Safe Semi-Supervised Learning

Andrea Bertazzi

Faculty EEMCS - TU Delft

TU Delft Delft University of Technology

Challenge the future

Safe Semi-Supervised Learning

by

Andrea Bertazzi

in partial fulfillment of the requirements for the degree of

Master of Science
in Applied Mathematics

at the Delft University of Technology,
to be defended publicly on Monday December 10, 2018 at 10:00 AM.

Student number:	4632729	
Project duration:	March 1, 2018 – December 3, 2018	
Thesis committee:	Prof. dr. ir. G. Jongbloed,	TU Delft, Chair of the DIAM
	Prof. dr. M. Loog,	TU Delft, main supervisor
	Dr. ir. J. Bierkens,	TU Delft, supervisor

TU Delft Delft University of Technology

Abstract

Semi-supervised algorithms have been shown to possibly have a worse performance than the corresponding supervised model. This may be due to a violation of the assumptions on the data that are introduced in most classification systems. We study an approach that was previously shown to have guarantees of improvement for the LDA classifier in terms of log-likelihood on the full data-set of labeled and unlabeled observations. This method is based on two key concepts: contrast and pessimism. We extend this approach to a broader class of probabilistic generative models, in which the class conditional distributions can be modeled with any parametric class that belongs to the exponential family. In this case, we prove that the classifier is never worse and, under mild assumptions, strictly improves the log-likelihood on the complete data-set. The case of Gaussian densities is analyzed in detail, both for LDA and QDA. Moreover, we study this method in the case of least squares classification. In terms of square loss we prove the contrastive pessimistic classifier is guaranteed not to degrade the performance and, with a further requirement on the data, it strictly outperforms the supervised model.
Finally, we apply contrast and pessimism to the task of parameter estimation of a multivariate Gaussian density in a missing data framework. We fully characterize the case of a missing block of data and we show that a strictly increased likelihood on the complete data-set is obtained for any monotone sample. In other terms, the contrastive pessimistic estimates are guaranteed to fit better the complete data-set composed of both observed and hidden components.

Keywords: semi-supervised learning, contrast, pessimism, parameter estimation, monotone sample, maximum likelihood.

Preface

This thesis is the result of a nine months project at the Patter Recognition Laboratory of TU Delft under the supervision of prof. dr. Marco Loog and dr. ir. Joris Bierkens. The project started with a three months literature review that touched several topics, among which a theoretical investigation of the use of margin based losses in machine learning and the study of several algorithms for semi-supervised learning. Afterwards, robust algorithms for semi-supervised learning became the main focus and a large part of this thesis is dedicated to this particular problem. In this work we investigate the possibility of building semi-supervised classifiers that are guaranteed to improve, or at least not to be worse, than their supervised counterpart. Two principles are recurrent in this thesis: *contrast* and *pessimism*. These are fundamental also in the second part of this work, in which we focus on parameter estimation in a statistical inference framework with an incomplete data-set.

The results of this thesis could not have been possible without Marco Loog, whose guidance was crucial. I would like to thank him for everything I have learned in these months and for the several constructive discussions we had. I had a lot of fun. Many thanks also go to Joris for the help in some intricate mathematical problems I encountered.

I would like to thank my mum for the unconditional support she has given me from the first day of my bachelor's degree. I also thank my father for always making sure that everything was alright. A final mention goes to the people that have made the last two years special.

Andrea Bertazzi
Delft, December 2018

Contents

1

Outline of the Thesis

The field of Machine Learning has grown rapidly in the last decades and is currently one of the most active research areas. The range of applications of machine learning techniques is very wide and spans from speech recognition and computer vision, to medical imaging and autonomous systems. In this thesis we focus on a particular branch of this broad discipline that is known as semi-supervised learning (SSL). SSL concerns the study of algorithms that solve the problem of classification when the available data-set is composed of labeled and unlabeled observations. This framework is very frequent in real world applications for which the classification of the observed samples must be performed manually by humans and is thus considerably expensive. Examples of typical SSL applications include speech recognition and webpage classification.
In Chapter 2 we define the fundamental notions and terminology required to analyze this problem. Section 2.1 is dedicated to supervised learning (SL), which is the study of classification tasks in case we have fully labeled data to train a learning system. SL represents the foundation upon which we obtain SSL by the addition of an unlabeled data-set. An overview of SSL is formally discussed in Section 2.2. It will become clear that a crucial problem of present SSL algorithms is the possible performance degradation that may be observed. This highly undesirable behaviour is the main subject of this work and motivates the following research questions:

> *Is it possible to build a semi-supervised classifier that is guaranteed not to be worse than the corresponding supervised model in a finite sample setting? In which cases is this possible? And when can we expect the semi-supervised classifier to strictly outperform its supervised counterpart?*

In Section 2.3 we first focus on an example in which a performance degradation takes place and then on two algorithms that propose a solution. From our point of view, these classifiers represent interesting first steps towards a solution but do not address some of the key issues. In Section 2.3.4 we discuss a *contrastive pessimistic* approach to SSL, which was first introduced in [33]. This method was shown to have the properties we look for in the specific setting of linear discriminant analysis. In this thesis we extend this approach to other general frameworks and we formally and rigorously investigate the guarantees on its performance compared to the corresponding supervised model. Chapter 3 is dedicated to least squares classification. In this case, the contrastive pessimistic approach is shown to have a non degradation guarantee and, under an additional assumption, a property of strict performance enhancement in terms of square loss on the full data-set for any labeling of the unlabeled observations. In Chapters 4 and 5 we study the maximum contrastive pessimistic likelihood (MCPL) approach, which is a semi-supervised generative probabilistic model that considerably extends and completes the analysis conducted in [33]. In particular, we consider classifiers that model the class conditional densities with any parametric class that is in

the exponential family. With Theorem 4.5 we show that the MCPL approach is never worse than its supervised counterpart in terms of likelihood on the full data-set. This once again holds for any labeling of the unlabeled observations. Then, the entire Section 4.2.5 is dedicated to the investigation of the possibilities of strict performance improvement. In Corollary 4.11 we show that under mild assumptions this happens almost surely with respect to the underlying distribution generating the unlabeled data. Considerable attention is then devoted to a thorough study of the main features of the proposed algorithm. For instance we point to Sections 4.2.7 and 4.3. The setting of Gaussian densities is then specifically studied in Chapter 5. The results of our analysis provide a clear and formal vision of when safe semi-supervised learning is possible and how we can build such a safe classifier.
The results obtained in SSL by the contrastive pessimistic approach motivate an additional research question:

> *Assume we are interested in estimating a multivariate probability density and we decide to use a parametric model. Suppose we have a data-set which contains both complete and incomplete observations. Is it possible to take advantage of the incomplete observations to find parameters that fit the data better than an estimate computed by using the subset of complete observations only?*

This question is addressed in Chapter 6 for the case of Gaussian densities. A brief answer to the question above is that this is the case at least for a data-set with a specific structure. In Theorem 6.3 we formally state and prove that this holds for a particular block-structure of the missingness, while in Corollary 6.5 we extend this result to more general data-sets.
In Chapter 7 we conclude the thesis with a discussion of the results we obtained and with some suggestions for future research.

2

Preliminaries on Supervised and Semi-Supervised Learning

Pattern Recognition is the field that studies the task of automatically classifying *objects* into categories, or *classes*. Objects can be, depending on the application, images or vectors of measurements that we are interested in assigning to a class. This problem is faced by learning from a set of available examples and thus defining a *decision rule* that can be applied on unseen measurements. The goal is then to have an accurate prediction for the unobserved output value y given an observed input vector x. For instance, suppose we are interested in learning a decision function that predicts whether a papaya is tasty or not [43]. We can select a number of criteria that in our experience are indicators of the tastiness, such as the color and the papaya's softness. A set of papayas is then measured and the color and softness of each fruit are grouped in a corresponding *feature vector* $x_i = (x_{i1}, x_{i2})$. Afterwards, we can taste each papaya to establish whether it is tasty, in which case we assign class label $y_i = +1$, or whether it is not, which is encoded as $y_i = -1$. The data-set we have obtained is denoted by $\mathscr{D}_l = \{x_i, y_i\}_{i=1}^N$ and is used for the training phase of a decision function. The trained *classifier* is then ready to be applied on new observations, for which the class label is hidden. This can result in an error, which is referred to as *misclassification*, or in a correct prediction. The ultimate goal of a classifier is a low percentage of misclassified test examples and more generally a small probability of error:

$$L(f) = \mathbb{P}_{(x,y)\sim p_{xy}}\Big(\{x\colon f(x) \neq y\}\Big),$$

where $f(x)$ is the prediction of the decision rule $f(\cdot)$ for the observation x, y is the true corresponding label and p_{xy} the underlying distribution that generates the sample (x, y). An instance of a data-set with a corresponding decision function is shown in Figure 2.1, in which it can be observed how some points of the training set are misclassified.

The setting we described is called *supervised learning* and is characterized by the availability of a *labeled data-set*, which means that every training measurement comes with its corresponding class label. In some cases, the class tags are unknown and the classifier has to be trained with *unlabeled observations* only. This setting is known as *unsupervised learning*. In the previous example an unsupervised classification would consist in separating the papayas without knowing their true membership in two classes, one of which is the group of tasty papayas and the other is that of non-tasty papayas. Furthermore, there is a third framework in between supervised and unsupervised learning. This setting is called *semi-supervised learning* (SSL). SSL consists in learning a classifier under partial supervision, which means that some of the observed papayas are labeled while others are not. The task of learning a classifier under partial supervision is the focus of this thesis and in Section 2.2 we present an overview of the main algorithms for SSL. On the other hand, the specific

problem we treat in this work is stated and motivated in Section 2.3. However, let us first describe the fundamental aspects of Pattern Recognition and, in particular, of supervised learning.

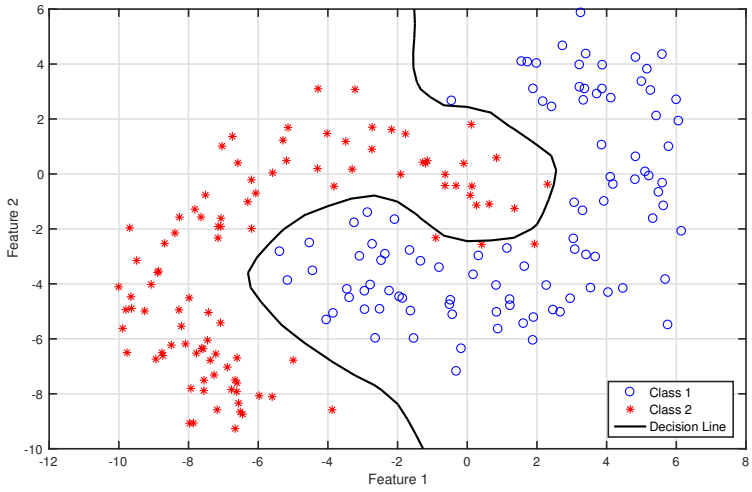


Figure 2.1: *Example of a classification task, showing a training set in $\mathbb{R}^2$ and a possible decision line. The plot is obtained with a Matlab toolbox called PRTools [16].*

2.1. Fundamentals of Pattern Recognition

The design of a classification system is conceptually divided in distinct phases, as illustrated in Figure 2.2. Given a data-set of measurements, the first two stages are called *feature generation* and *feature extraction*. The former concerns the use of algorithms such as the Principal Component Analysis (PCA) or the Singular Values Decomposition (SVD), that reduce the number of features while retaining a certain percentage of the information. On the other hand, feature selection consists in choosing a subset of the features that were present in the original measurements. In this phase, the discrimination effectiveness of the features is measured and the least representative are discarded. After these two stages, the data-set is defined and the focus is on the design of the classifier. For this task, there are several techniques that are based on diverse ideas and approaches to

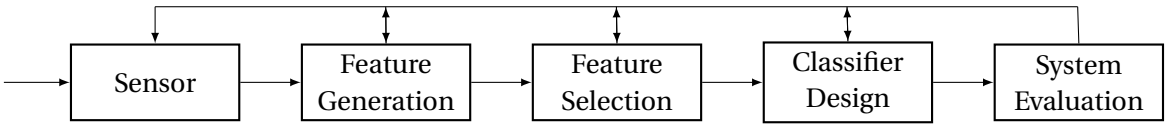


Figure 2.2: *Diagram of the basic stages in the design of a classifier [46].*

the problem. It has to be noted that the performance of a classifier depends on the data-set and it is in general necessary to select the classifier and/or tuning its parameters for the specific task. After having fixed the data-set and the classifier, the last stage consists in an evaluation of the system. This can be done for instance by applying a technique called cross validation (CV), in which the data-set is divided in K parts, leaving one out in the training phase and later using it to evaluate the goodness of the model. This process is then repeated K times changing the set that is used for testing, hence obtaining an estimate of the classification error that the classification system has on unseen data. This method is called K-folded CV, but several different variations of it exist. As explicitly illustrated in Figure 2.2, at the end of the evaluation stage it is possible, and recommended, to go back to one of the precedent stages and repeat the process. For instance, different classifiers or

different representations of the data-set can be tested and the classification system that seems the most accurate can finally be selected.

The design of a classifier is the focus of this thesis and from now on the data-set is assumed to be fixed. In the following subsections we discuss some broad groups of classifiers in a supervised framework, which is then used as a starting point for SSL algorithms.

2.1.1. Generative Probabilistic Models

Generative models tackle the classification problem by estimating the joint probability over the feature vectors and the class labels, which we denote $p(x, y)$. Here we assume $x \in \mathbb{R}^D$ and $y \in \{-1, +1\}$, which means there are two classes. Now, the discrete distribution on the class labels are denoted by a capital letter $P(\cdot)$, while we use the notation $p(\cdot)$ for continuous densities. The Bayes formula can then be used to obtain the posterior probability that x is a sample from class y as

$$P(y|x) = \frac{p(x|y)\,P(y)}{p(x)}.$$

Here $p(x|y)$ is called *class conditional distribution* and $P(y)$ is referred to as *prior probability*.
An observed object x is then assigned class label $+1$ if

$$P(y=+1|x) > P(y=-1|x)$$

and label -1 otherwise. Note that this condition is equivalent to

$$p(x|+1)\,P(y=+1) > p(x|-1)\,P(y=-1).$$

Now, we can define the decision error in this setting as

$$P_e = \int_{R_{+1}} P(y=-1)\,p(x|y=-1)\,dx + \int_{R_{-1}} P(y=+1)\,p(x|y=+1)\,dx,$$

where

$$R_{+1} := \left\{x:\ p(x|+1)\,P(y=+1) > p(x|-1)\,P(y=-1)\right\}$$

and $R_{-1} = R_{+1}^C$, that is the complementary set. This scheme is known as the *Bayes classification rule* if we have full knowledge of the involved distributions. Notably, the Bayes classification rule achieves the lowest possible error, which is called *Bayes error*. We have then a lower bound for the accuracy that a classifier can reach. However, this bound is useful only on a theoretical level, since the true densities are unknown.

The prior probabilities are usually estimated by the fraction of observations from each class, thus the class conditionals are the quantity of interest in generative probabilistic models. Estimates of $p(x|y)$ can then be obtained through any statistical method for density estimation. For instance, parametric methods like the maximum likelihood (ML) and the maximum a posteriori (MAP), or nonparametric methods like kernel density estimators, can be used. A very famous example of maximum likelihood-based generative classifier is known as *linear discriminant analysis* (LDA), in which the conditional distributions are assumed to be Gaussian with equal covariance.

Generative models are clearly suitable for situations in which we know that a particular model can accurately describe the data.

2.1.2. Discriminative Probabilistic Models

Discriminative models estimate directly the posterior probability $p(y|x)$, which is the object of interest in classification. This agrees with Vapnik's principle, which suggests that one should always directly solve the problem of interest without trying to solve a more difficult one [49].

Several discriminative methods have been developed and there is a wide variety of different approaches that fall into this class of algorithms. The most prominent example of a discriminative probabilistic classifier is the *logistic discrimination*. In this setting, the logarithm of the likelihood ratios is modeled as a linear function:

$$\log \frac{P(y=+1|x)}{P(y=-1|x)} = w_0 + w^T x,$$

where we limit our analysis to the two class case for simplicity of exposition and we assume both x and w to be d-dimensional. It is then possible to take advantage from

$$P(y=+1|x) + P(y=-1|x) = 1$$

to obtain the following structure of the posterior probabilities

$$\begin{aligned} P(y=+1|x) &= \frac{\exp(w_0 + w^T x)}{1+\exp(w_0 + w^T x)}, \\ P(y=-1|x) &= \frac{1}{1+\exp(w_0 + w^T x)}. \end{aligned}$$

The parameters w, w_0 are then usually estimated with a maximum likelihood approach. The marginal densities do not influence the optimization and thus the optimal parameters are obtained by

$$(\widehat{w}, \widehat{w}_0) = \underset{(\widehat{w},\widehat{w}_0)}{\arg\max} \sum_{i=1}^{N} \log P(y_i|x_i),$$

which define a linear classifier.

2.1.3. Surrogate Loss Functions

The logistic discriminant analysis can also be expressed in a different formulation, that is as the minimizer of an appropriately defined empirical risk. Indeed, let us define the *logistic loss* as

$$\ell_{logistic}(f(x), y) = \log\left(1+\exp(-yf(x))\right),$$

where $f(x)$ is the decision function and y is the label corresponding to observation x. We can then decide to restrict our attention to a specific *hypothesis class*, such as for instance that of linear classifiers. In this case $f(x) = w^T x$ and the logistic discriminant classifier is found by minimizing the following risk:

$$R_{logistic}(w|\mathscr{D}_l) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp(-y_i(w^T x_i))\right).$$

This principle goes by the name of *empirical risk minimization* (ERM) and can naturally be extended to a wide range of loss functions. As we are interested in minimizing the misclassification error, it is tempting to use the *0/1 loss*, which is defined as

$$\ell_{0/1}(f(x), y) = \mathbb{1}\big(\text{sign}(f(x)) \neq y\big),$$

where $\mathbb{1}\big(\text{sign}(f(x)) \neq y\big) = 0$ if $\text{sign}(f(x)) = y$, and $\mathbb{1}\big(\text{sign}(f(x)) \neq y\big) = 1$ otherwise. This loss seems the most appropriate, as the prediction of the linear classifier is defined by $\text{sign}(w^T x)$. However, performing ERM with the 0/1 loss is known to be a NP-hard problem even for very simple hypothesis classes [17]. For this reason it becomes necessary to use loss functions that lead to a computationally tractable optimization, while having a somewhat close relation to the 0/1 loss. Because of the former observation, surrogate loss functions are very often chosen to be convex, hence having a

well defined minimization task. Another desirable property of a loss function $\ell(f(x), y)$ would be to define an upper bound to the value of the 0/1 loss, that is

$$\ell_{0/1}(f(x), y) \leq \ell(f(x), y),$$

which then straightforwardly implies a bound on the empirical risk:

$$R_{0/1}(f\,|\mathscr{D}_l) \leq R(f\,|\mathscr{D}_l).$$

A well known class of such loss functions is that of *margin based losses*. These are losses of the par-

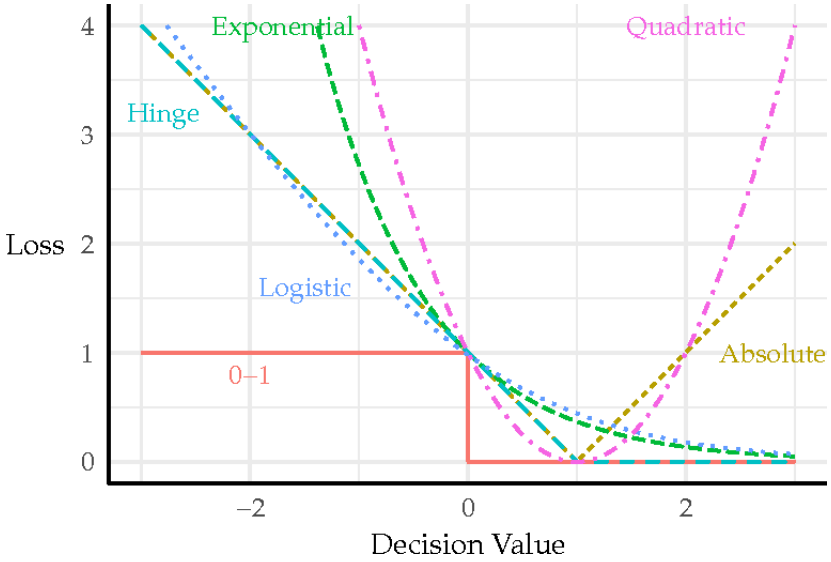


Figure 2.3: *Plot of several convex margin based surrogate loss functions, taken from [29].*

ticular form $\ell(f(x), y) = \phi(y\,f(x))$, where $y\,f(x)$ is usually called *margin*. Figure 2.3 shows some of the most common members of this family, among which we observe the logistic loss. A particularly celebrated margin based loss is the the *hinge loss*:

$$\phi_{hinge}(y\,w^T x) = \max(0, 1 - y\,w^T x).$$

The hinge loss is used to build, together with an additional convex regularization term, the support vector machine (SVM), which was first introduced in [12]. Intuitively, the SVM maximizes the margin between the decision line and the training observations. This simple idea turned out to be very effective and as a consequence this is one of the most well known classifiers.

It is reasonable to wonder whether there are some guarantees that minimizing an empirical risk based on a surrogate loss translates to a minimization of the classification error. This question has been addressed extensively in the literature, for instance in [4, 5, 40, 41, 52, 53]. It turns out that, under some assumptions on the loss function and on the flexibility of the hypothesis class, the ERM framework asymptotically converges to the Bayes rule [4]. Nonetheless, it has to be noted that the rate of convergence remains hard to characterize and thus in a finite sample setting this guarantee may not be too informative.

2.1.4. Other Classification Algorithms

There is a large number of approaches that do not fall into the previous categories. For instance, the *k-nearest-neighbour* (k-NN) [1] classifier assigns an unseen observation to the class to which correspond more than half of the k nearest neighbours in the data-set $\mathscr{D}_l$. Another very famous

alternative is known as *neural networks* [19] and use a combination of several layers and activation functions to predict the output of a feature vector. Then, we can cite the *decision trees* [39]. These are multistage decision systems in which a sequence of logical conditions is used to finally select and accept a class. More classifiers can be considered at once, resulting in a *combined classifier* [23]. This procedure combines several independently trained classifiers using some rule. An example of selection method is the *majority voting rule*, which defines the overall classifier by assigning a feature vector to the class that is the most frequently chosen by the considered classifiers. Finally, we cite *kernel methods* [14], in which the feature vectors are mapped to a different space in which a classifier is trained. Kernel methods can be ideally used to transform the data so that the points from each class can be separated more easily.

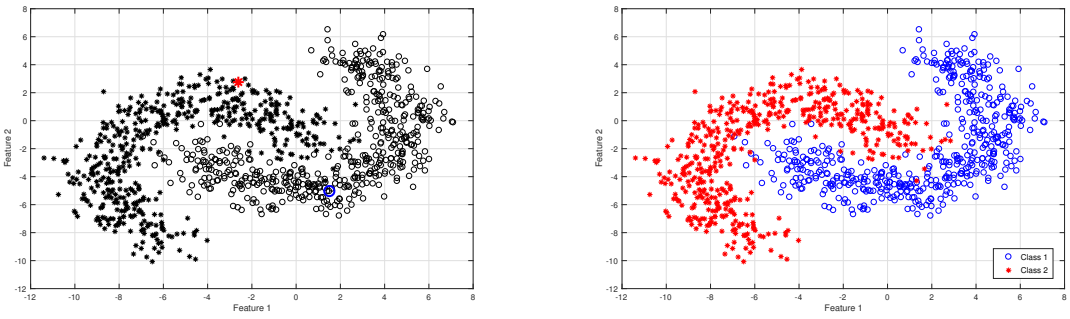
Other approaches to solve a classification problem are available, but are out of the scope of this work. A more detailed literature survey can be found, for instance, in [34].

2.2. Semi-Supervised Learning

In many real world applications the labeling of large amounts of observations can be very expensive and time consuming. This is the case in many applications of machine learning, such as for instance speech recognition and webpage classification [11]. Semi-supervised learning (SSL) is the field of research that studies the problem of learning from both labeled and unlabeled data. It is then evident how SSL is in between supervised and unsupervised learning. As such, SSL can be seen as an unsupervised task with partial supervision or as supervised learning with additional information on the distribution of the feature vectors. Most applications are easier to interpret in the former formulation. The particular structure of the data-set favors the definition of two different ways of building a semi-supervised classifier. These are *inductive* and *transductive learning*. In the former the goal is to learn a general decision rule which can be applied to unseen data, while in the latter the goal is to label (a subset of) the unlabeled observations that are available for the training phase. Either of these principles can be adopted depending on the specific application.

2.2.1. Assumptions in SSL

Most SSL algorithms rely on assumptions on the distribution of the data. Here we mean to highlight some of the most used. For the interested reader, an extensive introduction can be found in [11]. The meaning of these assumptions is in general that the distribution of the examples provides useful information for the classification task. In other terms, the knowledge of $p(x)$ that is obtained using



(a) *Plot of two labeled data-points, one for each class, and of several unlabeled observations.*

(b) *Plot of all the data-points with the corresponding membership to either class.*

Figure 2.4: *Typical example of the information that can be added by the unlabeled data-set. The scatter-plots are of the "banana data-set" from PRTools with* 500 *observations per class.*

the unlabeled observations should be meaningful in the inference of $p(y|x)$.
A first assumption is called the *semi-supervised smoothness assumption*, which states that if two points x_1, x_2 in a high density region are close, then so should be the corresponding outputs y_1, y_2. In other terms, if two points are linked by a path of high density, then their outputs are likely to be close. This can then be formulated slightly differently: if the points are in the same cluster, they are likely to be of the same class. This is known as the *cluster assumption* and is justified by the fact that classes should be identifiable in a classification problem. Clearly, we could consider a high density region as a cluster, hence the two assumptions above have a similar meaning. Equivalently, the *low density separation assumption* states that the decision boundary should lie in a low-density region. These hypotheses are evidently similar, but inspire different classes of algorithms.
Another important assumption is the *manifold assumption*: the (high dimensional) data lie roughly on a low dimensional manifold. This allows to avoid the curse of dimensionality. In some sense, the manifold assumption is again similar to the previous assumptions, as a manifold can be seen as an approximation of high density regions.

2.2.2. Classes of Algorithms for SSL

A first illustrative approach to SSL was introduced in the 1960s, then applied to the setting of Gaussian discriminant analysis in [37] and is generally referred to as *self-learning* or *self-training*. The simple idea is the following. First, suppose that N labeled observation $\mathscr{D}_l = \{(x_i, y_i)\}_{i=1}^N$ and M unlabeled observations $\{u_j\}_{j=1}^M$ are available, where x_i, u_j are feature vectors and y_i are the labels representing the membership of a data point to a class. In self-learning a supervised inductive classifier is trained on the labeled data-set of observations. This can be done in a generative model for instance by computing the parameters that maximize the likelihood on $\mathscr{D}_l$. Then, the obtained classifier is used to label the unlabeled feature vectors $\{u_j\}_{j=1}^M$, according for instance to the Bayes classification rule. This starts an iterative procedure in which a new supervised classifier is trained on the full data-set, where the labels of the unlabeled points are given by the previous classifier. The iterations stop when the method converges to a model that is then the returned semi-supervised classifier. Self learning has in some cases a puzzling behaviour and is known to sometimes lead to performance degradation with respect to the original supervised learner.

Since then, SSL has become a very active field of research and here we present some broad groups of algorithms. A first class of algorithms is composed by *generative models*. As defined earlier, these models choose to estimate the class conditional densities $p(x|y)$, thus information on $p(x)$ is useful. For instance, each conditional density can be modeled as a Gaussian and the optimal parameter can be computed with the EM algorithm [15], which is a standard tool for inference in missing data settings. A well known application of this framework to text classification can be found in [38]. Generative approaches have the useful property that knowledge of the structure of the data can be easily incorporated in the model. However, in Section 2.3.1 we discuss how the wrong modeling assumptions can negatively affect the classifier.
Another class of methods is composed by algorithms that try to directly implement the low-density separation assumption we introduced in Section 2.2.1, that is by pushing the decision boundary away from the unlabeled observations. The underlying idea is evidently very similar to SVM, for which a semi-supervised version called transductive SVM was introduced in [21]. Because of the transductive nature of the algorithm, predictions are made only at a fixed number of test points. A different SVM-based approach that takes advantage of the low-density assumption is [31] and is discussed in more detail in Section 2.3.
The last class we discuss is that of *graph-based methods*. The basic idea is to represent the observations as nodes of a graph and to use the edges to represent the distance between the incident nodes. Then, a decision function is chosen by minimizing the preferred measure on labeled and unlabeled

data. A possible choice is for instance the quadratic energy function, as discussed in [56], [54]:

$$E(f) = \frac{1}{2}\sum_{i,j} w_{ij}\big(f(i)-f(j)\big)^2,$$

where w_{ij} is the distance between nodes i and j, while $f(\cdot)$ is the function that is learned by minimizing the energy. Several methods rely on the graph Laplacian [50]. For instance, in [6] the Laplacian is used to define a Bayesian approach to graph-based learning that allows for uncertainty quantification in predictions. Graph based methods are naturally transductive, but generalizations to inductive classifiers have been studied.

It should by now be clear that there are several approaches to semi-supervised learning. We point to [55] for a comprehensive and detailed survey, which is out of our scope. In this work however we focus on a specific problem that arises in SSL, which is discussed in the next section.

2.3. Safe Semi-Supervised Learning

Semi-supervised learning has been shown to be a useful tool in many areas of application. However, semi-supervision can lead to a performance degradation with respect to the corresponding supervised classifier. This unwanted and worrying behaviour goes against the fundamental idea that the addition of unlabeled data results in a more accurate model. There has then been increasing attention on this issue and here we discuss the most interesting works in this direction. We start with a theoretical analysis of the case in which a wrong model is chosen in a generative setting, as discussed in [13]. Then, we introduce two methods that try to mitigate this effect in different ways. These are [22, 31]. Finally, we dedicate a longer section to the approach that is the focus of this thesis [29, 33].

2.3.1. Risks of Semi-Supervised Learning

The literature of semi-supervised learning contains a number of papers guaranteeing benefits from unlabeled data for generative models that are based on a correct model. For instance, we can cite [9], [10]. On the other hand, there is strong empirical evidence that unlabeled data may cause a performance degradation. As stated in [13] "performance degradation may occur whenever the modeling assumptions adopted for a particular classifier do not match the characteristics of the distribution generating the data". This is a crucial problem, as it is generally very difficult to show that a model is correct and the same holds for the assumptions that are very often made in semi-supervised learning algorithms. This claim is supported by an asymptotic analysis of the maximum likelihood estimates. To understand this better, we assume a generative model is adopted and thus that the joint probability is modeled by a parametric family of distributions $p(x,y|\theta)$. Suppose now that the labels are missing completely at random (MCAR) [32]. This is assumed through the entire thesis. Suppose then and that the semi-supervised estimates are computed by maximizing the following log-likelihood:

$$L(\theta) = \prod_{i=1}^{N} p(x_i,y_i|\theta)\prod_{j=1}^{M} p(u_j|\theta),$$

where the marginal density of x is a mixture $p(x|y=+1,\theta)p(y=+1|\theta)+p(x|y=-1,\theta)p(y=-1|\theta)$ that is assumed to be identifiable. Furthermore, we call λ the probability that a label is not hidden. Then, it is possible to show that the limiting value of the maximum likelihood estimates is the following:

$$\theta^*_\lambda = \arg\max_\theta \Big(\lambda\mathbb{E}_{p(x,y)}\big(\log p(X,Y|\theta)\big)+(1-\lambda)\mathbb{E}_{p(x,y)}\big(\log p(X|\theta)\big)\Big). \tag{2.1}$$

It is possible to see that the semi-supervised objective function is asymptotically a combination of objective functions for supervised and unsupervised learning. We can then denote the labeled

and unlabeled limits respectively as θ_1^* and θ_0^*. Now, for a correct model we have that there is a parameter θ_T such that $p(x,y|\theta_T) = p(x,y)$. In this case the model is consistent by identifiability, thus $\theta_1^* = \theta_0^* = \theta_T$ and the classification error converges to the Bayes error, that is the lowest value attainable. Hence, the addition of unlabeled observations is useful to converge to the Bayes error if the model is correctly specified.
It can also happen that the true underlying distribution is not in the parametric family that is used by the model. In that case, let us denote by $e(\theta)$ the classification error with parameter θ and suppose $e(\theta_0^*) > e(\theta_1^*)$. Then, for a large number of labeled objects the error converges to $e(\theta_1^*)$. But the limit approaches $e(\theta_0^*)$ if we add a large number of unlabeled observations. Therefore, the classification error is higher using the complete data-set compared to using labeled data only. In this sense we can expect this undesired asymptotic behaviour of the semi-supervised classifier. It has however to be noted that this analysis does not imply that, in a finite sample setting, the addition of unlabeled data cannot result in an improvement even with a wrong model.
In the following sections we discuss three proposals to overcome this issue.

2.3.2. Safe Semi-Supervised Learning Based on Weighted Likelihood

We present here a first approach to safe semi-supervised learning, that is discussed in [22] and is an extension of [45]. Here we mean to highlight the main ideas, assumptions and results.
Given labeled and unlabeled data, the authors choose the following statistical models

$$\mathscr{P}_{y|x} = \left\{ p(y|x,\alpha) : \alpha \in \mathscr{A} \subset \mathbb{R}^d \right\}$$

and

$$\mathscr{P}_x = \left\{ g(x|\eta) : \eta \in \mathscr{N} \subset \mathbb{R}^k \right\}.$$

A key assumption is that $\mathscr{P}_x$ is assumed to be correctly specified, which means that it contains the true marginal density. Nonetheless, $\mathscr{P}_{y|x}$ is not necessarily correctly specified. This is then a weaker assumption than [10] and it implies that the model for the joint density may be incorrect.
Now, the approach uses a weight factor, that is

$$w_N(x|\eta,\eta') := \frac{g(x|\eta') + \rho N^{\nu}}{g(x|\eta) + \rho N^{\nu}},$$

where $-1/2 < \nu < 0$ and $\rho \geq 0$ are used for regularization purposes. The model is then defined by

$$\begin{aligned}
\widehat{\eta} &:= \arg\max_{\eta} \sum_{i=1}^{N} \log(g(x_i|\eta)), \\
\widehat{\eta}' &:= \arg\max_{\eta} \sum_{j=1}^{M} \log(g(u_j|\eta)), \\
\tilde{\alpha} &:= \arg\max_{\alpha} \sum_{i=1}^{N} w_N(x_i|\widehat{\eta},\widehat{\eta}') \log(p(y_i|x_i,\alpha)).
\end{aligned}$$

The estimator $\tilde{\alpha}$ is called DRESS (Density Ratio Estimation-based Semi-Supervised estimator) I. A different estimator can be obtained by slightly changing the previous formulation by

$$\widehat{\eta}' := \arg\max_{\eta} \sum_{i=1}^{N} \log(g(x_i|\eta)) + \sum_{j=1}^{M} \log(g(u_j|\eta)),$$

while the remaining estimates are left unchanged. This alternative is called DRESS II and is showed geometrically to be better than DRESS I. It is then proved by asymptotic theory that DRESS I never

performs worse asymptotically than the respective supervised estimator in terms of parameter estimation if $g(x|\eta)$ is correctly specified and there are more unlabeled than labeled points, i.e. $M > N$. Furthermore, DRESS I improves the supervised model if $\mathscr{P}_{y|x}$ is not correctly specified, while if it is then DRESS I is equivalent to the supervised estimate. Under these assumptions, it is also shown that DRESS II is asymptotically better than DRESS I, hence it improves even more the supervised model if $\mathscr{P}_{y|x}$ does not contain the true density. It is worth noting that the improvement happens in the asymptotic variance.
While these results are interesting, some arguably strong assumptions are required. Moreover, in practical applications the amount of data is limited and an asymptotic result of improvement can be unreliable, thus meaning that a degraded performance can still be observed in a finite sample setting. In the next section we discuss a method with a finite sample guarantee.

