Multiclass Learning with Simplex Coding: Supplementary Material

Youssef Mroueh[♯,‡], **Tomaso Poggio**[♯], **Lorenzo Rosasco**[♯,‡] **Jean-Jacques E. Slotine**[†]
♯ - *CBCL, McGovern Institute, MIT;* † - IIT; † - *ME, BCS, MIT*
ymroueh, lrosasco,jjs@mit.edu tp@ai.mit.edu

A Loss Functions and Risk Minimization in a General Supervised Learning Setting

In this section we give a broader discussion on the minimization of the expected risk induced by a loss function. We consider a general setting and introduce minimal assumptions on the loss function allowing to exploit results from convex analysis and variational calculus for integral functionals [5]. Our approach follows the ideas in [12] and can be contrasted on the usual approach based on the study of the inner risk see [2, 13, 9, 15, 14, 11] and in particular Section 3.1 in [10]. We refer to appendix (D) for basic convex analysis tools used throughout this section.

A.1 A General Supervised Learning setting

Let (X, Y) be two random variables with values in a measurable space $\mathcal{X}$ and a Polish space $\mathcal{Y}$, respectively. We denote by μ the law of (X, Y) on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, by $\rho_{\mathcal{X}}$, the law of X on $\mathcal{X}$. For any measurable function g, we have

$$\int \mu(z)g(z) = \int \mu(x, y)g(x, y) = \int d\rho_{\mathcal{X}}(x) \int d\rho_y(x)g(x, y),$$

where $\rho_y(x)$ defines a measure on $\mathcal{Y}$ for almost all $x \in \mathcal{X}$, see Lemma A.3.16 in [10].

Let $(\mathcal{G}, \langle \cdot, \cdot \rangle)$, be a separable Hilbert space. We introduce some function spaces naturally associated to this setting. For $p \in [1, \infty[$, let

$$L^p(\mathcal{X}, \rho_{\mathcal{X}}) = \left\{ f : \mathcal{X} \to \mathcal{G} \mid \|f\|_{\rho,p} = \left(\int \|f(x)\|^p \, d\rho_{\mathcal{X}}(x) \right)^{\frac{1}{p}} < \infty \right\},$$

and $\|f\|_{\rho} = \|f\|_{\rho,2}$ for $f \in L^2(\mathcal{X}, d\rho_{\mathcal{X}})$. Similarly, let

$$L^p(\mathcal{Z}, d\mu) = \left\{ f : \mathcal{Z} \to \mathcal{G} \mid \|f\|_{\mu,p} = \left(\int \|g(x, y)\|^p \, d\mu(x, y) \right)^{\frac{1}{p}} < \infty \right\},$$

and $\|f\|_{\mu} = \|f\|_{\mu,2}$ for $f \in L^2(\mathcal{Z}, d\mu)$.
Consider the embedding $j : L^p(\mathcal{X}, \rho_{\mathcal{X}}) \to L^p(\mathcal{Z}, \mu)$ defined by

$$(jf)(z) = f(x), \qquad \text{for almost all } (x, y) \in \mathcal{Z},$$

and $f \in L^p(\mathcal{X}, d\rho)$. Clearly j is linear and bounded by one. For $q \in]1, \infty]$, such that $\frac{1}{p} + \frac{1}{q} = 1$, let $j^* : L^q(\mathcal{Z}, \mu) \to L^q(\mathcal{X}, \rho_{\mathcal{X}})$ be the adjoint of j. A version of the Riesz representation theorem ensures that

$$(j^*g)(x) = \int d\rho_y(x)g(x, y), \qquad \text{for almost all } x \in \mathcal{X},$$

and $g \in L^q(\mathcal{Z}, d\mu)$ (see Theorem 4.11-4.14 in [3] for the case $\mathcal{G} = \mathbb{R}$ and Theorem 2.3 in [7] for general $\mathcal{G}$).

A.2 General Nemitski Loss Functions

The class of loss functions we consider has been proposed in [12] for $\mathcal{G} = \mathbb{R}$ and further used in [10]. Here we consider the extension to the case of $\mathcal{G}$ being a Hilbert space, the only difference being that continuity is no longer implied by convexity and needs to be separately assumed.

Definition 1 (Nemitski Loss Function)**.** *Given $p \in [1, +\infty[$, a measurable function $V : \mathcal{Y} \times \mathcal{G} \to [0, +\infty[$ such that*

1. *for all $y \in \mathcal{Y}$ the function $V(y, \cdot)$ is convex and continuous on $\mathcal{G}$;*
2. *there are $b \in [0, +\infty[$ and $a : \mathcal{Y} \to [0, +\infty[$ such that*

$$V(y, w) \leq a(y) + b \|w\|^p \qquad \forall w \in \mathcal{G},\ y \in \mathcal{Y} \tag{1}$$

$$\int a(y) d\mu(x, y) \quad < \quad +\infty, \tag{2}$$

is called a Nemitski p-loss function with respect to μ.

Loss functions satisfying the above conditions include the following classes.

1. Let $C : \mathcal{Y} \to \mathcal{G}$, be a measurable map, such that

$$\int \|C(y)\|^2 d\mu(x, y) < +\infty,$$

   then $V(y, w) = \|C(y) - w\|^2$ is Nemitski 2-loss function with respect to μ.
2. Let $V(y, \cdot)$ be Lipschitz on $\mathcal{G}$ with a Lipschitz constant independent of y and

$$\int V(y, 0) d\mu(x, y) < +\infty,$$

   then V is Nemitski 1-loss function with respect to μ.

A.3 Risk Minimization

Given a loss function, the corresponding expected risk is defined by,

$$\mathcal{E}(f) = \int V(y, f(x)) d\mu(x, y).$$

The following theorem characterizes the properties of the expected risk defined by a Nemitski loss function and provides an explicit expression for its subdifferential.

Theorem 1. *Let $V : \mathcal{Y} \times \mathcal{G} \to \mathbb{R}^+$, be a p-Nemitski loss function with respect to ρ. Then,*

- *the expected risk $\mathcal{E} : L^p(\mathcal{X}, \rho_{\mathcal{X}}) \to \mathbb{R}^+$ is a well defined, convex continuous functional.*
- *Moreover $v \in (\partial \mathcal{E})(f)$ if and only if*

$$v(x) = \int d\rho_y(x) u(x, y), \tag{3}$$

  for almost all $x \in \mathcal{X}$, where $u(x, y) \in (\partial V)(y, f(x))$, for almost all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Proof. Let $\bar{V} : \mathcal{Z} \times \mathcal{G} \to \mathbb{R}^+$ be such that $\bar{V}(z, w) = V(y, w)$, with $w \in \mathcal{G}$, then the functional $\bar{\mathcal{E}} : L^p(\mathcal{Z}, \mu) \to \mathbb{R}^+$ defined by

$$\bar{\mathcal{E}}(g) = \int d\mu(z) \bar{V}(z, g(z))$$

is the Nemitski functional associated to $\bar{V}$ and from Proposition 2 it is a well defined convex and continuous. Moreover we have $\mathcal{E} = \bar{\mathcal{E}} \circ j$ so that the expected risk is a well defined convex continuous

functional in $L^p(\mathcal{X}, \rho_\mathcal{X})$ since it is a composition of a convex continuous functional and a linear map. Then, from Proposition (1) Item 6,

$$(\partial\mathcal{E})(f) = \{v \in L^q(\mathcal{X}, \rho_\mathcal{X}) \mid v(x) = (j^* u)(z), u \in (\partial\bar{\mathcal{E}})(jf)\}.$$

and from Proposition (3) we have

$$(\partial\bar{\mathcal{E}})(g) = \{u \in L^q(\mathcal{Z}, \mu) \mid u(z) \in (\partial\bar{V})(z, g(z))\}$$

for $g \in L^p(\mathcal{Z}, \mu)$, so that (3) follows combinining the above results and using the definition of j. □

A.4 Special Cases

The above setting is general enough to recover a number of different scenarios. The standard regression setting is given by $\mathcal{Y} = \mathcal{G} = \mathbb{R}$, while more generally vector -valued and functional regression setting correspond to $\mathcal{Y} = \mathcal{G}$ where is a Euclidean and a Hilbert space respectively. More generally the case where we have general $\mathcal{Y}$ and $\mathcal{G}$ a Hilbert space can be related to structured learning where $\mathcal{G}$ can be seen as the RKHS induced by reproducing kernel $K_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Binary classification case corresponds to $\mathcal{Y} = \{\pm 1\}$ and $\mathcal{G} = \mathbb{R}$. More generally multi-category classification in the simplex coding framework corresponds to $\mathcal{Y} = \{1, \dots, T\}$, $T \geq 2$, $\mathcal{G} = \mathbb{R}^{T-1}$ and in particular we have

$$v \in (\partial\mathcal{E})(f) \iff v(x) = \sum_{y \in \mathcal{Y}} \rho_y(x) u(x, y), \tag{4}$$

for almost all $x \in \mathcal{X}$, where $u(x, y) \in (\partial V)(y, f(x))$ for almost all $(z, y) \in \mathcal{Z}$.

