Contrastive Pessimistic Likelihood Estimation for Semi-Supervised Classification

Marco Loog

Pattern Recognition Laboratory
Delft University of Technology
The Netherlands
e-mail: m.loog@tudelft.nl

The Image Section
University of Copenhagen
Denmark
http: prlab.tudelft.nl

June 10, 2014

Abstract

Improvement guarantees for semi-supervised classifiers can currently only be given under restrictive conditions on the data. We propose a general way to perform semi-supervised parameter estimation for likelihood-based classifiers for which, on the full training set, the estimates are never worse than the supervised solution in terms of the log-likelihood. We argue, moreover, that we may expect these solutions to really improve upon the supervised classifier in particular cases. In a worked-out example for LDA, we take it one step further and essentially prove that its semi-supervised version is strictly better than its supervised counterpart. The two new concepts that form the core of our estimation principle are contrast and pessimism. The former refers to the fact that our objective function takes the supervised estimates into account, enabling the semi-supervised solution to explicitly control the potential improvements over this estimate. The latter refers to the fact that our estimates are conservative and therefore resilient to whatever form the true labeling of the unlabeled data takes on. Experiments demonstrate the improvements in terms of both the log-likelihood and the classification error rate on independent test sets.

Keywords: maximum likelihood, semi-supervised learning, contrast, pessimism, linear discriminant analysis.

1 Introduction

A century after its inception [1–3], parameter estimation through maximum likelihood (ML) is still one of the most widely used statistical estimation techniques. In a more rudimentary form, maximum likelihood can even be traced back as far as the 18th century [4]. ML estimation is employed in fields as diverse as genealogy, imaging, genetics, astrophysics, physiology, and quantum communication, as is illustrated by many recent research works such as [5–17]. Moreover, new tools and techniques based on or related to ML are still being developed within modern statistics and related fields. Some recent examples are [18–23]. A satisfactory approach to ML-based estimation for semi-supervised classifiers, however, has not been developed so far.

In general, the aim of semi-supervised learning is to improve supervised classifiers by exploiting additional, typically easier to obtain, unlabeled data [24, 25]. Up to now, however, the literature has reported mixed results when it comes to such improvements; it is not always the case that semi-supervision leads to lower expected error rates or the like. On the contrary, severely deteriorated performances have been observed in empirical studies and theory shows that improvement guarantees can often only be provided under rather stringent conditions on the data we are dealing with [26–29].

In this work, we demonstrate when and how ML estimators for classification can be improved in the semi-supervised setting. We show that semi-supervised estimates can be constructed that are essentially closer to the estimates that would be obtained when also all the labels for all unlabeled data would be available in the training phase. That is, the semi-supervised estimates are closer to the estimates obtained with all labels available than the supervised estimates that rely on the same labeled instances as semi-supervision does, but that do not use the additional unlabeled data set.

In order to show the potential improvements semi-supervised classifiers can deliver, we introduce a novel, generally applicable estimation principle that extends likelihood estimation to the semi-supervised case in a consistent way. In particular, our method is *contrastive*, which refers to the fact that the objective function takes into account the original supervised solution in an explicit way. This enables the semi-supervised solution to explicitly control the potential improvements over the supervised solution. In addition, our method is *pessimistic*, which refers to the fact that the unlabeled data is treated as if it behaves in a worst kind of way, i.e., such that the semi-supervised estimates benefit the least from it. It makes the estimates conservative, but resilient to any possible state in which the unlabeled data can be encountered. We therefore refer to this principle as maximum contrastive pessimistic likelihood estimation or MCPL estimation for short.

1.1 Outline

In Section 3, the main theory is introduced, contrast and pessimism are further elucidated, and our core, general estimation principle, MCPL, is presented. In that same section, we also sketch the possibility of improved semi-supervised estimation by means of MCPL. Sections 4 and 5 provide a worked-out illustration and a further specification of our theory.

The former section introduces the MCPL-based version of LDA, proves in what way the semi-supervised LDA parameters are expected to really improve over the regular supervised ones, and sketches the heuristic employed to tackle the related optimization problem. The latter section, Section 5, provides extensive results on a range of data sets, comparing regular supervised LDA and an earlier proposed semi-supervised approach to LDA [30] with the novel semi-supervised LDA introduced here. Section 6 puts the results in a somewhat broader perspective and raises some open issues. Finally, Section 7 concludes. To begin with, however, we put our work in context, provide some preliminaries, introduce ML estimation and LDA, give an overview of the principal related works, and discuss related earlier findings.

2 Background and Preliminaries

The log-likelihood objective function for a K-class supervised classification problem takes on the general form

$$L(\theta|X) = \sum_{i=1}^{N} \log p(x_i, y_i|\theta) = \sum_{k=1}^{K} \sum_{j=1}^{N_k} \log p(x_{kj}, k|\theta)\,, \tag{1}$$

where class k contains a total of N_k samples, $N = \sum_{k=1}^{K} N_k$ is the total number of samples,

$$X = \{(x_i, y_i)\}_{i=1}^{N}$$

is the set of all labeled training pairs with $x_i \in \mathbb{R}^d$ d-dimensional feature vectors[1], and

$$y_i \in C = \{1, \ldots, K\}$$

are their corresponding labels. Denoted with x_{kj} is the jth sample from class $k \in C$. Here, every model parameter—specific to a particular class or not—is absorbed in $\theta \in \Theta$. The set Θ contains all parameter settings possible, thus defining the full class of models under consideration. Now, the supervised ML estimate, $\hat{\theta}_{\text{sup}}$, maximizes the above criterion:

$$\hat{\theta}_{\text{sup}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta|X)\,. \tag{2}$$

What follows is an overview of the main approaches to semi-supervised learning with a particular focus on likelihood-based methods. Specific attention will furthermore be given to semi-supervised approaches to LDA. For broader and more extensive literature reviews, we refer to [24] and [31].

[1] As is also common in many mathematical statistics and analysis textbooks, plain italic lowercase letters may indicate vectors and not only scalars.

2.1 Self-Learning and Expectation Maximization

With the current work, we in essence revisit a problem in ML estimation that has already been considered as early as the late 1960s. In 1968, Hartley and Rao sketched a general way of exploiting unlabeled data

$$U = \{u_i\}_{i=1}^{M}$$

in likelihood estimation of model parameters for the analysis of variance [32]. The basic idea is to consider all possible labelings that the unlabeled data could have and choose that labeling that achieves the largest log-likelihood. As such, this procedure still relies on ML estimation, but where the fully supervised model would merely optimize the log-likelihood of the parameters of the model, here the unobserved labels

$$V = \{v_i\}_{i=1}^{M}$$

of the unlabeled data in U are considered parameters over which the likelihood is maximized as well:

$$\underset{\theta \in \Theta}{\operatorname{argmax}} \left[L(\theta|X) + \max_{V \in C^M} \sum_{i=1}^{M} \log p(u_i, v_i|\theta) \right]. \tag{3}$$

Clearly, as the number of possible labelings grows exponentially with the number of unlabeled data points, even for fairly small sample sizes M this procedure is generally intractable.

A learning strategy that is often referred to as self-learning or self-teaching approaches the problem in a similar though greedy way. In its most most simple form, the classifier of choice is trained on the available labeled data in an initial step. Using this trained classifier, all unlabeled data or part of it are assigned a label. Then, in a next step, this now labeled data is added to the training set and the classifier is retrained with this enlarged set. Given the newly trained classifier, one can relabel the initially unlabeled data and retrain the classifier again with these updated labels. This process is then iterated until convergence, i.e., when the labeling of the initially unlabeled data remains unchanged.

McLachlan [33], in 1975, was probably the first to apply this procedure and indeed suggested it as a computationally more tractable alternative to the one in [32]. Similar procedures have been reintroduced throughout the last couple of decades (see, for instance, [34–36]). Outside of the literature on likelihood estimation, a procedure reminiscent of McLachlan's had already been proposed. In 1966, while dealing with an issue slightly different from semi-supervised learning, Nagy and Shelton proposed a general technique similar to self-learning [37]. One of the crucial differences is that the labeled data is only used to train the initial classifier. It does not play a role in any of the subsequent self-learning iterations. Also this procedure has been reconsidered many years after it was initially suggested, e.g. in [34].

Possibly the best known semi-supervised likelihood-based approach treats the absence of labels as a classical missing-data problem and integrates out these nuisance parameters

to come to a new, full model likelihood [38–40]

$$L(\theta|X) + \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} p(u_i, k|\theta) \right).$$

Its maximization over θ typically relies on the classical technique of expectation maximization (EM). In 1973, [41] and [42] were possibly the first to consider this specific problem explicitly, though [43] had already employed such formulation in its applied work in 1972.

At a first glance, self-learning and EM may seem different ways of tackling the semi-supervised classification problem, but there are clear parallels. Indeed, where EM provides soft class assignments to all unlabeled data, self-learning just assigns every such instance in a hard way to one unique class in every iteration. In fact, [34] effectively shows that self-learners optimize the same objective as EM does. Similar observations have been made in [44] and [45].

