Sharp Analysis of Learning with Discrete Losses

Supplementary Material

Alex Nowak-Vila, Francis Bach, Alessandro Rudi

INRIA - Département d'Informatique de l'École Normale Supérieure
PSL Research University
Paris, France

Organization of the Appendix

1 Calibration and fast rates for surrogate methods

The goal of Sec. 1 is to provide a generic method to sistematically improve the relation between excess risks of surrogate methods. Our analysis is a generalization of the one in [1], which was done for binary classification under 0-1 loss, to the case of general discrete losses.

In Sec. 1.1, we introduce the basic quantities used for the analysis of surrogate methods. Then, in Sec. 1.2 we focus on the central concept of *calibration*, which is key to study the statistical properties of these methods. In particular, we will re-derive the calibration properties of the Quadratic Surrogate (QS), which were proved in [2]. Finally, in Sec. 1.3, we derive our main result, which generalizes the Tsybakov condition, existing for multiclass and binary [3, 4] classification.

1.1 Prerequisites on surrogate methods

Given a loss $L : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ and a probability measure P on $\mathcal{X} \times \mathcal{Y}$, recall that the goal of supervised learning is to find the function f^* that minimizes the *expected risk* $\mathcal{E}(f)$ of the loss,

$$f^*(x) = \arg\min_{z \in \mathcal{Z}} \ell(z, x), \quad \mathcal{E}(f) = \mathbb{E}\, \ell(f(X), X), \tag{1}$$

where $\ell(z, x) = \int L(z, Y) dP(Y|x)$ is the *Bayes risk*. The goal of surrogate methods is to design a tractable *surrogate loss* $S : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$ defined on a *surrogate space* $\mathcal{C}$, such that when approximately minimized by a *surrogate function* $\widehat{g} : \mathcal{X} \to \mathcal{C}$, then it produces a good estimator $\widehat{f}$ of f^*. The mapping from $\widehat{g}$ to $\widehat{f}$ is performed with a *decoding function* $d : \mathcal{C} \to \mathcal{Z}$.

For a given surrogate S, we define the following quantities,

$$g^*(x) = \arg\min_{v \in \mathcal{C}} W(v, x), \quad W(v, x) = \int S(v, Y) dP(Y|x), \quad \mathcal{R}(g) = \mathbb{E}\, W(g(X), X), \tag{2}$$

were here, g^* is the *optimal surrogate function*, $W(v, x)$ is the *Bayes surrogate risk* and $\mathcal{R}(g)$ is the *expected surrogate risk* of g.

An important requirement for a surrogate method is the so-called *Fisher consistency*, which says that the optimum g^* of the surrogate S gives the optimum f^* of the loss L. It can be written as $f^* = d \circ g^*$.

Example 1.1 (Surrogate elements for the QS)**.** *In the case of the QS, we have that ,*

$$S(v, y) = \|v - U_y\|^2_{\mathbb{R}^r}, \quad \mathcal{C} = \mathbb{R}^r, \quad d(v) = \arg\min_{z \in \mathcal{Z}} F_z \cdot v, \tag{3}$$

and its Bayes excess risk $W(\widehat{g}(x), x) - W(g^*(x), x)$ has the following form,

$$W(\widehat{g}(x), x) - W(g^*(x), x) = \|\widehat{g}(x) - g^*(x)\|_2^2. \tag{4}$$

Moreover, it is Fisher consistent by construction ([2]).

Example 1.2 (Surrogate elements for the CRFs and SSVMs)**.** *(Assume $\mathcal{Z} = \mathcal{Y}$) Conditional Random Fields (CRFs) and Structural SVMs (SSVMs) are also surrogate methods for structured prediction. In this case, they split the output into a set of parts/cliques C as $\{\mathcal{Y}_c\}_{c \in C}$, which encode the structure of the output set. Then, they both consider*

$$\mathcal{C} = \mathbb{R}^r, \quad d(v) = \arg\max_{z \in \mathcal{Z}} \sum_{c \in C} v_{z_c}, \tag{5}$$

where $r = \sum_{c \in C} |\mathcal{Y}_c|$. The surrogate for CRFs has the following form (note that it does not dependent on any L),

$$S(v, y) = \log \left(\sum_{y' \in \mathcal{Y}} \exp \left(\sum_{c \in C} v_{y'_c} \right) \right) - \sum_{c \in C} v_{y_c}. \tag{6}$$

For SSVM, one assumes that the loss decomposes accordingly to the structure given by C. Then, it takes the following form,

$$S(v, y) = \max_{y' \in \mathcal{Y}} \left\{ \sum_{c \in C} \left(\{ L(y_c, y'_c) + v_{y'_c} \} \right) \right\} - \sum_{c \in C} v_{y_c}. \tag{7}$$

1.2 Calibration

Fisher consistency is an essential property of a surrogate method, nevertheless, it is only a property at the optimum. In practice the surrogate will be never optimized exactly, this is why it is important to study the concept of *calibration*, i.e, how the excess risk of the surrogate relates to the excess risk of the loss of interest.

This concept is formalized through the following definition 1.3.

Definition 1.3 (Calibration and Calibration function)**.** *We say that a surrogate S is* calibrated *w.r.t a loss L if there exists a convex function $H_{L,S} : \mathbb{R} \to \mathbb{R}$ with $H_{L,S}(0) = 0$ and positive in $(0, \infty)$, such that,*

$$H_{L,S}(\ell(d \circ g(x), x) - \ell(f^*(x), x)) \leq W(g(x), x) - W(g^*(x), x), \tag{8}$$

for every $x \in \mathcal{X}$.