B Relaxation Error Analysis

In this section we study the properties of the loss function we introduced in the paper and quantify their relaxation error in terms of Fisher consistency and comparison inequalities. We let $L^p(\mathcal{X}, \rho_\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R}^{T-1} \mid \|f\|_{\rho,p} = (\int \|f(x)\|^p d\rho_\mathcal{X}(x))^{\frac{1}{p}} < \infty\}$, with $p \in \mathbb{N}$ and $\|f\|_\rho = \|f\|_{\rho,2}$ for $f \in L^2(\mathcal{X}, \rho_\mathcal{X})$. Given a function $f \in L^p(\mathcal{X}, d\rho_\mathcal{X})$, with some abuse of notation, we will denote by $D(f)$ the function with values $D(f(x))$, for almost all $x \in \mathcal{X}$. In this section we use the tools introduced in (A).

B.1 Simplex Square Loss

Theorem 2. *The expected risk of the the simplex square loss is a convex, continuous functional $\mathcal{E} : L^2(\mathcal{X}, \rho_\mathcal{X}) \to \mathbb{R}^+$.*

1. *The minimizer of the expected risk on $L^2(\mathcal{X}, \rho_\mathcal{X})$ is the regression function $f_\rho(x) = \mathbb{E}[c_Y | X = x]$ and the square loss is Fisher consistent.*
2. *Moreover for any $f \in L^2(\mathcal{X}, \rho_\mathcal{X})$ we have the following comparison inequality,*

$$R(D(f)) - R(D(f_\rho)) \leq \sqrt{\frac{2(T-1)}{T}(\mathcal{E}(f) - \mathcal{E}(f_\rho))}. \tag{5}$$

Proof of theorem 2. Let $f_\rho(x) = \mathbb{E}[c_Y | X = x]$ for almost all $x \in \mathcal{X}$, then by definition of the simplex coding

$$\|f_\rho(x)\|^2 \leq \mathbb{E}[\|c_Y\|^2 | X = x] = \sum_{y \in \mathcal{Y}} \|c_y\|^2 \rho_y(x) = 1,$$

so that f_ρ is almost surely bounded and belongs to $L^2(\mathcal{X}, \rho_\mathcal{X})$. Moreover,

$$\mathcal{E}(f_\rho) = \mathbb{E}(\|Y - f_\rho(X)\|^2) \leq \mathbb{E}[\|c_Y\|^2] - \|f_\rho\|_\rho^2 \leq 1,$$

and, for $f \in L^2(\mathcal{X}, \rho_\mathcal{X})$,

$$\mathcal{E}(f) = \mathbb{E}[\|Y - f(X)\|^2] = \mathbb{E}[\|(Y - f_\rho) + (f_\rho(X) - f(X))\|^2] = \|f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho), \tag{6}$$

since

$$-2\,\mathbb{E}[\langle Y - f_\rho, f_\rho - f\rangle_\rho] = \int d\rho_{\mathcal{X}}(x) \left\langle \sum_{y\in\mathcal{Y}} c_y \rho_y(x) - f_\rho(x), f_\rho(x) - f(x)) \right\rangle = 0.$$

Then the expected risk induced by the square loss is a Nemitski functional on $L^2(\mathcal{X}, \rho_{\mathcal{X}})$, since it is convex and $||c_y - f(x)||^2 \leq 2(1 + ||f(x)||^2)$, for all $f \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$.

Then, using (4) we have that, since the square loss is differentiable, in this case there is a unique $u(x, y) \in (\partial V)(y, f(x))$ given by $u(x, y) = \nabla \|y - f(x)\| = 2(c_y - f(x))$ and setting the (sub) gradient to be almost surely zero we have

$$0 = \sum_{y\in\mathcal{Y}} \rho_y(x) 2(c_y - f(x)) \Longrightarrow f(x) = \sum_{y\in\mathcal{Y}} \rho_y(x) c_y.$$

almost surely, where $u(x, y) \in (\partial V)(y, f(x))$.

Moreover, since $f_\rho(x) = \sum_{y\in\mathcal{Y}} c_y \rho_y(x)$ then

$$\langle f_\rho(x), c_y\rangle = f_\rho(x) = \sum_{y'\in\mathcal{Y}} \rho_y(x) \langle c_{y'}, c_y\rangle = \rho_y(x) - \frac{1}{T-1}(1 - \rho_y(x)) = \frac{T\rho_y(x) - 1}{T-1}$$

so that

$$\rho_y(x) = \frac{T-1}{T} \langle f_\rho(x), c_y\rangle + \frac{1}{T}, \forall y \in \mathcal{Y}. \tag{7}$$

Fisher consistency easily follows from the definition of the decoding map D.

We next derive the comparison inequality (5). We begin noting that by definition,

$$R(D(f)) = \int_{\mathcal{X}} d\rho_{\mathcal{X}}(x) \sum_{y\in\mathcal{Y}} \mathbb{1}_{y\neq D(f(x)))}(x, y) = \int_{\mathcal{X}} d\rho_{\mathcal{X}}(x) \sum_{y\neq D(f(x))} \rho_{y(x)} = \int_{\mathcal{X}} d\rho_{\mathcal{X}}(x)(1-\rho_{D(f(x))}(x)))$$

so that

$$R(D(f))-R(D(f_\rho)) = \int \left(\rho_{D(f_\rho(x))}(x) - \rho_{D(f(x))}(x)\right) d\rho_{\mathcal{X}}(x) = \int_{\mathcal{X}_f} \left(\rho_{D(f_\rho(x))}(x) - \rho_{D(f(x))}(x)\right) d\rho_{\mathcal{X}}(x) \tag{8}$$

where $\mathcal{X}_f = \{x \in \mathcal{X} \mid D(f(x)) \neq D(f_\rho(x))\}$. Moreover using equation (7) we can write,

$$\begin{aligned} R(D(f)) - R(D(f_\rho)) &= \frac{T-1}{T} \int_{\mathcal{X}_f} \left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x)\right\rangle d\rho_{\mathcal{X}}(x) \\ &= \frac{T-1}{T} \int_{\mathcal{X}_f} (\left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x) - f(x)\right\rangle d\rho_{\mathcal{X}}(x) \\ &+ \frac{T-1}{T} \int_{\mathcal{X}_f} \left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f(x)\right\rangle) d\rho_{\mathcal{X}}(x). \end{aligned} \tag{9}$$

The last term in the above expression can be shown to be negative since

$$\left\langle c_{D(f_\rho(x))}, f(x)\right\rangle \leq \left\langle c_{D(f(x))}, f(x)\right\rangle = \max_{y\in\mathcal{Y}} \langle c_y, f(x)\rangle, \quad \forall x \in \mathcal{X}_f, \tag{10}$$

by definition of D. Then using Jensen and Cauchy-Schwarz and inequalities we have,

$$\begin{aligned} (R(D(f)) - R(D(f_\rho)))^2 &\leq \frac{(T-1)^2}{T^2} \int_{\mathcal{X}_f} \left(\left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x) - f(x)\right\rangle\right)^2 d\rho_{\mathcal{X}}(x) \\ &\leq \frac{(T-1)^2}{T^2} \int_{\mathcal{X}_f} \left\|c_{D(f_\rho(x))} - c_{D(f(x))}\right\|^2 \|f(x) - f_\rho(x)\|^2 d\rho_{\mathcal{X}}(x) \\ &\leq \frac{(T-1)^2}{T^2} \frac{2T}{T-1} \int_{\mathcal{X}_f} \|f(x) - f_\rho(x)\|^2 d\rho_{\mathcal{X}}(x) \\ &\leq \frac{2(T-1)}{T} \int \|f(x) - f_\rho(x)\|^2 d\rho_{\mathcal{X}}(x), \end{aligned} \tag{11}$$

where we used the fact that $||c_y - c'_y||^2 = \frac{2T}{T-1}$, for $y \neq y'$. The result follows plugging (6) in the above expression in (11) and taking square roots.

