

Sample Complexity of Nonparametric Semi-Supervised Learning

Chen Dan, Liu Leqi, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing

Carnegie Mellon University

September 11, 2018

Abstract

We study the sample complexity of semi-supervised learning (SSL) and introduce new assumptions based on the mismatch between a mixture model learned from unlabeled data and the true mixture model induced by the (unknown) class conditional distributions. Under these assumptions, we establish an $\Omega(K \log K)$ labeled sample complexity bound without imposing parametric assumptions, where K is the number of classes. Our results suggest that even in nonparametric settings it is possible to learn a near-optimal classifier using only a few labeled samples. Unlike previous theoretical work which focuses on binary classification, we consider general multiclass classification ($K > 2$), which requires solving a difficult permutation learning problem. This permutation defines a classifier whose classification error is controlled by the Wasserstein distance between mixing measures, and we provide finite-sample results characterizing the behaviour of the excess risk of this classifier. Finally, we describe three algorithms for computing these estimators based on a connection to bipartite graph matching, and perform experiments to illustrate the superiority of the MLE over the majority vote estimator.

1 Introduction

With the rapid growth of modern datasets and increasingly passive collection of data, labeled data is becoming more and more expensive to obtain while unlabeled data remains cheap and plentiful in many applications. Leveraging unlabeled data to improve the predictions of a machine learning system is the problem of semi-supervised learning (SSL), which has been the source of many empirical successes (Blum and Mitchell, 1998; Kingma et al., 2014; Dai et al., 2017) and theoretical inquiries (Azizyan et al., 2013; Castelli and Cover, 1995, 1996; Cozman et al., 2003; Kääriäinen, 2005; Niyogi, 2013; Rigollet, 2007; Singh et al., 2009; Wasserman and Lafferty, 2008; Zhu et al., 2003). Commonly studied assumptions include identifiability of the class conditional distributions (Castelli and Cover, 1995, 1996), the cluster assumption (Rigollet, 2007; Singh et al., 2009) and the manifold assumption (Zhu et al., 2003; Wasserman and Lafferty, 2008; Niyogi, 2013). In this work, we propose a new type of assumption that loosely combines ideas from both the identifiability and cluster assumption perspectives. Importantly, we consider the general multiclass ($K > 2$) scenario, which introduces significant complications. In this setting, we study the sample complexity and rates of convergence for SSL and propose simple algorithms to implement the proposed estimators.

The basic question behind SSL is to connect the marginal distribution over the unlabeled data $\mathbb{P}(X)$ to the regression function $\mathbb{P}(Y | X)$. We consider multiclass classification, so that $Y \in \mathcal{Y} = \{\alpha_1, \dots, \alpha_K\}$ for some $K \geq 2$. In order to motivate our perspective, let F^* denote the marginal density of the unlabeled samples and suppose that F^* can be written as a mixture model

$$F^*(x) = \sum_{b=1}^K \lambda_b f_b(x). \quad (1)$$

Crucially, we *do not* assume that each f_b corresponds to some f_k^* , where f_k^* is the density of the k th class conditional $\mathbb{P}(X | Y = \alpha_k)$. Nor do we assume that λ_b corresponds to some λ_k^* where $\lambda_k^* = \mathbb{P}(Y = \alpha_k)$. We

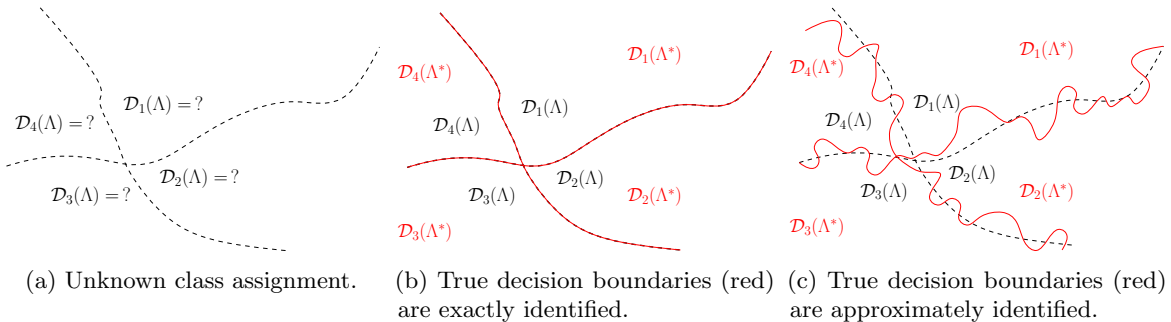


Figure 1: Illustration of the main idea for $K = 4$. The decision boundaries learned from the unlabeled data (cf. (1)) are depicted by the dashed black lines and the true decision boundaries are depicted by the solid red lines. (a) The unlabeled data is used to learn some approximate decision boundaries through the mixture model Λ . Even with the decision boundaries, it is not known which class each region corresponds to. The labeled data is used to learn this assignment. (b) Previous work assumes that the true and approximate decision boundaries are the same. (c) In the current work, we assume that the true decision boundaries are unknown, but that it is possible to learn a mixture model that approximates the true boundaries using unlabeled data.

assume that the number of mixture components K is the same as the number of classes. Assuming the unlabeled data can be used to learn the mixture model (1), the question becomes *when is this mixture model useful for predicting Y ?* Figure 1 illustrates an idealized example.

In an early series of papers, [Castelli and Cover \(1995, 1996\)](#) considered this question under the following assumptions: (a) For each b there is some k such that $f_b = f_k^*$ and $\lambda_b = \lambda_k^*$, (b) F^* is known, and (c) $K = 2$. Thus, they assumed that the true components and weights were known but it was unknown which class each mixture component represents. In Figure 1, this corresponds to the case (b) where the decision boundaries are identical. Given labeled data, the special case $K = 2$ reduces to a simple hypothesis testing problem which can be tackled using the Neyman-Pearson lemma. In this paper, we are interested in settings where each of these three assumptions fail:

- (a) *What if the class conditionals f_k^* are unknown?* Although we can always write $F^*(x) = \sum_k \lambda_k^* f_k^*(x)$, it is generally not the case that this mixture model is learnable from unlabeled data alone. In practice, what is learned will be different from this ideal case, but the hope is that it will still be useful. In this case, the argument in [Castelli and Cover \(1995\)](#) breaks down. Motivated by recent work on nonparametric mixture models ([Aragam et al., 2018](#)), we study the general case where the true mixture model is not known or even learnable from unlabeled data.
- (b) *What if F^* is unknown?* In a follow-up paper, [Castelli and Cover \(1996\)](#) studied the case where F^* is unknown by assuming that $K = 2$ and the class conditional densities $\{f_1^*, f_2^*\}$ are known up to a permutation. In this setting, the unlabeled data is used to ascertain the relative mixing proportions, but estimation error in the densities is not considered. We are interested in the general case in which a finite amount of unlabeled data is used to estimate both the mixture weights and densities.
- (c) *What if $K > 2$?* If $K > 2$, once again the argument in [Castelli and Cover \(1995\)](#) no longer applies, and we are faced with a challenging permutation learning problem. Permutation learning problems have gained notoriety recently owing to their applicability to a wide variety of problems, including statistical matching and seriation ([Collier and Dalalyan, 2016](#); [Fogel et al., 2013](#); [Lim and Wright, 2014](#)), graphical models ([van de Geer and Bühlmann, 2013](#); [Aragam et al., 2016](#)), and regression ([Pananjady et al., 2016](#); [Flammarion et al., 2016](#)), so these results may be of independent interest.

With these goals in mind, we study the MLE and majority voting (MV) rules for learning the unknown class assignment introduced in the next section. Our assumptions for MV are closely related to recent work based

on the so-called cluster assumption (Seeger, 2000; Singh et al., 2009; Rigollet, 2007; Azizyan et al., 2013); see Section 4.2 for more details.

