Multiclass Learning with Simplex Coding

Authored by:

Tomaso Poggio
Lorenzo Rosasco
Jean-jeacques Slotine
Youssef Mroueh

Abstract

In this paper we dicuss a novel framework for multiclass learning, defined by a suitable coding/decoding strategy, namely the simplex coding, that allows to generalize to multiple classes a relaxation approach commonly used in binary classification. In this framework a relaxation error analysis can be developed avoiding constraints on the considered hypotheses class. Moreover, we show that in this setting it is possible to derive the first provably consistent regularized methods with training/tuning complexity which is {em independent} to the number of classes. Tools from convex analysis are introduced that can be used beyond the scope of this paper.

1 Paper Body

As bigger and more complex datasets are available, multiclass learning is becoming increasingly important in machine learning. While theory and algorithms for solving binary classification problems are well established, the problem of multi-category classification is much less understood. Practical multiclass algorithms often reduce the problem to a collection of binary classification problems. Binary classification algorithms are often based on a relaxation approach: classification is posed as a non-convex minimization problem and then relaxed to a convex one, defined by suitable convex loss functions. In this context, results in statistical learning theory quantify the error incurred by relaxation and in particular derive comparison inequalities explicitly relating the excess misclassification risk to the excess expected loss. We refer to [2, 27, 14, 29] and [18] Chapter 3 for an exhaustive presentation as well as generalizations. Generalizing the above approach and results to more than two classes is not straightforward. Over the years, several computational solutions have been proposed (among others, see [10, 6, 5, 25, 1, 21]). Indeed, most of these methods can be interpreted as a kind of relaxation. Most proposed methods have complexity which is more than linear in the number of classes and simple one-vs all in practice offers a good alternative both in terms of performance and speed [15]. Much fewer works have focused on deriving theoretical guarantees. Results in this sense have been pioneered by [28, 20], see also [11, 7, 23]. In these works the error due to relaxation is studied asymptotically and under constraints on the function class to be considered. More quantitative results in terms of comparison inequalities are given in [4] under similar restrictions (see also [19]). Notably, the above results show that seemingly intuitive extensions of binary classification algorithms might lead to methods which are not consistent. Further, it is interesting to note that the restrictions on the function class, needed to prove the theoretical guarantees, make the computations in the corresponding algorithms more involved and are in fact often ignored in practice. In this paper we dicuss a novel framework for multiclass learning, defined by a suitable coding/decoding strategy, namely the simplex coding, in which a relaxation error analysis can be developed avoiding constraints on the considered hypotheses class. Moreover, we show that in this framework it is possible to derive the first provably consistent regularized method with training/tuning complexity that is independent to the number of classes. Interestingly, using the simplex coding, we can naturally generalize results, proof techniques and methods from the binary case, which is recovered as a special case of our theory. Due to space restriction in this paper we focus on extensions of least squares, and SVM loss functions, but our analysis can be generalized to a large class 1

of simplex loss functions, including extensions of the logistic and exponential loss functions (used in boosting). Tools from convex analysis are developed in the supplementary material and can be useful beyond the scope of this paper, in particular in structured prediction. The rest of the paper is organized as follow. In Section 2 we discuss the problem statement and background. In Section 3 we discuss the simplex coding framework which we analyze in Section 4. Algorithmic aspects and numerical experiments are discussed in Section 5 and Section 6, respectively. Proofs and supplementary technical results are given in the appendices.