□

B.1.1 Improved rates under noise condition

The above theorem can be improved for 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 ρ satisfies the classification noise condition with parameter B_q, if*

$$\rho_{\mathcal{X}}\left(\left\{x \in \mathcal{X} \mid 0 \leq \min_{j \neq D(f_\rho(x))} \frac{T-1}{T}(\langle c_{D(f_\rho(x))} - c_j, f_\rho(x)\rangle) \leq s\right\}\right) \leq B_q s^q, \tag{12}$$

where $s \in [0,1]$.

If a distribution ρ is characterized by a very large q, then, for each $x \in \mathcal{X}$, $f_\rho(x)$ is arbitrarily close to one of the coding vectors.
For $T = 2$, the above condition reduces to the binary Tsybakov noise. Indeed, let $c_1 = 1$, and $c_2 = -1$, if $f_\rho(x) > 0$, $\frac{1}{2}(c_1 - c_2)f_\rho(x) = f_\rho(x)$, and if $f_\rho(x) < 0$, $\frac{1}{2}(c_2 - c_1)f_\rho(x) = -f_\rho(x)$.

We have the following result:.

Theorem 3. *For each $f \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$, if (12) holds, then we have the following inequality,*

$$R(D(f)) - R(D(f_\rho)) \leq K\left(\frac{2(T-1)}{T}(\mathcal{E}(f) - \mathcal{E}(f_\rho))\right)^{\frac{q+1}{q+2}}, \tag{13}$$

for a constant $K = \left(2\sqrt{B_q + 1}\right)^{\frac{2q+2}{q+2}}$.

We start proving the following lemma:

Lemma 1. *The generalised Tsybakov condition is equivalent to that for all $f \in L^2_{\rho_{\mathcal{X}}}$:*

$$\rho(\mathcal{X}_f) \leq C_\alpha \left(R(D(f)) - R(D(f_\rho))\right)^\alpha \tag{14}$$

where

$$\alpha = \frac{q}{q+1} < 1 \text{ and } C_\alpha = B_q + 1 > 1$$

Proof of lemma 1. Let $m_\rho(x) = \frac{T-1}{T}\left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x)\right\rangle$

$$\begin{aligned}
R(D(f)) - R(D(f_\rho)) &= \int_{\mathcal{X}_f} m_\rho(x) d\rho_{\mathcal{X}}(x) \geq \int_{\mathcal{X}_f} m_\rho(x) \mathbb{1}_{m_\rho(x) \geq t} d\rho_{\mathcal{X}}(x) \\
&\geq t\left(\int_{\mathcal{X}} \mathbb{1}_{m_\rho(x) \geq t} d\rho_{\mathcal{X}}(x) - \int_{\mathcal{X}/\mathcal{X}_f} \mathbb{1}_{m_\rho(x) \geq t} d\rho_{\mathcal{X}}(x)\right) \\
&\geq t\left(1 - \mathbb{P}\{x \in \mathcal{X}, m_\rho(x) \leq t\} - \mathbb{P}\{x \in \mathcal{X}/\mathcal{X}_f\}\right) \\
&\geq t(1 - B_q t^q - \rho_{\mathcal{X}}(\mathcal{X}/\mathcal{X}_f)) \\
&= t(\rho_{\mathcal{X}}(\mathcal{X}_f) - B_q t^q)
\end{aligned}$$

Now taking the minimum of the above bound with respect to t, we get $t^* = \left(\frac{1}{B_q+1}\rho_{\mathcal{X}}(\mathcal{X}_f)\right)^{\frac{1}{q}}$.
Finally plugging t^* in the bound we get,

$$\rho_{\mathcal{X}}(\mathcal{X}_f) \leq C_\alpha \left(R(D(f)) - R(D(f_\rho))\right)^\alpha \tag{15}$$

where $\alpha = \frac{q}{q+1} < 1$ and $c_\alpha = B_q + 1 > 1$. □

Proof of theorem 3. Let $0 < t < 1$ if $t \leq m_\rho(x)$ we have $tm_\rho(x) \leq m_\rho^2(x)$ and therefore $m_\rho(x) \leq \frac{m_\rho^2(x)}{t}$.

$$\begin{aligned} R(D(f)) - R(D(f_\rho)) &= \int_{\mathcal{X}_f} m_\rho(x) d\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}_f} m_\rho(x) \mathbb{1}_{m_\rho(x) \leq t} d\rho_{\mathcal{X}}(x) + \int_{\mathcal{X}_f} m_\rho(x) \mathbb{1}_{m_\rho(x) > t} d\rho_{\mathcal{X}}(x) \\ &\leq t\rho_{\mathcal{X}}(\mathcal{X}_f) + \frac{1}{t} \int_{\mathcal{X}_f} m_\rho(x)^2 d\rho_{\mathcal{X}}(x) \\ &\leq tC_\alpha \left((R(D(f)) - R(D(f_\rho)) \right)^\alpha + \frac{1}{t} \frac{2(T-1)}{T} (\mathcal{E}(f) - \mathcal{E}(f_\rho)) \end{aligned}$$

In the last inequality we used lemma 1 and the fact:.

$$m_\rho(x) = \frac{T-1}{T} \left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x) \right\rangle \leq \frac{T-1}{T} (\left\langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x) \right\rangle - f(x)).$$

Squaring both sides of the inequality and using Cauchy Schwartz:

$$m_\rho^2(x) \leq (\frac{T-1}{T})^2 2 \frac{T}{T-1} ||f(x) - f_\rho(x)||^2 = 2 \frac{T-1}{T} (\mathcal{E}(f) - \mathcal{E}(f_\rho))$$

Minimizing the right hand side of the above inequality over t,we get the result (13). □

B.2 SVM Loss functions

Next we consider extensions of the SVM's hinge loss to a multiclass setting. We let

$$\text{co}(\mathcal{C}) = \{u \in \mathbb{R}^{T-1} \mid u = \sum_{y \in \mathcal{Y}} \lambda_y c_y, \sum_{y \in \mathcal{Y}} \lambda_y = 1, 0 \leq \lambda_y \leq 1, c_y \in \mathcal{C}, \forall y \in \mathcal{Y}\}.$$

Moreover we define the inner risk

$$\varepsilon : \mathbb{R}^{T-1} \times \mathbb{K}^{T-1} \to \mathbb{R}^+, \quad \varepsilon(w, p) = \sum_{y \in \mathcal{Y}} p_y V(y, w)$$

for $w \in \mathbb{R}^{T-1}$ and $p = (p_y)_{y \in \mathcal{Y}} \in \mathbb{K}^{T-1}$, so that $\mathcal{E}(f) = \int d\rho_{\mathcal{X}}(x) \varepsilon(f(x), p_x)$, where $p_x = (\rho_y(x))_{y \in \mathcal{Y}}$.

