

# Bias-Variance tradeoff in Hybrid Generative-Discriminative models

Guillaume Bouchard, Xerox Research Center Europe  
6, chemin de Maupertuis, 38240 Meylan, France

## Abstract

Given any generative classifier based on an inexact density model, we can define a discriminative counterpart that reduces its asymptotic error rate, while increasing the estimation variance. An optimal bias-variance balance might be found using Hybrid Generative-Discriminative (HGD) approaches. In these paper, these methods are defined in a unified framework. This allow us to find sufficient conditions under which an improvement in generalization performances is guaranteed. Numerical experiments illustrate the well fondness of our statements.

## 1 Introduction

In many Machine Learning applications, the overall goal is to find a classification rule with the smallest error rate on unseen data. Generative classification is a generic term to identify approaches defining a probabilistic model on the joint distribution of the inputs and the outputs and classifying new data points to the category with highest posterior probability. Many classifiers are built under this paradigm, including the widely used Naive Bayes classifier and Fisher Linear Discriminant. However, for theoretical and practical reasons, the usual way to learn the parameters of these model – the Maximum A Posteriori estimation – suffer from an asymptotic bias in the classification performances [1] and several methods have been proposed to improve their discriminative power.

A common discriminative framework is first to learn a generative model based on the data, and then apply a discriminative method in a feature space based on sufficient statistics of the learned generative model. A general methodology to build the features is called Augmented Statistical models [2], for which the well-known *Fisher Kernel* approach [3] is a special case. In principle, these methods involving two optimization problems could be improved by reducing them to a single maximization procedure.

Another approach, called *discriminative learning*, minimizes the classification loss of a generative classifier – usually the conditional likelihood. However, this type of parameter learning is prone to overfitting, and several

hybrid Generative-discriminative approach have been proposed [4, 5, 6] and often leads to significant improvements in real-world applications [7, 8]. There is nowadays a strong need of a theoretical foundations for these hybrid generative-discriminative approaches. This motivated the present work. In section 2, a theoretical framework is proposed, allowing existing hybrid generative-discriminative methods to be formally defined. we also propose a new robust hybrid generative-discriminative technique. Some theoretical insights to understand why hybrid techniques might improve prediction performances are given in section 3 and finally, we performed some experiments to compare the hybrid models.

## 2 Hybrid Generative-Discriminative models

We assume that the training dataset is composed by  $n$  labelled data  $(X, C) = \{(x_i, c_i)\}_{i=1}^n$ . A generative model defines a joint distribution  $\mathbf{P}(x, c|\theta)$  on the input variables  $x$  and output variables  $c$  where  $\theta$  denotes a vector of parameters. A common way of estimating the parameters of the generative models is to find the Maximum A Posteriori (MAP) solution:

$$\hat{\theta}_G = \arg \max_{\theta \in \Theta} L_G(\theta) + \log \mathbf{P}(\theta), \quad (1)$$

where  $L_G(\theta) = \log \mathbf{P}(X, C|\theta)$ . The discriminative learning problem corresponds to the estimation of the parameter  $\theta$  using a classification-based objective function. If we use the classification entropy loss function  $E_\theta[\log \mathbf{P}(c|x, \theta)]$  and a regularization term equal to  $\log \mathbf{P}(\theta)$ , we obtain the following optimization problem:

$$\hat{\theta}_D = \arg \max_{\theta \in \Theta} L_D(\theta) + \log \mathbf{P}(\theta), \quad (2)$$

where  $L_D(\theta) = \log \mathbf{P}(C|X, \theta)$  and  $L_D$  is called *discriminative log-likelihood*. Discriminative learning is known to converge to the parameter giving the smallest classification loss [9, 1] and empirical comparisons on real datasets show that it often – but not always – leads to a significant improvement of the classification performances [10].

However, for small datasets, the discriminative approach is prone to overfitting and recent work highlighted the

need of hybrid approaches to further improve the classification performances. Several proposals are model-specific, e.g. [5] proposed a hybrid generative-discriminative Naive Bayes classifier. Recently, some general solutions that smoothly interpolate between generative and discriminative learning has been proposed [4, 7, 6, 8].

One simple natural approach is to maximize a convex combination of the generative and the discriminative log-likelihoods plus the penalty [4, 7]:

$$\hat{\theta}_\lambda = \arg \max L_\lambda(\theta) + \log \mathbf{P}(\theta) \quad , \quad (3)$$

where  $L_\lambda(\theta) = (1 - \lambda)L_D(\theta) + \lambda L_G(\theta)$  and  $0 \leq \lambda \leq 1$  is a scalar interpolating between the generative ( $\lambda = 1$ ) and the discriminative solution ( $\lambda = 0$ ).

