Sharp Analysis of Learning with Discrete Losses

Alex Nowak-Vila, Francis Bach, Alessandro Rudi

Sharp Analysis of Learning with Discrete Losses

Alex Nowak-Vila, Francis Bach, Alessandro Rudi

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

October 16, 2018

The problem of devising learning strategies for discrete losses (e.g., multilabeling, ranking) is currently addressed with methods and theoretical analyses *ad-hoc* for each loss.

In this paper we study a least-squares framework to systematically design learning algorithms for discrete losses, with quantitative characterizations in terms of statistical and computational complexity. In particular we improve existing results by providing explicit dependence on the number of labels for a wide class of losses and faster learning rates in conditions of low-noise.

Theoretical results are complemented with experiments on real datasets, showing the effectiveness of the proposed general approach.

1. Introduction

Structured prediction with discrete labels of high cardinality is ubiquitous in machine learning, e.g., in multiclass problems, multilabel learning, ranking, ordinal regression, etc. [1, 2, 3, 4].

These supervised learning problems typically come with computational and theoretical challenges:

(1) how to design efficient algorithms dealing with potentially large number of data and labels?

(2) even if learning is computationally feasible, how to make sure that the resulting algorithm leads to improved accuracy on the test set*?*

Many special cases are often addressed in an *ad-hoc* fashion in terms of consistency, algorithms and convergence rates, depending on the specific loss used in each application to quantify the performance of predictors.

A few generic learning frameworks exist: (a) conditional random fields [5, 6] use conditional probabilistic modelling typically combined with maximum likelihood estimation, but may lead to intractable probabilistic inference and cannot easily incorporate structured losses which are needed in applications [7]; (b) Structured SVM [8, 9] extended the class of problems where a systematic max-margin framework can be applied, with the incorporation of arbitrary losses, but they are not consistent in general, that is, even with infinite amounts of data, they would not lead to optimal predictions [10]; (c) more recently, least-squares (or quadratic surrogate) frameworks [11, 12] have emerged. Such approaches can tackle arbitrary discrete losses producing consistent estimators and have the potential to provide a systematic way to design learning algorithms with both statistical and computational

guarantees. However, no sharp analyses exist yet, quantifying the impact of crucial quantities like the number of labels or the level of noise on the statistical and computational properties of the resulting algorithms. The goal of this paper is to characterize explicitly such impact for a number of widely used loss functions in the context of multilabeling and ranking, showing the effectiveness of least-squares frameworks for structured prediction with discrete labels.

We make the following contributions:

- We provide quantitative characterizations of the statistical and computational complexity for the least-squares framework of [11] depending on the number of labels and the number of examples. The characterization is explicit for a wide family of common losses in ranking and multilabel learning (see Secs. 3.1, 3.2 and 4).
- We propose a margin condition for discrete losses (generalizing the Tsybakov condition for binary classification [13]) and obtain fast learning rates for the framework of [11], that are adaptive to the proposed condition (see Sec. 3.3).
- Our analysis encompasses many previous results on special cases and provides improved learning rates over existing generic structured prediction frameworks (see Sec. 6).
- We conduct a series of experiments highlighting the benefits of the considered least-squares framework on ranking and multilabel problems (Sec. 5).

2. Background

The problem of *supervised learning* consists in learning from examples the function relating inputs with observations/labels. More specifically, let $\mathcal{Y}$ be the space of observations, denoted *observation space* or *label space* and $\mathcal{X}$ be the *input space*. The quality of the predicted output is measured by a given *loss function* L. In many scenarios the output of the function is in a different space than the observations (see Sec. 3.2 for some examples). We denote by $\mathcal{Z}$ the *output space*, so

$$L : \mathcal{Z} \times \mathcal{Y} \longrightarrow \mathbb{R}, \tag{1}$$

where $L(z, y)$ measures the cost of predicting z when the observed value is y. Finally the data are assumed to be distributed according to a probability measure P on $\mathcal{X} \times \mathcal{Y}$. The goal of supervised learning is then to recover the function $f^\star$ minimizing the *expected risk* $\mathcal{E}(f)$ of the loss,

$$f^\star = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Z}} \mathcal{E}(f), \quad \mathcal{E}(f) = \mathbb{E}\, L(f(X), Y), \tag{2}$$

given only a number of examples $(x_i, y_i)_{i=1}^n$, with $n \in \mathbb{N}$, sampled independently from P. The quality of an estimator f for $f^\star$ is measured in terms of the *excess risk* $\mathcal{E}(f) - \mathcal{E}(f^\star)$.

2.1. Quadratic Surrogate method

A systematic way to solve the problem in Eq. (2) is to consider that $f^\star$ is characterized as follows [14, 11]:

$$f^\star(x) = \operatorname*{arg\,min}_{z\in\mathcal{Z}} \ \ell(z, x),$$

where $\ell(z, x) = \int_{\mathcal{Y}} L(z, y) dP(y|x)$ is the *Bayes risk*, defined as the conditional expectation of y given $x \in \mathcal{X}$. The quadratic surrogate (QS) for structured prediction, introduced in [11], is a natural estimator that has the following form,

$$\widehat{f}(x) = \operatorname*{arg\,min}_{z\in\mathcal{Z}} \ \widehat{\ell}(z, x), \tag{3}$$

Multilabel and Ranking measures					
Measure	$\mathcal{Z}$	Definition	r	A	$\mathrm{INF}_F(\lvert\mathcal{Z}\rvert)$
0-1 ($\downarrow$)	$\mathcal{P}_m$	$1(z \neq y)$	2^m	$2^{m/2}$	$\mathcal{O}(n \wedge 2^m)$
Block 0-1 ($\downarrow$)	$\mathcal{P}_m$	$1(z \in B_j, y \notin B_j,\ j \in [b])$	b	$\sqrt{b}$	$\mathcal{O}(b)$
Hamming ($\downarrow$)	$\mathcal{P}_m$	$\frac{1}{m}\sum_{j=1}^m 1([z]_j \neq [y]_j)$	m	$\frac{1}{2}$	$\mathcal{O}(m)$
F-score ($\uparrow$)	$\mathcal{P}_m$	$2\frac{\lvert z\cap y\rvert}{\lvert z\rvert+\lvert y\rvert}$	m^2+1	$\sqrt{2}m$	$\mathcal{O}(m^2)$
Prec@k ($\uparrow$)	$\mathcal{P}_{m,k}$	$\frac{\lvert z\cap y\rvert}{k}$	m	$\sqrt{\frac{m}{k}}$	$\mathcal{O}(m\log k)$
NDCG ($\uparrow$)	$\mathfrak{S}_m$	$\frac{1}{N(r)}\sum_{j=1}^m G([r]_j)D_{\sigma(j)}$	m	$\sqrt{m}\ (\sum_j D_j^2)^{\frac{1}{2}} G_{\max}$	$\mathcal{O}(m\log m)$.
PD ($\downarrow$)	$\mathfrak{S}_m$	$\frac{1}{N(y)}\sum_{j,\ell=1}^m 1_{([y]_j<[y]_\ell)}1_{(\sigma(j)>\sigma(\ell))}$	$\frac{m(m-1)}{2}$	$\frac{m}{4}$	MWFAS(m).
MAP ($\uparrow$)	$\mathfrak{S}_m$	$\frac{1}{\lvert y\rvert}\sum_{j=1}^m \frac{[y]_j}{\sigma(j)}\sum_{\ell=1}^{\sigma(j)} y_{\sigma^{-1}(\ell)}$	$\frac{m(m+1)}{2}$	$\frac{1}{2}m\sqrt{\log(m+1)}$	QAP(m).

Table 1: Upper bounds for A in Thms. 3.1 and 3.5 and Cor. 3.6 and computational complexity of evaluating the QS estimator in Eq. (9), for a number of widely-used losses for multilabel/ranking problems. See Sec. 3.2 for notation, Sec. 4 for computational considerations and Appendix B for the full derivation of the results.

where $\widehat{\ell}(z,x) := \sum_{i=1}^n \alpha_i(x)L(z,y_i)$. Here $(\alpha_i)_{i=1}^n$ are suitable functions defined explicitly in terms of the observed data (not on L) and will be discussed later (see Eqs. (8) and (9)). Informally, the closer $\widehat{\ell}(z,x)$ is to $\ell(z,x)$, the closer $\widehat{f}$ will be to $f^\star$ in terms of the excess risk. In [11] a detailed statistical framework analyzes the generalization properties of the derived estimator, that will be recalled in the next paragraph. Here we point out that a crucial aspect of the algorithm in Eq. (3), that makes it appealing from a practical viewpoint, is that we can directly apply it given the loss at hand, without the need to devise a different surrogate (and consequently a different algorithm and theoretical analysis) *ad-hoc* for each specific loss.

