Stable Mixing of Complete and Incomplete Information

by

Adrian Corduneanu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2002

Author .....
Department of Electrical Engineering and Computer Science
February 5, 2002

Certified by .....
Tommi Jaakkola
Assistant Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Stable Mixing of Complete and Incomplete Information

by

Adrian Corduneanu

Submitted to the Department of Electrical Engineering and Computer Science
on February 5, 2002, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

An increasing number of parameter estimation tasks involve the use of at least two information sources, one complete but limited, the other abundant but incomplete. Standard algorithms such as EM used in this context are unfortunately not stable in the sense that they can lead to dramatic loss of accuracy with the inclusion of incomplete observations. We cast estimation with incomplete information as a source allocation problem, and show that instability occurs at a well-defined data-dependent source allocation. We introduce an algorithm that finds the critical allocation by tracing local maxima of the likelihood from full weight on complete data to full weight on incomplete data. Theoretical results support that such tracing is always possible in the exponential family by following the differential equation that governs the evolution of local maxima while changing allocation. Our approach finds in $O(n^3)$ in the number of model parameters the critical allocation along with a stable estimate. We demonstrate the effectiveness of the algorithm on artifically generated mixtures of Gaussians and on text classification with naïve Bayes. The approach readily generalizes to other problems that feature from multiple sources of information.

Thesis Supervisor: Tommi Jaakkola
Title: Assistant Professor

Acknowledgments

This work was supported in part by Nippon Telegraph and Telephone Corporation, by ARO MURI grant DAAD 19-00-1-0466, and by NSF ITR grant IIS-0085836.

Contents

List of Figures

Chapter 1

Introduction

Virtually any estimation problem must cope with incomplete information. It may be that data originates in fragmented experiments that produce samples with missing entries; or that a variable is missing altogether from the dataset, but its existance is conjectured from physical knowledge about the domain; yet, in other situations latent variables are artificially introduced to increase efficiency of estimation. Estimation algorithms that operate under incomplete information have been an active area of research for decades.

Many modern application areas involve estimation estimation in which incomplete data is prevalent, but complete samples also exist. A common instance of this type of problem is classification. The goal of classification is to construct from a collection of labeled training samples a classifier that predicts class labels of unseen data, but often training labels are expensive to generate, and most of the training samples are available with unknown class labels. This is the case in text classification, where documents in electronic form are readily available, but labeling requires human intervention. In other classification problems, such as those linked to the medical domain or bioinformatics, each labeling requires extensive expert knowledge, or is the result of laborious experiments, therefore datasets with only a few labeled samples are common. Similarly, in image classification images are widely available, but users construct queries by presenting only a few images from the intended class. Natural Language Processing is another domain that generates many problems amenable to

estimation with incomplete knowledge. To name a few, part-of-speech tagging, parsing, word sense disambiguation, all have access to large numbers of sentences, but labeling large corpora is expensive and often unreliable.

One may argue why unlabeled data can be useful at all for classification. It is impossible to build a classifier that performs better than random with unlabeled data alone, therefore the first thought is that incomplete data is not helpful in this context. However, in the presence of even very few labeled samples, unlabeled data can noticeably improve classification. Consider for example the named entity classification problem, in which names occurring in a text must be identified as belonging to one of few classes (*person*, *organization*, *location*, and so on). Collins [7] shows how document structure can be used to learn an accurate classifier from very few labeled samples. He uses two types of rules that determine the class of named entities: spelling rules, such as *Mr.*, a sub-word, or the entire entity, and context rules, such as the modifier *president*. A key *co-occurrence* assumption enables derivation of new rules over unlabeled data: spelling and context rules occurring at the same time on some named entity must detect the same class. Initially, the classifier knows only a few rules, but can derive more rules of each type from the other type due to co-occurrence. After all the possible rules have been derived, the result is a much better classifier. In general, like in any learning problem, unlabeled data can improve classification due to a priori model assumptions, that effectively correlate class label information with incomplete information.

Current approaches to classification with labeled and unlabeled data fall in three categories: constraint-based, generative, and regularization methods. We provide a brief account of each paradigm.

The solution to named entity classification discussed above is a constraint-based method. Algorithms from this category equate labels of different data points through hard constraints trained from unlabeled data. The constraints establish equivalence sets of data points that must belong to the same class, so that eventually each set needs only one labeled training sample. For example, Yarowsky [24] improves his word-sense-disambiguation classifier based on the local context of the word by en-

forcing the constraint that all instances of a word occurring in the same document must have the same meaning. Unlabeled occurrences of a word of interest can than be labeled at the document level, and used to train a better local-context classifier. Note that such hard constraints are often unreasonable, limiting the use of constraint-based methods to only a few applications.

Blum and Mitchell formalize constraint-based methods in their work on co-training [4]. In co-training the key assumption is that data can be redundantly represented by two independent sets of features, each being sufficient for classification, and that feature sets co-occurring in unlabeled data must have the same label. Each set of features yields a different classifier, and the co-occurrence condition along with unlabeled samples can be used to iteratively train one classifier over the other. Co-training can be very powerful, but like other constraint-based methods requires quite demanding assumptions.

In contrast, generative methods are less restrictive and are applicable to a wide range of problems. Such methods model the full joint probability distribution of data and labels, and learn its parameters from both labeled and unlabeled samples. Estimation usually resumes to finding parameters that maximize a likelihood-based criterion, such as the probability of the labeled and unlabeled training set under the model. The EM algorithm mentioned in the beginning specifically deals with missing information under generative models, and is directly applicable to the labeled/unlabeled setting. Intuitively, EM runs as an iteration that labels all incomplete data with the best estimate under the current model in one step, and re-estimates the model from the artificially completed data in the next step. EM has been extensively used to combine complete and incomplete information (see [12, 20] for a mixtures of Gaussians context, [16] for mixtures of experts, or [17] for document classification with naïve Bayes). Generative methods are very popular due to their flexibility in constructing complex models and incorporating prior information. Their disadvantage lies in the fact that they attempt to solve a more general problem than classification, and that optimization by EM is prone to local optima.

The last class of algorithms include a variety of methods that use unlabeled data

through regularization that modifies classification, without modeling the full joint of data and labels. Because such methods optimize classification error directly, they usually require less training data than generative methods. Transductive inference [21] for example advocates minimization of classification error on the given unlabeled samples only, without attempting to generalize on unseen data. For an instance of transductive inference in practice see Joachims' transductive SVM [13]. Other regularization methods include using unlabeled data to define the kernels for large margin classifiers, or metrics that can used in building classifiers [19].

The motivation for our work is that current estimation algorithms with labeled and unlabeled data are often unstable in the sense that inclusion of unlabeled data may dramatically improve but also reduce performance compared to labeled-data only estimates. This is the case because estimation with incomplete data heavily relies on model assumptions, and the addition of unlabeled data may amplify the effect of their violation. For example, consider EM in generative modeling: when model assumptions are inaccurate, incorrect labeling of incomplete samples followed by re-estimation based on the inaccurate labeling may eventually attenuate labeling errors, but may also accumulate them, and drastically diverge to an erroneous solution. Intuition suggests that there exists an *error level* [1] that achieves a transition between convergence to correct parameters and divergence to a flawed solution. Below that critical error level, inclusion of unlabeled data is beneficial, but above that level incomplete information is very damaging to estimation. Because current generative algorithms for estimation with incomplete information cannot distinguish between the two regimes, they are inherently unstable.

Central to our work is the view of estimation with complete and incomplete information as an allocation problem between two heterogeneous data sources providing often conflicting information about the variable of interest. The complete source is reliable and preferred, but scarce, while the incomplete source is less robust to model

[1] We introduce the notion of *error level* for intuitive purposes only, without claiming a formal definition of the concept. In practice we use other well-defined quantities with properties similar to those of the error level.

violations, but is abundant. Generative algorithms combine the two sources of information under an implicit weighting that achieves a certain level of robustness to model misfit. For example, because EM optimizes log-likelihood, which is additive in labeled and unlabeled samples, the algorithm allocates the two data sources according to the ratio of their sample counts. However, we can easily generalize such algorithms to work with any source allocation on the scale from complete-data only to incomplete-data only estimation.

There is a direct link between source allocation and sensitivity to model misfit; the less weight on unlabeled data, the less we trust it, and model assumptions are less important; thus source allocation effectively trades off robustness to model assumptions and the potential benefits brought by unlabeled data. This tradeoff has been recognized by other researchers [17], but currently there are no principled methods of choosing the allocation.

In this work we introduce an algorithm that explicitly identifies in a data-dependent manner the maximal source allocation that achieves stable estimation with incomplete information. Larger allocations presume model assumptions inconsistent with available training data, while for smaller values the misfit is small enough that the algorithm converges in the correct direction. The sudden change in performance at a well-defined allocation is due to the critical error level property mentioned above. Our algorithm continuously traces estimation parameters from complete-data only to incomplete-data only allocation by following the differential equation that governs the evolution of local optima. The significant feature of this approach is that it provides an explicit way of identifying critical points that arise in the estimation problem as discontinuities in the tracing process.

While we present our method in the context of generative modeling within the exponential family, we envision a wealth of applications of the algorithm to other problems that deal with multiple sources of information.

1.1 Outline

The thesis is organized as follows. In Chapter 2 we give a formal description of the problem, possible likelihood-based optimization criteria, and standard algorithms that optimize them. In Chapter 3 we turn to information geometry and present a geometrical analysis of mixing complete and incomplete data, providing intuition of why instability occurs, and how continuation methods can detect it. Chapter 4 presents a naïve approach to continuation, showing how it can achieve stability, but also why it is problematic. Chapter 5 is the formal presentation of our method, with complete theoretical arguments. To illustrate the use of homotopy continuation for mixing complete and incomplete information, Chapter 6 comprises its application to discrete naïve Bayes models, and mixtures of multivariate Gaussians. Experiments that support our method are presented in Chapter 7, and Chapter 8 concludes with directions of further research.

Chapter 2

Estimation with Incomplete Information

The purpose of this chapter is to give a formal treatment of generative methods for estimation with incomplete information. In the generative framework, one models the full probability distribution over complete data (both data and labels in a labeled/unlabeled setting), and than uses available complete and incomplete information to estimate the parameters of the model. We address issues such as what model family should the distribution come from, what criteria could be used to define the goal of estimation, and what are the standard algorithms that optimize them.

While we keep our presentation to the highest level of generality to account for a wide range of model families and types of incompleteness, for valuable intuition the reader should refer to the special case of classification with labeled and unlabeled data. Here the goal is to build a classifier that generalizes well from few labeled (complete) samples, and many unlabeled (incomplete) samples. Generative methods model a probability distribution over both data and labels, whose parameters are estimated from training data. To classify a new unlabeled sample we select the label that has the highest probability under the model given that sample.

Note that although we learn the full joint over data and labels, classification uses only the conditional of label given data. As opposed to fully generative methods, *discriminative* techniques use training data more efficiently by estimating only the

parameters of the conditional while ignoring the marginal. Although discriminative methods are superior in fully supervised classification, no algorithm that combines complete and incomplete information can be purely discriminative, because unlabeled samples provide direct information about the marginal only. Moreover, unlabeled data is much more available, which makes estimation of the marginal without inefficient use of labeled training data feasible; therefore, for the purpose of estimation with incomplete information we can safely focus on fully generative models.

2.1 Density Estimation within the Exponential Family

The goal in generative modeling is to find a density $p(z|\boldsymbol{\eta})$ parameterized by $\boldsymbol{\eta}$ on the space of complete samples $\mathcal{Z}$ from a chosen model family $\mathcal{M}$ that best matches the true but unknown density $p^*(z)$ in the following sense:

$$\boldsymbol{\eta}^* = \arg\min_{\boldsymbol{\eta}\in\mathcal{M}} D(p^*(z) \,\|\, p(z|\boldsymbol{\eta})) \tag{2.1}$$

where $D(f \,\|\, g) = \int f \log \frac{f}{g}$ is the Kullback Leibler distance (or KL-divergence) between densities [8]. The use of KL-divergence is motivated by its relation to maximum likelihood which will be apparent later in this chapter. Nevertheless, the actual distance metric is of no practical use because the true distribution p^* is never available; therefore, we see (2.1) only as a guiding principle. Also, note that p cannot match p^* exactly only because of possible model misfit of the true distribution.

The choice of $\mathcal{M}$ is important, because incomplete data provides information about the conditional only through the specifics of the model family. Restricting the model family gives more power to incomplete data over estimation, but it is potentially dangerous if real data violates model assumptions. To accommodate a wide variety of model families, we formalize our algorithm in the context of any exponential family [3]. Any distribution family that can be described by a finite number of sufficient statistics is exponential, therefore this class contains most of the families used in

practice.

Specifically, $\mathcal{M}$ an exponential family of natural parameters $\boldsymbol{\theta}$ has the following standard form:

$$p(z|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}\mathbf{t}(z)^T + k(z) - \psi(\boldsymbol{\theta})\right) \tag{2.2}$$

where $\mathbf{t}(z)$ the vector of sufficient statistics, and $\psi(\boldsymbol{\theta})$ the partition function.

We can equivalently specify the exponential family by its mean parameterization $\boldsymbol{\eta}$, that can be easily converted to natural parameters:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{t}(z)] = \frac{d}{d\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) \tag{2.3}$$

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \frac{d}{d\boldsymbol{\eta}}\int p(z|\boldsymbol{\eta}) \log p(z|\boldsymbol{\eta}) dx \tag{2.4}$$

Because of the one-to-one mapping between mean parameters and natural parameters we will specify the same density in either parameterization: $p(z|\boldsymbol{\eta})$ or $p(z|\boldsymbol{\theta})$. Note that the space of natural parameters is restricted to the domain of the partition function, while the space of mean parameters to its range. Moreover, both natural and mean parameters lie in convex domains.

2.2 Formalization of Incompleteness

In order to introduce estimation with incomplete information in a general setting, we must give a flexible formalization of incompleteness. Following [10], we define the relationship between complete and incomplete data by a many-to-one map $\mathbf{z} \rightarrow \mathbf{x}(\mathbf{z})$ between the space of complete samples $\mathcal{Z}$ and the space of incomplete samples $\mathcal{X}$. An incomplete sample $\mathbf{x}$ comes from an unknown complete sample from the inverse image $\mathcal{Z}(\mathbf{x})$ under the mapping. This general notion of incompleteness covers all situations in which complete data decomposes in independent observed and unobserved variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, where the mapping is projection to the first component of $\mathbf{z}$. For example, in the labeled/unlabeled data setting $\mathbf{x}$ is unlabeled data, and $\mathbf{y}$ is the class label.

Under this general view of incompleteness we can define marginalization by integrating over all consistent complete samples. For any density $p(\mathbf{z})$ on the complete-data space we can define an incomplete-data density on $\mathcal{X}$ by integration:

$$p_X(\mathbf{x}) = \int_{\mathcal{Z}(\mathbf{x})} p(\mathbf{z}) d\mathbf{z} \tag{2.5}$$

Using marginalization, we can define for each incomplete sample $\mathbf{x}$ a conditional density on $\mathcal{Z}(\mathbf{x})$ by $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})/p_X(\mathbf{x})$. This definition is actually the well-known Bayes rule, that provides in our setting a separation between information provided by incomplete data (the *marginal*) and other information in the complete density (the *conditional*).

Indeed, when no information is assumed a priori about the complete density, the conditional is independent of the marginal and thus of any incomplete samples. However, often we are interested only in the conditional, as it is the case in classification with labeled and unlabeled data, and fortunately prior assumptions about the model correlate conditionals with marginals. Even without independence the marginal/conditional decomposition is useful because the marginal estimate contains all information about incomplete samples, while the impact of incomplete data on estimation with complete data depends only on model restrictions that govern the relationship between the marginal and the conditional.

2.3 Setup of the Problem

In this section we merely introduce useful notation for describing estimation with incomplete information. We are presented with complete samples $\mathbf{z}^C = (z_1^C, z_2^C, \ldots, z_N^C)$ and incomplete samples $\mathbf{x}^I = (x_1^I, x_2^I, \ldots, x_M^I)$ generated i.i.d. from an unknown density $p^*(z)$ on $\mathcal{Z}$ and its marginalization, where typically $N \ll M$. The goal is to use both $\mathbf{z}^C$ and $\mathbf{x}^I$ to estimate the mean parameters $\boldsymbol{\eta}$ of the best approximation to $p^*(z)$ from the exponential family $\mathcal{M}$.