2

Problem Statement and Previous Work

Let (X, Y) be two random variables with values in two measurable spaces X and Y = {1 . . . T }, T ? 2. Denote by ?X , the law of X on X , and by ?j (x), the conditional probabilities for j ? Y. The data is a sample S = (xi , yi)ni=1 , from n identical and independent copies of (X, Y). We can think of X as a set of possible inputs and of Y as a set of labels describing a set of semantic categories/classes the input can belong to. A classification rule is a map b : X ? Y, and its error is measured by the misclassification risk R(b) = P(b(X) 6= Y) = E(1I[b(x)6=y] (X, Y)). The optimal classification rule that minimizes R is the Bayes rule b? (x) = arg maxy?Y ?y (x), x ? X . Computing the Bayes rule by directly minimizing the risk R is not possible since the probability Pn distribution is unknown. One might think of minimizing the empirical risk (ERM) RS (b) = n1 i=1 1I[b(x)6=y] (xi , yi), which is an unbiased estimator of the R, but the corresponding optimization problem is in general not feasible. In binary classification, one of the most common ways to obtain computationally efficient methods is based on a relaxation approach. We recall this approach in the next section and describe its extension to multiclass in the rest of the paper. Relaxation Approach to Binary Classification. If T = 2, we can set Y = {?1}. Most modern machine learning algorithms for binary classification consider a convex relaxation of the ERM functional RS . More precisely: 1) the indicator function in RS is replaced by non negative loss V : Y ? R ? R+ which is convex in the second argument and is sometimes called a surrogate loss; 2) the classification rule b replaced by a real valued measurable function f : X ? R. A classification rule is then obtained by considering the sign of f . It often suffices to consider a special class of loss functions, namely large margin loss functions V : R ? R+ of the form V (?yf (x)). This last expression is suggested by the observation that the misclassification risk, using the labels ?1, can be written as R(f) = E(?(?Y f (X))), where ? is the Heaviside step function. The quantity m = ?yf (x), sometimes called the margin, is a natural point-wise measure of the classification error. Among other examples of large margin loss functions (such as the logistic and exponential loss), we recall the hinge loss V (m) = —1 + m—+ = max{1 + m, 0} used in the support vector machine, and the square loss V (m) = (1 + m)2 used in regularized least squares (note that (1 ? yf (x))2 = (y ? f (x))2). Using large margin loss functions it is possible to design effective learning algorithms replacing the empirical risk with regularized empirical risk minimization n

ES? (f) =

1X V (yi , f (xi)) + ?R(f), n i=1

(1)

where R is a suitable regularization functional and ? is the regularization parameter, (see Section 5). 2.1

Relaxation Error Analysis

As we replace the misclassification loss with a convex surrogate loss, we are effectively changing the problem: the misclassification risk is replaced by the expected loss, E(f) = E(V (?Y f (X))) . The expected loss can be seen as a functional on a large space of functions F = FV,? , which depend on V and ?. Its minimizer, denoted by f? , replaces the Bayes rule as the target of our algorithm. The question arises of the price we pay by a considering a relaxation approach: ?What is the relationship between f? and b? ?? More generally, ?What is the price we incur into by estimating the expected risk rather than the misclassification risk?? The relaxation error for a given loss function can be quantified by the following two requirements: 1) Fisher Consistency. A loss function is Fisher consistent if sign(f? (x)) = b? (x) almost surely (this property is related to the notion of classification-calibration [2]). 2

2) Comparison inequalities. The excess misclassification risk, and the excess expected loss are related by a comparison inequality R(sign(f)) ? R(b?) ? ?(E(f) ? E(f?)), for any function f ? F, where ? = ?V,? is a suitable function that depends on V , and possibly on the data distribution. In particular ? should be such that ?(s) ? 0 as s ? 0, so that if fn is a (possibly random) sequence of functions, such that E(fn) ? E(f?) (possibly in probability), then the corresponding sequences of classification rules cn = sign(fn) is Bayes

consistent, i.e. R(cn) ? R(b?) (possibly in probability). If ? is explicitly known, then bounds on the excess expected loss yield bounds on the excess misclassification risk. The relaxation error in the binary case has been thoroughly studied in [2, 14]. In particular, Theorem 2 in [2] shows that if a large margin surrogate loss is convex, differentiable and decreasing in a neighborhood of 0, then the loss is Fisher consistent. Moreover, in this case it is possible to give an explicit expression for the function ?. In particular, for the hinge loss the target function?is exactly the Bayes rule and ?(t) = —t—. For least squares, f? (x) = 2?1 (x) ? 1, and ?(t) = t. The comparison inequality for the square loss can be improved for a suitable class of probability distribution satisfying the so-called Tsybakov noise condition [22], ?X ({x ? X , —f? (x)— ? s}) ? Bq sq , s ? [0, 1], q ¿ 0. Under this condition the probability of points such that ?y (x) ? 21 decreases q+1

