Mach Learn (2012) 86:209–231
DOI 10.1007/s10994-011-5262-7

Linear classifiers are nearly optimal when hidden variables have diverse effects

Nader H. Bshouty · Philip M. Long

Received: 8 November 2010 / Accepted: 12 August 2011 / Published online: 5 September 2011

Abstract We analyze classification problems in which data is generated by a two-tiered random process. The class is generated first, then a layer of conditionally independent hidden variables, and finally the observed variables. For sources like this, the Bayes-optimal rule for predicting the class given the values of the observed variables is a two-layer neural network. We show that, if the hidden variables have non-negligible effects on many observed variables, a linear classifier approximates the error rate of the Bayes optimal classifier up to lower order terms. We also show that the hinge loss of a linear classifier is not much more than the Bayes error rate, which implies that an accurate linear classifier can be found efficiently.

Keywords Learning theory · Bayes-optimal · Linear classification · Hidden variables

1 Introduction

In many classification problems, groups of features are positively associated, even among examples of a given class. For example, when classifying news articles as to whether they are about sports or not, words about soccer tend to appear in the same articles. Similarly, diseases often coordinately affect the production rates of members of biomolecular pathways.

One way to model this phenomenon is to use a probability distribution with hidden variables (Hofmann 2001; Blei et al. 2003; Zhang 2004a; Papadimitriou et al. 2000; Langseth and Nielsen 2006). In one model of this type, the class designation directly and conditionally independently affects the hidden variables, each of which in turn drives a set

Editor: Nicolo Cesa-Bianchi.

N.H. Bshouty
Department of Computer Science, Technion, 32000 Haifa, Israel
e-mail: bshouty@cs.technion.ac.il

P.M. Long (✉)
Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
e-mail: plong@google.com

of observed variables (see Fig. 1). Each hidden variable can be interpreted as indicating whether a group of observed variables have been collectively affected by the class of the item. For example, a hidden variable could indicate whether an article is about soccer or not. Its descendents would include words that are especially common in articles about soccer, like "corner" and "striker". It is intuitive that the Bayes optimal classifier for a source like this is a two-layer feed-forward neural network, with the hidden layer of the neural network corresponding to the layer of hidden variables in the generative model. (We provide a proof because we are not aware of a reference for this in the literature.)

Despite this fact, for many problems clearly possessing such hierarchical structure, learning algorithms that use linear hypotheses achieve excellent, often even state-of-the-art, performance (see, e.g. Joachims 1998; Schapire and Singer 2000; Tibshirani et al. 2002; Shalev-Shwartz et al. 2007; Hsieh et al. 2008). This might appear paradoxical, because one might think that such algorithms must be doomed to fail because they use an inordinately limited hypothesis space.

In this paper, we show that, despite the fact that the optimal classifier has a more complex structure, a linear classifier can provide a good approximation. We also show that the hinge loss of the linear classifier is not much more than the Bayes error rate; this can be combined with known tools (Zhang 2004b; Bartlett et al. 2006) to imply that nearly optimal accuracy can be obtained efficiently. Both results hold when the hidden variables influence many observed variables—this is to be expected for example in text classification problems, where subtopics may have many constituent words.

Here is the rough idea of the proof. When a hidden variable affects many observed features, a linear combination of those observed features should be expected to be highly concentrated—the combination will be close to one value when the hidden variable takes one value, and close to another value when the hidden variable takes the other value. Consequently, this linear combination of the observed features can be viewed as an approximation to a rescaling of the hidden variable. If we replace each hidden variable with the appropriate linear combination of the observed variables that it affects, and construct a linear classifier using the replacements, the result is a linear classifier of the original features.

2 Definitions

2.1 The structure of the source

In a *hidden variable model*, the joint distribution of the class label Y, hidden variables $H_1, \ldots, H_k$, and observed variables $X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k}$ (all of which take values in $\{-1, 1\}$) satisfies the conditional independence constraints shown in the Bayes Net of Fig. 1. The hidden variables $H_1, \ldots, H_k$ are collectively conditionally independent given the class designation Y. Each hidden variable H_i in turn has a collection of observed variables $X_{i1}, \ldots, X_{im_i}$ that are conditionally independent given H_i.

We can think of the model as generating labeled random examples $(\mathbf{x}, y)$ in stages, by

- generating the class label y, and fixing it, then
- independently sampling the hidden variables $h_1, \ldots, h_k$ using the appropriate class conditional distribution, fixing them, and finally
- independently sampling the observed variables

$$x_{11}, \ldots, x_{1m_1}, \ldots, x_{k1}, \ldots, x_{km_k},$$

each from the appropriate conditional distribution given the values of its parent.

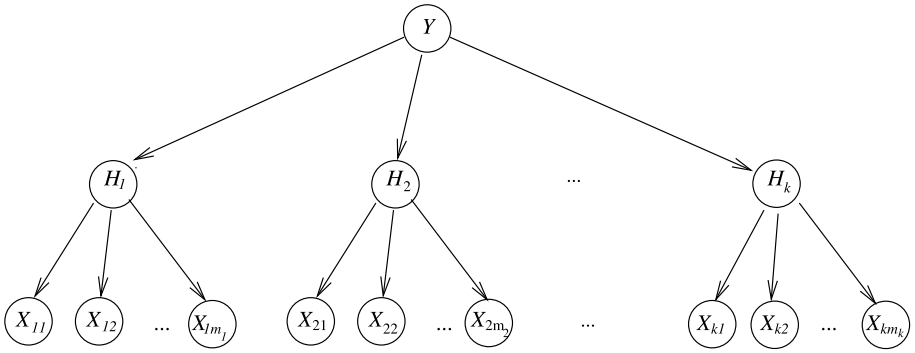

Fig. 1 A probability model in which the dependence of the observed variables on the class designation is mediated by a layer of hidden variables

Note that we may assume without loss of generality that for any indices i and j, we have

$$\mathbf{Pr}(X_{ij} = 1|H_i = 1) > \mathbf{Pr}(X_{ij} = 1|H_i = -1),$$

since otherwise, we could replace X_{ij} with its negation.

Definition 1 (β-effect) We say that a hidden variable H_i *β-affects* observed variable X_{ij} if

$$\mathbf{Pr}(X_{ij} = 1|H_i = 1) - \mathbf{Pr}(X_{ij} = 1|H_i = -1) > \beta.$$

2.2 Other probability tools

Definition 2 (Total variation distance) The *total variation distance* between probability distributions P and Q over a common domain U, denoted by $d_{TV}(P, Q)$, is $\max_{E \subseteq U} |P(E) - Q(E)|$.

Lemma 1 (Hoeffding bound (Hoeffding 1963), see (Pollard 1984)) *Let $U_1, \ldots, U_\ell$ be independent real random variables, each of which takes values in an interval of length κ. Then*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{\ell} U_i - \mathbf{E}\left(\sum_{i=1}^{\ell} U_i\right)\right| \geq \gamma\right] \leq 2e^{-\frac{2\gamma^2}{\kappa^2 \ell}}.$$

3 Linear approximation

Here is our main result.

Theorem 1 *Suppose that a hidden variable model P satisfies, for $\beta > 0$, that each hidden variable β-affects at least m observed variables, for*

$$m = \omega\left(\frac{k \log^2(k/\text{opt}) \log(1/\text{opt})}{\beta^2}\right),$$

where opt *is the error rate of the Bayes optimal classifier. Suppose*

$$\mathbf{X} = (X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k})$$

are the observed variables. Then there is a linear classifier f *such that*

$$\mathbf{Pr}_{(\mathbf{X},Y)\sim P}(f(\mathbf{X}) \neq Y) \leq \text{opt} + o(\text{opt}). \tag{1}$$

Note that, as opt gets smaller, (1) guarantees a closer approximation. This explains why the bound on m grows with $1/\text{opt}$.

We prove Theorem 1 through a series of lemmas. We will establish the stronger guarantee that the linear classifier approximates the behavior of an idealized classifier that has access to the hidden variables. The optimal classifier f_{opt} that uses the values $h_1, \ldots, h_k$ of the hidden variables $H_1, \ldots, H_k$ along with $x_{11}, \ldots, x_{km_k}$ is at least as accurate as the optimal classifier that only uses $x_{11}, \ldots, x_{km_k}$, since when optimizing over classifiers that have access to $h_1, \ldots, h_k$, one possibility is use a classifier that ignores them.

