UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Iterative proportional fitting and exponential families
Fall 2011

1 Introduction

In previous lectures, we have considered exponential family distributions of the form

$$p(x;\theta) = h(x)\ \exp\big\{\sum_{j=1}^{s}\theta_j T_j(x) - A(\theta)\big\}, \tag{1}$$

where $\{T_j\}_{j=1}^s$ are a family of sufficient statistics, and $\{\theta_j\}_{j=1}^s$ are the associated exponential parameters. Recall that we have showed that the function A has moment-generating properties, in that it is (infinitely) differentiable, with partial derivatives

$$\frac{\partial A}{\partial \theta_j}(\theta) = \mathbb{E}_\theta[T_j(X)], \tag{2}$$

where $\mathbb{E}_\theta$ denotes expectation under the distribution $p(\cdot\,;\theta)$. In these notes, we show how the IPF updates can be understood as a coordinate optimization procedure for an exponential family representation of an undirected graphical model.

2 Parameter estimation for undirected graphical models

We now turn to the following problem. Suppose that we have an undirected graphical model, based on some known graph $G = (V, E)$, and involving discrete random variables $X_s \in \mathcal{X} = \{0, 1, \ldots, m-1\}$ at each vertex $s \in V$. If vertex $V = \{1, 2, \ldots, d\}$, then the random vector $(X_1, \ldots, X_d)$ takes values in the space

$$\mathcal{X}^d = \underbrace{\{0,1,\ldots,m-1\} \times \cdots \times \{0,1,\ldots,m-1\}}_{d\text{-times}}$$

By our previous results, any such graphical model has the factorization

$$p(x_1, \ldots, x_d; \psi) = \frac{1}{Z}\prod_{C\in\mathfrak{C}} \psi_C(x_C), \tag{3}$$

where $\mathfrak{C}$ is the set of all cliques of the graph, and for each clique $C \in \mathfrak{C}$, the potential function ψ_C maps any configuration x_C to a non-negative real number $\psi_C(x_C)$.

Now suppose that we do not know the potential functions $\psi = \{\psi_C, C \in \mathfrak{C}\}$, but that we have some way of drawing a collection of samples $x_i \in \mathcal{X}^m$, for $i = 1, 2, \ldots, n$, where each $x_i \sim p(\cdot\,;\psi)$. Our goal is to use this collection of samples in order to compute a "reasonable" estimate of the unknown potential functions.

2.1 Iterative proportional fitting (IPF)

For each fixed configuration $\bar{x}_C$, we use $x_C \mapsto \delta(x_C = \bar{x}_C)$ to denote a 0-1-valued indicator function for the event $\{x_C = \bar{x}_C\}$—that is, the function

$$\delta(x_c = \bar{x}_C) = \begin{cases} 1 & \text{if } x_C = \bar{x}_C \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

For future reference, we note that the marginal probability $\mu_C(\bar{x}_C)$ has a convenient expression in terms of this function as

$$\mu_C(\bar{x}_C) = \sum_{x \in \mathcal{X}^d} p(x; \psi)\, \delta(x_C = \bar{x}_C). \tag{5}$$

Using this notation, the iterative proportional fitting algorithm consists of the following steps:

1. For all cliques $C \in \mathfrak{C}$:

   - For iteration $t = 0$, initialize the potential function $\psi_C^0(x_C) = 1$ for all configurations x_C.
   - Compute the *empirical marginal distributions* with values

   $$\widehat{\mu}_C(\bar{x}_C) = \frac{1}{n} \sum_{i=1}^{n} \delta\big(x_{C,i} = \bar{x}_C\big) \qquad \text{for each } x_C \in \mathcal{X}^{|C|}. \tag{6}$$

2. For iterations $t = 0, 1, 2, \ldots$ and with a fixed order of the cliques C:

   - Compute *model marginal distribution* μ_C^t over clique C using current set $\psi^t = \{\psi_C^t, C \in \mathfrak{C}\}$ of potentials:

   $$\mu_C^t(\bar{x}_C) = \sum_{x \in \mathcal{X}^m} p(x; \psi^t)\, \delta\big(x_C = \bar{x}_C\big) \tag{7}$$

   - Update potential function on clique C:

   $$\psi^{t+1}(x_C) = \psi_C^t(x_C)\, \frac{\widehat{\mu}_C(x_C)}{\mu_C^t(x_C)}. \tag{8}$$

   Leave all other clique potentials fixed (i.e., $\psi_D^{t+1} = \psi_D^t$ for all cliques $D \in \mathfrak{C} \backslash \{C\}$.)

2.2 Some properties

To gain some intuition for the IPF algorithm, let's consider some of its properties.

