \nu_{-n+1}, \ldots, \nu_{-1}, \nu_1, \nu_2, \cdots, \nu_n$ where

$$\nu_j = \frac{\mathcal{M} u_{|j|} + \mathcal{U}\mathrm{sign}(j)}{\mathcal{Q} u_{|j|} - \mathcal{U} q_{|j|}} \ . \tag{17}$$

The lines for which the denominator is zero correspond to infeasible intersection and can be discarded. The smallest value ν_j which is larger than the current traced value of ν corresponds to the next line intersecting ℓ_0.

While the above description is mathematically sound, we devised an equivalent intersection inspection scheme which is more numerically stable and efficient. We keep track of partition I_-, I_0, I_1 through the vector,

$$s_j = \begin{cases} -1 & j \in I_- \\ 0 & j \in I_0 \\ +1 & j \in I_+ \end{cases} \ .$$

Initially $s_1 = s_2 = \cdots = s_n = 0$. What kind of intersection does ℓ_0 have with ℓ_j? Recall that $\frac{\mathcal{Q}}{\mathcal{U}}$ is the slope of ℓ_0 while $\frac{q_{|j|}}{u_{|j|}}$ is the slope of ℓ_j. Thus $\frac{\mathcal{Q}}{\mathcal{U}} > \frac{q_{|j|}}{u_{|j|}}$ means that the $|j|$'th constraint is moving "up" from I_- to I_0 or from I_0 to I_+. When $\frac{\mathcal{Q}}{\mathcal{U}} < \frac{q_{|j|}}{u_{|j|}}$ the $|j|$'th constraint is moving "down" from I_+ to I_0 or from I_0 to I_-. See also Fig. 4 for an illustration of the possible transitions between the sets. For instance, the slope of $\mu(\nu)$ on the bottom left part of the figure is larger than the slope the line it intersects. Since this line defines the boundary between I_- and I_0, we transition from I_- to I_0. We need only consider $1 \le |j| \le n$ of the following types. Moving "up" from I_- to I_0 requires

$$s_{|j|} = -1 \qquad j < 0 \qquad \mathcal{Q} u_{|j|} - \mathcal{U} q_{|j|} > 0 \ .$$

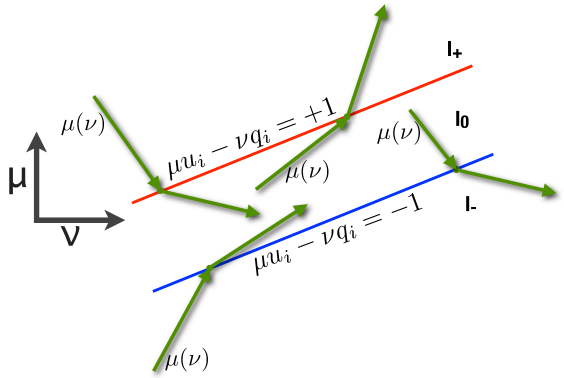

Figure 4: Illustration of the possible intersections between $\mu(\nu)$ and ℓ_j and the corresponding transition between the sets $I_\pm, I_0$.

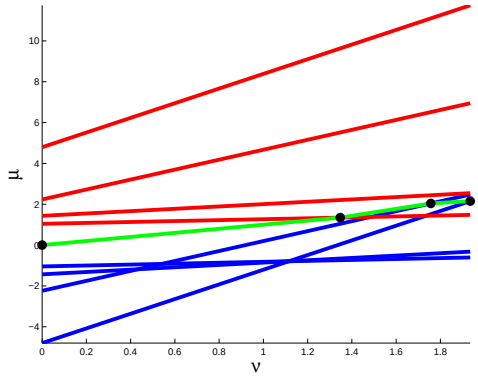

Figure 5: The result of the homotopy tracking for a 4 dimensional problem. The lines ℓ_j for $j < 0$ are drawn in blue and for $j > 0$ in red. The function $\mu(\nu)$ is drawn in green and its change points in black. Note that although the dimension is 4 the number of change points is rather small and does not exceed 4 either in this simple example.