Statistical properties of Quadratic Surrogate. Here we recall some generalization properties of the QS estimator from [11], that will be extended in Sec. 3. First, assume that the loss L is a structure encoding loss function (SELF), i.e., it can be written as,

$$L(z,y) = \langle \varphi(z), V\psi(y)\rangle_{\mathcal{H}}, \tag{4}$$

where $\mathcal{H}$ is a separable Hilbert space with $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ the associated inner product, $V : \mathcal{H} \to \mathcal{H}$ is a bounded linear operator and $\varphi : \mathcal{Z} \to \mathcal{H}$, $\psi : \mathcal{Y} \to \mathcal{H}$.

Note that by assuming $\mathcal{Z}, \mathcal{Y}$ discrete and finite, then every loss function on $\mathcal{Z}, \mathcal{Y}$ is SELF. Indeed Eq. (4) is recovered by setting $\mathcal{H} = \mathbb{R}^{|\mathcal{Z}|}$, $V = (L(z,y))_{z\in\mathcal{Z},y\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Y}|}$ the loss matrix, and $\varphi(z) = e_z, \psi(y) = e_y$ the vectors of the canonical basis in $\mathbb{R}^{|\mathcal{Z}|}$ and $\mathbb{R}^{|\mathcal{Y}|}$, respectively (For the case of continuous $\mathcal{Z}, \mathcal{Y}$ see [11]).

The key property of a loss being SELF is that, by linearity of the inner product,

$$\begin{aligned}\ell(z,x) &= \int L(z,y)dP(y|x)\\ &= \int \langle \varphi(z), V\psi(y)\rangle_{\mathcal{H}} dP(y|x)\\ &= \langle \varphi(z), Vg^*(x)\rangle_{\mathcal{H}},\end{aligned}$$

with $g^*(x) = \int \psi(y)dP(y|x)$ being the conditional expectation of $\psi(y)$, given x. This means that in order to estimate ℓ, we just need to find an estimator $\widehat{g}$ for the conditional expectation g^*, and then

define $\widehat{\ell}(z,x) = \langle \varphi(z), V\widehat{g}(x)\rangle_{\mathcal{H}}$. To find a suitable estimator for g^*, note that g^* can be written as the minimizer of the following quadratic surrogate (QS),

$$g^* = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathcal{H}} \mathcal{R}_\psi(g), \tag{5}$$

where $\mathcal{R}_\psi(g) := \int \|g(x) - \psi(y)\|^2_{\mathcal{H}} dP(x,y)$ is the *expected surrogate risk* of g. The quality of the surrogate estimator g is measured in terms of the *surrogate excess risk* $\mathcal{R}_\psi(g) - \mathcal{R}_\psi(g^*)$. In particular, denote by $d : \mathcal{H} \to \mathcal{Z}$ the decoding function $d(u) = \arg\min_{z\in\mathcal{Z}} \langle \varphi(z), Vu\rangle_{\mathcal{H}}$. In [11] it is proven that by construction, the QS estimator is *Fisher consistent*, i.e., $f^\star = d \circ g^*$, with $f^\star, g^*$ as above. Moreover, for any $g : \mathcal{X} \to \mathcal{H}$, the *comparison inequality* holds

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) \leq 2c_{V,\varphi}\sqrt{\mathcal{R}_\psi(g) - \mathcal{R}_\psi(g^*)}, \tag{6}$$

where $c_{V,\varphi} = \sup_{z\in\mathcal{Z}} \|V^*\varphi(z)\|_{\mathcal{H}}$. In the next paragraph we recall how to devise a suitable estimator of g^*.

The QS estimator depends only on L. Given a finite dataset $(x_i, y_i)_{i=1}^n$, an estimator $\widehat{g}$ for g^* can be found by considering the characterization of g^* in terms of Eq. (5). Indeed the problem in Eq. (5) can be solved using kernel ridge regression (KRR) [15]. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on $\mathcal{X}$ and $\mathcal{H}_\mathcal{X}$ the associated reproducing kernel Hilbert space (RKHS). Then given $\lambda > 0$, KRR reads

$$\widehat{g}_n \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}} + \lambda\|g\|^2_{\mathcal{G}}, \tag{7}$$

where $\mathcal{G}$ is the space of Hilbert-Schmidt operators from $\mathcal{H}_X$ to $\mathcal{H}$, which is isometric to $\mathcal{H} \otimes \mathcal{H}_\mathcal{X}$. The minimizer $\widehat{g}_n$ can be written in closed form as $\widehat{g}_n(\cdot) = \sum_{i=1}^n \alpha_i(\cdot)\psi(y_i) \in \mathcal{G}$ where $\alpha(x) = (\alpha_1(x), \ldots, \alpha_n(x)) \in \mathbb{R}^n$ is defined by

$$\alpha(x) = (K + n\lambda I)^{-1} K_x, \tag{8}$$

with $K_x = (k(x, x_1), \ldots, k(x, x_n)) \in \mathbb{R}^n$ and $K \in \mathbb{R}^{n\times n}$ is defined by $K_{ij} = k(x_i, x_j)$.

The key property here, is that due to the fact that $\widehat{g}_n$ is linear in the $\psi(y_i)$'s, then $\widehat{\ell}(z,x)$ does not explicitly depend on the surrogate space $\mathcal{H}$, indeed $\widehat{\ell}(z,x) = \langle \varphi(z), V\left(\sum_{i=1}^n \alpha_i(x)\psi(y_i)\right)\rangle_{\mathcal{H}} = \sum_{i=1}^n \alpha_i(x)L(z, y_i)$. Hence, the final estimator $\widehat{f}_n = d \circ \widehat{g}_n$ can be written as

$$\widehat{f}_n(x) = \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i=1}^n \alpha_i(x)L(z, y_i). \tag{9}$$

Finally, by combining the comparison inequality with results on the convergence of $\widehat{g}_n$ to g^* (see e.g. [15]), the following theorem holds.

Theorem 2.1 (Thm. 5 of [11])**.** *Let $n \in \mathbb{N}$, $\lambda_n = n^{-1/2}$ and $\tau > 0$. If L is SELF and $g^* \in \mathcal{G}$, then the following holds with probability at least $1 - 8e^{-\tau}$,*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq C\ c_{V,\varphi}\kappa\|g^*\|_{\mathcal{G}}\tau^2 n^{-1/4}. \tag{10}$$

where $\kappa^2 = \sup_x k(x,x)$ and C a universal constant.

Positioning of our contribution. From a theoretical viewpoint the result above holds for any loss on discrete and finite $\mathcal{Z}, \mathcal{Y}$, and shows a learning rate that is $\mathcal{O}(n^{-1/4})$. Moreover, from a practical viewpoint, to define and evaluate the QS estimator in Eq. (9) is enough to know only the loss L and a kernel k for $\mathcal{X}$ (no knowledge of $\mathcal{H}, \psi, \varphi$ is required). These considerations show that the QS framework could be a good candidate to systematically solve learning problems with discrete outputs.

However, note that constants of the bound depend on the specific SELF decomposition for L. If we use the one in [11], $\mathcal{H} = \mathbb{R}^{|\mathcal{Z}|}, \varphi(z) = e_z, \psi(y) = e_y, V = L$, then the constant $c_{V,\varphi}$ equals the spectral norm of the loss matrix $\|L\|$, which is exponentially large even for highly structured loss functions such as Hamming. In that case $\|L\| = 2^{m-1}$, where m is the number of labels (and a similar behaviour could affect $\|g^*\|_{\mathcal{G}}$). Then Eq. (10) can be totally uninformative if the constants of the rate are exponentially large [12].

In the next section, we prove that by using a suitable SELF-decomposition it is possible to find a version of Eq. (10), that depends only polynomially on the number of labels m. In particular we find the explicit constants for a number of widely used loss functions for ranking and multilabel learning. Finally we provide a refined generalization bound adaptive to the noise-level of the learning problem.

3. Main Results

In this section we study a specific SELF-decomposition for discrete losses, providing a generalization bound in the form of Eq. (10), with explicit constants depending on the specific loss chosen (Thm. 3.1). In Thm. 3.2 and Table 1 we quantify the constants for a number of widely used loss functions for multilabeling and ranking problems, showing that they are always polynomial with respect to the number of labels and in many cases optimal (Remark 3.3). Finally in Thm. 3.5 we generalize Eq. (10) (and so the learning rate in [11]), introducing a Tsybakov-like noise condition for the structured prediction problem.