2.3.3. SV4Ms

In [31] Li and Zhou proposed a safe semi-supervised version of the support vector machine called SV4Ms. The method is first based on the assumption that the data-set is such that there exist multiple low density separators $\mathscr{M} = \{\widehat{y}_t\}_{t=1}^T$. This is claimed to be usually the case with large amounts of unlabeled data. Denote now the supervised inductive SVM as w_{svm}, its output on the unlabeled data as y_{svm} and the ground truth labels as y^*. Then, for any label assignment $y \in \{-1,+1\}^M$ introduce $earn(y, y^*, y_{svm})$ and $lose(y, y^*, y_{svm})$, which are respectively the number of points in which y is right and y_{svm} is not and the other way around, where the truth is y^*. We can then define

$$J(y, y^*, y_{svm}) = earn(y, y^*, y_{svm}) - \lambda\, lose(y, y^*, y_{svm}),$$

where λ is a parameter that determines the amount of risk that the user would like to undertake. Intuitively, we would like to maximize $J(y, y^*, y_{svm})$. However, y^* is unknown. To mitigate this difficulty the ground truth labeling is assumed to be realized by a low density separator in $\{\widehat{y}_t\}_{t=1}^T$, i.e. $y^* \in \mathscr{M}$. The semi-supervised SVM classifier is then defined as

$$\bar{y} = \underset{y\in\{-1,+1\}^M}{\operatorname{argmax}} \min_{\widehat{y}\in\mathscr{M}} J(y, \widehat{y}, y_{svm}).$$

It can then be proved that if $y^* \in \mathscr{M}$ and $\lambda \geq 1$, then the accuracy of $\bar{y}$ is never worse than that of y_{svm} in the sense that $earn(\bar{y}, y^*, y_{svm}) \geq lose(\bar{y}, y^*, y_{svm})$. This is equivalent to saying that the classification error of $\bar{y}$ on the unlabeled data-set is lower than that of the supervised classifier y_{svm}. Note that on the labeled data y_{svm} is likely a better predictor than $\bar{y}$ as it is chosen on the labeled data-set. Moreover, there are no guarantees on the generalization performance of SV4M on unseen data. Finally, it has to be noted that the assumption that the true labeling of the unlabeled observations is assigned by one of the low density separators is very strong. This assumption can in no way be checked and is a key argument in the analysis in [31]. Therefore, violations of this assumption have a crucial effect on the validity of the improvement guarantee. It is then unclear whether one can expect an improvement given a real data-set.

2.3.4. Contrastive Pessimistic Semi-Supervised Learning

A different approach to safe semi-supervised learning was proposed in [33] for a specific generative classifier, and then extended to margin based losses in [29]. We do not discuss [33] in this section, as we extend its analysis in Chapters 4 and 5. Therefore, here we report the main findings in [29] and we use this as a first introduction to the main concepts.
The idea is to explicitly compare the performance of the semi-supervised with that of the supervised classifier on both the labeled and unlabeled data in the worst case scenario. This pessimistic choice can in some cases be used to establish whether we can conclude our classifier is never worse than the supervised model. The performance measure that is adopted to compare the two models is

the same that is used for the training of the supervised classifier. For instance, if a specific loss is selected, then the model minimizes a semi-supervised version of that loss which includes the unlabeled feature vectors. This makes sure that the result is not affected by other factors and that really reflects a possible improvement.
Here we discuss how the application of these principles proves that there is no safe semi-supervised learner for some choices of margin based losses, as discussed in [29].

We can then consider the task of binary classification with a restriction to linear classifiers. This means that the vector of labels is $y \in \{-1,+1\}^N$, while the labeled feature vectors are grouped in the $N \times d$ design matrix X. Then, any new measurement $x \in \mathbb{R}^d$ is classified according to $\text{sign}(x^T w)$.
We can now define a supervised linear classifier by minimizing the following empirical risk:

$$R_\phi(w|X,y) = \sum_{i=1}^{N} \phi(y_i x_i^T w) + \lambda(w),$$

where ϕ is a margin based loss and $\lambda(w)$ is a regularization term that is assumed convex in w. Therefore, the supervised estimate of w is

$$\widehat{w}_{sup} = \underset{w\in\mathbb{R}^d}{\arg\min}\, R_\phi(w|X,y).$$

We can then add M unlabeled observations u_j, which define a $M \times d$ dimensional design matrix X_u. In order to take advantage of this additional data-set, a semi-supervised risk is defined:

$$R_\phi^{semi}(w,q|X,y,X_u) = R_\phi(w|X,y) + \sum_{j=1}^{M} \left(q_i \phi(u_i^T w) + (1-q_i)\phi(-u_i^T w) \right), \tag{2.2}$$

in which $q \in [0,1]^M$ are coefficients that expresses the unknown and possibly soft membership of each unlabeled object to a class. Naturally, the true labels would correspond to hard memberships $q_{true} = \{-1,+1\}$. Now, we can compare a classifier w with the supervised classifier by defining the following difference

$$D_\phi(w,q|\widehat{w}_{sup},X,y,X_u) = R_\phi^{semi}(w,q|X,y,X_u) - R_\phi^{semi}(\widehat{w}_{sup},q|X,y,X_u).$$

A semi-supervised classifier $\widehat{w}_{semi}$ is then said to be safe if the following condition is satisfied:

$$\max_{q\in[0,1]^M} D_\phi(\widehat{w}_{semi},q|\widehat{w}_{sup},X,y,X_u) \leq 0. \tag{2.3}$$

In other words, the classifier is safe if it has a lower semi-supervised risk than $\widehat{w}_{sup}$ even for the worst possible labeling. This guarantees that in any other case the semi-supervised classifier can never result in a degradation in terms of the chosen loss. We can then note that $q_{true} \in \{-1,+1\}^M \subset [0,1]^M$, so this property holds in particular on the full data.
It is useful to define the set of classifiers induced by different responsibilities as

$$\mathscr{C}_\phi = \left\{ \underset{w\in\mathbb{R}^d}{\arg\min}\, R_\phi^{semi}(w,q|X,y,X_u) : \quad q \in [0,1]^M \right\}. \tag{2.4}$$

This set is called the constraint set and plays an important role in determining whether a safe classifier can be built at all. Indeed, Lemma 1 in [29] states that if $R_\phi(w|X,y)$ is strictly convex, $\widehat{w}_{sup} \in \mathscr{C}_\phi$ and $\widehat{w}_{semi} \neq \widehat{w}_{sup}$, then there exists a q^* such that

$$R_\phi^{semi}(\widehat{w}_{semi}|X,y,X_u,q^*) > R_\phi^{semi}(\widehat{w}_{sup}|X,y,X_u,q^*).$$

It is easy to see that this happens for q^* such that

$$\widehat{w}_{sup} = \underset{w\in\mathbb{R}^d}{\arg\min}\, R_\phi^{semi}(w|X,y,X_u,q^*).$$

This result serves as an intermediate step to prove that $\widehat{w}_{sup}$ is always in the constraint set if the chosen margin based loss is decreasing and convex, while the supervised risk is strictly convex. For such losses the supervised vector of coefficients can always be retrieved by choosing

$$q_i = \frac{\phi'(-u_i^T\widehat{w}_{sup})}{\phi'(u_i^T\widehat{w}_{sup})+\phi'(-u_i^T\widehat{w}_{sup})},$$

which is sufficient to set to zero the gradient of the semi-supervised risk evaluated at $\widehat{w}_{sup}$. This result shows that for decreasing convex margin based losses it is not possible to build a semi-supervised classifier that satisfies Condition (2.3) while strictly improving over the supervised classifier for at least a choice of q. It is then clear that several well known losses cannot be used to build a safe semi-supervised learner. Examples of decreasing convex losses are the hinge loss, which is used to build the support vector machine, the logistic loss and the exponential loss. On the other hand, losses that do not fall into this analysis are for instance the quadratic and the absolute loss.
In order to understand when an improvement is to be expected we can focus on the following minimax problem:

$$\min_{w\in\mathbb{R}^d}\max_{q\in[0,1]^M} D_\phi(w,q|\widehat{w}_{sup},X,y,X_u,q). \tag{2.5}$$

This is equivalent to searching for the classifier that minimizes the risk for the most adversarial choice of the soft assignments of each unlabeled observation to a class. Now, in [29] it is claimed that for convex losses Sion's minimax theorem (Corollary 3.3 in [44]) can be applied to show that indeed no degradation can occur if $\widehat{w}_{semi}$ is computed by (2.5). We comment and clarify this claim rigorously in Chapter 3. Nonetheless, a strict improvement is not guaranteed if the supervised classifier is in the constraint set. It is then necessary to define at least some sufficient conditions that certify that this does not happen. We can for instance consider the gradient of R_ϕ^{semi}, evaluate it in $\widehat{w}_{sup}$ and check when it is null. In the cases where this condition does not hold, it is possible to state that the semi-supervised classifier comes with a performance improvement. This line of reasoning results in the following equation:

$$\sum_{j=1}^{M} q_i\phi'(u_i^Tw)u_i - (1-q_i)\phi'(-u_i^Tw)u_i = 0. \tag{2.6}$$

If only one unlabeled observation is available , that is $X_u = u^T$, Equation (2.6) is equivalent to requiring that there is no $q\in[0,1]$ such that

$$\left(\phi'(u^Tw)+\phi'(-u^Tw)\right)uq = \left(\phi'(-u^Tw)\right)u. \tag{2.7}$$

In addition, a different condition is expressed for convex losses that are decreasing to the left of 1 and increasing to the right of 1, which means

$$\phi'(a)\begin{cases}\leq 0, & \text{if } a\leq 1,\\ >0, & \text{if } a>1.\end{cases}$$

Theorem 3 in [29] states that for these losses an improved semi-supervised estimator is obtained if $|u_i^T\widehat{w}_{sup}|>1$ for all unlabeled observations. As the authors discuss, this is a very strong requirement and is not in general usable in practical applications. The quadratic loss is in particular included in

this theorem, but a sharper result is obtained. An improved semi-supervised classifier is expected in case $d \leq M$ if

$$||X_u \widehat{w}_{sup}||_2^2 > U .$$

It is then clear that for the quadratic loss it is not necessary that all unlabeled points lie outside of the margin, but it is sufficient that some observations are sufficiently far outside of it.

This work illustrates that guarantees of performance improvement are very hard to obtain in the setting of empirical risk minimization with margin based losses. For decreasing convex losses it even shows that it does not exist a safe classifier that improves over the supervised estimator for some choices of q. Nonetheless, classifiers with such properties do exist for non-decreasing losses if the data-set satisfies some requirements.
The use of contrast and pessimism seems to be the only way to verify whether safe improvements are at all possible. While being a strong restriction, we previously discussed that determining the existence of this type of classifiers is of great importance in semi-supervised learning. For instance, if it is proved that a safe improved classifier cannot be found using a certain loss function, then we should avoid using that particular loss. Finally, it is crucial to observe that no assumptions on the data are made. This differentiates this method from the previous works and makes its analysis more reliable and meaningful.

In the next chapter we formulate the contrastive pessimistic semi-supervised version of least squares classification. As we will see, in that setting we are able both to show a non degradation property and to define a condition for strict performance improvement.

3

Contrastive Pessimistic Least Squares Classification for Semi-Supervised Learning

We start our analysis with the extension of the contrastive pessimistic approach to *least squares classification*. This setting was studied in [27, 28] and therefore allows also an interesting comparisons of the proposed method with other algorithms. The first sections are dedicated to the introduction of the necessary theoretical tools on minimax problems and on least squares classification. We then move to the *contrastive pessimistic least squares* (CPLS) classifier, for which properties of non-degradation and strict improvement are proven in Theorems 3.6 and 3.8.

3.1. Theory of Minimax Problems

Minimax problems play an important role in the analysis of the contrastive pessimistic approach, so we dedicate this section to an introduction of the necessary theory.

In order to define a minimax problem we first introduce two non-empty sets $\mathcal{X}$ and $\mathcal{Y}$ and a generic function $K : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We are then interested in taking for each $x \in \mathcal{X}$ the infimum of $K(x, y)$ over $y \in \mathcal{Y}$ and then taking the supremum of the infimum as a function of $x \in \mathcal{X}$, which results in

$$\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} K(x, y) .$$

Symmetrically, we could need to take the supremum before the infimum

$$\inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} K(x, y) .$$

The two minimax formulations generally have different values. This is illustrated by the following inequality, which is Lemma 36.1 in [42]:

$$\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} K(x, y) \leq \inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} K(x, y) . \tag{3.1}$$

In some cases equality is attained in (3.1), that is equivalent to saying that the order of the optimization does not influence the result. The common value is called minimax or saddle value of function K. In case the supremum on the left hand side of inequality (3.1) and the infimum on the right are attained, then the solution is said to be a *saddle point*. This is a point $(\bar{x}, \bar{y})$ such that

$$K(x, \bar{y}) \leq K(\bar{x}, \bar{y}) \leq K(\bar{x}, y) \qquad \forall x \in \mathcal{X}, \quad \forall y \in \mathcal{Y} .$$

The existence of a saddle point is a necessary and sufficient condition for the *minimax equality*:

$$\max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} K(x,y) \;=\; \min_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}} K(x,y)\,. \tag{3.2}$$

Nonetheless, there may not be a saddle point of the minimax problem unless the objective function satisfies some assumptions. It is still important to establish whether equality holds in (3.1), and even more if a saddle point exists. An important result in this direction is Lemma 36.2 in [42], which we now state.

Lemma 3.1. *Let $K : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and let $\mathcal{X} \times \mathcal{Y}$ be a non-empty product set. A point $(\bar{x}, \bar{y})$ is a saddle point if and only if the following conditions hold:*

i. the supremum in $\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} K(x,y)$ is attained at $\bar{x}$,

ii. the infimum in $\inf_{y\in\mathcal{Y}} \sup_{x\in\mathcal{X}} K(x,y)$ is attained at $\bar{y}$,

iii. equality in (3.1) holds.

Moreover, if $(\bar{x}, \bar{y})$ is a saddle point, then the saddle value of K is $K(\bar{x}, \bar{y})$.

There are several theorems in the literature that give conditions under which the supremum and the infimum are interchangeable. The first result of this kind was von Neumann's theorem [51], which was restricted to compact sets $\mathcal{X}$, $\mathcal{Y}$. Since then, many researchers have proved that equality in (3.1) holds under weaker assumptions. In particular, Sion proved a very interesting generalization that holds when only one of the two sets is compact. This is Corollary 3.3 in [44]. The result is stated in a rather technical form and uses advanced notions as topological spaces and semicontinuous functions. However, the settings that are treated in this thesis do not require such an abstraction. We will deal with vector spaces on which it is trivial to define a topology and it will be sufficient to know that a continuous convex function is also semicontinuous and quasi-convex. The interested reader can find the more abstract definitions in Appendix A.

Theorem 3.2. *Let $\mathcal{X}$ be a convex subset of a linear topological space and $\mathcal{Y}$ be a compact convex subset of a linear topological space. Let $K : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function which satisfies the following properties:*

i. $K(\cdot, y)$ is upper semicontinuous and quasi-concave on $\mathcal{X}$ for each $y \in \mathcal{Y}$,

ii. $K(x, \cdot)$ is lower semicontinuous and quasi-convex on $\mathcal{Y}$ for each $x \in \mathcal{X}$.

Then

$$\sup_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} K(x,y) \;=\; \min_{y\in\mathcal{Y}} \sup_{x\in\mathcal{X}} K(x,y). \tag{3.3}$$

Note that the theorem is here stated similarly as in [25], where the infimum is replaced by a minimum because $\mathcal{Y}$ is a compact set and we can use the Weierstrass theorem for lower semicontinuous functions (see Theorem 2.8 in [3]).

3.2. Least Squares Classification

In Section 2.1 we discussed several alternative approaches to solve the problem of supervised learning. Here we focus on learning a classifier by minimizing the square loss:

$$\ell_{ls}(y, f(x)) \;=\; (y - f(x))^2\,.$$

The training set in the supervised setting is assumed to be composed of N labeled points $\{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^D$ and $y_i \in \{0,1\}^N$ for any $i = 1,\dots,N$. We consider the class of linear classifiers with intercept, that is

$$f(x) = w_s^T x + w_0,$$

where $w_s \in \mathbb{R}^D$, $w_0 \in \mathbb{R}$. In order to lighten the notation, the intercept can be embedded in the slope coefficient as $w = (w_s^T, w_0)^T$. The feature vectors are also modified as $x_i = (x_i^T, 1)^T$, with a clear abuse of notation. Therefore, the dimensionality of w and of each x_i becomes $(D+1)$. It is quite useful to express the problem in matrix notation, thus we call $X \in \mathbb{R}^{N\times(D+1)}$ the design matrix that has x_i^T at the i-th row. Similarly, $y_i \in \{0,1\}^N$ is the vector of labels.
The least squares classifier is obtained by minimizing the following empirical risk:

$$\begin{aligned} R(w|X,y) &= \sum_{i=1}^N \ell(y_i, w^T x_i) \\ &= \sum_{i=1}^N (y_i - x_i^T w)^2 \\ &= \left\| y - Xw \right\|^2, \end{aligned} \tag{3.4}$$

that is the sum of the squared errors. Then, the optimal vector w is

$$\widehat{w}_{sup} = \underset{w\in\mathbb{R}^{D+1}}{\operatorname{argmin}} R(w|X,y). \tag{3.5}$$

By convexity of the objective function, we can compute $\widehat{w}_{sup}$ by setting the gradient of $R(w|X,y)$ to 0. The resulting solution is the following

$$\begin{aligned} \widehat{w}_{sup} &= \left(\sum_{i=1}^N x_i x_i^T\right)^{-1} \sum_{i=1}^N y_i x_i \\ &= \left(X^T X\right)^{-1} X^T y. \end{aligned} \tag{3.6}$$

Note that the matrix $X^T X = \sum_{i=1}^N x_i x_i^T$ is usually referred to as the sample correlation matrix, while $X^\dagger = (X^T X)^{-1} X^T$ is the pseudo-inverse of X. Clearly, this notation for the pseudo-inverse is meaningful if $X^T X$ is invertible, which is not the case for instance if $d > N$. In case invertibility does not hold, the optimal vector of parameters can be computed either using the singular value decomposition to compute the pseudo-inverse $X^\dagger$ or adding a regularization term in the loss function. In particular, the most common choices of regularization are the Lasso [48] and Ridge, which consist in adding respectively the L^1 or L^2 norm of w.

3.2.1. Semi-Supervised Least Squares Classification

In the semi-supervised setting we incorporate M additional unlabeled observations $\{u_j\}_{j=1}^M$, which are contained in the design matrix $X_u \in \mathbb{R}^{M\times(D+1)}$. The complete design matrix is then defined as $X_e = (X^T, X_u^T)^T$. In order to avoid any issue, in the following theoretical framework we assume both $X^T X$ and $X_e^T X_e$ to be invertible. We can then define the complete vector of labels as $y_e^* = (y^T, (y_u^*)^T)^T$, where $y_u^* \in \{0,1\}^M$ is unknown to the user in SSL. Let us now define the *oracle solution* as the minimizer of the empirical risk on the complete data-set (X_e, y_e^*):

$$\widehat{w}_{oracle} = \left(X_e^T X_e\right)^{-1} X_e^T y_e^*. \tag{3.7}$$

The oracle solution is then the optimal classifier and serves as a benchmark solution. Semi-supervised algorithms should aim to have a performance as close to the oracle as possible, while outperforming the supervised solution.
In the next section we discuss two least squares-based approaches to semi-supervised learning.

3.3. Preexisting Least Squares Methods for Semi-Supervised Learning

The two approaches we discuss in this section are the *projection method* [28] and the *implicitly constrained least squares* (ICLS) approach [27]. These methods possess some properties on their robustness that are very desirable in practical applications.

3.3.1. Projection Method

The projection method [28] is an approach based on the intuitive idea that it is possible to get closer to the oracle solution by defining a set that contains it and then projecting the supervised model onto this set. Clearly, there are several choices as of how to measure the distance between two parameter vectors. As a consequence, different choices of the metric lead to different results. In one of the available formulations the semi-supervised solution is guaranteed to be at least as good as the supervised solution in terms of the square loss on both labeled and unlabeled data. This robustness property shows that for some classifiers it is possible to build a semi-supervised version that is at least as good as their supervised counterpart on the full data set for any labeling of the unlabeled observations.

First, it is necessary to define a set that contains the oracle solution. The key observation is that by varying the labels on the unlabeled observations it is possible to obtain the optimal vector of parameters using the closed form solution to the least squares criterion for any specific labeling. Then, choosing y_e^* we obtain $\widehat{w}_{oracle}$. This set can be defined as

$$\mathscr{C}_w := \left\{ w = (X_e^T X_e)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} : \quad y_u \in [0,1]^M \right\}. \tag{3.8}$$

Note that soft labels are allowed.
Then, a notion of distance has to be chosen. The authors consider the following metric:

$$d(w_1, w_2) = \sqrt{(w_1 - w_2)^T X_o^T X_o (w_1 - w_2)}, \tag{3.9}$$

where matrix $X_o^T X_o$ is assumed to be positive definite. The semi-supervised estimate is then defined as

$$\widehat{w}_{proj} = \underset{w \in \mathscr{C}_w}{\operatorname{argmin}}\, d(w, \widehat{w}_{sup}). \tag{3.10}$$

Different choices of matrix X_o lead to different solutions of the minimization task. Selecting the full design matrix X_e it is possible to show that the obtained vector of parameters is robust as discussed earlier. More precisely, given the design matrices X and X_e and assuming $X_e^T X_e$ positive definite, the projected estimator satisfies the following inequality:

$$\left\| y_e^* - X_e \widehat{w}_{proj} \right\|^2 \leq \left\| y_e^* - X_e \widehat{w}_{sup} \right\|^2.$$

This means that in terms of the least squares criterion the projected solution performs always at least as well as the supervised parameter vector. The empirical tests performed in [28] corroborate this result and also show that the projection method succeeds in lowering the classification error.

There are other choices for matrix X_o that could seem natural. One of the most natural is certainly the design matrix of labeled observations. This choice seems to lead to a less conservative method known as the Implicitly Constrained Least Squares (ICLS) classifier, which is discussed in the next section.

3.3.2. Implicitly Constrained Least Squares

The ICLS approach [27] builds a least squares-based semi-supervised classifier using a particular choice of distance in the setting of the projection method, that is choosing X_o to be the design

matrix of labeled observations. This results in an estimate that performs better empirically, but that does not have the robustness guarantee discussed above.

The ICLS estimate is defined as the element of $\mathscr{C}_w$ which minimizes the supervised least squares criterion. This idea can be formulated mathematically as:

$$\begin{aligned} \min_{w \in \mathbb{R}^{D+1}} \quad & \left\| y - Xw \right\|^2 \\ \text{subject to} \quad & w \in \mathscr{C}_w . \end{aligned} \tag{3.11}$$

This results in a standard quadratic programming problem which is stated to be efficiently solvable using a quasi-Newton approach.
Now, we can show the claim that the ICLS approach is in fact equivalent to the projection method when the metric is induced by the design matrix X. Indeed, we can express the minimization as

$$\min_{w \in \mathscr{C}_w} \left\| y - Xw \right\|^2 = \min_{w \in \mathscr{C}_w} \left(\|Xw\|^2 - 2y^T Xw \right) + \left\| y \right\|^2 ,$$

where the last term is out of the minimum because it does not depend on w. On the other hand, the minimization task of the projection method can be written as

$$\min_{w \in \mathscr{C}_w} (w - \widehat{w}_{sup})^T X^T X (w - \widehat{w}_{sup}) = \min_{w \in \mathscr{C}_w} \left(\|Xw\|^2 - 2y^T Xw \right) + y^T X (X^T X)^{-1} X^T y ,$$

where we used that $\widehat{w}_{sup}$ is of the form defined in (3.6). It is then clear that the semi-supervised vector of parameters is the same for both methods, hence the equivalence between them.

The ICLS has an interesting theoretical result which is now discussed to close this subsection. Assume the measurements x_i and u_i are scalars and assume exact knowledge of the marginal distribution $p(x)$. This assumption is basically equivalent to having unlimited unlabeled data, as this means solving a density estimation problem with an infinite amount of observations. Then, consider a linear model without intercept, i.e. $y = xw$. In this setting, the expected risk is defined as

$$R^*(w) = \sum_{y \in \{0,1\}} \int_{-\infty}^{\infty} (y - xw)^2 p(x,y) dx$$

and the optimal value is $w^* = \operatorname{argmin}_{w \in \mathbb{R}} R^*(w)$. It is then proven that the ICLS estimate has an equal or lower risk than the supervised solution:

$$R^*(\widehat{w}_{semi}) \leq R^*(\widehat{w}_{sup}) .$$

This result shows how in this specific setting the ICLS approach always obtains an estimate which is at least as good as the supervised one. While this property is shown in a very specific and limited setting, empirical tests indicate that the performance of this semi-supervised classifier is usually better than that of the supervised case one.

3.4. Contrastive Pessimistic Semi-Supervised Least Squares Classification

We now extend the principles discussed in Section 2.3.4 in order to define a semi-supervised least squares-based classifier that is at least guaranteed to be at least as good as the corresponding supervised model. This property is formally stated in Theorem 3.6, while a condition for strict improvement is Theorem 3.8. Moreover, in Proposition 3.4 the CPLS approach is showed to be equivalent to the projection method for a specific choice of matrix X_o.

3.4.1. Definition of the CPLS Classifier

First, a risk that includes both labeled and unlabeled data has to be defined. We define the semi-supervised risk as

$$R(w,q|X_e,y) \;=\; R(w|X,y) \;+\; \sum_{j=1}^{M}\Big(q_j(u_j^T w)^2 + (1-q_j)(1-u_j^T w)^2\Big)\,, \tag{3.12}$$

where each $q_j = p(y=0|u_j)$ is the conditional probability of observing label 0 given measurement u_j. Each u_j is partially assigned to either class where the weight is assigned by vector q. Alternatively, we could define a soft label $y_j^u = 1-q_j$ that describes partial membership. We will switch from one formulation to the other, depending on the current situation.
The semi-supervised risk of w has to be compared to that of $\widehat{w}_{sup}$, thus defining the *contrastive risk*:

$$CR(w,q|\widehat{w}_{sup},X_e,y) \;=\; R(w,q|X_e,y) \;-\; R(\widehat{w}_{sup},q|X_e,y)\,. \tag{3.13}$$

This quantity allows us to explicitly compare our semi-supervised solution with the supervised one. Note that Equations (3.12) and (3.13) are functions of the vector of posterior probabilities q, which is clearly unknown. We choose the most pessimistic posteriors, i.e. the vector q that minimizes the improvement of our estimate w with respect to $\widehat{w}_{sup}$. This results in the *contrastive pessimistic risk* (CPR):

$$CPR(w|\widehat{w}_{sup},X_e,y) = \max_{q\in[0,1]^M} CR(w,q|\widehat{w}_{sup},X_e,y)\,. \tag{3.14}$$

The CPLS semi-supervised solution is then defined as follows.

Definition 3.3. *Let $\widehat{w}_{sup}$ be the minimizer of the supervised risk $R(w|X,y)$. Let X_e be the full design matrix, which includes the M unlabeled observations. The CPLS solution is the minimizer of the CPR:*

$$\widehat{w}_{CPLS} \;=\; \operatorname*{arginf}_{w\in\mathbb{R}^{D+1}} CPR(w|\widehat{w}_{sup},X_e,y)\,.$$

We have thus defined a minimax problem that can be written as

$$\inf_{w\in\mathbb{R}^{D+1}}\max_{q\in[0,1]^M}\Big(R(w,q|X_e,y) \;-\; R(\widehat{w}_{sup},q|X_e,y)\Big)\,. \tag{3.15}$$

In the next subsection we show that this formulation can be interpreted differently.

3.4.2. Interchangeability of the Minimization and the Maximization

In Section 3.1 we stated the very powerful Sion's minimax theorem, which was Theorem 3.2. This result proves to be extremely useful in the analysis of the CPLS, for instance in the proof of Theorem 3.6. The assumptions on the objective function are perfectly satisfied by the contrastive risk a defined in (3.13). Indeed, $CR(w,q|\widehat{w}_{sup},X_e,y)$ is continuous in both variables, and in addition concave in q and convex in w. Moreover, the vector of posteriors q varies in a compact set, that is the multi-dimensional box $[0,1]^M$. As a consequence, the infimum and the maximum can be freely interchanged:

$$\inf_{w\in\mathbb{R}^{D+1}}\max_{q\in[0,1]^M}\Big(R(w,q|X_e,y) \;-\; R(\widehat{w}_{sup},q|X_e,y)\Big) \;=\; \max_{q\in[0,1]^M}\inf_{w\in\mathbb{R}^{D+1}}\Big(R(w,q|X_e,y) \;-\; R(\widehat{w}_{sup},q|X_e,y)\Big)\,.$$

This property allows us to compute the structure of w that results by solving the infimum first. Since the objective function is strictly concave in w, we can simply set its gradient to 0 to obtain

$$-2\sum_{i=1}^{N}(y_i - x_i^T w)x_i^T + 2\sum_{j=1}^{M}\Big(q_j w^T u_j u_j^T + (1-q_j)(w^T u_j u_j^T - u_j^T)\Big) \;=\; 0\,.$$

We can then rearrange the above equation as

$$\begin{aligned}\widehat{w}(y_u) &= \left(\sum_{i=1}^{N} x_i x_i^T + \sum_{j=1}^{M} u_j u_j^T\right)^{-1} \left(\sum_{i=1}^{N} y_i x_i + \sum_{j=1}^{M} y_j^u u_j\right) \\ &= \left(X_e^T X_e\right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix}.\end{aligned} \tag{3.16}$$

The second equality is obtained by noting that $\sum_{i=1}^{N} x_i x_i^T + \sum_{j=1}^{M} u_j u_j^T = X_e^T X_e$. It is then evident that $\widehat{w}(y_u)$ is the least squares solution of a supervised classification task with design matrix X_e and vector of labels $y_e = (y^T, y_u^T)^T$. Therefore, $\widehat{w}_{CPLS}$ is in the constraint set $\mathscr{C}_w$ we introduced in (3.8) independently of the labeling that is chosen in the maximization task.

3.4.3. Equivalence with the Projection Method

We can now prove that the CPLS solution as in Definition 3.3 is equivalent to the projected estimator for a particular choice of X_0, which is the matrix that defines the distance measure according to (3.9).

Proposition 3.4. *Let the data-set be represented by a design matrix $X_e \in \mathbb{R}^{(N+M)\times(D+1)}$ and a partial labeling $y \in \{0,1\}^N$. Suppose the projected classifier $\widehat{w}_{proj}$ is trained as in (3.10), where the distance $d(w, \widehat{w}_{sup})$ is defined setting $X_o = X_e$. Then both $\widehat{w}_{CPLS}$ and $\widehat{w}_{proj}$ are computed by solving the following optimization:*

$$\inf_{w\in\mathbb{R}^{D+1}} \max_{y_u\in[0,1]^M} \Big(R(w|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e)\Big), \tag{3.17}$$

hence

$$\widehat{w}_{CPLS} = \widehat{w}_{proj}.$$

Proof. First, we show that the CPLS estimate can be obtained by solving the optimization defined in Equation (3.17). To this end, we focus on the unsupervised part of $R(w, q|X_e, y)$, which can be written as

$$\sum_{j=1}^{M} \left(q_j (u_j^T w)^2 + (1-q_j)(1-u_j^T w)^2\right) = \sum_{j=1}^{M} \left((u_j^T w)^2 + (1-q_j) - 2(1-q_j) u_j^T w\right).$$

Then, we switch to $y_j^u = 1 - q_j$ in order to obtain

$$\sum_{j=1}^{M} \left((u_j^T w)^2 - 2y_j^u u_j^T w + \left(y_j^u\right)^2\right) + C = \sum_{j=1}^{M} \left(y_j^u - u_j^T w\right)^2 + C,$$

where $C = \sum_{j=1}^{M} y_j^u - \sum_{j=1}^{M} \left(y_j^u\right)^2$ is used to obtain the square loss. Note that C is present also in the unsupervised part of the risk evaluated in $\widehat{w}_{sup}$, therefore it cancels out in the expression for the contrastive risk:

$$\begin{aligned}CR(w, \mathbb{1} - y_u|\widehat{w}_{sup}, X_e, y) &= \|y - Xw\|^2 - \left\|y - X\widehat{w}_{sup}\right\|^2 + \|y_u - X_u w\|^2 - \left\|y_u - X_u \widehat{w}_{sup}\right\|^2 \\ &= \|y_e - X_e w\|^2 - \left\|y_e - X_e \widehat{w}_{sup}\right\|^2,\end{aligned}$$

where $\mathbb{1}$ is the M-dimensional vector of ones.
It then follows that the CPLS minimax problem can be expressed using the formulation of the contrastive risk as above:

$$\inf_{w\in\mathbb{R}^{D+1}} \max_{y_u\in[0,1]^M} \Big(R(w|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e)\Big).$$

Now, it remains to prove that the projection method can be written in this form as well. This task can be done by expanding the distance, which is the objective function of the problem defined in (3.10), as follows:

$$\begin{aligned} d(w,\widehat{w}_{sup})^2 &= (w-\widehat{w}_{sup})^T X_e^T X_e (w-\widehat{w}_{sup}) \\ &= \left\|X_e w\right\|^2 - 2\langle X_e w, X_e \widehat{w}_{sup}\rangle + \left\|X_e \widehat{w}_{sup}\right\|^2 \\ &= \left\|X_e w\right\|^2 - 2\langle y_e, X_e \widehat{w}_{sup}\rangle + \left\|X_e \widehat{w}_{sup}\right\|^2 \\ &= -\left\|X_e w\right\|^2 + 2\langle y_e, X_e w\rangle - 2\langle y_e, X_e \widehat{w}_{sup}\rangle + \left\|X_e \widehat{w}_{sup}\right\|^2 \\ &= \left\|y_e - X_e \widehat{w}_{sup}\right\|^2 - \left\|y_e - X_e w\right\|^2, \end{aligned}$$

where we used that $w = \widehat{w}(y_u)$. In particular, by expanding the expression for w in the third equality we took advantage of $\langle X_e w, X_e \widehat{w}_{sup}\rangle = \langle y_e, X_e \widehat{w}_{sup}\rangle$, in the second to last equality we used $\left\|X_e w\right\|^2 = \langle y_e, X_e w\rangle$, while to obtain the last equality we added and subtracted $\left\|y_e\right\|^2$.
It is now clear that the semi-supervised estimate of the projection methods can be expressed as

$$\min_{y_u \in [0,1]^M} \quad \left(\left\|y_e - X_e \widehat{w}_{sup}\right\|^2 - \left\|y_e - X_e \widehat{w}(y_u)\right\|^2\right).$$

Now we can change the sign of the objective function, hence transforming the minimum to a maximum. We can finally observe that the constraint that $w = \widehat{w}(y_u)$ is obtained by the minimization of $R(w|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e)$ with respect to w. It then follows that we can reformulate the optimization as

$$\max_{y_u \in [0,1]^M} \inf_{w \in \mathbb{R}^{D+1}} \left(\left\|y_e - X_e w\right\|^2 - \left\|y_e - X_e \widehat{w}_{sup}\right\|^2\right).$$

Note then that in Section 3.4.2 we showed that the infimum and the maximum can be swapped. Therefore, the CPLS solution and the projected estimate computed with $X_0 = X_e$ are equal. □

3.4.4. Robustness of the CPLS solution

The same theoretical result reported for the projection method can be proved by taking advantage of the CPLS formulation. This result is obvious by the equivalence between the two methods, but here we are interested in proving this property using only the definition of the CPLS approach and the theory of minimax problems. In order to do so, we first introduce the following assumption.