The major problem with the aforementioned methods is that they can suffer from severely deteriorated performance with increasing numbers of unlabeled samples. This behavior, already extensively studied [30,46–48], is often caused by model misspecification, i.e., the statistical class of models with parameters θ is not able to properly fit the actual data distribution. We note that this is in contrast with the supervised setting, where most classifiers are capable of handling mismatched data assumptions rather well and adding more labeled data typically improves performance. The latter is in line with the behavior many misspecified likelihood models display [49].

2.2 Density-Ratio Correction

A rather different approach to semi-supervised estimation for likelihood-based models is offered in [50], in which the problem of semi-supervised learning is basically treated as one of learning under covariate shift [51]. Covariate shift is the setting in which the posterior distribution of the labels given the data, $p(y|x)$, remains the same, while the marginal $p(x)$ might change when going from the training to the testing phase. Following [52], the main idea in [50] is that the marginal distribution over the feature space can be better estimated based on all data, both labeled and unlabeled. Subsequently, the density ratio between this estimate and the marginal estimate based on labeled data only can be exploited to weight the training data by means of their importance, as generally suggested in [51].

In their work, the authors prove that, asymptotically, this semi-supervised learning procedure works better than its regular, supervised counterpart. Next to the fact that results hold only asymptotically, the behavior of this semi-supervised learner seems to depend strongly on the way the density ratio is determined. In the finite sample setting, one may run into similar kind of problems as those sketched in the previous subsection: choosing the incorrect model for estimating the density ratio of the marginal feature distributions, could lead to deteriorated performance instead of performance improvements. Experimental results in both [50] and [52] seem to reflect this.

2.3 Intrinsically Constrained Estimation

In recent years, the author proposed an essentially different take on semi-supervised learning [53, 54]. On a conceptual level, the idea is that the available unlabeled data indirectly puts restrictions on the parameters possible, i.e., it basically allows us to look at a set that is smaller than the initial set Θ. A first operationalization of this idea has been studied for the simple nearest mean classifier (NMC, [53]). It exploits constraints that are known to hold for this classifier, defining relationships between the class-specific parameters and certain statistics that are independent of the specific labeling. In particular, for the NMC the following constraint can be exploited:

$$\hat{\mu} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k \,, \tag{4}$$

with $\hat{\mu}$ the estimated overall sample mean of the data, $\hat{\mu}_k$ the sample means of the K classes, and $\hat{\pi}_k = \frac{N_k}{N}$ the estimates of the class priors. In the supervised setting this constraint is automatically fulfilled [55]. Its benefit only becomes apparent, therefore, with the arrival of unlabeled data that can be used to improve the label-independent estimate $\hat{\mu}$. Using this more accurate estimate results in a violation of the constraint. Fixing the constraint by properly adjusting the $\hat{\mu}_k$s, these label-dependent estimates become more accurate as well.

Supervised LDA can be improved in a similar way. The same constraint in Equation (4) holds, but for LDA additional ones involving the class-conditional covariance matrix apply. Notably, we have that the covariance matrix of all the data, the total covariance $\hat{\Sigma}_T$, equals the sum of the covariance between the class means, the between-class covariance $\hat{\Sigma}_B$, and the class-conditional covariance matrix $\hat{\Sigma}$ (which is also referred to as the within-class covariance) [55]:

$$\hat{\Sigma}_T = \hat{\Sigma}_B + \hat{\Sigma} \,. \tag{5}$$

These additional constraints further restrict the possible semi-supervised solutions, allowing for more significant improvements over the regular supervised classifier [30, 54].

The aforementioned works enforce the constraints imposed in a rather ad hoc way. A somewhat more principled constrained likelihood approach is suggested in [56, 57]. Generally, given any constraint $h(\theta) = 0$ that the parameters of the semi-supervised classifier should comply with, the idea is to maximize the original likelihood from Equation (1)—as in Equation (2), but subject to the constraint, i.e., we solve

$$\begin{aligned} &\underset{\theta \in \Theta}{\operatorname{argmax}} \; L(\theta | X) \\ &\text{subject to } h(\theta) = 0 \,. \end{aligned}$$

[57] shows, for instance, how to formulate the constrained NMC from [53] in this way. A major shortcoming of this approach is that such constraints must be identified for each classifier. For this reason, its applicability is currently limited.

A second and more recent instantiation of our general idea, coined in [53], does allow for broader applicability [58, 59]. The optimization suggests to find those parameters that maximize the likelihood on the labeled data set X, but only allows solutions that can be achieved with a data set that includes labeled versions of the initially unlabeled instances as well. In terms of a likelihood formulation, what it suggests to solve is the following:

$$\underset{\theta \in T}{\operatorname{argmax}} \; L(\theta|X) \text{with } T = \left\{ \underset{t \in \Theta}{\operatorname{argmax}} L(t|X_V) \middle| V \in C^M \right\}. \tag{6}$$

The first important ingredient is the set X_V, which is the labeled data set X augmented with the unlabeled data U combined with the labels in V. So

$$X_V = X \cup \{(u_i, v_i)\}_{i=1}^{M}$$

is a fully labeled data set for all $V \in C^M$. The second important ingredient is the set T, which typically is a proper subset of the original parameter set Θ. This set T contains all possible classifier parameters t that are obtained by training classifiers on all of the possible fully labeled data sets X_V. As we need to consider all possible labelings for the unlabeled data, this brings us back to Hartley and Rao's intractable method [32]. In [58] and [59], this problem is overcome by introducing the possibility of fractional or soft labels, resulting in a well-behaved quadratic programming problem for the case of the least squares classifier considered there.

Putting our earlier work further in the appropriate context, we should finally mention [60] and [61], where likelihood-based semi-supervised learning guided by particular constraints is considered as well. The crucial difference is that the constraints proposed in these works are typically derived from domain knowledge and very task specific. If these a priori constraints are correct, a learner can obviously benefit from them, even in the supervised case. If they are incorrect they may lead to severely deteriorated performance. So where these constraints are classifier-extrinsically motivated, any other method in this subsection relies on intrinsically motivated constraints, which are fixed as soon as the data is available and the choice of classifier is made.

2.4 Supervised and Semi-Supervised LDA

As our worked-out example in Sections 4 and 5 concerns LDA, this subsection turns to its associated likelihood and the specific semi-supervised solutions that have been proposed for this classical technique.

Compared to Equation (1), the log-likelihood objective function for K-class LDA takes on a more specific form. We can write [62]

$$\begin{aligned} L_{\text{LDA}}(\theta|X) &= \sum_{i=1}^{N} \log p(x_i, y_i|\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{N_k} \log p(x_{kj}, k|\pi_k, \mu_k, \Sigma) = \sum_{k=1}^{K} \sum_{j=1}^{N_k} \log \pi_k g(x_{kj}|\mu_k, \Sigma)\,, \end{aligned} \tag{7}$$

with $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma)$, π_k the class priors, μ_k the class means, and Σ the class-conditional covariance matrix. The g, on the last line, denotes the normal (or $g$aussian) probability density function. Of course, to find the supervised solution, we solve the maximization already noted in Equation (2), which leads to the well-known ML estimates of the parameters of regular supervised LDA.

Semi-supervised LDA has been considered both in theoretical and methodological work. The main example in Hartley and Rao's work [32] treats univariate LDA in the semi-supervised setting. Also McLachlan [33] focusses on LDA. Following these contributions, other early studies of the use of unlabeled data in LDA can be found in [39,40,63] and [64]. Self-learned and intrinsically constrained versions of LDA have been compared in [54] and [30]. As LDA is a likelihood-based classifier, most of the general theory from papers quoted throughout this section applies.

Let us finally remark that various contributions from a large number of disciplines still employ classical, supervised LDA as their decision rule of choice. A handful of recent examples from the applied and natural sciences can be found in some of the earlier-mentioned references: [5, 6, 8, 9, 13]. Semi-supervised versions of LDA, however, have not been widely applied. The general shortcoming mentioned in Subsection 2.1, the fact that self-learned and EM-versions can give sharply inferior performance, probably contributes to this.

3 Contrastive Pessimistic ML

For none of the aforementioned semi-supervised learning schemes and classifiers, there are currently any generally applicable guarantees when it comes to performance improvements. The learning strategy that we devise in this section does allow for such a guarantee on the training set in a strict way. This we will show in Section 4. The main, general theory is provided in the current section.

