Helsinki University of Technology
Publications in Computer and Information Science
Report E1
February 2006

DISCRIMINATIVE MCMC

Kai Puolamäki Jarkko Salojärvi Eerika Savia Samuel Kaski

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Distribution:
Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
P.O. Box 5400
FI-02015 TKK, Finland
Tel. +358-9-451 3267
Fax +358-9-451 3277

This report is downloadable at
http://www.cis.hut.fi/Publications/

ISBN 951-22-8094-9
ISSN 1796-2803

Discriminative MCMC

Kai Puolamäki, Jarkko Salojärvi, Eerika Savia and Samuel Kaski
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland
firstname.lastname@hut.fi

Abstract

We discuss Bayesian modeling in the case where the model is incorrect. Standard posterior distribution is optimal for inference if the true model is within the model family. In the case of an incorrect model, we show that for inference on conditioned distribution, a different posterior-type distribution is optimal. We provide here an axiomatic justification of previously suggested supervised posterior distribution, introduce Markov Chain Monte Carlo -type methods for computing with the posterior, and demonstrate empirically that it works as expected.

1 Introduction

In generative modeling tasks, it is well-known that usual Bayesian inference is not optimal for generalizing to new data if the model family is incorrect, that is, if the data does not come from any of the models within the model family. Arguably the best solution then is to improve the model family by taking more of the prior knowledge into account. This is not always possible or feasible, however, and simplified models are being generally used, often with good results. There are good reasons for still applying standard Bayesian or Bayesian-style techniques [1] but the general problem of how to best do inference with incorrect model families is still open.

In discriminative modeling, here meaning inference on the distribution $p(y|x)$, the question of using discriminative vs. generative models has attracted a lot of interest. In essence, the question has been whether to model $p(y|x)$ directly or to build a generative model for the joint distribution $p(y, x)$ and compute the conditional distribution from that. It is easy to show that, for instance, the point estimates computed by mazimizing the joint likelihood and the conditional likelihood differ. Maximum conditional likelihood, coupled with suitable regularization to avoid overfitting, works better asymptotically, and it can be optimized with expectation-maximization-type procedures [2]. Some other related point estimates have been proposed but while point estimates have been studied thoroughly, much fewer results exist on trying to extend from point estimates to posterior distributions. The standard posterior distribution is optimal for discriminative modeling if the model family is correct, but is there an extension that would be analogous to standard Bayesian inference but work better for incorrect model families?

We are aware of only one suggestion, the so-called supervised posterior [3], which is known to work well in practice [4]. The posterior has however, as far as we know, only been sug-

gested more or less heuristically. We give an axiomatic justification, introduce Markov Chain Monte Carlo -type methods for computing with the posterior, and demonstrate empirically that it works as expected. In the practical examples of this paper we apply the supervised posterior to data where y is a categorical (class) variable, and thus the task is to discriminate the different classes. In this experimental setup it is more natural to refer to the posterior as discriminative.

There exists another well-established line of research on using Bayesian methods for discriminative learning, namely Bayesian regression, where the x is considered as covariates of the model for y. From the generative modeling perspective such regression ignores any information about y supplied by x. This is justified if (i) the covariates are explicitly chosen when designing the experimental setting and hence are not noisy, or (ii) a different set of parameters generates x and $y|x$, and the sets are assumed to be independent in their prior distribution. Then the posterior factors out into two parts, and the parameters used for generating x are not needed or useful in the regression task. See for instance [5] for more details.

For the purpose of regression, the discriminative posterior makes it possible to use model structures where the parameters are not constrained, that is, any form of generative model. The gained advantage compared to using the standard non-discriminative posterior should be that the predictions should be more accurate assuming the model family is incorrect. Compared to Bayesian regression the predictions should be better if the introduced generative model for x is informative.

2 Theory

2.1 All Distributions are Multinomial

It is useful to work with discrete distributions, where the number of possible data points is finite. Continuous distributions can (almost) always be represented with a discrete distribution, with discretization error. The discretization error can be made arbitrarily small by increasing the number of possible data points within the discrete distribution.

We denote by X the set of possible data points. The most general discrete distribution for X is the multinomial distribution that can be parameterized by probabilities $\theta_i = p(x|\theta)$, $x \in X$, $i = 1, \ldots, |X|$. The dimensionality of the parameter space Θ is thus $\dim(\Theta) = |X| - 1$.

The multinomial distribution describes all distributions for draws from X. The dimensionality of a multinomial distribution is, however, usually too large to be of any practical use. To solve the dimensionality problem we can define a lower-dimensional subspace of the full parameter space $\overline{\Theta} \subseteq \Theta$. Equivalently, we can define a prior distribution $p(\theta)$ for the parameters that satisfies $p(\theta) \neq 0$, $\theta \in \overline{\Theta}$, and $p(\theta) = 0$ otherwise.

Consider for example a Gaussian distribution on a compact real axis, $R = [a, b]$. The probability density function of the Gaussian can be parameterized by

$$p(x|\overline{x}, \tilde{x}) = \frac{1}{Z} \exp\left(-\frac{(x - \overline{x})^2}{2\tilde{x}}\right) \quad , \tag{1}$$

where Z is the normalization factor. At the limit $a, b \to \pm\infty$ the normalization factor obeys $Z \to \sqrt{2\pi\tilde{x}}$, where $\tilde{x} = \sigma^2$ is the variance.

The axis R can be divided into $|X|$ partitions, denoted by $x_i = [c_{i-1}, c_i]$, where $c_0 = a$ and $c_{|X|} = b$ and $x \in X$. We can now approximate the Gaussian with multinomial distribution over X, $x \sim \text{Multinomial}(\theta; |X|)$. The multinomial distribution (without any additional

constrains) has a $|X| - 1$ dimensional parameter space, while the Gaussian has a two-dimensional parameter space. If we are to approximate a Gaussian, we must therefore select an appropriate two-dimensional subset $\overline{\Theta}$ of the parameter space. This subset can be defined by projection,

$$\theta_i = \frac{c_i - c_{i-1}}{Z} \exp\left(-\frac{(y_i - \overline{x})^2}{2\tilde{x}} \right) \quad , \tag{2}$$

where $y_i \in x_i$ and $\overline{x}$ is over real axis and $\tilde{x}$ is over positive-definite real axis.

We can approximate the Gaussian with a multinomial in $\overline{\Theta}$ with arbitrary precision by making the partitions x_i smaller and increasing their number.

Therefore, almost all distributions can be represented, with arbitrary accuracy, as a multinomial distribution in a subset of the full multinomial parameter space.

2.2 The Best Point Solution is Usually not MAP

The multinomial distribution includes all possible distributions for $x \in X$. We can therefore safely assume that our data has been generated by this multinomial distribution with some parameters $\tilde{\theta}$. We do not know $\tilde{\theta}$, but we can assume that there exists such parameters.

The parameter space of any realistic model, $\overline{\Theta}$, is usually much smaller than the parameter space of the full model, Θ. It follows that in almost any practical application the model is incorrect, i.e., $\tilde{\theta} \notin \overline{\Theta}$. Usually in Bayesian statistics one just hopes that the model that generated the data, $\tilde{\theta}$ is close enough to the parameter space of the practical model $\overline{\Theta}$.