For Lasserre, Bishop and Minka [6], this estimator does not correspond to the MAP solution of a well-defined probabilistic model. They proposed instead to find the MAP solution of a new joint distribution  $\mathbf{Q}^{(1)}(X, C|\theta, \tilde{\theta}, \lambda) = \mathbf{P}(C|X, \theta)\mathbf{P}(X|\tilde{\theta})$  which depends only on the initial generative model  $\mathbf{P}(X, C|\theta)$ . The new prior distribution  $\mathbf{Q}^{(1)}(\tilde{\theta}|\theta, \lambda)$  correlates  $\theta$  and  $\tilde{\theta}$  with a strength depending on  $\lambda$  to obtain the tradeoff between the discriminative solution  $\mathbf{Q}^{(1)}(\tilde{\theta}|\theta, \lambda = 0) = \mathbf{P}(\tilde{\theta})$  and the generative one  $\mathbf{P}(\tilde{\theta}|\theta, \lambda = 1) = \delta(\theta, \tilde{\theta})$ ,  $\delta$  denoting the Dirac distribution. One interest of this approach is that it corresponds to the MAP estimation of a well defined generative model. However, we show in the following that there exists also a joint distribution for which the linear interpolation of the likelihoods can also be a well-founded MAP solution.

Let  $\mathbf{U}$  be a distribution defined on the input space  $\mathcal{X}$ . Similarly to [6], we introduce a new distribution  $\mathbf{Q}^{(2)}$  whose MAP estimator interpolates between the generative and the discriminative settings:

$$\mathbf{Q}^{(2)}(X, C|\theta, \lambda) = (1 - \lambda)\mathbf{P}(C|X, \theta)\mathbf{U}(X) + \lambda\mathbf{P}(X, C|\theta) \quad .$$

One can see that the maximum likelihood estimator of this model corresponds exactly to the discriminative solution for  $\lambda = 0$  and to the generative solution for  $\lambda = 1$ . Interestingly, the path  $\hat{\theta}_\lambda$  followed by the estimator while moving  $\lambda$  from 0 to 1 is exactly the same path as the estimator of [4] defined by  $\hat{\theta}_\eta = \arg \max_\theta (1 - \eta)L_D(\theta) + \eta L_G(\theta)$  while moving  $\eta$  from 0 to 1. To prove it, first choose a value for  $\eta$  in  $[0, 1]$  and compute  $\hat{\theta}_\eta$ . If we choose  $\lambda(\eta)$  such that:  $\frac{1}{\lambda(\eta)} = 1 + \frac{1-\eta}{\eta} \frac{\mathbf{P}(X|\hat{\theta}_\eta)}{\mathbf{U}(X)}$ , then the maximum likelihood (ML) equation  $(1 - \lambda)\frac{\partial}{\partial \theta}\mathbf{Q}(C|X, \theta)\mathbf{U}(X) + \lambda\frac{\partial}{\partial \theta}\mathbf{Q}(X, C|\theta) = 0$  is satisfied, proving that any estimator  $\hat{\theta}_\eta$  is also the ML estimator of the model  $\mathbf{Q}^{(2)}$  with tradeoff parameter  $\lambda(\eta)$ .

Hence the two main hybrid generative-discriminative models proposed in the literature [4, 6] correspond to a well defined MAP estimator. This motivates the following definition for the *Hybrid Generative-Discriminative* models:

**Definition 1** A distribution  $\mathbf{Q}(X, C|\theta, \lambda)$  indexed by  $\lambda \in [0, 1]$  such that for all  $(X, C) \in (\mathcal{X}, \mathcal{C})^n$  and all  $\theta \in \Theta$ ,

$$(i) \quad \mathbf{Q}(X, C|\theta, \lambda = 1) = \mathbf{P}_1(X, C|\theta),$$

$$(ii) \quad \mathbf{Q}(C|X, \theta, \lambda = 0) = \mathbf{P}_2(C|X, \theta) \text{ and}$$

is called a *Hybrid Generative-Discriminative (HGD) model* based on  $\mathbf{P}_1(X, C|\theta)$  and  $\mathbf{P}_2(C|X, \theta)$ .

The important point in this definition is that the joint distribution  $\mathbf{P}_1$  and the conditional distribution  $\mathbf{P}_2$  share the same parameter  $\theta$ . A useful specific case is obtained when the conditional distribution of  $\mathbf{P}_1$  matches the distribution  $\mathbf{P}_2$  (which is a conditional distribution by definition):

**Proposition 1** Let  $\mathbf{P}(X, C|\theta)$  be a generative model. The MAP estimator of the HGD model based on  $\mathbf{P}_1(X, C|\theta) = \mathbf{P}(X, C|\theta)$  and  $\mathbf{P}_2(C|X, \theta) = \mathbf{P}(C|X, \theta)$  is the generative solution (1) for  $\lambda = 1$  and the discriminative solution (2) for  $\lambda = 0$ .