polynomially. In this case the comparison inequality for the square loss is given by ?(t) = cq t q+2 , see [2, 27]. Previous Works in Multiclass Classification. From a practical perspective, over the years, several computational solutions to multiclass learning have been proposed. Among others, we mention for example [10, 6, 5, 25, 1, 21]. Indeed, most of the above methods can be interpreted as a kind of relaxation of the original multiclass problem. Interestingly, the study in [15] suggests that the simple one-vs all schemes should be a practical benchmark for multiclass algorithms as it seems experimentally to achive performance that is similar or better than more sophisticated methods. As we previously mentioned from a theoretical perspective a general account of a large class of multiclass methods has been given in [20], building on results in [2] and [28]. Notably, these results show that seemingly intuitive extensions of binary classification algorithms can lead to inconsistent methods. These results, see also [11, 23], are developed in a setting where a classification rule is found by applying a suitable prediction/decoding map to a function f : X ? RT where f is found considering function V : Y ? RT ? R+ . The considered functions have to satisfy P a loss y the constraint y?Y f (x) = 0, for all x ? X . The latter requirement is problematic as it makes the computations in the corresponding algorithms more involved. It is in fact often ignored, so that practical algorithms often come with no consistency guarantees. In all the above papers relaxation is studied in terms of Fisher and Bayes consistency and the explicit form of the function ? is not given. More quantitative results in terms of explicit comparison inequality are given in [4] and (see also [19]), but also need to to impose the ?sum to zero? constraint on the considered function class.

3

A Relaxation Approach to Multicategory Classification

In this section we propose a natural extension of the relaxation approach that avoids constraining the class of functions to be considered, and allows us to derive explicit comparison inequalities. See Remark 1 for related approaches.

c2

? c1

c3

Figure 1: Decoding with simplex coding T = 3. Simplex Coding. We start

by considering a suitable coding/decoding strategy. A coding map turns a label y ? Y into a code vector. The corresponding decoding map given a vector returns a label in 3

Y. Note that this is what we implicitly did while treating binary classification,we encoded the label space Y = {1, 2} using the coding ?1, so that the naturally decoding strategy is simply sign(f (x)). The coding/decoding strategy we study here is described by the following definition. Definition 1 (Simplex Coding). The simplex coding is a map C : Y ? RT ?1 , C(y) = cy , 2 where the code vectors C = {cy — y ? Y} P ? RT ?1 satisfy: 1) kcy k = 1, ?y ? Y, 2)hcy , cy0 i = 1 0 0 ? T ?1 , for y 6= y with y, y ? Y, and 3) y?Y cy = 0. The corresponding decoding is the map D : RT ?1 ? {1, . . . , T }, D(?) = arg maxy?Y h?, cy i , ?? ? RT ?1 . The simplex coding has been considered in [8],[26], and [16]. It corresponds to T maximally separated vectors on the hypersphere ST ?2 in RT ?1 , that are the vertices of a simplex (see Figure 1). For binary classification it reduces to the ?1 coding and the decoding map is equivalent to taking the sign of f . The decoding map has a natural geometric interpretation: an input point is mapped to a vector f (x) by a function f : X ? RT ?1 , and hence assigned to the class having closest code 2 2 vector (for y, y 0 ? Y and ? ? RT ?1 , we have kcy ? ?k ? kcy0 ? ?k ? hcy0 , ?i ? hcy , ?i). Relaxation for Multiclass Learning. We use the simplex coding to propose an extension of binary classification. Following the binary case, the relaxation can be described in two steps: 1. using the simplex coding, the indicator function is upper bounded by a non-negative loss function V : Y ? RT ?1 ? R+ , such that 1I[b(x)6=y] (x, y) ? V (y, C(b(x))), for all b : X ? Y, and x ? X , y ? Y,