Let's first consider the limit where the number of data samples is infinite. In Bayesian statistics, at the limit of infinite data the posterior distribution $g(\theta)$ becomes a point solution, $g(\theta) = \delta(\theta - \hat{\theta})$.[1] At the limit of infinite data the the effect of the prior in $\overline{\Theta}$, in which the prior is non-zero, vanishes. The only effect the prior has is due to the division of the parameter space into $\overline{\Theta}$ vs. $\Theta \setminus \overline{\Theta}$ (non-zero prior vs. zero prior).

The best point solution $\hat{\theta} \in \overline{\Theta}$ (not necessarily the MAP solution) can usually (at least in principle) be found by minimizing some suitable cost function $K(\tilde{\theta}, \theta)$,

$$\hat{\theta} = \arg\min_{\theta \in \overline{\Theta}} K(\tilde{\theta}, \theta) \quad . \tag{3}$$

The form of $K(\tilde{\theta}, \theta)$ depends on the quantity we are approximating.

If we want to have an accurate approximation of the likelihood the suitable cost function can be written in terms of Kullback-Leibler (KL) divergence as

$$K_{MAP}(\tilde{\theta}, \theta) = \sum_{x \in X} p(x|\tilde{\theta}) \log \frac{p(x|\tilde{\theta})}{p(x|\theta)} \quad . \tag{4}$$

If the "model is correct," i.e., $\tilde{\theta} \in \overline{\Theta}$, equation (4) is minimized to zero (at the limit of infinite data) and the resulting point estimate is the known MAP solution. If $\tilde{\theta} \notin \overline{\Theta}$ the resulting point estimate is the best estimate (in KL-sense) for the likelihood, $p(x|\hat{\theta})$. $K_{MAP}(\tilde{\theta}, \hat{\theta})$ describes the KL-divergence between the "true" distribution at $\tilde{\theta}$ and the MAP estimate at $\hat{\theta}$.

However, The MAP estimate may not be optimal if we are interested in approximating some other quantity than the likelihood. Assume, for example, that the data space X can

[1] Actually the posterior can also have many modes (point estimates) and in some special cases the posterior is a manifold in $\overline{\Theta}$.

be expressed as a product $X = Y \times Z$, i.e., all $x \in X$ can be decomposed into $x = (y, z)$, where $y \in Y$ and $z \in Z$. Consider the problem of finding the best point estimate for the conditional likelihood $p(y|z, \theta)$. The conditional likelihood actually has $|Z|$ multinomial distributions for y, $y \sim$ Multinomial$(\theta_z; |Y|)$. The average KL-divergence between the true conditional likelihood at $\tilde{\theta}$ and estimate at θ is given by

$$K_{COND}(\tilde{\theta}, \theta) = \sum_{z \in Z} p(z|\tilde{\theta}) \sum_{y \in Y} p(y|z, \tilde{\theta}) \log \frac{p(y|z, \tilde{\theta})}{p(y|z, \theta)} \quad . \tag{5}$$

The relationship between eqs. (4) and (5) can be written as

$$K_{MAP}(\tilde{\theta}, \theta) - K_{COND}(\tilde{\theta}, \theta) = \sum_{z \in Z} p(z|\tilde{\theta}) \log \frac{p(z|\tilde{\theta})}{p(z|\theta)} \geq 0 \quad . \tag{6}$$

If the model is correct, or $\tilde{\theta} \in \overline{\Theta}$, the solutions are identical. However, if the model does not accurately predict the marginal distribution $p(z|\tilde{\theta})$ (implying $\tilde{\theta} \notin \overline{\Theta}$), the MAP estimate is worse in estimating the conditional likelihood.

Furthermore, we can decompose K_{COND} as

$$K_{COND}(\tilde{\theta}, \theta) = S(\tilde{\theta}) - R(\tilde{\theta}, \theta) \quad , \tag{7}$$

where $S(\tilde{\theta}) = \sum_x p(x|\tilde{\theta}) \log p(y|z, \tilde{\theta})$ and $R(\tilde{\theta}, \theta) = \sum_x p(x|\tilde{\theta}) \log p(y|z, \theta)$ and $x = (y, z)$.

2.3 Posterior Distribution

In real world situations the amount of data is finite. We do not obtain a point estimate, $\hat{\theta}_{COND}$, but a *posterior distribution* over θ, given data D, denoted by $g(\theta|D)$.

We require the posterior distribution to satisfy the following.

1. When no data is observed the posterior equals the prior distribution, $g(\theta|\emptyset) = p(\theta)$.
2. The posterior is multiplicative, i.e., $g(\theta|D) = Z_D^{-1} p(\theta) \prod_{x \in D} h(x, \theta)$, where the data-dependant normalization constant Z_D is chosen so that $\int d\theta g(\theta|D) = 1$, and $h(x, \theta)$ are smooth (C^∞) functions in $\theta \in \overline{\Theta}$.
3. For all $\tilde{\theta} \in \Theta$ and $\theta_1, \theta_2 \in \overline{\Theta}$, the following condition is satisfied:

   $$g(\theta_1|D_{\tilde{\theta}}) \leq g(\theta_2|D_{\tilde{\theta}}) \Leftrightarrow K(\tilde{\theta}, \theta_1) \geq K(\tilde{\theta}, \theta_2) \quad ,$$

   where $D_{\tilde{\theta}}$ is a very large data set sampled from $p(x|\tilde{\theta})$. We further assume that if the equality holds on the other side, it also holds on the other.

Proposition 2.1 *Given these axioms, the discriminative or supervised posterior is of the form*

$$p_d(\theta|D) = \frac{1}{Z_D} p(\theta) \prod_{(y,z) \in D} p(y|z, \theta)^A \quad , \tag{8}$$

where A is an arbitrary positive constant. We can fix $A = 1$, e.g., by requiring that the discriminative posterior reduces to the conventional Bayesian posterior for each fixed z.

The rest of this section is used for the proof of proposition 2.1.

At the limit of $n = |D_{\tilde{\theta}}| \gg 1$, and using the second axiom, the posterior can be written as

$$\log g(\theta|D_{\tilde{\theta}}) = \sum_{x \in D_{\tilde{\theta}}} \log h(x,\theta) + \log p(\theta) - \log Z_{D_{\tilde{\theta}}} \simeq n \sum_{x \in X} p(x|\tilde{\theta}) \log h(x,\theta) - \log Z_{D_{\tilde{\theta}}} \quad . \tag{9}$$

That is, the contribution of the prior $p(\theta)$ can be neglected.

The third axiom states that asymptotically (at the limit of large but finite data set) the shape of the posterior is such that posterior is always smaller if the "distance" $K_{COND}(\tilde{\theta}, \theta)$ is larger.

In other words, given a large data set $D_{\tilde{\theta}}$, the last axiom can be rewritten as

$$\sum_{x \in X} p(x|\tilde{\theta}) \log h(x,\theta_1) \leq \sum_{x \in X} p(x|\tilde{\theta}) \log h(x,\theta_2) \tag{10}$$

$$\Updownarrow$$

$$\sum_{x \in X} p(x|\tilde{\theta}) \log p(y|z,\theta_1) \leq \sum_{x \in X} p(x|\tilde{\theta}) \log p(y|z,\theta_2) \quad , \tag{11}$$

since the normalization terms $Z_{D_{\tilde{\theta}}}$ cancel out each other.

The problem then reduces to finding a functional form for $h(x,\theta)$.

Proposition 2.2 *From axiom 3. it follows that*

$$\log h(x,\theta) = f_C(\log p(y|z,\theta)) \tag{12}$$

where f_C is a monotonically increasing function.

Proof In the appendix.