Fixed points: First, what does a fixed point ψ^* look like? (A fixed point of the algorithm is a set of potential functions for which the updates (8) produce no change, and hence they remain fixed.) At any fixed point, we are guaranteed that the empirical marginal distribution $\widehat{\mu}_C$ is equal to the marginal distribution computed under the model $p(\cdot;\, \psi^*)$—that is

$$\text{For all } C \in \mathfrak{C}: \qquad \underbrace{\widehat{\mu}_C(\bar{x}_C)}_{\text{Emp. marginal}} \quad = \quad \underbrace{\mu_C^t(\bar{x}_C)}_{\text{Model marginal}} \quad := \sum_{x \in \mathcal{X}^N} p(x; \psi^*)\, \delta\big(x_C = \bar{x}_C\big) \tag{9}$$

This is an intuitively reasonable property, and as we will see, it underlies the connection between IPF and maximum likelihood in a suitable exponential family.

Marginal matching: At iteration t the IPF algorithm is maintaining its current best estimate of the distribution (3)—namely,

$$p(x;\psi^t) = \frac{1}{Z^t} \prod_{C\in\mathfrak{C}} \psi_C^t(x_C), \tag{10}$$

where $Z^t = \sum_x \prod_{C\in\mathfrak{C}} \psi_C^t(x_C)$ is the normalization constant. Suppose that in iteration $t \mapsto t+1$, clique D is updated. We then claim that after this update, we have

$$\mu_D^{t+1}(\bar{x}_B) = \widehat{\mu}_D(\bar{x}_D), \quad \text{and} \tag{11a}$$

$$Z^{t+1} = Z^t. \tag{11b}$$

To verify this claim, note that by definition (8) of the IPF update, we have

$$\frac{p(x;\psi^{t+1})}{p(x;\psi^t)} = \frac{Z^t}{Z^{t+1}} \frac{\widehat{\mu}_D(x_D)}{\widehat{\mu}_D^t(x_D)}. \tag{12}$$

Therefore, we can write

$$\begin{aligned}
\mu_D^{t+1}(\bar{x}_D) &\overset{(i)}{=} \sum_{x\in\mathcal{X}^d} p(x;\psi^{t+1})\, \delta(x_D = \bar{x}_D) \\
&\overset{(ii)}{=} \sum_{x\in\mathcal{X}^d} \frac{Z^t}{Z^{t+1}} \frac{\widehat{\mu}_D(x_D)}{\mu_D^t(x_D)} p(x;\psi^t)\delta(x_D = \bar{x}_D) \\
&= \frac{Z^t}{Z^{t+1}} \frac{\widehat{\mu}_D(\bar{x}_D)}{\mu_D^t(x_D)} \sum_{x\in\mathcal{X}^d} p(x;\psi^t)\delta(x_D = \bar{x}_D) \\
&\overset{(iii)}{=} \frac{Z^t}{Z^{t+1}} \widehat{\mu}_D(\bar{x}_D),
\end{aligned} \tag{13}$$

where step (i) uses the definition of the marginal μ_D^{t+1}, step (ii) uses the ratio (12), and step (iii) uses the definition of the marginal μ_D^t. This relation holds for each configuration $\bar{x}_D$, so we may sum both sides over all choices of $\bar{x}_D$, Doing so yields

$$1 \;=\; \sum_{\bar{x}_D} \mu_D^{t+1}(\bar{x}_D) \;=\; \frac{Z^t}{Z^{t+1}} \sum_{\bar{x}_D} \widehat{\mu}_D(\bar{x}_D) \;=\; \frac{Z^t}{Z^{t+1}}.$$

Consequently, we conclude that $Z^t = Z^{t+1}$. Using this fact, the relation (13) implies the equivalence (11a), as claimed.

Convergence: Given that we have some understanding of the fixed points and properties of the updates, the next natural question to ask is whether the IPF algorithm is guaranteed to converge. In order to do so, it turns out to be convenient to return to the formalism of exponential families.

3 IPF and exponential families

Let us now write the general undirected graphical model (3) as an exponential family in terms of the indicator functions $x_C \mapsto \delta(x_C = \bar{x}_C)$. In order to do so, note the potential function ψ_C can

be written in the form

$$\psi_C(x_C) = \prod_{\bar{x}_C \in \mathcal{X}^{|C|}} \left[\psi_C(\bar{x}_C)\right]^{\delta(x_C = \bar{x}_C)}. \tag{14}$$

For instance, in the case of a singleton clique $C = \{s\}$, we have

$$\psi_s(x_s) = \prod_{\bar{x}_s=0}^{m-1} \left[\psi_s(\bar{x}_s)\right]^{\delta(x_s=\bar{x}_s)} \;=\; \begin{cases} \psi_s(0) & \text{if } x_s = 0 \\ \psi_s(0) & \text{if } x_s = 1 \\ \vdots & \vdots \\ \psi_s(m-1) & \text{if } x_s = m-1. \end{cases}$$