Assumption 3.5. *The following infimum is attained for any value of $y_u \in [0,1]^M$:*

$$\inf_{w \in \mathbb{R}^{D+1}} \left(\left\|y_e - X_e w\right\|^2 - \left\|y_e - X_e \widehat{w}_{sup}\right\|^2\right).$$

We can expect this assumption to hold by using the notion of generalized inverse, but we meant to explicitly state it because it is useful to show the non-degradation property of the CPLS estimates.

Theorem 3.6. *Let $X_e \in \mathbb{R}^{(N+M)\times(D+1)}$ be the design matrix of labeled and unlabeled observations. Let $y \in \{0,1\}^N$ be the labels on the labeled data-points and $y_e = \left(y^T, y_u^T\right)^T$ a vector of labelings on X_e. Let Assumption 3.5 hold. Denote the vector containing the true labeling of the unlabeled data-points as y_e^*. Then*

$$R(\widehat{w}_{CPLS}|X_e, y_e^*) \leq R(\widehat{w}_{sup}|X_e, y_e^*).$$

Proof. Proposition 3.4 states that the minimax problem (3.15) can be written as

$$\inf_{w \in \mathbb{R}^{D+1}} \max_{y_u \in [0,1]^M} \left(R(w|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e)\right),$$

where $R(w|X_e, y_e) = \|y_e - X_e w\|^2$, that is the risk on the full data-set with labeling y_e.
We now take advantage of Sion's minimax theorem, that is Theorem 3.2, to show that the maximum and the infimum are interchangeable and their order does not affect the solution. Furthermore, Lemma 3.1 is applicable as the infimum is attained for any value of y_u by Assumption 3.5. It then follows that a solution of the minimax problem, which is denoted by $(\widehat{w}_{CPLS}, \widehat{y}_u)$ is a saddle point, i.e.

$$CR(\widehat{w}_{CPLS}, y_u|\widehat{w}_{sup}, X_e, y) \leq CR(\widehat{w}_{CPLS}, \widehat{y}_u|\widehat{w}_{sup}, X_e, y) \leq CR(w, \widehat{y}_u|\widehat{w}_{sup}, X_e, y)$$

for any $y_u \in [0,1]^M$ and $w \in \mathbb{R}^{D+1}$. The first inequality can be rewritten as

$$R(\widehat{w}_{CPLS}|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e) \leq R(\widehat{w}_{CPLS}|X_e, \widehat{y}_e) - R(\widehat{w}_{sup}|X_e, \widehat{y}_e) \qquad \forall y_u \in [0,1]^M,$$

in which $\widehat{y}_e = (y^T, \widehat{y}_u^T)^T$. In other words, if we fix $\widehat{w}_{CPLS}$ then for any choice of labels that is not $\widehat{y}_e$ the value of the objective function decreases. The last argument that is needed to end the proof is the following:

$$R(\widehat{w}_{CPLS}|X_e, \widehat{y}_e) - R(\widehat{w}_{sup}|X_e, \widehat{y}_e) \leq 0.$$

This holds as $\widehat{w}_{CPLS}$ is the optimal vector for labels $\widehat{y}_e$ and by strict convexity it has the lowest risk. Equality holds if and only if $\widehat{w}_{sup}$ is itself the optimum.
Finally, we can conclude that

$$R(\widehat{w}_{CPLS}|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e) \leq 0 \qquad \forall y_u \in [0,1]^M,$$

where $y_e = (y^T, y_u^T)^T$. This clearly means that $\widehat{w}_{CPLS}$ has a lower risk than $\widehat{w}_{sup}$ for any labeling of the unlabeled data. As the inequality is satisfied for any labeling y_u, it holds in particular for the true labels y_e^*:

$$R(\widehat{w}_{CPLS}|X_e, y_e^*) \leq R(\widehat{w}_{sup}|X_e, y_e^*).$$

□

It is interesting to note how differently this robustness result can be proven in the two formulations. In the CPLS case it basically follows from convexity-concavity of the objective function in its two variables and by convexity and compactness of (one of) the sets where a solution is searched. On the other hand, if the method is thought in its projection formulation the result follows by properties of projections on convex subspaces of a Hilbert space. However the result is obtained, the classifier is guaranteed not to be worse than the supervised model independently of the labeling on the unlabeled data.

3.4.5. On a Condition for Strict Improvement

Theorem 3.6 ensures that the CPLS estimate is at least as good as the supervised estimate in terms of the square loss on the full data. We may observe that this does not guarantee a strict improvement and the method can indeed select $\widehat{w}_{sup}$, thus sticking to the supervised model. In the next example we qualitatively describe an example in which this happens.

Example 3.7. Assume we restrict our class of decision functions to linear models without an intercept, that is $y = wx$. In addition, suppose the feature vectors are uni-dimensional values. Figure 3.1 shows an example with these characteristics. The two labeled observations, indicated by the blue circles, define the linear classifier. Now, we add an unlabeled data point, denoted by the red circle. If this point is within the margin, as in the figure, we can observe that there is a choice of the label such that the unlabeled observation lies on the decision line. For that specific labeling, denoted by the red cross in the figure, the supervised model is retrieved. This is evident as the sum of the squared errors is null for that choice, hence that remains the optimal classifier. It is then clear that in this case the CPLS is forced to select the supervised model and cannot improve over it.

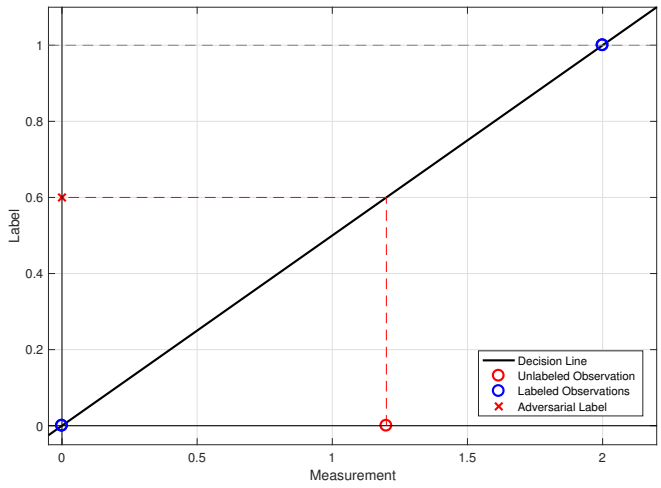


Figure 3.1: *Example with uni-dimensional features. The dashed red line identifies the most adversarial choice of the label for the unlabeled point.*

This example illustrated that it is not possible to obtain a strictly improved model if the unobserved feature is in the margin defined by the supervised classifier. Nonetheless, improvement is possible if the additional data-point is sufficiently outside of the margin. This reasoning can be easily extended to more complex models that with an intercept and multi-dimensional feature vectors. In case of several unlabeled observations it is sufficient to ask that some of the points are far enough from the margin, while the others are allowed to lie within it. In the next theorem we make this reasoning explicit and we state a condition for strict improvement in the uni-dimensional case.

Theorem 3.8. *Let the design matrix X_e be composed of N labeled and M unlabeled uni-dimensional features. Assume the linear model to be of the form $y = wx$, i.e. without intercept, and define $U = \left(\sum_{j=1}^{M} u_i^2\right)$. Moreover, assume $\widehat{w}_{sup}$ satisfies the following condition:*

$$\widehat{w}_{sup} \begin{cases} > \dfrac{\sum_{\{j:u_j>0\}} u_j}{U} & \text{if } \widehat{w}_{sup} > 0\,, \\ < \dfrac{\sum_{\{j:u_j<0\}} u_j}{U} & \text{if } \widehat{w}_{sup} < 0\,. \end{cases} \tag{3.18}$$

Then, the CPLS estimate strictly outperforms the supervised model:

$$R(\widehat{w}_{CPLS}|X_e, y_e^*) < R(\widehat{w}_{sup}|X_e, y_e^*)\,.$$

Proof. The idea is to verify whether it is possible to select q so that the supervised model is retrieved. In order to establish this possibility, we check if the risk on the full data reaches its minimum in $\widehat{w}_{sup}$ for some value of $y_u \in [0,1]^M$. By strict convexity of the empirical risk this is expressed by the following equation:

$$\left(\nabla R(w|X_e, y_e)\right)|_{w=\widehat{w}_{sup}} = 0\,.$$

Intuitively, if we can choose values of q such that $\widehat{w}_{sup}$ satisfies this condition, then it is the CPLS solution. We can now write the condition and evaluate it in $\widehat{w}_{sup}$:

$$0 + \sum_{j=1}^{M} q_i\Big(2(u_i\,\widehat{w}_{sup})u_i\Big) + (1-q_i)\Big(2(1-u_i\,\widehat{w}_{sup})u_i\Big) = 0\,,$$

where the initial 0 follows from the fact that $\widehat{w}_{sup}$ is the minimizer of the supervised risk, thus $\left(\nabla R(w|X)\right)|_{w=\widehat{w}_{sup}} = 0$. We can rearrange the equation above as

$$\sum_{j=1}^{M} u_j^2 \widehat{w}_{sup} = \sum_{j=1}^{M} (1-q_j) u_j \,. \tag{3.19}$$

If the condition holds for some choice of q, then we can retrieve $\widehat{w}_{sup}$.
Now, we can first consider the case of a positive slope coefficient of the supervised model, that is $\widehat{w}_{sup} > 0$. In this case, the term on the left hand side is always positive, while the sign of the term on the right hand side is determined by the choice of q. However, if

$$\left(\sum_{j=1}^{M} u_j^2\right) \widehat{w}_{sup} > \sum_{\{j:\, u_j > 0\}} u_j \,,$$

then there is no choice of q that satisfies equality in (3.19), hence $\widehat{w}_{sup}$ is not the CPLS solution and by Theorem 3.6 we have strict improvement for any y_u. The same line of reasoning can be applied to the opposite case, that is $\widehat{w}_{sup} < 0$, thus leading to the condition in (3.18). □

Observation 3.9. Condition (3.18) expresses the minimum slope that the supervised linear model must have in order to allow for a strictly improved semi-supervised solution. In other terms, it expresses quantitatively the intuitive reasoning we illustrated in Example 3.7. Indeed, the denominator $\left(\sum_{j=1}^{M} u_j^2\right)$ increases as unlabeled points outside of the margin get more numerous and as their distance from the margin grows. The numerator increases as well, but with a lower order. Therefore, the requirement becomes weaker as the unlabeled points follow the forementioned trend.

As a final comment to this section, we admit that the assumption that the model does not have an intercept is indeed quite strong. Nonetheless, it hints that it may indeed be possible to obtain a strictly improved model by using contrast and pessimism. An extension to multi-dimensional feature vectors is complicated from a theoretical standpoint, but empirical results in [28] and [27] suggest that a strict improvement is usually observed. This may be a consequence of the use of soft labels in our analysis, when in practice one observes hard labels only.

3.4.6. A Deviation to Non-degradation for Convex (Margin Based) Losses

Theorem 3.6 is proved by tools from minimax theory that rely mainly on the structure of the objective function, which is linear in the adversarial labels and convex in the parameter. This line of reasoning can then be extended to other settings that share these properties. We can thus use the same arguments to formally show the robustness property of the contrastive pessimistic solution for any convex loss. For instance, in Section 2.3.4 we discussed a theoretical analysis of the contrastive pessimistic approach with margin based losses, i.e. the class of losses of the form $\ell(y, f(x)) = \phi(y f(x))$. However, in [29] it is not formally proved that the contrastive pessimistic solution is robust. Here we do so by considering a general convex loss.
First, let us define the semi-supervised risk as

$$R_\ell^{semi}(w|X_e, y_e, q) = \sum_{i=1}^{N} \ell(y_i,\, w^T x_i) + \sum_{j=1}^{M} \left(q_i\, \ell(+1,\, u_i^T w) + (1-q_i)\, \ell(-1,\, u_i^T w)\right).$$

We then know that we can define the contrastive risk $CR_\ell(w, y_u|\widehat{w}_{sup}, X_e, y)$ by subtracting $R_\ell^{semi}(\widehat{w}_{sup}|X_e, y, q)$ to the previous equation. Let us now postulate the following assumption, which is an adaptation of Assumption 3.5 to general losses.

Assumption 3.10. *Suppose we are in the same setting as Assumption 3.5. Then, the following infimum is attained for any $y_u \in [0,1]^M$:*

$$\inf_{w \in \mathbb{R}^{D+1}} CR(w, y_u | \widehat{w}_{sup}, X_e, y).$$

Theorem 3.11. *Let $\ell : [0,1] \times \mathbb{R} \to \mathbb{R}_+$ be a continuous convex loss function and let Assumption 3.10 hold. Call $y_e^* \in \{0,1\}^{N+M}$ the vector of the true labels on X_e. Then the contrastive pessimistic estimate $\widehat{w}_{CPL}$ satisfies the following inequality:*

$$R_\ell(\widehat{w}_{CPL} | X_e, y_e^*) \leq R_\ell(\widehat{w}_{sup} | X_e, y_e^*).$$

Proof. Analogous to Theorem 3.6. □

In particular, this result holds for convex margin based losses. Observe however that in [29] the authors show that for decreasing margin based loss strict improvement is not possible, therefore in that case we have $\widehat{w}_{CPL} = \widehat{w}_{sup}$ independently of the data-set.

3.4.7. Quadratic Programming Formulation

It is meaningful to study whether the CPLS classifier, which we proved to have very interesting properties, can be computed efficiently. It has been shown that the minimization with respect to w can be solved first, thus resulting in the constrained structure of w as expressed in (3.16). Now, we focus on the structure of the remaining maximization of $CR(\widehat{w}(y_u), \mathbb{1} - y_u | \widehat{w}_{sup}, X_e, y)$ with respect to y_u. The following proposition states that the final task is a quadratic programming problem [7].

Proposition 3.12. *Let the data-set be X_e, y and suppose Assumption 3.5 holds. Then, the optimal labels $y_u \in [0,1]^M$ are defined by $y_j^u = v_{j+N}$ for $j = 1, \ldots, M$, where v is the solution of a quadratic programming problem:*

$$\begin{aligned} \text{minimize} \quad & v^T Q v + q^T v \\ \text{subject to} \quad & v_i = y_i \qquad \text{for } i = 1, \ldots, N \\ & 0 \leq v_j \leq 1 \qquad \text{for } j = N+1, \ldots, N+M, \end{aligned}$$

with $Q = X_e (X_e^T X_e)^{-1} X_e^T$ symmetric and semi-definite positive, $q = X_e \widehat{w}_{sup}$.

Proof. As a first step, we can rewrite the contrastive risk as

$$CR(w, \mathbb{1} - y_u | \widehat{w}_{sup}, X_e, y) = w^T X_e^T X_e w - \widehat{w}_{sup}^T X_e^T X_e \widehat{w}_{sup} - 2 \left(\sum_{i=1}^N y_i x_i^T + \sum_{j=1}^M y_j^u u_j^T \right) (w - \widehat{w}_{sup}).$$

Noting that $\sum_{i=1}^N y_i x_i^T + \sum_{j=1}^M y_j^u u_j^T = (y^T, y_u^T) X_e$, we can rewrite the former equation as

$$CR(w, \mathbb{1} - y_u | \widehat{w}_{sup}, X_e, y) = w^T X_e^T X_e w - \widehat{w}_{sup}^T X_e^T X_e \widehat{w}_{sup} - 2 \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e (w - \widehat{w}_{sup}).$$

We now take advantage of the fact that w is in $\mathscr{C}_w$:

$$\begin{aligned} CR(\widehat{w}(y_u), \mathbb{1} - y_u | \widehat{w}_{sup}, X_e, y) = & \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \left(X_e^T X_e \right)^{-1} \left(X_e^T X_e \right) \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} - \left\| X_e \widehat{w}_{sup} \right\|^2 \\ & - 2 \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} + 2 (y^T, y_u^T) X_e \widehat{w}_{sup}, \end{aligned}$$

which simplifies to

$$CR(\widehat{w}(y_u), \mathbb{1} - y_u | \widehat{w}_{sup}, X_e, y) = - \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} + 2 \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \widehat{w}_{sup} - \left\| X_e \widehat{w}_{sup} \right\|^2.$$

The norm $\left\| X_e \widehat{w}_{sup} \right\|$ is not relevant in the maximization task, hence it is omitted and the resulting optimization problem is

$$\max_{y_u \in [0,1]^M} -\begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} + 2 \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \widehat{w}_{sup} .$$

Finally, by changing the sign the maximum becomes a minimum:

$$\begin{aligned} \text{minimize} \quad & \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} - 2 \begin{pmatrix} y^T & y_u^T \end{pmatrix} X_e \widehat{w}_{sup} \\ \text{subject to} \quad & y_u \in [0,1]^M . \end{aligned}$$

Observe that $X_e \left(X_e^T X_e \right)^{-1} X_e^T$ is clearly symmetric, and also positive semi-definite as

$$z^T X_e \left(X_e^T X_e \right)^{-1} X_e^T z = \left(X_e^T z \right)^T \left(X_e^T X_e \right)^{-1} \left(X_e^T z \right) \geq 0$$

for any z by positive semi-definiteness of $X_e^T X_e$. □

The most adversarial posterior probabilities $\widehat{y}_u$ can thus be computed through a favorable optimization. The last step is then to plug the obtained value in $\widehat{w}(\widehat{y}_u)$, which is then the CPLS parameter vector of the linear classifier.

3.4.8. The Importance of the Constraint Set

The CPLS approach was shown to select the item of $\mathscr{C}_w$ that minimizes the improvement of the current estimates against $\widehat{w}_{sup}$. This can be translated to the following constrained optimization:

$$\begin{aligned} & \max_{y_u \in [0,1]^M} \left(R(\widehat{w}(y_u) | X_e, y_e) - R(\widehat{w}_{sup} | X_e, y_e) \right) \\ & \text{where} \quad \widehat{w}(y_u) = \left(X_e^T X_e \right)^{-1} X_e^T \begin{pmatrix} y \\ y_u \end{pmatrix} . \end{aligned}$$

This formulation can be studied to better understand some dynamics of the CPLS approach.

For a fixed $y_u \in [0,1]^M$, the objective function compares the risk of the best solution on that labeling with the risk of the supervised solution. Then, we can distinguish two different situations, either $\widehat{w}_{sup}$ is the optimal parameter vector for some y_u, or it is not:

- $\widehat{w}_{sup} \in \mathscr{C}_w$. By the definition of $\mathscr{C}_w$ there is at least one choice of y_u such that

  $$\widehat{w}_{sup} = \operatorname*{arginf}_{w \in \mathbb{R}^{D+1}} R(w | X_e, y_e) ,$$

  where $y_e = (y^T, y_u^T)^T$. Let us group all the vectors of labels that make this happen in a new set, which we call $\bar{\mathscr{Y}}$. It follows that for all $\bar{y}_e \in \bar{\mathscr{Y}}$

  $$R(w | X_e, \bar{y}_e) - R(\widehat{w}_{sup} | X_e, \bar{y}_e) > 0 ,$$

  for any choice of $w \in \mathbb{R}^{D+1} \setminus \widehat{w}_{sup}$. Therefore, the only possibility to satisfy the non degradation property is to select $\widehat{w}_{sup}$ as our semi-supervised estimate. Otherwise, the solution is worse than the supervised model on any labeling in $\bar{\mathscr{Y}}$.
  It follows that the CPLS estimate can at most match the performance of the supervised solution:

  $$R(w | X_e, y_e) = R(\widehat{w}_{sup} | X_e, y_e) .$$

- $\widehat{w}_{sup} \notin \mathscr{C}_w$. In this case there does not exist a labeling of the unlabeled observations that retrieves $\widehat{w}_{sup}$. By the saddle point property it follows that

$$R(\widehat{w}_{CPLS}|X_e, y_e) - R(\widehat{w}_{sup}|X_e, y_e) \leq R(\widehat{w}_{CPLS}|X_e, \widehat{y}_e) - R(\widehat{w}_{sup}|X_e, \widehat{y}_e) \qquad \forall y_u \in [0,1]^M.$$

  Note that the term on the right hand side is strictly negative because $\widehat{w}_{CPLS}$ is the chosen to be the optimal classifier on $\widehat{y}_e$ and $\widehat{w}_{sup} \neq \widehat{w}_{CPLS}$. As a consequence the term on the right hand side is itself strictly negative, hence

$$R(\widehat{w}_{CPLS}|X_e, y_e) < R(\widehat{w}_{sup}|X_e, y_e) \qquad \forall y_u \in [0,1]^M. \tag{3.20}$$

  Therefore, if $\widehat{w}_{sup} \notin \mathscr{C}_\omega$ the CPLS estimate obtains outperform the supervised model in terms of the square loss on the full data.

In conclusion, the only option when $\widehat{w}_{sup} \in \mathscr{C}_w$ is to copy the supervised solution, as there does not exist a choice of w that is always at least as good as $\widehat{w}_{sup}$ while strictly improving on a subset of labelings. On the other hand, when $\widehat{w}_{sup} \notin \mathscr{C}_w$ the CPLS approach finds a solution that is strictly better than the supervised model for any vector of labels y_u and thus also for the true labeling. We could then see the requirements in Theorem 3.8 also as a condition that makes sure that the supervised classifier is not part of the constraint set. However, requiring $\widehat{w}_{sup} \notin \mathscr{C}_w$ may be a weaker condition, so we state this result in the following proposition.

Proposition 3.13. *Suppose a data-set with design matrix $X_e \in \mathbb{R}^{(N+M)\times(D+1)}$ and a partial labeling $y \in \{0,1\}^N$ is available. In addition, suppose $\widehat{w}_{sup} \notin \mathscr{C}_w$. Then, $\widehat{w}_{CPLS}$ outperforms $\widehat{w}_{sup}$ for any labeling on the unlabeled observations $y_u \in [0,1]^M$, meaning that (3.20) is satisfied.*

3.4.9. Comparing the CPLS with the ICLS

Our interest is now in investigating and understanding why and how the choice of the metric that is used to project the solution on the constraint set affects the classifier. Recall the metric introduced in [28]:

$$\begin{aligned} d_{X_o}(w_1, w_2) &= \sqrt{(w_1 - w_2)^T X_o^T X_o (w_1 - w_2)} \\ &= \left\| X_o(w_1 - w_2) \right\|. \end{aligned}$$

As discussed in previous sections, the projection method consists in minimizing the distance to the supervised parameter vector under the constraint that the semi-supervised solution is in $\mathscr{C}_w$. In particular, in the ICLS the distance is defined by the design matrix that contains the measurements of labeled observations only, X, while the contrastive and pessimistic approach uses the full design matrix X_e. The empirical tests shown in [28] hint that the former choice shows relevant improvements over the supervised solution and seems to outperform the latter. In order to understand why this is the case, we start by observing that the minimization task can be rewritten as

$$\min_{w \in \mathscr{C}_w} \left\| X_o w - X_o \widehat{w}_{sup} \right\|.$$

If the matrix X_o contains an observation on each row, then the matrix-vector product $X_o w$ returns the outputs of the linear classifier defined by w on the corresponding observations. The same argument clearly applies to $X_o \widehat{w}_{sup}$. Since the Euclidean norm is the sum of square distances, the weight vector that solves the optimization task is then forced to generate outputs for the considered feature vectors as similarly to $\widehat{w}_{sup}$ as possible while being part of the constraint set.
The ICLS method uses the feature vectors of labeled observations only and chooses w so that its output is close to the supervised one on these observations. As a consequence, the ICLS estimate has more freedom in assigning the outputs on the unlabeled data. On the other hand, the CPLS

solution is computed with $X_o = X_e$, so $\widehat{w}_{CPLS}$ is built to have an output on the unlabeled observations that is similar to $\widehat{y}_u^{sup} = X_u \widehat{w}_{sup}$. It is thus clear that the CPLS estimate will in general be more related to $\widehat{w}_{sup}$ than the ICLS solution. This represents a limit when the unlabeled data-set carries considerable information and can thus be used to significantly improve the classifier, in which case the ICLS method can outperform the CPLS.

We have thus illustrated that the CPLS approach has a conservative nature. This is because it explicitly compares the outputs of the current estimate with those defined by the supervised model on the full design matrix X_e. The CPLS estimate is then the parameter vector that minimizes the distance from the outputs of the supervised model while being in the constraint set. While being able to obtain significant performance improvements, the CPLS approach may also result in a performance degradation in a finite sample setting. Theorems 3.6 and 3.8 suggest that this is never the case for the contrastive pessimistic approach. This approach then seems to answer our research question. Therefore, we are motivated to study the application of the notions of contrast and pessimism in other settings, such as that of generative probabilistic models.

4

Maximum Contrastive Pessimistic Likelihood Estimation with Exponential Families

We now focus on a contrastive and pessimistic semi-supervised generative classifier. A very specific classifier of this kind, i.e. the LDA, was previously studied in [33]. Here we extend the analysis to a much broader framework in which we can model the class conditional probabilities with any exponential family. This classifier is shown to have a non-degradation guarantee in Theorem 4.5. Moreover, in Section 4.2.5 we analyze the possibilities of strict performance improvement of the contrastive pessimistic generative classifier. As we will see, in this framework we can expect the semi-supervised learner to strictly outperform the supervised model under mild assumptions.
Let us start our analysis with a few fundamental results on the exponential family.

4.1. Fundamentals of Exponential Families

Exponential families are recurrent in the statistics literature because of their handy formulation and for the generalizations they provide. Several well known distributions fall into this family. For instance, the Bernoulli and the binomial probability mass functions are examples of discrete exponential families, while the Gaussian and the exponential distribution are continuous examples of this category of densities. A very useful reference for our analysis is [8], which is cited several times in this section.

4.1.1. Basic Definitions and Properties

The general definition of the exponential family is as follows.

Definition 4.1. *A family of densities* $\{p(x|\theta) : \theta \in \Theta\}$ *with respect to a measure* ν *on a probability space* $(\mathcal{X}, \mathcal{B})$ *is said to be an exponential family if it admits the following decomposition*

$$p(x|\theta) = c(x)\exp(\langle t(x), \eta(\theta)\rangle - F(\theta)),$$

where $t : \mathcal{X} \to \mathbb{R}^K$ *and* $c : \mathcal{X} \to [0, \infty)$ *are measurable, while* $\mathcal{X}$ *does not depend on* θ.

Any exponential family can be reduced to the so called canonical form. This means that the following simplified decomposition of the density holds:

$$p(x|\eta) = c(x)\exp(\langle t(x), \eta\rangle - F(\eta)), \tag{4.1}$$

where we have parametrised the family by

$$\eta = (\eta_1(\theta), \dots, \eta_K(\theta)).$$

We will mainly use the canonical form in the following discussion.
The function $t(x)$ is determined up to a multiplicative constant and is called *natural sufficient statistic*. On the other hand, $F(\eta)$ is usually referred to as *cumulant function*, or log-normalizer, and it is the logarithm of the normalizing factor. The cumulant function is then of the following form:

$$F(\eta) = \log \int_{\mathcal{X}} c(x) \exp(\langle t(x), \eta \rangle) d\nu(x).$$

It is important to define the subset of admissible parameters, which is called *natural parameter space*. This space is the subset of $\mathbb{R}^K$ where $F(\eta) < \infty$, or equivalently

$$\mathcal{N} = \left\{ \eta \in \mathbb{R}^K : \int_{\mathcal{X}} c(x) \exp(\langle t(x), \eta \rangle) d\nu(x) < \infty \right\}.$$

Another useful concept is that of minimal exponential family. We say that the family is minimal if there are no linear constraints among the components of the parameter vector or of the sufficient statistic. This means that the dimensions of the vectors involved in the scalar product cannot be reduced. The dimension of the parameter of a minimal family is then called order of the family. By Theorem 1.9 in [8] it is always possible to reduce an exponential family to a minimal family through sufficiency, reparametrization, and a proper choice of ν. Then, we can confine our analysis to minimal exponential families without loss of generality.
Finally, we discuss some properties of the cumulant function and of the parameter space. These results will be useful in the discussion of the MCPL and correspond to Theorems 1.13 and 2.2 in [8].

Theorem 4.2. *Let $\{p(x|\eta) : \eta \in \Omega\}$ be an exponential family. Then, the following statements hold:*

i. $\mathcal{N}$ is a convex set and $F(\eta)$ is a convex function on $\mathcal{N}$;

ii. If the family is minimal, then $F(\eta)$ is strictly convex on $\mathcal{N}$ and the family is identifiable, i.e. $p(x|\eta_1) = p(x|\eta_2)$ if and only if $\eta_1 = \eta_2 \in \mathcal{N}$;

iii. $F(\eta)$ is lower-semi-continuous on $\mathbb{R}^K$ and continuous on the interior of the parameter space;

iv. $F(\eta)$ is differentiable in the interior of $\mathcal{N}$ and the derivatives are obtained by differentiating under the integral sign.

Convexity of the parameter space and of the cumulant function easily follow from Holder's inequality, which for instance shows that for any $\alpha \in [0,1]$

$$\exp(F(\alpha\eta_1 + (1-\alpha)\eta_2)) \leq \exp(\alpha F(\eta_1) + (1-\alpha)F(\eta_2))$$

and by monotonicity we have convexity of the cumulant function. Similarly we can prove that $\mathcal{N}$ is convex.

4.1.2. Maximum Likelihood Estimates for Exponential Families

In Maximum Likelihood (ML) estimation the density that generates the N available observations is modeled with a parametric family, which in this section we assume to be an exponential family of the form (4.1). We are then interested in the structure of the ML estimates in this setting.

Suppose data $X = (x_1, \dots, x_N)$ are an independent, identically distributed sample from an exponential family with parameter η. In general, η is multidimensional, $\eta = (\eta_1, \dots, \eta_d) \in \Omega \subseteq \mathbb{R}^d$ for some

$d \geq 1$, and the observations can be multidimensional as well.
We then defined the likelihood function by

$$L(\eta|X) = \prod_{i=1}^{N} p(x_i|\eta) = \prod_{i=1}^{N} c(x_i)\exp(\langle t(x_i),\eta\rangle - F(\eta)).$$

We can then define the log-likelihood as

$$LL(\eta|X) = \log L(\eta|X) = \sum_{i=1}^{N} \big(\log c(x_i) + \langle t(x_i),\eta\rangle - F(\eta)\big).$$

The ML estimate is then the parameter that maximizes the log-likelihood, that is

$$\widehat{\eta}_{ML} = \underset{\eta\in\Omega}{\arg\max}\left(-NF(\eta) + \sum_{i=1}^{N}\big(\log c(x_i) + \langle t(x_i),\eta\rangle\big)\right).$$

As we discussed in Section 4.1 the cumulant function is convex, strictly if the family is minimal. Then, the log-likelihood is concave in η and the maximization task can be solved by equating the gradient of $LL(\eta|X)$ to 0. This results in the following equality

$$\nabla_\eta F(\widehat{\eta}_{ML}) = \frac{1}{N}\sum_{i=1}^{N} t(x_i). \tag{4.2}$$

Note that $\nabla_\eta F(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and this is thus a set of d equations. The ML estimator can then be found by inverting equations (4.2), which is possible by (strict) convexity of the cumulant function:

$$\widehat{\eta}_{ML} = \big(\nabla_\eta F\big)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} t(x_i)\right), \tag{4.3}$$

where by $\big(\nabla_\eta F\big)^{-1}$ we denote the inverse function of the gradient.
A precise result on the invertibility of the cumulant function is Corollary 2.5 in [8].

We conclude this section with a few observations. First, we observe that an exponential family can be parametrised with the so called mean parametrization, which is defined as $\mu = \nabla_\eta F(\eta)$. It is clear from the results above that the mean parametrization plays an important part in ML estimation for any exponential family. For some exponential families this parametrization is very similar or equal to the natural parametrization, but this is not the case for others classes of distributions. It is then interesting to note that the estimate depends on the sample only through the sufficient statistic. In addition, the derivatives of the cumulant function are the cumulants of the sufficient statistic. This means that the first derivative is the mean of $t(X)$, while the second derivative is the covariance. In mathematical terms:

$$\begin{aligned}\nabla_\eta F(\eta) &= \mathbb{E}\big(t(X)\big),\\ \nabla^2_\eta F(\eta) &= \mathrm{Cov}\big(t(X)\big).\end{aligned}$$

In (4.2) the ML estimate for the mean of the sufficient statistic is then the empirical mean computed on the sample.

4.2. MCPL Estimation for Safe Semi-Supervised Learning

We are now going to take advantage of the properties of exponential families that were introduced in Section 4.1 to build a semi-supervised generative classifier that safely incorporates an unlabeled data-set. This means that in this approach we are interested in estimating the joint density of (x, y),

Notation	Meaning
N	number of labeled observations
M	number of unlabeled observations
K	number of classes
N_k	number of labeled observations in class k
$k(i)$	mapping to the true index of the i-th labeled observation
$\mathscr{D}_l, \mathscr{D}_u$	labeled/unlabeled data-set
$\mathscr{D}^*_{full}$	true full data-set
U	matrix of unlabeled feature vectors
Δ_K	K-dimensional probability simplex
$t_k(\cdot)$	sufficient statistic of class k
$F_k(\cdot)$	cumulant function of class k
Ω_k	parameter space of the statistical model for the k-th class
$\eta_k \in \Omega_k$	parameter vector of the statistical model for the k-th class
$\mathscr{P}_k$	parametric family for the k-th class conditional distribution
$p_k(\cdot\|\eta_k) \in \mathscr{P}_k$	probability density function for class k
$\pi_k \in \Delta_K$	prior probability of class k
Ψ	vector containing the parameters of all classes
$\widehat{\Psi}_{sup}$	supervised estimates Ψ
$\widehat{\pi}_k^{sup}, \widehat{\eta}_k^{sup}$	supervised estimates of π_k, η_k
$\widehat{\Psi}_{semi}$	MCPL semi-supervised estimate of Ψ
$\widehat{\pi}_k^{semi}, \widehat{\eta}_k^{semi}$	MCPL semi-supervised estimates of π_k, η_k
$q_{\cdot j}$	vector of posterior probabilities for j-th unlabeled observation
Q	matrix containing all the vectors of posterior probabilities $q_{\cdot j}$
$\widehat{Q}_{semi}$	adversarial posterior probabilities chosen by the MCPL
$L(\Psi, Q\|\mathscr{D}_l, U)$	semi-supervised log-likelihood
$CL(\Psi, Q\|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$	contrastive log-likelihood
$\nabla_{q_{\cdot j}} CL(\cdot, \cdot\|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$	gradient of the contrastive likelihood with respect to $q_{\cdot j}$
$L(\Psi\|\mathscr{D}_{full})$	standard log-likelihood on a full data-set $\mathscr{D}_{full}$

Table 4.1: *Table of notation for the MCPL approach.*

where x is a measurement and y is the corresponding label. This is done by first learning a parametric model for the class conditional distributions and estimating the prior probabilities of observing a data-point from a specific class. The decision boundary is then defined by assigning an observation to the class that assigns the highest estimated posterior probability to that object, as discussed in Section 2.1.1. In this section we define a framework that allows the user to model each class conditional distribution with a parametric class that admits the decomposition in the exponential form. We first introduce the notation, which is recapitulated in Table 4.1, and we discuss the formulation the method.