Furthermore, utilizing both the equality part and the inequality part of axiom 3. we can derive the following proposition

Proposition 2.3 *For a continuous increasing function $f_C(t)$ for which*

$$\log h(x,\theta) = f_C\left(\log p(y|z,\theta)\right)$$

it follows from axiom 3. that

$$f_C(t) = At + \beta \quad ,$$

or, equivalently,

$$h(x,\theta) = \exp(\beta)\ p(y|z,\theta)^A \quad \textit{with} \quad A > 0. \tag{13}$$

Proof In the appendix.

To restrict ourselves further, we set a fourth axiom, which in effect states that our choice of utility function coincides with the posterior probability if the correct model is within the model family.

4. When the true $\tilde{\theta}$ belongs to the model family and the prior distribution is even (i.e. $p(\theta) = \text{const.}$) on the simplex: $\tilde{\theta}_i \in [0,1], \quad \sum_i \tilde{\theta}_i = 1$, for each fixed z the posterior distributions $g(\tilde{\theta} \mid D, z)$ and $p(\tilde{\theta} \mid D, z)$ are equal.

From this axiom we can derive the following proposition

Proposition 2.4 *From axiom 4. it follows that $A = 1$ in (13), which means that*

$$h(x, \theta) = \exp(\beta)\ p(y|z, \theta)$$

with constant β.

Proof In the appendix.

Finally, requiring such normalization that

$$\begin{cases} \inf h(x, \theta) & = 0 \\ \sup h(x, \theta) & = 1 \end{cases}$$

is equivalent to requiring that $\exp(\beta) = 1 \implies \beta = 0$.

So, with this normalization, we get

$$h(x, \theta) = p(y \mid z, \theta) \tag{14}$$

for all possible values of $x = (y, z)$ and for all parameter values θ. □

3 Practice

For point estimates it is known that for the same model structure, in case of an incorrect model, the maximum joint likelihood estimate differs from the conditional likelihood estimate. Algorithms for obtaining point estimates have included gradient ascent-based methods or discriminative versions of the EM algorithm.

So far, MCMC methods have been used for sampling from the posterior of a joint density model, thus making the implicit assumption that the model is very close to being correct. However, as shown in Section 2, in case of an incorrect model, a *discriminative* posterior, at least asymptotically, gives better predictions.

In the following, we discuss discriminative MCMC sampling in the case of predicting the value c of a class variable C, given observations $\mathbf{x}$. We denote the set of paired observations by $D = \{x_i, c_i\}_{i=1}^N$ in the following.

We assume that the discriminative posterior $p_d(\theta|D)$ is

$$p_d(\theta|D) \propto \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \theta)p(\theta). \tag{15}$$

The conditional probability in Eq. (15) can be derived from a joint density model using the Bayes formula:

$$\prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \theta) = \prod_{i=1}^{N} \frac{p(c_i, \mathbf{x}_i|\theta)}{\sum_{c'} p(c', \mathbf{x}_i|\theta)} \quad . \tag{16}$$

The above equations can be used for implementing a Metropolis-Hastings (M-H) sampling scheme.

Example: Mixture Model. As a simple example, we apply the discriminative posterior to obtain predictions from a discriminative joint density mixture model. A discriminative version of a joint density model can be obtained by changing the objective function to conditional likelihood. Consider the likelihood of an ordinary joint density mixture model,

$$\prod_i \sum_j p(\mathbf{x}_i|j, \theta_j)\pi(j), \tag{17}$$

where $\pi(j)$ is the prior probability of selecting the component j, and $p(\mathbf{x}_i|j,\theta_j)$ is the probability distribution of data $\mathbf{x}$, given the component j and the associated parameters θ_j. When given pairwise data $(\mathbf{x},c)_i$, it is common to assume a deterministic mapping from c to mixture component(s) j. In other words, a subset $\mathcal{C}_c$ of values of j is associated with the given class label c. In the simplest case, with one component per class, the task is trivial, since the class label gives us directly the component. The task for test data is then, given $\mathbf{x}$, to predict the component that generated the data.

A discriminative version of a joint density model can be obtained by switching to conditional likelihood

$$\prod_i p(c_i|\mathbf{x}_i,\theta) = \prod_i \frac{\prod_{C_c}\left(\sum_{j\in C_c} p(\mathbf{x}_i|j,\theta_j)\pi(j)\right)^{\delta(c_i,C_c)}}{\sum_{j'} p(\mathbf{x}_i|j',\theta_{j'})\pi(j')}. \tag{18}$$

The delta function in the exponent picks the components associated with the class c_i of the sample i. Equation (18) is plugged in to Eq. (15) to obtain M-H acceptance probabilities.

An alternative way would be to define a mixture model

$$\prod_i p(x_i,c_i|\theta,\psi) = \prod_i \sum_j p(\mathbf{x}_i|j,\theta_j)p(c_i|j,\psi_j)\pi(j). \tag{19}$$

In this case the mixture component assigns a (multinomial) probability distribution over class labels, $p(c|j,\psi_j)$. The corresponding conditional likelihood is then

$$\prod_i p(c_i|x_i,\theta,\psi) = \prod_i \frac{\sum_j p(\mathbf{x}_i|j,\theta_j)p(c_i|j,\psi_j)\pi(j)}{\sum_{j'} p(\mathbf{x}_i|j',\theta_{j'})\pi(j')} \quad . \tag{20}$$

3.1 Convergence of sampling from discriminative posterior

According to [6], in order to be valid, the MCMC sampling has to fulfill these conditions:

1. Simulated sequence is a Markov chain with unique stationary distribution.
2. The stationary distribution equals the discriminative posterior.

The conditions can be shown to hold in a similar manner as for joint likelihood sampling [6].

Condition 1. A Markov chain has a unique stationary distribution, when the chain is irreducible, aperiodic, and not transient. A random walk on any proper distribution is aperiodic and not transient [6].

The dMCMC chain is irreducible (=ergodic), since every value of θ has a non-zero probability of being sampled. The jumping distribution J_t must thus eventually be able to jump to any state with positive probability.

We may thus conclude that the Markov chain has a unique stationary distribution.

Condition 2. Assume that the sample θ^{t-1} at time $t-1$ is drawn from the target distribution $p_d(\theta|D)$. We further assume a labeling of samples such that

$$p_d(\theta_b|D)J_t(\theta_a|\theta_b) \geq p_d(\theta_a|D)J_t(\theta_b|\theta_a) \quad .$$

The joint probability of a transition is now

$$p(\theta^t=\theta_b,\theta^{t-1}=\theta_a) = p_d(\theta_a|D)J_t(\theta_b|\theta_a)\cdot r = p_d(\theta_a|D)J_t(\theta_b|\theta_a) \quad , \tag{21}$$

where r is the M-H probability, $r = \min\left(\frac{p_d(\theta_b|D)J_t(\theta_a|\theta_b)}{p_d(\theta_a|D)J_t(\theta_b|\theta_a)}, 1\right)$. Due to our labeling, we have $r = 1$.

On the other hand,

$$p(\theta^t = \theta_a, \theta^{t-1} = \theta_b) = p_d(\theta_b|D)J_t(\theta_a|\theta_b) \cdot r = p_d(\theta_a|D)J_t(\theta_b|\theta_a) \quad , \tag{22}$$

where $r = \min\left(\frac{p_d(\theta_a|D)J_t(\theta_b|\theta_a)}{p_d(\theta_b|D)J_t(\theta_a|\theta_b)}, 1\right)$. Due to our labeling, we have $r = \frac{p_d(\theta_a|D)J_t(\theta_b|\theta_a)}{p_d(\theta_b|D)J_t(\theta_a|\theta_b)}$.