Calibration means that for every x, one can control the excess of the Bayes risk by the excess Bayes risk of the surrogate.

Let's re-derive the form of the calibration function for the QS.

Lemma 1.4 (Calibration function for QS [2])**.** *Assumption 1 holds for the QS with*

$$H_{L,S}(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2} \tag{9}$$

Proof. Let's first decompose the Bayes risk into two terms A and B:

$$\begin{aligned}\ell(\widehat{f}(x), x) - \ell(f^*(x), x) &= \{\ell(\widehat{f}(x), x) - \widehat{\ell}(\widehat{f}(x), x)\} \\ &\quad + \{\widehat{\ell}(\widehat{f}(x), x) - \ell(f^*(x), x)\} \\ &= A + B.\end{aligned}$$

The first term, clearly $A \leq \sup_{z\in\mathcal{Z}} |\widehat{\ell}(z, x) - \ell(z, x)|$. For the second term , we use the fact that for any given two functions $\eta, \zeta : \mathcal{Z} \to \mathbb{R}$, it holds that $|\min_z \eta(z) - \min_z \zeta(z)| \leq \sup_z |\eta(z) - \zeta(z)|$. As $\widehat{f}(x)$ minimizes $\widehat{\ell}(\cdot, x)$ and $f^*(x)$ minimizes $\ell(\cdot, x)$, we can conclude also that $B \leq \sup_{z\in\mathcal{Z}} |\widehat{\ell}(z, x) - \ell(z, x)|$. Using the fact that $\widehat{\ell}(z, x) = F_z\widehat{g}(x)$ and $\ell(z, x) = F_z g^*(x)$, we can conclude that,

$$(\ell(f(x), x) - \ell(f^*(x), x))^2 \leq 2 \sup_{z\in\mathcal{Z}} (\widehat{\ell}(z, x) - \ell(z, x))^2 = 4\|F\|_\infty^2 \|\widehat{g}(x) - g^*(x)\|_2^2. \tag{10}$$

Re-arranging and using Eq. (4) gives the final result. □

The following important Theorem shows how Eq. (9) translates into a relation between excess risks, which are the quantities that we are ultimately interested at.

Theorem 1.5 (From Bayes risks to risks)**.** *Suppose Assumption 1 holds. Then,*

$$H_{L,S}(\mathcal{E}(f) - \mathcal{E}(f^*)) \leq \mathcal{R}(g) - \mathcal{R}(g^*) \tag{11}$$

Proof. This is a simple application of Jensen inequality.

$$\begin{aligned}H_{L,S}(\mathcal{E}(f) - \mathcal{E}(f^*)) &= H_{L,S}(\mathbb{E}_X(\ell(d \circ g(x), x) - \ell(f^*(x), x))) \\ &\leq \mathbb{E}_X\, H_{L,S}(\ell(d \circ g(x), x) - \ell(f^*(x), x)) \\ &= \mathbb{E}_X\, W(g(x), x) - W(g^*(x), x) \\ &= \mathcal{R}(g) - \mathcal{R}(g^*)\end{aligned}$$

□

If we combine Thm. 1.5 with Lemma 1.4, we obtain the comparison inequality for the QS.

Corollary 1.6 (Comparison inequality for QS [2])**.** *For the QS, we have that*

$$\mathcal{E}(f) - \mathcal{E}(f^*) \leq 2\ \|F\|_\infty \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)} \tag{12}$$

1.3 Improved calibration under low noise

Thm. 1.5 gives the ability to translate learning rates of the surrogate to learning rates of the full risk. However, as we will show, Eq. (11) can be loose in the presence of low noise at the boundary decision.

To formalize this, we will improve the result from the relation given by Thm. 1.5 under the p-noise assumption. We recall that the p-noise condition states that

$$P_{\mathcal{X}}(\gamma(X) < \varepsilon) = o(\varepsilon^p), \tag{13}$$

where $\gamma(x) = \min_{z' \neq f^*(x)} \ell(z', x) - \ell(f^*(x), x)$, is called the margin, and is defined as the minimum suboptimality gap between labels.

We have the following Lemma 1.7.

Lemma 1.7. *If the p-noise condition holds, then* $1/\gamma \in L_p(P_{\mathcal{X}})$.

Proof.

$$\begin{aligned}\|1/\gamma\|^p_{L_p(P_X)} &= \mathbb{E}\, 1/\gamma(X)^p \\ &= \int_0^\infty pt^{p-1} P_X(1/\gamma(X) > t)dt \\ &= \int_0^\infty pt^{p-1} P_X(\gamma(X) < t^{-1})dt.\end{aligned}$$

The integral converges if $P_X(\gamma(X) < t^{-1})$ decreases faster than t^{-p}. □

Let's now define the error set as $X_f = \{x \in \mathcal{X} \mid f(x) \neq f^*(x)\}$. The following Lemma 1.8, which bounds the probability of error by a power of the excess risk, is a generalization of the Tsybakov Lemma [4, Prop.1] for general discrete losses.