Similarly, moving "down" from I_+ to I_0 requires

$$s_{|j|} = 1 \qquad j > 0 \qquad \mathcal{Q}u_{|j|} - \mathcal{U}q_{|j|} < 0 \ .$$

Finally, moving "up" or "down" from I_0 entails

$$s_{|j|} = 0 \qquad j(\mathcal{Q}u_{|j|} - \mathcal{U}q_{|j|}) > 0 \ .$$

If there are no eligible ν_j's, we have finished traversing $\mu()$. Otherwise let index j belong to the the smallest eligible ν_j. Infinite accuracy guarantees that $\nu_j \geq \nu$. In practice we perform the update

$$\begin{aligned}
\nu &\leftarrow \max(\nu, \nu_j) \\
\mathcal{M} &\leftarrow \mathcal{M} + \operatorname{sign}(\mathcal{Q}u_{|j|} - \mathcal{U}q_{|j|})\, m_{|j|} \\
\mathcal{U} &\leftarrow \mathcal{U} + \left(2\left|s_{|j|}\right| - 1\right)\, m_{|j|}\, u_{|j|} \\
\mathcal{Q} &\leftarrow \mathcal{Q} + \left(2\left|s_{|j|}\right| - 1\right)\, m_{|j|}\, q_{|j|} \\
s_j &\leftarrow s_j + \operatorname{sign}(\mathcal{Q}u_{|j|} - \mathcal{U}q_{|j|}) \ .
\end{aligned}$$

We are done with the tracking process when I_0 is empty, i.e. for all j $s_j \neq 0$.

The local homotopy algorithm takes $O(n)$ memory and $O(nk)$ operations where k is the number of change points in the function $\mu(\nu)$. This algorithm is simple to implement, and when k is relatively small it is efficient. An illustration of the tracking result, $\mu(\nu)$, along with the lines $\ell_{\pm j}$, that provide a geometrical description of the problem, is given in Fig. 5.

7 Uniform Prior

We chose to denote the prior distribution as $\boldsymbol{u}$ to underscore the fact that in the case of no prior knowledge $\boldsymbol{u}$ is the *uniform* distribution,

$$\mathrm{u} \stackrel{\text{def}}{=} u_j = \left(\sum_{i=1}^{n} m_i\right)^{-1} \ .$$

In this case the objective function amounts to the negative entropy and by flipping the sign of the objective we obtain the classical maximum entropy problem. The fact that the prior probability is the same for all possible observations infuses the problem with further structure which we show how to exploit in this section. Needless to say though that all the results we obtained thus far are still valid.

Let us consider a point (ν, μ) on the boundary between I_0 and I_+, namely, there exist a line ℓ_{+i} such that,

$$\mu u_i - \nu q_i = \mu \mathrm{u} - \nu q_i = 1 \ .$$

By definition, for any $j \in I_0$ we have

$$\mu u_j - \nu q_j = \mu \mathrm{u} - \nu q_j < 1 = \mu \mathrm{u} - \nu q_i \ .$$

Thus, $q_i < q_j$ for all $j \in I_0$ which implies that

$$m_j \,\mathrm{u}\, q_j \ > \ m_j \,\mathrm{u}\, q_i \ . \tag{18}$$

Summing over $j \in I_0$ we get that

$$\mathcal{Q}\,\mathrm{u} \ = \ \sum_{j \in I_0} m_j \, q_j \,\mathrm{u} \ > \ \sum_{j \in I_0} m_j \,\mathrm{u}\, q_i \ = \ \mathcal{U} q_i \ ,$$

hence,

$$\frac{q_i}{u_i} = \frac{q_i}{\mathrm{u}} < \frac{\mathcal{Q}}{\mathcal{U}}$$

and we must be moving "up" from I_0 to I_+ when the line ℓ_0 hits ℓ_i. Similarly we must be moving "down" from when ℓ_0 hits on the boundary between I_0 and I_-. We summarize these properties in the following theorem.