Contributions A key aspect of our analysis is to establish conditions that connect the mixture model (1) to the true mixture model. Under these conditions we prove nonasymptotic rates of convergence for learning the class assignment (Figure 1a) from labeled data when $K > 2$, establish an $\Omega(K \log K)$ sample complexity for learning this assignment, and prove that the resulting classifier converges to the Bayes classifier. We then propose simple algorithms based on a connection to bipartite graph matching, and illustrate their performance on real and simulated data.

2 SSL as permutation learning

In this section, we formalize the ideas from the introduction using the language of mixing measures. We adopt this language for several reasons: 1) It makes it easy to refer to the parameters in the mixture model (1) by wrapping everything into a single, coherent statistical parameter Λ , 2) We can talk about convergence of these parameters via the Wasserstein metric, and 3) It simplifies discussions of identifiability in mixture models. Before going into technical details, we summarize the main idea as follows (see also Figure 1):

1. Use the unlabeled data to learn a K -component mixture model that approximates F^* , which is represented by the mixing measure Λ defined below;
2. Use the labeled data to determine the correct assignment π of classes α_k to the decision regions $\mathcal{D}_b(\Lambda)$ defined by Λ ;
3. Based on the pair (Λ, π) , define a classifier $g_{\Lambda, \pi} : \mathcal{X} \rightarrow \mathcal{Y}$ by (3) below.

Mixing measures and mixture models For concreteness, we will work on $\mathcal{X} = \mathbb{R}^d$, however, our results generalize naturally to any space \mathcal{X} with a dominating measure and well-defined density functions. Let $\mathcal{P} = \{f \in L^1(\mathbb{R}^d) : \int f dx = 1\}$ be the set of probability density functions on \mathbb{R}^d , and $\mathcal{M}_K(\mathcal{P})$ denote the space of probability measures over \mathcal{P} with precisely K atoms. An element $\Lambda \in \mathcal{M}_K(\mathcal{P})$ is called a (*finite*) *mixing measure*, and can be thought of as a convenient mathematical device for encoding the weights $\{\lambda_k\}$ and the densities $\{f_k\}$ into a single statistical parameter. By integrating against this measure, we obtain a new probability density which is denoted by

$$m(\Lambda) := \sum_{b=1}^K \lambda_b f_b(x), \quad (2)$$

where f_b is a particular enumeration of the densities in the support of Λ and λ_b is the probability of the b th density. Thus, (1) can be written as $F^* = m(\Lambda)$. By metrizing \mathcal{P} via the total variation distance $d_{\text{TV}}(f, g) = \frac{1}{2} \int |f - g| dx$, the distance between two finite K -mixtures can be computed via the Wasserstein metric:

$$W_1(\Lambda, \Lambda') = \inf \left\{ \sum_{i,j} \sigma_{ij} d_{\text{TV}}(f_i, f'_j) : 0 \leq \sigma_{ij} \leq 1, \sum_{i,j} \sigma_{ij} = 1, \sum_i \sigma_{ij} = \lambda'_j, \sum_j \sigma_{ij} = \lambda_i \right\}.$$

Decision regions, assignments, and classifiers Any mixing measure Λ defines K decision regions given by $\mathcal{D}_b = \mathcal{D}_b(\Lambda) := \{x \in \mathcal{X} : \lambda_b f_b(x) > \lambda_j f_j(x) \forall j \neq b\}$ (Figure 1). This allows us to assign an index from $1, \dots, K$ to any $x \in \mathcal{X}$, and hence defines a classifier $\check{g}_\Lambda : \mathcal{X} \rightarrow [K] := \{1, \dots, K\}$. This classifier does not solve the original labeled problem, however, since the output is an uninformative index $b \in [K]$ as opposed to a proper class label $\alpha_k \in \mathcal{Y}$. The key point is that even if we know Λ , we still must identify each label α_k with a decision region $\mathcal{D}_b(\Lambda)$, i.e. we must learn a permutation $\pi : \mathcal{Y} \rightarrow [K]$. With some abuse of notation, we

will sometimes write $\pi(k)$ instead of $\pi(\alpha_k)$ for any permutation π . Together a pair (Λ, π) defines a classifier $g_{\Lambda, \pi} : \mathcal{X} \rightarrow \mathcal{Y}$ by

$$g_{\Lambda, \pi}(x) = \pi(\hat{g}_{\Lambda}(x)) = \sum_{b=1}^K \pi^{-1}(b) 1(x \in \mathcal{D}_b(\Lambda)). \quad (3)$$

This mixing measure perspective helps to clarify the role of the unknown permutation in supervised learning: The unlabeled data is enough to learn Λ (and hence the decision regions $\mathcal{D}_b(\Lambda)$), however, labeled data are necessary to learn an assignment π between classes and decision regions.

This formulates SSL as a coupled mixture modeling and permutation learning problem: Given unlabeled and labeled data, learn a pair $(\hat{\Lambda}, \hat{\pi})$ which yields a classifier $\hat{g} = g_{\hat{\Lambda}, \hat{\pi}}$. The target is the *Bayes classifier*, which can also be written in the form (3): Let Λ^* denote the mixing measure that assigns probability λ_k^* to the density f_k^* and note that $F^* = m(\Lambda^*)$, which is the *true mixture model* defined previously. Let $\pi^* : \mathcal{Y} \rightarrow [K]$ be the permutation that assigns each class α_k to the correct decision region $\mathcal{D}_b^* = \mathcal{D}_b(\Lambda^*)$ (Figure 1). Then it is easy to check that g_{Λ^*, π^*} is the Bayes classifier.

Identifiability Although the true mixing measure Λ^* may not be identifiable from F^* , some other mixture model may be. In other words, although it may not be possible to learn Λ^* from unlabeled data, it may be possible to learn some other mixing measure $\Lambda \neq \Lambda^*$ such that $m(\Lambda) = F^* = m(\Lambda^*)$ (Figure 1c). This essentially amounts to a violation of the cluster assumption: High-density clusters are identifiable, but in practice the true class labels may not respect the cluster boundaries. Assumptions that guarantee a mixture model are identifiable are well-studied (Teicher, 1961, 1963; Yakowitz and Spragins, 1968), including both parametric Barndorff-Nielsen (1965) and nonparametric (Aragam et al., 2018; Teicher, 1967; Hall and Zhou, 2003) assumptions. In particular, Aragam et al. (2018) have proved general conditions under which mixture models with arbitrary, overlapping nonparametric components are identifiable and estimable, including examples where each component f_k has the same mean. Since this problem is well-studied, we focus hereafter on the problem of learning the permutation π^* . Thus, in the sequel we will assume that we are given an arbitrary mixing measure Λ which will be used to estimate π^* . We do not assume that $\Lambda = \Lambda^*$ or even that these mixing measures are close. The idea is to elicit conditions on Λ that ensure consistent estimation of π^* .

3 Two estimators

Assume we are given a mixing measure Λ along with the labeled samples $(X^{(i)}, Y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$. Two natural estimators of π^* are the MLE and majority vote. Although both estimators depend on Λ , this dependence will be suppressed for brevity.

Maximum likelihood Define $\ell(\pi; \Lambda, X, Y) := \log \lambda_{\pi(Y)} f_{\pi(Y)}(X)$. We will work with the following *misspecified* MLE (i.e. $\Lambda \neq \Lambda^*$)

$$\hat{\pi}_{\text{MLE}} \in \arg \max_{\pi} \ell_n(\pi; \Lambda), \quad \ell_n(\pi; \Lambda) := \frac{1}{n} \sum_{i=1}^n \ell(\pi; \Lambda, X^{(i)}, Y^{(i)}). \quad (4)$$

When $\Lambda = \Lambda^*$, this is the correctly specified MLE of the unknown permutation π^* , however, the definition above allows for the general misspecified case $\Lambda \neq \Lambda^*$.

