Learning From Non-iid Data: Fast Rates for the One-vs-All Multiclass Plug-in Classifiers

Vu Dinh[1,*] **Lam Si Tung Ho**[2,*] **Nguyen Viet Cuong**[3] **Duy Nguyen**[4] **Binh T. Nguyen**[5]
[1]Department of Mathematics, Purdue University
[2]Department of Biostatistics, University of California, Los Angeles
[3]Department of Computer Science, National University of Singapore
[4]Department of Statistics, University of Wisconsin-Madison
[5]Department of Computer Science, University of Science, Vietnam

Abstract

We prove new fast learning rates for the one-vs-all multiclass plug-in classifiers trained either from exponentially strongly mixing data or from data generated by a converging drifting distribution. These are two typical scenarios where training data are not iid. The learning rates are obtained under a multiclass version of Tsybakov's margin assumption, a type of low-noise assumption, and do not depend on the number of classes. Our results are general and include a previous result for binary-class plug-in classifiers with iid data as a special case. In contrast to previous works for least squares SVMs under the binary-class setting, our results retain the optimal learning rate in the iid case.

1 Introduction

Fast learning of plug-in classifiers from low-noise data has recently gained much attention [2, 3, 4, 5]. The first fast/super-fast learning rates[1] for plug-in classifiers were proven in [2] under Tsybakov's margin assumption [6], a type of low-noise condition. Their plug-in classifiers employ the local polynomial estimator to estimate the conditional probability of a label Y given an observation X and use it in the plug-in rule. Subsequently, [3] proved the fast learning rate for plug-in classifiers with a relaxed condition on the density of X and investigated the use of kernel, partitioning, and nearest neighbor estimators instead of the local polynomial estimator. Monnier [4] suggested to use local multi-resolution projections to estimate the conditional probability of Y and proved super-fast rates of the corresponding plug-in classifier under the same margin assumption.

These previous analyses of plug-in classifiers typically focus on the binary-class setting with iid data assumption. This is a limitation of the current theory for plug-in classifiers since (1) many classification problems are multiclass in nature and (2) data may also violate the iid data assumption in practice. In this paper, we contribute to the theoretical understandings of plug-in classifiers by proving novel fast learning rates of a multiclass plug-in classifier trained from non-iid data. In particular, we prove that the multiclass plug-in classifier constructed using the *one-vs-all method* can achieve fast learning rates, or even super-fast rates, with the following two types of non-iid training data: data generated from an *exponentially strongly mixing sequence* and data generated from a *converging drifting distribution*. This is the first result that proves fast learning rates for multiclass classifiers with non-iid data, and our rates do not depend on the number of classes.

Our results assume a multiclass version of Tsybakov's margin assumption that assumes the events in which the most probable label of an example is ambiguous with the second most probable label have

[*]The authors contributed equally to this work. The full paper was previously published at TAMC 2015 [1].

[1]Fast learning rate means the trained classifier converges with rate faster than $n^{-1/2}$, while super-fast learning rate means the trained classifier converges with rate faster than n^{-1}.

small probabilities. This margin assumption was previously considered for multiclass empirical risk minimization (ERM) classifiers with iid data [7] and in the context of active learning with cost-sensitive multiclass classifiers [8]. Our results are natural generalizations for both the binary-class and the iid data settings. As special cases of our results, we can obtain fast learning rates for the one-vs-all multiclass plug-in classifiers in the iid data setting and for the binary-class plug-in classifiers in the non-iid data setting. Our results can also be used to obtain the previous fast learning rates [2] for the binary-class plug-in classifiers in the iid data setting.

In terms of theory, the extension from binary class to multiclass problem is usually not trivial and depends greatly on the choice of the multiclass classifiers. Our results show that this extension can be achieved with plug-in classifiers and the one-vs-all method. The one-vs-all method is a practical way to construct a multiclass classifier using binary-class classification [9]. This method trains a model for each class by converting multiclass data into binary-class data and then combines them into a multiclass classifier.