2.4 Criteria for Estimation with Incomplete Information

In practice, we must work with an approximation of (2.1) that is computable from available training data, and whose optimization finds an estimate of $\boldsymbol{\eta}^*$. We build on the standard maximum likelihood criterion without incomplete data:

$$\arg\max_{\boldsymbol{\eta}\in\mathcal{M}} \sum_{1\leq i\leq N} \log p(z_i^C|\boldsymbol{\eta}) \tag{2.6}$$

whose estimate is well-understood and well-behaved (of minimum variance if such estimator exists, and asymptotically unbiased), but likely to be poor because of very limited complete data. The challenge is to improve the maximum likelihood $\boldsymbol{\eta}^*$ estimate by augmenting the criterion with incomplete data.

Note that if we replace $p^*(z)$ in (2.1) with the complete-data empirical estimate $\hat{p}^C(z) = \frac{1}{N}\sum_{1\leq i\leq N}\delta(z - z_i^C)$, we obtain an equivalent formulation of the standard maximum likelihood criterion. This relationship to maximum likelihood is the reason we chose KL-divergence as the distance measure between true and modeled density. In order to benefit from geometrical intuition we will use the KL-divergence formulation of likelihood whenever possible.

At the other extreme the likelihood of incomplete data provides an estimate of the full density:

$$\arg\max_{\boldsymbol{\eta}\in\mathcal{M}} \sum_{1\leq j\leq M} \log p_X(x_i^C|\boldsymbol{\eta}) \tag{2.7}$$

or its equivalent KL-divergence formulation $D(\hat{p}^I(x) \,\|\, p_X(x|\boldsymbol{\eta}))$, where $\hat{p}^I(x) = \frac{1}{M}\sum_{1\leq j\leq M}\delta(x - x_j^I)$ is the empirical estimate from incomplete data.

We must make a clear distinction between the standard usage of the incomplete-data likelihood, density estimation from a latent variable model, and its usage in the context of mixing complete and incomplete information. In the standard setting, although maximum likelihood gives an estimate of the full density, only the marginal

on the incomplete space is relevant; however, here the focus is on the conditional (in classification for instance), and a good estimate of the marginal is important only to the extent of improving the conditional.

While maximum likelihood of incomplete data provides a good estimate for the marginal, it is not as powerful for estimating the conditional. Even in the limit of infinite incomplete samples, when maximizing the likelihood amounts to having a perfect estimate for the marginal, there is more than one compatible conditional. For example, if incomplete data is unlabeled, a label permutation of the conditional achieves the same likelihood on incomplete data. The power of incomplete likelihood on evaluating estimates of the conditional can be as low as providing no information at all, when the model family of the estimates is too unrestricted.

The two estimators we have introduced, maximum likelihood with respect to complete data and with respect to incomplete data, represent sources of information about the conditional density that are often conflicting. For instance, allowing a slight reduction in incomplete likelihood may permit a sudden increase in complete likelihood, through a major reorganization of the conditional. The relationship between the two estimators stems from the specifics of the model family, but it is hard to characterize analytically in general.

We introduce a family of estimators that combine complete and incomplete log-likelihoods in a linear fashion with a source allocation parameter λ. Expressed in terms of KL-divergences, the criterion to be minimized is given by:

$$(1-\lambda)D(\hat{p}^C(z) \,\|\, p(z|\boldsymbol{\eta})) + \lambda D(\hat{p}^I(x) \,\|\, p_X(x|\boldsymbol{\eta})) \tag{2.8}$$

The linear mixing of the log-likelihoods is not only an artificial trick that aids estimation, but it has an important interpretation very relevant for the problem of estimation with incomplete information. Specifically, it can be shown that (2.8) is equivalent to the following criterion:

$$\arg\min_{\boldsymbol{\eta}\in\mathcal{M}} \{D(\hat{p}^I(x) \,\|\, p_X(x|\boldsymbol{\eta})) \text{ for } D(\hat{p}^C(z) \,\|\, p(z|\boldsymbol{\eta})) \leq \beta\} \tag{2.9}$$

where β is a monotonically increasing function of λ. To give some intuition into why the two criteria are equivalent note that if we put a Lagrange multiplier on the inequality constraint in $(eq : equiv_k ldiv_c riterion)$ we derive a criterion that is a linear combination of the likelihoods. For a formal proof see Theorem A.4 in the Appendix.

Therefore, by optimizing the mixed likelihood we effectively minimize the distance to the incomplete empirical estimate while bounding the distance to the complete empirical estimate. If $\lambda = 0$, the bound is so tight that only the complete-only estimate satisfies it. In contrast, if $\lambda = 1$ the bound on distance to $\hat{p}^C$ is vacuous, and we effectively perform an unconstrained minimization of the incomplete-data only likelihood.

Although equivalent, criterion (2.9) is harder to optimize than (2.8), and the mapping $\beta(\lambda)$ cannot be computed explicitly; therefore, in practice we must use the linear combination criterion.

Setting λ to $M/(M+N)$ retrieves the usual likelihood of the entire data set, the sum of complete and incomplete likelihoods. This is the usual criterion currently used in generative learning whenever training data contains incomplete samples, partly because the standard algorithm for training latent variable models (EM) directly optimizes this criterion. EM for mixing complete and incomplete information has been extensively used empirically (see for instance [17]).

However, due to complex interactions between the two estimators, fixing λ to $M/(M+N)$ is by no means optimal. Our goal is to find an a better data-dependent allocation parameter that provides stable estimation, in the sense that small changes in complete information lead to small changes in the estimated density. Instead of performing a model-by-model analysis which is often intractable, we use homotopy continuation methods (Chapter 5) to provide a general solution to the problem.

The algorithm we provide is not restricted to the likelihood criterion, but can be applied to any estimator that trades off complete vs. incomplete information, and that exhibits major changes with small variations in allocation. For example we can exhibit a different estimator from a geometrical perspective, that minimizes the distance to a set of distributions that combines the empirical complete and empirical

incomplete estimates. Formally, we define $\mathcal{P}_\lambda$ to be the set of densities obtained by completing the incomplete data with any conditional:

$$\mathcal{P}_\lambda = \left\{p : \; p(z) = (1-\lambda)\hat{p}^C(z) + \lambda \hat{p}^I(\mathbf{x}(z)) r_{\mathbf{x}(z)}(z) \text{ for some densities } r_x \text{ on } \mathcal{Z}(x)\right\} \tag{2.10}$$

The estimate $\hat{\boldsymbol{\theta}}$ is now found by minimizing the KL-divergence to the set $\mathcal{P}_\lambda$ or

$$D(\mathcal{P}_\lambda \,\|\, p(z|\boldsymbol{\theta})) \equiv \min_{q \in \mathcal{P}_\lambda} D(q(z) \,\|\, p(z|\boldsymbol{\theta})) \tag{2.11}$$

The first criterion leads to a weighted version of the EM-algorithm and the latter one to the em-algorithm [1]. The two algorithms are often identical but not here.

2.4.1 Identifiability with Incomplete Information

An important property of model families relevant for estimation with complete and incomplete information is identifiability:

Definition A model family $\mathcal{M}$ is *identifiable with incomplete information* if for any densities $p, q \in \mathcal{M}$ with $p_X \equiv q_X$ there exist one-to-one mappings $\sigma_x : \mathcal{Z}(x) \to \mathcal{Z}(x), \forall x \in \mathcal{X}$ such that $p(z) \equiv q(\sigma_{\mathbf{x}(z)}(z))$.

When $\mathcal{Z}(x)$ is discrete, identifiability amounts to the fact that the marginal determines the conditional up to a permutation of the missing component. The importance of identifiability is that for such models it is enough to estimate p_X from incomplete data, and than compute the posterior of σ_x given complete evidence, in order to produce a robust estimator.

Many common model families are identifiable. For instance mixtures of normal distributions are known to be identifiable up to a relabeling of the components [15]. Also, discrete naïve Bayes models with binary features and binary class variable are identifiable if they have more than 3 features [11]. If the class variable is not binary the exact number of features that make the naïve Bayes model identifiable is not known.

Nevertheless, even identifiable model families are virtually unidentifiable for the standard likelihood maximization algorithms, because convergence to local maxima introduces many degrees of freedom.

2.5 Maximum Likelihood with EM

Normally, incomplete likelihood is optimized by the EM algorithm [10]. Intuitively, the EM formulation is simple: initialize the algorithm with some parameters of the density, and then iteratively update them in a two-step iteration: the E step that completes incomplete training data with its expected value under current parameters, and the M step that re-estimates parameters by maximizing likelihood of completed data. Without any modification, if both complete and incomplete data are present EM optimizes the combined likelihood of complete and incomplete information, that is (2.8) with $\lambda = M/(M+N)$.

Extending EM to optimizing (2.8) with arbitrary λ is straightforward. For our analysis it is convenient to view each EM iteration as an operator that acts on the parameters of the current estimate, therefore this presentation of EM merges E and M steps into a single operator:

$$\mathrm{EM}_\lambda(\boldsymbol{\eta}) = (1-\lambda)\frac{1}{N}\sum_{1\le i\le N}\mathbf{t}(z_i^C) + \lambda\frac{1}{M}\sum_{1\le j\le M}\mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z)=x_j^I,\boldsymbol{\eta}] \tag{2.12}$$

where $\boldsymbol{\eta}$ are the mean parameters of our exponential family (2.2). For a derivation of EM in this general setting see Theorem A.1 in the appendix.

To retrieve the classical EM algorithm from its general version, it is enough to set $\lambda = 1$,thus ignoring all complete training data. Consequently, we will often use EM_1 as a baseline. Moreover, we can express the general algorithm in terms of EM_1:

$$\mathrm{EM}_\lambda(\boldsymbol{\eta}) = (1-\lambda)\boldsymbol{\eta}^C + \lambda\mathrm{EM}_1(\boldsymbol{\eta}) \tag{2.13}$$

where $\boldsymbol{\eta}^C$ are the mean parameters of the maximum likelihood density based on

complete data only.

The linear separation between complete and incomplete information in the mean parameterization will play an important role in proving regularity properties of the algorithm for finding stable allocation.

Chapter 3

Geometry of Incomplete Information

In this chapter we cast in terms of concepts from information geometry [2] the problem of estimation with incomplete information, and the associated optimization criteria and algorithms as described in Chapter 2. The goal is to provide a better understanding of the current algorithms, that reveals their stability issues, and suggests a direction for improvement.

We distinguish two spaces of distributions in our analysis: the space of complete distributions over $\mathcal{Z}$, and the space of marginal distributions over $\mathcal{X}$. Marginalization defines a mapping between the two spaces. The distance between distributions in each of these spaces is KL-divergence, even if it is not a metric and not even symmetric.

Families of densities form manifolds in these spaces. Relevant manifolds are the model family $\mathcal{M}$, viewed both in the complete and marginal spaces ($\mathcal{M}_X$), and the data manifold $\mathcal{D}$, the family of all complete densities consistent with incomplete evidence. Exponential-family manifolds are endowed with a differential geometrical structure, with Fisher information as the Riemannian metric, and the exponential and mixture affine connections.

We summarize complete and incomplete evidence by two points that represent the empirical complete estimate from complete data $\hat{p}^C$, and the empirical marginal estimate from incomplete data $\hat{p}^I$. As mentioned above, the inverse image of incomplete

evidence under marginalization $q(z)\hat{p}^I(\mathbf{x}(z))/q_X(\mathbf{x}(z)), \forall q$ on $\mathcal{Z}$, forms the data manifold. Optimization criteria for estimation with incomplete information are functions of distances (KL-divergences) between $\hat{p}^I$, $\hat{p}^C$, $\mathcal{M}$, and $\mathcal{D}$.

We call the point that achieves the minimum distance between a point and a manifold projection. Because of the asymmetry of KL-divergence we distinguish two types of projection: the m-projection $\arg\min_{p\in\mathcal{M}} D(q \,\|\, p)$, and the e-projection $\arg\min_{q\in\mathcal{D}} D(q \,\|\, p)$. Normally both projections are analytically tractable in the complete space (e-projections may be problematic, but not for our choice of $\mathcal{D}$), and not tractable in the marginal space. Maximum likelihood with respect to complete data is just the m-projection of $\hat{p}^C$ to $\mathcal{M}$, and the maximum likelihood estimate with respect to incomplete data is the m-projection of $\hat{p}^I$ to $\mathcal{M}_X$ in the incomplete space, which is not analytically tractable.

3.1 Geometry of Different Criteria and Their Optimization

3.1.1 Maximum Likelihood and EM

No Complete Data

We can m-project in the incomplete space with the EM algorithm. First we see that the m-projection to $\mathcal{M}_X$ is equivalent to minimizing the distance between $\mathcal{D}$ and $\mathcal{M}$ in the complete space. This minimization cannot be carried out directly, but can be approximated by an iteration of successive e- and m-projections between $\mathcal{M}$ and $\mathcal{D}$. This is in fact Amari's *em* algorithm as in [1], which is the same as EM in the no-complete-data setting, but may differ in more complex situations, as we shall see it is the case when we incorporate complete information.

Csiszár [9] shows that if both $\mathcal{D}$ and $\mathcal{M}$ are convex this alternating minimization procedure necessarily converges to the true distance between the manifolds. In our setting $\mathcal{D}$ is convex, but $\mathcal{M}$ is not. Because of lack of convexity, the alternating minimization may converge only to a local minimum of the distance between the

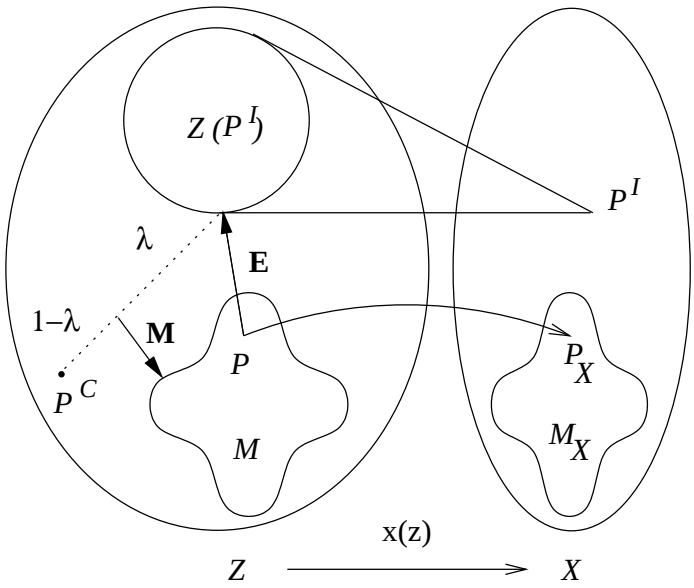

Figure 3-1: EM_λ as alternating minimization

manifolds. Even if the global minimum is found, there are many pairs of complete densities that achieve this minimum, at least up to a label permutation when incomplete data is unlabeled. Thus based only on incomplete information it is impossible to distinguish between these alternatives.

Complete and Incomplete Data

As we have seen in Chapter 2, EM generalizes easily to mixing complete and incomplete information. The EM_λ version of EM for any allocation parameter, optimizes (2.8), whose geometrical interpretation is finding a density in $\mathcal{M}$ that minimizes the λ linear combination between the distance to $\hat{p}^C$ and the distance to $\mathcal{D}$. There are two equivalent ways to geometrically describe the operation of EM_λ:

- E step e-projects to $\mathcal{D}$ then constructs a λ linear combination with $\hat{p}^C$; M step m-projects back to $\mathcal{M}$ (Figure 3-1).

- E step e-projects to $\mathcal{D}$; M step m-projects to $\mathcal{M}$ then constructs a λ linear combination in the mean parameterization of $\mathcal{M}$ with the m-projection of $\hat{p}^C$ on $\mathcal{M}$.

Therefore EM keeps iterates close to the empirical complete density by averaging with $\hat{p}^C$ at each step. Indeed, Theorem A.3 in the appendix shows that such linear combinations bound the distance to $\hat{p}^C$. In fact, criterion (2.9) states that the objective of EM_λ effectively is to minimize the distance between $\mathcal{M}$ and $\mathcal{D}$ subject to bounding the distance to $\hat{p}^C$.