Majority vote The majority vote estimator (MV) is given by a simple majority vote over each decision region. Formally, we define a permutation $\hat{\pi}_{\text{MV}}$ as follows: The inverse assignment $\hat{\pi}_{\text{MV}}^{-1} : [K] \rightarrow \mathcal{Y}$ is defined by

$$\hat{\pi}_{\text{MV}}^{-1}(b) = \arg \max_{\alpha \in \mathcal{Y}} \sum_{i=1}^n 1(Y^{(i)} = \alpha, X^{(i)} \in \mathcal{D}_b(\Lambda)). \quad (5)$$

If there is no majority class in a given decision region, we consider this a failure of MV and treat it as undefined. Note that when $K = 2$, the MV classifier defined by (3) with $\pi = \hat{\pi}_{\text{MV}}$ is essentially the same as the three-step procedure described in Rigollet (2007), which focuses on bounding the excess risk under the cluster assumption. In contrast, we are interested in the consistency of the unknown permutation π^* when $K > 2$, which is a more difficult problem.

4 Statistical results

Our main results establish rates of convergence for both the MLE and MV introduced in the previous section. We will use the notation $\mathbb{E}_* h(X, Y)$ to denote the expectation with respect to the true distribution $(X, Y) \sim \mathbb{P}(X, Y)$. Without loss of generality, we assume that $\pi^*(\alpha_k) = k$ and $f_b = f_b^* + h_b$ for some h_b . Then $\hat{\pi} = \pi^*$ if and only if $\hat{\pi}(\alpha_k) = k$, which helps to simplify the notation in the sequel.

4.1 Maximum likelihood

Given Λ , the notation $\mathbb{E}_* \ell(\pi; \Lambda, X, Y) = \mathbb{E}_* \log \lambda_{\pi(Y)} f_{\pi(Y)}(X)$ denotes the expectation of the *misspecified* log-likelihood with respect to the *true* distribution. Define the “gap”

$$\Delta_{\text{MLE}}(\Lambda) := \mathbb{E}_* \ell(\pi^*; \Lambda, X, Y) - \max_{\pi \neq \pi^*} \mathbb{E}_* \ell(\pi; \Lambda, X, Y). \quad (6)$$

For any function $a : \mathbb{R} \rightarrow \mathbb{R}$, define the usual Fenchel-Legendre dual $a^*(t) = \sup_{s \in \mathbb{R}} (st - a(s))$. Let $U_b = \log \lambda_b f_b(X)$ and $\beta_b(s) = \log \mathbb{E}_* \exp(sU_b)$. Finally, let $n_k := |\{i : Y^{(i)} = \alpha_k\}|$ denote the number of labeled samples with the k th label.

Theorem 4.1. *Let $\hat{\pi}_{\text{MLE}}$ be the MLE defined in (4). If $\Delta_{\text{MLE}} := \Delta_{\text{MLE}}(\Lambda) > 0$ then*

$$\mathbb{P}(\hat{\pi}_{\text{MLE}} = \pi^*) \geq 1 - 2K^2 \exp\left(-\inf_k n_k \cdot \inf_b \beta_b^*(\Delta_{\text{MLE}}/3)\right).$$

The condition $\Delta_{\text{MLE}}(\Lambda) > 0$ is a crucial condition that ensures that π^* is learnable from Λ , and the size of $\Delta_{\text{MLE}}(\Lambda)$ quantifies “how easy” it is to learn π^* is given Λ . A bigger gap implies an easier problem. Thus, it is of interest to understand this quantity better. The following proposition shows that when $\Lambda = \Lambda^*$, this gap is always nonnegative:

Proposition 4.2. *For any permutation π and any Λ ,*

$$\mathbb{E}_* \ell(\pi; \Lambda, X, Y) \leq \mathbb{E}_* \ell(\pi^*; \Lambda^*, X, Y)$$

and hence $\Delta_{\text{MLE}}(\Lambda^*) \geq 0$.

In general, assuming $\Delta_{\text{MLE}}(\Lambda) > 0$ is a weak assumption, but bounds on $\Delta_{\text{MLE}}(\Lambda)$ are difficult to obtain without making additional assumptions on the densities f_k and f_k^* . A brief discussion of this can be found in Appendix 4.5; we leave it to future work to study this quantity more carefully.

4.2 Majority vote

For any Λ , define $m_b := |\{i : X^{(i)} \in \mathcal{D}_b(\Lambda)\}|$ and $\chi_{bj}(\Lambda) := \frac{1}{m_b} \sum_{i=1}^n 1(Y^{(i)} = j, X^{(i)} \in \mathcal{D}_b(\Lambda))$, where $1(\cdot)$ is the indicator function. Similar to the MLE, our results for MV depend crucially on a “gap” quantity, given by

$$\Delta_{\text{MV}}(\Lambda) := \inf_b \left\{ \mathbb{E}_* \chi_{bb}(\Lambda) - \max_{j \neq b} \mathbb{E}_* \chi_{bj}(\Lambda) \right\}. \quad (7)$$

This quantity essentially measures how much more likely it is to sample the b th label in the b th decision region than any other label, averaged over the entire region. Thus, conditions on $\Delta_{\text{MV}}(\Lambda)$ are closely related to the well-known cluster assumption (Seeger, 2000; Singh et al., 2009; Rigollet, 2007; Azizyan et al., 2013).

Theorem 4.3. Let $\hat{\pi}_{\text{MV}}$ be the MV defined in (5). If $\Delta_{\text{MV}} := \Delta_{\text{MV}}(\Lambda) > 0$ then

$$\mathbb{P}(\hat{\pi}_{\text{MV}} = \pi^*) \geq 1 - 2K^2 \exp\left(\frac{-2\Delta_{\text{MV}}^2 \min_b m_b}{9}\right).$$

As with the MLE, the gap $\Delta_{\text{MV}}(\Lambda)$ is a crucial quantity. Fortunately, when $\Lambda = \Lambda^*$ it is always positive:

Proposition 4.4. For each $b = 1, \dots, K$,

$$\mathbb{E}_* \chi_{bb}(\Lambda^*) > \max_{j \neq b} \mathbb{E}_* \chi_{bj}(\Lambda^*)$$

and hence $\Delta_{\text{MV}}(\Lambda^*) > 0$.

When $\Lambda \neq \Lambda^*$, $\Delta_{\text{MV}}(\Lambda)$ has the following interpretation: $\Delta_{\text{MV}}(\Lambda)$ measures how well the decision regions defined by Λ match up with the decision regions defined by Λ^* . When Λ defines decision regions that assign high probability to one class, $\Delta_{\text{MV}}(\Lambda)$ will be large. If Λ defines decision regions where multiple classes have approximately the same probability, however, then it is possible that $\Delta_{\text{MV}}(\Lambda)$ will be small. In this case, our experiments in Section 6 indicate that the MLE performs much better by managing overlapping decision regions more gracefully.

4.3 Sample complexity

Theorems 4.1 and 4.3 imply upper bounds on the minimum number of samples required to learn the permutation π^* : For any $\delta \in (0, 1)$, as long as

$$\text{(MLE)} \quad \inf_k n_k := n_0 \geq \frac{\log \frac{2K^2}{\delta}}{\inf_b \beta_b^*(\Delta_{\text{MLE}}/3)} \quad (8)$$

$$\text{(MV)} \quad \inf_b m_b := m_0 \geq \frac{9 \log \frac{2K^2}{\delta}}{2\Delta_{\text{MV}}^2} \quad (9)$$

we recover π^* with probability at least $1 - \delta$.

To derive the sample complexity in terms of the total number of labeled samples n , it suffices to determine the minimum number of samples per class given n draws from a multinomial random variable. For the general case with unequal probabilities, Lemma B.2 provides a precise answer. For simplicity here, we summarize the special case where each class (resp. decision region) is equally probable for the MLE (resp. MV).

Corollary 4.5 (Sample complexity of MLE). Suppose that $\lambda_k^* = 1/K$ for each k , $\Delta_{\text{MLE}} > 0$, and

$$n \geq K \log(K/\delta) \left[1 + \frac{4}{\inf_b \beta_b^*(\Delta_{\text{MLE}}/3)} \right].$$

Then $\mathbb{P}(\hat{\pi}_{\text{MLE}} = \pi^*) \geq 1 - \delta$.