Our paper considers two types of non-iid data. Exponentially strongly mixing data is a typical case of identically but not independently distributed data. Fast learning from exponentially strongly mixing data has been previously analyzed for least squares support vector machines (LS-SVMs) [10, 11] and ERM classifiers [11]. On the other hand, data generated from a drifting distribution (or drifting concept) is an example of independently but not identically distributed data. Some concept drifting scenarios and learning bounds were previously investigated in [12, 13, 14, 15]. In this paper, we consider the scenario where the parameters of the distributions generating the training data converge uniformly to those of the test distribution with some polynomial rate.

We note that even though LS-SVMs can be applied to solve a classification problem with binary data, the previous results for LS-SVMs cannot retain the optimal rate in the iid case [10, 11]. In contrast, our results in this paper still retain the optimal learning rate for the Hölder class in the iid case. Besides, the results for drifting concepts can also achieve this optimal rate.

2 Preliminaries

Let $\{(X_i, Y_i)\}_{i=1}^n$ be labeled training data where $X_i \in \mathbb{R}^d$ and $Y_i \in \{1, 2, \ldots, m\}$ for all i. The binary-class case corresponds to $m = 2$, while the multiclass case corresponds to $m > 2$. For now we do not specify how $\{(X_i, Y_i)\}_{i=1}^n$ are generated, but we assume that test data are drawn iid from an unknown distribution $\mathbf{P}$ on $\mathbb{R}^d \times \{1, 2, \ldots, m\}$.

Given the training data, our aim is to find a classification rule $f : \mathbb{R}^d \to \{1, 2, \ldots, m\}$ whose risk is as small as possible. The risk of a classifier f is defined as $R(f) \triangleq \mathbf{P}(Y \neq f(X))$. One minimizer of the above risk is the Bayes classifier $f^*(X) \triangleq \arg\max_j \eta_j(X)$, where $\eta_j(X) \triangleq \mathbf{P}(Y = j|X)$ for all $j \in \{1, 2, \ldots, m\}$. For any classifier $\widehat{f}_n$ trained from the training data, it is common to characterize its accuracy via the excess risk $\mathcal{E}(\widehat{f}_n) \triangleq \mathbf{E}R(\widehat{f}_n) - R(f^*)$, where the expectation is with respect to the randomness of the training data. A small excess risk for $\widehat{f}_n$ is thus desirable as the classifier will perform close to the optimal classifier f^* on average.

For a classifier f, we write $\eta_f(X)$ as an abbreviation for $\eta_{f(X)}(X)$, which is the value of the function $\eta_{f(X)}$ at X. We have the following useful property of the excess risk in the multiclass setting.

Proposition 1. *For any classifier $\widehat{f}_n$, we have $\mathcal{E}(\widehat{f}_n) = \mathbf{E}[\eta_{f^*}(X) - \eta_{\widehat{f}_n}(X)]$, where the expectation is with respect to the randomness of both the training data and the testing example X.*

Following [2], we assume all the functions η_j's are in the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ and the marginal distribution $\mathbf{P}_X$ of X satisfies the strong density assumption (see [1, 2] for details).

• **Margin Assumption for Multiclass Setting:**

As in the binary-class case, fast learning rates for the multiclass plug-in classifier can be obtained under a low-noise assumption similar to Tsybakov's margin assumption [6]. In particular, we assume that the conditional probabilities η_j's satisfy the following margin assumption, which is an extension of Tsybakov's margin assumption to the multiclass setting.

Assumption (Margin Assumption)**.** *There exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for all $t > 0$, $\mathbf{P}_X(\eta_{(1)}(X) - \eta_{(2)}(X) \leq t) \leq C_0 t^\alpha$, where $\eta_{(1)}(X)$ and $\eta_{(2)}(X)$ are the largest and second largest conditional probabilities among all the $\eta_j(X)$'s.*