3.1. Affine decomposition

Motivated by the limitations given by the possible exponential magnitude of the constants in the generalization bound in Eq. (10), we consider another SELF-decomposition of the loss, based on the following *affine decomposition* of the loss matrix,

$$L = FU^\top + c\mathbf{1}. \tag{11}$$

where $F \in \mathbb{R}^{|\mathcal{Z}|\times r}, U \in \mathbb{R}^{|\mathcal{Y}|\times r}$, $c \in \mathbb{R}$ 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 $r \in \mathbb{N}$. The minimum r for which there exists a decomposition as Eq. (11) is called the *affine dimension* of the loss L and is denoted $\operatorname{affdim}(L)$.

Note that the"centered" loss $L - c$ is SELF with

$$\mathcal{H} = \mathbb{R}^r,\ \varphi(z) = F_z,\ \psi(y) = U_y,\ V = I_{r\times r}, \tag{12}$$

where F_z is the z-th row of F and U_y the y-th row of U. Using the decomposition above, the following theorem gives a new version of the bound Eq. (10) specialized to discrete losses. Before giving the result, note that when we use the SELF-decomposition above for a loss, the conditional expectation g^* is characterized by $g^* : \mathcal{X} \to \mathbb{R}^r$, $g^*(x) = (g_j^*(x))_{j=1}^r$, for $g_j^* : \mathcal{X} \to \mathbb{R}$ defined as $g_j^*(x) = U^{j\top}\Pi(x)$, with $U^j \in \mathbb{R}^{|\mathcal{Y}|}$ the j-th column of U and $\Pi : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, $\Pi(x)_y = P(y|x)$ the conditional probability of y given x. In particular $g_j^*(x) \le \max_{k\in\mathcal{Y}} |U_{kj}|$, ($U_{kj}$ is the k, j-th element of U). Finally, $\mathcal{G}$ is isometric to $\mathcal{H}_{\mathcal{X}}^r$, since $\mathcal{H} = \mathbb{R}^r$.

Theorem 3.1 (Statistical complexity)**.** *Let $n \in \mathbb{N}, \tau > 0$ and $\lambda_n = n^{-1/2}$. Assume that the loss L decomposes as Eq.* (11)*. If $g^* \in \mathcal{G}$, we have that with probability $1 - 8e^{-\tau}$,*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \ \le\ \mathsf{A}Q\ C\kappa\tau^2\ n^{-1/4}, \tag{13}$$

where C, κ are as in Thm. 2.1,

$$\mathsf{A} = \sqrt{r}\|F\|_\infty U_{\max}, \tag{14}$$

$Q = \max_{1\le j\le r} \|g_j^*/U_{\max}\|_{\mathcal{H}_{\mathcal{X}}}$, $U_{\max} = \max_{j,k} |U_{kj}|$.

Proof. First, note that the excess risk $\mathcal{E}(f) - \mathcal{E}(f^*)$ is the same for L and for $L - c$ for any $c \in \mathbb{R}$, moreover both the definition of $f^\star$ and $\widehat{f}$ are invariant when $L - c$ is used instead of L. So we bound $\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*)$ with Eq. (10) applied to $L - c$, with c as in Eq. (11).

Applying Thm. 2.1 with the affine decomposition in Eq. (12) for $L - c$ and the definition of $c_{V,\varphi}$ by [11], we have that $c_{V,\varphi} := \sup_{z \in \mathcal{Z}} \|F_z\|_2 = \|F\|_\infty$ and $\|g^*\|_{\mathcal{G}}^2 = \sum_{j=1}^r \|g_j^*\|_{\mathcal{H}_\mathcal{X}}^2 \leq r \max_{1 \leq j \leq r} \|g_j^*\|_{\mathcal{H}_\mathcal{X}}^2$. The final result is obtained by multiplying and dividing by $U_{\max}$. □

The theorem above is essentially a version of Thm. 2.1 where we use the affine decomposition in Eq. (11) for the loss L, making explicit the dependence of the constants on structural properties of the loss, like the *affine dimension*.

In particular, we explicitly identify three distinct terms A, Q and $C\kappa\tau^2 n^{-1/4}$. The third term is completely explicit and does not dependent on the loss nor on the data distribution. It expresses the dependence of the statistical error with respect to the number of examples n and the high probability confidence τ (C is a universal constant and κ the constant of the kernel). The second term depends on the data distribution P and measures in a sense the "regularity" of the most difficult regression scalar function g_j^* defining the surrogate conditional expectation for the given loss. Note that the Q is renormalized by $U_{\max}$ so is invariant to the magnitude of the representation vector ψ.

Finally A depends only on the chosen loss and measures the cost of using the QS method as surrogate approach. In the next subsection we give sharp bounds on the constant A for many discrete losses used in practice, together with the computational complexity required to evaluate the QS estimator. In particular, we prove that, contrary to what suggested by [12], A depends only polynomially on the number of labels, making the QS method a good systematic approach to deal with discrete losses.

3.2. Sharp constants for multilabel and ranking losses

In this section (Thm. 3.2, Table 1) we characterize explicitly the constants introduced in Thm. 3.1, for a number of widely used losses for multi-labeling and ranking problems. In particular we show that they depend only polynomially on the number of labels (or equivalently polylog($|\mathcal{Y}|, |\mathcal{Z}|$)). Moreover in Remark 3.3 we show that the bounds obtained for many of the considered losses are sharp in a precise sense [16]. Finally we characterize the computational complexity of evaluating the QS estimator in Eq. (9) for such losses.

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 $[m] = \{1, \ldots, m\}$ 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 $\mathcal{Y} = \{1, \ldots, R\}^m = [R]^m$, the set of observed relevance scores for the m documents where R is the highest relevance [17]. 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 $\sigma \in \mathfrak{S}_m$, $j \in [m]$.

Theorem 3.2. *The constant* A *and the computational complexity of the QS estimator for the multilabel losses: 0-1, block 0-1, Hamming, Prec@k, F-score and ranking losses: NDCG-type, PD and MAP, appearing in Table 1 hold.*

Proof. We sketch here the analyses for the Hamming loss and the NDCG-type ranking measures. The complete analysis for all the losses is in Appendix B.

Hamming. Let $m \in \mathbb{N}$ be the number of labels. We represent each output element as a binary vector ($\mathcal{Z} = \mathcal{Y} = \{0,1\}^m$). We re-write the Hamming loss as

$$L(y', y) = \frac{1}{2m} - \frac{1}{2m} \sum_{j=1}^{m} s_j(y') s_j(y), \tag{15}$$

where $s_j(y) = 2[y]_j - 1$. Hence, this corresponds to an affine decomposition by setting

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

We have that $r = m, \|F\|_\infty = \frac{1}{2\sqrt{m}}$, $U_{\max} = 1$. This implies that $A = \frac{1}{2}$. Finally, inference corresponds to $\widehat{f}_j(x) = (\operatorname{sign}(\widehat{g}_j(x)) + 1)/2$ where $\widehat{g}_j(x) = \sum_{i=1}^n s_j(y_i)\alpha_i(x)$. This is done in $\mathcal{O}(m)$.

NDCG-type. [18, 17, 19] Let $\mathcal{Z} = \mathfrak{S}_m$ be the set of permutations of m elements and $\mathcal{Y} = [R]^m$ the set 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. The NDCG-type losses are defined as the normalized discounted sum of the gain of the relevance scores ordered by the predicted permutation:

$$L(\sigma, r) = 1 - \frac{1}{N(r)} \sum_{j=1}^m G([r]_j) D_{\sigma(j)}, \tag{16}$$

where $N(r) = \max_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^m G([r]_j) D_{\sigma(j)}$ is a normalizer.

