Structured prediction with theoretical guarantees

Alex Nowak Vila

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Structured Prediction with Theoretical Guarantees

Soutenue par
Alex Nowak Vila
Le 28/09/2021

École doctorale n°386
Sciences Mathématiques de Paris Centre

Spécialité
Informatique

Composition du jury :

Ingo Steinwart Universität Stuttgart	*Rapporteur*
Robert C. Williamson Universität Tübingen	*Rapporteur*
Alexandre d'Aspremont École Normale Supérieure, Inria	*Examinateur*
Florence D'Alché-Buc Telecom Paris	*Présidente du jury*
Francis Bach École Normale Supérieure, Inria	*Directeur de thèse*
Alessandro Rudi École Normale Supérieure, Inria	*Directeur de thèse*

ENS | PSL

Abstract

Classification is the branch of supervised learning that aims at estimating a discrete valued mapping from data made of input-output pairs. The most classical and well studied setting is binary classification, where the discrete predictor takes zero or one as value. However, most of the practical classification settings deal with large structured output spaces such as sequences, grids, graphs, permutations, matchings, etc. There are many fundamental differences between structured prediction and vanilla binary or multi-class classification, such as the exponentially large size of the output space with respect to the natural dimension of the output objects and the cost-sensitive nature of the learning task. This thesis focuses on surrogate methods for structured prediction, whereby the typically intractable discrete problem is approached using a convex continuous surrogate problem which in turn can be addressed using techniques from regression. The large size of the output space and the cost sensitivity of the task in the structured setting defines new challenges not present in the binary case. Much effort in this thesis is towards a general theory of surrogate methods whereby statistical and computational properties such as Fisher consistency, calibration, complexity of training and complexity of inference are studied. More specifically, two main types of surrogate strategies show up in this thesis: probabilistic methods, also known as plug-in classifiers, and non-probabilistic methods. The main contribution on the first type is a quantitative calibration analysis for both existing and recently proposed structured probabilistic surrogates, which is a key ingredient to obtain guarantees for those estimators. The main contributions on the second type include a statistical and computational analysis of the Max-Min loss, a recently proposed calibrated polyhedral surrogate loss for structured prediction and a consistency analysis of the structured Max loss, also known as structural SVM, which is the classical extension of the binary SVM to structured output spaces.

Résumé

La classification est la branche de l'apprentissage supervisé qui vise à estimer une fonction à valeurs discrètes à partir de données constituées de paires d'entrées et de sorties. Le cadre le plus classique et le plus étudié est celui de la classification binaire, où le prédicteur discret prend pour valeur zéro ou un. Cependant, la plupart des problèmes de classification qu'on retrouve en pratique sont definis sur de grands espaces de sortie structurés tels que des séquences, des grilles, des graphs, des permutations, etc. Il existe des différences fondamentales entre la prédiction structurée et la classification multiclasse ou binaire non structurée: la grandeur exponentielle de l'espace de sortie par rapport à la dimension naturelle des objets à prédire et la sensibilité des coûts de la tâche de classification. Cette thèse se concentre sur les méthodes de substitution pour la prédiction structurée, dans lesquelles le problème discret typiquement insoluble est abordé à l'aide d'un problème continu convexe qui, à son tour, peut être résolu à l'aide de techniques de régression. La grandeur de l'espace de sortie et la sensibilité des coûts de classification de la tâche dans le cadre structuré définissent de nouveaux défis qui ne sont pas présents dans le cas binaire. Un soin tout particulier dans cette thèse porte sur une théorie générale des méthodes de substitution à partir de l'étude de propriétés statistiques et computationelles telles que la calibration à la tâche de prédiction discrète, la complexité de l'apprentissage et la complexité de l'inférence. En particulier, deux types principaux de stratégies de substitution sont présentés dans cette thèse : les méthodes probabilistes, également appelées classifieurs plug-in, et les méthodes non probabilistes. La principale contribution pour le premier type est une analyse quantitative de la calibration des substituts probabilistes structurés existants et aussi récemment proposés; cette analyse est un ingrédient clé pour obtenir des garanties pour ces estimateurs. Les principales contributions pour le deuxième type incluent une analyse statistique et computationelle de la fonction de perte max-min, une perte de substitution polyédrique calibrée récemment proposée pour la prédiction structurée et finalement une analyse de calibration de la fonction de perte max-margin structurée, également connue sous le nom de structural SVM, qui est l'extension classique de la SVM binaire aux espaces de sorties structurées.

Remerciements

First of all, I would like to thank my advisors Francis and Alessandro for giving me with such a unique opportunity for conducting my PhD and for providing me with their guidance and wisdom through all the ups and downs during these years. In particular, I want to emphasize how much the freedom you gave me with regards to the research topic was valuable to me. I want to thank my two reviewers Pr. Williamson and Pr. Steinwart for reading my thesis, Pr. Florence d'Alché-Buc to accept the presidency of the jury and Pr. d'Aspremont to accept being part of it. This experience would not have been possible without Joan Bruna, who introduced me to the world of machine learning when I was still in my undergraduate degree, and who recommended me Paris as a great place to pursue my graduate studies.

Finalment, vull donar les gràcies a totes les persones que m'han fet costat durant aquest petit fragment de la meva trajectòria vital. Sense la vostra companyia en els bons moments i el vostre recolzament en els mal moments, de ben segur no seria pas on estic ara.

Acknowledgements. This thesis was funded by "La Caixa" Foundation and Inria.

Outline

PhD advisors: Francis Bach and Alessandro Rudi
Location: SIERRA project-team, INRIA & DI-ENS, Paris, France

Contributions and Thesis Outline

We briefly summarize the structure of the thesis, which is divided into three main parts.

Part I: Introduction.

- **Chapter 1: Introduction.** The goal of the introductory chapter is to provide a general picture of surrogate methods and its application to structured prediction. Hopefully, this will prove to be useful for establishing the necessary background and to give a brief taste of the contributions presented in depth in the following chapters. More specifically, the chapter starts with a brief introduction to structured prediction and a presentation of the very basics of statistical learning theory. Then, surrogate methods are presented in its generality together with their desirable theoretical properties and the main types. After that, the surrogate strategy is studied under the point-wise framework for cost-sensitive losses, and finally all these tools are deployed in the structured prediction setting. A final section of related works and summary of contributions closes the introductory chapter.

Part II: Probabilistic Surrogates. This part focuses on probabilistic surrogate methods, i.e., based on conditional risk estimation. It is divided into the following two chapters:

- **Chapter 2: Sharp analysis of learning with discrete losses.** This chapter focuses on the statistical and computational analysis of the quadratic surrogate loss. We explicitly show that the statistical cost of minimizing low-rank structured losses is logarithmic in the potentially exponential size of the output space, showing thus the feasibility of the learning problem.
- **Chapter 3: A general theory for structured prediction with smooth convex surrogates.** This chapter extends the above analysis to generic smooth convex surrogate losses for structured prediction such as conditional random fields. We leverage tools from property elicitation theory to study the generic structure of probabilistic based surrogate losses, which allows us to provide statistical results on the performance of these methods when approached using stochastic gradient techniques.

Part III: Non-probabilistic Surrogates. This part focuses on non-probabilistic surrogate methods, i.e., not based on conditional risk estimation. Tt has the following two chapters:

- **Chapter 4: Consistent structured prediction with Max-Min Margin Markov Networks.** We introduce a novel non-probabilistic consistent convex surrogate method for structured prediction called Max-Min Margin Markov Networks ($\mathrm{M}^4\mathrm{N}$), based on the Max-Min loss. We provide a full-stack analysis of the method; calibration, generalisation bounds and a generalised version of Block-Coordinate Frank-Wolfe as an efficient algorithm to minimize the regularized empirical risk. The theoretical findings are validated with reproducible experiments available in a public repository.

- **Chapter 5: Max-Margin is Dead, Long Live Max-Margin!** In this chapter we study margin maximization methods and their inability to provide guarantees for general structured prediction. In particular, we provide necessary and sufficient conditions on the discrete loss to ensure consistency of the Max loss, which are not satisfied by most of the losses used in practice. Moreover, we propose a novel restricted version of the Max loss, called Restricted-Max loss, which significantly weakens those conditions and it is consistent for discrete losses such as the 0-1 loss.

Finally, we conclude the thesis by summarizing the contributions and providing guidelines for future work.

Publications and Preprints. The main bulk of the thesis is based on the following published and not yet published scientific papers:

- Part II is based on:
  - **Sharp Analysis of Learning with Discrete Losses**, Nowak-Vila, A., Bach, F., and Rudi, A., published in *International Conference on Artificial Intelligence and Statistics 2019* (cited as Nowak et al. (2019)).
  - **A General Theory for Structured Prediction with Smooth Convex Surrogates**, Nowak-Vila, A., Bach, F., and Rudi, A., technical report (cited as Nowak-Vila et al. (2019)).
- Part III is based on:
  - **Consistent Structured Prediction with Max-Min Markov Networks**, Nowak-Vila, A., Bach, F., and Rudi, A., published in the *International Conference on Machine Learning 2020* (cited as Nowak-Vila et al. (2020)).
  - **Max-Margin is Dead, Long Live Max-Margin!**, Nowak-Vila, A., Bach, F., and Rudi, A., submitted to NeurIPS 2021.

The paper Eboli et al. (2020) is a project on structured prediction applied to image restoration. The work presents a novel approach to image restoration that leverages ideas from localized structured prediction and non-linear multi-task learning by optimizing a learned penalized energy function. The resulting estimator comes with strong statistical guarantees leveraging local dependency properties of overlapping patches and its practical efectiveness is demonstrated on different image restoration problems using standard benchmarks. This research work does not appear in this manuscript.

I

Introduction

1 Introduction

1 A Brief Introduction to Structured Prediction

Machine Learning. Classical computer science deals with input-output tasks for which there exists an explicit description of the underlying mechanism generating the output from the input. The sequence of operations the computer performs to check the correctness of our credentials when logging into a session or to sort an array of natural numbers is directly programmed into the computer as a set of rules derived from a complete understanding of the procedure. Unfortunately, expert knowledge is not available for most complex tasks. The lack of mathematical description for recognizing objects in an image or to translate a sentence from one language to another does not allow us to write down a set of rules to generate the label of the object or the translated sentence from the corresponding input. While the exponential increase in computer power of these last decades cannot provide us the explicit rules to solve these tasks, it gives us an indirect solution through *learning*. Indeed, a massive amount of data related to these complex tasks, such as labelled images or translated text in the case of the examples above, is generated daily due to our every day exposition to electronic devices and can be leveraged by computers as raw material for learning. Machine learning is the scientific discipline lying at the intersection of computer science and statistics aiming at learning such complex tasks from data.

Supervised learning. The setting we consider in this thesis is supervised learning (Vapnik, 1995; Devroye et al., 1996), which is the branch of machine learning where the data is made of input-output pairs. More specifically, the goal is to predict an unobserved output y from an output space $\mathcal{Y}$ given an input x from an input space $\mathcal{X}$, by estimating a function

$$f : \mathcal{X} \longrightarrow \mathcal{Y},$$

such that $f(x) \approx y$ learned from finite data made of pairs (x, y). In the examples above, $\mathcal{X}$ is the space of images and $\mathcal{Y} = \{-1, 1\}$ are binary labels specifying whether there is a dog or a cat present in an image and for translation $\mathcal{X}$ and $\mathcal{Y}$ are the space of sentences in two different languages, respectively. The error between the observed value y and the prediction $f(x)$ is measured by a non-negative loss function as $L(f(x), y)$, where

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

is a measure of error satisfying $L(y, y) = 0$ for all output elements y. Other well-known branches of machine learning are unsupervised learning (Friedman et al., 2001), where only input data is available and the goal is to solve the task by finding patterns in the

data, and reinforcement learning (Sutton and Barto, 2018), where the data is generated in an online fashion by interaction with an environment.

There are two classical settings in supervised learning depending on whether the output willing to predict is *continuous* or *discrete*.

Continuous Prediction. This setting is also known as *regression*. The output space $\mathcal{Y}$ is generally a vector space. This setting includes scalar regression, where the output lives in $\mathcal{Y} = \mathbb{R}$ and vector regression, where the output is in $\mathcal{Y} = \mathbb{R}^k$. A classical loss used in those settings is least squares, defined as $L(y, y') = \|y - y'\|_2^2$.

Discrete Prediction. While we will extensively use continuous prediction as a means to predict discrete labels, the main goal in this thesis is to study discrete prediction problems. We classify them into two main groups depending on the type of output.

- **Unstructured Prediction.** Informally, this setting corresponds to the case where the different labels to be predicted have the same importance. Thus, we can visualise the elements of the output space as being independent objects. This includes vanilla *binary classification* $\mathcal{Y} = \{-1, 1\}$ and *multi-class classification* $\mathcal{Y} = \{1, 2, \dots, k\}$. The loss computing the error between predictions and observations is the 0-1 loss $L(y, y') = 1(y \neq y')$, defined as zero if both elements are equal and one otherwise.
- **Structured Prediction.** This is the more general setting in supervised learning. The labels are *structured objects* (BakIr et al., 2007; Nowozin et al., 2014) in the sense that some pairs of labels are closer in the output space than others. For instance, if $\mathcal{Y}$ is the space of sequences of a fixed length in a given dictionary of characters, a sequence is closer to another one differing in a single character than to a sequence with all characters being different. Furthermore, the output space can be of *exponentially large size* with respect to the natural dimensional of the output objects (e.g., the number of possible sequences is exponential with the length), which can lead in many learning settings to larger number of potential outputs than available data for learning. More concretely, we say that a prediction problem is structured if it has *at least* one of the two following distinctive features:

Structured prediction distinctive features

1. **Cost-sensitivity of predictions.** The loss measuring the error between observations and predictions is different than the 0-1 loss $L(y, y') = 1(y' \neq y)$.
2. **Large output space.** The cardinality of the output space is much larger than the natural dimension of the output elements.

While all cost-sensitive learning problems are not always considered as structured in the literature, we include it in our definition as the developed framework and analysis in this thesis will naturally include this setting. The following are examples of structured output spaces and some discrete losses defining its geometry:

  - **Ordered elements.** The output set consists of ordered elements $1 \prec 2 \prec \cdots \prec k$. A classical measure of error is the absolute deviation loss $L(y, y') = |y - y'|$.

- **Subsets.** The output space is the set of subsets of $\{1, \dots, k\}$. The space has exponential size 2^k and a classical measure of error is the F-1 score, defined as the harmonic mean between precision and recall $L(y, y') = 2|y \cap y'|/(|y| + |y'|)$.
- **Permutations.** The output space is the set of permutations acting on a set of k elements. The space has $k! = k \times (k-1) \times \cdots \times 2 \times 1$ elements. There exist several losses between pairs of permutations σ, σ' (Deza and Deza, 2009), such as the Kendall's τ distance $L(\sigma, \sigma') = \sum_{i<j} 1((\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0)$ and the Hamming loss $L(\sigma, \sigma') = \frac{1}{k} \sum_{j=1}^{k} 1(\sigma(j) \neq \sigma'(j))$.
- **Sequences.** The output space is the set of sequences of length k with characters in a dictionary of size R. The space has R^k elements. Classical losses for sequences are the edit distance (Jurafsky, 2000), which counts the minimum number of required editions to transform one sequence to another and the Hamming loss measuring the proportion of position-wise errors.
- **Trees.** The output space is the set of trees, i.e, acyclic undirected graphs. Typical metrics focus on the number of 'constituent errors' between two trees (Black et al., 1991), and it is generally measured using the F1-score.
- **Grids.** The output space is a $k \times k$ - dimensional grid, where every element of the grid is a character in a dictionary of size R. The space has R^{k^2} elements and error is generally measured using the Hamming loss.

Remark 1 .1 (Continuous structured problems). In some cases continuous prediction problems are also considered as structured. For instance, if the continuous output space is structured as a manifold (Rudi et al., 2018). This setting will not be considered in this thesis.

Example 1 .2. There exists a vast zoo of structured classification problems in many different scientific and engineering areas (see Figure 1.1). The following are examples from three particular domains.

Natural Language Processing. Data related to language is highly structured in nature. Structure prediction tasks in this field include part-of-speech tagging, handwritten recognition, parsing, named-entity recognition, text summarization, translation and word alignment, amongst others (Smith, 2011).

Computer Vision. Images are structured objects made of pixels laid out in a grid and many tasks dealing with images require predicting other types of structures. Some examples are object detection, segmentation, motion tracking, 3D reconstruction from video and stereo (Nowozin and Lampert, 2011).

Computational Biology. The emergence of modern molecular genetics has changed modern biology. In particular, recent advancements such as large-scale DNA sequencing are producing massive amount of structured data that requires

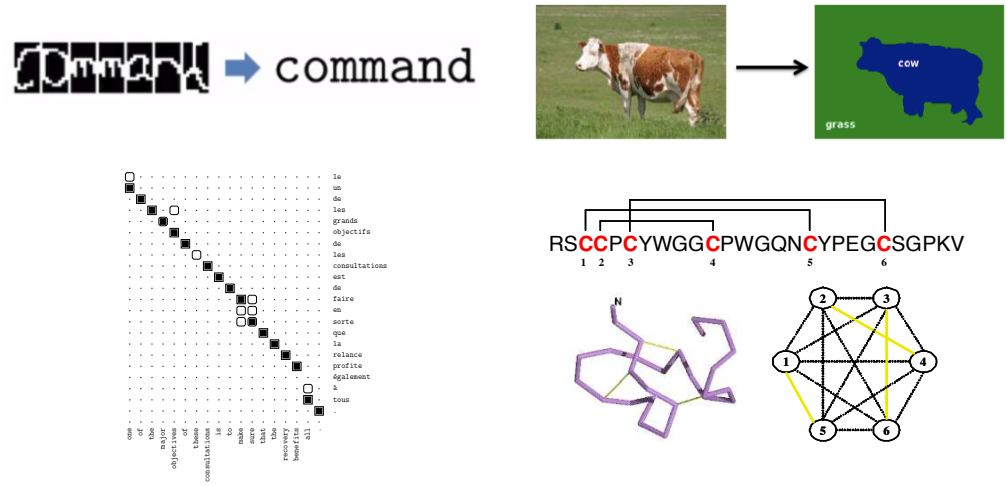


Figure 1.1: Examples of structured prediction tasks. **Top-left:** In optical character recognition (OCR) (Taskar et al., 2004), the goal is to predict the word written in the sequence of images. The structured output is a sequence. **Top-Right:** In image segmentation (Forsyth and Ponce, 2012), the goal is to predict the object every pixel of the image belongs to. The structured output is a grid of pixels. **Bottom-left:** In word alignment (Lacoste-Julien et al., 2006), the goal is to predict an alignment between the words of two sentences of different languages, such that every word is mapped to its semantic equivalent in the other language. The structured output is an alignment. **Bottom-Right:** In disulfide connectivity prediction (Chatalbashev et al., 2005), the goal is to predict a set of pairs of elements of an aminoacid chain describing the disulfide bridges (intra-chain covalent bonds). The structured output is a set of pairs where each element belongs to a single pair.

machine learning techniques to make sense of. In particular, prediction tasks in computational biology include gene-finding, alignment of sequences, protein structure prediction, DNA sequence labelling, molecular pathway discovery and bond structure discovery, amongst others (Durbin et al., 1998).

2 Mathematical Framework

In this section, we give a brief introduction to the mathematical framework we will use throughout the thesis. In particular, we show why continuous prediction tasks are both statistically and computationally more attractive than its discrete counterparts for learning directly. This serves as a motivation for the introduction of surrogate methods in the next section, which is the central methodology studied in this thesis whereby a surrogate continuous problem is designed to solve the discrete task of ultimate interest.

Statistical learning. The statistical learning setting (Vapnik, 1995) is the classical mathematical framework to study supervised learning methods. Let ρ be an *unknown distribution* of input-output pairs in $\mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ and L a *loss* measuring the error between a predicted output y and an observed output y' as $L(y, y')$. The quality of a *predictor* $f : \mathcal{X} \to \mathcal{Y}$ is measured using the so-called *expected risk*,

$$\mathcal{E}(f) \coloneqq \mathbb{E}_{(x,y)\sim\rho}\ L(f(x), y), \tag{1.1}$$

defined as the averaged incurred cost of predicting y using $f(x)$ measured by L, where the pair (x, y) comes from the distribution ρ. Thus, the problem consists in finding a predictor f with low expected risk $\mathcal{E}(f)$. A central mathematical object of the learning task is the *Bayes predictor* $f^\star$ defined as the exact minimizer of the expected risk over all measurable functions. This is precisely the mapping we want to approximate. The Bayes predictor can be characterized point-wise in $x \in \mathcal{X}$ as

$$f^\star(x) \in \operatorname*{arg\,min}_{y'\in\mathcal{Y}}\ \mathbb{E}_{y\sim\rho_x}\ L(y', y), \tag{1.2}$$

where $\rho_x(y) = \rho(y \mid x)$ is the conditional probability of y given x. Note that the Bayes predictor is always a function of the conditional distribution ρ_x.

Example 2 .1 (Bayes predictors). The following are two classical examples of Bayes predictors for classification and regression, respectively.

- **Binary classification.** If $L(y, y') = 1(y' \neq y)$ is the 0-1 loss and $\mathcal{Y} = \{-1, 1\}$:

$$f^\star(x) = \mathrm{sign}(\mathbb{E}_{y\sim\rho_x}\ y).$$

- **Least-squares regression.** If $L(y, y') = (y - y')^2$ and $\mathcal{Y} = \mathbb{R}$:

$$f^\star(x) = \mathbb{E}_{y\sim\rho_x}\ y.$$

Estimators. In general, the exact minimization of the expected risk cannot be solved in practice as we only have access to the distribution ρ through a *finite dataset* $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ made of n input-output pairs sampled independent and identically distributed (i.i.d.) from the distribution ρ. The goal is to design an *estimator* $\mathcal{A} : \cup_{n\geq 1}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ taking as input the dataset $\mathcal{D}_n$ and generating a function from a *hypothesis space* $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$, also known as class of predictors, having low expected risk (1.1). More concretely, we want to design an estimator $\mathcal{A}$ minimizing the

quantity $\mathcal{E}(\mathcal{A}(\mathcal{D}_n))$ in expectation or in high probability over the n samples $\mathcal{D}_n \sim \rho^n$ [1]. In particular, typical guarantees in expectation take the form $\mathbb{E}_{\mathcal{D}_n \sim \rho^n}\ \mathcal{E}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{E}(f^\star) \leq \mathcal{O}(n^{-\gamma})$, where $\gamma > 0$ is positive scalar specifying the polynomial rate of convergence to the Bayes predictor $f^\star$.

> *Remark* 2 .2. **A note on the i.i.d. assumption.** The statistical learning framework assumes that the distribution of the training data is the same as the distribution in which the estimator is evaluated. In practice this assumption is rarely satisfied, as in many settings the learned estimators are applied in domains different than the ones used for training (Torralba and Efros, 2011). The problem of out-of-distribution generalization is out of the scope of this thesis, but it is important to keep in mind that the i.i.d. assumption is usually a deliberate simplification differing from real practices in the field.

An insightful way to look at the error incurred by an estimated predictor f is to decompose it into the so-called *estimation error* $\mathcal{E}_{\text{est}}$ and *approximation error* $\mathcal{E}_{\text{app}}$ as

$$\mathcal{E}(f) - \mathcal{E}(f^\star) = \underbrace{\mathcal{E}(f) - \mathcal{E}(f^\star_{\mathcal{F}})}_{\mathcal{E}_{\text{est}}} + \underbrace{\mathcal{E}(f^\star_{\mathcal{F}}) - \mathcal{E}(f^\star)}_{\mathcal{E}_{\text{app}}},$$

where $f^\star_{\mathcal{F}}$ is the minimizer of the expected risk over $\mathcal{F}$. The estimation error is a random quantity (f is random as it is an estimated quantity) and it is the source of error coming from the fact that we only have access to the distribution ρ through the finite dataset $\mathcal{D}_n$. The approximation error is a deterministic quantity coming from the fact we seek a solution over a restricted set of hypothesis $\mathcal{F}$ and not the set of all possible measurable input-output mappings. In the context of least-squares regression, the errors $\mathcal{E}_{\text{est}}$ and $\mathcal{E}_{\text{app}}$ correspond to the classical variance and bias terms, respectively. Undesired unbalance between the two sources of error above leads to important phenomena in statistical learning theory known as *overfitting* and *underfitting*.

Overfitting and underfitting. If the hypothesis space $\mathcal{F}$ is large, the approximation error will be small as the minimizer of the expected risk over this space will be closer to the exact minimizer $f^\star$. However, picking $\mathcal{F}$ 'too large' with respect to the amount of available data will increase the statistical error as the estimator will pick 'complicated' functions very specific to the observed data but unstable to unobserved data under the distribution. This phenomena is known as overfitting. On the other hand, choosing a small hypothesis space will reduce the statistical error as small amount of data will be enough to pick a stable solution in $\mathcal{F}$ but choosing it too small will move the potential predictors away from the Bayes predictor $f^\star$. This phenomena is known as underfitting. See Figure 1.2 for a visualization of those phenomena in a one dimensional regression task.

[1] High probability learning bounds are also known as 'Probably approximately correct' (PAC) (Valiant, 1984).

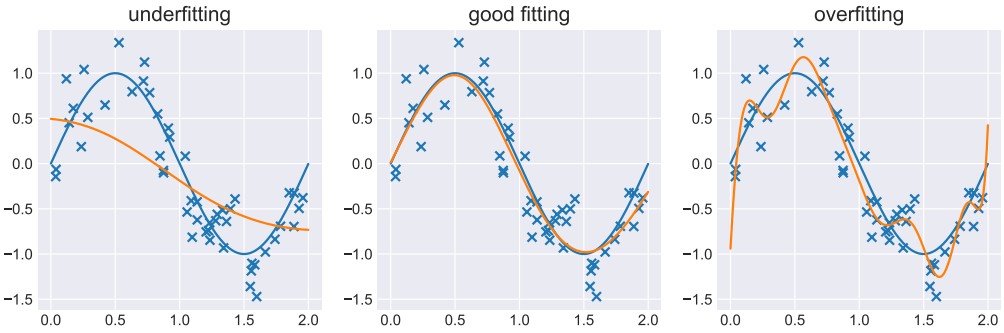


Figure 1.2: We depict in a blue line the Bayes predictor $f^\star$, with crosses the elements of the dataset and in an orange line the estimator f. **Left (underfitting):** The estimator gives a too simple explanation of the data, incurring in large approximation error. On the other hand, the estimation error is small as the solution is stable under additional data. **Right (overfitting):** The hypothesis space $\mathcal{F}$ is large enough to contain $f^\star$ but too large as it provides a solution too specialized to the observations and not stable under additional data from the distribution. **Middle (good fitting):** The hypothesis space $\mathcal{F}$ is small enough to keep the estimation error small and good enough to approximate well $f^\star$.

> The crux of designing a good estimator is to choose the right function class $\mathcal{F}$ *small enough* to control the estimation error and *good enough* to approximate the Bayes predictor $f^\star$.

Below we briefly formalize what we mean by large and small hypothesis space by studying the estimation error of the minimizer of the average error on the dataset $\mathcal{D}_n$.

Empirical Risk Minimization (ERM). A natural approach to minimize the expected risk having only access to a finite dataset is to consider the *empirical risk* $\mathcal{E}_n(f)$ defined as

$$\mathcal{E}_n(f) := \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i), \tag{1.3}$$

and minimize it over the hypothesis space $\mathcal{F}$. The estimation error $\mathcal{E}_{\text{est}}$ of the minimizer f_n of the empirical risk (1.3) can be upper bounded as follows

$$\begin{aligned}\mathcal{E}(f_n) - \mathcal{E}(f_{\mathcal{F}}^\star) &= \{\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\} + \underbrace{\{\mathcal{E}_n(f_n) - \mathcal{E}_n(f_{\mathcal{F}}^\star)\}}_{\leq 0} + \{\mathcal{E}_n(f_{\mathcal{F}}^\star) - \mathcal{E}(f_{\mathcal{F}}^\star)\} \\ &\leq 2\sup_{f\in\mathcal{F}} |\mathcal{E}_n(f) - \mathcal{E}(f)|,\end{aligned}$$

where we have used that $\mathcal{E}_n(f_n) = \min_{f\in\mathcal{F}} \mathcal{E}_n(f) \leq \mathcal{E}_n(f)$ for all $f \in \mathcal{F}$. Thus, the estimation error can be bounded by the maximal deviation of the empirical risk to the expected risk over the hypothesis space $\mathcal{F}$. Upper bounds on the expectation of this random object with respect to the number of samples n and complexity quantities associated to the hypothesis space $\mathcal{F}$ are known as *uniform bounds*. These bounds provide us with a notion of "size" of the hypothesis space $\mathcal{F}$ that will allow us to control the estimation error.

Uniform bounds through empirical process theory. Uniform bounds for binary (thus discrete) classification problems with the 0-1 loss date back from the first works of statistical learning theory (Vapnik, 1999). In this setting, the hypothesis space is a class of sets from the input space defining the decision boundary of the predictor f. The main result is that the estimation error of the ERM estimator f_n can be bounded in expectation as

$$\lesssim \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}},$$

where $\mathrm{VC}(\mathcal{F})$ is the so-called Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ and $\lesssim$ means greater or equal up to universal multiplicative constants independent of $\mathcal{F}$ and n. This dimension is *combinatorial* in nature and it is defined as the maximum number of points in $\mathcal{X}$ that can be shattered [2] using sets from the hypothesis class. Put in simple terms, it is a measure of the capability of the decision boundaries to separate elements of the input space. As we will argue below, in practice discrete prediction problems are solved using vector-valued functional spaces, and while the VC dimension can be extended to this setting (Pollard, 1984), other measures of complexity have appeared which are more natural in the continuous setting and sensitive to the distribution of the data, such as the *Rademacher complexity* $\mathrm{Rad}(n, \mathcal{F})$ (Koltchinskii and Panchenko, 2000; Bartlett et al., 2002). In particular, if the loss L is K-Lipschitz [3], the estimation error can be upper bounded in expectation as

$$\lesssim K \underbrace{\mathbb{E}_{\sigma,\rho}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right]}_{\mathrm{Rad}(n,\mathcal{F})},$$

where σ are i.i.d. Rademacher variables $\mathbb{P}\{\sigma_i = \pm 1\} = 1/2$. The Rademacher complexity can usually be bounded as $\leq RB/\sqrt{n}$, where B is a measure of the maximal complexity of the predictors, typically a functional norm $\|\cdot\|_{\mathcal{F}}$ of the hypothesis space, and R is the size of the inputs. [4] We informally call the *capacity* of $\mathcal{F}$ the hypothesis specific term controlling the estimation error. These bounds on the statistical error of the form $\frac{\text{capacity of } \mathcal{F}}{\sqrt{n}}$ suggest that one should *control the capacity of $\mathcal{F}$ in terms of the number of available samples*. The technique to adaptively control the complexity of the predictors with the number of samples to avoid large estimation error (thus overfitting) is called *regularization*.

Capacity control for continuous output spaces. When the output space $\mathcal{Y}$ is continuous, a classical way to implement such a regularization strategy is to choose a large hypothesis space $\mathcal{F}$, potentially a dense set in the continuous space of functions, and define a regularizer on $\mathcal{F}$, typically a norm $\|\cdot\|_{\mathcal{F}}$ controlling the capacity. Indeed, as we have seen above, the capacity in the continuous setting can be usually controlled by the functional norm of the hypothesis space. The estimator is then defined as the minimizer of the following penalized optimization problem

$$\min_{f\in\mathcal{F}} \mathcal{E}_n(f) + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2,$$

[2] If A is a discrete set and $\mathcal{C}$ a class of sets, we say that $\mathcal{F}$ shatters A if for each subset $a \in A$, there exists $c \in \mathcal{C}$ such that $a = A \cap c$.

[3] Note that this excludes its direct applicability to discrete learning problems. As we will see later, this is not a problem as those are usually tackled with surrogate continuous prediction problems.

[4] This is the case for reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950).

where $\lambda > 0$ is the so-called *regularization parameter*. If λ is large the penalization will guide the solution to small capacity predictors, while small penalization will encourage minimizers of larger norm, thus more complex. On other words, λ controls the trade-off between estimation and approximation error. While it is clear that the regularization parameter should decrease with the number of samples, choosing the right parameter is a hard problem in practice and one usually additional validation data is required for this task. Note that the *convexity* of L in the first argument is key to be able to tractably minimize the regularized functional. In the next section we will briefly discuss about computational issues and general gradient based techniques to efficiently solve the regularized ERM problem.

Example 2 .3 (Capacity control for linear hypothesis spaces). The classical example of capacity control for regression problems is that of Tikhonov regularization (Tikhonov, 1963) for least-squares loss and linear hypothesis space $\mathcal{F}_{\text{lin}} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = w^\top x,\ w, x \in \mathbb{R}^d\}$. In this case, the estimator is defined as the minimizer of the following penalized convex minimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2,$$

where $\|w\|_2$ can be seen as a norm of the function $f(x) = w^\top x$ in the space $\mathcal{F}_{\text{lin}}$ (Schölkopf et al., 2002).

Problems of ERM for discrete output spaces. While the study of empirical risk minimization for binary classification can be found in the first works of statistical learning theory using the combinatorial VC-dimension introduced above, in practice this estimator has several issues both of computational and statistical nature due to the discreteness of the output space $\mathcal{Y}$.

(i) **ERM is generally intractable.** In binary classification, a binary-valued predictor f can be viewed as a set in $\mathcal{X}$, where the points x satisfy $f(x) = 1$ if they belong to the set and $f(x) = -1$ otherwise. For instance, the hypothesis space of linear decision functions corresponds to the set of halfplanes. Unfortunately, minimizing the empirical risk (1.3) over this type of space of functions is a combinatorial problem which is known to be computationally intractable for many non-trivial classes of predictors (Arora et al., 1997).

(ii) **Lack of regularization techniques.** While for real or vector valued functions the capacity of the hypothesis space can be controlled by a norm, the capacity of spaces of sets is controlled by combinatorial quantities such as the VC-dimension, for which there is no clear way to adaptively control it with the number of samples n. There exist capacity measures of spaces of sets depending continuously on a regularity parameter of the decision boundaries (Mammen and Tsybakov, 1995), but the resulting algorithm remains abstract and impossible to implement in practice (Tsybakov, 2004).

The classical solution to problems (i) and (ii) with discrete empirical risk minimization is to consider a surrogate continuous prediction problem for which the solution can

be mapped to a solution of the discrete one. Regularization techniques such as penalization with a norm can then be applied to adaptively control the capacity of the space of continuous surrogate predictors. This thesis is devoted to the design and study of both the statistical and computational properties of such surrogate strategies for tackling discrete structured prediction problems.

3 Surrogate Methods

From now on, we assume the output space $\mathcal{Y}$ to be *discrete* and f to be discrete valued. The goal of surrogate methods is to substitute the discrete prediction problem into a continuous prediction problem such that the solution of the first can be written in terms of a solution of the second. More concretely, the surrogate approach is made of the following two objects:

1. **Surrogate loss S.** A *surrogate loss* $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ is a loss bounded from below and continuously depending on the first argument. The vector space $\mathbb{R}^k$ is the *surrogate space*, in contrast with the discrete output space $\mathcal{Y}$.

2. **Decoding function d.** The decoding is a discrete-valued function $d : \mathbb{R}^k \to \mathcal{Y}$ mapping points from the surrogate space $\mathbb{R}^k$ to the discrete output space $\mathcal{Y}$.

The goal of the surrogate approach is to find a *surrogate predictor* $g : \mathcal{X} \to \mathbb{R}^k$ minimizing the *expected surrogate risk* $\mathcal{R}$ defined as

$$\mathcal{R}(g) \coloneqq \mathbb{E}_{(x,y)\sim\rho}\ S(g(x), y). \tag{1.4}$$

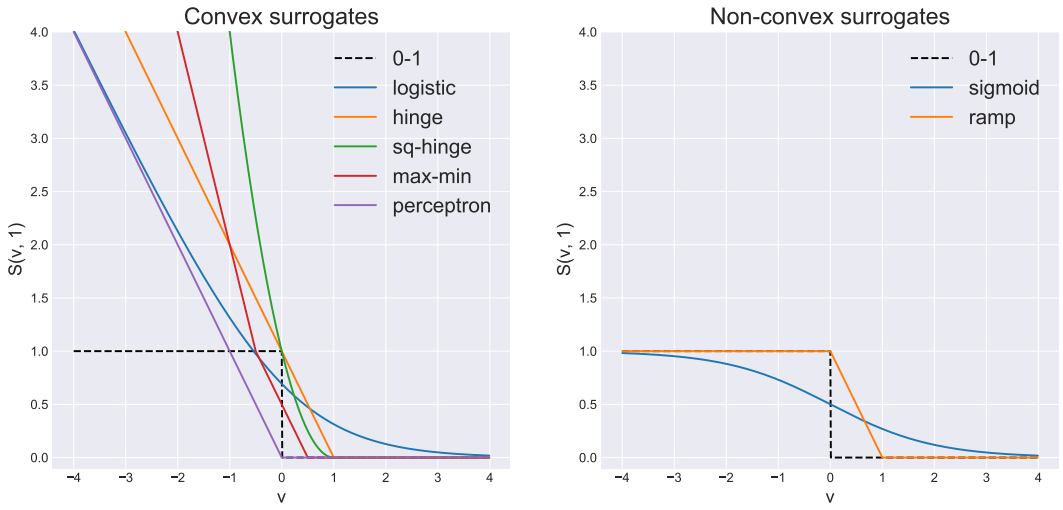
Then, a discrete predictor $f : \mathcal{X} \to \mathcal{Y}$ is constructed using the decoding as $f = d \circ g$ [5]. Analogously to the Bayes predictor $f^\star$ associated to the discrete problem, we define the *surrogate Bayes predictor* $g^\star$ as the exact minimizer of the expected surrogate risk $\mathcal{R}$. It can be characterized point-wise in terms of the conditional distribution ρ_x as

$$g^\star(x) = \operatorname*{arg\,min}_{v\in\mathbb{R}^k}\ \mathbb{E}_{y\sim\rho_x} S(v, y). \tag{1.5}$$

See Figure 1.3 for examples of classical surrogate losses and decodings in binary classification and structured prediction.

Surrogate hypothesis space $\mathcal{G}$. Given a hypothesis space of vector valued functions $\mathcal{G} \subseteq \{g : \mathcal{X} \to \mathbb{R}^k\}$, the goal is to find the minimizer of the expected surrogate risk (1.4) over $\mathcal{G}$. The associated hypothesis space $\mathcal{F}$ of discrete predictors corresponds to the set of discrete-valued functions $f : \mathcal{X} \to \mathcal{Y}$ that can be written as $f = d \circ g$, where $g \in \mathcal{G}$. As an example, in binary classification with a sign decoding $d(v) = \operatorname{sign}(v), v \in \mathbb{R}$ and $\mathcal{G}$ the space of affine functions, the discrete hypothesis space corresponds to linear decision boundaries.

[5] The notation $f = d \circ g$ stands for $f(x) = d(g(x))$ for all $x \in \mathcal{X}$.



Binary Classification with decoding $d(v) = \text{sign}(v)$

Name	$S(v, y) : \mathbb{R} \times \{1, 1\} \to \mathbb{R}$	Convex	Smooth
Hinge	$\max(1 - yv, 0)$	✓	✗
Squared hinge	$\max(1 - yv, 0)^2$	✓	✓
Logistic	$\log(1 + e^{-yv})$	✓	✓
Max-Min	$\max(\lvert v\rvert, 1/2) - yv$	✓	✗
Ramp	$\min(1, \max(1 - yv, 0))$	✗	✗
Sigmoid	$(1 + e^{yv})^{-1}$	✗	✓
Perceptron	$\max(-yv, 0)$	✓	✗

Structured Prediction with decoding $d(v) = \arg\max_{y \in \mathcal{Y}} \ v_y$

Name	$S(v, y) : \mathbb{R}^{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$	Convex	Smooth
Structured Max	$\max_{y' \in \mathcal{Y}} \ L(y, y') + v_{y'} - v_y$	✓	✗
Structured Squared	$\lVert v + L_y\rVert_2^2/2$	✓	✓
Multinomial Logistic	$\log\left(\sum_{y' \in \mathcal{Y}} \exp v_{y'}\right) - v_y$	✓	✓
Max-Min	$\max_{q \in \Delta_{\mathcal{Y}}} \min_{y' \in \mathcal{Y}} \ L_{y'}^\top q + v^\top q - v_y$	✓	✗
Ramp	$\max_{y' \in \mathcal{Y}} \ L(y, y') + v_{y'} - \max_{z \in \mathcal{Y}} \ v_z$	✗	✗
Probit	$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0, I_{\mathcal{Y}})} \ L(\arg\max_{y' \in \mathcal{Y}} \ v_{y'} + \varepsilon_{y'}, y)$	✗	✓
Structured perceptron	$\max_{y' \in \mathcal{Y}} \ v_{y'} - v_y$	✓	✗

Figure 1.3: **Plots and table at top:** Convex and non-convex binary surrogates losses $S(v, 1)$, $v \in \mathbb{R}$ plotted against the 0-1 binary loss $L(\text{sign}(v), 1) = 1(\text{sign}(v) \neq 1)$. The classical decoding in binary classification is the sign function. **Table at bottom:** Surrogate losses used in structured prediction for minimizing a discrete loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The classical decoding used in structured prediction is the argmax function. Here, L_y stands for the vector $L_y = (L(y, y'))_{y' \in \mathcal{Y}}$ and $\mathcal{N}(0, I_{\mathcal{Y}})$ for the $|\mathcal{Y}|$-dimensional gaussian isotropic distribution.

Regularized ERM for the surrogate problem. As the problem is now continuous, given a finite dataset $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of pairs of points sampled i.i.d. from ρ, we can define the regularized ERM estimator g_n as the minimizer of the following mini-

mization problem

$$g_n = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} S(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_{\mathcal{G}}^2, \tag{1.6}$$

where $\|g\|_{\mathcal{G}}$ is the norm of the function g in the hypothesis space $\mathcal{G}$ and $\lambda > 0$ is the regularization parameter controlling the capacity of the solution. The above functional can be tractably minimized using convex optimization gradient-based methods (Boyd et al., 2004) if S is *convex* in the first argument and the surrogate predictors are *linearly parametrized*.

Remark 3 .1 (Stochastic gradient methods). Note that one has to look at all data points of the dataset in order to computing a gradient of the regularized ERM functional. Unfortunately, this can be extremely costly for massively large datasets. Most optimization algorithms used in machine learning use stochastic gradient algorithms (Robbins and Monro, 1951), where the parameters are updated at every iteration using information of individual elements of the dataset. Despite being a central part of the design of machine learning estimators, this is not the main topic of this thesis and we avoid going into much detail in the introductory section. We explicitly use these algorithms in Chapter 3 and more significantly in Chapter 4 where we design a specific algorithm for the strucutured Max-Min loss.

Fisher consistency. In order to make sure that the surrogate strategy is well posed with respect to the original task of interest, we ask that the exact minimizer $g^\star$ of the expected surrogate risk $\mathcal{R}$ gives the exact minimizer $f^\star$ of the expected risk $\mathcal{E}$ through the decoding as $f^\star = d \circ g^\star$. This is properly formalized by the notion of *Fisher consistency* introduced by Lin (2004) in the context of binary classification.

Definition 3 .2 (Fisher consistency)**.** *We say that S is Fisher consistent to L under the decoding d if for every probability distribution $\rho \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ it is verified that the minimizer $g^\star$ of the expected surrogate risk reaches optimal expected risk, that is,*

$$\mathcal{R}(g) = \mathcal{R}(g^\star) \implies \mathcal{E}(d \circ g) = \mathcal{E}(f^\star). \tag{1.7}$$

Example 3 .3. **Quadratic surrogate for binary classification.** From Example 2 .1, we know that the Bayes predictor for binary classification is $f^*(x) = \mathrm{sign}(\mathbb{E}_{y \sim \rho_x}\ y)$ and for least-squares regression is $g^\star(x) = \mathbb{E}_{y \sim \rho_x}\ y$ if the space of (continuous) observations in $\mathbb{R}$ is set to $\mathcal{Y} = \{-1, 1\} \subset \mathbb{R}$. Thus, one can estimate f^* by estimating the conditional expectation using least-squares regression. More specifically, the surrogate method defined by the surrogate loss $S(v, y) = (v - y)^2$ and decoding $d(v) = \mathrm{sign}(v)$ is trivially Fisher consistent to L, as by construction $f^\star = d \circ g^\star$.

Calibration. Fisher consistency is a property of the Bayes predictors $g^\star$ and $f^\star$. In practice, however, we are interested in the usual notion of consistency, which states that if the surrogate predictor g is arbitrarily close to $g^\star$, then the associated discrete predictor $d \circ g$ will be also close to $f^\star$. Indeed, this is precisely the notion we are ultimately interested in as $g^\star$ is never attained in practice. This is captured by the concept of *calibration* (Bartlett et al., 2006; Zhang, 2004a; Steinwart, 2007).

Definition 3 .4 (Calibration)**.** *We say that S is calibrated to L under the decoding d if for every probability distribution $\rho \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ there exists a non-negative convex function $\zeta : \mathbb{R} \to \mathbb{R}$ such that*

$$\zeta(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star), \tag{1.8}$$

satisfying $\zeta(0) = 0$ and $\zeta(\varepsilon) > 0$ for all $\varepsilon > 0$.

Note that calibration trivially implies Fisher consistency. From the form of ζ in Definition 3 .4, it can easily be seen that the function is invertible for non-negative values and the inequality can also be written as $\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) \leq \zeta^{-1}(\mathcal{R}(g) - \mathcal{R}(g^\star))$. Inequalities of the form (1.8) are called *comparison inequalities* and are useful to translate performance guarantees of the surrogate problem into the original task. In particular, it directly follows that if S is calibrated to L under the decoding d and $(g_t)_{t \geq 1}$ is a sequence of surrogate predictors the following is satisfied

$$\lim_{t \to \infty} \mathcal{R}(g_t) = \mathcal{R}(g^\star) \implies \lim_{t \to \infty} \mathcal{E}(d \circ g_t) = \mathcal{E}(f^\star).$$

As we will see in the next section, Fisher consistency does not always imply calibration for a given decoding d. On the other hand, if Fisher consistency holds for a decoding, in general it is possible to design decodings for which calibration also holds.

Example 3 .5. **Calibration for the quadratic surrogate.** By construction, the quadratic surrogate method introduced in Example 3 .3 is Fisher consistent. One can also prove calibration:

$$\begin{aligned}
&\mathcal{E}(\mathrm{sign}(g)) - \mathcal{E}(f^\star) \\
&= \mathbb{E}_{x \sim \rho_\mathcal{X}} \{1_{g^\star(x) g(x) < 0} |g^\star(x)|\} && \text{(Theorem 2.2 of (Devroye et al., 1996))} \\
&\leq \mathbb{E}_{x \sim \rho_\mathcal{X}} |g^\star(x) - g(x)| && (ab \leq 0 \implies |a| \leq |a - b|) \\
&\leq \sqrt{\mathbb{E}_{x \sim \rho_\mathcal{X}} (g^\star(x) - g(x))^2} && \text{(Cauchy-Schwartz)} \\
&= \sqrt{\mathbb{E}_{(x,y) \sim \rho} (g(x) - y)^2 - \mathbb{E}_{(x,y) \sim \rho} (g^\star(x) - y)^2} && (g^\star(x) = \mathbb{E}_{y \sim \rho_x} y) \\
&= \sqrt{\mathcal{R}(g) - \mathcal{R}(g^\star)},
\end{aligned}$$

where $\rho_\mathcal{X}$ stands for the marginal distribution of ρ over the input space $\mathcal{X}$. Thus, the comparison inequality (1.8) holds for $\zeta(\varepsilon) = \varepsilon^2$.

$\mathcal{G}$-consistency. The consistency notions of the surrogate method we have introduced so far deal with the relationship between both excess risks around its exact minimizers $f^\star$ and $g^\star$. In practice, however, the surrogate risk (1.4) is minimized over a surrogate hypothesis space $\mathcal{G}$ not containing the surrogate Bayes predictor $g^\star$. In this case, we would like to know whether $f^\star$ can be found by minimizing the surrogate risk under the assumption that there exists $g \in \mathcal{G}$ satisfying $f^\star = d \circ g$, i.e., $f^\star$ belongs to the hypothesis space $\mathcal{F}$ associated to $\mathcal{G}$. This is formalized with the notion of $\mathcal{G}$-consistency (Long and Servedio, 2013).

Definition 3 .6 ($\mathcal{G}$-consistency)**.** *We say that S is $\mathcal{G}$-consistent to L under the decoding d if for every probability distribution $\rho \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ the following is satisfied:*

$$\mathcal{E}(f^\star) = \inf_{g \in \mathcal{G}} \mathcal{E}(d \circ g) \implies \mathcal{E}(f^\star) = \mathcal{E}(d \circ g^\star_{\mathcal{G}}),$$

where $g^\star_{\mathcal{G}}$ is the minimizer of the surrogate risk over $\mathcal{G}$, i.e., $\mathcal{R}(g^\star_{\mathcal{G}}) = \min_{g \in \mathcal{G}} \mathcal{R}(g)$.

Note that $\mathcal{G}$-consistency corresponds to Fisher consistency when $\mathcal{G}$ is the space of measurable functions. Unfortunately, convex surrogate methods are in general not $\mathcal{G}$-consistent with respect to an hypothesis space other than the full set of measurable functions. For instance, if $f^\star$ corresponds to a linear decision boundary, in general it cannot be recovered using a convex surrogate method with affine surrogate estimators. Note that this is not surprising, as minimizing $\mathcal{E}(d \circ g)$ over $\mathcal{G}$ is known to be NP-Hard for most non-trivial class of functions (Arora et al., 1997). On the other hand, $\mathcal{G}$-consistency holds for scale-invariant spaces $\mathcal{G}$ [6] and certain non-convex surrogate losses which are Lipschitz continuous approximations of the discrete loss such as the ramp and probit surrogate losses (see Figure 1.3) (Keshet and McAllester, 2011).

Example 3 .7 ($calG$-consistency of the non-convex ramp loss). The ramp loss defined as $S(v, y) = \min(1, \max(1 - yv, 0))$ satisfies $\lim_{\gamma \to 0} S(v/\gamma, y) = L(\mathrm{sign}(v), y)$. Thus, if $\mathcal{G}$ is scale invariant, then it directly follows $\inf_{g \in \mathcal{G}} \mathcal{E}(\mathrm{sign}(g)) = \inf_{g \in \mathcal{G}} \mathcal{R}(g)$ and so $\mathcal{G}$-consistency is satisfied. Moreover, as it upper-bounds the 0-1 loss as $L(\mathrm{sign}(v), y) \leq S(v, y)$, the following comparison inequality holds:

$$\mathcal{E}(\mathrm{sign}(g)) - \mathcal{E}(f^\star) \leq \mathcal{R}(g) - \inf_{g' \in \mathcal{G}} \mathcal{R}(g').$$

Thus, the ramp loss is $\mathcal{G}$-consistent to the 0-1 loss.

$\mathcal{G}$-consistency can hold for convex surrogates under strong conditions on the data distribution. For instance, if the problem is realizable in $\mathcal{G}$, i.e., there exists $g \in \mathcal{G}$ for which $\mathcal{E}(d \circ g) = 0$, then there are convex surrogates for which $\mathcal{G}$-consistency holds for scale-invariant $\mathcal{G}$. This is the case for the hinge loss and logistic in binary classification and the corresponding extensions to multi-class (Long and Servedio, 2013). It is an interesting open problem to characterize or provide useful sufficient conditions on the data distribution for $\mathcal{G}$-consistency to hold for specific convex surrogate losses.

Theoretical guarantees of surrogate-based estimators. As we have seen, the properties of surrogate methods described above are key to derive theoretical guarantees of surrogate-based estimators. Let's now make the use of these properties more precise depending on the setting of the learning problem.

(i) **The surrogate estimator converges to $g^\star$ and/or $g^\star \in \mathcal{G}$.** This is the most favorable setting. In particular, if the surrogate method is calibrated, then one can readily deduce convergence on the discrete predictors from convergence on the surrogate risks as $\lim_{t \to \infty} \mathcal{E}(d \circ g_t) = \mathcal{E}(f^\star)$. Moreover, if $g^\star \in \mathcal{G}$, generalization

[6] $\mathcal{G}$ is said to be scale-invariant if $\alpha g \in \mathcal{G}$ for every $g \in \mathcal{G}$ and $\alpha \in \mathbb{R}$.

bounds can be obtained by means of the *comparison inequality* (1.8). Note, however, that comparing risks is an *undirect* strategy to obtain guarantees on the excess risk $\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star)$. Fast rates under low-noise settings are often obtained by directly bounding this quantity using convergence directly on the surrogate predictors and not on the surrogate risk (Audibert and Tsybakov, 2007).

(ii) **There exists $g \in \mathcal{G}$ such that $f^\star = d \circ g$, i.e., $f^\star \in \mathcal{F}$.** In this setting, proving convergence to the Bayes predictor $f^\star$ requires $\mathcal{G}$-consistency, which in general is only possible for intractable (non-convex) surrogate methods. Some works prove convergence to $f^\star$ in this setting without assuming $g^\star \in \mathcal{G}$ under the assumption that there exists a $\lambda > 0$, for which $g_\lambda = \arg\min_{g \in \mathcal{G}} \mathcal{R}(g) + \frac{\lambda}{2}\|g\|^2_{\mathcal{G}} \in \mathcal{G}$ satisfies $f^\star = d \circ g_\lambda$ (Koltchinskii and Beznosova, 2005; Pillaud-Vivien et al., 2018a). Another option is to show that $\mathcal{G}$-consistency holds for a specific subset of data distributions containing the learning problem of study.

(iii) **There is no $g \in \mathcal{G}$ such that $f^\star = d \circ g$, i.e., $f^\star \notin \mathcal{F}$.** In this case, there is no hope in obtaining convergence to the Bayes predictor as $\mathcal{E}(f^\star) \neq \inf_{g \in \mathcal{G}} \mathcal{E}(d \circ g)$ and one would like to obtain guarantees relative to the best predictor in $\mathcal{G}$ as

$$\mathcal{E}(d \circ g_n) - \inf_{g \in \mathcal{G}} \mathcal{E}(d \circ g).$$

These results can be obtained using non-convex surrogates such as the one described in Example 3 .7, but remain unaccessible for tractable estimators.

Beyond margin bounds. There exists a vast literature in discrete prediction on the so-called *margin bounds* (Mohri et al., 2018) (see also Taskar et al. (2004); Cortes et al. (2016); London et al. (2016)). We argue that even though the practice of analyzing surrogate methods using margin bounds is still quite extended in today's theoretical research in structured prediction, the obtained guarantees are vacuous for analyzing tractable (i.e., convex) surrogate methods. In a nutshell, these bounds are obtained by bounding the loss with a $1/\gamma$-Lipschitz-continuous *non-convex* surrogate loss S^γ such that $L(d(v), y) \leq S_\gamma(v, y)$ for all $\gamma > 0$, such as the ramp loss defined in Figure 1.3. Then, the classical methodology is to prove generalization bounds of estimators g_n minimizing the penalized functional $\mathcal{R}^\gamma_n(g) + \lambda\Omega(g)$, $\lambda > 0$, where the penalization term controls the capacity of the estimator and $\mathcal{R}^\gamma_n$ is the empirical version of the non-convex risk $\mathcal{R}^\gamma(g) = \mathbb{E}_{(x,y)\sim\rho}\ S^\gamma(g(x), y)$. As computing g_n is intractable due to non-convexity, convex surrogate losses are usually justified by further upper bounding the non-convex risk by their corresponding convex risk. However, we stress that all theoretical guarantees are lost in this last step, as minimizing an upper bound of the risk does not guarantee its minimization.

Probabilistic and non-probabilistic estimators. We now make a key distinction between two types of surrogate losses which is central to this thesis. Recall that the Bayes estimator $f^\star$ can be characterized point-wise in x as the minimizer of the conditional risk. We classify surrogate methods into two groups depending on whether the surrogate predictor g is also an estimator of the conditional risk $\mathbb{E}_{y' \sim \rho_x}\ L(y, y')$ or not.