Our first lemma is that f_{opt} depends only on the hidden variables.

Lemma 2 *For any realization* $\mathbf{h}$ *of the hidden variables, and any realization* $\mathbf{x}$ *of the observed variables,*

$$\mathbf{Pr}(Y = 1|\mathbf{H} = \mathbf{h}, \mathbf{X} = \mathbf{x}) = \mathbf{Pr}(Y = 1|\mathbf{H} = \mathbf{h}), \tag{2}$$

so that

$$f_{\text{opt}}(\mathbf{h}, \mathbf{x}) = \operatorname*{argmax}_y \mathbf{Pr}(Y = y|\mathbf{H} = \mathbf{h}). \tag{3}$$

Proof Since

$$f_{\text{opt}}(\mathbf{h}, \mathbf{x}) = \operatorname*{argmax}_y \mathbf{Pr}(Y = y|\mathbf{H} = \mathbf{h} \text{ and } \mathbf{X} = \mathbf{x})$$

and (2) follows from the fact that $H_1, \ldots, H_k$ form a Markov blanket for Y, we get (3). □

Our next lemma, which is proved using established techniques (Duda et al. 2000), characterizes f_{opt}.

Lemma 3 *There is a* $\mathbf{w} \in \mathbf{R}^k$ *and a* $w_0 \in \mathbf{R}$ *such that, for all realizations* $\mathbf{h} = (h_1, \ldots, h_k)$ *of the hidden variables, and all realizations* $\mathbf{x} = (x_{11}, \ldots, x_{k,m_k})$ *of the observed variables,*

$$\frac{\mathbf{Pr}(Y = y, \mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y, \mathbf{H} = \mathbf{h})} = \exp(y(w_0 + \mathbf{w} \cdot \mathbf{h})), \tag{4}$$

and therefore

$$f_{\text{opt}}(\mathbf{h}, \mathbf{x}) = \text{sign}(w_0 + \mathbf{w} \cdot \mathbf{h}).$$

Proof Maximizing the right-hand side of (3) is equivalent to maximizing

$$\frac{\mathbf{Pr}(Y = y|\mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y|\mathbf{H} = \mathbf{h})} = \frac{\mathbf{Pr}(Y = y, \mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y, \mathbf{H} = \mathbf{h})}, \tag{5}$$

which decomposes nicely, facilitating analysis, as we will see.

The odds ratio (5) can be written as follows

$$\frac{\mathbf{Pr}(Y=y,\mathbf{H}=\mathbf{h})}{\mathbf{Pr}(Y=-y,\mathbf{H}=\mathbf{h})}=\frac{\mathbf{Pr}(Y=y)}{\mathbf{Pr}(Y=-y)}\prod_{i=1}^{k}\frac{\mathbf{Pr}(H_i=h_i|Y=y)}{\mathbf{Pr}(H_i=h_i|Y=-y)}$$

and a case analysis verifies that for each i, we have

$$\frac{\mathbf{Pr}(H_i=h_i|Y=y)}{\mathbf{Pr}(H_i=h_i|Y=-y)}=\exp\left(\frac{y}{2}\ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=1)}{\mathbf{Pr}(H_i=-1|Y=-1)\mathbf{Pr}(H_i=1|Y=-1)}\right)+\frac{yh_i}{2}\ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=-1)}{\mathbf{Pr}(H_i=-1|Y=1)\mathbf{Pr}(H_i=1|Y=-1)}\right)\right).$$

Thus, if for each $i\in\{1,\ldots,k\}$, we define

$$w_i=\frac{1}{2}\ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=-1)}{\mathbf{Pr}(H_i=-1|Y=1)\mathbf{Pr}(H_i=1|Y=-1)}\right),$$

let

$$w_0=\ln\left(\frac{\mathbf{Pr}(Y=y)}{\mathbf{Pr}(Y=-y)}\right)+\frac{1}{2}\sum_{i=1}^{k}\ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=1)}{\mathbf{Pr}(H_i=-1|Y=-1)\mathbf{Pr}(H_i=1|Y=-1)}\right)\tag{6}$$

and set $\mathbf{w}=(w_1,\ldots,w_k)$, we get (4) which immediately implies that $f_{\mathrm{opt}}(\mathbf{x},\mathbf{h})=\mathrm{sign}(w_0+\mathbf{w}\cdot\mathbf{h})$. □

Our next lemma will concern estimates of the hidden variables constructed from the observed variables. First, let us define some notation.

Definition 3 ($\mathcal{X}_i$) For each i, let $\mathcal{X}_i$ consist of all indices j such that

$$\mathbf{Pr}(X_{ij}=1|H_i=1)-\mathbf{Pr}(X_{ij}=1|H_i=-1)>\beta.\tag{7}$$

Definition 4 (H_i^+ and H_i^-) For each i, define

$$H_i^+=\frac{1}{|\mathcal{X}_i|}\sum_{j\in\mathcal{X}_i}\mathbf{E}(X_{ij}|H_i=1)$$

$$H_i^-=\frac{1}{|\mathcal{X}_i|}\sum_{j\in\mathcal{X}_i}\mathbf{E}(X_{ij}|H_i=-1).$$

Note that (7) implies that $H_i^+-H_i^->2\beta$.

Definition 5 (ϕ_i) For each i, define $\phi_i:\mathbf{R}\to\mathbf{R}$ to be the affine transformation of the real line that maps H_i^+ to 1, and H_i^- to -1; that is,

$$\phi_i(x)=\frac{2x-(H_i^++H_i^-)}{H_i^+-H_i^-}.$$

Definition 6 ($\hat{H}_i$) For each i, define

$$\hat{H}_i = \phi_i \left(\frac{1}{|\mathcal{X}_i|} \sum_{j \in \mathcal{X}_i} X_{ij} \right),$$

so that

$$\mathbf{E}(\hat{H}_i | H_i = h_i) = h_i. \tag{8}$$

Our next definition is the linear approximation to f_{opt} that we will analyze.

Definition 7 (f) Define

$$f(\mathbf{X}) = \text{sign}(w_0 + \mathbf{w} \cdot \hat{\mathbf{H}}),$$

where $\hat{\mathbf{H}} = (\hat{H}_1, \ldots, \hat{H}_k)$.

As we have discussed, a key aspect of the analysis will be to argue that $\sum_{i=1}^k w_i \hat{H}_i$ is likely to be a good approximation to $\sum_{i=1}^k w_i H_i$. To control the variance of

$$\sum_{i=1}^k w_i \left(H_i - \hat{H}_i \right)$$

for this purpose, we need to show that we can assume without loss of generality that each weight w_i is not very big. This is the subject of the next lemma.

Lemma 4 *There is a hidden variable source Q whose total variation distance from P is at most ϵ, and for which $W \stackrel{def}{=} \max_{1 \le i \le k} |w_i| = O(\log(k/\epsilon))$ and $|w_0| = O(k \log(k/\epsilon))$.*

Proof Suppose we modified P by adding a secondary label $\tilde{Y}$ and a layer of hidden variables $\tilde{H}_1, \ldots, \tilde{H}_k$ so that

- $\tilde{Y}$ is obtained by flipping Y with probability $\epsilon/(k+1)$,
- the conditional distribution of $H_1, \ldots, H_k$ given $\tilde{Y}$ was the same as the old conditional distribution given Y, and
- each $\tilde{H}_i$ was obtained by negating the value of H_i with probability $\epsilon/(k+1)$, and
- the conditional distributions of $X_{i1}, \ldots, X_{im_i}$ given $\tilde{H}_i$ were the same as the old conditional distributions of $X_{i1}, \ldots, X_{im_i}$ given H_i.

