Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2011-043
CBCL-305

September 27, 2011

Multi-Class Learning: Simplex Coding And Relaxation Error

Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jacques E. Slotine

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

MULTI-CLASS LEARNING: SIMPLEX CODING AND RELAXATION ERROR

YOUSSEF MROUEH$^{\sharp,\ddagger}$, TOMASO POGGIO$^{\sharp}$, LORENZO ROSASCO$^{\sharp,\ddagger}$ JEAN-JACQUES E. SLOTINE†
$\sharp$ - *CBCL, MCGOVERN INSTITUTE, BCS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY*
$\dagger$ - ISTITUTO ITALIANO DI TECNOLOGIA
$\dagger$ - *ME, BCS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY*
YMROUEH,TP, LROSASCO,JJS@MIT.EDU

Abstract. *We study multi-category classification in the framework of computational learning theory. We show how a relaxation approach, which is commonly used in binary classification, can be generalized to the multi-class setting. We propose a vector coding, namely the simplex coding, that allows to introduce a new notion of multi-class margin and cast multi-category classification into a vector valued regression problem. The analysis of the relaxation error be quantified and the binary case is recovered as a special case of our theory. From a computational point of view we can show that using the simplex coding we can design regularized learning algorithms for multi-category classification that can be trained at a complexity which is independent to the number of classes.*

1. Problem Setting. We consider an input space $\mathcal{X} \subset \mathbb{R}^d$, and output space $\mathcal{Y} = \{1, \dots, T\}$. Given a probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ we let $\rho_{\mathcal{X}}$ be the marginal probability on $\mathcal{X}$ and $\rho_j(x) = \rho(j|x)$ the conditional distribution of class j given x, for each $j = 1, \dots, T$ and $x \in \mathcal{X}$. A training set is a sequence $(x_i, y_i)_{i=1}^n$ sampled i.i.d. with respect to ρ. A classification rule is a map $c : \mathcal{X} \to \mathcal{Y}$ and its properties can be measured via the misclassification error (or misclassification risk),

$$R(c) = \mathbb{P}(c(x) \neq y),$$

which is minimized, by the Bayes rule $b_\rho(x) = \arg\max_{j=\{1,\dots,T\}} \rho_j(x)$. This risk functional cannot be directly minimized for two reasons: 1) the true probability distribution is unknown, 2) it requires optimizing a non convex functional over a set of discrete valued functions, in fact the risk can be written as $R(c) = \int \Theta(yc(x)) d\rho(x, y)$ where $\Theta(x) = 1$ if $x < 0$ and 0 otherwise. While we can tackle the first issue looking at the empirical error on the data– rather than the risk, in this work we consider the second issue.

The typical approach in binary classification, i.e. $T = 2$, is based on the following steps. First real valued functions are considered in place of binary valued ones so that a classification rule is defined defined by the sign of a function. Second, the *margin* of a function is defined to be the quantity $m = yf(x)$ and $\Theta(m)$ is replaced by a *margin loss* function $V(m)$ where V is a non-negative and convex. This *relaxation* approach introduces an error which can be quantified. In fact, if we define $\mathcal{E}(f) = \int V(yf(x)) d\rho(x, y)$, and let f_ρ be its minimizer, it is possible to prove [2] that if V is decreasing in a neighborhood of 0, and differentiable in 0, then $b_\rho(x) = \text{sign}(f_\rho)(x)$, namely the loss is *classification calibrated*. Moreover, for any measurable function $f : \mathcal{X} \longmapsto \mathbb{R}$ and probability distribution ρ we can derive a so called *comparison theorem*, that is, there exits a function $\psi_V : [0, 1] \mapsto [0, \infty)$

$$\psi_V(R(\text{sign}(f)) - R(\text{sign}(f_\rho))) \leq \mathcal{E}(f) - \mathcal{E}(f_\rho).$$

For example for the the square loss $V(m) = (1 - m)^2$ we have $\psi_V(t) = t^2$ and for the hinge loss $V(m) = |1 - m|_+$ we have $\psi_V(t) = t$. In this note we discuss how the above approach can be extended to $T \geq 2$.