Two strategies for estimating the Bayes predictor $f^\star$

$$f^\star(x) = \underbrace{\arg\min_{y \in \mathcal{Y}} \overbrace{\mathbb{E}_{y' \sim \rho_x} L(y, y')}^{\text{probabilistic estimators}}}_{\text{non-probabilistic estimators}} . \tag{1.9}$$

- **Probabilistic estimators.** In the literature they are also called *plug-in estimators* because the discrete estimator is constructed by plugging-in an estimator of the conditional distribution ρ_x into the right hand side of Eq. (1.9). We call it *probabilistic estimators* because they estimate the conditional expected risk $x \mapsto (\mathbb{E}_{y' \sim \rho_x} L(y, y'))_{y \in \mathcal{Y}}$, and as we will see throughout this thesis, it does not all always correspond to estimating the conditionals ρ_x. More precisely, a surrogate loss S is a probabilistic estimator if $g^\star$ can be mapped to the conditional risk as $g^\star(x) \mapsto (\mathbb{E}_{y' \sim \rho_x} L(y, y'))_{y \in \mathcal{Y}}$ for all $x \in \mathcal{X}$. Note that these estimators are solving a *harder* problem than classification, as estimating the conditional risk is strictly more than estimating the decision boundaries. In general, if the associated surrogate loss S is convex, these types of surrogates are **smooth** and the comparison inequality (1.8) is quadratic, i.e., $\zeta(\varepsilon) \propto \varepsilon^2$ as in Example 3 .5. Part II of this thesis focuses on this type of surrogates. More specifically, in Chapter 2 we analyze both statistically and computationally the quadratic surrogate for structured prediction by leveraging the low dimensional structure of the loss, and in Chapter 3 we go beyond the quadratic setting and derive a general theory for smooth convex surrogates based on regression of low dimensional representations of the discrete conditional risk.
- **Non-probabilistic estimators.** These type of predictors directly tackle the classification task, i.e., the decision boundary, without estimating the conditional risk as an intermediate step. If the associated surrogate loss S is convex, these types of surrogates are generally **non-smooth** and the calibration function ζ is linear, i.e., $\zeta(\varepsilon) \propto \varepsilon$. Part III of this thesis focuses on polyhedral surrogates, which fall into this category. More specifically, in Chapter 4 we study the Max-Min loss (see Figure 1.3) in the context of structured prediction and derive an efficient algorithm with generalization guarantees to the Bayes predictor. In Chapter 5 we provide both necessary and sufficient conditions for Fisher consistency of the structured Max loss (see Figure 1.3) and we derive a novel loss called Restricted-Max loss partially overcoming these consistency limitations and maintaining the maximization structure.

Example 3 .8 (Binary probabilistic and non-probabilistic estimators). Recall that in binary classification $f^\star(x) = \operatorname{sign}(\mathbb{E}_{y \sim \rho_x} y)$, where $\mathbb{E}_{y \sim \rho_x} y = 2\rho_x(1) - 1$ and $\rho_x(1)$ is the conditional probability of $y = 1$ given x. In this case, probabilistic estimators are the ones estimating the conditional $\rho_x(1)$ and non-probabilistic estimators are the ones directly estimating $f^\star(x)$. From tables in Figure 1.3, all binary *smooth convex*

surrogates are probabilistic estimators. For instance, the conditional distribution can be written in terms of the Bayes surrogate predictor as $\rho_x(1) = (g^\star(x) - 1)/2$ for the quadratic surrogate and $\rho_x(1) = (1+e^{-g^\star(x)})^{-1}$ for logistic. On other hand, the *convex non-smooth* hinge and Max-Min are non-probabilistic estimators and satisfy $g^\star(x) = f^\star(x)$ and $g^\star(x) = f^\star(x)/2$ respectively.

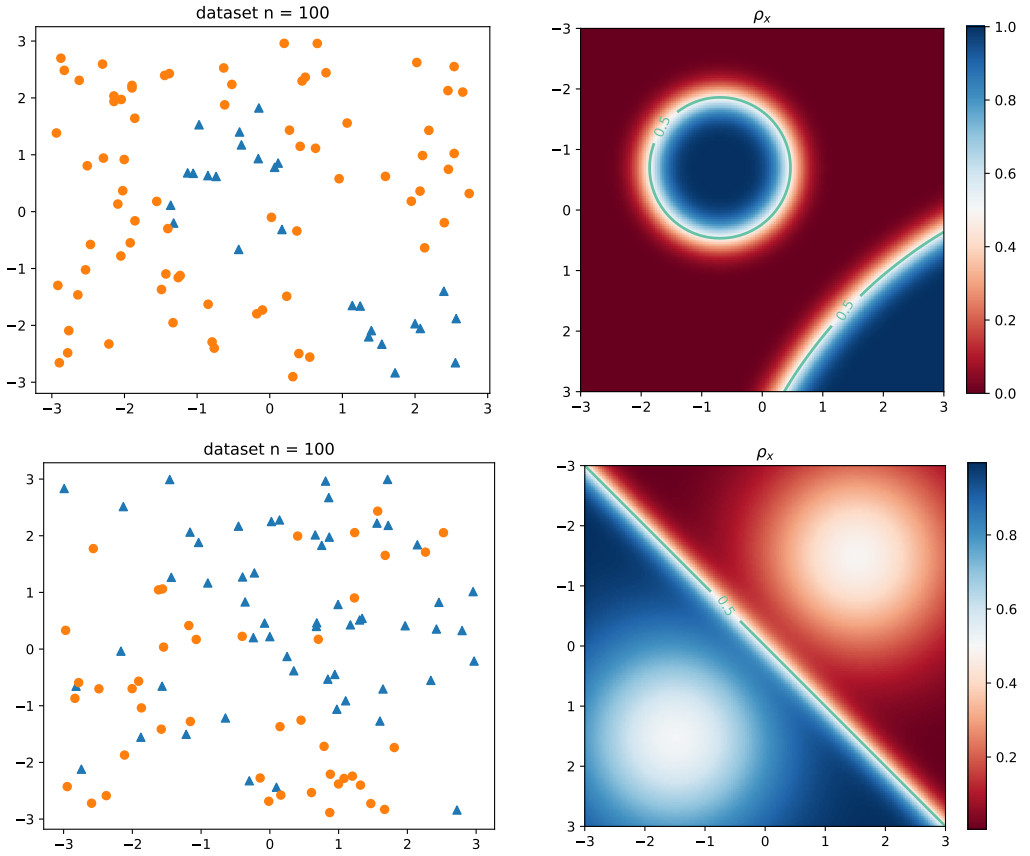


Figure 1.4: Two binary classification settings in $\mathcal{X} = [-3, 3]^2$ with uniform marginal distribution over the input space. We depict on the left $n = 100$ samples of the dataset and on the right the conditional distribution $\rho_x(1) \in [0, 1]$. **Top:** The problem has low-noise but $f^\star$ defines two disconnected regions which cannot be recovered with linear decision boundaries. **Bottom:** $f^\star$ is a linear decision boundary but there is a lot of mass at points where $\rho_x \approx 0.5$.

When is a discrete prediction problem easy/hard? The difficulty of learning discrete prediction tasks comes essentially from two different sources:

- **Shape of the decision boundary $f^\star$.** If the decision boundary is hard to approximate as $d \circ g$ with a function g from the hypothesis space $\mathcal{G}$, then the approximation error will be large. Note that contrary to regression problems, where typically the regularity of the function to predict charaterizes the difficulty of the problem, it is

not clear whether the regularity of the decision boundary in $\mathcal{X}$ is important for having small approximation error. A remarkable example is given by Steinwart and Scovel (2007), where fast convergence rates are obtained for binary classification with smooth surrogate functions independently of the regularity of the decision boundary.

- **Large noise $\mathcal{E}(f^\star)$.** The learning rate $\gamma > 0$ of a discrete estimator converging at $\mathcal{O}(n^{-\gamma})$ highly depends on the amount of noise present in the learning task. This is in contrast to regression, where the noise generally appears as a multiplicative constant in the learning bounds. For instance, in binary classification where $f^\star(x) = \operatorname{sign}(2\rho_x(1) - 1)$, the amount of mass for which $\rho_x(1) \approx 0.5$ plays an important role in the learning rates (Tsybakov, 2004).

4 Point-wise Analysis of Surrogate Methods

In the previous section we presented surrogate methods in its generality and their key required properties to guarantee the surrogate strategy is theoretically sound. In this section, we study Fisher consistency and calibration using a point-wise analysis on the surrogate losses S and decodings d. In particular, two main types of losses we already discussed above naturally show up in the analysis: *probabilistic estimators*, which essentially require a form of smoothness of the loss, and *non-probabilistic estimators*, which estimate directly the decision boundaries. By now, we do not care about the potential exponentially large size of the output space appearing in structured prediction - this will be addressed in the following Section 5 .

Point-wise analysis. What makes Fisher consistency and calibration easy to analyze is the fact that $f^\star(x)$ and $g^\star(x)$ depend on $x \in \mathcal{X}$ only through the conditional distribution $\rho_x \in \Delta_\mathcal{Y}$, where $\Delta_\mathcal{Y}$ is the $|\mathcal{Y}|$-dimensional simplex. Thus, we can remove the dependence of the input x and study these properties in terms of distributions $q \in \Delta_\mathcal{Y}$ over the output space. As we will see, it turns out that the resulting mathematical objects can be studied using convex analysis.

> **We drop the dependence on x in the conditional distribution ρ_x and write q.**

Let's now define the mathematical objects associated to the losses L and S under the point-wise framework. Let $L_y^\top q = \mathbb{E}_{y' \sim q}\ L(y, y')$ be the conditional risk of the discrete loss L where $L_y = (L(y, y'))_{y' \in \mathcal{Y}}$ is the y-th row of the loss matrix L. We define the *Bayes risk* $\ell : \Delta_\mathcal{Y} \to \mathbb{R}$ and the *excess conditional risk* $\delta\ell : \mathcal{Y} \times \Delta_\mathcal{Y} \to \mathbb{R}$ as

$$\ell(q) = \min_{y' \in \mathcal{Y}}\ L_{y'}^\top q, \qquad \delta\ell(y, q) = L_y^\top q - \ell(q).$$

The Bayes risk ℓ is the minimum possible error measured by L when the output is distributed as q, and the excess risk $\delta\ell$ is the error relative to the Bayes risk for a given prediction y. Note that ℓ is a *concave polyhedral* function as it is defined as the point-wise minimum of finite $|\mathcal{Y}|$ linear functions. We also define the *optimal predictor* $y^\star : \Delta_\mathcal{Y} \to 2^\mathcal{Y}$ [7] as

$$y^\star(q) = \arg\min_{y' \in \mathcal{Y}}\ L_{y'}^\top q \subseteq \mathcal{Y},$$

which corresponds to the set of output elements optimal for the distribution q. In particular, the Bayes error and the excess risk can be written in terms of these quantities as $\mathcal{E}(f^\star) = \mathbb{E}_x\ \ell(\rho_x)$ and $\mathcal{E}(f) - \mathcal{E}(f^\star) = \mathbb{E}_x\, \delta\ell(f(x), \rho_x)$.

Let's now define the objects for the surrogate loss. Let $S(v)^\top q = \mathbb{E}_{y' \sim q}\, S(v, y')$ be the conditional surrogate risk, where $S(v) = (S(v, y))_{y \in \mathcal{Y}}$. We analogously define the *Bayes surrogate risk* $s : \Delta_\mathcal{Y} \to \mathbb{R}$ and the *conditional excess surrogate risk* $\delta s : \mathbb{R}^k \times \Delta_\mathcal{Y} \to \mathbb{R}$ as

$$s(q) = \min_{v' \in \mathbb{R}^k}\ S(v')^\top q, \qquad \delta s(v, q) = S(v)^\top q - s(q).$$

[7] The set 2^A denotes the power set of the set A.

Similarly, s is *concave* (but not polyhedral) as it is defined as the point-wise minimum of a *non-finite* set of linear functions. The *surrogate optimal predictor* $v^\star : \Delta_\mathcal{Y} \to 2^{\mathbb{R}^k}$ reads

$$v^\star(q) = \arg\min_{v' \in \mathbb{R}^k} S(v')^\top q \subseteq \mathbb{R}^k.$$

The Bayes surrogate error and the excess surrogate risk can be written in terms of these quantities as $\mathcal{R}(g^\star) = \mathbb{E}_x\, s(\rho_x)$ and $\mathcal{R}(g) - \mathcal{R}(g^\star) = \mathbb{E}_x\, \delta s(g(x), \rho_x)$.

Prediction sets. The polyhedral concave structure of the Bayes risk ℓ naturally defines a partition of the simplex defined by the affine regions of ℓ. More concretely, we define the *prediction sets* $\Delta_L : 2^\mathcal{Y} \to 2^{\Delta_\mathcal{Y}}$ as

$$\Delta_L(Y) = \{q \in \Delta_\mathcal{Y} \mid Y \subseteq y^\star(q)\} \subseteq \Delta_\mathcal{Y}.$$

The prediction set $\Delta_L(Y)$ associated to a subset $Y \subseteq \mathcal{Y}$ corresponds to the probability vectors $q \in \Delta_\mathcal{Y}$ for which all elements in Y are optimal. When $Y = \{y\}$ is a singleton we just write $\Delta_L(y)$. By construction, the simplex can be decomposed into the union of $|\mathcal{Y}|$ prediction sets as $\cup_{y \in \mathcal{Y}} \Delta_L(y) = \Delta_\mathcal{Y}$.

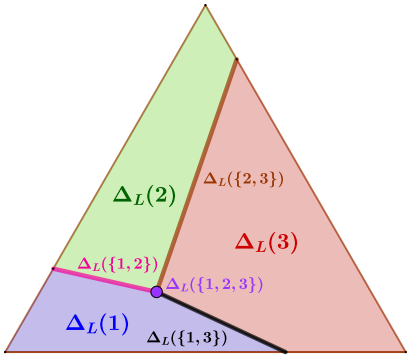


Figure 1.5: Visualization of the prediction sets in the simplex of a generic discrete loss over an output space with three elements. Note that $q \in \Delta_L(Y)$ if and only if $Y \subseteq y^\star(q)$.

> *Remark* 4 .1 (Power diagram associated to L). The finite set $\mathcal{P}_L = \{\Delta(y^\star(q)\}_{q \in \Delta_\mathcal{Y}}$ defines a *cell complex* in the simplex, i.e., a set of faces which (i) union to $\Delta_\mathcal{Y}$, (ii) have pairwise disjoint relative interiors and (iii) any nonempty intersection of faces F, F' in $\mathcal{P}_L$ is a face of F and F' and an element of $\mathcal{P}_L$. Moreover, this cell complex can be shown to be a *power diagram* (Aurenhammer, 1987; Finocchiaro et al., 2019), which is a generalized version of a Voronoi diagram.

In the following, we provide the point-wise versions of the notions of Fisher consistency and calibration introduced in the previous section.

Fisher consistency. Recall that Fisher consistency states that $\mathcal{E}(f^\star) = \mathcal{E}(d \circ g^\star)$ for any data distribution ρ. This can be re-written using the point-wise objects introduced above as $\mathbb{E}_x\, \delta\ell(d \circ v, \rho_x) = 0$ for any $v \in v^\star(\rho_x)$ and ρ. It can readily be seen that this is equivalent to the point-wise characterization provided by the following Proposition 4 .2.

Proposition 4 .2 (Point-wise characterization of Fisher consistency)**.** *S is Fisher consistent to L under the decoding d if and only if*

$$v \in v^\star(q) \implies d(v) \in y^\star(q), \quad \forall q \in \Delta_\mathcal{Y}. \tag{1.10}$$

Put it simply, every minimizer of the conditional surrogate risk can be mapped to a minimizer of the conditional risk through a decoding function.

Calibration. Calibration has a similar point-wise characterization as Fisher consistency given by the following proposition.

Proposition 4 .3 (Point-wise characterization of calibration)**.** *S is calibrated to L under the decoding d if and only if:*

$$\inf_{\substack{v\in\mathbb{R}^k \\ y=d(v)}} \delta s(v,q) = 0 \implies y = d(v) \in y^\star(q), \quad \forall y \in \mathcal{Y}, \forall q \in \Delta_\mathcal{Y}.$$

See Theorem 3 of Zhang (2004a) for a proof of this result in the context of multi-class classification. In other words, y is an optimal prediction if the excess surrogate risk can be made zero under the constraint that the decoding predicts output y. Note that calibration trivially implies Fisher consistency.

Remark 4 .4. Fisher consistency and calibration are not equivalent for a fixed decoding d. Note that Fisher consistency only depends on the decoding restricted to the set $V = \{v \in \mathbb{R}^k \mid v \in v^\star(q), q \in \Delta_\mathcal{Y}\}$. In particular, if $V \neq \mathbb{R}^k$, the decoding can be extended to $\mathbb{R}^k$ in such a way that calibration does not hold.

A comparison inequality relating both excess risks can be computed from point-wise quantities by means of the so-called *calibration function* (Steinwart, 2007) defined as,

$$\zeta(\varepsilon) = \inf_{\substack{v\in\mathbb{R}^k \\ q\in\Delta_\mathcal{Y}}} \delta s(v,q) \quad \text{s.t.} \quad \delta\ell(d(v),q) \geq \varepsilon. \tag{1.11}$$

Note that the minimization problem defining the calibration function is generally not jointly convex with respect to v and q and it is straightforward from Proposition 4 .3 that if calibration holds then $\zeta(0) = 0$ and $\zeta(\varepsilon) > 0$ for all $\varepsilon > 0$, i.e., δs cannot be made zero if $\delta\ell$ is bounded away from zero. The following result, which corresponds to Theorem 2.13 by Steinwart (2007), establishes the link between the calibration function (1.11) and a comparison inequality between the risks.

Proposition 4 .5. *If S is calibrated to L under the decoding d, the Fenchel bi-conjugate ζ^{**} [8] of the calibration function gives a convex comparison inequality of excess risks as in* (1.8)*.*

The Fenchel bi-conjugate of ζ corresponds precisely to its lower convex envelope in the non-negative real numbers, thus maintaining the property $\zeta^{**}(\varepsilon) > 0$ for $\varepsilon > 0$. Proposition 4 .5 follows from Jensen's inequality as

$$\zeta^{**}(\underbrace{\mathbb{E}_x\, \delta\ell(d\circ g(x), y)}_{\mathcal{E}(d\circ g)-\mathcal{E}(f^\star)}) \leq \mathbb{E}_x\, \zeta^{**}(\delta\ell(d\circ g(x), y)) \leq \underbrace{\mathbb{E}_x\, \delta s(g(x), y)}_{\mathcal{R}(g)-\mathcal{R}(g^\star)},$$

[8] The Fenchel conjugate (Rockafellar, 1997) of a function $h : \mathcal{C} \to \mathbb{R}$ is the convex function h^* defined as $h^*(v) = \sup_{u\in\mathcal{C}}\ v^\top u - h(u)$. The Fenchel bi-conjugate h^{**} of h is $(h^*)^*$.

where we have used that $\zeta^{**}(\delta\ell(d(v),q)) \leq \zeta(\delta\ell(d(v),q)) \leq \delta s(v,q)$ by construction. Note that the same holds for any convex lower bound of the calibration function. In particular, most comparison inequalities we will obtain are lower bounds of ζ, which is in general hard to compute exactly.

We now present the two main types of losses, i.e., probabilistic and non-probabilistic estimators, but restricting ourselves to losses with a particular structure for the sake of presentation. The general case will be briefly discussed at the end of this section with references into the corresponding parts of this thesis in which are studied.

Fenchel-Young (FY) surrogate losses. Fenchel-Young losses (Blondel et al., 2020) are a central type of surrogate losses constructed using Fenchel conjugates, which makes them easy to analyze. Let $\Omega : \mathcal{C} \to \mathbb{R}$ be a convex function defined in a *compact* [9] convex set $\mathcal{C} \supset \Delta$ containing the simplex. The Fenchel-Young loss associated to Ω is defined as

$$S(v, y) = \Omega^*(v) - v_y, \tag{1.12}$$

where $\Omega^*(v) = \sup_{u \in \mathcal{C}} \; u^\top v - \Omega(u)$ is the Fenchel-conjugate of the function Ω. Note that the surrogate loss S is defined in the vector space $\mathbb{R}^k$ due to the compactness of the domain $\mathcal{C}$ and it is lower bounded by $-\Omega(e_y) > -\infty$ because of the Fenchel-Young inequality [10] (Rockafellar, 1997). The key property of these surrogates is that the Bayes risk can be written as

$$s(q) = \min_{v \in \mathbb{R}^k} \; \Omega^*(v) - v^\top q = -\Omega(q), \quad \forall q \in \Delta_\mathcal{Y}. \tag{1.13}$$

Moreover, the set of surrogate predictors has the following form:

$$v^\star(q) = \partial\Omega(q), \quad \forall q \in \Delta_\mathcal{Y}, \tag{1.14}$$

where $\partial h(q) \subseteq \mathbb{R}^k$ denotes the subgradient of the convex function h at the point q.

FY smooth surrogates are probabilistic estimators. These losses are also known in the literature as proper composite losses with canonical link (Williamson et al., 2016). If Ω^* is smooth, then the sub-gradient is a singleton $\partial\Omega^*(v) = \{\nabla\Omega^*(v)\}$. Moreover, by Fenchel duality, we have that $\nabla\Omega(v) = q$ if and only if $v \in v^\star(q) = \partial\Omega(q)$. Hence, the decoding defined as

$$d(v) = y^\star(\nabla\Omega^*(v)) = \arg\min_{y \in \mathcal{Y}} \; L_y^\top \nabla\Omega^*(v), \tag{1.15}$$

is Fisher consistent to L by construction. In Eq. (1.15), the ties (i.e., $|y^\star(q)| > 1$) are broken arbitrarily. Moreover, if Ω^* is β-smooth [11], then we can also prove calibration using the decoding Eq. (1.15). More specifically, a direct application of the general results provided in Chapter 3 to this particular case shows that the calibration function can be lower bounded as

$$\zeta(\varepsilon) \geq \frac{\varepsilon^2}{8c_L^2\beta},$$

[9] The construction also works for functions Ω such that $\forall v \in \mathbb{R}^k$ exists $u \in \mathcal{C}$ such that $v \in \partial\Omega(u)$. In particular, this is always satisfied if the domain of Ω is compact and convex, where the sub-gradients at the boundary correspond to the normal cone.

[10] The Fenchel-Young inequality states that $h(u) + h^*(v) \geq u^\top v$ for all $u \in \mathrm{dom}(h), v \in \mathrm{dom}(h^*)$.

[11] We say that a differentiable function $h : \mathbb{R}^k \to \mathbb{R}$ is β-smooth with respect to a norm $\|\cdot\|$ if it holds $|h(v) - h(v')| \leq \beta\|v - v'\|$ for all $v, v' \in \mathbb{R}^k$.

where c_L is a constant only depending on the discrete loss L. As we will see later, quadratic-type calibration functions are a feature of probabilistic estimators. The classical example of smooth FY loss is when $\Omega(q) = \sum_{y\in\mathcal{Y}} q_y \log q_y$ is the negative Shannon entropy, defined in the simplex $\Delta_\mathcal{Y}$. The resulting loss is the well known multinomial logistic loss defined as $S(v, y) = \log\Big(\sum_{y'\in\mathcal{Y}} \exp v_{y'}\Big) - v_y$. In this particular case, when L is the 0-1 loss the decoding (1.15) can be made independent of the mapping $\nabla\Omega^*$ as it holds $d(v) = \arg\max_{y'\in\mathcal{Y}} v_{y'} = y^\star(\nabla\Omega^*(v))$ [12].

Remark 4 .6 (Loss-independent decodings). As smooth FY losses are Fisher consistent to any discrete loss, it is the decoding d that contains all information about the discrete task (see Eq. (1.15)). However, this is not desirable in practice as it is preferred that the surrogate loss contains all the information about the task so that the minimization of the expected surrogate risk is guided to minimize the discrete loss of interest (Lacoste-Julien et al., 2011), and thus make the approximation error smaller. This can be easily achieved in cost-sensitive binary classification (Scott, 2012), but remains an open question to do it systematically in the structured prediction case.

FY polyhedral surrogates are non-probabilistic estimators. Strictly speaking, non-probabilistic FY surrogates are those for which Ω^* is non-smooth. In this thesis, however, we we will implicitly think of them as polyhedral. This is because polyhedral losses are at the other end of the spectrum: they cannot be used to estimate conditional probabilities in any neighborhood of the simplex. Let Ω^* be a polyhedral function, i.e., a function that can be written as a point-wise maximum of a finite number of affine functions. Note that if Ω^* is polyhedral, then Ω is also polyhedral in $\mathcal{C}$. While consistency is trivially met from the smoothness of Ω^*, this is not anymore the case for polyhedral surrogates as $\partial\Omega^*(v)$ is not always a singleton. Let Ω be the polyhedral convex function defined as $\Omega(q) = -\ell(q) + i_{\Delta_\mathcal{Y}}(q)$ [13] for all $q \in \Delta_\mathcal{Y}$. If q is in the interior of the simplex $\Delta_\mathcal{Y}$, then

$$v^\star(q) = \partial\Omega(q) = -\operatorname{hull}(\{L_y\}_{y\in y^\star(q)}),$$

where hull stands for the convex hull. Under the argmax decoding $d(v) = \arg\max_{y'\in\mathcal{Y}} v_{y'}$ Fisher consistency holds in the interior of the simplex. To see this, let $v = -\sum_{y\in y^\star(q)} \alpha_y L_y$ in $v^\star(q)$ with $\alpha^\top 1 = 1, \alpha \succeq 0$. Then:

$$d(v) = \arg\max_{y'\in\mathcal{Y}} \; -\sum_{y\in y^\star(q)} \alpha_y L(y, y') = \arg\min_{y\in y^\star(q)} \sum_{y\in y^\star(q)} \alpha_y L(y, y') \in y^\star(q),$$

whenever $L(y, y') = 0 \iff y = y'$ (i..e, L is not degenerate). The resulting FY loss is the so-called *Max-Min loss*, which was introduced by Fathony et al. (2016) for cost-sensitive learning and in the general structured setting in Chapter 4. It takes the following form:

$$S(v, y) = \underbrace{\max_{q\in\Delta_\mathcal{Y}} \min_{y'\in\mathcal{Y}} L_{y'}^\top q + v^\top q}_{\Omega^*(v) = (-\ell + i_{\Delta_\mathcal{Y}})^*(v)} - v_y. \tag{1.16}$$

[12] This statement for the 0-1 loss can be extended to any functions Ω which are Schur-convex.

[13] Let A be a set. The function i_A is defined as 0 if $u \in A$ and ∞ otherwise.

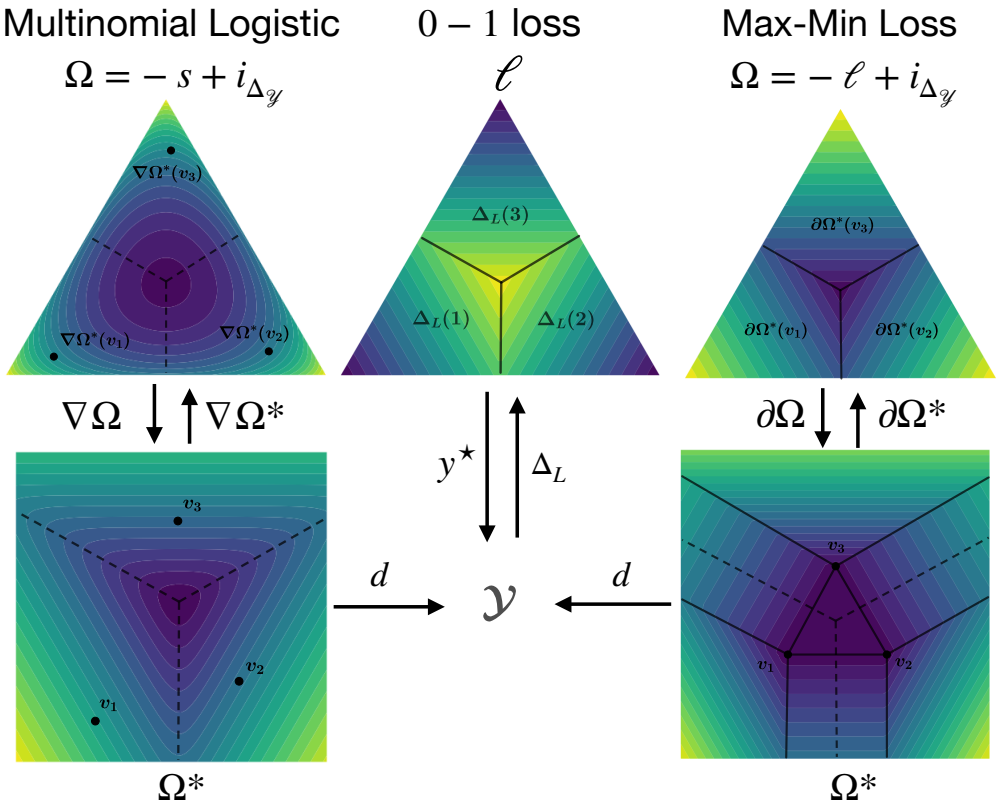


Figure 1.6: Diagram of the smooth multinomial logistic and polyhedral Max-Min Fenchel-Young surrogate losses and the discrete 0-1 loss for an output space of three elements. The dashed lines correspond to the decision boundaries of the argmax function d that make both losses Fisher consistent. **Middle:** Visualization of the Bayes risk $\ell(q) = 1 - \|q\|_\infty$ of the 0-1 loss and its prediction sets $\{\Delta_L(j)\}_{j=1}^3$ partitioning the simplex. **Left:** Visualization of the function $\Omega(q) = \sum_{j=1}^3 q_j \log q_j$ and its Fenchel-conjugate $\Omega^* = \log(\sum_{j=1}^3 e^{v_j})$ in the affine space $v^\top 1 = 0$. The gradient mapping $\nabla\Omega^*$ is a bijection between $v^\top 1 = 0$ and the simplex. **Right:** Visualization of the function $\Omega(q) = \|q\|_\infty - 1 + i_{\Delta_3}(q)$ and its Fenchel-conjugate Ω^* in $v^\top 1 = 0$. The sub-gradient mapping at the non-differentiable points $\{v_j\}_{j=1}^3$ correspond precisely to the convex prediction sets $\{\Delta_L(j)\}_{j=1}^3$ of the 0-1 loss.

This loss is studied in detail in Chapter 4 of this thesis, where we rigorously show that calibration holds under the argmax decoding. In particular, it can be shown that under mild assumptions on the discrete loss L, the calibration function under the argmax decoding $d(v) = \arg\max_{y\in\mathcal{Y}} v_y$ can be lower bounded by a linear function as

$$\zeta(\varepsilon) \geq \frac{\varepsilon}{c_L},$$

where c_L is a constant depending on the discrete loss L. This result corresponds to Theorem 3 .3 in Chapter 4.

Finally, we briefly discuss about the general case where the surrogate loss does not necessarily have a FY structure.

Non-FY probabilistic estimators. In contrast to FY losses, smooth losses without a FY structure do not necessary correspond to probabilistic estimators. More specifically, the required smoothness is not on the surrogate loss S but rather in the convex function $\Omega^* = (-s + i_{\Delta_\mathcal{Y}})^*$ [14], where recall that s stands for the surrogate Bayes risk. Note that both forms of smoothness are equivalent for FY losses. In the general case with smooth Ω^*, the decoding is not computed using the gradient $\nabla\Omega^*$ but rather the inverse of the function $v^\star$. Moreover, it is known that these losses have a composite proper structure in the set of minimizers $V = \{v \in \mathbb{R}^k \mid v \in v^\star(q), q \in \Delta_\mathcal{Y}\}$ (Williamson et al., 2016) (which we also name Bregman representation in Chapter 3), i.e., the excess risk can be written as $\delta s(v, q) = D_{-s}(q, t(v))$ where $D_h(u, u') = h(u) - h(u') - \nabla h(u')^\top(u - u')$ is the Bregman divergence between u, u' with respect to the convex function h and t is a mapping such that $t(v) = q \iff v \in v^\star(q)$. In Chapter 3 we further study this setting and provide quadratic lower bounds of the calibration function.

Non-FY polyhedral non-probabilistic estimators. The classical polyhedral loss used for classification is the hinge loss (or SVM loss) and its generalization in structured prediction called the Max loss (also known as Max-Margin, structured hinge or structural SVM (Tsochantaridis et al., 2005; Taskar et al., 2004)) defined as:

$$S_{\mathrm{M}}(v, y) = \max_{y' \in \mathcal{Y}} L(y, y') + v_{y'} - v_y. \tag{1.17}$$

Note that this loss does not have a FY structure. It is known that the Max loss is not Fisher consistent to the 0-1 loss for output spaces with more than two labels (Liu, 2007). On the other hand, it was not known of any necessary or sufficient conditions on the discrete loss L for Fisher consistency to hold. One of the contributions of this thesis (see Chapter 5) is the derivation of highly restrictive necessary conditions for consistency of S_{M}, which are essentially only satisfied by discrete losses corresponding to a shortest path distance in an acyclic graph (for which we prove Fisher consistency). In the same section, we also introduce another non-FY polyhedral loss called the *Restricted-Max loss*, which maintains the maximization structure of (1.17) but the maximization domain is restricted to the prediction sets Δ_L and has better consistency guarantees. The novel introduced loss reads,

$$S_{\mathrm{RM}}(v, y) = \max_{q \in \Delta_L(y)} L_y^\top q + v^\top q - v_y. \tag{1.18}$$

There exist other non-FY polyhedral surrogate losses which are Fisher consistent to a generic L, such as the Lin-Lin-Wahba SVM (Lee et al., 2004) and Simplex-coding SVM (Mroueh et al., 2012). Unfortunately, the structure of these losses does not allow them to be used in large structured output spaces studied in the following Section 5 .

5 Structured Prediction

The surrogate losses we have studied in the previous section had a surrogate space dimension k potentially as large as the size of the output space $\mathcal{Y}$. In structured prediction, this is generally intractable due to the exponentially large cardinality of this space. In this section, we introduce the classical methodology of structured surrogate methods to deal

[14] In particular, this is equivalent to say that the surrogate Bayes risk s is strictly concave.

with both the statistical and computational challenges associated to the size of $\mathcal{Y}$. More specifically, we impose two basic computational requirements on the surrogate method: (I) a model for which decoding is tractable, and (II) a tractable subgradient of the surrogate loss. These requirements will be introduced along with the two classical surrogate methods for structured prediction - conditional random fields and structural support vector machines. Then, we will discuss the statistical properties of these methods by introducing classical parameter tying techniques to design output size independent hypothesis spaces $\mathcal{G}$ allowing for size-varying outputs. Finally, we extend the point-wise analysis of surrogate methods from the previous section to the general structured setting along with the study of theoretical properties such as Fisher consistency and calibration.

Decoding models for structured prediction. Given $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a continuous scalar-valued function, we write a generic decoding as the following discrete maximization problem,

$$f(x) = \arg\max_{y \in \mathcal{Y}} \; g(x, y). \tag{1.19}$$

Note that without further structure on g, the computational complexity of the argmax is linear in the size of the output space $|\mathcal{Y}|$. Unfortunately, as we have seen at the beginning of the introduction, this is intractable due to the exponentially large size of the output space. The classical strategy in structured prediction is to consider predictors g for which (1.19) can be tractably computed. More concretely, we assume that the predictor can be decomposed as follows,

$$g(x, y) = g(x)^\top \varphi(y), \tag{1.20}$$

where $\varphi : \mathcal{Y} \to \mathbb{R}^k$ is an embedding of the output space to $\mathbb{R}^k$ with $k \ll |\mathcal{Y}|$. We can now formulate the first requirement (I).

I. We require that the following decoding can be computed:

$$d(v) = \arg\max_{y \in \mathcal{Y}} \; v^\top \varphi(y).$$

Let's define the *marginal polytope* (Wainwright and Jordan, 2008) $\mathcal{M} \subseteq \mathbb{R}^k$ as the convex hull generated by the embeddings $\varphi(\mathcal{Y}) = \{\varphi(y)\}_{y \in \mathcal{Y}}$,

$$\mathcal{M} = \text{hull}(\varphi(\mathcal{Y})) = \left\{ \sum_{y \in \mathcal{Y}} q_y \varphi(y) \mid q \in \Delta_{\mathcal{Y}} \right\}. \tag{1.21}$$

The decoding can be seen as a maximization of a linear function over $\mathcal{M}$ as $d(v) = \arg\max_{\mu \in \mathcal{M}} \; v^\top \mu$. We can classify the type of structures into two classes depending on whether the structure comes from a graphical model defined on the parts of the output, or whether it comes from a combinatorial problem in the output space.

1. **Factor graphs / Graphical models.** The structured output is seen as composed by a set of p parts as $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_p$. These parts are grouped into a (possibly overlapping) set of factors $\mathcal{T}_t$ for $t = 1, \ldots, T$ as $\mathcal{T}_t = \Pi_{j \in N(t)} \mathcal{Y}_j$, where $N(t) \subseteq$

$\{1, \dots, p\}$ is the subset of the parts corresponding to the factor t. The assumption is that g decomposes as

$$g(x, y) = \sum_{t=1}^{T} g_t(x, y_t), \qquad y_t \in \mathcal{T}_t.$$

The predictor g can be written in the form (1.20) by defining

$$g(x) = (g_t(x, y_t))_{\substack{t=1,\dots,T \\ y_t \in \mathcal{T}_t}} \in \mathbb{R}^k, \qquad \varphi(y) = (e_{y_t})_{\substack{t=1,\dots,T \\ y_t \in \mathcal{T}_t}} \in \mathbb{R}^k,$$

where e_{y_t} is the y_t-vector of the natural basis of $\mathbb{R}^{\mathcal{Y}_t}$. The dimension of the embedding space is $k = \sum_{t=1}^{T} |\mathcal{Y}_t|$. The computational complexity of the decoding highly depends on the factor graph. For pairwise factor models, it is known that the decoding can be efficiently computed whenever the associated graph is a tree. If the graph has cycles, then iterative methods such as loopy belief propagation (also called max-product algorithm) can be used to approximate the solution. Another set of techniques are linear programming relaxation techniques, where the goal is to find tractable relaxations of the marginal polytope (Sontag, 2010).

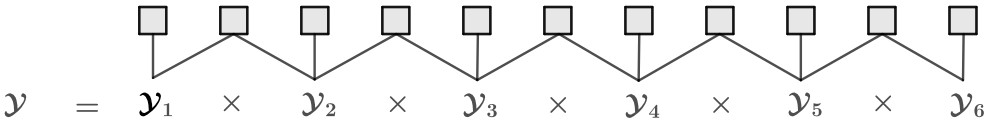


Figure 1.7: Output sequence with unary and pair-wise adjacent factors.

Example 5 .1 (Sequence prediction). In sequence prediction it is typically assumed that adjacent characters in the sequence are more correlated that characters which are at a long distance of each other. In particular, this motivates the factor graph structure depicted in Figure 1.7, where pair-wise adjacent factors are considered in order to model these dependencies. If each character belongs to a dictionary of size R, then the dimension k of the surrogate space is $NR + (N-1)R^2$ where N is the length of the sequence. Note that this is much smaller than the number of total sequences R^N. The decoding can be solved using the *Viterbi algorithm* in $\mathcal{O}(MR^2)$ operations (Viterbi, 1967), which corresponds precisely to the max-product algorithm applied to this sequential factor graph.

2. **Combinatorial problem structure.** These are structures derived from specific combinatorial problems on $\mathcal{Y}$ defined as $\arg\max_{y \in \mathcal{Y}} \; c^\top \varphi(y)$, with $c \in \mathbb{R}^k$. Many graph-related combinatorial problems over different output spaces can be formulated in this form, where the vector $c \in \mathbb{R}^k$ is a representation of the graph, such as maximum weight bipartite and perfect matching, spanning tree, graph-cut, edge-cover, and many others (Papadimitriou and Steiglitz, 1998).

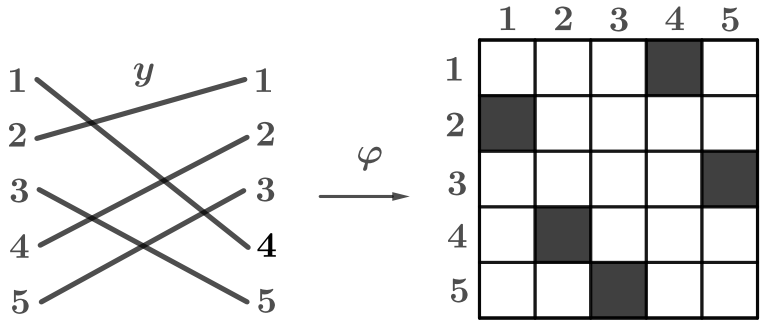


Figure 1.8: Matrix representation of the permutation.

Example 5 .2 (Permutations and matchings). When $\mathcal{Y}$ is the space of permutations σ acting on a set of N elements, one can map each permutation to the permutation matrix $P_\sigma \in \mathbb{R}^{N\times N}$, where the (i,j)-th coordinate is one if and only if $\sigma(i) = j$ and zero otherwise (see Figure 1.8). The size of the surrogate space is N^2, which is much smaller than the total number of permutations $N!$. The decoding problem corresponds to the *linear assignment problem* (Burkard et al., 2012). This matrix representation can be also used for more complex structures such as matchings in a graph (Chatalbashev et al., 2005), where further structural constraints are imposed to the binary valued matrix, such as zeros in the diagonal or graph specific constraints (Lacoste-Julien et al., 2006). For each variation, the decoding algorithm changes accordingly.

Finally, note that the distinction above between the two types of structures is made regarding the most natural representation of the problem. Indeed, many combinatorial structures such as the one in Example 5 .2 can also be written as graphical models (Sontag, 2010).

Surrogate losses for structured prediction. Now that we have defined a structure on the surrogate predictors, we move to the next step: the design of tractable surrogate losses to estimate it. In order to do that, let's work in the point-wise setting as done in Section 4 and proceed with the following identification $v = g(x) \in \mathbb{R}^k$. Given a surrogate loss $\widetilde{S} : \mathbb{R}^{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, we consider the following low-dimensional parametrization $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ associated to the embedding φ as

$$S(v, y) = \widetilde{S}((v^\top \varphi(y))_{y\in\mathcal{Y}}, y). \tag{1.22}$$

In order to be able to minimize S using gradient descent techniques, we formulate our second requirement (II).

II. We require that a sub-gradient $u \in \partial_v S(v, y)$ can be computed.

The two classical surrogates used in structured prediction are the multinomial logistic and the Max loss, also called *conditional random fields* (CRFs) (Lafferty et al., 2001) and *Structural SVMs* (SSVMs) (Tsochantaridis et al., 2005; Taskar et al., 2004). Both losses are typically used with the decoding $d(v) = \arg\max_{y \in \mathcal{Y}} \ v^\top \varphi(y)$. We briefly introduce these methods and discuss the computational complexity of computing a sub-gradient of the corresponding loss.

Conditional Random Fields. CRFs are *smooth* losses constructed by parametrizing the multinomial logistic loss as (1.22). The resulting loss reads as follows,

$$S(v, y) = \log\Big(\sum_{y' \in \mathcal{Y}} \exp v^\top \varphi(y') \Big) - v^\top \varphi(y).$$

The CRF loss is smooth and its gradient can be written as a variational problem taking the form of an entropy penalized decoding as

$$\nabla S(v, y) = \underbrace{\arg\max_{\mu \in \mathcal{M}} \ v^\top \mu + H(\mu)}_{\text{marginal inference}} - \varphi(y), \tag{1.23}$$

where $H : \mathcal{M} \to \mathbb{R}$ is the entropy on the marginal polytope $\mathcal{M}$ defined as

$$H(\mu) = \max_{q \in \Delta_{\mathcal{Y}}} \ -\sum_{y \in \mathcal{Y}} q_y \log q_y \quad \text{s.t.} \quad \mathbb{E}_{y \sim q} \, \varphi(y) = \mu. \tag{1.24}$$

The entropy penalized decoding is also known as *marginal inference* as one can easily check that it is equivalent to computing the expectation $\mathbb{E}_{y \sim q_v} \, \varphi(y)$, where $q_v(y) = \exp v^\top \varphi(y) / (\sum_{y' \in \mathcal{Y}} \ \exp v^\top \varphi(y'))^{-1}$ is an element of the exponential family over $\mathcal{Y}$ with statistic φ (Fahrmeir et al., 1994). Thus, the required oracle for computing the gradient of the CRF loss is the following:

Oracle for Conditional Random Fields: Marginal Inference

$$\arg\max_{\mu \in \mathcal{M}} \ v^\top \mu + H(\mu).$$

In the following we discuss about the tractability of marginal inference depending on the structured model we presented above. As we will see, although having a tractable decoding does not always guarantee a tractable sub-gradient, it is generally true for factor graph models.

- **Marginal inference for factor graphs.** There exist plenty of algorithms for exact inference for factor graphs. The most naive method is the elimination algorithm, whereby the marginal of a factor $\mathcal{T}_t$ is computed by picking an ordering of the other factors and eliminate them, i.e., marginalise out, according to that order. However, this procedure is very expensive as one has to repeat the elimination sequence for every factor without re-using computations. The sum-product algorithm (or belief propagation) is a dynamic algorithm leveraging intermediate terms

by 'message passing' operations between the factors, where the messages are precisely shared. However, this method only works for acyclic graphs. The junction tree algorithm (Lauritzen and Spiegelhalter, 1988) is a combination of both eliminaton and dynamic programming for graphs with small tree-width (the computational complexity is exponential with the tree-width of the graph). For large tree-width graphs such as grids, approximate inference methods are required. Variational methods aim at approximating the marginals using the variational form Eq. (1.23) and approximating the intractable $\mathcal{M}$ and entropy H. Outer approximations include loopy belief propagation (relaxation of the polytope and approximation of H with the non-convex Bethe entropy) (Yedidia et al., 2005), tree-reweighted sum-product (Wainwright et al., 2005) (concave upper-bound of entropy) and log-determinant relaxation (Wainwright and Jordan, 2008), while inner approximations include mean-field methods (Baxter, 2016). Other classical approximation methods include Monte Carlo sampling (Robert and Casella, 2013), and methods leveraging the submodularity of the energy defined by the factors (Djolonga and Krause, 2014). Finally, note that when the decoding is also intractable, both problems can have very different complexities. As an example, 2-dimensional pair-wise factor graph grids (also known as Ising models) have an NP-Hard decoding and #P-hard [15] marginal inference (Barahona, 1982).

- **Marginal inference for combinatorial structures.** Generally speaking, having a tractable decoding in this setting does not imply tractability of marginal inference. An example of this is the matrix representation of permutations in Example 5 .2. While the decoding can be solved using the polynomial-time Hungarian method (Kuhn, 1955), performing marginal inference corresponds to computing the permanent matrix which is known to be #P-complete (Valiant, 1979).

Calibration properties of CRFs are studied below when discussing about structured Fenchel-Young surrogates. We just disclose here that in contrast to the 0-1 loss case, which corresponds to the multiomial logistic loss, CRFs are generally not calibrated in the structured setting to any discrete loss L using the argmax decoding presented in the beginning of this section (also known as MAP inference in the literature). On the other hand, they can be made consistent by changing the decoding.

Structural support vector machines. The loss of structural SVMs is the *polyhedral* Max loss defined in Eq. (1.17) parametrized as (1.22),

$$S(v, y) = \max_{y' \in \mathcal{Y}} L(y, y') + v^\top \varphi(y') - v^\top \varphi(y).$$

The sub-gradient of the loss can be written as

$$\partial S(v, y) = \varphi\Big(\underbrace{\arg\max_{y' \in \mathcal{Y}} L(y, y') + v^\top \varphi(y')}_{\text{loss-augmented decoding}} \Big) - \varphi(y),$$

where the maximization problem is known as *loss-augmented decoding*, as it has the structure of the decoding with an additive term corresponding to the discrete loss to be minimized. Thus, the required oracle for computing the sub-gradient is the following.

[15] #P is the set of counting problems associated with the decision problems in NP.

Oracle for Structural SVM: Loss-augmented Decoding

$$\arg\max_{y' \in \mathcal{Y}} \; L(y, y') + v^\top \varphi(y').$$

The loss-augmented decoding is tractable whenever the loss can be decomposed using the same structure as the surrogate predictor. This is formalized by means of an *affine decomposition* of the loss L.

Definition 5 .3 (Affine decomposition of the loss)**.** *Let $\psi, \varphi : \mathcal{Y} \to \mathbb{R}^k$ be two embeddings of the discrete output space $\mathcal{Y}$ to the vector space $\mathbb{R}^k$ and $c \in \mathbb{R}$. We say that the triplet (ψ, φ, c) is an affine decomposition of the loss L if it can be decomposed as*

$$L(y, y') = \psi(y)^\top \varphi(y') + c, \qquad \forall y, y' \in \mathcal{Y}. \tag{1.25}$$

Many examples of affine decompositions of classical losses used in structured prediction can be found in Chapter 2. If the discrete loss satisfies the above decomposition, then the loss-augmented decoding can be written in terms of the decoding d as

$$\arg\max_{y' \in \mathcal{Y}} \; (\psi(y) + v)^\top \varphi(y') = d(\psi(y) + v).$$

Thus, computing a sub-gradient has exactly the same computational complexity as the decoding. This is a considerable advantage of the structural SVM over CRFs, together with the fact that the surrogate loss is dependent on the full discrete loss to minimize. Unfortunately, as we have seen disclosed at the end of Section 4 and will be studied in detail in Chapter 5, the SSVM loss is not Fisher consistent for the vast majority of losses used in practice (and cannot be made consistent by changing the decoding).

Design of the hypothesis space $\mathcal{G}$ for structured prediction. For the sake of presentation, let's assume that the model has a factor graph structure. The surrogate predictor $g : \mathcal{X} \to \mathbb{R}^k$ is made of $k = \sum_{t=1}^T |\mathcal{Y}_t|$ scalar-valued predictors

$$g_{t,y_t} : \mathcal{X} \longrightarrow \mathbb{R}. \tag{1.26}$$

Even though we have considerably reduced the number of scalar-valued predictors to learn from $|\mathcal{Y}|$ to $k \ll |\mathcal{Y}|$, this number can still be very large. Moreover, in many settings the size of the output elements are varying in size, and this cannot be taken into account if the hypothesis space $\mathcal{G}$ grows with $|\mathcal{Y}|$. The following are common settings in which additional structure can be enforced on the hypothesis space.

1. **Problems with some form of stationarity in the output space.** The factors are grouped by different types, such as unary and pairwise adjacent potentials, respectively, as in Figure 1.7. Then, within each factors t, t' belonging to the same group, the predictors are tied as $g_{t,y_t} = g_{t',y_{t'}}$. This parameter-tying technique is widely used in practice in structured prediction.

2. **The input has the same parts as the output.** In this case, it is generally assumed that the domain of g_{t,y_t} is $\mathcal{X}_t$, where $\mathcal{X}_t$ is the input part associated to the output part $\mathcal{Y}_t$.

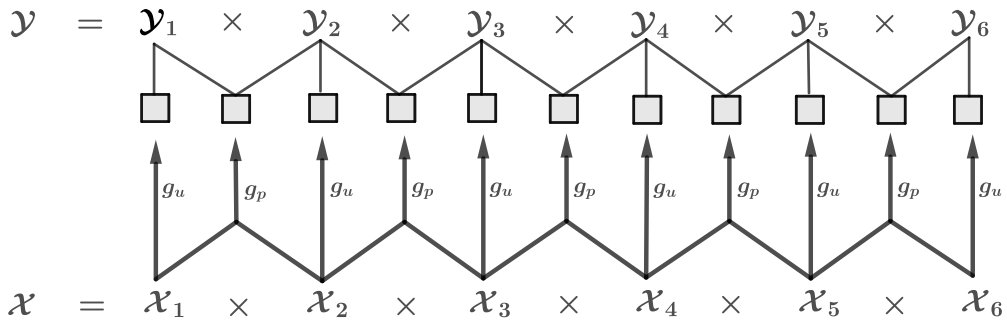


Figure 1.9: The predictors g_u, g_p corresponding to unary and pair-wise potentials are tied, respectively. Moreover, the predictors only depend on the corresponding input factors. Note that the size of the hypothesis space is *independent* of the output size. Thus, these tied predictors can be used to learn using datasets of sequences of variable size.

Both techniques above allow us to apply surrogate approaches to problems with varying size of the outputs (using 1) and varying size of both the input and the output (using also 2). The extension to this setting is straightforward and corresponds to make the output space depend on the input as $\mathcal{Y}(x)$. Furthermore, these techniques also provide the possibility of proving generalization bounds from one example as done by Ciliberto et al. (2019) for the quadratic surrogate. Indeed, with sufficient correlation decay between the output parts, tending the size of the observed outputs to infinity in a fixed-size dataset corresponds to increasing the information present in the data for learning. Finally, note that these considerations extend naturally to problems with a combinatorial structure.

Example 5 .4 (Linear hypothesis space). In practice we generally use a linear hypothesis space over shared input-output features $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as

$$g(x, y) = w^\top \Phi(x, y).$$

In the factor graph setting, the predictor is assumed to decompose as $g(x, y) = \sum_{t=1}^{T} w_t^\top \Phi_t(x, y_t)$, which corresponds precisely to our setting by letting $g(x) = (w_t^\top \Phi_t(x, y_t))_{y_t \in \mathcal{Y}_t, t=1,\dots,T} \in \mathbb{R}^k$. Parameter tying in the linear hypothesis case corresponds to set $w_t = w_{t'}$ if the factors t, t' belong to the same group.

We now extend the point-wise analysis of the previous section to discrete losses satisfying the affine decomposition and surrogate losses defined over structured surrogate spaces introduced above. More concretely, the geometry will be studied on the marginal polytope $\mathcal{M}$ instead of the simplex $\Delta_\mathcal{Y}$.

Point-wise analysis of structured surrogate losses. Assuming an affine decomposition of the loss (see Definition 5 .3), we define the Bayes risk $\ell : \mathcal{M} \to \mathbb{R}$ as

$$\ell(\mu) = \operatorname*{arg\,min}_{y' \in \mathcal{Y}} \ \psi(y')^\top \mu + c, \qquad \forall \mu \in \mathcal{M}, \tag{1.27}$$

which corresponds to a parametrization in the marginal polytope of the Bayes risk defined in the previous section. Analogously, we define the set of optimal predictors $y^\star$:

$\mathcal{M} \to 2^{\mathcal{Y}}$ and the prediction sets $\mathcal{M}_L : 2^{\mathcal{Y}} \to 2^{\mathcal{M}}$ as

$$y^\star(\mu) = \operatorname*{arg\,min}_{y' \in \mathcal{Y}} \ \psi(y')^\top \mu \subseteq \mathcal{Y}, \quad \mathcal{M}_L(Y) = \{\mu \in \mathcal{M} \mid Y \subseteq y^\star(\mu)\}. \subseteq \mathcal{M}. \tag{1.28}$$

Unfortunately, the surrogate Bayes risk $s(q) = \arg\min_{v \in \mathbb{R}^k} \ S(v)^\top q$ of a parametrized surrogate loss of the form $S(v, y) = \widetilde{S}((v^\top \varphi(y))_{y \in \mathcal{Y}}, y)$ in general cannot be written in terms of elements μ from the marginal polytope as in (1.27). This is the case of the Max loss. We now present structured Fenchel-Young losses, which have the property that the surrogate Bayes risk can be defined in the marginal polytope.