B.2.1 Simplex Cone SVM

Theorem 4. *The expected risk of the the simplex cone hinge loss is a convex, continuous functional $\mathcal{E} : L^1(\mathcal{X}, \rho_{\mathcal{X}}) \to \mathbb{R}^+$.*

1. *The minimizer of the expected risk on $L^1(\mathcal{X}, \rho_{\mathcal{X}})$ is the function $f_\rho(x) = c_{k(x)}$ where $k(x) = \arg\max_{y' \in \mathcal{Y}} \rho_{y'}(x)$ for almost all $x \in \mathcal{X}$, and the simplex cone hinge loss is Fisher consistent.*
2. *Moreover for any $f \in L^1(\mathcal{X}, \rho_{\mathcal{X}})$ we have the following comparison inequality,*

$$R(D(f)) - R(D(f_\rho)) \leq (T-1)(\mathcal{E}(f) - \mathcal{E}(f_\rho)). \tag{16}$$

Proof. We first show that V is a Nemitski loss function with $p = 1$. V is convex since it is the sum of convex loss functions $\phi(y, w) = \left| \frac{1}{T-1} + \langle c_y, w \rangle \right|_+$ and satisfies,

$$V(y, w) = \sum_{j \neq y} \left| \frac{1}{T-1} + \langle c_j, w \rangle \right|_+ \leq \sum_{j \neq y} |\frac{1}{T-1} + \langle c_j, w \rangle| \leq \sum_{j \neq y} \frac{1}{T-1} + |\langle c_j, w \rangle| \leq 1 + (T-1)||w||,$$

where we used Cauchy Schwartz inequality and the properties of the simplex coding. Then theorem 1 in the appendix ensures that $\mathcal{E} : L^1(\mathcal{X}, \rho_{\mathcal{X}}) \to \mathbb{R}^+$ is a well defined, convex and continuous

functional.
Define $f_\rho(x) = c_{k(x)}$, where $k(x) = \arg\max_{y\in Y} \rho_y(x)$ for almost all $x \in \mathcal{X}$.
We claim that f_ρ is a minimizer of $\mathcal{E}(f)$. Since $\mathcal{E}(f)$ is continuous and convex it is sufficient to show that $0 \in \partial\mathcal{E}(f_\rho)$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, j : L^1(\mathcal{X}, \rho_\mathcal{X}) \to L^1(\mathcal{Z}, \rho_\mathcal{Z})$, $j^* : L^\infty(\mathcal{Z}, \rho_\mathcal{Z}) \to L^\infty(\mathcal{X}, \rho_\mathcal{X})$, and $u(x, y) \in \partial V(y, f(x))$, $u \in L^\infty(\mathcal{Z}, \rho_\mathcal{Z})$. By appendix C we have,

$$\partial\mathcal{E}(f) = \{v \in L^\infty(\mathcal{X}, \rho_\mathcal{X}) \mid v(x) = j^*u(x, y) = \sum_{y\in\mathcal{Y}} \rho_y(x)u(x, y)\}.$$

With some abuse of notation let $w = f(x), k = k(x), p_y = \rho_y(x), u_y = u(x, y)$, therefore we want to show that $0 \in \partial\mathcal{E}(c_k)$. We start first computing the sub-gradient $\partial V(y, w)$:

$$\partial V(y, w) = \sum_{y'\neq y} \partial \left| \frac{1}{T-1} + \langle c_{y'}, w\rangle \right|_+ . \tag{17}$$

Where we used Proposition 1.5, since $V(y, w) = \sum_{y'\neq y} \phi(y, w)$ is convex continuous and $\phi(y, 0) = \frac{1}{T-1}$. Then for all $y' \in \mathcal{Y}$,

$$\partial \left| \frac{1}{T-1} + \langle c_{y'}, w\rangle \right|_+ = \begin{cases} c_{y'} & \text{if } \langle c_{y'}, w\rangle > \frac{-1}{T-1} \\ \lambda_{y'}c_{y'}, 0 < \lambda_{y'} < 1, & \text{if } \langle w, c_{y'}\rangle = -\frac{1}{T-1} \\ 0, & \text{if } \langle w, c_{y'}\rangle < -\frac{1}{T-1}. \end{cases}$$

Let us compute $\partial V(y, c_k)$.

$$\partial V(y, c_k) = \begin{cases} \sum_{y'\neq k} \lambda_{y'}c_{y'}, 0 < \lambda_{y'} < 1 & \text{if } y = k \\ c_k + \sum_{y'\in\mathcal{Y}/\{y,k\}} \lambda_{y'}c_{y'}, 0 < \lambda_{y'} < 1 & \text{if } y \neq k. \end{cases}$$

Let $u_y \in \partial V(y, c_k)$, let $v = j^*u = \sum_{y\in\mathcal{Y}} p_y u_y$. We will use the following fact:

$$\sum_{y\in\mathcal{Y}} p_y \sum_{y'\in\mathcal{Y}/\{y,k\}} \alpha(y') = \sum_{y\in\mathcal{Y}} \alpha(y) \sum_{y'\in\mathcal{Y}/\{y,k\}} p'_y = \sum_{y\in\mathcal{Y}} (1 - p_y - p_k)\alpha(y). \tag{18}$$

Then,

$$\begin{aligned} v = \sum_{y\in\mathcal{Y}} p_y\partial V(y, c_k) &= p_k\partial V(k, c_k) + \sum_{y\neq k} p_y\partial V(y, c_k) \\ &= p_k \sum_{y\neq k} \lambda_y c_y + \sum_{y\neq k} p_y(c_k + \sum_{y'\in\mathcal{Y}/\{y,k\}} \lambda_{y'}c_{y'}) \\ &= p_k \sum_{y\neq k} \lambda_y c_y + c_k \sum_{y\neq k} p_y + \sum_{y\neq k} p_y \sum_{y'\in\mathcal{Y}/\{y,k\}} \lambda_{y'}c_{y'} \\ &= p_k \sum_{y\neq k} \lambda_y c_y + (1 - p_k)c_k + \sum_{y\neq k} \lambda_y(\sum_{y'\in\mathcal{Y}/\{y,k\}} p_{y'})c_y \\ &= (1 - p_k)c_k + p_k \sum_{y\neq k} \lambda_y c_y + \sum_{y\neq k} \lambda_y(1 - p_y - p_k)c_y \\ &= (1 - p_k)c_k + \sum_{y\neq k} (1 - p_y)\lambda_y c_y. \end{aligned}$$

Let $\lambda^*_y = \frac{1-p_k}{1-p_y}$. Since $p_k > p_y$, for all $y \in \mathcal{Y}/\{k\}$ we indeed have $0 < \lambda^*_y < 1$.
So that, setting λ_y to λ^*_y, we get:

$$\begin{aligned} v^* &= (1 - p_k)c_k + \sum_{y\neq k} (1 - p_j)\lambda^*_y c_y = (1 - p_k)a_k + (1 - p_k) \sum_{y\neq k} c_j \\ &= (1 - p_k)c_k + (1 - p_k)(-c_k) \\ &= 0. \end{aligned}$$

Then $0 \in \partial\mathcal{E}(f_\rho)$, and the hinge loss is consistent since $D(f_\rho) = b_\rho$ a .s.
Next we prove the comparison inequality (16). Note that the point wise risk can be written as:

$$\begin{aligned}\varepsilon(w,p) &= \sum_{y\in\mathcal{Y}} p_y \sum_{y'\neq y} \left|\frac{1}{T-1} + \langle c_{y'}, w\rangle\right|_+ \\ &= \sum_{y=1}^{T}(1-p_y)\left|\frac{1}{T-1} + \langle c_{y'}, w\rangle\right|_+ .\end{aligned}$$

Note that the sub gradient of $\partial\varepsilon(c_k,p)$ has the same form of v. It follows that c_k is a minimizer of the point-wise risk , and $\varepsilon(c_k,p) = \frac{T}{T-1}(1-p_k)$. Let w be a vector in $\mathbb{R}^{T-1}$, such that $D(w) = \ell$. Let $I = \{i \in \mathcal{Y} | \langle c_i, w\rangle \geq -\frac{1}{T-1}\}$. ℓ belongs to the set I, as it is the maximum of $\langle c_j, w\rangle$ and the equation $\sum_{j=1}^{T}\langle c_j, w\rangle = 0$ is always satisfied so that $\langle c_\ell, w\rangle$ is necessarily positive.

$$\begin{aligned}\varepsilon(w,p) - \varepsilon(c_k,p) &= \sum_{i=1}^{T}(1-p_i)\left|\langle c_i, w\rangle + \frac{1}{T-1}\right|_+ - (1-p_k)\sum_{i=1}^{T}(\langle c_i, w\rangle + \frac{1}{T-1}) \\ &= \sum_{i\in I}(p_k - p_i)(\langle a_i, w\rangle + \frac{1}{T-1}) - (1-p_k)\sum_{i\notin I}(\langle c_i, w\rangle + \frac{1}{T-1}) \\ &= \frac{1}{T-1}(p_k - p_\ell) + \left((p_k - p_\ell)\langle c_\ell, w\rangle + \sum_{i\in I, i\neq l}(p_k - p_i)(\langle c_i, w\rangle + \frac{1}{T-1})\right) \\ &- (1-p_k)\sum_{i\notin I}(\langle c_i, w\rangle + \frac{1}{T-1}) \\ &\geq \frac{1}{T-1}(p_k - p_\ell).\end{aligned}$$

The last inequality is due to the positivity of the other terms. Then if we let $w = f(x)$, let $f_\rho(x) = c_{k(x)}$ and $p = (\rho_y(x))_{y\in\mathcal{Y}}$ and integrate the above inequality over over x, we obtain

$$\begin{aligned}&\frac{1}{T-1}\int_{\mathcal{X}}[\rho_{D(f_\rho)(x)}(x) - \rho_{D(f)(x))(x)}]d\rho_{\mathcal{X}}(x) = \frac{1}{T-1}(R(f) - R(f_\rho)) \\ &\leq \int_{\mathcal{X}}\left(\varepsilon(f(x),(\rho(x))_{y\in\mathcal{Y}}) - \varepsilon(f_\rho(x),(\rho_y(x))_{y\in\mathcal{Y}})\right)d\rho_{\mathcal{X}}(x) = \mathcal{E}(f) - \mathcal{E}(f_\rho)\end{aligned}$$

where we used (8).