**Proof** The objective functions are the same:  $\mathbf{Q}(X, C|\theta, \lambda = 1)\mathbf{P}(\theta) = \mathbf{P}(X, C|\theta)\mathbf{P}(\theta)$  for generative learning and  $\mathbf{Q}(C|C, \theta, \lambda = 0)\mathbf{P}(\theta) = \mathbf{P}(C|X, \theta)\mathbf{P}(\theta)$  for the discriminative learning.  $\square$

With this definition, most of the hybrid generative-discriminative methods proposed in the literature [4, 6, 5] can be defined using the HGD formalism. For example, in the model of Lasserre et al. [6], the distribution  $\mathbf{Q}^{(1)}$  is a HGD model since we have:

$$\mathbf{Q}^{(1)}(X, C|\theta, \lambda) = \mathbf{P}(C|X, \theta) \int_{\Theta} \mathbf{P}(X|\tilde{\theta})\mathbf{Q}^{(1)}(\tilde{\theta}|\theta, \lambda)d\tilde{\theta} \quad ,$$

and using the definition of the prior distribution for  $\tilde{\theta}$ , we get:  $\mathbf{Q}^{(1)}(X, C|\theta, \lambda = 1) = \mathbf{P}(C|X, \theta)\mathbf{P}(X|\theta) = \mathbf{P}(X, C|\theta)$  and  $\mathbf{Q}^{(1)}(C|X, \theta, \lambda = 0)$ .

The other large class of methods designed to improve the discriminative power of generative classifiers defines a distribution conditional to *score space* of the generative model [2]:  $\mathbf{P}_2(C|X, \tilde{\theta}, \theta) = \mathbf{R}(C|\frac{\partial}{\partial \tilde{\theta}} \log \mathbf{P}_1(\mathbf{X}|\theta), \tilde{\theta})$  where  $\tilde{\theta}$  are some regression parameters (the weights of the linear logistic regression for example). The well-known Fisher Kernel approach is a special case of this method. The main problem with this approach is that the learning is generally done sequentially: first  $\theta$  is learned using the generative model only to obtain the scores  $\frac{\partial}{\partial \tilde{\theta}} \log \mathbf{P}_1(\mathbf{X}|\theta)$ , and in second phase the classifier parameters  $\tilde{\theta}$  are estimated. However, in Section 3 the convergence results are based only on the true MAP estimator which jointly maximizes  $\theta$  and  $\tilde{\theta}$ .

**Semi-supervised learning** We already mentioned that the main interest of generative classification comes from the

fact that unlabelled data can really improve the performances. In addition to the labelled data, we observe  $\bar{n}$  unlabelled data  $(\bar{X}) = \{(x_{i+n})\}_{i=1}^{\bar{n}}$ . The extension of the previous approaches to the semi-supervised case is straightforward:

$$\hat{\theta}_\lambda = \arg \max_{\theta \in \Theta} \log \mathbf{Q}(X, C|\theta, \lambda) + \nu \log \mathbf{P}(\bar{X}|\theta) + \log \mathbf{P}(\theta)$$

where  $\nu$  is a non-negative weight useful to control the influence of unlabelled data, so that the generative learning corresponds exactly to the standard mixture-type semi-supervised EM algorithm [11]. This estimation is slightly different from [6], since they assumed that the unlabelled data are generated by the surrogate model having parameters  $\tilde{\theta}$ .

**A mixture-type Generative-Discriminative Blending model** First considering the pure discriminative learning, we can see that the discriminative learning corresponds to the maximization of the joint distribution  $\mathbf{Q}(x, c|\theta) = \mathbf{P}(c|x, \theta)\tilde{\mathbf{P}}(x)$  for any choice of the distribution  $\tilde{\mathbf{P}}$ . Using the same prior as before, it is straightforward to see that the MAP solution of the generative model  $\mathbf{Q}(x, c|\theta)$ , is exactly the discriminative learning problem (2) for any distribution  $\tilde{\mathbf{P}}(x)$ . We could choose a outlier distribution for  $\tilde{\mathbf{P}}(x)$ , for example a uniform distribution or the original model  $\mathbf{P}(x|\tilde{\theta})$  but with different parameters  $\tilde{\theta}$  for  $\mathbf{Q}(X, C|\theta, \tilde{\theta}, \lambda)$ , similarly to [6]:

$$\mathbf{P}(\theta)\mathbf{P}(\tilde{\theta}) \prod_{i=1}^n \left( (1-\lambda)\mathbf{P}(c_i|x_i, \theta)\mathbf{P}(x_i|\tilde{\theta}) + \lambda\mathbf{P}(c_i, x_i|\theta) \right). \quad (4)$$