Consider the fully labeled data set

$$X_{V^*} = X \cup \{(u_i, v_i^*)\}_{i=1}^M .$$

It is similar to X_V considered in Subsection 2.3, but we now assume that V^* contains the true labels v_i^* belonging to the feature vectors in U. Define

$$\hat{\theta}_{\text{opt}} = \operatorname*{argmax}_{\theta \in \Theta} L(\theta | X_{V^*}) ,$$

which gives the classifier's parameter estimates on the full training set in which also the unlabeled data is labeled. With respect to this enlarged training set X_{V^*}, the estimate $\hat{\theta}_{\text{opt}}$ is optimal by construction and cannot be improved upon. As the supervised parameters in $\hat{\theta}_{\text{sup}}$ are estimated merely on a subset X of X_{V^*}, we have

$$L(\hat{\theta}_{\text{sup}} | X_{V^*}) \leq L(\hat{\theta}_{\text{opt}} | X_{V^*}) .$$

In the semi-supervised setting, both X and U are at our disposal, but V^* has not been observed. We have more information than in the supervised setting, but less than in

the optimal, fully labeled case. The principal result obtained in this section is that, for likelihood-based classifiers, semi-supervised parameter estimates $\hat{\theta}_{\text{semi}}$ obtained by means of MCPL are essentially in between the corresponding supervised and the optimal estimates:

$$L(\hat{\theta}_{\text{sup}}|X_{V^*}) \leq L(\hat{\theta}_{\text{semi}}|X_{V^*}) \leq L(\hat{\theta}_{\text{opt}}|X_{V^*})\,.$$

In itself, this result might not seem all too helpful as we can easily come up with a semi-supervised parameter estimate for which these inequalities are trivially fulfilled: take $\hat{\theta}_{\text{semi}}$ to equal $\hat{\theta}_{\text{sup}}$. However, we first want to clarify that the inequality holds generally for MCPL before we proceed and make the claim that strict improvements by means of MCPL over regular supervised estimation can be expected. That is, we argue, at least for particular classifiers, that

$$L(\hat{\theta}_{\text{sup}}|X_{V^*}) < L(\hat{\theta}_{\text{semi}}|X_{V^*})\,,$$

i.e., the log-likelihood on the fully labeled set X_{V^*} obtained by the semi-supervised estimates is strictly larger than that obtained under supervision. For LDA, this is proven in Section 4.

3.1 Contrast and Pessimism

To be able to construct a semi-supervised learner that improves upon its supervised counterpart, we take the supervised estimate into account explicitly. Therefore, we consider the difference in loss incurred by $\hat{\theta}_{\text{semi}}$ and $\hat{\theta}_{\text{sup}}$.

Before doing so, however, we first introduce some notation. We define q_{ki} to be the hypothetical posterior $P(k|u_i)$ of observing a particular label k given the feature vector u_i. We may interpret the q_{ki} as soft labels for every u_i and will also refer to them as such. This respects the fact that classes may be overlapping and not every u_i can be be assigned unambiguously to a single class. By definition, $\sum_{k\in C} q_{ki} = 1$. More precisely, we can state that the K-dimensional vector $q_{\cdot i}$ is an element of the $K-1$-simplex Δ_{K-1} in $\mathbb{R}^K$:

$$q_{\cdot i} \in \Delta_{K-1} = \left\{ (\rho_1 \dots \rho_K)^T \in \mathbb{R}^K \,\middle|\, \sum_{i=1}^{K} \rho_i = 1, \rho_i \geq 0 \right\}.$$

Provided that these posteriors are given, we can express the log-likelihood on the complete data set for any θ as

$$L(\theta|X,U,q) = L(\theta|X) + \sum_{i=1}^{M}\sum_{k=1}^{K} q_{ki} \log p(u_i,k|\theta)\,, \tag{8}$$

in which the dependence on the q_{ki}s is explicitly indicated also on the left-hand side by means of the variable q. Note that use of these soft labels in q allows more flexibility than just using a set of hard labels $V \in C^M$, such as was for instance done in Equations (3) and (6).

For a given q, the relative improvement of any semi-supervised estimate θ over the supervised solution can now be expressed as follows:

$$CL(\theta, \hat{\theta}_{\text{sup}}|X, U, q) = L(\theta|X, U, q) - L(\hat{\theta}_{\text{sup}}|X, U, q)\,. \tag{9}$$

This contrasts the semi-supervised solution with the regular supervised solution obtained on the data set X, enabling us to explicitly check to what extent semi-supervised improvements are possible in terms of log-likelihood. As we are dealing with a semi-supervised problem, q is unknown and we cannot use Equation (9) directly for optimization. The choice we make now is the most pessimistic one: we are going to assume that the true (soft) labeling is most adverse against any semi-supervised approach and consider the q that minimizes the gain in likelihood. That is, our objective function becomes

$$CPL(\theta, \hat{\theta}_{\text{sup}}|X, U) = \min_{q \in \Delta_{K-1}^{M}} CL(\theta, \hat{\theta}_{\text{sup}}|X, U, q)\,, \tag{10}$$

where

$$\Delta_{K-1}^{M} = \prod_{i=1}^{M} \Delta_{K-1}\,,$$

the Cartesian product of M simplices.

3.2 MCPL Estimation

We are now ready to define MCPL estimation, which extends general likelihood estimation for supervised learners to the general semi-supervised case.

Definition 1 (MCPL)**.** *Let $\hat{\theta}_{\text{sup}}$ be the supervised ML estimate maximizing $L(\theta|X)$ and let U be a set of unlabeled data. A maximum contrastive pessimistic likelihood estimate, $\hat{\theta}_{\text{semi}}$, is an estimate that maximizes the criterion $CPL(\theta, \hat{\theta}_{\text{sup}}|X, U)$ in Equation (10), i.e.,*

$$\hat{\theta}_{\text{semi}} = \operatorname*{argmax}_{\theta \in \Theta} CPL(\theta, \hat{\theta}_{\text{sup}}|X, U)\,. \tag{11}$$

Maximizing the objective function CPL for θ leads to a rather conservative estimate, because of the pessimistic choice of q. But we need this choice, in combination with the contrastive nature of the objective function, to be able to guarantee that the following holds.

Lemma 1.

$$L(\hat{\theta}_{\text{sup}}|X_{V^*}) \leq L(\hat{\theta}_{\text{semi}}|X_{V^*}) \leq L(\hat{\theta}_{\text{opt}}|X_{V^*})\,. \tag{12}$$

There is, however, no necessity to consider true hard labels for U, as is done in the set X_{V^*}. Similarly, for the true soft labeling $q^* \in \Delta_{K-1}^{M}$, we can guarantee the following.

Lemma 2.

$$L(\hat{\theta}_{\text{sup}}|X, U, q^*) \leq L(\hat{\theta}_{\text{semi}}|X, U, q^*) \leq L(\hat{\theta}_{\text{opt}}|X, U, q^*)\,, \tag{13}$$

To see that both lemmas indeed hold, consider Equation (11). Because we can take $\theta = \hat{\theta}_{\text{sup}}$, 0 is always among the minimizers in this equation. As a consequence, the maximum will never be smaller than 0:

$$\max_{\theta \in \Theta} CPL(\theta, \hat{\theta}_{\text{sup}}|X, U) \geq 0\,.$$

Looking at Equation (9), this means that the difference between the semi-supervised and the supervised log-likelihood is larger than 0, but as this holds even for the worst choice of q, it must also hold for the true soft labeling q^* and the true hard labeling considered in X_{V^*}. From this, both Equations (12) and (13) follow, which shows the lemmas to hold.

3.3 Prospects of Improved Estimates

If we can show for a classifier that we can expect the inequalities in Lemmas 1 and 2 to be strict, then we can conclude that the semi-supervised parameter estimates are essentially better than those obtained under supervision. When can we expect this to happen? There are at least two different ways.

Firstly, a semi-supervised classifier can be better if the true underlying soft labeling q^* is less adversarial than the worst-case that is considered in MCPL estimation. Even though we cannot give any general quantitative statement on how often this happens, we can imagine that this is quite likely. Secondly, we can expect improvements in case the set of feature vectors of the labeled instances, X, is an ill representation of the complete set of labeled and unlabeled data, X and U. It is clear that nothing can be gained in the other extreme, where the feature vectors in U are just exact copies of those in X. In that case, MCPL estimation would just recover the supervised estimate. In the next section, we use such ill-representation argument to show that semi-supervised LDA typically outperforms its supervised counterpart.

4 MCPL Version of LDA

Combining MCPL estimation as defined in Subsection 3.2 with the log-likelihood formulation of regular supervised LDA from Equation (7) leads to our proposal of a proper semi-supervised version of LDA. Following the previous section, we have

$$L_{\text{LDA}}(\hat{\theta}_{\text{sup}}|X_{V^*}) \leq L_{\text{LDA}}(\hat{\theta}_{\text{semi}}|X_{V^*})\,.$$

Here and in what follows, the subscripted LDA makes explicit that we are specifically considering LDA. Subsection 4.3 briefly presents the heuristic we used to carry out the necessary maximinimization to actually obtain $\hat{\theta}_{\text{semi}}$. But first, in the next two subsections, we demonstrate that we can expect improved semi-supervised estimation.