Corollary 4.6 (Sample complexity of MV). Suppose that $\mathbb{P}(X \in \mathcal{D}_b(\Lambda)) = 1/K$ for each k , $\Delta_{\text{MV}} > 0$, and

$$n \geq K \log(K/\delta) \left[1 + \frac{18}{\Delta_{\text{MV}}^2} \right].$$

Then $\mathbb{P}(\hat{\pi}_{\text{MV}} = \pi^*) \geq 1 - \delta$.

Coupon collector's problem and SSL To better understand these bounds, consider arguably the simplest possible case: Suppose that each density f_k^* has disjoint support, $\lambda_k^* = 1/K$, and that we know Λ^* . Under these very strong assumptions, an alternative way to learn π^* is to simply sample from $\mathbb{P}(X)$ until we have visited each decision region \mathcal{D}_k^* at least once. This is the classical *coupon collector's problem* (CCP), which is known to require $\Theta(K \log K)$ samples (Newman, 1960; Flajolet et al., 1992). Thus, under these assumptions the expected number of samples required to learn π^* is $\Theta(K \log K)$. By comparison, our results indicate that *even if the f_k^* have overlapping supports and we do not know Λ^** , as long as $\Delta_{\text{MLE}} = \Omega(1)$ (resp. $\Delta_{\text{MV}} = \Omega(1)$) then $\Omega(K \log K)$ samples suffice to learn π^* . In other words, SSL is approximately as difficult as CCP in very general settings.

4.4 Classification error

So far our results have focused on the probability of recovery of the unknown permutation π^* . In this section, we bound the classification error of the classifier (3) in terms of the Wasserstein distance $W_1(\Lambda, \Lambda^*)$ between Λ and Λ^* . We assume the following general set-up: We are given m unlabeled samples from which we estimate Λ by $\hat{\Lambda}_m$. Based on this mixing measure, we learn a permutation $\hat{\pi}_{m,n}$ from n labeled samples, e.g. using either MLE (4) or MV (5). Together, the pair $(\hat{\Lambda}_m, \hat{\pi}_{m,n})$ defines a classifier $\hat{g}_{m,n}$ via (3). We are interested in bounding the probability of misclassification $\mathbb{P}(\hat{g}_{m,n}(X) \neq Y)$ in terms of the Bayes error.

Theorem 4.7 (Classification error). *Suppose $W_1(\hat{\Lambda}_m, \Lambda) = O(r_m)$ for some $r_m \rightarrow 0$ where m is the number of unlabeled samples. Let $g^* = g_{\Lambda^*, \pi^*}$ denote the Bayes classifier. Then there is a constant $C > 0$ depending on K and Λ^* such that if $\hat{\pi}_{m,n} = \pi^*$,*

$$\mathbb{P}(\hat{g}_{m,n}(X) \neq Y) \leq \mathbb{P}(g^*(X) \neq Y) + Cr_m + C \cdot W_1(\Lambda, \Lambda^*).$$

This theorem allows for the possibility that the mixture model learned from the unlabeled data (i.e. $\hat{\Lambda}_m$) does not converge to the true mixing measure Λ^* . In this case, there is an irreducible error quantified by the Wasserstein distance $W_1(\Lambda, \Lambda^*)$. When $W_1(\Lambda, \Lambda^*) = 0$, however, we can improve this upper bound considerably to yield nonasymptotic rates of convergence to the Bayes error rate:

Corollary 4.8. *If $W_1(\hat{\Lambda}_m, \Lambda^*) = O(r_m)$ for some $r_m \rightarrow 0$, then the excess risk of $\hat{g}_{m,n}$ converges to zero at the same rate as $W_1(\hat{\Lambda}_m, \Lambda^*)$:*

$$\mathbb{P}(\hat{g}_{m,n}(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y) = O(r_m).$$

Clairvoyant SSL Previous work (Castelli and Cover, 1995, 1996; Singh et al., 2009) has studied the so-called *clairvoyant* SSL case in which it is assumed that we know (1) perfectly. This amounts to taking $\hat{\Lambda}_m = \Lambda$ in the previous results, or equivalently $m = \infty$. Under this assumption, we have perfect knowledge of the decision regions and only need to learn the label permutation π^* . Then Corollary 4.8 implies that with high probability, we can learn a Bayes classifier for the problem using finitely many labeled samples.

Convergence rates The convergence rate r_m used here is essentially the rate of convergence in estimating an identifiable mixture model, which is well-studied for parametric mixture models (Heinrich and Kahn, 2015; Ho and Nguyen, 2016a,b). In particular, for so-called *strongly* identifiable parametric mixture models, the minimax rate of convergence attains the optimal root- m rate $r_m = m^{-1/2}$ (Heinrich and Kahn, 2015).¹ Asymptotic consistency theorems for nonparametric mixtures can be found in Aragam et al. (2018).

Comparison to supervised learning (SL). Previous work (Singh et al., 2009) has compared the sample complexity of SSL to SL under a cluster-type assumption. While a precise characterization of these trade-offs is not the main focus of this paper, we note in passing here the following: If the minimax risk of SL for a particular problem is larger than $W_1(\Lambda, \Lambda^*)$, then Theorem 4.7 implies that SSL provably outperforms SL on finite samples.

4.5 Discussion of conditions

Here we have a simple experiment with the underlying distribution being a mixture of two Gaussians:

$$F = \frac{1}{2}\lambda_1^* + \frac{1}{2}\lambda_2^* = \frac{1}{2}\mathcal{N}(-\mu, 1) + \frac{1}{2}\mathcal{N}(\mu, 1)$$

where μ is a small positive number indicating the separation between two Gaussians. We would like to compare the number of samples needed to recover the true permutation π^* with probability $(1 - \delta)$ for both MLE and MV.

¹This paper corrects an earlier result due to Chen (1995) that claimed an $m^{-1/4}$ minimax rate.

Our experiments show that both estimators have roughly $O(\mu^{-2})$ sample complexity when $\mu \rightarrow 0^+$, but MV needs about **4 times** as many samples as the MLE. In fact, our theory can verify the sample complexity of MV: The gap Δ_{MV} is $\Phi(\mu) - \Phi(-\mu) = O(\mu)$ and the sample complexity has $\log(K/\delta)/\Delta_{\text{MV}}^2$ dependence with Δ_{MV} , which gives exactly $O(\mu^{-2})$. Here $\Phi(\mu)$ is the cumulative distribution function of standard normal random variable. Unfortunately, the intractable form of the dual functions β_b^* makes similar analytical comparisons difficult.

5 Algorithms

One of the significant appeals of MV (5) is its simplicity. It is conceptually easy to understand and trivial to implement. The MLE (4), on the other hand, is more subtle and difficult to compute in practice. In this section, we discuss two algorithms for computing the MLE: 1) An exact algorithm based on finding the maximum weight perfect matching in a bipartite graph by the Hungarian algorithm (Kuhn, 1955), and 2) Greedy optimization.

Define $C_k = \{i : Y^{(i)} = \alpha_k\}$. Consider the weighted complete bipartite graph $G = (V_{K,K}, w)$ with edge weights

$$w(k, k') = \sum_{i \in C_k} \log(\lambda_{k'} f_{k'}(X^{(i)})), \quad \forall k, k' \in [K]$$

Since a permutation π defines a perfect matching on G , the log-likelihood can be rewritten as

$$\ell_n(\pi; \Lambda) = \sum_{k=1}^K \sum_{i \in C_k} \log(\lambda_{\pi(\alpha_k)} f_{\pi(\alpha_k)}(X^{(i)})) = \sum_{k=1}^K w(k, \pi(\alpha_k)),$$

the right side of which is the total weight of the matching π . Hence, the maximizer $\hat{\pi}_{\text{MLE}}$ can be found by finding a perfect matching for this graph that has maximum weight. This can be done in $O(K^3)$ using the well-known Hungarian algorithm (Kuhn, 1955).