4.2.1. Formulation of the Method

Assume there are N labeled observations $\mathscr{D}_l = \{(x_i, y_i)\}_{i=1}^N$, in which $x_i \in \mathbb{R}^D$ and $y_i \in \{1, \dots, K\}$ for each $i = 1, \dots, N$. This defines a K-class classification problem. Suppose M additional unlabeled feature vectors $U = \{u_i\}_{j=1}^M$ are observed, with $u_j \in \mathbb{R}^D$ for every $j = 1, \dots, M$. Now, let us model each class conditional distribution with a parametric class that is a minimal exponential family. This restriction can be done without loss of generality as discussed in Section 4.1. In practice we want to find for each class k the best estimate within a set $\mathscr{P}_k = \{p_k(x|\eta_k) : \eta_k \in \Omega_k\}$, which satisfies Definition 4.1 and is minimal. This means the structure of the densities is the following

$$p_k(x|\eta_k) \;=\; c_k(x) \exp\left(\langle t_k(x), \eta_k \rangle - F_k(\eta_k)\right),$$

where Ω_k is the parameter space of class k, i.e. $\eta_k \in \Omega_k$. The probability mass function of the random variable y is defined by $\pi_k = P(y = k)$, where clearly $\sum_{k=1}^{K} \pi_k = 1$.
It is now necessary to define the supervised classifier, because as we know it is used as benchmark in the MCPL approach. To this end we introduce the log-likelihood on the labeled observations

$$\begin{aligned} L(\Psi|\mathscr{D}_l) &= \sum_{i=1}^{N} \log p(x_i, y_i|\eta_{y_i}, \pi_{y_i}) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log p(x_{k(i)}, k|\eta_k, \pi_k) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log\left(\pi_k \, p_k(x_{k(i)}|\eta_k)\right). \end{aligned} \tag{4.4}$$

Here $\Psi = (\pi_1, \ldots, \pi_K, \eta_1, \ldots, \eta_K)$ contains all the parameters of the model, while the function $k(i) : \mathbb{N} \to \mathbb{N}$ is a map to the index of the i-th element in class k, and N_k is the number of labeled observations from class k. We follow the maximum likelihood approach and select the estimate as

$$\widehat{\Psi}_{sup} = \underset{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K}{\arg\max} \; L(\Psi|\mathscr{D}_l), \tag{4.5}$$

where $\Delta_K = \left\{(\pi_1, \ldots, \pi_K) : \sum_{i=1}^{K} \pi_i = 1, \pi_i \geq 0 \quad \forall i \in [K]\right\}$ is the K-dimensional simplex. As discussed in Section 4.1.2, the supervised estimates for the parameters of each class k are of the following form:

$$\widehat{\eta}_k^{sup} = \left(\nabla F_k\right)^{-1}\left(\frac{1}{N_k}\sum_{i=1}^{N_k} t_k\left(x_{k(i)}\right)\right),$$

while it is easy to show that the estimates for the prior probabilities are

$$\widehat{\pi}_k^{sup} = \frac{N_k}{N}.$$

Note that the estimated prior probabilities sum to 1: $\sum_{k=1}^{K} \widehat{\pi}_k^{sup} = \sum_{k=1}^{K} \frac{N_k}{N} = 1$, because $\sum_{k=1}^{K} N_k = N$.
Now, we add the unlabeled data U in order to define the semi-supervised classifier. The labels corresponding the the M measurements u_j are unknown, but it is still possible to add a term that is the expectation of the log-likelihood, where the randomness lies in the labels. This term is defined as

$$\begin{aligned} \mathbb{E}_{y_1,\ldots,y_M}\left(\sum_{j=1}^{M} \log p(u_j, y_j|\eta_{y_j}, \pi_{y_j}) \middle| (u_1, \ldots, u_M)\right) &= \sum_{j=1}^{M} \mathbb{E}_{y_j|u_j}\left(\log p(u_j, y_j|\eta_{y_j}, \pi_{y_j})\right) \\ &= \sum_{j=1}^{M} \sum_{k=1}^{K} q_{kj} \log\left(\pi_k p_k(u_j|\eta_k)\right), \end{aligned}$$

where we used linearity of the expectation, while $q_{kj} = \mathbb{P}(y_j = k|u_j)$ are the true posterior probabilities. The *semi-supervised log-likelihood* is then defined as

$$L(\Psi, Q|\mathscr{D}_l, U) = L(\Psi|\mathscr{D}_l) + \sum_{j=1}^{M} \sum_{k=1}^{K} q_{kj} \log\left(\pi_k p_k(u_j|\eta_k)\right), \tag{4.6}$$

in which we grouped all the conditional probabilities in Q. Clearly, we do not have access to the true underlying distribution, so the correct values in Q are unavailable. Note also that $q_{\cdot j} \in \Delta_K$ for $j = 1, \ldots, M$.
The MCPL approach has three main aspects that drive its formulation. The first characteristic is that

the method is contrastive, as it explicitly contrasts the semi-supervised estimates with the corresponding supervised ones, i.e. $\widehat{\Psi}_{sup}$. This comparison is in terms of log-likelihood on the complete data-set and it is done by subtraction. For this purpose, we define the *contrastive log-likelihood* as

$$CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) = L(\Psi, Q|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U).$$

The contrastive log-likelihood allows us to control and quantify the improvement over the supervised estimates. At this point a pessimistic choice is made in order to prepare our classifier even to the most adversarial choice of the simplices in Q. This translates into a minimization task over all possible choices of conditional probabilities:

$$CPL(\Psi|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) = \min_{Q \in \Delta_K^M} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U),$$

where by Δ_K^M we mean the Cartesian product between M simplices. Note that we can talk about minimum because the optimization is on a compact space and the contrastive log-likelihood is continuous in each element of Q.
The final step to complete the formulation of the MCPL approach is to maximize the modified likelihood that we have built. This means that the semi-supervised estimates are computed by maximizing $CPL(\Psi|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ over the set of parameters.

Definition 4.3. *Let $\mathscr{D}_l$ contain the labeled observations with their corresponding labels and U contain the unlabeled feature vectors. Let $\widehat{\Psi}_{sup}$ be the supervised estimate computed on $\mathscr{D}_l$ as in (4.5). The maximum contrastive pessimistic likelihood estimate is*

$$\widehat{\Psi}_{semi} := \operatorname*{arg\,sup}_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} CPL(\Psi|\widehat{\Psi}_{sup}, \mathscr{D}_l, U). \tag{4.7}$$

It is clear from Definition 4.3 that in order to find the MCPL solution we have to solve the following minimax problem

$$\sup_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} \min_{Q \in \Delta_K^M} \left\{ L(\Psi, Q|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U) \right\}, \tag{4.8}$$

where $L(\Psi, Q|\mathscr{D}_l, U)$ was defined in (4.6). This shows that the semi-supervised estimates are obtained by maximizing the likelihood on the most adversarial soft labels.
In the next section we show that the MCPL solution is under a mild assumption guaranteed not to result in a performance degradation.

4.2.2. Robustness of the MCPL Solution

In Section 4.2.1 the MCPL estimate is introduced as the solution of the minimax problem (4.8). The objective function of the optimization is the contrastive log-likelihood, which in the case of exponential families has an attractive form that allows to prove strong theoretical results. In order to have a better understanding, we first write the semi-supervised log-likelihood in the following meaningful form:

$$\begin{aligned} L(\Psi, Q|\mathscr{D}_l, U) = & \sum_{k=1}^{K} \left(\left(\sum_{i=1}^{N_k} \langle t_k(x_{k(i)}), \eta_k \rangle + \log c_k(x_{k(i)}) \right) - N_k F_k(\eta_k) + N_k \log \pi_k \right) \\ & + \sum_{k=1}^{K} \sum_{j=1}^{M} q_{kj} \left(\langle t_k(u_j), \eta_k \rangle + \log c_k(u_j) - F_k(\eta_k) + \log \pi_k \right). \end{aligned} \tag{4.9}$$

By subtraction we can then explicitly write the contrastive log-likelihood $CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ similarly. It is worth noting that the measures $c_k(\cdot)$ appear in both the semi-supervised log-likelihoods

in an identical fashion, thus they are not part of the contrastive log-likelihood.
Now, we are most interested in a guarantee that, independently of which are the true labels on the unlabeled observations, the proposed semi-supervised solution does not lead to a degraded performance. In this sense the semi-supervised learner is said to be *safe*. The most natural choice of performance measure in this setting is the log-likelihood on the full data-set. It is then necessary to introduce the true complete data:

$$\mathscr{D}^*_{full} = \mathscr{D}_l \cup \left\{(u_j, y^*_j)\right\}_{j=1}^{M},$$

in which y^*_j are the true, unknown labels. In the next theorem, we state and prove that the MCPL estimate is guaranteed to be a safe classifier in terms of the likelihood on $\mathscr{D}^*_{full}$. In order to obtain this result it is necessary to first postulate a requirement on the data.

Assumption 4.4. *For any* $Q \in \Delta_K^M$, $\sup_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ *is attained.*

This assumption is necessary in order to prove the robustness of the MCPL estimates. We will discuss its meaning and what it actually requires after the proof of the following theorem.

Theorem 4.5. *Let each class conditional distribution be modeled by a parametric class of densities* $\{p_k(x|\eta_k) : \eta_k \in \Omega_k\}$ *that is a minimal exponential family. Let* $\mathscr{D}^*_{full}$ *be the true complete data-set. Assume the data-set is such that Assumption 4.4 is satisfied. Then, the semi-supervised estimates* $(\widehat{\Psi}_{semi}, \widehat{Q}_{semi})$ *obtained by the MCPL approach are a saddle point of the minimax problem (4.8). Moreover,* $\widehat{\Psi}_{semi}$ *is guaranteed not to result in a performance degradation:*

$$L(\widehat{\Psi}_{semi}|\mathscr{D}^*_{full}) \geq L(\widehat{\Psi}_{sup}|\mathscr{D}^*_{full}). \tag{4.10}$$

Proof. The robustness guarantee can be proved with a two step argument. First, we apply Theorem 3.2 to show that the order in which the supremum and the minimum are taken does not influence the solution. Lemma 3.1 then implies that MCPL solution is a saddle point. The definition of saddle point finally gives us the needed property.
First, we verify that Theorem 3.2 can be applied to the minimax problem defined in (4.8). We start by checking that the assumptions on the sets that define the optimization are satisfied:

- the supremum is on $\Delta_K \times \Omega_1 \times \ldots \times \Omega_K$ and each set in the Cartesian product is convex. This follows from the convexity of the probability simplex and of the natural parameter spaces of each exponential family (see Theorem 4.2). The Cartesian product of convex spaces is itself convex;
- the infimum is on Δ_K^M, which again is a Cartesian product of convex spaces and thus itself convex. Moreover, this set closed and bounded, hence it is compact.

Then, the requirements on the objective function $CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ need to be checked:

- $CL(\cdot, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ is upper-semi-continuous and strictly concave, as by Theorem 4.2 each cumulant function is lower-semi-continuous and strictly convex and $\log(\pi_k)$ is concave;
- $CL(\Psi, \cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ is linear and thus continuous and convex.

Convexity and continuity are stronger conditions than quasi-convexity and semi-continuity, thus all the requirements are met. Therefore, we can apply Sion's minimax theorem to state that

$$\sup_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} \min_{Q \in \Delta_K^M} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) = \min_{Q \in \Delta_K^M} \sup_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U).$$

Assumption 4.4 is now necessary to have the first two conditions in Lemma 3.1. Indeed, the assumption that the supremum is attained for any choice of Q, together with the minimax equality, allows us to state that there exists a solution $\left(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}\right)$ that solves the minimax independently of the order of the optimization. Then by Lemma 3.1 the solution of the minimax problem is a saddle point that gives the saddle value:

$$CL(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) = \max_{\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K} \min_{Q \in \Delta_K^M} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U).$$

In order to prove (4.10) we use the fact that $(\widehat{\Psi}_{semi}, \widehat{Q}_{semi})$ is saddle point:

$$CL(\Psi, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \leq CL(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \leq CL(\widehat{\Psi}_{semi}, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \tag{4.11}$$

for any $\Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K$ and any $Q \in \Delta_K^M$. The first inequality is trivial as $\widehat{\Psi}_{semi}$ is the best estimate for the labeling $\widehat{Q}_{semi}$. On the other hand, the second inequality can be rearranged to state that

$$L(\widehat{\Psi}_{semi}, Q|\mathscr{D}_l, U) \geq L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U) + \left(L(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, \widehat{Q}_{semi}|\mathscr{D}_l, U)\right)$$

$\forall Q \in \Delta_K^M$. This implies that

$$L(\widehat{\Psi}_{semi}, Q|\mathscr{D}_l, U) \geq L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U) \qquad \forall Q \in \Delta_K^M, \tag{4.12}$$

as the term in the parentheses is non-negative again because $\widehat{\Psi}_{semi}$ is the maximizing choice for the labeling $\widehat{Q}_{semi}$, thus results in a larger log-likelihood.
Finally, the true labels of the unlabeled observations are certainly included in Δ_K^M, so we can in particular conclude that

$$L(\widehat{\Psi}_{semi}|\mathscr{D}^*_{full}) \geq L(\widehat{\Psi}_{sup}|\mathscr{D}^*_{full}),$$

which is the robustness guarantee we needed to conclude the proof. □

Observation 4.6. In the proof above we showed that the solution of the minimax task is a saddle point, as showed in (4.11). By strict concavity of the log-likelihood we can conclude that the following inequality is strict

$$L(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\mathscr{D}_l, U) > L(\Psi, \widehat{Q}_{semi}|\mathscr{D}_l, U) \qquad \forall\, \Psi \in \Delta_K \times \Omega_1 \times \ldots \times \Omega_K \setminus \widehat{\Psi}_{semi}.$$

Observation 4.7. We can define the so called *oracle solution* as the optimal estimate on the full data-set, that is $\widehat{\Psi}_{opt} = \arg\max_{\Psi} L(\Psi|\mathscr{D}^*_{full})$. The log-likelihood on the full data evaluated in $\widehat{\Psi}_{opt}$ then serves as an upper bound to the log-likelihood of the MCPL semi-supervised estimate:

$$L(\widehat{\Psi}_{opt}|\mathscr{D}^*_{full}) \geq \quad L(\widehat{\Psi}_{semi}|\mathscr{D}^*_{full}) \qquad L(\widehat{\Psi}_{sup}|\mathscr{D}^*_{full}).$$

It is clear that a consequence of the strict concavity of the log-likelihood is the following equivalence

$$L(\widehat{\Psi}_{opt}|\mathscr{D}^*_{full}) \geq L(\widehat{\Psi}_{semi}|\mathscr{D}^*_{full}) \quad \Longleftrightarrow \quad \widehat{\Psi}_{opt} = \widehat{\Psi}_{semi}.$$

The condition on the right is satisfied if and only if the labeling that solves the minimax turns out to be the true one. This would represent a fortunate coincidence, thus the oracle solution intuitively will remain an unreachable benchmark.

The MCPL solution is guaranteed not to produce a performance degradation as long as Assumption 4.4, or equivalently Assumption 4.8, holds. In other words, the information contained in the unlabeled observations is carefully used to produce a safe semi-supervised learner. However, the estimates $\widehat{\Psi}_{semi}$ can still in principle result in no improvement at all. We will address this issue extensively in Section 4.2.5.

4.2.3. A Reformulation of the Assumption

Theorem 4.5 illustrates the desirable robustness of the MCPL approach. As we have seen, Assumption 4.4 plays a crucial role in the proof above and it is crucial to conclude that the MCPL estimates are a saddle point, which in turn means that $\widehat{\Psi}_{semi}$ performs at least as good as $\widehat{\Psi}_{sup}$ for any soft labeling of the unlabeled observations. From a practical point of view Assumption 4.4 makes sure that the log-likelihood is well behaved and that taking the supremum does not result in infinite values. Note that a sufficient condition so that the assumption is satisfied is that each parameter space is compact. If that is the case, we know that a concave function attains a maximum on a compact set. However, the parameter space of an exponential family is in general not compact. We could still interpret it as such if the data are reasonable, in the sense that the parameter space could be restricted to a compact set without excluding any value of interest. Anyway, here we reformulate Assumption 4.4 in order to express it in a more comprehensible form.

Let us analyze the contrastive log-likelihood in more detail. We can immediately see that the supremum affects only $L(\Psi, Q|\mathscr{D}_l, U)$, so the term $L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U)$ can be ignored. We can then concentrate on (4.9) only, which is a sum of several terms. The prior probabilities can be isolated and their estimates are the solutions of the following supremum

$$\sup_{(\pi_1,\dots,\pi_K)\in\Delta_K} \sum_{k=1}^{K} \left(N_k \log(\pi_k) + \sum_{j=1}^{M} q_{kj} \log(\pi_k) \right).$$

An explicit solution to this maximization can be found in (4.15), from which it is clear that the supremum is attained if $N_k \geq 1$ for any $k = 1,\dots,K$. Note that that supervised estimates $\widehat{\eta}_k^{sup}$ would not be defined if this was not true. We can then ignore this term in the following.
The remaining part of the semi-supervised log-likelihood can be seen as a sum of K independent optimization tasks:

$$\sum_{k=1}^{K} \sup_{\eta_k\in\Omega_k} \left(\langle \sum_{i=1}^{N_k} t_k(x_{k(i)}) + \sum_{j=1}^{M} q_{kj}\, t_k(u_j), \eta_k \rangle - \left(N_k + \sum_{j=1}^{M} q_{kj} \right) F_k(\eta_k) \right), \tag{4.13}$$

where the terms that include the carrier measures $c_k(x)$ are not included because they have no influence on the solution of the supremum.
It is then clear that Assumption 4.4 is equivalent to asking that each independent supremum (4.13) is attained. In other terms, this means that for any $k = 1,\dots,K$ the parameters of each exponential family can be estimated by maximizing a modified log-likelihood, which can be defined for any k as

$$L_k(\eta_k, q_{k\cdot}|\mathscr{D}_l, U) \;=\; \sum_{i=1}^{N_k} \log p_k(x_{k(i)}|\eta_k) + \sum_{j=1}^{M} q_{kj} \log p_k(u_j|\eta_k),$$

where $q_{k\cdot} = \left(q_{k1},\dots,q_{kM}\right) \in [0,1]^M$ is a vector containing the posterior probabilities of assigning each unlabeled observation to class k.
We can then reformulate Assumption 4.4 as follows:

Assumption 4.8. *For any value of Q, each supremum in (4.13) is attained, i.e. for any class k and for any $Q \in \Delta_K^M$*

$$\exists \widehat{\eta}_k(q_{k\cdot}) \in \Omega_k : \qquad L_k(\widehat{\eta}_k(q_{k\cdot}), q_{k\cdot}|\mathscr{D}_l, U) \;=\; \sup_{\eta_k\in\Omega_k} L_k(\eta_k, q_{k\cdot}|\mathscr{D}_l, U).$$

We have thus shown the following result.

Proposition 4.9. *Assume $N_k \geq 1$ for any $k = 1,\dots,K$. Then Assumption 4.8 is equivalent to Assumption 4.4.*

This kind of hypotheses are common in standard maximum likelihood estimation, where L_k is the sum of the logarithms of the density evaluated in each observation. In particular, the ML estimator is consistent under the assumption that the model is correct and that the supremum is attained. We can then expect Assumption 4.8 to be satisfied in common real world applications. Nonetheless, the assumption can be verified in practice for instance by checking that each $-L_k(\widehat{\eta}_k(q_{k\cdot}), q_{k\cdot}|\mathscr{D}_l, U)$ is coercive for any $Q \in \Delta_K^M$. In that case the supremum is guaranteed to be attained.

4.2.4. Solving the Minimax Problem

In the proof of Theorem 4.5 we showed that we can apply Sion's minimax theorem to show that the optimization task (4.8) satisfies the minimax equality. We are thus allowed to interchange the supremum with the minimum without affecting the result. We can take advantage of this result to define a strategy to solve the minimax problem, while understanding what is the structure of the MCPL estimates. We then focus on the following formulation of the optimization task:

$$\min_{Q\in\Delta_K^M} \max_{\Psi\in\Delta_K\times\Omega_1\times\ldots\times\Omega_K} \left(L(\Psi, Q|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U)\right),$$

in which we talk about maximum because we assume the supremum to be attained, i.e. either Assumption 4.4 or Assumption 4.8 hold.

The maximization task is affected by the first log-likelihood only, as $L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U)$ does not depend on Ψ. Therefore, we can first focus on solving

$$\max_{\Psi\in\Delta_K\times\Omega_1\times\ldots\times\Omega_K} L(\Psi, Q|\mathscr{D}_l, U), \tag{4.14}$$

where $L(\Psi, Q|\mathscr{D}_l, U)$ was explicitly defined in Equation (4.9). The objective function is strictly concave in Ψ for any $Q \in \Delta_K^M$ and thus we expect to have a unique solution. The estimates for the prior probabilities are computed by a constrained optimization, the constraint being $\sum_{k=1}^K \widehat{\pi}_k = 1$. The resulting estimates are

$$\widehat{\pi}_k(q_{k\cdot}) = \frac{N_k + \sum_{j=1}^M q_{kj}}{N+M} \qquad \text{for } k = 1,\ldots,K. \tag{4.15}$$

The constraint is satisfied as $\sum_{k=1}^K \sum_{j=1}^M q_{kj} = \sum_{j=1}^M \left(\sum_{k=1}^K q_{kj}\right) = M$ and again $\sum_{k=1}^K N_k = N$. Note also that it is clear from this expression that the observations are partially assigned to potentially more than one class by the soft labels.

Now, we can estimate the parameters that define each class conditional distribution simply by equating the gradient of the semi-supervised log-likelihood to 0. This is possible as the cumulant function is differentiable in the interior of the parameter space by Theorem 4.2 and leads to the following equations

$$\nabla_{\eta_k} F_k(\widehat{\eta}_k) = \frac{1}{N_k + \sum_{j=1}^M q_{kj}} \left(\sum_{i=1}^{N_k} t_k(x_{k(i)}) + \sum_{j=1}^M q_{kj}\, t_k(u_j)\right) \qquad \text{for } k = 1,\ldots,K. \tag{4.16}$$

For each class, we have then defined a number of equations that is equal to the order of the corresponding exponential family. It is interesting to compare the MCPL estimates with the supervised ones, which are of the form introduced in Equation (4.2). We can clearly see once again that the unlabeled feature vectors appear in the weighted sum preceded by their corresponding soft labels. It is important to note that, as we assumed the exponential families to be minimal, the cumulant function is strictly convex. As a consequence, its gradient is strictly increasing and we can uniquely determine the MCPL estimate either by (4.16) or by inverting it:

$$\widehat{\eta}_k\left(q_{k\cdot}\right) = \left(\nabla F_k\right)^{-1} \left(\frac{1}{N_k + \sum_{j=1}^M q_{kj}} \left(\sum_{i=1}^{N_k} t_k(x_{k(i)}) + \sum_{j=1}^M q_{kj}\, t_k(u_j)\right)\right) \tag{4.17}$$

for $k = 1, \dots, K$.
The last optimization problem we need to solve is then the minimum over Q, that is the Cartesian product of M simplices:

$$\min_{Q \in \Delta_K^M} \left(L(\widehat{\Psi}(Q), Q | \mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, Q | \mathscr{D}_l, U) \right). \tag{4.18}$$

This is a constrained optimization task that can be solved numerically. Note that the space on which the optimum is searched is compact and grows with the number of classes and with the number of unlabeled observations. It is not immediately clear how the objective function and its derivatives behave, as this depends on the values that the sufficient statistics take on the observations.
As we can see from Equations (4.15), (4.16) the choice of the posterior probabilities determines the MCPL estimates. If we call $\widehat{Q}_{semi}$ the solution of the minimum (4.18), then the MCPL estimates are

$$\begin{aligned} \widehat{\pi}_k^{semi} &= \widehat{\pi}_k(\widehat{q}_{k\cdot}^{semi}) \qquad \text{for } k = 1, \dots, K, \\ \widehat{\eta}_k^{semi} &= \widehat{\eta}_k\left(\widehat{q}_{k\cdot}^{semi}\right) \qquad \text{for } k = 1, \dots, K. \end{aligned} \tag{4.19}$$

The MCPL solution is then the oracle if the labels that are selected in the minimization (4.18) are the true ones, while in any other case by Theorem 4.5 we know the MCPL can never result in a performance degradation.

4.2.5. Conditions for Strict Performance Improvement

In this section we investigate under which assumptions we can expect the MCPL estimates to result in a strict performance improvement. The following theorem represents a first step in this direction. We take advantage of Equation (4.16) to show that the MCPL estimates are almost surely different from the supervised ones if we choose the class conditional distributions from the same family.

Theorem 4.10. *Let all the assumptions in Theorem 4.5 hold, where in addition the class conditional distributions are modeled with the same minimal exponential family, which is assumed to have an injective sufficient statistic. Fix a labeled data-set $\mathscr{D}_l$ on which the supervised solution is computed as in (4.5). Suppose the observed unlabeled feature vectors $U = \{u_j\}_{j=1}^M$ are a sample from a continuous underlying density $p(x)$. Then, the semi-supervised MCPL solution $\widehat{\Psi}_{semi}(U)$ built using the unlabeled data-set is almost surely different from the supervised estimate:*

$$\mathbb{P}_{u_1, \dots, u_M \sim p(x)} \left(\widehat{\Psi}_{semi}(U) = \widehat{\Psi}_{sup} \right) = 0.$$

Proof. First, observe that the assumption that each class conditional distribution is modeled by the same exponential family means in particular that $t_k(\cdot) = t(\cdot)$ for any k. In Section 4.1.2 we discussed that an exponential family can be parametrised with the so called mean parametrization, which is defined by $\mu = \nabla F(\eta) = \mathbb{E}\big(t(X)\big)$. The contrastive pessimistic estimate of this parameter can be obtained for each class by (4.16), as $\widehat{\mu}_k^{semi} = \nabla_{\eta_k} F_k(\widehat{\eta}_k^{semi})$. Then, we can estimate the average of this parameter over the classes with a weighted sum, where the weights are the estimates of the prior probabilities:

$$\widehat{\mu}_{semi}(U) = \sum_{k=1}^{K} \widehat{\pi}_k^{semi} \widehat{\mu}_k^{semi}.$$

This can be rewritten as

$$\begin{aligned}\widehat{\mu}_{semi}(U) &= \sum_{k=1}^{K}\left(\frac{N_k+\sum_{j=1}^{M}\widehat{q}_{kj}^{semi}}{N+M}\cdot\frac{\sum_{i=1}^{N_k}t(x_{k(i)})+\sum_{j=1}^{M}\widehat{q}_{kj}^{semi}t(u_j)}{N_k+\sum_{j=1}^{M}\widehat{q}_{kj}^{semi}}\right)\\ &= \frac{1}{N+M}\sum_{k=1}^{K}\left(\sum_{i=1}^{N_k}t(x_{k(i)})+\sum_{j=1}^{M}\widehat{q}_{kj}^{semi}t(u_j)\right)\\ &= \frac{1}{N+M}\left(\sum_{i=1}^{N}t(x_i)+\sum_{j=1}^{M}t(u_j)\right),\end{aligned}$$

where in the first equality we used Equations (4.15) and (4.16), while in the last equality we took advantage of the fact that $t(u_j)$ does not depend on k and then that $\sum_{k=1}^{K}\widehat{q}_{kj}^{semi}=1$.
It is now clear that $\widehat{\mu}_{semi}(U)$ is independent of the posterior probabilities Q and depends only on the data. On the other hand, the supervised optimal estimates are such that $\nabla_{\eta_k}F_k(\widehat{\eta}_k^{sup})=\frac{1}{N}\sum_{i=1}^{N}t(x_i)$. Note that there is no randomness in this estimate because we consider the labeled data-set to be fixed. With computations similar to the ones above we can show that the supervised estimate for the average mean parameter is

$$\widehat{\mu}_{sup} = \sum_{k=1}^{K}\widehat{\pi}_k^{sup}\widehat{\mu}_k^{sup} = \frac{1}{N}\sum_{i=1}^{N}t(x_i)\,.$$

Now, observe that the sufficient statistic has a continuous distribution since the data are assumed to be generated by a continuous density. Then by setting $\widehat{\mu}_{semi}(U)=\widehat{\mu}_{sup}$ we obtain the following condition

$$\sum_{j=1}^{M}t(u_j) = \frac{M}{N}\sum_{i=1}^{N}t(x_i)\,, \tag{4.20}$$

which almost surely does not hold because we assumed the sufficient statistic to be injective and distribution of the feature vectors $u_1,\dots,u_M$ to be continuous.
The final argument is that equality of the MCPL estimates and the supervised estimates, i.e. $\widehat{\Psi}_{semi}=\widehat{\Psi}_{sup}$, implies equality of the estimated average mean parameters. In other words

$$\{U\in\mathbb{R}^d:\widehat{\Psi}_{semi}(U) = \widehat{\Psi}_{sup}\}\subseteq\{U\in\mathbb{R}^d:\widehat{\mu}_{semi}(U)=\widehat{\mu}_{sup}\}\,.$$

However, we have just seen that $\widehat{\mu}_{semi}(U)\neq\widehat{\mu}_{sup}$ almost surely. This in turn implies that $\widehat{\Psi}_{semi}(U)\neq\widehat{\Psi}_{sup}$ almost surely, which is the thesis. □

It is important to observe that it is a common choice to use the same statistical model for all the classes. This framework for instance includes two very well known classifiers, such as LDA and QDA, in which the parametric family is that of Gaussian densities. In Chapter 5 we explicitly derive the MCPL estimates for both learners.
Another important assumption is that the sufficient statistic of the chosen exponential family is injective. This again is a very mild assumption and is in general verified.
We have thus shown that the MCPL approach almost surely produces a different estimate than the supervised one under the forementioned choice of the statistical model. In particular, we consider the randomness to be in the unlabeled observations, while the labeled data-set is considered to be fixed. We may consider the labeled data-set as random as well, but the result remains unchanged.

We now discuss an important consequence of this property in the following corollary of Theorems 4.5 and 4.10. Theorem 4.10 in particular was the missing piece to prove that the MCPL results almost surely in a strict improvement in terms of the log-likelihood on the full data-set. To this end we introduce the complete additional data-set $\mathscr{D}_u=\{(u_j,v_j)\}_{j=1}^{M}$, in which we included the hidden labels associated with each feature vector u_j. We can now state the result.

Corollary 4.11. *Let the assumptions of Theorem 4.10 be verified. Assume that the labels corresponding to the unlabeled observations are generated from $p(y|x)$. Define the complete data-set as $\mathscr{D}_{full} = \mathscr{D}_l \cup \mathscr{D}_u$.*
Then, the MCPL solution results almost surely in a strict performance improvement:

$$\mathbb{P}_{\mathscr{D}_u \sim p(x,y)} \left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) > L(\widehat{\Psi}_{sup}|\mathscr{D}_{full}) \right) = 1. \tag{4.21}$$

Proof. In Theorem 4.5 we showed that the solution of the minimax problem $(\widehat{\Psi}_{semi}, \widehat{Q}_{semi})$ is a saddle point of the contrastive log-likelihood. In particular, this implies that

$$CL(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \leq CL(\widehat{\Psi}_{semi}, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \qquad \forall Q \in \Delta_K^M. \tag{4.22}$$

Now, we know by Theorem 4.10 that $\mathbb{P}_{u_1,\dots,u_M \sim p(x)} \left(\widehat{\Psi}_{semi}(U) = \widehat{\Psi}_{sup} \right) = 0$, which means that adding a random unlabeled sample implies that the MCPL solution is almost surely different than $\widehat{\Psi}_{sup}$. In addition, the fact that the exponential family is minimal means that the semi-supervised log-likelihood is strictly concave. Therefore, there exists a unique optimal solution for each set of posterior probabilities, that is equivalent to the following inequality

$$L(\widehat{\Psi}_{semi}(U), \widehat{Q}_{semi}|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, \widehat{Q}_{semi}|\mathscr{D}_l, U) > 0 \qquad \text{a.s.}$$

In other terms, the contrastive likelihood for the labeling $\widehat{Q}_{semi}$ is almost surely strictly positive. Then using this and (4.22) we have that

$$\mathbb{P}_{u_1,\dots,u_M \sim p(x)} \left(CL(\widehat{\Psi}_{semi}(U), Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) > 0 \qquad \forall Q \in \Delta_K^M \right) = 1.$$

Noting again that Δ_K^M contains all possible hard labels, we can conclude that the improvement is strict however the labels are chosen. Therefore, we might as well take them to be randomly generated from $p(y|x)$, thus having a probability with respect to the joint density, which is our thesis. □

We have then proved that for continuous data the semi-supervised solution obtained by the MCPL approach strictly improves the supervised model for any labeling Q if we model all the class conditional distributions with the same exponential family. The improvement is then almost sure with respect to the joint distribution $p(x, y)$, which generates the unlabeled observations. In other words, there is zero probability of drawing a sample from $p(x, y)$ which does not imply a strict performance improvement. Therefore, the unlabeled data are guaranteed to have the desired effect, that is to add useful information to the model. This is a very strong guarantee which shows that the MCPL approach succeeds in finding a safe semi-supervised learner in this setting under very mild assumptions on the data-set.
A similar result can be proved for the case in which $\mathscr{D}_u$ is drawn from a discrete distribution. In this setting we have to settle for a strict improvement in expectation.

Theorem 4.12. *Let Assumption 4.8 hold. Suppose the class conditional distributions are modeled with the same exponential family, which has injective sufficient statistic $t(\cdot)$. Assume we observe a labeled data-set $\mathscr{D}_l$ and its corresponding supervised solution $\widehat{\Psi}_{sup}$. Suppose then to draw M additional unlabeled feature vectors U. In addition, assume the underlying density that generates U to be discrete.*
Then, the following strict inequality holds:

$$\mathbb{E}_{\mathscr{D}_u \sim p(x,y)} \left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) \right) > \mathbb{E}_{\mathscr{D}_u \sim p(x,y)} \left(L(\widehat{\Psi}_{sup}|\mathscr{D}_{full}) \right). \tag{4.23}$$

Proof. Proceeding as in the proof of Theorem 4.10, we define the supervised and semi-supervised mean parameters $\widehat{\mu}_{sup}$ and $\widehat{\mu}_{semi}$. With the same steps we arrive at Equation (4.20). Now, the distribution of the data is assumed to be discrete, so also the sufficient statistic has a discrete distribution and so does $\sum_{j=1}^{M} t(u_j)$. There may exist a sample U that is observed with strictly positive probability that satisfies (4.20), so we cannot rule this event out. However, there are in general several possible values that the sufficient statistic can take with non-zero probability. Moreover, adding M evaluations of the sufficient statistic implies a number of values that are taken from $\sum_{j=1}^{M} t(u_j)$ with non-zero probability. We can then conclude that, defining the scalar δ as

$$\delta \;=\; \mathbb{P}_{u_1,\dots,u_M\sim p(x)}\left(\sum_{j=1}^{M} t(u_j) = \frac{M}{N}\sum_{i=1}^{N} t(x_i)\right),$$

we have

$$\mathbb{P}_{u_1,\dots,u_M\sim p(x)}\left(\widehat{\Psi}_{semi}(U) \neq \widehat{\Psi}_{sup}\right) \;\geq\; 1-\delta\,.$$

The inequality sign follows from the fact that (4.20) is only a necessary condition, thus it can be that $\widehat{\Psi}_{semi}(U) \neq \widehat{\Psi}_{sup}$ even if (4.20) holds. In addition, it is clear by the precedent remark that $\delta < 1$.
But we have showed in the proof of Corollary 4.11 that we have strict improvement if the semi-supervised and the supervised solution are different. This means that

$$\mathbb{P}_{\mathscr{D}_u\sim p(x,y)}\left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) > L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})\right) \;\geq\; 1-\delta\,.$$

Now, we know that the complementary event of $L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) > L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})$ is

$$L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) \;=\; L(\widehat{\Psi}_{sup}|\mathscr{D}_{full}),$$

thus in expectation we have

$$\mathbb{E}_{\mathscr{D}_u\sim p(x,y)}\left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) - L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})\right) \;>\; 0,$$

which is our thesis. □

We have then shown that in the discrete case we can expect a strict improvement in expectation. This result could be expressed in terms of the probability δ that the sample satisfies (4.20). However, this probability is unknown. We could also formulate the result as

$$\mathbb{P}_{\mathscr{D}_u\sim p(x,y)}\left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) > L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})\right) \;>\; 0\,,$$

which is meaningful because otherwise we have that $L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) = L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})$.