4.1 Preliminaries

As the normal density $g(x|\mu_k, \Sigma)$ makes up an exponential family, it can be reparameterized into a so-called canonical parametrization such that it is concave in its parameters [65,66]. Denote this reparametrization by ϑ. For fixed q, $L_{\mathrm{LDA}}(\vartheta|X, U, q)$ is also concave. Now, by definition of the MCPL estimate

$$\begin{aligned}
&\max_{\vartheta\in\Theta} \; CPL_{\mathrm{LDA}}(\vartheta, \hat{\vartheta}_{\mathrm{sup}}|X, U) = \\
&\max_{\vartheta\in\Theta} \; \min_{q\in\Delta_{K-1}^{M}} CL_{\mathrm{LDA}}(\vartheta, \hat{\vartheta}_{\mathrm{sup}}|X, U, q) = \\
&\max_{\vartheta\in\Theta} \; \min_{q\in\Delta_{K-1}^{M}} \left[L_{\mathrm{LDA}}(\vartheta|X, U, q) - L_{\mathrm{LDA}}(\hat{\vartheta}_{\mathrm{sup}}|X, U, q)\right].
\end{aligned}$$

From this, it is not difficult to see that for fixed q, CL_{LDA} is concave in ϑ and for fixed ϑ, CL_{LDA} is linear in q. So CL_{LDA} is in fact concave-convex on $\Theta \times \Delta_{K-1}^{M}$. In addition, Δ_{K-1}^{M} is compact and so we can invoke the important minimax corollary by Sion [67] that allows us to interchange the maximization and minimization, which in turn means that the solution to the above maximiminimzation is a saddle point [68]. Moreover, the estimate $\hat{\vartheta}_{\mathrm{semi}}$ is unique if CL_{LDA} is strictly concave in ϑ [68]. For this to hold, Σ should be positive definite. This is generally the case if at least $d + K$ feature vectors are in the training set [55], which, from now on, we assume to hold.

Going back from the canonical parametrization ϑ to our original θ, we see that the the maximiminimzation also leads to a unique solution for $\hat{\theta}_{\mathrm{semi}}$. This will be important in what follows.

4.2 Semi-Supervised Improvements

We consider $CL_{\mathrm{LDA}}(\theta, \hat{\theta}_{\mathrm{sup}}|X, U, q)$, which is Equation (9) with the particular choice of the likelihood from Equation (7). Leaving q fixed, we saw that there is a unique maximizer for CL_{LDA}. Fixing q, the supervised part of the contrastive likelihood does not play an essential role in the objective function. It merely provides an offset, and the maximizer of CL_{LDA} is equal to the maximizer of $L_{\mathrm{LDA}}(\theta|X, U, q)$. Now, the latter is a weighted version of standard LDA—the weights are provided by q—and it is not difficult to show that, for every class $k \in C$, the optimal ML parameter estimates are given by

$$\begin{aligned}
\hat{\pi}_k &= \frac{N_k + \sum_{i=1}^{M} q_{ki}}{N + M}, \\
\hat{\mu}_k &= \frac{\sum_{j=1}^{N_k} x_{kj} + \sum_{i=1}^{M} q_{ki} u_i}{N_k + \sum_{i=1}^{M} q_{ki}},
\end{aligned} \tag{14}$$

while the estimate of the average class-conditional covariance matrix becomes

$$\hat{\Sigma} = \frac{1}{N + M} \sum_{k=1}^{K} \left[\sum_{j=1}^{N_k} (x_{kj} - \hat{\mu}_k)(x_{kj} - \hat{\mu}_k)^T + \sum_{i=1}^{M} q_{ki} (u_i - \hat{\mu}_k)(u_i - \hat{\mu}_k)^T \right]. \tag{15}$$

Note that the total data mean equals

$$\hat{\mu}^{\text{semi}} = \frac{1}{N+M}\left[\sum_{i=1}^{N} x_i + \sum_{i=1}^{M} u_i\right], \tag{16}$$

which is independent of the soft labels q. We now additionally note that also for weighted LDA, for any choice of q, the constraint in Equation (4) holds. The MCPL solution $\hat{\theta}_{\text{semi}}$ will have corresponding pessimistic soft labels $\hat{q}^{\text{semi}}$ and therefore satisfies the constraint as well: $\hat{\mu}^{\text{semi}} = \sum_{k=1}^{K} \hat{\pi}_k^{\text{semi}} \hat{\mu}_k^{\text{semi}}$.

Now, if semi-supervised learning does not improve over the supervised estimate, $\hat{\theta}_{\text{semi}}$ should equal the initial supervised solution $\hat{\theta}_{\text{sup}}$, because the estimate is unique (see Subsection 4.1). This, in turn, implies that we also have $\hat{\mu}^{\text{semi}} = \sum_{k=1}^{K} \hat{\pi}_k^{\text{sup}} \hat{\mu}_k^{\text{sup}}$. But as the supervised solution is trained on X only, it should simultaneously fulfil the constraint in Equation (4) with the total data mean equal to

$$\hat{\mu}^{\text{sup}} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \tag{17}$$

i.e., the sample average of X. We therefore have:

$$\hat{\mu}^{\text{sup}} = \sum_{k=1}^{K} \hat{\pi}_k^{\text{sup}} \hat{\mu}_k^{\text{sup}} = \hat{\mu}^{\text{semi}}\,.$$

If the feature vectors of our classification problem come from a continuous distribution then, unless U is empty, the probability that $\hat{\mu}^{\text{sup}}$ equals $\hat{\mu}^{\text{semi}}$ is zero. This, in turn, implies that we can expect $\hat{\theta}_{\text{semi}}$ to be different from $\hat{\theta}_{\text{sup}}$ and, therefore, improve upon it. With this, we have proven our first main result concerning semi-supervised LDA.

Theorem 1. *Given $M \geq 1$ and $N \geq d + K$, if the feature vectors are continuously distributed, the strict inequality*

$$L_{\text{LDA}}(\hat{\theta}_{\text{semi}}|X_{V^*}) > L_{\text{LDA}}(\hat{\theta}_{\text{sup}}|X_{V^*})$$

holds almost surely.

We should note that if the feature distribution is discrete, the inequality holds with a probability smaller than one. Nonetheless, when either the number of discrete elements of the distribution, the number N of labeled points, or the number M of unlabeled feature vectors is large, the probability that the inequality is strict typically gets close to one. We dare to conjecture that Theorem 1 will be accurate for many practical purposes, even in the discrete case.

What we can say in the discrete case is that the probability that $\hat{\mu}^{\text{sup}}$ does not equal $\hat{\mu}^{\text{semi}}$ is nonzero and, therefore, we at least have strict improvement in expectation.

Theorem 2. *Given $M \geq 1$ and $N \geq d + K$, if the feature distribution is nondegenerate, in the sense that there is not a single feature vector occurring with probability 1, we have*

$$E[L_{\text{LDA}}(\hat{\theta}_{\text{opt}}|X_{V^*})] \geq E[L_{\text{LDA}}(\hat{\theta}_{\text{semi}}|X_{V^*})] > E[L_{\text{LDA}}(\hat{\theta}_{\text{sup}}|X_{V^*})]\,,$$

where the expectation is taken over U.

Hence, LDA parameter estimation by means of MCPL is, in the average, always better than classical supervised log-likelihood estimation.

4.3 Solving the Maximinimization

As was discussed in Subsection 4.1 already, the objective function, as provided by Equation (9), is linear in q and strictly concave in θ. As a result, we know that we are looking for a saddle point solution with a unique optimizer for θ. Moreover, we know there are no other local saddle point solutions for this maximinimization problem [68]. The basis of our heuristic to come to an MCPL estimate for the parameters of semi-supervised LDA are the following two steps between which the optimization alternates.

1. Given a soft labeling q, the optimal, maximizing LDA parameters θ are estimated by means of Equations (14) and (15).

2. Given LDA parameters θ, the gradient ∇ for q is calculated, and q is changed to $q - \alpha\nabla$, with $\alpha > 0$ the step size. The following should be noted:

   (a) $q - \alpha\nabla$ is not guaranteed to be in Δ_{K-1}^{M}, so we project back into this set in every iteration [69];

   (b) the objective function is linear in q, so the gradient ∇ is easily obtained:

   $$\nabla_{ki} = \log \pi_k g(x_{ki}|\mu_k, \Sigma) - \log \hat{\pi}_{k\,\text{sup}} g(x_{ki}|\hat{\mu}_{k\,\text{sup}}, \hat{\Sigma}_{\text{sup}})\,;$$

   (c) we want to minimize for q, so we change its value in the direction opposite of the gradient, i.e., with $-\alpha$.

In our experiments in Section 5, the step size α is decreased as one over the number of iterations. Furthermore, we limit the maximum number of iterations to 1000. In addition, if the maximin objective does not change more than 10^{-6} in one iteration, the optimization is halted. With these settings, in our experiments, the maximum number of iterations is reached seldom (in less than one in every thousand cases).