3.1.2 Amari's *em*

Surprisingly, in the presence of complete information Amari's *em* algorithm [1] differs from EM_λ. *em* is a straight alternation of *e*- and *m*-projections between data and model manifolds, while in EM the E step is defined as an expectation that may differ from *e*-projection, though for usual models it does not.

In order to present the *em* algorithm we need to define a data manifold that incorporates both complete and incomplete evidence. The obvious choice is the following linear combination:

$$\mathcal{D}' = \{q' : q' = (1-\lambda)\hat{p}^C + \lambda q, \text{ for some } q \in \mathcal{D}\} \tag{3.1}$$

The *em* algorithm attempts to minimize the distance between $\mathcal{D}'$ and $\mathcal{M}$ by alternate projections:

$$q^n = \arg\min_{q \in \mathcal{D}'} D(q \,\|\, p(z|\boldsymbol{\eta}^n)) \text{ and } \boldsymbol{\eta}^{n+1} = \arg\min_{\boldsymbol{\eta} \in \mathcal{M}} D(q^n \,\|\, p(z|\boldsymbol{\eta})). \tag{3.2}$$

While in EM_λ, the density $\tilde{q}^n$ after the mixing step always resides in $\mathcal{D}'$, it is not the one that minimizes the distance to $\mathcal{D}'$ as in *em*. Therefore in principle EM_λ is suboptimal.

However, unfortunately the set $\mathcal{D}'$ is too rich with respect to the conditional part of its elements, because $\mathcal{D}$ itself consists of the marginal $\hat{p}^I$ combined with any conditional. As a consequence, if $q^{n\prime} = p(z|y, \boldsymbol{\eta}^n)[(1-\lambda)p_X^C(x) + \lambda\hat{p}^I(x)]$ is a member of $\mathcal{D}'$, which is the case for moderately large λ, the *e*-projection of $p(z|\boldsymbol{\eta}^n)$ to $\mathcal{D}'$ will be $q^{n\prime}$. Therefore the *e*-projection would not depend at all on conditional complete

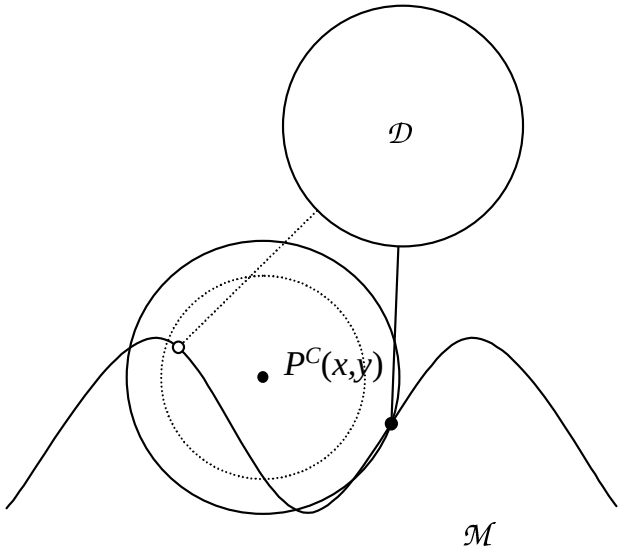

Figure 3-2: Discontinuity introduced by increasing source allocation

evidence, even if $\mathcal{D}'$ does. Thus *em* is not relevant for mixing incomplete and complete information. It seems that EM_λ that forcefully moves the E-step density towards complete evidence make better use of that evidence.

3.2 Instability of Maximum Likelihood with respect to Allocation

The interpretation of maximum likelihood with incomplete information as minimizing distance to the data manifold while bounding distance to the complete empirical estimate permits us to justify that small changes in allocation may result in large changes in the estimate. As in Figure 3-2, $\mathcal{M}$ is not convex and usually features many local minima of the distance to $\mathcal{D}$. The complete evidence $\hat{p}^C$ is close to one of the local minima but far from the global, and while relaxing the bound of distance to $\hat{p}^C$ constraint with increasing λ, the estimation moves from the local to the global minimum. Moreover, this move features discontinuities, because as we can see from the figure, transitions from one convex region to another are sudden.

Bringing more intuition, the local minimum of the distance between $\mathcal{D}$ and $\mathcal{M}$

closest to $\hat{p}^C$ is *related* to complete evidence, while the global minimum or other local minima are *unrelated*. Therefore, estimates of the conditional for allocations larger than values that pass through sudden changes must be unreliable, because the jump removes the dependency on complete evidence. Small changes in the conditional evidence do not reflect changes in estimation after the jump, thus the estimation of the conditional becomes unstable.

We have seen that normal maximum likelihood allocates between distances to complete and incomplete data in the fixed ratio of the number of incomplete to total samples. Since the interesting problem is when the number of incomplete samples dominates significantly, typically maximum likelihood allocation is close to 1. Therefore maximum likelihood focuses on an accurate estimate of the marginal, and less accurate conditional. Thus maximum likelihood assigns too large a weight to incomplete data for no theoretically grounded reason. Moreover, if sudden changes in estimation are present for smaller allocation as in Figure 3-2, the maximum likelihood estimate will be unrelated to complete evidence, and thus it offers no guarantee on the stability of estimation of the conditional.

Therefore it is important to allow for unrestricted source allocation in combining the likelihoods of complete and incomplete data. Fixing allocation to any single value may lead to suboptimal estimation – unstable if it is too large, or without using the potential of incomplete data if it is too small. Our aim is to find the maximum allocation such that the estimate is stable and remains related to complete evidence. The algorithm will consider a range of allocations rather than focusing on a single one.

Chapter 4

Stable Estimation with Path Following

4.1 Motivation

Standard maximum likelihood algorithms, that combine complete and incomplete data by simply summing the log-probability of each sample, bias estimation towards an accurate marginal over the observed variables at the expense of the conditional density of unobserved variables. More incomplete samples should only improve and not hurt estimation, otherwise we would be better off ignoring them. But this is not the case with maximum likelihood, because when the number of complete samples is much smaller than available incomplete data, the complete evidence becomes insignificant for the purpose of likelihood maximization. In the limit of infinite incomplete information, the maximum likelihood estimate does not depend at all on complete evidence. This tendency of maximum likelihood to focus on the marginal and ignore the conditional contradicts the typical goal of estimation with limited complete and abundant incomplete sources of information, where the variable of interest is usually the missing component from incomplete samples.

One may argue that maximizing the likelihood of incomplete data, and than using the small number of complete samples only to estimate missing components without modifying the estimate from incomplete data is a legitimate strategy. After all, in

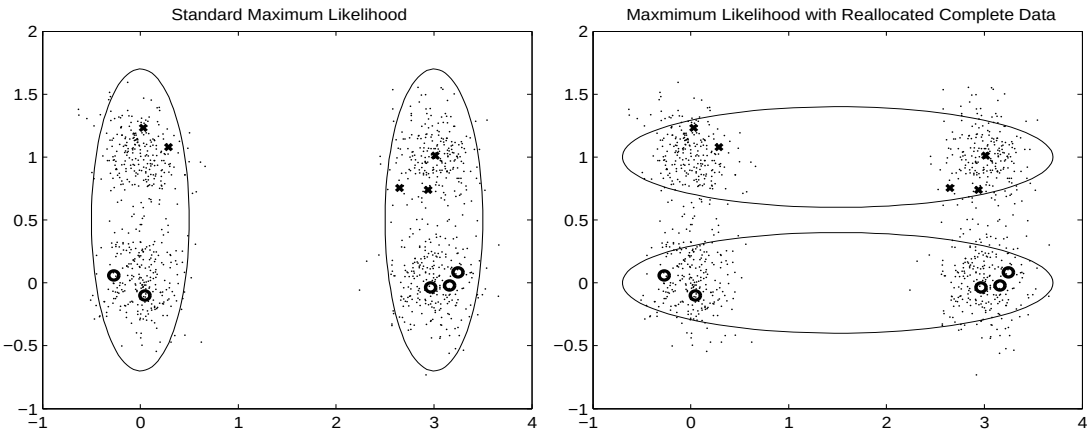

Figure 4-1: Standard maximum likelihood overvalues abundant incomplete data to the expense of classification.

many classification problems unlabeled data alone can correctly estimate the model up to a permutation of the labels, and any such permutation achieves maximum likelihood. In such situations the role of unlabeled and labeled data is clearly separable, with labeled data being responsible only for estimating the label permutation.

However, not all model families are identifiable with incomplete data, and even if they are, when empirical data violates model assumptions maximum likelihood of incomplete data may conflict with conditional estimation. For example, Figure 4-1 presents a binary classification problem in which the model family of each class is bivariate Gaussian. There are only 10 labeled samples, therefore the 1000 unlabeled dominate the likelihood, such that the maximum likelihood solution (left) is incompatible with the classification problem. The intended solution that achieves better classification is the one to the right, which is only a local maxima of the likelihood. However, if we allocate more weight to the labeled samples in the likelihood to offset the number disproportion, the solution to the right becomes a global maximum of the reallocated likelihood.

Therefore a weighted likelihood such as (2.8) is more suited for estimation with incomplete information, because it can assign a significant weight to complete data even if the number of incomplete samples is very large. But the question of what smaller source allocation to use is difficult, and should not be answered for a model family in general, but in a case-by-case data-dependent manner.

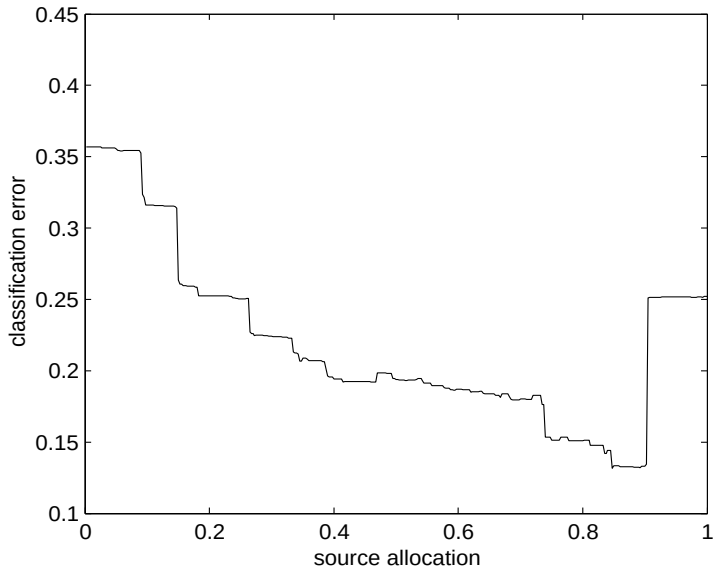

Figure 4-2: EM classification error for different weightings of complete and incomplete information in a naïve Bayes model with 20 binary features and 3 classes.

The important observation is that the change from the incorrect to the correct local maximum is sudden and occurs at a well-defined source allocation. While initially the left configuration of Figure 4-1 achieves maximum likelihood, as we decrease the weight of the incomplete source, the right configuration becomes more and more likely. Then the two configurations become equally likely at a given source allocation, and for smaller allocations only the right picture achieves global maximum likelihood. The geometrical considerations in Chapter 3 explain the sudden change in parameters with varying allocation.

Such discontinuity in maximum likelihood parameters can be observed empirically. In Figure 4-2 we show the error rates at different source allocations in a naïve Bayes classification problem (20 binary features, 3 classes, 10 labeled, 1000 unlabeled samples), where for each allocation we have run the EM algorithm initialized to complete evidence. The sudden change in parameters corresponds to a dramatic decrease in performance.

Assuming a large number of incomplete samples in comparison with available complete data, the allocation at which the parameter configuration switches should also be near optimal. All estimates with smaller allocation optimize the correct configu-

ration, therefore it is safe to use an allocation that maximizes the use of incomplete data, without switching to a maximum unrelated to complete evidence. Therefore we advocate the usage of critical allocation for the purpose of mixing complete and incomplete information.

4.2 Path Following

In what follows we describe an algorithm that detects allocations that feature sudden changes in estimation parameters. The focus is towards an intuitive description, while a rigorous numerically stable version of the algorithm will be presented in Chapter 5.

Consider the EM_λ algorithm (2.12) that optimizes maximum likelihood for any given source allocation $\lambda \in [0, 1]$. We define parameter paths to be continuous segments $\boldsymbol{\eta}(\lambda)$ of density parameters that are fixed points of the EM_λ operator for every λ along the path. The allocation λ can be thought as the variable that parameterizes the continuous path. Sudden changes in maximum likelihood parameters are in fact discontinuities, that is allocations beyond which the parameter path cannot be continuously extended.

We can formalize the concept of continuous path extendibility through a differential equation. Differentiating with respect to λ the fixed point condition $\mathrm{EM}_\lambda(\boldsymbol{\eta}(\lambda)) = \boldsymbol{\eta}(\lambda)$, we get:

$$\left[I - \nabla_{\boldsymbol{\eta}(\lambda)} \mathrm{EM}_\lambda(\boldsymbol{\eta}(\lambda))\right] \frac{d\boldsymbol{\eta}(\lambda)}{d\lambda} = \frac{\partial \mathrm{EM}_\lambda}{\partial \lambda}(\boldsymbol{\eta}(\lambda)) \tag{4.1}$$

As long as the transformed Jacobian $I - \nabla_{\boldsymbol{\eta}(\lambda)} \mathrm{EM}_\lambda(\boldsymbol{\eta}(\lambda))$ is invertible, a unique continuous extension of the fixed point $\boldsymbol{\eta}(\lambda)$ exists in both directions of smaller and larger allocation. On the other hand if the transformed Jacobian is singular irregularities such as bifurcations or jump discontinuities must occur (Figure 4-3). Therefore the singularity of the transformed Jacobian detects critical allocations.

Bifurcations of continuous parameter paths may be a serious problem, because while discontinuities indicate sudden changes in parameters, bifurcation are smooth, and choosing the correct branch to follow is difficult. Fortunately, almost never sin-

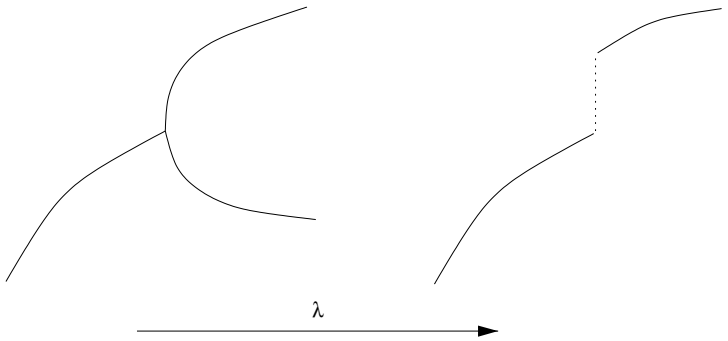

Figure 4-3: Parameter bifurcation or discontinuity at critical allocation.

gularities of the EM algorithm correspond to bifurcations; if bifurcations do emerge, a very small change in complete evidence removes them. The reason for this strong mathematical result will be apparent in Chapter 5.

The path following algorithm proceeds by starting at the complete-only maximum likelihood estimate at $\lambda = 0$, and extends this fixed point continuously by adding small fractions of $d\boldsymbol{\eta}(\lambda)/d\lambda$ as long as the transformed Jacobian remains invertible. Singularities indicate that no further increase in source allocation is possible without introducing a discontinuity, thus they are a stopping condition. Besides numerical consideration in solving the differential equation, the path following algorithm has no convergence issue because at any step the parameters define a fixed point of the EM_λ operator.

Chapter 5

Path Following with Homotopy Continuation

We introduce a powerful method for path following that is numerically stable and does not suffer from bifurcations or discontinuities along the path from $\lambda = 0$ to $\lambda = 1$. In principle, the method permits following through critical allocations, although in practice this feature is irrelevant to our algorithm for stable estimation, that does not need estimates beyond critical allocation. From a practical point of view the method only solves numerical difficulties in finding fixed points, while from a theoretical perspective it brings a simple and insightful description of the set of fixed points of the EM operator.

The method builds on the theory of globally convergent with probability one homotopy continuation, introduced in mathematics by Chow [5]. Watson gives a careful and readable description of this important type of homotopy continuation and its various extensions in [22], and he is also the author of HOMPACK [23], a numerical package for homotopy continuation that implements mature algorithms for continuation. Homotopy continuation ideas have been used before in neural networks [6], but are largely unknown to the machine learning community.