Lemma 1.8 (Bounding the size of the error set)**.** *If $1/\gamma \in L_p(P_X)$, then*

$$P_X(X_f) \leq \gamma_p^{\frac{1}{p+1}} (\mathcal{E}(f) - \mathcal{E}(f^*))^{\frac{p}{p+1}} \tag{14}$$

Proof. By the definition of the margin $\gamma(x)$, we have that:

$$1(f(x) \neq f^*(x)) \leq 1/\gamma(x) \Delta\ell(f(x), x) \tag{15}$$

By taking the $\frac{p}{p+1}$-th power on both sides, taking the expectation w.r.t P_X and finally applying Hölder's inequality, we obtain the desired result. □

Before proving Thm. 1.10, we will need the following useful Lemma 1.9 of convex functions.

Lemma 1.9 (Property of convex functions)**.** *Suppose $h : \mathbb{R} \to \mathbb{R}$ is convex and $h(0) = 0$. Then, for all $x > 0$, $0 \leq y \leq x$,*

$$h(y) \leq \frac{y}{x} h(x) \quad \textit{and} \quad h(x)/x \textit{ is increasing on } (0, \infty). \tag{16}$$

Proof. Take $\alpha = \frac{y}{x} < 1$. The result follows directly by definition of convexity, as $h(y) = h((1-\alpha)0 + \alpha x) \leq (1-\alpha)h(0) + \alpha h(x) = \frac{y}{x} h(x)$. For the second part, re-arrange the terms in the above inequality. □

The following Thm. 1.10, is an adaptation of Thm. 10 of [1], which was specific for binary 0-1 loss, now adapted to the case of general discrete losses.

Theorem 1.10 (Improved Calibration)**.** *Suppose that the surrogate S is calibrated with calibration function $H_{L,S}$ (see Eq. (8)) and the p-noise condition holds. Then, we have that*

$$H_{L,S,p}(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)) \leq \mathcal{R}(g) - \mathcal{R}(g^*), \tag{17}$$

where

$$H_{L,S,p}(\varepsilon) = (\gamma_p \varepsilon^p)^{\frac{1}{p+1}} H_{L,S}\left(\frac{1}{2}(\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}\right). \tag{18}$$

Moreover, we have that $H_{L,S,p}(\varepsilon) \geq \gamma_p^{\frac{1}{p+1}} H_{L,S}(\varepsilon/(2\gamma_p^{\frac{1}{p+1}}))$. Hence, $H_{L,S,p}$ never provides a worse rate than $H_{L,S}$.

Proof. (Of Thm. 1.10). Write the excess Bayes risk as $\Delta\ell(z', x) = \ell(z', x) - \ell(f^*(x), x)$.

We split the excess Bayes risk into a part with low noise $\Delta\ell(f(x), x) \leq t$ and a part with high noise $\Delta\ell(f(x), x) \geq t$. The first part will be controlled by the p-noise assumption and the second part by Eq. (8).

$$\begin{aligned}\mathcal{E}(d \circ g) - \mathcal{E}(f^*) &= \mathbb{E}_X\, \Delta\ell(f(X), X) \\ &= \mathbb{E}\,\{1(X_f)\Delta\ell(f(X), X)\} \\ &= \mathbb{E}\,\{\Delta\ell(f(X), X) 1(X_f \cap \{\Delta\ell(f(X), X) \leq t\}\} \\ &+ \mathbb{E}\,\{\Delta\ell(f(X), X) 1(X_f \cap \{\Delta\ell(f(X), X) \geq t\}\} \\ &= A + B.\end{aligned}$$

- *Bounding the error in the region with low noise A:*

$$A \leq tP_{\mathcal{X}}(X_f) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}}, \tag{19}$$

  where in the last inequality we have used Lemma 1.8.

- *Bounding the error in the region with high noise B:*

  We have that

$$\Delta\ell(f(x), x) 1(\Delta\ell(f(x), x) \geq t) \leq \frac{t}{H_{L,S}(t)} H_{L,S}(\Delta\ell(f(x), x)) \tag{20}$$

  In the case $\Delta\ell(f(x), x) < t$, inequality in Eq. (20) follows from the fact that $H_{L,S}$ is nonnegative. For the case $\Delta\ell(f(x), x) > t$, apply Lemma 1.9 with $h = H_{L,S}, x = \Delta\ell(f(x), x)$ and $y = t$.

  From Eq. (9), we have that $\mathbb{E}\{1(X_f) H_{L,S}(\Delta\ell(f(X), X))\} \leq \mathcal{R}(g) - \mathcal{R}(g^*)$. Hence,

$$B \leq \frac{t}{H_{L,S}(t)} (\mathcal{R}(g) - \mathcal{R}(g^*)) \tag{21}$$

Putting everything together,

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}} + \frac{t}{H_{L,S}(t)} (\mathcal{R}(g) - \mathcal{R}(g^*)), \tag{22}$$

and hence,

$$\left(\frac{\mathcal{E}(d \circ g) - \mathcal{E}(f^*)}{t} - \gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}} \right) H_{L,S}(t) \leq \mathcal{R}(g) - \mathcal{R}(g^*). \tag{23}$$

Choosing $t = \frac{1}{2} \gamma_p^{\frac{-1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{1}{p+1}}$ and substituting finally gives Eq. (18). The second part of the Theorem follows because $\frac{H_{L,S}(t)}{t}$ is non-decreasing by Lemma 1.9. □

Finally, if we apply Thm. 1.10 to the QS, we get the desired result as Cor. 1.11.