Finally, we remark that care should be taken when calculating the necessary log-likelihoods or any of the related quantities. For example, the logarithm of the determinant of the average class covariance matrices can, especially for moderate- and high-dimensional problems, easily results in numerical infinities. Fairly reliable results can, in this instance, be obtained by determining the singular values of the covariance matrix through an SVD and taking the sum of the logarithm of these values.

5 Experiments and Results with LDA

Having presented the specific theory for semi-supervised LDA and a heuristic approach to find its MCPL parameters in Section 4, there are four main issues we want to investigate experimentally. To start with, the theory states that semi-supervised LDA estimates are better on the training data at hand given the log-likelihood as the performance measure. The two questions this raises are, firstly, how do these estimates compare to the supervised estimates on new and previously unseen test data? And secondly, how do they perform and compare in terms of the 0-1 loss, i.e., the classification error? Concerning the second point, we remark that the relation between likelihood and error rate is not necessarily monotonic and a higher likelihood does not necessarily lead to a lower error. It is only in recent years that considerable effort has been spent on understanding the nontrivial relationship between the criterion a classifier optimizes (here the likelihood) and how that classifier performs in terms of any other criterion of interest (here the error rate). Refer, for instance, to [70–75]. Thirdly, we measure the log-likelihood for the various parameter estimates also on the training set. This gives us a basic check on the performance of our optimization heuristic: we should find that the semi-supervised solutions never deteriorates the supervised solution and typically even improves upon it. The final, fourth point is to compare our theoretically underpinned method to the semi-supervised LDA technique from [30], which enforced the constraints in Equations (4) and (5) in an ad hoc way. It puts our novel method in a broader perspective, as the earlier method has been studied extensively already. Among others, this constrained LDA has been shown to perform much better than self-learning or EM approaches to LDA and to be competitive with transductive SVM [76] and even entropy regularized logistic regression [77], especially in the small sample setting.

5.1 Data Sets and Preprocessing

We chose 16 data sets from the UCI Machine Learning Repository [78] to perform our experiments on. The full names can be found in Table 1. The same table contains abbreviated names that we use to refer to the data sets in the other tables and throughout the text.

A main criterion for choosing these particular data sets was their size. We wanted to be able to easily generate labeled and unlabeled training sets from them plus independent test sets and we wanted especially the last two sets to have a fair size. In addition, we wanted to limit the computational burden and therefore did not choose too high-dimensional sets. Moreover, in order to rid ourselves of potential problems with singular class-conditional covariance matrices or numerical challenges related to this, the complete data sets were preprocessed in the following way. In a first step, the variance of every individual feature was normalized to one. A feature was removed altogether if its variance was numerically zero. In a second step, PCA was applied to the full sets and 999‰ of the variance was retained in order to remove linearly dependent features. We note that reducing the dimensionality essentially changes the likelihood of a data set, but that any nonsingular linear

Table 1: Full names and abbreviations of the 16 data sets from the UCI Machine Learning Repository [78]. Also included are the requested references to some of the original papers related to these data sets.

full data set name	abbreviated name	citation
banknote authentication	banknote	
climate model simulation crashes	climate	[79]
first-order theorem proving	first-order	[80]
gas sensor array drift	gas	[81]
landsat satellite	landsat	
letter recognition	letter	
low resolution spectrometer	low	
magic gamma telescope	magic	
miniboone particle identification	miniboone	
optical recognition of handwritten digits	optical	
pen-based recognition of handwritten digits	pen-based	
qsar biodegradation	qsar	[82]
shuttle	shuttle	
skin segmentation	skin	[83]
spambase	spambase	
spectf heart	spectf	

transformation merely offsets the log-likelihood attained by LDA.

Table 2 provides various statistics for the 16 data sets. It also indicates, in the last column, which 6 of the 16 data sets consist purely of discrete feature values. The fourth-to-last to second-to-last column in the table gives the different sizes of labeled (N), unlabeled (M), and test sets we used in every run of our experiments. We do not expect much gain from employing unlabeled data if the number of labeled points is large. We therefore kept the labeled set small, choosing a size of twice the dimensionality plus once the number of classes: $2d + K$. We also took care that every class has at least one labeled instance in the training set. The remainder of the data was then randomly divided in two, more or less, equally sized sets that make up the unlabeled and test sets, respectively.

5.2 Performance Criteria and Results

With the labeled, unlabeled, and test sets as described above, we determined $\hat{\theta}_{\text{sup}}$, $\hat{\theta}_{\text{semi}}$, and $\hat{\theta}_{\text{opt}}$. In addition, we calculated $\hat{\theta}_{\text{hoc}}$, which are the parameters of the constrained LDA estimated by means of the more ad hoc procedure in [30]. For $\hat{\theta}_{\text{opt}}$, we of course had to use the true labels belonging to the unlabeled data. The parameters in $\hat{\theta}_{\text{hoc}}$ can be estimated in closed form. For details, we refer to the original work in [30].

For every data set the experiments were repeated 1000 times. Using the estimates $\hat{\theta}_{\text{sup}}$, $\hat{\theta}_{\text{semi}}$, and $\hat{\theta}_{\text{opt}}$, we calculated the following twelve criteria based on the log-likelihood

for Table 3: the three average log-likelihoods (denoted L_{sup}, L_{semi}, and L_{opt}) on the independent test data; the same three average log-likelihoods on the labeled plus unlabeled data, i.e., the training data X_{V^*}; the percentage of times that the log-likelihood of the semi-supervised learner is strictly larger than the log-likelihood of the supervised learner ($\overset{\text{semi}}{\text{sup}}$, read: semi-supervised over supervised); the percentage that the log-likelihood of the optimal classifier is strictly larger than the semi-supervised one (this number, denoted $\overset{\text{opt}}{\text{semi}}$, as well as the previously defined $\overset{\text{semi}}{\text{sup}}$ are calculated both on the test and the training set); and finally we expressed the relative improvement of the semi-supervised approach over the supervised approach in comparison with the optimal estimates by $\frac{L_{\text{semi}}-L_{\text{sup}}}{L_{\text{opt}}-L_{\text{sup}}}$. Again this is done both on the test and the training set. The same quantities are also calculated for the corresponding error rates ε_{sup}, $\varepsilon_{\text{semi}}$, and ε_{opt} (see Table 4), with the only difference that we check numbers to be strictly smaller, instead of larger, to determine $\overset{\text{semi}}{\text{sup}}$ and $\overset{\text{opt}}{\text{semi}}$. Finally, Table 5 contains averaged log-likelihoods L_{hoc} and error rates ε_{hoc}, both on training and test sets, for the more ad hoc semi-supervised approach. Similar to those in Tables 3 and 4, in the last four columns, comparisons to the corresponding log-likelihoods and classification errors of the supervised and our novel semi-supervised approach were made.

A permutation test on all different paired results [84], both for the four log-likelihoods L_{sup}, L_{semi}, L_{opt}, and L_{hoc} and the four errors ε_{sup}, $\varepsilon_{\text{semi}}$, ε_{opt}, and ε_{hoc}, showed that for almost all cases we cannot retain the hypothesis that their averages are the same (at $p \ll 0.001$). There are a few exceptions though. For the test error rates ε_{sup} and $\varepsilon_{\text{semi}}$ on `spectf`, we cannot reject the null hypothesis of equality of expectation (at $p = 0.68$). On `optical` and `qsar` there is no statistically significant difference between L_{semi} and L_{opt} for the test log-likelihoods (at $p = 0.01$ and 0.50, respectively). Finally, L_{sup} and L_{hoc} are, both in training and testing, not significantly different on `shuttle` (at $p = 0.25$ and 0.25) and `spambase` (at $p = 0.76$ and 0.99), while ε_{sup} and ε_{hoc} are not significantly different on `skin` (at $p = 0.03$ and 0.03). For easy reference, the related performance numbers are underlined in the respective result tables.

6 Discussion

6.1 Guarantees on the Training Set

The results in Table 3 show that, on the training set, MCPL-based semi-supervised LDA is in between the regular supervised and the optimal estimate. That this happens to be the case in a strict sense, in all experiments we carried out, can be most readily deduced from the values under $\overset{\text{semi}}{\text{sup}}$ and $\overset{\text{opt}}{\text{semi}}$ on the training set. These numbers equal 100.0 in all cases. This, in turn, indicates that in all of the 16,000 experiments we ran, the strict inequality from Theorem 1 was satisfied. Even for the discrete data sets this holds true, which was to be expected, given the number of different discrete vectors these data sets take on. `Spectf` has the smallest number, 267, implying that every feature vector in `spectf` is unique. With 267 distinct values, chances are indeed very small that the means from Equation (16) and (17) coincide.