□

B.2.2 SH-SVM Loss

Theorem 5. *The expected risk of the the simplex cone hinge loss is a convex, continuous functional $\mathcal{E} : L^1 \to \mathbb{R}^+$. Let $\mathcal{F} = \{f \in L^1(\mathcal{X}, \rho_{\mathcal{X}}) \mid f(x) \in co(\mathcal{C}),$ for almost all $x \in \mathcal{X}\}$.*

1. *The minimizer of the expected risk on $\mathcal{F}$ is the function $f_\rho(x) = c_{k(x)}$ where $k(x) = \arg\max_{y'\in\mathcal{Y}} \rho_{y'}(x)$ for almost all $x \in \mathcal{X}$ and the constrained single margin hinge loss is Fisher consistent.*

2. *Moreover for any $f \in \mathcal{F}$ we have the following comparison inequality,*

$$R(D(f)) - R(D(f_\rho)) \leq (T-1)(\mathcal{E}(f) - \mathcal{E}(f_\rho)). \tag{19}$$

Proof. We use the same notations of the proof in theorem 4, that is : $w = f(x), k = k(x), p_y = \rho_y(x)$, for almost $x \in \mathcal{X}$.
Note that for $w \in \text{co}(\mathcal{C})$, using the definition of simplex coding, we have

$$\langle w, c_y\rangle = \lambda_y - \frac{1}{T-1}\sum_{y'\neq y}\lambda_{y'} = \frac{T\lambda_y - 1}{T-1}, \Longrightarrow -\frac{1}{T-1} \leq \langle c_y, w\rangle \leq 1, \tag{20}$$

for all $y \in \mathcal{Y}$. Then $|1 - \langle c_y, w\rangle|_+ = 1 - \langle c_y, w\rangle\,, \quad \forall w \in \text{co}(\mathcal{C})$. Then the inner risk can be written as

$$\begin{aligned}\varepsilon(w,p) &= \sum_{y\in\mathcal{Y}} p_y(1 - \langle c_y, w\rangle) \\ &= 1 - \left\langle w, \sum_{y\in\mathcal{Y}} p_y c_y \right\rangle.\end{aligned}$$

Minimizing the inner risk ε under the convex hull constraint is equivalent to the linear programming (LP) problem,

$$\max_{w\in Co(\mathcal{C})} F(w) = \left\langle w, \sum_{y\in\mathcal{Y}} p_y c_y \right\rangle.$$

It is a standard result that in an LP maximization over a convex polytope, the solution is achieved on a vertex.

The vertices in our case are c_y, so it is sufficient to find the vertex that maximizes the loss f:

$$\max_{w\in\{c_1 \ldots c_T\}} F(w).$$

$$F(c_j) = \left\langle c_j, \sum_{y\in\mathcal{Y}} p_y c_y \right\rangle = p_j - \frac{1}{T-1}\sum_{y\neq j} p_y = \frac{T}{T-1} p_j, \forall j \in \mathcal{Y}.$$

Let $k = \arg\max p_j$, it follows that $F(c_k) > F(c_j), \forall j \in \mathcal{Y}/\{k\}$.

Thus, $w^* = c_k$ is the minimizer of the inner risk. From inner risk to the expected risk minimization:

Then minimization of the inner risk yields a minimizer of the expected risk setting f_ρ such that $f_\rho(x) = c_T$ and the simplex hinge loss is Fisher calibrated. We next derive the comparison inequality (19). Let $w \in Co(\mathcal{C})$ such that $D(w) = \ell \neq k$, $w^* = c_k$ and $p \in \mathbb{K}^{T-1}$, then

$$\varepsilon(w,p) = \sum_{y\in\mathcal{Y}} (1 - z_y)p_y, \qquad \varepsilon(w^*,p) = \frac{T}{T-1}(1 - p_k),$$

where $z_y = \langle c_y, w\rangle$, for $y \in \mathcal{Y}$.

Note that $\sum_{y\in\mathcal{Y}} z_y = \left\langle \sum_{y\in\mathcal{Y}} c_y, w \right\rangle = 0$ by definition of simplex coding, and $\sum_{y\in\mathcal{Y}} p_y = 1$ from (21), so that

$$\begin{aligned}\varepsilon(w,p) - \varepsilon(w^*,p) &= \sum_{y\in\mathcal{Y}} (1 - z_y - \frac{T}{T-1})p_y + \sum_{y\in\mathcal{Y}} (z_y + \frac{1}{T-1})p_k = -\sum_{y\in\mathcal{Y}}^{T} (z_y + \frac{1}{T-1})p_y + \sum_{y\in\mathcal{Y}} (z_y + \frac{1}{T-1})p_k \\ &= \sum_{y\in\mathcal{Y}} (z_y + \frac{1}{T-1})(p_k - p_y).\end{aligned}$$

We showed in (21) that $-\frac{1}{T-1} \leq z_y \leq 1$, for all y and $w \in \text{co}(\mathcal{C})$. Moreover, $p_k - p_y > 0$, $\forall\, y$, since $D(w^*) = k$, and $z_\ell > 0$ since $D(w) = \ell$. Then we have,

$$\begin{aligned}\varepsilon(w,p) - \varepsilon(w^*,p) &= (\frac{1}{T-1} + z_\ell)(p_k - p_\ell) + \sum_{y\neq\ell} (z_y + \frac{1}{T-1})(p_k - p_y) \\ &\geq \frac{1}{T-1}(p_k - p_\ell).\end{aligned}$$

Then if we let $w = f(x)$, let $w^* = f_\rho(x)$ and $p = \rho_y(x))_{y\in\mathcal{Y}}$ and integrate over over x, we obtain

$$\frac{1}{T-1} R(D(f)) - R(D(f_\rho)) = \frac{1}{T-1}\int_X [\rho_{D(f_\rho)(x)} - \rho_{D(f)(x))}](x) d\rho_{\mathcal{X}}(x) \leq \mathcal{E}(f) - \mathcal{E}(f_\rho)$$

where we used (8). □

An alternative proof is given here bypassing the inner risk minimization:

Proof. We use the same notations of the proof in theorem 4, that is : $w = f(x), k = k(x), p_y = \rho_y(x), g_\rho(x) = \sum_y \rho_y(x) c_y$. , for almost $x \in \mathcal{X}$.
Note that for $w \in \text{co}(\mathcal{C})$, using the definition of simplex coding, we have

$$\langle w, c_y \rangle = \lambda_y - \frac{1}{T-1} \sum_{y' \neq y} \lambda_{y'} = \frac{T\lambda_y - 1}{T-1}, \Longrightarrow -\frac{1}{T-1} \leq \langle c_y, w \rangle \leq 1, \tag{21}$$

for all $y \in \mathcal{Y}$. Then $|1 - \langle c_y, w \rangle|_+ = 1 - \langle c_y, w \rangle$, $\quad \forall w \in \text{co}(\mathcal{C})$. Then for all $f \in \mathcal{F}$, the functional we are minimizing can be written as:

$$\begin{aligned} \mathcal{E}(f) &= \int (1 - \langle c_y, f(x) \rangle) d\rho(x, y) \\ &= 1 - \int \left\langle \sum_y \rho_y(x) c_y, f(x) \right\rangle d\rho_{\mathcal{X}}(x) \\ &= 1 - \langle g_\rho, f \rangle_{L^2(\mathcal{X}, \rho_{\mathcal{X}})}, f \in \mathcal{F}. \end{aligned} \tag{22}$$