5.1 Theoretical Background

Homotopy continuation is a numerical method for solving general equations that has its roots in mathematical topology. In order to solve a difficult equation $F(x) = 0$, start with a trivial equation $G(x) = 0$ whose solution is known, and track its root while morphing G into F. Formally, given $F, G : E^n \to E^n$, define a homotopy function to be a smooth map $H : E^n \times [0,1] \to E^n$ such that $H(x,0) = G(x)$ and $H(x,1) = F(x)$. The goal becomes to find solutions of $H(x,\lambda) = 0$ for all $\lambda \in [0,1]$.

We can track (x,λ) roots of the homotopy while changing λ by following the differential equation the governs continuous paths of solutions:

$$\nabla_x H(x,\lambda)\frac{dx}{d\lambda} + \frac{\partial}{\partial\lambda}H(x,\lambda) = 0 \tag{5.1}$$

By solving the differential equation with the initial condition $G(x) = 0$ and $\lambda = 0$, we hope to find a continuous path of solutions up to $\lambda = 1$. However, following the path continuously and uniquely is possible if and only if $\nabla_x H(x,\lambda)$ is never singular. This condition is rarely satisfied, and the path of solutions may end prematurely in a critical point with a discontinuity or bifurcation.

Surprisingly, a simple change to the naive method, introduced by Chow in [5], is able to remove all critical points, producing a robust optimization algorithm. In order to consider the modification, we extend the analysis of homotopy to a *family* of trivial equations $G_a(x) = 0$, with $a \in E^m$ an initial point that we are free to choose. The homotopy $H_a(x,\lambda) = H(a,x,\lambda)$ now depends also on the initial point. The new algorithm will have the property that for almost all a there are no critical points on the path from $\lambda = 0$ to $\lambda = 1$.

The main idea is to follow continuous paths of solutions in the extended space (x,λ), parameterized by a new variable s. The differential equation that governs the evolution of solutions becomes:

$$\nabla_{(x,\lambda)} H_a(x,\lambda) \cdot \begin{pmatrix} dx/ds \\ d\lambda/ds \end{pmatrix} = 0 \tag{5.2}$$

which is well-behaved as long as the Jacobian $\nabla_{(x,\lambda)} H_a$ has maximal rank. Theorem 5.1, proved in [5] from *Parametrized Sard's Theorem*, is fundamental to homotopy continuation because it shows that under certain conditions, easy to satisfy in general situations, the Jacobian always has maximal rank.

Theorem 5.1 *Let $H : E^m \times E^n \times [0,1) \to E^n$ be a C^2 map such that the $n \times (m+n+1)$ Jacobian $\nabla_{(a,x,\lambda)} H$ has rank n on the set $H^{-1}(0)$, and $H_a^{-1}(0)$ is bounded for fixed $a \in E^m$. Then $\nabla_{(x,\lambda)} H_a$ has full rank on $H_a^{-1}(0)$ except for possibly on a set of measure 0 of a's.*

Corollary 5.2 *Under the conditions of Theorem 5.1, if $H_a(x,0) = 0$ has a unique solution x_0, there exists a unique continuous path in $H_a^{-1}(0)$ from $(x_0, 0)$ to $\lambda = 1$. All other elements in $H_a^{-1}(0)$ are on closed continuous loops with $0 < \lambda < 1$, or on paths that start and end at $\lambda = 1$.*

5.1.1 The Fixed-Point Homotopy

The most important assumption of Theorem 5.1, the full rank of $\nabla_{(a,x,\lambda)} H$, is trivially satisfied for all fixed-point problems of C^2 functions under the following standard *fixed-point homotopy*:

$$H_a(x,\lambda) = (1-\lambda)(a-x) + \lambda(f(x)-x) \tag{5.3}$$

Therefore we can find fixed points of $f(x)$ by starting at almost any initial point a and continuously following a path of (x,λ) solutions of the homotopy from $\lambda = 0$ to $\lambda = 1$. We follow the path by solving a differential equation with respect to a path parameterization s:

$$[\lambda \nabla_x f(x) - I \quad f(x) - a] \cdot \begin{pmatrix} dx/ds \\ d\lambda/ds \end{pmatrix} = 0 \tag{5.4}$$

Several methods for numerically solving such equation exist [23]. One of the simplest ones is to take sufficiently small steps in the direction of $(dx/ds,\, d\lambda/ds)$. However, note that equation (5.4) determines the derivative along the path only

up to a multiplicative factor. Therefore, in practice we need to choose a specific parameterization of the curve, such as the unit length parameterization. Even in the context a specific parameterization, the direction is still unspecified up to a sign. We can resolve this ambiguity by following the curve in the direction that makes the smallest angle with the previous direction, but more principled methods exist [23].

5.2 Homotopy Continuation of the EM Operator

In this section we apply homotopy continuation results to the problem of finding fixed points of the EM_λ operator. Due to the linearity of the operator with respect to source allocation (2.13), $H_{\boldsymbol{\eta}^C}(\boldsymbol{\eta}, \lambda) = \mathrm{EM}_\lambda(\boldsymbol{\eta}) - \boldsymbol{\eta}$ is exactly the standard fixed-point homotopy (5.3) of EM_1 at initial point $\boldsymbol{\eta}^C$. Moreover, the EM_λ operator is C^2 on the domain of valid parameters (as long as the partition function is C^2), thus we only need boundedness to claim the results of Theorem 5.1.

A necessary condition for the boundedness of the set of fixed points at all allocations is that conditional expectations of sufficient statistics are always bounded:

$$\{\mathrm{E}[\mathbf{t}(z)|\mathbf{x}, \boldsymbol{\eta}] : \forall \boldsymbol{\eta}\} \text{ bounded } \forall \mathbf{x} \in \mathcal{X} \tag{5.5}$$

The above condition is not satisfied by all exponential families, but most interesting cases that occur in practice will satisfy it. In particular, if $\mathcal{Z}(\mathbf{x})$ is discrete for all $\mathbf{x} \in \mathcal{X}$, the conditional expectation becomes a finite convex combination of functions of $\mathbf{x}$, thus it is bounded. This is the case for all classification problems, in which missing information comes from finitely many labels. Many problems with continuous missing data will also satisfy property (5.5). [1]

Therefore all the strong results of homotopy continuation theory naturally apply to EM with incomplete information. In particular, for almost any initial points $\boldsymbol{\eta}^C$ there is a unique continuous path of homotopy solutions from $\lambda = 0$ to $\lambda = 1$,

[1] Even if (5.5) does not hold, it is unlikely that homotopy will fail in an infinite continuous path that never reaches $\lambda = 1$. Moreover, the condition is detectable when it occurs, because in the limit the homotopy matrix will not have full rank.

and there are no discontinuities or bifurcations along the way. Thus all numerical problems of naïve path following are removed by simply performing the continuation in the extended $(\boldsymbol{\eta}, \lambda)$ space.

According to the theory, homotopy continuation amounts to solving the following differential equation

$$\begin{bmatrix} \lambda \nabla_{\boldsymbol{\eta}} \mathrm{EM}_1(\boldsymbol{\eta}) - I & \mathrm{EM}_1(\boldsymbol{\eta}) - \boldsymbol{\eta}^C \end{bmatrix} \cdot \begin{pmatrix} d\boldsymbol{\eta}/ds \\ d\lambda/ds \end{pmatrix} = 0 \tag{5.6}$$

subject to $(d\boldsymbol{\eta}/ds)^2 + (d\lambda/ds)^2 = 1$ and initial condition $(\boldsymbol{\eta}^C, 0)$.

The homotopy matrix can be computed explicitly for exponential families in general. From (2.12) we know that $\mathrm{EM}_1(\boldsymbol{\eta}) = \frac{1}{M} \sum_{1 \le j \le M} \mathrm{E}[\mathbf{t}(z) | \mathbf{x}(z) = x_j^I, \boldsymbol{\eta}]$, and the Jacobian of EM_1 is given by

$$\nabla_{\boldsymbol{\eta}} \mathrm{EM}_1(\boldsymbol{\eta}) = \left[\frac{1}{M} \sum_{1 \le j \le M} \mathrm{cov}(\mathbf{t}(z) | \mathbf{x}(z) = x_j^I, \boldsymbol{\eta}) \right] F(\boldsymbol{\eta})^{-1} \tag{5.7}$$

where cov is the covariance matrix of the vector of sufficient statistics, and F is the Fisher information matrix $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \psi(\boldsymbol{\theta})$. For a derivation of the Jacobian see Theorem A.2 in the Appendix.

Figure 5-1 gives a high level description of simplest realization of homotopy continuation for stable estimation with incomplete information. Note that besides solving the differential equation directly there exist other numerical methods for homotopy continuation [23]. Also, Euler's method only advantage is its simplicity, thus in practice better ODE solvers should be used [18].

5.3 Qualitative Analysis of EM Continuation

5.3.1 General Configuration of Fixed Points

Homotopy theory provides a general characterization of the configuration of fixed points of EM. According to Corollary 5.2, for almost all starting points $\boldsymbol{\eta}^C$, the set of parameters $(\boldsymbol{\eta}, \lambda)$ that are fixed points of the EM_λ operator is composed of only

Input: Complete data $\mathbf{z}^C = (z_1^C, z_2^C, \ldots, z_N^C)$, incomplete data $\mathbf{x}^I = (x_1^I, x_2^I, \ldots, x_M^I)$

Output: Mean parameters $\boldsymbol{\eta}$ at critical allocation

1. Initialize $(\boldsymbol{\eta}, \lambda)$ to $(\boldsymbol{\eta}^C, 0)$ and $(d\boldsymbol{\eta}/ds, d\lambda/ds)_{\text{old}}$ to $(0, 1)$.
2. Compute $\text{EM}_1(\boldsymbol{\eta})$ and $\nabla_{\boldsymbol{\eta}}\text{EM}_1(\boldsymbol{\eta})$.
3. Compute $(d\boldsymbol{\eta}/ds, d\lambda/ds)$ as the kernel of $\begin{bmatrix}\lambda\nabla_{\boldsymbol{\eta}}\text{EM}_1(\boldsymbol{\eta}) - I & \text{EM}_1(\boldsymbol{\eta}) - \boldsymbol{\eta}^C\end{bmatrix}$.
4. Normalize $(d\boldsymbol{\eta}/ds, d\lambda/ds)$ and choose sign such that $(d\boldsymbol{\eta}/ds, d\lambda/ds) \cdot (d\boldsymbol{\eta}/ds, d\lambda/ds)_{\text{old}}^T \geq 0$.
5. Update parameters: $(\boldsymbol{\eta}, \lambda)_{\text{new}} = (\boldsymbol{\eta}, \lambda) + \epsilon \cdot (d\boldsymbol{\eta}/ds, d\lambda/ds)$, $(d\boldsymbol{\eta}/ds, d\lambda/ds)_{\text{old}} = (d\boldsymbol{\eta}/ds, d\lambda/ds)$.
6. Repeat 2 until $\lambda \geq 1$ or $\det(\lambda\nabla_{\boldsymbol{\eta}}\text{EM}_1(\boldsymbol{\eta}) - I) = 0$.

Figure 5-1: Simple homotopy continuation algorithm for stable estimation with incomplete information using Euler's method

the following types of structures:

- a unique continuous path from $\lambda = 0$ to $\lambda = 1$;
- some continuous paths with both endpoints at $\lambda = 1$;
- some continuous loops with all λ's in $(0, 1)$;
- continuous regions with all λ's equal to 1.

Intuitively, because the homotopy differential equation is never degenerate for $\lambda \in (0, 1)$, every fixed point can be extended continuously and uniquely in both directions, therefore the set of fixed points consists of continuous paths only. In addition, the only fixed point at $\lambda = 0$ is $\boldsymbol{\eta}^C$, therefore the path from 0 allocation to one 1 allocation is unique, and there are no paths with both endpoints at $\lambda = 0$. For a generic configuration of the fixed points of EM_λ see Figure 5-2. Note that even if the path from $\boldsymbol{\eta}^C$ to $\lambda = 1$ is always continuous, it is not necessarily monotonic in allocation, and this is the reason why naïve path following, that always increases the allocation, finds discontinuities.

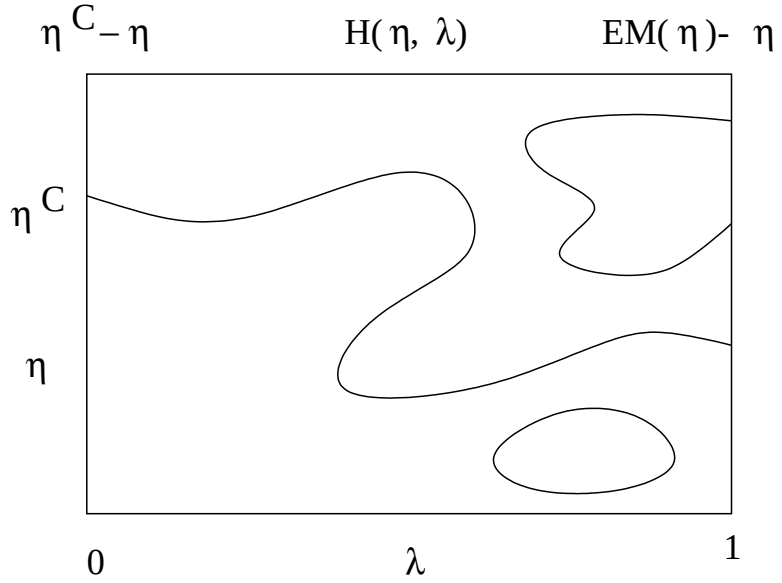

Figure 5-2: Generic configuration of fixed points of the EM_λ operator

Continuous regions entirely within $\lambda = 1$ must be included because according to the theory the homotopy matrix is guaranteed to have maximal rank only for $\lambda < 1$. At $\lambda = 1$, when fixed points do not depend on the initial point, degeneracies may occur. This is probably not likely to happen often, but it is conceivable that entire continuous paths may stay at $\lambda = 1$. Indeed, likelihood with no complete data can achieve its global maximum not only at multiple points, but even at a continuum of points. For instance, this is the case for discrete naïve Bayes models with only two independent features. Nevertheless, as soon as λ is less than 1, linearly combining with $\boldsymbol{\eta}^C$ removes any irregularities.

It is worth mentioning that homotopy continuation can be used to follow any continuous component of the set of fixed points. We may start at a fixed point at $\lambda = 1$ and follow it until we reach 1 allocation again, or 0 allocation if we happen to be on the unique path, or we may trace loops of fixed points. Nevertheless, the only fixed point that we know a priori is $(\boldsymbol{\eta}^C, 0)$, and for a different initialization of continuation other fixed-point finding algorithms must be used in advance. For the purpose of this work we are interested only in the continuous path that relates fixed points to complete evidence, therefore we do not study other initializations.

Even standard EM without complete samples benefits from the properties revealed by homotopy theory because we can construct an EM_λ operator by setting $\boldsymbol{\eta}^C$ to whatever starting point EM is initialized to. The unique path from 0 to 1 allocation looses its meaning in this context, but the connections established through continuation between different local maxima of likelihood at $\lambda = 1$ might prove insightful.

5.3.2 Types of Critical Points

Because Theorem 5.1 is not valid without exception for all initial points, homotopy continuation is not free from irregularities. In this section we give a detailed description of such irregularities so that we can use them to our advantage. We begin with a definition:

Definition Given an initial point and its associated EM_λ operator, $(\boldsymbol{\eta}, \lambda)$ is a *strongly critical point* if $[\lambda \nabla_{\boldsymbol{\eta}} \mathrm{EM}_1(\boldsymbol{\eta}) - I \quad \mathrm{EM}_1(\boldsymbol{\eta}) - \boldsymbol{\eta}]$ does not have maximal rank, and λ for which this happens is a *strongly critical allocation*. An initial point $\boldsymbol{\eta}^C$ is *strongly critical* if the continuous path of fixed points starting at $\boldsymbol{\eta}^C$ contains a strongly critical point.

Strongly critical points are places where bifurcations or discontinuities may occur. Existence of such points may hinder continuation, but fortunately they may occur only for a measure 0 set of initial points. Therefore, almost never the initial point is strongly critical, and if it is, a small perturbation removes the critical character.