We can also approximately solve the matching problem by a greedy method: Assign the k th class to

$$\hat{\pi}_{\text{G}}(\alpha_k) = \arg \max_{k' \in [K]} w(k, k') = \arg \max_{k' \in [K]} \sum_{i \in C_k} \log(\lambda_{k'} f_{k'}(X^{(i)})),$$

This greedy heuristic isn't guaranteed to achieve optimal matching, however, it is simple to implement and can be viewed as a "soft interpolation" of $\hat{\pi}_{\text{MLE}}$ and $\hat{\pi}_{\text{MV}}$ as follows: If we define $w_{\text{MV}}(k, k') = \sum_{i \in C_k} 1(X^{(i)} \in \mathcal{D}_{k'}(\Lambda))$, we can see that a training example $(X^{(i)}, Y^{(i)} = \alpha_k)$ contributes 1 to $w_{\text{MV}}(k, k')$ if $k' = \arg \max_j \lambda_j f_j(X^{(i)})$, and contributes 0 to $w_{\text{MV}}(k, k')$ otherwise. By comparison, for the greedy heuristic, a training example $(X^{(i)}, Y^{(i)} = \alpha_k)$ contributes $\log(\lambda_{k'} f_{k'}(X^{(i)}))$ to $w(k, k')$. Therefore, the greedy estimator can be seen as a "soft" version of MV that also greedily optimizes the MLE objective.

6 Experiments

In order to evaluate the performance of the proposed estimators in practice, we implemented each of the three methods described in Section 5 on simulated and real data. Our experiments considered three settings: (i) Parametric mixtures of Gaussians, (ii) A nonparametric mixture model, and (iii) Real data from MNIST. All three experiments followed the same pattern: A random mixture model Λ^* was generated, and then $N = 99$ labeled samples were drawn from this mixture model. We generated Λ^* under different separation conditions, from well-separated to overlapping. Then, Λ was generated in two ways: (a) $\Lambda = \Lambda^*$, corresponding to a setting where the true decision boundaries are known, and (b) $\Lambda \neq \Lambda^*$ by perturbing the components and weights of Λ^* by a parameter $\eta > 0$ (see below for details). Then Λ was used to estimate π^* using each of the three algorithms described in the previous section for the first $n = 3, 6, 9, \dots, 99$ labeled samples. This procedure was repeated $T = 50$ times (holding Λ^* and Λ fixed) in order to estimate $\mathbb{P}(\hat{\pi} = \pi^*)$. Figure 2 depicts some examples of the mixtures used in our experiments.

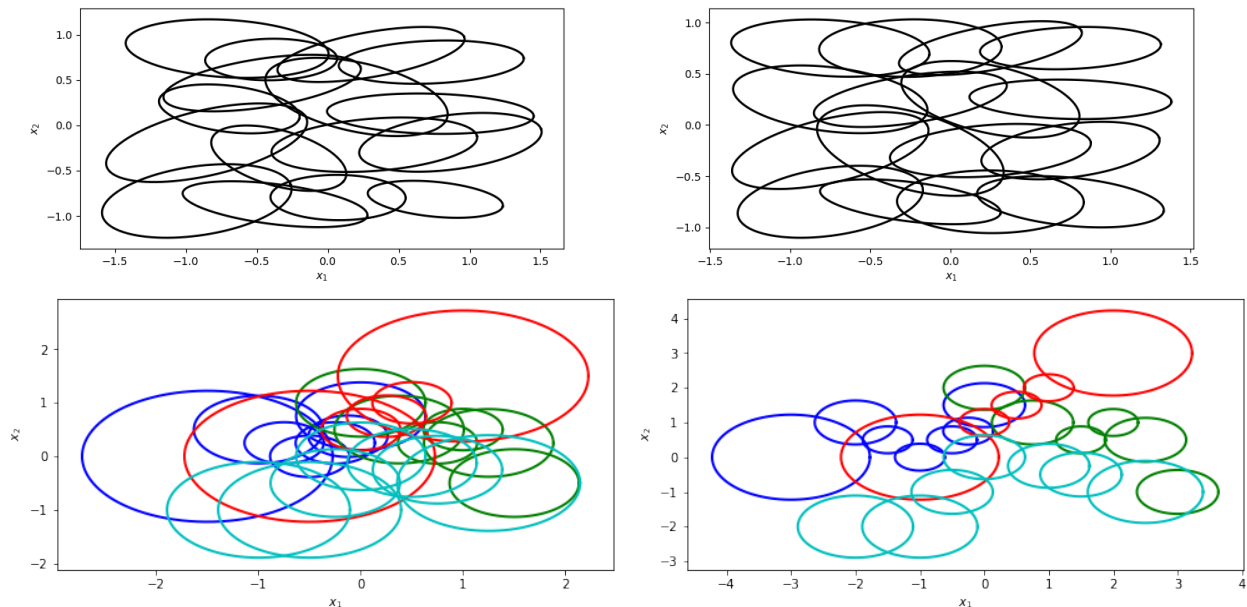


Figure 2: Examples of some examples used in the experiments. Depicted are contour lines of the densities for one standard deviation from the mean. (top) Mixture of Gaussians with $K = 16$. (bottom) Nonparametric mixture of Gaussian mixtures; each Gaussian component is coloured according to the class label it generates.

Mixture of Gaussians The first experiment uses synthetic data where $F = \sum_k \lambda_k^* f_k^*$ is a mixture of Gaussians with λ_k^* being randomly drawn from a uniform distribution $\mathcal{U}(0, 1)$ (normalized afterwards) and f_k^* being a Gaussian density. The f_k^* were arranged on a square grid with randomly generated positive-definite covariance matrices.

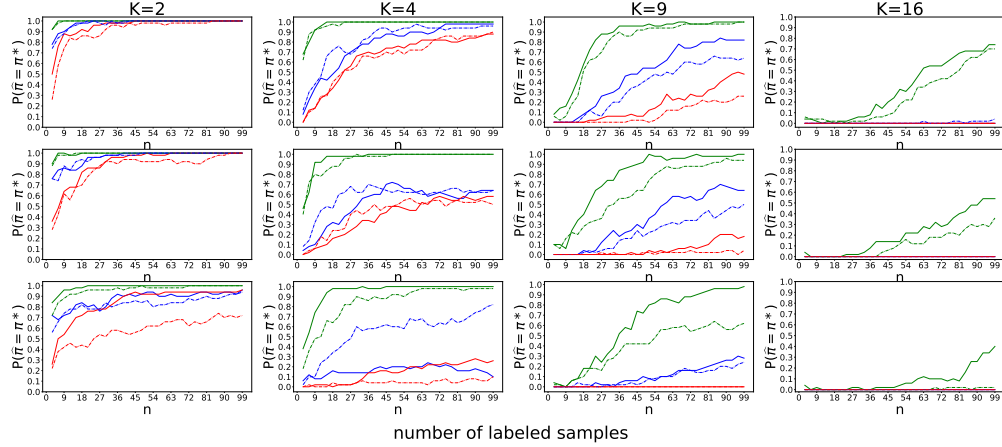
To explicitly control how well-separated the Gaussians are, we shrink the expectations of the Gaussians towards the origin using a parameter η where $\eta \in \{1, 0.75, 0.5\}$. We design the means of the Gaussians so that they are on a grid centered at the origin. The mean of each Gaussian component is thus given by $\eta\mu_k^*$, where μ_k^* is the mean of the k th density. When $\eta = 1$, components in the mixture are well-separated where $\{f_k^*\}_{k=1}^K$ have no or very little overlap within one standard deviation. The smaller the η is, the more overlapping the components are. For each choice of dimension $d \in \{2, 10\}$, K is varied between $\{2, 4, 9, 16\}$.

Perturbed mixture of Gaussians In this setting, we test the case where Λ^* is unknown and the algorithms only have access to its perturbed version Λ . Similar to the above setups, we sample n labeled data using Λ^* . However, instead of feeding the algorithms the true mixture Λ^* , we input Λ where mixture weights are shifted: Each dimension of the means of the Gaussians are shifted by a random number drawn from $\mathcal{N}(0, 0.1)$ and the variance of each Gaussians is scaled by either 0.5 or 2 (chosen at random).