Let:

$$F(f) = \langle g_\rho, f \rangle_{L^2(\mathcal{X}, \rho_{\mathcal{X}})}.$$

Minimizing the functional $\mathcal{E}$ for $f \in \mathcal{F}$, is equivalent to the linear programming (LP) in $L^2(\mathcal{X}, \rho_{\mathcal{X}})$:

$$\max_{f \in \mathcal{F}} F(f).$$

It is a standard result that in an LP maximization over a convex polytope, the solution is achieved on an extremal point. It is easy to see that $\mathcal{F}$ is a convex set and that its extremal set $\mathcal{V}(\mathcal{F})$ has the following form:

$$\mathcal{V}(\mathcal{F}) = \{f | f(x) = c_{\pi(x)}, \quad \pi : \mathcal{X} \to \mathcal{Y}\}$$

Let $f_\rho(x) = c_{b_\rho(x)}$, where $b_\rho(x) = \arg\max_{j \in \mathcal{Y}} \rho_j(x)$ for almost $x \in \mathcal{X}$. It is easy to see that $f_\rho \in \mathcal{V}(\mathcal{F})$, and for all $f \in \mathcal{V}(f)$:

$$\begin{aligned} F(f) &= \langle g_\rho, f \rangle_{L^2(\mathcal{X}, \rho_{\mathcal{X}})} = \int \left\langle \sum_{j \in \mathcal{Y}} \rho_j(x) c_j, c_{\pi(x)} \right\rangle d\rho_{\mathcal{X}}(x) \\ &= \int \left(\rho_{\pi(x)} - \frac{1}{T-1} \sum_{j \neq \pi(x)} \rho_j(x) \right) d\rho_{\mathcal{X}}(x) \\ &= \int (\frac{T}{T-1} \rho_{\pi(x)} - \frac{1}{T-1}) d\rho_{\mathcal{X}}(x). \end{aligned}$$

Then:

$$F(f_\rho) = \int (\frac{T}{T-1} \rho_{b_\rho(x)} - \frac{1}{T-1}) d\rho_{\mathcal{X}}(x)$$

It follows that: $F(f) \leq F(f_\rho), \forall f \in \mathcal{V}(\mathcal{F})$. Therefore f_ρ is the minimizer of the functional $\mathcal{E}$, and $D(f_\rho) = b_\rho$. The Half spaces hinge loss is therefore Fisher consistent. □

C Appendix to section 5

In this section we present the complementary material of section 5. We derive an algorithm for computing the simplex code and show its correctness. As well as the dual formulation for SC-SVM and 2 relaxation of the original SH-SVM.

C.1 Computing Simplex Coding

In this section we give the proof of lemma 1.

Algorithm 1 Simplex Code

SET: $C[2] = [1 \; -1]$,
FOR $i = 2, \cdots, T-1$
$u = (-\frac{1}{i} \cdots -\frac{1}{i})$ (column vector in $\mathbb{R}^i$)
$v = (0, \ldots, 0)$(column vector in $\mathbb{R}^{i-1}$)
$C[i+1] = \begin{pmatrix} 1 & u^\top \\ v & C[i] \times \sqrt{1-\frac{1}{i^2}}, \end{pmatrix}$
ENDFOR
OUTPUT:$C[T]$

Lemma 2. *The T columns of $C[T]$ are a set of $T-1$ dimensional vectors satisfying the properties of Definition 1.*

Proof. The statement is proved by induction. The base case is trivially true. Let $c_1 \ldots c_i$ be the columns of $C[i]$ and $b_1 \ldots b_{i+1}$ be the columns of $C[i+1]$. By construction $b_1 = (1, 0, \ldots, 0)$ and $b_m = (-\frac{1}{i}, \sqrt{1-\frac{1}{i^2}} c_m)$ for all $m = 2, \ldots, i+1$.
Assume $C[i]$ to satisfy definition 1, that is $\|c_m\|^2 = 1$, for $1 \leq m \leq i$ and $\langle c_m, c_n \rangle = -\frac{1}{i-1} \forall m \neq n$. Indeed, a direct calculation shows that $\|b_1\| = 1$ and $\langle b_1, b_m \rangle = -\frac{1}{i}, \forall m \neq 1$. Moreover, for $m \neq n$ such that $2 \leq m, n \leq i+1$ we have:

$$\|b_m\|^2 = \frac{1}{i^2} + (1 - \frac{1}{i^2}) \|c_m\|^2 = 1,$$

and

$$\langle b_m, b_n \rangle = \frac{1}{i^2} + (1 - \frac{1}{i^2}) \langle c_m, c_n \rangle = \frac{1}{i^2} - (1 - \frac{1}{i^2}) \frac{1}{i-1} = \frac{1}{i^2} - \frac{i+1}{i} = -\frac{1}{i}.$$

□

C.2 Support Vector Machine

C.3 SC-SVM

We sketch the derivation of problem 4. Following the notation for binary SVM we write (**??**) as

$$\min_{f \in \mathcal{H}} \left\{ C_0 \sum_{i=1}^{n} \sum_{y \neq y_i} \left| \langle a_y, f(x_i) \rangle + \frac{1}{T-1} \right|_+ + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right\}$$

where $C_0 = \frac{1}{2n\lambda}$. Using the representer theorem, and introducing the slack variables $\xi_i = (\xi_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^T$, for $i = 1, \ldots, n$, we can write the above problem as,

$$\min_{c_1, \ldots, a_n; \xi_1, \ldots, \xi_n \in \mathbb{R}^T} \left\{ C_0 \sum_{i=1}^{n} \sum_{y \neq y_i} \xi_i^y + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} \langle a_i, a_j \rangle \right\}$$

$$\frac{1}{T-1} + \sum_{j=1}^{n} K_{ij} \langle a_j, c_y \rangle \leq \xi_i^y, \;\; \forall \, i = 1 \ldots n, \text{ and } y \neq y_i,$$

$$\xi_i^y \geq 0, \;\; \forall \, i = 1 \ldots n, \text{ and } y \neq y_i.$$

Let $\Xi = \{\xi_i = (\xi_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^T, \; i = 1, \ldots, n\}$, $\alpha = \{\alpha_i = (\alpha_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^T, i = 1, \ldots, n\}$, $\nu = \{\nu_i = (\nu_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^T \; i = 1, \ldots, n\}$ and consider the Lagrangian corresponding to the above

problem given by,

$$\begin{aligned}\mathcal{L}(A,\Xi,\alpha,\nu) &= C_0\sum_{i=1}^{n}\sum_{y\neq y_i}\xi_i^y + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{ij}\left\langle a_i,a_j\right\rangle + \sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_i^y\left(\frac{1}{T-1}+\sum_{j=1}^{n}K_{ij}\left\langle a_j,c_y\right\rangle - \xi_i^y\right)\\ &- \sum_{i=1}^{n}\sum_{y\neq y_i}\nu_i^y\\ &= \frac{1}{T-1}\sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_i^y + \sum_{i=1}^{n}\sum_{y\neq y_i}(C_0-\alpha_i^y-\nu_i^y)\xi_i^y + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{ij}\left\langle a_i,a_j\right\rangle\\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n}K_{ij}\left\langle\sum_{y\neq y_i}\alpha_i^y c_y, a_j\right\rangle\end{aligned}$$

Considering the first order condition of optimality, a direct computation gives:

$$a_i = -\sum_{y\neq y_i}\alpha_i^y c_y, \quad \text{with } 0\leq\alpha_i^y\leq C_0,\ \forall\, i=1\dots n, \text{ and } y\neq y_i. \tag{23}$$

Let $\alpha_i = (\alpha_i^1\dots\alpha_i^{y_i}\dots\alpha_i^T)$ a T dimensional vector, with $\alpha_i^{y_i}=0$, and $\alpha=(\alpha_1,\dots,\alpha_n)$ a $n\times T$ dimensional vector. In the following we let $\alpha_i^{y_i}$ free in the objective and add a constraint such that $\alpha_i^{y_i}=0, \quad \forall i$.