This mixture-type model allows each data point to be generated either by the full joint model or by a sequence of two step: first the input  $x$  is chosen according to  $\tilde{\mathbf{P}}(x)$ . Then an output  $c$  is sampled from a multinomial using  $\mathbf{P}(c|x, \theta)$ . This approach is expected to be more robust than the others since only the points with high input densities contribute to the generative learning of the classifier. The following table summarizes the three different generative-discriminative blending models we are aware of:

| method            | $\mathbf{Q}(X, C, \theta, \tilde{\theta} \lambda)$  |
|-------------------|---|
| interpolation [4] | $\mathbf{P}(\theta)\mathbf{P}(\tilde{\theta}) \left( (1-\lambda) \prod_{i=1}^n \mathbf{P}(c_i x_i, \theta)\mathbf{U}(x_i \tilde{\theta}) + \lambda \prod_{i=1}^n \mathbf{P}(c_i, x_i \theta) \right)$ |
| mixture           | $\mathbf{P}(\theta)\mathbf{P}(\tilde{\theta}) \prod_{i=1}^n \left( (1-\lambda)\mathbf{P}(c_i x_i, \theta)\mathbf{U}(x_i \tilde{\theta}) + \lambda\mathbf{P}(c_i, x_i \theta) \right)$                 |
| LBM [6]           | $\mathbf{P}(\theta, \tilde{\theta}, \lambda) \prod_{i=1}^n \mathbf{P}(c_i x_i, \theta)\mathbf{P}(x_i \tilde{\theta})$   |

Despite its simplicity, the LBM model contains about twice as many parameters as the other methods. Moreover, this method requires additional care when choosing the prior probability  $\mathbf{P}(\theta, \tilde{\theta}, \lambda)$  because scaling problems might occur between different parameters<sup>1</sup>.

<sup>1</sup>The spherical prior  $\mathbf{P}(\theta)\mathbf{P}(\tilde{\theta}) \exp\{-\frac{1}{2\sigma^2}\|\theta - \tilde{\theta}\|^2\}$  proposed in [6] might not be appropriated if the parameters have different meanings, for example the mean and variance parameters of Gaussian distributions.

### 3 Theoretical results

Given any parameter  $\theta \in \Theta$ , the classification loss of the classifier is the *classification entropy* defined by  $C(\theta) = \mathbb{E}[-\log \mathbf{P}(c|x, \theta)]$  where expectation  $\mathbb{E}[\cdot]$  the expectation over the exact data distribution  $\mathbf{P}^*(x, c)$ . For the sake of simplicity, we do not consider the 0-1 here, but the experimental results at the end of the paper show that they are close connection between these two losses.

Given the tradeoff value  $\lambda$ , the MAP estimator  $\hat{\theta}_\lambda$  of  $\mathbf{Q}(X, C|\theta, \lambda)$  is estimated on the training sample  $(X, C)$  of size  $n$ , where we assume that the data points are i.i.d. A decision rule for a test data  $x$  is obtained using the conditional distribution  $\mathbf{Q}(c|x, \hat{\theta}_\lambda, \lambda)$ . We would like to know what is best choice of  $\lambda$  to obtain on average the best classification performances, or equivalently we look for the value  $\lambda_n^*$  minimizing  $\mathbb{E}[C(\hat{\theta}_\lambda)]$  over  $\lambda$ . To achieve this goal, we decompose the objective into the sum of a *bias* term independent of the training sample and a *variance* term which tends to zero with the training sample size:

$$\mathbb{E}[C(\hat{\theta}_\lambda)] = \underbrace{C(\theta_\lambda^*)}_{\text{bias}(\lambda)} + \underbrace{\mathbb{E}[C(\hat{\theta}_\lambda)] - C(\theta_\lambda^*)}_{\text{variance}(\lambda)}, \quad (5)$$

where  $\theta_\lambda^* = \arg \max_{\theta \in \Theta} \mathbb{E}[\log \mathbf{Q}(x, c|\theta, \lambda)]$  is the value of the HGD estimator if we had an infinite training sample size. Intuitively, we can expect that the discriminative estimation  $\lambda = 0$  has a small bias but a high variance compared to the generative estimation. This means that the function  $\text{bias}(\lambda)$  is increasing while the function  $\text{variance}(\lambda)$  is decreasing. In the remaining, we find the conditions under which this intuition is true.

Defining the matrices  $K_\lambda(\theta) = \mathbb{E}\left[\frac{\partial}{\partial \theta} \log \mathbf{Q}(x, c|\theta, \lambda) \frac{\partial}{\partial \theta^T} \log \mathbf{Q}(x, c|\theta, \lambda)\right]$  and  $J_\lambda(\theta) = \mathbb{E}\left[-\frac{\partial^2 \theta}{\partial \log \partial \log^T} \mathbf{Q}(x, c|\theta, \lambda)\right]$ , a classical result of asymptotic statistics gives:

**Theorem 1** (*asymptotic sampling distribution*)

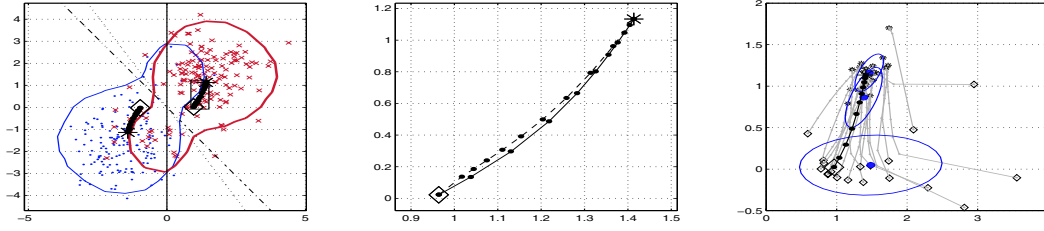
C1 For any  $\lambda \in [0, 1]$ , The third order partial derivatives of  $\mathbb{E}[\log \mathbf{Q}(x, c|\theta, \lambda)]$  according to  $\theta$  exist and are bounded in  $\Theta$ ,

C2 For all  $\theta \in \Theta$ ,  $K_\lambda(\theta)$  and  $J_\lambda(\theta)$  are (symmetric) definite positive.

Let  $J_\lambda = J_\lambda(\theta_\lambda^*)$  and  $K_\lambda = K_\lambda(\theta_\lambda^*)$ . If the regularity conditions C1 and C2 hold, then

$$\sqrt{n} \left( \mathbb{E}[\hat{\theta}_\lambda] - \theta_\lambda^* \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, K_\lambda^{-1} J_\lambda K_\lambda^{-1})$$

Here,  $\xrightarrow{\mathcal{D}}$  denotes the convergence in distribution. For references and a formal proof of this result, see *e.g.* Ripley's book [12]. This result is used to find an approximation of



**Figure 1.** Left: 300 data points from the simulation experiment, with class-conditional densities isocontours and asymptotic classification boundaries for the generative, discriminative and hybrid interpolation ( $\lambda = \frac{1}{2}$ ) methods. Middle: zoom on the mean parameter of the second class, where the continuous (resp. dashed) line corresponds to the interpolation (resp. LBM) HGD. The points '\*' and '◇' are the generative and discriminative solutions. The right plot shows the estimations of the mean for 20 different training sets. The ellipsoids show the estimation variances for  $\lambda \in \{0, \frac{1}{2}, 1\}$ .

the variance term for large training sample sizes since the expected classification loss around the asymptotic solution  $\theta_\lambda^*$  can be approximated locally by a quadratic function with Jacobian  $J_{D,\lambda} = \mathbb{E} \left[ -\frac{\partial^2 \theta}{\partial \log \partial \log^T} \mathbf{Q}(c|x, \theta_\lambda^*, \lambda) \right]$ . This is stated in the following proposition

**Proposition 2** Under conditions C1 and C2,  $E \left[ C(\hat{\theta}_\lambda) \right] = \gamma_n(\lambda) + o\left(\frac{1}{n}\right)$ , where  $\gamma_n(\lambda) = C(\theta_\lambda^*) + \frac{1}{2n} \text{tr} (J_{D,\lambda} K_\lambda^{-1} J_\lambda K_\lambda^{-1})$ . and  $o$  is a function satisfying  $\lim_{n \rightarrow \infty} n o\left(\frac{1}{n}\right) = 0$ .

**Proof** We have  $n \left( \mathbb{E} \left[ \log \mathbf{P}(x, c|\hat{\theta}_\lambda, \lambda) \right] - \log \mathbf{P}(x, c|\theta_\lambda^*, \lambda) \right) \xrightarrow{D} \frac{1}{2} Z^T J_{D,\lambda} Z$  where  $Z \sim \mathcal{N}(0, K_\lambda^{-1} J_\lambda K_\lambda^{-1})$  using Theorem 1 and  $\theta_\lambda^* = \arg \max_\theta \mathbb{E} [\log \mathbf{P}(x, c|\theta, \lambda)]$ . Finally, remarking that  $\mathbb{E} [Z^T J_{D,\lambda} Z] = \frac{1}{2n} \text{tr} (J_{D,\lambda} K_\lambda^{-1} J_\lambda K_\lambda^{-1})$  gives the result (see [9] for a similar derivation).  $\square$

We need to identify HGD methods for which the classification bias increases with  $\lambda$ :

**Definition 2** A *nested HGD model* is a HGD model such that for all  $\lambda > \tilde{\lambda}$  and all  $\theta \in \Theta$  there exist  $\tilde{\theta} \in \Theta$  such that  $\mathbf{Q}(C|X, \theta, \lambda) = \mathbf{Q}(C|X, \tilde{\theta}, \tilde{\lambda})$ .

**Proposition 3** (correct model assumptions) For any nested HGD model  $\mathbf{Q}(X, C|\theta, \lambda)$ , if there exists a unique  $\theta^*$  in  $\Theta$  such that  $\mathbf{Q}(x, c|\theta^*, \lambda = 1) = \mathbf{P}^*(X, C)$ , then,

- the expected classification loss is  $C(\theta_\lambda^*) + \frac{1}{2n} \text{tr} (J_{D,\lambda} J_\lambda^{-1}) + o\left(\frac{1}{n}\right)$  and,
- the generative learning ( $\lambda = 1$ ) has asymptotically the best expected classification loss compared to other HGD methods, i.e.  $\min_\lambda \gamma_n(\lambda) = \gamma_n(1) + o\left(\frac{1}{n}\right)$ .