Theorem 7.1 *When the prior distribution $\boldsymbol{u}$ is uniform, $I_-(\nu)$ and $I_+(\nu)$ are monotonically nondecreasing and $I_0(\nu)$ is monotonically nonincreasing in $\nu > 0$. Further, the piecewise linear function $\mu(\nu)$ consists of at most $n+1$ line segments.*

The homotopy tracking procedure when the prior is uniform is particularly simple and efficient. Intuitively, there is a sole condition which controls the order in which indices would enter $I_\pm$ from I_0, which is simply how "far" each q_i is from u, the single prior value. Therefore, the algorithm starts by sorting $\boldsymbol{q}$. Let $q_{\pi_1} > q_{\pi_2} > \cdots > q_{\pi_n}$ denote the sorted vector. Instead of maintaining the vector of set indicators $\boldsymbol{s}$, we merely maintain two indices j_- and j_+ which designate the size of I_- and I_+ that were constructed thus far. Due to the monotonicity property of the sets $I_\pm$ as ν grows, the two sets can be written as, $I_- = \{\pi_j \,|\, 1 \le j < j_-\}$ and $I_+ = \{\pi_j \,|\, j_+ < j \le n\}$. The homotopy tracking procedure starts as before with $\nu = 0$, $\mathcal{M} = 0$, $\mathcal{U} = \mathcal{Q} = 1$. We also set $j_- = 1$ and $j_+ = n$ which by definition imply that $I_\pm$ are empty and $I_0 = [n]$. In each tracking iteration we need to compare only two values which we compactly denote as,

$$\nu_\pm = \frac{\mathcal{M}\mathrm{u} \pm \mathcal{U}}{\mathcal{Q}\mathrm{u} - \mathcal{U} q_{\pi_{j_\pm}}} \ .$$

When $\nu_- \le \nu_+$ we just encountered a transition from I_0 to I_- and as we encroach I_- we perform the updates, $\nu \leftarrow \nu_-$, $\mathcal{M} \leftarrow \mathcal{M} - m_{\pi_{j_-}}$, $\mathcal{U} \leftarrow \mathcal{U} - m_{\pi_{j_-}}\mathrm{u}$, $\mathcal{Q} \leftarrow \mathcal{Q} - m_{\pi_{j_-}} q_{\pi_{j_-}}$, $j_- \leftarrow j_- + 1$. Similarly when $\nu_- > \nu_+$ we perform the updates $\nu \leftarrow \nu_+$, $\mathcal{M} \leftarrow \mathcal{M} + m_{\pi_{j_+}}$, $\mathcal{U} \leftarrow \mathcal{U} - m_{\pi_{j_+}}\mathrm{u}$, $\mathcal{Q} \leftarrow \mathcal{Q} - m_{\pi_{j_+}} q_{\pi_{j_+}}$, $j_+ \leftarrow j_+ - 1$.

The tracking process stops when $j_- > j_+$ as we exhausted the transitions out of the set I_0 which becomes empty. Homotopy tracking for a uniform prior takes $O(n)$ memory and $O(n \log(n))$ operations and is very simple to implement.

We also devised a global homotopy tracking algorithms that requires a priority queue which facilitates insertions, deletions, and finding the largest element in the queue in $O(\log(n))$ time. The algorithm requires $O(n)$ memory and $O(n^2 \log(n))$ operations. Clearly, if the number of line segments constituting $\mu(\nu)$ is greater than $n \log(n)$ (recall that the upper bound is $O(n^2)$) then the global homotopy procedure is faster than the local one. However, as we show in Sec. 8, in practice the number of line segments is merely linear and it thus suffices to use the local homotopy tracking algorithm.