3 The One-vs-All Multiclass Plug-in Classifier

We now introduce the one-vs-all multiclass plug-in classifier which we will analyze in this paper. Let $\widehat{\eta}_n(X) = (\widehat{\eta}_{n,1}(X), \widehat{\eta}_{n,2}(X), \ldots, \widehat{\eta}_{n,m}(X))$ be an m-dimensional function where $\widehat{\eta}_{n,j}$ is a non-parametric estimator of η_j from the training data. The corresponding multiclass plug-in classifier $\widehat{f}_n$ predicts the label of an observation X by: $\widehat{f}_n(X) = \arg\max_j \widehat{\eta}_{n,j}(X)$. In this paper, we consider plug-in classifiers where $\widehat{\eta}_{n,j}$'s are estimated using the one-vs-all method and the local polynomial regression function as follows. For each class $j \in \{1, 2, \ldots, m\}$, we first convert the training data $\{(X_i, Y_i)\}_{i=1}^n$ to binary class by considering all (X_i, Y_i)'s such that $Y_i \neq j$ as negative (label 0) and those such that $Y_i = j$ as positive (label 1). Then we construct $\widehat{\eta}_{n,j}$ from the new binary-class training data using the local polynomial regression function with an appropriate bandwidth h and kernel K [1]. Specifically, K has to satisfy the assumptions similar to those in [2] (see [1] for details). The conditions for h are given in Section 4 and 5.

4 Fast Learning For Exponentially Strongly Mixing Data

In this section, we consider the case where training data are generated from an exponentially strongly mixing sequence [10, 16]. Let $Z_i = (X_i, Y_i)$ for all i. Assume that $\{Z_i\}_{i=1}^\infty$ is a stationary sequence of random variables on $\mathbb{R}^d \times \{1, 2, \ldots, m\}$ with stationary distribution $\mathbf{P}$. That is, $\mathbf{P}$ is the marginal distribution of any random variable in the sequence. For all $k \geq 1$, we define the $\boldsymbol{\alpha}$-mixing coefficients [10]: $\boldsymbol{\alpha}(k) \triangleq \sup_{A_1 \in \sigma_1^t, A_2 \in \sigma_{t+k}^\infty, t \geq 1} |\mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_1)\mathbf{P}(A_2)|$, where σ_a^b is the σ-algebra generated by $\{Z_i\}_{i=a}^b$. The sequence $\{Z_i\}_{i=1}^\infty$ is exponentially strongly mixing if there exist positive constants C_1, C_2 and C_3 such that for every $k \geq 1$, we have

$$\boldsymbol{\alpha}(k) \leq C_1 \exp(-C_2 k^{C_3}). \tag{1}$$

We now state some key lemmas for proving the convergence rate of the multiclass plug-in classifier in this setting. Let $n_e \triangleq \left\lfloor \frac{n}{\lceil \{8n/C_2\}^{1/(C_3+1)} \rceil} \right\rfloor$ be the effective sample size. The following lemma is about the convergence rate of the local polynomial regression functions using the one-vs-all method.

Lemma 1. *Let β, r_0, and c be the constants in the Hölder assumption, the strong density assumption, and the assumption for the kernel K respectively. Then there exist constants $C_4, C_5, C_6 > 0$ such that for all $\delta > 0$, all bandwidth h satisfying $C_6 h^\beta < \delta$ and $0 < h \leq r_0/c$, all $j \in \{1, 2, \ldots, m\}$ and $n \geq 1$, we have $\mathbf{P}^{\otimes n}(|\widehat{\eta}_{n,j}(x) - \eta_j(x)| \geq \delta) \leq C_4 \exp(-C_5 n_e h^d \delta^2)$ for almost surely x with respect to $\mathbf{P}_X$, where d is the dimension of the observations (inputs).*

Given the above convergence rate of the local polynomial regression functions, Lemma 2 below gives the convergence rate of the excess risk of the one-vs-all multiclass plug-in classifier.

Lemma 2. *Let α be the constant in the margin assumption. Assume that there exist $C_4, C_5 > 0$ such that $\mathbf{P}^{\otimes n}(|\widehat{\eta}_{n,j}(x) - \eta_j(x)| \geq \delta) \leq C_4 \exp(-C_5 a_n \delta^2)$ for almost surely x with respect to $\mathbf{P}_X$, and for all $j \in \{1, 2, \ldots, m\}$, $\delta > 0$. Then there exists $C_7 > 0$ such that for all $n \geq 1$, $\mathcal{E}(\widehat{f}_n) = \mathbf{E}R(\widehat{f}_n) - R(f^*) \leq C_7 a_n^{-(1+\alpha)/2}$.*

Using Lemma 1 and 2, we can obtain the following theorem about the convergence rate of the one-vs-all multiclass plug-in classifier when training data are exponentially strongly mixing. This theorem is a direct consequence of Lemma 1 and 2 with $h = n_e^{-1/(2\beta+d)}$ and $a_n = n_e^{2\beta/(2\beta+d)}$.