1.1. Simplex Coding and Relaxation Error. The following definition is at the core of our approach.

DEFINITION 1.1. *The simplex coding is a map $C : \{1, \dots, T\} \to \mathbb{R}^{T-1}$ such that for $i = 1, \dots, T$, $C(i) = a_i$, where the code vectors $\mathcal{A} = \{a_1, \dots, a_T\} \subset \mathbb{R}^{T-1}$ satisfy*

$$\|a_i\|^2 = 1, \quad \forall i = 1, \dots, T, \quad \langle a_i, a_j \rangle = -\frac{1}{T-1}, \quad \forall i, j = 1, \dots, T, \ i \neq j,$$

and $\sum_{i=1}^T a_i = 0$. The corresponding decoding is the map $D : \mathbb{R}^{T-1} \to \{1, \dots, T\}$ $D(\alpha) = \arg\max_{i=1,\dots,T} \langle \alpha, a_i \rangle$, $\forall \alpha \in \mathbb{R}^{T-1}$.

The simplex coding corresponds to the T most separated vectors on the hypersphere $\mathbb{S}^{T-2}$ in $\mathbb{R}^{T-1}$, which are the vertices of the simplex. For binary classification it reduces to the ± 1 coding. The decoding map has a natural geometric interpretation: an input point is mapped to a vector $f(x)$ by a vector regressor and hence assigned to the class having closer code vector. The projection $\langle f(x), a_j \rangle$ is precisely the multi-class extension of the notion of margin that we discussed in binary classification and allows to extend the relaxation approach. Using the simplex coding the misclassification risk can be written as

$$R(D(f)) = \int \Theta(\langle f(x), a \rangle) d\rho(a, x) = \sum_{j=1}^{T} \int \Theta(\langle f(x), a_j \rangle) \rho_j(x) d\rho_{\mathcal{X}}(x).$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^{T-1}$. Then, we can simply consider any margin loss, e.g. hinge or logistic loss, and can replace the misclassification risk by the expected risk $\mathcal{E}(f) = \int V(\langle f(x), y \rangle) d\rho(x, y)$. Note that the square loss can be seen as margin loss if f is on the sphere.

1.2. Relaxation error analysis. As in the binary case, it is natural to ask what is the error we incur into by considering a convex relaxation of the classification problem. Interestingly, the results in the binary case can be now extended to the multiclass setting. In fact, also in this case if V is decreasing in a neighborhood of 0, and differentiable in 0, then $b_\rho(x) = D(f_\rho)(x)$, where the sign is replaced by the decoding map. Comparison theorems can also be proved. For example, for the the square loss $V(m) = (1 - m)^2$ we have $\psi_V(t) = t^2/(T-1)^2$ and for the hinge loss $V(m) = |1 - m|_+$ we have $\psi_V(t) = t/(T-1)$, where we see the price to pass from $T = 2$ to $T \geq 2$. While we omit further details we mention here that a notion of (multi) classification noise related to the one used in binary classification [2] can also be defined, which allows to improve the above results. Compared to previous works [7, 8] we see that the simplex coding allows to avoid any further constraint to the function class.

1.3. Computing the simplex coding. The simplex coding can be easily implemented and can induce regularized learning methods for multi-category classification that can be trained at the same computational complexity of a binary classification problem, hence independently to the number of classes.

We start discussing a simple algorithm for the generation of the simplex coding. We use a recursive projection of simpleces, by observing that the simplex in $\mathbb{R}^{T-1}$, can be obtained projecting the simplex in $\mathbb{R}^T$ on the hyperplane orthogonal to the element $(1, \ldots, 0)$ of the canonical basis in $\mathbb{R}^T$. Let $C[T-1]$ be the simplex code associated to T classes, $C[T-1]$ is a matrix of dimension $T \times (T-1)$. We have the following recursion, where at each step we add a dimension, and backproject:

$$C[T] = \begin{pmatrix} 1 & u \\ v & C[T-1] \times \sqrt{1 - \frac{1}{T^2}} \end{pmatrix} \tag{1.1}$$

$$C[1] = [1; -1]$$

Where u is row vector of dimension T, $u = (-\frac{1}{T} \cdots -\frac{1}{T})$, and v a column vector of dimension T, $v = (0, \ldots, 0)$.