Since Eq. (21)= Eq. (22), the joint distribution is symmetric. The θ^{t-1}, θ^t thus have same marginal distributions, and so $p_d(\theta|D)$ is the stationary distribution of the Markov chain of θ.

3.2 Predictions from discriminative posterior

Predictions for test data are obtained by

$$p(c|x) \approx \frac{1}{K}\sum_{k=1}^{K} p(c|\theta^{(k)}, x), \tag{23}$$

where K is the number of MCMC samples.

4 Experiments

4.1 Toy example

In order to demonstrate the difference between joint density sampling and discriminative sampling, we constructed a simple example using a mixture of two 1-dimensional Gaussians. The model is defined by

$$p(x) = \sum_j \pi(j)p(x|\mu_j, \sigma_j)\prod_j p(\mu_j|m, s), \tag{24}$$

where $\pi(j)$ is the mixing parameter, in this example fixed to 0.5. The distribution $p(x|\mu_j, \sigma_j)$ is a Gaussian with mean μ_j and standard deviation σ_j. The $p(\mu_j|m, s)$ is a Gaussian prior for μ_j, having hyperparameters $m = 7, s = 7$. The hyperparameters were fixed in this experiment. The index j runs over the mixture components, in our case $j \in \{1, 2\}$.

Toy data were generated from the mixture model using $\mu_1 = 5$, $\mu_2 = 9$, and $\sigma_1 = \sigma_2 = 2$. The data were then labeled according to the generating mixture component j.

The task is to predict the mixture component responsible for generating the data, i.e., the label of the data. In this demonstration we deliberately choose an incorrect model, where the standard deviation is restricted to be the same as the mean, that is, $\sigma_j = \mu_j$.

4.1.1 Sampling

The Metropolis-Hastings sampling scheme was adopted. We implemented two sampling methods: sampling from the standard posterior and from the supervised posterior. The methods were implemented to be as similar as possible; the only difference is in the sample selection criteria. Both methods used the same symmetric proposal distribution $q(\mu^{(new)}|\mu^{(old)})$, namely a Gaussian $\mathcal{N}(\mu^{(old)}, 0.3)$, centered around the old parameter value $\mu^{(old)}$ (and with a standard deviation of 0.3). The difference is in the M-H step where the samples are selected according to the joint likelihood or conditional likelihood criterion.

M-H for joint density sampling The basic principle for computing with mixture models is to introduce unobserved indicators ζ, random variables which specify the mixture component from which each particular observation is drawn [6]. M-H sampling from a joint density mixture model is straightforward, since for training data we know the identity (label) of the generating component, that is, the values of ζ. Since the likelihood of the other components is zero, we therefore sample from the augmented joint density $p(\theta|x,c) \sim p(x,\delta(j,c),\theta)$, where $\delta(\cdot)$ picks the mixture component corresponding to value $C = c$. Notice that also a variant using Gibbs sampling can be applied [6], however we restrict our sampling to M-H in order to have as similar a sampling as in the discriminative MCMC case.

1. Draw a proposal sample $\mu^{(new)} \sim \mathcal{N}(\mu^{(old)}, 0.3)$
2. Accept sample with probability $r \sim \min\left\{1, \frac{\prod_i \prod_j \left(\pi(j)p(x_i|\mu_j^{(new)})\right)^{\delta(j,c_i)} p(\mu_j^{(new)}|m,s)}{\prod_i \prod_j \left(\pi(j)p(x_i|\mu_j^{(old)})\right)^{\delta(j,c_i)} p(\mu_j^{(old)}|m,s)}\right\}$

M-H for conditional density sampling In case of discriminative MCMC, the M-H acceptance probability is the resulting conditional likelihood of the data.

1. Draw a proposal sample $\mu^{(new)} \sim \mathcal{N}(\mu^{(old)}, 0.3)$
2. Accept sample with probability $r \sim \min\left\{1, \frac{\prod_i \prod_j \left(p(j|x_i,\mu_j^{(new)})\right)^{\delta(j,c_i)} p(\mu_j^{(new)}|m,s)}{\prod_i \prod_j \left(p(j|x_i,\mu_j^{(old)})\right)^{\delta(j,c_i)} p(\mu_j^{(old)}|m,s)}\right\}$, where $p(j|x_i,\mu_j) = \frac{\pi(j)p(x_i|\mu_j)}{\sum_j \pi(j)p(x_i|\mu_j)}$.

4.1.2 Results

In the experiments, the training data set size was varied in $N_{Tr} \in \{2,4,6,8,10,12,14,16,20,30,100\}$; the test data set was always 200 samples. For each size of training data, 1000 test and training data sets were generated. MCMC sampling was carried out for each data set for 2000 iterations, with 400 samples as a burn-in period length. After burn-in, every fifth sample was selected. After sampling, the perplexity of the test data set was computed using the retained samples.

Figure 1 shows that the distributions produced by the two sampling methods are different. The joint density sampling concentrates on the area of the support of the data (not visible in the Figure), whereas discriminative sampling obtains parameters outside of the space spanned by data.

For each size of training data sets, 1000 training and test data sets were generated. The number of times the discriminative sampling outperformed the joint density sampling is reported in Fig. 2

In the final experiment, we compared the performance of the MCMC sampling methods in a case where the model is incorrect vs. when the model is correct. The results are shown in Figure 3. In case of an incorrect model the discriminative sampling outperforms joint density sampling, whereas in case of a correct model the performance of both methods are roughly equal with discriminative sampling being slightly worse for small data sets.

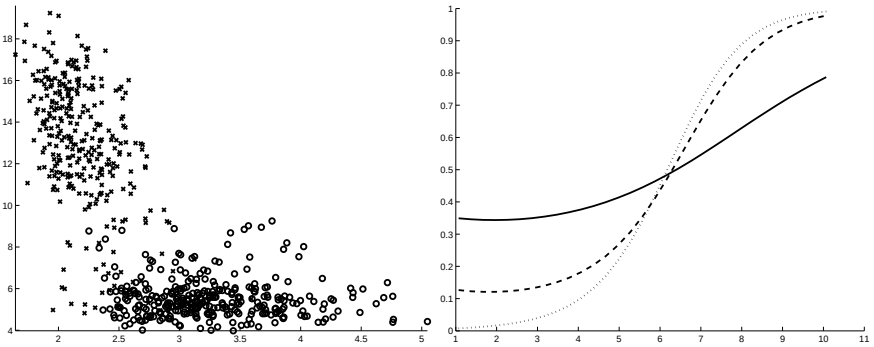

Figure 1: MCMC sampling in the case of an incorrect model. Left: MCMC samples, plotted by μ_1 (horizontal axis) against μ_2 (vertical). 'x' – samples from discriminative sampling, 'o' – samples from joint likelihood sampling. The number of data points in the learning set was 30. Right: Plot of conditional density $p(c|x)$ (vertical axis) as a function of the value of x (horizontal axis). Solid line: joint likelihood sampling, Dashed: discriminative sampling, dotted: true conditional density. True values were $\mu_1 = 5,\ \mu_2 = 9$.