FY losses in structured prediction. Structured Fenchel-Young losses were first introduced by Blondel et al. (2020), and the first calibration analysis is given by the work presented in Chapter 3. Let $\Omega : \mathcal{C} \subseteq \mathbb{R}^k \to \mathbb{R}$ be a convex function where $\mathcal{C}$ is a convex set containing the marginal polytope $\mathcal{M}$. We define the Fenchel-Young loss $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ as

$$S(v, y) = \Omega^*(v) - v^\top \varphi(y). \tag{1.29}$$

The surrogate bayes risk s and the set of minimizers $v^\star : \mathcal{M} \to 2^{\mathbb{R}^k}$ take the following form:

$$s(\mu) = -\Omega(\mu), \quad v^\star(\mu) = \partial\Omega(\mu), \quad \forall \mu \in \mathcal{M}.$$

The sub-gradient of the FY loss reads $\partial S(v, y) = \partial\Omega^*(v) - \varphi(y)$. Thus, the computational complexity of computing the sub-gradient boils down to the complexity of computing a sub-gradient of Ω^*. The requirement for the tractability of structured FY losses can be summarized with the following oracle:

Oracle for Structured Fenchel-Young losses: Projection Oracle

$$\partial\Omega^*(v) = \operatorname*{arg\,max}_{\mu \in \mathcal{C}} \ v^\top \mu - \Omega(\mu).$$

We call it projection as it corresponds to an euclidean projection into the marginal polytope if Ω is a quadratic function. Both Fisher consistency and calibration properties between S and L introduced in the previous section can therefore be written using the marginal polytope parametrization. In particular, Fisher consistency corresponds to the following statement:

$$v \in v^\star(\mu) \implies d(v) \in y^\star(\mu), \ \forall \mu \in \mathcal{M},$$

and the calibration function can be written as

$$\zeta(\varepsilon) = \inf_{\substack{v \in \mathbb{R}^k \\ \mu \in \mathcal{M}}} \delta s(v, \mu) \quad \text{s.t.} \quad \delta\ell(d(v), \mu) \geq \varepsilon,$$

where $\delta s(v, \mu)$ and $\delta\ell(y, \mu)$ are the parametrizations of the excess risks in $\mathcal{M}$. Analogously to the presentation of FY losses for the simplex in the previous section, we study the two main types of FY surrogate losses: smooth and polyhedral.

Structured FY smooth losses are probabilistic estimators. Smooth structured Fenchel-Young losses are always Fisher consistent to any discrete loss satisfying an affine decomposition of the form $L(y, y') = \psi(y')^\top \varphi(y) + c$ for some $\psi : \mathcal{Y} \to \mathbb{R}^k$ and $c \in \mathbb{R}$. To see this, note that $\mu = \nabla\Omega^*(v) \iff v \in v^\star(\mu) = \partial\Omega(\mu)$ by construction of the Fenchel conjugate. Hence, Fisher consistency holds with the decoding defined as

$$d(v) = y^\star(\nabla\Omega^*(v)) = \arg\min_{y' \in \mathcal{Y}} \psi(y')^\top \nabla\Omega^*(v) \in y^\star(\mu), \quad \forall v \in v^\star(\mu),$$

where ties are broken arbitrarily. These losses are probabilistic estimators as they estimate the conditional risk:

$$\begin{aligned} L_y^\top q &= \mathbb{E}_{y' \sim q}\, \psi(y)^\top \varphi(y') + c \\ &= \psi(y)^\top \mathbb{E}_{y' \sim q}\, \varphi(y') + c \\ &= \psi(y)^\top \nabla\Omega^*(v) + c, \quad \forall v \in v^\star(\mathbb{E}_{y' \sim q}\, \varphi(y')). \end{aligned}$$

The following are some examples of functions Ω defining structured smooth FY losses:

- **Entropy on the marginal polytope.** The function $\Omega : \mathcal{M} \to \mathbb{R}$ is defined as $\Omega = -H$, where H is the entropy of the marginal polyope defined in Eq. (1.24). The resulting surrogate loss is the one of conditional random fields presented above.

- **$\mathcal{M}$-constrained entropy.** As we already seen above, marginal inference (oracle required for the sub-gradient of CRFs) can be intractable for combinatorial structures such as the matrix representation of permutations in Example 5 .2. One can relax the computational intractability of the previous entropy by defining an entropy directly on the surrogate space. More specifically, if $\mathcal{M} \subseteq \mathbb{R}^k_{\geq 0}$ is included in the positive orthant, we can define the negative entropy $\Omega : \mathcal{M} \to \mathbb{R}$ as

  $$\Omega(\mu) = \sum_{j=1}^{k} \mu_j \log \mu_j + i_{\mathcal{M}}(\mu), \quad \forall \mu \in \mathcal{M}. \tag{1.30}$$

  In this case, the sub-gradient of Ω^* using the structure of permutations from Example 5 .2 can be efficiently approximated to precision ε with complexity $\mathcal{O}(k^2/\varepsilon)$ using the Sinkhorn algorithm (Sinkhorn and Knopp, 1967). This loss was introduced by Blondel (2019).

- **Quadratic.** Let $\Omega : \mathbb{R}^k \to \mathbb{R}$ defined as

  $$\Omega(u) = \frac{1}{2}\|u\|_2^2, \quad \forall u \in \mathbb{R}^k.$$

  The resulting loss corresponds to the quadratic surrogate $\frac{1}{2}\|v - \varphi(y)\|_2^2$ up to additive constant terms. This loss is studied in depth in Chapter 2. It was introduced in the structured setting by Ramaswamy et al. (2013) and further generalized to continuous structured spaces by Ciliberto et al. (2020).

- **$\mathcal{M}$-constrained quadratic.** Let $\Omega : \mathcal{M} \to \mathbb{R}$ defined as

  $$\Omega(\mu) = \frac{1}{2}\|\mu\|_2^2 + i_{\mathcal{M}}(\mu), \quad \forall \mu \in \mathcal{M}.$$

  In this case, the sub-gradient corresponds to an Euclidean projection into the marginal polytope. The resulting loss is known as SparseMAP (Niculae et al., 2018).

Finally, using the β-smoothness of the function Ω^*, in Chapter 3 we also prove calibration results by deriving lower bounds on the calibration function ζ of the form

$$\zeta(\varepsilon) \geq \frac{\varepsilon^2}{8c_\psi^2\beta},$$

where c_ψ is a constant depending on the embedding ψ. An important observation is that both constants c_ψ and β are *not* exponentially large even if it is the case for the size of the output space, but rather depend polynomially with the affine dimension of the discrete loss decomposition.

Structured FY polyhedral losses are non-probabilistic estimators. The FY polyhedral Max-Min loss Eq. (1.17) presented in the previous section under the parametrization Eq. (1.22) reads as follows:

$$S(v, y) = \max_{\mu \in \mathcal{M}} \min_{y' \in \mathcal{Y}} \ \psi(y')^\top \mu + v^\top \mu - v^\top \varphi(y). \tag{1.31}$$

The entire Chapter 4 of this thesis is devoted to the statistical and computational analysis of this surrogate loss. For the sake of presentation, let's now assume that the affine decomposition takes the form $L(y, y') = \varphi(y')^\top A \varphi(y) + c$ where $A \in \mathbb{R}^{k \times k}$ is a linear operator from the surrogate space $\mathbb{R}^k$ to itself. In this case, the sub-gradient of Ω^* can be written as a bi-linear max-min problem in the marginal polytope as

$$\partial\Omega^*(v) = \arg\max_{\mu \in \mathcal{M}} \min_{\nu \in \mathcal{M}} \ \nu^\top A \mu + v^\top \mu, \quad \forall v \in \mathbb{R}^k. \tag{1.32}$$

The computation of the sub-gradient (1.32) is different than the maximization oracle from the Max loss and the projection oracle from smooth FY loss presented above. In particular, the above sub-gradient can be approximated using both maximization and projection oracles. However, as we see below, even though using maximization oracles seems more attractive at first sight (as we only rely on the decoding d), it is not practical to implement into the regularized ERM optimization procedure.

- **Using maximization oracles.** Having in mind that by assumption we have access to a linear maximization oracle over the marginal polytope, the first idea that comes into mind is to use a saddle-point version of the conditional gradient method (a.k.a. Frank-Wolfe) (Frank et al., 1956), whereby the maximization oracle is performed in an alternated fashion over the marginal polytope as in (Gidel et al., 2017). It has been proven that the method converges (Robinson, 1951) with worst-case complexity of $\mathcal{O}(t^{-1/(2|\mathcal{Y}|-2)})$ (Shapiro, 1958), while it is an open conjecture to prove it actually converges at a better rate of $\mathcal{O}(t^{-1/2})$. Despite being an attractive algorithm due to the required oracle its main drawback lies in its inability to warm-start the method with a solution close to the optimum. In particular, in Chapter 4 this will prove to be a key property of the algorithm in order to reuse past gradients during the iterative optimization procedure.
- **Using projection oracles.** Let $\widetilde{\Omega} : \mathcal{C} \supseteq \mathcal{M} \to \mathbb{R}$ be a strongly convex [16] differentiable function defined in a convex set $\mathcal{C}$ containing $\mathcal{M}$ such that $\nabla\widetilde{\Omega}(\mathcal{C}) = \mathbb{R}^k$

[16] A differentiable convex function h is said to be strongly convex with constant c if $h(u) \geq h(v) + \nabla h(v)^\top (u - v) + \frac{c}{2}\|u - v\|_2^2$ holds for all u, v in the domain.

and $\lim_{u \to \partial \mathcal{C}} \|\nabla \widetilde{\Omega}(u)\| = \infty$, where $\partial \mathcal{C}$ denotes the boundary of $\mathcal{C}$. We can show that with t projection oracles

$$\arg\min_{\mu \in \mathcal{M}} -\eta \mu^\top u + D_{\widetilde{\Omega}}(\mu, \mu'), \tag{1.33}$$

a sub-gradient (1.32) can be approximated up to precision $\varepsilon = \mathcal{O}(t^{-1})$, where $D_h(\mu, \mu') = h(\mu) - h(\mu') - \nabla h(\mu')^\top (\mu - \mu')$ is the Bregman divergence associated to the convex function h. A part from the fast convergence to a sub-gradient, this method can be easily warm-started with past approximated sub-gradients. In particular, under the assumption that the sub-gradients do not change much during the full optimization procedure, it can be shown experimentally that a constant number of projections at each iteration is enough to make the algorithm convergence.

Remark 5 .5. A sub-gradient of the Max-Min loss can thus be approximated using several gradients of smooth FY losses corresponding to $\Omega = \widetilde{\Omega}$. One might ask what are the advantages of using the Max-Min loss rather than a smooth FY loss if both are calibrated. Being a non-probabilistic estimator instead of a probabilistic estimator is not generally a valid argument (i.e., decision boundaries are estimated directly), as plug-in estimators can have super fast rates under certain low-noise assumptions (Audibert and Tsybakov, 2007). On the other hand, note that the Max-Min surrogate is dependent of the discrete loss L, whereas smooth FY surrogates generally only depend on the embedding φ and remaining information about the loss is contained in the decoding function. This means that the approximation error $\mathcal{E}(d \circ g_{\mathcal{G}}) - \mathcal{E}(f^\star)$ of the Max-Min loss is likely to be smaller than the smooth loss due to the full dependence of $g_{\mathcal{G}}$ in L.

In Chapter 4 we also derive *linear* lower bounds on the calibration function for the argmax decoding $d(v) = \arg\max_{y' \in \mathcal{Y}} v^\top \varphi(y')$ under mild conditions on the learning problem. They take the following form:

$$\zeta(\varepsilon) \geq \frac{\varepsilon}{c_{\varphi, A}},$$

where the constant $c_{\varphi,A}$ depends on the matrix A and embedding φ corresponding to the affine decomposition of the structured discrete loss. Similarly as for smooth FY loss, the constant $c_{\varphi,A}$ is shown to be polynomial with the affine dimension of the loss.

Structured Prediction Topics Not Covered

We conclude this section by pointing out two topics on structured prediction which are not covered in this thesis: the learning to search approach, and the use of deep learning techniques.

Other approaches to structured prediciton: learning to search. We now briefly discuss a completely different approach to structured output space prediction from the one presented in this manuscript. This approach is known in the literature as *learning to search* (Daumé et al., 2009; Daumé III and Marcu, 2005; Doppa et al., 2014), and it is closely related to imitation learning (Osa et al., 2018) and inverse reinforcement learning (Russell, 1998). The main idea behind is to view the structured outputs as objects whose parts are built iteratively by an agent navigating in a search space or environment. For instance, in sequence prediction, a learning agent is navigating the positions from left to write and at every step it takes an action using both input and past information to choose the following character in the sequence. The input-output examples from the training dataset are thus realizations of an expert policy that the learning agent wants to use in order to minimize the expected risk of the discrete loss L. The potential advantages of this framework over the surrogate based approach is that it can be used to minimize discrete losses without a low-dimensional structure (i.e., non-decomposable losses) and it can take into account complicated dependencies in the output space while keeping the tractability of the search procedure. On the other hand, many structured output spaces such as the space of permutations do not have a natural associated search space in which the outputs can be iteratively constructed. Indeed, this constrains the application of the learning to search approach to structured spaces where there is a natural order of the parts of the output such as sequences.

Structured prediction with deep learning. Deep learning is the branch of machine learning that uses deep artificial neural networks (also known as deep networks) to learn a representation of the data (LeCun et al., 2015). In deep learning, discrete prediction problems are tackled using precisely the same surrogate strategies described in this thesis. This means that most theory developed here is complementary to the theoretical understanding of the performance of deep networks. More specifically, finite-sample generalization bounds on the excess risk of a deep network (if obtained), can be directly plugged into the calibration functions studied in this thesis to obtain guarantees on the Bayes predictor. When the output elements have a natural order such as sequences, deep learning techniques are usually used under the learning to search framework presented above (Sutskever et al., 2014).

6 Related works and Summary of Contributions

In this final section of the introduction, we provide a brief historical review of the related works on discrete prediction learning until the moment where the research developed in this thesis began. The literature review is not fully exhaustive and it is biased by the personal views of the author of the thesis. In particular, we want to convey the need for a sound general framework for structured output prediction that goes beyond the margin-based paradigm that was present in the early phases of statistical learning theory and served as basis to design classification methods such as the structural SVM. After the related works exposition, we present the main questions that were asked at the beginning of the PhD and finally a summary of the contributions provided in this thesis.

From margin bounds to quantitative guarantees to the Bayes predictor. As it constantly happens during the scientific entreprise, paradigms in which phenomena, problems or methods are seen change (Kuhn, 1970). In particular, it is common that procedures are developed under a certain view and later are reformulated differently. This is the case for classification methods in statistical learning, which were developed using the central concept of *margin* for binary linearly separable data, which corresponds to the maximum distance between the linear classifier and the set of positive and negative labels, respectively. More specifically, binary classification and some of the first multi-class and structured prediction methods were designed and studied as margin maximization procedures. The perceptron algorithm (Rosenblatt, 1961) was proved to converge at a rate depending on the margin, and the (hard) support vector machine (SVM) (Vapnik and Lerner, 1963; Boser et al., 1992) was designed to find precisely the maximum margin hyperplane. Later, SVMs were generalized to non-separable training data (Cortes and Vapnik, 1995) by extending the notion of margin to its soft version where data points are allowed to be missclassified. Probabilistic methods for classification such as logistic regression (Friedman et al., 2000) and boosting techniques (Freund and Schapire, 1997), also known as plug-in rules, where also seen through the lens of margin maximization (Bartlett et al., 1998), and their theoretical analysis was performed under this paradigm (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002) using the so-called *margin bounds* (see also related paragraph in Section 2). Even if this vision has been partially abandoned at least for binary classification, most introductory courses in machine learning motivate those binary convex surrogate methods with the notion of margin and their name margin losses has remained. The extension of the binary margin to larger output spaces led to generalizations of the SVM such as the Weston-Watkins SVM (Weston and Watkins, 1999) and Crammer-Singer SVM (Crammer and Singer, 2001) for multi-class classification with 0-1 loss and the structural SVM (Taskar et al., 2004; Tsochantaridis et al., 2004) for the general structured prediction setting.

Soon after, some parts of the community started to realize that margin-based analysis was not able to address in a sound way the two basic sources of error of a statistical procedure introduced in Section 2 , namely, the estimation error and the approximation error. Quoting Zhang (2004b): 'The margin idea mixes the two aspects together (...) so that it is not clear which aspect is the main contribution to the success of the so-called margin maximization methods.' In other words, margin bounds were missing a quantitative control on the closeness of the surrogate predictor to the discrete Bayes predictor, which is the ultimate quantity of interest. Although there were works addressing this question for specific instances of these methods such as Devroye et al. (1996) for general-

ized linear models, Lugosi and Vayatis (2004) for boosting methods and Steinwart (2002) for SVMs, the first general framework for studying binary convex surrogates for the 0-1 loss with respect to the discrete Bayes predictor was done by Zhang (2004b); Lin (2004); Bartlett et al. (2006). In particular, these works introduced the concepts of Fisher consistency and calibration, which are central in this thesis (see Section 3). Soon after, this analysis was applied to multi-class methods with 0-1 loss, both providing conditions for calibration (Tewari and Bartlett, 2007) and a corresponding quantitative analysis (Zhang, 2004a). In particular, the former work and Liu (2007) proved that many of the multi-class extensions of the SVM derived from margin generalizations were not consistent for the 0-1 loss. Steinwart (2007) introduced the concept of calibration function in its generality and a point-wise analysis of probabilistic-based classifiers was studied in depth by Reid and Williamson (2010) in the binary case and Williamson et al. (2016) for the multi-class setting under the name of composite proper losses. Despite existing calibration analysis of surrogate methods for structured output spaces, those were specialized to a particular setting. For instance, Gao and Zhou (2011) for multi-label, Duchi et al. (2010) and Ravikumar et al. (2011) for ranking, Pires et al. (2013) for some specific surrogates, and Ramaswamy et al. (2013) for low-rank tasks with a quadratic surrogate. The first formalization of surrogate methods for general structured prediction can be traced back to Ramaswamy and Agarwal (2016), but the authors do not provide a quantitative analysis required to obtain finite-sample guarantees of the estimators, such as calibrated functions. The first quantitative study of calibration for structured prediction is by Ciliberto et al. (2016); Osokin et al. (2017), but both works are very specific to the quadratic surrogate.

The remaining existing generic quantitative analysis for structured prediction (London et al., 2016; Cortes et al., 2016) are based on margin bounds, and as already discussed above and in Section 3 , these bounds are unable to provide theoretically sound quantitative guarantees to the discrete Bayes predictor.

Open questions at the beginning of the PhD. As we have seen in the literature review above, there has been a continuous effort to provide a general framework to theoretically understand surrogate methods for structured prediction. Although there are still some analysis working under the margin-based framework, most of the research is towards providing quantitative calibration guarantees, which serve as basis to understand the capabilities of the method to approximate the Bayes predictor. In particular, the way towards this general framework comes by understanding calibration properties of existing surrogate losses designed as margin maximizers such as the structural SVM and by understanding how calibrated surrogate losses can be systematically constructed from the learning problem. More specifically, these are some of the open questions at the moment where the research investigations appearing in this thesis started:

- There was lacking a general and user-friendly strategy to explicitly compute comparison inequalities between excess risks in the general context of structured prediction. Moreover, as pointed out by Osokin et al. (2017), it is key to show that these inequalities do not hide large constants when dealing with exponentially large output spaces.
- Despite being known that the structural SVM is not consistent to the 0-1 loss in multi-class classification (Liu and Shen, 2006; Tewari and Bartlett, 2007), nothing

was known regarding its calibration for general structured output spaces, which is the learning setting in which the method is most often used in practice.

- Two multi-class extensions of the SVM were known to be calibrated (Lee et al., 2004; Mroueh et al., 2012). However, they are not useful for structured prediction with large output spaces due to its additive structure. It was not known of any calibrated non-probabilistic method applicable to structured prediction.

Summary of contributions. During this thesis, these open questions have been addressed and have been partially or completely answered. The following is a summary of the contributions of this thesis:

- In Chapter 2, we provide a statistical and computational analysis of the quadratic surrogate for structured prediction. In particular, we show that the constants appearing in the calibration function are logarithmic with respect to the size of large structured output spaces by leveraging the affine decomposition of the task loss, thus making learning possible. Moreover, a sharp analysis of constants appearing in the generalization bounds of kernel ridge regression is given, together with the computational complexity of the decoding procedure. In Chapter 3, this analysis is extended to generic smooth convex surrogates beyond the quadratic surrogate and special emphasis is made on calibration functions, which are lower bounded by quadratics using the strong convexity of the Bayes risk. Comparison inequalities between excess risks are obtained for smooth surrogates ranging from one-vs-all methods to conditional random fields.

- In Chapter 5, we provide a general calibration analysis of the Max loss (structural SVM) for general symmetric task losses. In particular, we provide strong necessary conditions for its Fisher consistency, which are not satisfied by most losses used in practice. Moreover, we show that losses defined as shortest path distances in acyclic graphs such as the absolute deviation loss used in ordinal regression are Fisher consistent, thus providing the first consistent losses for the Max loss for output spaces with more than two labels. In particular, this work shows how the maximum margin extension to multi-class problems led to the design of ill-posed surrogate methods for most discrete prediction tasks.

- In Chapter 4, we present the Max-Min loss, a polyhedral non-probabilistic surrogate loss for structured prediction that it is calibrated to any task loss and it is applicable to structured output spaces. The minimization of the surrogate requires projection oracles to the marginal polytope. We provide an efficient algorithm and corresponding finite-sample generalization bounds with respect to the Bayes predictor of the corresponding regularized ERM estimator, where the constants are shown to be logarithmic with respect to the size of large structured output spaces. This loss, however, does not correspond to the SVM in the binary classification setting. In Chapter 5, we introduce the Restricted-Max loss, which is a generalization of the binary SVM to structured prediction, it has a maximization additive structure such as the Max loss, but despite having better consistency guarantees than the latter, it is not consistent to any task loss as it is the case for the Max-Min loss.

II

Probabilistic Estimators

2 Sharp Analysis of Learning with Discrete Losses

Abstract

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 (BakIr et al., 2007; Crammer and Singer, 2001; Read et al., 2011; Pedregosa et al., 2017). 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 (Lafferty et al., 2001; Settles, 2004) 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 (Volkovs et al., 2011); (b) Structured SVM (Tsochantaridis et al., 2004; Joachims, 2006) 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 (Tewari and Bartlett, 2007); (c) more recently, least-squares (or quadratic surrogate) frameworks (Ciliberto et al., 2016; Osokin et al., 2017) 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 Ciliberto et al. (2016) 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 Sections 3 .1, 3 .2 and 4).
- We propose a margin condition for discrete losses (generalizing the Tsybakov condition for binary classification (Tsybakov, 2004)) and obtain fast learning rates for the framework of Ciliberto et al. (2016), that are adaptive to the proposed condition (see Section 3 .3).
- Our analysis encompasses many previous results on special cases and provides improved learning rates over existing generic structured prediction frameworks (see Section 6).
- We conduct a series of experiments highlighting the benefits of the considered least-squares framework on ranking and multilabel problems (see Section 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 Section 3 .2 for some examples). We denote by $\mathcal{Z}$ the *output space*, so

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

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 ρ 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,

$$f^\star = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Z}} \mathcal{E}(f), \quad \mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho} L(f(x), y),$$

given only a number of examples $(x_i, y_i)_{i=1}^n$, with $n \in \mathbb{N}$, sampled independently from ρ. 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 Section 2 is to consider that $f^\star$ is characterized as follows (Steinwart and Christmann, 2008; Ciliberto et al., 2016):

$$f^\star(x) = \arg\min_{z \in \mathcal{Z}} \ \ell(z, x),$$

where $\ell(z, x) = \mathbb{E}_{y \sim \rho_x} L(z, y)$ is the *conditional risk*, where $\rho_x(y)$ is the conditional distribution of y given $x \in \mathcal{X}$. The quadratic surrogate (QS) for structured prediction, introduced by Ciliberto et al. (2016), is a natural estimator that has the following form,

$$\widehat{f}(x) = \arg\min_{z \in \mathcal{Z}} \ \widehat{\ell}(z, x), \tag{2.1}$$

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. (2.6) and (2.7)). 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. Ciliberto et al. (2016) 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. (2.1), 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 Ciliberto et al. (2016), that will be extended in Section 3 . First, assume that the loss L is a structure encoding loss function (SELF), i.e., it can be written as,

$$L(z, y) = \langle \varphi(z), V \varphi(y) \rangle_{\mathcal{H}}, \tag{2.2}$$

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 $\varphi : \mathcal{Z} \to \mathcal{H}$, $\varphi : \mathcal{Y} \to \mathcal{H}$.

Note that by assuming $\mathcal{Z}, \mathcal{Y}$ discrete and finite, then every loss function on $\mathcal{Z}, \mathcal{Y}$ is SELF. Indeed Eq. (2.2) 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 $\varphi(z) = e_z, \varphi(y) = e_y$ 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 Ciliberto et al. (2016)). The key property of a loss being SELF is that, by linearity of the inner product,

$$\begin{aligned} \ell(z, x) &= \mathbb{E}_{y \sim \rho_x} \ L(z, y) \\ &= \mathbb{E}_{y \sim \rho_x} \ \langle \varphi(z), V \varphi(y) \rangle_{\mathcal{H}} \\ &= \langle \varphi(z), V g^\star(x) \rangle_{\mathcal{H}}, \end{aligned}$$

with $g^\star(x) = \mathbb{E}_{y \sim \rho_x} \varphi(y)$ being the conditional expectation of $\varphi(y)$, given x. This means that in order to estimate ℓ, we just need to find an estimator $\widehat{g}$ for the conditional expectation $g^\star$, and then define $\widehat{\ell}(z, x) = \langle \varphi(z), V \widehat{g}(x) \rangle_{\mathcal{H}}$. To find a suitable estimator for $g^\star$,

Multilabel and Ranking measures					
Measure	$\mathcal{Z}$	Definition	r	A	$\text{INF}_F(\lvert\mathcal{Z}\rvert)$
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 2.1: Upper bounds for A in Theorems 3 .1 and 3 .5 and Corollary 3 .6 and computational complexity of evaluating the QS estimator in Eq. (2.7), for a number of widely-used losses for multilabel/ranking problems. See Section 3 .2 for notation, Section 4 for computational considerations and Section B for the full derivation of the results.

note that $g^\star$ can be written as the minimizer of the following quadratic surrogate (QS),

$$g^\star = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathcal{H}} \mathcal{R}_\varphi(g), \tag{2.3}$$

where $\mathcal{R}_\varphi(g) := \mathbb{E}_{(x,y)\sim\rho}\,\|g(x)-\varphi(y)\|^2_{\mathcal{H}}$ is the *expected surrogate risk* of g. The quality of the surrogate estimator g is measured in terms of the *surrogate excess risk* $\mathcal{R}_\varphi(g)-\mathcal{R}_\varphi(g^\star)$. In particular, denote by $d:\mathcal{H}\to\mathcal{Z}$ the decoding function $d(u)=\arg\min_{z\in\mathcal{Z}}\langle\psi(z),Vu\rangle_{\mathcal{H}}$. Ciliberto et al. (2016) prove that by construction the QS estimator is *Fisher consistent*, i.e., $f^\star = d\circ g^\star$, with $f^\star, g^\star$ as above. Moreover, for any $g:\mathcal{X}\to\mathcal{H}$, the *comparison inequality* holds

$$\mathcal{E}(d\circ g)-\mathcal{E}(f^\star) \leq 2c_{V,\psi}\sqrt{\mathcal{R}_\varphi(g)-\mathcal{R}_\varphi(g^\star)}, \tag{2.4}$$

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

The QS estimator depends only on L. Given a finite dataset $(x_i,y_i)_{i=1}^n$, an estimator $\widehat{g}$ for $g^\star$ can be found by considering the characterization of $g^\star$ in terms of Eq. (2.3). Indeed the problem in Eq. (2.3) can be solved using kernel ridge regression (KRR) (Caponnetto and De Vito, 2007). 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

$$\widehat{g}_n \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \|g(x_i)-\varphi(y_i)\|^2_{\mathcal{H}} + \lambda\|g\|^2_{\mathcal{G}}, \tag{2.5}$$

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 $\widehat{g}_n(\cdot)=\sum_{i=1}^n\alpha_i(\cdot)\varphi(y_i)\in\mathcal{G}$ where $\alpha(x)=(\alpha_1(x),\dots,\alpha_n(x))\in\mathbb{R}^n$ is defined by

$$\alpha(x) = (K+n\lambda I)^{-1}K_x, \tag{2.6}$$

with $K_x = (k(x, x_1), \dots, k(x, x_n)) \in \mathbb{R}^n$ 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 $\varphi(y_i)$'s, then $\widehat{\ell}(z, x)$ does not explicitly depend on the surrogate space $\mathcal{H}$. Indeed, we have that $\widehat{\ell}(z, x) = \langle \psi(z), V\left(\sum_{i=1}^n \alpha_i(x)\varphi(y_i)\right)\rangle_{\mathcal{H}} = \sum_{i=1}^n \alpha_i(x) L(z, y_i)$, so the final estimator $\widehat{f}_n = d \circ \widehat{g}_n$ can be written as

$$\widehat{f}_n(x) = \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^n \alpha_i(x) L(z, y_i). \tag{2.7}$$

Finally, by combining the comparison inequality with results on the convergence of $\widehat{g}_n$ to $g^\star$ (see for example Caponnetto and De Vito (2007)), the following theorem holds.

Theorem 2 .1 (Thm. 5 of Ciliberto et al. (2016))**.** *Let $n \in \mathbb{N}$, $\lambda_n = n^{-1/2}$ and $\tau > 0$. If L is SELF and $g^\star \in \mathcal{G}$, then the following holds with probability at least $1 - 8e^{-\tau}$,*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \leq C\, c_{V,\psi} \kappa \|g^\star\|_{\mathcal{G}} \tau^2 n^{-1/4}. \tag{2.8}$$

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. (2.7) is enough to know only the loss L and a kernel k for $\mathcal{X}$ (no knowledge of $\mathcal{H}, \varphi, \psi$ 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 by Ciliberto et al. (2016), $\mathcal{H} = \mathbb{R}^{|\mathcal{Z}|}, \psi(z) = e_z, \varphi(y) = e_y, V = L$, then the constant $c_{V,\psi}$ 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 $\|g^\star\|_{\mathcal{G}}$). Then Eq. (2.8) can be totally uninformative if the constants of the rate are exponentially large (Osokin et al., 2017).

In the next section, we prove that by using a suitable SELF-decomposition it is possible to find a version of Eq. (2.8), 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. (2.8), with explicit constants depending on the specific loss chosen (Theorem 3 .1). In Theorem 3 .2 and Table 2.1 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 Theorem 3 .5 we generalize Eq. (2.8) (and so the learning rate obtained by Ciliberto et al. (2016)), 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. (2.8), we consider another SELF-decomposition of the loss, based on the following *affine decomposition* of the loss matrix:

$$L = FU^\top + c\mathbf{1}, \tag{2.9}$$

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. (2.9) 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

$$\mathcal{H} = \mathbb{R}^r,\ \psi(z) = F_z,\ \varphi(y) = U_y,\ V = I_{r\times r}, \tag{2.10}$$

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. (2.8) specialized to discrete losses. Before giving the result, note that when we use the SELF-decomposition above for a loss, the conditional expectation $g^\star$ is characterized by $g^\star : \mathcal{X} \to \mathbb{R}^r$, $g^\star(x) = (g_j^\star(x))_{j=1}^r$, for $g_j^\star : \mathcal{X} \to \mathbb{R}$ defined as $g_j^\star(x) = U^{j\top}\rho_x$, with $U^j \in \mathbb{R}^{|\mathcal{Y}|}$ the j-th column of U and $\rho_x(y)$ is the conditional probability of y given x. In particular $g_j^\star(x) \le \max_{k\in\mathcal{Y}} |U_{kj}|$, ($U_{kj}$ 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.* (2.9)*. If $g^\star \in \mathcal{G}$, we have that with probability $1 - 8e^{-\tau}$,*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \ \le\ \mathsf{A}Q\, C\kappa\tau^2\, n^{-1/4}, \tag{2.11}$$

where C, κ are as in Theorem 2 .1,

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

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

Proof. First, note that the excess risk $\mathcal{E}(f) - \mathcal{E}(f^\star)$ 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 $\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star)$ with Eq. (2.8) applied to $L - c$, with c as in Eq. (2.9).

Applying Theorem 2 .1 with the affine decomposition in Eq. (2.10) for $L-c$ and the definition of $c_{V,\psi}$ by Ciliberto et al. (2016), we have that $c_{V,\psi} := \sup_{z\in\mathcal{Z}} \|F_z\|_2 = \|F\|_\infty$ and $\|g^\star\|_\mathcal{G}^2 = \sum_{j=1}^r \|g_j^\star\|_{\mathcal{H}_\mathcal{X}}^2 \le r \max_{1\le j\le r} \|g_j^\star\|_{\mathcal{H}_\mathcal{X}}^2$. The final result is obtained by multiplying and dividing by $U_{\max}$. □

The theorem above is essentially a version of Theorem 2 .1 where we use the affine decomposition in Eq. (2.9) 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 $g_j^\star$ 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 φ.

Finally A depends only on the chosen loss and measures the cost of using the QS method as surrogate approach. In the next section 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 Osokin et al. (2017), 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 (Theorem 3 .2, Table 2.1) we characterize explicitly the constants introduced in Theorem 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 (Ramaswamy and Agarwal, 2016). Finally we characterize the computational complexity of evaluating the QS estimator in Eq. (2.7) 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 $[m] = \{1, \dots, m\}$ 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 $\mathcal{Y} = \{1, \dots, R\}^m = [R]^m$, the set of observed relevance scores for the m documents where R is the highest relevance (Ravikumar et al., 2011). 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 2.1 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 Section 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

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

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

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

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 $\widehat{f}_j(x) = (\text{sign}\,(\widehat{g}_j(x))+1)/2$ 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. (Valizadegan et al., 2009; Ravikumar et al., 2011; Wang et al., 2013) 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:

$$L(\sigma, r) = 1 - \frac{1}{N(r)} \sum_{j=1}^{m} G([r]_j) D_{\sigma(j)}, \tag{2.12}$$

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. (2.12) 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 $\|F\|_\infty = \|D\|_2$, and $U_{\max} = D_{\max} G_{\max}$. Hence, we have that $A = \sqrt{m} G_{\max} D_{\max} (\sum_{j=1}^m D_j^2)^{1/2}$. For Table 2.1, assume $D_1 = 1$. If we define the vector $u \in \mathbb{R}^m$ as

$$u_j = \sum_{i=1}^{n} \frac{G([r_i]_j) \alpha_i(x)}{N(r_i)}, \quad 1 \leq j \leq m, \tag{2.13}$$

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

The key result in Table 2.1 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 was conjectured by Osokin et al. (2017)).

Remark 3 .3 (On the sharpness of the QS estimator). It is natural to ask to what extent the statistical rates provided by Theorem 3 .1 can be considered representative of the statistical difficulty of solving the problem in Section 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 of Ramaswamy and Agarwal (2016), 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 Section B).

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 Osokin et al. (2017), where they obtain a constant in the order of $\mathcal{O}(m^2)$. For Prec@k, we obtain 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 provided by Waegeman et al. (2014). 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 by Ramaswamy et al. (2013).

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

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

The margin function γ measures the minimum suboptimality gap in terms of the conditional 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

$$\rho_{\mathcal{X}}(\{x \in \mathcal{X} \mid \gamma(x) \leq \varepsilon\}) = o(\varepsilon^p), \tag{2.14}$$

where $\rho_{\mathcal{X}}$ 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. (2.14) is a generalization of the Tsybakov condition for binary classification (Tsybakov, 2004) and of the condition provided by Mroueh et al. (2012) 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 ρ_x in the simplex, one associates the corresponding optimal label as $z^\star(x) = \arg\min_z \ell(z, x)$. Figure 2.1 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 Figure 2.1 and so $\{\rho_x \mid \gamma(x) < \varepsilon\}$ corresponds to the yellow area. Eq. (2.14) says that the probability of falling in that region vanishes as $o(\varepsilon^p)$.

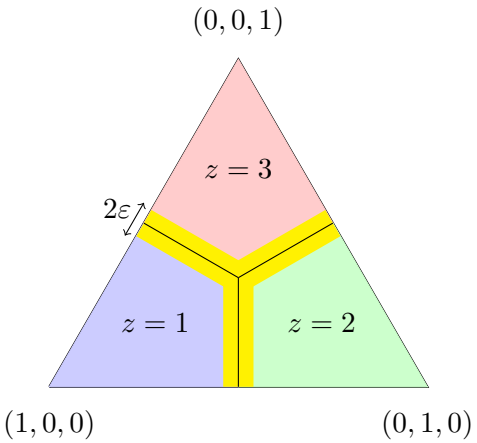


Figure 2.1: Generalized Tsybakov condition for discrete losses, Eq. (2.14), in the case of multi-class. See Example 3 .4 for more details.

In the next theorem we improve the comparison inequality of Eq. (2.4) to take into account the generalized Tsybakov condition for discrete losses of Eq. (2.14).

Theorem 3 .5 (Improved comparison inequality)**.** *Assume $\mathcal{Y}, \mathcal{Z}$ to be finite and γ to satisfy Eq.* (2.14) *for $p > 0$. Then the following holds*

1. $1/\gamma \in L_p(\rho_{\mathcal{X}})$.
2. *Assume a decomposition as in Eq.* (2.9) *for the loss L. Then, for any bounded measurable $g : \mathcal{X} \to \mathcal{H}$,*
$$\mathcal{E}(f) - \mathcal{E}(f^\star) \leq q\gamma_p^{\frac{1}{p+2}} \left(\mathcal{R}_\varphi(g) - \mathcal{R}_\varphi(g^\star)\right)^{\frac{p+1}{p+2}},$$
where $\gamma_p = \|1/\gamma\|_{L_p(\rho_{\mathcal{X}})}$, $q = (16\|F\|_\infty^2)^{\frac{p+1}{p+2}}$.

The proof of the first part can be found in Lemma A .7, while the second part is Corollary A .11, both in the Section A . As you can note, the comparison inequality of Eq. (2.4) 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 Theorem 3 .1.

Corollary 3 .6 (Improved rates)**.** *Under the p-noise condition, we have the following improvement on the generalization bound in Eq.* (2.11)*,*

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^\star) \leq C\gamma_p^{\frac{1}{p+2}} \left(\mathsf{A}^2 Q^2 \kappa^2 \tau^4 n^{-\frac{1}{2}}\right)^{\frac{p+1}{p+2}}, \tag{2.15}$$

with C universal constant and A, Q, κ as in Theorem 3 .1.

Note that the result of Theorem 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. $p \gg 0$).

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 2.2: Numerical results on real-world multilabeling and ranking datasets comparing our QS estimator, THBM (Zhang and Zhou, 2014) and SSVM (Joachims, 2006). 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 Section 5 for more details.

4 Computational Considerations

As already observed by Ciliberto et al. (2016): (1) the computation of the QS estimator (Eq. (2.7)) 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 Section 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. (2.6).

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. (2.6), i.e. $\alpha(x) = WK_x$, with the notation in Eq. (2.6); second predicting the output $z \in \mathcal{Z}$ associated to the test input x, by solving Eq. (2.7).

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 Ciliberto et al. (2016) in terms of consistency and learning rates. Indeed note that the SELF-decomposition in terms of $\mathcal{H}, \psi, \varphi$ for the loss and in particular the affine decomposition of Eq. (2.9) is needed only to prove the sharper guarantees in Theorems 3 .1, 3 .2 and 3 .5 and Corollary 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 $\varphi(y) = U_y$, so we can compute explicitly the solution to Eq. (2.5) (Caponnetto and De Vito, 2007), $\widehat{g}_n : \mathcal{X} \to \mathbb{R}^r$ 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

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

with $\varphi = (\varphi(y_1), \ldots, \varphi(y_n)) \in \mathbb{R}^{r\times n}$. 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 (Smola and Schökopf, 2000; Rahimi and Recht, 2008; Rudi et al., 2017).

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 (Smola and Schökopf, 2000; Rahimi and Recht, 2008; Rudi et al., 2017)). Then we use the characterization of $\widehat{f}_n(x) = (d \circ \widehat{g}_n)(x) = d \circ \theta$, to obtain the equivalent problem

$$\min_{z\in\mathcal{Z}} F_z \cdot \theta, \tag{2.16}$$

where $F_z \in \mathbb{R}^r$ is the z-th row of F (see Eqs. (2.9) and (2.10)) and $(\cdot)$ the dot-product. The computational complexity of Eq. (2.16) (we denote it by $\mathrm{INF}_F(|\mathcal{Z}|)$) has been devised for a number of widely used losses in Theorem 3 .2, Table 2.1 (see Section B for the proofs).

5 Numerical Experiments

We perform numerical experiments for the QS-estimator on multilabeling (9 datasets provided by Tsoumakas et al. (2011)) and ranking problems (1 dataset provided by Hersh et al. (1994)), see Table 2.2. We use three evaluation measures for multilabel, namely, 0-1, Hamming and F-score, and NDCG@k for ranking (in the NDCG-type family provided by Valizadegan et al. (2009)), which have been theoretically analysed in Section 3 and Table 2.1. 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 (Zhang and Zhou, 2014), and the Structural SVM (Joachims, 2006) (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 of Zhang and Zhou (2014)). The SSVM corresponds to the multilabel-SVM (Finley and Joachims, 2008), 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 2.2 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 (Gao and Zhou, 2011), as the experimental result empirically shows. For the ranking experiment, we have used the SSVM from Joachims (2006) called RankSVM as baseline to compare with the QS-estimator. The algorithm corresponding to the QS-estimator for NDCG, which corresponds to the one by Ravikumar et al. (2011) 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) (Tsochantaridis et al., 2004; Crammer and Singer, 2001) and Conditional Random Fields (CRFs) (Lafferty et al., 2001; Settles, 2004; Sutton and McCallum, 2012) 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 Section 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 (Sutton and McCallum, 2012). This lack of calibration to a given loss is an important drawback of this method (Volkovs et al., 2011). 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 $1/2$ (Zhang, 2004a).

QS for structured prediction. Ramaswamy et al. (2013) 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 Section B), PD and MAP. As Fisher consistency is a property only at the optimum, their analysis is not able to provide any statistical guarantees. Ravikumar et al. (2011) 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, Osokin et al. (2017) 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) $\mathbb{E}_{(x,y)\sim\rho} \|Fg(x) - L(\cdot, y)\|^2_{\mathbb{R}^{|\mathcal{Z}|}}$, and not in the low dimensional space of the decomposition $\mathbb{E}_{(x,y)\sim\rho} \|g(x) - U_y\|^2_{\mathbb{R}^r}$. 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 Corollary 3 .6.

To conclude, Ramaswamy and Agarwal (2016) 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.

Appendices

A Calibration and fast rates for surrogate methods

The goal of Section 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 provided by Bartlett et al. (2006), which was done for binary classification under 0-1 loss, to the case of general discrete losses.

In Section A .1, we introduce the basic quantities used for the analysis of surrogate methods. Then, in Section 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 by Ciliberto et al. (2016). Finally, in Section A .3, we derive our main result, which generalizes the Tsybakov condition, existing for multiclass and binary (Mroueh et al., 2012; Tsybakov, 2004) classification.

A .1 Prerequisites on surrogate methods

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

$$f^\star(x) = \operatorname*{arg\,min}_{z \in \mathcal{Z}} \ell(z, x), \quad \mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho}\, \ell(f(x), x),$$

where $\ell(z, x) = \mathbb{E}_{y\sim\rho_x}\, L(z, y)$ is the *conditional 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 $f^\star$. 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,

$$g^\star(x) = \operatorname*{arg\,min}_{v \in \mathcal{C}} s(v, x), \quad s(v, x) = \mathbb{E}_{y\sim\rho_x}\, S(v, y), \quad \mathcal{R}(g) = \mathbb{E}_{(x,y)\sim\rho}\, s(g(x), x),$$

were $g^\star$ is the *optimal surrogate function*, $s(v, x)$ is the *conditional 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 $g^\star$ of the surrogate S gives the optimum $f^\star$ of the loss L. It can be written as $f^\star = d \circ g^\star$.

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

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

and its conditional excess risk $s(\widehat{g}(x), x) - s(g^\star(x), x)$ has the following form,

$$s(\widehat{g}(x), x) - s(g^\star(x), x) = \|\widehat{g}(x) - g^\star(x)\|_2^2. \tag{2.17}$$

Moreover, it is Fisher consistent by construction (Ciliberto et al., 2016).

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

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

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),

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

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

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

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 $\zeta : \mathbb{R} \to \mathbb{R}$ with $\zeta(0) = 0$ and positive in $(0, \infty)$, such that,*

$$\zeta(\ell(d \circ g(x), x) - \ell(f^\star(x), x)) \leq s(g(x), x) - s(g^\star(x), x), \tag{2.18}$$

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

Calibration means that for every x, one can control the excess of the conditional risk by the excess conditional risk of the surrogate. Let's re-derive the form of the calibration function for the QS.

Lemma A .4 (Calibration function for QS (Ciliberto et al., 2016)). *Assumption 1 holds for the QS with*

$$\zeta(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2} \tag{2.19}$$

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

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

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 $\eta, \eta' : \mathcal{Z} \to \mathbb{R}$, it holds that $|\min_z \eta(z) - \min_z \eta'(z)| \leq \sup_z |\eta(z) - \eta'(z)|$. As $\widehat{f}(x)$ minimizes $\widehat{\ell}(\cdot, x)$ and $f^\star(x)$ 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 $\ell(z, x) = F_z g^\star(x)$, we can conclude that,

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

Re-arranging and using Eq. (2.17) gives the final result. □

The following important Theorem shows how Eq. (2.19) translates into a relation between excess risks, which are the quantities that we are ultimately interested at.

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

$$\zeta(\mathcal{E}(f) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star) \tag{2.20}$$

Proof. This is a simple application of Jensen inequality.

$$\begin{aligned}\zeta(\mathcal{E}(f) - \mathcal{E}(f^\star)) &= \zeta(\mathbb{E}_{x\sim\rho_\mathcal{X}}(\ell(d \circ g(x), x) - \ell(f^\star(x), x))) \\ &\leq \mathbb{E}_{x\sim\rho_\mathcal{X}} \zeta(\ell(d \circ g(x), x) - \ell(f^\star(x), x)) \\ &= \mathbb{E}_{x\sim\rho_\mathcal{X}} s(g(x), x) - s(g^\star(x), x) \\ &= \mathcal{R}(g) - \mathcal{R}(g^\star)\end{aligned}$$

□

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

Corollary A .6 (Comparison inequality for QS (Ciliberto et al., 2016)). *For the QS, we have that*

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

A .3 Improved calibration under low noise

Theorem 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. (2.20) 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 Theorem A .5 under the p-noise assumption. We recall that the p-noise condition states that

$$\rho_{\mathcal{X}}(\{x \in \mathcal{X} \mid \gamma(x) < \varepsilon\}) = o(\varepsilon^p),$$

where $\gamma(x) = \min_{z' \neq f^\star(x)} \ell(z', x) - \ell(f^\star(x), x)$, 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* $1/\gamma \in L_p(\rho_{\mathcal{X}})$.

Proof.

$$\begin{aligned}
\|1/\gamma\|^p_{L_p(\rho_{\mathcal{X}})} &= \mathbb{E}_{x\sim\rho_{\mathcal{X}}}\, 1/\gamma(x)^p \\
&= \int_0^\infty p t^{p-1} P_{\mathcal{X}}(1/\gamma(x) > t) dt \\
&= \int_0^\infty p t^{p-1} P_{\mathcal{X}}(\gamma(x) < t^{-1}) dt.
\end{aligned}$$

The integral converges if $\rho_{\mathcal{X}}(\{x \in \mathcal{X} \mid \gamma(x) < t^{-1}\})$ decreases faster than t^{-p}. □

Let's now define the error set as $X_f = \{x \in \mathcal{X} \mid f(x) \neq f^\star(x)\}$. 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 (Tsybakov, 2004, Prop.1) for general discrete losses.

Lemma A .8 (Bounding the size of the error set)**.** *If* $1/\gamma \in L_p(\rho_{\mathcal{X}})$*, then*

$$\rho_{\mathcal{X}}(X_f) \leq \gamma_p^{\frac{1}{p+1}} (\mathcal{E}(f) - \mathcal{E}(f^\star))^{\frac{p}{p+1}}$$

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

$$1(f(x) \neq f^\star(x)) \leq 1/\gamma(x) \Delta\ell(f(x), x)$$

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

Before proving Theorem 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* $x > 0$, $0 \leq y \leq x$,

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

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 Theorem A .10, is an adaptation of Thm. 10 of Bartlett et al. (2006), 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 ζ (see Eq.* (2.18)*) and the p-noise condition holds. Then, we have that*

$$\zeta_p(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star),$$

where

$$\zeta_p(\varepsilon) = (\gamma_p \varepsilon^p)^{\frac{1}{p+1}} \zeta\left(\frac{1}{2}(\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}\right). \tag{2.21}$$

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

Proof. (Of Theorem A .10). Write the excess conditional risk as $\delta\ell(z', x) = \ell(z', x) - \ell(f^\star(x), x)$. We split the excess conditional 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. (2.18).

$$\begin{aligned}
\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) &= \mathbb{E}_{x\sim\rho_\mathcal{X}} \ \delta\ell(f(x), x) \\
&= \mathbb{E}_{x\sim\rho_\mathcal{X}} \ \{1(X_f)\delta\ell(f(x), x)\} \\
&= \mathbb{E}_{x\sim\rho_\mathcal{X}} \ \{\delta\ell(f(x), x)1(X_f \cap \{\delta\ell(f(x), x) \leq t\}\} \\
&+ \mathbb{E}_{x\sim\rho_\mathcal{X}} \ \{\delta\ell(f(x), x)1(X_f \cap \{\delta\ell(f(x), x) \geq t\}\} \\
&= A + B.
\end{aligned}$$

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

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

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

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

  $$\delta\ell(f(x), x)1(\delta\ell(f(x), x) \geq t) \leq \frac{t}{\zeta(t)}\zeta(\delta\ell(f(x), x)) \tag{2.22}$$

  In the case $\delta\ell(f(x), x) < t$, inequality in Eq. (2.22) follows from the fact that ζ is nonnegative. For the case $\delta\ell(f(x), x) > t$, apply Lemma A .9 with $h = \zeta, x = \delta\ell(f(x), x)$ and $y = t$.
  From Eq. (2.19), we have that $\mathbb{E}_{x\sim\rho_\mathcal{X}}\{1(X_f)\zeta(\delta\ell(f(x), x))\} \leq \mathcal{R}(g) - \mathcal{R}(g^\star)$. Hence,

  $$B \leq \frac{t}{\zeta(t)}(\mathcal{R}(g) - \mathcal{R}(g^\star))$$

  Putting everything together,

  $$\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)\right)^{\frac{p}{p+1}} + \frac{t}{\zeta(t)}(\mathcal{R}(g) - \mathcal{R}(g^\star)),$$

  and hence,

  $$\left(\frac{\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)}{t} - \gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)\right)^{\frac{p}{p+1}}\right)\zeta(t) \leq \mathcal{R}(g) - \mathcal{R}(g^\star).$$

Choosing $t = \frac{1}{2}\gamma_p^{\frac{-1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)\right)^{\frac{1}{p+1}}$ and substituting finally gives Eq. (2.21). The second part of the Theorem follows because $\frac{\zeta(t)}{t}$ is non-decreasing by Lemma A .9. □

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

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

$$\mathcal{E}(f) - \mathcal{E}(f^\star) \leq \gamma_p^{\frac{1}{p+2}} \left(16\|F\|_\infty^2 (\mathcal{R}(g) - \mathcal{R}^\star)\right)^{\frac{p+1}{p+2}}. \tag{2.23}$$

Proof. Substituting $\zeta(\varepsilon) = \frac{\varepsilon^2}{4\|F\|_\infty^2}$ in Eq. (2.21), gives that,

$$\zeta_p = \frac{\varepsilon^{\frac{p+2}{p+1}}}{\gamma_p^{\frac{1}{p+1}} 16\|F\|_\infty^2}.$$

Reversing the relation gives the comparison inequality in Eq. (2.23). □

B Multilabel and ranking losses

The goal of this section is to derive all of the constants from Table 2.1. In Section B .1, we recall the elements that we need in order to derive the constants. In Section B .2, we introduce the main tool from Ramaswamy and Agarwal (2016) that we use in order to study the optimality of the QS. Finally, the main bulk is in Section 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

$$\widehat{f}_n(x) = \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^n \alpha_i(x) L(z, y_i). \tag{2.24}$$

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

$$L = FU^\top + c\mathbf{1}. \tag{2.25}$$

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. (2.25) holds. Recall that the quantity of interest for the statistical complexity is

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

The inference complexity corresponds to the computational complexity of solving Eq. (2.24).

B .2 On the optimality of the QS.

We use results from Ramaswamy and Agarwal (2016) 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 of Ramaswamy and Agarwal (2016)), which is defined as the minimum dimension over all consistent convex surrogates w.r.t L. In the following Theorem B .1 (their Thm. 18), they provide a sufficient condition to lower bound this dimension.

Theorem B .1 ((Ramaswamy and Agarwal, 2016)). *Let* $L \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Y}|}$ *the loss matrix. If* $\exists q \in \operatorname{relint}(\Delta_{|\mathcal{Y}|})$, $c \in \mathbb{R}$, *such that* $Lq = c1$, *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, Theorem B .1 says that if there exists at least one distribution q at the interior of the simplex for which the conditional 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 Theorem 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 (Pedregosa et al., 2017) 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 (Ravikumar et al., 2011). 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

$$L(z, y) = 1(z \neq y).$$

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

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

We have that

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

Hence,

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

- **Inference.** Inference corresponds to

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

which can be done in

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

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

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

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

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

We have that

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

Hence,

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

- **Inference.** Inference corresponds to

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

which can be done in

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

- **Optimality of r.** Taking $q_y = \frac{1}{b|B(y)|}$, where $B(y)$ is the partition where $y \in \mathcal{Y}$ belongs to and applying Theorem 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

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

- **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

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

  We have that

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

  Hence,

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

- **Inference.** Inference corresponds to

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

  which can be done in

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

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

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

- **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,

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

- **Inference.** Inference corresponds to

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

  which can be done in

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

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

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

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

$$2\frac{|z \cap y|}{|z| + |y|} = \begin{cases} 2\sum_{j=1}^{m}\sum_{\ell=0}^{m} \frac{[z]_j}{\ell+|z|} 1([y]_j = 1, |y| = \ell) & y \neq 0 \\ 1(z = 0) & y = 0 \end{cases}. \tag{2.26}$$

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

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

Then, the conditional risk reads

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

Hence, for every x, one needs no more than $r = m^2 + 1$ parameters to compute the F-score conditional risk. We have the following Lemma G .2.

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

$$A_{jk}(x) = \sum_{\ell=0}^{m} \frac{P_{j\ell}(x)}{\ell + k} \tag{2.27}$$

   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:

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

Then, $f^\star(x)$ 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. (2.27). As the estimator is the same, the statistical complexity is the minimum of both.

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

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

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

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

Hence,

$$\mathsf{A}_1 \leq \sqrt{m^2 + 1} \leq \sqrt{2}m$$

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

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

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

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

Hence,

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

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

$$\mathsf{A} \leq \sqrt{2}m.$$

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

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

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

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

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

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

- **Optimality of r.** We can't say anything about the potential existence of a convex calibrated surrogate with smaller dimension than affdim$(L) - 1$. This is because the sufficient condition from Theorem 18 of Ramaswamy and Agarwal (2016) does not hold for any q 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:

$$L(\sigma, r) = 1 - \frac{1}{N(r)} \sum_{j=1}^{m} G([r]_j) D_{\sigma(j)} \tag{2.28}$$

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

  It follows that,

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

  hence,

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

- **Inference.**

  The inference corresponds to,

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

  which can be done in

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

  operations.

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

$$\|D\|_2 \sim \left(\int_2^m \frac{1}{\log^2(t)} dt \right)^{1/2} \sim \sqrt{\frac{m}{\log m}}.$$

Thus,

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

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 $\|D\|_2 \leq \frac{2}{\sqrt{3}}$,

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

The QS-estimator estimates the marginals of the normalized relevance scores and sorts the estimates at inference. Ravikumar et al. (2011) showed that 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,

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

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

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

  Hence,

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

  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,

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

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

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

with

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

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 Ramaswamy and Agarwal (2016).