Note that looking at Eq. (16) we can directly write that $r = m$ and

$$F_\sigma = -(D_{\sigma(j)})_{j=1}^m, U_r = \left(\frac{G([r]_j)}{N(r)}\right)_{j=1}^m, c = 1. \tag{17}$$

It follows that, $\|F\|_\infty = \|D\|_2, U_{\max} = D_{\max} G_{\max}$, so $A = \sqrt{m} G_{\max} D_{\max} (\sum_{j=1}^m D_j^2)^{1/2}$. For Table 1, assume $D_1 = 1$. If we define the vector $v \in \mathbb{R}^m$ as

$$v_j = \sum_{i=1}^n \frac{G([r_i]_j)\alpha_i(x)}{N(r_i)}, \quad 1 \le j \le m, \tag{18}$$

then inference corresponds to $f^*(x) = \operatorname{argsort}_{\sigma \in \mathfrak{S}_m}(v)$, which can be done in $\mathcal{O}(m \log m)$ operations. □

The key result in Table 1 is that the generalization properties and the computational complexity of the algorithm are both polynomial in the number of labels m (or equivalently polylog($|\mathcal{Y}|, |\mathcal{Z}|$)) for all considered losses except the 0-1, which does not provide any structural information of the observation/output spaces $\mathcal{Y}, \mathcal{Z}$. This theoretically explains why in discrete structured prediction and in particular multi-labeling and ranking, learning is possible even if the size of the output space is exponentially large compared to the number labels and, potentially, to the number of examples. Moreover this result shows that the Quadratic Surrogate is a valid candidate for systematically addressing learning problems with discrete losses both from a statistical and from a computational viewpoint (in contrast with what conjectured in [12]).

Remark 3.3 (On the sharpness of the QS estimator)**.** *It is natural to ask to what extent the statistical rates provided by Thm. 3.1 can be considered representative of the statistical difficulty of solving the problem in Eq.* (2)*. Of course, formally answering this question necessarily requires a study of the corresponding minimax rates under certain priors. In particular, one would be interested in studying the dependence of those rates both in the number of samples and the size of the output space $\mathcal{Z}$.*

Although far from answering this question, we can provide a weaker notion of optimality on the framework of surrogate-based methods. In particular, by using the results in [16], we prove that cannot exist a consistent convex surrogate that maps the discrete problem in a vector valued problem of lower dimension than r (the one used by the QS estimator through the affine-decomposition) for the following losses: 0-1, block 0-1, Hamming, Prec@k, NDCG, PD and MAP (see Appendix B).

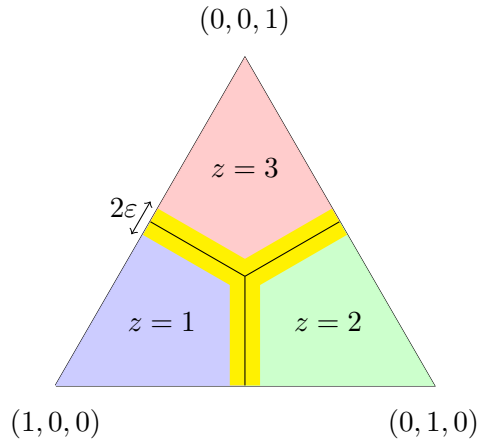

Figure 1: Generalized Tsybakov condition for discrete losses, Eq. (20), in the case of multi-class. See Example 3.4 for more details.

More in detail, for the Hamming loss we obtain that the statistical complexity of the problem is independent of the number of labels. Intuitively this is explained by the fact that the QS estimator corresponds to estimating the m marginals independently. Our result is to be compared with [12], where they obtain a constant in the order of $\mathcal{O}(m^2)$. For Prec@k, we obtain $\mathsf{A} = \sqrt{\frac{m}{k}}$, which is coherent with the intuition that the problem becomes more challenging when k is fixed and m increases. For the F-score the computational bound of the resulting algorithm is essentially in [20]. For the NDCG-type losses, $G : \mathbb{R} \to \mathbb{R}$, the *gain* is an increasing function and $D = (D_j)_{j=1}^m \in [0,1]^m$, the *discount*, is a coordinate-wise decreasing vector. For this family of losses A depends crucially on the discount factor D_j, tending to $\sqrt{m}$ (the constant of Prec@1) for fast decaying D_j and to m for low decaying ones. For PD and MAP, estimating the surrogate function is statistically tractable, but both inference algorithms are NP-Hard (Minimum Weight Feedback Arcset problem (MWFAS) for PD and an instance of Quadratic Assignment Problem (QAP) for MAP), as was already noted in [21].

3.3. Improved rates under low-noise assumption

Intuitively, if there is small noise at the decision boundary between different labels, then it should be statistically easier to discriminate between them. To formalize this intuition, we define the margin $\gamma(x)$ as

$$\gamma(x) = \min_{z' \neq f^*(x)} \ell(z', x) - \ell(f^*(x), x). \tag{19}$$

The margin function γ measures the minimum suboptimality gap in terms of the Bayes risk. If for a given x the margin is small, then its cost at the optimum is very close to the cost at a suboptimal label. We will say that the *p-noise condition* is satisfied if

$$P_{\mathcal{X}}(\gamma(X) \leq \varepsilon) = o(\varepsilon^p), \tag{20}$$

where $P_{\mathcal{X}}$ is the marginal of P over $\mathcal{X}$, with $p \geq 0$. The parameter p characterizes how fast the noise vanishes at the boundary and corresponds to no assumption when $p = 0$.

Note that Eq. (20) is a generalization of the Tsybakov condition for binary classification [13] and of the condition in [22] for multi-class classification, to general discrete losses. Indeed, for the binary 0-1 loss ($\mathcal{Y} = \{-1, 1\}$), $\gamma(x) = |\mathbb{E}[Y|x]|$, so we recover the classical Tsybakov condition.

Example 3.4 (Generalized Tsybakov for multiclass)**.** *For every $P(y|x)$ in the simplex, one associates the corresponding optimal label as $z^* = \arg\min_z \ell(z, x)$. Fig. 1 represents the partition of the simplex*

corresponding to the 0-1 loss for $\mathcal{Z} = \mathcal{Y} = \{1, 2, 3\}$. *In this case,* $\gamma(x)$ *corresponds to the distance to the boundary decision depicted in Fig. 1 and so* $\{P(y|x) \mid \gamma(x) < \varepsilon\}$ *corresponds to the yellow area. Eq.* (20) *says that the probability of falling in that region vanishes as* $o(\varepsilon^p)$.

In the next theorem we improve the comparison inequality of Eq. (6) to take into account the generalized Tsybakov condition for discrete losses of Eq. (20).

Theorem 3.5 (Improved comparison inequality)**.** *Assume* $\mathcal{Y}, \mathcal{Z}$ *to be finite and* γ *to satisfy Eq.* (20) *for* $p > 0$. *Then the following holds*

1. $1/\gamma \in L_p(P_{\mathcal{X}})$.
2. *Assume a decomposition as in Eq.* (11) *for the loss* L. *Then, for any bounded measurable* $g : \mathcal{X} \to \mathcal{H}$,

$$\mathcal{E}(f) - \mathcal{E}(f^\star) \leq q\gamma_p^{\frac{1}{p+2}} \left(\mathcal{R}_\psi(g) - \mathcal{R}_\psi(g^*)\right)^{\frac{p+1}{p+2}}, \tag{21}$$

where $\gamma_p = \|1/\gamma\|_{L_p(P_{\mathcal{X}})}$, $q = (16\|F\|_\infty^2)^{\frac{p+1}{p+2}}$.

The proof of the first part can be found in Lemma A.7, while the second part is Cor. A.11, both in the Appendix A. As you can note, the comparison inequality of Eq. (6) is recovered when $p = 0$ (i.e. when the generalized Tsybakov condition is always verified), while an exponent close to 1, instead of $1/2$ is obtained when $p \gg 0$. Finally, by using the improved comparison inequality we refine the rates for the QS estimator in Thm. 3.1.

Corollary 3.6 (Improved rates)**.** *Under the p-noise condition, we have the following improvement on the generalization bound in Eq.* (13)*,*

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^\star) \leq C\gamma_p^{\frac{1}{p+2}} \left(\mathsf{A}^2 Q^2 \kappa^2 \tau^4 n^{-\frac{1}{2}}\right)^{\frac{p+1}{p+2}}, \tag{22}$$

with C *universal constant and* A, Q, κ *as in Thm. 3.1.*

Note that the result of Thm. 3.1 is recovered for $p = 0$ (always verified), while we obtain a learning rate essentially in the order of $n^{-1/2}$, instead of $n^{-1/4}$, in conditions of low-noise (i.e. $p \gg 0$).

4. Computational Considerations

As already observed in [11]: (1) the computation of the QS estimator (Eq. (9)) is divided in *training step* and *inference step* (or *evaluation step*), (2) the SELF-decomposition of the loss is not needed to run the algorithm, but only to derive the theoretical guarantees. Here we show how the explicit knowledge of the affine decomposition of the loss can be useful to improve also the computational complexity of the method (its theoretical implications have been studied in Sec. 3.1). First we recall the training and test steps.