Why is the existence of strongly critical initial points unavoidable? To answer this question consider an identifiable model, or any model for which likelihood can have only finitely many local maxima. Homotopy continuation establishes a mapping between non-critical initial points and fixed points of the EM_1 operator. This is a mapping between a continuous set to a discrete set, therefore continuation partitions the set of all non-critical initial points into continuous regions that map to the same fixed point. The boundaries of these regions realize a sudden transition between fixed points that can happen only if boundaries contain strongly critical initial points.

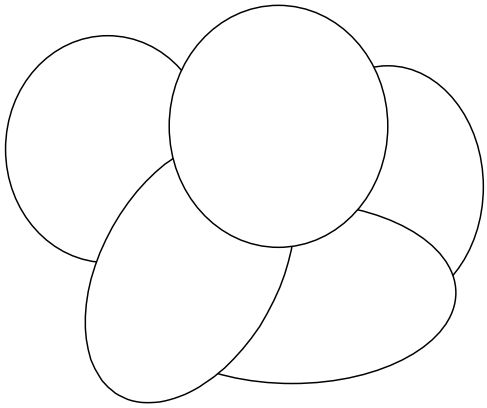

Figure 5-3: Initial points partitioned into regions of convergence bounded by strongly critical points.

Figure 5-3 pictures a typical partition of initial points into regions of convergence, in which the boundaries which are strongly critical are of course of measure 0.

Misestimation occurs when the empirical initial point $\boldsymbol{\eta}^C$ and the true parameters $\boldsymbol{\eta}^*$ of the m-projection of the true distribution to $\mathcal{M}$ fall in different regions of convergence. While no criteria can verify membership to the same region of convergence, jumps in path following can detect that the initial point is not consistent with the marginal, and limit the extent to which the final estimate falls into a distant region.

In order to achieve stable estimation the true initial point $\boldsymbol{\eta}^*$ cannot be on a boundary. Otherwise even with abundant complete evidence the estimate may be easily biased towards distant fixed points. Theoretically we can guarantee that if $\hat{p}^I$ is the true marginal and is an element of $\mathcal{M}_X$, a small region around the true $\boldsymbol{\eta}^*$ contains no strongly critical initial point.

To illustrate the effect of the initial point being close to a strongly critical value consider an operator that has exactly three fixed points as in Figure 5-4 (for instance $f(x) = x^3$ with fixed points -1, 0, 1, and strongly critical initial point at 0). A path splits into branches at a strongly critical point, and each branch leads to a different fixed point of the operator at $\lambda = 1$. If we perturb $\boldsymbol{\eta}^C$ by a small amount, the path proceeds without any criticality from $\lambda = 0$ to $\lambda = 1$, and the fixed points that were on other branches become connected through continuous paths starting and ending at $\lambda = 1$. But depending on the direction of the small perturbation of the starting point,

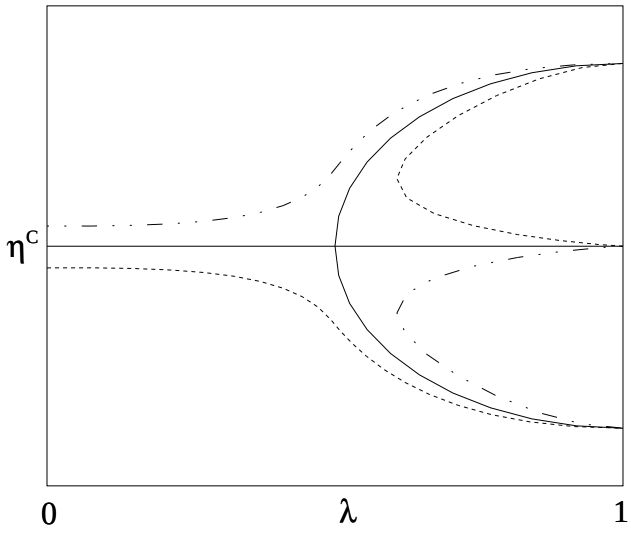

Figure 5-4: Dynamics of fixed point paths as the initial point passes through a strongly critical value. Similar line styles correspond to the same initial point.

the final fixed point at $\lambda = 1$ may end up on any of the initial branches. Therefore even if the path is not critical, it is not stable at $\lambda = 1$ because a small perturbation in the initial point results into a large change at $\lambda = 1$. The further the initial point is from the boundary, the more room there is for a poor complete-date estimate, and the more stable the estimation.

If strongly critical points are unlikely to occur, other types of critical situations are more frequent:

Definition Given an initial point and its associated EM_λ operator, $(\boldsymbol{\eta}, \lambda)$ is a *weakly critical point* if $\lambda \nabla_{\boldsymbol{\eta}} \mathrm{EM}_1(\boldsymbol{\eta}) - I$ is singular, and λ for which this happens is a *weakly critical allocation.*

Weakly critical points include not only all strongly critical points, but also places where λ must change its monotonicity in order to continue following the path. In short, weakly critical points are the places at which continuity would break had we tried to follow the $\boldsymbol{\eta}$ path naively under its λ parameterization.

Although less obvious, weakly critical points also have a negative impact of stability of estimation. From the geometrical considerations in Chapter 3, weakly critical

allocation is the place where a sudden change in parameters marks a shift from one convex region to another convex region of $\mathcal{M}$. For larger allocations the fixed point looses its connection to $\boldsymbol{\eta}^C$, and moves away rapidly from $\boldsymbol{\eta}^C$ and possible $\boldsymbol{\eta}^*$. While the estimate after critical allocation in some instances may be better, we observe a sudden increase in variance of estimation, compromising stability.

5.3.3 Relationship of Continuation to Standard EM

Both EM and homotopy continuation find fixed points of the same operator. While EM offers no guarantees on the outcome of the estimation, homotopy continuation follows a preferential fixed point which is related to complete evidence. The question is whether in practice the two local optima will indeed be different.

From a theoretical point of view, one can establish that when allocation is in a small region near 0, EM_λ initialized to $\boldsymbol{\eta}^C$ must converge to the fixed point found by continuation. The reason is that the uniqueness of fixed points at $\lambda = 0$ and the continuous-path structure prove that in a region near $\lambda = 0$ there are no multiple fixed points at the same allocation. However, the region may be quite small, and beyond it theory does not guarantee similarities between EM and continuation.

Empirically, as it will emerge from Chapter 7, EM_λ started at $\boldsymbol{\eta}^C$ and homotopy continuation follow each other for moderate allocation, but at larger allocation may diverge from each other without regard to weakly critical λ's. When EM_λ and continuation follow the same fixed point up to critical allocation, often EM_λ and continuation give the same estimate for slightly larger allocation. Of course, continuation must first decrease λ and increase it again in order to follow the path, but surprisingly when it reaches slightly larger than critical allocation it meets again the solution found by EM. Note that we never perform true path following with standard EM_λ, but for each value of λ we run it separately starting from $\boldsymbol{\eta}^C$. In conclusion, EM_λ and homotopy continuation tend to stay together for moderate allocation, but diverge unpredictably for larger; path following indeed finds a preferential fixed point compared to EM_λ.

5.4 Alternate Homotopy of the EM Operator

To illustrate the power of the generality of homotopy continuation, we exhibit a different homotopy of the EM_λ operator useful for stability analysis. Fixing the λ allocation between complete and incomplete information, we want to characterize how the fixed point reached by continuation changes while the initial point changes. $\boldsymbol{\eta}^C$ is very noisy because it is an empirical estimate from limited complete data, and such analysis brings insight into the impact of uncertainty in $\boldsymbol{\eta}^C$ over uncertainty in estimation.

The goal of alternate homotopy is to track fixed points while changing the initial point and holding λ constant. For this purpose we introduce a second allocation parameter $\mu \in [0, 1]$ and look for fixed points $(\boldsymbol{\eta}, \mu)$ of the following linear combination:

$$F(\boldsymbol{\eta}) = (1-\lambda)[(1-\mu)\boldsymbol{\eta}_1^C + \mu\boldsymbol{\eta}_2^C] + \lambda \mathrm{EM}_1(\boldsymbol{\eta}) \tag{5.8}$$

assuming that we are given a starting fixed point at $\mu = 0$.

The fixed-point homotopy does not apply in this situation, but we can still use the general homotopy theory by defining the following homotopy function:

$$\begin{aligned} H_{\boldsymbol{\eta}_1^C}(\boldsymbol{\eta}, \mu) = (1-\mu)[(1-\lambda)\boldsymbol{\eta}_1^C + \lambda \mathrm{EM}_1(\boldsymbol{\eta}) - \boldsymbol{\eta}] + \\ \mu[(1-\lambda)\boldsymbol{\eta}_2^C + \lambda \mathrm{EM}_1(\boldsymbol{\eta}) - \boldsymbol{\eta}] \end{aligned} \tag{5.9}$$

In order to apply the theory, we need to verify that the conditions of Theorem 5.1 are satisfied. The most important condition, $\nabla_{(\boldsymbol{\eta}_1^C, \boldsymbol{\eta}, \mu)} H$ has full rank at all fixed points, is trivially satisfied because the following square submatrix of maximal dimension

$$\nabla_{\boldsymbol{\eta}_1^C} H_{\boldsymbol{\eta}_1^C}(\boldsymbol{\eta}, \mu) = (1-\lambda)(1-\mu) I \tag{5.10}$$

is always invertible.

Therefore tracing fixed points with homotopy continuation while changing the initial point is possible without reaching bifurcations or discontinuities. However, μ need not be monotonic along the path, and the fixed points found at $\mu = 1$ are not

necessarily on the unique continuous path of fixed points that passes through $\boldsymbol{\eta}_2^C$. Another problem is that at $\mu = 0$ fixed points other than $\boldsymbol{\eta}_1^C$ may exist, therefore the continuous path may return to one of those points and never reach $\mu = 1$.

The importance of the alternate homotopy is only theoretical as it provides insight into stability of fixed points found by homotopy while the starting point varies.

Chapter 6

Specific Models

We illustrate the application of homotopy continuation for mixing complete and incomplete information to a number of specific exponential families. In all cases data decomposes into unobserved and observed components $z = (x, y)$, so that (x, y) are complete and x incomplete samples. The missing information y is the class label, and the class description $p(x|y)$ is model specific.

6.1 Discrete Naïve Bayes

We detail homotopy continuation on a naïve Bayes model, in which observed data decomposes into features independent given class. Both the class label and features are discrete. Discrete naïve Bayes is one of the simplest graphical model, and yet is applicable to many practical problems, such as document classification.

Formally, observed data $\mathbf{x} \in \mathcal{X}$ consists of features $(x_1, x_2, \ldots, x_k) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k$, and the model family $\mathcal{M}$ contains probabilities that satisfy the naïve Bayes independence assumption:

$$P(\mathbf{x}, y) = \left[\prod_{i=1}^{k} P_i(\mathbf{x}_i|y)\right] P_Y(y) = \left[\prod_{i=1}^{k} P_i(\mathbf{x}_i, y)\right] P_Y(y)^{1-k}, \forall \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y} \qquad (6.1)$$

Assessing the modeling power of naïve Bayes is important for understanding the impact of incomplete data in label estimation. The more features, the more restricted

is the model family. Given enough features, the model is identifiable, in the sense that $P_X(\mathbf{x})$ determines $P(\mathbf{x}, y)$ up to a label permutation. If features and classes are all binary, three features suffice for identifiability [11]. In other settings, the exact number of features needed for identifiability is unclear. Also, the more class labels, the less restricted the model is: if features are binary and there are more than 2^k labels, $\mathcal{M}$ contains all distributions. In practical applications to document classification the number of classes is small and there is a large number of binary features, thus the model is probably identifiable.

Naïve Bayes, like any discrete graphical model, is in the exponential family [11]. In order to exhibit explicit parameters and sufficient statistics, let's characterize the label set and feature sets completely:

$$\mathcal{Y} = \{y_0, y_1, \ldots, y_Y\} \text{ and } \mathcal{X}_i = \{x_0^{(i)}, x_1^{(i)}, \ldots, x_{K_i}^{(i)}\}, 1 \leq i \leq k$$

Discrete naïve Bayes with positive probabilities can be put into exponential form (2.2) with Kronecker delta sufficient statistics: $\{\delta_{(x_j^{(i)}, y)}\}_{1 \leq i \leq k, 1 \leq j \leq K_i, y \in \mathcal{Y}}$ and $\{\delta_{y_i}\}_{1 \leq i \leq Y}$. Note that symbols of the type $\delta_{(x_0^{(i)}, y)}$ and δ_{y_0} have been omitted, otherwise the parameterization would be over-complete. The expectation of sufficient statistics produces the mean parameterization of naïve Bayes:

$$\{P_Y(y_i)\}_{1 \leq i \leq Y}, \{P_i(x_j^{(i)}, y)\}_{1 \leq i \leq k, 1 \leq j \leq K_i, y \in \mathcal{Y}} \tag{6.2}$$

Under this parameterization the EM_λ operator is:

$$\begin{array}{rcl} P_Y(y) & \leftarrow & (1-\lambda)P^C(y) + \lambda \sum_{\mathbf{x} \in \mathcal{X}} P^I(\mathbf{x}) P(y|\mathbf{x}) \\ P_i(x, y) & \leftarrow & (1-\lambda)P^C(x, y) + \lambda \sum_{\mathbf{x} \in \mathcal{X}, \mathbf{x}_i = x} P^I(\mathbf{x}) P(y|\mathbf{x}) \end{array} \tag{6.3}$$

where

$$P(y|\mathbf{x}) = \frac{\left[\prod_{i=1}^k P_i(\mathbf{x}_i, y)\right] P_Y(y)^{1-k}}{\sum_{0 \leq j \leq Y} \left[\prod_{i=1}^k P_i(\mathbf{x}_i, y_j)\right] P_Y(y_j)^{1-k}} \tag{6.4}$$

and P^C and P^I are empirical frequencies computed from complete and incomplete

samples respectively.

Mean parameters must satisfy certain validity constraints: they must be positive, $\sum_{1\leq i\leq Y} P_Y(y_i) < 1$, and $\sum_{1\leq j\leq K_i} P_i(x_j^{(i)}, y) < P_Y(y)$. Because we perform homotopy continuation in an unconstrained fashion, we must ensure these conditions are automatically satisfied.

Positivity is the most important condition, because it is necessary for the differentiability of $P(y|\mathbf{x})$, thus of the EM_λ operator. We show that all fixed points along the unique continuous path with $\lambda \in [0, 1)$ have positive coordinates. Assume this is not the case. Let $\lambda_0 \in (0, 1)$ be the minimum allocation on the continuous path starting at P^C for which the positivity constraint is violated. Because of continuity, the fixed point at λ_0 must have some 0 coordinates, but no negative coordinate yet. Therefore for $\lambda < \lambda_0$, fixed points on the curve have positive coordinates, and some of them approach 0 when $\lambda \to \lambda_0$. But from (6.3), if all coordinates of a fixed point are positive, they all must be greater than $(1-\lambda)\min P^C(\mathbf{x}, y)$. In practice we use smoothed P^C so that the previous minimum is positive, therefore no coordinate can converge to 0. In conclusion, homotopy continuation cannot produce non-positive coordinates for $\lambda < 1$. The other constraints on mean parameters are automatically satisfied for any positive fixed point given (6.3), because by just summing up the equations we get $\sum_{y\in\mathcal{Y}} P_Y(y) = 1$ and $\sum_{x\in\mathcal{X}_i} P_i(x, y) = P_Y(y)$.

It remains to give an explicit formula for the homotopy matrix, or more exactly $\nabla_{\boldsymbol{\eta}}\mathrm{EM}_\lambda$. Rather than using (5.7), we compute the Jacobian directly. It is enough to produce the derivatives of $P(y|\mathbf{x})$ with respect to all parameters, and then we can use (6.3) to construct the Jacobian:

$$\begin{aligned}
\frac{\partial P(y|\mathbf{x})}{\partial P_Y(y')} &= (k-1)\left[\frac{P(y|\mathbf{x})P(\mathbf{x}|y')}{P_X(\mathbf{x})} - \delta_{y'}(y)\frac{P(\mathbf{x}|y')}{P_X(\mathbf{x})}\right] \\
\frac{\partial P(y|\mathbf{x})}{\partial P_i(x, y')} &= \delta_x(\mathbf{x}_i)\left[-\frac{P(y|\mathbf{x})P(\mathbf{x}\setminus\mathbf{x}_i|y')}{P_X(\mathbf{x})} + \delta_{y'}(y)\frac{P(\mathbf{x}\setminus\mathbf{x}_i|y')}{P_X(\mathbf{x})}\right] \qquad (6.5)
\end{aligned}$$

where $P(\mathbf{x}\setminus\mathbf{x}_i|y')$ is marginalization with respect to i'th feature.