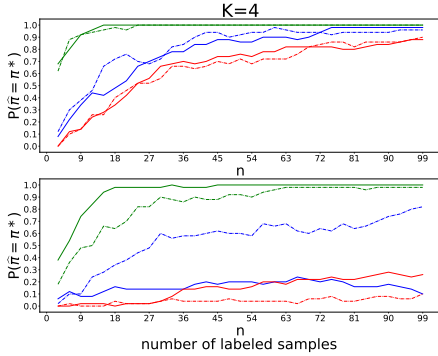
Mixture of Gaussian mixtures and its perturbation This experiment is similar to the first experiment with a mixture of Gaussians except each f_k^* is itself a Gaussian mixture. We also controlled the degree of separation by shrinking the expectation of each Gaussian towards the origin with $\eta \in \{1, 0.5\}$.

MNIST and corrupted MNIST We trained 10 kernel density estimators (one for each digit) for $\{f_k\}_{k=1}^{10}$. These mixtures are used to define the true mixture Λ^* . We then tested, under corruption of the labeled samples from the test set, how the three algorithms behave. With probability 0.1, the label of the sampled data is changed to an incorrect label.

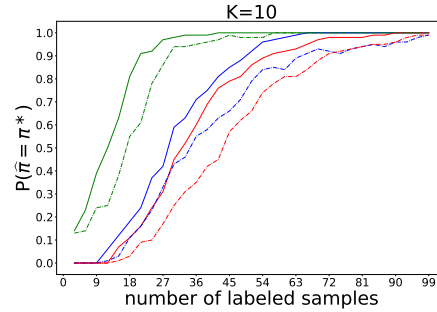
The results are depicted in Figure 3. As expected, the MLE performs by far the best, obtaining near perfect recovery of π^* with fewer than $n = 20$ labeled samples on synthetic data, and fewer than $n = 40$ on MNIST. Unsurprisingly, the most difficult case was $K = 16$, in which only the MLE was able to recover the true permutation $> 50\%$ of the time. By increasing n , the MLE is eventually able to learn this most difficult case, in accordance with our theory. Furthermore, the MLE is much more robust to misspecification $\Lambda \neq \Lambda^*$ and component overlap compared to the others. This highlights the advantage of leveraging density information in the MLE, which is ignored by the MV estimator (i.e. MV only uses decision regions).



(a) Mixture of Gaussians



(b) Mixture of Gaussian mixtures



(c) MNIST

Figure 3: Performance of MLE (Hungarian - Green; Greedy - Blue) and MV (Red). Solid line and dashed line correspond to the performance when $\Lambda^* = \Lambda$ and $\Lambda^* \neq \Lambda$, respectively. Columns correspond to the number of classes K ; rows correspond to decreasing separation; e.g. the bottom rows in each figure are the least separated.

A Proofs

A.1 Proof of Theorem 4.1

Proof. Denote a maximizer of the expected log-likelihood by $\tilde{\pi} \in \arg \max \mathbb{E}_* \ell(\pi; \Lambda)$ and define $\Delta(\pi) = \mathbb{E}_* \ell(\tilde{\pi}; \Lambda, X, Y) - \mathbb{E}_* \ell(\pi; \Lambda, X, Y)$. Note that $\Delta(\pi) \geq \Delta > 0$ for all $\pi \neq \tilde{\pi}$. Define $\mathcal{A}_\pi(t) = \{|\ell(\pi; \Lambda, X, Y) - \mathbb{E}_* \ell(\pi; \Lambda, X, Y)| < t\}$.

Then for any $t < \Delta/2 \leq \Delta(\pi)/2$, on the event $\cap_{\pi} \mathcal{A}_{\pi}(t)$ we have

$$\begin{aligned} \ell(\tilde{\pi}; \Lambda, X, Y) &> \mathbb{E}_* \ell(\tilde{\pi}; \Lambda, X, Y) - t \\ &> \mathbb{E}_* \ell(\pi; \Lambda, X, Y) + \Delta(\pi) - 2t \\ &> \ell(\pi; \Lambda, X, Y) \quad \forall \pi \neq \tilde{\pi}. \end{aligned}$$

Invoking Lemma B.1 with $g_k(X, Y) = \log \lambda_k f_k(X, Y)$, we have

$$\begin{aligned} \mathbb{P}(\cap_{\pi} \mathcal{A}_{\pi}(t)) &= \mathbb{P}\left(\forall \pi, \left| \frac{1}{n} \sum_{i=1}^n \ell(\tilde{\pi}; \Lambda, X^{(i)}, Y^{(i)}) - \mathbb{E}_* \ell(\tilde{\pi}; \Lambda, X^{(i)}, Y^{(i)}) \right| \leq t\right) \\ &\geq 1 - 2K^2 \exp(-\inf_k \inf_b n_k \beta_b^*(t)) \end{aligned}$$

Therefore, making the arbitrary choice of $t = \Delta/3$,

$$\begin{aligned} \mathbb{P}(\hat{\pi} = \tilde{\pi}) &= \mathbb{P}(\ell(\tilde{\pi}; \Lambda, X, Y) > \ell(\pi; \Lambda, X, Y) \quad \forall \pi \neq \tilde{\pi}) \\ &\geq 1 - 2K^2 \exp(-\inf_k \inf_b n_k \beta_b^*(\Delta/3)). \end{aligned}$$

Since $\Delta > 0 \implies \pi^* = \tilde{\pi}$, the desired result follows. \square

A.2 Proof of Proposition 4.2

Proof. Let $p(x, y) = \lambda_{\pi^*(y)}^* f_{\pi^*(y)}^*(x)$, $q(x, y) = \lambda_{\pi(y)} f_{\pi(y)}(x)$, so that

$$\begin{aligned} \mathbb{E}_* \ell(\pi^*; \Lambda^*, X, Y) - \mathbb{E}_* \ell(\pi; \Lambda, X, Y) &= \mathbb{E}_* \log(p(x, y)) - \mathbb{E}_* \log(q(x, y)) \\ &= \int_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} dx \\ &= \text{KL}(p \parallel q) \\ &\geq 0. \end{aligned}$$

The equality holds if and only if $p(x, y) = q(x, y)$ holds for all x, y . \square

A.3 Proof of Theorem 4.3

Proof. We have

$$\mathbb{P}(\hat{\pi} = \pi) = \mathbb{P}\left(\underbrace{\hat{\pi}(b)}_{\mathcal{E}_b} = b \quad \forall b \in [K]\right) = \mathbb{P}\left(\bigcap_{b=1}^K \mathcal{E}_b\right),$$

where

$$\mathcal{E}_b = \left\{ \sum_{i=1}^n 1(Y^{(i)} = b, X^{(i)} \in \mathcal{D}_b(\Lambda)) > \sum_{i=1}^n 1(Y^{(i)} = j, X^{(i)} \in \mathcal{D}_b(\Lambda)) \quad \forall j \neq b \right\}.$$

Let $U_{bj}^{(i)} := 1(Y^{(i)} = j, X^{(i)} \in \mathcal{D}_b(\Lambda))$ so that $\chi_{bj} = \frac{1}{n_b} \sum_i U_{bj}^{(i)}$. It suffices to control the event

$$\left\{ \sum_{i=1}^n U_{bb}^{(i)} > \sum_{i=1}^n U_{bj}^{(i)} \quad \forall j \neq b \right\} = \{\chi_{bb} > \chi_{bj} \quad \forall j \neq b\} \quad (10)$$

where $U_j^{(i)} \in \{0, 1\}$ are i.i.d. random variables. Thus, we are interested in the probability $\mathbb{P}(\chi_{bb} > \chi_{bj} \ \forall j \neq b)$. Note that

$$\mathbb{E}_* \chi_{bj} = \frac{1}{n_b} \sum_{i=1}^n \mathbb{E}_* U_{bj}^{(i)} = \frac{1}{n_b} \sum_{i: X^{(i)} \in \mathcal{D}_b} \mathbb{P}(Y^{(i)} = j, X^{(i)} \in \mathcal{D}_b(\Lambda)).$$