Theorem 1. *Let α and β be the constants in the margin assumption and the Hölder assumption respectively, and let d be the dimension of the observations. Let $\widehat{f}_n$ be the one-vs-all multiclass plug-in classifier with bandwidth $h = n_e^{-1/(2\beta+d)}$ that is trained from an exponentially strongly mixing sequence. Then there exists some constant $C_8 > 0$ such that for all n large enough that satisfies $0 < n_e^{-1/(2\beta+d)} \leq r_0/c$, we have: $\mathcal{E}(\widehat{f}_n) = \mathbf{E}R(\widehat{f}_n) - R(f^*) \leq C_8 n_e^{-\beta(1+\alpha)/(2\beta+d)}$.*

The convergence rate in Theorem 1 is expressed in terms of the effective sample size n_e rather than the sample size n since learning with dependent data typically requires more data to achieve the same level of accuracy as learning with independent data (see e.g., [10, 17, 18]). However, Theorem 1 still implies the fast rate for the one-vs-all multiclass plug-in classifier in terms of the sample size n. Indeed, the rate in the theorem can be rewritten as $O(n^{-\frac{\beta(1+\alpha)}{2\beta+d}\cdot\frac{C_3}{C_3+1}})$, so the fast learning rate is achieved when $2(\alpha - 1/C_3)\beta > (1 + 1/C_3)d$ and the super-fast learning rate is achieved when $(\alpha - 1 - 2/C_3)\beta > d(1 + 1/C_3)$.

5 Fast Learning From a Drifting Concept

In this section, we consider the case where training data are generated from a drifting concept that converges to the test distribution $\mathbf{P}$. Unlike the setting in Section 4 where the training data form a stationary sequence of random variables, the setting in this section may include training data that are not stationary. Formally, we assume the training data $\{Z_i\}_{i=1}^n = \{(X_i, Y_i)\}_{i=1}^n$ are generated as follows. The observations X_i are generated iid from the marginal distribution $\mathbf{P}_X$ satisfying the strong density assumption. For each $i \geq 1$, the label Y_i of X_i is generated from a categorical distribution on $\{1, 2, \ldots, m\}$ with parameters $\eta^i(X_i) \triangleq (\eta_1^i(X_i), \eta_2^i(X_i), \ldots, \eta_m^i(X_i))$. That is, the probability of $Y_i = j$ conditioned on X_i is $\eta_j^i(X_i)$, for all $j \in \{1, 2, \ldots, m\}$. Note that from our setting, the training data are independent but not identically distributed. To prove the convergence rate of the multiclass plug-in classifier, we assume that $\|\eta_j^n - \eta_j\|_\infty \triangleq \sup_{x\in\mathbb{R}^d} |\eta_j^n(x) - \eta_j(x)| = O(n^{-(\beta+d)/(2\beta+d)})$ for all j, i.e., η_j^n converges uniformly to the label distribution η_j of test data with rate $O(n^{-(\beta+d)/(2\beta+d)})$. The following lemma states the convergence rate of the local polynomial regression functions in this setting. Note that the constants in this section may be different from those in Section 4.

Lemma 3. *Let β, r_0, and c be the constants in the Hölder assumption, the strong density assumption, and the assumption for the kernel K respectively. Let $\widehat{\eta}_{n,j}$ be the estimator of η_j estimated using the local polynomial regression function with $h = n^{-1/(2\beta+d)}$. If $\|\eta_j^n - \eta_j\|_\infty = O(n^{-(\beta+d)/(2\beta+d)})$ for all j, then there exist constants $C_4, C_5, C_6 > 0$ such that for all $\delta > 0$, all n satisfying $C_6 n^{-\beta/(2\beta+d)} < \delta < 1$ and $0 < n^{-1/(2\beta+d)} \leq r_0/c$, and all $j \in \{1, 2, \ldots, m\}$, we have $\mathbf{P}^{\otimes n}(|\widehat{\eta}_{n,j}(x) - \eta_j(x)| \geq \delta) \leq C_4 \exp(-C_5 n^{2\beta/(2\beta+d)}\delta^2)$ for almost surely x with respect to $\mathbf{P}_X$, where d is the dimension of the observations.*

The following theorem is a direct consequence of Lemma 2 and 3 with $a_n = n^{2\beta/(2\beta+d)}$. We note that the convergence rate in Theorem 2 is fast when $\alpha\beta > d/2$ and is super-fast when $(\alpha-1)\beta > d$.