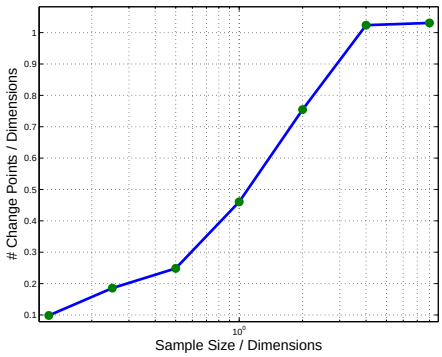

Figure 6: The number of line-segments in the homotopy as a function of the number of samples used to build the observed distribution $\boldsymbol{q}$.

8 Number of line segments in practice

The focus of the paper is the design and analysis of a novel homotopy method for maximum entropy problems. We thus left with relatively little space to discuss the empirical aspects of our approach. In this section we focus on one particular experimental facet that underscores the usability of our apparatus. We briefly discuss current natural language applications that we currently work on in the next section.

The practicality of our approach hinges on the number of line segments that occur in practice. Our bounds indicate that this number can scale quadratically with the dimension, which would render the homotopy algorithm impractical when the size of the alphabet is larger than a few thousands. We therefore extensively tested the *actual* number of line segments in the resulting homotopy when $\boldsymbol{u}$ and $\boldsymbol{q}$ are Zipf (1949) distributions. We used an alphabet of size $50,000$ in our experiments. The distribution $\boldsymbol{u}$ was set to be the Zipf distribution with an offset parameter of 2, that is, $u_i \sim 1/(i+2)$. We defined a "mother" distribution for $\boldsymbol{q}$, denoted $\bar{\boldsymbol{q}}$, which is

a plain Zipf distribution without an offset, namely $\bar{q}_i \sim 1/i$. We then sampled $n/2^l$ letters according to the distribution $\bar{\boldsymbol{q}}$ where $l \in -3, \ldots, 3$. Thus the smallest sample was $n/2^3 = 6,250$ and the largest sample was $n/3^{-3} = 40,000$. Based on the sample we defined the observed distribution $\boldsymbol{q}$ such that q_i is proportional to the number of times the i'th letter appeared in the sample. We repeated the process 100 times for each sample size and report average results. Note that when the sample is substantially smaller than the dimension the observed distribution $\boldsymbol{q}$ tends to be "simple" as it consists of many zero components. In Fig. 6 we depict the average number line segments for each sample size. When the sample size is one eighth of the dimension we average st most $0.1\,n$ line segments. More importantly, even when the size of the sample is fairly large, the number of lines segments is linear in the dimension with a constant close to one. We also performed experiments with large sample sizes for which the empirical distribution $\boldsymbol{q}$ is very close to the mother distribution $\bar{\boldsymbol{q}}$. We seldom found that the number of line segments exceeds $4n$ and the mode is around $2n$. These findings render our approach usable even in the very large natural language applications.

9 Conclusions

We presented a novel efficient apparatus for tracking the entire relaxation path of maximum entropy problems. We currently study natural language processing applications. In particular, we are in the process of devising homotopy methods for domain adaptation Blitzer (2008) and language modeling based on context tree weighting (Willems et al., 1995). We also examine generalization of our approach in which the relative entropy objective is replaced with a separable Bregman (Censor and Zenios, 1997) function. Such a generalization is likely to distill further connections to the other homotopy methods, in particular the least angle regression algorithm of Efron et al. (2004) and homotopy methods for the Lasso in general (Osborne et al., 2000). We also plan to study separable Bregman functions in order to derive entire path solutions for less explored objectives such as the Itakura-Saito spectral distance (Rabiner and Juang, 1993) and distances especially suited for natural language processing.

References

A.L. Berger, S.A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22 (1):39–71, 1996.

John Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania, 2008.

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

Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York, NY, USA, 1997.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:179–190, 1997.

M. Dudík, S. J. Phillips, and R. E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, June 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4 (3):227–241, September 1968.

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2): 319–337, 2000.

L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.