(See Fig. 2.) If we did this, the joint distribution of

$$\tilde{Y}, \tilde{H}_1, \ldots, \tilde{H}_k, X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k}$$

would have total variation distance at most ϵ from the distribution over

$$Y, H_1, \ldots, H_k, X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k},$$

because the probability that any of flips are executed is at most $(k+1)(\epsilon/(k+1)) = \epsilon$. This means that the probability of error of any classifier with respect to the original source P is at most ϵ more than its error probability with respect to the modified source Q. Furthermore,

$$\tilde{Y}, \tilde{H}_1, \ldots, \tilde{H}_k, X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k}$$

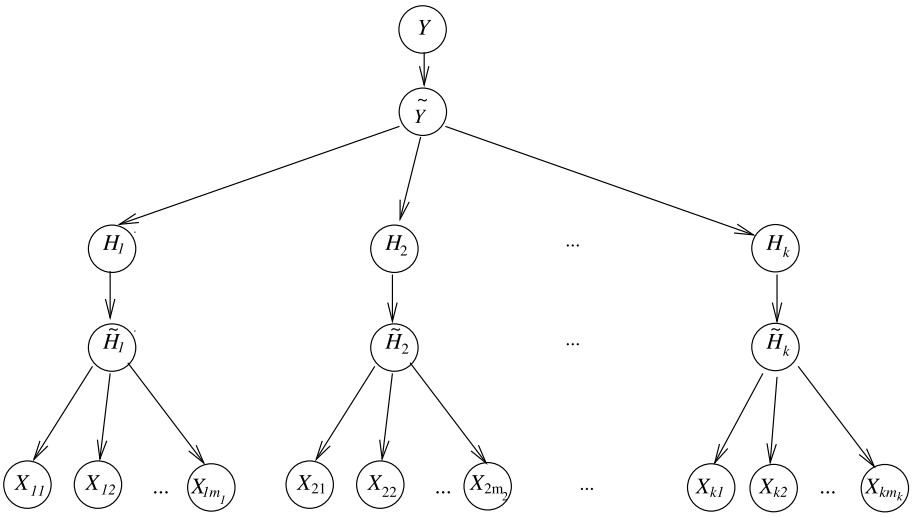

Fig. 2 The dependence structure of the probability distribution used in the proof of Lemma 4

have the same conditional independence structure as the original source, but $\mathbf{Pr}(\tilde{H}_i = h|\tilde{Y} = y)$ is always in the interval $[\epsilon/(k+1), 1-\epsilon/(k+1)]$, and so is $\mathbf{Pr}(\tilde{Y} = y)$. Thus, for $i \geq 1$, since, for the modified source Q,

$$w_i = \frac{1}{2}\ln\left(\frac{\mathbf{Pr}(\tilde{H}_i = 1|\tilde{Y} = 1)\mathbf{Pr}(\tilde{H}_i = -1|\tilde{Y} = -1)}{\mathbf{Pr}(\tilde{H}_i = -1|\tilde{Y} = 1)\mathbf{Pr}(\tilde{H}_i = 1|\tilde{Y} = -1)}\right)$$

we have

$$\begin{aligned}&\frac{1}{2}\ln\left(\frac{(\epsilon/(k+1)) \times (\epsilon/k+1)}{1 \times 1}\right)\\ &\quad\leq w_i\\ &\quad\leq \frac{1}{2}\ln\left(\frac{1 \times 1}{(\epsilon/(k+1)) \times (\epsilon/k+1)}\right)\end{aligned}$$

and, similarly,

$$\begin{aligned}&\ln\left(\frac{\epsilon/(k+1)}{1}\right) + \frac{1}{2}\sum_{i=1}^{k}\ln\left(\frac{\epsilon/(k+1) \times \epsilon(k+1)}{1 \times 1}\right)\\ &\quad\leq w_0\\ &\quad\leq \ln\left(\frac{1}{\epsilon/(k+1)}\right) + \frac{1}{2}\sum_{i=1}^{k}\ln\left(\frac{1 \times 1}{\epsilon/(k+1) \times \epsilon(k+1)}\right),\end{aligned}$$

and simplifying completes the proof. □

Now our goal is to approximate the optimal classifier for Q. To keep the notation simple, until further notice, let us reuse the notation f_{opt}, $\mathbf{w}$, etc. to refer to optimal classification for

Q, and dispense with the tildes. Another way to think of this is that we are assuming without loss of generality (modulo ϵ-approximation) that the weights of the optimal classifier for P have magnitude at most W.

From here, our analysis will make use of the standard notion of a margin.

Definition 8 (μ and ρ) Define

$$\mu(\mathbf{h}, y) = y\,(w_0 + \mathbf{w} \cdot \mathbf{h})$$

so that (4) can be rewritten as

$$\frac{\mathbf{Pr}(Y = y, \mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y, \mathbf{H} = \mathbf{h})} = \exp\left(\mu(\mathbf{h}, y)\right).$$

We can think of $\rho(\mathbf{h}) = \max_y \mu(\mathbf{h}, y)$ as a measure of the extent to which $\mathbf{h}$ determines the value of y.

Our analysis will proceed by showing that, for any $\mathbf{h}$, conditioned on the event that $\mathbf{H} = \mathbf{h}$, the linear classifier f obtained by using $\hat{H}_1, \ldots, \hat{H}_k$ makes prediction errors at a slightly larger rate than f_{opt}. This will be achieved through two bounds. The first bound will capture the intuition that, when $\rho(\mathbf{h})$ is not too small, then approximating $\mathbf{H}$ by $\hat{\mathbf{H}}$ is unlikely to perturb correct classifications. The second bound will capture the intuition that, when $\rho(\mathbf{h})$ is small, even f_{opt} is inaccurate enough that any classifier approximates its accuracy to within a small factor.

We begin with the case in which ρ is big.

Lemma 5 *For any* $\mathbf{h}$,

$$\begin{aligned}\mathbf{Pr}(f(\mathbf{X}) \neq f_{\text{opt}}(\mathbf{X}, \mathbf{H}) | \mathbf{H} = \mathbf{h}) &\leq \mathbf{Pr}\left(\left|\sum_{i=1}^{k} w_i \left(H_i - \hat{H}_i\right)\right| \geq \rho(\mathbf{h}) \;\middle|\; \mathbf{H} = \mathbf{h}\right) \\ &\leq \exp\left(-\Omega\left(\frac{\rho(\mathbf{h})^2 \beta^2 m}{k \ln^2(k/\epsilon)}\right)\right).\end{aligned}$$

Proof The first inequality follows directly from the definition of ρ.

By (8), after conditioning on $\mathbf{H} = \mathbf{h}$, we have $\mathbf{E}(\hat{H}_i) = h_i$, so the expectation of

$$S = \sum_{i=1}^{k} w_i \left(H_i - \hat{H}_i\right)$$

is 0. To apply the Hoeffding bound (Lemma 1), we need to show that S is the sum of independent random variables, each of which takes values in a small interval. Unwinding the definition of $\hat{H}_i$, we get

$$\begin{aligned}S &= \sum_{i=1}^{k} w_i \left(h_i - \phi_i\left(\frac{1}{|\mathcal{X}_i|}\sum_{j \in \mathcal{X}_i} X_{ij}\right)\right) \\ &= \sum_{i=1}^{k} w_i \left(h_i + \frac{H_i^+ + H_i^-}{H_i^+ - H_i^-} - \frac{1}{|\mathcal{X}_i|(H_i^+ - H_i^-)}\sum_{j \in \mathcal{X}_i} X_{ij}\right).\end{aligned}$$

Moving out the sum over j,

$$S = \sum_{i=1}^{k} \sum_{j \in \mathcal{X}_i} w_i \left(\frac{h_i}{|\mathcal{X}_i|} + \frac{H_i^+ + H_i^-}{|\mathcal{X}_i|(H_i^+ - H_i^-)} - \frac{1}{|\mathcal{X}_i|(H_i^+ - H_i^-)} X_{ij} \right). \tag{9}$$