2. rather than C ? b we consider functions with values in f : X ? RT ?1 , so that V (y, C(b(x))) ? V (y, f (x)), for all b : X ? Y, f : X ? RT ?1 and x ? X , y ? Y. In the next section we discuss several loss functions satisfying the above conditions and we study in particular the extension of the least squares and SVM loss functions. Multiclass Simplex Loss Functions. Several loss functions for binary classification can be naturally extended to multiple classes using the simplex coding. Due to space restriction, in this paper we focus on extensions of the least squares and SVM loss functions, but our analysis can be generalized to a large class of loss functions, including extensions of logistic and exponential loss functions 2 (used in boosting). The Simplex Least Square loss (S-LS) is given by V (y, f (x)) = kcy ? f (x)k , and reduces to the usual least square approach to binary classification for T = 2. One natural extension of the SVM?s hinge loss in this setting would be to consider the Simplex Half space SVM loss (SH-SVM) V (y, f (x)) = —1 ? hcy , f (x)i—+ . We will see in the following that while this loss function would induce efficient algorithms in general is not Fisher consistent unless further constraints are assumed. These latter constraints would considerably slow down the computations. We then consider Simplex Cone SVM (SC-SVM), which is defined as a second loss function

P

1 V (y, f (x)) = y0 6=y T ?1 + hcy0 , f (x)i . The latter loss function is related to the one considered +

in the multiclass SVM proposed in [10]. We will see that it is possible to quantify the relaxation error of the loss function without requiring further constraints. Both of the above SVM loss functions reduce to the binary SVM hinge loss if T = 2. Remark 1 (Related approaches). An SVM loss is considered in [8] where Vq (y, f (x)) = P cy ?cy0 1 T ? 0 0 0 y 0 6=y —? ? hf (x), vy (y)i—+ and vy (y) = kcy ?c 0 k , with ? = hcy , vy (y)i = T ?1 . More 2 y recently [26] considered the loss function V (y, f (x)) = —kcy ? f (x)k ? ?—+ , and a simplex multiP class boosting loss was introduced in [16], in our notation V (y, f (x)) = e?hcy ?cy0 ,f (x)i . j6=y

While all those losses introduce a certain notion of margin that makes use of the geometry of the simplex coding, it is not to clear how to derive explicit comparison theorems and moreover the computational complexity of the resulting algorithms scales linearly with the number of classes in the case of the losses considered in [16, 26] and O((nT)?), ? ? {2, 3} for losses considered in [8] . 4

Figure 2: Level sets of the different losses considered for T = 3. A classification is correct if an input (x, y) is mapped to a point f (x) that lies in the neighborhood of the vertex cy . The shape of the neighborhood is defined by the loss. It takes the form of a cone supported on a vertex, in the case of SC-SVM, a half space delimited by the hyperplane orthogonal to the vertex in the case of the SH-SVM, and a sphere centered on the vertex, in the case of S-LS.

4

Relaxation Error Analysis

T ?1 If we consider the simplex coding, a function f taking values in R , and the decoding operator R D, the misclassification risk can also be written as: R(D(f)) = X (1 ? ?D(f (x)))d?X (x). Then, following a relaxation approach, we replace the misclassification loss by the expected risk induced by one of the loss functionsR V defined in the previous section. As in the binary case we consider p the expected loss E(f) = V (y, f (x))d?(x, y). Let Lp (X , ?X) = {f : X ? RT ?1 — kf k? = R p kf (x)k d?X (x) ¡ ?}, p ? 1.