Kernels and Regularization Algorithms.. Next we need to recall some basic concepts in the theory of reproducing kernel Hilbert spaces (RKHS) of vector valued functions. The definition of RKHS for vector valued functions parallels the one in the scalar [1], with the main difference that the reproducing kernel is now *matrix* valued – see [3] and references therein. A reproducing kernel is a symmetric function $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{D \times D}$, such that for any $x, x' \in \mathcal{X}$ $\Gamma(x, x')$ is a positive semi-definite *matrix*. A vector valued RKHS is a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}^D$, such that for very $c \in \mathbb{R}^D$, and $x \in \mathcal{X}$, $\Gamma(x, \cdot)c$ belongs to $\mathcal{H}$ and moreover Γ has the reproducing property $\langle f, \Gamma(\cdot, x)c \rangle_{\mathcal{H}} = \langle f(x), c \rangle$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. The choice of the kernel corresponds to the choice of the representation (parameterization) for the functions of interest. In fact it can be shown that any function in a RKHS with kernel Γ, is in the completion of the span of $\Gamma(x_i, \cdot)$ with $c_j \in \mathbb{R}^D$. Given the reproducing property, the norm of f can be written as $\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{\infty} \langle c_j, \Gamma(x_i, x_j) c_j \rangle$. Note that for $D = 1$ we recover the classic theory of scalar valued RKHS. In the following we restrict our attention to kernels of the form $\Gamma(x, x') = k(x, x')A, \quad A = I$, where k is a scalar reproducing kernel. As we discuss elsewhere [6] the choice of A corresponds to a prior belief that different components can be related. In fact, if we let $f = (f_1, \ldots, f_D)$ it is possible to see that the entry $A_{t,t'}$ defines the relation between f_t and $f_{t'}$. For the sake of simplicity we restrict ourselves to $A = I$, hence treating each component as independent. This case is directly comparable to the one-vs-all approach.

Next, we discuss the properties of different learning algorithms using the simplex coding. We use the following notation, $Y \in \mathbb{R}^{n\times(T-1)}, Y = (y_1, ..., y_n),\ y_i \in \mathcal{A},\ i = 1, \dots, n$; $K \in \mathbb{R}^{n\times n}, K_{ij} = k(x_i, x_j)$; $C \in \mathbb{R}^{n\times(T-1)}, C = (c_1, c_2, ..., c_n)$. We consider algorithms defined by the minimization of a Tikhonov functional

$$\mathcal{E}_n^{\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} V(\langle y_i, f(x_i)\rangle) + \frac{\lambda}{2}\left\| f \right\|_{\mathcal{H}}^2 ,$$

where in particular $V(\langle y_i, f(x_i)\rangle)$ will be either the square loss or a margin loss (in particular the SVM's hinge loss). It is well known [5] that the representer theorem [4] can be easily extended to a vector valued setting to show that minimizer of the above functional over $\mathcal{H}$ can be written as $f(x) = \sum_{j=1}^{n} k(x, x_j)c_j, \quad c_j \in \mathbb{R}^{T-1}$. The choice of different loss functions induce different strategy to compute C.

If we choose let the loss to be $\left\| y - f(x) \right\|^2$ it is easy to see that, $(K + \frac{\lambda}{2}nI)C = Y$. If we want to compute a solution for N values of λ, by using SVD to perform the matrix inversion, we can still compute a regularized inverse in essentially $O(n^3)$ but the multiplication $(K + \frac{\lambda}{2}nI)^{-1}Y$ is going to be $O(n^2TN)$, which is linear in T. Note that this complexity is still much better than the one-vs-all approach that would give a $O(n^3TN)$. If we choose let the loss to be $|1 - \langle y, f(x)\rangle|_+$, following standard reasonings from the binary case [9] to see that we have to solve the problem

$$\max_{\alpha\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\alpha^{\top}Q\alpha\right\}, \qquad\qquad 0 \le \alpha_i \le \frac{1}{n\lambda}, i = 1, \dots, n$$

where $Q_{ij} = K(x_i, x_j)y_i^T y_j$ and $c_k = \alpha_k y_k$ where $\alpha_k \in \mathbb{R}$, for $k = 1, \dots, n$. Note that the optimization is now only over the n dimensional vector α and T appears only in the computation of the matrix Q. Training for fixed C is hence independent of the number of classes and is essentially $O(n^3)$ in the worst case. If we are interested into N different values of λ we would get a complexity $O(n^3N)$. Note that more sophisticated strategy to compute the whole regularization path could be coupled with the use of simplex coding.

REFERENCES

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] P.L. Bartlett, M.I. Jordan, and J.D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2005. To appear. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

[3] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl. (Singap.)*, 4(4):377–408, 2006.

[4] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.

[5] C.A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

[6] Y. Mroueh and L. Poggio, T.and Rosasco. Regularization predicts while discovering taxonomy. Technical Report MIT-CSAIL-TR-2011-029/CBCL-299, Massachusetts Institute of Technology,cambridge,MA, june 2011.

[7] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, May 2007.

[8] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 2004.

[9] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.