The independence structure of the source implies that, after conditioning on the event that $H_1 = h_1, \ldots, H_k = h_k$, the various variables X_{ij} are mutually independent, and therefore so are the various terms of the double sum in (9). Recall that the definition of β-effect implies that $(H_i^+ - H_i^-) \geq 2\beta$, and that we assumed that $|\mathcal{X}_i| \geq m$. Since $|w_i| \leq W$, each term in (9) can be upper and lower bounded as follows:

$$\begin{aligned}
& w_i \left(\frac{h_i}{|\mathcal{X}_i|} + \frac{H_i^+ + H_i^-}{|\mathcal{X}_i|(H_i^+ - H_i^-)} \right) - \frac{W}{2\beta m} \\
& \quad \leq w_i \left(\frac{h_i}{|\mathcal{X}_i|} + \frac{H_i^+ + H_i^-}{|\mathcal{X}_i|(H_i^+ - H_i^-)} - \frac{1}{|\mathcal{X}_i|(H_i^+ - H_i^-)} X_{ij} \right) \\
& \quad \leq w_i \left(\frac{h_i}{|\mathcal{X}_i|} + \frac{H_i^+ + H_i^-}{|\mathcal{X}_i|(H_i^+ - H_i^-)} \right) + \frac{W}{2\beta m},
\end{aligned}$$

so that each term of the right-hand side of (9) is bounded in an interval of length $\frac{W}{\beta m}$. Applying the Hoeffding bound (Lemma 1) to the terms of the right-hand side of (9), we get

$$\begin{aligned}
& \mathbf{Pr}\left[\left| \sum_{i=1}^{k} w_i \left(h_i - \hat{H}_i \right) \right| > \rho(\mathbf{h}) \,\middle|\, \mathbf{H} = \mathbf{h} \right] \\
& \quad \leq 2 \exp\left(\frac{-2\rho(\mathbf{h})^2 \beta^2 m}{k W^2} \right) \\
& \quad \leq \exp\left(-\Omega\left(\frac{\rho(\mathbf{h})^2 \beta^2 m}{k \ln^2(k/\epsilon)} \right) \right).
\end{aligned} \tag{10}$$

which completes the proof. □

Now, let us work on a bound for small $\rho(\mathbf{h})$. Here is the basic idea. We can pair each borderline case with its counterpart in which the label is negated. Lemma 3 implies that the two cases are nearly equally likely. Since both the linear classifier and the Bayes optimal classifier make an incorrect classification in one of the cases, the linear classifier approximates the accuracy of the Bayes optimal classifier, on average, over borderline cases.

Lemma 6 *For all* $\mathbf{h}$,

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \leq e^{\rho(\mathbf{h})} \mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y | \mathbf{H} = \mathbf{h}). \tag{11}$$

Proof When a pair of examples differs only in the label, any classifier, in particular, the linear classifier f, must classify one example of each pair correctly, thus

$$\begin{aligned}
& \mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \\
& \quad \leq \sum_{\mathbf{x}} \max\{\mathbf{Pr}(Y = 1 | \mathbf{X} = \mathbf{x}, \mathbf{H} = \mathbf{h}), \mathbf{Pr}(Y = -1 | \mathbf{X} = \mathbf{x}, \mathbf{H} = \mathbf{h})\} \\
& \qquad \times \mathbf{Pr}(\mathbf{X} = \mathbf{x} | \mathbf{H} = \mathbf{h}).
\end{aligned}$$

Since $H_1, \ldots, H_k$ form a Markov blanket for Y,

$$\begin{aligned}&\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \\ &\quad \leq \sum_{\mathbf{x}} \max\{\mathbf{Pr}(Y = 1 | \mathbf{H} = \mathbf{h}), \mathbf{Pr}(Y = -1 | \mathbf{H} = \mathbf{h})\} \times \mathbf{Pr}(\mathbf{X} = \mathbf{x} | \mathbf{H} = \mathbf{h}). \end{aligned} \tag{12}$$

Suppose y maximizes $\mathbf{Pr}(Y = y | \mathbf{H} = \mathbf{h})$. Then

$$\begin{aligned}&\frac{\mathbf{Pr}(Y = y | \mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y | \mathbf{H} = \mathbf{h})} \\ &\quad = \frac{\mathbf{Pr}(Y = y, \mathbf{H} = \mathbf{h})}{\mathbf{Pr}(Y = -y, \mathbf{H} = \mathbf{h})} \\ &\quad = \exp(\mu(\mathbf{h}, y)) \quad \text{(by Lemma 3)} \\ &\quad \leq \exp(\rho(\mathbf{h})).\end{aligned}$$

Putting this together with (12), we get

$$\begin{aligned}&\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \\ &\quad \leq \sum_{\mathbf{x}} e^{\rho(\mathbf{h})} \min\{\mathbf{Pr}(Y = 1 | \mathbf{H} = \mathbf{h}), \mathbf{Pr}(Y = -1 | \mathbf{H} = \mathbf{h})\} \times \mathbf{Pr}(\mathbf{X} = \mathbf{x} | \mathbf{H} = \mathbf{h}).\end{aligned}$$

The Bayes optimal classifier cannot avoid making a mistake on one label or the other which implies (11). □

Proof of Theorem 1 Let us condition on the event that

$$H_1 = h_1, \ldots, H_k = h_k. \tag{13}$$

We have

$$\begin{aligned}&\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \\ &\quad \leq \mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y | \mathbf{H} = \mathbf{h}) + \mathbf{Pr}(f(\mathbf{X}) \neq f_{\text{opt}}(\mathbf{X}, \mathbf{H}) | \mathbf{H} = \mathbf{h}).\end{aligned}$$

Let $\kappa > 0$ be a parameter, independent of $\mathbf{h}$ (but possibly depending on the source), that will be fixed later in the argument. We will use κ as a dividing line between large and small margin cases. In particular, if $\rho(\mathbf{h}) \leq \ln(1 + \kappa)$, then Lemma 6 implies

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \leq (1 + \kappa) \mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y | \mathbf{H} = \mathbf{h}).$$

If $\rho(\mathbf{h}) > \ln(1 + \kappa)$, then Lemma 5 implies that

$$m = \Omega\left(\frac{k \ln^2(k/\epsilon) \ln(1/\text{opt})}{\beta^2 \ln^2(1 + \kappa)}\right)$$

suffices for

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y | \mathbf{H} = \mathbf{h}) \leq \mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y | \mathbf{H} = \mathbf{h}) + \text{opt}^2.$$

So, in either case,

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y|\mathbf{H} = \mathbf{h}) \leq (1+\kappa)\mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y|\mathbf{H} = \mathbf{h}) + \text{opt}^2.$$

Since κ was chosen independently of $\mathbf{h}$, averaging over $\mathbf{h}$ yields

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y) \leq (1+\kappa)\mathbf{Pr}(f_{\text{opt}}(\mathbf{X}, \mathbf{H}) \neq Y) + \text{opt}^2.$$

Letting κ go to zero arbitrarily slowly with opt and setting $\epsilon = \text{opt}^2$ completes the proof. □

The hidden variables can afford to be much less influential if they have similar degrees of association with the class designation. This is illustrated by considering the idealized case in which all associations are equally strong.

Theorem 2 *Suppose there is* $0 < \alpha < 1/4$ *such that each hidden variable* H_i *has* $\mathbf{Pr}(H_i = y|Y = y) = 1/2 + \alpha$ *for both* $y \in \{-1, 1\}$.

If in addition for $\beta > 0$, *each hidden variable* β*-affects at least* m *observed variables, and* opt *is the error rate of the Bayes optimal classifier, for*

$$m = \omega\left(\frac{\log^2(1/\text{opt})}{\beta^2}\right),$$

then there is a linear classifier whose error rate is

$$\text{opt} + o(\text{opt}).$$

Proof The proof is a modification of the proof of Theorem 1; we only describe the modifications that are needed.

First, $w_i = \ln\frac{1+2\alpha}{1-2\alpha}$ for all i, so $w_i = \Theta(\alpha)$.

Replacing (10) with

$$\mathbf{Pr}\left[Y\left(\sum_{i=1}^{k} w_i\left(H_i - \hat{H}_i\right)\right) > \rho(\mathbf{h})\right] \leq \exp\left(\frac{-c\rho(\mathbf{h})^2\beta^2 m}{\alpha^2 k}\right) \tag{14}$$

and otherwise arguing as in Theorem 1 leads to the conclusion that

$$m = \omega\left(\frac{k\alpha^2\log(1/\text{opt})}{\beta^2}\right)$$

suffices for the linear classifier to have error $\text{opt} + o(\text{opt})$.