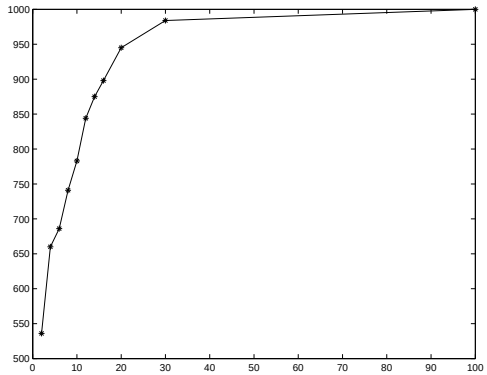

Figure 2: The number of times (out of a total of 1000) that a discriminative sampling method resulted in better perplexity than joint likelihood sampling for a test data set, plotted as a function of training set size.

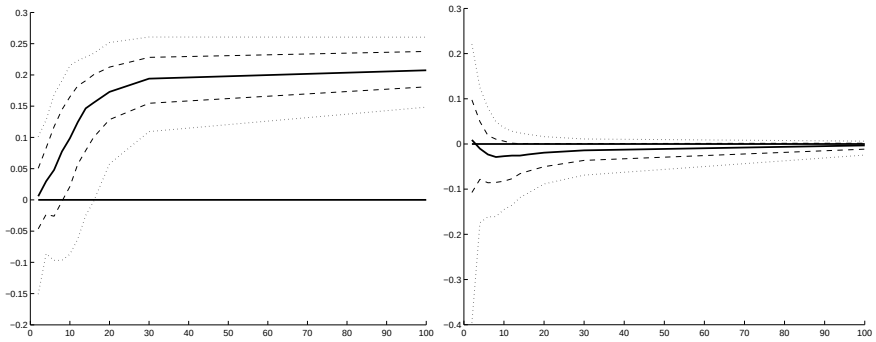

Figure 3: Difference of discriminative vs. joint density sampling when (left) the data was modeled with an incorrect model, or (right) the correct model was used. Difference between perplexities was computed for 1000 different test sets, with models learned for 1000 separate training sets. When the difference is positive, discriminative method is better. The sample size of the training set was varied by $N_{Tr} \in \{2, 4, 6, 8, 10, 12, 14, 16, 20, 30, 100\}$ (x-axis). The plot shows the median difference (solid line) between the test set perplexities, along with 25% and 75% fractiles (dashed lines), and 10% and 90% fractiles (dotted lines).

4.2 Discriminative document modeling

As a practical application we consider the task of predicting the category of a given document. We used the Reuters data set [7], of which we selected a subset of 400 documents from four categories, 100 from each category. The categories were:

1. CCAT: Corporate-Industrial
2. ECAT: Economics and Economic Indicators
3. GCAT: Government and Social
4. MCAT: Securities and Commodities Trading and Markets

The selected documents were each classified to only one of the four classes. The words that occurred less than 10 times in the whole subset were left out, thus leaving 2144 words. The data set was then split into equal-sized training and test sets.

4.2.1 Mixture of Unigrams Model

We apply the mixture of unigrams model (MUM) in our experiments. The MUM [8] is a hidden variable model that generates word counts for documents. The model assumes that each document is generated from a mixture of M hidden "topics," $\sum_{j=1}^{M} \pi(j) p(\mathbf{x}_i|\beta_j)$, where j is the index of the topic, and β_j the multinomial parameters that generate words from the topic. The vector $\mathbf{x}_i$ is the observed word counts for document i, and $\pi(j)$ the probability of generating the words from the topic j. Notice that the model is clearly incorrect, since it places unrealistic assumptions on the documents; it assumes that each document is generated from only one topic.

The full joint density of the MUM is

$$p(D_x, \alpha, \pi|\gamma, \beta_{1\ldots J}) = \prod_i \left(\sum_j p(\mathbf{x}_i|\alpha_j)\pi(j) \right) \prod_j p(\alpha_j|\beta) p(\pi|\gamma) \quad , \tag{25}$$

where $D_x = \{\mathbf{x}_i\}_{i=1}^N$ is the data. We assume a uniform prior for β, γ. The generating process of the MUM is

- Draw the component probabilities from a Dirichlet, $\pi \sim p(\pi|\gamma)$.
- For each component j: Draw word probabilities α_j from a Dirichlet $p(\alpha_j|\beta)$.
- For each document i:
  - Draw a sample vector ζ_i from Multin$(1, \pi)$ (i.e., pick one component).
  - Draw the document vector $\mathbf{x}_i$ from Multin$(n_i, \alpha_{\delta(\zeta_i,1)})$.

We assume here that the number of words n_i in a document i is given. We indicate the sampled component using a vector ζ_i, consisting of zeros and a one at ζ_{ij}.

The usual appproach for modeling paired data $\{\mathbf{x}_i, c_i\}_{i=1}^N$ by a joint density mixture model is to associate c with the label of a mixture component from which the data is assumed to be generated. In its simplest form the discriminative version of MUM thus contains one topic vector per class (and thus, for teaching data, the component that generated that data is known). When the number of components per class grows we modify our model to

$$p(D_x, \alpha, \pi|\gamma, \beta_{1\ldots J}) = \prod_i \left(\sum_j p(\mathbf{x}_i|\alpha_j)\pi(j) \right) \prod_c \prod_{j \in \mathcal{C}_c} p(\alpha_j|\beta_c) p(\pi|\gamma) \quad , \tag{26}$$

that is, we have a different prior β_c for components within each class, $c \in \{1 \ldots C\}$.

4.2.2 Joint likelihood MCMC for a mixture of unigrams model

We next consider joint density MCMC sampling for the simple case of a multinomial mixture model. We implement the Gibbs sampling scheme presented in [6]. In sampling step 1 we form a Jensen lower bound of the posterior of the mixture model:

$$\sum_j \pi(j) p(x_i|\beta_j) \geq \prod_i \left(\pi(j) p(x_i|\beta_j)\right)^{z_{ij}} \quad , \tag{27}$$

where z_{ij} is the posterior probability $p(j|x_i, \alpha)$ of the data $\mathbf{x}_i$ being generated from component j. The lower bound approximation makes the model parameters separable so that β and π can be drawn by Gibbs sampling. For the case of MUM, Gibbs sampling will then proceed as follows:

- Compute $z_{ij} = \frac{\pi(j) p(\mathbf{x}(i)|\alpha_j)}{\sum_{j'} \pi(j') p(\mathbf{x}(i)|\alpha_{j'})}$
- Draw $\zeta_i \sim$ Multin$(1, z_i)$
- Draw $\alpha_j \sim \prod_i p(\mathbf{x}(i)|\alpha_j)^{\delta(\zeta_{ij}, c_i)} p(\alpha_j|\beta_j) =$ Dirichlet $(\beta_j + \sum_i \delta(\zeta_{ij}, c_i)\mathbf{x}_i)$
- Draw $\pi \sim \prod_i \pi(j)^{\delta(\zeta_{ij}, c_i)} p(\pi|\gamma) =$ Dirichlet$(\gamma + \sum_i \delta(\zeta_{ij}, c_i)\zeta_i)$.
- Draw proposal $\beta_{c,new} \sim \mathcal{N}(\beta_{c,old}, \sigma_\beta)$, accept with M-H probability $\min\left(\frac{\prod_{j \in \mathcal{C}_c} p(\alpha_j|\beta_{c,new}) q(\beta_{old}|\beta_{new})}{\prod_{j \in \mathcal{C}_c} p(\alpha_j|\beta_{c,old}) q(\beta_{new}|\beta_{old})}, 1\right)$.
- Draw proposal $\gamma_{new} \sim \mathcal{N}(\gamma_{old}, \sigma_\gamma)$, accept with M-H probability $\min\left(\frac{p(\pi|\gamma_{new}) q(\gamma_{old}|\gamma_{new})}{p(\pi|\gamma_{old}) q(\gamma_{new}|\gamma_{old})}, 1\right)$,

where $q(\beta_{new}|\beta_{old})$, $q(\gamma_{old}|\gamma_{new})$ are proposal kernels for β, γ, respectively. In the experiments we applied a Gaussian proposal kernel, which is symmetric and thus cancels out

from M-H acceptance probabilities. In case of several components per class the delta function $\delta(\zeta_{ij}, c_i)$ picks those data items i where the mapping from the sampled component j to a corresponding class label ($j \in \mathcal{C}_c$) is correct.