The following theorem studies the relaxation error for SH-SVM, SC-SVM, and S-LS loss functions. Theorem 1. For SH-SVM, SC-SVM, and S-LS loss functions, there exists a p such that E : Lp (X , ?X) ? R+ is convex and continuous. Moreover, 1. The minimizer f? of E over F = {f ? Lp (X , ?X) — f (x) ? K a.s.} exists and D(f?) = b? . 2. For any f ? F, R(D(f)) ? R(D(f?)) ? CT (E(f) ? E(f?))? , where the expressions of p, K, f? , CT , and ? are given in Table 1. Loss SH-SVM SC-SVM

p 1 1

K conv(C) RT ?1

S-LS

2

RT ?1

f? cb? cb? P

y?Y

?y cy

CT T ?1 T q? 1

2(T ?1) T

? 1 1 1 2

Table 1: conv(C) is the convex hull of the set C defined in (1). The proof of this theorem is given, in Theorems 1 and 2 for S-LS, and Theorems 3, and 4 for SCSVM and SH-SVM respectively, in Appendix B. The above theorem can be improved for Least Squares under certain classes of distribution . Toward this end we introduce the following notion of misclassification noise that generalizes Tsybakov?s noise condition. Definition 2. Fix q ¿ 0, we say that the distribution ? satisfies the multiclass noise condition with parameter Bq , if

T ?1

?X x?X —0? min (cD(f? (x)) ? cj , f? (x)) ? s ? Bq sq , (2) T j6=D(f? (x)) where s ? [0, 1]. 5

If a distribution ? is characterized by a very large q, then, for each x ? X , f? (x) is arbitrarily close to one of the coding vectors. For T = 2, the above condition reduces to the binary Tsybakov noise. Indeed, let c1 = 1, and c2 = ?1, if f? (x) ¿ 0, 21 (c1 ? c2)f? (x) = f? (x), and if f? (x) ¡ 0, 1 2 (c2 ? c1)f? (x) = ?f? (x). The following result improves the exponent of simplex-least square to

q+1 q+2

¿

1 2

:

Theorem 2. For each f ? L2 (X , ?X), if (2) holds, then for S-LS we have the following inequality,

R(D(f)) ? R(D(f?)) ? K

2(T ? 1) (E(f) ? E(f?)) T

q+1 q+2

,

(3)

2q+2 p with K = 2 Bq + 1 q+2 . Remark 2. Note that the comparison inequalities show a tradeoff between the exponent ? and the constant C(T) for S-LS and SVM losses. While the constant is order T for SVM it is order 1 for SLS, on the other hand the exponent is 1 for SVM losses and 21 for S-LS. The latter could be enhanced to 1 for close to separable classification problems by virtue of the Tsybakov noise condition. Remark 3. The comparison inequalities given in Theorems 1 and 2 can be used to derive generalization bounds on the excess misclassification risk. For least squares min-max sharp bound, for vector valued regression are known [3]. Standard techniques for deriving sample complexity bounds in binary classification extended for multi-class SVM losses, are found in [7] and could be adapted to our setting. The obtained bound are not known to be tight. Better bounds akin to those in [18], will be subject of future work.

5

Computational Aspects and Regularization Algorithms

The simplex coding framework allows us to extend batch and online kernel methods to the Multiclass setting. Computing the Simplex Coding. We begin by noting that ! the simplex coding can be easily com¿ 1 uq puted via the recursion: C[i+1] = , C[2] = [1?1], where u = (? 1i ? ? ?? 1i) v C[i] ? 1 ?