As shown by Calauzenes et al. (2012), 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 conditional 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:

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

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,

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

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

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

  Denote by

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

  We have that,

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

  where q_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 Ramaswamy and Agarwal (2016).

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.

3 A General Theory for Structured Prediction with Smooth Convex Surrogates

Abstract

In this work we provide a theoretical framework for structured prediction that generalizes the existing theory of surrogate methods for binary and multiclass classification based on estimating conditional probabilities with smooth convex surrogates (e.g. logistic regression). The theory relies on a natural characterization of structural properties of the task loss and allows to derive statistical guarantees for many widely used methods in the context of multilabeling, ranking, ordinal regression and graph matching. In particular, we characterize the smooth convex surrogates compatible with a given task loss in terms of a suitable Bregman divergence composed with a link function. This allows to derive tight bounds for the calibration function and to obtain novel results on existing surrogate frameworks for structured prediction such as conditional random fields and quadratic surrogates.

1 Introduction

In statistical machine learning, we are usually interested in predicting an unobserved output element y from a discrete output space $\mathcal{Y}$ given an observed value x from an input space $\mathcal{X}$. This is done by estimating a function f such that $f(x) \approx y$ from a finite set of example pairs (x, y).

In many practical domains such as natural language processing (Smith, 2011), computer vision (Nowozin and Lampert, 2011) and computational biology (Durbin et al., 1998), the outputs are structured objects, such as sequences, images, graphs, etc. This structure is implicitly characterized by the loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ used to measure the error between the prediction and the observed output as $L(f(x), y)$. Unfortunately, as the outputs are discrete, the direct minimization of the loss function is known to be intractable even for the simplest losses such as the binary 0-1 loss (Arora et al., 1997). A common approach to the problem is to design a surrogate loss $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ defined in a continuous surrogate space $\mathcal{V}$ that can be minimized in practice and construct the functional f by "decoding" the values from the continuous space to the discrete space of outputs.

In this paper, we construct a general theory for structured output prediction using smooth convex surrogates based on estimating the conditional risk of the task loss. The methods we consider can be seen as a generalization of binary and multiclass methods based on estimating the conditional probabilities (Bartlett et al., 2006; Zhang, 2004a,b) to general discrete losses, and correspond to proper composite losses (Reid and Williamson,

2010; Vernet et al., 2011) for multiclass classification. Our construction is based on two main ingredients; first, the characterization of the structural properties of a loss function L by means of an affine decomposition of the loss (Ramaswamy et al., 2013; Nowak et al., 2019), which we present in Section 2 , and second, the Bregman divergence characterization of proper scoring rules for eliciting linear properties of a distribution (Abernethy and Frongillo, 2012; Frongillo and Kash, 2015a), which has already been noted to have strong links with the design of consistent surrogate losses (Agarwal and Agarwal, 2015).

We put these two ideas together in Section 3 to construct calibrated surrogates, which are *consistent* smooth convex surrogates with two basic elements, namely, a differentiable and strictly convex *potential* h and a continuous invertible *link function* t, which can be easily obtained from the surrogate loss. We showcase the generality of our construction by showing how general methods for structured prediction such as the quadratic surrogate (Ciliberto et al., 2016; Osokin et al., 2017; Ciliberto et al., 2019) and conditional random fields (CRFs) (Lafferty et al., 2001; Settles, 2004), and widely used methods in multiclass classification (Zhang, 2004a), multilabel classification (Read et al., 2011), ordinal regression (Pedregosa et al., 2017), amongst others, fall into our framework. Hinge-type surrogates such as the structured SVM (Crammer and Singer, 2001), which is known to be inconsistent (Tewari and Bartlett, 2007), are not included.

This theoretical framework allows to derive guarantees by relating the surrogate risk associated to S (object that we can minimize) to the actual risk associated to L (object that we want to minimize) by means of convex lower bounds on the *calibration function* ζ_h (Bartlett et al., 2006; Steinwart, 2007; Osokin et al., 2017), which is a mathematical object that only depends on the surrogate loss through the potential h. In Section 4 , we provide an exact formula for the calibration function (Theorem 4 .3) and a user-friendly quadratic lower bound for strongly convex potentials (Theorem 4 .4).

There, we also analyze the role of the link function on the complexity of the surrogate method by studying the learning guarantees when the convex surrogate is minimized with a stochastic learning algorithm (Theorem 4 .7). In particular, we show that, while the relation between excess risks is related to the potential h, the approximation error is crucially related to the link function. More specifically, we discuss the benefits of logistic-type surrogates with respect to the quadratic-type ones.

Finally, those results are then used in Section 5 to derive learning guarantees for specific methods on multiple tasks for the first time, while also recovering existing results. The most significant novel results on this direction being an exact expression for the calibration function for the quadratic surrogate (Theorem 5 .1) and a quadratic lower bound for CRFs (Proposition 5 .2).

2 Setting

2 .1 Supervised Learning

The problem of *supervised learning* consists in learning from examples a 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 lies in a different space than the observations, for instance in subset ranking losses (Chen et al., 2009) or losses with an abstain option (Ramaswamy et al., 2015). We

denote then by $\mathcal{Z}$ the *output space*, so

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

where $L(z, y)$ measures the cost of predicting z when the observed value is y. We assume that $\mathcal{Y}$ and $\mathcal{Z}$ are discrete. Finally the data are assumed to be distributed according to a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The goal of supervised learning is then to recover the function $f^\star$ [1] minimizing the *expected risk* $\mathcal{E}(f)$ of the loss,

$$f^\star = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Z}} \mathcal{E}(f), \quad \mathcal{E}(f) = \mathbb{E}_{(X,Y)\sim\rho} L(f(X), Y), \tag{3.1}$$

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

It is known that $f^\star$ is characterized as (Steinwart and Christmann, 2008; Ciliberto et al., 2016),

$$f^\star(x) = \arg\min_{z\in\mathcal{Z}} \ \ell(z, \rho(\cdot|x)), \tag{3.2}$$

where for any $q \in \mathrm{Prob}(\mathcal{Y})$, the quantity $\ell(z, q) = \mathbb{E}_{Y\sim q} L(z, Y)$ is the *conditional risk*, defined as the expectation of the loss with respect to the distribution q on the labels. We also define the *excess conditional risk* as $\delta\ell(z, q) = \ell(z, q) - \min_{z'\in\mathcal{Z}} \ell(z', q) \geq 0$.

2 .2 Affine Decomposition of Discrete Losses and Marginal Polytope

Consider the following *affine decomposition* of a loss L (Ramaswamy et al., 2013; Nowak et al., 2019),

$$L(z, y) = \langle \psi(z), \varphi(y) \rangle + c,$$

where $\psi : \mathcal{Z} \to \mathcal{H}$ and $\varphi : \mathcal{Y} \to \mathcal{H}$ are embeddings to a vector space $\mathcal{H}$ with Euclidean scalar product $\langle \cdot, \cdot \rangle$ and $c \in \mathbb{R}$ is a scalar constant. Note that by linearity of the inner product,

$$\ell(z, q) = \mathbb{E}_{Y\sim q} \langle \psi(z), \varphi(Y) \rangle + c = \langle \psi(z), \mu(q) \rangle + c,$$

with $\mu(q) = \mathbb{E}_{Y\sim q} \varphi(Y)$ the vector of moments of the statistic φ. If we denote $\mu^\star(x) = \mu(\rho(\cdot|x))$, then the excess conditional risk takes the following form $\delta\ell(z, \rho(\cdot|x)) = \langle \psi(z) - \psi(f^\star(x)), \mu^\star(x) \rangle \geq 0$. Note that the affine decomposition always exists, is not unique and it corresponds to a low-rank decomposition of the "centered" loss matrix $L - c \in \mathbb{R}^{\mathcal{Z}\times\mathcal{Y}}$. The image of $\mu^\star$ lies inside the convex hull of the $\varphi(y)'s$, that is,

$$\mathrm{Im}(\mu^\star) \subseteq \mathcal{M} \coloneqq \mathrm{hull}(\varphi(\mathcal{Y})) \subset \mathcal{H}.$$

The set $\mathcal{M}$ is the polytope corresponding to the convex hull of the finite set $\varphi(\mathcal{Y}) \subset \mathcal{H}$. We will refer to $\mathcal{M}$ as the *marginal polytope* associated to the statistic φ, making an analogy to the literature on graphical models (Wainwright and Jordan, 2008). We denote by $k = \dim(\mathcal{H})$ the dimension of the embedding space and by $r = \dim(\mathcal{M})$ the dimension of the marginal polytope, defined as the dimension of its affine hull. Note that it can be the case that $r < k$, which means that $\mathcal{M}$ is not full-dimensional in $\mathcal{H}$.

[1] In general $f^\star$ is not unique, as there might be $x \in \mathcal{X}$ with more than one optimal outputs. For simplicity, we assume that we have a method to choose a unique output between the optimal ones. Note that this is always possible as $\mathcal{Z}$ is discrete, so one can always construct this method using an ordering of the elements of $\mathcal{Z}$.

Example 2 .1 (Multiclass and multilabel classification). The 0-1 loss used for k-multiclass classification ($\mathcal{Z} = \mathcal{Y} = \{1, \dots, k\}$) can be decomposed as $L(z, y) = 1(z \neq y) = 1 - \langle e_z, e_y \rangle$, where $\mathcal{H} = \mathbb{R}^k$ and e_z is the z-th vector of the canonical basis in $\mathbb{R}^k$. In this case, the loss matrix is full-rank and the marginal polytope is the simplex in k dimensions, $\mathcal{M} = \mathrm{hull}(\{e_y\}_{y=1}^k) = \Delta_k$, which is not full-dimensional and has dimension $r = k - 1$. Another example is the Hamming loss used for multilabel classification ($\mathcal{Z} = \mathcal{Y} = \{-1, 1\}^k$). In this case, the loss matrix is extremely low-rank and can be decomposed as $L(z, y) = \frac{1}{k} \sum_{j=1}^k 1(z_j \neq y_j) = \frac{1}{2} - \langle z/(2k), y \rangle$, where $k = \log |\mathcal{Y}|$. The marginal polytope is the cube $\mathcal{M} = \mathrm{hull}(\{-1, 1\}^k) = [-1, 1]^k$ which is full dimensional in $\mathcal{H} = \mathbb{R}^k$.

3 Surrogate Framework

3 .1 Estimation of the Conditional Risk with Surrogate Losses

The construction in Section 2 .2 leads to a natural method in order to estimate $f^\star$ based on estimating the conditional expectation $\mu^\star$. Indeed, given an estimator $\widehat{\mu}$ of $\mu^\star$, one can first construct an estimator of the conditional risk as $\widehat{\ell}(z, \rho(\cdot|x)) \coloneqq \langle \psi(z), \widehat{\mu}(x) \rangle + c$, and then define the resulting estimator as

$$\widehat{f}(x) \coloneqq \operatorname*{arg\,min}_{z \in \mathcal{Z}} \ \widehat{\ell}(z, \rho(\cdot|x)) = \operatorname*{arg\,min}_{z \in \mathcal{Z}} \ \langle \psi(z), \widehat{\mu}(x) \rangle.$$

In the following, we study a framework to construct estimators of $\mu^\star$ using surrogate losses. We consider estimators which are based on the minimization of the *expected surrogate risk* $\mathcal{R}(g)$ of a *surrogate loss* $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ defined in a (unconstrained) vector space $\mathcal{V}$,

$$g^\star = \operatorname*{arg\,min}_{g : \mathcal{X} \to \mathcal{V}} \mathcal{R}(g), \quad \mathcal{R}(g) = \mathbb{E}_{(X,Y) \sim \rho} \, S(g(X), Y). \tag{3.3}$$

An estimator $\widehat{f}$ of $f^\star$ is built from an estimator $\widehat{g}$ of $g^\star$ using a *decoding mapping* $d : \mathcal{V} \to \mathcal{Z}$ as $\widehat{f} = d \circ \widehat{g}$. The pair (S, d) constitutes a *surrogate method* and we say that it is *Fisher consistent* (Lin, 2004) to the loss L if the minimizer of the expected surrogate risk (3.3) leads to the minimizer of the true risk (3.1) as $f^\star = d \circ g^\star$.

Analogously to the quantities defined in Section 2 .1 for the discrete loss L, we define the *excess surrogate risk* as $\mathcal{R}(\widehat{g}) - \mathcal{R}(g^\star)$, the *conditional surrogate risk* $s(v, q) = \mathbb{E}_{Y \sim q} \, S(v, Y)$ and the *excess conditional surrogate risk* as $\delta s(v, q) = s(v, q) - \min_{v' \in \mathcal{V}} s(v', q) \geq 0$. Similarly to Eq. (3.2), $g^\star$ is characterized by $g^\star(x) = \arg\min_{v \in \mathcal{V}} s(v, \rho(\cdot|x))$, which we assume unique.

We will now focus on surrogate losses for which $\mu^\star$ can be computed from the minimizer $g^\star$ through a continuous *injective* mapping $t : \mathcal{M} \to \mathcal{V}$ called the *link function*. More precisely, we ask

$$t(\mu(q)) = \operatorname*{arg\,min}_{v \in \mathcal{V}} \ s(v, q), \quad \forall q \in \mathrm{Prob}(\mathcal{Y}). \tag{3.4}$$

Although Eq. (3.4) is the only property that we need from S in order to build the theoretical framework, we assume in the following that S is *smooth* and *convex*. This is justified in Remark 3 .1.

Remark 3 .1 (On smoothness and convexity requirement on S). Although we do not formalize any statement of that kind, the smoothness of S is closely related to the injectivity of t. For instance, in the multiclass case where $\mu(q) = q$, if $v_0 = \arg\min_{v\in\mathcal{V}} s(v, q_0)$ and $s(\cdot, q_0)$ is not differentiable at v_0, then one can find $q' \neq q_0$ such that $s(v_0, q') = s(v_0, q_0)$, and so the link is not injective. This is the case for hinge-type surrogates, which do not estimate conditional probabilities. A proper analysis in this direction can be formalized in terms of supporting hyperplanes on the so-called superdiction set associated to S (see Sec. 5.3 of Vernet et al. (2011)). The convexity requirement is made in order to be able to minimize in a tractable way the expected surrogate risk.

If a surrogate loss satisfies Eq. (3.4), then one can relate $f^\star$ and $g^\star$ using the decoding mapping $d_{\psi,t} : t(\mathcal{M}) \to \mathcal{Z}$ defined as

$$d_{\psi,t}(v) = \arg\min_{z\in\mathcal{Z}} \langle \psi(z), t^{-1}(v)\rangle.$$

The role of the link function here is to deal with the fact that the image of $\mu^\star$ lives in $\mathcal{M}$, which is a constrained, bounded, and possibly non full-dimensional set of $\mathcal{H}$. As in general it is not easy to impose a structural constraint on the hypothesis space, the goal of the link function is to encode this geometry by mapping points from a "simpler" $t(\mathcal{M}) \subseteq \mathcal{V}$ to $\mathcal{M}$. Note that $d_{\psi,t}$ is defined in $t(\mathcal{M})$, so if $t(\mathcal{M}) \neq \mathcal{V}$, we do not know how to map points from $\mathcal{V}\backslash t(\mathcal{M})$ to $\mathcal{H}$. In the next Section 3 .2 we show that in the cases where $t(\mathcal{M}) \neq \mathcal{V}$, the link can be sometimes naturally extended to cover the whole vector space $\mathcal{V}$. In order to do this, we first show that surrogates satisfying Eq. (3.4) have a very rigid structure in $t(\mathcal{M})$ in the form of a Bregman divergence representation. Then, we define φ-calibrated surrogates as the ones such that the corresponding Bregman divergence representation can be extended to $\mathcal{V}$.

Assume for now that $t(\mathcal{M}) = \mathcal{V}$. The surrogate method $(S, d_{\psi,t})$ works as follows; in the learning phase, an estimator $\widehat{g}$ is found by (regularized) empirical risk minimization on the smooth convex surrogate loss S, and then, given a new input element x, the decoding mapping $d_{\psi,t}$ computes the prediction $\widehat{f}(x)$ from $\widehat{g}$. Note that the computational complexity of inference can vary depending on the loss L (see Nowak et al. (2019); Ciliberto et al. (2016)). See boxes below.

Learning

- *Given*: a functional hypothesis space $\mathcal{G} \subset \{g : \mathcal{X} \to \mathcal{V}\}$, dataset $(x_i, y_i)_{1\leq i\leq n}$ and surrogate loss $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$.
- *Goal*: Minimize the expected surrogate risk $\mathcal{R}(g)$ as:

$$\widehat{g} = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} S(g(x_i), y_i) + \lambda\|g\|^2_{\mathcal{G}}. \tag{3.5}$$

Inference

- *Given*: an input element $x \in \mathcal{X}$, an estimator $\widehat{g} \in \mathcal{G}$, the inverse of the link function $t^{-1} : \mathcal{V} \to \mathcal{H}$ and an embedding $\psi : \mathcal{Z} \to \mathcal{H}$.
- *Goal*: Construct prediction $\widehat{f}(x) \in \mathcal{Z}$ as:

$$\widehat{f}(x) = d_{\psi,t} \circ \widehat{g}(x) = \arg\min_{z \in \mathcal{Z}} \left\langle \psi(z), t^{-1}(\widehat{g}(x)) \right\rangle . \tag{3.6}$$

3 .2 Bregman Divergence Representation

Let $\mathcal{D} \subseteq \mathcal{H}$ be a convex set. Recall that the Bregman divergence (BD) associated to a convex and differentiable function $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$ is defined as

$$D_h(u', u) = h(u') - h(u) - \langle u' - u, \nabla h(u) \rangle.$$

We will say that a surrogate loss S has a BD representation if the excess conditional surrogate risk $\delta s(v, q)$ can be written as a BD by composition with the link function.

Definition 3 .2 (BD Representation)**.** *The surrogate loss S has a (h, t, φ)-BD representation in $\mathcal{V}' \subset \mathcal{V}$, if there exists a set $\mathcal{D} \supseteq \mathcal{M}$ containing the marginal polytope, a strictly convex and differentiable potential $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$ and continuous invertible link $t : \mathcal{D} \to \mathcal{V}'$, such that the excess conditional surrogate risk can be written as*

$$\delta s(v, q) = D_h(\mu(q), t^{-1}(v)), \quad \forall v \in \mathcal{V}' \subset \mathcal{V}, \forall q \in \mathrm{Prob}(\mathcal{Y}).$$

The following Theorem 3 .3 states that any surrogate loss satisfying Eq. (3.4) has a BD representation in $t(\mathcal{M})$, which justifies why we focus on these representations of losses.

Theorem 3 .3 (BD Representation in $t(\mathcal{M})$)**.** *If the surrogate loss $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ is continuous and satisfies Eq.* (3.4) *for a continuous injective mapping $t : \mathcal{M} \to \mathcal{V}$, then it has a (h, t, φ)-BD representation in $t(\mathcal{M}) \subseteq \mathcal{V}$.*

The proof of Theorem 3 .3 can be found in Section A and it is based on a characterization of scoring rules for linear properties of a distribution as Bregman divergences associated to strictly convex functions (Abernethy and Frongillo, 2012; Frongillo and Kash, 2015b). The differentiability of h is derived from the continuity of the link t and S.

It is important to highlight the fact that the function h is defined up to an additive affine term, as the BD is invariant under this transformation. Hence, we will say that h and h' are equivalent if and only if $h - h'$ is an affine function. Note that the function h given by Theorem 3 .3 can be computed as

$$h(\mu(q)) = \delta s(v_0, q), \tag{3.7}$$

for any $v_0 \in t(\mathcal{M})$. Indeed, by Theorem 3 .3, the dependence on q of $\delta s(v_0, q)$ is only through the vector of moments $\mu(q)$ and $\delta s(v, q) - \delta s(v', q)$ is an affine function of $\mu(q)$ for all $v, v' \in t(\mathcal{M})$.

Observe that different surrogate losses can yield the same BD representation in $t(\mathcal{M})$. For instance, in binary classification, the square, squared hinge and modified Huber margin losses have the same BD representation in $[-1, 1]$ (Zhang, 2004b) (see Section F).

3 .3 φ-Calibrated Surrogates

Now, we define the concept of a φ-calibrated loss by asking the surrogate loss satisfying Eq. (3.4) to extend (in the case that $t(\mathcal{M}) \neq \mathcal{V}$) its (h, t, φ)-BD representation in $t(\mathcal{M})$ given by Theorem 3 .3 to $\mathcal{V}$, which will allow us to define the decoding mapping $d_{\psi,t}$ to the whole vector space $\mathcal{V}$.

Definition 3 .4 (φ-Calibrated Surrogates)**.** *Let $\varphi : \mathcal{Y} \to \mathcal{H}$. A smooth convex surrogate loss $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ is φ-calibrated if it has a (h, t, φ)-BD representation in the vector space $\mathcal{V}$.*

There are many ways of building a continuous extension of t to an invertible mapping in $\mathcal{V}$ (and thus to extend $d_{\psi,t}$), however, the BD representation extension allows to prove guarantees for estimators $\widehat{g}$ with $\mathrm{Im}(\widehat{g}) \not\subset t(\mathcal{M})$ (see Section 4). In general, it is not true that any surrogate loss satisfying Eq. (3.4) with $t(\mathcal{M}) \subsetneq \mathcal{V}$ has an extended BD representation in $\mathcal{V}$, this is the case for squared hinge and modified Huber margin losses in binary classification (see Section F).

Example 3 .5 (Quadratic, logistic and hinge surrogates). Let us provide some examples in binary classification where $\mathcal{M} = \Delta_2 \subset \mathcal{H} = \mathbb{R}^2$. The quadratic surrogate is defined as $S(v, y) = 1/2 \cdot \|v - e_y\|_2^2$ with $\mathcal{V} = \mathbb{R}^2$ and satisfies $q = \arg\min_{v \in \mathbb{R}^2} s(v, q)$. It has $h(u) = 1/2 \cdot \|u\|_2^2$ and the link is $t = Id$. Note that although $t(\Delta_2) = \Delta_2 \subsetneq \mathbb{R}^2$, the BD representation can be extended to $\mathbb{R}^2$, and in this case $\mathcal{D} = \mathbb{R}^2$ and so it is φ-calibrated. The logistic corresponds to $S(v, y) = \log(1 + e^{-yv})$ with $\mathcal{V} = \mathbb{R}$ and satisfies $\log(q_1/(1 - q_1)) = \arg\min_{v \in \mathbb{R}} s(v, q)$. In this case the potential is minus the entropy $h(q) = -\mathrm{Ent}(q)$ and the link is $t(q) = \log(q_1/(1 - q_1))$ with inverse $t^{-1}(v) = (1 + e^{-v})^{-1}$. Note that we have $t(\Delta_2) = \mathbb{R}$, so it is φ-calibrated. Finally, consider the hinge margin loss $S(v, y) = \max(1 - yv, 0)$ with $\mathcal{V} = \mathbb{R}$, which satisfies $\mathrm{sign}(2q_1 - 1) = \arg\min_{v \in \mathbb{R}} s(v, q)$, hence, t is not injective, so S is not φ-calibrated.

Note that if a loss S is φ-calibrated for a statistic φ, then the surrogate method $(S, d_{\psi,t})$ is Fisher consistent w.r.t $L(z, y) = \langle \psi(z), \varphi(y) \rangle + c$. This implies that a φ-calibrated loss can be used to consistently minimize different losses by simply changing the embedding ψ at inference time. For instance, if S is φ-calibrated for the statistic $\varphi(y) = e_y \in \mathbb{R}^{\mathcal{Y}}$, then it can be made consistent for any cost-sensitive matrix loss $L \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}}$ by setting $\psi(z) = L_z$, where L_z is the z-th row of L. Indeed, in this case $\mathcal{M} = \Delta_{\mathcal{Y}}$, and so one can estimate the conditional risk of any loss with labels $\mathcal{Y}$.

Summary. The surrogate loss has two components; the potential $h : \mathcal{D} \to \mathbb{R}$ and the invertible link function $t : \mathcal{D} \to \mathcal{V}$, which compose the surrogate loss S. In the learning phase (see Eq. (3.5)), only the surrogate loss is needed to minimize $\mathcal{R}(g)$, while in the inference phase (see Eq. (3.6)), one needs the inverse of the link to construct an estimate of $\mu^\star$, and the rest of the inference only depends on ψ. The potential function h is not needed to define the surrogate method but it is the mathematical object providing the guarantees in order to relate both excess risks in Section 4 . The link function also has implications in terms of learning complexity (see discussion in Section 4 .4). See Figure 3.2 in Section A for an illustrative diagram.

We now provide a recipe on how to check whether a surrogate loss is φ-calibrated and to compute its corresponding (h, t, φ)-BD representation if applicable.

Computing the BD representation and checking φ-calibration. Given a statistic $\varphi : \mathcal{Y} \to \mathcal{H}$ and a surrogate $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$, the first thing to do is to check whether the minimizer of $s(v, q)$ satisfies Eq. (3.4) for a continuous injective t. If this is the case, the potential h can be found up to an additive affine term by Eq. (3.7). If $t(\mathcal{M}) = \mathcal{V}$, then S is φ-calibrated. Otherwise, one has to check if there exists an extension of t and h such that $\delta s(v, q) = D_h(\mu(q), t^{-1}(v))$ for all $v \in \mathcal{V}, q \in \mathrm{Prob}(\mathcal{Y})$. We provide numerous examples in Section 5 and in the Appendix.

We present now a special group of φ-calibrated surrogates, whose potential h is a function of Legendre-type and the link is the gradient of the potential ∇h.

φ-Calibrated surrogates of Legendre-type. A function h is of Legendre-type in $\mathcal{D} \subseteq \mathcal{H}$ if it is strictly convex in $\mathrm{int}(\mathcal{D})$ and essentially smooth, which in particular requires $\lim_{u \to \partial\mathcal{D}} \|\nabla h(u)\|_2 = +\infty$, where $\partial\mathcal{D}$ is the boundary of $\mathcal{D}$. Given a Legendre-type function h with domain $\mathcal{D} \supseteq \mathcal{M}$ including the marginal polytope, one can set the link function to $t = \nabla h$. We call it the *canonical link*. It has the nice property that if h is of Legendre-type in $\mathcal{D}$, then its Fenchel conjugate h^* is also of Legendre-type and its gradient is the inverse of the link function $\nabla h^* = (\nabla h)^{-1}$. We denote the resulting loss $S : \mathrm{dom}(h^*) \times \mathcal{Y} \to \mathbb{R}$ a *surrogate loss of Legendre-type*, which is *convex* and has the form:

$$S(v, y) = D_h(\varphi(y), \nabla h^*(v)) = h^*(v) + h(\varphi(y)) - \langle \varphi(y), v \rangle, \tag{3.8}$$

The excess conditional surrogate risk can be written as a BD also in $\mathrm{dom}(h^*)$ as:

$$\delta s(v, q) = D_h(\mu(q), \nabla h^*(v)) = D_{h^*}(v, \nabla h(\mu(q))).$$

Moreover, $\mathcal{D}$ is bounded if and only $\mathrm{dom}(h^*)$ is a vector space and h^* is Lipschitz. Those losses were studied by Blondel et al. (2020) as a subset of Fenchel-Young losses, but without providing learning guarantees. The most important examples are the quadratic surrogate, where $\mathcal{D} = \mathcal{H}$, and CRFs, where $\mathcal{D} = \mathcal{M}$, both studied in detail in Section 5 .1. Further details on this construction can be found in Section B .

4 Theoretical Analysis

We know by construction that φ-calibrated surrogate losses lead to Fisher consistent surrogate methods $(S, d_{\psi,t})$, which means that the minimizer of the surrogate risk $\mathcal{R}$ provides the minimizer of the true risk $\mathcal{E}$ as $f^\star = d \circ g^\star$. However, in practice we will never be able to minimize the surrogate risk to optimality. The goal of this section is to *calibrate* the excess surrogate risk to the true excess risk, i.e., quantify how much the excess surrogate risk has to be minimized so that the excess true risk is smaller than ε. This quantification is made by means of the *calibration function* (Steinwart, 2007; Osokin et al., 2017; Bartlett et al., 2006; Duchi et al., 2010), which is the mathematical object that will allow us to relate the quantity we can directly minimize to the one that we are ultimately interested in. All the proofs from this section can be found in Section C .

4 .1 Calibrating Risks with the Calibration Function

The calibration function is defined as the largest function $\zeta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ that relates both excess conditional risks as $\zeta(\delta\ell(d(v), q)) \leq \delta s(v, q)$, $\forall v \in \mathcal{V}, \forall q \in \mathrm{Prob}(\mathcal{Y})$. The calibration function for general losses is thus defined as follows.

Definition 4 .1 (Calibration function (Osokin et al., 2017)). *he calibration function* $\zeta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ *is defined for* $\varepsilon \geq 0$ *as the infimum of the excess conditional surrogate risk when the excess conditional risk is at least* ε*:*

$$\zeta(\varepsilon) = \inf \delta s(v, q) \quad \textit{such that} \quad \delta\ell(d(v), q) \geq \varepsilon,\ q \in \mathrm{Prob}(\mathcal{Y}),\ v \in \mathcal{V}.$$

We set $\zeta(\varepsilon) = \infty$ *when the feasible set is empty.*

Note that ζ is non-decreasing on $[0, +\infty)$, not necessarily convex (see Example 5 by Bartlett et al. (2006)) and also $\zeta(0) = 0$. Note that a larger ζ is better because we want a large $\delta s(v, q)$ to incur small $\delta\ell(d(v), q)$.

The calibration function ζ relates conditional risks. In order to calibrate risks $\mathcal{R}$ and $\mathcal{E}$ one needs to impose convexity so that the expectation with respect to the marginal distribution $\rho_{\mathcal{X}} \in \mathrm{Prob}(\mathcal{X})$ can be moved outside of the calibration function. In Theorem 4 .2 we calibrate the risks by taking a convex lower bound of ζ (Steinwart, 2007).

Theorem 4 .2 (Calibration between risks (Steinwart, 2007)). *Let* $\bar{\zeta}$ *be a convex lower bound of* ζ*. We have*

$$\bar{\zeta}(\mathcal{E}(d \circ \widehat{g}) - \mathcal{E}(f^\star)) \leq \mathcal{R}(\widehat{g}) - \mathcal{R}(g^\star) \tag{3.9}$$

for all $\widehat{g} : \mathcal{X} \to \mathcal{V}$*. The tightest convex lower bound* $\bar{\zeta}$ *of* ζ *is its lower convex envelope which is defined by the Fenchel bi-conjugate* ζ^{**} [2]*.*

Note that a surrogate method is Fisher consistent if and only if $\zeta^{**}(\varepsilon) > 0$ for all $\varepsilon > 0$, as this implies $f^\star = d \circ g^\star$. In the case that $\zeta^{**} \neq \zeta$, this property also translates to ζ. See Figure 3.1.

4 .2 Calibration Function for φ-Calibrated Losses

The computation of ζ (or a convex lower bound thereof) is known not to be easy and has been a central topic of study for many past works (Bartlett et al., 2006; Pires et al., 2013; Osokin et al., 2017). One of the main contributions of this work is to provide an exact formula for ζ for φ-calibrated losses based on Bregman divergences between pairs of sets in $\mathcal{H}$. This geometric interpretation of the calibration function will be used to compute the calibration function for existing surrogates which are widely used in practice.

First, let us define the *calibration sets* $\mathcal{H}_\varepsilon(z)$ for every $\varepsilon \geq 0$ and $z \in \mathcal{Z}$ as

$$\mathcal{H}_\varepsilon(z) = \{u \in \mathcal{H} \mid \langle \psi(z) - \psi(z'), u \rangle \leq \varepsilon, \forall z' \in \mathcal{Z}\} \subset \mathcal{H}.$$

The points in $\mathcal{H}_\varepsilon(z)$ are the ones whose conditional risk is at least ε-close to have z as optimal prediction. In particular, $\mathcal{H}_0(z)$ is the set of points with optimal prediction z, which can be equivalently written as $\mu^\star(x) \in \mathcal{H}_0(f^\star(x))$, $\forall x \in \mathcal{X}$. Note that $\mathcal{H}_\varepsilon(z)$ is convex $\forall \varepsilon \geq 0, \forall z \in \mathcal{Z}$. See Figure 3.1 for a visualization of the calibration sets for the Hamming loss in the context of multilabel classification (and Figure 3.4 in Section G for the 0-1 loss for multiclass classification).

Theorem 4 .3 (Calibration function for φ-calibrated losses). *Let* S *be a* φ*-calibrated surrogate with potential function* $h : \mathcal{D} \to \mathbb{R}$ *and let* $L(z, y) = \langle \psi(z), \varphi(y) \rangle + c$*. The calibration function only depends on* S *through* h *and we denote it by* ζ_h*. Moreover, it can be written as*

$$\zeta_h(\varepsilon) = \min_{z \in \mathcal{Z}} D_h(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z) \cap \mathcal{D}), \tag{3.10}$$

[2] The Fenchel bi-conjugate is characterized by $\mathrm{epi}(\zeta^{**}) = \overline{\mathrm{hull}(\mathrm{epi}(\zeta))}$, where $\mathrm{epi}(\zeta)$ denotes the epigraph of the function ζ and $\overline{\mathrm{hull}(A)}$ is the closure of the convex hull of the set A.

where the Bregman divergence between sets A, B is defined as $D_h(A, B) = \inf_{u\in A, v\in B} D_h(u, v)$.

Note that the Bregman divergence inside the minimum in Eq. (3.10) does not lead to a convex minimization problem since $\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}$ is not convex and $D_h(u, v)$ is in general not jointly convex in (u, v), with notable exceptions such as the KL-divergence and squared distance (Bauschke and Borwein, 2001). In general, the exact computation of ζ_h using Theorem 4 .3 can still be hard to perform, for instance, when the embeddings ψ are not simple to work with or the problem lacks symmetries. In the following Theorem 4 .4 we provide a user-friendly lower bound when the potential h is strongly convex. Recall that a function h is $(1/\beta_{\|\cdot\|})$-strongly convex w.r.t a norm $\|\cdot\|$ in $\mathcal{D}$ if it satisfies $h(u) \geq h(v) + \langle u - v, \nabla h(v)\rangle + \frac{1}{2\beta_{\|\cdot\|}}\|u - v\|^2, \forall u, v \in \mathcal{D}$.

Theorem 4 .4 (User-friendly lower bound on ζ_h)**.** *Let $\zeta_h(\varepsilon)$ be the calibration function given by Eq.* (3.10)*. If h is $(1/\beta_{\|\cdot\|})$-strongly convex w.r.t a norm $\|\cdot\|$ in $\mathcal{D}$, then:*

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{8c^2_{\psi,\|\cdot\|_*}\beta_{\|\cdot\|}},$$

where $c_{\psi,\|\cdot\|_*} = \sup_{z\in\mathcal{Z}} \|\psi(z)\|_*$ and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

The proof is provided in Section D .2, together with Theorem D .4, that gives a tighter bound in the case of strong convexity w.r.t the Euclidean norm. Finally, the following Theorem 4 .5, states that ζ_h can never be larger than a quadratic for φ-calibrated surrogates.

Theorem 4 .5 (Existence of quadratic upper bound)**.** *Assume h is twice differentiable. Then, the calibration function ζ_h is upper bounded by a quadratic close to the origin, i.e., $\zeta_h(\varepsilon) = O(\varepsilon^2)$.*

4 .3 Improved Calibration under Low Noise Assumption

The result of Theorem 4 .2 can be further improved under low noise assumptions on the marginal distribution $\rho_\mathcal{X}$. Following the definition from classification (Bartlett et al., 2006; Mroueh et al., 2012; Zhang, 2004a), we define the *margin function* $\gamma : \mathcal{X} \to \mathbb{R}$ as $\gamma(x) = \min_{z'\neq f^\star(x)} \delta\ell(z', \rho(\cdot|x))$. We say that the *p-noise condition* is satisfied if

$$\rho_\mathcal{X}(\gamma(X) \leq \varepsilon) = o(\varepsilon^p). \tag{3.11}$$

A simple computation shows that Eq. (3.11) holds if and only if $\|1/\gamma\|_{L_p(\rho_\mathcal{X})} = \gamma_p < \infty$ (Steinwart and Christmann, 2011).

Theorem 4 .6 (Calibration of Risks under low noise and hard margin assumption)**.** *Let $\bar{\zeta}$ be a convex lower bound of ζ.*

(1) If the p-noise condition (3.11) *is satisfied, we have that $\bar{\zeta}^{(p)}$ defined as*

$$\bar{\zeta}^{(p)}(\varepsilon) = (\gamma_p\varepsilon^p)^{\frac{1}{p+1}}\bar{\zeta}((\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}/2), \tag{3.12}$$

satisfies Eq. (3.9) *where $\|1/\gamma\|_{L_p(\rho_\mathcal{X})} = \gamma_p$. Moreover, we have that $\bar{\zeta}^{(p)} \gtrsim \bar{\zeta}$. Hence, $\bar{\zeta}^{(p)}$ never provides a worse rate than $\bar{\zeta}$.*

(2) If $\gamma(x) \geq \delta > 0$ $\rho_\mathcal{X}$-a.s. Then, $\delta v(\widehat{g}(x), \rho(\cdot|x)) < \zeta(\delta),\ \rho_\mathcal{X}$-a.s. $\implies \mathcal{E}(d \circ \widehat{g}) = \mathcal{E}(f^\star)$.

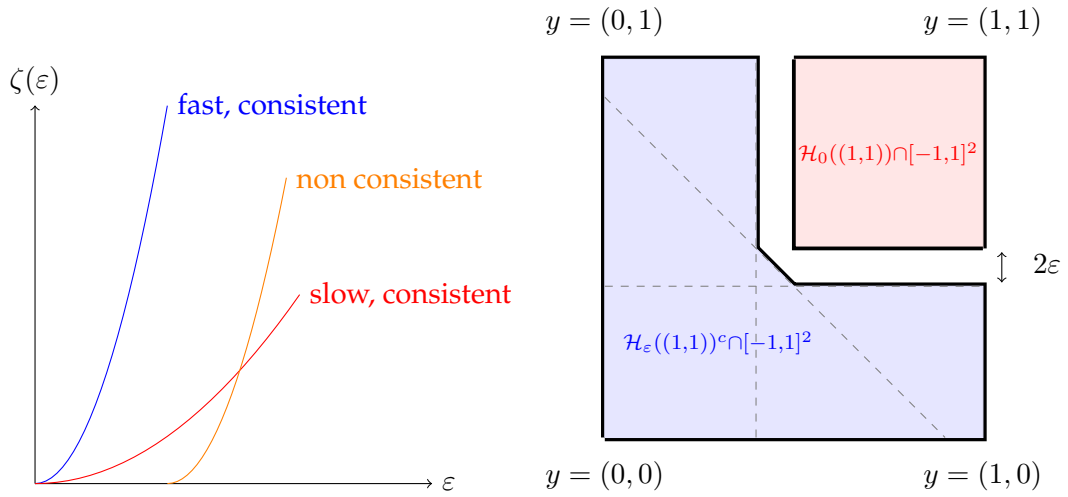


Figure 3.1: **Left:** Both red and blue curves correspond to calibration functions of Fisher consistent surrogate methods as $\zeta(\varepsilon) > 0$ for all $\varepsilon > 0$, which is not the case for the orange curve. However, the blue curve has better guarantees because the same surrogate excess risk leads to smaller excess true risk. **Right:** Illustration of the sets $\mathcal{H}_0(z) \cap \mathcal{M}$ and $\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}$ for the Hamming loss with $k = 2$ labels and $z = (1,1)$. In this case, by symmetry, the calibration function is computed as the Bregman divergence between these two sets.

The first part of Theorem 4 .6 is a generalization of Thm. 10 by Bartlett et al. (2006) to general discrete losses. Note that combining Eq. (3.12) with the lower bound given by Theorem 4 .4 gives $\bar{\zeta}^{(p)} \gtrsim \varepsilon^{(\frac{p+2}{p+1})}$. Indeed, $p = 0$ corresponds to no assumption at all and so $\zeta^{(0)}$ stays quadratic, while $p \to \infty$ corresponds to having less and less noise at the boundary decision and $\zeta^{(p)}$ tends to be linear. Note that $p = \infty$ corresponds to having $\delta > 0$ such that $\gamma(x) \geq \delta > 0$ $\rho_{\mathcal{X}}$-a.s, and so from the second part of Theorem 4 .6, one obtains zero excess risk if the excess surrogate conditional risk is smaller than $\zeta(\delta)$ almost surely. This fact has been used in binary classification together with high probability bounds on the estimator to obtain exponential rates of convergence for the risk $\mathcal{E}$ (Audibert and Tsybakov, 2007; Koltchinskii and Beznosova, 2005; Pillaud-Vivien et al., 2018b), and our result could be used in the same way for the structured case. Finally, note that Theorem 4 .2 and Theorem 4 .6 are not specific to φ-calibrated surrogates and apply to any surrogate method.

4 .4 Minimizing the Surrogate Loss with Averaged Stochastic Gradient Descent (ASGD)

In this section, for simplicity, we assume $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ is a loss of Legendre-type (see Eq. (3.8)) with associated Legendre-type potential $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$ and $\mathcal{V} = \mathrm{dom}(h^*) = \mathbb{R}^{k'}$. Following Osokin et al. (2017), we provide a statistical analysis of the minimization of the expected risk of S using online projected averaged stochastic gradient descent (ASGD) (Nemirovski et al., 2009) on a reproducing kernel Hilbert space (RKHS) [3] (Aron-

[3]Recall that a scalar RKHS $\bar{\mathcal{G}}$ is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ with an associated kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, \cdot) \in \bar{\mathcal{G}}$ for all $x \in \mathcal{X}$ and $g(x) = \langle g, k(x, \cdot) \rangle_{\bar{\mathcal{G}}}$ for all $g \in \bar{\mathcal{G}}$.

szajn, 1950) $\mathcal{G} = (\bar{\mathcal{G}})^{\otimes k'} = \mathcal{V} \otimes \bar{\mathcal{G}}$, where $\bar{\mathcal{G}}$ is a scalar RKHS. This will give us insight on the important quantities for the design of the surrogate method when minimizing a discrete loss L. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel associated to the RKHS $\bar{\mathcal{G}}$ and $\sup_{x \in \mathcal{X}} k(x,x) \leq \kappa^2$. The n-th update of ASGD reads

$$\widehat{g}_n = \frac{1}{n} \sum_{i=1}^n g_i, \quad g_i = \Pi_D \left(g_{i-1} - \eta_i \nabla S(g_{i-1}(x_i), y_i) \otimes k(x_i, \cdot) \right),$$

where ∇S denotes the gradient of S w.r.t the first coordinate, η_i is the step size and Π_D is the projection onto the ball of radius D w.r.t the norm induced by $\mathcal{G}$. We have the following theorem.

Theorem 4 .7. *Let $S : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ be a loss of Legendre-type with associated $(1/\beta_{\|\cdot\|_2})$-strongly convex h. Let $(x_i, y_i)_{i=1}^n$, with $n \in \mathbb{N}$ be independently and identically distributed according to ρ and assume $g^\star \in \mathcal{G}$ and $\|g^\star\|_{\mathcal{G}}^2 = \sum_{j=1}^{k'} \|g_j^\star\|_{\bar{\mathcal{G}}}^2 \leq D^2$. Let $C^2 = 1 + \sup_{y \in \mathcal{Y}} \|\varphi(y) - \nabla h^*(0)\|_2 / (\kappa \beta_{\|\cdot\|_2} D)$, then, by using the constant step size $\eta = 2/(\beta_{\|\cdot\|_2} \kappa^2 C^2 \sqrt{n})$, we have*

$$\mathbb{E}\left[\mathcal{E}(d \circ \widehat{g}_n) - \mathcal{E}(f^\star)\right] \leq \frac{4 \cdot \kappa \cdot c_{\psi, \|\cdot\|_2} \cdot \beta_{\|\cdot\|_2} \cdot D \cdot C}{n^{1/4}}. \tag{3.13}$$

Proof. Let's first compute a uniform bound on the gradients as

$$\begin{aligned}
\|\nabla S(g(x), y) \otimes k(x, \cdot)\|_{\mathcal{G}} &\leq \kappa \|\nabla S(g(x), y)\|_2 \\
&= \kappa \|\nabla h^*(g(x)) - \varphi(y)\|_2 \\
&\leq \kappa (\|\nabla h^*(g(x)) - \nabla h^*(0)\|_2 + \|\varphi(y) - \nabla h^*(0)\|_2) \\
&\leq \kappa (\kappa \beta_{\|\cdot\|_2} D + \sup_{y \in \mathcal{Y}} \|\varphi(y) - \nabla h^*(0)\|_2) = M,
\end{aligned}$$

where at the first step we have used that $\|k(x, \cdot)\|_{\bar{\mathcal{G}}} \leq \kappa$ and at the last step that h^* is $\beta_{\|\cdot\|_2}$-smooth because h is $(1/\beta_{\|\cdot\|_2})$-strongly convex, $\|g(x)\|_2^2 = \sum_{j=1}^{k'} \langle g_j, k(x, \cdot) \rangle_{\bar{\mathcal{G}}}^2 \leq \sum_{j=1}^{k'} \kappa^2 \|g_j\|_{\bar{\mathcal{G}}}^2 = \kappa^2 \|g\|_{\mathcal{G}}^2 \leq \kappa^2 D^2$ and that $\nabla h^*(v) - \nabla h^*(0)$ vanishes at the origin. Using classical results on ASGD (Nemirovski et al., 2009), we know that using the constant step size $\eta = 2D/(M\sqrt{n})$, we have that $\mathbb{E}[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^\star)] \leq 2DM/n^{1/2}$ after n iterations of ASGD. Finally, applying the lower bound on ζ in Theorem 4 .4, re-arranging terms, and using the fact that $\mathbb{E}[w] \leq \sqrt{\mathbb{E}[w^2]}$, for $w = \mathcal{E}(d \circ \widehat{g}_n) - \mathcal{E}(f^\star) \geq 0$, we obtain the bound (3.13). □

Note that Eq. (3.13) is bounded by $8\kappa^{1/2} c_{\psi, \|\cdot\|_2} \max(\kappa^{1/2} \beta_{\|\cdot\|_2} D, (c_{\varphi,h} \beta_{\|\cdot\|_2} D)^{1/2}) / n^{1/4}$, where $c_{\varphi,h} = \sup_{y \in \mathcal{Y}} \|\varphi(y) - \nabla h^*(0)\|_2$. There are essentially 4 quantities appearing in the bound (3.13): $c_{\psi, \|\cdot\|_*}$ that depends on L, $c_{\varphi,h}$ that bounds the marginal polytope centered at $\nabla h^*(0)$, $\beta_{\|\cdot\|}$ that depends on h and D, which is an upper bound on the norm of the optimum $\|g^\star\|_{\mathcal{G}} = \|\nabla h(\mu^\star)\|_{\mathcal{G}}$ which depends on the link, in this case ∇h, the RKHS $\mathcal{G}$, and $\mu^\star$. Note that the image of $\mu^\star$ lies in the marginal polytope, which is bounded and potentially non full-dimensional in $\mathcal{H}$, so if one directly estimates $\mu^\star$, the hypothesis space $\mathcal{G}$ has to model this constraint. The role of the link function is to remove this additional complexity from $\mathcal{G}$ by mapping the marginal polytope (or a superset $\mathcal{D}$ of it) to the vector space $\mathcal{V}$, and consequently smoothing out $\mu^\star$ close to the boundary of $\mathcal{M}$, leading to a smaller $\|g^\star\|_{\mathcal{G}}$. The types of surrogates that directly estimate $\mu^\star$ are of quadratic-type (see Section 5 .1), which have $\mathcal{D} = \mathcal{H}$ and the link is the identity. In this case, $\mathcal{G}$ has to be able to model the fact that $\mathrm{Im}(\mu^\star) \subseteq \mathcal{M}$. The second types are of logistic-type (see CRFs

in Section 5 .1), which have $\mathcal{D} = \mathcal{M}$, and as $\lim_{\mu \to \partial \mathcal{M}} \|\nabla h(\mu)\| = +\infty$, the link smooths out $\mu^\star$ close to the boundary. In this case, $\|g^\star\|_{\mathcal{G}}$ can potentially be much smaller as $\mathcal{G}$ does not have to model the polytope constraint. This generalizes the idea that the logistic link is preferable for estimating class-conditional probabilities, for instance, when using linear hypothesis spaces. In between the two, there are methods with bounded $\mathcal{D}$ but different than $\mathcal{M}$, such as one-vs-all methods in multiclass classification, where $\mathcal{M} = \Delta_k \subsetneq \mathcal{D} \subsetneq \mathcal{H}$ (see Section G).

5 Analysis of Existing Surrogate Methods

In this section we apply the theory developed so far to derive new results on multiple surrogate methods used in practice. In Section 5 .1, we study two generic methods for structured prediction, namely, the quadratic surrogate (Ciliberto et al., 2016; Nowak et al., 2019; Ciliberto et al., 2019) and conditional random fields (CRFs) (Lafferty et al., 2001; Settles, 2004). Then, in Section 5 .2, we present new theoretical results on multiple tasks in supervised learning which can be derived using results from Section 4 . The proofs of the results and further details can be found in Section E for Section 5 .1, and from Section F to Section K for Section 5 .2.

5 .1 Optimizing generic losses: Quadratic surrogate vs. CRFs

Quadratic Surrogate for Structured Prediction. The quadratic surrogate for structured prediction (Ciliberto et al., 2016, 2019) has the form

$$\mathcal{V} = \mathcal{H}, \quad S(v, y) = \frac{1}{2}\|v - \varphi(y)\|_2^2.$$

This is a loss of Legendre-type with $\mathcal{D} = \mathcal{H}$, $h(u) = \frac{1}{2}\|u\|_2^2$ and $t^{-1}(u) = \nabla h^*(u) = u$. We can exactly compute the calibration function when $\mathcal{M}$ is full-dimensional. Theorem 5 .1 is a simpler version of the result which holds for ε small enough, the complete result can be found in Section E .1.

Theorem 5 .1. *If $\mathcal{M}$ is full-dimensional, there exists $\varepsilon_0 > 0$ such that*

$$\zeta_h(\varepsilon) = \frac{\varepsilon^2}{2 \max_{(z,z') \in A} \|\psi(z) - \psi(z')\|_2^2}, \quad \forall \varepsilon \leq \varepsilon_0,$$

where $A = \{(z, z') \in \mathcal{Z}^2 \mid z' \neq z, \mathcal{H}_0(z) \cap \mathcal{H}_0(z') \neq \varnothing\}$.

Note that in this case as h is 1-strongly convex with respect to the Euclidean norm, Theorem 4 .4 gives $\zeta_h(\varepsilon) \geq \varepsilon^2 \cdot (8 c_{\psi, \|\cdot\|_2}^2)^{-1}$, and we recover the comparison inequality from Ciliberto et al. (2016). In Section E .1 we compare this result with the lower bounds on the calibration function for the quadratic-type surrogates studied by Osokin et al. (2017). An interesting property of the quadratic surrogate is that one can build the estimator $\widehat{f}$ independently of the affine decomposition of L by minimizing the expected surrogate risk with kernel ridge regression. In particular, this allows to extend the framework to continuous losses defined in compact sets $\mathcal{Z}, \mathcal{Y}$ where $\mathcal{H}$ can be infinite-dimensional (Ciliberto et al., 2016).

Conditional Random Fields. Recall that $r = \dim(\mathcal{M})$. CRFs correspond to

$$\mathcal{V} = \mathbb{R}^r, \quad S(v, y) = \log(\textstyle\sum_{y' \in \mathcal{Y}} \exp(\langle v, \varphi(y') \rangle)) - \langle v, \varphi(y) \rangle. \tag{3.14}$$

This is a loss of Legendre-type with $\mathcal{D} = \mathcal{M}$ and $h(u) = -\max_{q \in \mathrm{Prob}(\mathcal{Y})} \mathrm{Ent}(q)$ such that $\mu(q) = u$, where $\mathrm{Ent}(q) = -\sum_{y \in \mathcal{Y}} q(y) \log q(y)$ is the Shannon entropy of the distribution $q \in \mathrm{Prob}(\mathcal{Y})$ [4]. In this case, the inverse of the link function $t^{-1} = \nabla h^*$ corresponds to performing marginal inference on the exponential family with sufficient statistics φ. A well-known important drawback of CRFs is the fact that they are in general not calibrated to any specific loss. The reason being that inference in CRFs is done using MAP assignment, which corresponds to $\widehat{f}_{\mathrm{MAP}}(x) = \arg\max_{y \in \mathcal{Y}} \langle \varphi(y), \widehat{g}(x) \rangle$. It can be written in terms of ∇h^* as $\widehat{f}_{\mathrm{MAP}}(x) = \varphi^{-1}(\lim_{\gamma \to \infty} \nabla h^*(\gamma \widehat{g}(x)))$ (see Section E .2 for the computation), which is different from the decoding $d_{\psi, \nabla h}$ we propose: $\widehat{f}(x) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \nabla h^*(\widehat{g}(x)) \rangle$. Note that $\nabla h^*(\gamma \widehat{g}(x))$ converges to a vertex of $\mathcal{M}$ as $\gamma \to \infty$, while decoding $d_{\psi, \nabla h}$ partitions $\mathcal{M}$ into $|\mathcal{Z}|$ regions using ψ and assigns a different output to each of those. With our proposed decoding, we can calibrate CRFs to any loss that decomposes with the cliques of the probabilistic model. For instance, for linear chains, the method consistently minimizes any loss that depends on the neighbors. Moreover, we can compute a lower bound on ζ_h.

Proposition 5 .2 (Calibration of CRFs)**.** *The calibration function of CRFs can be lower bounded as*

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{8 c^2_{\psi, \|\cdot\|_2} c^2_{\varphi, \|\cdot\|_2}},$$

where $c_{\varphi, \|\cdot\|_2} = \sup_{y \in \mathcal{Y}} \|\varphi(y)\|_2$.

5 .2 Specific Problems

Binary Classification. In this case $\mathcal{Z} = \mathcal{Y} = \{-1, 1\}$. See Example 2 .1 setting $k = 2$ for the structure of the loss. We consider *margin losses*, which are losses of the form $S(v, y) = \Phi(yv)$ with $\mathcal{V} = \mathbb{R}$, where $\Phi : \mathbb{R} \to \mathbb{R}$ is a non-increasing function with $\Phi(0) = 1$. The decoding simplifies to $d(v) = \mathrm{sign}(v)$. The logistic, exponential ($\mathcal{D} = [0, 1] = \Delta_2$) and square ($\mathcal{D} = \mathbb{R} \supsetneq \Delta_2$) margin losses are φ-calibrated. The calibration function is $\zeta_h(\varepsilon) = h\left(\frac{1+\varepsilon}{2}\right) - h\left(\frac{1}{2}\right)$ (Bartlett et al., 2006; Scott, 2012), where $h : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$ is the associated potential.

Multiclass Classification. In this case $\mathcal{Z} = \mathcal{Y} = \{1, \ldots, k\}$. See Example 2 .1 for the structure of the loss. The *one-vs-all method* corresponds to $S(v, y) = \Phi(v_y) + \sum_{j \neq y}^{k} \Phi(-v_j)$ with $\mathcal{V} = \mathbb{R}^k$. If the margin loss $\Phi(y_j v_j)$ is φ-calibrated for binary classification with potential $\bar{h} : \bar{\mathcal{D}} \subset \mathbb{R} \to \mathbb{R}$, then S is φ-calibrated with $h : \mathcal{D} = \bar{\mathcal{D}}^k \subset \mathbb{R}^k \to \mathbb{R}$ defined as $h(u) = \sum_{j=1}^{k} \bar{h}(u_j)$. Note that the marginal polytope is strictly included in $\mathcal{D}$: $\Delta_k \subsetneq [0, 1]^k \subseteq \mathcal{D}$. The decoding can be simplified to $d(v) = \arg\max_{j \in [k]} v_j$ and the calibration function has the form given by Proposition 5 .3.