Theorem 2. *Let α and β be the constants in the margin assumption and the Hölder assumption respectively, and let d be the dimension of the observations. Let $\widehat{f}_n$ be the one-vs-all multiclass plug-in classifier with bandwidth $h = n^{-1/(2\beta+d)}$ that is trained from data generated from a drifting concept converging uniformly to the test distribution. Then there exists some constant $C_8 > 0$ such that for all n large enough that satisfies $0 < n^{-1/(2\beta+d)} \leq r_0/c$, we have $\mathcal{E}(\widehat{f}_n) = \mathbf{E}R(\widehat{f}_n) - R(f^*) \leq C_8 n^{-\beta(1+\alpha)/(2\beta+d)}$.*

6 Remarks

The rates in Theorem 1 and 2 do not depend on the number of classes m. They are both generalizations of the previous result for binary-class plug-in classifiers with iid data [2]. More specifically, $C_3 = +\infty$ in the case of iid data, thus we have $n_e = n$ and the data distribution also satisfies the condition in Theorem 2. Hence, we can obtain the same result as in [2]. Our results for the one-vs-all multiclass plug-in classifiers retain the optimal rate $O(n^{-\beta(1+\alpha)/(2\beta+d)})$ for the Hölder class in the iid case [2] while the previous results in [10, 11] for LS-SVMs with smooth kernels do not (see Example 4.3 in [11]). Besides, from Theorem 2, the one-vs-all multiclass plug-in classifiers trained from a drifting concept can also achieve this optimal rate. We note that for LS-SVMs with Gaussian kernels, Hang and Steinwart [11] proved that they can achieve the essentially optimal rate in the iid scenario (see Example 4.4 in [11]). That is, their learning rate is n^ζ times of the optimal rate for any $\zeta > 0$. Although this rate is very close to the optimal rate, it is still slower than $\log n$ times of the optimal rate.[2]

[2]The optimal rates in Example 4.3 and 4.4 of [11] may not necessarily be the same as our optimal rate since Hang and Steinwart considered Sobolev space and Besov space instead of Hölder space.

References

[1] Vu Dinh, Lam Si Tung Ho, Nguyen Viet Cuong, Duy Nguyen, and Binh T. Nguyen. Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers. In *Theory and Applications of Models of Computation*, pages 375–387. 2015.

[2] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[3] Michael Kohler and Adam Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory*, 53(5):1735–1742, 2007.

[4] Jean-Baptiste Monnier. Classification via local multi-resolution projections. *Electronic Journal of Statistics*, 6:382–420, 2012.

[5] Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13:67–90, 2012.

[6] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, pages 135–166, 2004.

[7] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

[8] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *Proceedings of the International Conference on Machine Learning*, 2013.

[9] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[10] Ingo Steinwart and Andreas Christmann. Fast learning from non-iid observations. In *Advances in Neural Information Processing Systems*, pages 1768–1776, 2009.

[11] H. Hang and I. Steinwart. Fast learning from alpha-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014.

[12] Peter L. Bartlett. Learning with a slowly changing distribution. In *COLT 1992*.

[13] Philip M. Long. The complexity of learning according to two models of a drifting environment. *Machine Learning*, 37(3):337–354, 1999.

[14] Rakesh D. Barve and Philip M. Long. On the complexity of learning from drifting distributions. In *COLT 1996*.

[15] Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*, pages 124–138, 2012.

[16] Dharmendra S. Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145, 1996.

[17] Nguyen Viet Cuong, Lam Si Tung Ho, and Vu Dinh. Generalization and robustness of batched weighted average algorithm with V-geometrically ergodic Markov data. In *Algorithmic Learning Theory*, pages 264–278, 2013.

[18] C. Ané. Analysis of comparative data with hierarchical autocorrelation. *The Annals of Applied Statistics*, 2(3):1078–1102, 2008.

[19] VV Yurinskiĭ. Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis*, 6(4):473–499, 1976.