6.2 Mixture of Multivariate Normals

Consider a classification problem in which we model each class as a multivariate normal distribution. Therefore the joint model of observed data and class label is a mixture of multivariate normals. Formally, if $\mathcal{Y} = \{y^0, y^1, \ldots, y^Y\}$ is the set of labels, and d the dimensionality of data, $\mathcal{M}$ consists of densities of the type

$$p(\mathbf{x}, y) = \pi_y \mathcal{N}(\mathbf{x}|y) = \pi_y \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_y|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \mathbf{\Sigma}_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right] \tag{6.6}$$

where $\boldsymbol{\mu}_{y^i}$, $\mathbf{\Sigma}_{y^i}$, and π_{y^i}, $0 \leq i \leq Y$, are the means, covariances, and mixing coefficients of the mixture.

Mixtures of normal distributions form an exponential family. Simple computations show that (6.6) can be put into exponential form with the following sufficient statistics: $\{\delta_{y^i}(y), \delta_{y^i}(y)\mathbf{x}, \delta_{y^i}(y)\mathbf{x}\mathbf{x}^T\}_{1\leq i\leq Y}$, $\mathbf{x}$, and $\mathbf{x}\mathbf{x}^T$. Taking expectations of sufficient statistics we obtain the following mean parameterization of a mixture of normal distributions: $\{\pi_{y^i}, \pi_{y^i}\boldsymbol{\mu}_{y^i}, \pi_{y^i}\mathbf{S}_{y^i}\}_{1\leq i\leq Y}$, and $\sum_{0\leq i\leq Y} \pi_{y^i}\boldsymbol{\mu}_{y^i}, \sum_{0\leq i\leq Y} \pi_{y^i}\mathbf{S}_{y^i}$, where $\mathbf{S}_y = \mathrm{E}_{p(\mathbf{x}|y)}[\mathbf{x}\mathbf{x}^T] = \mathbf{\Sigma}_y + \boldsymbol{\mu}_y\boldsymbol{\mu}_y^T$. However, we will use a simpler equivalent parameterization:

$$\{\pi_{y^i}\}_{1\leq i\leq Y}, \text{ and } \{\pi_{y^i}\boldsymbol{\mu}_{y^i}, \pi_{y^i}\mathbf{S}_{y^i}\}_{0\leq i\leq Y} \tag{6.7}$$

where π's are positive of sum less than 1, and $\pi\mathbf{S}$'s positive definite.

Given complete samples $(\mathbf{x}_i, y_i)_{1\leq i\leq N}$ and incomplete samples $(\mathbf{x}'_j)_{1\leq j\leq M}$, the EM_λ algorithm mixing the two sources of information has the following updates:

$$\pi_y = (1-\lambda)\frac{1}{N}\sum_{1\leq i\leq N} \delta_y(y_i) + \lambda\frac{1}{M}\sum_{1\leq j\leq M} p(y|\mathbf{x}'_j) \tag{6.8}$$

$$\pi_y\boldsymbol{\mu}_y = (1-\lambda)\frac{1}{N}\sum_{1\leq i\leq N} \delta_y(y_i)\mathbf{x}_i + \lambda\frac{1}{M}\sum_{1\leq j\leq M} p(y|\mathbf{x}'_j)\mathbf{x}'_j \tag{6.9}$$

$$\pi_y\mathbf{S}_y = (1-\lambda)\frac{1}{N}\sum_{1\leq i\leq N} \delta_y(y_i)\mathbf{x}_i\mathbf{x}_i^T + \lambda\frac{1}{M}\sum_{1\leq j\leq M} p(y|\mathbf{x}'_j)\mathbf{x}'_j\mathbf{x}'^T_j \tag{6.10}$$

Here the mean parameterization is crucial for the linearity in allocation of the operator. Nevertheless, if we use the standard parameterization (means, covariances, mixing), we can still derive a valid homotopy continuation algorithm, as the regularity properties of linear homotopy translate under a reparameterization. Because taking gradients with respect to mean parameters (or computing Fisher information) is more involved from the point of view of symbolic computation, we detail homotopy continuation for the standard parameterization.

To apply homotopy continuation we need the homotopy matrix, therefore explicit formulas for ∇EM. Derivatives with respect to π_y, $\boldsymbol{\mu}_y$, and $\boldsymbol{\Sigma}_y$ can be computed from the EM updates once we know the derivatives of $p(y|\mathbf{x})$. These can be obtained from the derivatives of the full joint

$$\begin{aligned}
\frac{\partial p(\mathbf{x}, y)}{\partial \pi_y} &= \frac{1}{\pi_y} p(\mathbf{x}, y) - \frac{1}{\sum_{y'} \pi_{y'}} p(\mathbf{x}, y) && (6.11)\\
\frac{\partial p(\mathbf{x}, y)}{\partial \boldsymbol{\mu}_y} &= \frac{1}{2} p(\mathbf{x}, y) \left[\boldsymbol{\Sigma}_y^{-1} + (\boldsymbol{\Sigma}_y^{-1})^T\right] (\mathbf{x} - \boldsymbol{\mu}_y) && (6.12)\\
\frac{\partial p(\mathbf{x}, y)}{\partial \boldsymbol{\Sigma}_y} &= -\frac{1}{2} p(\mathbf{x}, y) \left[(\boldsymbol{\Sigma}_y^{-1})^T - (\boldsymbol{\Sigma}_y^{-1})^T (\mathbf{x} - \boldsymbol{\mu}_y)(\mathbf{x} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1}\right] && (6.13)
\end{aligned}$$

using the following formula:

$$\frac{\partial p(y|\mathbf{x})}{\partial t} = \frac{1}{p_X(\mathbf{x})} \frac{\partial p(\mathbf{x}, y)}{\partial t} - \frac{p(y|\mathbf{x})}{p_X(\mathbf{x})} \sum_{y'} \frac{\partial p(\mathbf{x}, y')}{\partial t} \tag{6.14}$$

For numerical reasons, in practice it is better to work with logarithms $\tilde{\pi}_y = \log \pi_y$ rather than the true mixing coefficients:

$$\frac{\partial p(\mathbf{x}, y)}{\partial \tilde{\pi}_y} = p(\mathbf{x}, y) - \frac{\pi_y}{\sum_{y'} \pi_{y'}} p(\mathbf{x}, y) \tag{6.15}$$

The above derivative has the right scale. Moreover, when working with logarithms it is trivial to show that the positiveness of mixing coefficients is maintained along the path.

Chapter 7

Experiments

We implement homotopy continuation for finding fixed points of the EM algorithms on mixtures of multivariate normals, and discrete naïve Bayes families. To solve the differential equation we use a second order Runge-Kutta algorithm with a small step size. The computationally dominant step is solving equation (5.4) by a QR factorization of the homotopy matrix. Due to this operation the algorithm is $O(n^3)$ in the dimension of the homotopy matrix, i.e. the total number of parameters.

7.1 Multivariate Normal Models

We analyze the performance of homotopy continuation for estimation with complete and incomplete data under Gaussian model assumptions on artificially generated binary classification problems. Because of the high variability due to limited complete data, we run each experiment many times with randomly selected labeled and unlabeled samples from generated data. In homotopy continuation, we extend a continuous path of fixed points started from complete evidence until we reach a weakly critical allocation, or maximum-likelihood allocation, whichever is the minimum. Because maximum likelihood allocation is typically almost 1, stopping a path with no critical points at maximum-likelihood allocation or at 1 makes little difference. Also, if the path features no critical points EM and homotopy continuation have equal performance in practice, therefore we compare EM and continuation mostly on the runs

with critical points.

As expected, weakly critical allocation occurs when there is disagreement between labeled and unlabeled evidence, either because unlabeled data violates model assumptions, or because unlabeled data is so limited that the complete-data only estimate is far from the truth. To emphasize the impact of homotopy continuation, we choose data generated mostly from models that disagree with the family used by the algorithm, but we also show an instance in which data is generated from exactly the model family of the algorithm.

In Figure 7-1 we show homotopy continuation and EM runs on three different cases, each generated from four Gaussian clusters. In the first two cases one class consists of three of the clusters, and the other class of the remaining cluster, while in the third case two clusters are assigned to each class and classes cross each other. In each case we show a plot of the classification error of homotopy and EM in terms of the complete-data only classification error for each run. Points above the main diagonal are to be avoided, because in such cases addition of unlabeled data hurts classification. The plots feature only runs in which homotopy stopped at a critical allocation, because in the rest of the runs EM and continuation achieve the same classification error. In each case we performed 500 runs with 20 labeled and 200 unlabeled randomly selected samples, and critical allocation appeared in 20%, 27%, and 24% of the runs, respectively.

In the first experiment we can see from Figure 7-1 that EM dramatically increases the error when the labeled-data only estimate is accurate, and it reduces it when the initial error is large. Homotopy is able to pick up in the instances that feature critical allocation all situations in which EM dramatically increases the error. All critical homotopy runs are below or not far above the main diagonal. However, stopping at critical allocation is more conservative than EM, not reducing the error by as much as EM when the initial error is high to begin with.

The second experiment differs from the first only in that clusters have higher variance, therefore there is a higher degree of model misfit. The impact is that the percentage of critical homotopy runs is higher (27% compared to 20%), and the

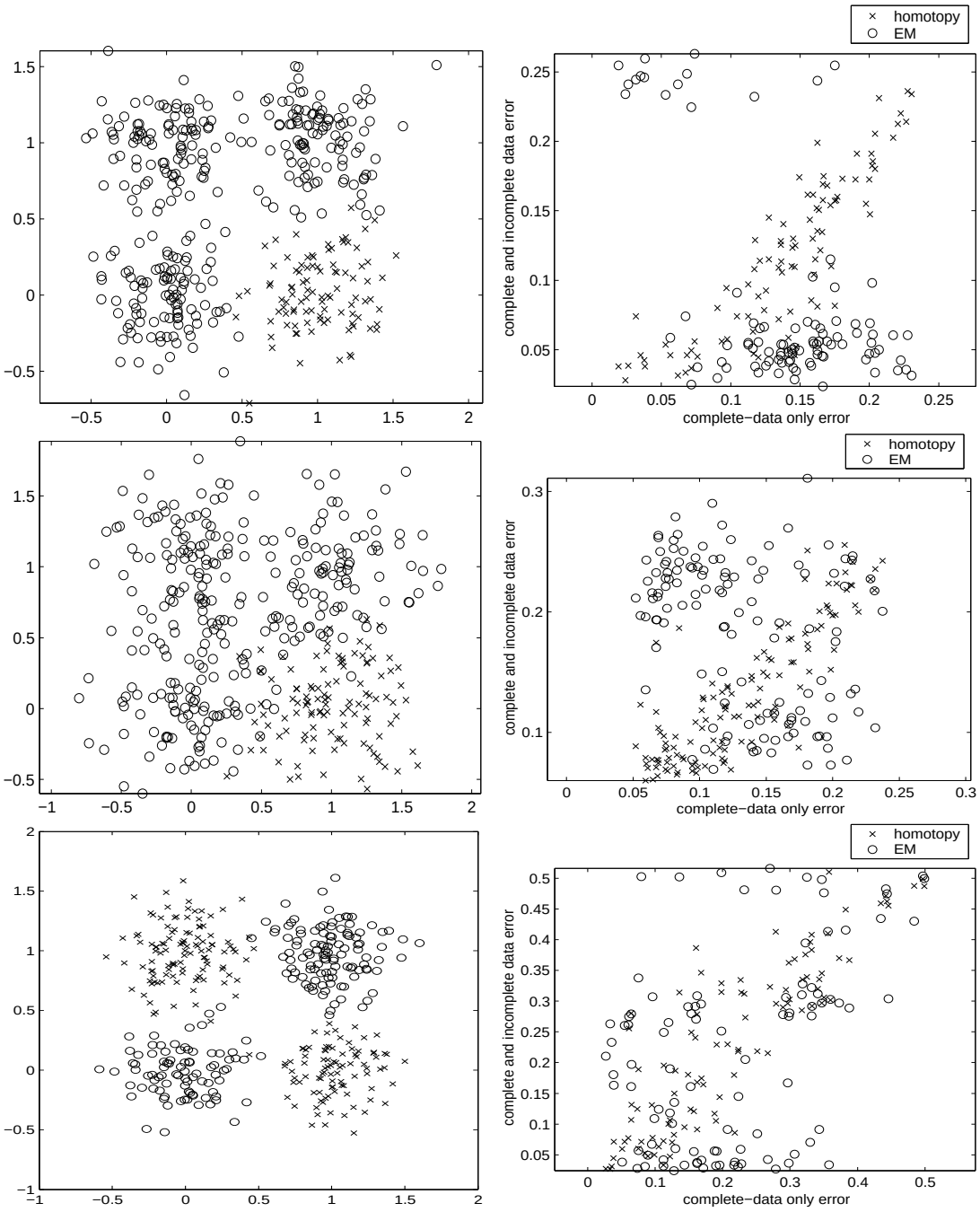

Figure 7-1: Comparison of EM and homotopy continuation on 500 random selections of 20 labeled and 200 unlabeled samples on each of three different problems. Left: complete data from which to choose samples. Right: classification error vs. complete-data only classification error on each run that had a critical allocation.

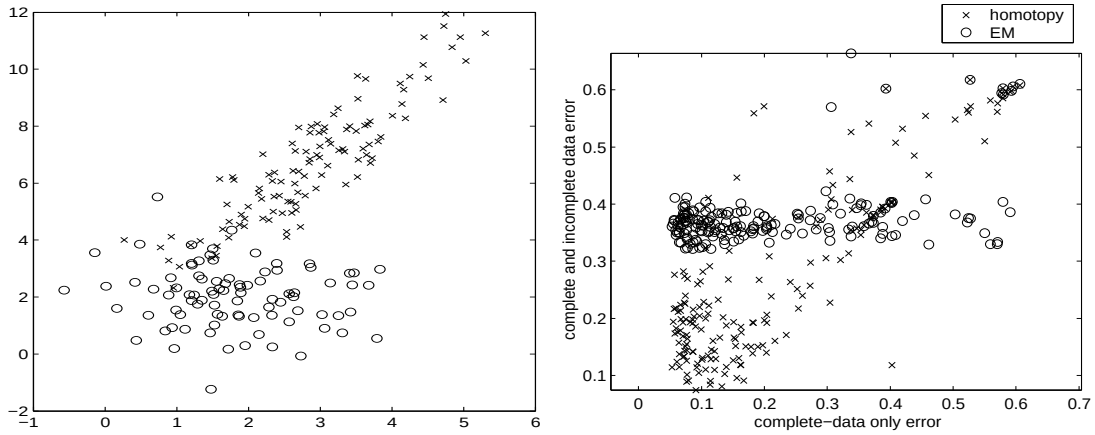

Figure 7-2: EM and homotopy continuation on data that violates the *diagonal* Gaussian class model assumption. Left: complete data from which to choose samples. Right: classification error vs. complete-data only error on 500 random selections of 8 labeled and 1000 unlabeled samples (critical allocation only).

number of runs in which EM fails is also higher, while EM reduces the error by a smaller amount in the better runs. Homotopy is again able to pick up all situations in which EM fails, increasing the gain from the addition of unlabeled data.

Because the performance of EM seems to be very related to the degree of model misfit, we run an experiment in which there is high disagreement between data samples and the model. In Figure 7-2 data comes from two Gaussian clusters with correlated coordinates and with a large overlap. We model each class by naïve Bayes Gaussians, i.e. normals with diagonal covariances. We sample 8 labeled and 1000 unlabeled points in 500 runs, and we achieve a higher percentage of critical homotopies than in the previous experiments, 48%. In this experiment EM fails dramatically with almost all runs increasing error with the addition of unlabeled data, while homotopy is able to not only limit but also reduce the error on the problematic runs.