Proposition 5 .3 (One-vs-all calibration function)**.** *Assume $\bar{h}''$ is non-decreasing in $\bar{\mathcal{D}} \cap [1/2, +\infty)$. Then, the calibration function for the one-vs-all method is $\zeta_h(\varepsilon) = 2 \cdot \zeta_{\bar{h}}(\varepsilon)$.*

[4] Note that here, for simplicity, we do not consider the constant term $h(\varphi(y))$ from Eq. (3.8).

Note that the assumption on $\bar{h}''$ is met by the logistic, exponential and square binary margin losses. Another important example is the *multinomial logistic* loss, which corresponds to (3.14) for multiclass. In this case the decoding is also simplified to $d(v) = \arg\max_{j\in[k]} v_j$, and $\zeta_h(\varepsilon) \geq \varepsilon^2/8$ by using strong convexity of the entropy w.r.t $\|\cdot\|_1$ norm on Theorem 4 .4.

Multilabel Classification. In this case $\mathcal{Z} = \mathcal{Y} = \{-1, 1\}^k$. See Example 2 .1 for the structure of the loss. We consider *independent classifiers*, which have the form: $S(v, y) = \sum_{j=1}^k \Phi(y_j v_j)$, with $\mathcal{V} = \mathbb{R}^k$. In this case the potential has the form $h(u) = \sum_{j=1}^k \bar{h}((u_j + 1)/2)$, where $\bar{h}$ is the potential for the individual classifier. In this case $\mathcal{M}$ equals $\mathcal{D}$ for logistic and exponential classifiers, as they have $\bar{\mathcal{D}} = [0, 1]$. The decoding is simplified to $d(v) = (\text{sign}(v_j))_{j=1}^k$ and the calibration function ζ_h can be computed exactly and it is linear in the number of labels $\zeta_h(\varepsilon) = k \cdot \zeta_{\bar{h}}(\varepsilon)$ (see Proposition H .1 in Section H).

Ordinal Regression. In this case $\mathcal{Z} = \mathcal{Y} = \{1, \ldots, k\}$ with an ordering: $1 \prec \cdots \prec k$. We consider the absolute error loss function defined as $L(z, y) = |z - y|$. We analyze two types of surrogates, the *all thresholds (AT)* (Lin and Li, 2006) and the *cumulative link (CL)* (McCullagh, 1980), both studied by Pedregosa et al. (2017). AT methods correspond to independent classifiers $S(v, y) = \sum_{j=1}^{k-1} \Phi(\varphi_j(y)v_j)$ with $\mathcal{V} = \mathbb{R}^{k-1}$, where $\varphi(y) = (2 \cdot 1(y_j \geq j) - 1)_{j=1}^{k-1}$. In this case, using results from the Hamming loss we show that $\zeta_h(\varepsilon) \geq (k-1) \cdot \zeta_{\bar{h}}(\varepsilon/(k-1)) \sim \varepsilon^2/(k-1)$, where $\bar{h}$ is the potential for the individual classifier. CL methods, instead, are based on applying a link to the cumulative probabilities, and it can be shown that the associated potential is the negative entropy on the simplex in k dimensions. Using the strong convexity of the entropy w.r.t the $\|\cdot\|_1$ norm, we can show that $\zeta_h(\varepsilon) \geq 1/8 \cdot \varepsilon^2/(k-1)^2 \sim \varepsilon^2/(k-1)^2$. In particular, this explains the experiment of Fig. 1 from (Pedregosa et al., 2017), where they provide empirical evidence that the calibration function for AT is larger than the one for CL, which they are not able to compute.

Ranking with NDCG loss. We provide guarantees for learning permutations with the NDCG loss. In this case, we recover the results from Ravikumar et al. (2011).

Graph matching. In graph matching, the goal is to map the nodes from one graph to another. In this case, the outputs are "matchings" (permutations) and the goal is to minimize the Hamming loss between permutations $L(\sigma, \sigma') = \frac{1}{m}\sum_{j=1}^m 1(\sigma(j) \neq \sigma'(j)) = 1 - \langle X_\sigma, X_{\sigma'}\rangle_F/m$, where $X_\sigma \in \mathbb{R}^{m\times m}$ is the permutation matrix associated to the permutation σ. In this case, $k = m^2$, the marginal polytope is the polytope of doubly stochastic matrices which has dimension $r = \dim(\mathcal{M}) = k^2 - 2k + 1$, and the decoding mapping corresponds to perform linear assignment. As CRFs are intractable in this case (Petterson et al., 2009), a common approach is to learn the probabilities of each row independently by casting this problem as k multiclass problems (one for each row of the matrix). Doing multinomial logistic regression independently at each row corresponds to set $\mathcal{D}$ as the polytope of row-stochastic matrices, which strictly includes $\mathcal{M}$ and has dimension $k^2 - k$. A direct application of Theorem 4 .4 gives $\zeta_h(\varepsilon) \geq m^2\varepsilon^2/8$.

Appendices

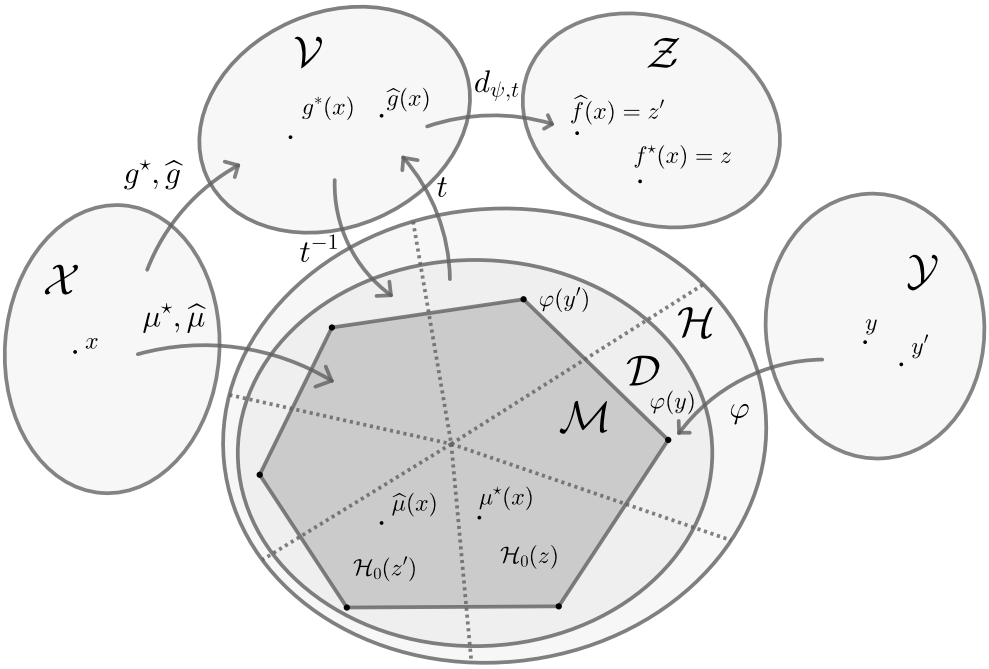


Figure 3.2: Diagram representing the surrogate method $(S, d_{\psi,t})$. The dashed lines separate the calibration sets $\mathcal{H}_0(z)$. If $\widehat{\mu}(x) = t^{-1}(\widehat{g}(x)) \in \mathcal{H}_0(z')$, then $\widehat{f}(x) = d_{\psi,t} \circ \widehat{g}(x) = z'$.

A Bregman Divergence Representation of Surrogate Losses

Proof of Theorem 3 .3. As $t : \mathcal{M} \to \mathcal{V}$ is injective, we have that for any $v \in t(\mathcal{M})$ there exists a unique $\mu \in \mathcal{M}$ such that $t(\mu) = v$. We then consider the loss $\bar{S} : \mathcal{M} \times \mathcal{Y} \to \mathbb{R}$ defined as $\bar{S}(\mu, y) = S(t(\mu), y)$. Moreover, $\bar{S}$ is a continuous function of μ because t and S are continuous. We define the quantities $\bar{s}(\mu, q) = \mathbb{E}_{Y\sim q}\, \bar{S}(\mu, Y)$ and $\delta\bar{s}(\mu, q) = \bar{s}(\mu, q) - \min_{\mu' \in \mathcal{M}} \bar{s}(\mu', q) \geq 0$.

Furthermore, we have that if S satisfies $t(\mu(q)) = \arg\min_{v\in\mathcal{V}}\ s(v, q)$, then $\bar{S}$ satisfies

$$\mu(q) = \operatorname*{arg\,min}_{\mu' \in \mathcal{M}}\ \bar{s}(\mu', q). \tag{3.15}$$

A loss $\bar{S}$ satisfying Eq. (3.15) is said to elicit the function $\mu(\cdot) = \mathbb{E}_{Y\sim\cdot}\, \varphi(Y)$. It is a known result from the theory of property elicitation (Abernethy and Frongillo,

2012; Frongillo and Kash, 2015b) that if a loss elicits a linear function of a distribution, then there exists a strictly convex function h such that,

$$\delta\bar{s}(\mu, q) = D_h(\mu(q), \mu), \quad \forall q \in \mathrm{Prob}(\mathcal{Y}), \forall \mu \in \mathcal{M}.$$

Here, the Bregman divergence of a strictly convex (and potentially non-differentiable) function h is defined as

$$D_h(\mu', \mu) = h(\mu') - h(\mu) - \langle \mu' - \mu, dh_\mu \rangle,$$

where $\{dh_\mu\}_{\mu \in \mathcal{M}}$ are a selection of subgradients of h. Note that $dh_\mu \in T_\mu \mathcal{M}$, where $T_\mu \mathcal{M} = \mathbb{R}^r$ is the tangent space of the marginal polytope $\mathcal{M}$ at the point $\mu \in \mathcal{M}$, where $r = \dim(\mathcal{M})$.

Finally, we prove that h is differentiable in $\mathcal{M}$. Let's first note that as $\bar{S}(\mu, q)$ is a continuous function of μ, then $\delta\bar{s}(\mu, q) = D_h(\mu(q), \mu)$ and consequently $A_{\mu'}(\mu) \coloneqq h(\mu) - \langle \mu - \mu', dh_\mu \rangle$ are also continuous in μ.

Now assume that h is not differentiable at μ_0 and consider $dh^{(1)}_{\mu_0}$ and $dh^{(2)}_{\mu_0}$ two different subgradients of h at the point $\mu_0 \in \mathcal{M}$. In particular, this means that there exists at least a point $\mu_1 \in \mathcal{M}$ such that $\langle \mu_0 - \mu_1, dh^{(1)}_{\mu_0} \rangle \neq \langle \mu_0 - \mu_1, dh^{(2)}_{\mu_0} \rangle$. Assume without loss of generality that $\langle \mu_0 - \mu_1, dh^{(1)}_{\mu_0} \rangle < \langle \mu_0 - \mu_1, dh^{(2)}_{\mu_0} \rangle$ and that $\mu_0 + (\mu_0 - \mu_1) \in \mathcal{M}$. Then, consider the parametrization of the segment $\mathrm{hull}(\{\mu_0 + (\mu_0 - \mu_1), \mu_1\})$ as

$$\mu(t) = t \cdot \mu_1 + (1 - t) \cdot (\mu_0 + (\mu_0 - \mu_1)).$$

We have that

$$\begin{aligned}
&\lim_{t \to (1/2)^-} A_{\mu_1}(\mu(t)) \\
&\leq h(\mu_0) - \langle \mu_0 - \mu_1, dh^{(1)}_{\mu_0} \rangle \\
&< h(\mu_0) - \langle \mu_0 - \mu_1, dh^{(2)}_{\mu_0} \rangle \\
&\leq \lim_{t \to (1/2)^+} A_{\mu_1}(\mu(t)).
\end{aligned}$$

Hence, $A_{\mu_1}(\mu(\cdot))$ is not continuous at $t = 1/2$, which means that $A_{\mu_1}(\cdot)$ is not continuous at μ_0, which is a contradiction. □

B Functions of Legendre-type and Canonical Link

In this section, we first introduce functions of Legendre-type and provide some of the most representative examples, namely, the quadratic function and negative maximum-entropy. We then show how Fenchel duality applied to this group of functions can be used to construct φ-calibrated surrogates by taking the gradient as the link function. In particular, we show that the surrogate loss resulting from this construction, which we refer to as Legendre-type loss function, has desirable properties such as convexity and the fact that the surrogate excess conditional risk can be written as a Bregman divergence directly at the surrogate space $\mathcal{V}$.

First, we recall the concept of *essentially smooth* functions (see Rockafellar (2015) for more details).

Definition B .1 (Essentially Smooth Functions). *A function $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$ is called essentially smooth if*

(1) $\mathcal{D}$ is non-empty,

(2) h is differentiable throughout $\mathrm{int}(\mathcal{D})$,

(3) and $\lim_{u \to \partial\mathcal{D}} \nabla h(u) = +\infty$, where $\partial\mathcal{D}$ is the boundary of the set $\mathcal{D}$.

Definition B .2 (Legendre-type Functions). *A function $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$ is of Legendre-type if it is strictly convex in $\mathrm{int}(\mathcal{D})$ and essentially smooth.*

The two most important examples of such functions are the quadratic loss $h(u) = \frac{1}{2}\|u\|_2^2$ with domain $\mathcal{D} = \mathcal{H} \supsetneq \mathcal{M}$ and the negative maximum-entropy

$$h(u) = -\max_{q \in \mathrm{Prob}(\mathcal{Y})} \mathrm{Ent}(q) \quad \text{s.t} \quad \mu(q) = u,$$

with $\mathcal{D} = \mathcal{M} \subsetneq \mathcal{H}$, where $\mathrm{Ent}(q) = -\sum_{y \in \mathcal{Y}} q(y) \log q(y)$ is the Shannon entropy of the distribution $q \in \mathrm{Prob}(\mathcal{Y})$. Now, recall the concept of *Fenchel conjugate* of a function $h : \mathcal{D} \subseteq \mathcal{H} \to \mathbb{R}$, which is defined as the function h^* computed as

$$h^*(v) = \sup_{u \in \mathcal{D}} \{\langle v, u \rangle - h(u)\},$$

with domain $\mathrm{dom}(h^*)$. The following Proposition B .3 states that Legendre-type functions behave specially well with Fenchel duality.

Proposition B .3 (Fenchel conjugate of a Legendre-type function (Rockafellar, 2015)). *The Fenchel conjugate h^* of a Legendre-type function h is also of Legendre-type, with $\mathrm{int}(\mathrm{dom}(h^*)) = \mathrm{int}(\mathrm{Im}(\nabla h))$. Moreover, the gradient functions $\nabla h : \mathrm{int}(\mathcal{D}) \to \mathrm{int}(\mathrm{dom}(h^*))$ and $\nabla h^* : \mathrm{int}(\mathrm{dom}(h^*)) \to \mathrm{int}(\mathcal{D})$ are inverse of each other $\nabla h^* = (\nabla h)^{-1}$. Furthermore, we also have that*

$$D_h(u, u') = D_{h^*}(\nabla h(u'), \nabla h(u)).$$

The Fenchel conjugate of the quadratic function is the same quadratic function with $\mathrm{dom}(h^*) = \mathcal{D} = \mathcal{H}$, while the Fenchel conjugate for the negative maximum-entropy is the log-sum-exp function $h^*(v) = \log(\sum_{y \in \mathcal{Y}} \exp(\langle \psi(y), v \rangle))$ with domain $\mathrm{dom}(h^*) = \mathbb{R}^r$ where $r = \dim(\mathcal{M})$.

An interesting consequence of Proposition B .3 for our framework is that one has a systematic way of constructing a surrogate method from a function of Legendre-type with domain including the marginal polytope. More specifically, we define the surrogate loss associated to h as the loss with $(h, \nabla h, \varphi)$-BD representation, which takes the following form

$$S(v, y) = D_h(\varphi(y), \nabla h^*(v)) = h^*(v) + h(\varphi(y)) - \langle \varphi(y), v \rangle.$$

Note that $S(v, y)$ is always convex and the excess conditional surrogate risk $\delta s(v, q)$ has the form of a Bregman divergence both in $\mathcal{D}$ and $\mathrm{dom}(h^*)$,

$$\delta s(v, q) = D_h(\mu(q), \nabla h^*(v)) = D_{h^*}(v, \nabla h(\mu(q))),$$

where we have used the last property of Proposition B .3. Moreover, we have that

- $\mathcal{D}$ is bounded if and only if $\mathrm{dom}(h^*)$ is a vector space and h^* is globally Lipschitz.
- If h is $(1/\beta_{\|\cdot\|})$-strongly convex w.r.t the norm $\|\cdot\|$, then $S(\cdot, y)$ is $(\beta_{\|\cdot\|})$-smooth w.r.t the dual norm $\|\cdot\|_*$.

The surrogate loss associated to the quadratic function is the quadratic surrogate and has the form $S(v, y) = \frac{1}{2}\|v - \varphi(y)\|_2^2$ with $\mathcal{V} = \mathcal{H}$, while the surrogate loss associated to the entropy corresponds to conditional random fields (CRFs) and has the form $S(v, y) = \log(\sum_{y' \in \mathcal{Y}} \exp(\langle \psi(y'), v\rangle)) - \langle \psi(y), v\rangle$ with $\mathcal{V} = \mathbb{R}^r$. Both surrogates are studied in detail in Section 5 .1.

C Calibration of Risks

In this section we study the implications of the calibration function ζ for relating both excess risks. In particular, we first prove in Section C .1 that a convex lower bound $\bar{\zeta}$ of ζ satisfies $\bar{\zeta}(\mathcal{E}(d \circ \widehat{g}) - \mathcal{E}(f^\star)) \leq \mathcal{R}(\widehat{g}) - \mathcal{R}(g^\star)$, which corresponds to Theorem B .4. Then, in Section C .2 we improve the calibration between risks by imposing a low noise assumption at the decision boundary.

C .1 Calibration of Risks without Noise Assumption

Proof of Theorem B .4. Note that by the definition of the calibration function, we have that

$$\zeta(\delta\ell(d \circ \widehat{g}(x), \rho(\cdot|x))) \leq \delta s(\widehat{g}(x), \rho(\cdot|x)). \tag{3.16}$$

The comparison between risks is then a consequence of Jensen's inequality:

$$\begin{aligned}
\bar{\zeta}(\mathcal{E}(d \circ \widehat{g}) - \mathcal{E}(f^\star)) &= \bar{\zeta}(\mathbb{E}_{X \sim \rho_\mathcal{X}} \delta\ell(d \circ \widehat{g}(X), \rho(\cdot|X))) \\
&\leq \mathbb{E}_{X \sim \rho_\mathcal{X}} \bar{\zeta}(\delta\ell(d \circ \widehat{g}(X), \rho(\cdot|X))) && \text{(Jensen ineq.)} \\
&\leq \mathbb{E}_{X \sim \rho_\mathcal{X}} \zeta(\delta\ell(d \circ \widehat{g}(X), \rho(\cdot|X))) && (\bar{\zeta} \leq \zeta) \\
&\leq \mathbb{E}_{X \sim \rho_\mathcal{X}} \delta s(\widehat{g}(X), \rho(\cdot|X)) && (Eq.\ (3.16)) \\
&= \mathcal{R}(\widehat{g}) - \mathcal{R}(g^\star).
\end{aligned}$$

□

C .2 Calibration of Risks with Low Noise Assumption

In this section we prove Theorem 4 .6 by imposing assumptions on the the behavior of the margin function $\gamma(x) = \min_{z' \neq f^\star(x)} \delta\ell(z', \rho(\cdot|x))$ under the marginal distribution of the data $\rho_\mathcal{X} \in \mathrm{Prob}(\mathcal{X})$. Note that the p-noise condition $\rho_\mathcal{X}(\gamma(X) \leq \varepsilon) = o(\varepsilon^p)$ is a generalization of the Tsybakov condition for binary classification (Tsybakov, 2004) and of the condition by Mroueh et al. (2012) for multiclass classification to general discrete losses. Indeed, for the binary 0-1 loss ($\mathcal{Y} = \{-1, 1\}$), $\gamma(x) = |2\eta(x) - 1|$ with $\eta(x) = \rho(Y = 1|x)$, so we recover the classical Tsybakov condition.

We first prove Lemma C .1, which states the equivalence between the p-noise condition $\rho_\mathcal{X}(\gamma(X) \leq \varepsilon) = o(\varepsilon^p)$ and $1/\gamma \in L_p(\rho_\mathcal{X})$.

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

Proof.

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

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

Let's now define the error set as $X_f = \{x \in \mathcal{X} \mid f(x) \neq f^\star(x)\}$. The following Lemma C .2, which bounds the probability of error by a power of the excess risk, is a generalization of the Tsybakov Lemma (Tsybakov, 2004, Prop.1) for general discrete losses.

Lemma C .2 (Bounding the size of the error set)**.** *If $1/\gamma \in L_p(\rho_{\mathcal{X}})$, then*

$$\rho_{\mathcal{X}}(X_f) \leq \gamma_p^{\frac{1}{p+1}}(\mathcal{E}(f) - \mathcal{E}(f^\star))^{\frac{p}{p+1}}.$$

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

$$1(f(x) \neq f^\star(x)) \leq 1/\gamma(x)\delta\ell(f(x), \rho(\cdot|x)).$$

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

We will need the following useful Lemma C .3 of convex functions.

Lemma C .3 (Property of convex functions)**.** *Suppose $\bar{\zeta} : \mathbb{R} \to \mathbb{R}$ is convex and $\bar{\zeta}(0) = 0$. Then, for all $y' > 0$, $0 \leq y \leq y'$,*

$$\bar{\zeta}(y) \leq \frac{y}{y'}\bar{\zeta}(y') \quad \textit{and} \quad \bar{\zeta}(y')/y' \textit{ is increasing on } (0, \infty).$$

Proof. Take $\alpha = \frac{y}{y'} \leq 1$. The result follows directly by definition of convexity, as

$$\bar{\zeta}(y) = \bar{\zeta}((1-\alpha)0 + \alpha y') \leq (1-\alpha)\bar{\zeta}(0) + \alpha\bar{\zeta}(y') = \frac{y}{y'}\bar{\zeta}(y').$$

For the second part, re-arrange the terms in the above inequality. □

We now have the tools to prove Theorem 4 .6, which is an adaptation of the proof of Thm. 10 of Bartlett et al. (2006) which was specific to binary 0-1 loss.

Proof of part 1 Theorem 4 .6. The intuition of the proof is to split the conditional excess risk into a part with low noise $\delta\ell(f(x), \rho(\cdot|x)) \leq t$ and a part with high noise $\delta\ell(f(x), \rho(\cdot|x)) \geq t$. The first part will be controlled by the p-noise assumption and the second part by the convex lower bound of the calibration function $\bar{\zeta}$.

$$\begin{aligned}
\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) &= \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\, \delta\ell(f(X), \rho(\cdot|X)) \\
&= \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\, \{1(X_f) \cdot \delta\ell(f(X), \rho(\cdot|X))\} \\
&= \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\, \{\delta\ell(f(X), \rho(\cdot|X)) \cdot 1(X_f \cap \{\delta\ell(f(X), \rho(\cdot|X)) \leq t\}\} \\
&+ \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\, \{\delta\ell(f(X), \rho(\cdot|X)) \cdot 1(X_f \cap \{\delta\ell(f(X), \rho(\cdot|X)) \geq t\}\} \\
&= A + B.
\end{aligned}$$

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

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

  where in the last inequality we have used Lemma C .2.

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

$$\delta\ell(f(x), \rho(\cdot|x)) \cdot 1(\delta\ell(f(x), \rho(\cdot|x)) \geq t) \leq \frac{t}{\bar{\zeta}(t)} \bar{\zeta}(\delta\ell(f(x), \rho(\cdot|x))). \tag{3.17}$$

  In the case $\delta\ell(f(x), \rho(\cdot|x)) < t$, inequality in Eq. (3.17) follows from the fact that $\bar{\zeta}$ is nonnegative. For the case $\delta\ell(f(x), \rho(\cdot|x)) > t$, apply Lemma C .3 with $y' = \delta\ell(f(x), \rho(\cdot|x))$ and $y = t$.
  From Eq. (4.18), we have that $\mathbb{E}_{X\sim\rho_{\mathcal{X}}}\{1(X_f) \cdot \bar{\zeta}(\delta\ell(f(X), \rho(\cdot|X)))\} \leq \mathcal{R}(g) - \mathcal{R}(g^*)$. Hence,

$$B \leq \frac{t}{\bar{\zeta}(t)}(\mathcal{R}(g) - \mathcal{R}(g^*)).$$

  Putting everything together,

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq t\gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}} + \frac{t}{\bar{\zeta}(t)}(\mathcal{R}(g) - \mathcal{R}(g^*)),$$

and hence,

$$\left(\frac{\mathcal{E}(d \circ g) - \mathcal{E}(f^*)}{t} - \gamma_p^{\frac{1}{p+1}} \left(\mathcal{E}(d \circ g) - \mathcal{E}(f^*)\right)^{\frac{p}{p+1}}\right) \bar{\zeta}(t) \leq \mathcal{R}(g) - \mathcal{R}(g^*).$$

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. (3.12).

The fact that $\bar{\zeta}^{(p)}$ never provides a worse rate than $\bar{\zeta}$ is because we have

$$\varepsilon^{\frac{p}{p+1}} \cdot \bar{\zeta}((\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}/2) \geq \bar{\zeta}(\varepsilon \cdot (\gamma_p^{-1})^{\frac{1}{p+1}}/2). \tag{3.18}$$

To see this, re-arrange the terms in Eq. (3.18) to,

$$\frac{\bar{\zeta}((\gamma_p^{-1}\varepsilon)^{\frac{1}{p+1}}/2)}{\varepsilon^{\frac{1}{p+1}}} \geq \frac{\bar{\zeta}(\varepsilon \cdot (\gamma_p^{-1})^{\frac{1}{p+1}}/2)}{\varepsilon}. \tag{3.19}$$

Then, Eq. (3.19) follows from the fact that $\frac{\bar{\zeta}(t)}{t}$ is non-decreasing by Lemma C .3. □

Proof of part 2 of Theorem 4 .6. If $\delta s(g^\star(x), \rho(\cdot|x)) \leq \zeta(\delta)$ $\rho_{\mathcal{X}}$-a.s implies by the definition of the calibration function that $\delta\ell(\widehat{f}(x), \rho(\cdot|x)) < \delta$ $\rho_{\mathcal{X}}$-a.s.

As $\gamma(x) = \min_{z' \neq f^*(x)} \delta\ell(z', \rho(\cdot|x)) \geq \delta > 0$ $\rho_{\mathcal{X}}$-a.s., then we necessarily have $\widehat{f}(x) = f^\star(x)$ $\rho_{\mathcal{X}}$-a.s., which implies that $\mathcal{R}(\widehat{f}) = \mathcal{R}(f^\star)$. □

D Calibration Function for φ-Calibrated Losses

In this section we study the calibration function for φ-calibrated losses. In Section D .1, we compute exact expressions for ζ, in Section D .2 we provide lower bounds and in Section D .3 we prove the existence of an upper bound.

D .1 Exact Formula for the Calibration Function

This section contains three results. The first is Lemma D .1, which re-writes the calibration function from its definition by leveraging the BD representation of the surrogate loss. In particular, it shows that ζ only depends on S through the potential h, and it can be written as a (constrained) minimization problem where the Bregman divergence associated to h is minimized. In the following we will denote

$$z(u) = \arg\min_{z' \in \mathcal{Z}} \ \langle \psi(z'), u \rangle.$$

Lemma D .1. *The calibration function for a φ-calibrated surrogate can be written as*

$$\zeta_h(\varepsilon) = \inf_{u' \in \mathcal{M}, u \in \mathcal{D}} D_h(u', u) \quad s.t \quad \langle \psi(z(u)) - \psi(z(u')), u' \rangle \geq \varepsilon. \tag{3.20}$$

Proof. As S is φ-calibrated, we can write

$$\delta s(v, q) = D_h(\mu(q), t^{-1}(v)), \quad \forall v \in \mathcal{V}, \ \forall q \in \mathrm{Prob}(\mathcal{Y}).$$

Using the affine decomposition of the loss L, we can write

$$\delta \ell(z, q) = \langle \psi(z) - \psi(z(\mu(q))), \mu(q) \rangle, \quad \forall z \in \mathcal{Z}, \ \forall q \in \mathrm{Prob}(\mathcal{Y}).$$

Hence, the constrained minimization problem only depends on q through $\mu(q)$, so the minimization over $\mathrm{Prob}(\mathcal{Y})$ can be done over $\mathcal{M}$. Moreover, we can write $\delta \ell(d(v), q) = \delta \ell(z(t^{-1}(v)), q)$ and $t^{-1}(\mathcal{V}) = \mathcal{D}$. Hence, applying the inverse of the link to the problem one obtains (3.20). □

The second result is Theorem 4 .3, which uses the result of Lemma D .1 to view the problem (3.20) as the minimum over z of the Bregman divergence between the sets $\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}$ and $\mathcal{H}_0(z) \cap \mathcal{D}$.

Proof of Theorem 4 .3. Use the fact that $\mathcal{H} = \bigcup_{z \in \mathcal{Z}} \mathcal{H}_0(z)$ with $z = z(u) \iff u \in \mathcal{H}_0(z)$ to re-write the calibration function as

$$\zeta_h(\varepsilon) = \min_{z \in \mathcal{Z}} \inf_{\substack{u' \in \mathcal{M} \\ u \in \mathcal{H}_0(z) \cap \mathcal{D}}} D_h(u', u) \quad \text{s.t} \quad \langle \psi(z) - \psi(z(u')), u' \rangle \geq \varepsilon.$$

Now, the minimization over the first coordinate is made on the set (now independent of u)

$$\{u' \in \mathcal{H} \mid \langle \psi(z) - \psi(z(u')), u' \rangle \geq \varepsilon\} \cap \mathcal{M} = \{u' \in \mathcal{H} \mid \langle \psi(z) - \psi(z(u')), u' \rangle \leq \varepsilon\}^c \cap \mathcal{M}.$$

Now let's show that

$$\{u' \in \mathcal{H} \mid \langle \psi(z) - \psi(z(u')), u' \rangle \leq \varepsilon\} = \mathcal{H}_\varepsilon(z),$$

where $\mathcal{H}_\varepsilon(z) = \{u' \in \mathcal{H} \mid \langle \psi(z) - \psi(z'), u' \rangle \leq \varepsilon, \forall z' \in \mathcal{Z}\}$. Note that the inclusion $(\supset)$ is trivial. For $(\subset)$, note that for any $z' \in \mathcal{Z}$, we have that

$$\langle \psi(z'), u' \rangle + \varepsilon \geq \langle \psi(z(u')), u' \rangle + \varepsilon \geq \langle \psi(z), u' \rangle.$$

Hence, we obtain the final result,

$$\zeta_h(\varepsilon) = \min_{z \in \mathcal{Z}} \inf_{\substack{u' \in \mathcal{H}_\varepsilon(z)^c \cap \mathcal{M} \\ u \in \mathcal{H}_0(z) \cap \mathcal{D}}} D_h(u', u) = \min_{z \in \mathcal{Z}} D_h(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z) \cap \mathcal{D}).$$

□

Finally, the following Proposition D .2 provides an exact formula for the Euclidean distance between the sets $\mathcal{H}_\varepsilon(z)^c$ and $\mathcal{H}_0(z)$. This result will be useful to derive an improved lower bound on ζ_h using strong convexity of the potential w.r.t the Euclidean norm and an exact expression in the case of the quadratic surrogate when the marginal polytope $\mathcal{M}$ is full-dimensional. In the following, we denote by $d_2(A, B) = \inf_{u' \in A, u \in B} \|u' - u\|_2$ the Euclidean distance between the sets A and B.

Proposition D .2. *We have that*

$$d_2(\mathcal{H}_\varepsilon(z)^c, \mathcal{H}_0(z)) = \min_{z' \neq z} \frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2} + \delta_{z,z'}, \tag{3.21}$$

where $\delta_{z,z'} = d_2(\mathcal{H}_0(z), \{\langle \psi(z) - \psi(z'), u \rangle = 0\}) > 0$ *if and only if* $\mathcal{H}_0(z) \cap \mathcal{H}_0(z') = \varnothing$.

Proof. Write $\mathcal{H}_\varepsilon(z)^c = \bigcup_{z' \neq z} \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}$. Then, we have that

$$d_2(\mathcal{H}_\varepsilon(z)^c, \mathcal{H}_0(z)) = \min_{z' \neq z} d_2(\mathcal{H}_0(z), \{u \in \mathcal{H} \mid \langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}),$$

where we have used that $d_2(\cup_{i=1}^k A_i, B) = \min_{i=1,\dots,k} d_2(A_i, B)$. Recall that the distance between two convex bodies is characterized by the minimum distance between two parallel supporting hyperplanes. Let's split the analysis in two cases:

- $\mathcal{H}_0(z) \cap \mathcal{H}_0(z') \neq \varnothing$. In this case $\{\langle \psi(z) - \psi(z'), u \rangle = 0\}$ is a supporting hyperplane of $\mathcal{H}_0(z)$. Hence, the distance betwen $\mathcal{H}_0(z)$ and $\{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}$ is equal to the distance between $\{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}$ and $\{\langle \psi(z) - \psi(z'), u \rangle \leq 0\}$, which is equal to $\frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2}$.
- $\mathcal{H}_0(z) \cap \mathcal{H}_0(z') = \varnothing$. In this case $\{\langle \psi(z) - \psi(z'), u \rangle = 0\}$ is not a supporting hyperplane of $\mathcal{H}_0(z)$. The supporting hyperplane parallel to $\{\langle \psi(z) - \psi(z'), u \rangle = \varepsilon\}$ has the form $\{\langle \psi(z) - \psi(z'), u \rangle = -\varepsilon'\}$ for $\varepsilon' > 0$. The distance between both hyperplanes is then $\frac{\varepsilon + \varepsilon'}{\|\psi(z) - \psi(z')\|_2} = \frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2} + \delta_{z,z'}$, where $\delta_{z,z'} = d_2(\mathcal{H}_0(z), \{\langle \psi(z) - \psi(z'), u \rangle = 0\})$.

Hence, we obtain the final result. □

Note that many losses satisfy $\mathcal{H}_0(z) \cap \mathcal{H}_0(z') \neq \varnothing$ for all $z, z' \in \mathcal{Z}$, such as the 0-1, Hamming and absolute error in ordinal regression. In this case, Eq. (3.21) is simplified to $\frac{\varepsilon}{\max_{z' \neq z} \|\psi(z) - \psi(z')\|_2}$.

D .2 Lower Bounds on the Calibration Function

In this section, we provide two lower bounds on the calibration function. The first one is given by Theorem 4 .4 and allows to exploit strong convexity of the potential w.r.t an arbitrary norm in $\mathcal{H}$. The second one is given by Theorem D .4 and provides a tighter bound but it is special to strong convexity w.r.t the Euclidean norm. The proof of Theorem 4 .4 relies on the following Lemma D .3.

Lemma D .3 (Generic bound on the conditional excess risk)**.** *Let* $\|\cdot\|$ *be a norm in* $\mathbb{R}^k$ *and denote by* $\|\cdot\|_*$ *its dual norm. We have that,*

$$\delta\ell(z(u), q) \leq 2\, c_{\psi, \|\cdot\|_*}\, \|u - \mu(q)\|,$$

where $c_{\psi, \|\cdot\|_*} = \sup_{z \in \mathcal{Z}} \|\psi(z)\|_*$.

Proof. Decompose the excess conditional risk $\delta\ell(z(u), q)$ into two terms A and B:

$$\begin{aligned}\delta\ell(z(u), q) &= \langle \psi(z(u)), \mu(q) - u\rangle \\ &\quad + \langle \psi(z(u)), u\rangle - \langle \psi(z(\mu(q))), \mu(q)\rangle \\ &= A + B.\end{aligned}$$

For the first term, we directly have $A \leq \sup_{z\in\mathcal{Z}} |\langle \psi(z), \mu(q) - u\rangle|$. For the second term, we use the fact that for any given two functions $\eta, \eta' : \mathcal{Z} \to \mathbb{R}$, it holds that $|\min_z \eta(z) - \min_{z'} \eta'(z')| \leq \sup_z |\eta(z) - \eta'(z)|$. As $z(\mu(q))$ minimizes $\langle \psi(z), \mu(q)\rangle$ and $z(u)$ minimizes $\langle \psi(z), u\rangle$, we can conclude also that $B \leq \sup_{z\in\mathcal{Z}} |\langle \psi(z), \mu(q) - u\rangle|$. Hence, we obtain

$$\delta(z(u), q) \leq 2 \sup_{z\in\mathcal{Z}} |\langle \psi(z), \mu(q) - u\rangle| \leq 2\, c_{\varphi,\|\cdot\|_*}\, \|u - \mu(q)\|,$$

where at the last step we have used Cauchy-Schwarz inequality. □

We now proceed to the proof of Theorem 4 .4.

Proof of Theorem 4 .4. Starting from Lemma D .1, we see that the constrains are $\langle \psi(z(u)) - \psi(z(u')), u'\rangle \geq \varepsilon$. From the definition of Bregman divergence and strong convexity we have that if h is $(1/\beta_{\|\cdot\|})$-strongly convex in $\mathcal{D}$, then

$$D_h(u', u) \geq \frac{1}{2\beta_{\|\cdot\|}} \|u' - u\|^2, \quad \forall u, u' \in \mathcal{D}.$$

From Lemma D .3 we have that $\langle \psi(z(u)) - \psi(z(u')), u'\rangle \leq 2c_{\psi,\|\cdot\|_*} \|u' - u\|$. Putting all these things together, we obtain,

$$D_h(u', u) \geq \frac{1}{2\beta_{\|\cdot\|}} \|u' - u\|^2 \geq \frac{\langle \psi(z(u)) - \psi(z(u')), u'\rangle^2}{8c^2_{\psi,\|\cdot\|_*}\beta_{\|\cdot\|}} \geq \frac{\varepsilon^2}{8c^2_{\psi,\|\cdot\|_*}\beta_{\|\cdot\|}}.$$

□

Finally, we present Theorem D .4, which is based on Proposition D .2 and provides a tighter lower bound under strong convexity w.r.t the Euclidean distance.

Theorem D .4 (Improved lower bound for L_2-strong convexity)**.** *If h is $(1/\beta_{\|\cdot\|_2})$-strongly convex w.r.t the L_2 norm $\|\cdot\|_2$, then:*

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{2\beta_{\|\cdot\|_2} \max_{z'\neq z} \|\psi(z) - \psi(z')\|_2^2}. \tag{3.22}$$

Proof.

$$\begin{aligned}\zeta_h(\varepsilon) &\geq \frac{1}{2\beta_{\|\cdot\|_2}} \min_{z\in\mathcal{Z}} d_2(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z) \cap \mathcal{D})^2 && (D_h(u', u) \geq \frac{1}{2\beta_{\|\cdot\|_2}} \|u' - u\|_2^2) \\ &\geq \frac{1}{2\beta_{\|\cdot\|_2}} \min_{z\in\mathcal{Z}} d_2(\mathcal{H}_\varepsilon(z)^c, \mathcal{H}_0(z))^2 && \text{(minimization on larger domain)} \\ &= \frac{1}{2\beta_{\|\cdot\|_2}} \min_{z\in\mathcal{Z}} \left(\min_{z'\neq z} \frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2} + \delta_{z,z'}\right)^2 && \text{(Proposition D .2)}\end{aligned}$$

$$\geq \frac{1}{2\beta_{\|\cdot\|_2}} \min_{z' \neq z} \frac{\varepsilon^2}{\|\psi(z) - \psi(z')\|_2^2} \qquad (\delta_{z,z'} \geq 0)$$
$$= \frac{\varepsilon^2}{2\beta_{\|\cdot\|_2} \max_{z' \neq z} \|\psi(z) - \psi(z')\|_2^2}.$$

□

Note that the lower bound (3.22) is tighter than the one given by Theorem 4 .4:

$$\frac{\varepsilon^2}{2\beta_{\|\cdot\|_2} \max_{z' \neq z} \|\psi(z) - \psi(z')\|_2^2} \geq \frac{\varepsilon^2}{8\beta_{\|\cdot\|_2} \max_{z \in \mathcal{Z}} \|\psi(z)\|_2^2},$$

using that $\|\psi(z) - \psi(z')\|_2^2 \leq \|\psi(z)\|_2^2 + \|\psi(z')\|_2^2 \leq 2 \cdot \max_{z \in \mathcal{Z}} \|\psi(z)\|_2^2$.

D .3 Upper bound on the Calibration Function

In this section we prove the result of existence of a quadratic upper bound on ζ_h. The idea of the proof is to show that there exists a point $u_0 \in \mathcal{H}_0 \cap \text{relint}(\mathcal{M})$ and a continuous path $(u_\varepsilon)_{\varepsilon \leq \varepsilon_0}$ such that $u_\varepsilon \in \mathcal{H}_\varepsilon^c \cap \text{relint}(\mathcal{M})$ with $\|u_\varepsilon - u_0\| \lesssim \varepsilon$ for all $\varepsilon \leq \varepsilon_0$. Then, the norm of the Hessian of h can be uniformly bounded in this compact continuous path and the result follows. It is important to take this sequence at the relative interior of the marginal polytope because $\|\nabla^2 h\|_2$ could explode at the boundary. Note that if $\mathcal{M}$ is full dimensional, the result follows easily from Proposition D .2. We begin by constructing this path as a segment in Lemma D .5.

Lemma D .5. *There exists $z \in \mathcal{Z}$ and a closed segment $I = \text{hull}(\{u_{\varepsilon_0}, u_0\}) \subset \text{relint}(\mathcal{M})$ with $u_0 \in \mathcal{H}_0(z)$ such that the point $u_\varepsilon = u_0 \cdot (1 - \varepsilon/\varepsilon_0) + u_{\varepsilon_0} \cdot (\varepsilon/\varepsilon_0) \in I$ satisfies for a constant $C \in \mathbb{R}$:*

$$u_\varepsilon \in \mathcal{H}_\varepsilon(z)^c \qquad \textit{and} \qquad \|u_\varepsilon - u_0\|_2 \leq C \cdot \varepsilon, \quad \forall \varepsilon \leq \varepsilon_0.$$

Proof. We will first assume $\mathcal{M}$ is *non full-dimensional*. Hence, it lies in an affine subspace of $\mathcal{H}$. Take $u_0 \in \partial\mathcal{H}_0(z) \cap \text{relint}(\mathcal{M})$ and take $z' \in \mathcal{Z}$ corresponding to a supporting hyperplane of $\mathcal{H}_0(z)$ at u_0, i.e, $\langle \psi(z) - \psi(z'), u_0 \rangle = 0$.

Using that $\mathcal{H}_\varepsilon(z)^c = \bigcup_{z' \neq z} \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}$, we have that

$$\begin{aligned} &d_2(u_0, \mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}) \\ &= \min_{z'' \neq z} d_2(u_0, \mathcal{M} \cap \{\langle \psi(z) - \psi(z''), u \rangle \geq \varepsilon\}) \\ &\leq d_2(u_0, \mathcal{M} \cap \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}). \end{aligned}$$

Now, consider a convex neighborhood $U \subset \text{relint}(\mathcal{M})$ of u_0. We have that for ε_0 small enough, the distance in **??** is achieved at U at a point $u_{\varepsilon_0} \in U \subset \text{relint}(\mathcal{M})$. Moreover, we have that

$$u_\varepsilon = u_0(1 - \varepsilon/\varepsilon_0) + u_{\varepsilon_0}(\varepsilon/\varepsilon_0) = \underset{u \in U \cap \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}}{\arg\min} \|u_0 - u\|_2, \quad \forall \varepsilon \leq \varepsilon_0,$$

and

$$\|u_\varepsilon - u_0\|_2$$

$$
\begin{aligned}
&= d_2(u_0, U \cap \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}) \\
&= L \cdot d_2(u_0, \{\langle \psi(z) - \psi(z'), u \rangle \geq \varepsilon\}) \\
&= L \cdot \frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2} = C \cdot \varepsilon.
\end{aligned}
$$

For the *full-dimensional* case, the proof follows the same with $L = 1$. □

Proof of Theorem 4 .5. We will show that for a sufficiently small ε_0, there exists $C' \in \mathbb{R}$ such that $\zeta_h(\varepsilon) \leq C' \cdot \varepsilon^2$ for all $\varepsilon \leq \varepsilon_0$. First use Lemma D .5 and define (using that h is twice differentiable) $C_I = \sup_{u' \in I} \|\nabla^2 h(u')\|_2 < +\infty$ which is finite because $I = \bar{I} \subset \operatorname{int}(\mathcal{M})$. Then, for all $\varepsilon \leq \varepsilon_0$, the proof follows as:

$$
\begin{aligned}
\zeta_h(\varepsilon) &= \min_{z' \in \mathcal{Z}} D_h(\mathcal{H}_\varepsilon(z')^c \cap \mathcal{M}, \mathcal{H}_0(z') \cap \mathcal{D}) \\
&\leq \min_{z' \in \mathcal{Z}} D_h(\mathcal{H}_\varepsilon(z')^c \cap \mathcal{M}, \mathcal{H}_0(z') \cap \mathcal{M}) \\
&\leq D_h(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z) \cap \mathcal{M}) \\
&\leq D_h(u_\varepsilon, u_0) \\
&\leq C_I \cdot \|u_\varepsilon - u_0\|_2^2 \\
&\leq C_I \cdot C \cdot \varepsilon^2 = C' \cdot \varepsilon^2.
\end{aligned}
$$

□

E Generic Methods for Structured Prediction

In this section we present results on two generic methods for structured prediction: the quadratic surrogate in Section E .1 and conditional random fields (CRFs) in Section E .2.

E .1 Quadratic Surrogate

We first provide an exact formula for the calibration function of the quadratic surrogate when the marginal polytope $\mathcal{M}$ is full-dimensional. Note that in this case, one can directly apply Proposition D .2 if one makes sure that the distances are achieved inside $\mathcal{M}$.

Theorem E .1 (Exact Calibration for Quadratic Surrogate)**.** *Let $h : \mathcal{D} = \mathcal{H} \to \mathbb{R}$ be $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ corresponding to the quadratic surrogate. If $\mathcal{M}$ is full-dimensional, then*

$$
\zeta_h(\varepsilon) = \frac{1}{2}\left(\min_{z' \neq z} \frac{\varepsilon}{\|\psi(z) - \psi(z')\|_2} + \delta_{z,z'}\right)^2, \quad \forall \varepsilon \leq \min_{z \neq z'} \|L_z - L_{z'}\|_\infty, \tag{3.23}
$$

where $\delta_{z,z'} = d_2(\mathcal{H}_0(z), \{\langle \psi(z) - \psi(z'), u \rangle = 0\}) > 0$ if and only if $\mathcal{H}_0(z) \cap \mathcal{H}_0(z') = \varnothing$, and L_z is the z-th row of the loss matrix $L \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}}$.

Proof. Note that the calibration for the quadratic surrogate is

$$
\zeta_h(\varepsilon) = \frac{1}{2} \min_{z' \neq z} d_2(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z))^2.
$$

Hence, the goal of the proof is to show that $d_2(\mathcal{H}_\varepsilon(z)^c, \mathcal{H}_0(z)) = d_2(\mathcal{H}_\varepsilon(z)^c \cap \mathcal{M}, \mathcal{H}_0(z))$ if $\varepsilon \leq \min_{z \neq z'} \|L_z - L_{z'}\|_\infty$ and then use Proposition D .2.

Remember from the proof of Proposition D .2 that the term $\frac{\varepsilon}{\|\psi(z)-\psi(z')\|_2}$ is the distance between the half-spaces $\{\langle\psi(z)-\psi(z'),u\rangle\leq 0\}$ and $\{\langle\psi(z)-\psi(z'),u\rangle\geq\varepsilon\}$. This distance is achieved inside of the marginal polytope if $\varepsilon\leq\sup_{\mu\in\mathcal{M}}|\{\langle\psi(z)-\psi(z'),\mu\rangle|$. Hence, we have that $d_2(\mathcal{H}_\varepsilon(z)^c\cap\mathcal{M},\mathcal{H}_0(z))=d_2(\mathcal{H}_\varepsilon(z)^c,\mathcal{H}_0(z))$ if

$$\begin{aligned}\varepsilon &\leq \min_{z'\neq z}\sup_{\mu\in\mathcal{M}}|\{\langle\psi(z)-\psi(z'),\mu\rangle|\\ &= \min_{z'\neq z}\sup_{q\in\mathrm{Prob}(\mathcal{Y})}|\ell(z,q)-\ell(z',q)|\\ &= \min_{z'\neq z}\sup_{q\in\mathrm{Prob}(\mathcal{Y})}|(L_z-L_{z'})\cdot q|\\ &= \min_{z\neq z'}\|L_z-L_{z'}\|_\infty.\end{aligned}$$

□

Now we prove Theorem E .1, which states that if one takes ε small enough, then the expression can be simplified by removing the $\delta_{z,z'}$'s from Eq. (3.23).

Proof of Theorem 5 .1. Take a 3-tuple (z,z',z'') such that $\mathcal{H}_0(z)\cap\mathcal{H}_0(z')\neq\varnothing$ and $\mathcal{H}_0(z')\cap\mathcal{H}_0(z'')=\varnothing$. Then, it is clear that there exists $\varepsilon_0'>0$ such that for all $\varepsilon\leq\varepsilon_0'$:

$$\frac{\varepsilon}{\|\psi(z)-\psi(z')\|_2}\leq\frac{\varepsilon}{\|\psi(z')-\psi(z'')\|_2}+\delta_{z',z''},$$

because $\delta_{z',z''}>0$ as $\mathcal{H}_0(z')\cap\mathcal{H}_0(z'')=\varnothing$. Taking ε_0 as the minimum of the ε_0''s over all 3-tuples of this type gives the desired result. □

Kernel ridge regression as an estimator independent of the affine decomposition of L. It was shown by Ciliberto et al. (2016) that if one minimizes the expected risk of the quadratic surrogate using kernel ridge regression, then one can construct an estimator independent of the affine decomposition of the loss. Indeed, given n data points $\{(x_i,y_i)\}_{i\leq n}$ and a kernel $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ with corresponding RKHS $\mathcal{G}$, the kernel ridge regression estimator $\widehat{g}_n\in\mathcal{G}\otimes\mathcal{H}$ of $g^\star$ can be written in closed form as $\widehat{g}_n(\cdot)=\sum_{i=1}^n\alpha_i(\cdot)\psi(y_i)$ where $\alpha(x)=(\alpha_1(x),\ldots,\alpha_n(x))\in\mathbb{R}^n$ is defined by $\alpha(x)=(K+n\lambda I)^{-1}K_x$ with $K\in\mathbb{R}^{n\times n}$ is defined by $K_{ij}=k(x_i,x_j)$ and $K_x=(k(x,x_1),\ldots,k(x,x_n))\in\mathbb{R}^n$. Note that $\widehat{g}_n$ is linear in the embeddings $\varphi(y_i)_{i\leq n}$ and the link function is the identity, so the estimator is independent of the choice of the embedding of the loss because

$$\widehat{f}(x)=\operatorname*{arg\,min}_{z\in\mathcal{Z}}\langle\psi(z),t^{-1}(\widehat{g}(x)\rangle=\operatorname*{arg\,min}_{z\in\mathcal{Z}}\sum_{i=1}^n\alpha_i(x)L(z,y_i).$$

In the following, we compare our calibration results on the quadratic surrogate with the work by Osokin et al. (2017).

Comparison with related work on the quadratic surrogate for structured prediction. In the work by Osokin et al. (2017), they study the calibration properties of a quadratic-type surrogate, which is constructed differently than ours. In order to understand their construction under our framework, let's consider the following decomposition of the loss function, $L(z,y)=\langle\psi(z),\psi(y)\rangle=\langle e_z,(L\cdot e_y)\rangle$, where L is the loss matrix. Note

that in this case the quadratic surrogate associated to this decomposition is $S(v, y) = \frac{1}{2}\|v - L_{:,y}\|_2^2$ with decoding $d(v) = \arg\min_{z\in\mathcal{Z}} v_z$. As $v \in \mathbb{R}^{\mathcal{Y}}$ can be an exponentially large vector, they consider the parametrization $v = F \cdot w$, where $F \in \mathbb{R}^{\mathcal{Y}\times k}$ is a score matrix and $w \in \mathbb{R}^k$, where k can be potentially much smaller than $\mathcal{Y}$. The surrogate loss they consider is

$$S(w, y) = \frac{1}{2}\|F \cdot w - L_{:,y}\|_2^2, \tag{3.24}$$

where $L_{:,y}$ is the y-th column of L. In their work they normalize the surrogate loss by $|\mathcal{Y}|$, but we remove this factor in order to properly compare calibration functions[5]. It is important to note that this loss does not fall into our framework for F different than the identity. They provide the following lower bound on the calibration function

$$\zeta(\varepsilon) \geq \frac{\varepsilon^2}{2\max_{z'\neq z}\|P_F(e_z - e_{z'})\|_2^2}, \tag{3.25}$$

where $P_F = F(F^TF)^\dagger F^T$ is the orthogonal projection to the subspace generated by the columns of F. In order to compare with our work, we follow (Nowak et al., 2019) and consider a decomposition $L = F\cdot U^T$ with $F \in \mathbb{R}^{\mathcal{Z}\times k}, U \in \mathbb{R}^{\mathcal{Y}\times k}$ and $S(v, y) = \frac{1}{2}\|v - U_y\|_2^2$ with decoding $d(v) = \arg\min_{z\in\mathcal{Z}} F_z \cdot v$, where $F_z = \psi(z) = F^T \cdot e_z$ and $U_y = \varphi(y) = U^T \cdot e_y$. For this surrogate method, Theorem D .4 provides the following lower bound:

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{2\max_{z'\neq z}\|F^T(e_z - e_{z'})\|_2^2}. \tag{3.26}$$

Note the similarity between expressions (3.26) and (3.25). In particular, if $F \in \mathbb{R}^{\mathcal{Y}\times\mathcal{Y}}$ is the identity, (so that surrogate (3.24) enters our framework), both expressions are equal. For other F's, both calibration functions are not comparable since their surrogate is larger than ours. For instance, if one takes $F \in \mathbb{R}^{\mathcal{Z}\times k}$ with the smallest k such that $L = F \cdot U^T$ for the Hamming loss, their calibration function is proportional to $\frac{|\mathcal{Y}|}{k}$ (Osokin et al., 2017), while ours is linear in k (see Proposition H .1). Indeed, their surrogate is larger by construction because it is defined in $\mathbb{R}^{\mathcal{Y}}$, while ours is defined in $\mathbb{R}^k$. It is important to note that our surrogates are the ones used in practice while theirs require a summation over $|\mathcal{Y}|$ elements (see (3.24)), which in structured prediction is in general exponentially large.

E .2 Conditional Random Fields

This section has two parts. In the first one, we show how changing the decoding procedure in CRFs from MAP assignment (what it is used in practice) to the decoding we propose, it is possible to calibrate CRFs to any discrete loss with affine decomposition $L(z, y) = \langle\psi(z), \varphi(y)\rangle + c$, where $\varphi(y)$ are the sufficient statistics of the CRF. At the second part, we prove the convex lower bound on the calibration function.

On calibration of CRFs and MAP assignment. Using MAP as a decoding mapping does not calibrate CRFs to any discrete loss in general. This is not the case for multinomial logistic regression (which is the equivalent method in multiclass classification),

[5] If you multiply a surrogate by a factor, the associated calibration function gets multiplied by the same factor.