We have thus proved that if we use the same statistical model for all classes we expect that the MCPL approach takes advantage of the unlabeled observations successfully. Now, we focus on the case of different statistical models for the class conditional distributions and we investigate whether we can draw similar conclusions.
The proof of Corollary 4.11 illustrates that in order to have strict improvement we just need to be sure that the MCPL and the supervised solutions are different. An immediate result is that given a data-set of labeled and unlabeled observations we can check whether $\widehat{\Psi}_{semi}$ and $\widehat{\Psi}_{sup}$ are equal to determine if we have strict improvement. We state this result in the following proposition.

Proposition 4.13. *Let Assumption 4.8 hold. Assume each class conditional distribution is modeled with a minimal exponential family. Call $\mathscr{D}_{full} = \mathscr{D}_l \cup \mathscr{D}_u$ the full data-set and $\widehat{\Psi}_{semi}$ the estimates obtained through the MCPL approach. It then follows that*

$$\mathbb{P}_{\mathscr{D}_u\sim p(x,y)}\left(L(\widehat{\Psi}_{semi}(U)|\mathscr{D}_{full}) > L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})\,\middle|\,\widehat{\Psi}_{semi}(U) \neq \widehat{\Psi}_{sup}\right) \;=\; 1\,.$$

It is interesting to note that the condition $\widehat{\Psi}_{semi} \neq \widehat{\Psi}_{sup}$ can be checked without any knowledge of the true labels. Hence, we know if our semi-supervised estimates strictly outperform the supervised model as soon as the MCPL estimates have been computed.
Clearly, Proposition 4.13 is a weaker result than for instance Corollary 4.11, as in principle $\widehat{\Psi}_{semi}$ might be equal to the supervised solution, thus leading to no improvement at all. However, setting the estimator $\widehat{\mu}_{semi}$ equal to $\widehat{\mu}_{sup}$ in the general case results in the following necessary condition:

$$\frac{1}{N}\sum_{i=1}^{N} t_{y_i}(x_i) \;=\; \frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{K} \widehat{q}_{kj}^{\,semi}\, t_k(u_j)\,,$$

where we assume that the sufficient statistics have the same dimensionality. In this case we cannot get rid of the probabilities $\widehat{q}_{kj}^{\,semi}$ as we did in Corollary 4.11. Note that the posterior probabilities computed with the MCPL approach depend on the unlabeled data, thus making the characterization complicated. It is thus unclear whether this equality holds in general since there is not a closed form solution for every term $\widehat{q}_{kj}^{\,semi}$. Nevertheless, we can obtain a simple sufficient condition for strict improvement as

$$\nexists Q \in \Delta_K^M : \qquad \frac{1}{N}\sum_{i=1}^{N} t_{y_i}(x_i) \;=\; \frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{K} q_{kj}\, t_k(u_j)\,. \tag{4.24}$$

In this sense, it is possible to obtain strict improvement also if one chooses to use different statistical models for different classes. It has to be noted that this is a rather strict requirement. However, this is only a sufficient condition and thus strict improvement can be obtained even if (4.24) does not hold.

In conclusion, in the case of a common statistical model we proved with Corollary 4.11 that the improvement happens almost surely in the continuous case and in expectation in the discrete setting. Then, we have discussed two conditions for strict improvement in case more than one statistical model is used. These conditions are (4.24) and the requirement that $\widehat{\Psi}_{semi} \neq \widehat{\Psi}_{sup}$. In both cases we know the improvement is strict and thus the semi-supervised learner is indeed better for any labeling of the unlabeled data. In other terms, it is possible to build a classifier that uses safely the available unlabeled data and in some cases is guaranteed to strictly outperform the supervised model. Such strong results are even more important because they do not require any assumption on the structure of the data, unlike the large majority of semi-supervised algorithms.

4.2.6. The Constraint Set

In Chapter 3 we introduced the set of all optimal solutions for different values of the labels on the unlabeled observations. This was called the *constraint set*. It became clear that this set, which is here denoted by $\mathscr{C}$, is crucial in determining whether we can expect that the semi-supervised estimate results in a strict performance improvement, or if the contrastive pessimistic approach returns the supervised model. In the present chapter, we moved from discriminative models to generative models. The constraint set maintains its relevance also in this framework, as we now show.
In order to understand this, recall that in Section 4.2.4 we derived the structure of the MCPL estimates by solving the maximum before the minimum. From this derivation, and in particular from Equations (4.15), (4.16), we can see that the MCPL approach can be interpreted in terms of $\mathscr{C}$. First, let us define a labeled data-set as $\mathscr{D}_l$. We can then formally define the constraint set $\mathscr{C} \subset \Delta_K \times \Omega_1 \times \ldots \times \Omega_K$ as follows

$$\mathscr{C} := \left\{ \Psi : \Psi = \widehat{\Psi}\left(Q\right), \quad Q \in \Delta_K^M \right\}, \tag{4.25}$$

where the expressions for $\widehat{\Psi}\left(Q\right)$ are as in Equations (4.15) and (4.17). These expressions are thus what we find by solving the maximization task for a fixed Q.

We can formulate the MCPL approach as

$$\min_{Q \in \Delta_K^M} \left(L(\widehat{\Psi}(Q), Q|\mathscr{D}_l, U) - L(\widehat{\Psi}_{sup}, Q|\mathscr{D}_l, U) \right), \tag{4.26}$$

which means that we look for a solution in $\mathscr{C}$ that solves the minimum. The result is very similar to what we showed in Chapter 3. Once again, this formulation is similar to the method introduced in [26], which is the LDA version of the Implicitly Constrained Least Squares classifier. In [26] the objective function is the supervised log-likelihood, similarly to the least squares case. Therefore, the same comments that we made for that case hold in this setting.
Now, the objective function of the minimization (4.26) is non-negative for any Q by strict convexity of the log-likelihood and since $\widehat{\Psi}(Q)$ is the optimal estimate for Q. As a consequence if $\widehat{\Psi}_{sup}$ is in the constraint set, then the objective function is null and thus minimized. This means that the minimizing choice of Q is the one that retrieves $\widehat{\Psi}_{sup}$. It is clear that the objective function is strictly positive if $\widehat{\Psi}_{sup} \notin \mathscr{C}$.
This intuitive reasoning is used to prove the following corollary, which is a direct consequence of Theorem 4.10 and Equation (4.25).

Corollary 4.14. *Let the class conditional distributions be modeled with the same exponential family and assume a labeled data-set $\mathscr{D}_l$ is available. Denote the supervised solution by $\widehat{\Psi}_{sup}$. Call U the matrix containing the additional unlabeled feature vectors and suppose Assumption 4.8 is satisfied. In addition, assume the unlabeled observations to be drawn from a continuous density $p(x)$. Then*

$$\mathbb{P}_{u_1,\dots,u_M \sim p(x)} \left(\widehat{\Psi}_{sup} \notin \mathscr{C}(U) \right) = 1. \tag{4.27}$$

Proof. By Theorem 4.10 we know that under the same assumptions on the underlying distribution and on the statistical model the MCPL estimate is almost surely not equal to the supervised one. Reasoning by contradiction, suppose that $\widehat{\Psi}_{sup} \in \mathscr{C}$. This means that $\exists \tilde{Q} : \widehat{\Psi}_{sup} = \widehat{\Psi}(\tilde{Q})$. In other words $\widehat{\Psi}_{sup}$ is the optimal solution for the choice of labels in $\tilde{Q}$, that is

$$\widehat{\Psi}_{sup} = \underset{\Psi \in \Delta_K \times \Omega_1 \times \dots \times \Omega_K}{\arg\max} L(\Psi, \tilde{Q}|\mathscr{D}_l, U).$$

Moreover, the optimum is unique by strict concavity of the semi-supervised log-likelihood. But this would mean that $L(\widehat{\Psi}_{semi}, \tilde{Q}|\mathscr{D}_l, U) < L(\widehat{\Psi}_{sup}, \tilde{Q}|\mathscr{D}_l, U)$, which contradicts Theorem 4.5 and in particular Equation (4.12). The supervised solution is then almost surely not contained in the constraint set. □

We have thus shown that we can interpret the MCPL using the constraint set, similarly to the case of least squares classification.

4.2.7. Theoretical Analysis of the Adversarial Posterior Probabilities

In Section 4.2.4 we solved first the maximum over the parameters, thus obtaining the structure of the MCPL estimates for fixed values of the posterior probabilities Q. The final estimates $\widehat{\Psi}_{semi}$ are then selected by the optimization with respect to Q. It is then evident that the adversarial probabilities $\widehat{Q}_{semi}$ play a key role in the choice of the MCPL estimates. For this reason, we now focus on this particular aspect of the proposed method from a theoretical point of view. In particular, we investigate how the adversarial probabilities are chosen, if they are in general hard labels or soft labels and when we fall in either case. The ultimate goal here is to have a deeper understanding of the method. Our findings are then tested in Section 4.3, in which we analyze in detail some simple examples in order to clarify our theoretical analysis.

We have previously highlighted that, after having solved the maximization, the adversarial posterior probabilities can be chosen by the following optimization:

$$\widehat{Q}_{semi} = \underset{Q\in\Delta_K^M}{\arg\min}\, CL(\widehat{\Psi}(Q), Q\,|\,\widehat{\Psi}_{sup}, \mathscr{D}_l, U).$$

Unfortunately, this formula does not clarify what values of Q we might expect. Indeed, the expression of $CL(\widehat{\Psi}(Q), Q\,|\,\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ is very complicated and no conclusions on convexity nor on the position of the optimal Q can be drawn for general cases. We thus approach the problem from a different perspective, that is the following:

$$\widehat{Q}(\Psi) = \underset{Q\in\Delta_K^M}{\arg\min}\, CL(\Psi, Q\,|\,\widehat{\Psi}_{sup}, \mathscr{D}_l, U).$$

In this case we know that by definition $CL(\Psi, Q\,|\,\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ is linear in each posterior probability. We can thus imagine Ψ to be fixed and focus on solving the minimization.
The contrastive likelihood can be thought to be divided in two parts. The first part is the difference on the labeled data $\mathscr{D}_l$ between the log-likelihood defined by Ψ and the one defined by the $\widehat{\Psi}_{sup}$. In this analysis this term is constant and represents a bias term that does not influence the minimization. The second part is the following:

$$\sum_{j=1}^{M}\sum_{k=1}^{K} q_{kj}\left(\langle t_k(u_j), \eta_k - \widehat{\eta}_k^{sup}\rangle - \left(F_k(\eta_k) - F_k(\widehat{\eta}_k^{sup})\right) + \log\left(\frac{\pi_k}{\widehat{\pi}_k^{sup}}\right)\right). \tag{4.28}$$

Note that the order of the summations is on purpose inverted compared to previous sections. In this second term we clearly have linearity in each posterior probability. Therefore, the first term is ignored in what follows and we can consider the objective function of the minimization task to be (4.28).
We are then interested in the gradient of $CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ in direction q_{kj}. This is clearly the slope coefficient:

$$\begin{aligned}\frac{\partial CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)}{\partial q_{kj}} &= \langle t_k(u_j), \eta_k - \widehat{\eta}_k^{sup}\rangle - \left(F_k(\eta_k) - F_k(\widehat{\eta}_k^{sup})\right) + \log\left(\frac{\pi_k}{\widehat{\pi}_k^{sup}}\right) \\ &= \log\left(\pi_k\, p(u_j|\eta_k)\right) - \log\left(\widehat{\pi}_k^{sup}\, p(u_j|\widehat{\eta}_k^{sup})\right)\end{aligned} \tag{4.29}$$

for any $k = 1,\dots,K$ and $j = 1,\dots,M$. The second equality can be obtained by adding and subtracting $c_k(u_j)$. The gradient in direction q_{kj} is then the difference between the log-likelihoods with parameters Ψ_k and $\widehat{\Psi}_k^{sup}$ evaluated in u_j. Moreover, we can observe that at this stage the posterior probabilities q_{kj} of each observation are chosen independently as they do not directly influence each other. The consequence is then that the minimization can be separated in M independent tasks, one for each unlabeled observation. We take advantage of this and we focus on a fixed observation with index $j \in \{1,\dots,M\}$. Then, the gradient with respect to the posterior probabilities $q_{\cdot j} = (q_{1j},\dots,q_{Kj})$ is

$$\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) = \begin{pmatrix} \frac{\partial CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)}{\partial q_{1j}} \\ \vdots \\ \frac{\partial CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)}{\partial q_{Kj}} \end{pmatrix},$$

which by (4.29) is a constant vector because, as we said previously, we consider Ψ to be fixed.
Now, the set on which the minimum is searched in is the probability simplex corresponding to the j-th unlabeled observation $\Delta_K = \left\{q_{kj} : \sum_{k=1}^K q_{kj} = 1,\, q_{kj} \geq 0\right\}$. We can encode this constraint by

computing the projection of $\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ on this set. This allows us to understand in what direction the contrastive likelihood increases in this subspace and thus to identify where we can expect to find the optimum. The simplex defines a bounded region in the hyper-plane that is defined by the normal vector $n = \mathbf{1}_K$, which is the K-dimensional vector having each component equal to one. We can then divide the gradient in the sum of two components that are respectively orthogonal and parallel to the simplex:

$$\begin{aligned}\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) &= \left(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)\right)_{\perp} n \\ &+ \left(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)\right)_{\parallel} v,\end{aligned}$$

where vector v is such that $\langle v, n\rangle = 0$, with $v \neq 0$. The component of the gradient that is orthogonal to the simplex can be calculated by projecting and normalizing:

$$\begin{aligned}\left(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)\right)_{\perp} &= \frac{\langle \nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U), n\rangle}{||n||^2} \\ &= \frac{1}{K}\sum_{k=1}^{K}\Bigg\{\langle t_k(u_j), \eta_k - \widehat{\eta}_k^{sup}\rangle - \left(F_k(\eta_k) - F_k(\widehat{\eta}_k^{sup})\right) \\ &+ \log\left(\frac{\pi_k}{\widehat{\pi}_k^{sup}}\right)\Bigg\},\end{aligned}$$

where we used that $\|n\|^2 = \sum_{k=1}^{K} 1 = K$.
It immediately follows that the component of the gradient that is parallel to the simplex can be obtained subtracting the orthogonal part:

$$\begin{aligned}\left(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)\right)_{\parallel} v &= \nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \\ &- \left(\frac{\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)\cdot \mathbf{1}_K}{K}\right)\mathbf{1}_K.\end{aligned}$$

In other words, the mean value among the K classes is subtracted to each component of the gradient.
We are then interested in understanding what are the possible scenarios and what type of solutions we might expect in each case. For instance, we can observe that Ψ can in principle be such that the gradient results to be orthogonal to the simplex, in which case its parallel component is null and hence we have that for each k

$$\frac{\partial CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)}{\partial q_{kj}} = \frac{\sum_{k=1}^{K}\left(\langle t_k(u_j), \eta_k - \widehat{\eta}_k^{sup}\rangle - \left(F_k(\eta_k) - F_k(\widehat{\eta}_k^{sup})\right) + \log\left(\frac{\pi_k}{\widehat{\pi}_k^{sup}}\right)\right)}{K}. \tag{4.30}$$

Note that this defines a set of $K-1$ linearly independent conditions, while the K-th can always be retrieved by summing the other $K-1$. It follows that the parameters are allowed to vary on a subspace and there is an infinite number of possible solutions.
In addition, Equation (4.30) implies that the gradient has the same value for each of the K classes. We can rewrite this as

$$\log(\pi_k\, p_k(u_j|\eta_k)) - \log(\widehat{\pi}_k^{sup}\, p_k(u_j|\widehat{\eta}_k^{sup})) = C \qquad \forall k = 1, \ldots, K, \tag{4.31}$$

where C is the average value among the classes. Basically, the values of the joint probabilities defined by Ψ evaluated at u_j are just a shifted version of the supervised ones, where the shift in terms

of log-likelihoods is C. It is evident that in this case the choice of $q_{\cdot j}$ is irrelevant. Indeed, the effect on the objective function is the same independently of the choice of $q_{\cdot j}$ because all the terms in the summation are equal and the weights must sum to one. Therefore, any $q_{\cdot j} \in \Delta_K$ is the solution of the minimization task if Ψ satisfies (4.31). Note that this event occurs for instance when $\widehat{\Psi}_{semi} = \widehat{\Psi}_{sup}$, in which case $C = 0$ and every component of the gradient is null.

Now, we assume that the projection of the gradient onto the simplex is not null in order to analyze the further possible scenarios. We have observed that the gradient is a constant vector, thus its projection in the direction that is parallel to the simplex is constant as well. This means that the solution of the minimization is the last point of the simplex that we find going in the opposite direction of $(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U))_{\parallel} v$, that is the direction in which the objective function decreases. The consequence of this is that the optimal posterior probabilities have to be a point on the boundary of the simplex, which is either a vertex or an edge. These points have a clear interpretation in terms of probabilities. A vertex is defined as a point where one of the K posterior probabilities is equal to one, while all the others are null. On the other hand, an edge is a sub-simplex, that is an L-dimensional simplex, with $1 < L < K$, in which all the mass is divided among L of the K classes. The edges can then be defined as

$$\Delta_L^I = \left\{ (q_1, \dots, q_K) : \; q_i = 0 \text{ if } i \in I, \; \sum_{\substack{i=1 \\ i \notin I}}^{K} q_i = 1, \; q_i \geq 0 \quad \forall i \notin I \right\},$$

where $I \subset \{1, \dots, K\}$ is a set that contains the indices to which is assigned null posterior probability. Note that $L = K - |I|$, where $|I|$ is the number of indices in the set. For any choice of I we define a corresponding unit norm vector that is orthogonal to the sub-simplex Δ_L^I, while being parallel to Δ_K and pointing towards the interior of the K-dimensional simplex. We call each of these vectors v_I, where I represents the edge to which the vector is orthogonal. These vectors play an important role in determining whether the solution is an edge or a vertex. Indeed, if the projected gradient is parallel to a vector v_I and points to the same direction, then the minimization task is solved by any element of the sub-simplex Δ_L^I. In that case the whole edge is on a level curve of the objective function, thus all the points of the edge are valid solutions. It is then worth investigating when this event happens. In order to answer to this question, suppose that there exists a set of indices I such that (4.31) is satisfied for any $k \in I$, that is the gradient has equal values in components identified by I. We refer to the common value of the components in positions I as C_I. Moreover, assume that the following condition holds

$$C_I < \log(\pi_l \, p_l(u_j|\eta_l)) - \log(\widehat{\pi}_l^{sup} \, p_l(u_j|\widehat{\eta}_l^{sup})) \qquad \forall \, l \notin I.$$

There are then $|I|$ equal components that are strictly smaller than the other $L = K - |I|$ ones. Clearly, the objective function is then minimum for any choice of $q_{\cdot j} \in \Delta_L^I$. We have then showed that the solution is any element of Δ_L^I if the components at the positions I are tied as smallest in the gradient. In this case it makes no difference how the mass is distributed among the indices in I.

The projected gradient can also be neither null nor parallel to any of the vectors v_I. It is then clear that the solution has to be a vertex of the K-dimensional simplex. Note that a vertex can be seen as a degenerate simplex of dimension one, so this case can in principle be included in the previous setting. If we fall in this situation, the optimal choice of posterior probabilities is a vector of the standard basis, i.e. $\widehat{q}_{\cdot j} = e_k$ for some k. In this case the difference of log-likelihoods is minimized by a single class and thus all the mass is placed on the corresponding class.

We have thus shown the three scenarios that can happen when solving the minimization task. It is now helpful to recap which is the solution to the minimization task for different values of Ψ:

- if Ψ is such that $(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U))_{\parallel} v \perp \Delta_K$, then the solution is any $q_{\cdot j} \in \Delta_K$;

- if Ψ is such that $(\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U))_{\|} \, v \parallel v_I$ for some set of indices $I \subset \{1,\dots,K\}$, then the solution is any $q_{\cdot j} \in \Delta_L^I$, where $L = K - |I|$;
- otherwise, the solution is a vector of the standard basis e_k for some $k \in \{1,\dots,K\}$.

We have limited our analysis to the posterior probabilities that a specific observation is in each class, but as we have observed previously each minimum can be solved separately at this stage. As a consequence, this analysis applies independently to each of the M minimization tasks. It is thus clear that in order to have soft labels associated to a specific unlabeled observation there are orthogonality conditions that must be satisfied. Conversely, hard labels are picked if these conditions do not hold. It remains however unclear from this analysis whether all these scenarios can possibly take place in practice. Intuitively one could expect the adversarial probabilities to follow in some sense the supervised model. In other terms, we could imagine that the unlabeled observations that are predominantly assigned to one class by the supervised learner are then tagged with a large adversarial probability by the MCPL approach. On the other hand, the orthogonality conditions we have inferred seem to be quite strict and thus hardly satisfied. These intuitive arguments are the object of scrutiny in the next section, where we use a series of illustrative toy problems.

To conclude our discussion, we point out that the findings of this section are insightful also from a different perspective. Consider the inequality that represents a part of the definition of a saddle point:

$$CL(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \;\leq\; CL(\widehat{\Psi}_{semi}, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \qquad \forall Q \in \Delta_K^M.$$

We proved that this inequality holds for any data-set and any choice of the statistical models in Theorem 4.5. First of all, the reader can convince himself that our analysis of the minimization task is perfectly coherent with this property. Indeed, $\widehat{Q}_{semi}$ can be interpreted as the minimizer of $CL(\widehat{\Psi}_{semi}, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ independently of the particular choice of $\widehat{\Psi}_{semi}$. However, if we assume that for at least one unlabeled observation we have $\widehat{q}_{\cdot j}^{\,semi} \in \Delta_L^I$ then it follows from our precedent analysis that

$$CL(\widehat{\Psi}_{semi}, \widehat{Q}_{semi}|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \;=\; CL(\widehat{\Psi}_{semi}, Q^*|\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \qquad \forall Q^* \in \mathscr{S},$$

where the set $\mathscr{S}$ is defined as

$$\mathscr{S} = \left\{ (q_{\cdot 1}, \dots, q_{\cdot M}) \in \Delta_K^M : \quad q_{\cdot j} = \widehat{q}_{\cdot j}^{\,semi} \;\text{ if }\; \widehat{q}_{\cdot j}^{\,semi} \in \{0,1\}^M, \;\text{ for } j = 1,\dots,M \right\}$$

In other words, the value of the contrastive likelihood is not affected by moving on that sub-simplex, so we cannot expect uniqueness of the saddle value. This result is caused by the fact that the contrastive likelihood is not strictly convex in Q, thus it is possible that this inconvenience takes place. On a more qualitative level, in this case there are several values that true posterior probabilities can take while minimizing the performance improvement of the semi-supervised learner. It is however worth observing that not all the points $(\widehat{\Psi}_{semi}, Q^*)$ are saddle points, as there is no reason why $\widehat{\Psi}_{semi}$ should be the optimum for any point in the sub-simplex. This claim can be justified from a theoretical standpoint using Lemma 3.1. Indeed, either the first and the second conditions are verified for the pair $(\widehat{\Psi}_{semi}, \widehat{Q}_{semi})$, while

$$Q \neq \underset{Q \in \Delta_K^M}{\arg\min}\, CL(\widehat{\Psi}(Q), Q\,|\,\widehat{\Psi}_{sup}, \mathscr{D}_l, U) \qquad Q \in \mathscr{S} \setminus \widehat{Q}_{semi}$$

unless the following holds

$$\exists \bar{Q} \in \mathscr{S} \setminus \widehat{Q}_{semi} : \qquad \widehat{\Psi}_{semi} = \widehat{\Psi}(\bar{Q}).$$

It can be seen from Equations (4.15) and (4.16) that $\widehat{\Psi}(Q)$ is in general not constant on $\mathscr{S}$, thus showing that not all the points in $\mathscr{S}$ are saddle points. This however does not exclude that there is

more than one saddle point. Curiously, this analysis shows that performing first the optimization with respect to the variable in which the function is not strictly convex leads to the entire set of points where the saddle value is attained, while the opposite order brings us to the set of saddle points.

4.3. Empirical Study of the Adversarial Posterior Probabilities

The MCPL approach has been so far studied from a theoretical perspective. In particular, it was shown to have interesting properties of robustness and strict performance improvement. In the previous section we discussed a key aspect of the method, that is the choice of the posterior probabilities Q. These values determine the semi-supervised estimates and hence the decision surface and performance of the method. It is thus worth investigating more in depth this aspect of the proposed approach.
A theoretical analysis in this direction was conducted in Section 4.2.7, where we showed the conditions on Ψ that lead to a hard or soft labeling of the unlabeled data-points. However, it remains unclear in practice when the method falls in either case. It is possible to obtain some insights by solving a series of simple problems from an empirical point of view. Here focus in particular on the choice of the adversarial posterior probabilities, but occasionally we refer to other properties that were discussed in Section 4.2. In order to have simple computations, the univariate Gaussian distribution is used throughout this section. We start with the easiest case, then increasingly adding some parameters, classes or observations.

4.3.1. A Set of Simple Examples

Suppose there are two classes and one labeled observation for each class. We assume the variance of each class conditional distribution to be unitary so that it does not have to be estimated. This setting is illustrated graphically in Figure 4.1, where the labeled observations are $x_1 = -1$, $x_2 = 1$ and the resulting supervised estimates for the means are used to define the class conditional distributions. We can then add an unlabeled observation u and see how different values of it influence the

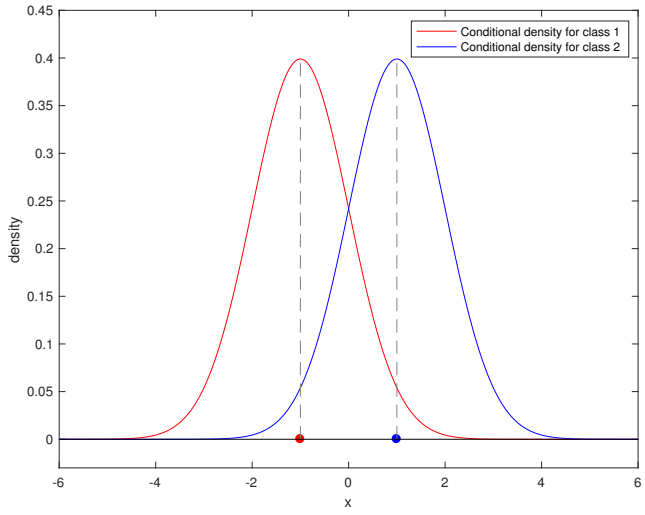


Figure 4.1: *Plot of the supervised class conditional distributions in the setting described Section 4.3.1*

MCPL estimates. Note that $\widehat{\mu}_1^{sup}$, $\widehat{\mu}_2^{sup}$ are the empirical means computed with the labeled observations. Therefore, in this simple setting we have that $\widehat{\mu}_1^{sup} = x_1 = -1$ and $\widehat{\mu}_2^{sup} = x_2 = 1$ and the estimates for the prior probabilities are $\widehat{\pi}_1^{sup} = \widehat{\pi}_2^{sup} = 1/2$. This framework is usually called nearest

mean classifier (NMC). It is understandable that the NMC gets its name from the fact that a new observation is assigned to the class that has the nearest mean, as the classes are equiprobable and the variances are equal.
On the other hand, the MCPL semi-supervised estimates have the following structure

$$\widehat{\pi}_1(q_1) = \frac{N_1+q_1}{N+1}, \qquad \widehat{\pi}_2(q_1) = \frac{N_2+(1-q_1)}{N+1}, \tag{4.32}$$

and

$$\widehat{\mu}_1(q_1) = \frac{x_1+q_1u}{N_1+q_1}, \qquad \widehat{\mu}_2(q_1) = \frac{x_2+(1-q_1)u}{N_2+(1-q_1)}, \tag{4.33}$$

where in this example $N_1 = N_2 = 1$ are the number of labeled observations for each class, $N = N_1 + N_2 = 2$, and we used that $q_2 = 1 - q_1$ to reduce the number of unknowns while including the constraint. We will focus more specifically on Gaussian Discriminant Analysis in Chapter 5, but for now it is sufficient to know that these formulas follow from Equations (4.15) and (4.16). In other terms, solving the maximization results in the optimal estimates $\widehat{\Psi}(q_1)$, which are a function of q_1. We take this for granted and we focus on the issues described above.
Now, we have discussed in Section 4.2.7 that the gradient of the contrastive likelihood is the quantity that regulates the type of optimal probabilities that solve the minimization for a fixed Ψ. In particular, we showed that we can expect soft probabilities if and only if $\nabla_{q_{\cdot j}} CL(\Psi, Q|\widehat{\Psi}_{sup}, \mathscr{D}_l, U)$ is orthogonal to the simplex, while if this is not the case then the result are hard labels. In this case there are two classes and one unlabeled observation, which means that there is only one linearly independent equation that determines the orthogonality. This is

$$\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_1 = \left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_2,$$

where $\Psi = (\mu_1, \mu_2, \pi_1, \pi_2)$, $q_\cdot = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$ are the posterior probabilities that u is respectively from class 1 or class 2, while by $(\cdot)_i$ we denote the i-th component of the vector in brackets. It is easy to show that for $k = 1, 2$ we have

$$\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_k = \log\left(\frac{\pi_k}{\widehat{\pi}_k^{sup}}\right) + \frac{1}{2}\left(-(u-\mu_k)^2 + (u-\widehat{\mu}_k^{sup})^2\right),$$

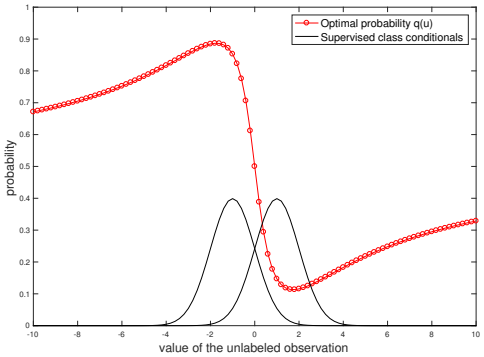
by which it follows that the equation that has to be satisfied in order to have orthogonality to the simplex is

$$\log\left(\frac{\pi_1}{\pi_2}\right) + \frac{1}{2}\left((u-\mu_2)^2 - (u-\mu_1)^2\right) = \frac{1}{2}\left((u-\widehat{\mu}_2^{sup})^2 - (u-\widehat{\mu}_1^{sup})^2\right). \tag{4.34}$$

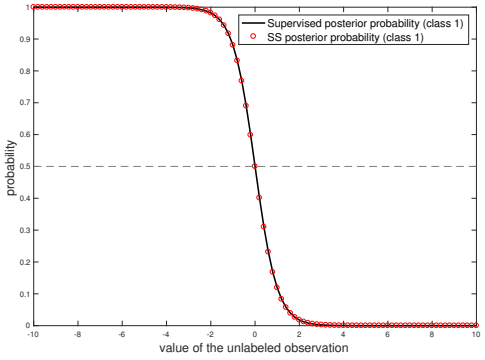
If the semi-supervised estimates satisfy this equation, then the difference of log-likelihoods evaluated at u between semi-supervised and supervised estimates is equal for both classes. As a consequence, the contrastive likelihood evaluated in the semi-supervised estimates should understandably in that case be constant in q.
Our experiment works as follows: we fix two labeled observations and we let the unlabeled observation u vary in a support, that here is $[-10, 10]$. The MCPL estimates are then computed for each value of u as $\widehat{\Psi}_{semi} = \widehat{\Psi}(\widehat{q}_1^{semi})$, where

$$\widehat{q}_1^{semi} = \underset{q_1 \in [0,1]}{\arg\min}\, CL(\widehat{\Psi}(q_1), q_1|\widehat{\Psi}_{sup}, \mathscr{D}_l, u). \tag{4.35}$$

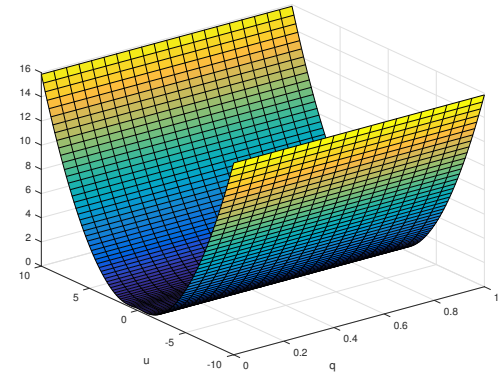
Then, the orthogonality condition is checked by plugging $\widehat{\Psi}_{semi}$ in (4.34). Note that numerical errors are taken into account, thus we consider the semi-supervised estimates to satisfy the orthogonality



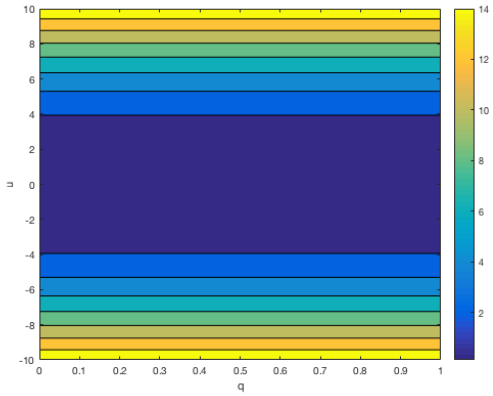
(a) *Optimal q_1 over the supervised class conditional densities.*



(b) *Posterior probabilities of class* 1 *for different values of u.*



(c) *Surface of $CL(\widehat{\Psi}_{semi}, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)$ for different values of the unlabeled observation.*



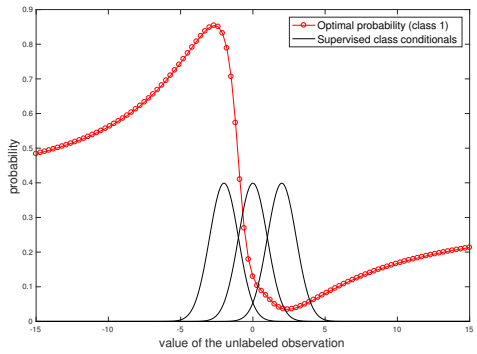
(d) *Level curves of $CL(\widehat{\Psi}_{semi}, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)$ for different values of the unlabeled observation.*

Figure 4.2: *Plots in the setting with* 2 *labeled observations and* 2 *classes.*

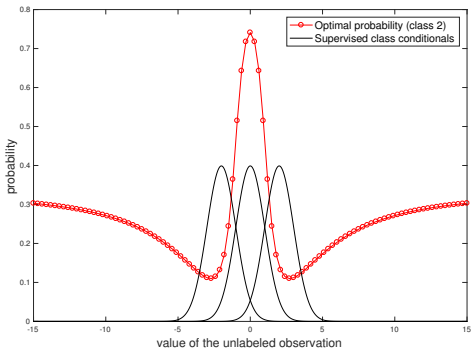
condition if the difference between the terms on the right and left hand sides is less than $\varepsilon = 10^{-4}$. In Figure 4.2a we show how $\widehat{q}_1^{semi}$ varies when it is computed with different values of u. It is clear that $\widehat{q}_1^{semi}$ is never a point on the boundary boundary of the 2-dimensional simplex. This is in turn justified by the fact that the semi-supervised estimates are such that the gradient of the contrastive likelihood is orthogonal to Δ_2 for any value of the unlabeled observation. This is a somewhat unexpected result as the orthogonality condition in principle seemed to be verified only in degenerate cases. Moreover, it is clear that $\widehat{q}_1^{semi}$ is proportional to the difference between the two supervised estimates of the class conditional distributions. In other terms, u is assigned predominantly to the class with the highest supervised posterior distribution, thus favoring the supervised classifier and resulting in the most adversarial setting for our semi-supervised learner. Figure 4.2b illustrates a peculiar behaviour: the MCPL estimates trained with a specific u are such that the posterior probability in that point is copied from the supervised model. Note that the posterior densities are computed as

$$p(y = 1|u, \Psi) = \frac{\pi_1 p(u|\mu_1)}{\pi_1 p(u|\mu_1) + \pi_2 p(u|\mu_2)},$$

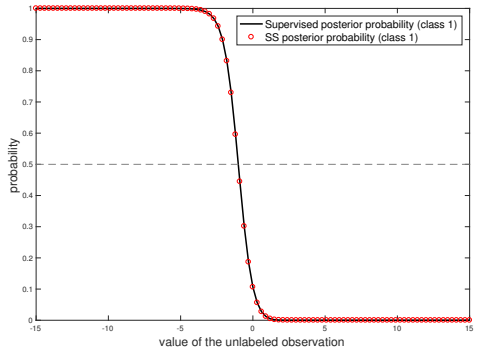
where $p(u|\mu_i)$ is a Gaussian with unitary variance and mean μ_i. In other terms, the unlabeled observation that is used to train the semi-supervised model is then classified and assigned exactly the same posterior probability as in the supervised learner. This means that the supervised and semi-



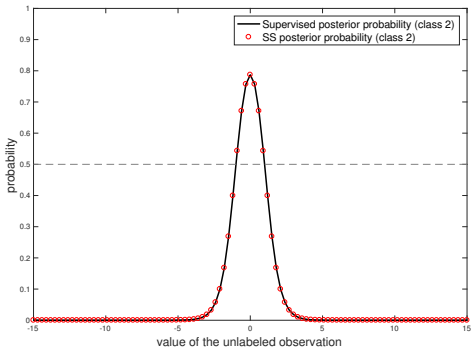
(a) *Adversarial posterior probability* $\widehat{q}_1^{semi}$ *over the supervised densities for different values of* u.