Corollary 1.11 (Improved comparison inequality for QS)**.** *For the QS, we have that*

$$\mathcal{E}(f) - \mathcal{E}(f^*) \leq \gamma_p^{\frac{1}{p+2}} \left(16 \|F\|_\infty^2 (\mathcal{R}(g) - \mathcal{R}^*)\right)^{\frac{p+1}{p+2}}. \tag{24}$$

Proof. Substituting $H_{L,S}(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2}$ in Eq. (18), gives that,

$$H_{L,S,p} = \frac{\varepsilon^{\frac{p+2}{p+1}}}{\gamma_p^{\frac{1}{p+1}} 16 \|F\|_\infty^2}. \tag{25}$$

Reversing the relation gives the comparison inequality in Eq. (24). □

2 Multilabel and ranking losses

The goal of this section is to derive all of the constants appearing on Table 1 of the paper.

In Sec. 2.1, we recall the elements that we need in order to derive the constants. In Sec. 2.2, we introduce the main tool from [5] that we use in order to study the optimality of the QS. Finally, the main bulk is in Sec. 2.3, where we analyse each loss separately.

2.1 Prerequisites.

Remember that the goal here is to study the statistical and computational properties of the QS-estimator $\widehat{f}_n : \mathcal{X} \to \mathcal{Z}$ defined as

$$\widehat{f}_n(x) = \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{n} \alpha_i(x) L(z, y_i). \tag{26}$$

Recall that the statistical complexity is determined by the following quantity,

$$L = FU^\top + c\mathbf{1}. \tag{27}$$

where $F = (F_z)_{z \in \mathcal{Z}} \in \mathbb{R}^{|\mathcal{Z}| \times r}, U = (U_y)_{y \in \mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}| \times r}$, $c \in \mathbb{R}$ is a scalar and $\mathbf{1} \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Y}|}$ is the matrix of ones, i.e. $\mathbf{1}_{ij} = 1$ and $r \in \mathbb{N}$. Here, F_z is the z-th row of F and U_y the y-th row of U. We denote by $\operatorname{affdim}(L)$ the *affine dimension* of the loss L, which is defined as the minimum r for which Eq. (27) holds.

Recall that the quantity of interest for the statistical complexity is

$$\mathsf{A} = \sqrt{r} \|F\|_\infty U_{\max}. \tag{28}$$

The inference complexity corresponds to the computational complexity of solving Eq. (26).

2.2 On the optimality of the QS.

We use results from [5] in order to study the optimality of the dimension of the QS. We implicitly use the concept of *convex calibration dimension of a loss L* (see Def. 10 in [5]), which is defined as the minimum dimension over all consistent convex surrogates w.r.t L. In the following Thm. 2.1 (their Thm. 18), they provide a sufficient condition to lower bound this dimension.

Theorem 2.1. *(In [5]) Let $L \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Y}|}$ the loss matrix. If $\exists \Pi \in \operatorname{relint}(\Delta_{|\mathcal{Y}|})$, $c \in \mathbb{R}$, such that $L\Pi = c1$, then there cannot exist any consistent convex surrogate with dimension less than* $\operatorname{affdim}(L) - 1$. *Here, $\Delta_{|\mathcal{Y}|}$ is the simplex of $|\mathcal{Y}|$ dimensions and* $\operatorname{relint}(A)$ *denotes the relative interior of the set A.*

In particular, Thm. 2.1 says that if there exists at least one distribution Π at the interior of the simplex for which the Bayes risk is the same for all labels, then one can't hope to be consistent by estimating less than $\operatorname{affdim}(L) - 1$ scalar functions. In particular, this means that the QS is essentially optimal over all surrogate methods, in the sense that it estimates $\operatorname{affdim}(L)$ scalar functions.

For each loss, we test the condition given by Thm. 2.1 to show the optimality (or not) of the Quadratic Surrogate approach.

Note that there exist problems for which you can find consistent surrogates with dimension much smaller than $\operatorname{affdim}(L)$. In ordinal regression, where the discrete labels have a natural order, there exist one dimensional surrogates [6] despite the loss matrix being full rank.

2.3 Analysis of the losses

Notation. In the following we denote by $m \in \mathbb{N}$ the number of classes of a multilabel/ranking problem, by $\mathcal{P}_m$ the power-set of $[m] = \{1, \dots, m\}$ and by $\mathfrak{S}_m$ the set of permutations of m-elements. In particular note that in the multilabel problems both the output space $\mathcal{Z}$ and the observation space $\mathcal{Y}$ are equal to $\mathcal{P}_m$, while in ranking $\mathcal{Z} = \mathfrak{S}_m$ and $\mathcal{Y} = [R]^m$, the set of observed relevance scores for the m documents where R is the highest relevance [7]. Finally we denote by $[v]_j$ the j-th element of a vector v and we identify $\mathcal{P}_m$ with $\{0,1\}^m$, moreover $\sigma(j)$ is the j-th element of the permutation σ, for $\sigma \in \mathfrak{S}_m, j \in [m]$.

0-1 loss

The 0-1 loss is defined as 0 if the subsets are exactly equal and 1 otherwise, i.e, it does not provide any structural information. In this case, $\mathcal{Y} = \mathcal{Z} = \{0,1\}^m$ and

$$L(z, y) = 1(z \neq y). \tag{29}$$

- **Statistical complexity.** We can decompose it as

$$F_z = -(1_{[z=z']})_{z' \in \{0,1\}^m},\ U_y = (1_{[y=y']})_{y' \in \{0,1\}^m},\ c = 1. \tag{30}$$

  We have that

$$r = 2^m,\ \|F\|_\infty = 1,\ U_{\max} = 1. \tag{31}$$

  Hence,

$$\mathsf{A} = 2^{m/2}. \tag{32}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \arg\max_{z \in \mathcal{P}_m} \sum_{i | y_i = z} \alpha_i(x), \tag{33}$$

  which can be done in

$$\mathcal{O}(2^m \wedge n). \tag{34}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. 2.1, one has that $\mathrm{affdim}(L) = 2^m$ is optimal.