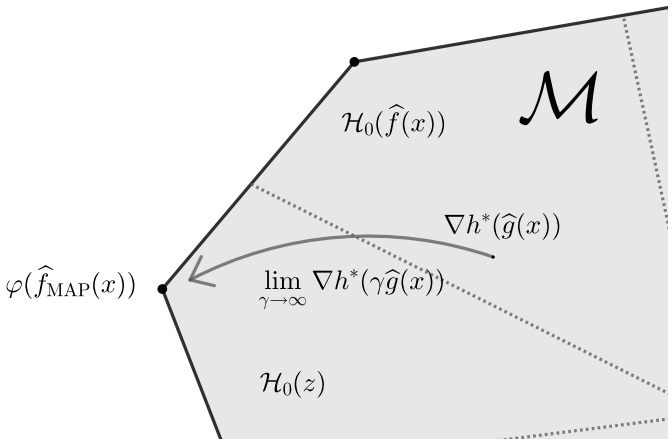


Figure 3.3: MAP assignment can be written as $\widehat{f}_{\mathrm{MAP}}(x) = \varphi^{-1}\left(\lim_{\gamma\to\infty} \nabla h^*(\gamma\widehat{g}(x))\right)$, which corresponds to continuously move the vector of predicted marginals $\nabla h^*(\widehat{g}(x))$ to a vertex of the marginal polytope $\lim_{\gamma\to\infty} \nabla h^*(\gamma\widehat{g}(x))$. The decoding $d_{h,\nabla h}$ corresponds to assign to $\nabla h^*(\widehat{g}(x))$ an output element $z \in \mathcal{Z}$ such that $\nabla h^*(\widehat{g}(x)) \in \mathcal{H}_0(z)$.

where our decoding corresponding to the multiclass 0-1 loss is exactly MAP assignment (see Section G .2). A way to understand the difference between both decoding mappings is to write MAP assignment in terms of ∇h^* as:

$$\begin{aligned}
\widehat{f}_{\mathrm{MAP}}(x) &= \arg\max_{y\in\mathcal{Y}} \ \langle \varphi(y), \widehat{g}(x)\rangle \\
&= \varphi^{-1}\left(\lim_{\gamma\to\infty} \arg\max_{\mu\in\mathcal{M}} \{\langle \gamma\widehat{g}(x), \mu\rangle - h(\mu)\}\right) \\
&= \varphi^{-1}\left(\lim_{\gamma\to\infty} \nabla h^*(\gamma\widehat{g}(x))\right).
\end{aligned}$$

With this form we can compare it to the decoding of our framework which is

$$\widehat{f}(x) = \arg\min_{z\in\mathcal{Z}} \ \langle \psi(z), \nabla h^*(\widehat{g}(x))\rangle.$$

See Figure 3.3 with explanation.

Finally, we provide the proof of the lower bound on ζ_h given by Theorem 4 .4, which is based on the computation of the strong convexity constant w.r.t the Euclidean distance of the negative maximum-entropy potential.

Proof of Proposition 5 .2. Recall that the strong convexity constant of a Legendre-type function h w.r.t a norm $\|\cdot\|$ is the inverse of the Lipschitz constant of ∇h^* w.r.t $\|\cdot\|_*$. In this case, h^* corresponds to the partition function

$$h^*(v) = \log\Big(\sum_{y'\in\mathcal{Y}} \exp(\langle v, \varphi(y')\rangle\Big),$$

and the Hessian corresponds to the Fisher Information matrix which in this case is equal to the covariance $\Sigma(v)$ of $\varphi(y)$ under $p_\varphi(\cdot|v)$, where

$$p_\varphi(\cdot|v) = \frac{\exp\ \langle \varphi(y), v\rangle}{\sum_{y'\in\mathcal{Y}} \exp\ \langle \varphi(y'), v\rangle},$$

is the exponential family with sufficient statistics φ and parameter vector v. Hence, the strong convexity constant of h under the Euclidean norm is the maximal spectral norm of the covariance, which can be upper bounded as

$$\begin{aligned}\sup_{v\in\mathcal{V}} \|\Sigma(v)\|_2 &= \sup_{v\in\mathcal{V}} \left\| \sum_{y\in\mathcal{Y}} q_\varphi(y|v)\cdot\varphi(y)\varphi(y)^T \right\|_2 \\ &\leq \sup_{q\in\mathrm{Prob}(\mathcal{Y})} \left\| \sum_{y\in\mathcal{Y}} q(y)\cdot\varphi(y)\varphi(y)^T \right\|_2 \\ &= \sup_{y\in\mathcal{Y}} \|\varphi(y)\|_2^2.\end{aligned}$$

□

F Binary Classification

We will present the cost-sensitive case to highlight the fact that a φ-calibrated loss can be calibrated to multiple losses. In this case $\mathcal{Z} = \mathcal{Y} = \{-1, 1\}$ and consider the following cost-sensitive loss L defined as $L(-1, 1) = 2 - c, L(1, -1) = c$ and 0 otherwise with $0 < c \leq 1$. We consider the following embeddings $\psi(1) = (0, c)^T$, $\psi(-1) = (2 - c, 0)^T$, $\varphi(1) = (1, 0)^T$, $\varphi(-1) = (0, 1)^T$. In this case $\mathcal{H} = \mathbb{R}^2$, $\mathcal{M} = \Delta_2$, $r = 1$, $k = 2$. Hence, the marginal polytope is not full-dimensional. The decoding corresponds to $d(v) = \mathrm{sign}(2q - c)$, where we will abuse notation and set $q = q(Y = 1) = \mu_1 = \mathbb{E}_{Y\sim q}\,\varphi_1(Y)$.

We will focus on surrogate *margin losses* (Bartlett et al., 2006), which are losses of the form $S(v, y) = \Phi(yv)$ with $\mathcal{V} = \mathbb{R}$, where $\Phi : \mathbb{R} \to \mathbb{R}$ is a non-increasing function with $\Phi(0) = 1$. The link function is computed as

$$t(q) = \arg\min_{v\in\mathcal{V}} \ \mathbb{E}_{Y\sim q}\,\Phi(Yv). \tag{3.27}$$

Note that it is always the case that the link is symmetric around $q = 1/2$, i.e., $t(q)(2q - 1) > 0$ for $q \neq 1/2$. Hence, in the non-cost-sensitive case ($c = 1$), the decoding can be simplified to $d(v) = \mathrm{sign}(2t^{-1}(v) - 1) = \mathrm{sign}(v)$. Moreover, the potential function can be computed as

$$-h(q) = \min_{v\in\mathcal{V}} \ \mathbb{E}_{Y\sim q}\,\Phi(Yv), \tag{3.28}$$

and it is also symmetric around $1/2$. In the following, we prove that logistic, exponential and square margin losses are φ-calibrated, squared hinge and modified Huber satisfy Eq. (3.27) for injective $t : \Delta_2 \to \mathcal{V}$ but don't have a BD representation extension to $\mathcal{V}$, and hinge loss does not satisfy Eq. (3.27) because the corresponding t is not injective.

Here we present those examples and provide the corresponding BD representation (if applicable) and the calibration function using Proposition F .3, which states that $\zeta_h(\varepsilon) = h((1 + \varepsilon)/2) - h(1/2)$ when $c = 1$.

Remark F .1 (Notation). Throughout this section, we will identify Δ_2 with $[0, 1]$ and $\mathrm{affhull}(\Delta_2)$ to $\mathbb{R}$ by projecting onto the first coordinate.

Logistic. The logistic loss corresponds to $\Phi(u) = \log(1 + \exp(-u))$.

$$\mathcal{D} = [0, 1], \quad h(q) = -\operatorname{Ent}(q), \quad t(q) = \log\left(\frac{q}{1-q}\right), \quad \zeta_h(\varepsilon) = 1 - \operatorname{Ent}\left(\frac{1+\varepsilon}{2}\right).$$

The link is the canonical link and corresponds to $t(q) = h'(q) = \log(q/(1-q))$ with inverse $t^{-1}(v) = (h^*)'(v) = (1 + e^{-v})^{-1}$.

Exponential. The exponential loss corresponds to $\Phi(u) = \exp(-u)$.

$$\mathcal{D} = [0, 1], \quad \Phi(u) = \exp(-u), \quad h(q) = -2\sqrt{q(1-q)}$$
$$t(q) = \frac{1}{2}\log\left(\frac{q}{1-q}\right), \quad \zeta_h(\varepsilon) = 1 - \sqrt{1-\varepsilon^2}.$$

The link corresponds to $t(q) = \frac{1}{2}\log(q/(1-q))$ with inverse $t^{-1}(v) = (1 + e^{-2v})^{-1}$. It does not correspond to the canonical link, which is $(h^*)'(v) = \frac{1}{2}\left(1 - \frac{u}{\sqrt{4+u^2}}\right)$ and $h'(q) = \frac{1-2q}{\sqrt{q(1-q)}}$.

Square. The square loss corresponds to $\Phi(u) = (1-u)^2$.

$$\mathcal{D} = \mathbb{R}, \quad h(q) = -4q(1-q), \quad t(q) = 2q - 1, \quad \zeta_h(\varepsilon) = \varepsilon^2.$$

The link corresponds to $t(q) = 2q - 1$ with inverse $t^{-1}(v) = (v+1)/2$. It does correspond to the canonical link up to a multiplicative factor because $(h^*)'(v) = (4+v)/8$ and $h'(p) = 4(2q-1)$.

Squared Hinge. The squared hinge loss corresponds to $\Phi(u) = (\max(1-u, 0))^2$. The link and potential is the same in Δ_2 as the square margin loss. However, in this case the excess conditional surrogate risk reads (see Zhang (2004b))

$$\delta s(v, q) = (2q - 1 - v)^2 - q\max(v-1, 0)^2 - (1-q)\min(0, v+1)^2.$$

We know that for $v \in t(\Delta_2) = [-1, 1]$, $\delta s(v, q) = 4(q - t^{-1}(v))^2 = D_h(q, t^{-1}(v))$ as the square loss. However this BD representation can't be extended to $\mathcal{V} = \mathbb{R}$ (see Proposition F .2).

Modified Huber loss. The Modified Huber loss (Zhang, 2004b) corresponds to

$$\Phi(u) = \begin{cases} 0 & \text{if } u \geq 1 \\ -4u & \text{if } u \leq -1 \\ (1-u)^2 & \text{otherwise} \end{cases}.$$

The excess conditional surrogate risk reads (see Zhang (2004b))

$$\delta s(v, q) = (2q - 1 - T(v))^2 + 2|2q - 1 - T(v)||q - T(v)|,$$

where $T(v) = \min(\max(v, -1), 1)$. As the squared hinge, it has the same BD representation as the squared margin loss in $t(\Delta_2) = [-1, 1]$ but it can't be extended to $\mathcal{V} = \mathbb{R}$ (see Proposition F .2).

Hinge. The hinge loss corresponds to $\Phi(u) = \max(1-u, 0)$. We have that

$$t(q) = \operatorname{sign}(2q-1).$$

Note that in this case t is not injective.

Proposition F .2. *The Bregman divergence representation of the squared hinge can not be extended to* $\mathbb{R}$.

Proof. Observe that $\delta s(v, 1) = 0$ for $v \geq 1$ and so for any right continuous extension of $t(q) = 2q-1$ to $[1, +\infty)$, $\delta s(t(q'), 1) = 0$ for $q' \geq 1$. In particular, this means that the extension of h must be linear for all $q' \geq 1$. And so $\delta s(v, q)$ should be independent of $v \geq 1$ for any $q < 1$. However, this is not the case. Hence, squared hinge does not have a BD representation extension to $\mathcal{V} = \mathbb{R}$. □

Finally, we prove the form of the calibration function for binary margin losses, which can be found at Bartlett et al. (2006) for $c = 1$ and at Scott (2012) for the asymmetric case.

Proposition F .3 (Binary 0-1 calibration function)**.** *The calibration function for a φ-calibrated margin loss can be written as $\zeta_h(\varepsilon) = \min_{\alpha \in \{-\varepsilon, +\varepsilon\}} D_h((c+\alpha)/2, c)$. Moreover, if $c = 1$, then the calibration function simplifies to $\zeta_h(\varepsilon) = h\left(\frac{1+\varepsilon}{2}\right) - h\left(\frac{1}{2}\right)$.*

Proof. Note that $\langle \psi(1) - \psi(-1), u \rangle = cu_2 - (2-c)u_1$. Hence,

$$\mathcal{H}_\varepsilon(1) = \{u \in \mathbb{R}^2 \mid cu_2 - (2-c)u_1 \leq \varepsilon\}.$$

Taking the intersection with $\mathcal{M}$ gives

$$\mathcal{H}_\varepsilon(1) \cap \Delta_2 = \{q \in [0,1] \mid c - 2q \leq \varepsilon\} = \left[\left(\frac{c-\varepsilon}{2}\right), 1\right].$$

Analogously, we obtain $\mathcal{H}_\varepsilon(-1) = \{u \in \mathbb{R}^2 \mid (2-c)u_1 - cu_2 \leq \varepsilon\}$ and $\mathcal{H}_\varepsilon(-1) \cap \Delta_2 = [0, (\frac{c+\varepsilon}{2})]$. Recall that $\mathcal{D} \supseteq [0,1]$. We obtain

$$D_h(\mathcal{H}_\varepsilon(1)^c \cap \Delta_2, \mathcal{H}_0(1) \cap \mathcal{D}) = D_h([0, (c-\varepsilon)/2], [c/2, \infty) \cap \mathcal{D}) = D_h((c-\varepsilon)/2, c/2),$$

and

$$D_h(\mathcal{H}_\varepsilon(-1)^c \cap \Delta_2, \mathcal{H}_0(-1) \cap \mathcal{D}) = D_h([(c+\varepsilon)/2, 1], \mathcal{D} \cap (-\infty, c/2]) = D_h((c+\varepsilon)/2, c/2).$$

Hence, we obtain the desired result:

$$\zeta_h(\varepsilon) = \min_{\alpha \in \{-\varepsilon, +\varepsilon\}} D_h((c+\alpha)/2, c/2).$$

Finally, setting $c = 1$ gives $\zeta_h(\varepsilon) = \min_{\alpha \in \{-\varepsilon, +\varepsilon\}} D_h((1+\varepsilon)/2, 1/2)$. Note that if h is convex differentiable and symmetric around $1/2$, then $h((1+\varepsilon)/2) = h((1-\varepsilon)/2)$ and $h'(1/2) = 0$, which simplifies the expression to

$$\zeta_h(\varepsilon) = h\left(\frac{1+\varepsilon}{2}\right) - h\left(\frac{1}{2}\right).$$

□

G Multiclass Classification

In this case, the loss considered is the 0-1 loss with $\mathcal{Z} = \mathcal{Y} = \{1, \dots, k\}$. The 0-1 loss can be written as $L(z, y) = 1 - \langle e_z, e_y \rangle$, where $e_z, e_y \in \mathbb{R}^k$ are vectors of the natural basis in $\mathbb{R}^k$. Setting $\psi(z) = -e_z$, $\varphi(y) = e_y$, the marginal polytope $\mathcal{M} = \Delta_k$ corresponds to the simplex in $\mathbb{R}^k$, $r = k - 1$, which means that the marginal polytope is not full-dimensional. We write $q_j = \mu_j = q(Y_j = 1)$. The decoding mapping can be written as $d(v) = \arg\max_{j \in [k]} t_j^{-1}(v)$. See Figure 3.4 for a visualization of the calibration sets.

G .1 One-vs-all Method

The *one-vs-all method* (Zhang, 2004a) corresponds to $S(v, y) = \Phi(v_y) + \sum_{j \neq y}^k \Phi(-v_j)$ with $\mathcal{V} = \mathbb{R}^k$. The surrogate conditional risk reads $s(v, q) = \sum_{j=1}^k \{q_j \Phi(v_j) + (1 - q_j)\Phi(-v_j)\}$. Note that one can compute h and t as in (3.27) and (3.28) independently for each coordinate. Hence, we obtain

$$h(q) = \sum_{j=1}^k \bar{h}(q_j) \qquad t(q) = (\bar{t}(q_j))_{j=1}^k, \tag{3.29}$$

where $\bar{h}, \bar{t}$ are the potential and the link corresponding to the associated margin loss. As the individual link $\bar{t}$ is invertible and $\bar{t}(q)(2q - 1) > 0$ for $q \neq 1/2$, it means that it is increasing and so $t = (\bar{t})_{j=1}^k$ is order preserving. This implies that the decoding can be simplified to $d(v) = \arg\max_{j \in [k]} v_j$. Note that if the margin loss is φ-calibrated, then the associated one-vs-all method is φ-calibrated for multiclass with h, t given by (3.29) and $\mathcal{D} = \bar{\mathcal{D}}^k$, where $\bar{\mathcal{D}}$ is the (extended) domain of the margin loss. Note that in this case the marginal polytope is always a strict subset of $\mathcal{D}$: $\Delta_k \subsetneq [0, 1]^k \subseteq \bar{\mathcal{D}}^k$.

We know provide the proof of Proposition 5 .3, which computes the exact calibration function for the one-vs-all method.

Proof of Proposition 5 .3. Note that by exploiting the symmetries of the problem, one can considerably simplify it to

$$\zeta_h(\varepsilon) = D_h(\{p_2 \geq p_1 + \varepsilon, p_2 \geq p_j\} \cap \Delta_k, \{q_1 \geq q_2\} \cap \bar{\mathcal{D}}^k). \tag{3.30}$$

Indeed, as all of the quantities in Eq. (3.10) are invariant by permutation, hence, one can get rid of the minimization over $\mathcal{Z}$ and set $z = 1$. Then, also by symmetry, one can reduce to problem to the comparison between $z = 1$ and $z = 2$.

The idea of the proof is to show that the minimizer of the following minimization problem

$$\min_{\substack{q_1 \geq q_2 \\ p_2 \geq p_1 + \varepsilon}} D_h(p, q) \tag{3.31}$$

is achieved at the points $q^\star = (1/2, 1/2, 0, \dots, 0)$ and $p^\star = ((1+\varepsilon)/2, (1-\varepsilon)/2, 0, \dots, 0)$. Then, as the constraints in Eq. (3.31) are included in Eq. (3.30) and $q^\star \in \mathcal{H}_0(1) \cap \bar{\mathcal{D}}^k \subset \{q_1 \geq q_2\}$ and $p^\star \in \mathcal{H}_\varepsilon(1)^c \cap \Delta_k \subset \{p_2 \geq p_1 + \varepsilon\}$, the result follows. Note that as $D_h(p, q) = \sum_{j=1}^k D_{\bar{h}}(p_j, q_j)$, we already have $p_j^\star = q_j^\star = 0$ for $j = 3, \dots, k$. Hence, we reduce the problem to minimizing $D_{\bar{h}}(p_1, q_1) + D_{\bar{h}}(p_2, q_2)$ in $\{q_1 \geq q_2\} \cap \{p_2 \geq p_1 + \varepsilon\}$. Note that the minimum must be necessarily achieved at the boundary. So by setting $\bar{q} = q_1 = q_2$ and $\bar{p} = p_1$, one has the following

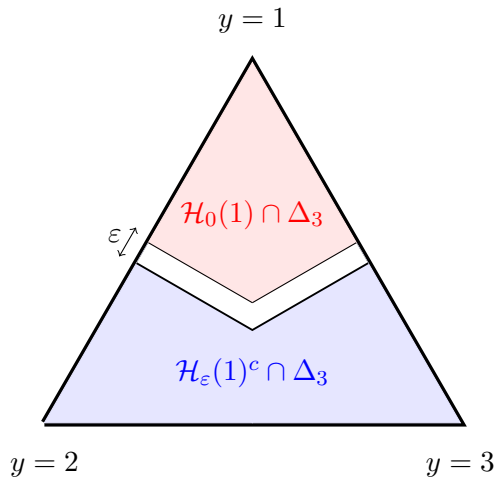


Figure 3.4: Illustration of calibration sets for the multiclass with 0-1 loss and $k = 3$ labels.

unconstrained problem $D_{\bar{h}}(\bar{p}, \bar{q}) + D_{\bar{h}}(\bar{p} + \varepsilon, \bar{q})$. Now, using the fact that $\bar{h}$ is symmetric around $1/2$ and its Hessian is non-decreasing in $\bar{\mathcal{D}} \cap [1/2, \infty)$, we have that $\bar{q} = 1/2$ and $\bar{p} = (1 - \varepsilon)/2$ is a minimizer. Hence, the result follows. □

Comparison with lower bounds. Note that if $\bar{h}$ is $(1/\beta_{\|\cdot\|_2})$-strongly convex in $\bar{\mathcal{D}}$, then h is $(1/\beta_{\|\cdot\|_2})$-strongly convex in $\mathcal{D} = \bar{\mathcal{D}}^k$. Moreover, using that $\max_{z' \neq z} \|e_z - e_{z'}\|_2^2 = 2$, Theorem D .4 gives

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{4\beta}.$$

Note that for square margin loss, where $\bar{h}(q) = -4q(1-q)$ with $\beta_{\|\cdot\|_2} = \frac{1}{8}$, this lower bound is tight.

G .2 Multinomial Logistic

Another important example is the *multinomial logistic* surrogate, which corresponds to the loss (3.14) with $\mathcal{M} = \Delta_k$, which is $S(v, y) = \log(\sum_{j=1}^k \exp(v_j)) - v_y$ where $\mathcal{V} = \{v \in \mathbb{R}^k \mid \sum_{j=1}^k v_j = 0\} \cong \mathbb{R}^{k-1}$. In this case $\mathcal{D} = \Delta_k$, $t_j^{-1}(v) = \frac{\exp(v_j)}{\sum_{\ell=1}^k \exp(v_\ell)}$ and so the decoding is also simplified to $d(v) = \arg\max_{j \in [k]} v_j$ by taking the logarithm coordinate-wise, which is a monotone function.

Lower bound on calibration function. Note that the entropy is 1-strongly convex w.r.t the $\|\cdot\|_1$ norm over the simplex. As $\|\cdot\|_\infty$ is the associated dual norm and $\max_z \|\psi(z)\|_\infty = \max_z \|e_z\|_\infty = 1$, Theorem 4 .4 gives

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{8}.$$

H Multilabel Classification with Hamming Loss

In this case $\mathcal{Z} = \mathcal{Y} = \{-1, 1\}^k$. We have that

$$L(z, y) = \frac{1}{k}\sum_{j=1}^{k} 1(z_j \neq y_j) = \frac{1}{2} + \sum_{j=1}^{k} \psi_j(z)\varphi_j(y), \tag{3.32}$$

with $\psi(z) = -z/(2k)$ and $\varphi(y) = y$. In this case $\mathcal{M} = [-1, 1]^k$ which corresponds to the cube.

H .1 Independent Classifiers

The surrogates considered are $S(v, y) = \sum_{j=1}^{m} \Phi(y_j v_j)$, with $\mathcal{V} = \mathbb{R}^k$. The surrogate conditional risk reads $s(v, q) = \sum_{j=1}^{k}\{\frac{\mu_j+1}{2}\Phi(v_j) + \frac{1-\mu_j}{2}\Phi(-v_j)\}$. Note the similarity with the one-vs-all method from multiclass classification. If $\bar{h}$ and $\bar{t}$ are the potential and link for the associated margin loss, we have:

$$h(\mu) = \sum_{j=1}^{k} \bar{h}\left(\frac{\mu_j + 1}{2}\right) \qquad t(\mu) = \left(\bar{t}\left(\frac{\mu_j + 1}{2}\right)\right)_{j=1}^{k}. \tag{3.33}$$

The decoding simplifies to $d(v) = (\text{sign}(v_j))_{j=1}^{k}$. The calibration function ζ_h can be computed exactly and it is k times the calibration function of the margin loss: $\zeta_h(\varepsilon) = k \cdot \zeta_{\bar{h}}(\varepsilon)$.

Proposition H .1 (Calibration function for Hamming loss)**.** *The calibration function for the Hamming loss is*

$$\zeta(\varepsilon) = k \cdot \zeta_{\bar{h}}(\varepsilon).$$

Proof. The proof consists of two parts. First, we show that the lower bound $\zeta_h(\varepsilon) \geq k \cdot \zeta_{\bar{h}}(\varepsilon)$ holds, and then we prove that it is actually tight, by showing it is achieved at a pair of points on the minimization problem (3.10). The excess conditional risk can be written as $\delta\ell(z, q) = \frac{1}{k}\sum_{j|z_j \neq z_j(\mu)} |\mu_j|$, where $z(\mu)$ denotes the optimal prediction. Note that $|\mu_j|$ is the excess conditional risk of the binary 0-1 loss, hence, $\zeta_{\bar{h}}(|\mu_j|) \leq \delta s_j(\bar{t}(\mu_j), q)$, where $\delta s_j(v_j, q) = \frac{\mu_j+1}{2}\Phi(v_j) + \frac{1-\mu_j}{2}\Phi(-v_j)$ is the excess surrogate conditional risk of the j-th independent classifier. Hence,

$$\begin{aligned}
\zeta_{\bar{h}}(\delta\ell(d(v), q)) &= \zeta_{\bar{h}}\left(\frac{1}{k}\sum_{j|z_j \neq d_j(v)} |2\bar{t}^{-1}(v_j) - 1|\right) && (\mu_j = 2t^{-1}(v_j) - 1) \\
&\leq \frac{1}{k}\sum_{j|z_j \neq d_j(v)} \zeta_{\bar{h}}(|2\bar{t}^{-1}(v_j) - 1|) && \text{(Jensen ineq.)} \\
&\leq \frac{1}{k}\sum_{j|z_j \neq d_j(v)} \delta s_j(v_j, q) && (\zeta_{\bar{h}}\text{calibrates single classifiers.}) \\
&\leq \frac{1}{k}\delta s(v, q).
\end{aligned}$$

Hence, $\zeta_h(\varepsilon) \geq k \cdot \zeta_{\bar{h}}(\varepsilon)$. To prove tightness, consider the point $\mu_0 = 0 = (0)_{j=1}^{k}$ and the point $\mu_\varepsilon = (-\varepsilon)_{j=1}^{k}$ for all $0 \leq \varepsilon \leq 1$. If we denote by $\mathbf{1}$ the output $(1, \ldots, 1) \in$

$\mathcal{Z}$, we have that $\mu_0 \in \mathcal{H}_0(\mathbf{1}) \cap \mathcal{M} \subset \mathcal{H}_0(\mathbf{1}) \cap \mathcal{D}$ and $\mu_\varepsilon \in \mathcal{H}_\varepsilon(\mathbf{1})^c \cap \mathcal{M}$ for all $0 \leq \varepsilon \leq 1$ because

$$\mathcal{H}_\varepsilon(\mathbf{1}) = \left\{ u \in \mathcal{H} \mid -\frac{1}{k} \sum_{j \mid b_j = 1} u_j \leq \varepsilon, \forall b \in \{0,1\}^k \right\}.$$

Moreover, its Bregman divergence is $D_h(\mu_\varepsilon, \mu_0) = \sum_{j=1}^k D_{\bar{h}}(((\mu_\varepsilon)_j + 1)/2, (1 + (\mu_0)_j)/2) = \sum_{j=1}^k D_{\bar{h}}((1-\varepsilon)/2, 1/2) = k \cdot \zeta_{\bar{h}}(\varepsilon)$. Hence, the lower bound is tight. □

I Ordinal Regression

In this case $\mathcal{Z} = \mathcal{Y} = \{1, \ldots, k\}$ and these are ordinal output variables instead of categorical. Which means that there is an intrinsic order between them: $1 \prec \cdots \prec k$. This is captured by the absolute error loss function defined as

$$L(z, y) = |z - y|.$$

Note that in this case the loss matrix is full-rank, because it is a Toeplitz matrix. Hence, it can be seen as a "structured" cost-sensitive multiclass loss.

Let's consider the following embedding $\varphi(y) = (2 \cdot 1(y \geq j) - 1)_{j=1}^{k-1} \in \{-1, 1\}^{k-1}$ for both $\mathcal{Z}$ and $\mathcal{Y}$. In this embedding, we have that

$$L(z, y) = \frac{k-1}{2} - \frac{1}{2} \sum_{j=1}^{k-1} \varphi_j(z) \varphi_j(y). \tag{3.34}$$

Comparing the expression above with the affine decomposition of the Hamming loss from the section above, we observe that Eq. (3.32) and Eq. (3.34) are proportional by a factor $k - 1$.

The decoding from $\mathcal{H}$ can be written as

$$z(\mu) = 1 + \sum_{j=1}^{k-1} 1(\mu_j \geq 0) \in \mathcal{Z}.$$

The excess conditional risk is

$$\delta\ell(z, q) = \sum_{\substack{z(\mu(q)) < j \leq z \\ z < j \leq z(\mu(q))}} |\mu_j(q)|.$$

By choosing an embedding φ different than the canonical one used in multiclass classification ($\varphi(y) = e_y$), we have performed an affine transformation to the simplex in k dimensions and project it inside the cube $[-1, 1]^{k-1}$. What we have gained is that under this transformation the calibration sets have the same structure as for the Hamming loss for $k - 1$ labels. Note however, that the marginal polytope for the Hamming loss is the entire cube, while here is a strict subset of the cube. See Figure 3.5 for the calibration sets $\mathcal{H}_0(z)$ for $k = 3$ using the canonical embedding.

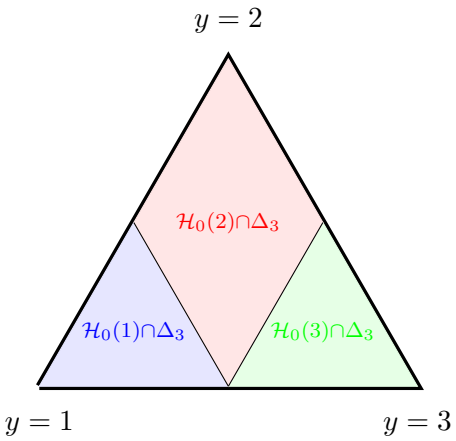


Figure 3.5: Partition of the simplex corresponding to the absolute loss for ordinal regression with $\mathcal{Z} = \mathcal{Y} = \{1, 2, 3\}$.

I .1 All thresholds (AT)

AT methods (Lin and Li, 2006) correspond to apply an independent classifier (see Section H .1) to the embedding φ. It corresponds to

$$S(v, y) = \sum_{j=1}^{k-1} \Phi(\varphi_j(y)v_j),$$

with $\mathcal{V} = \mathbb{R}^{k-1}$. We have that $h(\mu) = \sum_{j=1}^{k-1} \bar{h}((\mu_j+1)/2)$ and $t(\mu) = (\bar{t}((\mu_j+1)/2))_{j=1}^{k-1}$, exactly as for the Hamming loss. Note that in this case however $\mathcal{M} \subsetneq \mathcal{D}$ and the decoding mapping is $d(v) = 1 + \sum_{j=1}^{k-1} 1(v_j \geq 0) \in \mathcal{Z}$.

Using the fact that with the embedding φ, the ordinal loss is a $(k-1)$ factor away from the Hamming loss, and that the marginal polytope is included in the cube, so the minimization is done in a smaller domain, we have that

$$\zeta_h(\varepsilon) \geq \zeta_{\text{ham},h}(\varepsilon/(k-1)) = (k-1) \cdot \zeta_{\bar{h}}(\varepsilon/(k-1)), \tag{3.35}$$

where $\zeta_{\text{ham},h}$ is the calibration function of the Hamming loss. With this, we recover the calibration results from Pedregosa et al. (2017).

I .2 Cumulative link (CL)

These methods are of the form (McCullagh, 1980)

$$S(v, y) = \begin{cases} -\log(\bar{t}^{-1}(v_1)) & \text{if } y = 1 \\ -\log(\bar{t}^{-1}(v_y) - \bar{t}^{-1}(v_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \bar{t}^{-1}(v_{k-1})) & \text{if } y = k \end{cases},$$

with $\mathcal{V} = \mathbb{R}^{k-1}$. In this case, it corresponds to the following decomposition $L(z, y) = \langle L^T \cdot e_z, e_y \rangle$, (i.e., $\varphi(y) = e_y \in \mathbb{R}^k$), $h(q) = -\operatorname{Ent}(q)$ and $\mathcal{D} = \Delta_k$ with inverse link,

$$t^{-1}(v) = \begin{cases} \bar{t}^{-1}(v_1) & \text{if } y = 1 \\ \bar{t}^{-1}(v_y) - \bar{t}^{-1}(v_{y-1}) & \text{if } 1 < y < k \\ 1 - \bar{t}^{-1}(v_{k-1}) & \text{if } y = k \end{cases},$$

which is not the canonical one. It is called cumulative link because the link is applied to the cumulative probabilities $\bar{t}^{-1}(v_y) = \sum_{j=1}^{y} q_j = (\mu_y+1)/2$. The decoding can be written as $d(v) = 1 + \sum_{j=1}^{k-1} 1(\bar{t}^{-1}(v_j) \geq 1/2)$. In the case that $\bar{t}(q)(2q-1) > 0$ for $p \neq \frac{1}{2}$, then one can directly write (as for AT) $d(v) = 1 + \sum_{j=1}^{k-1} 1(v_j \geq 0)$. The most common link is the logistic link $\bar{t}^{-1}(v) = 1/(1+e^{-v}))$. With this link CL surrogate is convex (see Lemma 8 by Pedregosa et al. (2017)). In this case, the exact calibration function is not easy to calculate due to the lack of symmetry of the calibration sets (see Figure 3.5). However, it is straightforward to apply the lower bound by using the fact that the entropy is 1-strongly convex w.r.t the $\|\cdot\|_1$ norm and $c_{\psi,\|\cdot\|_\infty} = \max_{z\in\mathcal{Z},y\in\mathcal{Y}} |z-y| = k-1$ using the fact that $F_z = L^T \cdot e_z = L_z$. Hence, applying Theorem 4 .4 we obtain:

$$\zeta_h(\varepsilon) \geq \frac{\varepsilon^2}{8(k-1)^2}. \tag{3.36}$$

Note that this lower bound has a factor $(k-1)^{-2}$ instead of the $(k-1)^{-1}$ of Eq. (3.35).

This explains the experiment of Fig. 1 from Pedregosa et al. (2017), where they show that the calibration function (3.35) of AT is larger than the calibration function for CL. However, they were not able to provide any result such as (3.36).

J Ranking with NDCG Measure

Let $\mathcal{Z} = \mathfrak{S}_m$ be the set of permutations of m elements and $\mathcal{Y} = [\bar{R}]^m$ 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:

$$L(\sigma, r) = 1 - \frac{1}{N(\bar{r})} \sum_{j=1}^{m} G([\bar{r}]_j) D_{\sigma(j)}, \tag{3.37}$$

where $N(\bar{r}) = \max_{\sigma\in\mathfrak{S}_m} \sum_{j=1}^m G([\bar{r}]_j) D_{\sigma(j)}$ is a normalizer. Note that looking at Eq. (3.37), we have the following affine decomposition (Ramaswamy et al., 2013; Nowak et al., 2019):

$$\psi(\sigma) = -(D_{\sigma(j)})_{j=1}^m, \quad \varphi(\bar{r}) = \left(\frac{G([\bar{r}]_j)}{N(\bar{r})}\right)_{j=1}^m. \tag{3.38}$$

Inference from $\mathcal{H}$ corresponds to $z(\mu) = \operatorname{argsort}_{\sigma\in\mathfrak{S}_m}(\mu_j)_{j=1}^m$. If we now consider a strictly convex potential defined in $\mathbb{R}^m$ and the canonical link, we recover the group of surrogates presented by Ravikumar et al. (2011). With our framework, Fisher consistency comes for free by construction, we recover the same lower bound on the calibration function of their Thm. 10 from Theorem 4 .4 and the same improvement under low noise of their Thm. 11 from Theorem 4 .6.

K Graph Matching

In graph matching, the input space $\mathcal{X}$ encodes features of two graphs G_1, G_2 with the same set of nodes, and the goal is to map the nodes from G_1 to the nodes of G_2. The loss

used for graph matching is the Hamming loss between permutations defined as

$$L(\sigma, \sigma') = \frac{1}{m} \sum_{j=1}^{m} 1(\sigma(j) \neq \sigma'(j)) = 1 - \frac{\langle X_\sigma, X_{\sigma'} \rangle_F}{m} = 1 + \langle \psi(\sigma), \varphi(\sigma') \rangle_F,$$

where $X_\sigma \in \mathbb{R}^{m \times m}$ is the permutation matrix associated to the permutation σ and the embeddings are $\psi(\sigma) = -X_\sigma/m$ and $\varphi(\sigma) = X_\sigma$. In this case, the conditional risk reads

$$\ell(z, q) = 1 - \frac{\langle X_\sigma, P(q) \rangle_F}{m},$$

where $P(q) = \sum_{\sigma'} q(\sigma') X_{\sigma'}$ and $\mathcal{H} = \mathbb{R}^k$ with $k = m^2$. The Bayes optimum is computed through *linear assignment* as

$$z(P) = \arg\max_{\sigma' \in \mathfrak{S}} \ \langle X_{\sigma'}, P \rangle_F.$$

In this case, the marginal polytope corresponds to the polytope of doubly stochastic matrices (also called Birkhoff polytope),

$$\mathcal{M} = \{P \in \mathbb{R}^{m \times m} \mid P^T \mathbf{1} = \mathbf{1}, P\mathbf{1} = \mathbf{1}, 0 \leq P_{ij} \leq 1, 1 \leq i, j \leq m\},$$

which has dimension $r = \dim(\mathcal{M}) = k^2 - 2k + 1 < k^2$. One might consider CRFs (Petterson et al., 2009), however, the inverse of the canonical link requires performing inference to the associated exponential family (see Section E .2) and this corresponds to computing the permanent matrix which is a #P-complete problem. A possible workaround is to estimate the rows of the matrix P independently with a multiclass classification algorithm and then perform linear assignment with the estimated probabilities. For instance, if one performs multinomial logistic regression independently at each row, it corresponds to the potential $h(P) = -\sum_{j=1}^{m} \mathrm{Ent}(P_j)$ where P_j is the j-th row of the matrix P and $\mathcal{D}$ is the polytope of row-stochastic matrices,

$$\mathcal{D} = \{P \in \mathbb{R}^{m \times m} \mid P\mathbf{1} = \mathbf{1}, 0 \leq P_{ij} \leq 1, 1 \leq i, j \leq m\} = \prod_{j=1}^{m} \Delta_m \supsetneq \mathcal{M},$$

which has dimension $k^2 - k < k^2$ strictly larger than the dimension of the marginal polytope. As the sum of entropies is 1-strongly convex w.r.t the $\|\cdot\|_1$ norm and $c_{\psi, \|\cdot\|_\infty} = \frac{1}{m}$, Theorem 4 .4 gives,

$$\zeta_h(\varepsilon) \geq \frac{m^2 \varepsilon^2}{8}.$$

III

Non-probabilistic Estimators

4 Consistent Structured Prediction with Max-Min Margin Markov Networks

Abstract

Max-margin methods for binary classification such as the support vector machine (SVM) have been extended to the structured prediction setting under the name of max-margin Markov networks ($\mathrm{M}^3\mathrm{N}$), or more generally structural SVMs. Unfortunately, these methods are statistically inconsistent when the relationship between inputs and labels is far from deterministic. We overcome such limitations by defining the learning problem in terms of a "max-min" margin formulation, naming the resulting method max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$). We prove consistency and finite sample generalization bounds for $\mathrm{M}^4\mathrm{N}$ and provide an explicit algorithm to compute the estimator. The algorithm achieves a generalization error of $O(1/\sqrt{n})$ for a total cost of $O(n)$ projection-oracle calls (which have at most the same cost as the max-oracle from $\mathrm{M}^3\mathrm{N}$). Experiments on multi-class classification, ordinal regression, sequence prediction and ranking demonstrate the effectiveness of the proposed method.

1 Introduction

Many classification tasks in machine learning lie beyond the classical binary and multi-class classification settings. In those tasks, the output elements are structured objects made of interdependent parts, such as sequences in natural language processing (Smith, 2011), images in computer vision (Nowozin and Lampert, 2011), permutations in ranking or matching problems (Caetano et al., 2009) to name just a few (BakIr et al., 2007). The structured prediction setting has two key properties that makes it radically different from multi-class classification, namely, the exponential growth of the size of the output space with the number of its parts, and the cost-sensitive nature of the learning task, as prediction mistakes are not equally costly. In sequence prediction, for instance, the number of possible outputs grows exponentially with the length of the sequences, and predicting a sequence with one incorrect character is better than predicting the whole sequence wrong.

Classical approaches in binary classification such as the *non-smooth* support vector machine (SVM), and the *smooth* logistic and quadratic plug-in classifiers have been extended to the structured setting under the name of max-margin Markov networks ($\mathrm{M}^3\mathrm{N}$) (Taskar et al., 2004) (or more generally structural SVM (SSVM) (Tsochantaridis et al., 2005)), conditional random fields (CRFs) (Lafferty et al., 2001) and quadratic surrogate (QS) (Ciliberto et al., 2016, 2019), respectively. Theoretical properties of CRF and QS are well-understood. In particular, it is possible to obtain finite-sample generaliza-

tion bounds of the resulting estimator on the cost-sensitive structured loss (Nowak-Vila et al., 2019). Unfortunately, these guarantees are not satisfied by $\mathrm{M}^3\mathrm{N}$s even though the method is based on an upper bound of the loss. More precisely, it is known that the upper bound can be not tight (and lead to inconsistent estimation) when the relationship between input and output labels is far from deterministic (Liu, 2007), which it is essentially always the case in structured prediction due to the exponentially large output space. This means that the estimator does not converge to the minimizer of the problem leading to inconsistency.

Recently, a line of work (Fathony et al., 2016, 2018a,b) proposed a consistent method based on an adversarial game formulation on the structured problem. However, their analysis does not allow to get generalization bounds and their proposed algorithm is specific for every setting with at least a complexity of $O(n^2)$ to obtain optimal statistical error when learning from n samples. In this paper, we derive this method in the generic structured output setting from first principles coming from the binary SVM. We name this method max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$), as it is based on a correction of the max-margin of $\mathrm{M}^3\mathrm{N}$ to a 'max-min' margin. The proposed algorithm has essentially the same complexity as state-of-the-art methods for $\mathrm{M}^3\mathrm{N}$ on the regularized empirical risk minimization problem, but it comes with consistency guarantees and finite sample generalization bounds on the discrete structured prediction loss, with constants that are polynomial in the number of parts of the structured object and do not scale as the size of the output space. More precisely, the algorithm requires a constant number of projection-oracles at every iteration, each of them having at most the same cost as the max-oracle of $\mathrm{M}^3\mathrm{N}$. We also provide experiments on multiple tasks such as multi-class classification, ordinal regression, sequence prediction and ranking, showing the effectiveness of the algorithm. We make the following contributions:

- We introduce max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$) in Definition 3 .1 and prove consistency, linear calibration and finite sample generalization bounds for the regularized ERM estimator in Thms. 3 .2, 3 .3 and 3 .4, respectively.

- We generalize the BCFW algorithm (Lacoste-Julien et al., 2013) used for $\mathrm{M}^3\mathrm{N}$s to $\mathrm{M}^4\mathrm{N}$s and solve the max-min oracle iteratively with projection oracle calls using Saddle Point Mirror Prox (Nemirovski, 2004). We prove bounds on the expected duality gap of the regularized ERM problem in Theorem 5 .1 and statistical bounds in Theorem 5 .2.

- In Section 6 , we perform a thorough experimental analysis of the proposed method on classical unstructured and structured prediction settings.

2 Surrogate Methods for Classification

In this section, we review the first principles underlying surrogate methods starting from binary classification and moving into structured prediction. We put special attention to the difference between plug-in (e.g., logistic) and direct (e.g., SVM) classifiers to show that while there is a complete picture in the binary setting, existing direct classifiers in structured prediction lack the basic properties of binary SVMs. The first goal of this paper is to complete this picture in the structured output setting.

2 .1 A Motivation from Binary Classification

Let $\mathcal{Y} = \{-1, 1\}$ and $(x_1, y_1), \dots, (x_n, y_n)$ be n input-output pairs sampled from a distribution ρ. The goal in binary classification is to estimate a binary-valued function $f^\star : \mathcal{X} \to \mathcal{Y}$ that minimizes the classification error

$$\mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho} 1(f(x) \neq y).$$

We can avoid working with binary-valued functions by considering instead real-valued functions $g : \mathcal{X} \to \mathbb{R}$ and use the prediction model $f(x) = d \circ g(x) \coloneqq \operatorname{sign}(g(x))$ (Bartlett et al., 2006) where d stands for *decoding*. The resulting problem reads

$$g^\star \in \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}} \mathcal{E}(d \circ g). \tag{4.1}$$

Unfortunately, directly estimating a $g^\star$ from (4.1) is intractable for many classes of functions (Arora et al., 1997).

Convex surrogate methods. The source of intractability of minimizing the classification error (4.1) comes from the discreteness and non-convexity of the loss. The idea of surrogate methods (Bartlett et al., 2006) is to consider a *convex surrogate loss* $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ such that $g^\star$ can be written as

$$g^\star = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}} \mathcal{R}(g) \coloneqq \mathbb{E}_{(x,y)\sim\rho} S(g(x), y). \tag{4.2}$$

In this case, $g^\star$ can be tractably estimated from n samples over a family of functions $\mathcal{G}$ using regularized ERM. The resulting estimator g_n has the form

$$g_n = \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} S(g(x_i), y_i) + \frac{\lambda_n}{2}\|g\|_{\mathcal{G}}^2, \tag{4.3}$$

where $\lambda_n > 0$ is the regularization parameter and $\|\cdot\|_{\mathcal{G}}$ is the norm associated to the hypothesis space $\mathcal{G}$. If not stated explicitly, our analysis of the surrogate method holds for any function space, such as reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950) or neural networks (LeCun et al., 2015), where we lose global theoretical convergence guarantees of problem (4.3).

The classical theoretical requirements of such a surrogate strategy are *Fisher consistency* (i) and a *comparison inequality* (ii):

$$\begin{aligned} &\text{(i)} \quad \mathcal{E}(f^\star) = \mathcal{E}(d \circ g^\star) \\ &\text{(ii)} \quad \zeta(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star), \end{aligned}$$

for all measurable functions g, where $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ is such that $\zeta(\varepsilon) \to 0$ when $\varepsilon \to 0$. Note that Condition (i) is equivalent to (4.1). Condition (ii) is needed to prove consistency results, to show that $\mathcal{R}(g) \to \mathcal{R}(g^\star)$ implies $\mathcal{E}(d \circ g) \to \mathcal{E}(f^\star)$. The existence of ζ satisfying (ii) is derived from (i) and the continuity and lower boundedness of $S(v, y)$, see Thm. 3 by Zhang (2004a). Even though the explicit form of ζ is not needed for a consistency analysis, it is necessary to prove finite sample generalization bounds, as it is the mathematical object relating the suboptimality of the surrogate problem to the suboptimality of the original task. Note that the larger $\zeta(\varepsilon)$, the better.

Plug-in classifiers. It is known that (i) is satisfied for any function $g^\star$ that continuously depends on the conditional probability $\rho(1|x)$ as $g^\star(x) \coloneqq t(\rho(1|x))$, where $t : \mathbb{R} \to \mathbb{R}$ is a suitable continuous bijection of the real line[1]. In this case, Eq. (4.2) can be satisfied using *smooth* losses. Some examples are the logistic loss $\log(1 + e^{-yv})$, the squared hinge loss $\max(0, 1-yv)^2$ and the exponential loss e^{-yv}. In this case, the convexity and smoothness of $S(\cdot, y)$ imply that (ii) is satisfied with $\zeta(\varepsilon) \sim \varepsilon^2$ (Bartlett et al., 2006). Combining this with standard convergence results of regularized ERM estimators g_n on RKHS, the resulting statistical rates are of the form $\mathbb{E}\,\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star) \sim \|g^\star\|_{\mathcal{G}} n^{-1/4}$. Even if the binary learning problem is easy, $g^\star$ can be highly non-smooth away from the decision boundary, resulting in large $\|g^\star\|_{\mathcal{G}}$. It is known that the dependence on the number of samples can be improved under low noise conditions (Audibert and Tsybakov, 2007).

Support vector machines (SVM). Plug-in classifiers indirectly estimate the conditional probability as $\rho(1|x) = t^{-1}(g^\star(x))$, which is more than just falling in the right binary decision set. SVMs directly tackle the classification task by estimating $g^\star \coloneqq f^\star = \operatorname{sign}(\rho(1|x) - 1/2)$. In this case, the *non-smooth* hinge loss $S(v, y) = \max(0, 1 - yv)$ satisfies (4.2). Moreover, (ii) is satisfied with $\zeta(\varepsilon) = \varepsilon$ and statistical rates are of the form $\mathbb{E}\,\mathcal{E}(d \circ \widehat{g}_n) - \mathcal{E}(f^\star) \sim \|f^\star\|_{\mathcal{G}} n^{-1/2}$. Note that $f^\star$ is piece-wise constant on the support of ρ, but it can be shown $f^\star \in \mathcal{G}$, (i.e., $\|f^\star\|_{\mathcal{G}} < \infty$), for standard hypothesis spaces $\mathcal{G}$ such as Sobolev spaces with input space $\mathbb{R}^d$ and smoothness $s > d/2$ under low noise conditions (Pillaud-Vivien et al., 2018a).

2 .2 Structured Prediction Setting

In binary classification, the output data are naturally embedded in $\mathbb{R}$ as $\mathcal{Y} = \{-1, 1\} \subset \mathbb{R}$. However, as this is not necessarily the case in structured prediction, it is classical (Taskar et al., 2005a) to represent the output with an embedding $\varphi : \mathcal{Y} \to \mathbb{R}^k$ encoding the parts structure with $k \ll |\mathcal{Y}|$. Let $g : \mathcal{X} \to \mathbb{R}^k$ and define the following linear prediction model

$$f(x) = d \circ g(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \ \varphi(y)^\top g(x). \tag{4.4}$$

The above decoding (4.4) corresponds to the classical linear prediction model over factorized joint features $\Phi(x, y) = \varphi(y) \otimes \Phi(x)$ when $g(x)$ is linear in some input features $\Phi(x)$ (BakIr et al., 2007). The form in (4.4) is required to perform the consistency analysis but the algorithm developed in Section 5 can be readily extended to joint features that do not factorize. Non-linear prediction models have been recently proposed by Belanger and McCallum (2016), but this is out of the scope of this paper.

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function between structured outputs encoding the cost-sensitivity of predictions. For instance, it is common to take L to be the Hamming loss over the parts of the structured object. The goal in structured prediction is to estimate $f^\star : \mathcal{X} \to \mathcal{Y}$ that minimizes the *expected risk*:

$$\mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho} L(f(x), y). \tag{4.5}$$

Loss-decoding compatibility. It is classical to assume that the loss decomposes over the structured output parts (Joachims, 2006). This can be generalized as the following

[1] It must satisfy $(u - 1/2)t(u - 1/2) \geq 0$ for all $u \in \mathbb{R}$.

affine decomposition of the loss (Ramaswamy et al., 2013; Nowak et al., 2019)

$$L(y, y') = \varphi(y)^\top A \varphi(y') + a, \tag{4.6}$$

for a matrix $A \in \mathbb{R}^{k \times k}$ and scalar $a \in \mathbb{R}$. Indeed, assumption (4.6) together with the tractability of (4.4) is essentially equivalent to the tractability of *loss-augmented inference* in structural SVMs (Joachims, 2006). For the sake of notation, we drop the constant a and work with the 'centered' loss $L(y, y') - a$. We provide some examples below.

Example 2 .1 (Structured prediction with factor graphs). Let $\mathcal{Y} = [R]^M$ be the set of objects made of M parts, each in a vocabulary of size R. In order to model interdependence between different parts, we consider embeddings that decompose over (overlapping) subsets of indices $\alpha \subseteq \{1, \dots, M\}$ (Taskar et al., 2004) as $\varphi(y) = (\varphi_\alpha(y_\alpha))_\alpha$. More precisely, the prediction model corresponds to

$$\arg\max_{y \in \mathcal{Y}} \ \sum_\alpha \varphi_\alpha(y_\alpha)^\top v_\alpha, \tag{4.7}$$

where $\varphi_\alpha(y_\alpha) = e_{y_\alpha} \in \mathbb{R}^{R^{|\alpha|}}$ with e_j being the j-th vector of the canonical basis and the dimension of the full-embedding φ is $k = \sum_\alpha R^{|\alpha|} \ll |\mathcal{Y}| = R^M$. It is common (Tsochantaridis et al., 2005) to assume that the loss decomposes additively over the coordinates as $L(y, y') = \frac{1}{M} \sum_{m=1} L_m(y_m, y'_m)$ and so the matrix A associated to the loss decomposition of L is low-rank. Problem (4.7) can be solved efficiently for low tree-width structures using the junction-tree algorithm (Wainwright and Jordan, 2008). More specifically, if the objects are sequences with embeddings modelling individual and adjacent pairwise characters, Problem (4.7) can be solved in time $O(MR^2)$ using the Viterbi algorithm (Viterbi, 1967).

Example 2 .2 (Ranking and matching). The output space is the group of permutations $\mathcal{S}_M$ acting on $\{1, \dots, M\}$. This setting also includes the task of matching the nodes of two graphs of the same size (Caetano et al., 2009). We represent a permutation $\sigma \in \mathcal{S}_M$ using the corresponding permutation matrix $\varphi(\sigma) = P_\sigma \in \mathbb{R}^{M \times M}$. The prediction model corresponds to the linear assignment problem (Burkard et al., 2012)

$$\arg\max_{\sigma \in \mathcal{S}_M} \ \langle P_\sigma, v \rangle_F, \tag{4.8}$$

where $v \in \mathbb{R}^{M \times M}$, $\langle \cdot, \cdot \rangle_F$ is the Frobenius scalar product and $k = M^2 \ll |\mathcal{Y}| = M!$. The Hamming loss on permutations satisfies Eq. (4.6) as $L(\sigma, \sigma') = \frac{1}{M} \sum_{m=1}^M 1(\sigma(m) \neq \sigma'(m)) = 1 - \frac{1}{M} \langle P_\sigma, P_{\sigma'} \rangle_F$. The linear assignment problem (4.8) can be solved in time $O(M^3)$ using the Hungarian algorithm (Kuhn, 1955).

Plug-in classifiers in structured prediction. Let $\mu(x) = \mathbb{E}_{y \sim \rho(\cdot|x)} \varphi(y)$ be the conditional expectation of the output embedding. Using the fact that $f^\star$ can be characterized pointwise in x as the minimizer in y of $\varphi(y)^\top A \mu(x)$ (Nowak et al., 2019), it directly follows that (i) is satisfied for $g^\star(x) = -A\mu(x)$ and, analogously to binary classification, it can be estimated using *smooth* surrogates. Some examples are the quadratic surrogate (QS) $\|v + A\varphi(y)\|_2^2$ (Ciliberto et al., 2016) that estimates $g^\star$ and conditional random

fields (CRF) (Lafferty et al., 2001) defined by $\log\left(\sum_{y'\in\mathcal{Y}}\exp v^\top\varphi(y')\right) - v^\top\varphi(y)$ that estimate an invertible continuous transformation of $\mu(x)$. Although CRFs have a powerful probabilistic interpretation, they cannot incorporate the cost-sensitivity matrix A into the surrogate loss, and it must be added a posteriori in the decoding (4.4) to guarantee consistency. It was shown by Nowak-Vila et al. (2019) that these methods satisfy condition (ii) with $\zeta(\varepsilon)\sim\varepsilon^2$ and achieve the analogous statistical rates of binary plug-in classifiers $\sim\|g^\star\|_{\mathcal{G}}n^{-1/4}$.

SVMs for structured prediction. The extension of binary SVM to structured outputs is the structural SVM (SSVM) (Joachims, 2006) (denoted $\mathrm{M^3Ns}$ (Taskar et al., 2004) in the factor graph setting described in Example 2 .1). It corresponds to the following surrogate loss

$$S(v,y)=\max_{y'\in\mathcal{Y}}\ \varphi(y)^\top A\varphi(y')+v^\top\varphi(y')-v^\top\varphi(y). \tag{4.9}$$

In the multi-class case with $\mathcal{Y}=\{1,\dots,k\}$ and $L(y,y')=1(y\neq y')$ it is also known as the Crammer-Singer SVM (CS-SVM) (Crammer and Singer, 2001) and reads $S(v,j)=\max_{r\neq j}\ 1+v_r-v_j$. It shares some properties of the binary SVM such as the upper bound property, i.e., $L(d\circ v,y)\leq S(v,y)$ for all $y\in\mathcal{Y}$. However, an important drawback of this loss is that while the upper bound property holds, the minimizer of the surrogate expected risk $g^\star$ and the one of the expected risk $f^\star$ do not coincide when the problem is far from deterministic, as shown by the following Proposition 2 .3.