(b) *Adversarial posterior probability* $\widehat{q}_2^{semi}$ *over the supervised densities for different values of* u.



(c) *SS and supervised posterior probabilities of class* 1 *for different values of* u.



(d) *SS and supervised posterior probabilities of class* 2 *for different values of* u.

Figure 4.3: *Plots in the setting with* 3 *labeled observations and* 3 *classes.*

supervised sigmoid functions, which represent the partial assignments to each class of both models, intersect in u and are free to differ elsewhere.
The orthogonality of $\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)$ to the simplex has an additional consequence. The contrastive likelihood evaluated in $\widehat{\Psi}_{semi}$, thus seen as a function of q_1 only, is constant for any value of q_1. In mathematical terms, this means the following:

$$CL(\widehat{\Psi}_{semi}, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u) \;=\; \text{const} \qquad \forall\, q_\cdot \in \Delta_2\,.$$

This is illustrated in Figure 4.2c, in which one axis represents the different values that are, one at a time, used to train the semi-supervised classifier, while the other spans the feasible values of q_1. The semi-supervised likelihood obtained with a specific u and evaluated in $\widehat{\Psi}_{semi}$ is then greater than the supervised one by a constant. In addition, this holds independently of the value of the true q_1. The reader can observe that the contrastive likelihood is indeed strictly positive, even if we there is a unique additional observation. Clearly, in this simple setting we cannot expect a significant improvement if the only unlabeled observation is near the labeled feature vectors. It is anyway interesting to note that the improvement increases as u moves further away from the labeled observations.

We have shown that in this simple setting the orthogonality condition is verified independently of the value of the unlabeled observation. This hints that the most adversarial posterior probabilities $(q_1, 1-q_1)$ can indeed be soft labels and are thus not limited in any way to be on the boundary of the

simplex. Now, we would like to understand the extent to which these conclusions can be applied. To this end we iterate the same procedure to slightly different settings in order to have a generalization of the peculiarities of the method. For instance, we might add a labeled observation from a third additional class, while leaving unchanged the remaining aspects. Even this modification does not deviate the orthogonality condition from being satisfied, as one can see from Figure 4.3, where the plots for classes 1 and 2 are shown. The additional class introduces a second orthogonality condition, which has to be satisfied in order to have soft labels. The two conditions can now be expressed as

$$\begin{cases} +\frac{2}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_1 - \frac{1}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_2 - \frac{1}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_3 = 0, \\ -\frac{1}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_1 + \frac{2}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_2 - \frac{1}{3}\left(\nabla_{q_\cdot} CL(\Psi, q_\cdot|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)\right)_3 = 0. \end{cases}$$

However, the third class introduces two more parameters, so this does not represent a restriction. The plots for class 3 are omitted as they can be obtained from the first two by taking advantage of the fact that both the adversarial probabilities and the posterior distributions sum to one. Note that the semi-supervised posterior probabilities are one more time a copy of the supervised ones. This happens for all the three classes independently of the value of u. Similarly, we observe the same phenomenon for the posterior distributions as in the previous simpler setting.

We have studied two settings in which a single additional unlabeled observation is available to train the semi-supervised classifier. The MCPL approach uses it to build the estimated vector of parameters in such a way that the posterior probability assigned to the unlabeled observation is as in the supervised model. This curious behaviour is observed in either of the cases we analyzed. It is reasonable to say that this is desirable as in this simple case the safest way to use the additional observation is to copy the output from the supervised classifier, even more so if at the same time a strict improvement is obtained. It can also be observed that the adversarial probabilities converge to a neutral value, that is 0.5 in the first case and 0.3 in the second. This illustrates that when the supervised joint probability that is assigned to u is similar for all classes, then $\widehat{q}_{semi}$ reflects uncertainty. We will examine these aspect in the next section for some more complicated examples.
In addition, the orthogonality condition is always verified and the corresponding adversarial probabilities are not hard labels. These statements hold for any value of the unlabeled observation and corroborate our analysis of Section 4.2.7, which highlighted the relation between the orthogonality condition and the soft assignments. Interestingly, the orthogonality condition, which seemed to hold in degenerate cases only, is always verified in the studied setting. We can expect this behaviour to change when the data-set is larger, as the conditions become increasingly strict. In the next section we add one more unlabeled observation and we will discuss some interesting differences.

4.3.2. Two More Complicated Examples

The examples that we analyzed in the previous section can be made more realistic by considering the variance of each univariate Gaussian class conditional density as a parameter to estimated, rather than as a fixed value intrinsic of the parametric family. Clearly, at least two labeled observations per class are needed in order to define a meaningful estimate for each variance. In this section we thus add one labeled observation per class, for a total of four. Later, we assume that two unlabeled feature vectors are available to the semi-supervised learner, hence studying an interesting combination effect.

We begin our discussion by analyzing the influence of adding the variance as a parameter. This permits us to study how the model is influenced by the introduction of two parameters in a simple setting. To this end, we assume that the labeled feature vectors are $x_{11} = -2$ and $x_{12} = 1$ for class 1,

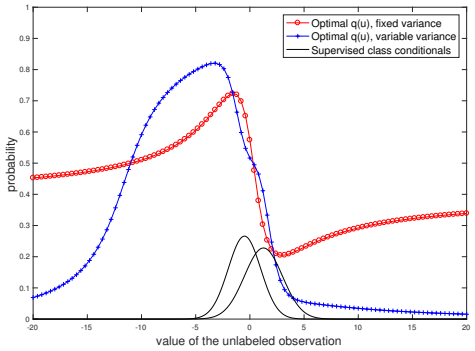
$x_{21} = -0.5$ and $x_{22} = 3$ for class 2. This choice is maintained throughout this section in order to perform a meaningful comparison. Different choices of the labeled sample can understandably lead to some dissimilarities, but the conclusions that are drawn here have general validity.
The supervised estimates based on this labeled data-set are then

$$\widehat{\mu}_1^{sup} = \frac{x_{11}+x_{12}}{2} = -\frac{1}{2}, \qquad \widehat{\pi}_1^{sup} = \frac{N_1}{4} = \frac{1}{2},$$
$$\widehat{\mu}_2^{sup} = \frac{x_{21}+x_{22}}{2} = \frac{5}{4}, \qquad \widehat{\pi}_2^{sup} = \frac{N_2}{4} = \frac{1}{2},$$

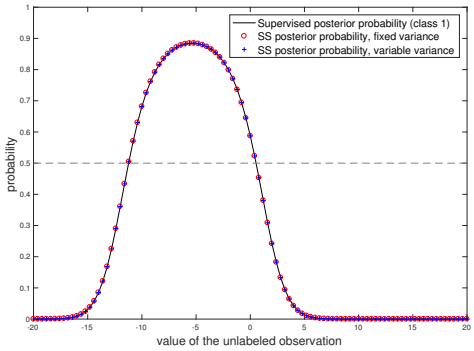
while the supervised estimates of the variances are

$$\widehat{\sigma}^2_{sup,1} = \frac{1}{N_1}\left((x_{11}-\widehat{\mu}_1^{sup})^2 + (x_{12}-\widehat{\mu}_1^{sup})^2\right) = 2.25,$$
$$\widehat{\sigma}^2_{sup,2} = \frac{1}{N_2}\left((x_{21}-\widehat{\mu}_2^{sup})^2 + (x_{22}-\widehat{\mu}_2^{sup})^2\right) = 3.0625.$$

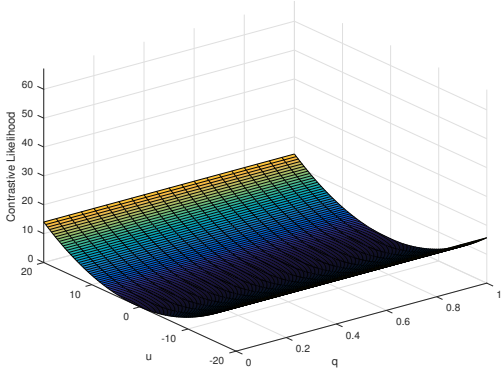
We underline how the supervised variances are here different, hence the supervised class conditional distributions have different shapes. Now, the structure of the semi-supervised estimates that



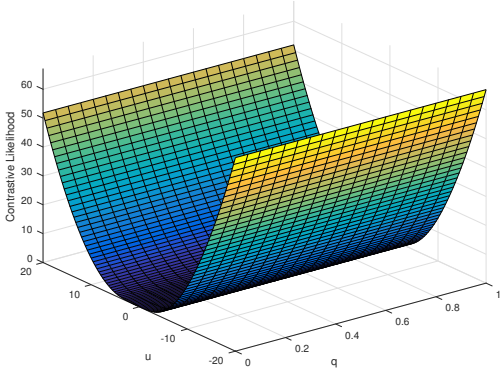
(a) *Adversarial probabilities with fixed (red) and variable (blue) variance for different values of u.*



(b) *Semi-supervised posterior probabilities at point u, used for training, with fixed (red) and variable (blue) variance.*



(c) *Surface of* $CL(\widehat{\Psi}_{semi}, q_1|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)$ *as a function of* q_1 *for fixed variance.*



(d) *Surface of* $CL(\widehat{\Psi}_{semi}, q_1|\widehat{\Psi}_{sup}, \mathscr{D}_l, u)$ *as a function of* q_1 *for variable variance.*

Figure 4.4: *Comparison fixed / variable variance for the setting with* 4 *labeled and* 1 *unlabeled observations.*

is obtained by solving the maximization task is similar to Equations (4.32), (4.33). In particular, in

this setting the priors have an identical structure as in (4.32) but where $N_1 = N_2 = 2$, while the estimates for the means have the following structure:

$$\widehat{\mu}_1(q_1) = \frac{x_{11} + x_{12} + q_1 u}{N_1 + q_1}, \qquad \widehat{\mu}_2(q_1) = \frac{x_{21} + x_{22} + (1 - q_1) u}{N_2 + (1 - q_1)}. \tag{4.36}$$

On the other hand, the MCPL estimates of the variances are

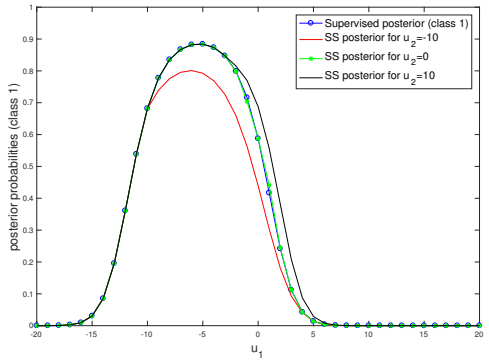
$$\begin{aligned} \widehat{\sigma}_1(q_1) &= \frac{(x_{11} - \widehat{\mu}_1(q_1))^2 + (x_{12} - \widehat{\mu}_1(q_1))^2 + q_1(u - \widehat{\mu}_1(q_1))^2}{N_1 + q_1}, \\ \widehat{\sigma}_2(q_1) &= \frac{(x_{21} - \widehat{\mu}_2(q_1))^2 + (x_{22} - \widehat{\mu}_2(q_1))^2 + (1 - q_1)(u - \widehat{\mu}_2(q_1))^2}{N_2 + (1 - q_1)}. \end{aligned} \tag{4.37}$$

These estimates are used when the variance is a parameter of the problem, while when it is fixed we impose that the variances are as in the supervised model. This once again allows us to perform a sensible comparison.
The minimax optimization is then solvable following the procedure discussed in the previous section. This means first plugging in the contrastive likelihood the estimates for the priors, the means and variances as in (4.36), (4.37), then minimizing the obtained objective function with respect to q_1. The minimization then has an identical formulation as in (4.35), hence resulting in the most adversarial value of $q_1 \in [0, 1]$. The results are shown and compared in Figure 4.4. It is clear that either for fixed or variable variance we have roughly that

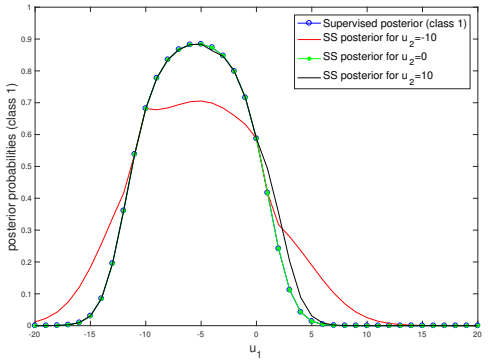
$$\widehat{q}_1^{\,semi} \begin{cases} > \frac{1}{2} & \text{if } \ \widehat{\pi}_1^{sup}\, p(u|\widehat{\mu}_1^{sup}, \widehat{\sigma}^2_{sup,1}) > \widehat{\pi}_2^{sup}\, p(u|\widehat{\mu}_2^{sup}, \widehat{\sigma}^2_{sup,2}), \\ \leq \frac{1}{2} & \text{if } \ \widehat{\pi}_1^{sup}\, p(u|\widehat{\mu}_1^{sup}, \widehat{\sigma}^2_{sup,1}) \leq \widehat{\pi}_2^{sup}\, p(u|\widehat{\mu}_2^{sup}, \widehat{\sigma}^2_{sup,2}). \end{cases}$$

Note that the two lines in Figure 4.4a meet exactly when their value is approximately 0.5 and thus assign u predominantly to the same class in any case. However, when the variance is considered to be a parameter we observe that $\widehat{q}_1^{\,semi}$ goes to zero for $\widehat{\pi}_1^{sup}\, p(u|\widehat{\mu}_1^{sup}, \widehat{\sigma}^2_{sup,1}) \ll \widehat{\pi}_2^{sup}\, p(u|\widehat{\mu}_2^{sup}, \widehat{\sigma}^2_{sup,2})$. On the contrary, the limit is non-zero if the variance is fixed. Nonetheless, this different behaviour does not affect the orthogonality of the gradient of the contrastive likelihood to the simplex, which indeed holds for any u. The orthogonality condition is once again as in Equation (4.30) and the same criterion to check its validity is used.
It is also interesting to observe that, in the case of fixed variances, the adversarial probabilities have a different limit to the left and to the right, contrarily to the setting illustrated in Figure 4.2a. This is a consequence of the dissimilarity between the supervised estimates of the variances. Indeed, the class to which is associated the higher variance has heavier tails and thus prevails from some point on.
As a final note we underline the method achieves a larger improvement when considering the variance as a parameter to be estimated. This can be explained by observing that adding the variance to the parameter vector increases the flexibility of the classifier. In addition, it is evident that in both settings the contrastive likelihood evaluated at $\widehat{\Psi}_{semi}$ is constant in q_1 for a fixed u. This once again follows from the orthogonality condition, as claimed in Section 4.2.7.

It is then interesting to check how these considerations generalize to a more complicated setting with two unlabeled observations. In order to do so, the labeled observations are left unchanged and are thus equal to the previous analysis. The two unlabeled observations are now referred to as u_1 and u_2, while we indicate the posterior probabilities as $q_1 = p(y = 1|u_1)$ and $q_2 = p(y = 1|u_2)$. The reader can once again trust that in this setting the MCPL estimates after the maximization task have



(a) *Supervised (blue) and SS posterior probabilities* $p(y=1|u_1)$ *for different values of* u_2 *for fixed variance.*



(b) *Supervised (blue) and SS posterior probabilities* $p(y=1|u_1)$ *for different values of* u_2 *for variable variance.*

Figure 4.5: *Plots of the posterior probabilities in* u_1 *for different values of* u_2 *for the case of* 4 *labeled and* 2 *unlabeled observations, with fixed (left) and variable (right) variances.*

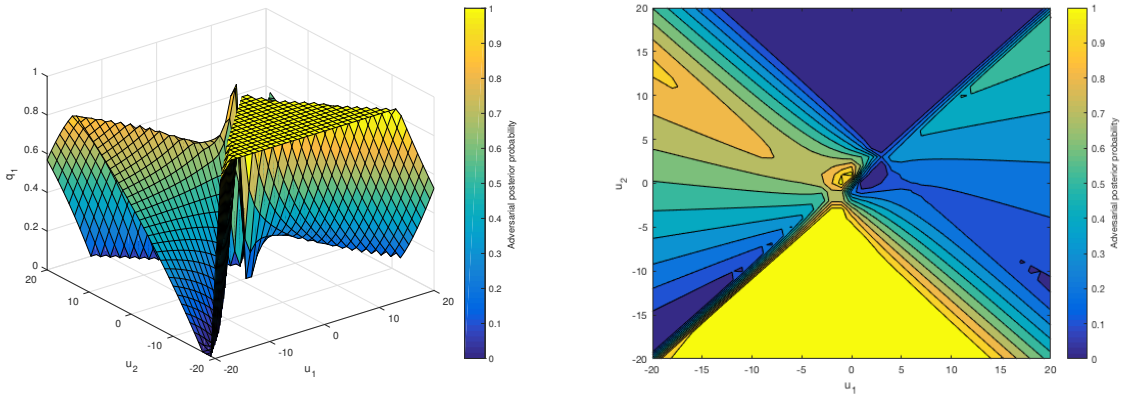
the following structure

$$\widehat{\mu}_1(q_1,q_2) = \frac{x_{11}+x_{12}+q_1u_1+q_2u_2}{N_1+q_1+q_2}, \qquad \widehat{\pi}_1(q_1,q_2) = \frac{N_1+q_1+q_2}{N+M},$$
$$\widehat{\mu}_2(q_1,q_2) = \frac{x_{21}+x_{22}+(1-q_1)u_1+(1-q_2)u_2}{N_1+(1-q_1)+(1-q_2)}, \quad \widehat{\pi}_1(q_1,q_2) = \frac{N_2+(1-q_1)+(1-q_2)}{N+M}.$$

and the semi-supervised variances are similar to Equation (4.37), but including also the second unlabeled feature vector as in the previous formulas.
The same experiment as in earlier examples is now performed. First, it has to be noted that there are now two separate orthogonality conditions, one for each unlabeled data point:

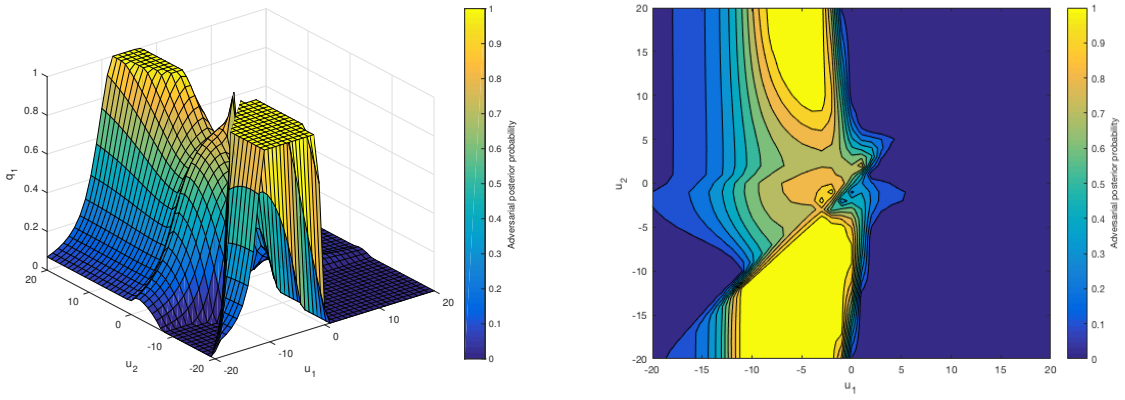
$$\log(\pi_1\, p(u_l|\mu_1,\sigma_1^2))-\log(\widehat{\pi}_1^{sup}\, p(u_l|\widehat{\mu}_1^{sup},\widehat{\sigma}^2_{\text{sup},1})) = \log(\pi_2\, p(u_l|\mu_2,\sigma_2^2))-\log(\widehat{\pi}_2^{sup}\, p(u_l|\widehat{\mu}_2^{sup},\widehat{\sigma}^2_{\text{sup},2}))$$

for $l=1,2$. This is certainly expected to make it more difficult to have both conditions simultaneously satisfied, but it is possible that only one of the two is verified. It should be clear that when the variances are fixed we assume $\sigma_1^2=\widehat{\sigma}^2_{\text{sup},1}$ and $\sigma_2^2=\widehat{\sigma}^2_{\text{sup},2}$.
Figure 4.5 shows how the semi-supervised estimates of the posterior probabilities $p(y=1|u_1)$ vary in u_1 for different values of the second unlabeled observation. It can be observed that the posterior overlaps the supervised line when u_2 takes values that are near the labeled observations, but can change significantly when u_2 is more extreme. In any case it is clear that when it is possible the MCPL estimates are again chosen so that the supervised and semi-supervised posteriors are close in the unlabeled observation.
Considering the variance as a parameter that needs to be estimated reveals to be a major factor on the profile of the adversarial probabilities. This can be seen in Figure 4.6, where the surface and the level curves of $\widehat{q}_1^{semi}$ in both settings are shown. The reader can convince himself that the plots for $\widehat{q}_2^{semi}$ are identical to Figure 4.6, as the order of the observations is not relevant. In the case of a variable variance, $\widehat{q}_1^{semi}$ follows approximately the behaviour that was observed in the setting of a single unlabeled observation, see Figure 4.4a. However, it is clear that the value of u_2 influences the shape of $\widehat{q}_1^{semi}$. In particular, for extreme values of u_2 there are some clear differences, but the trend remains similar. The behaviour is on the contrary closer to the precedent setting when u_2 is nearer to the labeled points. This line of reasoning does not apply to the fixed variance setting, thus making the interpretation in that case more difficult. It is in any case clear that for reasonable values of the



(a) *Adversarial probabilities with fixed variance for different values of u_1 and u_2.*

(b) *Level curves of adversarial probabilities with fixed variance.*



(c) *Adversarial probabilities with variable variance for different values of u_1 and u_2.*

(d) *Level curves of adversarial probabilities with variable variance.*

Figure 4.6: *Comparison fixed / variable variance for the setting with* 4 *labeled and* 2 *unlabeled observations.*

unlabeled points the most adversarial probabilities are chosen similarly to what was previously discussed. This can be observed for instance by focusing on the interval $[-6,6]$ on the y-axis in Figure 4.6d. The value of u_2 has in any case a relevant influence, but it is nonetheless evident that $\widehat{q}_1^{semi}$ follows the supervised posterior probability assigned to class 1.
Once again, the orthogonality conditions are satisfied when the adversarial probabilities are on the boundary of their respective simplex. We can then state that, for a variable variance and non-extreme unlabeled observations, at least one of the two orthogonality conditions is satisfied if the observations are reasonably close to the supervised feature vectors. This observation holds also in the fixed variance setting, even if it is harder to fully understand the reasons that lead to the profile in Figure 4.6b. However, either the two adversarial probabilities are soft if the variance is variable and the observations are in approximately in the range of the labeled ones.

We have thus seen how adding an unlabeled observation complicates the analysis. Nevertheless, we can claim that also in this setting the adversarial probabilities follow the behaviour of the supervised posteriors, at least in a more realistic case. This means that we can expect to have soft labels for reasonable values of the unlabeled data and that in general it is possible to expect this situation.

4.3.3. Final Comments

We have analyzed a series of illustrative examples with a focus on the adversarial probabilities. This study has served not only to confirm some properties that were found theoretically, but also new findings were made. In particular, it was confirmed that soft labels are to be expected if the semi-supervised estimate satisfy some orthogonality conditions. Moreover, the contrastive likelihood is in that case constant in the corresponding simplex when evaluated in $\widehat{\Psi}_{semi}$. The examples of this section make it clear that it can indeed happen that the adversarial probabilities are not hard labels. This seems to hold in general settings if the unlabeled sample is not too far from the labeled one. Furthermore, we have seen that the semi-supervised learner tries to copy the supervised output in the unlabeled points. Our analysis hints that this is the case for reasonable values of the unlabeled data-set, while it may not be possible to do so for more extreme observations. Finally, it was discovered that the adversarial probabilities follow in some sense the supervised class conditional distributions. This is a sensible finding as it is clear that this choice favors the supervised solution. An effect of this choice is that the MCPL estimates can hardly result in a significant performance improvement if the unlabeled features are close to the labeled ones. Indeed, the MCPL approach (partially) assigns those observations to a specific class in order to make the supervised classifier a good predictor on the entire data-set. This result suggests that the MCPL approach is in this case very conservative. On the other hand, considerable improvements seem possible if some of the unlabeled features are far from the labeled sample. This is illustrated by Figures 4.2c, 4.4c and 4.4d. This behaviour seems to find its motivation in the fact that the supervised model does not fit well extreme data points, rather than in a less conservative approach of the proposed method. Nonetheless, we could expect the MCPL estimates to be conservative because of the very strong guarantees that they provide on the unlabeled data.

In conclusion, this section has highlighted some interesting aspects of the MCPL approach. The theoretical analysis of Section 4.2.7 is corroborated by our findings and it seems now to be clear that the adversarial probabilities are in no way limited to be points on the boundary of the simplex.

4.4. Conclusion

We have described the adaptation of the principles of contrast and pessimism to the case of a generative probabilistic model. The resulting classifier, which we called MCPL, satisfies several interesting guarantees on performance improvement. In this sense, it became clear that we could obtain considerably more powerful results for this generative classifier than for margin based losses or the least squares classifier. This motivates us to continue developing this theory in the remaining bits of this thesis. In the next chapter we focus on the use of Gaussian densities in the MCPL approach.

5

MCPL for Gaussian Discriminant Analysis

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are two very popular generative classifiers that are based on an original work by Sir Ronald Fisher. These methods have been widely studied in the past decades and have found applications in several areas such as classification and feature selection. In this chapter we build the semi-supervised counterpart of the LDA and QDA classifiers by applying the framework developed in Chapter 4. This means that once again we assume to have N labeled and M unlabeled observations and we want to safely incorporate them in LDA and QDA using the MCPL approach.
In both classifiers the class conditional distributions are modeled as (multivariate) Gaussian densities. The difference between the two lies in the fact that in LDA the covariance matrix is assumed to be common to all classes, while in QDA there is no such constraint. This results in different type of decision surface, which is linear for LDA and quadratic for QDA. Section 5.1.2 is dedicated to a brief explanation of this property in a supervised setting.
We start with a description of the Gaussian family and its decomposition to the exponential form. This formulation is then used to define the semi-supervised counterparts in Sections 5.2.1 and 5.2.2. In Section 5.2.2 we discuss an extension of the MCPL approach to the case of a subset of shared parameters with same statistical models. This subsumes LDA and is thus a more general setting. Results on strict improvement hold almost surely for this scenario and trivially for the MCPL semi-supervised version of QDA.

5.1. Preliminaries

5.1.1. The Gaussian Distribution as an Exponential Family

We have anticipated that in LDA and QDA we model each class conditional distribution with a Gaussian density, which is defined as:

$$g(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

In this formula, d is the dimension of the observed feature vectors, while μ and Σ indicate the mean and the covariance matrix respectively. We can then define the family of Gaussian densities as

$$\mathscr{P} = \left\{ g(\cdot|\mu,\Sigma): \quad \mu \in \mathbb{R}^d, \quad \Sigma \in \mathbb{R}^{d\times d}, \quad \Sigma_{ij} = \Sigma_{ji}, \quad \Sigma_{ii} > 0 \quad \forall i,j \right\}. \tag{5.1}$$

Now, we are interested in the decomposition of a minimal Gaussian family in its canonical form, which was defined in Section 4.1. To this end we take advantage of the following observations

$$\mu^T\Sigma^{-1}x = x^T\Sigma^{-1}\mu,$$
$$x^T\Sigma^{-1}x = \mathrm{tr}\left(x^T\Sigma^{-1}x\right) = \mathrm{tr}\left(\Sigma^{-1}xx^T\right),$$

where the first equation is obtained by simply noting that the term on the left hand side is a scalar and hence it is equal to its transpose, while the second equation is obtained by observing that a scalar is equal to its trace and the trace operator is invariant under cyclic permutations. In both equations the symmetry of the covariance matrix, and hence of its inverse, is crucial.
Taking advantage of the two previous equations we can now write the sufficient statistic and the cumulant function as

$$\begin{aligned} t(x) &= \begin{pmatrix} t_1(x) \\ t_2(x) \end{pmatrix}, \\ F(\mu,\Sigma) &= \frac{1}{2}\left(\mu^T\Sigma^{-1}\mu - \log|\Sigma^{-1}|\right), \end{aligned}$$

where the two sub-vectors that define the sufficient statistic are

$$\begin{aligned} t_1(x) &= x, \\ t_2(x) &= \left\{ \frac{x_i x_j}{1+\delta_{ij}} : i \le j, \quad i,j \in [M] \right\}, \end{aligned}$$

in which δ_{ij} is the Kronecker delta. On the other hand, the canonical parameters are defined by

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},$$

where

$$\begin{aligned} \eta_1 &= \left(\left(\Sigma^{-1}\mu\right)_1, \ldots, \left(\Sigma^{-1}\mu\right)_d \right), \\ \eta_2 &= \left\{ \left(-\Sigma^{-1}\right)_{ij} : \quad i \le j, \quad i,j \in [M] \right\}. \end{aligned}$$

The second part of the vector of canonical parameters contains the components of the negative, inverse covariance matrix that lie on and above its diagonal. These are sufficient as the covariance matrix, and hence its inverse, is symmetric. Note that the constant term is absorbed in the carrier measure, i.e. $c(x) = (2\pi)^{-d/2}$.
Now, in Section 4.1.2 we have discussed that the mean parametrization, in which the parameter is the gradient of the cumulant generating function, is important in the computation of the maximum likelihood estimates. We are then interested in the expression of the mean parameter for Gaussian distributions. First observe that we can complete the parameter η_2 by including all the elements of $-\Sigma^{-1}$, so that the expressions we are interested in are easier to obtain and interpret. This does not affect our analysis as we are basically duplicating the conditions on the upper triangular part of $-\Sigma^{-1}$ to the lower part. It is clear that in order to use this notation the second part of the sufficient statistic has to be modified accordingly to $t_2(x) = xx^T$. The gradient of the cumulant generating function can then be divided in two parts as follows

$$\nabla F(\eta) = \left(\mu, \mu\mu^T + \Sigma\right), \tag{5.2}$$

where the first part is the derivative with respect to η_1 and the second part with respect to η_2. This division will be handy in the following sections.
We conclude this section by observing that the MCPL estimates of the prior probabilities $(\pi_1, \ldots, \pi_K)$ do not depend on the statistical model and are thus computed as in Equation (4.15). In the following sections we can then focus only on the MCPL estimates of the mean vectors and of the covariance matrices.

5.1.2. Supervised LDA and QDA

This section serves as a small introduction to get acquainted to Gaussian Discriminant Analysis. The two classifiers are defined and the shape of the decision boundary is derived following [47].

LDA and QDA are generative classifiers based on the Bayes classification rule, that in a two-class case is the following:

$$\widehat{y} = \begin{cases} y_1, & \text{if } p(y_1|x) > p(y_2|x), \\ y_2, & \text{if } p(y_1|x) < p(y_2|x). \end{cases}$$

We have discussed in Chapter 2 that the Bayes rule can be used to state that the decision is equivalently based on

$$p(y_1)p(x|y_1) \;>\; p(y_2)p(x|y_2).$$

In the current framework the class conditional distributions are modelled with Gaussian distributions and thus it is easier to deal with the logarithm of these quantities. We can then define the discriminant functions as

$$\begin{aligned} g_i(x) \;&=\; \log(p(y_i)p(x|y_i)) = \log p(y_i) + \log p(x|y_i) \\ &=\; \log p(y_i) - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + c_i. \end{aligned}$$

It is then clear that the decision boundary is in general quadratic in x because of the term $x^T\Sigma_i^{-1}x$. However, the quadratic term is null if the covariance matrix is shared by all classes. Now, the former is the case of QDA, while the latter is that of LDA.
The parameters of the class of densities can be estimated in several ways. The most common choice is the maximum likelihood method, which is chosen in the MCPL approach.

5.2. Semi-Supervised LDA and QDA

We are now ready to define the semi-supervised counterpart of the LDA and QDA classifiers. We begin with the QDA setting, which is a straightforward application of the results in Chapter 4, and later the focus is on LDA.