Block 0-1 loss

Assume that the prediction space $\mathcal{P}_m$ is partitioned into b regions $\mathcal{P}_m = \sqcup_{j=1}^b B_j$. The block 0-1 loss is defined as 0 if the subsets belong to the same region and 1 otherwise. In this case, $\mathcal{Y} = \mathcal{Z} = \{0,1\}^m$ and

$$L(z, y) = 1(z \in B_j, y \notin B_j, \text{for some } j \in [b]). \tag{35}$$

- **Statistical complexity.** We can decompose it as

$$F_z = -(1_{[z \in B_j]})_{j=1}^b,\ U_y = (1_{[y \in B_j]})_{j=1}^b,\ c = 1. \tag{36}$$

  We have that

$$r = b,\ \|F\|_\infty = 1,\ U_{\max} = 1. \tag{37}$$

  Hence,

$$\mathsf{A} = \sqrt{b}. \tag{38}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \arg\max_{1 \leq j \leq b} \sum_{i | y_i \in B_j} \alpha_i(x), \tag{39}$$

  which can be done in

$$\mathcal{O}(b) \tag{40}$$

- **Optimality of r.** Taking $\Pi_y = \frac{1}{b|B(y)|}$, where $B(y)$ is the partition where $y \in \mathcal{Y}$ belongs to and applying Thm. 2.1, one has that $\mathrm{affdim}(L)$ is optimal.

Hamming

The Hamming loss counts the average number of classes that disagree. In this case, $\mathcal{Y} = \mathcal{Z} = \{0,1\}^m$ and

$$L(z,y) = \frac{1}{m}\sum_{j=1}^{m} 1([z]_j \neq [y]_j). \tag{41}$$

- **Statistical complexity.** If we define $s_j(y) = 2[y]_j - 1$, we can re-write the Hamming loss as

$$L(z,y) = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1 - s_j(z)s_j(y)}{2}\right) = \frac{1}{2} - \frac{1}{2m}\sum_{j=1}^{m} s_j(z)s_j(y).$$

  This implies that

$$F_z = -\frac{1}{2m}(s_j(z))_{j=1}^m,\ U_y = (s_j(y))_{j=1}^m,\ c = \frac{1}{2}. \tag{42}$$

  We have that

$$\|F\|_\infty = \frac{1}{2\sqrt{m}},\ U_{\max} = 1. \tag{43}$$

  Hence,

$$\mathsf{A} = \frac{1}{2}. \tag{44}$$

- **Inference.** Inference corresponds to

$$\widehat{f}_j(x) = \left(\frac{\operatorname{sign}\left(\widehat{g}_j(x)\right) + 1}{2}\right), \quad \text{where} \quad \widehat{g}_j(x) = \sum_{i=1}^{n} s_j(y_i)\alpha_i(x), \tag{45}$$

  which can be done in

$$\mathcal{O}(m). \tag{46}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. 2.1, one has that $\operatorname{affdim}(L) = m$ is optimal.

Prec@k

Prec@k (Precision at k) measures the average number of elements in the predicted k-set that also belong to the ground truth. In this case, the prediction space is $\mathcal{Z} = \mathcal{P}_{m,k}$, i.e, subsets of $[m]$ of size k, and $\mathcal{Y} = \mathcal{P}_m$.

$$L(z,y) = 1 - \frac{|y \cap z|}{k} = 1 - \frac{1}{k}\sum_{j=1}^{m}[z]_j[y]_j. \tag{47}$$

- **Statistical complexity.** We have that $r = m, F_z = -\frac{1}{k}([z]_j)_{j=1}^m,\ U_y = ([y]_j)_{j=1}^m,\ c = 1, \|F\|_\infty = \frac{1}{\sqrt{k}},\ U_{\max} = 1$. Hence,

$$\mathsf{A} = \sqrt{\frac{m}{k}}. \tag{48}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \underset{z \in \mathcal{P}_{m,k}}{\arg\operatorname{top}_k}\left(\left(\sum_{i|[y_i]_j=1}\alpha_i(x)\right)_{j=1}^{m}\right), \tag{49}$$

  which can be done in

$$\mathcal{O}(m \log k). \tag{50}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. 2.1, one has that $\operatorname{affdim}(L) = m$ is optimal.

F-score

The F-score is defined as the harmonic mean of precision and recall. In this case $\mathcal{Z} = \mathcal{Y} = \mathcal{P}_m$ and

$$L(z, y) = 1 - 2\frac{|z \cap y|}{|z| + |y|}, \tag{51}$$

where we treat the case $y = 0$ as follows:

$$2\frac{|z \cap y|}{|z| + |y|} = \begin{cases} 2\sum_{j=1}^{m}\sum_{\ell=0}^{m} \frac{[z]_j}{\ell+|z|} 1([y]_j = 1, |y| = \ell) & y \neq 0 \\ 1(z = 0) & y = 0 \end{cases}. \tag{52}$$

Let's define the matrix $P(x) \in \mathbb{R}^{m\times m}$ and $p_0(x) \in \mathbb{R}$ as,

$$P_{j\ell}(x) = P([Y]_j = 1, |Y| = \ell | X = x), \qquad p_0(x) = P(Y = 0|X = x). \tag{53}$$

Then, the Bayes risk reads

$$\ell(z, x) = \begin{cases} 2\sum_{j=1}^{m}\sum_{\ell=0}^{m} \frac{[z]_j}{\ell+|z|} P_{j\ell}(x) & y \neq 0 \\ p_0(x) & y = 0 \end{cases}. \tag{54}$$

Hence, for every x, one needs no more than $r = m^2 + 1$ parameters to compute the F-score Bayes risk.