**Proof** Under correct model assumptions,  $K_\lambda = J_\lambda$  (see e.g. [12]), so the first point is proved. The fact that the models are

nested implies that  $J_{D,\lambda} = J_{D,1}$  for all  $\lambda$ , so the best choice is obtained for the minimum of  $\text{tr} (J_{D,1} J_\lambda^{-1})$  for large  $n$ . Moreover, the generative estimator  $\hat{\theta}_1$  is an unbiased estimator of  $\theta^*$  since the model is exact. The Cramer-Rao bound tells us that the minimal variance of an unbiased estimator is the inverse of the Fisher information matrix  $J_1(\theta^*)^{-1}$ , which is precisely the asymptotic variance of the generative estimator. So  $J_\lambda^{-1} - J_1^{-1}$  is semi-definite positive. This implies  $\text{tr} (J_{D,1} J_\lambda^{-1}) \geq \text{tr} (J_{D,1} J_1^{-1})$  and proves the second point.  $\square$

Based only on definitions 1 and 2, it is difficult to find theoretical results under wrong model assumptions since there is no guaranty that two models with close trade-off parameters are similar. To identify HGD models with smoothed transition between the discriminative and generative extremes, we also need the following definition:

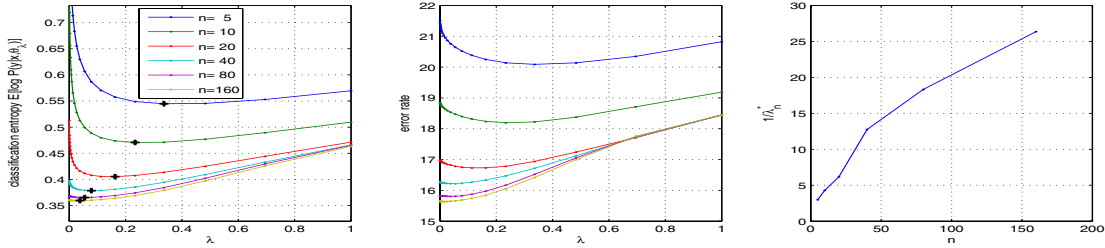
**Definition 3** A *HGD model for which  $H(\lambda) = \mathbf{Q}(X, C|\theta, \lambda)$  is  $C^k$  on  $]0, 1[$  is  $C^k$ -HGD model.*

A simple differentiation according to  $\lambda$  shows that the GDT models  $\mathbf{Q}^{(1)}$  and  $\mathbf{Q}^{(2)}$  are  $C^2$ -HGD models but it is not the case for the hybrid Naive Bayes model of [5] since it defines a finite set of maximization problems and the continuity cannot be guaranteed.

The main result of this paper gives sufficient conditions for the HGD estimation to be useful:

**Theorem 2** (wrong model assumptions) Let  $\mathbf{Q}(X, C|\theta, \lambda)$  be a nested  $C^\infty$ -HGD model and  $a = \frac{1}{2} \frac{d}{d\lambda} \text{tr} (J_{D,\lambda} K_\lambda^{-1} J_\lambda K_\lambda^{-1}) \Big|_{\lambda=0}$ . If  $\theta_1^* \neq \theta_0^*$  and  $a < 0$ , then there exists a HGD estimator better than the generative and the discriminative estimators for  $n$  sufficiently large. Moreover, if  $b = \frac{d^2}{d\lambda^2} \mathbb{E} [C(\theta_\lambda^*)] > 0$  then the optimal choice of the tradeoff parameter is asymptotically  $\lambda_n^* = \frac{1}{n} \frac{a}{b}$ .

**Proof** The optimal choice for lambda is  $\lambda_n^* =$



**Figure 2. Results of the simulated dataset. The left (middle) plot shows the expected classification entropy (test error rate) vs. the generative-discriminative tradeoff parameter  $\lambda$ . The optimal  $\lambda$  are marked with '+' (left plot) and show relation between  $\lambda_n^*$  and  $\frac{1}{n}$  (right plot).**