Training. The training step requires only to have a kernel function k over $\mathcal{X}$ and to have access to the training input examples $(x_i)_{i=1}^n$. It consists essentially in computing the inverse of the kernel matrix necessary for the second step, i.e. $W = (K + \lambda n I)^{-1}$, with K defined in Eq. (8).

Evaluation. The evaluation step requires only the knowledge of the loss L and to have access to the train observations $(y_i)_{i=1}^n$. Given a test input point $x \in \mathcal{X}$, it consists in: first, computing the coefficients $(\alpha_i(x))_{i=1}^n$ according to Eq. (8), i.e. $\alpha(x) = WK_x$, with the notation in Eq. (8); second predicting the output $z \in \mathcal{Z}$ associated to the test input x, by solving Eq. (9).

Multilabel		bibtex	birds	CAL500	corel5k	enron	mediamill	medical	scene	yeast
	n	7395	645	502	5000	1702	43907	978	2407	2417
	d	1836	260	68	499	1001	120	1449	294	103
	m	159	19	174	374	53	101	45	6	14
0-1 (↓)	THBM	0.82	0.57	1.0	0.99	0.92	0.93	0.31	0.49	0.93
	SSVM	0.91	0.53	1.0	0.99	0.90	1.0	0.35	0.51	0.95
	QS	**0.78**	**0.52**	1.0	**0.95**	**0.86**	**0.86**	**0.29**	**0.34**	**0.76**
Ham (↓)	THBM	1.3e-2	7.9e-2	0.14	1.1e-2	**5.9e-2**	**3.1e-2**	**9.4e-3**	0.11	**0.26**
	SSVM	1.3e-2	6.4e-2	0.13	1.0e-2	7.1e-2	8.7e-2	1.07e-2	0.11	0.40
	QS	1.3e-2	**4.9e-2**	**0.14**	**9.4e-3**	8.6-2	**3.1e-2**	9.6e-3	0.11	0.42
F-score (↑)	THBM	0.44	0.25	0.46	0.25	0.51	**0.56**	0.80	0.63	**0.48**
	SSVM	0.19	0.16	0.33	0.11	0.49	0.40	0.74	0.57	**0.48**
	QS	**0.47**	**0.28**	**0.47**	**0.26**	**0.52**	**0.56**	**0.83**	**0.68**	0.47

Ranking	Ohsumed	
	n	106
	d	25
	m	150
NDCG@3 (↑)	SSVM	0.47
	QS	**0.51**
NDCG@5 (↑)	SSVM	0.45
	QS	**0.48**
NDCG@10 (↑)	SSVM	0.43
	QS	**0.46**

Table 2: Numerical results on real-world multilabeling and ranking datasets comparing our QS estimator, THBM [23] and SSVM [9]. n is the size of the full dataset, d the dimensionality of the data and m the number of classes (multilabel), or the avg. number of query-document pairs (ranking). See Sec. 5 for more details.

4.1. Using the affine decomposition to speed up the QS estimator

Note that, to run the algorithm described above, only the loss L and kernel k are needed. This makes the QS-method (1) systematically applicable to any supervised learning problem with discrete loss, since it does not require to devise a specific surrogate for each loss (2) theoretically grounded with basic guarantees from [11] in terms of consistency and learning rates. Indeed note that the SELF-decomposition in terms of $\mathcal{H}, \varphi, \psi$ for the loss and in particular the affine decomposition of Eq. (11) is needed only to prove the sharper guarantees in Thms. 3.1, 3.2 and 3.5 and Cor. 3.6.

However it is possible to additionally exploit the affine decomposition to have even a computational benefit for the presented algorithm, as we are going to show in the rest of the section.

Improved training when U is known. When we know the affine decomposition of the loss, we have $\mathcal{H} = \mathbb{R}^r$ and $\psi(y) = U_y$, so we can compute explicitly the solution to Eq. (7) [15], $\widehat{g}_n : \mathcal{X} \to \mathbb{R}^r$ that is $\widehat{g}_n(x) = \sum_{i=1}^n k(x, x_i) C_i$, where $C_i \in \mathbb{R}^r$ is the i-th row of $C \in \mathbb{R}^{n \times r}$, the solution of the linear system

$$(K + \lambda n I) C = \Psi^\top,$$

with $\Psi = (\psi(y_1), \dots, \psi(y_n)) \in \mathbb{R}^{r \times n}$. This is the same as solving r scalar KRR problems independently and its computation can be efficiently reduced from essentially $\mathcal{O}(n^3 r)$ to $\mathcal{O}(n\sqrt{n}r)$ via suitable random projection techniques [24, 25, 26].

Improved evaluation when F is known. Given a test point $x \in \mathcal{X}$, first we evaluate $\theta := \widehat{g}_n(x) \in \mathbb{R}^r$, requiring essentially $\mathcal{O}(nr)$ (up to $\mathcal{O}(\sqrt{n}r)$ by using random projection techniques [24, 25, 26]). Then we use the characterization of $\widehat{f}_n(x) = (d \circ \widehat{g}_n)(x) = d \circ \theta$, to obtain the equivalent problem

$$\min_{z \in \mathcal{Z}} \; F_z \cdot \theta, \tag{23}$$

where $F_z \in \mathbb{R}^r$ is the z-th row of F (see Eqs. (11) and (12)) and $(\cdot)$ the dot-product. The computational complexity of Eq. (23) (we denote it by $\mathrm{INF}_F(|\mathcal{Z}|)$) has been devised for a number of widely used losses in Thm. 3.2, Table 1 (see Appendix B for the proofs).

5. Numerical Experiments

We perform numerical experiments for the QS-estimator on multilabeling (9 datasets [27]) and ranking problems (1 dataset [28]), see Table 2. We use three evaluation measures for multilabel, namely, 0-1, Hamming and F-score, and NDCG@k for ranking (in the NDCG-type family [18]), which have been

theoretically analysed in Sec. 3 and Table 1. All experiments are performed using 60% of the dataset for training, 20% for validation and 20% for testing. We compare the performance of the QS-estimator with a threshold-based method, which we denote by THBM [23], and the Structural SVM [9] (SSVM). THBM is a common method for multilabelling where learning is done in two stages. The method first estimates the m marginals $\widehat{g}_j(\cdot)$ and then learns the best threshold function $\widehat{t}(\cdot)$ minimizing via least-squares the measure of interest. The inference is performed via thresholding the estimated marginals by $\widehat{t}(x)$ (see Sec. 2 in [23]). The SSVM corresponds to the multilabel-SVM [29], which is an instance of the SSVM with unary potentials that optimizes the Hamming loss. Note that we have used the same multilabel-SVM for all multilabel losses; for the F-score, there is no principled way of optimizing the measure with SSVMs. The experimental results in Table 2 show that the QS-estimator outperforms the other methods for 0-1 loss and F-score. Indeed, the method depends on the loss and is designed to be consistent with it. THBM achieves approximatively the same accuracy for Hamming as it is based on estimating the marginals, while the SSVM is proven to be inconsistent even in this case [30], as the experimental result empirically shows. For the ranking experiment, we have used the SSVM from [9] called RankSVM as baseline to compare with the QS-estimator. The algorithm corresponding to the QS-estimator for NDCG, which corresponds to the one in [17] for this measure, outperforms the SSVM. This highlights the importance of consistency in learning, and the importance of making the algorithm dependent on the measure willing to use for evaluation.

6. Related Works & Discussion

While the QS for structured prediction generalizes the QS for binary classification, Structural SVMs (SSVMs) [8, 2] and Conditional Random Fields (CRFs) [5, 6, 31] generalize the binary SVM and logistic regression to the structured case. All of them are surrogate methods based on minimizing the expected risk of a certain surrogate loss $S(v, y) : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$ in a convex surrogate space $\mathcal{C}$. The corresponding surrogates are $S_{\text{QS}}(v, y) = \|v - U_y\|^2_{\mathbb{R}^r}$, $S_{\text{SSVM}}(v, y) = \max_{y' \in \mathcal{Y}}(v_{y'} + L(y', y)) - v_y$ and $S_{\text{CRF}}(v, y) = \log(\sum_{y' \in \mathcal{Y}} \exp v_{y'}) - v_y$ (See Examples in Appendix A.1) for QS, SSVM and CRF, respectively. SSVMs and CRFs exploit the structure of the problem by decomposing each output element into cliques and considering only the features on this parts. This is necessary for the tractability of the methods. Moreover, for SSVMs, the loss L must decompose into these cliques to make possible the maximization inside the surrogate, often called augmented inference. The clique decomposability of the loss, can be seen as a low rank decomposition, analogous to our SELF-decomposition. While the QS has attractive statistical properties, it is generally not the case for the other surrogate methods. CRFs are only consistent for the 0-1 loss in the case that the model is well-specified [31]. This lack of calibration to a given loss is an important drawback of this method [7]. SSVMs are in general not Fisher consistent, even for the 0-1 loss, for which is only consistent if the problem is deterministic, i.e, there always exists a majority label y with probability larger than 1/2 [32].