It is feasible to choose one of the classes to be known explicitly (i.e., we do not sample ζ for those samples). All those samples that have $\delta(\zeta_{ij}, c_i) = 0$ (that is, we have sampled an incorrect component for the sample) are then assumed to belong to that one class. The associated hypothesis is that the classes cannot be separated, and will thus be pooled into one large class. This alternative, however, was not applied.

After sampling, class predictions for new data can be obtained by

$$\mathrm{E}_{p(\theta|D)}\{c\} = \int p(c|\mathbf{x},\theta)p(\theta|D)d\theta \approx \frac{1}{K}\sum_{i=1}^{K}\frac{p(c,\mathbf{x}|\theta^{(k)})}{\sum_c p(c,\mathbf{x}|\theta^{(k)})}, \tag{28}$$

where K is the number of samples from posterior and $\theta^{(k)}$ is the posterior sample of parameters at sampling iteration k.

4.2.3 Discriminative MCMC for a mixture of unigrams model

Not all samples obtained from the posterior of the joint model are optimal for averaging over $p(c|x)$. This happens in the case where our model is incorrect.

Discriminative sampling from $p(\theta|D)$ needs to be carried out using Metropolis-Hastings (M-H) algorithm, since unlike the joint MCMC case, a Jensen lower bound approximation is now not available. In the M-H algorithm, we assume that the discriminative posterior (15) is

$$p(\theta|D) \propto \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \alpha_{1\ldots j}, \pi) \prod_{j=1}^{M} p(\alpha_j|\beta_j)p(\beta_j)p(\pi|\gamma)p(\gamma)\,, \tag{29}$$

where

$$p(c_i|\mathbf{x}_i, \alpha_{1\ldots j}, \pi) = \frac{\prod_{C_c}\left(\sum_{j\in C_c}\pi(j)p(\mathbf{x}_i|\alpha_j)\right)^{\delta(c_i,C_c)}}{\sum_{j'}\pi(j')p(\mathbf{x}_i|\alpha_{j'})}. \tag{30}$$

Here C_c denotes the states associated with the class c. The delta function at the exponent thus picks the components associated with the class c_i of the sample i.

For the case of the mixture of unigrams model, the discriminative M-H sampling then proceeds as follows:

1. For each α_j: Draw a proposal α^{new}, accept with probability $\min\left(\frac{\prod_i p(c_i,\alpha_j^{new}|\beta,\alpha,\mathbf{x}_i)q(\alpha_j^{old}|\alpha_j^{new})}{\prod_i p(c_i,\alpha_j^{old}|\beta,\alpha,\mathbf{x}_i)q(\alpha_j^{new}|\alpha_j^{old})}, 1\right)$.
2. Draw a proposal π^{new}, accept with probability $\min\left(\frac{\prod_i p(c_i,\pi^{new}|\gamma,\mathbf{x}_i)q(\pi^{old}|\pi^{new})}{\prod_i p(c_i,\pi^{old}|\gamma,\mathbf{x}_i)q(\pi^{new}|\pi^{old})}, 1\right)$.
3. For each γ_j: Draw $\gamma^{(new)} \sim \text{Gauss}(\gamma^{(old)}, \sigma_\gamma^2)$, such that $\gamma^{(new)} \in [0.01,\ 9.99]$. Accept with probability $\min(\frac{p(\pi|\gamma^{new})}{p(\pi|\gamma^{old})}, 1)$.
4. For each β_{ij}: Draw $\beta^{(new)} \sim \text{Gauss}(\beta^{(old)}, \sigma_\beta^2)$, such that $\beta^{(new)} \in [0.01,\ 9.99]$. Accept with probability $\min(\frac{p(\alpha|\beta^{new})}{p(\alpha|\beta^{old})}, 1)$.

The M-H acceptance probabilities above result from canceling the common terms in the numerator and denominator, both formed using Eq. (29).

Class predictions for test data are obtained by

$$p(c|\mathbf{x}) \approx \frac{1}{K} \sum_k p(c|\theta^{(k)}, \mathbf{x}), \tag{31}$$

and using Equation (30).

4.2.4 Proposal (Jump) distribution

Notice that the proposal distributions in the above formulas may be of any form. In case of an incorrect model it is typical that the model parameters of a discriminative model lie outside of the support of the sufficient statistics of the data. This can be seen for example in the toy data, Section 4.1. It is therefore not recommended to use a proposal distribution using sufficient statistics of the data. The best alternative is then to draw samples around previous sample of parameter values. In the experiments the proposal distributions are

$$\begin{aligned} \alpha_j^{new} &\sim \exp\left\{\mathcal{N}(\log \alpha_j^{old}, 0.01)\right\} - Z_{\alpha_j} \\ \pi^{new} &\sim \exp\left\{\mathcal{N}(\log \pi^{old}, 0.01)\right\} - Z_{\pi}, \end{aligned} \tag{32}$$

where $Z_{\alpha_j},\ Z_\pi$ are constants such that $\sum \alpha_j^{new} = 1$ and $\sum \pi^{new} = 1$, respectively.

4.2.5 Including test data into sampling stage

If test data is included into the sampling process, we first need to sample the class of each of the test data items before the M-H steps. This adds two preliminary steps to the above sampling scheme:

1. Compute $z_{ij}^{test} = \frac{\pi(j)p(x_i|\alpha_j)}{\sum_{j'} \pi(j')p(x_i|\alpha_{j'})}$, where $j = 1 \dots M$.
2. Draw $\mathbf{c}_i^{test} \sim \text{Multin}(1, \mathbf{z}_i^{test})$.

Class predictions for test data can then be obtained directly by

$$p(c^{test}|x_i) \approx \frac{1}{K} \sum_k z_{i\cdot}^{test,(k)}. \tag{33}$$

4.3 Results

For discriminative MCMC, the Metropolis-Hastings sampler was run for 4100 iterations, with 100 burn-in iterations. After burn-in, every 80th sample was retained. Discriminative sampling was initiated from the maximum conditional likelihood point estimate.

Joint MCMC was implemented as a Gibbs sampler, run for 210 iterations. The burn-in was 10 iterations. After burn-in every 4th sample was retained. Sampling was initiated from the maximum likelihood estimate of a naive Bayes model.

We measure the model performance in terms of perplexity,

$$\text{perplexity} = e^{-\frac{\mathcal{L}}{N}}\ ,\ \text{where } \mathcal{L} = \sum_{i=1}^{N} \log p(c_i|x_i)\ .$$

Here $p(c_i|x_i)$ is the MCMC prediction, and N is the size of the test set. The perplexities are reported in Table 1. Judging from the results, we can conclude that the MCMC sampling works. Notice that with 1 component per class the mixture of unigrams model corresponds to a logistic regression model where the parameter space is constrained [9]. The model thus has one global maximum which seems to perform fairly well for test data. However, in the multi-component case the point estimate clearly overfits whereas the MCMC avoids this problem, as expected.