$\arg \min_{[0,1]} \gamma_n(\lambda) + o(\frac{1}{n})$ . The fact that the unique minimizer of  $E[C(\theta)]$  is unique and  $\theta_1^* \neq \theta_0^*$  imply that  $\gamma_n(0) < \gamma_n(1)$  so the generative method cannot be optimal. Moreover,  $\frac{d}{d\lambda} C(\theta_\lambda^*)|_{\lambda=0} = \frac{\partial C(\theta)}{\partial \theta}|_{\theta_0^*} \frac{d\theta_\lambda^*}{d\lambda} = 0$  because  $\theta_0^* = \arg \max_{\theta} C(\theta)$ , so that the condition  $a < 0$  implies that  $\gamma_n(\lambda)$  decreases in the neighborhood of 0. This proves that the discriminative solution is not optimal neither, so that  $0 < \lambda_n^* < 1$  for large  $n$ . The optimal  $\lambda$  can be estimated by minimizing the Taylor approximation of  $\gamma$  around 0:  $\gamma_n(\lambda) = C(\theta_0^*) + \frac{a}{n}\lambda + \frac{b+c/n}{2}\lambda^2 + o(\lambda^2)$  where  $c = \frac{d^2}{d\lambda^2} \text{tr}(J_{D,\lambda} K_\lambda^{-1} J_\lambda K_\lambda^{-1})|_{\lambda=0}$  is bounded, and the minimum of  $\gamma_n(\lambda)$  is asymptotically  $\frac{1}{n} \frac{a}{b}$ .  $\square$

## 4 Experiments

**Simulations** To illustrate the different HGD models already mentioned, we a binary ( $c \in \{-1, 1\}$ ) classification problem in two dimension where the inputs  $x$  are distributed according to a mixture of two scaled Gaussian distributions  $\mathbf{P}(x|\theta) = \frac{1}{2}\phi(x; (-1.5, -1.5); I_d) + \frac{1}{2}\phi(x; (1.5, 1.5); I_d)$  where  $\phi(x; \mu, \Sigma)$  denotes the pdf of a Gaussian distribution. The conditional probability of the output is a logistic link where the discriminative direction corresponds to the first variable:  $\mathbf{P}^*(c|x) = 1/(1 + e^{2cx_1})$

To estimate the classification boundary, we define a generative model using one Gaussian per class, but with the means constrained to be symmetric relatively to  $(0,0)$ :  $\mathbf{P}(x, c|\theta) = \frac{1}{2}\phi(x; c\theta, I_d)$ . An example of simulated data is given in Figure 1. On this figure, the mean parameter estimated using different values of  $\lambda$  is plotted, for the interpolation and the LBM HGD models. We numerically computed the fraction  $\frac{a}{b} \approx 0.183$  defined in Theorem 2, showing that the conditions are satisfied for the HGD method to give asymptotically better results than the generative and the discriminative estimators. The Figure 2 illustrates this. It is not plotted here, but the two HGD models give very similar results in terms of classification performances.

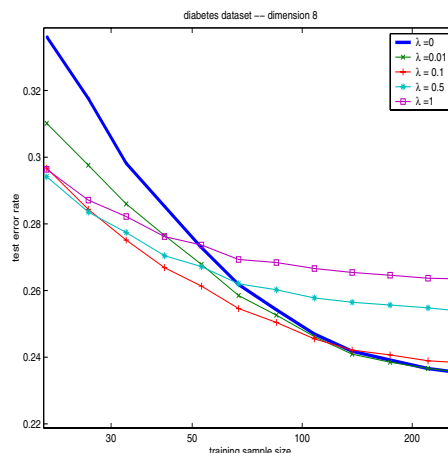
**Real data** We performed a test on 14 datasets from the UCI Machine Learning repository, using the Naive Bayes model as generative classifier, whose discriminative version corresponds to the multinomial logistic regression. We use the interpolation HGD method of [4] to compute the hybrid classifier. Since small training samples favor the generative classifier and large training samples favor the discriminative ones, we selected a training size between these two extremes<sup>2</sup> to show the possible improvements of the HGD estimation. On Table 1, we show the average and standard deviation of the test error rate (in %) for several real datasets (over 50 random test/train splits). The variables  $d$  and  $n$  are the data dimension and the training sample size. The columns  $\text{err}_G$ ,  $\text{err}_{\hat{\lambda}}$  and  $\text{err}_D$  show the generative estimator, the interpolation HGD estimator with  $\hat{\lambda}$  selected by CV10 and the discriminative estimator. The last column is the average of  $\hat{\lambda}$ .

The results show that substantial improvements in the classification rate can be obtained for intermediate values of  $\lambda$ . One can see that the training sample size is generally of the same order as the number of parameters, approximately 3 times the dimension  $d$ . We also note that the optimal  $\lambda$  can be very close to 1, especially when the classes are well separated. This is mainly due to the fact that the conditional and joint likelihoods are not at the same scale. Cross-validation over a predefined set of  $\lambda$  values enables us to find a good balance between the generative and discriminative objective functions.