Recalling the definition of the matrices G and K, the dual problem corresponding to (23) is given by, the maximization of

$$\begin{aligned}L_D(\alpha) &= -\frac{1}{2}\sum_{i,j}K_{ij}\sum_{y\neq y_i,y'\neq y_j}\alpha_i^y G_{yy'}\alpha_j^{y'} + \frac{1}{T-1}\sum_{i=1}^{n}\sum_{y\neq y_i}\alpha_i^y\\ &= -\frac{1}{2}\sum_{y,y',i,j}\alpha_i^y K_{ij}G_{yy'}\alpha_j^{y'} + \frac{1}{T-1}\sum_{i=1}^{n}\sum_{y}\alpha_i^y\end{aligned}$$

subject to the constraints (23). Let $H_{ijyy'}=K_{ij}G_{yy'}$, H is a $(n\times T)\times(n\times T)$ matrix. So that we can write:

$$L_D(\alpha) = -\frac{1}{2}\alpha^\top H\alpha + \frac{1}{T-1}1_{n\times T}^\top\alpha \tag{24}$$

$$\max_\alpha -\frac{1}{2}\alpha^\top H\alpha + \frac{1}{T-1}1_{n\times T}^\top\alpha$$

$$0\leq\alpha_i^y\leq C_0(1-\delta_{y,y_i}) \quad \forall y\in\mathcal{Y},\forall i=1\dots n$$

C.4 SH-SVM

The convex hull assumption needed for the consistency of SH-SVM, introduces $2Tn$ constraints, and leads to a heavy optimization problem. Instead of considering $-\frac{1}{T-1}\leq\langle c_y,f(x_i)\rangle\leq 1,\forall\, y,\forall i$, it is reasonable to consider $\langle c_y,f(x_i)\rangle\geq-\frac{1}{T-1}\forall i=1\dots n,\forall y\neq y_i$, since the SH-SVM loss ensures a soft version of the left hand-side of the original inequality. Consider the SH-SVM loss with the reduced constrained.

$$\min_{f\in\mathcal{H}} C_0\sum_{i=1}^{n}\max(1-\langle c_{y_i},f(x_i)\rangle,0) + \frac{1}{2}||f||_{\mathcal{H}}^2$$

$$\langle c_y,f(x_i)\rangle\geq-\frac{1}{T-1}\forall i=1\dots n,\forall y\neq y_i.$$

By the representer theorem we have: $f(x) = \sum_{j=1}^{n} K(x, x_j)a_j, a_j \in \mathbb{R}^{T-1}$, and $||f||_{\mathcal{H}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j)\langle a_i, a_j\rangle$. We introduce slack variable as in the SVM litterature:

$$\min_{C_i,\Xi} C_0 \sum_{i=1}^{n} \xi_i + \frac{1}{2}\sum_{i,j} K_{ij}\langle a_i, a_j\rangle$$
$$1 - \sum_{j=1}^{n} K_{ij}\langle a_j, c_{y_i}\rangle \leq \xi_i, \forall i = 1 \dots n$$
$$-\frac{1}{T-1} - \sum_{j=1}^{n} K_{ij}\langle a_j, c_y\rangle \leq 0, \forall y \neq y_i$$
$$\xi_i \geq 0, \forall i = 1 \dots n.$$

Introducing the lagrangian, we have:

$$\begin{aligned}
L(\alpha, \xi, C) &= C_0 \sum_{i=1}^{n} \xi_i + \frac{1}{2}\sum_{ij} K_{ij}\langle a_i, a_j\rangle + \sum_{i=1}^{n} \alpha_i^{y_i}(1 - \sum_{j=1}^{n} K_{ij}\langle a_j, c_{y_i}\rangle - \xi_i) \\
&+ \sum_{i=1}^{n}\sum_{y\neq y_i} \alpha_i^y(\frac{-1}{T-1} - \sum_{j=1}^{n} K_{ij}\langle a_j, c_y\rangle) - \sum_{i=1}^{n} \nu_i\xi_i \\
&= \sum_{i=1}^{n}(C_0 - \alpha_i^{y_i} - \nu_i)\xi_i + \frac{1}{2}\sum_{ij} K_{ij}\langle a_i, a_j\rangle + \sum_{i=1}^{n}(\alpha_i^{y_i} - \frac{1}{T-1}\sum_{y\neq y_i}\alpha_i^y) \\
&- \sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}\left\langle a_j, \alpha_i^{y_i}c_{y_i} + \sum_{y\neq y_i}\alpha_i^y c_y\right\rangle.
\end{aligned}$$

Setting the optimality conditions we get:

$$0 \leq \alpha_i^{y_i} \leq C_0, \forall i, \quad \alpha_i^y \geq 0, y \neq y_i, \forall i, \quad a_i = \sum_y \alpha_i^y c_y$$
$$L_D(\alpha) = \sum_{i=1}^{n}(\alpha_i^{y_i} - \frac{1}{T-1}\sum_{y\neq y_i}\alpha_i^y) - \frac{1}{2}\sum_{ij} K_{ij}\left\langle \sum_y \alpha_i^y c_y, \sum_{y'} \alpha_j^y c_{y'}\right\rangle$$

Setting $G_{yy'} = \langle a_y, a_{y'}\rangle$, we can write the equivalent dual:

$$\max_{\alpha} -\frac{1}{2}\sum_{i,j,y,y'} \alpha_i^y K_{ij} G_{yy'} \alpha_j^{y'} + \sum_{i=1}^{n}(\alpha_i^{y_i} - \frac{1}{T-1}\sum_{y\neq y_i}\alpha_i^y)$$
$$0 \leq \alpha_i^{y_i} \leq C_0, \forall i = 1 \dots n$$
$$\alpha_i^y \geq 0, \forall y \neq y_i, \forall i = 1 \dots n$$

Therefore:

$$f(x) = \sum_{i=1}^{n} K(x, x_i)(\sum_{y=1}^{T} \alpha_i^y c_y).$$

If we relax the convex hull constraint, it is easy to see that the equivalent dual is therefore:

$$\max_{\alpha} \sum_{i,j} \alpha_i K_{ij} G_{y_i y_j} \alpha_j - \sum_{i=1}^{n} \alpha_i$$
$$0 \leq \alpha_i \leq C_0$$
$$f(x) = \sum_{i=1}^{n} K(x, x_i)\alpha_i c_{y_i}.$$

The latter formulation could be trained at the same complexity of the binary SVM but lacks fisher consistency.

C.5 Algorithms

Algorithm 2 Multi-Class Pegasos

INPUT: S, λ, L
INITIALIZE: $W_0 = 0$
FOR $i = 1 \cdots L$
$\eta_i = \frac{1}{\lambda i}$
$W_{\text{tmp}} = (1 - \eta_i \lambda) W_i - \eta_i \partial(V(y_i, f_{W_i}(x_i)))$
$W_i = \min(1, \frac{1}{\sqrt{\lambda}||W_{tmp}||_F}) W_{\text{tmp}}$
OUTPUT: W_L

C.6 Datasets

	ntrain	p	T	ntest
Landsat	4435	36	6	2000
Optdigit	3823	64	10	1797
Pendigit	7494	16	121	3498
Letter	10000	16	26	10000
Isolet	6238	617	26	1559
Pubfig83	7470	25600	83	830
Ctech101	3060	8192	102	6084

D Mathematical tools

We collect in this appendix basic tools from convex analysis and Vector Reproducing Kernel Hilbert Spaces.

D.1 Elements of Convex Analysis and Variational Calculus

In this appendix we report an appendix from [12] collecting some basic results from [5] – see also [6].

Let $\mathcal{H}$ be a Banach space and $\mathcal{H}^*$ its dual. A function $F : \mathcal{H} \to \Re$ is *convex* if

$$F(tv + (1-t)w) \leq tF(v) + (1-t)F(w),$$

for all $v, w \in \mathcal{H}$ and $t \in [0,1]$ (if the strict inequality holds for $t \in (0,1)$, F is called *strictly convex*).