As argued in the proof of Theorem 1, the error rate of the Bayes optimal classifier that uses only the observed variables is at least as large as the error rate of the optimal classifier that also uses the hidden variables, and, for sources considered in this theorem, the latter classifier simply takes a majority vote over the values of the hidden variables. This classifier is incorrect when a majority of the hidden variables take values different from the label. Applying the Hoeffing bound, this happens with probability $\exp(-\Omega(\alpha^2 k))$, and thus, $\alpha^2 k = O(\log(1/\text{opt}))$ which completes the proof. □

4 A convex loss bound

In this section, we show that a convex upper bound on the error rate of a linear classifier can in turn be bounded in terms of the error rate of the Bayes-optimal classifier. We will use the hinge loss.

Definition 9 (Hinge loss) For $z \in \mathbf{R}$, define the hinge loss $\ell(z)$ by $\max\{1 - z, 0\}$.

Next is our bound on the hinge loss of a linear classifier. Note that the bound is *not* in terms of the optimal hinge loss, but rather in terms of the optimal prediction error rate. In fact, as before, we actually prove a bound in terms of the error rate of a classifier that has access to the hidden variables.

Theorem 3 *Suppose that a hidden variable model satisfies, for $\beta > 0$, that each hidden variable β-affects at least m observed variables, for*

$$m = \Omega\left(\frac{k \ln^2\left(\frac{k}{\beta \mathrm{opt}}\right)}{\beta^2 \mathrm{opt}^2}\right)$$

where opt *is the error rate of the Bayes optimal classifier. Suppose*

$$\mathbf{X} = (X_{11}, \ldots, X_{1m_1}, \ldots, X_{k1}, \ldots, X_{km_k})$$

are the observed variables. Then there is a weight vector

$$\mathbf{v} = (v_{11}, \ldots, v_{1m_1}, \ldots, v_{k1}, \ldots, v_{km_k})$$

and $v_0 \in \mathbf{R}$ such that

$$\mathbf{E}(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) = O\left(\mathrm{opt} \log \frac{1}{\mathrm{opt}}\right).$$

Our proof of Theorem 3 has two parts. First, we bound the expected loss of the classifier that minimizes the classification error rate. Then we show that the linear classifier constructed in Theorem 1 approximates this loss.

Lemma 7 *If $w_0, \ldots, w_k$ are the weights of the Bayes optimal classifier (for minimizing error rate using the hidden variables $\mathbf{h}$) and $\mathbf{w} = (w_1, \ldots, w_k)$, then*

$$\mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H}))) = O(\mathrm{opt} \log(1/\mathrm{opt})).$$

Proof Recall the following definitions:

$$\mu(\mathbf{h}, y) = y\,(w_0 + \mathbf{w} \cdot \mathbf{h})\,, \qquad \rho(\mathbf{h}) = \max_y \mu(\mathbf{h}, y).$$

Since the Bayes optimal classifier picks the more likely value of y, Lemma 3 implies that

$$\frac{\mathbf{Pr}(Y \neq f_{\mathrm{opt}}(\mathbf{X}, \mathbf{H}) | \mathbf{H} = \mathbf{h})}{1 - \mathbf{Pr}(Y \neq f_{\mathrm{opt}}(\mathbf{X}, \mathbf{H}) | \mathbf{H} = \mathbf{h})} = \exp(-\rho(\mathbf{h}))$$

which in turn implies

$$\mathbf{Pr}(Y \neq f_{\text{opt}}(\mathbf{X}, \mathbf{H}) | \mathbf{H} = \mathbf{h}) = \frac{1}{1 + \exp(\rho(\mathbf{h}))} \tag{15}$$

and therefore

$$\text{opt} = \sum_{\mathbf{h} \in \{-1,1\}^k} \frac{1}{1 + \exp(\rho(\mathbf{h}))} \times \mathbf{Pr}(\mathbf{h}). \tag{16}$$

Furthermore,

$$\begin{aligned}
&\mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H}))) \\
&\quad = \mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})) \times 1_{\rho(\mathbf{H})>1}) + \mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})) \times 1_{\rho(\mathbf{H})\leq 1}),
\end{aligned}$$

where $1_{\rho(\mathbf{H})>1}$ is the indicator function for the event that $\rho(\mathbf{H}) > 1$. Since for any realization $\mathbf{h}$ of $\mathbf{H}$, $\rho(\mathbf{h}) \leq 1$ implies $\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{h})) \leq 2$, we have

$$\mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H}))) \leq \mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})) \times 1_{\rho(\mathbf{H})>1}) + 2\mathbf{Pr}(\rho(\mathbf{H}) \leq 1). \tag{17}$$

Let us start by bounding the first term of (17). Since

- $\mu(\mathbf{h}, y) > 1$ implies $\ell(y(w_0 + \mathbf{w} \cdot \mathbf{h})) = 0$,
- $\mu(\mathbf{h}, y) \in \{-\rho(\mathbf{h}), \rho(\mathbf{h})\}$, and
- always, $\ell(y(w_0 + \mathbf{w} \cdot \mathbf{h})) \leq 1 + \rho(\mathbf{h})$,

we have

$$\begin{aligned}
&\mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})) \times 1_{\rho(\mathbf{H})>1}) \\
&\quad \leq \sum_{\mathbf{h} \in \{-1,1\}^k : \rho(\mathbf{h})>1} (1 + \rho(\mathbf{h})) \times \mathbf{Pr}(\mu(\mathbf{H}, Y) = -\rho(\mathbf{h}) | \mathbf{H} = \mathbf{h}) \times \mathbf{Pr}(\mathbf{h}) \\
&\quad \leq \sum_{\mathbf{h} \in \{-1,1\}^k} (1 + \rho(\mathbf{h})) \times \mathbf{Pr}(\mu(\mathbf{H}, Y) = -\rho(\mathbf{h}) | \mathbf{H} = \mathbf{h}) \times \mathbf{Pr}(\mathbf{h}) \\
&\quad \leq \sum_{\mathbf{h} \in \{-1,1\}^k} \frac{1 + \rho(\mathbf{h})}{1 + \exp(\rho(\mathbf{h}))} \times \mathbf{Pr}(\mathbf{h}) \quad \text{(by (15))} \\
&\quad = \left(\sum_{\mathbf{h} \in \{-1,1\}^k : \rho(\mathbf{h}) \leq 2\ln(1/\text{opt})} \frac{1 + \rho(\mathbf{h})}{1 + \exp(\rho(\mathbf{h}))} \times \mathbf{Pr}(\mathbf{h}) \right) \\
&\qquad + \left(\sum_{\mathbf{h} \in \{-1,1\}^k : \rho(\mathbf{h}) > 2\ln(1/\text{opt})} \frac{1 + \rho(\mathbf{h})}{1 + \exp(\rho(\mathbf{h}))} \times \mathbf{Pr}(\mathbf{h}) \right) \\
&\quad \leq \left(\sum_{\mathbf{h} \in \{-1,1\}^k : \rho(\mathbf{h}) \leq 2\ln(1/\text{opt})} \frac{1 + 2\ln(1/\text{opt})}{1 + \exp(\rho(\mathbf{h}))} \times \mathbf{Pr}(\mathbf{h}) \right) \\
&\qquad + \left(\sum_{\mathbf{h} \in \{-1,1\}^k : \rho(\mathbf{h}) > 2\ln(1/\text{opt})} \frac{1 + 2\ln(1/\text{opt})}{1 + 1/\text{opt}^2} \times \mathbf{Pr}(\mathbf{h}) \right)
\end{aligned}$$

since opt $\leq 1/2$, and $(1+z)/(1+\exp(z))$ is nonincreasing in z when $z > 2\ln 2$. Applying (16) to bound the first sum, and noting that the second sum is at most $\frac{1+2\ln(1/\text{opt})}{1+1/\text{opt}^2}$ which is $o(\text{opt}\log(1/\text{opt}))$, we get

$$E(\ell(Y(w_0 + \mathbf{w}\cdot\mathbf{H})) \times 1_{\rho(\mathbf{H})>1}) = O(\text{opt}\log(1/\text{opt})). \tag{18}$$

Now, let us turn to the second term of (17). By (15), the conditional probability that the Bayes optimal algorithm makes a prediction error, given that $\rho(\mathbf{h}) \leq 1$, is at least $1/(1+e)$. Thus

$$\text{opt} \geq \frac{\mathbf{Pr}(\rho(\mathbf{H}) \leq 1)}{1+e},$$

which implies $\mathbf{Pr}(\rho(\mathbf{H}) \leq 1) = O(\text{opt})$. Putting this together with (18) and (17) completes the proof. □

What remains is to show that a linear classifier in the observed variables can approximate the ℓ-loss of the Bayes optimal classifier. It will be useful for this to use a conversion from tail bounds to bounds on the expectation. While results of this sort are known, we include a proof because we don't know a reference for precisely this statement.