5.2.1. Contrastive Pessimistic Quadratic Discriminant Analysis

In QDA the class conditional distributions are modeled with K independent Gaussian densities. There is then only one parametric family that models each class conditional distribution, or in other terms

$$p_k(x|\eta_k) \in \mathscr{P} \qquad \forall k = 1,\dots,K, \tag{5.3}$$

where $\mathscr{P}$ is the set of Gaussian distributions defined in (5.1). Using the notation of Section 4.2 we then have

$$\Psi_{QDA} \;=\; \left(\pi_1,\dots,\pi_K,\mu_1,\dots,\mu_K,\Sigma_1,\dots,\Sigma_K\right).$$

This means that there are $K\left(d + d(d+1)/2\right)$ parameters for the Gaussian densities and K additional prior probabilities. We have already observed that the prior probabilities have the same structure as in (4.15), so here we can focus on the remaining parameters.
Now, we observe that this setting is precisely part of the more general framework we have introduced in Section 4.2, so we can apply the conditions we have already derived in order to compute the structure of the MCPL estimates for the means and the covariances. In particular, Equation (4.16) is exactly what we need. Taking advantage of (5.2) we find that the MCPL estimates of the means have the following structure:

$$\widehat{\mu}_k^{semi}(Q) \;=\; \frac{1}{N_k + \sum_{j=1}^M q_{kj}} \left(\sum_{i=1}^{N_k} x_{k(i)} + \sum_{j=1}^{M} q_{kj}u_j \right) \qquad \text{for } k = 1,\dots,K. \tag{5.4}$$

The second part of $\nabla F(\eta)$ implies that the MCPL estimates of the covariance matrices satisfy the relation

$$\widehat{\Sigma}_k^{semi}(Q) = \frac{1}{N_k + \sum_{j=1}^M q_{kj}} \left(\sum_{i=1}^{N_k} x_{k(i)} x_{k(i)}^T + \sum_{j=1}^M q_{kj} u_j u_j^T \right) - \widehat{\mu}_k^{semi}(Q) \left(\widehat{\mu}_k^{semi}(Q) \right)^T ,$$

which using the distributive property of the matrix product on the last term of the right hand side can be rearranged as

$$\begin{aligned} \widehat{\Sigma}_k^{semi}(Q) = \frac{1}{N_k + \sum_{j=1}^M q_{kj}} \Bigg(& \sum_{i=1}^{N_k} \left(x_{k(i)} - \widehat{\mu}_k^{semi} \right) \left(x_{k(i)} - \widehat{\mu}_k^{semi} \right)^T \\ & + \sum_{j=1}^M q_{kj} \left(u_j - \widehat{\mu}_k^{semi} \right) \left(u_j - \widehat{\mu}_k^{semi} \right)^T \Bigg) \end{aligned} \tag{5.5}$$

for any k. Note that in the last equation we omitted the dependency of $\widehat{\mu}_k^{semi}$ from Q to lighten the notation.
The estimates (5.4) and (5.5) still depend on the probabilities Q, which are then selected solving the minimization task (4.18). However, the minimum is attained by compactness of the space of the posterior probabilities, so the final estimates are simply obtained by plugging the optimal Q in the expressions for the MCPL estimates of the means and of the covariance matrices.
It is also interesting to note that we need to have a set of observations such that the covariance matrix has full rank, which makes it a meaningful estimate. This happens in general when $N_k > d$ if the observations in the data-set are non redundant. The same condition is required to estimate the covariance matrices in the supervised setting, thus this is not really a new requirement on the data. It has to be noted that the need for large data-sets is a common criticism to both the LDA and QDA classifiers.

The QDA semi-supervised estimates computed with the MCPL approach come with the guarantees we proved in Section 4.2. In particular the statistical model is common to all the classes, so we can conclude that $\mathbb{P}_{u_1,\dots,u_M \sim p(x)}(\widehat{\Psi}_{semi}(U) \neq \widehat{\Psi}_{sup}) = 1$ and most importantly that

$$\mathbb{P}_{\mathscr{D}_u \sim p(x,y)} \left(L_{QDA}(\widehat{\Psi}_{semi}(U) | \mathscr{D}_{full}) > L_{QDA}(\widehat{\Psi}_{sup} | \mathscr{D}_{full}) \right) = 1 ,$$

where $\mathscr{D}_{full} = \mathscr{D}_l \cup \mathscr{D}_u$, $\mathscr{D}_u = \{(u_j, v_j)\}_{j=1}^M$ and $p(x,y)$ is the underlying joint density.
In conclusion, a straightforward application of the framework we have introduced in Section 4.2 can be used to build a semi-supervised version of QDA that almost surely improves over its supervised counterpart.

5.2.2. Contrastive Pessimistic Linear Discriminant Analysis

We have previously discussed that in LDA the parameters that need to be estimated are the mean vector of each class, the shared covariance matrix, and the prior probabilities. These parameters can be united similarly to Section 5.2.1 as follows

$$\Psi_{LDA} = (\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma) .$$

This implies that $K(d+1) + d(d+1)/2$ parameters have to be estimated in order to learn the LDA classifier.
Contrary to the case of QDA, the MCPL approach cannot be applied to LDA as it was described in Section 4.2. The reason is that the additional constraint on the covariance matrices cannot be incorporated in the parametric family itself. This has either to be added as an explicit constraint or

be incorporated in the objective function. The latter choice is examined in this section in a more general setting where part of the parameter is shared between the classes. Later, this is applied to the LDA setting.

Assume every class conditional distribution is modeled with the same exponential family. Suppose there is a part of the parameter vector that is chosen to be equal for each class. In other terms the classes have a shared parameter that has to be estimated in a suitable way. For instance in LDA the constraint is on the covariance matrix, while the mean vectors are class specific. It is then necessary to split the canonical parameter vector of each class as

$$\eta_k = \begin{pmatrix} \eta_1^k \\ \eta_2 \end{pmatrix} \in \Omega,$$

where η_2 does not depend on the class as it is indeed the shared parameter, and Ω is the parameter space. Similarly, we can split the sufficient statistic vector in two parts, which we denote $t_1(\cdot)$ and $t_2(\cdot)$. We call the dimension of each η_1^k and η_2 respectively d_1 and d_2. Then, the maximization task that needs to be solved is the following formula:

$$\begin{aligned} \max_{(\eta_1^1,\dots,\eta_1^K,\eta_2)} \quad & \sum_{k=1}^{K} \Bigg\{ \langle \sum_{i=1}^{N_k} t_1(x_{k(i)}) + \sum_{j=1}^{M} q_{kj}\, t_1(u_j)\,, \eta_1^k \rangle + \langle \sum_{i=1}^{N_k} t_2(x_{k(i)}) + \sum_{j=1}^{M} q_{kj}\, t_2(u_j)\,, \eta_2 \rangle \\ & - \left(N_k + \sum_{j=1}^{M} q_{kj} \right) F(\eta_k) \Bigg\}, \end{aligned} \tag{5.6}$$

where we have divided the scalar product in two parts to explicitly separate η_1^k and η_2. Note that the likelihood evaluated in $\widehat{\Psi}_{sup}$ is not taken into account because it does not affect the result of this first step. The only term where both parts are present is the cumulant generating function, which has to be differentiated carefully. We denote the two parts of its gradient as $\nabla_{\eta_1} F(\cdot)$ and $\nabla_{\eta_2} F(\cdot)$, which are defined as

$$\left(\nabla_{\eta_j} F(\cdot) \right)_i = \frac{\partial F(\cdot)}{\partial (\eta_j)_i} \qquad \text{for } j = 1,2; \quad i = 1,\dots,d_j\,.$$

First, we differentiate with respect to η_1^k and equate the result to zero. We thus obtain for each k the first set of d_1 equations:

$$\nabla_{\eta_1} F(\widehat{\eta}_k) = \frac{1}{N_k + \sum_{j=1}^{M} q_{kj}} \left(\sum_{i=1}^{N_k} t_1(x_{k(i)}) + \sum_{j=1}^{M} q_{kj}\, t_1(u_j) \right), \tag{5.7}$$

Observe that this condition is equivalent to what was derived in Equation (4.16). This is reasonable as the first part of the parameter varies from class to class, which was exactly the setting in Section 4.2. On the other hand, we expect a different behaviour for the shared parameter. Indeed, equating the derivative with respect to η_2 to zero we obtain the remaining d_2 equations

$$\sum_{k=1}^{K} \left\{ \nabla_{\eta_2} F(\widehat{\eta}_k) \left(N_k + \sum_{j=1}^{M} q_{kj} \right) \right\} = \sum_{k=1}^{K} \left\{ \sum_{i=1}^{N_k} t_2(x_{k(i)}) + \sum_{j=1}^{M} q_{kj}\, t_2(u_j) \right\}. \tag{5.8}$$

As expected, setting the second derivative to zero results in a different condition, where every class is present. Note that (5.7) gives us Kd_1 equations, while (5.8) are the last d_2.

We have then extended the MCPL approach to the case of same statistical models for the class conditional distributions where a part of the parameter vector is shared between the classes. Any combination of constrained parameters can in principle be studied and no restrictions are made. Now,

the robustness of these estimates has yet to be confirmed. In Section 4.2 we proved some results in this direction, but here the structure of the problem was modified and as a consequence the application of the same results must be discussed. Fortunately, the same arguments that were used in Section 4.2 can be applied in this context. Indeed, the objective function in (5.6) has indeed the properties that are needed to prove the non-degradation guarantee of the estimates. This is because the contrastive likelihood is again a sum of functions that are strictly concave in the parameters and linear, thus convex, in the adversarial probabilities. As a consequence, Sion's minimax theorem can be applied as long as we assume the supremum to be attained for any value of Q. Then, the following holds under Assumption 4.4:

$$L(\widehat{\Psi}_{semi}|\mathscr{D}_{full}) \geq L(\widehat{\Psi}_{sup}|\mathscr{D}_{full})$$

for any choice of the labels on the unlabeled data in $\mathscr{D}_{full}$. Now, it remains to determine whether we can expect the improvement to be strict. This is addressed in the next theorem.

Theorem 5.1. *Consider the setting described in this section where each class conditional distribution is modeled with the same exponential family and a part of the parameters is shared between the classes. Moreover, suppose that Assumption 4.4 holds and that the data $\mathscr{D}_u$ are sampled from a continuous distribution. Then the MCPL estimates have $p(x,y)$-almost surely a strictly improved performance.*

Proof. Using the same arguments as in the proof of Theorem 4.10, we can define the estimated mean parameter as

$$\widehat{\mu}_{semi}(U) = \sum_{k=1}^{K} \widehat{\pi}_k^{semi} \widehat{\mu}_k^{semi},$$

where the dependency on the unlabeled sample is made explicit. Now, we can limit our attention the the d_1 components that correspond to the non-shared parameters:

$$\left(\widehat{\mu}_{semi}(U)\right)_1 = \frac{1}{N+M}\left(\sum_{i=1}^{N} t_1(x_i) + \sum_{j=1}^{M} t_1(u_j)\right).$$

Equality follows from similar computations to the proof of Theorem 4.10. We can then state that $(\widehat{\mu}_{semi}(U))_1 = (\widehat{\mu}_{sup})_1$ if the semi-supervised and supervised estimates are equal. However, this happens if and only if

$$\frac{1}{N+M}\left(\sum_{i=1}^{N} t_1(x_i) + \sum_{j=1}^{M} t_1(u_j)\right) = \frac{1}{N}\left(\sum_{i=1}^{N} t_1(x_i)\right),$$

which because of the continuity of the random sample is a zero probability event with respect to the underlying marginal density p_x. The consequence is then that p_x-almost surely the supervised and the semi-supervised estimates are not equal.
It is now sufficient to observe that the contrastive likelihood is strictly concave in the parameters and by the same arguments of Corollary 4.11 we have the thesis. □

We have then proved that also in this setting the MCPL estimates almost surely improve over the supervised model. Now, we can finally focus specifically on the LDA framework. In particular, the application of Equation (5.7) to the case of Gaussian distributions results in MCPL estimates of the mean vectors that are equal to Equation (5.4). The parameters that are free to be different are thus estimated equivalently to the unconstrained setting, which is QDA. The estimate of the covariance matrix can be computed by manipulating the left hand side of (5.8):

$$\sum_{k=1}^{K}\left\{\left(\widehat{\Sigma} + \widehat{\mu}_k\widehat{\mu}_k^T\right)\left(N_k + \sum_{j=1}^{M} q_{kj}\right)\right\} = (N+M)\,\widehat{\Sigma} + \sum_{k=1}^{K}\left(N_k + \sum_{j=1}^{M} q_{kj}\right)\widehat{\mu}_k\widehat{\mu}_k^T,$$

thus obtaining by the same computations as in Section 5.2.1 the following estimate

$$\begin{aligned}\widehat{\Sigma}_{semi}(Q) \;=\; \frac{1}{N+M}\sum_{k=1}^{K}\Bigg(&\sum_{i=1}^{N_k}\left(x_{k(i)}-\widehat{\mu}_k^{semi}\right)\left(x_{k(i)}-\widehat{\mu}_k^{semi}\right)^T \\ &+\sum_{j=1}^{M} q_{kj}\left(u_j-\widehat{\mu}_k^{semi}\right)\left(u_j-\widehat{\mu}_k^{semi}\right)^T\Bigg).\end{aligned} \tag{5.9}$$

Note that at this stage $\widehat{\mu}_k = \widehat{\mu}_k(Q)$, but we dropped the explicit dependence on Q not to burden the notation.
The estimates for the means and the covariance matrix that we derived are the same that were found in [33]. This highlights that the more general formulation we introduced is indeed applicable to this setting.

We can then apply Theorem 5.1 to the LDA setting, hence obtaining the following guarantee

$$\mathbb{P}_{\mathscr{D}_u\sim p_{xy}}\left(L_{LDA}(\widehat{\Psi}_{semi}|\mathscr{D}_{full}) > L_{LDA}(\widehat{\Psi}_{sup}|\mathscr{D}_{full})\right) = 1\,.$$

In conclusion, the MCPL approach is capable of building a safe semi-supervised version of the LDA. This shows that the theoretical results discussed in [33] indeed hold in this and in more general settings where a subset of the parameters is shared between the classes.

6

Maximum Contrastive Pessimistic Likelihood Approach for Missing Data Problems

In previous chapters the contrastive pessimistic approach was introduced and studied for the problem of classification. This method is motivated by the improvement guarantees that were proved for instance in Theorem 3.8 and in Corollary 4.11. In particular, we showed that the MCPL approach is guaranteed to make safe use of incomplete observations, which in semi-supervised learning correspond to feature vectors with a hidden class label. We could then wonder whether the MCPL algorithm can be used in a more general missing data setting [30]. In this chapter we address this question in the situation in which a parametric family is used to model the distribution of a random vector. The goal is then to find the optimal estimates for the parameters of the family, i.e. to select a density in the parametric class. It is not uncommon that for some of the available observations there is a subset of the components that is hidden. In this situation it is useful to have an estimator that is guaranteed to be a better alternative than what is obtained by simply discarding the incomplete observations. It is then necessary to select a baseline approach to establish in which sense the parameters we look for are optimal. Here we decide to measure the goodness of fit of an estimator with the log-likelihood on the full data set, by which we mean the data-set composed of both the observed and the hidden components. Therefore, we study the setting known as maximum likelihood (ML) estimation and the benchmark for the proposed estimator is the ML solution computed on the fully observed vectors only. This may in some cases be a naive choice, but serves to build our estimator.

Maximum likelihood estimators for the Gaussian density have been studied in a missing data framework first by T.W. Anderson in [2] in the case where the missingness takes place in a single block of components. In Section 6.1 we show how the estimates of the MCPL approach and those in [2] are related. A more complex situation is that of a *monotone sample*, which is a hierarchically nested block-structure of missing components. Closed form ML estimates for the Gaussian density were derived for a general monotone sample in this setting in [20], or considerably more clearly in [18]. In Section 6.2 the MCPL approach in this setting is studied.

6.1. A First Structure of the Missingness

Suppose N complete observations $\{(x_i, y_i)\}_{i=1}^{N}$ and M incomplete observations $\{x_j\}_{j=N+1}^{M}$ are available, where $x_i \in \mathbb{R}^{D-d}$ and $y_j \in \mathbb{R}^d$ for any $i = 1,\dots,N+M$ and $j = 1,\dots,N$. We refer to the data-set

as $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N} \cup \{x_j\}_{j=N+1}^{M}$. This is equivalent to formulating the pattern as in [2]:

$$\begin{aligned} &x_1, \dots, x_N, x_{N+1} \dots, x_{N+M} \\ &y_1, \dots, y_N . \end{aligned} \tag{6.1}$$

In other terms, there are $N+M$ observations for the first $D-d$ components of the random vector, while the last d are observed in only N samples. The missingness is thus block-shaped.
We can now model the observations as realizations of a D-dimensional Gaussian distribution:

$$p(x, y) \sim \mathcal{N}(\mu, \Sigma) .$$

where the mean and the covariance can be decomposed as

$$\Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{x,y} \\ \Sigma_{y,x} & \Sigma_y \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}. \tag{6.2}$$

Normality of the joint distribution implies that also the marginal distribution $p(x)$ and the conditional distribution $p(y|x)$ are Gaussian:

$$\begin{aligned} p(x) &\sim \mathcal{N}(\mu_x, \Sigma_x) , \\ p(y|x) &\sim \mathcal{N}(\mu_{y|x}, \Sigma_{y|x}) , \end{aligned}$$

where we introduced the following notation

$$\begin{aligned} \mu_{y|x} &= \mathbb{E}(y|x) &&= \mu_y + \Sigma_{y,x}\Sigma_x^{-1}(x - \mu_x) , \\ \Sigma_{y|x} &= \mathrm{Cov}(y|x) = &&\Sigma_y - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y} . \end{aligned} \tag{6.3}$$

In this framework we are interested in the guarantees that the MCPL approach can provide. Specifically, we would like to investigate if it is possible to find a classifier that better fits the data in the case of a finite sample. It will become clear from our analysis that this is indeed possible.

6.1.1. Maximum Likelihood Approach

A Maximum Likelihood (ML) approach to this incomplete data problem was proposed in [2]. Here we briefly discuss its formulation in order to introduce a few notions of interest.

First, the log-likelihood on (6.1) can be defined as

$$L(\mu, \Sigma|\mathscr{D}) = \sum_{i=1}^{N} \log g(x_i, y_i|\mu, \Sigma) + \sum_{j=N+1}^{N+M} \log g(x_j|\mu_x, \Sigma_x) , \tag{6.4}$$

where by $g(x_i, y_i|\mu, \Sigma)$ and $g(x_j|\mu_x, \Sigma_x)$ we denote the Gaussian density with mean and covariance of μ, Σ and μ_x, Σ_x respectively.
Then, $g(x, y|\mu, \Sigma)$ can be expressed as a product of the marginal and the conditional distributions, thus we rewrite (6.4) as

$$L(\mu, \Sigma|\mathscr{D}) = \sum_{i=1}^{N+M} \log g(x_i|\mu_x, \Sigma_x) + \sum_{i=1}^{N} \log g(y_i|\mu_{y|x_i}, \Sigma_{y|x_i}) . \tag{6.5}$$

It is interesting to observe that the conditional covariance does not depend on the specific value of the observed x, i.e. $\Sigma_{y|x_i} = \Sigma_{y|x}$. The ML estimates are then defined by maximizing (6.5) with respect to the parameters of the Gaussian density:

$$(\widehat{\mu}, \widehat{\Sigma}) = \underset{\mu, \Sigma}{\arg\max}\, L(\mu, \Sigma|\mathscr{D}) . \tag{6.6}$$

Now, the values of μ_x and Σ_x that solve the maximization defined in (6.6) are obtained by maximizing the first sum in (6.5), leading to the following estimates

$$\begin{aligned}\widehat{\mu}_x &= \frac{1}{N+M}\sum_{i=1}^{N+M} x_i\,,\\ \widehat{\Sigma}_x &= \frac{1}{N+M}\sum_{i=1}^{N+M}(x_i-\widehat{\mu}_x)(x_i-\widehat{\mu}_x)^T\,.\end{aligned} \tag{6.7}$$

Note that these are the standard ML estimates that would have been obtained if we observed only the first $D-d$ components of the random vector. On the other hand, the remaining parameters are estimated by maximizing the second sum in (6.5), which leads to

$$\begin{aligned}\widehat{\mu}_y &= \mu_y^C + \widehat{B}\,(\widehat{\mu}_x-\mu_x^C)\,,\\ \widehat{\Sigma}_{y,x} &= \widehat{\Sigma}_{x,y}^T = \widehat{B}\,\widehat{\Sigma}_x\,,\\ \widehat{\Sigma}_y &= \Sigma_y^C - \widehat{B}\,\Sigma_x^C\widehat{B}^T + \widehat{B}\,\widehat{\Sigma}_x\widehat{B}^T\,,\end{aligned} \tag{6.8}$$

where

$$\widehat{B} = \Sigma_{y,x}^C\big(\Sigma_x^C\big)^{-1}.$$

Here we denoted by μ^C, Σ^C the ML estimates obtained by maximizing the likelihood on the first N observations, that are the complete ones:

$$\begin{aligned}\mu_x^C &= \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \mu_y^C = \frac{1}{N}\sum_{i=1}^{N} y_i,\\ \Sigma_x^C &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x^C)(x_i-\mu_x^C)^T, \qquad \Sigma_y^C = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_y^C)(y_i-\mu_y^C)^T,\\ \Sigma_{x,y}^C &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x^C)(y_i-\mu_y^C)^T.\end{aligned} \tag{6.9}$$

We have thus derived the ML solutions to this particular missing data problem according to [2] and we can now concentrate on the MCPL method.

6.1.2. The MCPL Approach in the One Block Missing Data Setting

The MCPL approach was conceived for the task of semi-supervised learning, which can indeed be seen as a very specific missing data setting. We show how to extend this approach to a parameter estimation framework with a data-set with structure as in (6.1). We base our analysis on the same principles we have become familiar with, that are *contrast* and *pessimism*. In particular, we first define a particular complete log-likelihood that includes the incomplete observations. On the defined likelihood we contrast our estimates with the ML estimates computed by maximizing the likelihood on the complete observations only. These estimates were defined in (6.9). Afterwards, we assume the hidden components to be generated from the most adversarial distribution, i.e. the distribution that minimizes the improvement of our estimates on the complete-case estimates in terms of the previously defined likelihood. The final step is to maximize the objective function with respect to the parameters, which in this case are the mean and the covariance of the Gaussian density.

Let us first introduce some necessary notation. We call the unobserved parts that complete the corresponding incomplete observations as u_j for $j = N+1,\dots,N+M$. Hence, the complete j-th observation is denoted as (x_j,u_j). Then, we indicate the conditional probabilities of observing u_j given x_j by $p_j(u_j) = p(u_j|x_j)$ for any j. Now we can define the *complete log-likelihood* as

$$L\big(\mu,\Sigma,p_{N+1},\dots,p_{N+M}|\mathscr{D}\big) = \sum_{i=1}^{N}\log g(x_i,y_i|\mu,\Sigma) + \sum_{j=N+1}^{N+M}\int_{\mathbb{R}^d} p_j(u_j)\log g(x_j,u_j|\mu,\Sigma)\,du_j\,. \tag{6.10}$$

Since u_j appears as an integrating variable in this expression, we might as well ease the notation and write u instead. The idea is now to compare the pair (μ, Σ) with the ML estimates defined in (6.9), which we denoted by (μ^C, Σ^C). To this end, we define the contrastive likelihood as

$$\begin{aligned} CL(\mu, \Sigma, p_{N+1}, \dots, p_{N+M} \,|\, \mu^C, \Sigma^C, \mathscr{D}) \;=\; & \sum_{i=1}^{N} \log g(x_i, y_i|\mu, \Sigma) - \sum_{i=1}^{N} \log g(x_i, y_i|\mu^C, \Sigma^C) \\ + \sum_{j=N+1}^{N+M} \int_{\mathbb{R}^d} p_j(u) \log g(x_j, u_j|\mu, \Sigma)\, du - & \sum_{j=N+1}^{N+M} \int_{\mathbb{R}^d} p_j(u) \log g(x_j, u_j|\mu^C, \Sigma^C)\, du. \end{aligned} \tag{6.11}$$

The conditional probabilities $p_j(u)$ are chosen to be the most adversarial, which in mathematical terms translates to the following minimization task

$$CPL(\mu, \Sigma \,|\, \mu^C, \Sigma^C, \mathscr{D}) = \inf_{p_j \in \mathscr{P}} CL(\mu, \Sigma, p_{N+1}, \dots, p_{N+M} \,|\, \mu^C, \Sigma^C, \mathscr{D}), \tag{6.12}$$

where $\mathscr{P}$ is the set that contains any probability distribution function, i.e. the set of non-negative functions that sum to 1 on their domain. We are then ready to define the MCPL estimates.

Definition 6.1. *Let (μ^C, Σ^C) be the ML estimates based on the N complete observations $\{(x_i, y_i)\}_{i=1}^N$. Then the MCPL estimates $(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP})$ are defined as*

$$(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP}) \coloneqq \operatorname*{arg\,sup}_{\mu, \Sigma} CPL(\mu, \Sigma \,|\, \mu^C, \Sigma^C, \mathscr{D}). \tag{6.13}$$

Once again, the MCPL estimates are computed by solving a minimax problem, that in this setting is the following:

$$\sup_{\mu, \Sigma} \inf_{p_j \in \mathscr{P}} CL(\mu, \Sigma, p_{N+1}, \dots, p_{N+M} \,|\, \mu^C, \Sigma^C, \mathscr{D}).$$

This formulation is evidently similar to the previously treated cases, but there is a crucial difference. In earlier chapters, the pessimism was expressed by a labeling, which was a real number in $[0, 1]$ encoding the partial membership of an unlabeled observation to one of the classes. In this setting on the other hand the missing components are continuously distributed and thus the pessimism is represented by an adversarial distribution. The minimization with respect to these distributions, which we denoted by $p_j(\cdot)$, entails a considerable problem to the applicability of Sion's minimax theorem, i.e. Theorem 3.2. Indeed, the space of all probability densities is clearly not compact and this was one of the required properties. Therefore, we cannot take advantage of the same arguments we previously relied on. Fortunately, it is possible to show the results we are interested in without using minimax theorems. Indeed, we can plug the ML estimates (6.7), (6.8) in the contrastive likelihood and show that they solve the minimax problem. In the next theorem we state this result.

Theorem 6.2. *Suppose a data-set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \cup \{x_j\}_{j=N+1}^M$ with structure as in (6.1) is observed. Then, the estimates $(\widehat{\mu}, \widehat{\Sigma})$ defined in (6.7), (6.8) are such that*

$$(\widehat{\mu}, \widehat{\Sigma}) \;=\; \operatorname*{arg\,sup}_{\mu, \Sigma} CPL(\mu, \Sigma \,|\, \mu^C, \Sigma^C, \mathscr{D}).$$

Proof. First, note that we can follow Anderson's reasoning and decomposing the Gaussian density as follows

$$\begin{aligned} g(x, y|\mu, \Sigma) \;&=\; g(x|\mu_x, \Sigma_x) \cdot g(y|\mu_y + \Sigma_{y,x}\Sigma_x^{-1}(x - \mu_x), \Sigma_y - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}) \\ &=\; g(x|\mu_x, \Sigma_x) \cdot g(y|\mu_{y|x_j}, \Sigma_{y|x}). \end{aligned}$$

Similarly, we can define $\mu^C_{y|x_j}$ and $\Sigma^C_{y|x}$ as

$$\begin{aligned}\mu^C_{y|x_j} &= \mu^C_y + \Sigma^C_{y,x}\left(\Sigma^C_x\right)^{-1}\left(x_j - \mu^C_x\right),\\ \Sigma^C_{y|x} &= \Sigma^C_y - \Sigma^C_{y,x}\left(\Sigma^C_x\right)^{-1}\Sigma^C_{x,y}.\end{aligned}$$

Then, we can take advantage of this and rewrite (6.11) as

$$\begin{aligned}CL\big(\mu,\Sigma,p_{N+1},\ldots,p_{N+M}\,|\,\mu^C,\Sigma^C,\mathscr{D}\big) &= \sum_{i=1}^{N}\log g(x_i,y_i|\mu,\Sigma) + \sum_{j=N+1}^{N+M}\log g(x_j|\mu_x,\Sigma_x)\\ &+ \sum_{j=N+1}^{N+M}\int_{\mathbb{R}^d} p_j(u)\log g(u|\mu_{y|x_j},\Sigma_{y|x})du - \sum_{i=1}^{N}\log g(x_i,y_i|\mu^C,\Sigma^C)\\ &- \sum_{j=N+1}^{N+M}\log g(x_j|\mu^C_x,\Sigma^C_x) - \sum_{j=N+1}^{N+M}\int_{\mathbb{R}^d} p_j(u)\log g(u|\mu^C_{y|x_j},\Sigma^C_{y|x})du.\end{aligned} \tag{6.14}$$

It is then clear that the infimum affects only the difference of the two integrals. Moreover, we can further simplify the infimum by noting that the following holds

$$\begin{aligned}&\inf_{\{p_j\in\mathscr{P},j=N+1,\ldots,N+M\}}\sum_{j=N+1}^{N+M}\int_{\mathbb{R}^d} p_j(u)\left(\log g(u|\mu_{y|x_j},\Sigma_{y|x}) - \log g(u|\mu^C_{y|x_j},\Sigma^C_{y|x})\right)du\\ &= \sum_{j=N+1}^{N+M}\inf_{p_j\in\mathscr{P}}\int_{\mathbb{R}^d} p_j(u)\left(\log g(u|\mu_{y|x_j},\Sigma_{y|x}) - \log g(u|\mu^C_{y|x_j},\Sigma^C_{y|x})\right)du.\end{aligned}$$

and thus each minimization can be solved independently. As a consequence, the minimax problem can be written as

$$\max_{\mu,\Sigma}\Big\{L(\mu,\Sigma|\mathscr{D}) - L(\mu^C,\Sigma^C|\mathscr{D}) + \sum_{j=N+1}^{N+M}\inf_{p_j\in\mathscr{P}}\int_{\mathbb{R}^d} p_j(u)\big(\log g(u|\mu_{y|x_j},\Sigma_{y|x}) - \log g(u|\mu^C_{y|x_j},\Sigma^C_{y|x})\big)du\Big\},$$

where $L(\mu,\Sigma|\mathscr{D})$ was defined in (6.4), or equivalently in (6.5). Since each $p_j(u) = p(u|x_j)$ is a probability density function, we know that $p_j(u)\geq 0$ for any u in the domain and for any $j = N+1,\ldots,N+M$. On the other hand, neither $g(u|\mu_{y|x_j},\Sigma_{y|x})$ nor $g\big(u|\mu^C_{y|x_j},\Sigma^C_{y|x}\big)$ can dominate the other on all the domain. This is a straightforward consequence of the fact that densities have to sum to 1. This observation, together with the monotonicity of the logarithm, implies that there must exist a region where

$$\log g(u|\mu_{y|x_j},\Sigma_{y|x}) - \log g(u|\mu^C_{y|x_j},\Sigma^C_{y|x}) \leq 0,$$

in which equality holds if and only if the conditional distributions share the same parameters. It is then clear that it is always possible to choose a distribution $p_j(\cdot)$ that positions enough mass on this region in such a way that

$$\inf_{p_j\in\mathscr{P}}\int_{\mathbb{R}^d} p_j(u)\left(\log g(u|\mu_{y|x_j},\Sigma_{y|x}) - \log g\big(u|\mu^C_{y|x_j},\Sigma^C_{y|x}\big)\right)du \leq 0 \qquad j = N+1,\ldots,N+M,$$

where again equality holds iff $\mu_{y|x_j} = \mu^C_{y|x_j}$ for $j = N+1,\ldots,N+M$ and $\Sigma_{y|x} = \Sigma^C_{y|x}$.
Intuitively, the minimax problem is solved for a value of the estimates of the mean and covariance such that the infimum is null and the remaining terms in (6.14) are maximized. Choosing the parameters as (6.7) and (6.8) maximizes by definition the difference $L(\mu,\Sigma) - L(\mu^C,\Sigma^C)$, as $L(\mu^C,\Sigma^C)$ does not depend on the parameters. On the other hand, it can be seen that the conditional parameters are equal to the complete-case values by plugging $\left(\widehat{\mu},\widehat{\Sigma}\right)$ in the expressions for $\mu_{y|x_j}$ and $\Sigma_{y|x}$. The infimum is equal to 0 and thus maximized.
This therefore concludes the proof. □

We have thus showed that

$$\left(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP}\right) = \left(\widehat{\mu}, \widehat{\Sigma}\right).$$

Nonetheless, the formulation of the MCPL approach can be used to prove that the obtained solution fits the data strictly better than the complete-case estimates. The measure of performance we use here is the log-likelihood on the full data-set, which we define as follows

$$L\left(\mu, \Sigma \,|\, \mathscr{D}_{full}\right) = \sum_{i=1}^{N} \log g(x_i, y_i | \mu, \Sigma) + \sum_{i=N+1}^{N+M} \log g(x_i, y_i^* | \mu, \Sigma).$$

Here we denoted the full data-set as $\mathscr{D}_{full} = \left\{(x_i, y_i)\right\}_{i=1}^{N} \cup \left\{(x_i, y_i^*)\right\}_{i=N+1}^{N+M}$, where y_i^* are the true unobserved components. We can now state and prove the result in the next theorem.

Theorem 6.3. *The MCPL estimates, or alternatively the ML estimates derived in [2], are guaranteed to strictly improve on the complete case estimates in the following sense:*

$$L\left(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} \,|\, \mathscr{D}_{full}\right) > L\left(\mu^C, \Sigma^C \,|\, \mathscr{D}_{full}\right). \tag{6.15}$$

Proof. Define the adversarial probabilities that place mass 1 on the true value of the unobserved components, i.e.

$$p_j^*(u) = \delta_{y_j^*}(u) \qquad \text{for } j = N+1, \ldots, N+M,$$

where by $\delta_{y_j^*}(\cdot)$ we indicate Dirac's delta with center in y_j^*. Now, in the proof of Theorem 6.2 we showed that the conditional distributions for $p(y|x)$ are modeled with the same Gaussian density and are thus equal. Therefore, their difference is null independently of the chosen distributions $p_j(\cdot)$. As a consequence, the following equality holds:

$$\begin{aligned} L\left(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} \,|\, \mathscr{D}_{full}\right) - L\left(\mu^C, \Sigma^C \,|\, \mathscr{D}_{full}\right) = &\sum_{i=1}^{N} \log g(x_i, y_i | \widehat{\mu}_{CP}, \widehat{\Sigma}_{CP}) + \sum_{j=N+1}^{N+M} \log g(x_j | \widehat{\mu}_{CP}^x, \widehat{\Sigma}_{CP}^x) \\ &- \sum_{i=1}^{N} \log g(x_i, y_i | \mu^C, \Sigma^C) - \sum_{j=N+1}^{N+M} \log g(x_j | \mu_x^C, \Sigma_x^C). \end{aligned}$$

We know by Theorem 6.2 that the MCPL estimates are equivalent to the ML estimates and thus maximize the log-likelihood on the incomplete data-set $\mathscr{D}$. Moreover, the log-likelihood of a Gaussian density is strictly concave, hence the maximizer is unique. This implies that

$$L\left(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} \,|\, \mathscr{D}_{full}\right) - L\left(\mu^C, \Sigma^C \,|\, \mathscr{D}_{full}\right) > 0,$$

which is our thesis. □

This result holds for any choice of the distributions $p_j(\cdot)$ and thus for any value of the hidden components. Moreover, in [18] the author shows that $\widehat{\mu}$ is an unbiased estimator of the mean vector.

We have thus shown that it is possible to adapt the MCPL approach to this setting, and the result is a choice of parameters that fits the data strictly better than what we would obtain by discarding the incomplete samples. This meaningful result can also be interpreted as a property of the ML estimate discussed in [2]. In other terms, we have shown that the ML estimates are robust even to the most adversarial conditional distribution that may generate the hidden components.
In the next section we try to extend this result to a general monotone sample.

6.2. The MCPL Approach for Monotone Missing Data

We have thus studied and characterized the behaviour of the MCPL approach in the simplest case of a monotone sample, that is when a block of data is missing. It is then natural to investigate more complex settings. In order to do this, we continue using the family of Gaussian densities to model the joint probability distribution of the random vector. This restriction simplifies the analysis, which still becomes rather complicated in this setting.