i12 (column vector in Ri) and v = (0, . . . , 0)(column vector in Ri?1) (see Algorithm C.1). Indeed we have the following result (see the Appendix C.1 for the proof). Lemma 1. The T columns of C[T] are a set of T ? 1 dimensional vectors satisfying the properties of Definition 1. The above algorithm stems from the observation that the simplex in RT ?1 can be obtained by projecting the simplex in RT onto the hyperplane orthogonal to the element (1, . . . , 0) of the canonical basis in RT . Regularized Kernel Methods. We consider regularized methods of the form (1), induced by simplex loss functions and where the hypothesis space is a vector-valued reproducing kernel Hilbert space H(VV-RKHS) the regularizer is the corresponding norm ——f ——2H . See Appendix D.2 for a brief introduction to VV-RKHS. In the following, we consider that if f minimizes (1) for R(f) = ——f ——2H Pn a class of kernels K such T ?1 [12], where we note that the coefficients we have that f (x) = i=1 K(x, xi)ai , ai ? R are vectors in RT ?1 . In the case that the kernel is induced by a finite dimensional feature map, k(x, x0) = h?(x), ?(x0)i , where ? : X ? Rp , and h?, ?i is the inner product in Rp , we can write each function in H as f (x) = W ?(x), where W ? R(T ?1)?p . It is known [12] that the representer theorem [9] can be easily extended to a vector valued setting, a simplex version of Tikhonov regularization is given by fS? (x) = Pn so that that minimizerTof ?1 , for all x ? X , where the explicit expression of the coefficients j=1 k(x, xj)aj , aj ? R depends on the considered loss function. We use the following notation: K ? Rn?n , Kij = k(xi , xj), ?i, j ? {1 . . . n}, A ? Rn?(T ?1) , A = (a1 , ..., an)T . Simplex Regularized Least squares (S-RLS). S-RLS is obtained by substituting the simplex least square loss in the Tikhonov functional. It is easy to see [15] that in this case the coefficients 6

?T X ? + ?nI)W = X ? T Y? in the linear case, where must satisfy either (K + ?nI)A = Y? or (X n?p ? ¿ n?(T ?1) ? ? ? X?R , X = (?(x1), ..., ?(xn)) and Y ? R , Y = (cy1 , ..., cyn)¿ . Interestingly, the classical results from [24] can be extended to show that the value fSi (xi), obtained computing the solution fSi removing the i ? th point from the training set (the leave one out so? ? lution), can be computed in closed form. Let floo ? Rn?(T ?1) , floo = (fS?1 (x1), . . . , fS?n (xn)). Let K(?) = (K + ?nI)?1 and C(?) = K(?)Y? . Define M (?) ? Rn?(T ?1) , such that: ? M (?)ij = 1/K(?)ii , ? j = 1 . . . T ? 1. One can show that floo = Y? ? C(?) M (?), where P n 1 is the Hadamard product [15]. Then, the leave-one-out error n i=1 1Iy6=D(fSi (x)) (yi , xi), can ? T X). ? be minimized at essentially no extra cost by precomputing the eigen decomposition of K (or X Simplex Cone Support Vector Machine (SC-SVM). Using standard reasoning it is easy to show that (seePAppendix C.2), for the SC-SVM the coefficients in the representer theorem are given by ai = ? y6=yi ?iy cy , i = 1, . . . , n, where ?i = (?iy)y?Y ? RT , i = 1, . . . , n, solve the quadratic programming (QP) problem ? ? n X T ? 1 X ? X 0 1 max ? ?iy Kij Gyy0 ?jy + ?iy (4) ? T ?1 ?1 ,...,?n ?RT ? 2 0 i=1 y=1

y,y ,i,j

subject to 0 ?

?iy

? C0 ?y,yi , ? i = 1, . . . , n, y ? Y

1 , ?i = (?iy)y?Y ? RT , for i = 1, . . . , n and ?i,j where Gy,y0 = hcy , cy0 i ?y, y 0 ? Y and C0 = 2n? is the Kronecker delta. Simplex Halfspaces Support Vector Machine (SH-SVM). A similar, yet more more complicated procedure, can be derived for the SH-SVM. Here, we omit this derivation and observe instead that if we neglect the convex hull constraint from Theorem 1, that requires f (x) ? co(C) for almost all x ? X , then the SH-SVM has an especially simple formulation at the price of loosing consistency guarantees. In fact, in this case the coefficients are given by ai = ?i cyi , i = 1, . . . , n, where ?i ? R, with i = 1, . . . , n solve the quadratic programming (QP) problem n

max

?1 ,...,?n ?R

?