As pointed out by Ng and Jordan [1] and [5], the strong model assumptions of the Naive Bayes model make it suitable to experiment HGD techniques. The graph on the left of Table 1 plots the test error rate vs. the training sample size for two datasets. This experiment is very similar to [1] in the pure generative or discriminative case. It shows that

<sup>2</sup>The hold out error rates of the Naive Bayes classifier and the multinomial logistic regression has been computed for various sample size and we selected the sample size wich gave approximately an equal error rate.

| dataset    | $d$ | $n$ | $\text{err}_D$  | $\text{err}_{\hat{\lambda}}$ | $\text{err}_G$  | $\hat{\lambda}$ |
|------------|-----|-----|-----------------|------------------------------|-----------------|-----------------|
| abalone    | 8   | 99  | $22.9 \pm 1.5$  | $22.1 \pm 1.1$               | $22.7 \pm 0.34$ | 0.012           |
| contracep. | 9   | 475 | $37.3 \pm 1.5$  | $37.3 \pm 1.5$               | $40.3 \pm 2.2$  | 0.01            |
| ionosphere | 34  | 97  | $16.8 \pm 3.2$  | $13.2 \pm 1.5$               | $20.1 \pm 5.6$  | 0.103           |
| optdigits  | 64  | 48  | $1.65 \pm 1.2$  | $1.52 \pm 1.1$               | $1.73 \pm 1.2$  | 0.174           |
| pageblocks | 10  | 60  | $4.21 \pm 2.1$  | $3.11 \pm 0.73$              | $3.69 \pm 1.3$  | 0.215           |
| spam       | 57  | 420 | $11.8 \pm 1.3$  | $10 \pm 0.96$                | $10.8 \pm 0.78$ | 0.262           |
| australian | 14  | 142 | $15.3 \pm 1.6$  | $13.6 \pm 0.97$              | $13.5 \pm 0.9$  | 0.54            |
| diabetes   | 8   | 45  | $27.7 \pm 3.1$  | $27 \pm 2.4$                 | $27.6 \pm 2.4$  | 0.588           |
| dna        | 180 | 514 | $5.82 \pm 0.83$ | $3.59 \pm 0.44$              | $5.64 \pm 0.52$ | 0.030           |
| german     | 20  | 73  | $31.4 \pm 2.7$  | $30 \pm 2.7$                 | $31.4 \pm 2.5$  | 0.5             |
| heart      | 13  | 179 | $17.1 \pm 3.3$  | $16.7 \pm 2.7$               | $16.8 \pm 2.9$  | 0.426           |
| letter     | 16  | 102 | $3.12 \pm 1.6$  | $2.63 \pm 0.84$              | $7.12 \pm 1.5$  | 0.005           |
| shuttle    | 9   | 31  | $9.76 \pm 3.9$  | $7.23 \pm 2.9$               | $9.79 \pm 3.2$  | 0.143           |
| vehicle    | 18  | 288 | $4.37 \pm 1.5$  | $3.9 \pm 1.5$                | $31.5 \pm 4.1$  | 0.001           |



**Table 1. Classification results on real datasets.**

the HGD classifier with  $\hat{\lambda} = 0.1$  outperforms both generative and discriminative classifiers for large range of training sample sizes.

## 5 Conclusion

Through the formal definition of a Hybrid Generative-Discriminative models, we proposed a unified framework in which most of the existing probabilistic methods which mix generative and discriminative learning can be identified and compared. This framework helped us to design new hybrid techniques, such as a robust HGD estimator or the tradeoff between the discriminative Fisher Kernel and the generative classification.

Contrary to most of the existing work on hybrid-generative models which focus on empirical studies, we tried to identify the necessary conditions allowing a HGD technique to be valid. Our results illustrate the importance of the geometry of the statistical model around the discriminative solution. Finally, numerical experiments confirmed theoretical statement and illustrate that the classification performance gains are potentially high, even with simple generative models.

## References

- [1] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 609–616, Cambridge, MA, 2002. MIT Press.
- [2] N.D. Smith. *Using Augmented Statistical Models and Score Spaces for Classification*. PhD thesis, University of Cambridge, 2003.
- [3] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In S. Solla M. Kearns and D. Cohn, editors, *Advances in Neural Information Processing Systems 11*, 1998.
- [4] G. Bouchard and B. Triggs. The tradeoff between generative and discriminative classifiers. In J. Antoch, editor, *Proc. of COMPSTAT'04, 16th Symposium of IASC*, volume 16. Physica-Verlag, 2004.
- [5] R. Raina, Y. Shen, A. Y. Ng, and A. K. McCallum. Classification with hybrid generative/discriminative models. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- [6] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, 2006*. to appear.
- [7] M. Kelm, C. Pal, and A. K. McCallum. Combining generative and discriminative methods for pixel classification with multi-conditional learning. In *Proc. of ICPR '06*. IEEE Computer Society, 2006.
- [8] A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to semi-supervised classifier design. In *AAAI*, pages 764–769, 2005.
- [9] T.J. O’Neil. A general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *JASA*, 75:154–160, 1980.
- [10] R. Greiner and W. Zhou. Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. In *Eighteenth national conference on Artificial intelligence*, pages 167–173, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.
- [11] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [12] B. D. Ripley. *Pattern Recognition and Neural Networks*. University Press, Cambridge, 1996.