Table 1: Classification accuracies for discriminative pLSI (d-pLSI) and naive Bayes classifier for a test set of 200 documents from four different classes.

Method	Perplexity
point estimate, 1 component/class	2.94
point estimate, 5 components/class	3.96
joint MCMC, 1 component/class	4.55
joint MCMC, 5 components/class	4.45
discriminative MCMC, 1 component/class	3.52
discriminative MCMC, 5 components/class	1.95

5 Discussion

The discriminative posterior makes it possible to take generative modeling of x into account in conditional modeling of $y|x$. In many applications "unlabeled" samples of x are common whereas obtaining labeled samples, pairs (x, y) is costly. The problem of using the unlabeled samples in a discriminative task has been coined semisupervised learning. The next question is whether the unlabeled samples could be useful in connection with the discriminative posterior. Hansen [10] has given a Bayesian treatment of semisupervised learning; in our future work we will study in detail whether the discriminative posterior could be embedded in this idea.

References

[1] L. K. Hansen. Bayesian averaging is well-temperated. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 265–271. MIT Press, 2000.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

[3] Peter Grünwald, Petri Kontkanen, Petri Myllymäki, Teemu Roos, Henry Tirri, and Hannes Wettig. Supervised posterior distributions. presentation at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 2002. `http://homepages.cwi.nl/~pdg/presentationpage.html`.

[4] Jesús Cerquides and Ramon López Mántaras. Robust Bayesian linear classifier ensembles. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, Lecture Notes in Artificial Intellligence 3720, pages 72–83, Berlin, Germany, 2005. Springer-Verlag.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL, 1995.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[7] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[8] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

[9] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. On discriminative joint density modeling. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo,

editors, *Machine Learning: ECML 2005*, Lecture Notes in Artificial Intellligence 3720, pages 341–352, Berlin, Germany, 2005. Springer-Verlag.

[10] Lars Kai Hansen. How useful are unlabeled examples for supervised learning? Unpublished manuscript, March 2002.

A Appendix: Proofs

A.1 Mapping from $p(y \mid, z, \theta)$ to $h(x, \theta)$ is monotonically increasing

Proposition A.1 *From axiom 3. it follows that*

$$\log h(x, \theta) = f_C(\log p(y|z, \theta)) \tag{34}$$

where f_C is a monotonically increasing function.

Proof Denoting $\tilde{\theta}_x = p(x|\tilde{\theta})$ in inequalities (10) and (11) we can write in the following form

$$\sum_{x \in X} \tilde{\theta}_x \ \log h(x, \theta_1) \leq \sum_{x \in X} \tilde{\theta}_x \ \log h(x, \theta_2) \tag{35}$$

$$\Updownarrow$$

$$\sum_{x \in X} \tilde{\theta}_x \ \log p(y|z, \theta_1) \leq \sum_{x \in X} \tilde{\theta}_x \ \log p(y|z, \theta_2) \quad . \tag{36}$$

Consider the points in the parameter space Θ, where $\tilde{\theta}_k = 1$ and $\tilde{\theta}_i = 0$ for $i \neq k$ ("corner points"). In these points the linear combinations vanish and the equivalent inequalities (35) and (36) become

$$\begin{cases} \log h(x_k, \theta_1) \leq \log h(x_k, \theta_2) \\ \Updownarrow \\ \log p(y_k|z_k, \theta_1) \leq \log p(y_k|z_k, \theta_2) \quad , \end{cases} \quad . \tag{37}$$

Since the functional form of f_C must be the same regardless of the choice of $\tilde{\theta}$, equivalence (37) holds everywhere in the parameter space, not just in the corners.

From the equivalence (37) (and the symmetry of the models with respect of re-labeling the data items) it follows that $h(x, \theta)$ must be of the form

$$\log h(x, \theta) = f_C(\log p(y|z, \theta))$$

where f_C is a monotonically increasing function. □

A.2 Mapping is of form $h(x,\theta) = c\, p(y \mid z,\theta)^A$

Proposition A.2 *For a continuous increasing function $f_C(t)$ for which*

$$\log h(x,\theta) = f_C\left(\log p(y|z,\theta)\right)$$

it follows from axiom 3. that

$$f_C(t) = At + \beta \quad ,$$

or, equivalently,

$$h(x,\theta) = \exp(\beta)\; p(y|z,\theta)^A \quad \textit{with} \quad A > 0. \tag{38}$$

Proof Consider any $\tilde{\theta}$, and the set of points θ that satisfy $R(\tilde{\theta},\theta) = t$, where t is some constant. From the last axiom (the equality part) it follows that there must exist a constant $f_{\tilde{\theta}}(t)$ that defines the same set of points θ, defined by $\sum_{x\in X} p(x|\tilde{\theta}) \log h(x,\theta) = f_{\tilde{\theta}}(t)$. From the inequality part of the same axiom it follows that $f_{\tilde{\theta}}$ is a monotonically increasing function. Hence,

$$\sum_{x\in X} p(x|\tilde{\theta}) \log h(x,\theta) = f_{\tilde{\theta}}\left(\sum_{x\in X} p(x|\tilde{\theta}) \log p(y|z,\theta)\right) \quad . \tag{39}$$

On the other hand, from (12) we know that we can write

$$\log h(x,\theta) = f_C\left(\log p(y|z,\theta)\right) \quad . \tag{40}$$

So, (39) and (40) lead to

$$f_{\tilde{\theta}}\left(\sum_{x\in X} p(x|\tilde{\theta}) \log p(y|z,\theta)\right) = \sum_{x\in X} p(x|\tilde{\theta}) f_C\left(\log p(y|z,\theta)\right) \quad . \tag{41}$$

If we make a variable change $u_i = \log p(y_i \mid z_i,\theta)$ and denote $p(x_i|\tilde{\theta}) = \tilde{\theta}_i$ for short equation (41) becomes

$$f_{\tilde{\theta}}\left(\sum_i \tilde{\theta}_i\; u_i\right) = \sum_i \tilde{\theta}_i\; f_C\left(u_i\right) \quad . \tag{42}$$

Not all u_i are independent, however: for each fixed z, one of the variables u_l is determined by the other u_i's

$$\exp(u_i) = p(y_i \mid z_i,\theta) \quad ,$$

and

$$\sum_{fixed\ z} p(y_i \mid z,\theta) = 1 \Leftrightarrow \sum_{fixed\ z} \exp(u_i) = 1 \quad .$$

So the last u_l for each z is

$$u_l = \log\left(1 - \sum_{\substack{fixed\ z \\ indep.\ u_m}} \exp(u_m)\right) \quad , \tag{43}$$

where the sum only includes the independent varibles u_m for the fixed z.