X 1X ?i ?i Kij Gyi yj ?j + 2 i,j i=1

subject to 0 ? ?i ? C0 , ? i = 1 . . . n, 1 2n? . 3

where C0 = The latter formulation can be solved at the same complexity of the binary SVM (worst case O(n)) but lacks consistency. Online/Incremental Optimization The regularized estimators induced by the simplex loss functions can be computed by means of online/incremental first order (sub) gradient methods. Indeed, when considering finite dimensional feature maps, these strategies offer computationally feasible solutions to train estimators for large datasets where neither a p by p or an n by n matrix fit in memory. Following [17] we can alternate a step of stochastic descent on a data point : Wtmp = (1 ? ?i ?)Wi ? ?i ?(V (yi , fWi (xi))) and a projection on the Frobenius ball Wi = min(1, ??——W1 ——)Wtmp (See tmp F Algorithn C.5 for details.) The algorithm depends on the used loss function through computation of the (point-wise) subgradient ?(V). The latter can be easily computed for all the loss functions previously discussed. For the SLS loss we haveP ?(V (yi , fW (xi))) = 2(cyi ? W xi)x¿ i , while for the SC1 SVM loss we have ?(V (yi , fW (xi))) = (k?Ii ck)x¿ where I = {y = 6 y — hc i i y , W xi i ¿ ? T ?1 }. i For the SH-SVM loss we have: ?(V (y, fW (xi))) = ?cyi x¿ i if cyi W xi ¡ 1 and 0 otherwise . 5.1

Comparison of Computational Complexity

The cost of solving S-RLS for fixed ? is in the worst case O(n3) (for example via Cholesky decomposition). If we are interested in computing the regularization path for N regularization parameter values, then as noted in [15] it might be convenient to perform an eigendecomposition of the kernel matrix rather than solving the systems N times. For explicit p?dimensional feature maps the cost is O(np2), so that the cost of computing the regularization path for simplex RLS algorithm is O(min(n3 , np2)) and hence independent of T . One can contrast this complexity with that of a n?aive One Versus All (OVA) approach that would lead to a O(N n3 T) complexity. Simplex SVMs can be solved using solvers available for binary SVMs that are considered to have complexity O(n?) with ? ? {2, 3}(the actual complexity scales with the number of support vectors) . For SC-SVM, though, 7

we have nT rather than n unknowns and the complexity is (O(nT)?). SH-SVM in which we omit the constraint, can be trained with the same complexity as the binary SVM (worst case O(n3)) but lacks consistency. Note that unlike for S-RLS, there is no straightforward way to compute the regularization path and the leave one out error for any of the above SVM. The online algorithms induced by the different simplex loss functions are essentially the same. In particular, each iteration depends linearly on the number of classes.

6

Numerical Results

We conduct several experiments to evaluate the performance of our batch and online algorithms, on 5 UCI datasets as listed in Table 2, as well as on Caltech101 and Pubfig83. We compare the performance of our algorithms to one versus all svm (libsvm) , as well as simplex- based boosting [16]. For UCI datasets we use the raw features, on Caltech101 we use hierarchical features (hmax), and on Pubfig83 we use the feature maps from [13]. In all cases the parameter selection is based either on a hold out (ho) (80% training ? 20% validation) or a leave one out error (loo). For the model selection of ? in S-LS, 100 values are chosen in the range [?min , ?max],(where ?min and ?max , correspond to the smallest and biggest eigenvalues of K). In the case of a Gaussian kernel (rbf) we use a heuristic that sets the width of the Gaussian ? to the 25-th percentile of pairwise distances between distinct points in the training set. In Table 2 we collect the resulting classification accuracies. SC-SVM Online (ho) SH-SVM Online (ho) S-LS Online (ho) S-LS Batch (loo) S-LS rbf Batch (loo) SVM batch ova (ho) SVM rbf batch ova (ho) Simplex boosting [16]

Landsat 65.15% 75.43% 63.62% 65.88% 90.15% 72.81% 95.33% 86.65%

Optdigit 89.57% 85.58% 91.68% 91.90% 97.09% 92.13% 98.07% 92.82%

Pendigit 81.62% 72.54% 81.39% 80.69% 98.17% 86.93% 98.88% 92.94%

Letter 52.82% 38.40% 54.29% 54.96% 96.48% 62.78% 97.12% 59.65%

Isolet 88.58% 77.65% 92.62% 92.55% 97.05% 90.59% 96.99% 91.02%

Ctech 63.33% 45% 58.39% 66.35% 69.38% 70.13% 51.77% ?