6.2 Likelihood Behavior on the Test Set

The aforementioned guarantees are on the training set that includes the unlabeled samples in U, but of course we are interested in the performance on independent test data as well. We are unaware of any theoretical results for the log-likelihood that provide a precise connection between performance on the training set and the test set, though we do expect that with more training data the likelihood of the supervised model on the test set becomes better in expectation. We need to consider such improvement in expectation, simply because, for a single instantiation of a classification problem, we might be unlucky in our draw of training or test set. In contrast with the situation in the training phase, we can therefore only get improvements in the average. Comparing the test log-likelihood in Table 3 for the supervised method with the one for the semi-supervised approach, we see the same as on the training data: for every data set, L_{sup} is smaller than L_{semi}. Also if we look at $\overset{\text{semi}}{\underset{\text{sup}}{}}$, we see that there are only two cases out of 16,000 in which the supervised estimate was better: we find a percentage of 99.8 instead of 100.0 on `miniboone`.

The story is different, however, if we compare the semi-supervised and the optimal estimates. First of all, $\overset{\text{opt}}{\underset{\text{semi}}{}}$ indicates that, on the independent test set, the semi-supervised estimate is better than the optimal one in about 5% of the cases. In itself, this does not have to be at odds with what we expect for the likelihood, as it concerns the number of wins or losses and not the average log-likelihood. Our results on `gas`, `optical`, and `qsar`, however, indicate that also when it comes to the expected log-likelihood, $\hat{\theta}_{\text{semi}}$ may outperform $\hat{\theta}_{\text{opt}}$. Only the result on `gas` is statistically significant though. Moreover, the differences are anyway relatively small, as also the second-to-last column in Table 3 illustrates, where we find values basically equal to 1 for these sets.

Regarding the log-likelihood, we generally note the following. Overall, the relative improvements, as provided in the last two columns of Table 3, are considerable, sometimes enormous even. None of them is lower than 0.9 and many are virtually 1. This shows that the semi-supervised log-likelihood is, relative to the supervised value, very close to the optimal estimate. The immense improvements are probably explained by the fact that the averaged class-conditional covariance matrix Σ is much more stably estimated in case of semi-supervision. The supervised estimate relies on $N = 2d + K$ samples, while the semi-supervised estimate, as can be readily seen from Equation (15), is based on all $N+M$ in the training set. In our experiments $N + M$ is considerably larger than N. The latter is only slightly larger than twice the dimensionality, resulting in unstable covariance estimates. Clearly, the extreme difference in behavior for the various estimates will disappear with increasing numbers of labeled data.

6.3 Error Rates

Unlike the log-likelihood, the 0-1 loss is bounded and the differences and relative improvements stated in Table 4 are not that large. In almost all cases, $\varepsilon_{\text{semi}}$ is smaller than ε_{sup} and ε_{opt} is smaller than $\varepsilon_{\text{semi}}$ in turn. On the test set, the maximum relative improvement reported is 0.426 on `optical`, with a good second of 0.415 on `shuttle`.

There are three settings, however, in which no improvements of semi-supervised over supervised learning are attained: the first one is on the training set for `low` and the two others are in the training and test phase for `spectf`. In all cases, L_{semi} is better than L_{sup}. So we have the, possibly, somewhat counterintuitive behavior that the estimates improve in terms of the expected log-likelihood, but that the expected error rate still deteriorates. Similar phenomena for other classifiers have been described in [71, 72], where simple artificial examples are provided of how such behavior can be realized. It is a glimpse of the earlier mentioned difficult interrelationship two different performance criteria can display [70, 73–75], which we alluded to earlier on in Section 5. We checked the learning curves for `low` and `spectf` and they just showed the regular behavior: with increasing labeled sample sizes, the expected error rate of the supervised classifier decreases.

Finally, we remark that the increase in error rate going from the training to the test set is less for the semi-supervised classifier than for the supervised one. This shows that the semi-supervised classifier is less overtrained on the training set than supervised LDA.

6.4 Comparison to Constrained LDA

Looking at Table 5, we see that also the ad hoc approach can work well. Especially when looking at the likelihood and comparing it to the supervised estimates, we see that, both on the training and the test set, the estimated likelihood is often better than the one obtained by the regular supervised parameters. The reason for the constrained approach to often be so much better than the supervised approach is probably similar to the one given in Subsection 6.2 to explain why the new approach comes so close to the optimal log-likelihoods. The large improvements are probably due to the fact that the averaged class-conditional covariance matrix Σ is much more stably estimated in case of semi-supervision. The estimated covariance matrix might still not be very good, but at least it is substantially better than the volatile and not so well conditioned supervised estimate. Nonetheless, the novel approach clearly outperforms the more ad hoc technique in most of the cases where the likelihood is concerned. In fact, compared to the constrained approach, MCPL provides the best average test log-likelihood on all data sets. The only expected log-likelihood that is worse during training is the one for `spectf`.

Looking at the error rate, we see that the ad hoc procedure does very bad on `optical` and `shuttle`. Still, $\hat{\theta}_{\text{hoc}}$ leads to the best error rate on the test set on seven data sets. On the other nine data sets $\hat{\theta}_{\text{semi}}$ turns out to be preferred.

6.5 MCPL for Other Classifiers

MCPL is proposed as a general estimation principle, which delivers semi-supervised estimates that are at least as good as the regular supervised parameter estimates for any log-likelihood based classifier. To come to results such as Theorems 1 and 2, additional knowledge about the class-conditional distributions is needed. Because they are very similar to LDA and the same kind of mean constraints hold, classifiers for which it is almost immediate that strict or expected improvements can be obtained through semi-supervision,

are the NMC (nearest mean classifier), quadratic discriminant analysis (QDA), and all kinds of kernelized or flexibilized versions of NMC, LDA, and QDA [85]. We speculate that also many classifiers constructed on the basis of exponential families [65, 66] allow for theorems making equivalent statements. These include, for instance, the Bernoulli, multinomial, and exponential density.

Another interesting group of classifiers to study in the context of MCPL is that for which every class may consist of a mixture model. As the analysis of mixture models is in itself already rather difficult [86]—for one, the likelihood function is not concave, such classifiers may be outside the reach of any helpful theoretical analysis. We do, however, expect to benefit, if only from the regularizing effect our semi-supervised approach has, similar to the situation mentioned at the end of Subsection 6.2. What does seem a problem still, is to find an appropriate solution to the optimization that needs to be carried out in order to find an MCPL estimate. It seem worthwhile, though, to try to get to the nearest saddle point that can be found by means of a combined gradient ascent (in θ) and descent (in q).

All in all, what we consider the main open research issue is finding and characterizing all those classifiers for which statements similar to Theorem 2 and possibly Theorem 1 hold.

7 Conclusion

We presented a well-founded approach to likelihood-based semi-supervised learning. Our principle of maximum contrastive pessimistic likelihood (MCPL) estimation is generally applicable to supervised classifiers whose parameters are estimated by means of a maximization of the likelihood. Moreover, under certain concavity assumptions, improvements of the semi-supervised estimates can be expected and, in particular cases, even be guaranteed. A worked-out illustration based on classical LDA demonstrates the significant improvements that can be obtained by our novel approach.

Acknowledgments

First of all, Marleen de Bruijne (Erasmus Medical Center and University of Copenhagen) is wholeheartedly acknowledged for truly scrutinizing an initial version of this article beginning to end, weeding out illegible prose, ambiguities, and plain errors (both with respect to content and form). Jesse H. Krijthe (Leiden University Medical Center and Delft University of Technology) and David M. J. Tax (Delft University of Technology) are kindly thanked for their critical assessment of this work and for their proofreading of parts of the text. I also thank all three for the various helpful discussions. Joris Mooij (University of Amsterdam) is acknowledged for inviting me to give a talk that, eventually, triggered insights into a simplification and generalization of the theory. Are C. Jensen (Halfwave AS and University of Oslo) is warmly thanked for all the semi-supervised inspiration he

provided me with. Finally, thanks go out to Mads Nielsen (University of Copenhagen) who gave me some great opportunities throughout the past decade.

References

[1] Ronald A. Fisher. An absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

[2] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[3] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge Univ Press, 1925.

[4] Stephen M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620, 2007.

[5] Markus Ackermann, M. Ajello, A. Allafort, L. Baldini, J. Ballet, G. Barbiellini, et al. Detection of the characteristic pion-decay signature in supernova remnants. *Science*, 339(6121):807–811, 2013.

[6] Jenny Allen, Mason Weinrich, Will Hoppitt, and Luke Rendell. Network-based diffusion analysis reveals cultural transmission of lobtail feeding in humpback whales. *Science*, 340(6131):485–488, 2013.

[7] Hu Cang, Anna Labno, Changgui Lu, Xiaobo Yin, Ming Liu, Christopher Gladden, Yongmin Liu, and Xiang Zhang. Probing the electromagnetic field of a 15-nanometre hotspot by single molecule imaging. *Nature*, 469(7330):385–388, 2011.

[8] Hoi Sung Chung and William A Eaton. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature*, 2013.

[9] Bingni W. Brunton, Matthew M. Botvinick, and Carlos D. Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–98, 2013.

[10] Angélique D'Hont, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217, 2012.

[11] Yuannian Jiao, Norman J Wickett, Saravanaraj Ayyampalayam, André S Chanderbali, Lena Landherr, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100, 2011.