We have the following Lemma 2.2.

Lemma 2.2. *Given the matrix $P(x) \in \mathbb{R}^{m\times m}$ and the scalar $p_0(x)$, inference can be performed through the following two-step procedure:*

1. *Compute the matrix $A(x) \in \mathbb{R}^{m\times m}$:*

$$A_{jk}(x) = \sum_{\ell=0}^{m} \frac{P_{j\ell}(x)}{\ell + k} \tag{55}$$

   This is a matrix-by-matrix multiplication that takes $\mathcal{O}(m^3)$.

2. *From $A(x)$ and $p_0(x)$, the prediction $f(x)$ can be computed in $\mathcal{O}(m^2)$ through an iterated maximization procedure.*

Proof. Suppose we have already computed $A(x) \in \mathbb{R}^{m\times m}$ and $p_0(x) \in \mathbb{R}$. Now, we perform the following m maximizations:

$$f^{(k)}(x) = \arg\max_{z\in\mathcal{P}_{m,k}} A^T_{\cdot,k}(x)z, \qquad \text{for } k = 1, \ldots, m. \tag{56}$$

Then, $f^*(x)$ is computed by taking the maximum over the $f^{(k)}(x)$'s together with $p_0(x)$, which corresponds to $z = 0$. □

- **Statistical complexity.**

  Note that depending on whether we approximate P or directly A, we have different computational complexities. In particular, if the surrogate approximates directly A, then it avoids the operation Eq. (55). As the estimator is the same, the statistical complexity is the minimum of both.

  Decomposition 1. Estimating $P(x)$, corresponds to the following decomposition:

$$F_{z,m(\ell-1)+j} = -\left(\frac{1([z]_j = 1)}{|z| + \ell}\right), \qquad 1 \leq j, \ell \leq m$$
$$U_{y,m(\ell-1)+j} = 1([y]_j = 1, |y| = \ell), \qquad 1 \leq j, \ell \leq m$$

and $F_{z,m^2+1} = 1(z=0), U_{y,m^2+1} = 1(y=0)$. In this case,

$$r = m^2 + 1,\ \frac{1}{2} \le \|F\|_\infty \le 1,\ U_{\max} = 1. \tag{57}$$

Hence,

$$\mathsf{A}_1 \le \sqrt{m^2+1} \le \sqrt{2}m \tag{58}$$

Decomposition 2. Estimating $A(x)$, corresponds to the following decomposition:

$$F_{z,m(\ell-1)+j} = -1([z]_j = 1, |z| = \ell), \qquad 1 \le j, \ell \le m$$
$$U_{y,m(\ell-1)+j} = \left(\frac{1([y]_j = 1)}{|y| + \ell}\right), \qquad 1 \le j, \ell \le m$$

and $F_{z,m^2+1} = 1(z=0), U_{y,m^2+1} = 1(y=0)$.
In this case,

$$r = m^2 + 1,\ \|F\|_\infty = \sqrt{m},\ U_{\max} = 1. \tag{59}$$

Hence,

$$\mathsf{A}_2 = \sqrt{m(m^2+1)} \le m\sqrt{2m}. \tag{60}$$

We take $\mathsf{A} = \min(\mathsf{A}_1, \mathsf{A}_2)$, hence,

$$\mathsf{A} \le \sqrt{2}m. \tag{61}$$

- **Inference.** The quadratic surrogate approximates $A(x)$ and $P(x)$ as:

$$\widehat{P}_{j\ell}(x) = \sum_{i|[y_i]_j=1, |y_i|=\ell} \alpha_i(x), \quad \widehat{A}_{jk}(x) = \sum_{\ell=0}^{m} \frac{\widehat{P}_{j\ell}(x)}{\ell + k},\ \widehat{p}_0(x) = \sum_{i|y_i=0} \alpha_i(x). \tag{62}$$

If we use *Decomposition 1*, i.e, $\widehat{g}(x) = (\widehat{P}(x), \widehat{p}_0(x))$, then we have cubic inference,

$$\mathcal{O}(m^3). \tag{63}$$

If we use *Decomposition 2*, i.e, $\widehat{g}(x) = (\widehat{A}(x), \widehat{p}_0(x))$, then we have quadratic inference,

$$\mathcal{O}(m^2). \tag{64}$$

- **Optimality of r.** We can't say anything about the potential existence of a convex calibrated surrogate with smaller dimension than $\text{affdim}(L) - 1$. This is because the sufficient condition from Theorem 18 of [5] does not hold for any Π even for $m = 2$.

