# Multi-Conditional Learning for Joint Probability Models with Latent Variables

**Chris Pal, Xuerui Wang, Michael Kelm and Andrew McCallum**
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{pal,xuerui,mccallum}@cs.umass.edu

## Abstract

We introduce Multi-Conditional Learning, a framework for optimizing graphical models based *not* on joint likelihood, or on conditional likelihood, but based on a product of several marginal conditional likelihoods each relying on common sets of parameters from an underlying joint model and predicting different subsets of variables conditioned on other subsets. When applied to undirected models with latent variables, such as the Harmonium, this approach can result in powerful, structured latent variable representations that combine some of the advantages of conditional random fields with the unsupervised clustering ability of popular topic models, such as latent Dirichlet allocation and its successors. We present new algorithms for parameter estimation using expected gradient based optimization and develop fast approximate inference algorithms inspired by the contrastive divergence approach. Our initial experimental results show improved cluster quality on synthetic data, promising results on a vowel recognition problem and significant improvement inferring hidden document categories from multiple attributes of documents.

## 1 Introduction

Recently, there has been substantial interest in Conditional Random Fields (CRFs) [6] for sequence processing. CRFs are random fields for a joint distribution globally conditioned on feature observations. The CRF construction can be contrasted with MRFs which have been used in the past to define and model the joint distribution of both labels *and* features. CRFs are also optimized using a Maximum Conditional Likelihood objective as opposed to a Maximum (Joint) Likelihood objective which is a traditional objective function used for MRFs. In the Machine Learning community the Boltzmann machine [1] is another well known example of an MRF and recent attention has been given to a restricted type of Boltzman machine [3] known as a Harmonium [8]. We are interested in more deeply exploring the relationships between these models, the distributions they define and the ways in which they can be optimized.

In the approach we propose and develop here, one begins by specifying a joint probability model for all quantities one wishes to consider as random quantities, for our discussion here we can think of this as a joint model for: "labels", "features", hidden variables and parameters if desired. However, to optimize the joint model we propose finding point estimates for parameters using an objective function consisting of the product of select

(marginal) conditional distributions. The goal of this objective is thus to obtain a Random Field that has been optimized to be good at modeling a number of conditional distributions that the modeler is particularly interested in focusing on capturing well within the context of a single joint model. Our experiments show that latent variable models obtained by optimization with respect to this type of objective can produce both qualitatively better clusters as well as quantitatively better structured latent topic spaces.

## 2 Multi-Conditional Learning in Joint Models

### 2.1 A Simple Illustrative Example in a Locally Normalized Model

Here we present an example of how one can optimize a joint probability model under a number of different objectives. Consider a Gaussian mixture model (GMM) for real valued random observed variables $\mathbf{x}$ (e.g., observed 2D values) with an unobserved sub-class, $s$ associated with each observed class label $c$. We will use the notation $\tilde{\mathbf{x}}$ and $\tilde{c}$ to denote observations or instantiations of continuous and discrete random variables. We can write a model for the joint distribution of these random variables as $P(\mathbf{x}, c, s) = p(\mathbf{x}|s)P(s|c)P(c)$, where $P(s|c)$ is a sparse matrix associating a number of sub-classes. We shall use $\Theta$ to denote all the parameters of the model. Now consider that it is possible to optimize the GMM in a number of different ways. First, consider the log marginal *joint* likelihood $\mathcal{L}_{\mathbf{x},c}$ of such a model, which can be expressed as:

$$\mathcal{L}_{\mathbf{x},c}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) = \sum_i \log P(\tilde{\mathbf{x}}_i, \tilde{c}_i|\Theta) = \sum_i \log \sum_{s_i} P(\tilde{\mathbf{x}}_i, s_i, \tilde{c}_i|\Theta) = \mathcal{L}_{\mathbf{x},c}(\Theta) \quad (1)$$

Second, in contrast to the log marginal joint likelihood, the log marginal *conditional* likelihood $\mathcal{L}_{c|\mathbf{x}}$ can be expressed as:

$$\mathcal{L}_{c|\mathbf{x}}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) = \sum_i \log \sum_{s_i} P(\tilde{c}_i, s_i|\tilde{\mathbf{x}}_i, \Theta) = \mathcal{L}_{\mathbf{x},c}(\Theta) - \mathcal{L}_{\mathbf{x}}(\Theta) \quad (2)$$