QS for structured prediction. [21] proposed the QS through an affine decomposition of the loss and derived Fisher consistency of the corresponding surrogate method. They analyzed the inference algorithms for Prec@k, ERU (NDCG-type measure that we study in Appendix B), PD and MAP. As Fisher consistency is a property only at the optimum, their analysis is not able to provide any statistical guarantees. [17] analyses consistency and calibration properties for the QS specialized for NDCG-type losses. In particular, they highlight the fact that estimating the normalized relevance scores is key to be consistent, which is a property that follows directly from our framework.

As far as we know, [12] is the only work that addresses the learning complexity of general discrete losses for structured prediction. They consider a different QS surrogate than ours, which could be potentially intractable to compute since it is defined on the space of labels (even when the loss is low-rank) $\mathbb{E}\,\|Fg(X) - L(\cdot, Y)\|^2_{\mathbb{R}^{|\mathcal{Z}|}}$, and not in the low dimensional space of the decomposition $\mathbb{E}\,\|g(X) - U_Y\|^2_{\mathbb{R}^r}$. They also obtain rates of the form $\propto \mathsf{A} n^{-1/4}$, however, their constants are always

larger than ours and computed explicitly only for a small number of loss functions. In particular, for A, they obtain $\mathcal{O}(2^m), \mathcal{O}(b), \mathcal{O}(m^2)$, while we obtain $\mathcal{O}(2^{m/2}), \mathcal{O}(\sqrt{b}), \mathcal{O}(1)$ for the 0-1, block 0-1 and Hamming, respectively. In addition, our constants are interpretable and most of them can be proven to be optimal (in the sense explained in Remark 3.3). Finally we provide a refined bound adaptive to the noise of the problem as in Cor. 3.6.

To conclude, [16] introduces and studies the concept of convex calibration dimension. We use their lower bound on this quantity to study the optimality of the dimension of the QS as reported in Remark 3.3.

Acknowledgements

This work was supported by the European Research Council (project Sequoia 724063).

References

[1] Gökhan BakIr, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S.V.N. Vishwanathan. *Predicting Structured Data*. MIT press, 2007.

[2] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2(Dec):265–292, 2001.

[3] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, 2011.

[4] 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.

[5] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[6] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

[7] Maksims N. Volkovs, Hugo Larochelle, and Richard S. Zemel. Loss-sensitive training of probabilistic conditional random fields. *arXiv preprint arXiv:1107.1805*, 2011.

[8] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 104. ACM, 2004.

[9] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM, 2006.

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

[11] 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.

[12] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pages 302–313, 2017.

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

[14] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[15] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

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

[17] 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.

[18] Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. Learning to rank by optimizing ndcg measure. In *Advances in Neural Information Processing Systems*, pages 1883–1891, 2009.

[19] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on Learning Theory*, pages 25–54, 2013.

[20] Willem Waegeman, Krzysztof Dembczyński, Arkadiusz Jachnik, Weiwei Cheng, and Eyke Hüllermeier. On the bayes-optimality of f-measure maximizers. *The Journal of Machine Learning Research*, 15(1):3333–3388, 2014.

[21] Harish G. Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2013.

[22] 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.

[23] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[24] Alex J. Smola and Bernhard Schökopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 911–918, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[25] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[26] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901, 2017.

[27] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(Jul):2411–2414, 2011.

[28] William Hersh, Chris Buckley, T.J. Leone, and David Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR94*, pages 192–201. Springer, 1994.

[29] Thomas Finley and Thorsten Joachims. Training structural svms when exact inference is intractable. In *Proceedings of the 25th International Conference on Machine Learning*, pages 304–311. ACM, 2008.

[30] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 341–358, 2011.

[31] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

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

[33] 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.

[34] 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.

Sharp Analysis of Learning with Discrete Losses

Supplementary Material

A. Calibration and fast rates for surrogate methods

The goal of Appendix A 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 [33], which was done for binary classification under 0-1 loss, to the case of general discrete losses.

In Appendix A.1, we introduce the basic quantities used for the analysis of surrogate methods. Then, in Appendix A.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 [11]. Finally, in Appendix A.3, we derive our main result, which generalizes the Tsybakov condition, existing for multiclass and binary [22, 13] classification.

A.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,

$$f^*(x) = \operatorname*{arg\,min}_{z \in \mathcal{Z}} \ell(z, x), \quad \mathcal{E}(f) = \mathbb{E}\, \ell(f(X), X), \tag{24}$$

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,

$$g^*(x) = \operatorname*{arg\,min}_{v \in \mathcal{C}} W(v, x), \quad W(v, x) = \int S(v, Y) dP(Y|x) \quad \mathcal{R}(g) = \mathbb{E}\, W(g(X), X), \tag{25}$$

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 A.1 (Surrogate elements for the QS)**.** *In the case of the QS, we have that ,*

$$S(v, y) = \|v - U_y\|^2_{\mathbb{R}^r}, \quad \mathcal{C} = \mathbb{R}^r, \quad d(v) = \arg\min_{z \in \mathcal{Z}} F_z \cdot v, \tag{26}$$

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

$$W(\widehat{g}(x), x) - W(g^*(x), x) = \|\widehat{g}(x) - g^*(x)\|_2^2. \tag{27}$$

Moreover, it is Fisher consistent by construction ([11]).

Example A.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*

$$\mathcal{C} = \mathbb{R}^r, \quad d(v) = \arg\max_{z \in \mathcal{Z}} \sum_{c \in C} v_{z_c}, \tag{28}$$

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),

$$S(v, y) = \log \left(\sum_{y' \in \mathcal{Y}} \exp \left(\sum_{c \in C} v_{y'_c} \right) \right) - \sum_{c \in C} v_{y_c}. \tag{29}$$

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

$$S(v, y) = \max_{y' \in \mathcal{Y}} \left\{ \sum_{c \in C} \left(\{ L(y_c, y'_c) + v_{y'_c} \} \right) \right\} - \sum_{c \in C} v_{y_c}. \tag{30}$$

A.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 A.3.

Definition A.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,*

$$H_{L,S}(\ell(d \circ g(x), x) - \ell(f^*(x), x)) \leq W(g(x), x) - W(g^*(x), x), \tag{31}$$

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 A.4 (Calibration function for QS [11]). *Assumption 1 holds for the QS with*

$$H_{L,S}(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2} \tag{32}$$

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

$$\ell(\widehat{f}(x), x) - \ell(f^*(x), x) = \{\ell(\widehat{f}(x), x) - \widehat{\ell}(\widehat{f}(x), x)\} + \{\widehat{\ell}(\widehat{f}(x), x) - \ell(f^*(x), x)\} = A + B. \tag{33}$$

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,

$$(\ell(f(x), x) - \ell(f^*(x), x))^2 \leq 2 \sup_{z\in\mathcal{Z}} (\widehat{\ell}(z, x) - \ell(z, x))^2 = 4\|F\|_\infty^2 \|\widehat{g}(x) - g^*(x)\|_2^2. \tag{34}$$

Re-arranging and using Eq. (27) gives the final result. □

The following important Theorem shows how Eq. (32) translates into a relation between excess risks, which are the quantities that we are ultimately interested at.

Theorem A.5 (From Bayes risks to risks). *Suppose Assumption 1 holds. Then,*

$$H_{L,S}(\mathcal{E}(f) - \mathcal{E}(f^*)) \leq \mathcal{R}(g) - \mathcal{R}(g^*) \tag{35}$$

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. A.5 with Lemma A.4, we obtain the comparison inequality for the QS.

Corollary A.6 (Comparison inequality for QS [11]). *For the QS, we have that*

$$\mathcal{E}(f) - \mathcal{E}(f^*) \leq 2\|F\|_\infty \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)} \tag{36}$$

A.3. Improved calibration under low noise

Thm. A.5 gives the ability to translate learning rates of the surrogate to learning rates of the full risk. However, as we will show, Eq. (35) 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. A.5 under the p-noise assumption. We recall that the p-noise condition states that

$$P_{\mathcal{X}}(\gamma(X) < \varepsilon) = o(\varepsilon^p), \tag{37}$$

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 A.7.