This way we can make the dependency on each u_i explicit in equation (42):

$$f_{\tilde{\theta}}\left[\sum_{indep.\ u_j}\tilde{\theta}_j\ u_j + \sum_{dependent\ u_l}\tilde{\theta}_l\ \log\left(1 - \sum_{\substack{fixed\ z\\ indep.\ u_m}}\exp(u_m)\right)\right]$$

$$= \sum_{indep.\ u_j}\tilde{\theta}_j\ f_C(u_j) + \sum_{dependent\ u_l}\tilde{\theta}_l\ f_C\left(\log\left(1 - \sum_{\substack{fixed\ z\\ indep.\ u_m}}\exp(u_m)\right)\right) \quad . \quad (44)$$

Let us differentiate both sides with respect to a u_k:

$$\underbrace{f'_{\tilde{\theta}}\left(\sum_i \tilde{\theta}_i\ u_i\right)}_{\alpha}\left[\tilde{\theta}_k - \underbrace{\frac{\tilde{\theta}_l}{u_l}}_{c_z}\exp(u_k)\right]$$

$$= \quad \tilde{\theta}_k\ f'_C(u_k) - \underbrace{\frac{\tilde{\theta}_l}{u_l}}_{c_z}\ \underbrace{f'_C(u_l)}_{d_z}\ \exp(u_k) \quad . \quad (45)$$

For all such variables u_k that share the same z, we get

$$\alpha\ \left[\tilde{\theta}_k - c_z\ \exp(u_k)\right] = \tilde{\theta}_k\ f'_C(u_k) - c_z\ d_z\ \exp(u_k)$$

$$\Updownarrow$$

$$\tilde{\theta}_k\ \exp(-u_k)\ (f'_C(u_k) - \alpha) = c_z\ (d_z - \alpha) \quad .$$

Since the right-hand side only depends on z, not on individual u_k, the left-hand side must also only depend on z and the factors depending on u_k must cancel out.

$$f'_C(u_k) - \alpha = B_z\ \frac{\exp(u_k)}{\tilde{\theta}_k}$$

$$\Updownarrow$$

$$\underbrace{f'_{\tilde{\theta}}\left(\sum_i \tilde{\theta}_i\ u_i\right)}_{does\ not\ depend\ on\ u_k} = \underbrace{f'_C(u_k) - B_z\ \frac{\exp(u_k)}{\tilde{\theta}_k}}_{depends\ on\ u_k} \quad .$$

Since the left-hand side depends neither on u_k nor z, both sides must be constant

$$\implies f'_{\tilde{\theta}}(t) = A$$

$$\implies f_{\tilde{\theta}}(t) = A\ t + \beta \quad . \quad (46)$$

Subsituting (46) to (41) we get

$$A\ f_{\tilde{\theta}}\left(\sum_i \tilde{\theta}_i\ u_i\right) + \beta = \sum_i \tilde{\theta}_i\ f_C(u_i)$$

$$\Updownarrow$$

$$\sum_i \tilde{\theta}_i\ \left(A\ u_i - f_C(u_i)\right) = -\beta \quad ,$$

and since this must hold for any parameters $\tilde{\theta}$, it must also hold for the corner points:

$$\begin{aligned} & A\ u_i - f_C(u_i) = -\beta \\ \implies & f_C(t) = A\ t + \beta \quad . \end{aligned} \tag{47}$$

□

A.3 Axiom 4. implies exponent $A = 1$

Proposition A.3 *From axiom 4. it follows that $A = 1$ in (13), which means that*

$$h(x, \theta) = \exp(\beta)\ p(y|z, \theta) \tag{48}$$

with constant β.

Proof Let us first write down an expression for $g(\tilde{\theta} \mid D, z)$:

$$g(\tilde{\theta} \mid D, z) = \frac{p(\tilde{\theta})}{Z_h} \prod_{j=1}^{|D_z|} h(x_j, \tilde{\theta}) = \frac{p(\tilde{\theta})}{Z_h} \prod_{x \in X_z} p(y_j \mid z, \tilde{\theta})^{A\, n_x} \quad , \tag{49}$$

where

$$X_z = \{x_j = (y_j, z_j) \mid z_j = z\} \quad .$$

Now the data set that contributes to the posterior with fixed z is

$$D_z = \{x_j \in D \mid z_j = z\} \quad .$$

Taking logarithms on both sides of equation (49) yields

$$\begin{aligned} \log g(\tilde{\theta} \mid D, z) \quad & = \log p(\tilde{\theta}) - \log Z_h + |D_z|\ \textstyle\sum_{x \in X_z} A\ \frac{n_x}{|D_z|}\ \log p(y_j \mid z, \tilde{\theta}) \\ & = \log p(\tilde{\theta}) - \log Z_h + A\,|D_z|\ \textstyle\sum_{x \in X_z} p(y \mid \tilde{\theta}, z)\ \log h(x, \tilde{\theta}) \quad , \end{aligned} \tag{50}$$

where

$$Z_h = \underbrace{c}_{p(\tilde{\theta})} \left[\int_\theta \prod_{x \in X_z} p(y_j \mid z, \theta)^{A\, n_x}\ d\theta\right] \quad .$$

Whereas for the posterior $p(\tilde{\theta} \mid D, z)$ we get

$$p(\tilde{\theta} \mid D, z) = \frac{p(\tilde{\theta})}{Z_p} \prod_{j=1}^{|D_z|} p(y_j \mid z, \tilde{\theta}) = \frac{p(\tilde{\theta})}{Z_p} \prod_{x \in X_z} p(y \mid z, \tilde{\theta})^{n_x} \quad . \tag{51}$$

Again, taking logarithms on both sides yields

$$\begin{aligned}\log p(\tilde{\theta} \mid D, z) &= \log p(\tilde{\theta}) - \log Z_p + |D_z| \ \textstyle\sum_{x \in X_z} \frac{n_x}{|D_z|} \ \log p(y \mid z, \tilde{\theta}) \\ &= \log p(\tilde{\theta}) - \log Z_p + |D_z| \ \textstyle\sum_{x \in X_z} p(y \mid \tilde{\theta}, z) \ \log p(y \mid z, \tilde{\theta}) \quad ,\end{aligned} \tag{52}$$

where

$$Z_p = \underbrace{c}_{p(\tilde{\theta})} \left[\int_\theta \prod_{x \in X_z} p(y_j \mid z, \theta)^{n_x} \ d\theta \right] \quad .$$

Let us now demand that expressions (50) and (52) are equal for a fixed z as it says in axiom 4.

$$\begin{gathered}\log p(\tilde{\theta}) - \log Z_h + A \, |D_z| \ \textstyle\sum_{x \in X_z} p(y \mid \tilde{\theta}, z) \ \log p(y \mid z, \tilde{\theta}) \\ = \log p(\tilde{\theta}) - \log Z_p + |D_z| \ \textstyle\sum_{x \in X_z} p(y \mid \tilde{\theta}, z) \ \log p(y \mid z, \tilde{\theta}) \\ \Updownarrow \\ (A-1) \underbrace{\sum_{x \in X_z} p(y \mid \tilde{\theta}, z) \ \log p(y \mid z, \tilde{\theta})}_{\text{depends on } \tilde{\theta}} = \underbrace{\frac{1}{|D_z|} \log \left(\frac{Z_h}{Z_p} \right)}_{\text{does not depend on } \tilde{\theta}} \quad .\end{gathered} \tag{53}$$

Note that the right-hand side does not depend on the parameters $\tilde{\theta}$, only in the data set D and the functional forms of h and p. On the left-hand side we have an entropy that does depend on $\tilde{\theta}$. From this it follows that both sides equal to zero, resulting in $A = 1$. Let us substitute the formula of f_C into equation (40)

$$\log h(x, \theta) = f_C(\log p(y \mid z, \theta)) = A \ \log p(y \mid z, \theta) + \beta = \log p(y \mid z, \theta) + \beta \quad . \tag{54}$$

Taking exponentials on both sides gives

$$h(x, \theta) = \exp(\beta) p(y \mid z, \theta) \quad . \tag{55}$$

□