Third, consider the following multi-conditional objective function, $\mathcal{L}_{c|\mathbf{x},\mathbf{x}|c}$ which we express as:

$$\begin{aligned} \mathcal{L}_{c|\mathbf{x},\mathbf{x}|c}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) &= \sum_i \log P(\tilde{c}_i|\tilde{\mathbf{x}}_i, \Theta) + \sum_i \log P(\tilde{\mathbf{x}}_i|\tilde{c}_i, \Theta) \\ &= 2\mathcal{L}_{\mathbf{x},c}(\Theta) - \mathcal{L}_{\mathbf{x}}(\Theta) - \mathcal{L}_c(\Theta) \end{aligned} \quad (3)$$

Consider now the following simple example data set which is similar to the example presented in Jebara's work [4] to illustrate his Conditional Expectation Maximization (CEM) approach. Similarly, we generate data from two classes, each with four sub-classes drawn from 2D isotropic Gaussians. The data are illustrated by red $\circ$'s and blue $\times$'s in Figures 1. In contrast to [4], here we fit models with diagonal covariance matrices and we use the conditional expected gradient [7] optimization approach to update parameters. To illustrate the effects of the different optimization criteria we have fit models with two subclasses for each class. We run each algorithm with 30 random initializations using gradient based optimization for the three objective functions and choose the best model under the joint, conditional and multi-conditional objectives, (1), (2) and (3), respectively. We illustrate the model parameters using ellipses of constant probability under the model.

From this illustrative example, we see that the joint likelihood based objective encodes no element explicitly enforcing a good model of the conditional distribution of the class label and can thus place probability mass in poor locations with respect to classification. The conditional objective focuses completely on the decision boundary and can produce parameters with very little interpretability. Whereas our multi-conditional objective explicitly optimizes for a good class conditional distribution and a good setting of parameters for making classifications.

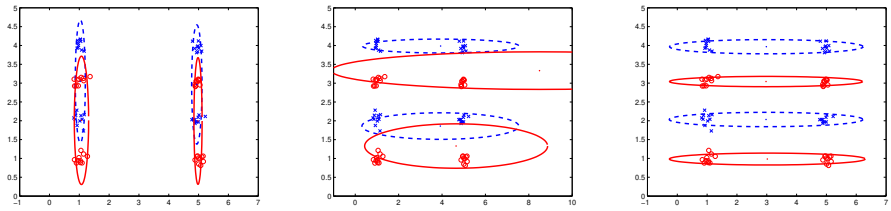

Figure 1: (Left) Joint likelihood optimization. (Middle) One of the many near optimal solutions found by conditional likelihood optimization. (Right) An optimal solution found by our multi-conditional objective.

Quantitatively, we have found that a similar multi-conditional optimization and model selection procedure for the 11 class, isolated vowel recognition problem in [2] leads to a test set error rate of .36 compared to .40 using the ML and CL objectives. In contrast, the best published result is .39 using multivariate adaptive regression splines (MARS) [2].

### 2.2 Our Proposed Multi-Conditional Objective

More generally, our objective function can be expressed as follows. Consider a data set consisting of $i = 1 \ldots N$ observation instances, hidden discrete and continuous variables $\{z\}_i$ and $\mathbf{z}_i$ respectively. We define $j = 1 \ldots M$ pairs of disjoint subsets of observed variables where $\{\tilde{x}\}_{i,j}$ represents the $i$th instance of the variables in subset $j$ and $\{\tilde{x}\}_{i,\bar{j}}$ is the other half of the pair (which we will condition upon). Using these definitions, the optimal parameter settings under our multi-conditional criterion are given by

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_i \prod_j \sum_{\{z\}_{i,j}} \int P(\{\tilde{x}\}_{i,j}, \{z\}_{i,j}, \mathbf{z}_{i,j} | \{\tilde{x}\}_{i,\bar{j}}; \boldsymbol{\theta}) \mathrm{d}\mathbf{z}_{i,j}, \tag{4}$$

where we derive these marginal conditional likelihoods from a single underlying joint model which itself may be normalized locally, globally or using some combination of the two.

### 2.3 A Structured, Globally Normalized, Latent Variable Model for Documents

A Harmonium model [8] is a Markov Random Field (MRF) consisting of observed variables and hidden variables. Like all MRFs the model we present here will be defined in terms of a globally normalized product of (un-normalized) potential functions defined upon subsets of variables. A Harmonium can also be described as a type of restricted Boltzmann machine [3] which can be written as an exponential family model. In particular, the exponential family Harmonium structured model we develop here can be written as

$$P(\mathbf{x}, \mathbf{y} | \boldsymbol{\Theta}) = \exp \left\{ \sum_i \boldsymbol{\theta}_i^T \boldsymbol{f}_i(\mathbf{x}_i) + \sum_j \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\mathbf{y}_j) + \sum_i \sum_j \boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{y}_j) - A(\boldsymbol{\Theta}) \right\}, \tag{5}$$