Lemma A.7. *If the p-noise condition holds, then* $1/\gamma \in L_p(P_{\mathcal{X}})$.

Proof.

$$\|1/\gamma\|^p_{L_p(P_X)} = \mathbb{E}\, 1/\gamma(X)^p = \int_0^\infty pt^{p-1} P_X(1/\gamma(X) > t)dt = \int_0^\infty pt^{p-1} P_X(\gamma(X) < t^{-1})dt. \tag{38}$$

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 A.8, which bounds the probability of error by a power of the excess risk, is a generalization of the Tsybakov Lemma [13, Prop.1] for general discrete losses.

Lemma A.8 (Bounding the size of the error set)**.** *If $1/\gamma \in L_p(P_X)$, then*

$$P_X(X_f) \leq \gamma_p^{\frac{1}{p+1}} (\mathcal{E}(f) - \mathcal{E}(f^*))^{\frac{p}{p+1}} \tag{39}$$

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

$$1(f(x) \neq f^*(x)) \leq 1/\gamma(x) \Delta\ell(f(x), x) \tag{40}$$

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. A.10, we will need the following useful Lemma A.9 of convex functions.

Lemma A.9 (Property of convex functions)**.** *Suppose $h : \mathbb{R} \to \mathbb{R}$ is convex and $h(0) = 0$. Then, for all $x > 0$, $0 \leq y \leq x$,*

$$h(y) \leq \frac{y}{x} h(x) \quad \textit{and} \quad h(x)/x \textit{ is increasing on } (0, \infty). \tag{41}$$

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. A.10, is an adaptation of Thm. 10 of [33], which was specific for binary 0-1 loss, now adapted to the case of general discrete losses.

Theorem A.10 (Improved Calibration)**.** *Suppose that the surrogate S is calibrated with calibration function $H_{L,S}$ (see Eq. (31)) and the p-noise condition holds. Then, we have that*

$$H_{L,S,p}(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)) \leq \mathcal{R}(g) - \mathcal{R}(g^*), \tag{42}$$

where

$$H_{L,S,p}(\varepsilon) = (\gamma_p \varepsilon^p)^{\frac{1}{p+1}} H_{L,S}\left(\frac{1}{2}(\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}\right). \tag{43}$$

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. A.10). To ease notation let's denote the excess Bayes risk by $\Delta\ell(z', x) = \ell(z', x) - \ell(f^*(x), x)$.

The intuition of the proof is to split the Bayes excess 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. (31).

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

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

$$A \leq tP_X(X_f) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}}, \tag{44}$$

where in the last inequality we have used Lemma A.8.

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

We have that

$$\Delta\ell(f(x), x)1(\Delta\ell(f(x), x) \geq t) \leq \frac{t}{H_{L,S}(t)} H_{L,S}(\Delta\ell(f(x), x)) \tag{45}$$

In the case $\Delta\ell(f(x), x) < t$, inequality in Eq. (45) follows from the fact that $H_{L,S}$ is nonnegative. For the case $\Delta\ell(f(x), x) > t$, apply Lemma A.9 with $h = H_{L,S}, x = \Delta\ell(f(x), x)$ and $y = t$.

From Eq. (32), we have that $\mathbb{E}\{1(X_f)H_{L,S}(\Delta\ell(f(X), X))\} \leq \mathcal{R}(g) - \mathcal{R}(g^*)$. Hence,

$$B \leq \frac{t}{H_{L,S}(t)}(\mathcal{R}(g) - \mathcal{R}(g^*)) \tag{46}$$

Putting everything together,

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}} + \frac{t}{H_{L,S}(t)}(\mathcal{R}(g) - \mathcal{R}(g^*)), \tag{47}$$

and hence,

$$\left(\frac{\mathcal{E}(d \circ g) - \mathcal{E}(f^*)}{t} - \gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}}\right) H_{L,S}(t) \leq \mathcal{R}(g) - \mathcal{R}(g^*). \tag{48}$$

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. (43). The second part of the Theorem follows because $\frac{H_{L,S}(t)}{t}$ is non-decreasing by Lemma A.9. □

Finally, if we apply Thm. A.10 to the QS, we get the desired result as Cor. A.11.

Corollary A.11 (Improved comparison inequality for QS)**.** *For the QS, we have that*

$$\mathcal{E}(f) - \mathcal{E}(f^*) \leq \gamma_p^{\frac{1}{p+2}} \left(16\|F\|_\infty^2 (\mathcal{R}(g) - \mathcal{R}^*)\right)^{\frac{p+1}{p+2}}. \tag{49}$$

Proof. Substituting $H_{L,S}(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2}$ in Eq. (43), gives that,

$$H_{L,S,p} = \frac{\varepsilon^{\frac{p+2}{p+1}}}{\gamma_p^{\frac{1}{p+1}} 16\|F\|_\infty^2}. \tag{50}$$

Reversing the relation gives the comparison inequality in Eq. (49). □

B. Multilabel and ranking losses

The goal of this section is to derive all of the constants from Table 1.

In Appendix B.1, we recall the elements that we need in order to derive the constants. In Appendix B.2, we introduce the main tool from [16] that we use in order to study the optimality of the QS. Finally, the main bulk is in Appendix B.3, where we analyse each loss separately.

B.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

$$\widehat{f}_n(x) = \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{n} \alpha_i(x) L(z, y_i). \tag{51}$$

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

$$L = FU^\top + c\mathbf{1}. \tag{52}$$

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. (52) holds.

Recall that the quantity of interest for the statistical complexity is

$$\mathsf{A} = \sqrt{r} \|F\|_\infty U_{\max}. \tag{53}$$

The inference complexity corresponds to the computational complexity of solving Eq. (51).

B.2. On the optimality of the QS.

We use results from [16] in order to study the optimality of the dimension of the QS as commented in Remark 3.3. We implicitly use the concept of *convex calibration dimension of a loss L* (see Def. 10 in [16]), which is defined as the minimum dimension over all consistent convex surrogates w.r.t L. In the following Thm. B.1 (their Thm. 18), they provide a sufficient condition to lower bound this dimension.

Theorem B.1. *(In [16]) 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. B.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. B.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 [4] despite the loss matrix being full rank.

B.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 [17]. 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

$$L(z, y) = 1(z \neq y). \tag{54}$$

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

$$F_z = -(1_{[z=z']})_{z' \in \{0,1\}^m},\ U_y = (1_{[y=y']})_{y' \in \{0,1\}^m},\ c = 1. \tag{55}$$

We have that

$$r = 2^m,\ \|F\|_\infty = 1,\ U_{\max} = 1. \tag{56}$$

Hence,

$$\mathsf{A} = 2^{m/2}. \tag{57}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \arg\max_{z \in \mathcal{P}_m} \sum_{i | y_i = z} \alpha_i(x), \tag{58}$$

which can be done in

$$\mathcal{O}(2^m \wedge n). \tag{59}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. B.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

$$L(z, y) = 1(z \in B_j, y \notin B_j, \text{for some } j \in [b]). \tag{60}$$

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

$$F_z = -(1_{[z \in B_j]})_{j=1}^b,\ U_y = (1_{[y \in B_j]})_{j=1}^b,\ c = 1. \tag{61}$$

We have that

$$r = b,\ \|F\|_\infty = 1,\ U_{\max} = 1. \tag{62}$$

Hence,

$$\mathsf{A} = \sqrt{b}. \tag{63}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \arg\max_{1 \leq j \leq b} \sum_{i | y_i \in B_j} \alpha_i(x), \tag{64}$$

which can be done in

$$\mathcal{O}(b) \tag{65}$$

- **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. B.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

$$L(z,y) = \frac{1}{m}\sum_{j=1}^{m} 1([z]_j \neq [y]_j). \tag{66}$$

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

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

  This implies that

$$F_z = -\frac{1}{2m}(s_j(z))_{j=1}^m,\ U_y = (s_j(y))_{j=1}^m,\ c = \frac{1}{2m}. \tag{67}$$