Let $v_0 \in \mathcal{H}$ such that $F(v_0) < +\infty$. The *subgradient* of F at point $v_0 \in \mathcal{H}$ is the subset of $\mathcal{H}^*$ given by

$$\partial F(v_0) = \{w \in \mathcal{H}^* \,|\, F(v) \geq F(v_0) + \langle w, v - v_0 \rangle \,,\ \forall v \in \mathcal{H}\}. \tag{25}$$

where $\langle \cdot, \cdot \rangle$ is the pairing between $\mathcal{H}^*$ and $\mathcal{H}$. If $F(v) = +\infty$, we let $\partial F(v_0) = \emptyset$.

In the following proposition we summarize the main properties of the subgradient we need.

Proposition 1. *The following facts hold:*

1. *If F is differentiable at v_0, the subgradient reduces to the usual gradient $F'(v_0)$.*
2. *If F is defined on $\mathbb{R}$ and $F(v_0) < +\infty$, then F admits left and right derivative and*
$$\partial F(v_0) = [F'_-(v_0), F'_+(v_0)].$$
3. *Assume that $F \neq +\infty$. A point v_0 is a minimizer of F if and only if $0 \in \partial F(v_0)$.*
4. *If F is continuous and*
$$\lim_{\|v\|_{\mathcal{H}} \to +\infty} F(v) = +\infty.$$
then F has a minimizer. If F is strictly convex, the minimizer is unique.

5. *Let G be another convex function on $\mathcal{H}$. Assume that there is $v_0 \in \mathcal{H}$ such that F and G are continuous and finite at v_0. Let $a, b \geq 0$, then $aF + bG$ is convex and, for all $v \in \mathcal{H}$,*

$$\partial(aF + bG)(v) = a(\partial F)(v) + b(\partial G)(v).$$

6. *Let $\mathcal{H}'$ be another Banach space and $\mathcal{J}$ be a continuous linear operator from $\mathcal{H}'$ into $\mathcal{H}$. Assume that there is $v'_0 \in \mathcal{H}'$ such that F is continuous and finite at $\mathcal{J}v'_0$. For all $v' \in \mathcal{H}'$*

$$(\partial F \circ \mathcal{J})(v') = \mathcal{J}^*(\partial F)(\mathcal{J}v'),$$

where $\mathcal{J}^* : \mathcal{H}^* \to \mathcal{H}'^*$ is the adjoint of $\mathcal{J}$ defined by

$$\langle v', \mathcal{J}^* v \rangle_{\mathcal{H}'} = \langle \mathcal{J}v', v \rangle_{\mathcal{H}} .$$

for all $v \in \mathcal{H}$ and $v' \in \mathcal{H}'$.

Proof. We simply give the references to the results in [5].

1. Prop. III.2.8
2. Prop. III.2.7
3. It is a simple consequence of Prop. III.3.1
4. It is a simple consequence of Prop. II.4.6.
5. Prop. III.2.13
6. Prop. III.2.12

□

Certain integral functionals naturally arise in the context of learning theory [12]. In particular, we recall the definition of *Nemitski* functional, adapted to our framework [5, p.41 and p.143]. Let $\mathcal{Z}$ be a locally compact second countable space, μ be a finite measure on $\mathcal{Z}$, and $(\mathcal{G}, \langle \cdot, \cdot \rangle)$ a separable Hilbert space.

Let $W : \mathcal{Z} \times \mathcal{G} \to \mathbb{R}^+$ be a measurable function on $\mathcal{Z} \times \mathcal{G}$.

The Nemitski functional associated to W is

$$I(g) = \int W(z, g(z)) d\mu(z)$$

for any measurable function $g : \mathcal{X} \to \mathcal{G}$. The following proposition collects the main properties of the Nemitski functional.

Proposition 2. *The following properties hold,*

- *If $W(z, \cdot)$ is lower semi-continuous (l.s.c.) for all $z \in \mathcal{Z}$ then $I : L^p(\mathcal{Z}, \mu) \to \mathbb{R}^+$ is well defined and l.s.c.*
- *If $W(z, \cdot)$ is continuous for all $z \in \mathcal{Z}$ and moreover*

$$W(z, w) \leq a(z) + b \|w\|^p , \qquad \text{for almost all } (x, y) \in \mathcal{Z},$$

where $b \in \mathbb{R}^+$ and $\int |a(z)| d\mu(z) < \infty$ and $p \geq 1$, then I is upper semicontinuous and hence continuous in $L^p(\mathcal{Z}, \mu)$.

- *If $W(z, \cdot)$ is convex for all $z \in \mathcal{Z}$, then I is convex.*

Proof. The proof of Item 1 is given in Proposition II.2.3 in [5] for the case $\mathcal{X} = \mathbb{R}$ and $\mathcal{G} = \mathbb{R}^n$.Similarly, the proof of Item 2 is given in Proposition III.5.1. Finally item 3 is proved in Theorem II.5.1. □

Next proposition provides us with a straightforward method to study the subgradient (∂I). Let $q \in]1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

Proposition 3. *Assume that there is an element* $u_0 \in L^p(\mathcal{Z}, \mu)$ *such that* $\sup_{z \in \mathcal{Z}} |u_0(z)| < +\infty$ *and* $I[u_0] < +\infty$. *Given* $u \in L^p(\mathcal{Z}, \mu)$

$$(\partial I)(u) = \{w \in L^q(\mathcal{Z}, \mu) \mid \ w(z) \in (\partial W)(z, u(z)) \text{ for almost all } \ (x, y) \in \mathcal{Z}\}\,. \tag{26}$$

Proof. See the proof of Prop. III.5.3 of [5]. The proof is for $\mathcal{Z}$ interval of $\mathbb{R}$, but can be easily extended to arbitrary $\mathcal{Z}$, compare with [6]. □

D.2 Reproducing Kernel Hilbert Spaces of Vector Valued Function

The framework of vector valued reproducing kernel Hilbert spaces (RKHSs) provides a natural choice for hypotheses spaces and regularizers in the multi-class setting. The definition of RKHS for vector valued functions parallels the one in the scalar case [1], with the main difference that the reproducing kernel is now *matrix* valued – see [8, 4] and references therein.

Let $\mathcal{X}$ be a set, a matrix valued 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*, and $\sum_{i,j=1}^{N} \langle a_j, \Gamma(x_i, x_j) a_j \rangle \geq 0$, for all $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_N \in \mathbb{R}^D$.

A vector valued RKHS is a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions $f : \mathcal{X} \to \mathbb{R}^D$, such that for every $a \in \mathbb{R}^D$, and $x \in \mathcal{X}$, $\Gamma(x, \cdot)c$ belongs to $\mathcal{H}$ and moreover Γ has the reproducing property $\langle f, \Gamma(x, \cdot)c \rangle_{\mathcal{H}} = \langle f(x), c \rangle$. The space $\mathcal{H}$is closure of the linear span $\{f(x) = \sum_{i=1}^{N} \Gamma(x_i, x) a_j,\ a_j \in \mathbb{R}^D, x_1, \ldots, x_n \in \mathcal{X}\}$. Then, the choice of the kernel can be interpreted as inducing a representation for the functions of interest. 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

$$\Gamma(x, x') = k(x, x')B, \quad B = I, \tag{27}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar valued reproducing kernel. One can see that the choice of B corresponds to imposing a prior assumption on how the different components can be related, so that by choosing $B = I$ we are treating each component to be independent. In the following we will discuss in particular the case where the kernel is induced by a finite dimensional feature map,

$$k(x, x) = \langle \Phi(x), \Phi(x') \rangle_p\,, \quad \text{where} \quad \Phi : \mathcal{X} \to \mathbb{R}^p, \tag{28}$$

and $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^p$. In this case we can write each function in $\mathcal{H}$ as $f(x) = W\Phi(x)$, where $W \in \mathbb{R}^{T \times p}$ matrix.

References

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

[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] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[4] 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.

[5] I. Ekeland and T. Turnbull. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1983.

[6] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. North-Holland Publishing Co., Amsterdam, 1976. Translated from the French, Studies in Mathematics and its Applications, Vol. 1.

[7] Serge Lang. *Real and functional analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1993.

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

[9] M.D. Reid and R.C. Williamson. Composite binary losses. *JMLR*, 11, September 2010.

[10] I. Steinwart and A. Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[11] 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.

[12] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Some properties of regularized kernel methods. *JOURNAL OF MACHINE LEARNING RESEARCH*, 5:2004, 2004.

[13] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

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

[15] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, Vol. 32, No. 1, 56134, 2004.