Lemma 8 *If Z is a real-valued random variable, $u > 0$, and for all $\eta > 0$, $\mathbf{Pr}(Z \geq \eta) \leq e^{-\eta^2 u}$, then*

$$\mathbf{E}(Z) \leq 3\sqrt{1/u}.$$

Proof We have

$$\begin{aligned}
\mathbf{E}(Z) &\leq \sum_{i=1}^{\infty} \sqrt{i/u}\,\mathbf{Pr}(Z \in (\sqrt{(i-1)/u}, \sqrt{i/u}]) \\
&\leq \sum_{i=1}^{\infty} \sqrt{i/u}\,\mathbf{Pr}(Z > \sqrt{(i-1)/u}) \\
&\leq \sum_{i=1}^{\infty} \sqrt{i/u}(1/e)e^{-i} \\
&= \sqrt{1/u}\,\frac{e^2}{(e-1)^2},
\end{aligned}$$

completing the proof. □

Now we are ready for the loss bound.

Lemma 9 *If*

$$m = \Omega\left(\frac{k\ln^2\left(\frac{k}{\beta\text{opt}}\right)}{\beta^2\text{opt}^2}\right)$$

then there is a weight vector

$$\mathbf{v} = (v_{11}, \ldots, v_{1m_1}, \ldots, v_{k1}, \ldots, v_{km_k})$$

and $v_0 \in \mathbf{R}$ *such that*

$$|\mathbf{E}(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))| = O(\text{opt}).$$

Proof Define the linear classifier f as in the proof of Theorem 1. Let $\mathbf{v}$ and v_0 be the parameters of f, as in the statement of this theorem. Recall that f was constructed by replacing each H_i with $\hat{H}_i$, a linear combination of some of the observed variables, and then applying the Bayes optimal classifier for using the hidden variables. Consequently, $\mathbf{v} \cdot \mathbf{X} = \mathbf{w} \cdot \hat{\mathbf{H}}$ and $v_0 = w_0$.

Let Q be the approximation to P constructed in Lemma 4. For now, let us continue our analysis for such a source Q, and return to treating the general case at the end of the proof.

Since ℓ is 1-Lipschitz,

$$\begin{aligned}
&|\mathbf{E}(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))| \\
&\quad \leq \mathbf{E}(|(v_0 + \mathbf{v} \cdot \mathbf{X}) - (w_0 + \mathbf{w} \cdot \mathbf{H})|) \\
&\quad = \mathbf{E}(|\mathbf{w} \cdot (\hat{\mathbf{H}} - \mathbf{H})|)
\end{aligned}$$

since $\mathbf{v} \cdot \mathbf{X} = \mathbf{w} \cdot \hat{\mathbf{H}}$ and $v_0 = w_0$.

The proof of Lemma 5 establishes that, for $\eta > 0$,

$$\mathbf{Pr}(|\mathbf{w} \cdot (\hat{\mathbf{H}} - \mathbf{H})| \geq \eta) \leq \exp\left(-\frac{c\eta^2\beta^2 m}{k \ln^2(k/\epsilon)}\right)$$

which means, using Lemma 8, that

$$\mathbf{E}(|\mathbf{w} \cdot (\hat{\mathbf{H}} - \mathbf{H})|) \leq c_1 \sqrt{\frac{k \ln^2(k/\epsilon)}{\beta^2 m}}$$

for a constant c_1.

We are almost there, but our analysis was for the ϵ-approximation Q to P that satisfies $|w_i| = O(\log(k/\epsilon))$ for $i \geq 1$ and $|w_0| = O(k \log(k/\epsilon))$. We have showed that

$$\left|\mathbf{E}_Q(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}_Q(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))\right| \leq c_1 \sqrt{\frac{k \ln^2(k/\epsilon)}{\beta^2 m}}. \tag{19}$$

Now we want to show that $\mathbf{E}_Q(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X})))$ cannot be too much less than $\mathbf{E}_P(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X})))$ and $\mathbf{E}_Q(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))$ cannot be too much more than $\mathbf{E}_P(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))$. Since

$$\left|\mathbf{E}_P(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}_Q(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X})))\right| \leq \left(\max_{\mathbf{X},Y} \ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))\right) d_{TV}(P, Q), \tag{20}$$

we need a bound on $\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))$. We have

$$\begin{aligned}
&\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X})) \\
&\quad \leq 1 + |Y(v_0 + \mathbf{v} \cdot \mathbf{X})| \\
&\quad \leq 1 + |v_0| + |\mathbf{v} \cdot \mathbf{X}|.
\end{aligned} \tag{21}$$

Recall that $\mathbf{v} \cdot \mathbf{X} = \mathbf{w} \cdot \hat{\mathbf{H}}$, so

$$|\mathbf{v} \cdot \mathbf{X}| = |\mathbf{w} \cdot \hat{\mathbf{H}}| \leq \|\mathbf{w}\|_\infty \|\hat{\mathbf{H}}\|_1 \leq O(\log(k/\epsilon))\|\hat{\mathbf{H}}\|_1.$$

Definition 6 (of $\hat{H}_i$) immediately implies that $|\hat{H}_i| \leq \frac{2}{\beta}$ for all i, so that

$$\|\hat{\mathbf{H}}\|_1 \leq \frac{2k}{\beta}. \tag{22}$$

Lemma 4 gives $v_0 = w_0 = O(k \log(k/\epsilon))$, and putting this together with (22), (21) and (20), we have

$$\left|\mathbf{E}_P(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}_Q(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X})))\right| \leq O\left(\frac{k\epsilon \log(k/\epsilon)}{\beta}\right).$$

We may similarly show that

$$\left|\mathbf{E}_P(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H}))) - \mathbf{E}_Q(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))\right| \leq O\left(k\epsilon \log(k/\epsilon)\right)$$

(note the absence of $\hat{\mathbf{H}}$). Applying (19) yields

$$|\mathbf{E}_P(\ell(Y(v_0 + \mathbf{v} \cdot \mathbf{X}))) - \mathbf{E}_P(\ell(Y(w_0 + \mathbf{w} \cdot \mathbf{H})))| \leq c_1 \sqrt{\frac{k \ln^2(k/\epsilon)}{\beta^2 m}} + O\left(\frac{k\epsilon \log(k/\epsilon)}{\beta}\right).$$

Setting $\epsilon = \Theta(\frac{\beta \mathrm{opt}^2}{k^2})$ makes the second term at most opt, and applying the bound on m completes the proof. □

5 Bayes optimal models are two-layer neural networks

In this section, we show that, even with further restrictions on the structure of the source, a two-layer neural network is needed to compute the exact Bayes optimal classifier.