We can analyze the experiments in even more detail by looking at what happens to the decision boundary while following a path of fixed points with homotopy continuation. In Figure 7-3 we see a typical run in which homotopy outperforms EM on the data from Figure 7-2. The left graph shows the evolution of allocation along the path with each iteration. Weakly critical points occur at local optima of the graph, with the first one at $\lambda = 0.11$, and sharp subsequent ones at 0.55 and 0.17. Although

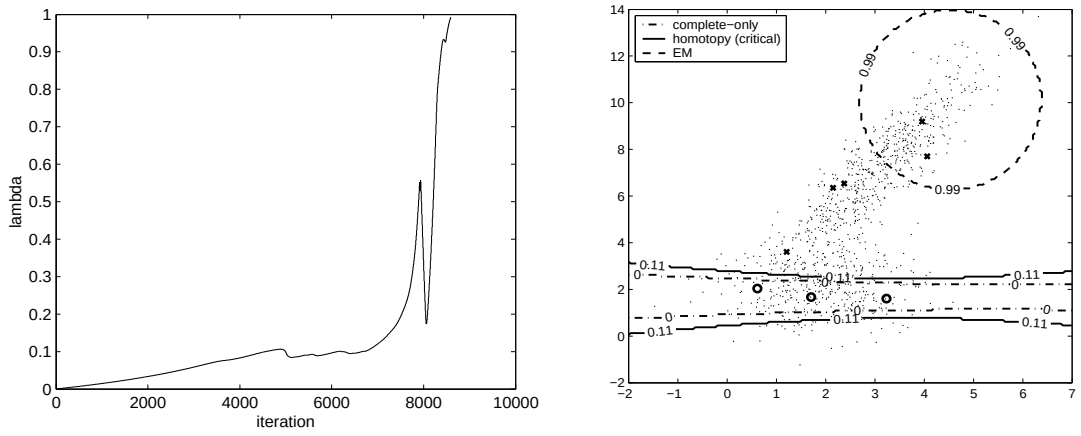

Figure 7-3: homotopy continuation run on experiment from Figure 7-2. Left: allocation with critical values. Right: decision boundary from complete-data only estimation, at the first critical allocation, and as trained by EM.

homotopy can follow through all critical points, we do not have a way to rank critical allocations, therefore we always stop at the first one. Note that the large number of iterations is due to a very small homotopy step size in following the gradient, and can be greatly reduced with adaptive step size methods [23].

The right graph in Figure 7-3 depicts the 8 labeled and all unlabeled samples, and the evolution of the decision boundary while following the path of fixed points. At the first critical allocation unlabeled data correctly adjust the decision boundary, but the solution provided by EM is very far from the truth, splitting a class in two. The transition from the initial boundary to the final one does not happen smoothly with increasing allocation, and that is why homotopy continuation is able to prevent it.

It is important to point out that EM fails mostly due to violation of model assumptions, thus when the training data matches well the model it is likely that homotopy continuation is too conservative. To illustrate such an example we consider a case in which classes are far apart Gaussian clusters, and the model family is also Gaussian with no restrictions (Figure 7-4). With sufficient unlabeled data there is little chance that EM with converge to locally optimal parameters. Therefore at large allocations EM will converge to the same parameters, up to a label permutation, indifferent of the complete evidence. In such situations even noisy labeled data should suffice for

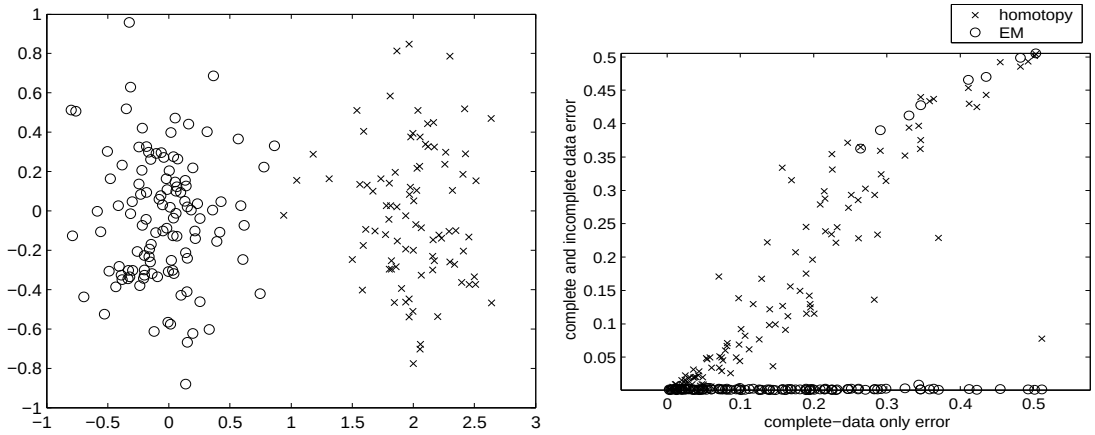

Figure 7-4: EM and homotopy continuation on a data set that agrees well with the Gaussian class model assumption. Left: complete data from which to choose samples. Right: classification error vs. complete-data only error on each run that had a critical allocation out of 500 random selections of 10 labeled and 100 unlabeled samples.

a correct identification of the classes through EM initialization. Indeed, in Figure 7-4 on 500 random selections of 10 labeled and 100 unlabeled samples EM achieves very good error reduction in all cases, while the conservatory homotopy almost does make use of unlabeled data on the critical runs. Average over all runs homotopy still reduces error from 8% to 7%, but EM achieves 3%.

Interestingly, there are still 26% critical runs even with such a perfect match between unlabeled data and model assumptions. This indicates that the critical allocations encountered in this experiment are due to the high variability of the labeled samples that causes an initial mismatch between complete-data only estimate and true distribution. To verify such claim we take a close look at the performance of a sample homotopy continuation (Figure 7-5). In the presented run it is clear that the complete-data only estimate is far from the true distribution, and cannot be continuously evolved into it. Indeed, the path has a critical point at $\lambda = 0.62$, and the decision boundary evolves into a negative direction before critical allocation until it is able to adjust to the unlabeled evidence. Path following before critical allocation does preserve a continuous link with complete evidence, and does not allow a large jump to a different configuration, as it cannot evaluate whether the new configuration will be a very positive or very negative one.

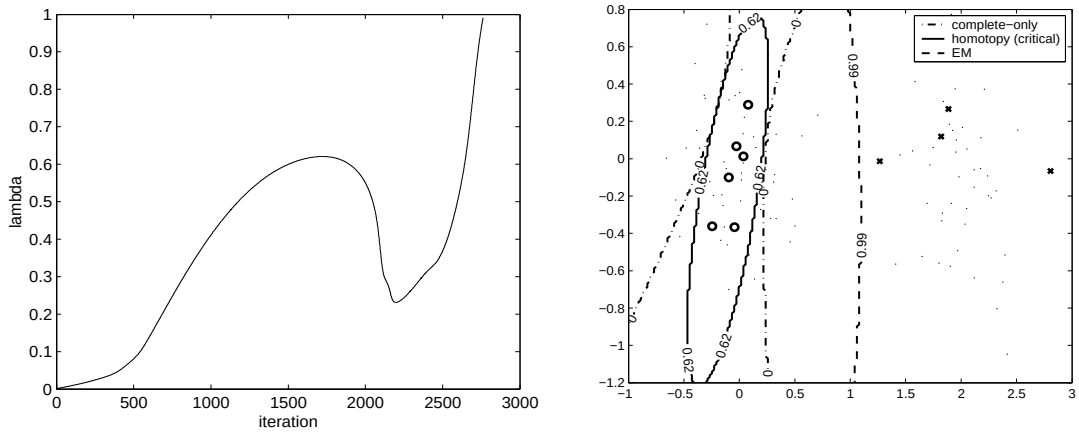

Figure 7-5: homotopy continuation run on experiment from Figure 7-4. Left: allocation with critical values. Right: decision boundary from complete-data only estimation, at the first critical allocation, and as trained by EM.

The artificial experiments on mixtures of normal distributions suggest that homotopy continuation trades off increased stability and connectedness with labeled evidence at the expense of less aggressive usage of incomplete information. This tradeoff is especially beneficial when model assumptions are questionable, as other algorithms such as EM may not only increase but also dramatically decrease performance, and homotopy continuation detects when this may happen.

7.2 Text Classification with Naïve Bayes

We apply homotopy continuation to a document classification problem that potentially poses more difficulties in estimation with labeled and unlabeled samples than the artificial mixture of normals. This is because real data usually has a higher degree of model misfit, and also the larger number of parameters offers more opportunity for local optima. The task is to construct from few documents with known topics and many uncategorized documents a classifier that correctly assigns new documents to one of the given topics. We use the *20-newsgroups* database for training [17], which is a collection of newsgroup articles labeled by the newsgroup on which they were posted.

We use a standard naïve Bayes model for documents, that performs well in practice

even though it discards all syntactical information and makes simplifying assumptions that are clearly violated. Specifically, we represent each document by a set of binary features that indicate word appearance in the document for each word in the vocabulary [14]. Note that only word presence matters, not the actual word count. Other naïve Bayes models that make use of word counts exist [17], but to illustrate homotopy continuation on a simple model we opted for binary features. The naïve Bayes assumption is that documents in each class have statistically independent features. Under this assumption the computations from Section 6.1 are directly applicable.

Because homotopy continuation with exact computations is $O(n^3)$ in the number of parameters, we reduce the size of the problem to control the complexity of experiments. The parameterization of discrete naïve Bayes has $YN + Y - 1$ parameters, where Y is the number of classes and N the number of features, and we reduce both Y and N. We run experiments on 3 randomly selected classes, and the 20 features that have the largest mutual information with the class label. We perform feature selection only once on all 20 newsgroups, hoping that the most significant 20 words do not vary too much for the three-class problem with limited labeled samples.

A difficulty that we may encounter in discrete settings, especially under scarce labeled data, is that empirical maximum likelihood probabilities of some features may be zero. Such zero probabilities do not generalize because the feature will eventually occur in some test sample, and also pose technical problems because they correspond to infinite natural parameters of the exponential family. To avoid such problems we smooth the labeled empirical estimate by adding a count of 1 to each (feature, class) pair (i. e. Laplace smoothing). Smoothing changes the starting point of homotopy continuation, hopefully to a sensible value.

In Table 7.2 we show results from an experiment with 10 randomly selected labeled documents, and the rest of 2916 available documents from the 3 classes selected (*talk.politics.mideast*, *soc.religion.christian*, and *sci.crypt*) as unlabeled data. The results combine 50 experiments out of which 45 featured homotopy paths with critical allocation. Maximum likelihood allocation is effectively 1, while the average critical allocation was 0.93. We see that homotopy continuation dramatically improves the

	labeled only	homotopy	EM
critical runs	35.8%	20.4%	28.0%
all runs	35.7%	21.4%	27.7%

Table 7.1: Error rates of maximum likelihood from labeled data, homotopy continuation, and EM on 50 random selections of 10 labeled documents from 2926 documents in three classes.

poor estimation based only on 10 labeled samples, and outperforms EM even if the critical allocation is close to the maximum likelihood allocation. This is due to the fact that even a small increase in allocation can result in a jump followed by large change in error rate.

We visualize the effect of critical allocation on classification error on a few typical instances of the experiment described above. In Figure 7-6 we show on the same graph in terms of homotopy iteration both the error rate and the allocation parameter. Critical values of allocation are followed by a sudden change in error rate, which is negative most of the time, but it can also be positive. In most situations EM is able to reduce the error rate of maximum likelihood from labeled data, but the reduction can be even more dramatic for lower allocation. In other cases even the slightest positive allocation of unlabeled data damages estimation, and homotopy continuation is able to bound the increase in error by stopping before the dramatic increase. Yet in other runs homotopy continuation does not detect the benefit or the loss from including unlabeled data because there is no critical allocation.

We have seen on a simple naïve Bayes model that homotopy continuation has the ability to bound allocation of incomplete data to maximize its use while maintaining a connection with complete evidence that disallows convergence to unsupported local maxima. Maintaining the homotopy connection is likely to be even more important with larger number of parameters because of a larger number of local optima. Nevertheless, the number of parameters can be prohibitive in computational complexity. Approximate methods that may alleviate this problem are a direction of future research.

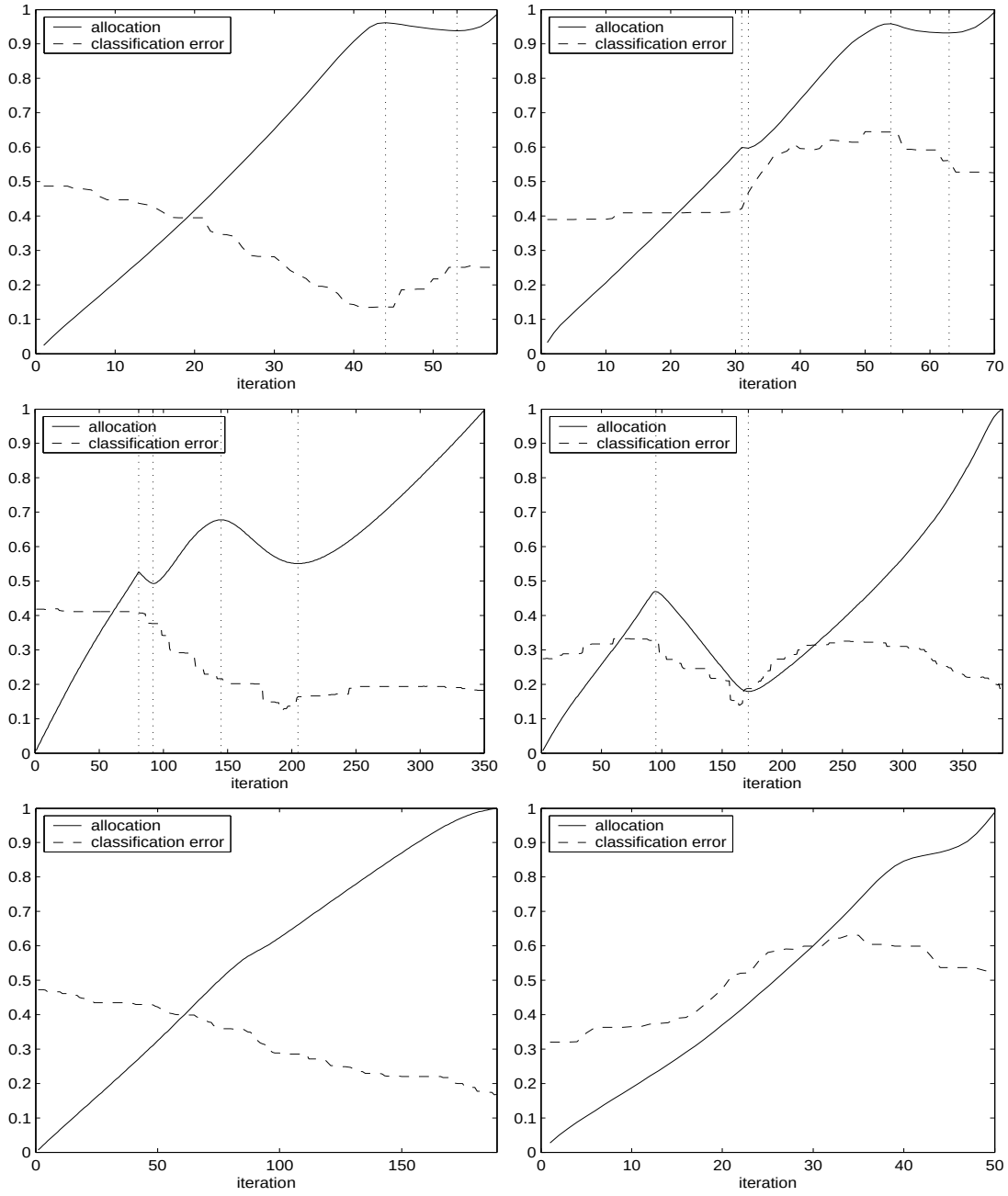

Figure 7-6: Possible evolutions of error rate and allocation with homotopy continuation iteration on a discrete naïve Bayes model with 20 binary features. Critical allocations may or may not be present, and may signal a negative or positive change in error rate.