[12] Lauri Nummenmaa, Enrico Glerean, Riitta Hari, and Jari K Hietanen. Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111(2):646–651, 2014.

[13] Dana C. Price, Cheong Xin Chan, Hwan Su Yoon, Eun Chan Yang, Huan Qiu, et al. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science*, 335(6070):843–847, 2012.

[14] E. Saglamyurek, N. Sinclair, J. Jin, J. A. Slater, D. Oblak, F. Bussières, M. George, R. Ricken, W. Sohler, and W. Tittel. Broadband waveguide quantum memory for entangled photons. *Nature*, 469(7331):512, 2011.

[15] Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28(10):2731–2739, 2011.

[16] J. Wang. An improvement on the maximum likelihood reconstruction of pedigrees from marker data. *Heredity*, 2013.

[17] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–314, 2012.

[18] Jacob Bien and Robert J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[19] Madeleine Cule, Richard Samworth, and Michael Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607, 2010.

[20] Yeojin Chung, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, pages 1–25, 2013.

[21] Ted A. Laurence and Brett A. Chromy. Efficient maximum likelihood estimator fitting of histograms. *Nature Methods*, 7(5):338–339, 2010.

[22] Jason D. Lee and Trevor J. Hastie. Learning mixed graphical models. *arXiv preprint arXiv:1205. 5012*, 2012.

[23] N. Simon and R. J. Tibshirani. Discriminant analysis with adaptively pooled covariance. *arXiv preprint arXiv:1111. 1687*, 2011.

[24] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[25] X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers, 2009.

[26] V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

[27] S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of COLT 2008*, pages 33–44, 2008.

[28] J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, volume 20, pages 801–808, 2007.

[29] A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[30] Marco Loog. Semi-supervised linear discriminant analysis through moment-constraint parameter estimation. *Pattern Recognition Letters*, 37(1):24–31, 2014.

[31] X. Zhu. Semi-supervised learning literature survey. Computer Sciences TR 1530, University of Wisconsin, 2008.

[32] H. O. Hartley and J. N. K. Rao. Classification and estimation in analysis of variance problems. *Review of the International Statistical Institute*, 36(2):141–147, 1968.

[33] G. J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

[34] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 19–26, 2002.

[35] J. N. Vittaut, M. R. Amini, and P. Gallinari. Learning classification with both labeled and unlabeled data. In *Machine Learning: ECML 2002*, pages 69–78, 2002.

[36] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, 1995.

[37] G. Nagy and G.L. Shelton. Self-corrective character recognition system. *IEEE Transactions on Information Theory*, 12(2):215–222, 1966.

[38] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 792–799, 1998.

[39] T. J. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, pages 821–826, 1978.

[40] D. M. Titterington. Updating a diagnostic system using unconfirmed cases. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):238–247, 1976.

[41] N. P. Dick and D. C. Bowden. Maximum-likelihood estimation for mixtures of two normal distributions. *Biometrics*, 29:781–791, 1973.

[42] D. W. Hosmer Jr. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, pages 761–770, 1973.

[43] W. Y. Tan and W. C. Chang. Convolution approach to genetic analysis of quantitative characters of self-fertilized population. *Biometrics*, 28:1073–1090, 1972.

[44] S. Abney. Understanding the Yarowsky algorithm. *Computational Linguistics*, 30(3):365–395, 2004.

[45] G. Haffari and A. Sarkar. Analysis of semi-supervised learning with the Yarowsky algorithm. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[46] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1553–1567, 2004.

[47] F. Cozman and I. Cohen. Risks of semi-supervised learning. In *Semi-Supervised Learning*, chapter 4. MIT Press, 2006.

[48] Ting Yang and Carey E Priebe. The effect of model misspecification on semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2093–2103, 2011.

[49] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[50] Masanori Kawakita and Junichi Takeuchi. Safe semi-supervised learning based on weighted likelihood. *Neural Networks*, 2014.

[51] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[52] Nataliya Sokolovska, Olivier Cappé, and François Yvon. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th International Conference on Machine learning*, pages 984–991. ACM, 2008.

[53] M. Loog. Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010)*, volume 6322 of *LNAI*, pages 291–304. Springer, 2010.

[54] M. Loog. Semi-supervised linear discriminant analysis using moment constraints. In *Partially Supervised Learning (PSL 2011)*, volume 7081 of *LNAI*, pages 32–41. Springer, 2012.

[55] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[56] M. Loog and A. C. Jensen. Constrained log-likelihood-based semi-supervised linear discriminant analysis. In *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626 of *LNCS*, pages 327–335. Springer, 2012.

[57] M. Loog and A. C. Jensen. Semi-supervised nearest mean classification through a constrained log-likelihood. *IEEE Transactions on Neural networks and Learning Systems*, *accepted*, 2014.

[58] J. H. Krijthe and M. Loog. Implicitly constrained semi-supervised least squares classification. submitted November 2013, *available through* `http://www.jessekrijthe.com/papers/krijthe2013.pdf`, 2013.

[59] J. H. Krijthe and M. Loog. Implicitly constrained semi-supervised linear discriminant analysis. In *Proceedings of the 22nd International Conference on Pattern Recognition*, volume 22, pages —, Stockholm, Sweden, *accepted*, 2014.

[60] Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Prague, Czech Republic, 2007.

[61] G.S. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *The Journal of Machine Learning Research*, 11:955–984, 2010.

[62] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

[63] G. J. McLachlan. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American Statistical Association*, 72(358):403–406, 1977.

[64] G. J. McLachlan and S. Ganesalingam. Updating a discriminant function on the basis of unclassified data. *Communications in Statistics - Simulation and Computation*, 11(6):753–767, 1982.

[65] Ole E. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, 1978.

[66] P. J. Bickel and K. A. Doksum. *Mathematical Statistics*, volume 1. Prentice-Hall, Inc., second edition, 2001.

[67] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

[68] M. Dresher. *Games of Strategy*. Prentice-Hall Inc., 1961.

[69] Nelson Maculan and Geraldo Galdino de Paula Jr. A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$. *Operations Research Letters*, 8(4):219–222, 1989.

[70] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[71] Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings of the 29th Annual International Conference on Machine Learning*, 2012.

[72] M. Loog and R. P. W. Duin. The dipping phenomenon. In *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626 of *LNCS*, pages 310–317. Springer, 2012.

[73] Mark D. Reid and Robert C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

[74] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, 12:731–817, 2011.

[75] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

[76] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 6th International Conference on Machine Learning*, pages 200–209, 1999.

[77] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 17:529–536, 2004.

[78] K. Bache and M. Lichman. Uci machine learning repository, 2013.

[79] D. D. Lucas, R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic, and Y. Zhang. Failure analysis of parameter-induced simulation crashes in climate models. *Geoscientific Model Development Discussions*, 6(1):585–623, 2013.

[80] J. P. Bridge, S. B. Holden, and L. C. Paulson. Machine learning for first-order theorem proving: learning to select a good heuristic. *submitted*, 2013.

[81] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A Ryan, Margie L Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.

[82] Kamel Mansouri, Tine Ringsted, Davide Ballabio, Roberto Todeschini, and Viviana Consonni. Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling*, 53(4):867–878, 2013.

[83] R. Bhatt and A. Dhall. Skin segmentation dataset.

[84] P. I. Good. *Permutation Tests*. Springer, 2000.

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

[86] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer-Verlag, second edition, 1998.

Table 2: Basic properties of the data sets: the number of objects, the dimensionality of the original feature vectors, the dimensionality after PCA (d), the number of classes K, the sizes of the largest and the smallest class, the number of labeled (N), unlabeled (M), and test objects in every run of our experiments, and whether the features are purely discrete.

data set (abbr.)	#objects	dim.	PCA/d	K	largest	(%)	smallest	(%)	N	M	#test	discr.
banknote	1372	4	4	2	762	(55.5)	610	(44.5)	10	681	681	no
climate	540	18	18	2	494	(91.5)	46	(8.5)	38	251	251	no
first-order	6118	51	41	6	2554	(41.7)	486	(7.9)	88	3015	3015	no
gas	13910	128	60	6	3009	(21.6)	1641	(11.8)	126	6892	6892	no
landsat	6435	36	33	6	1533	(23.8)	626	(9.7)	72	3182	3181	yes
letter	20000	16	16	26	813	(4.1)	734	(3.7)	58	9971	9971	yes
low	531	93	70	10	90	(16.9)	4	(0.8)	150	191	190	no
magic	19020	10	10	2	12332	(64.8)	6688	(35.2)	22	9499	9499	no
miniboone	130064	50	11	2	93565	(71.9)	36499	(28.1)	24	65020	65020	no
optical	5620	64	61	10	572	(10.2)	554	(9.9)	132	2744	2744	yes
pen-based	10992	16	16	10	1144	(10.4)	1055	(9.6)	42	5475	5475	yes
qsar	1055	41	38	2	699	(66.3)	356	(33.7)	78	489	488	no
shuttle	58000	9	6	7	45586	(78.6)	10	(0.0)	19	28991	28990	yes
skin	245057	3	3	2	194198	(79.2)	50859	(20.8)	8	122525	122524	no
spambase	4601	57	56	2	2788	(60.6)	1813	(39.4)	114	2244	2243	no
spectf	267	44	43	2	212	(79.4)	55	(20.6)	88	90	89	yes