NDCG-type

Let $\mathcal{Z} = \mathfrak{S}_m$ be the set of permutations of m elements and $\mathcal{Y} = \{1, \dots, R\}^m = [R]^m$ the space of relevance scores for m documents. Let the *gain* $G : \mathbb{R} \to \mathbb{R}$ be an increasing function and the *discount* vector $D = (D_j)_{j=1}^m$ be a coordinate-wise decreasing vector. NDCG-type losses are defined as the normalized discounted sum of the gain of the relevance scores ordered by the predicted permutation:

$$L(\sigma, r) = 1 - \frac{1}{N(r)} \sum_{j=1}^{m} G([r]_j) D_{\sigma(j)} \tag{65}$$

where $N(r) = \max_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^m G([r]_j) D_{\sigma(j)}$ is the normalizer. The discount is performed in order to give more importance to the relevance of the top ranked elements.

- **Statistical complexity.**

  Note that looking at Eq. (65) we can directly write that $r = m$ and $F_\sigma = -(D_{\sigma(j)})_{j=1}^m, U_r = \left(\frac{G([r]_j)}{N(r)}\right)_{j=1}^m, c = 1$.

  It follows that,

$$\|F\|_\infty = \sqrt{\sum_{j=1}^m D_j^2}, \quad U_{\max} = G_{\max} D_{\max}, \tag{66}$$

  hence,

$$\mathsf{A} = \sqrt{m} G_{\max} D_{\max} \sqrt{\sum_{j=1}^m D_j^2}. \tag{67}$$

- **Inference.**

  The inference corresponds to,

$$\widehat{f}(x) = \operatorname{argsort}_{\sigma \in \mathfrak{S}_m}(v), \qquad \text{where } v_j = \sum_{i=1}^n \frac{G([r_i]_j)\alpha_i(x)}{N(r_i)}, \tag{68}$$

  which can be done in

$$\mathcal{O}(m \log m) \tag{69}$$

  operations.

- **Optimality of r.** Optimal. As Hamming, the barycenter of the simplex satisfies Thm. 2.1.

Normalized Discounted Cumulative Gain (NDCG) This is the most widely used configuration, in this case, $G(t) = 2^t - 1$ and $D_j = \frac{1}{\log(j+1)}$. We have that $\|D\|_2 \sim \left(\int_2^m \frac{1}{\log^2(t)} dt\right)^{1/2} \sim \sqrt{\frac{m}{\log m}}$. And hence,

$$\mathsf{A} \leq c\ G_{\max} \frac{m}{\sqrt{\log m}}. \tag{70}$$

Expected Rank Utility (ERU) In this case, $G(t) = \max(t - \bar{r}, 0)$ and $D_j = 2^{1-j}$, where $\bar{r}$ corresponds to a neutral score. We have that $\|D\|_2 \leq \frac{2}{\sqrt{3}}$,

$$A \leq \frac{2}{\sqrt{3}}\ G_{\max} \sqrt{m}. \tag{71}$$

The QS-estimator estimates the marginals of the normalized relevance scores and sorts the estimates at inference. As it was shown in [7], in order to be consistent for NDCG, one has to estimate the *normalized* relevance scores and not the *unnormalized* ones as one would do at the first place. In particular, the QS-estimator for the NDCG that follows directly from our framework corresponds exactly to their proposed consistent algorithm.

Due to the discount factor, the statistical complexity grows with the number of elements to sort. In particular, faster the decay is, more samples you need to optimize the corresponding loss. This is shown in the two examples we have shown, where the NDCG is statistically easier to optimize than the ERU.

Pairwise Disagreement (PD)

The pairwise disagreement computes the cost associated to a given permutation in terms of pairwise comparisons using binary relevance scores. In this case, $\mathcal{Z} = \mathfrak{S}_m$, $\mathcal{Y} = [0,1]^m = \mathcal{P}_m$, and,

$$L(\sigma, y) = \frac{1}{N(y)} \sum_{j=1}^m \sum_{\ell \neq j} 1([y]_j < [y]_\ell) 1(\sigma(j) > \sigma(\ell)), \tag{72}$$

where $N(y) = \sup_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^m \sum_{\ell \neq j} 1([y]_j < [y]_\ell) 1(\sigma(j) > \sigma(\ell)) = |y|(m - |y|)$ is a normalizer.

- **Statistical complexity.** Note that we can re-write

$$1([y]_j < [y]_\ell) = \frac{\text{sign}([y]_\ell - [y]_j) + 1}{2}, \quad 1(\sigma(j) > \sigma(\ell)) = \frac{\text{sign}(\sigma(j) - \sigma(\ell)) + 1}{2}. \tag{73}$$

Hence,

$$L(\sigma, y) = \frac{1}{4} + \frac{1}{4N(y)} \sum_{j=1}^{m} \sum_{\ell \neq j} \text{sign}([y]_\ell - [y]_j)\text{sign}(\sigma(j) - \sigma(\ell)) \tag{74}$$

Note that $F_\sigma = 1/4(\text{sign}(\sigma(j) - \sigma(\ell)))_{j,\ell=1}^m$ and $U_y = (\frac{\text{sign}([y]_\ell - [y]_j)}{N(y)})_{j,\ell=1}^m$ are anti-symmetric matrices. Hence, they can be described with $m(m-1)/2$ numbers. We can then consider F_σ and U_y as vectors of $m(m-1)/2$ coordinates.

This implies that $r = m(m-1)/2,\ c = 1/4, \|F\|_\infty = 1/4\sqrt{m(m-1)/2},\ U_{\max} = \frac{2}{m-1}$. Hence,

$$\mathsf{A} = \frac{m}{4} \tag{75}$$

- **Inference.** In this case, the optimization problem reads

$$\widehat{f}(x) \in \arg\min_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^{m} \sum_{\ell \neq j} \gamma_{j\ell}(x) 1(\sigma(j) > \sigma(\ell)), \tag{76}$$

with

$$\gamma_{j\ell}(x) = \sum_{i | [y_i]_j < [y_i]_\ell} \frac{\alpha_i(x)}{N(y_i)}. \tag{77}$$

This is precisely a Minimum Weight Feedback Arcset (MWFAS) problem with associated directed graph having weights $\gamma_{j\ell}(x)$. This problem is known to be NP-Hard.