Define

$$\Delta_{bj} := \mathbb{E}_* \chi_{bb} - \mathbb{E}_* \chi_{bj} \tag{11}$$

and $\mathcal{A}_{bj}(t) = \{|\chi_{bj} - \mathbb{E}_* \chi_{bj}| < t\}$. Then for any $t < \Delta/2$, on the event $\cap_{j=1}^K \mathcal{A}_{bj}(t)$ we have

$$\chi_{bb} > \mathbb{E}_* \chi_{bb} - t > \mathbb{E}_* \chi_{bj} + \Delta - 2t > \chi_{bj} \quad \forall j \neq b.$$

In other words, making the arbitrary choice of $t = \Delta/3$, we deduce

$$\mathbb{P}(\mathcal{E}_b^c) \leq \mathbb{P}\left(\bigcup_{j=1}^K \mathcal{A}_j(\Delta/3)^c\right) \leq 2K \exp(-2n_b \Delta^2/9)$$

where we used Hoeffding's inequality to bound $\mathbb{P}(\mathcal{A}_j(\Delta/3)^c)$ for each j . Thus

$$\begin{aligned} \mathbb{P}\left(\bigcap_{b=1}^K \mathcal{E}_b\right) &= 1 - \sum_{b=1}^K \mathbb{P}\left(\bigcup_{j=1}^K \mathcal{A}_j(\Delta/3)^c\right) \\ &\geq 1 - 2K \sum_{b=1}^K \exp(-2n_b \Delta^2/9) \\ &\geq 1 - 2K^2 \exp\left(\frac{-2\Delta^2 \min_b n_b}{9}\right), \end{aligned}$$

as claimed. □

A.4 Proof of Proposition 4.4

Proof. We have for any $j \neq b$,

$$\begin{aligned} \mathbb{E}_* \chi_{bb}(\Lambda^*) &= \frac{1}{n_b} \sum_{i=1}^n \mathbb{E}_* 1(Y^{(i)} = b, X^{(i)} \in \mathcal{D}_b(\Lambda)) \\ &= \frac{1}{n_b} \sum_{i=1}^n \mathbb{P}(Y^{(i)} = b, X^{(i)} \in \mathcal{D}_b(\Lambda)) \\ &= \frac{1}{n_b} \sum_{i=1}^n \mathbb{P}(Y^{(i)} = b \mid X^{(i)} \in \mathcal{D}_b(\Lambda)) \mathbb{P}(X^{(i)} \in \mathcal{D}_b(\Lambda)) \\ &> \frac{1}{n_b} \sum_{i=1}^n \mathbb{P}(Y^{(i)} = j \mid X^{(i)} \in \mathcal{D}_b(\Lambda)) \mathbb{P}(X^{(i)} \in \mathcal{D}_b(\Lambda)) \\ &= \mathbb{E}_* \chi_{bj}(\Lambda^*). \end{aligned} \tag{□}$$

A.5 Proof of Corollaries 4.5 and 4.6

We prove Corollary 4.5; the proof of Corollary 4.6 is similar with n_k replaced by m_b and n_0 in (8) by m_0 in (9).

Proof. Using $p_k = 1/K$ in Lemma B.2, we deduce for any $m > 0$

$$\mathbb{P}(\min_k n_k \geq m) \geq 1 - K \exp\left(-\frac{2K}{n}(n/K - m)^2\right).$$

Thus, for any $\delta > 0$, we have

$$n \geq \frac{K}{2} \left[\log(K/\delta) + 4m \right] \implies \mathbb{P}(\min_k n_k \geq m) \geq 1 - \delta.$$

The desired result follows from replacing m with the lower bound on n_0 in (8) and invoking Theorem 4.1. \square

A.6 Proof of Theorem 4.7

Proof. To avoid notational clutter, we will suppress the dependence on m and n in the following, so that $\widehat{\Lambda} = \widehat{\Lambda}_m$, $\widehat{\pi} = \widehat{\pi}_{m,n}$, $\widehat{\mathcal{D}}_b = \mathcal{D}_b(\widehat{\Lambda}_m)$, and $\widehat{g} = \widehat{g}_{m,n}$. Write \widehat{f}_k for the components of $\widehat{\Lambda}$ and $\widehat{\lambda}_k$ for the corresponding weights. Since $\widehat{\pi} = \pi^*$, $\widehat{\mathcal{D}}_b$ corresponds to the decision region for label α_b , and hence (see e.g. §2.5 in Devroye et al., 2013)

$$\begin{aligned} \mathbb{P}(\widehat{g}(X) \neq Y) &\leq \mathbb{P}(g^*(X) \neq Y) + \sum_b \mathbb{P}(X \in \widehat{\mathcal{D}}_b \Delta \mathcal{D}_b^*) \\ &\leq \mathbb{P}(g^*(X) \neq Y) + \sum_b \int_{\mathcal{X}} |\widehat{\lambda}_b \widehat{f}_b(x) - \lambda_b^* f_b^*(x)| dx, \end{aligned} \quad (12)$$

where $\widehat{\mathcal{D}}_b \Delta \mathcal{D}_b^*$ is the symmetric difference between the estimated and true decision regions. Since $W_1(\widehat{\Lambda}_m, \Lambda) = O(r_m) \rightarrow 0$, we may assume without loss of generality that $d_{\text{TV}}(\widehat{f}_b, f_b) = O(r_m)$ and $|\widehat{\lambda}_b - \lambda_b| = O(r_m)$. Focusing on the second quantity on the right hand side above, we have

$$\sum_b \int_{\mathcal{X}} |\widehat{\lambda}_b \widehat{f}_b(x) - \lambda_b^* f_b^*(x)| dx \leq \sum_b \underbrace{\int_{\mathcal{X}} |\widehat{\lambda}_b \widehat{f}_b(x) - \lambda_b f_b(x)| dx}_{(A)} + \sum_b \underbrace{\int_{\mathcal{X}} |\lambda_b f_b(x) - \lambda_b^* f_b^*(x)| dx}_{(B)}.$$

Now, for any b ,

$$(A) \leq |\widehat{\lambda}_b - \lambda_b| + \lambda_b d_{\text{TV}}(\widehat{f}_b, f_b) = O(r_m),$$

and invoking Lemma B.3,

$$(B) \leq |\lambda_b - \lambda_b^*| + \lambda_b^* d_{\text{TV}}(f_b, f_b^*) \leq C(\Lambda^*) \cdot W_1(\Lambda, \Lambda^*).$$

Thus

$$\begin{aligned} \sum_b \int_{\mathcal{X}} |\widehat{\lambda}_b \widehat{f}_b(x) - \lambda_b^* f_b^*(x)| dx &\leq K [O(r_m) + C(\Lambda^*) \cdot W_1(\Lambda, \Lambda^*)] \\ &\leq Cr_m + C \cdot W_1(\Lambda, \Lambda^*) \end{aligned}$$

for some sufficiently large constant C depending on K and Λ^* . Plugging this back into (12) establishes the claim. \square

B Additional lemmas

B.1 Lemma B.1

For ease of notation in the following lemma, assume without loss of generality that $Y \in [K]$.