Chapter 8

Discussion

We have introduced a general algorithm for estimation within the exponential family from limited complete and many incomplete samples that is stable as opposed to standard EM-like algorithms. The algorithm proceeds by following a continuous path of fixed points of the EM operator, starting from full weight on complete data, and increasing the allocation of incomplete data until the path reaches a discontinuity. The algorithm is able to find the critical allocation in $O(n^3)$ complexity in the number of parameters. We present results on artificially generated mixtures of Gaussians, and on text classification with naïve Bayes, demonstrating that our algorithm is able to pick up instances in which EM drastically increases classification error with the inclusion of unlabeled data.

Our view of estimation with incomplete information as a source allocation problem between two heterogeneous data sources permits easy generalization to other problems involving estimation from different sources of information. An important consequence of the presented work is that error rates, and other quantities of interest that depend on the allocation between the different sources, feature discontinuities and sudden changes at specific values of the allocation. Such critical values are identifiable through homotopy continuation methods, and depending on the problem may have different interpretations. We have seen that in the labeled/unlabeled data setting critical allocation signals an unwanted decrease in stability, but in other contexts, such as novelty detection, it may signal that a desired departure from the initial model

has occurred.

The numerical algorithm used in path following, globally convergent with probability one homotopy continuation, is of important interest in itself, because it benefits from theoretically-proven strong regularity properties, that are widely applicable to any fixed-point optimization problem, provided that the conditions of the homotopy theorem are verifiable. Of particular interest are domains that feature natural homotopies in which the allocation parameter has a physical interpretation rather than being artificially introduced.

Possible directions of extending the current work include a homotopy continuation study of allocation between two sources without a priori bias over one of the sources, and further, the study of allocation between multiple sources of information (more than two). Another direction of research is estimation with incomplete information in which samples can exhibit more than one type of incompleteness (different incomplete samples may have different variables missing). Yet another direction of research is exploiting the ability to detect critical allocation in developing active learning algorithms.

Appendix A

Proof of Results

Theorem A.1 *Let* $\mathbf{z}^C = (z_1^C, z_2^C, \ldots, z_N^C)$ *and* $\mathbf{x}^I = (x_1^I, x_2^I, \ldots, x_M^I)$ *be sets of complete and incomplete samples. Given a distribution* $p(z|\boldsymbol{\eta})$ *on the complete space from an exponential family* $\exp\left(\boldsymbol{\theta}\mathbf{t}(z)^T + k(z) - \psi(\boldsymbol{\theta})\right)$*, the following update of the mean parameters:*

$$\mathrm{EM}_\lambda(\boldsymbol{\eta}) = (1-\lambda)\frac{1}{N}\sum_{1\le i\le N}\mathbf{t}(z_i^C) + \lambda\frac{1}{M}\sum_{1\le j\le M}\mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z) = x_j^I, \boldsymbol{\eta}] \tag{A.1}$$

monotonically increases the λ *convex combination of complete and incomplete log-likelihoods.*

Proof We begin with the following application of Jensen's inequality due to the concavity of log:

$$\log\frac{p_X(\mathbf{x}|\boldsymbol{\eta})}{p_X(\mathbf{x}|\hat{\boldsymbol{\eta}})} = \log\int_{\mathcal{Z}(\mathbf{x})}\frac{p(\mathbf{z}|\hat{\boldsymbol{\eta}})}{p_X(\mathbf{x}|\hat{\boldsymbol{\eta}})}\frac{p(\mathbf{z}|\boldsymbol{\eta})}{p(\mathbf{z}|\hat{\boldsymbol{\eta}})}d\mathbf{z} \ge \int_{\mathcal{Z}(\mathbf{x})}\frac{p(\mathbf{z}|\hat{\boldsymbol{\eta}})}{p_X(\mathbf{x}|\hat{\boldsymbol{\eta}})}\log\frac{p(\mathbf{z}|\boldsymbol{\eta})}{p(\mathbf{z}|\hat{\boldsymbol{\eta}})}d\mathbf{z} \tag{A.2}$$

With the above inequality we can derive a variational lower bound at $\hat{\boldsymbol{\eta}}$ on the combined log-likelihood:

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1-\lambda}{N}\sum_{1\le i\le N}\log p(z_i^C|\boldsymbol{\eta}) + \frac{\lambda}{M}\sum_{1\le j\le M}\log p(x_j^I|\boldsymbol{\eta}) \ge \mathcal{L}(\hat{\boldsymbol{\eta}}) +$$
$$+\frac{1-\lambda}{N}\sum_{1\le i\le N}\log\frac{p(z_i^C|\boldsymbol{\eta})}{p(z_i^C|\hat{\boldsymbol{\eta}})} + \frac{\lambda}{M}\sum_{1\le j\le M}\int_{\mathcal{Z}(x_j^I)} p(z|x_j^I,\hat{\boldsymbol{\eta}})\log\frac{p(z|\boldsymbol{\eta})}{p(z|\hat{\boldsymbol{\eta}})}dz \quad \text{(A.3)}$$

Therefore we can guarantee an increase in log-likelihood by maximizing with respect to $\boldsymbol{\eta}$ the following quantity:

$$\frac{1-\lambda}{N}\sum_{1\le i\le N}\log p(z_i^C|\boldsymbol{\eta}) + \frac{\lambda}{M}\sum_{1\le j\le M}\int_{\mathcal{Z}(x_j^I)} p(z|x_j^I,\hat{\boldsymbol{\eta}})\log p(z|\boldsymbol{\eta})dz \quad \text{(A.4)}$$

Representing p as an exponential distribution in terms of natural parameters we obtain the following equivalent formula that must be maximized:

$$\boldsymbol{\theta}\left[(1-\lambda)\frac{1}{N}\sum_{1\le i\le N}\mathbf{t}(z_i^C) + \lambda\frac{1}{M}\sum_{1\le j\le M}\int_{\mathcal{Z}(x_j^I)} p(z|x_j^I,\hat{\boldsymbol{\eta}})\mathbf{t}(z)\right]^T - \psi(\boldsymbol{\theta}) \quad \text{(A.5)}$$

Differentiating with respect to $\boldsymbol{\theta}$ we recover the necessary condition that must be satisfied at the maximum, if the maximum lies inside the domain of mean parameters:

$$\boldsymbol{\eta} = (1-\lambda)\frac{1}{N}\sum_{1\le i\le N}\mathbf{t}(z_i^C) + \lambda\frac{1}{M}\sum_{1\le j\le M}\int_{\mathcal{Z}(x_j^I)} p(z|x_j^I,\hat{\boldsymbol{\eta}})\mathbf{t}(z) \quad \text{(A.6)}$$

Because the domain of mean parameters is convex, and $\mathbf{t}(\mathbf{z})$ is in the domain of mean parameters (a typical exponential-family constraint), the above equation defines a valid set of mean parameters, thus the maximum is not achieved on the boundary. Because the partition function of an exponential family is convex, the above condition is also sufficient, defining the unique maximum. □

Theorem A.2 *Let $p(z|\boldsymbol{\eta})$ be a distribution on the complete space from an exponential family*

$\exp\left(\boldsymbol{\theta}\mathbf{t}(z)^T + k(z) - \psi(\boldsymbol{\theta})\right)$ *and* $\mathbf{x}^I = (x_1^I, x_2^I, \ldots, x_M^I)$ *a set of incomplete samples. The Jacobian of the* EM_1 *operator (with full weight on incomplete samples) is given by:*

$$\nabla_{\boldsymbol{\eta}}\mathrm{EM}_1(\boldsymbol{\eta}) = \left[\frac{1}{M}\sum_{1\le j\le M}\mathrm{cov}(\mathbf{t}(z)|\mathbf{x}(z) = x_j^I, \boldsymbol{\eta})\right] F(\boldsymbol{\eta})^{-1} \tag{A.7}$$

where cov *is the covariance matrix of the random variable* $\mathbf{t}(z)$ *under* $p(z|\mathbf{x}(x) = x_j^I, \boldsymbol{\eta})$*, and* $F(\boldsymbol{\eta})$ *is the Fisher information matrix* $\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\psi(\boldsymbol{\theta})$*.*

Proof We first write the EM_1 operator in the following way using our definition of incompleteness:

$$\mathrm{EM}_1(\boldsymbol{\eta}) = \frac{1}{M}\sum_{1\le j\le M}\mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z) = x_j^I, \boldsymbol{\eta}] = \frac{1}{M}\sum_{1\le j\le M}\frac{\int_{\mathcal{Z}(x_j^I)} p(z|\boldsymbol{\eta})\mathbf{t}(z)dz}{\int_{\mathcal{Z}(x_j^I)} p(z|\boldsymbol{\eta})dz} \tag{A.8}$$

Taking the derivative with respect to $\boldsymbol{\eta}$ we express the Jacobian in terms of $\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta})$:

$$\nabla_{\boldsymbol{\eta}}\mathrm{EM}_1(\boldsymbol{\eta}) = \frac{1}{M}\sum_{1\le j\le M}\left\{\frac{\int_{\mathcal{Z}(x_j^I)}\mathbf{t}(z)^T\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta})}{\int_{\mathcal{Z}(x_j^I)} p(z|\boldsymbol{\eta})} - \mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z) = x_j^I, \boldsymbol{\eta}]\frac{\int_{\mathcal{Z}(x_j^I)}\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta})}{\int_{\mathcal{Z}(x_j^I)} p(z|\boldsymbol{\eta})}\right\} \tag{A.9}$$

Using the fact the p is an exponential family we can derive $\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta})$ by first differentiating with respect to the natural parameters $\boldsymbol{\theta}$

$$\frac{d}{d\boldsymbol{\theta}}p(z|\boldsymbol{\theta}) = p(z|\boldsymbol{\theta})\left[\mathbf{t}(z) - \frac{d}{d\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\right] = p(z|\boldsymbol{\theta})\left[\mathbf{t}(z) - \boldsymbol{\eta}\right] \tag{A.10}$$

and changing the variable to $\boldsymbol{\eta}$

$$\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta}) = \frac{d}{d\boldsymbol{\theta}}p(z|\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\eta}}\boldsymbol{\theta} = p(z|\boldsymbol{\eta})\left[\mathbf{t}(z) - \boldsymbol{\eta}\right]\cdot F(\boldsymbol{\eta})^{-1} \tag{A.11}$$

where $F(\boldsymbol{\eta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}^2}\psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\eta}$ is the Fisher information matrix.

Substituting $\frac{d}{d\boldsymbol{\eta}}p(z|\boldsymbol{\eta})$ in (A.9) we get

$$\nabla_{\boldsymbol{\eta}}\mathrm{EM}_1(\boldsymbol{\eta}) = \frac{1}{M}\sum_{1\le j\le M}\{\mathrm{E}[\mathbf{t}(z)^T\mathbf{t}(z)|\mathbf{x}(z)=x_j^I,\boldsymbol{\eta}] - \mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z)=x_j^I,\boldsymbol{\eta}]^T\mathrm{E}[\mathbf{t}(z)|\mathbf{x}(z)=x_j^I,\boldsymbol{\eta}]\}F(\boldsymbol{\eta})^{-1} \tag{A.12}$$

which is the result stated in the theorem. □

Theorem A.3 *For any distributions P and Q, and for any $\lambda \in [0,1)$ the following inequality holds:*

$$D(P \,\|\, (1-\lambda)P + \lambda Q) \le \log \frac{1}{1-\lambda} \tag{A.13}$$

Proof $D(P \,\|\, (1-\lambda)P + \lambda Q) = \int P \log \frac{P}{(1-\lambda)P+\lambda Q} \le \int P \log \frac{P}{(1-\lambda)P} = \log \frac{1}{1-\lambda}$. □

Theorem A.4 *Let f and g be continuous functions on a compact set. There exists a monotonically increasing function $\beta : [0,1] \to [0,\infty)$ such that for every $\lambda \in [0,1]$, $(1-\lambda)f + \lambda g$ and $\{g|f \le \beta(\lambda)\}$ achieve their minimum at the same time. Moreover, we can choose $\beta(0) = \min f$ and any $\beta(1) \ge \max f$.*

Proof $(1-\lambda)f + \lambda g$ is continuous on a compact set thus it achieves its minimum. Let x_λ be one of the points achieving it. Moreover, let x'_λ be a point achieving the minimum of $\{g|f \le f(x_\lambda)\}$. Then $f(x'_\lambda) \le f(x_\lambda)$ and $g(x'_\lambda) \le g(x_\lambda)$. Therefore $(1-\lambda)f(x'_\lambda) + \lambda g(x'_\lambda) \le (1-\lambda)f(x_\lambda) + \lambda g(x_\lambda)$, and because x_λ is a minimum by definition, we must have equality. If follows that x'_λ achieves the minimum of both $\{g|f \le f(x'_\lambda)\}$ and $(1-\lambda)f + \lambda g$. Therefore we can define $\beta(\lambda) = f(x'_\lambda)$ to ensure that the functions in the statement achieve their minimum at the same time.

It remains to show that β is increasing. Let $\lambda_1 \le \lambda_2$. By definition

$$(1-\lambda_1)f(x'_{\lambda_1}) + \lambda_1 g(x'_{\lambda_1}) \le (1-\lambda_1)f(x'_{\lambda_2}) + \lambda_1 g(x'_{\lambda_2})$$
$$(1-\lambda_2)f(x'_{\lambda_1}) + \lambda_2 g(x'_{\lambda_1}) \ge (1-\lambda_2)f(x'_{\lambda_2}) + \lambda_2 g(x'_{\lambda_2})$$

Substracting the second inequality from the first we obtain:

$$(\lambda_2 - \lambda_1)(f(x'_{\lambda_1}) - g(x'_{\lambda_1})) \leq (\lambda_2 - \lambda_1)(f(x'_{\lambda_2}) - g(x'_{\lambda_2}))$$

therefore $f(x'_{\lambda_1}) - g(x'_{\lambda_1}) \leq f(x'_{\lambda_2}) - g(x'_{\lambda_2})$, or equivalently $\beta(\lambda_1) - \beta(\lambda_2) \leq g(x'_{\lambda_1}) - g(x'_{\lambda_2})$.

Assume by contradiction that $\beta(\lambda_1) > \beta(\lambda_2)$. It follows that $g(x'_{\lambda_1}) > g(x'_{\lambda_2})$, and since $f(x'_{\lambda_2}) \leq \beta(\lambda_2) < \beta(\lambda_1)$, x'_{λ_1} cannot be a minimum of $\{g | f \leq \beta(\lambda_1)\}$. This is a contradiction. The choice of values of β at 0 and 1 is trivial. □

Bibliography

[1] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000.

[3] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, New York, 1978.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[5] S.N. Chow, J. Mallet-Paret, and J.A. Yorke. Finding zeros of maps: Homotopy methods that are constructive with probability one. *Math. Comput.*, 32:887–899, 1978.

[6] F. M. Coetzee and V. L. Stonick. On a natural homotopy between linear and nonlinear single layer perceptron networks. *IEEE Transactions on Neural Networks*, 7:307–317, 1996.

[7] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, New York, 1991.

[9] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No. 1:205–237, 1984.

[10] 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 B*, 39:1–22, 1977.

[11] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29:505–529, 2001.

[12] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an EM approach. In *Neural Information Processing Systems*, volume 6, pages 120–127, 1994.

[13] T. Joachims. Transductive inference for text classification using s support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, 1999.

[14] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorisation task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50, 1992.

[15] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 1997.

[16] D. Miller and H. Uyar. A mixture of experts classifier with learning based on both labeled and unlabeled data. In *Adv. in Neural Information Processing Systems 9*, pages 571–577, 1997.

[17] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press, 2nd edition, 1992.

[19] D. Schuurmans and F. Southey. An adaptive regularization criterion for supervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 847–854, 2000.

[20] B. M. Shahshahani and D. A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Trans. on Geosc. and Remote Sensing*, 32(5):1087–1095, 1994.

[21] V. N. Vapnik. *Statistical Learning Theory*. Wiley & Sons, New York, 1998.

[22] L. T. Watson. Theory of globally convergent probability-one homotopies for nonlinear programming. *SIAM Journal on Optimization*, 11(3):761–780, 2000.

[23] L.T. Watson and A.P. Morgan. Algorithm 652. HOMPACK: A suite of codes for globally convergent homotopy algorithms. *ACM Trans. on Math. Software*, 13(3):281–310, 1987.

[24] D. Yarowsky. Usupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.