Proposition 2 .3 (Inconsistency of CS-SVM (Liu, 2007))**.** *The CS-SVM is Fisher-consistent if and only if for all $x\in\mathcal{X}$, there exists $y\in\{1,\dots,k\}$ such that $\rho(y|x)>1/2$.*

Note that the consistency condition from Proposition 2 .3 is much harder to be met in the structured prediction case as the size of the output space is exponentially large, and it is always satisfied in the binary case (the binary SVM is always consistent). Although there exist consistent extensions of the SVM to the cost-sensitive multi-class setting such as the ones from Lee et al. (2004); Mroueh et al. (2012), they cannot be naturally extended to the structured setting. In the following section we address this problem by introducing the max-min surrogate and studying its theoretical properties.

3 Max-Min Surrogate Loss

Assume that the loss is not degenerated, i.e., $L(y,y)<L(y,y')$ for all $y,y'\in\mathcal{Y}$ such that $y\neq y'$. In this case, $f^\star(x)$ is the minimizer in y of $\varphi(y)^\top A^\top\varphi(f^\star(x))$, which means that (4.1) is satisfied by

$$g^\star(x)\coloneqq -A^\top\varphi(f^\star(x))\in\mathbb{R}^k.$$

Note the analogy with SVMs, where we directly estimate $f^\star$ but now through the representation φ of the structured output, avoiding the full enumeration of $\mathcal{Y}$. We need to find a surrogate function $S(v,y)$ that satisfies Eq. (4.2) for this $g^\star$. Following the same notation as Nowak-Vila et al. (2019), we define the *marginal polytope* (Wainwright and Jordan, 2008) as the convex hull of the embedded output space $\mathcal{M}=\mathrm{hull}(\varphi(\mathcal{Y}))\subset\mathbb{R}^k$.

Definition 3 .1 (Max-min surrogate loss)**.** *Define the max-min loss as*

$$S(v,y)\coloneqq\max_{\mu\in\mathcal{M}}\min_{y'\in\mathcal{Y}}\ \varphi(y')^\top A\mu+v^\top\mu-v^\top\varphi(y). \tag{4.10}$$

The max-min loss is *non-smooth*, *convex* and can be cast as a Fenchel-Young loss (Blondel et al., 2020). More specifically, Eq. (4.10) can be written as $S(v, y) = \Omega^*(v) + \Omega(\varphi(y)) - v^\top \varphi(y)$ with

$$\Omega(\mu) = -\min_{y' \in \mathcal{Y}} \varphi(y')^\top A\mu + 1_{\mathcal{M}}(\mu), \tag{4.11}$$

where $\Omega(\varphi(y)) = 0$ for all $y \in \mathcal{Y}$, Ω^* denotes the Fenchel-conjugate of Ω, and $1_{\mathcal{M}}(\mu) = 0$ if $\mu \in \mathcal{M}$ and $+\infty$ otherwise.

Note that the dependence on y is only in the linear term $v^\top \varphi(y)$, while for SSVMs (4.9) it appears in the maximization. Thus, we can study the geometry of the loss through the non-smooth convex function $\Omega^*(v)$ (see Figure 4.1 for visualizations of some representative unstructured examples). Connections between surrogates (4.10) and (4.9) are discussed in Section 4 .

3 .1 Fisher Consistency

Fisher consistency is provided by the following Theorem 3 .2.

Theorem 3 .2 (Fisher Consistency (i))**.** *The surrogate loss (4.10) satisfies (i) for* $g^\star(x) = -A^\top \varphi(f^\star(x))$.

This result has been proven by Fathony et al. (2018a) in the cost-sensitive multi-class case. Our proof of Theorem 3 .2 is constructive and based on Fenchel duality.

Sketch of the proof. We want to show that $-A^\top \varphi(f^\star(x))$ is the minimizer of the conditional risk $\mathbb{E}_{y \sim \rho(\cdot|x)} S(v, y)$ almost surely for every x. The proof is constructive and based on Fenchel duality, using the Fenchel-Young loss representation of the max-min surrogate. First, note that the conditional surrogate risk can be written as $\mathbb{E}_{y \sim \rho(\cdot|x)} S(v, y) = \Omega^*(v) - v^\top \mu(x)$, where $\mu(x) = \mathbb{E}_{y \sim \rho(\cdot|x)} \varphi(y) \in \mathcal{M}$. Second, note that by Fenchel-duality, $\partial_\mu \Omega(\mu(x))$ is the set of minimizers of $\Omega^*(v) - v^\top \mu(x)$. Finally, if we assume that the set of $x \in \mathcal{X}$ such that $\mu(x)$ is in the boundary of $\mathcal{M}$ has measure zero, then

$$-A^\top f^\star(x) \in \partial_\mu \Omega(\mu(x)),$$

where Ω is defined in (4.11) and we have used that $f^\star(x)$ is the minimizer in y of $\varphi(y)^\top A\mu(x)$. A more detailed proof can be found in Section B .1. □

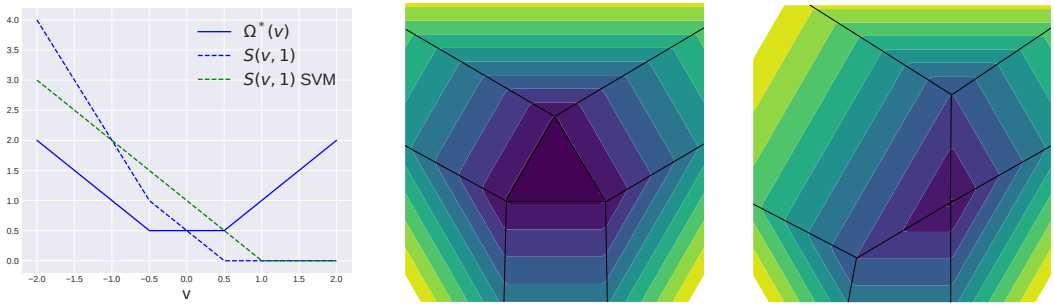


Figure 4.1: **Left:** The binary max-min loss has two symmetric kinks instead of one as the SVM. **Middle:** $\Omega^*(v)$ in $v^\top 1 = 0$ for multi-class 0-1 loss $1(y \neq y')$ with $k = 3$. **Right:** $\Omega^*(v)$ in $v^\top 1 = 0$ for ordinal regression with the absolute loss $|y - y'|$ with $k = 3$.

3 .2 Comparison Inequality

Fisher consistency is not enough to prove finite-sample generalization bounds on the excess risk $\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)$. For this, we provide in the following Theorem 3 .3 an explicit form of the comparison inequality.

Theorem 3 .3 (Comparison inequality (ii))**.** *Assume L is symmetric and that there exists $C > 0$ such that for any probability $\alpha \in \Delta_\mathcal{Y}$, it holds that $\alpha_y \geq 1/C$ for $y \in \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{z \sim \alpha} L(y, z)$. Then, the comparison inequality (ii) for the max-min loss* (4.10) *reads*

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^\star) \leq C(\mathcal{R}(g) - \mathcal{R}(g^\star)).$$

The second condition on the loss states that if y is optimal for x, then its conditional probability is bounded away from zero as $\rho(y|x) \geq 1/C$. This condition is used to obtain a simple quantitative lower bound on the function ζ of (ii) and more tight (albeit less explicit in general) expressions of the constant C can be found in Appendices C.3 and C.4.

Constant C for multi-class. When $L(y, y') = 1(y \neq y')$ with $\mathcal{Y} = \{1, \dots, k\}$, we have that $C = k$, as the minimum conditional probability of an optimal output is $1/k$. The constant for this specific setting was derived independently using a different analysis by Duchi et al. (2018).

Constant C for factor graphs (Example 2 .1). For a factor graph with separable embeddings and a decomposable loss $L = \frac{1}{M} \sum_{m=1}^{M} L_m(y_m, y'_m)$, we have that $C = \max_{m \in [M]} C_m$, where C_m is the constant associated to the individual loss L_m. This is proven in Proposition B .11.

Constant C for ranking and matching (Example 2 .2). In this setting, Theorem 3 .3 gives $C = M!$, and so the relation between both excess risks is not informative. The problem of exponential constants in the comparison inequality was pointed out by Osokin et al. (2017). We can weaken the assumption and change condition $\alpha_y \geq 1/C$ to

$$\max_{\beta \in \Delta_\mathcal{Y}} \beta_y \;\text{ s.t. }\; \mathbb{E}_{z \sim \beta} \varphi(z) = \mathbb{E}_{z' \sim \alpha} \varphi(z') \geq 1/C.$$

Under this assumption, we have that $C = M$, thus avoiding the exponentially large size of the output space.

3 .3 Generalization of Regularized ERM

In the following Theorem 3 .4, we use this result to prove a finite-sample generalization bound on the regularized ERM estimator (4.3) when the hypothesis space $\mathcal{G}$ is a vector-valued RKHS.

Theorem 3 .4 (Generalization of regularized ERM)**.** *Let $\mathcal{G}$ be a vector-valued RKHS, assume $g^\star \in \mathcal{G}$ and let g_n and $\lambda_n = \kappa L \log^{1/2}(1/\delta) n^{-1/2}$ as in* (4.3)*. Then, with probability $1 - \delta$:*

$$\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star) \leq M \|\varphi(f^\star)\|_\mathcal{G} \sqrt{\frac{\log(1/\delta)}{n}},$$

with $M = \kappa C L \|A\|$. Here, the constants are: $L = 2 \max_{y \in \mathcal{Y}} \|\varphi(y)\|_2$, $\|A\| = \sup_{\|v\|_2 \leq 1} \|Av\|_2$, $\kappa = \sup_{x \in \mathcal{X}} \operatorname{Tr} K(x, x)^{1/2}$ is the size of the features and C is the one of Theorem 3 .3.

Analogously to the binary case, the multivariate function $\varphi(f^\star)$ is piecewise constant on the support of the distribution ρ. In Theorem C .2 in Appendix C we prove that standard low noise conditions, analogous to the one discussed by Pillaud-Vivien et al. (2018a) for the binary case, are enough to guarantee $\|\varphi(f^\star)\|_{\mathcal{G}} < \infty$.

4 Comparison with Structural SVM

Max-min as a correction of the Structural SVM. We can re-write the maximization over the discrete output space $\mathcal{Y}$ in the definition of the SSVM (4.9) as a maximization over its convex hull $\mathcal{M} = \mathrm{hull}(\varphi(\mathcal{Y}))$

$$S(v, y) = \max_{\mu \in \mathcal{M}} \varphi(y)^\top A\mu + v^\top \mu - v^\top \varphi(y). \tag{4.12}$$

Note the similarity between (4.10) and (4.12). In particular, the max-min loss differs from the structural SVM in that the maximization is done using $\min_{y' \in \mathcal{Y}} \varphi(y')^\top A\mu$ and not the loss at the observed output y as $\varphi(y)^\top A\mu$. Hence, we can view the max-min surrogate as a *correction* of the SSVM so that basic statistical properties (i) and (ii) hold. Moreover, this connection might be used to properly understand the statistical properties of SSVM. This is left for future work.

Notion of max-min margin. Given $v \in \mathbb{R}^k$ and $y_i \in \mathcal{Y}$, the classical SSVM is motivated by a soft version of the following notion of margin:

$$v^\top \varphi(y_i) - v^\top \varphi(y) \geq L(y_i, y) = \varphi(y_i)^\top A\varphi(y),$$

for all $y \in \mathcal{Y}$, which is equivalent to $v^\top \varphi(y_i) - v^\top \mu \geq \varphi(y_i)^\top A\mu$ for all $\mu \in \mathcal{M}$. However, we have seen in Proposition 2 .3 that this condition is too strong and only leads to a consistent method if the problem is nearly deterministic, i.e., we observe the optimal y with large probability, which, as already mentioned, is generally far from true in structured prediction. The max-min surrogate (4.10) deals with the case where this strong condition is not met and works with a notion of margin that compares groups of outputs instead of just pairs. We define the max-min margin as

$$v^\top \varphi(y_i) - v^\top \mu \geq \min_{y' \in \mathcal{Y}} \varphi(y')^\top A\mu, \tag{4.13}$$

for all $\mu \in \mathcal{M}$. After introducing slack variables in (4.13) we obtain a soft version of the max-min margin that leads to the max-min regularized ERM problem (4.3).

5 Algorithm

In this section we derive a dual-based algorithm to solve the max-min regularized ERM problem (4.3) when the hypothesis space is a RKHS. The algorithm can be easily adapted to the case where g is parametrized using a neural network as commented at the end of Section 5 .3.

Algorithm 1: Generalized Block-Coordinate Frank-Wolfe (GBCFW), primal

Let $w^{(0)} := w_i^{(0)} := 0$;
for $t = 0$ ***to*** T **do**
  Pick i at random in $\{1, \dots, n\}$;
  $(\mu_i^\star, \nu_i^\star) \in \mathcal{O}_K(g_{w^{(t)}}(x_i), \mu_i^\star, \nu_i^\star)$;
  $w_s := \Phi(x_i)(\mu_i^\star - \varphi(y_i))^\top / (\lambda n)$;
  $w_i^{(t+1)} := (1 - \frac{2n}{t+2n}) w_i^{(t)} + \frac{2n}{t+2n} w_s$;
  $w^{(t+1)} := w^{(t)} + w_i^{(t+1)} - w_i^{(t)}$;
end

Algorithm 2: Saddle Point Mirror Prox (SP-MP) $\quad (\bar{\mu}^{(K)}, \bar{\nu}^{(K)}) \in \mathcal{O}_K(v, \mu^{(0)}\nu^{(0)})$

for $k = 0$ ***to*** $K - 1$ **do**
  $\mu_{1/2}^{(k+1)} \in \arg\min_{\mu \in \mathcal{M}} - \eta \mu^\top (A^\top \nu^{(k)} + v) + D_{-H}(\mu, \mu^{(k)})$;
  $\nu_{1/2}^{(k+1)} \in \arg\min_{\nu \in \mathcal{M}} \eta \nu^\top A \mu^{(k)} + D_{-H}(\nu, \nu^{(k)})$;
  $\mu^{(k+1)} \in \arg\min_{\mu \in \mathcal{M}} - \eta \mu^\top (A^\top \nu_{1/2}^{(k+1)} + v) + D_{-H}(\mu, \mu^{(k)})$;
  $\nu^{(k+1)} \in \arg\min_{\nu \in \mathcal{M}} \eta \nu^\top A \mu_{1/2}^{(k+1)} + D_{-H}(\nu, \nu^{(k)})$;
end
$\bar{\mu}^{(K)} := \frac{1}{K} \sum_{k=1}^K \mu^{(k)}$, $\bar{\nu}^{(K)} := \frac{1}{K} \sum_{k=1}^K \nu^{(k)}$

5 .1 Problem Formulation

Let $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R}^k\}$ be a vector-valued RKHS, which we assume of the form $\mathcal{G} = \mathbb{R}^k \otimes \overline{\mathcal{G}}$, where $\overline{\mathcal{G}}$ is a scalar RKHS with associated features $\Phi : \mathcal{X} \to \overline{\mathcal{G}}$. Every function in $\mathcal{G}$ can be written as $g_w(x) = w^\top \Phi(x) \in \mathbb{R}^k$ where $w_j, \Phi(x) \in \overline{\mathcal{G}}$. For the sake of presentation, we assume that $\mathcal{G} = \mathbb{R}^{d \times k}$ is finite dimensional, but our analysis also holds for the infinite dimensional case. The dual **(D)** of the regularized ERM problem (4.3) for the max-min surrogate loss (4.10) reads

$$\textbf{(D)} \quad \max_{\mu \in \mathcal{M}^n} \frac{1}{n} \sum_{i=1}^n \min_{y'} \varphi(y')^\top A \mu_i - \frac{\lambda}{2} \|\Phi_n(\mu - \varphi_n)\|_2^2,$$

where $\Phi_n = \frac{1}{\lambda n}(\Phi(x_1), \dots, \Phi(x_n))$ is the $d \times n$ scaled input data matrix and the matrix $\varphi_n = (\varphi(y_1), \dots, \varphi(y_n))^\top$ is the $n \times k$ output data matrix. The dual variables map to the primal variables through the mapping $w(\mu) = \frac{1}{\lambda n} \sum_{i=1}^n \Phi(x_i)(\mu_i - \varphi(y_i))^\top$. By strong duality, it holds $w^\star = w(\mu^\star)$. The dual formulation **(D)** is a constrained *non-smooth* optimization problem, where the non-smoothness comes from the first term of the objective function. In order to derive a learning algorithm, we leverage ideas from the block-coordinate Frank-Wolfe algorithm used for SSVMs.

5 .2 Derivation of the Algorithm

Background on BCFW for M^{3}Ns. The dual of the SSVM is the same as problem **(D)** but the first term is linear: $\frac{1}{n} \sum_{i=1}^n \varphi(y_i)^\top A \mu_i$, making the dual objective function *smooth*. The BCFW algorithm (Lacoste-Julien et al., 2013) minimizes a linearization of the smooth

dual objective function block-wise, using the separability of the compact domain. At each iteration t, the algorithm picks $i \in [n]$ at random, and updates $\mu_i^{(t+1)} = (1-\gamma)\mu_i^{(t)} + \gamma\bar{\mu}_i^{(t+1)}$ with $\bar{\mu}_i^{(t+1)} := \arg\max_{\mu_i' \in \mathcal{M}} \langle \mu_i', \nabla_{(i)} h(\mu^{(t)}) \rangle$ where h is the dual objective and γ is the step-size. Note that $\bar{\mu}_i^{(t+1)}$ is an extreme point of $\mathcal{M}$ and it can be written as a combinatorial maximization problem over $\mathcal{Y}$ that corresponds precisely to inference (4.4). In the next section, we generalize BCFW to the case where the dual is a sum of a non-smooth and a smooth function such as the dual **(D)** of our problem.

Generalized BCFW (GBCFW) for M^{4}N. Borrowing ideas from Bach (2015) in the non block-separable case, we only linearize the smooth-part of the function, i.e., the quadratic term. We change the computation of the direction to

$$\bar{\mu}_i^{(t+1)} := \arg\max_{\mu_i' \in \mathcal{M}} \langle \mu_i', \nabla_{(i)} \frac{-\lambda}{2} \|\Phi_n(\mu^{(t)} - \varphi_n)\|_2^2 \rangle + \min_{y'} \varphi(y')^\top A \mu_i' = \mathcal{O}(g_{w(\mu^{(t)})}(x_i)),$$

where the max-min oracle $\mathcal{O} : \mathbb{R}^k \to \mathcal{M}$ is defined as

$$\mathcal{O}(v) = \arg\max_{\mu \in \mathcal{M}} \min_{\nu \in \mathcal{M}} \nu^\top A \mu + v^\top \mu. \tag{4.14}$$

Note that the mapping $w(\mu)$ between primal and dual variables is affine. Hence, one can write the update of the primal variables without saving the dual variables as detailed in Algorithm 1. The following Theorem 5 .1 specifies the required number of iterations of Algorithm 1 to obtain an ε-optimal solution with an approximate oracle (4.14).

Theorem 5 .1 (Convergence of GBCFW with approximate oracle)**.** *Let $\varepsilon > 0$. If the approximate oracle provides an answer with error $\varepsilon/2$, then the final error of Algorithm 1 achieves an expected duality gap of ε when $T = \tilde{O}(n + \frac{2R^2}{\lambda\varepsilon} \operatorname{diam}(\mathcal{M})^2)$, where R is the maximum norm of the features.*

5 .3 Computation of the Max-Min Oracle

The max-min oracle (4.14) corresponds to a concave-convex bilinear saddle-point problem. We use a standard alternating procedure of ascent and descent steps on the variables μ and ν, respectively. Consider a strongly concave differentiable entropy $H : \mathcal{C} \supset \mathcal{M} \to \mathbb{R}$ defined in a convex set $\mathcal{C}$ containing $\mathcal{M}$ such that $\nabla H(\mathcal{C}) = \mathbb{R}^k$ and it is verified that $\lim_{\mu \in \partial\mathcal{C}} \|\nabla H(\mu)\| = +\infty$, where $\partial\mathcal{C}$ is the boundary of $\mathcal{C}$. Then, perform Mirror ascent/descent updates using $-H$ as the Mirror map. For instance, if $u = A^\top \nu + v$ is the gradient of (4.14) w.r.t μ, the update on μ takes the following form:

$$\arg\min_{\mu \in \mathcal{M}} -\eta \mu^\top u + D_{-H}(\mu, \mu^{(t)}), \tag{4.15}$$

where $D_{-H}(\mu, \mu') = -H(\mu) + H(\mu') + \nabla H(\mu')^\top(\mu - \mu')$ is the Bregman divergence associated to the convex function $-H$. The resulting ascent/descent algorithm has a convergence rate of $O(t^{-1/2})$, which can be considerably improved to $O(t^{-1})$ with essentially no extra cost by performing four projections instead of two at each iteration. This corresponds to the extra-gradient strategy, called *Saddle Point Mirror Prox* (SP-MP) when using a Mirror map and is detailed in Algorithm 2.

Projection for factor graphs (Example 2 .1). The entropy in $\mathcal{M}$ defined by the factor graph (Wainwright and Jordan, 2008) can be written explicitly in terms of the entropies of each part $\alpha \subset [M]$ if the factor graph has a junction tree structure (Koller and Friedman, 2009). For instance, in the case of a sequence of length M with unary and adjacent pairwise factors, we have $H(\mu) = \sum_{m=1}^{M-1} H_S(\mu_{m,m+1}) - \sum_{m=1}^{M} H_S(\mu_m)$, where H_S is the Shannon entropy and $\mu_m, \mu_{m,m+1}$ are the unary and pair-wise marginals, respectively. The projection (4.15) corresponds to marginal inference in CRFs and can be computed using the sum-product algorithm in time $O(MR^2)$. In this case, the complexity of the projection-oracle is the same the one of the max-oracle for SSVMs.

Projection for ranking and matching (Example 2 .2). In this setting, the projection using the entropy in $\mathcal{M}$ is known to be #P-complete (Valiant, 1979). Thus, CRFs are essentially intractable in this setting (Petterson et al., 2009). If instead we use the entropy $H(P) = -\sum_{i,j=1}^{M} P_{ij} \log P_{ij}$ defined over the marginals $P \in \mathcal{M}$, the projection can be computed up to precision δ in $O(M^2/\delta)$ iterations using the Sinkhorn-Knopp algorithm (Cuturi, 2013). This can be potentially much cheaper than the max-oracle of SSVMs, which has a cubic dependence in M. The projection with respect to the Euclidean norm has similar complexity but implementation is more involved (**?**).

Warm-starting the oracles. On the one hand, Algorithm 1 is guaranteed to converge as long as the error incurred in the oracle $\mathcal{O}$ decreases sublinearly with the number of global iterations as $\varepsilon_t \propto n/(t+n)$ (see Section E .1). On the other hand, Algorithm 2 can be naturally warm-started because it is an *any-time* algorithm as the step-size η does not depend on the current iteration or a finite horizon. Hence, we are in a setting where a warm-start strategy can be advantageous. More specifically, at every iteration t, we save the pairs $(\mu_i^\star, \nu_i^\star) \in \mathcal{O}(g_{w^{(t)}}(x_i))$ and the next time we revisit the i-th training example we initialize Algorithm 2 with this pair. Even though the formal demonstration of the effectiveness of the strategy is technically hard, we provide a strong experimental argument showing that a constant number of Algorithm 2 iterations are enough to match the allowed error ε_t.

Using the kernel trick. An extension to infinite-dimensional RKHS is straightforward to derive as Algorithm 1 is dual-based. In this case, the algorithm keeps track of the dual variables μ_i for $i = 1, \ldots, n$.

Connection to stochastic subgradient algorithms. It is known that (generalized) conditional gradient methods in the dual are formally equivalent to subgradient methods in the primal (Bach, 2015). Indeed, note that w_s in Algorithm 1 is a subgradient of the scaled surrogate loss $S(g_w(x_i), y_i)/\lambda n$. However, the dual-based analysis we provide in this paper allows us to derive guarantees on the expected duality gap and a line-search strategy, which we leave for future work. Viewing Algorithm 1 as a subgradient method is useful when learning the data representation with a neural network. More specifically, both Algorithm 1 and Algorithm 2 remain essentially unchanged by applying the chain rule in the update of w.

5 .4 Statistical Analysis of the Algorithm

Finally, the following Theorem 5 .2 shows that the full algorithm without the warm-start strategy achieves the same statistical error as the regularized ERM estimator (4.3) after at most $O(n\sqrt{n})$ projections oracle calls.

Theorem 5 .2 (Generalization bound of the algorithm)**.** *Assume the setting of Theorem 3 .4. Let $g_{n,T}$ be the T-th iteration of Algorithm 1 applied to problem* (4.3)*, where each iteration is computed with $K = O(\sqrt{n})$ iterations of Algorithm 2. Then, after $T = O(n)$ iterations, $g_{n,T}$ satisfies the bound of Theorem 3 .4 with probability $1 - \delta$.*

As we will show in the next section, in practice a constant number of iterations of Algorithm 2 are enough when using the warm-start strategy. Hence, the total number of required projection-oracles is $O(n)$.

6 Experiments

We perform a comparative experimental analysis for different tasks between $\mathrm{M^4Ns}$, $\mathrm{M^3Ns}$ and CRFs optimized with Generalized BCFW + SP-MP (Algorithm 1 + Algorithm 2), BCFW (Lacoste-Julien et al., 2013) and SDCA (Shalev-Shwartz and Zhang, 2013), respectively. All methods are run with our own implementation [2]. We use datasets of the UCI machine learning repository (Asuncion and Newman, 2007) for multi-class classification and ordinal regression, the OCR dataset from Taskar et al. (2004) for sequence prediction and the ranking datasets used by Korba et al. (2018). We use 14 random splits of the dataset into 60% for training, 20% for validation and 20% for testing. We choose the regularization parameter λ in $\{2^{-j}\}_{j=1}^{10}$ using the validation set and show the average test loss on the test sets in Table 4.1 of the model with the best λ. We use a Gaussian kernel and perform 50 passes on the data and set the number of iterations of Algorithm 2 to 20 and 10 times the length of the sequence for sequence prediction. The results are in Table 4.1. We perform better than $\mathrm{M^3Ns}$ in most of the datasets for multi-class classification, ordinal regression and ranking, while we obtain similar results in the sequence dataset with the three methods.

Effect of warm-start. We perform an experiment tracking the test loss and the average error in the max-min oracle for different iterations of Algorithm 2 with and without warm-starting. The experiments are done in two datasets for ordinal regression and they are shown in Table 2. We observe that both the test loss and average oracle error are lower for the warm-start strategy. Moreover, when warm-starting the final test error barely changes when increasing the iterations past the 50 iteration threshold.

7 Conclusion

In this paper, we introduced max-min margin Markov networks ($\mathrm{M^4Ns}$), a method for general structured prediction, that has the same algorithmic and theoretical properties as the regular binary SVM, that is, quantitative convergence bounds through a linear comparison inequality, as well as efficient optimization algorithms. Our experiments

[2] Code in `https://github.com/alexnowakvila/maxminloss`

Task	Dataset	(d, n, k)	$\mathrm{M^3N}$	CRF	$\mathrm{M^4N}$
MC	segment	(19, 2310, 7)	6.64%	6.43%	**6.09%**
	iris	(4, 150, 3)	3.33%	3.08%	3.33%
	wine	(13, 178, 3)	2.56%	2.14%	**2.35%**
	vehicle	(18, 846, 4)	24.6%	25.1%	**24%**
	satimage	(36, 4435, 6)	12.2%	11.5%	**11.9%**
	letter	(16, 15000, 26)	14.6%	13.2%	**13.5%**
	mfeat	(216, 2000, 10)	**3.94%**	4.35%	3.96%
ORD	wisconsin	(32, 193, 5)	**1.24**	1.26	1.26
	stocks	(9, 949, 5)	0.167	0.168	**0.160**
	machine	(6, 208, 10)	0.634	0.628	**0.628**
	abalone	(10, 4176, 10)	0.520	0.526	0.520
	auto	(7, 391, 10)	0.589	0.621	**0.585**
		(d, n, M)			
SEQ	ocr	(128, 6877, 26)	16.2%	16.3%	16.2%
RNK	glass	(9, 214, 6)	17.7%	-	**17.4%**
	bodyfat	(7, 252, 7)	79.6%	-	79.6%
	wine	(13, 178, 3)	5.06%	-	**4.34%**
	vowel	(10, 528, 11)	33.7%	-	**32.2%**
	vehicle	(18, 846, 4)	**14.8%**	-	15.0%

Table 4.1: Average test losses on the 14 splits for multi-class classification (first), ordinal regression (second), sequence prediction (third) and ranking (forth). We show in percentage the losses for multi-class, sequence prediction and ranking since they are between zero and one. We show in bold the lowest test loss between the direct classifiers $\mathrm{M^3N}$ and $\mathrm{M^4N}$.

Dataset	W-S	$K = 10$	$K = 30$	$K = 50$	$K = 100$
machine	yes	0.42 / 0.57	0.41 / 0.43	0.41 / 0.22	0.41 / 0.13
	no	0.50 / 4.41	0.50 / 2.74	0.44 / 1.25	0.42 / 0.63
auto	yes	0.56 / 1.55	0.55 / 1.29	0.51 / 0.81	0.50 / 0.44
	no	0.61 / 2.66	0.57 / 1.79	0.53 / 0.89	0.51 / 0.47

Table 4.2: We show the (final ordinal test loss / average oracle error at the last epoch) for $\mathrm{M^4N}$s trained with 100 passes on data with different iterations of Algorithm 2 with and without warm-starting.

show its performance on classical structured prediction problems when using RKHS hypothesis spaces. It would be interesting to extend the analysis of the proposed algorithm by rigorously proving the linear dependence in the number of samples when using the warm-start strategy and incorporating a line-search strategy.

Appendices

Notation on the structured prediction setting. Denote by $\mathcal{P}(A)$ the set of subsets of the set A. We define the following quantities

- *Marginal polytope*: $\mathcal{M} = \text{hull}(\varphi(\mathcal{Y})) = \{v \in \mathbb{R}^k \mid v = \sum_{y \in \mathcal{Y}} \alpha_y \varphi(y), \alpha \in \Delta_{\mathcal{Y}}\}$.
- *Normal cone of $\mathcal{M}$ at μ*: $\mathcal{N}_{\mathcal{M}}(\mu) = \{u \in \mathbb{R}^k \mid \langle \mu' - \mu, u \rangle \leq 0, \forall \mu' \in \mathcal{M}\}$.
- *Conditional moments*: $\nu(q) = \mathbb{E}_{y' \sim q}\, \varphi(y') \in \mathcal{M}$ where $q \in \text{Prob}(\mathcal{Y})$.
- *Conditional risk*: $\ell(y, \mu) \coloneqq \mathbb{E}_{y' \sim q}\, L(y, y') = \varphi(y)^\top A \mu$.
- *Bayes risk*: $\ell(\mu) \coloneqq \min_{y \in \mathcal{Y}} \ell(y, \mu)$
- *Minus Bayes risk*: $\Omega(\mu) \coloneqq -\ell(\mu) + 1_{\mathcal{M}}(\mu)$.
- *Excess conditional risk*: $\delta\ell(y, \mu) = \ell(y, \mu) - \ell(\mu)$
- *Optimal predictor set*: $y^\star(\mu) = \arg\min_{y \in \mathcal{Y}}\ \ell(y, \mu) \subseteq \mathcal{Y}$.
- *Marginal polytope cell complex*: $\mathcal{C}(\mathcal{M}) = (y^\star)^{-1} \circ y^\star(\mathcal{M}) \subset \mathcal{P}(\mathcal{M})$.

Notation on the max-min surrogate.

- *Partition function*: $\Omega^*(v) = \max_{\mu \in \mathcal{M}} \ell(\mu) + v^\top \mu$.
- *Surrogate loss*: $S(v, y) = \Omega^*(v) - v^\top \varphi(y)$.
- *Conditional surrogate risk*: $s(v, \mu) \coloneqq \mathbb{E}_{y \sim q}\, S(v, y) = \Omega^*(v) - v^\top \mu$.
- *Bayes surrogate risk*: $s(\mu) \coloneqq \min_{v \in \mathbb{R}^k} s(v, \mu) \quad (= \ell(\mu))$.
- *Excess surrogate conditional risk*: $\delta s(v, \mu) \coloneqq s(v, \mu) - s(\mu)$.
- *Optimal predictors*: $v^\star(\mu) = \arg\min_{v \in \mathbb{R}^k}\ s(v, \mu) \subset \mathbb{R}^k$.
- *Surrogate space cell complex*: $\mathcal{C}(\mathbb{R}^k) = v^\star(\mathcal{M}) \subset \mathcal{P}(\mathbb{R}^k)$.

A Geometrical Properties

In this section, we study the rich geometrical properties of the max-min surrogate construction. The geometrical interpretation provides a valuable intuition on different key mathematical objects appearing in the further analysis needed for the proofs of the main theorems. More precisely, we show that the max-min surrogate loss S defines a partition $\mathcal{C}(\mathbb{R}^k) \subset \mathcal{P}(\mathbb{R}^k)$ of the surrogate space $\mathbb{R}^k$ which is dual to the partition $\mathcal{C}(\mathcal{M}) \subset \mathcal{P}(\mathcal{M})$ of the marginal polytope $\mathcal{M}$ defined by L. Moreover, we show that the mapping between those partitions is the subgradient mapping $\partial\Omega$ with inverse $\partial\Omega^*$ (see Figure 4.2). Visualization for binary 0-1 loss, multi-class 0-1 loss, absolute loss for ordinal regression and Hamming loss are provided in Section A .4.

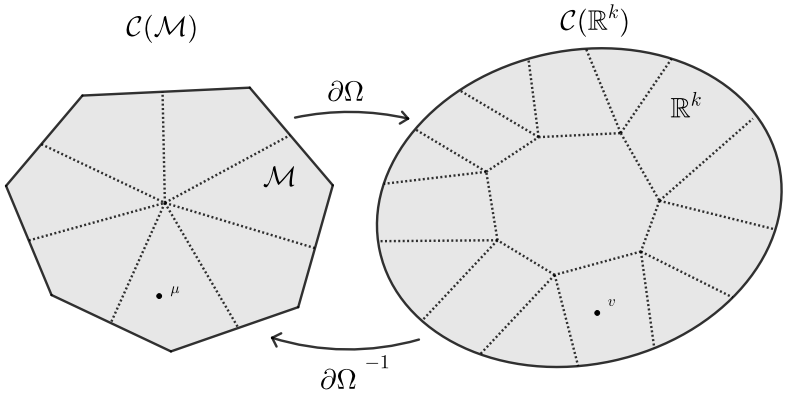


Figure 4.2: The cell complex $\mathcal{C}(\mathcal{M})$ on the marginal polytope $\mathcal{M}$ maps to the cell complex $\mathcal{C}(\mathbb{R}^k)$ on the surrogate space $\mathbb{R}^k$ through the subgradient mapping of the partition function $\partial\Omega$.

Following Finocchiaro et al. (2019), we now introduce the definition of a *cell complex*.

Definition A .1 (Cell complex)**.** *A cell complex in $\mathbb{R}^k$ is a set $\mathcal{C}$ of faces (of dimension $0, \ldots, k$) such that:*

(i) union to $\mathbb{R}^k$.

(ii) have pairwise disjoint relative interiors.

(iii) any nonempty intersection of faces F, F' in $\mathcal{C}$ is a face of F and F' and an element of $\mathcal{C}$.

Any convex affine-by-parts function has an associated cell complex defined by considering the polytope corresponding to the epigraph of the function and projecting the faces down to to the domain. Moreover, if $f(v)$ is convex affine-by-parts, the cell complex associated to $f(v)$ and $f(v) + v^\top a$ are the same for any a.

A .1 Geometry of the Loss L

The convex affine-by-parts function $\Omega(\mu) = -\ell(\mu) + 1_{\mathcal{M}}(\mu)$ naturally defines a cell complex of its domain $\mathcal{M}$ (see for instance Ramaswamy and Agarwal (2016); Nowak-Vila et al. (2019)). This can be constructed by considering the polyhedra corresponding to the epigraph of $-\ell(\mu)$ and then projecting the faces to $\mathcal{M}$. Each face corresponds to a

different group of active hyperplanes in the definition of $\ell(\mu)$. The cell complex can be defined as $\mathcal{C}(\mathcal{M}) = \{\overline{(y^\star)^{-1} \circ y^\star(\mu)} \mid \mu \in \mathcal{M}\} \subset \mathcal{P}(\mathcal{M})$, i.e., each face is defined as the set of moments that share the same set of optimal predictors. Note that $\mathcal{C}(\mathcal{M})$ contains faces of 0-dimensions (points) up to faces of k-dimensions.

A .2 Geometry of the Loss S

Recall that $S(v, y) = \Omega^*(v) - v^\top \mu$ and $\Omega(\mu) = -\ell(\mu) + 1_{\mathcal{M}}(\mu)$, where $\ell(\mu)$ is concave affine-by-parts. In particular, as Ω is convex affine-by-parts with compact domain, then Ω^* is convex affine-by-parts with full-dimensional domain $\mathbb{R}^k$. The projection of the faces of the convex polyhedron defined as the epigraph of Ω^* defines a cell complex $\mathcal{C}(\mathbb{R}^k) \subset \mathcal{P}(\mathbb{R}^k)$ in the (unbounded) vector space $\mathbb{R}^k$. The cell complex defined by $\Omega^*(v)$ is the same as the one defined by $s(v, \mu) = \Omega^*(v) - v^\top \mu$ for every $\mu \in \mathcal{M}$. The faces of $\mathcal{C}(\mathbb{R}^k)$ can be written as $v^\star(\mu) = \arg\min_{v \in \mathbb{R}^k} \; s(v, \mu) \in \mathcal{P}(\mathbb{R}^k)$ for a certain $\mu \in \mathcal{M}$, i.e., the faces are the minimizers of the conditional surrogate risk. Hence, we can write in a compact form $v^\star(\mu) \in \mathcal{C}(\mathbb{R}^k)$.

A .3 Relation between Cell Complexes

Recall that $\mathcal{C}(\mathcal{M})$ is generated by projecting the faces of the epigraph of Ω while $\mathcal{C}(\mathbb{R}^k)$ is generated by projecting the faces of the epigraph of Ω^*. The subgradients are well-defined in the cell complexes and define a bijection between them:

$$\partial\Omega : \mathcal{C}(\mathcal{M}) \to \mathcal{C}(\mathbb{R}^k), \qquad \partial\Omega^* : \mathcal{C}(\mathbb{R}^k) \to \mathcal{C}(\mathcal{M}), \qquad \partial\Omega^* = (\partial\Omega)^{-1}.$$

Moreover, we have that

$$\begin{aligned} \dim(\partial\Omega(F)) &= \dim(\mathcal{M}) - \dim(F), && \forall F \in \mathcal{C}(\mathcal{M}) \\ \dim(\partial\Omega^*(F')) &= \dim(\mathcal{M}) - \dim(F'), && \forall F' \in \mathcal{C}(\mathbb{R}^k), \end{aligned}$$

where F, F' are faces of $\mathcal{C}(\mathcal{M}), \mathcal{C}(\mathbb{R}^k)$, respectively.

A .4 Examples

Let's now provide some concrete examples of cell complexes and the associated mapping subgradient mapping for several classical tasks.

Binary Classification. The output space is $\mathcal{Y} = \{-1, 1\}$. The loss is $L(y, y') = 1(y \neq y')$ with affine decomposition $a = 1$, $\varphi(1) = (1, 0)^\top$, $\varphi(-1) = (0, 1)^\top$ and $A = -I_{2\times 2}$. The marginal polytope is $\mathcal{M} = \Delta_2$.

$$\Omega(p) = -\min(p, 1-p) + 1_{\Delta_2}(p), \quad \Omega^*(v) = \max(|v|, 1/2), \quad S(v, y) = \Omega^*(v) - yv.$$

See Figure 4.3. The mapping between cells is

$$\begin{aligned} &\partial\Omega(\{0\}) = (-\infty, -1/2], && \partial\Omega^*((-\infty, -1/2]) = \{0\} \\ &\partial\Omega(\{1/2\}) = [-1/2, 1/2], && \partial\Omega^*([-1/2, 1/2]) = \{1/2\} \\ &\partial\Omega(\{1\}) = [1/2, +\infty), && \partial\Omega^*([1/2, +\infty)) = \{1\} \\ &\partial\Omega([0, 1/2]) = \{-1/2\}, && \partial\Omega^*(\{-1/2\}) = [0, 1/2] \\ &\partial\Omega([1/2, 1]) = \{1/2\}, && \partial\Omega^*(\{1/2\}) = [1/2, 1] \end{aligned}$$

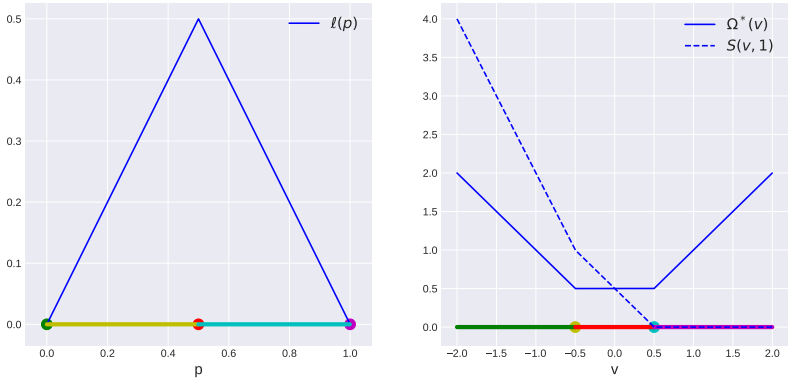


Figure 4.3: Binary 0-1 loss. **Left:** The Bayes risk $-\Omega$ is a concave polyhedral function defined in $\Delta_2 = [0,1]$. The faces of the induced cell-complex are the 0-dimensional faces $\{0\}, \{1/2\}, \{1\}$ and the 1-dimensional faces $[0,1/2]$, $[1/2,1]$. **Right:** The partition function is a convex polyhedral function defined in $\mathbb{R}$. The faces of the induced cell-complex are the 0-dimensional faces $\{-1/2\}$, $\{1/2\}$ and the 1-dimensional faces $(-\infty,-1/2]$, $[1/2,1/2]$, $[1/2,+\infty)$.

0-1 Multi-class Classification. The output space is $\mathcal{Y} = [k] = \{1,\cdots,k\}$. The loss is $L(y,y') = 1(y \neq y')$ with affine decomposition $a = 1$, $\varphi(y) = e_y$ and $A = -I_{k\times k}$, where $e_y \in \mathbb{R}^k$ is the y-th vector of the canonical basis in $\mathbb{R}^k$. The marginal polytope is the k-dimensional simplex $\mathcal{M} = \Delta_k$.

In this case, $\ell(p) = 1 - \|p\|_\infty$ and so

$$\Omega(p) = \|p\|_\infty - 1 + 1_{\Delta_k}(p), \qquad \Omega^*(v) = 1 + \max_{j\in[k]} \left\{ \frac{1}{j} \sum_{r=1}^{j} v_{(r)} - \frac{1}{j} \right\},$$

where $v_{(1)} \geq \cdots \geq v_{(k)}$ (see Figure 4.4).

Note that the subgradient mapping $\partial\Omega$ sends the 2^k 0-dimensional faces (points) and the full-dimensional faces of $\mathcal{C}(\Delta_k)$ to the full-dimensional faces and 0-dimensional faces of $\mathcal{C}(\mathbb{R}^{k-1})$, respectively.

Ordinal Regression. The output space is the same as for multiclass classification, but in this case there is an implicit ordering between outputs: $1 \prec 2 \prec \cdots \prec k$. This is encoded using the absolute difference loss $L(y,y') = |y - y'|$. We consider the affine decomposition $\varphi(y) = e_y \in \mathbb{R}^k$, $A = (|i-j|)_{i,j\in[k]^2}$ and $a = 0$. It is possible to obtain a closed form expression for the partition function (see Thm. 6 by Fathony et al. (2018a)):

$$\Omega^*(v) = \frac{1}{2} \max_{i,j\in[k]} v_i + v_j + j - i.$$

In Figure 4.5 we plot the Bayes risk and the partition function for the ordinal loss. Note that that the topology of the cell-complex is different from the previous example.

Multi-label Classification with Hamming Loss. This corresponds to Example 2 .1 with unary potentials. Let $\mathcal{Y} = \Pi_{m=1}^M \mathcal{Y}_m$ with $\mathcal{Y}_m = \{1,\ldots,R\}$. We consider the Hamming loss defined as an average of multi-class losses: $L(y,y') = \frac{1}{M}\sum_{m=1}^M 1(y_m \neq y'_m)$. The marginal polytope factorizes as $\mathcal{M} = \Pi_{m=1}^M \Delta_R$. The Bayes risk decomposes additively as

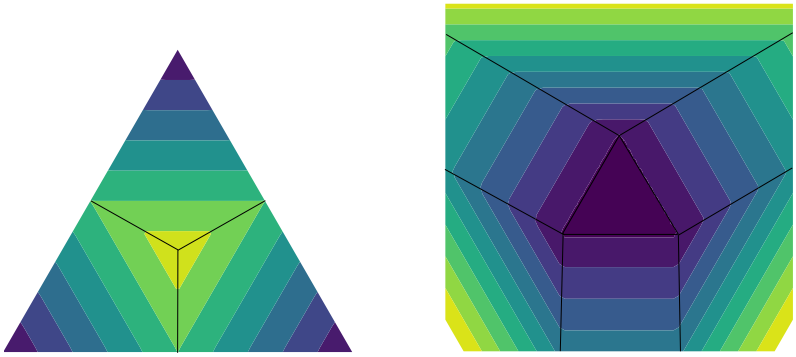


Figure 4.4: Multi-class 0-1 loss ($k = 3$). **Left:** The Bayes risk $-\Omega$ has a pyramid shape centered at the simplex. The cell-complex $\mathcal{C}(\Delta_k)$ has 2^k 0-dimensional faces (points) and k full-dimensional faces. In the figure, the set of points are the center point, the 3 vertices of the triangle and the 3 middle points in the triangle face. The 3 full-dimensional faces are the 3 colored zones. **Right:** The partition function Ω^* is a convex polyhedral function. The cell-complex has k 0-dimensional faces (points) and 2^k full-dimensional faces. In the figure, the set of points are the 3 vertices of the triangle in the center and the full-dimensional faces are the colored zones.

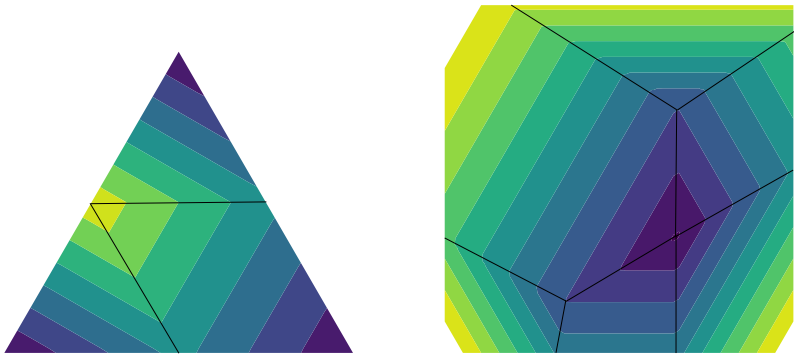


Figure 4.5: Absolute difference loss ($k = 3$). **Left:** The Bayes risk $-\Omega$ has an asymmetrical pyramid shape with the tip in one face of the simplex. **Right:** The partition function Ω^* has a different topology than the one from multi-class.

the sum of the Bayes risks of the individual multi-class losses and the partition function decomposes analogously. In Figure 4.6 we plot the Bayes risk and the partition function for $R = M = 2$.

B Theoretical Properties of the Surrogate

The goal of this section is to prove the two theoretical requirements for the surrogate method. These are *Fisher consistency (1)* and a *comparison inequality (2)*:

$$\begin{aligned} &(1) \quad \mathcal{E}(f^\star) = \mathcal{E}(d \circ g^\star) \\ &(2) \quad \zeta(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star). \end{aligned}$$

for all measurable $g : \mathcal{X} \to \mathcal{H}$, where $\zeta : \mathbb{R} \to \mathbb{R}$ is such that $\zeta(\varepsilon) \to 0$ if $\varepsilon \to 0$. Fisher consistency ensures that the optimum of the surrogate loss $g^\star$ provides the Bayes

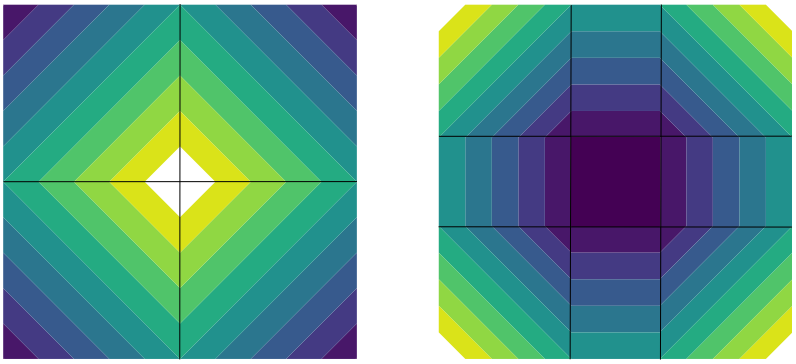


Figure 4.6: Hamming loss for $(R = M = 2)$. **Left:** The marginal polytope is the cube $\mathcal{M} = [0,1]^2$ and the Bayes risk $-\Omega$ has a pyramid shape centered in the cube. **Right:** The partition function Ω^*.

optimum $f^\star$ of the problem. However, in practice the optimum of the surrogate is never attained and so one wants to control how close $f = d \circ g$ is to $f^\star$ in terms of the estimation error of g to $g^\star$. The comparison inequality gives this quantification by relating the excess expected risk $\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)$ to the excess expected surrogate risk $\mathcal{R}(g) - \mathcal{R}(g^\star)$, which allows to translate rates from the surrogate problem to the original problem.

Let's start first by showing that $s(\mu) = \ell(\mu)$ for all $\mu \in \mathcal{M}$, i.e., that the minimizers of the conditional surrogate risks coincide.

Lemma B .1. *The Bayes risk and the surrogate Bayes risk are the same:*

$$s(\mu) = \min_{v \in \mathbb{R}^k} \; s(v, \mu) = \min_{y' \in \mathcal{Y}} \; \ell(y, \mu) = \ell(\mu), \qquad \forall \mu \in \mathcal{M}.$$

Proof. Note that $s(\mu) = \min_{v \in \mathbb{R}^k} \; s(v, \mu) = \min_{v \in \mathbb{R}^k} \; \Omega^*(v) - v^\top \mu = -\Omega = \ell(\mu) - 1_{\mathcal{M}}(\mu)$. □

Note that this is not the case for smooth surrogates. It was noted by Finocchiaro et al. (2019) (see Prop. 1 and 2) that consistent polyhedral surrogates necessarily satisfy the property of matching Bayes risks.

B .1 Fisher Consistency

The following Proposition B .2 characterizes the form of the exact minimizer of the conditional surrogate risk $s(v, \mu)$.

Proposition B .2. *Let $\mu \in \mathcal{M}$ and $y^\star(\mu) = \arg\min_{y \in \mathcal{Y}} \varphi(y)^\top A \mu$ be the set of optimal predictors. Then, we have that*

$$\operatorname{hull}(-A^\top \varphi(y))_{y \in y^\star(\mu)} + \mathcal{N}_{\mathcal{M}}(\mu) = \arg\min_{v \in \mathbb{R}^k} s(v, \mu). \tag{4.16}$$

Proof. The proof consists in noticing that $\operatorname{hull}(-A^\top \varphi(y))_{y \in y^\star(\mu)}$ is a subgradient at μ of the non-smooth convex function $\Omega = -\ell(\mu) + 1_{\mathcal{M}}(\mu)$ with compact domain $\mathcal{M}$. That is,

$$\Omega(\mu) = -\min_{y' \in \mathcal{Y}} \varphi(y')^\top A \mu + 1_{\mathcal{M}}(\mu)$$

$$= -\varphi(y)^\top A\mu + 1_{\mathcal{M}}(\mu), \qquad y \in y^\star(\mu).$$

The subgradient reads $\partial\Omega(\mu) = -\operatorname{hull}(A^\top\varphi(y))_{y\in y^\star(\mu)} + \mathcal{N}_{\mathcal{M}}(\mu)$, where $\mathcal{N}_{\mathcal{M}}(\mu)$ is the normal cone of $\mathcal{M}$ at the point μ. Then, using Fenchel duality we have that

$$\partial\Omega(\mu) = \operatorname*{arg\,min}_{v\in\mathbb{R}^k} \Omega^*(v) - v^\top\mu = \operatorname*{arg\,min}_{v\in\mathbb{R}^k} s(v,\mu).$$

□

Let $\rho(\cdot|x)$ be the conditional distribution of outputs and $\mu(x) = \mathbb{E}_{y\sim\rho(\cdot|x)}\varphi(y)$. If we assume that the set of points $x \in \mathcal{X}$ for which $\mathcal{N}_{\mathcal{M}}(\nu(x)) \neq \{0\}$ has measure zero, then we have that $g^\star(x) \in -\operatorname{hull}(A^\top\varphi(y))_{y\in y^\star(\rho(\cdot|x))}$ almost-surely. Thus, we can write $g^\star(x) = -\sum_{y\in y^\star(\rho(\cdot|x))}\alpha_y A^\top\varphi(y)$ with $\sum_{y\in y^\star(\rho(\cdot|x))}\alpha_y = 1$. We have Fisher consistency as

$$f^\star(x) \in \operatorname*{arg\,max}_{y'\in\mathcal{Y}} \varphi(y')^\top g^\star(x) = \operatorname*{arg\,min}_{y'\in\mathcal{Y}} \sum_{y\in y^\star(\rho(\cdot|x))} \alpha_y L(y',y).$$

B .2 Comparison Inequality and Calibration Function

The goal of this section is to explicitly compute a comparison inequality. We will show that the relation between both excess risks is linear and that the constants appearing scale nicely with the natural dimension of the structured problem and not with the total number of possible outputs $|\mathcal{Y}|$ which can potentially be exponential.

The main object of study will be the so-called *calibration function*, which is defined as the 'worst' comparison inequality between both excess conditional risks.

Definition B .3 (Calibration function (Steinwart, 2007))**.** *The calibration function $\zeta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is defined for $\varepsilon \geq 0$ as the infimum of the excess conditional surrogate risk when the conditional risk is at least ε:*

$$\zeta(\varepsilon) = \inf \delta s(v,\mu) \quad \textit{such that} \quad \delta\ell(d\circ v,\mu) \geq \varepsilon,\ \mu\in\mathcal{M},\ v\in\mathbb{R}^k.$$

We set $\zeta(\varepsilon) = \infty$ when the feasible set is empty.

Note that ζ is non-decreasing in $[0,+\infty)$, not necessarily convex (see Example 5 by Bartlett et al. (2006)) and also $\zeta(0) = 0$. Note that a larger ζ is better because we want a large $\delta s(v,\mu)$ to incur small $\delta\ell(d\circ v,\mu)$. The following Theorem B .4 justifies Definition B .3.

Theorem B .4 (Comparison inequality in terms of calibration function (Steinwart, 2007))**.** *Let $\bar{\zeta}$ be a convex lower bound of ζ. We have*

$$\bar{\zeta}(\mathcal{E}(d\circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star) \tag{4.17}$$

for all $g : \mathcal{X} \to \mathbb{R}^k$. The tightest convex lower bound $\bar{\zeta}$ of ζ is its lower convex envelope which is defined by the Fenchel bi-conjugate ζ^{**}.