where $\mathbf{y}$ is a vector of hidden variables, $\mathbf{x}$ is a vector of observations, $\boldsymbol{\theta}_i$ represents parameter vectors (or weights), $\boldsymbol{\theta}_{ij}$ represents a parameter vector on a cross product of states, $\boldsymbol{f}_i$ denotes potential functions, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j\}$ is the set of all parameters and $A$ is the log-partition function or normalization constant. A Harmonium model factorizes the third term of (5) into $\boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{y}_j) = \boldsymbol{f}_i(\mathbf{x}_i)^T \mathbf{W}_{ij}^T \boldsymbol{f}_j(\mathbf{y}_j)$, where $\mathbf{W}_{ij}^T$ is a parameter matrix with dimensions $a \times b$, i.e., with rows equal to the number of states of $\boldsymbol{f}_i(\mathbf{x}_i)$ and columns equal to the number of states of $\boldsymbol{f}_j(\mathbf{y}_j)$. Figure 2 (right) illustrates a Harmonium model as a factor graph [5]. Importantly, a Harmonium describes the factorization of a

joint distribution for observed and hidden variables into a globally normalized product of local functions. In our experiments here we shall use the Harmonium's factorization structure to define a MRF and we will then define sets of marginal conditionals distributions of some *observed* variables given others that are of particular interest so as to form our multi-conditional objective.

Importantly, using a globally normalized joint distribution with this construction it is also possible to derive two consistent conditional models, one for hidden variables given observed variables and one for observed variables given hidden variables [9]. The conditional distributions defined by these models can also be used to implement sampling schemes for various probabilities in the underlying joint model. However, is important to remember that the original model parameterization is not defined in terms of these conditional distributions. In our specific experiments below we use a joint model with a form defined by (5) with $\mathbf{W}^T = [\mathbf{W}_b^T \mathbf{W}_d^T]$ such that the (exponential family) conditional distributions consistent with the joint model are given by

$$P(\mathbf{y}_n|\tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{y}_n; \hat{\boldsymbol{\mu}}, \mathbf{I}), \quad \hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \mathbf{W}^T\tilde{\mathbf{x}} \quad \text{and} \tag{6}$$

$$P(\mathbf{x}_b|\tilde{\mathbf{y}}) = \mathcal{B}(\mathbf{x}_b; \hat{\boldsymbol{\theta}}_b), \quad \hat{\boldsymbol{\theta}}_b = \boldsymbol{\theta}_b + \mathbf{W}_b\tilde{\mathbf{y}} \quad \text{and} \tag{7}$$

$$P(\mathbf{x}_d|\tilde{\mathbf{y}}) = \mathcal{D}(\mathbf{x}_d; \hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\theta}}_d = \boldsymbol{\theta}_d + \mathbf{W}_d\tilde{\mathbf{y}} \tag{8}$$

Where $\mathcal{N}()$, $\mathcal{B}()$ and $\mathcal{D}()$ represent Normal, Bernoulli and Discrete distributions respectively. The following equation can be used to represent the marginal

$$P(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} - A(\boldsymbol{\theta}, \boldsymbol{\Lambda})\} \tag{9}$$

where $\boldsymbol{\Lambda} = \frac{1}{2}\mathbf{W}\mathbf{W}^T$. In an exponential family model with exponential function $\mathbf{F}(\mathbf{x}; \theta)$, it is easy to verify that the gradient of the log joint likelihood can be expressed as:

$$\frac{\partial \mathcal{L}(\theta; \tilde{\mathbf{x}})}{\partial \theta} = N\left[E_{\tilde{P}(\mathbf{x})}\left\langle \frac{\partial \mathbf{F}(\mathbf{x}; \theta)}{\partial \theta} \right\rangle - E_{P(\mathbf{x};\theta)}\left\langle \frac{\partial \mathbf{F}(\mathbf{x}; \theta)}{\partial \theta} \right\rangle\right] \tag{10}$$

where $E_{\tilde{P}(\mathbf{x})}$ denotes the expectation under the empirical distribution, $E_{P(\mathbf{x})}$ is an expectation under the models marginal distribution and $N$ is the number of data elements. We can thus compute the gradient of the log-likelihood under our construction using

$$\frac{\partial \mathcal{L}(\mathbf{W}^T; \tilde{\mathbf{X}})}{\partial \mathbf{W}^T} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\mathbf{W}^T\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T - \frac{1}{N_s}\sum_{j=1}^{N_s}\mathbf{W}^T\tilde{\mathbf{x}}_{i,(j)}\tilde{\mathbf{x}}_{i,(j)}^T\right) \tag{11}$$