Lemma B.1. *Let g_1, \dots, g_K be functions and $\psi_k(s) = \log \mathbb{E}_* \exp(sg_k(X, Y))$ be the log moment generating function of $g_k(X, Y)$. Then*

$$\mathbb{P}\left(\forall \pi : \frac{1}{n} \sum_{i=1}^n g_{\pi(Y_i)}(X_i) - \mathbb{E}g_{\pi(Y_i)}(X_i) \leq t\right) \geq 1 - K^2 \exp(-\inf_k \inf_b n_k \psi_b^*(t)).$$

Proof. Define $C_k := \{i : Y_i = k\}$, $n_k := |C_k|$, and note that

$$\{i : \pi(Y_i) = b\} = \{i : Y_i = \pi^{-1}(b)\} = C_{\pi^{-1}(b)}.$$

Then we have the following:

$$\begin{aligned} Z &:= \frac{1}{n} \sum_{i=1}^n g_{\pi(Y_i)}(X_i) - \mathbb{E}g_{\pi(Y_i)}(X_i) = \frac{1}{n} \sum_{k=1}^K \sum_{i:\pi(Y_i)=k} g_b(X_i) - \mathbb{E}g_b(X_i) \\ &= \frac{1}{n} \sum_{b=1}^K \sum_{i \in C_{\pi^{-1}(b)}} g_b(X_i) - \mathbb{E}g_b(X_i) \\ &= \frac{1}{n} \sum_{b=1}^K n_{\pi^{-1}(b)} \underbrace{\left\{ \frac{1}{n_{\pi^{-1}(b)}} \sum_{i \in C_{\pi^{-1}(b)}} g_b(X_i) - \mathbb{E}g_b(X_i) \right\}}_{:= \tilde{Z}_b(\pi)} \\ &= \sum_{b=1}^K \frac{n_b(\pi)}{n} \tilde{Z}_b(\pi). \end{aligned}$$

Now, for each π , $\tilde{Z}_b(\pi)$ is just a sum over one of K possible subsets of $[n]$, i.e. samples indices. To see this, define

$$Z_{b,k} := \frac{1}{n_k} \sum_{i \in C_k} g_b(X_i) - \mathbb{E}g_b(X_i)$$

and note that $\tilde{Z}_b(\pi) = Z_{b, \pi^{-1}(b)}$ for each b . It follows that

$$Z = \sum_{b=1}^K \frac{n_b(\pi)}{n} \tilde{Z}_b(\pi) = \sum_{b=1}^K \frac{n_{\pi^{-1}(b)}}{n} Z_{b, \pi^{-1}(b)}$$

Chernoff's inequality implies $\mathbb{P}(Z_{b,k} \geq t) \leq \exp(-n_k \psi_b^*(t))$ for each b and k , which implies that

$$\begin{aligned} \mathbb{P}(\sup_{b,k} Z_{b,k} < t) &= \mathbb{P}\left(\bigcap_k \bigcap_b \{Z_{b,k} < t\}\right) \\ &\geq 1 - \mathbb{P}\left(\bigcup_k \bigcup_b \{Z_{b,k} < t\}^c\right) \\ &\geq 1 - \sum_{k=1}^K \sum_{b=1}^K \mathbb{P}(Z_{b,k} \geq t) \\ &\geq 1 - \sum_{k=1}^K \sum_{b=1}^K \exp(-n_k \psi_b^*(t)) \\ &\geq 1 - K^2 \exp(-\inf_k \inf_b n_k \psi_b^*(t)). \end{aligned}$$

Now, if $\sup_{b,k} Z_{b,k} < t$, then

$$Z = \sum_{b=1}^K \frac{n_{\pi^{-1}(b)}}{n} Z_{b,\pi^{-1}(b)} < \sum_{b=1}^K \frac{n_{\pi^{-1}(b)}}{n} t = t$$

since $\sum_b n_b/n = 1$ and π is a bijection. The desired result follows. \square

B.2 Lemma B.2

The following lemma gives a precise bound on the minimum number of samples n required to ensure $\min_k n_k \geq m$ from a generic multinomial sample with high probability:

Lemma B.2. *Let Y_i be a multinomial random variable such that $\mathbb{P}(Y_i = k) = p_k$ and define $n_k = \sum_{i=1}^n 1(Y_i = k)$. Then for any $m > 0$,*

$$\mathbb{P}(\min_k n_k \geq m) \geq 1 - \sum_{k=1}^K \exp\left(-\frac{2}{np_k}(np_k - m)^2\right).$$

Proof. By standard tail bounds on $n_k \sim \text{Bin}(n, p_k)$, we have $\mathbb{P}(n_k \leq m) \leq \exp(-2(np_k - m)^2/(np_k))$. Thus

$$\mathbb{P}(\min_k n_k < m) = \mathbb{P}(\cup_{k=1}^K \{n_k < m\}) \leq \sum_{k=1}^K \mathbb{P}(n_k < m) \leq \sum_{k=1}^K \exp\left(-\frac{2}{np_k}(np_k - m)^2\right),$$

as claimed. \square

B.3 Lemma B.3

For any density $f \in L^1$, let δ_f denote the point mass concentrated at f , so that for any Borel subset $A \subset \mathcal{P}$,

$$\delta_f(A) = \begin{cases} 1, & f \in A \\ 0, & f \notin A. \end{cases}$$

Lemma B.3. *Let $\Lambda = \sum_{k=1}^K \lambda_k \delta_{f_k}$ and $\Lambda' = \sum_{k=1}^K \lambda'_k \delta_{f'_k}$. Then there is a constant $C = C(\Lambda', K)$ such that*

$$\sup_j \inf_i |\lambda_i - \lambda'_j| \leq C W_1(\Lambda, \Lambda'), \quad (13)$$

$$\sup_j \inf_i d_{\text{TV}}(f_i, f'_j) \leq C W_1(\Lambda, \Lambda'). \quad (14)$$

Proof. The first inequality (13) follows from Theorem 4 in [Gibbs and Su \(2002\)](#), and the second inequality (14) is standard. \square

References

- B. Aragam, A. A. Amini, and Q. Zhou. Learning directed acyclic graphs with penalized neighbourhood regression. *arXiv:1511.08963*, 2016.
- B. Aragam, C. Dan, P. Ravikumar, and E. Xing. Identifiability of nonparametric mixture models and bayes optimal clustering. *arXiv preprint*, arXiv:1802.04397, 2018.
- M. Azizyan, A. Singh, L. Wasserman, et al. Density-sensitive semisupervised inference. *The Annals of Statistics*, 41(2):751–771, 2013.
- O. Barndorff-Nielsen. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, 12(1):115–121, 1965.
- A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.
- V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.
- V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on information theory*, 42(6):2102–2117, 1996.
- J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, pages 221–233, 1995.
- O. Collier and A. S. Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.
- F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-supervised learning of mixture models. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 99–106, 2003.
- Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in Neural Information Processing Systems*, pages 6513–6523, 2017.
- L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.
- P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.
- N. Flammarion, C. Mao, and P. Rigollet. Optimal rates of statistical seriation. *arXiv preprint arXiv:1607.02435*, 2016.
- F. Fogel, R. Jenatton, F. Bach, and A. d’Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024, 2013.
- A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.
- P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, pages 201–224, 2003.
- P. Heinrich and J. Kahn. Minimax rates for finite mixture estimation. *arXiv preprint arXiv:1504.03506*, 2015.
- N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016a.
- N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *arXiv preprint arXiv:1609.02655*, 2016b.
- M. Kääriäinen. Generalization error bounds using unlabeled data. In *International Conference on Computational Learning Theory*, pages 127–142. Springer, 2005.
- D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative

- models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.
- H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2): 83–97, 1955.
- C. H. Lim and S. Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems*, pages 2168–2176, 2014.
- D. J. Newman. The double dixie cup problem. *The American Mathematical Monthly*, 67(1):58–61, 1960.
- P. Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *The Journal of Machine Learning Research*, 14(1):1229–1250, 2013.
- A. Pananjady, M. J. Wainwright, and T. A. Courtade. Linear regression with an unknown permutation: Statistical and computational limits. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*, pages 417–424. IEEE, 2016.
- P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007.
- M. Seeger. Learning with labeled and unlabeled data. Technical report, 2000.
- A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in neural information processing systems*, pages 1513–1520, 2009.
- H. Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1):244–248, 1961.
- H. Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.
- H. Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4): 1300–1302, 1967.
- S. van de Geer and P. Bühlmann. ℓ_0 -penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.
- L. Wasserman and J. D. Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.
- S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.
- X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.