Theorem 4 *Suppose that there are real constants $\alpha, \beta > 0$ and a positive integer m such that*

- *each H_i is independently equal to Y with probability $1/2 + \alpha$,*
- *$m_i = m$ for all $i > 0$, and*
- *each X_{ij} is independently equal to H_i with probability $1/2 + \beta$.*

Define $A = \frac{1+2\alpha}{1-2\alpha}$, $B = \frac{1+2\beta}{1-2\beta}$, and, for each $i \in \{1, \ldots, k\}$, $s_i(\mathbf{x}) = \sum_{j=1}^m x_{ij}$.
The Bayes optimal classifier is

$$h(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^k \log\left(\frac{B^{s_i(\mathbf{x})}A + 1}{B^{s_i(\mathbf{x})} + A}\right)\right). \tag{23}$$

Proof Notice that for any $y \in \{-1, 1\}$,

$$\begin{aligned}&\mathbf{Pr}[Y = y | (\forall i, j) X_{ij} = x_{ij}] \\ &\quad = \frac{\mathbf{Pr}[Y = y]\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = y]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij}]}\end{aligned}$$

$$= \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = y]}{2\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij}]}$$

and therefore

$$\mathbf{Pr}[Y = 1 | (\forall i, j) X_{ij} = x_{ij}] > \mathbf{Pr}[Y = -1 | (\forall i, j) X_{ij} = x_{ij}]$$

if and only if

$$\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] > \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1].$$

Therefore, the Bayes optimal classifier gives

$$h(\mathbf{x}) = \operatorname{sign}\left(\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] - \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]\right).$$

Since log is a monotone function we also have

$$\begin{aligned} h(\mathbf{x}) &= \operatorname{sign}\left(\log \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] - \log \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]\right) \\ &= \operatorname{sign}\left(\log \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]}\right). \end{aligned} \tag{24}$$

Let $S_i = H_i Y$ (so that S_i that is 1 with probability $\frac{1}{2} + \alpha$ and -1 with probability $\frac{1}{2} - \alpha$), and $T_{ij} = X_{ij} H_i$ (so T_{ij} is 1 with probability $\frac{1}{2} + \beta$ and -1 with probability $\frac{1}{2} - \beta$). Now since T_{ij} and S_i are independent of Y, and, the events $[(\forall j) T_{ij} S_i = x_{ij}]$ are independent

$$\begin{aligned} &\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] \\ &\quad = \mathbf{Pr}[(\forall i, j) T_{ij} S_i Y = x_{ij} | Y = 1] \\ &\quad = \mathbf{Pr}[(\forall i, j) T_{ij} S_i = x_{ij} | Y = 1] \\ &\quad = \mathbf{Pr}[(\forall i, j) T_{ij} S_i = x_{ij}] \\ &\quad = \prod_{i=1}^{k} \mathbf{Pr}[(\forall j) T_{ij} S_i = x_{ij}]. \end{aligned}$$

Similarly,

$$\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1] = \prod_{i=1}^{k} \mathbf{Pr}[(\forall j) T_{ij} S_i = -x_{ij}].$$

By (24) we get

$$\begin{aligned} &h(\mathbf{x}) \\ &\quad = \operatorname{sign}\left(\log \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]}\right) \\ &\quad = \operatorname{sign}\left(\sum_{i=1}^{k} \log\left(\frac{\mathbf{Pr}[(\forall j) T_{ij} S_i = x_{ij}]}{\mathbf{Pr}[(\forall j) T_{ij} S_i = -x_{ij}]}\right)\right). \end{aligned} \tag{25}$$

Now, since for every i,

$$\begin{aligned}&\frac{\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]}{\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]}\\&\quad=\prod_{j=1}^{m}\frac{\mathbf{Pr}[T_{ij}=x_{ij}]}{\mathbf{Pr}[T_{ij}=-x_{ij}]}\\&\quad=\prod_{j=1}^{m}B^{x_{ij}},\end{aligned}$$

we have

$$\begin{aligned}&\mathbf{Pr}[(\forall j)T_{ij}S_i=x_{ij}]\\&\quad=\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]\mathbf{Pr}[S_i=1]+\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\mathbf{Pr}[S_i=-1]\\&\quad=\left(\frac{1}{2}-\alpha\right)\left(\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]A+\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\right)\\&\quad=\left(\frac{1}{2}-\alpha\right)\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\left(\left(\frac{\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]}{\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]}\right)A+1\right)\\&\quad=\left(\frac{1}{2}-\alpha\right)\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\left(\left(\frac{\prod_j\mathbf{Pr}[T_{ij}=x_{ij}]}{\prod_j\mathbf{Pr}[T_{ij}=-x_{ij}]}\right)A+1\right)\\&\quad=\left(\frac{1}{2}-\alpha\right)\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\left(A\prod_j\frac{\mathbf{Pr}[T_{ij}=x_{ij}]}{\mathbf{Pr}[T_{ij}=-x_{ij}]}+1\right)\\&\quad=\left(\frac{1}{2}-\alpha\right)\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\left(A\prod_{j=1}^{m}B^{x_{ij}}+1\right)\end{aligned}$$

and, similarly,

$$\begin{aligned}&\mathbf{Pr}[(\forall j)T_{ij}S_i=-x_{ij}]\\&\quad=\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\mathbf{Pr}[S_i=1]\\&\qquad+\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]\mathbf{Pr}[S_i=-1]\\&\quad=\left(\frac{1}{2}-\alpha\right)\left(\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]A+\mathbf{Pr}[(\forall j)T_{ij}=x_{ij}]\right)\\&\quad=\left(\frac{1}{2}-\alpha\right)\mathbf{Pr}[(\forall j)T_{ij}=-x_{ij}]\left(A+\prod_{j=1}^{m}B^{x_{ij}}\right).\end{aligned}$$

Now by (25),

$$h(\mathbf{x})=\operatorname{sign}\left(\sum_{i=1}^{k}\log\left(\frac{A\prod_{j=1}^{m}B^{x_{ij}}+1}{A+\prod_{j=1}^{m}B^{x_{ij}}}\right)\right)$$

$$= \operatorname{sign}\left(\sum_{i=1}^{k} \log\left(\frac{A\exp\left(\sum_{j=1}^{m}(\ln B)x_{ij}\right)+1}{A+\exp\left(\sum_{j=1}^{m}(\ln B)x_{ij}\right)}\right)\right)$$

$$= \operatorname{sign}\left(\sum_{i=1}^{k} \log\left(\frac{B^{s_i(\mathbf{x})}A+1}{B^{s_i(\mathbf{x})}+A}\right)\right),$$

completing the proof. □

One useful representation uses the following Taylor series

$$\ln x = 2\left[\left(\frac{x-1}{x+1}\right)+\frac{1}{3}\left(\frac{x-1}{x+1}\right)^3+\frac{1}{5}\left(\frac{x-1}{x+1}\right)^5+\cdots\right]$$

and gives

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{k}\sum_{\ell=1}^{\infty}\frac{(2\alpha)^{2\ell-1}}{2\ell-1}\tanh^{2\ell-1}\left(\frac{1}{2}\sum_{j=1}^{m}(\ln B)x_{ij}\right)\right), \tag{26}$$

where $\tanh y = \frac{e^{2y}-1}{e^{2y}+1}$.

The hyperbolic tangent is a standard squashing function for the hidden nodes in a two-layer neural network (Hertz et al. 1991), and raising it to a positive odd power maintains the sigmoid shape. Thus the Bayes optimal classifier described in Theorem 4 can be thought of as a two-layer neural network.

The classifier of (23) approximately,

- for each i, computes an estimate V_i of H_i by taking a majority vote over $X_{i1},\ldots,X_{im}$, and
- outputs a vote over V_i.

Intuitively, this is not a linear classifier, since, for example, X_{im} matters less if the value of V_i is already more-or-less determined by the values of $X_{i1},\ldots,X_{i(m-1)}$. This is formalized in the following.

Theorem 5 *If $k=m=3$, for any $\alpha>0$, there is a value of $\beta\in(0,1/2)$, so that the classifier h defined in (23) is not linear.*

Proof Assume for contradiction that $\mathbf{w}\in\mathbf{R}^{km}$ is the weight vector of a linear classifier f equal to h, i.e.

$$\operatorname{sign}\left(\sum_i\sum_j w_{ij}x_{ij}\right) = h(\mathbf{x})$$

for all $\mathbf{x}\in\{-1,1\}^{km}$.