where $N_d$ are the number of vectors of observed data, $\tilde{\mathbf{x}}_{i,(j)}$ are samples indexed by $j$ and $N_s$ are the number of MCMC samples used per data vector and computed Gibbs sampling and conditionals (6), (7) and (8). In our experiments here we have found it possible to use either one or a small number of MCMC steps initialized from the data vector (the contrastive divergence approach) but a more standard MCMC approximation is also possible. Finally, for conditional likelihood and multi-conditional likelihood based learning, gradient values can be obtained from

$$\frac{\partial \mathcal{L}}{\partial \theta} = N\left[\sum_j\left[E_{\tilde{P}(x_j, x_{\bar{j}})}\left\langle \frac{\partial \mathbf{F}(x_j, x_{\bar{j}}; \theta)}{\partial \theta} \right\rangle - E_{\tilde{P}(x_{\bar{j}})}\left\langle E_{P(x_j|x_{\bar{j}};\theta)}\left\langle \frac{\partial \mathbf{F}(x_j, x_{\bar{j}}; \theta)}{\partial \theta} \right\rangle \right\rangle\right]\right] \tag{12}$$

## 3 Experiments, Results and Analysis

We are interested in examining the quality of the latent representations obtained when optimizing multi-attribute Harmonium structured models under ML, CL and MCL objectives. We use a similar testing strategy to [9] but focus on comparing the different latent spaces obtained with the different optimization objectives. For our experiments, we use the reduced "20newsgroups" dataset prepared in MATLAB by Sam Roweis. In this data

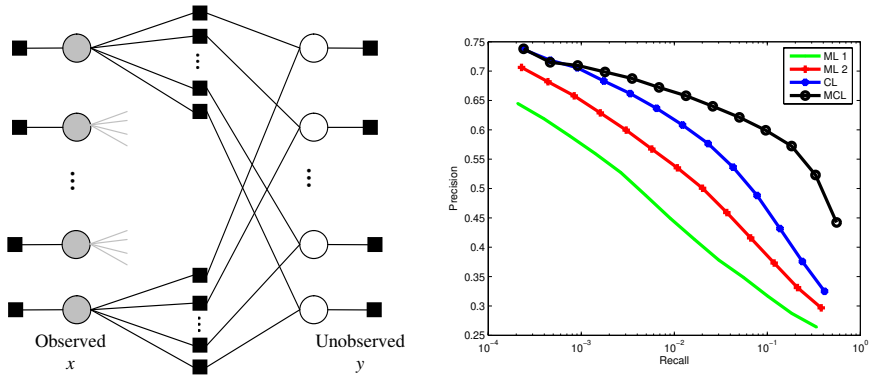


Figure 2: (Left) A factor graph for a Harmonium model. (Right) Precision-recall curves for the "20newsgroups" data using ML, CL and MCL with 20 latent variables. Random guessing is a horizontal line at .25.

set, 16242 documents are represented by 100 word vocabulary binary occurrences and are labeled as one of four domains. To evaluate the quality of our latent space, we retrieve documents that have the same domain label as a test document based on their cosine coefficient in the latent space when observing only binary occurrences. We randomly split data into a training set of 12,000 documents and a test set of 4242 documents. We use joint model with a corresponding full rank multivariate Bernoulli conditional for binary word occurrences and a discrete conditional for domains. Figure 2 shows precision-recall results. ML-1 is our model with no domain label information. ML-2 is optimized with domain label information. CL is optimized to predict domains from words and MCL is optimized to predict both words from domains and domains from words. From Figure 2 we see that the latent space captured by the model is more relevant for domain classification when the model is optimized under the CL and MCL objectives. Further, at low recall both the CL and MCL derived latent spaces produced similar precisions. However, as recall increases the precision for comparisons made in the MCL derived latent space is consistently better. In conclusion, these results lead us to believe that further investigation is warranted into the the use of Multi-Conditional Learning methods for deriving both more meaningful and more useful hidden variable models.

## References


[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[3] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[4] T. Jebara and A. Pentland. On reversing jensen's inequality. *NIPS 13*, 2000.

[5] F. R. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[6] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[7] Ruslan Salakhutdinov, Sam Roweis, and Zoubin Ghahramani. Optimization with EM and expectation-conjugate-gradient. *Proceedings of (ICML)*, 2003.

[8] P. Smolensky. *Information processing in dynamical systems: foundations of harmony theory*, chapter 2, pages 194–281. McGraw-Hill, New York, 1986.

[9] Max Welling, Michal Rosen-Zvi, and Geoffrey Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS17*, pages 1481–1488. 2005.

## 4 Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. This work was also supported in part by Microsoft Research under the Memex and eScience programs. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.