Table 3: Results calculated based on the log-likelihoods from the 1000 experiments per data set for the supervised and our semi-supervised approach. Refer to Subsection 5.2 for a description of the various criteria determined.

data set	estimated on test			estimated on full train			% test wins		% trn. wins		$\frac{L_{\text{semi}}-L_{\text{sup}}}{L_{\text{opt}}-L_{\text{sup}}}$	
(abbr.)	L_{sup}	L_{semi}	L_{opt}	L_{sup}	L_{semi}	L_{opt}	semi sup	opt semi	semi sup	opt semi	test	trn.
banknote	-11.7	-4.72	-4.51	-11.5	-4.69	-4.48	100.0	98.4	100.0	100.0	0.971	0.970
climate	-34.1	-26.5	-26.2	-32.6	-25.8	-25.5	100.0	100.0	100.0	100.0	0.964	0.961
first-order	-1.88e+03	-62.6	-60.3	-1.78e+03	-40.4	-39.2	100.0	100.0	100.0	100.0	0.999	0.999
gas	-4.46e+04	-4.4e+03	-4.41e+03	-4.37e+04	-13.1	-12.4	100.0	44.8	100.0	100.0	1.000	1.000
landsat	-33.2	-4.64	-3.73	-32.4	-4.35	-3.42	100.0	100.0	100.0	100.0	0.969	0.968
letter	-63.6	-22.3	-18.4	-63.3	-22.2	-18.3	100.0	100.0	100.0	100.0	0.914	0.913
low	-90.1	-19.8	-17.6	-37.8	11.7	13.9	100.0	99.9	100.0	100.0	0.969	0.957
magic	-30.6	-11.7	-11.1	-30.6	-11.6	-11.1	100.0	100.0	100.0	100.0	0.974	0.974
miniboone	-2.2e+09	-7.17e+07	-6.93e+07	-2.42e+09	-9.75	-9.48	99.8	93.1	100.0	100.0	0.999	1.000
optical	-6.24e+15	-5.66e+12	-6.35e+12	-6.06e+15	-61.1	-60.1	100.0	83.8	100.0	100.0	1.000	1.000
pen-based	-45.2	-15.9	-13.5	-44.9	-15.8	-13.5	100.0	100.0	100.0	100.0	0.927	0.926
qsar	-4.02e+14	-1.02e+03	-1.03e+03	-3.36e+14	-37.2	-36.9	100.0	99.7	100.0	100.0	1.000	1.000
shuttle	-5.42e+07	-9.81	-9.24	-6.8e+07	-9.37	-8.76	100.0	96.9	100.0	100.0	1.000	1.000
skin	-125	-3.84	-3.45	-125	-3.84	-3.45	100.0	100.0	100.0	100.0	0.997	0.997
spambase	-1.09e+16	-81.6	-81.3	-9.76e+15	-73.7	-73.4	100.0	100.0	100.0	100.0	1.000	1.000
spectf	-78.6	-53.6	-53.1	-54.5	-36.8	-36.5	100.0	97.5	100.0	100.0	0.982	0.985

Table 4: Results based on the error rates obtained from the 1000 experiments per data set for the supervised and our semi-supervised approach. Subsection 5.2 gives a description of the various criteria.

data set (abbr.)	estimated on test			estimated on full trn.			% test wins		% trn. wins		$\frac{\varepsilon_{\text{semi}}-\varepsilon_{\text{sup}}}{\varepsilon_{\text{opt}}-\varepsilon_{\text{sup}}}$	
	ε_{sup}	$\varepsilon_{\text{semi}}$	ε_{opt}	ε_{sup}	$\varepsilon_{\text{semi}}$	ε_{opt}	semi sup	opt semi	semi sup	opt semi	test	trn.
banknote	0.061	0.052	0.025	0.061	0.052	0.024	69.7	89.7	70.5	89.3	0.254	0.240
climate	0.150	0.143	0.053	0.133	0.129	0.034	63.9	99.8	56.0	100.0	0.071	0.033
first-order	0.666	0.658	0.529	0.652	0.650	0.514	75.9	100.0	55.3	100.0	0.055	0.015
gas	0.141	0.134	0.085	0.139	0.133	0.082	68.5	99.9	65.7	99.8	0.134	0.105
landsat	0.291	0.251	0.161	0.285	0.247	0.153	100.0	100.0	99.9	100.0	0.312	0.286
letter	0.618	0.599	0.299	0.615	0.595	0.294	97.5	100.0	97.1	100.0	0.061	0.060
low	0.763	0.747	0.696	0.475	0.501	0.334	70.0	91.5	2.2	100.0	0.233	-0.181
magic	0.317	0.303	0.216	0.316	0.303	0.216	90.3	100.0	89.4	99.8	0.136	0.134
miniboone	0.246	0.229	0.159	0.246	0.229	0.159	83.6	99.9	83.7	99.9	0.198	0.197
optical	0.161	0.113	0.049	0.154	0.111	0.042	100.0	100.0	100.0	100.0	0.426	0.385
pen-based	0.280	0.243	0.124	0.278	0.241	0.122	99.6	100.0	100.0	100.0	0.238	0.234
qsar	0.257	0.247	0.154	0.229	0.226	0.132	65.7	100.0	53.1	100.0	0.089	0.031
shuttle	0.134	0.103	0.059	0.134	0.103	0.059	82.1	83.7	81.7	83.7	0.415	0.413
skin	0.098	0.087	0.068	0.098	0.087	0.068	79.8	55.9	79.8	56.0	0.365	0.365
spambase	0.195	0.185	0.112	0.189	0.182	0.108	76.2	99.8	70.7	100.0	0.117	0.086
spectf	0.325	0.325	0.260	0.203	0.210	0.131	41.7	85.7	21.6	100.0	-0.006	-0.108

Table 5: Log-likelihood and error rate results obtained from the 1000 experiments per data set for the ad hoc semi-supervised approach and its comparison to our novel semi-supervised and regular supervised approach. Refer to Subsection 5.2 for an explanation of the various criteria.

data set (abbr.)	test L_{hoc}	trn. L_{hoc}	test ε_{hoc}	trn. ε_{hoc}	win test lik. hoc sup	win test lik. semi hoc	win trn. lik. hoc sup	win trn. lik. semi hoc	win test err. hoc sup	win test err. semi hoc	win trn. err. hoc sup	win trn. err. semi hoc
banknote	-9.38	-9.29	0.087	0.086	73.8	96.5	74.0	96.6	30.1	76.2	30.6	75.2
climate	-27	-26.2	0.117	0.102	100.0	93.7	100.0	93.3	79.9	22.4	81.1	17.5
first-order	-68	-43.7	0.626	0.616	100.0	100.0	100.0	100.0	96.8	7.6	95.0	5.8
gas	-5.66e+03	-21.1	0.145	0.143	100.0	99.9	100.0	100.0	44.7	68.3	42.9	67.9
landsat	-16.8	-16.2	0.308	0.302	99.4	100.0	99.5	100.0	29.8	98.6	27.9	98.0
letter	-53.1	-52.9	0.625	0.622	99.8	100.0	99.7	100.0	33.2	92.4	32.2	92.9
low	-27.9	9.42	0.744	0.485	100.0	100.0	100.0	100.0	74.9	39.3	26.1	16.4
magic	-12.4	-12.4	0.292	0.292	100.0	80.7	100.0	80.7	74.0	37.8	74.3	38.9
miniboone	-7.65e+07	-10.8	0.218	0.218	99.7	96.1	100.0	98.3	73.1	41.3	72.6	40.7
optical	-7.74e+15	-7.48e+15	0.900	0.900	29.5	99.0	32.7	100.0	0.0	100.0	0.0	100.0
pen-based	-35.4	-35	0.299	0.297	98.9	100.0	99.1	100.0	24.5	98.7	24.8	98.5
qsar	-1.51e+13	-1.1e+13	0.229	0.209	100.0	93.2	100.0	96.6	86.9	16.1	83.8	14.9
shuttle	-5.51e+05	-5.82e+05	0.822	0.822	1.6	100.0	1.6	100.0	1.6	99.1	1.6	99.1
skin	-40.4	-40.4	0.102	0.102	94.7	95.2	94.7	95.4	40.1	71.2	40.6	71.1
spambase	-1.66e+16	-8.65e+15	0.310	0.307	85.1	100.0	85.1	100.0	51.3	51.0	51.8	48.4
spectf	-53.8	-36.8	0.293	0.182	100.0	74.2	100.0	42.3	71.0	17.8	78.4	8.0