Suppose now a monotone sample is available and denote it as

$$\mathscr{D} = \{x_i\}_{i=1}^{N} \cup \left\{x_i^{d_{k(i)}}\right\}_{i=N+1}^{N+M}. \tag{6.16}$$

This means that there are N complete observations denoted by $\{x_i\}_{i=1}^{N}$ for $x_i \in \mathbb{R}^D$, and M incomplete observations with a specific structure. Their pattern is described by the notation $x_i^{d_{k(i)}}$, which indicates that the first $d_{k(i)}$ components of x_i are observed, while the others are hidden. If a K-step pattern is assumed, then we can express the incomplete part of the sample as

$$\left\{x_i^{d_1}\right\}_{i=N+1}^{N+M_1} \cup \left\{x_i^{d_2}\right\}_{i=N+M_1+1}^{N+M_1+M_2} \cup \ldots \cup \left\{x_i^{d_K}\right\}_{i=N+M-M_K+1}^{N+M}, \tag{6.17}$$

in which we have that $D > d_1 > \ldots > d_K$ are the corresponding dimensions. In other words the $d_{k(i)}$ is used to identify that the observation x_i is in the $k(i)$-th step of the sample, where $k(\cdot)$ is a properly defined mapping.
The overall pattern is then similar to (6.1), where the first N observations are complete and successively the samples lose the last $(D-d_i)$ components, which are hidden. Once again, we want to explicitly compare our estimate with the complete-case estimate that is trained by ML on $\{x_i\}_{i=1}^{N}$. The hidden components are then assumed to be generated by distributions $\{p_j\}_{j=N+1}^{N+M}$ and therefore the contrastive likelihood for the monotone sample is

$$\begin{aligned} CL\left(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M}|\mu^C,\Sigma^C,\mathscr{D}\right) &= L(\mu,\Sigma|\{x_i\}_{i=1}^{N}) - L(\mu^C,\Sigma^C|\{x_i\}_{i=1}^{N}) \\ &+ \sum_{j=N+1}^{N+M} \int p_j\left(u_j^{d_{k(j)}}\right)\left(\log g\left(x_j^{d_{k(j)}},u_j^{d_{k(j)}}|\mu,\Sigma\right) - \log g\left(x_j^{d_{k(j)}},u_j^{d_{k(j)}}|\mu^C,\Sigma^C\right)\right) du_j^{d_{k(j)}}. \end{aligned} \tag{6.18}$$

Note that the densities $p_j(\cdot)$ have dimension $d_{k(j)}$ for $j = N+1,\ldots,N+M$.
Now, we should be sufficiently familiar with the method to foresee the next definition.

Definition 6.4. *Suppose we have a K-step monotone sample $\mathscr{D}$ with a pattern as in (6.16). Let the underlying probability distribution be modeled by the family of Gaussian densities. Then, the MCPL estimates are*

$$\left(\widehat{\mu}_{CP},\widehat{\Sigma}_{CP}\right) := \underset{(\mu,\Sigma)\in\Theta}{\arg\sup} \inf_{p_{N+1},\ldots,p_{N+M}} CL\left(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M}|\mu^C,\Sigma^C,\mathscr{D}\right). \tag{6.19}$$

We would intuitively expect a similar behaviour as in Section 6.1, meaning that the maximum likelihood estimates obtained in [18, 20] are the solution of the minimax (6.19). However, the line of reasoning we used in Section 6.1 cannot be applied to this more general framework. We illustrate this in Appendix B for the case of two nested blocks of missing data. The issue is that the difference of the integrals in the contrastive likelihood is not set to 0 by the ML estimate, so it is difficult to draw any conclusion on whether the proposed estimates are actually a solution of the minimax problem. We can fortunately still prove an interesting result, which is a corollary of Theorems 6.2, 6.3.

Corollary 6.5. *Let $\left(\widehat{\mu}_{1B},\widehat{\Sigma}_{1B}\right)$ be the solution to the minimax problem (6.13) based on the following data-set:*

$$\mathscr{D}_{1B} = \{x_i\}_{i=1}^{N} \cup \left\{x_i^{d_K}\right\}_{i=N+1}^{N+M}.$$

Then, the log-likelihood of $(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B})$ *on the complete data-set* $\mathscr{D}_{full} = \{x_i\}_{i=1}^{N} \cup \left\{ x_j^{d_{k(j)}}, u_j \right\}_{j=N+1}^{N+M}$ *is strictly larger than that of the complete-case estimates:*

$$L(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B} | \mathscr{D}_{full}) > L(\mu^C, \Sigma^C | \mathscr{D}_{full}).$$

The theorem can be easily proved by noting that we can as earlier isolate the likelihood on the N complete observations and on the first d_K components, which are always observed. The remaining conditional distributions are equal to the complete case estimates as previously shown in Theorem 6.2, hence the difference in the integral is null for each observation. The contrastive likelihood is then equal to the difference of the log-likelihoods on $\mathscr{D}_{1B}$, which is strictly positive and thus gives the thesis.

As previously indicated, this trick does not work for the ML solution that is computed by maximizing the likelihood on the whole observed data-set $\mathscr{D}$. It then becomes very complicated to characterize the solution because of the particularly intricate and tangled computations that arise in the analysis. Nevertheless, the fact that the contrastive likelihood at $(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B})$ has a positive value means that the solution of the minimax problem (6.19) must have the same property. So it must hold that

$$L(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} | \mathscr{D}_{full}) \geq L(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B} | \mathscr{D}_{full}), \tag{6.20}$$

because otherwise we would find a better option in $(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B})$. Clearly by Corollary 6.5 this implies that

$$L(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} | \mathscr{D}_{full}) > L(\mu^C, \Sigma^C | \mathscr{D}_{full}). \tag{6.21}$$

We conclude this chapter with the following conjecture

Conjecture 6.6. *Assume a K-step monotone sample* (6.16) *is observed and call* $(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP})$ *the solution of the minimax* (6.19)*. Then,* $(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP})$ *strictly improves over* $(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B})$*, which means*

$$L(\widehat{\mu}_{CP}, \widehat{\Sigma}_{CP} | \mathscr{D}_{full}) > L(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B} | \mathscr{D}_{full}).$$

We then conjecture that the additional observed components are expected to contain information that can be used by the algorithm to build a better estimator than $(\widehat{\mu}_{1B}, \widehat{\Sigma}_{1B})$. This is left formally unproven, but the previous theorem indeed shows that strict improvement over the complete-case estimate is certainly attained.

7

Conclusions

In Chapters 3, 4, 5 and 6 we investigated the possibilities of performing robust statistical inference with incomplete data. In other terms, we looked for estimators that are never worse than their supervised counterparts, i.e. the estimators computed by discarding the incomplete samples. In particular, we dedicated a large part of our work to semi-supervised learning, that is a classification task in which the missingness is limited to the class labels. A briefer but denser chapter was reserved to a more general missing data setting, in which the interest was in estimating a distribution function with a set of incomplete samples. In both settings, we built our estimators using *contrast* and *pessimism*. These principles allowed us to formally prove that it is indeed possible to build *safe* semi-supervised classifiers and parameter estimates. We analyzed both non-degradation properties and conditions for strict performance improvement in a finite sample setting. It has become clear that the contrastive pessimistic estimates can either tie or outperform the benchmark in all the frameworks we treated. This is stated in several results, such as Theorems 3.6, 3.8, 6.3 and Corollary 4.11.
This chapter is meant to address and discuss some interesting aspects and questions that arise from our precedent analysis of the contrastive pessimistic approach.

7.1. On the Contrastive Pessimism Approach for Semi-Supervised Classification

In Chapter 2 we discussed the importance of building a semi-supervised algorithm that cannot perform worse than its supervised counterpart. From a practical perspective this is necessary because the estimation of the accuracy of a classifier requires a large number of labeled objects, which are usually not available in real world applications. Therefore, it is very important to define an algorithm that has this guarantee *a priori*. In addition, most SSL algorithms base their theoretical properties on certain assumptions on the data, which are generally impossible to verify. Our extensions of the contrastive pessimistic approach address most of these issues very clearly. First, we have proved that it is possible to build a semi-supervised learner that is always at least as accurate as the supervised model without posing any assumption on the data-set. This robustness property is valid in terms of a specific performance measure, as discussed in Section 7.2, and in a finite sample setting, that is most common in applications. We have also characterized in which cases the improvement is strict and we discovered that for generative probabilistic models this happens almost surely under mild assumptions on the model. Then, we have shown how to build such a classifier using the contrastive pessimistic approach, of which we carefully studied the main features.

7.1.1. A Comparison with Other Attempts to Safe SSL

In Section 2.3.3 we discussed the SV4Ms [31], a semi-supervised version of the support vector machines with a robustness guarantee. Our approach differs from SV4Ms in several aspects. The crucial difference lies in the assumption that is made in the construction of the SV4Ms: the ground truth labeling of the unlabeled observations is among K low density separators that are identifiable from the data-set. This is a rather strong assumption on the data and most importantly it cannot be verified. This violates one of the founding principles of our analysis, that is not assuming any particular structure of the data-set. In this sense, the robustness guarantees we proved are much more meaningful, as they hold on any data-set and do not depend on any assumption.
Then, the SV4Ms are shown to choose one of the K low density separators such that the fraction of misclassified samples among the unlabeled objects is smaller than that of the supervised SVM. This is different from our approach in two ways. First, our results are in terms of a chosen performance measure that is not the misclassification error. Then, the non degradation property of the contrastive pessimistic approach holds on the full data-set and not only on the unlabeled observations. While a result on the misclassification rate is desirable, the fact that it holds on the unlabeled data-set only does not entail that the classifier is never worse than the supervised SVM on the full data-set. This follows from observing that the supervised SVM classifier is optimal on the labeled data-set and thus generally outperforms the semi-supervised learner on those samples. The property of the contrastive pessimistic approach is then a stronger one. In addition, no conditions on strict improvement are shown for SV4Ms, while we proved several results in this direction throughout this work. We have then highlighted several points that are in favor of the contrastive pessimistic approach, while the only negative one is that SV4Ms is robust in terms of misclassification error.

Two other algorithms we discussed are the implicitly constrained least squares classifier [27] and the weighted likelihood based approach [22]. These methods build two learners that are asymptotically robust. In our work, we focused on proving this property in a finite sample setting, which does not assume an infinite number of samples and we believe is more meaningful.

7.2. On the Choice of the Performance Measure

In every setting we have studied, the robustness guarantees are expressed in terms of a specific quantity that represents the performance of the learner. The measure is chosen to be the most natural one in the particular setting of interest. The parameters of a supervised generative probabilistic classifier for instance can be learned by maximizing the likelihood on the labeled data-set. It is thus natural to express the possible improvement obtained with the addition of the unlabeled samples by the likelihood on the full data-set, which contains also the hidden class labels. An increased likelihood thus expresses an improved accuracy of the classifier. In other terms, the semi-supervised model is closer in terms of likelihood to the oracle solution, which is the optimal estimate trained on the fully labeled data-set. The same comments hold for least squares classification, but where the performance measure is the square loss. It has to be noted once again that the main interest in Pattern Recognition is a low misclassification error. The results we proved are then a surrogate of what we are really interested in. In our view an increased accuracy in the terms we discussed is a strong hint to a lower classification error. This is not straightforward and there is not a formal characterization of the relation between the two quantities. For a more detailed discussion on this we point to [35, 36]. Furthermore, there is empirical evidence [28, 33] to support the fact that the contrastive pessimistic approach successfully builds a more accurate classifier also in terms of the classification error.
There are significant indications that the guarantees we proved are meaningful also in terms of misclassification error, but the fact that this is not formally explained may still be considered as a weak feature of the proposed method. In a way, it would be interesting to adapt and modify the method in order to have this property. It is however unclear how this can be done and whether it is at all

possible. The only algorithm we are aware of with such a guarantee is [31], which as discussed has a rather strong condition on the data. A meaningful further research would then be an investigation on whether an equivalent result to that of [31] can be obtained with weaker or no assumptions.

7.2.1. Induction vs. Transduction

We have discussed the difference between inductive and transductive learning in Chapter 2. Basically, in a semi-supervised framework the former is concerned with finding a general classification rule that can be applied on unseen data, while the latter focuses on predicting the labels of some of the unlabeled observations that are available for training. It is then interesting to observe that the contrastive pessimistic approach defines an inductive decision rule which is guaranteed to never be worse than the supervised model on the training set. As a consequence, this holds in particular on the unlabeled data-set and hence the contrastive pessimistic approach is always at least as good as the supervised learner in a transductive sense.
An inductive semi-supervised classifier may be considered really safe when we know that on unseen data there is never a performance degradation compared the supervised model. It then follows that the results obtained for the contrastive pessimistic approach hint to this behaviour on unseen data, but do not formally imply it. We have to observe that to the best of our knowledge there are no semi-supervised classifiers with guarantees on unseen data in a finite sample setting. We could imagine that this may not be possible in some cases, but it is very difficult to draw any conclusion in this direction. Anyway, we would expect that the contrastive pessimistic classifier may hardly show a deteriorated performance on unseen data if it is trained on a large set of unlabeled observations.

7.3. On Parametric Inference with a Monotone Sample

The principles of contrast and pessimism were applied to a missing data setting in Chapter 6. We proved that the maximum likelihood estimates described in [2] are closely related to our approach. Indeed, in the case of a missing block of data they solve the minimax problem that defines the contrastive pessimistic approach. This result can be seen as an interesting property on the estimates found in [2]. In other terms, we showed that these estimates fit better the data and have a larger log-likelihood on the complete data-set than the ML estimates obtained by using the complete observations only. Once again, this result does not necessarily mean that our estimate is in absolute terms better than the complete case one, but it still represents an interesting property on the robustness of an estimate.

7.4. Future Research

We believe this work has addressed and answered a large part of the research questions that we formulated in Chapter 1. Nevertheless, our findings open some further questions that we believe would be worth investigating.

In the semi-supervised learning case, kernel methods represent a possible future direction. We would expect that, under some regularity conditions on the kernel, the contrastive pessimistic approach can be applied to build a robust non-linear classifier. This would in a way conclude the analysis that was started in [29, 33] and that we extended in this work.
We have argued that, for obvious reasons, the contrastive pessimistic approach is rather conservative. It would then be interesting to modify its formulation in order to relax this aspect. For instance, we have shown that the contrastive pessimistic semi-supervised version of a generative model based on exponential families outperforms its supervised counterpart with probability 1. We can then wonder if we can relax the requirements of the method while maintaining this guarantee. This would clearly enable greater improvements, while maintaining the key properties of the method.

The last indications for future research concern the missing data setting described in Chapter 6 and Appendix B. This setting in a way still has to be fully understood. For instance, a complete characterization of the contrastive pessimistic solution for the case of a monotone sample still has to be determined. This could be done either theoretically or empirically. From an empirical perspective this would mean testing whether the approximate solution turns out to be of a known form. On the other hand, a theoretical analysis seems to be rather difficult at the moment. Finally, the ideal destination of this research is the extension to any structure of the randomness.

A

Topological Spaces, Semicontinuous Functions and Quasi-convexity

The main result on minimax theory we discuss, which is Theorem 3.2, requires a few technical notions. In this appendix we give a brief introduction to some of the necessary definitions, which are based on [24]. We begin with the definition of topology.

Definition A.1. *Given a set X, by a topology is meant a system τ of subsets $G \subset X$, called open sets, with the following two properties:*

i) the set X itself and the empty set $\emptyset$ are in τ;

ii) Arbitrary unions $\cup_\alpha G_\alpha$ and finite intersections $\cap_{k=1}^n G_k$ of open sets belong to τ.

A topological space T is then a pair (X,τ), where X is a set and τ a topology defined in X. Note that every metric space is a topological space as well.
The definition of topological space T allows us to introduce the concept of neighborhood of a point $x \in T$ as any open set $G \subset T$ containing x. With this notion we can define the following class of functions.

Definition A.2. *A function f defined on a topological space T is said to be lower semicontinuous at a point x_0 if, given any $\epsilon > 0$, there exists a neighborhood of x_0 in which $f(x) > f(x_0) - \epsilon$. Similarly, f is said to be upper semicontinuous at a point $x_0 \in T$ if, given any $\epsilon > 0$, there exists a neighborhood of x_0 in which $f(x) < f(x_0) + \epsilon$.*

A function is then said lower (upper) semicontinuous if it is lower (upper) semicontinuous at each point of its domain. Note that semicontinuity is a weaker requirement than continuity, so a continuous function is automatically semicontinuous.
We can now define a topological linear space T as a space that satisfies the definitions of linear space, topological space, as well as satisfying continuity with respect to the topology of the operations of addition of elements of T and multiplication of elements of T by scalars. For what concerns our needs, it is not necessary to investigate the meaning of the last requirement any further. It is sufficient to know that every normed linear space is a topological linear space. Indeed, we will deal with vector spaces, on which it is enough to define the topology generated by the set of spheres centered in zero. This is indeed the Euclidean topology.
The second class of functions we introduce is the following.

Definition A.3. *A function f on $\mathscr{X}$, where $\mathscr{X}$ is a convex subset of a linear space as for instance $\mathbb{R}^n$, is said to be quasi-convex in $\mathscr{X}$ if the level sets $\{x \in \mathscr{X} : f(x) \leq \lambda\}$ are convex for any $\lambda \in \mathbb{R}$.*

Equivalently, we can define quasi-concave functions. A convex (concave) function is also quasi-convex (concave), as convexity (concavity) of the level sets follows by definition.

B

Additional Material on the MCPL in the Two Blocks Missing Data Setting

In Section 6.2 we discussed the application of the MCPL approach to a monotone sample, that is a data-set with a hierarchical structure of missing components as in (6.16). In the analysis we stated that the key argument used for the proof of Theorem 6.2 is not valid even in a particular monotone sample as the case of two blocks of missing data. Here we formally illustrate in which sense this is the case.

Assume the observations have the following structure

$$\begin{aligned} &x_1, \dots, x_N, x_{N+1}, \dots, x_{N+M}, x_{N+M+1}, \dots, x_{N+M+L} \\ &y_1, \dots, y_N, y_{N+1}, \dots, y_{N+M} \\ &z_1, \dots, z_N \end{aligned} \tag{B.1}$$

where $x_i \in \mathbb{R}^{d_x}$, $y_i \in \mathbb{R}^{d_y}$ and $z_i \in \mathbb{R}^{d_z}$, with $D = d_x + d_y + d_z$. We call this data-set $\mathscr{D}$, which is composed N complete observations, M incomplete samples of (x_j, y_j) and L of the form x_l.
The random vector (x, y, z) is modeled as a multivariate D-dimensional Gaussian density:

$$p(x, y, z) \sim \mathcal{N}(\mu, \Sigma),$$

for which the mean vector and the covariance matrix are to be estimated. This choice implies normality of the marginal density $p(x)$ and of the conditional $p(y|x)$ as described in (6.3), but also of other conditionals:

$$\begin{aligned} p(y, z|x) &\sim \mathcal{N}(\mu_{yz|x}, \Sigma_{yz|x}), \\ p(z|x, y) &\sim \mathcal{N}(\mu_{z|xy}, \Sigma_{z|xy}), \end{aligned} \tag{B.2}$$

where

$$\begin{aligned} \mu_{yz|x} &= \mu_{yz} + \Sigma_{yz,x}\Sigma_x^{-1}(x - \mu_x), & \Sigma_{yz|x} &= \Sigma_{yz,yz} - \Sigma_{yz,x}\Sigma_x^{-1}\Sigma_{x,yz}, \\ \mu_{z|xy} &= \mu_z + \Sigma_{z,xy}\Sigma_{xy,xy}^{-1}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \mu_{xy}\right), & \Sigma_{z|xy} &= \Sigma_z - \Sigma_{z,xy}\Sigma_{xy,xy}^{-1}\Sigma_{xy,z}. \end{aligned} \tag{B.3}$$

in which we used the same notation as in (6.2), but adapted to the current setting.

B.1. Maximum Likelihood Estimates in the Two Blocks Setting

The ML estimates for μ and Σ can be found either by working out the recursive close form solution in [18], or by applying the reasoning in [2]. Both options clearly lead to the same result. Here we

briefly extend the reasoning in [2] and we start by formulating the log-likelihood as

$$L(\mu,\Sigma|\mathscr{D}) = \sum_{i=1}^{N} \log g(x_i,y_i,z_i|\mu,\Sigma) + \sum_{j=N+1}^{N+M} \log g(x_j,y_j|\mu_{xy},\Sigma_{xy,xy}) + \sum_{l=N+M+1}^{N+M+L} \log g(x_l|\mu_x,\Sigma_x). \tag{B.4}$$

Then, we can rewrite (B.4) using (B.2) and (6.3) as

$$L(\mu,\Sigma|\mathscr{D}) = \sum_{i=1}^{N+M+L} \log g(x_i|\mu_x,\Sigma_x) + \sum_{j=1}^{N+M} \log g(y_j|\mu_{y|x_j},\Sigma_{y|x}) + \sum_{k=1}^{N} \log g(z_k|\mu_{z|x_k y_k},\Sigma_{z|xy}). \tag{B.5}$$

The idea is now to compute the MLE for the mean and covariance of random vector x from the first summation, then to compute the values of the conditional parameters that maximize the second and third summation in the log-likelihood. Now, this is made precise by the introduction of auxiliary parameters, but here we report the results only and we point to [2] for the precise argument.
By maximizing the first summation in (B.5), the following estimates are obtained

$$\begin{aligned} \widehat{\mu}_x &= \frac{1}{N+M+L} \sum_{i=1}^{N+M+L} x_i, \\ \widehat{\Sigma}_x &= \frac{1}{N+M+L} \sum_{i=1}^{N+M+L} (x_i-\widehat{\mu}_x)(x_i-\widehat{\mu}_x)^T. \end{aligned} \tag{B.6}$$

Maximizing the second summation in (B.5) we obtain

$$\begin{aligned} \widehat{\mu}_y &= \mu_y^{C_{NM}} + \widehat{B}_y(\widehat{\mu}_x - \mu_x^{C_{NM}}), \\ \widehat{\Sigma}_{y,x} &= \widehat{\Sigma}_{x,y}^T = \widehat{B}_y \widehat{\Sigma}_x, \\ \widehat{\Sigma}_y &= \Sigma_y^{C_{NM}} - \widehat{B}_y \Sigma_x^{C_{NM}} \widehat{B}_y^T + \widehat{B}_y \widehat{\Sigma}_x \widehat{B}_y^T, \end{aligned} \tag{B.7}$$

where $\widehat{B}_y = \Sigma_{y,x}^{C_{N+M}} (\Sigma_x^{C_{N+M}})^{-1}$ and we used the superscript C_{NM} to denote the sample mean and covariance computed on the first $N+M$ observations.
Finally, maximizing the third summation in (B.5) we find the following estimates

$$\begin{aligned} \widehat{\mu}_z &= \mu_z^{C_N} + \widehat{B}_z(\widehat{\mu}_{xy} - \mu_{xy}^{C_N}), \\ \widehat{\Sigma}_{z,xy} &= \widehat{\Sigma}_{xy,z}^T = \widehat{B}_z \widehat{\Sigma}_{xy,xy}, \\ \widehat{\Sigma}_z &= \Sigma_z^{C_N} - \widehat{B}_z \Sigma_{xy,xy}^{C_N} \widehat{B}_z^T + \widehat{B}_z \widehat{\Sigma}_{xy,xy} \widehat{B}_z^T, \end{aligned} \tag{B.8}$$

where $\widehat{B}_z = \Sigma_{z,xy}^{C_N} (\Sigma_{xy,xy}^{C_N})^{-1}$ and again the superscript C_N denotes the empirical mean and covariance on the first N observations, as shown in (6.9).

B.2. MCPL Approach in the Two Blocks Setting

The same reasoning discussed in Sections 6.1.2 and 6.2 is now used to define the MCPL approach in this setting.
First, we introduce densities $p_j(u) = p(u|x_j,y_j)$ and $q_l(u,v) = p(u,v|x_l)$ for $j = N+1,\dots,N+M$ and $l = N+M+1,\dots,N+M+L$. Then, the complete likelihood is defined as

$$L\big(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mathscr{D}\big) = \sum_{i=1}^{N} \log g(x_i,y_i,z_i|\mu,\Sigma) + \sum_{j=N+1}^{N+M} \int_{\mathbb{R}^{d_z}} p_j(v) \log g(x_j,y_j,v|\mu,\Sigma)\, dv + \sum_{l=N+M+1}^{N+M+L} \int_{\mathbb{R}^{d_y+d_z}} q_l(u,v) \log g(x_j,u,v|\mu,\Sigma)\, du dv.$$

The relations reported in (B.2) allow us to rewrite the previous equation as

$$L(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mathscr{D}) = L(\mu,\Sigma|\mathscr{D}) + \sum_{j=N+1}^{N+M}\int_{\mathbb{R}^{d_z}} p_j(v)\log g(v|\mu_{z|x_jy_j},\Sigma_{z|xy})\,dv + \sum_{l=N+M+1}^{N+M+L}\int_{\mathbb{R}^{d_y+d_z}} q_l(u,v)\log g(u,v|\mu_{yz|x_l},\Sigma_{yz|x})\,du\,dv.$$

This follows by expressing the joint densities in the logarithms as product of a marginal and a conditional distribution and pulling out of the integrals the known terms. We can then define the contrastive likelihood as

$$CL(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mu^{C_N},\Sigma^{C_N},\mathscr{D}) = L(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mathscr{D}) - L(\mu^{C_N},\Sigma^{C_N},\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mathscr{D}).$$

Finally, the MCPL estimate is computed as

$$\arg\sup_{\mu,\Sigma}\inf_{\substack{p_j\in\mathscr{P}\\ q_l\in\mathscr{Q}}} CL(\mu,\Sigma,\{p_j\}_{j=N+1}^{N+M},\{q_l\}_{l=N+M+1}^{N+M+L}|\mu^{C_N},\Sigma^{C_N},\mathscr{D}),$$

where $\mathscr{P}$ and $\mathscr{Q}$ are the sets containing any probability distribution function on $\mathbb{R}^{d_z}$ and $\mathbb{R}^{d_y+d_z}$ respectively.

B.2.1. Plugging the ML Solution in the CPL

Motivated by the one block case we try to plug the ML estimates obtained in section B.1 into the contrastive pessimistic likelihood so to check whether this is a solution of the minimax problem. It follows that ideally we would like to find parameters that maximize the standard log-likelihood $L(\mu,\Sigma|\mathscr{D})$, while setting the minimization tasks to 0 independently of the choice of the adversarial distributions. We show that this does not happen with the ML estimates.

Proposition B.1. *The ML estimates defined in Section B.1 are such that*

$$\int_{\mathbb{R}^{d_z}} p_j(v)\log g(v|\widehat{\mu}_{z|x_jy_j},\widehat{\Sigma}_{z|xy})\,dv = \int_{\mathbb{R}^{d_z}} p_j(v)\log g(v|\mu^{C_N}_{z|x_jy_j},\Sigma^{C_N}_{z|xy})\,dv.$$

for any $j = N+1,\dots,N+M$.

Proof. We have already argued that the equality of the integrals is attained if and only if the two conditional Gaussian densities are equal, i.e. have the same parameters. Hence, we show that this holds. Substituting the ML estimates derived in Section B.1, the conditional means take the form:

$$\begin{aligned}\widehat{\mu}_{z|x_jy_j} &= \mu_z^{C_N} + \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\big(\widehat{\mu}_{xy}-\mu^{C_N}_{xy}\big) + \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\widehat{\Sigma}_{xy,xy}\big(\widehat{\Sigma}_{xy,xy}\big)^{-1}\left(\begin{pmatrix}x_j\\y_j\end{pmatrix}-\widehat{\mu}_{xy}\right)\\ &= \mu_z^{C_N} + \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\left(\begin{pmatrix}x_j\\y_j\end{pmatrix}-\mu^{C_N}_{xy}\right) = \mu^{C_N}_{z|x_jy_j}.\end{aligned}$$

Analogously, the conditional covariance is

$$\begin{aligned}\widehat{\Sigma}_{z|xy} &= \Sigma_z^{C_N} - \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\big(\Sigma^{C_N}_{xy,xy}+\widehat{\Sigma}_{xy,xy}\big)\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\Sigma^{C_N}_{z,xy}\\ &\quad - \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\widehat{\Sigma}_{xy,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\widehat{\Sigma}_{xy,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\Sigma^{C_N}_{xy,z}\\ &= \Sigma_z^{C_N} - \Sigma^{C_N}_{z,xy}\big(\Sigma^{C_N}_{xy,xy}\big)^{-1}\Sigma^{C_N}_{xy,xy} = \Sigma^{C_N}_{z|xy},\end{aligned}$$

which concludes the proof. □

We have therefore shown that choosing the estimates as in section B.1 the first infimum attains its maximum value, that is 0. If this was the case also for the second set of integrals, then the we would have found the MCPL solution. However, this is not the case. In order to show this, let us focus on the integrals in two variables:

$$\int_{\mathbb{R}^{d_y+d_z}} q_l(u,v)\Big\{\log g(u,v|\widehat{\mu}_{yz|x_l},\widehat{\Sigma}_{yz|x}) - \log g(u,v|\mu^{C_N}_{yz|x_l},\Sigma^{C_N}_{yz|x})\Big\}dudv.$$

In this case the computations are much more complicated.

Proposition B.2. *The ML estimates defined in Section B.1 are such that*

$$\int_{\mathbb{R}^{d_y+d_z}} q_l(u,v)\log g(u,v|\widehat{\mu}_{yz|x_l},\widehat{\Sigma}_{yz|x})dudv \neq \int_{\mathbb{R}^{d_y+d_z}} q_l(u,v)\log g(u,v|\mu^{C_N}_{yz|x_l},\Sigma^{C_N}_{yz|x})\Big\}dudv,$$

for any $l = N+M+1,\dots,N+M+L$.

Proof. It is sufficient to prove that the mean vectors are not equal, i.e.

$$\widehat{\mu}_{yz|x_l} \neq \mu^{C_N}_{yz|x_l}.$$

for all $l = N+M+1,\dots,N+M+L$. We then consider

$$\widehat{\mu}_{yz|x} = \widehat{\mu}_{yz} + \widehat{\Sigma}_{yz,x}\widehat{\Sigma}_x^{-1}(x_l - \widehat{\mu}_x).$$

It is complicated to obtain a manageable expression for the last d_z components of $\widehat{\mu}_{yz|x}$, hence we focus on the first d_y only:

$$\widehat{\mu}_{y|x} = \widehat{\mu}_y + \widehat{\Sigma}_{y,x}\widehat{\Sigma}_x^{-1}(x_l - \widehat{\mu}_x), \tag{B.9}$$

where we used the block formulation of $\widehat{\Sigma}_{yz,x}$. This can be rewritten as

$$\widehat{\mu}_{y|x} = \mu_y^{C_{N+M}} + \Sigma_{y,x}^{C_{N+M}}\left(\Sigma_x^{C_{N+M}}\right)^{-1}(x_l - \mu_x^{C_{N+M}}),$$

which is obtained by substituting the ML estimates in (B.9). Since the complete-case estimates satisfy

$$\widehat{\mu}_{y|x}^{C_N} = \mu_y^{C_N} + \Sigma_{y,x}^{C_N}\left(\Sigma_x^{C_N}\right)^{-1}(x_l - \mu_x^{C_N}),$$

we see that the conditional means are not equal. □

This shows that the second infimum is not set to 0 by the ML estimates. Note that this does not mean that the ML solution solves the MCPL minimax problem, but it makes it much more complicated to prove.

Bibliography

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[2] T. W. Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203, 1957.

[3] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Mathematics and its applications (D. Reidel Publishing Company): East European series. Editura Academiei, 1978.

[4] P. L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity , classification , and risk bounds. *Journal of the American Statistical Association*, 2003.

[5] Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *ICML*, 2012.

[6] Andrea L. Bertozzi, Xiyang Luo, A. M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. 2017.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[8] Lawrence D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986.

[9] Vittorio Castelli and Thomas M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16:105–111, 1995.

[10] Vittorio Castelli and Thomas M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Information Theory*, 42: 2102–2117, 1996.

[11] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[13] Fabio Cozman and Ira Cohen. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 4, pages 57–72. MIT Press, September 2006.

[14] Nello Cristianini. Kernel methods for pattern analysis. In *ICTAI*, 2003.

[15] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.

[16] R. P. W. Duin. Prtools - version 3.0 - a matlab toolbox for pattern recognition. In *Proceedings of SPIE*, page 1331, 2000.

[17] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 385–394, 2009.

[18] Hironori Fujisawa. A note on the maximum likelihood estimators for multivariate normal distribution with monotone data. *Communications in Statistics - Theory and Methods*, 24(6):1377–1382, 1995.

[19] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[20] K.G. Jinadasa and D.S. Tracy. Maximum likelihood estimation for multivariate normal distribution with monotone sample. *Communications in Statistics - Theory and Methods*, 21(1):41–50, 1992.

[21] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

[22] Masanori Kawakita and Jun'ichi Takeuchi. Safe semi-supervised learning based on weighted likelihood. *Neural networks : the official journal of the International Neural Network Society*, 53:146–64, 2014.

[23] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:226–239, 1998.

[24] A.N. Kolmogorov, S.V. Fomin, and R.A. Silverman. *Introductory Real Analysis*. Dover Books on Mathematics. Dover Publications, 2012.

[25] Hidetoshi Komiya. Elementary proof for sion's minimax theorem. *Kodai Mathematical Journal*, 11(1):5–7, 1988.

[26] Jesse H. Krijthe and Marco Loog. Implicitly constrained semi-supervised linear discriminant analysis. *2014 22nd International Conference on Pattern Recognition*, pages 3762–3767, 2014.

[27] Jesse H. Krijthe and Marco Loog. Robust semi-supervised least squares classification by implicit constraints. *Pattern Recognition*, 63:115–126, 2017.

[28] Jesse H. Krijthe and Marco Loog. Projected estimators for robust semi-supervised classification. *Machine Learning*, 106:993–1008, 2017.

[29] Jesse H. Krijthe and Marco Loog. The pessimistic limits and possibilities of margin-based losses in semi-supervised learning. *NIPS*, 2018.

[30] Jesse H. Krijthe, Marco Loog, and Bo Markussen. Missing data and the maximum contrastive pessimistic likelihood. In *The 27th Nordic Conference in Mathematical Statistics: Abstracts*, page 47, Tartu, Estonia, June 2018.

[31] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:175–188, 2011.

[32] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-80254-9.

[33] Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:462–475, 2016.

[34] Marco Loog. Supervised classification: Quite a brief overview. *CoRR*, abs/1710.09230, 2017.

[35] Marco Loog and Are Charles Jensen. Semi-supervised nearest mean classification through a constrained log-likelihood. *IEEE Transactions on Neural Networks and Learning Systems*, 26 (5):995–1006, 2015.

[36] Marco Loog, Jesse H Krijthe, and Are C Jensen. On measuring and quantifying performance: Error rates, surrogate loss, and an example in ssl. In *Handbook of Pattern Recognition and Computer Vision*, pages 53–68. World Scientific, 2016.

[37] 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.

[38] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.

[39] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[40] Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *ICML*, 2009.

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

[42] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[43] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[44] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[45] Nataliya Sokolovska, Olivier Cappé, and François Yvon. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *ICML*, 2008.

[46] Sergios Theodoridis and Konstantinos Koutroumbas. Chapter 1 - introduction. In Sergios Theodoridis and Konstantinos Koutroumbas, editors, *Pattern Recognition (Fourth Edition)*, pages 1 – 12. Academic Press, Boston, fourth edition edition, 2009.

[47] Sergios Theodoridis and Konstantinos Koutroumbas. Chapter 2 - classifiers based on bayes decision theory. In Sergios Theodoridis and Konstantinos Koutroumbas, editors, *Pattern Recognition (Fourth Edition)*, pages 13 – 89. Academic Press, Boston, fourth edition edition, 2009.

[48] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[49] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[50] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[51] John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

[52] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 2004.

[53] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

[54] Xiaojin Zhu. *Semi-supervised Learning with Graphs*. PhD thesis, Pittsburgh, PA, USA, 2005.

[55] Xiaojin Zhu. Semi-supervised learning literature survey. 2006.

[56] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.