Pubfig83 84.70% 49.76% 83.61% 86.63% 86.75% 85.97% 85.60% ?

Table 2: Accuracies of our algorithms on several datasets. As suggested by the theory, the consistent methods SC-SVM and S-LS have large performance advantage over SH-SVM (where we omitted the convex hull constraint). Batch methods are overall superior to online methods. Online SC-SVM achieves the best results among online methods. More generally, we see that rbf S- LS has the best performance amongst the simplex methods, including simplex boosting [16]. We see that S-LS rbf achieves essentially the same performance as One Versus All SVM-rbf.

2 References

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multi-class to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113?141, 2000. [2] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Journal of

the American Statistical Association, 101(473):138?156, 2006. [3] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. Foundations of Computational Mathematics, 2006. [4] D. Chen and T. Sun. Consistency of multiclass empirical risk minimization methods based in convex loss. Journal of machine learning, X, 2006. [5] Crammer.K and Singer.Y. On the algorithmic implementation of multiclass kernel-based vector machines. JMLR, 2001. [6] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via errorcorrecting output codes. Journal of Artificial Intelligence Research, 2:263?286, 1995. [7] Yann Guermeur. Vc theory of large margin multi-category classiers. Journal of Machine Learning Research, 8:2551?2594, 2007. 8

[8] Simon I. Hill and Arnaud Doucet. A framework for kernel-based multicategory classification. J. Artif. Int. Res., 30(1):525?564, December 2007. [9] G. Kimeldorf and G. Wahba. A correspondence between bayesian estimation of stochastic processes and smoothing by splines. Ann. Math. Stat., 41:495?502, 1970. [10] Lee.Y, L.Yin, and Wahba.G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association, 2004. [11] Liu.Y. Fisher consistency of multicategory support vector machines. Eleventh International Conference on Artificial Intelligence and Statistics, 289-296, 2007. [12] C.A. Micchelli and M. Pontil. On learning vector?valued functions. Neural Computation, 17:177?204, 2005. [13] N. Pinto, Z. Stone, T. Zickler, and D.D. Cox. Scaling-up biologically-inspired computer vision: A case-study on facebook. 2011. [14] M.D. Reid and R.C. Williamson. Composite binary losses. JMLR, 11, September 2010. [15] Rifkin.R and Klautau.A. In defense of one versus all classification. journal of machine learning, 2004. [16] Saberian.M and Vasconcelos .N. Multiclass boosting: Theory and algorithms. In NIPS 2011, 2011. [17] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated subgradient solver for svm. In Proceedings of the 24th ICML, ICML ?07, pages 807?814, New York, NY, USA, 2007. ACM. [18] I. Steinwart and A. Christmann. Support vector machines. Information Science and Statistics. Springer, New York, 2008. [19] Van de Geer.S Tarigan.B. A moment bound for multicategory support vector machines. JMLR 9, 2171-2185, 2008. [20] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. In Proceedings of the 18th Annual Conference on Learning Theory, volume 3559, pages 143? 157. Springer, 2005. [21] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. JMLR, 6(2):1453?1484, 2005. [22] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. Annals of Statistics, 32:135?166, 2004. [23] Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In Proceedings of Neural Information Processing Systems (NIPS 2011), 2011. [24] G. Wahba. Spline models for observational data, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA, 1990. [25] Weston and Watkins. Support vector machine for multi class pattern recognition. Proceedings of the seventh european symposium on artificial neural networks, 1999. [26] Tong

Tong Wu and Kenneth Lange. Multicategory vertex discriminant analysis for highdimensional data. Ann. Appl. Stat., 4(4):1698?1721, 2010. [27] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. Constructive Approximation, 26(2):289?315, 2007. [28] T. Zhang. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5:1225?1251, 2004. [29] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. The Annals of Statistics, Vol. 32, No. 1, 56134, 2004.

9