  We have that

$$\|F\|_\infty = \frac{1}{2\sqrt{m}},\ U_{\max} = 1. \tag{68}$$

  Hence,

$$\mathsf{A} = \frac{1}{2}. \tag{69}$$

- **Inference.** Inference corresponds to

$$\widehat{f}_j(x) = \left(\frac{\operatorname{sign}\left(\widehat{g}_j(x)\right) + 1}{2}\right), \quad \text{where} \quad \widehat{g}_j(x) = \sum_{i=1}^{n} s_j(y_i)\alpha_i(x), \tag{70}$$

  which can be done in

$$\mathcal{O}(m). \tag{71}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. B.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 $\mathcal{Y} = \mathcal{P}_m$.

$$L(z,y) = 1 - \frac{|y \cap z|}{k} = 1 - \frac{1}{k}\sum_{j=1}^{m}[z]_j[y]_j. \tag{72}$$

- **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,

$$\mathsf{A} = \sqrt{\frac{m}{k}}. \tag{73}$$

- **Inference.** Inference corresponds to

$$\widehat{f}(x) \in \underset{z \in \mathcal{P}_{m,k}}{\arg\operatorname{top}_k}\left(\left(\sum_{i|[y_i]_j=1}\alpha_i(x)\right)_{j=1}^{m}\right), \tag{74}$$

  which can be done in

$$\mathcal{O}(m\log k). \tag{75}$$

- **Optimality of r.** Taking $\Pi_y = 1/2^m$ for every $y \in \mathcal{Y}$ and applying Thm. B.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

$$L(z, y) = 1 - 2\frac{|z \cap y|}{|z| + |y|}, \tag{76}$$

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

$$2\frac{|z \cap y|}{|z| + |y|} = \begin{cases} 2\sum_{j=1}^{m}\sum_{\ell=0}^{m} \frac{[z]_j}{\ell+|z|} 1([y]_j = 1, |y| = \ell) & y \neq 0 \\ 1(z = 0) & y = 0 \end{cases}. \tag{77}$$

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

$$P_{j\ell}(x) = P([Y]_j = 1, |Y| = \ell | X = x), \quad p_0(x) = P(Y = 0|X = x). \tag{78}$$

Then, the Bayes risk reads

$$\ell(z, x) = \begin{cases} 2\sum_{j=1}^{m}\sum_{\ell=0}^{m} \frac{[z]_j}{\ell+|z|} P_{j\ell}(x) & y \neq 0 \\ p_0(x) & y = 0 \end{cases}. \tag{79}$$

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 B.2.

Lemma B.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 $A(x) \in \mathbb{R}^{m\times m}$:*

$$A_{jk}(x) = \sum_{\ell=0}^{m} \frac{P_{j\ell}(x)}{\ell + k} \tag{80}$$

   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:

$$f^{(k)}(x) = \arg\max_{z\in\mathcal{P}_{m,k}} A_{\cdot,k}^T(x) z, \quad \text{for } k = 1, \ldots, m. \tag{81}$$

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. (80). As the estimator is the same, the statistical complexity is the minimum of both.

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

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

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

$$r = m^2 + 1,\ \frac{1}{2} \le \|F\|_\infty \le 1,\ U_{\max} = 1. \tag{82}$$

Hence,

$$\mathsf{A}_1 \le \sqrt{m^2 + 1} \le \sqrt{2}m \tag{83}$$

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

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

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

$$r = m^2 + 1,\ \|F\|_\infty = \sqrt{m},\ U_{\max} = 1. \tag{84}$$

Hence,

$$\mathsf{A}_2 = \sqrt{m(m^2 + 1)} \le m\sqrt{2m}. \tag{85}$$

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

$$\mathsf{A} \le \sqrt{2}m. \tag{86}$$

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

$$\widehat{P}_{j\ell}(x) = \sum_{i|[y_i]_j=1,|y_i|=\ell} \alpha_i(x), \quad \widehat{A}_{jk}(x) = \sum_{\ell=0}^{m} \frac{\widehat{P}_{j\ell}(x)}{\ell + k},\ \widehat{p}_0(x) = \sum_{i|y_i=0} \alpha_i(x). \tag{87}$$

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

$$\mathcal{O}(m^3). \tag{88}$$

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

$$\mathcal{O}(m^2). \tag{89}$$

- **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 [16] 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:

$$L(\sigma, r) = 1 - \frac{1}{N(r)} \sum_{j=1}^{m} G([r]_j) D_{\sigma(j)} \tag{90}$$

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. (90) 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,

$$\|F\|_\infty = \sqrt{\sum_{j=1}^{m} D_j^2}, \quad U_{\max} = G_{\max} D_{\max}, \tag{91}$$

  hence,

$$\mathsf{A} = \sqrt{m} G_{\max} D_{\max} \sqrt{\sum_{j=1}^{m} D_j^2}. \tag{92}$$

- **Inference.**

  The inference corresponds to,

$$\widehat{f}(x) = \operatorname{argsort}_{\sigma \in \mathfrak{S}_m}(v), \quad \text{where } v_j = \sum_{i=1}^{n} \frac{G([r_i]_j) \alpha_i(x)}{N(r_i)}, \tag{93}$$

  which can be done in

$$\mathcal{O}(m \log m) \tag{94}$$

  operations.

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

$$\mathsf{A} \leq c G_{\max} \frac{m}{\sqrt{\log m}}. \tag{95}$$

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

$$A \leq \frac{2}{\sqrt{3}} G_{\max} \sqrt{m}. \tag{96}$$

The QS-estimator estimates the marginals of the normalized relevance scores and sorts the estimates at inference. As it was shown in [17], 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,

$$L(\sigma, y) = \frac{1}{N(y)} \sum_{j=1}^{m} \sum_{\ell \neq j} 1([y]_j < [y]_\ell) 1(\sigma(j) > \sigma(\ell)), \tag{97}$$

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

$$1([y]_j < [y]_\ell) = \frac{\operatorname{sign}([y]_\ell - [y]_j) + 1}{2}, \quad 1(\sigma(j) > \sigma(\ell)) = \frac{\operatorname{sign}(\sigma(j) - \sigma(\ell)) + 1}{2}. \tag{98}$$

  Hence,

$$L(\sigma, y) = \frac{1}{4} + \frac{1}{4N(y)} \sum_{j=1}^{m} \sum_{\ell \neq j} \operatorname{sign}([y]_\ell - [y]_j) \operatorname{sign}(\sigma(j) - \sigma(\ell)) \tag{99}$$

  Note that $F_\sigma = 1/4(\operatorname{sign}(\sigma(j) - \sigma(\ell)))_{j,\ell=1}^m$ and $U_y = (\frac{\operatorname{sign}([y]_\ell - [y]_j)}{N(y)})_{j,\ell=1}^m$ 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,

$$\mathsf{A} = \frac{m}{4} \tag{100}$$

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

$$\widehat{f}(x) \in \arg\min_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^{m} \sum_{\ell \neq j} \gamma_{j\ell}(x) 1(\sigma(j) > \sigma(\ell)), \tag{101}$$

  with

$$\gamma_{j\ell}(x) = \sum_{i | [y_i]_j < [y_i]_\ell} \frac{\alpha_i(x)}{N(y_i)}. \tag{102}$$

  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 [16].

As it was shown in [34], 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:

$$L(\sigma, y) = 1 - \frac{1}{|y|} \sum_{j|[y]_j=1} \frac{1}{\sigma(j)} \sum_{\ell=1}^{\sigma(j)} [y]_{\sigma^{-1}(\ell)}. \tag{103}$$

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,

$$A = \frac{1}{2} m \sqrt{\log(m+1)} \tag{104}$$

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

$$\widehat{f}(x) = \arg\max_{\sigma \in \mathfrak{S}_m} \sum_{j=1}^{m} \sum_{\ell=1}^{j} \frac{1}{\max(\sigma(j), \sigma(\ell))} \sum_{i|[y_i]_j [y_i]_\ell = 1} \frac{\alpha_i(x)}{|y_i|} \tag{105}$$

  Denote by

$$W_{j\ell} = \begin{cases} \sum_{i|[y_i]_j [y_i]_\ell = 1} \frac{\alpha_i(x)}{|y_i|} & j \geq \ell \\ 0 & \text{otherwise} \end{cases}, \; D_{j\ell} = \begin{cases} \max(j, \ell)^{-1} & j \geq \ell \\ 0 & \text{otherwise} \end{cases} \tag{106}$$

We have that,

$$\widehat{f}(x) = \arg\max_{\sigma \in \mathfrak{S}_m} \sum_{j,\ell=1}^{m} W_{j\ell} D_{\sigma(j)\sigma(\ell)} \equiv \arg\max_{P \in \Pi_m} \ \mathrm{Tr}(W^T P D P^T), \tag{107}$$

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 [16].

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.