We claim that this implies that h computes a majority vote. By symmetry, for any $\mathbf{x}$, any permutation ϕ of $\{1,\ldots,k\}$ and any permutation ψ over $\{1,\ldots,m\}$, we have

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_i\sum_j w_{ij}x_{\phi(i)\psi(j)}\right). \tag{27}$$

In general, for real a and b, if $\text{sign}(a) = \text{sign}(b)$, then $\text{sign}(a + b) = \text{sign}(a) = \text{sign}(b)$. Thus, (27) implies

$$h(\mathbf{x}) = \text{sign}\left(\sum_{\phi}\sum_{\psi}\sum_{i}\sum_{j} w_{ij} x_{\phi(i)\psi(j)}\right).$$

This in turn implies

$$h(\mathbf{x}) = \text{sign}\left((k-1)!(m-1)!\left(\sum_{i,j} w_{ij}\right)\sum_{i,j} x_{ij}\right)$$

because the permutations ϕ and ψ pair each weight with each feature an equal number of times. Rescaling, we get

$$h(\mathbf{x}) = \text{sign}\left(\sum_{i,j} x_{ij}\right),$$

the majority function.

To arrive at a contradiction, suppose $k = m = 3$, and

$$\mathbf{x} = ((1,1,1),(1,-1,-1),(1,-1,-1)).$$

Note that the majority function evaluates to 1 on $\mathbf{x}$. On the other hand, using the definition in (23), we have

$$h(\mathbf{x}) = \text{sign}\left(\log\left(\frac{B^3A+1}{B^3+A}\right) + 2\log\left(\frac{B^{-1}A+1}{B^{-1}+A}\right)\right).$$

As β gets closer to $1/2$, B gets arbitrarily large. But

$$\begin{aligned}\lim_{B\to\infty} &\log\left(\frac{B^3A+1}{B^3+A}\right) + 2\log\left(\frac{B^{-1}A+1}{B^{-1}+A}\right)\\ &= \log A - 2\log A < 0\end{aligned}$$

and therefore there is a value of β such that $h(\mathbf{x}) = -1$, a contradiction. □

6 Some related work

A number of papers have considered why the Naive Bayes algorithm, which outputs a linear hypothesis, works well despite class-conditional dependencies among the features (Domingos and Pazzani 1997; Bickel and Levina 2004; Kuncheva 2006). While Naive Bayes works suprisingly well, other linear classifiers typically perform better (Caruana and Niculescu-Mizil 2006; Caruana et al. 2008). Note that Naive Bayes may not work for the sources considered in this paper.

The hidden variable model studied here is a generalization of the Neyman Model of Evolution (Neyman 1971). A PAC algorithm for learning the probability distribution over the leaves for such models is known (Cryan et al. 2001). Using known tools, this algorithm

can be used as a subroutine in a polynomial-time algorithm for approximating the Bayes-Optimal classifier for sources in which the class-conditional distributions are of this form (Anoulova et al. 1996; Devroye et al. 1996). The linear approximation pointed out in this paper could be a step toward a more efficient algorithm for this problem.

The proof of Theorem 1 used the observation that the Bayes optimal classifier that has access to the hidden variables can be approximated by the classifier using small weights. Some recent research (Servedio 2007; Diakonikolas and Servedio 2009) established a related result; the analogous statement in our setting would concern the case in which the marginal over the hidden variables is uniform.

7 Conclusion

The analysis of this paper illustrates the expressive power of linear models even in the presence of class-conditional dependence among the features. The exact mathematical statements of this paper are among many possible choices that trade off between a clean and interpretable analysis, and a broadly relevant one, in different ways.

For example, it would not be hard to extend the approximation to apply to sources in which some observed variables depend on multiple hidden variables. As long as each hidden variable has enough variables that depend on it alone, we can construct the linear approximation as a function only of the observed variables that depend on specific hidden variables. Our analysis may also easily be extended to the case in which an unlimited number of variables depend directly on the class designation (as was done explicitly in the preliminary version of this paper (Bshouty and Long 2009)).

If each hidden variable H_i can take on more than two values, it is not hard to see that the Bayes optimal classifier that has access to them is a linear function of binary-valued indicator functions for events like $H_i = h_i$, so our analysis should extend easily to this case (though we would appear to need a collection of observed variables for each hidden variable-value pair).

The Hoeffding bounds that we use to analyze concentration do not require that the variables in the sums are binary-valued, so our analysis can also be straightforwardly extended to real-valued observed variables.

The analysis can also be extended without much modification to handle limited conditional dependence among the observed variables associated with a given hidden variable, with some degradation in the bounds, by applying generalizations of the Hoeffding bound to this case (see Schmidt et al. 1993; Dubhashi and Ranjan 1998; Pemmaraju 2001).

We also provided a bound on the hinge loss of the linear classifier in terms of the Bayes error rate, thereby showing that the Bayes error rate can be approximated efficiently. It is not clear whether this approximation bound can be improved. General tools that have recently been developed for the analysis of learning with convex loss functions (Zhang 2004b; Bartlett et al. 2006) may be useful for this.

As we mentioned previously, our analysis guarantees a closer approximation to the Bayes optimal as opt, the Bayes optimal error, gets small, which explains why more resources are needed in this case. It would be interesting to determine the optimal dependence of parameters in bounds like ours, such as m in Theorem 1, on opt.

Finally, it may be interesting to explore the possible tradeoffs between the computational complexity of learning algorithms and the quality of their approximations to the Bayes optimal error rate for sources like this, possibly exploiting the linear approximation pointed out in this paper, among other things.

Acknowledgements We thank Fernando Pereira, Olcan Sercinoglu, Rocco Servedio and Dekang Lin for their guidance and advice.

We also thank the COLT'09 and *Machine Learning* reviewers for their careful reading of earlier versions of this paper, and thoughtful comments.

References

Anoulova, S., Fischer, P., Pölt, S., & Simon, H. U. (1996). Probably almost Bayes decisions. *Information and Computation*, *129*(1), 63–71.

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Bickel, P., & Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'Naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989–1010.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*.

Bshouty, N. H., & Long, P. M. (2009). Linear classifiers are nearly optimal when hidden variables have diverse effects. In *COLT*.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *ICML* (pp. 161–168).

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *ICML* (pp. 96–103).

Cryan, M., Goldberg, L. A., & Goldberg, P. W. (2001). Evolutionary trees can be learned in polynomial time in the two-state general Markov model. *SIAM Journal on Computing*, *31*(2), 375–397.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Berlin: Springer.

Diakonikolas, I., & Servedio, R. (2009). Improved approximation of linear threshold functions. In *24th conference on computational complexity (CCC)* (pp. 161–172).

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Dubhashi, D., & Ranjan, D. (1998). Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, *13*(2), 99–124.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley.

Hertz, J. A., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation*. Reading: Addison-Wesley.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Society*, *58*(301), 13–30.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1/2), 177–196.

Hsieh, C. J., Chang, K. W., Lin, C. J., Keerthi, S. S., & Sundararajan, S. (2008). A dual coordinate descent method for large-scale linear SVM. In *ICML*.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European conference on machine learning* (pp. 137–142).

Kuncheva, L. I. (2006). On the optimality of naive Bayes with dependent binary features. *Pattern Recognition Letters*, *27*(7), 830–837.

Langseth, H., & Nielsen, T. D. (2006). Classification using hierarchical naive Bayes models. *Machine Learning*, *63*(2), 135–159.

Neyman, J. (1971). *Molecular studies of evolution: A source of novel statistical problems* (pp. 1–27). New York: Academic Press.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(1), 217–235.

Pemmaraju, S. (2001). Equitable coloring extends Chernoff-Hoeffding bounds. In *RANDOM*.

Pollard, D. (1984). *Convergence of stochastic processes*. Berlin: Springer.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, *39*(2/3), 135–168.

Schmidt, J. P., Siegel, A., & Srinivasan, A. (1993). Chernoff-Hoeffding bounds for applications with limited independence. In *SODA*.

Servedio, R. A. (2007). Every linear threshold function has a low-weight approximator. *Computational Complexity*, *16*(2), 180–209.

Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). Pegasos: Primal estimated sub-gradient solver for SVM. In *ICML* (pp. 807–814).

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(10), 6567–6572.

Zhang, N. L. (2004a). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, *5*(6), 697–723.

Zhang, T. (2004b). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *32*(1), 56–85.