As long as $\psi_C(x_C) > 0$, we can define a new function

$$\theta_C(x_C) = \log \psi_C(x_C) \;=\; \sum_{\bar{x} \in \mathcal{X}^{|C|}} \theta_C(\bar{x}_C)\delta(x_C = \bar{x}_C).$$

Here the numbers $\{\theta_C(\bar{x}_C), \bar{x}_C \in \mathcal{X}^N\}$ are the *canonical parameters* of the exponential family, and the functions $\{x_C \mapsto \delta(x_C = \bar{x}_C)\}$ are the *sufficient statistics*. (That is, for a clique C, we have a total of $m^{|C|}$ different parameters and associated sufficient statistics.

With this notation, we can re-express the original factorization (3) as the exponential family

$$p(x;\theta) = \exp\Big\{\sum_{C \in \mathfrak{C}} \theta_C(x_C) - A(\theta)\Big\} \;=\; \exp\Bigg\{\sum_{C \in \mathfrak{C}} \Big\{\sum_{\bar{x}_C \in \mathcal{X}^{|C|}} \theta_C(\bar{x}_C)\delta(x_C = \bar{x}_C)\Big\} - A(\theta)\Bigg\}, \tag{15}$$

where $A(\theta) = \log \sum_x \exp\big\{\sum_{C \in \mathfrak{C}} \theta_C(x_C)\big\}$ is the log normalization constant. So the whole graph has a total of $\sum_{C \in \mathfrak{C}} m^{|C|}$ exponential parameters.

Let us now consider the form of maximum likelihood for this exponential family. From our development in previous lectures, if we have n i.i.d. samples $x_i \in \mathcal{X}^d$ from our model, then the MLE can be obtained by minimizing the objective function

$$J(\theta) = A(\theta) - \sum_{C \in \mathfrak{C}} \Bigg\{\sum_{x_C \in \mathcal{X}^N} \theta_C(x_c)\widehat{\mu}_C(x_C)\Bigg\}, \tag{16}$$

where $\widehat{\mu}_C(\bar{x}_C) = \frac{1}{n}\sum_{i=1}^n \delta(x_{C,i} = \bar{x}_C)$ is the empirical marginal on clique C.

One way in which to optimize is by *coordinate minimization*: namely, we choose a block of variables to update, and then perform a limited minimization over this block, with all the other variables held fixed. For the problem at hand, the cliques of graph provide a natural partitioning of the variables. Suppose that at iteration t, we choose to update clique D. Let us define the vector Then the update $t \mapsto t+1$ of coordinate minimization applied to the objective (16) takes the form

- Set $\theta_D^{t+1} = \arg\min_{\theta_D} J(\theta_D,\ \theta_C^t \text{ for all } C \neq D)$
- Set $\theta_C^{t+1} = \theta_C^t$ for all $C \neq D$.

By properties of exponential families, the function J is convex in each co-ordinate. Hence, we can find the optimum θ_D^{t+1} over block D by solving the equation $\nabla_D J(\theta)$, where ∇_D denotes the partial derivative of J with respect to the block θ_D. This condition reduces to the coupled collection of equations

$$\frac{\partial A}{\partial \theta_D(\bar{x}_D)}(\theta^{t+1}) = \widehat{\mu}_D(\bar{x}_D) \quad \text{for all configurations } \bar{x}_D. \tag{17}$$

But by the moment generating properties of A and our choice of sufficient statistics, we know that

$$\frac{\partial A}{\partial \theta_D(\bar{x}_D)}(\theta^{t+1}) = \underbrace{\sum_{x \in \mathcal{X}^N} p(x; \theta^{t+1}) \delta(x_D = \bar{x}_D)}_{\text{Model marginal } \mu_D^t(\bar{x}_D)}.$$

Consequently, the minimum over block D will be achieved when the marginal matching condition

$$\mu_D^{t+1}(\bar{x}_D) = \widehat{\mu}_D(\bar{x}_D) \tag{18}$$

holds. As we have seen, this is exactly the condition guaranteed by the IPF algorithm.

In summary, then, we have shown that IPF can be re-derived as an algorithm for *performing coordinate minimization of the negative log likelihood* (16). This is a useful result, because it allows us to understand the convergence properties of IPF by recourse to more general results about coordinate minimization. In general, it can be shown that for a convex and differentiable function (such as our objective (16)), a coordinatewise minimization algorithm will converge as long as each coordinate sub-problem is well-behaved in a certain sense. These conditions are satisfied for our problem, so we may conclude that IPF will converge to exactly the maximum likelihood estimate of the model parameters.