- **Optimality of r.** Optimal. See Corollary 19 and Proposition 20 from [5].

As it was shown in [8], there is no hope of devising a consistent convex surrogate method which is based on sorting an estimated vector of relevance scores. In particular, one needs to estimate $\frac{m(m-1)}{2}$ scalar functions corresponding to the weights of a graph between the classes. Although estimating the graph structure is statistically feasible, inference corresponds to finding a directed acyclic graph (DAG) with minimum cost. This is equivalent to the Minimum Weight Feedback Arcset Problem (MWFAS), which is known to be NP-Hard. Consequently, one can state that, unless $P = NP$, there does not exist any polynomial surrogate-based consistent algorithm for the PD loss. If it existed, one could solve the Bayes risk minimization problem, i.e., MFWAS, to ε-accuracy in poly$(\frac{1}{\varepsilon})$.

Mean Average Precision (MAP)

The mean average precision (MAP) is a widely used ranking measure in information retrieval. The precision associated to a relevant document j ($[y]_j = 1$) ranked at position $\sigma(j)$ is the Precision at $\sigma(j)$ of the $\sigma(j)$ retrieved documents ranked before (and including), j. In this case, $\mathcal{Z} = \mathfrak{S}_m$ and $\mathcal{Y} = [0,1]^m = \mathcal{P}_m$. The mean average precision corresponds to the mean over all relevant documents in y. Hence, MAP has the following form:

$$L(\sigma, y) = 1 - \frac{1}{|y|} \sum_{j | [y]_j = 1} \frac{1}{\sigma(j)} \sum_{\ell=1}^{\sigma(j)} [y]_{\sigma^{-1}(\ell)}. \tag{78}$$

Note that it can be re-written as

$$\begin{aligned} L(\sigma, y) &= 1 - \frac{1}{|y|} \sum_{j=1}^{m} \frac{[y]_j}{\sigma(j)} \sum_{\ell=1}^{\sigma(j)} [y]_{\sigma^{-1}(\ell)} \\ &= 1 - \frac{1}{|y|} \sum_{j=1}^{m} \sum_{\ell=1}^{j} \frac{[y]_{\sigma^{-1}(\ell)} [y]_{\sigma^{-1}(j)}}{j} \\ &= 1 - \frac{1}{|y|} \sum_{j=1}^{m} \sum_{\ell=1}^{j} \frac{[y]_\ell [y]_j}{\max(\sigma(j), \sigma(\ell))}. \end{aligned}$$

- **Statistical complexity.** We have that $r = \frac{m(m+1)}{2}, F_\sigma = \left(\max(\sigma(j), \sigma(\ell))^{-1}\right)_{j \geq \ell}$, $U_y = -\left(\frac{[y]_j [y]_\ell}{|y|}\right)_{j \geq \ell}$, $c = 1, \|F\|_\infty \leq \sqrt{\log(m+1)}$, $U_{\max} = 1/2$. Hence,

$$A = \frac{1}{2} m \sqrt{\log(m+1)} \tag{79}$$

- **Computational complexity.** The inference problem reads

$$\widehat{f}(x) = \arg\max_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^{m} \sum_{\ell=1}^{j} \frac{1}{\max(\sigma(j), \sigma(\ell))} \sum_{i | [y_i]_j [y_i]_\ell = 1} \frac{\alpha_i(x)}{|y_i|} \tag{80}$$

  Denote by

$$W_{j\ell} = \begin{cases} \sum_{i | [y_i]_j [y_i]_\ell = 1} \frac{\alpha_i(x)}{|y_i|} & j \geq \ell \\ 0 & \text{otherwise} \end{cases}, \; D_{j\ell} = \begin{cases} \max(j, \ell)^{-1} & j \geq \ell \\ 0 & \text{otherwise} \end{cases} \tag{81}$$

  We have that,

$$\widehat{f}(x) = \arg\max_{\sigma \in \mathfrak{S}_m} \sum_{j,\ell=1}^{m} W_{j\ell} D_{\sigma(j)\sigma(\ell)} \equiv \arg\max_{P \in \Pi_m} \operatorname{Tr}(W^T P D P^T), \tag{82}$$

  where Π_m is the set of permutation matrices of size m. This is an instance of the Quadratic Assignment Problem (QAP).

- **Optimality of r.** Optimal. See Corollary 19 and Proposition 21 from [5].

As for PD, inference for MAP corresponds to a NP-Hard problem, more specifically, to an instance of the Quadratic Assignment Problem (QAP). Consequently, one can conclude analogously as for the PD loss, i.e., that no efficient and consistent surrogate algorithm exists for MAP.

References

[1] 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.

[2] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pages 4412–4420, 2016.

[3] Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jeacques Slotine. Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2012.

[4] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[5] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

[6] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *The Journal of Machine Learning Research*, 18(1):1769–1803, 2017.

[7] Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On ndcg consistency of listwise ranking methods. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 618–626, 2011.

[8] Clément Calauzenes, Nicolas Usunier, and Patrick Gallinari. On the (non-) existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems*, pages 197–205, 2012.