Proof. Note that by the definition of the calibration function, we have that

$$\zeta(\delta\ell(d\circ g(x),\mu(x))) \leq \delta s(g(x),\mu(x)), \tag{4.18}$$

where $\mu(x) = \mathbb{E}_{y' \sim \rho(\cdot|x)} \varphi(y')$. The comparison between risks is then a consequence of Jensen's inequality:

$$\begin{aligned}
\bar{\zeta}(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) &= \bar{\zeta}(\mathbb{E}_{x \sim \rho_\mathcal{X}} \delta\ell(d \circ g(x), \mu(x))) \\
&\leq \mathbb{E}_{x \sim \rho_\mathcal{X}} \bar{\zeta}(\delta\ell(d \circ g(x), \mu(x))) && \text{(Jensen ineq.)} \\
&\leq \mathbb{E}_{\sim \rho_\mathcal{X}} \zeta(\delta\ell(d \circ g(x), \mu(x))) && (\bar{\zeta} \leq \zeta) \\
&\leq \mathbb{E}_{x \sim \rho_\mathcal{X}} \delta s(g(x), \mu(x)) \\
&= \mathcal{R}(g) - \mathcal{R}(g^\star).
\end{aligned}$$

□

B .3 Characterizing the Calibration Function for Max-Min Margin Markov Networks

Following Osokin et al. (2017), we write the calibration function in terms of pairwise interactions.

Lemma B .5 (Lemma 10)**.** *We can re-write the calibration function $\zeta(\varepsilon)$ as*

$$\zeta(\varepsilon) = \min_{y \neq y'} \zeta_{y,y'}(\varepsilon),$$

where

$$\zeta_{y,y'}(\varepsilon) = \left\{ \begin{array}{lll} \min_{v,\mu \in \mathbb{R}^k} & \delta s(v, \mu) & \\ \textit{s.t} & \delta\ell(y', \mu) \geq \varepsilon & (\varepsilon - \textit{suboptimality}) \\ & y = y^\star(\mu) & (\textit{optimal prediction}) \\ & y' = d \circ v & (\textit{prediction}) \\ & \mu \in \mathcal{M} & \end{array} \right. \tag{4.19}$$

Proof. The idea of the proof is to decompose the feasibility set of the optimization problem into a union of sets enumerated by the pairs (y, y') corresponding to the optimal prediction y and the prediction y'. Let's first define the sets $V(y) \subset \mathbb{R}^k$ and $\mathcal{M}_{y,y',\varepsilon} \subset \mathcal{M}$.

1. Define the prediction sets as $V(y) \coloneqq \{v \in \mathbb{R}^k \mid v^\top(\varphi(y) - \varphi(y')) > 0, \forall y' \in \mathcal{Y}\} \subset \mathbb{R}^k$ to denote the set of elements in the surrogate space $\mathbb{R}^k$ for which the prediction is the output element $y \in \mathcal{Y}$. Note that the sets $V(y)$ do not contain their boundary, but their closure can be expressed as

$$\overline{V}(y') \coloneqq \{v \in \mathbb{R}^k \mid v^\top(\varphi(y') - \varphi(y)) \geq 0, \forall y \in \mathcal{Y}\}.$$

   Note that $\bigcup_{y' \in \mathcal{Y}} \overline{V}(y') = \mathbb{R}^k$.
2. If $v \in V(y')$, the feasible set of conditional moments μ for which output y is one of the best possible predictions (i.e., $\ell(y', \mu) - \ell(y, \mu) \geq \varepsilon$) is

$$\mathcal{M}_{y,y',\varepsilon} = \{\mu \in \mathcal{M} \mid \ell(y, \mu) = \ell(\mu) \mid \ell(y', \mu) - \ell(y, \mu) \geq \varepsilon\}.$$

The union of the sets $\{\overline{V}(y') \times \mathcal{M}_{y,y',\varepsilon}\}_{y,y' \in \mathcal{Y}}$ exactly equals the feasibility set of the optimization problem Definition B .3. We can then re-write the calibration function as

$$\zeta(\varepsilon) = \min_{y \neq y'} \left\{ \begin{array}{ll} \min_{v,\mu} & \delta s(v, \mu) \\ \text{s.t.} & v \in V(y') \\ & \mu \in \mathcal{M}_{y,y',\varepsilon} \end{array} \right. . \tag{4.20}$$

Finally, by Lemma 27 of Zhang (2004a), the function $\delta s(v, \mu)$ is continuous w.r.t both μ and v, allowing to substitute the sets $V(y')$ in Eq. (4.20) by their closures $\overline{V}(y')$ without changing the minimum. □

Until now, the results were general for any calibration function. We will now construct a lower bound on the calibration function for $\mathrm{M}^4\mathrm{Ns}$. Let's first introduce some notation.

Notation.

- Let $\mathcal{M}_0$ be the finite set of 0-dimensional faces (points) of the cell complex $\mathcal{C}(\mathcal{M}) \subset \mathcal{P}(\mathcal{M})$, or equivalently (mapped by $\partial\Omega$), the full dimensional faces of the cell complex $\mathcal{C}(\mathbb{R}^k) \subset \mathcal{P}(\mathbb{R}^k)$. Note that $|\mathcal{M}_0|$ is finite.
- Let $w(y) = -A^\top \varphi(y)$.

Recall that in Lemma B .5 we split the optimization problem into $|\mathcal{Y}|(|\mathcal{Y}| - 1)$ optimization problems corresponding to all possible (ordered) pairs of different optimal prediction and prediction. The following Theorem B .6 further splits the inner optimization problems into some faces of the cell complex $\mathcal{C}(\mathbb{R}^k)$ and simplifies the objective function into an affine function.

Theorem B .6 (Calibration Function for the surrogate loss of $\mathrm{M}^4\mathrm{N}$)**.** *We have that*

$$\zeta_{y,y'}(\varepsilon) = \min_{\bar{\mu} \in \mathcal{M}_0(y,y')} \zeta_{y,y',\bar{\mu}}(\varepsilon),$$

where $\mathcal{M}_0(y, y') = \{\bar{\mu} \in \mathcal{M}_0 \mid w(y), w(y') \in \partial\Omega(\bar{\mu})\} \subseteq \mathcal{M}_0 \subset \mathcal{M}$*, and*

$$\zeta_{y,y',\bar{\mu}}(\varepsilon) = \begin{cases} \min\limits_{v,\mu \in \mathbb{R}^k} & \langle w(y) - v, \mu - \bar{\mu} \rangle & & \\ s.t & \langle w(y) - w(y'), \mu \rangle \geq \varepsilon & & (\varepsilon - \textit{suboptimality}) \\ & \langle w(y) - w(z), \mu \rangle \geq 0, & \forall z \in \mathcal{Y} & (\textit{optimal prediction}) \\ & y' = d \circ v & & (\textit{prediction}) \\ & v \in \partial\Omega(\bar{\mu}) & & (\textit{face in } \mathcal{C}(\mathbb{R}^k)) \\ & \mu \in \mathcal{M} & & \end{cases} \tag{4.21}$$

Proof. We split the proof into three steps. First, we split the optimization problem w.r.t $v \in \mathbb{R}^k$ among the faces $\partial\Omega(\mathcal{M}_0)$ of the complex cell $\mathcal{C}(\mathbb{R}^k) \subset \mathcal{P}(\mathbb{R}^k)$. Second, we show that the minimizer is achieved in a face $\partial\Omega(\bar{\mu})$ such that $w(y), w(y') \in \partial\Omega(\bar{\mu})$ and simplify the objective function. Finally, we update the notation of some constraints.

1st step. Split the optimization problem according to the affine parts. Recall that $s(v, \mu)$ is defined as a supremum of affine functions, where each affine function corresponds to a $\bar{\mu} \in \mathcal{M}_0$: $s(v, \mu) = \sup_{\bar{\mu} \in \mathcal{M}_0} \ell(\bar{\mu}) + v^\top(\bar{\mu} - \mu)$. Using that $\bigcup_{\bar{\mu} \in \mathcal{M}_0} \overline{\partial\Omega(\bar{\mu})} = \mathbb{R}^k$ and the continuity of the loss, we split problem (4.19) into $|\mathcal{M}_0|$ minimization problems and define

$$\zeta_{y,y'}(\varepsilon) = \min_{\bar{\mu} \in \mathcal{M}_0} \zeta_{y,y',\bar{\mu}}(\varepsilon),$$

where $\zeta_{y,y',\bar{\mu}}(\varepsilon)$ is given by problem (4.19) with the additional constraint $v \in \partial\Omega(\bar{\mu})$.

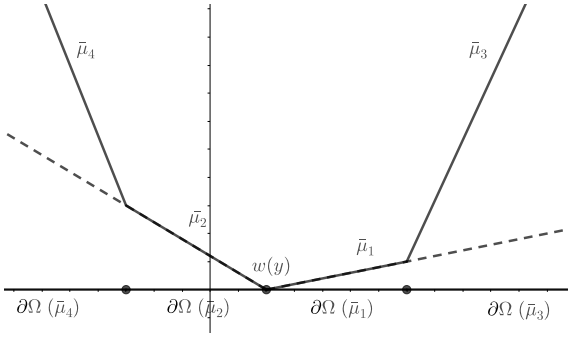


Figure 4.7: The minimizer of the objective $\delta s(v, \mu)$ over $\mathbb{R}^k$ is $w(y)$. In order to minimize the objective under the constraints, we only need to consider the faces which include the minimizer $w(y)$. In the figure above, we can safely remove the optimization over the faces $\partial\Omega(\bar{\mu}_4)$ and $\partial\Omega(\bar{\mu}_3)$.

2nd step. Reduce the number of considered affine parts and simplify objective. We will show that

$$\min_{\bar{\mu}\in\mathcal{M}_0} \zeta_{y,y',\bar{\mu}}(\varepsilon) = \min_{\bar{\mu}\in\mathcal{M}_0(y,y')} \zeta_{y,y',\bar{\mu}}(\varepsilon),$$

where $\mathcal{M}_0(y, y') = \{\bar{\mu} \in \mathcal{M}_0 \mid w(y), w(y') \in \partial\Omega(\bar{\mu})\} \subset \mathcal{M}_0$. Moreover, when $\bar{\mu} \in \mathcal{M}_0(y, y')$, the objective function in the definition of $\zeta_{y,y',\bar{\mu}}(\varepsilon)$ takes the affine form $\langle w(y) - v, \mu - \bar{\mu}\rangle$. In order to see this, let's make the following observations.

- We have that $w(y')$ must belong to the feasibility set as $y' = d \circ w(y')$. And so, we must have $w(y') \in \partial\Omega(\bar{\mu})$.
- Fix μ in the feasibility set of Eq. (4.19). As y is the optimal prediction, $w(y)$ is a minimizer of the conditional surrogate risk: $\min_{v'} \; s(v', \mu) = s(w(y), \mu)$. The objective function $s(v, \mu) - s(w(y), \mu) \geq 0$ is a convex affine-by-parts function with minimizer $w(y)$. We can lower bound this quantity by simply considering the affine parts $\bar{\mu} \in \mathcal{M}_0$ that include the minimizer, i.e., $w(y) \in \partial\Omega(\bar{\mu})$ (see Figure 4.7). Moreover, note that if $w(y) \in \partial\Omega(\bar{\mu})$, then $\Omega^*(v) = \Omega^*(w(y)) + \langle \bar{\mu}, v - w(y)\rangle$, as $\bar{\mu}$ is the slope of the affine part $\partial\Omega(\bar{\mu})$. Using that $\min_{v'\in\mathbb{R}^k} s(v', \mu) = s(w(y), \mu) = \Omega^*(w(y)) - \langle w(y), \mu\rangle$, we have that

$$\begin{aligned}\delta s(v, \mu) &= s(v, \mu) - s(w(y), \mu) \\ &= \Omega^*(v) - \Omega^*(w(y)) + \langle w(y) - v, \mu\rangle \\ &= \langle w(y) - v, \mu - \bar{\mu}\rangle.\end{aligned}$$

3rd step. Re-write constraints in terms of $w(y)$. The constraint $y = y^\star(\mu) = \arg\min_{y\in\mathcal{Y}} \varphi(y)^\top A\mu$ is equivalent to $\ell(z, \mu) - \ell(y, \mu) \geq 0$ for all $z \in \mathcal{Y}$, which can be written $\delta\ell(z, \mu) = \langle w(y) - w(z), \mu\rangle$, for all $z \in \mathcal{Y}$. Similarly, the constraint $\delta\ell(y', \mu) \geq \varepsilon$ reads $\langle w(y) - w(y'), \mu\rangle \geq \varepsilon$. □

In order to state Theorem B .7, let's first define the function $\lambda^\mu_{y'} : \partial\Omega(\mu) \to \mathbb{R}_{\geq 0}$. By Proposition B .2, we know that $\partial\Omega(\mu) = \text{hull}(w(y))_{y\in y^\star(\mu)} + \mathcal{N}_\mathcal{M}(\mu)$. In general, there exist multiple ways to describe a vector $v \in \partial\Omega(\mu)$ as $v = \sum_{y\in y^\star(\mu)} \lambda_y w(y) + n$ with $\lambda \in \Delta_\mathcal{Y}$ and $n \in \mathcal{N}_\mathcal{M}(\mu)$. The function $v \mapsto \lambda^\mu_{y'}(v)$ is defined as the maximal weight of

the vector $w(y')$ over all possible decompositions:

$$\lambda^{\mu}_{y'}(v) = \left\{ \begin{array}{ll} \max\limits_{\lambda,n} & \lambda_{y'} \\ \text{s.t} & v = \sum_{y \in y^\star(\mu)} \lambda_y w(y) + n \\ & \lambda \in y^\star(\mu) \\ & n \in \mathcal{N}_{\mathcal{M}}(\mu) \end{array} \right. . \tag{4.22}$$

The following Theorem B .7 gives a *constant positive* lower bound of the ratio $\zeta(\varepsilon)/\varepsilon$ as a minimization of $\lambda^{\bar{\mu}}_{y'}(v)$ over the prediction set of y'.

Theorem B .7. *We have that*

$$\zeta(\varepsilon) = \min_{y' \in \mathcal{Y}} \min_{\bar{\mu} \in \mathcal{M}_0(y')} \zeta_{y',\bar{\mu}}(\varepsilon)$$

, where $\mathcal{M}_0(y') = \{\bar{\mu} \in \mathcal{M}_0 \mid w(y') \in \partial\Omega(\bar{\mu})\} \subseteq \mathcal{M}_0 \subset \mathcal{M}$ *and*

$$\zeta_{y',\bar{\mu}}(\varepsilon)/\varepsilon \geq \left\{ \begin{array}{ll} \min\limits_{v \in \partial\Omega(\bar{\mu})} & \lambda^{\bar{\mu}}_{y'}(v) \\ s.t & y' = d \circ v \end{array} \right. . \tag{4.23}$$

Proof. We split the proof into four steps. First, we remove some constraints and write the optimization problem in terms of $\mu - \bar{\mu}$. Second, we construct the dual of the linear program associated to the minimization w.r.t. μ and extract the variable ε as a multiplying factor in the objective, thus showing the linearity of the calibration function. Then, we add a simplex constraint to simplify the problem and finally, we put everything together to obtain the desired result.

1st step. Write optimization w.r.t μ in terms of $\mu - \bar{\mu}$ by removing some constraints. Let's proceed with the following editions of the constraints of (4.21) to obtain a lower bound:

1. The cone $\mathcal{N}_{\mathcal{M}}(\bar{\mu})$ is polyhedral, as it is the normal cone of the convex polytope $\mathcal{M}$ at the point $\bar{\mu}$. Hence, it is a finitely generated cone (De Loera et al., 2012), which can be described as

   $$\mathcal{N}_{\mathcal{M}}(\bar{\mu}) = \{a_1 n_1 + \cdots + a_r n_r \mid a_i \geq 0, n_i \in \mathbb{R}^k\}.$$

   Let's now replace the constraint $\mu \in \mathcal{M}$ (last constraint of (4.21)) by the constraints $\langle -n_i, \mu - \bar{\mu} \rangle \geq 0$ where $1 \leq i \leq r$ and n_i are the generators of the cone $\mathcal{N}_{\mathcal{M}}(\bar{\mu})$.
2. Note that by construction, we have that

   $$\langle \varphi(z), A\bar{\mu} \rangle = \langle \varphi(z'), A\bar{\mu} \rangle, \quad \forall z, z' \in y^\star(\bar{\mu}), \tag{4.24}$$

   as z, z' are optimal for the conditional moments $\bar{\mu}$. Let's remove from the second line of constraints of (4.21) the ones corresponding to $z \in \mathcal{Y} \setminus y^\star(\bar{\mu})$ and use (4.24) for the remaining constraints. We obtain

   $$\zeta_{y,y',\bar{\mu}}(\varepsilon) \geq \left\{ \begin{array}{lll} \min\limits_{v,\mu \in \mathbb{R}^k} & \langle w(y) - v, \mu - \bar{\mu} \rangle & \\ \text{s.t} & \langle A^\top(\varphi(y') - \varphi(y)), \mu - \bar{\mu} \rangle \geq \varepsilon, & \\ & \langle A^\top(\varphi(z) - \varphi(y)), \mu - \bar{\mu} \rangle \geq 0, & \forall z \in y^\star(\bar{\mu}) \\ & \langle -n_i, \mu - \bar{\mu} \rangle \geq 0, & 1 \leq i \leq r \\ & \langle v, \varphi(y') - \varphi(z) \rangle \geq 0, & \forall z \in \mathcal{Y} \end{array} \right.$$

3. Do the change of variables $\mu' = \mu - \bar{\mu}$ and re-define $\mu \coloneqq \mu'$ to ease notation.

$$\zeta_{y,y',\bar{\mu}}(\varepsilon) \geq \begin{cases} \min\limits_{v,\mu \in \mathbb{R}^k} & \langle w(y) - v, \mu \rangle \\ \text{s.t} & \langle A^\top(\varphi(y') - \varphi(y)), \mu \rangle \geq \varepsilon, \\ & \langle A^\top(\varphi(z) - \varphi(y)), \mu \rangle \geq 0, \quad \forall z \in y^\star(\bar{\mu}) \\ & \langle -n_i, \mu \rangle \geq 0, \qquad 1 \leq i \leq r \\ & \langle v, \varphi(y') - \varphi(z) \rangle \geq 0, \qquad \forall z \in \mathcal{Y} \end{cases}$$

2nd step. Linearity in ε via duality. Define $\overline{\mathcal{Y}} = y^\star(\bar{\mu})$. Let's now study separately the linear program corresponding to the variables μ, which reads as

$$\textbf{(P)} \qquad \begin{cases} \min\limits_{\mu} & \langle w(y) - v, \mu \rangle \\ \text{s.t} & \langle A^\top(\varphi(y') - \varphi(y)), \mu \rangle \geq \varepsilon, \\ & \langle A^\top(\varphi(z) - \varphi(y)), \mu \rangle \geq 0, \quad \forall z \in \overline{\mathcal{Y}} \\ & \langle -n_i, \mu \rangle \geq 0, \qquad 1 \leq i \leq r \end{cases}$$

Let's consider the dual formulation **(D)** of **(P)**:

$$\textbf{(D)} \qquad \begin{cases} \max\limits_{\lambda \in \mathbb{R}^{\overline{\mathcal{Y}}+r}} & \varepsilon \lambda_{y'} \\ \text{s.t} & A^\top \sum_{z \in \overline{\mathcal{Y}}} \lambda_z (\varphi(y) - \varphi(z)) + \sum_{i=1}^r \lambda_i^n n_i = A^\top \varphi(y) + v \\ & \lambda_y \geq 0 \qquad\qquad y \in \overline{\mathcal{Y}} \\ & \lambda_i^n \geq 0 \qquad\qquad 1 \leq i \leq r \end{cases},$$

where we have used that $w(y) = -A^\top \varphi(y)$.

3rd step. Simplify by adding a simplex constraint (dependence of optimal prediction y disappears). As problem **(D)** is written as a maximization, we can lower bound the objective by adding constraints. If we add the constraint $\sum_{z \in \overline{\mathcal{Y}}} \lambda_z = 1$, the term $A^\top \varphi(y)$ simplifies and we obtain the following lower bound

$$\begin{cases} \max\limits_{\lambda \in \mathbb{R}^{\overline{\mathcal{Y}}+r}} & \varepsilon \lambda_{y'} \\ \text{s.t} & A^\top \sum_{z \in \overline{\mathcal{Y}}} \lambda_z w(z) + \sum_{i=1}^r \lambda_i^n n_i = v \\ & \lambda_y \geq 0 \qquad\qquad y \in \overline{\mathcal{Y}} \\ & \sum_{z \in \overline{\mathcal{Y}}} \lambda_z = 1 \\ & \lambda_i^n \geq 0 \qquad\qquad 1 \leq i \leq r \end{cases}, \tag{4.25}$$

Note that the term $\sum_{i=1}^r \lambda_i^n n_i$ with $\lambda_i^n \geq 0$ covers all possible normal cone vectors, and so the maximization can be written over vectors in $\mathcal{N}_{\mathcal{M}}(\bar{\mu})$. Hence, Eq. (4.25) can be written as

$$\begin{cases} \max\limits_{\lambda, n} & \varepsilon \lambda_{y'} \\ \text{s.t} & v = \sum_{y \in \overline{\mathcal{Y}}} \lambda_y w(y) + n \\ & \lambda \in \Delta_{\overline{\mathcal{Y}}} \\ & n \in \mathcal{N}_{\mathcal{M}}(\bar{\mu}) \end{cases}. \tag{4.26}$$

4th step. Putting everything together. Recall that problem (4.26) is a function of v. The desired lower bound is constructed by minimizing the quantity (4.26) under the constraints $v \in \partial\Omega(\bar{\mu})$ and $y' = d \circ v$. □

B .4 Quantitative Lower Bound.

The compressed form of the calibration function (4.23) given by Theorem B .7 is still far from a quantitative understanding on the value of the function. The following Theorem B .8 provides a quantitative lower bound under mild assumptions on the loss L.

Assumption on L. L is symmetric and there exists $C > 0$ such that

$$y \in \arg\min_{y' \in \mathcal{Y}} \mathbb{E}_{z\sim\alpha} L(y', z) \implies \alpha_y \geq 1/C > 0, \tag{4.27}$$

for all $\alpha \in \Delta_{\mathcal{Y}}$.

Theorem B .8. *Assume* (4.27). *Then, for any $\varepsilon > 0$, the calibration function is lower bounded by*

$$\zeta(\varepsilon) \geq \frac{\varepsilon}{D},$$

where $D = \max_{y' \in \mathcal{Y}} D_{y'}$ and

$$1/D_{y'} = \min_{\bar{\mu} \in \mathcal{M}_0(y')} \begin{cases} \min\limits_{\alpha \in \Delta_{\overline{\mathcal{Y}}}} \max\limits_{\beta \in \Delta_{\overline{\mathcal{Y}}}} & \beta_y \\ \textit{s.t} & A^\top \mathbb{E}_{z\sim\alpha} \varphi(z) = A^\top \mathbb{E}_{z'\sim\beta} \varphi(z') \\ & y' \in \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{z\sim\alpha} L(y, z) \end{cases}, \tag{4.28}$$

where $\overline{\mathcal{Y}} = y^\star(\bar{\mu})$.

Proof. We use the notation $\nu(\alpha) = \mathbb{E}_{y\sim\alpha} \varphi(y)$ for $\alpha \in \Delta_{\mathcal{Y}}$. Let's first show that

$$\zeta_{y',y,\bar{\mu}}(\varepsilon)/\varepsilon \geq \begin{cases} \min\limits_{\alpha \in \Delta_{\overline{\mathcal{Y}}}} & \lambda^{\bar{\mu}}_{y'}(-A^\top \nu(\alpha)) \\ \text{s.t} & y' \in \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{z\sim\alpha} L(y, z) \end{cases}. \tag{4.29}$$

In order to see this, note that as $v \in \partial\Omega(\bar{\mu})$, we can write $v = -A^\top \nu(\alpha) + n_v$ where $\alpha \in \Delta_{\overline{\mathcal{Y}}}$ and $n_v \in \mathcal{N}_{\mathcal{M}}(\bar{\mu})$. We will show that condition $y' = d \circ v$ implies

$$\langle \varphi(z) - \varphi(y'), A^\top \nu(\alpha) \rangle \geq 0, \quad \forall z \in \mathcal{Y}.$$

The condition $y' = d \circ v$ is equivalent to $\langle \varphi(z) - \varphi(y'), A^\top \nu(\alpha) - n_v \rangle \geq 0$ for all $z \in \mathcal{Y}$. By definition of $\mathcal{N}_{\mathcal{M}}(\bar{\mu})$, we have that n_v satisfies $\langle s - \bar{\mu}, n_v \rangle \leq 0$ for any $s \in \mathcal{M}$. Now let $z \in \mathcal{Y} \setminus \{y'\}$ and consider the representation $\bar{\mu} = c_{y',\bar{\mu}} \varphi(y') + c_{z,\bar{\mu}} \varphi(z) + (1 - c_{y',\bar{\mu}} - c_{z,\bar{\mu}}) r$ with $r \in \text{hull}(\mathcal{Y} \setminus \{y', z\})$ and $0 \leq c_{y',\bar{\mu}}, c_{z,\bar{\mu}} \leq 1$. Since $n_v \in \mathcal{N}_{\mathcal{M}}(\bar{\mu})$ satisfies $\langle s - \bar{\mu}, n_v \rangle \leq 0$ also for $s = (c_{y',\bar{\mu}} + c_{z,\bar{\mu}}) \varphi(y') + (1 - c_{y',\bar{\mu}} - c_{z,\bar{\mu}}) r \in \mathcal{M}$, when $c_{z,\bar{\mu}} > 0$ we have

$$0 \geq c^{-1}_{z,\bar{\mu}} \langle s - \bar{\mu}, n_v \rangle = \langle \varphi(y') - \varphi(z), n_v \rangle.$$

From (4.27), we know that $c_{z,\bar{\mu}} \geq 1/C > 0$ for all $z \in y^\star(\bar{\mu})$. Then we have

$$\langle \varphi(z) - \varphi(y'), n_v \rangle \geq 0, \quad \forall\, z \in y^\star(\bar{\mu}).$$

Note moreover that $y^\star(\nu(\alpha)) \subseteq y^\star(\bar{\mu})$. Indeed, by the assumption (4.27), we have that $\alpha_z = 0$ implies $z \notin y^\star(\nu(\alpha))$ and since $\alpha \in \Delta_{y^\star(\bar{\mu})}$ we have that $\alpha_z = 0$ for

$z \notin y^\star(\bar{\mu})$. Since $y' \in y^\star(\bar{\mu})$ by construction of μ and A is symmetric due to the symmetry of L

$$\begin{aligned}
&\mathbb{E}_{t\in\alpha} L(t,z) - \mathbb{E}_{t\in\alpha} L(t,y') \\
&= \langle \varphi(z) - \varphi(y'), A\nu(\alpha)\rangle \\
&= \langle \varphi(z) - \varphi(y'), A^\top \nu(\alpha)\rangle \\
&\geq \langle \varphi(z) - \varphi(y'), n_v\rangle \geq 0,
\end{aligned}$$

for all $z \in y^\star(\bar{\mu})$. Hence, Eq. (4.29) is proven. Finally, setting $n = 0$ in the definition (4.22) of $\lambda^{\bar{\mu}}_{y'}(-A^\top \nu(\alpha))$, we obtain the desired lower bound. □

Corollary B .9. *Under the same assumptions of Theorem B .8, we have that $C \geq D$, and so*

$$\zeta(\varepsilon) \geq \frac{\varepsilon}{C}.$$

Proof. This can be seen by setting $\alpha = \beta$ in Eq. (4.28). □

Exponential constants in the calibration function. We argue that the constant D from Theorem B .8 does not grow as the size of the output space $\mathcal{Y}$ when the problem is structured, i.e., $k \ll |\mathcal{Y}|$. On the other hand, the constant C from Assumption (4.27) and Corollary B .9 can take exponentially large values (of the order of $|\mathcal{Y}|$) when the problem is structured. We show this by studying the calibration function for Example 2 .1 and Example 2 .2 in the next section.

B .5 Computation of the Constant for Specific Losses

Calibration function for factor graphs (Example 2 .1). Assume that we only have unary potentials and the individual losses are the 0-1 loss, which means that $A = -Id$ is the negative identity. Assume also that each part takes binary values, i.e., $R = 2$. The constant C can be as large as $|\mathcal{Y}| = 2^M$, by considering the uniform distribution $\alpha_y = 1/2^M$ for all y, which is optimal for every output. On the other hand, as the marginals for the uniform distribution are $(1/2, 1/2)$ for every m, one can take $\beta = 1/2\delta_y + 1/2\delta_{-y}$, and so D is 2.

Calibration function for ranking and matching (Example 2 .2). In this case, the constant C can be as large as $|\mathcal{Y}| = M!$ by considering the uniform distribution $\alpha_y = 1/M!$ for all y. This corresponds to $\mathbb{E}_{z\sim\alpha}\, \varphi(z) = 1/M 1 1^\top$. For this distribution, the value of the constant D is M, because one can write $1/M 1 1^\top$ as the the uniform distribution over M different permutations.

Proposition B .10 (Lipschitz Multi-class)**.** *Let $\mathcal{Y} = \{1, \ldots, k\}$,$k \geq 2$ and assume that L be symmetric. If there exists $q \in [0, 1)$ such that for all $y, z \in \mathcal{Y}$*

$$\sum_{t\in\mathcal{Y}\setminus\{y,z\}} |L(t,y) - L(t,z)| \;\leq\; q\, L(y,z),$$

then the calibration function for $M^4 N$ is bounded by

$$\zeta(\varepsilon) \geq H\epsilon, \quad H \geq \frac{1-q}{k-q} > 0.$$

Proof. First we prove that L satisfies (4.27). Then we apply Theorem B .6. Let $\alpha \in \Delta_{\mathcal{Y}}$ and assume that $y \in \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{t\sim\alpha} L(y,t)$. This is equivalent to the following

$$\sum_{t\in\mathcal{Y}} \alpha_t L(t,y) \leq \sum_{t\in\mathcal{Y}} \alpha_t L(t,z), \quad \forall z \in \mathcal{Y}.$$

In particular fix as $z = \arg\max_{t\in\mathcal{Y}\setminus\{y\}} \alpha_t$. By symmetry of the loss, the equation above is equivalent to

$$(\alpha_z - \alpha_y) L(y,z) \leq \sum_{t\in\mathcal{Y}\setminus\{y,z\}} \alpha_t (L(t,z) - L(t,y)).$$

Let $s = \arg\max_{t\in\mathcal{Y}\setminus\{y,z\}} \alpha_t$, then

$$\sum_{t\in\mathcal{Y}\setminus\{y,z\}} \alpha_t (L(t,z) - L(t,y)) \leq (\max_{t\in\mathcal{Y}\setminus\{y,z\}} \alpha_t) \sum_{t\in\mathcal{Y}\setminus\{y,z\}} |L(t,z) - L(t,y)| \leq \alpha_s \, q \, L(y,z).$$

Note that by construction $\alpha_s \leq \alpha_z$, so

$$(\alpha_z - \alpha_y) L(y,z) \leq \alpha_s \, q \, L(y,z) \leq \alpha_z \, q \, L(y,z),$$

from which we have $\alpha_z(1-q) \leq \alpha_y$. Since α_z is the maximum probability over $\mathcal{Y} \setminus \{y\}$, then it can not be smaller than $(1-\alpha_y)/(k-1)$, so $\alpha_z \geq (1-\alpha_y)/(k-1)$. From which we derive

$$\alpha_y \geq \frac{1-q}{k-q}.$$

This holds for any $\alpha \in \Delta_{\mathcal{Y}}$, $y \in \mathcal{Y}$ and implies that (4.27) is valid for L, with $c \geq \frac{1-q}{k-q}$. Then we can apply Theorem B .6 obtaining the desired result. □

Proposition B .11 (Decomposable Multi-label Loss)**.** *Let $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_M$, $L(y,y') = \sum_{m=1}^M L_m(y_m, y'_m)$ and $\varphi(y) = (e_{y_m})_{m\in M}$. Let ζ_m be the calibration function of L_m and assume $\zeta_m(\varepsilon) \geq \varepsilon/C_m$, with $C_m > 0$. The calibration function ζ associated to $L(y,y')$ has the following form:*

$$\zeta(\varepsilon) \;\geq\; \varepsilon/(\max_{m\in[M]} C_m).$$

Proof. We have $\delta\ell(y,\mu) = \sum_{m=1}^M \delta\ell_m(y_m,\mu_m)$ and the surrogate conditional loss decomposes additively as

$$s(v,\mu) = \sum_{m=1}^M s_m(v_m,\mu_m), \qquad s_m(v_m,\mu_m) = \max_{q\in\Delta_{\mathcal{Y}_m}} \min_{y'_m\in\mathcal{Y}_m} L_{y'_m} q + v^\top q - v^\top \mu_m.$$

We recall that the calibration function satisfies $\zeta(\varepsilon) \geq \varepsilon/C$ iff $\delta\ell(y,\mu) \leq C\delta s(v,\mu)$, for all v, μ and y among the minimizers of the surrogate. Hence, for all v, μ and y among the minimizers of the surrogate:

$$\begin{aligned}(\max_{m\in[M]} C_m)\delta s(v,\mu) = (\max_{m\in[M]} C_m) \sum_{m=1}^M \delta s_m(v_m,\mu_m)) &\geq \sum_{m=1}^M C_m \delta s_m(v_m,\mu_m) \\ &\geq \sum_{m=1}^M \delta\ell_m(y_m,\mu_m) = \delta\ell(y,\mu).\end{aligned}$$

□

Proposition B .12 (Calibration function for high-order factor graphs (Example 2 .1))**.** *Assume* (4.27)*. The constant D from Theorem B .8 for embeddings for unary and high-order interactions is the same as the constant D with only unary potentials.*

Proof. As the loss is decomposable as $L(y, y') = \frac{1}{M}\sum_{m=1}^{M} L_m(y_m, y'_m)$, it only depends on the unary embeddings. This means that the constraint $A^\top \mathbb{E}_{z\sim\alpha}\varphi(z) = A^\top \mathbb{E}_{z'\sim\beta}\varphi(z')$ from Eq. (4.28) only affects the unary embeddings, and so the lower bound is the same. □

C Sharp Generalization Bounds for Regularized Objectives

For $\lambda > 0$ and $g \in \mathcal{G}$ where $\mathcal{G}$ is a vector valued reproducing kernel Hilbert space and with norm $\|\cdot\|_{\mathcal{G}}$ and $g(x) \in \mathbb{R}^k$ defined as $g(x)_i = \langle g, \Psi_i(x)\rangle_{\mathcal{G}}$ with $\Psi_i : X \to \mathcal{G}$ for $i = 1, \dots, k$. Note that in particular we have the identity

$$K(x, x') = \langle \Psi(x), \Psi(x')\rangle_{\mathcal{G}} \in \mathbb{R}^{k\times k},$$

where K is the associated *vector-valued reproducing Kernel*. A simple example is the following. Let $K_0 : X \times X \to \mathbb{R}$ be a scalar reproducing kernel, then the kernel $K(x, x') = \frac{1}{k}K_0(x, x')I_{k\times k}$ is a vector-valued reproducing kernel whose associated vector-valued reproducing kernel Hilbert space contains functions of the form $g : X \to \mathbb{R}^k$. Now define

$$\mathcal{R}(g) = \mathbb{E}_{(x,y)\sim\rho} S(g(x), y), \quad \mathcal{R}^\lambda(g) = \mathcal{R}(g) + \lambda\|g\|^2_{\mathcal{G}}, \quad g^\lambda = \arg\min_{g\in\mathcal{G}} \mathcal{R}^\lambda(g).$$

Define also the empirical versions given a dataset $(x_i, y_i)_{i=1}^n$

$$\mathcal{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} S(g(x_i), y_i), \quad \mathcal{R}^\lambda_n(g) = \mathcal{R}_n(g) + \lambda\|g\|^2_{\mathcal{G}}, \quad g^\lambda_n = \arg\min_{g\in\mathcal{G}} \mathcal{R}^\lambda_n(g).$$

We will use the following theorem that is a slight variation of Thm.1 from Sridharan et al. (2009).

Theorem C .1. *Let $\delta \in (0, 1)$. Let L be the Lipschitz constant of S and let $\|\Psi_i(x)\| \le B$ for all $x \in X$, $i = 1, \dots, k$. Assume that there exists $g^\star$ such that $\mathcal{R}(g^\star) = \inf_{g\in\mathcal{G}} \mathcal{R}(g)$. For any $g \in \mathcal{G}$ the following holds*

$$\mathcal{R}(g) - \mathcal{R}(g^\star) \le 2(\mathcal{R}^\lambda_n(g) - \mathcal{R}^\lambda_n(g^\lambda_n)) + \frac{16L^2B^2(32 + \log(1/\delta))}{\lambda n} + \frac{\lambda}{2}\|g\|^2_{\mathcal{G}}.$$

with probability $1 - \delta$.

Proof. We apply the following error decomposition:

$$\mathcal{R}(g) - \mathcal{R}(g^\star) = (\mathcal{R}^\lambda(g) - \mathcal{R}^\lambda(g^\lambda)) + (\mathcal{R}^\lambda(g^\lambda) - \mathcal{R}^\lambda(g^\star)) + \frac{\lambda}{2}(\|g^\star\|^2_{\mathcal{G}} - \|g\|^2_{\mathcal{G}}).$$

By applying Theorem 1 of Sridharan et al. (2009) on $\mathcal{R}^\lambda(g) - \mathcal{R}^\lambda(g^\lambda)$, we have

$$\mathcal{R}^\lambda(g) - \mathcal{R}^\lambda(g^\lambda) \le 2(\mathcal{R}^\lambda_n(g) - \mathcal{R}^\lambda_n(g^\lambda_n)) + \frac{16L^2B^2(32 + \log(1/\delta))}{\lambda n}.$$

Considering that $\mathcal{R}^\lambda(g^\lambda) - \mathcal{R}^\lambda(g^\star) \le 0$ by definition of g^λ, we obtain the desired result. □

Now we are ready to prove Theorem 3.4

Proof of Theorem 3.4. Apply the theorem above with $g = g_n^\lambda$. Note moreover that by Fisher consistency in Theorem 3 .2, $g^\star(x) = -A^\top \varphi(f^\star(x))$. Let $\mathcal{G}$ be the vector-valued reproducing kernel Hilbert space associated to the vector-valued kernel K, then $B = \sup_{x\in X} \|\Psi(x)\|_\mathcal{G} = \|K(x,\cdot)\|_\mathcal{G} = \sup_{x\in X} \mathrm{Tr}(K(x,x))^{1/2}$, where Tr is the trace. Moreover we have $L \leq 2\|A\| \max_{y\in\mathcal{Y}} \|\varphi(y)\|$. The result is obtained by minimizing the resulting upper bound in λ and then applying comparison inequality of Theorem 3 .3. □

To conclude we extend a result from Pillaud-Vivien et al. (2018b) to our case. In the following assume $X = \mathbb{R}^d$ and denote by $\mathcal{G}_m$ the vector-valued reproducing kernel induced by $K_m(x,x') = \frac{1}{k}\bar{K}_m(x,x')I_{k\times k}$ where $I_{k\times k}$ is the identity matrix and $\bar{K}_m(x,x')$ is the scalar kernel associated to the Sobolev space $W_s^m(\mathbb{R}^d)$ for $m > d/2$. Note that when $m = (d+1)/2$, $\bar{K}_m(x,x') = e^{-\|x-x'\|}$.

Theorem C .2. *Let $X = \mathbb{R}^d$ and ρ be such that $\overline{X_y} \cap \overline{X_{y'}} = \emptyset$, for every $y \neq y'$ where $X_y = \{x \in X \mid y \in \arg\min_{z\in\mathcal{Y}} \mathbb{E}_{y\sim\rho(\cdot|x)} L(z,y)\}$. When $\mathcal{G} \subseteq \mathcal{G}_m$ for $m > d/2$, we have that*

$$\|\varphi(f^\star)\|_\mathcal{G} < \infty.$$

Proof. Since $W_2^m(\mathbb{R}^d)$ contains the smooth and compactly supported functions $C_c^\infty(\mathbb{R}^d)$ by construction for any $m > 0$ and $\mathcal{G}_m = W_2^m(\mathbb{R}^d)^{\otimes k}$, when $\mathcal{G} \supseteq \mathcal{G}_m$ we have that $\mathcal{G}$ contains all the vector valued compactly supported smooth functions. Now note for any two sets A, B there exists a compactly supported smooth function $f_{A,B}$ that has value 1 on A and 0 on B (see Pillaud-Vivien et al. (2018b) for more details). Now we build

$$g = -A^\top \sum_{y\in\mathcal{Y}} \varphi(y) f_{X_y, \cup_{z\neq y} X_z}.$$

Note that $d \circ g \in \arg\min_{f:X\to\mathcal{Y}} \mathbb{E}_{(x,y)\sim\rho} L(f(x),y)$, since for any $x \in \cup_{y\in X_y} = \mathrm{support}(\rho)$ we have $d\circ g \in \arg\min_{z\in\mathcal{Y}} \mathbb{E}_{y\sim\rho(\cdot|x)} L(z,y)$ by construction. I.e. $d\circ g = f^\star$ and so $g = -A^\top\varphi(f^\star)$. To conclude the theorem, note that $\|g\|_\mathcal{G} < \infty$ since $g \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^k) \subseteq \mathcal{G}_m \subseteq \mathcal{G}$. □

D Max-min margin and dual formulation

D .1 Derivation of the Dual Formulation

Let us first define

$$H_i(\mu, w) = \min_{y\in\mathcal{Y}} \varphi(y)^\top A\mu + g_w(x_i)^\top \mu - g_w(x_i)^\top \varphi(y_i).$$

Let's denote $w(\mu) = \frac{1}{\lambda n}\Phi_n(\mu - \varphi_n)$ where $\Phi_n = \frac{1}{\lambda n}(\Phi(x_1), \dots, \Phi(x_n))$ is the $d \times n$ scaled input data matrix and $\varphi_n = (\varphi(y_1), \dots, \varphi(y_n))^\top$ is the $n \times k$ output data matrix. Note that $w^\top w(\mu) = \frac{1}{\lambda n}\sum_{i=1}^n g_w(x_i)^\top(\mu_i - \varphi(y_i))$. The dual formulation **(D)** of $\mathrm{M^4Ns}$ can be derived as follows:

$$\begin{aligned}
&\min_{w\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \max_{\mu_i\in\mathcal{M}} H_i(\mu_i, w) + \frac{\lambda}{2}\|w\|_\mathcal{G}^2 \\
&= \min_{w\in\mathcal{G}} \frac{1}{n}\Big(\sum_{i=1}^n \max_{\mu_i\in\mathcal{M}} g_w(x_i)^\top(\mu_i - \varphi(y_i)) + \min_{y'\in\mathcal{Y}} \varphi(y')^\top A\mu_i\Big) + \frac{\lambda}{2}\|w\|_\mathcal{G}^2
\end{aligned}$$

$$
\begin{aligned}
&= \max_{\mu \in \mathcal{M} \times \cdots \times \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \Big(\min_{w \in \mathcal{G}} \; g_w(x_i)^\top (\mu_i - \varphi(y_i)) + \min_{y' \in \mathcal{Y}} \varphi(y')^\top A \mu_i \Big) + \frac{\lambda}{2} \|w\|_{\mathcal{G}}^2 \\
&= \max_{\mu \in \mathcal{M} \times \cdots \times \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \min_{y' \in \mathcal{Y}} \varphi(y')^\top A \mu_i + \min_{w \in \mathcal{G}} \; - w^\top \frac{1}{n} w(\mu) + \frac{\lambda}{2} \|w\|_{\mathcal{G}}^2 \\
&= \max_{\mu \in \mathcal{M} \times \cdots \times \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \min_{y' \in \mathcal{Y}} \varphi(y')^\top A \mu_i + \lambda \min_{w \in \mathcal{G}} \; - w^\top w(\mu) + \frac{1}{2} \|w\|_{\mathcal{G}}^2 \\
&= \max_{\mu \in \mathcal{M} \times \cdots \times \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \min_{y'} \varphi(y')^\top A \mu_i - \frac{\lambda}{2} \|w(\mu)\|_2^2,
\end{aligned}
$$

where the maximization and minimization have been interchanged using strong duality. We have $w^\star = w(\mu^\star)$.

D .2 Computation of the Dual Gap

The dual gap g at the pair $(w(\mu), \mu)$ decomposes additively in individual dual gaps as $g(w, \mu) = \frac{1}{2} \sum_{i=1}^{n} g_i(w, \mu_i)$:

$$
\begin{aligned}
g(w, \mu_i) &= \frac{1}{n} \sum_{i=1}^{n} \max_{\mu_i' \in \mathcal{M}} \; H_i(\mu_i', w) + \frac{\lambda}{2} w^\top w - \Big(\frac{1}{n} \sum_{i=1}^{n} \min_{y'} \varphi(y')^\top A \mu_i - \frac{\lambda}{2} w^\top w \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(\max_{\mu_i'} H_i(\mu_i', w) - \min_{y'} \varphi(y')^\top A \mu_i \Big) + \lambda w^\top w \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(\max_{\mu_i'} H_i(\mu_i', w) - \min_{y'} \varphi(y')^\top A \mu_i \Big) + \frac{1}{n} \sum_{i=1}^{n} w^\top (\lambda n w_i(\mu_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} \max_{\mu_i'} H_i(\mu_i', w) + w^\top (\lambda n w_i(\mu_i)) - \min_{y'} \varphi(y')^\top A \mu_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \max_{\mu_i'} H_i(\mu_i', w) - H_i(\mu_i, w) = \frac{1}{n} \sum_{i=1}^{n} g_i(w, \mu_i),
\end{aligned}
$$

where $w_i(\mu_i) = \frac{1}{\lambda n} \Phi(x_i)(\mu_i - \varphi(y_i))^\top$.

E Generalized Block-Coordinate Frank-Wolfe

E .1 General Convergence Result

In order to prove a convergence bound, following Lacoste-Julien et al. (2013), we will consider a more general optimization problem, and combine their proof with the proof of generalized conditional gradient from Bach (2015), with an additional support from approximate oracles.

We thus consider a product domain $\mathcal{K} = \mathcal{K}_1 \times \cdots \mathcal{K}_n$, and a smooth function f defined on $\mathcal{K}$, as well as n functions $h_1, \ldots, h_n$. We assume that f is L_i-smooth with respect to the i-th block. The optimization problem reads

$$
\min_{z \in \mathcal{K}_1 \times \cdots \times \mathcal{K}_n} g(z) \coloneqq f(z) + \sum_{i=1}^{n} h_i(z_i). \tag{4.30}
$$

Algorithm 3: Generalized Block-Coordinate Frank-Wolfe (GBCFW)

Let $z^{(0)} \in \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$;
for $t = 0$ ***to*** T **do**
  Pick $i(t)$ at random in $\{1, \ldots, n\}$;
  $\bar{z}^{\star}_{i(t)} \in \arg\min_{z'_{i(t)} \in \mathcal{K}_i} \nabla_{i(t)} f(z^{(t)})^\top z'_{i(t)} + h_{i(t)}(z'_{i(t)})$ *(solve oracle)*;
  $\gamma \coloneqq \frac{2n}{t+2n}$ or optimize γ by line-search.;
  $z^{(t+1)}_{i(t)} \coloneqq (1-\gamma) z^{(t)}_{i(t)} + \gamma \bar{z}^{\star}_{i(t)}$ *(Update only the i-th coordinate)*;
end

The algorithm, described in Algorithm 3, proceeds as follows. Starting from $z^{(0)} \in \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$, for $t \geq 0$, select $i(t)$ uniformly at random and find $\bar{z}_{i(t)} \in \mathcal{K}_{i(t)}$ such that minimizes a convex lower bound of the objective function on the $\mathcal{K}_{i(t)}$'th block. This convex lower bound is constructed by linearizing only the smooth part of the objective function. Hence, the minimization of the lower bound reads:

$$h_{i(t)}(\bar{z}_{i(t)}) + \nabla_{i(t)} f(z^{(t)})^\top \bar{z}_{i(t)} \leqslant \inf_{z_{i(t)} \in \mathcal{K}_{i(t)}} h_{i(t)}(z_{i(t)}) + \nabla_{i(t)} f(z^{(t)})^\top z_{i(t)} + \varepsilon_t. \tag{4.31}$$

Note that we allow an error of at most ε_t on the computation of the generalized Frank-Wolfe oracle. This is key in our analysis as in our setting we only have access to an approximate oracle. Finally, define $z^{(t+1)}$ by copying $z^{(t)}$ except the $i(t)$-th coordinate, which is taken to be

$$z^{(t+1)}_{i(t)} = (1-\gamma_t) z^{(t)}_{i(t)} + \gamma_t \bar{z}_{i(t)}.$$

We have, using the convexity of $h_{i(t)}$ and the smoothness of f, and denoting z^* a minimizer of $h(z) = f(z) + \sum_{i=1}^n h_i(z_i)$:

$$\begin{aligned}
& f(z^{(t+1)}) + \sum_{i=1}^n h_i(z_i^{(t+1)}) \\
\leq\ & f(z^{(t)}) + (z^{(t+1)}_{i(t)} - z^{(t)}_{i(t)})^\top \nabla_{i(t)} f(z^{(t)}) + \frac{L_{i(t)}}{2} \|z^{(t+1)}_{i(t)} - z^{(t)}_{i(t)}\|^2 \\
& \qquad + \sum_{i=1}^n h_i(z_i^{(t)}) + h_{i(t)}(z^{(t+1)}_{i(t)}) - h_{i(t)}(z^{(t)}_{i(t)}) \\
\leq\ & f(z^{(t)}) + \gamma_t(\bar{z}_{i(t)} - z^{(t)}_{i(t)})^\top \nabla_{i(t)} f(z^{(t)}) + \frac{L_{i(t)}}{2} \gamma_t^2 \|\bar{z}_{i(t)} - z^{(t)}_{i(t)}\|^2 \\
& \qquad + \sum_{i=1}^n h_i(z_i^{(t)}) + (1-\gamma_t) h_{i(t)}(z^{(t)}_{i(t)}) + \gamma_t h_{i(t)}(\bar{z}_{i(t)}) - h_{i(t)}(z^{(t)}_{i(t)}) \\
\leq\ & f(z^{(t)}) + \sum_{i=1}^n h_i(z_i^{(t)}) + \gamma_t \big[h_{i(t)}(\bar{z}_{i(t)}) + \bar{z}_{i(t)}^\top \nabla_{i(t)} f(z^{(t)})\big] \\
& \qquad - \gamma_t \big[h_{i(t)}(z^{(t)}_{i(t)}) + (z^{(t)}_{i(t)})^\top \nabla_{i(t)} f(z^{(t)})\big] + \frac{L_{i(t)}}{2} \gamma_t^2 \mathrm{diam}(\mathcal{K}_{i(t)})^2
\end{aligned}$$

Now, we use Eq. (4.31), i.e., the fact that $\bar{z}_{i(t)}$ is an approximate solution of the minimization problem $\inf_{z_{i(t)} \in \mathcal{K}_{i(t)}} \{h_{i(t)}(z_{i(t)}) + \nabla_{i(t)} f(z^{(t)})^\top z_{i(t)}\}$. In this case, we can continue upper bounding the above quantity as

$$
\begin{aligned}
\leq \quad & f(z^{(t)}) + \sum_{i=1}^{n} h_i(z_i^{(t)}) + \gamma_t \big[\inf_{z_{i(t)} \in \mathcal{K}_{i(t)}} \{ h_{i(t)}(z_{i(t)}) + \nabla_{i(t)} f(z^{(t)})^\top \, z_{i(t)} \} \big] \\
& \qquad -\gamma_t \big[h_{i(t)}(z_{i(t)}^{(t)}) - (z_{i(t)}^{(t)})^\top \nabla_{i(t)} f(z^{(t)}) \big] + \frac{L_{i(t)}}{2} \gamma_t^2 \mathrm{diam}(\mathcal{K}_{i(t)})^2 + \gamma_t \varepsilon_t \\
\leq \quad & f(z^{(t)}) + \sum_{i=1}^{n} h_i(z_i^{(t)}) + \gamma_t \big[h_{i(t)}(z_{i(t)}^{*}) + \nabla_{i(t)} f(z^{(t)})^\top \, z_{i(t)}^{*} \\
& \qquad -h_{i(t)}(z_{i(t)}^{(t)}) - (z_{i(t)}^{(t)})^\top \nabla_{i(t)} f(z^{(t)}) \big] + \frac{L_{i(t)}}{2} \gamma_t^2 \mathrm{diam}(\mathcal{K}_{i(t)})^2 + \gamma_t \varepsilon_t.
\end{aligned}
$$

Let's now define $C \geq 0$ as

$$
C = \sum_{i=1}^{n} L_i \mathrm{diam}(\mathcal{K}_i)^2.
$$

Finally, if we denote by $\mathcal{F}_t$ the information up to time t, we have that

$$
\begin{aligned}
& \mathbb{E}\Big[f(z^{(t+1)}) + \sum_{i=1}^{n} h_i(z_i^{(t+1)}) \mid \mathcal{F}_t \Big] \\
\leq \quad & f(z^{(t)}) + \sum_{i=1}^{n} h_i(z_i^{(t)}) \\
+ \quad & \gamma_t \big[\frac{1}{n} \sum_{i=1}^{n} h_i(z_i^{*}) + \frac{1}{n} \nabla f(z^{(t)})^\top \, z^{*} - \frac{1}{n} \sum_{i=1}^{n} h_i(z_i^{(t)}) - \frac{1}{n} (z^{(t)})^\top \nabla f(z^{(t)}) \big] \\
+ \quad & \frac{\gamma_t^2}{2} \frac{C}{n} + \gamma_t \varepsilon_t,
\end{aligned}
$$

where we have used $C/n = \mathbb{E}[L_{i(t)} \mathrm{diam}(\mathcal{K}_{i(t)})^2]$, $\mathbb{E}\left[h_{i(t)}(z_{i(t)}) \right] = \frac{1}{n} \sum_{i=1}^{n} h_i(z_i)$ and $\mathbb{E}[z_{i(t)}] = \frac{1}{n} z$ for any $z \in \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$. Thus, if g is the objective function as defined in (4.30), we get:

$$
\mathbb{E}[g(z^{(t+1)}) - g(z^{*})] \quad \leq \quad (1 - \frac{\gamma_t}{n}) [\mathbb{E} g(z^{(t)}) - g(z^{*})] + \frac{\gamma_t^2}{2} \frac{C}{n} + \gamma_t \varepsilon_t.
$$

Note that the above inequality is the same appearing in Jaggi (2013) but with the key difference of the factor $1/n$, which stems from the random block-coordinate procedure of the algorithm. If we define $G_t = \mathbb{E}[g(z^{(t)}) - g(z^{*})]$, we can re-write the recursion as

$$
G_{t+1} \leq (1 - \frac{\gamma_t}{n}) G_t + \frac{\gamma_t^2 C}{2n} + \gamma_t \varepsilon_t.
$$

Let's first set $\varepsilon_t = 0$, i.e., $G_{t+1} \leq (1 - \frac{\gamma_t}{n}) G_t + \frac{\gamma_t^2 C}{2}$, and prove by induction if $\gamma_t = \frac{2n}{t+2n} \in [0, 1]$, we obtain

$$
G_t \leq \frac{2n(C + G_0)}{t + 2n} \qquad t \geq 0.
$$

Let's proceed by induction. The *base-case* $k = 0$ is satisfied as $C \geq 0$.

$$
G_{t+1} \leq (1 - \frac{\gamma_t}{n}) G_t + \frac{\gamma_t^2 C}{2n}
$$

$$
\begin{aligned}
&= (1 - \frac{2}{t+2n})G_t + (\frac{2n}{t+2n})^2 \frac{C}{2n} \\
&\leq (1 - \frac{2}{t+2n})\frac{2n(C+G_0)}{t+2n} + (\frac{1}{t+2n})^2 2nC
\end{aligned}
$$

Rearranging the terms gives

$$
\begin{aligned}
G_{t+1} &\leq \frac{2nC}{t+2n}(1 - \frac{2}{t+2n} + \frac{1}{t+2n}) \\
&= \frac{2nC}{t+2n}\frac{t+2n-1}{t+2n} \\
&\leq \frac{2nC}{t+2n}\frac{t+2n}{t+2n+1} \\
&= \frac{2nC}{t+2n+1},
\end{aligned}
$$

which is the claimed bound for $k+1$. If we now we use an error

$$
\varepsilon_t = \frac{1}{2}\delta\gamma_t L_{i(t)} \mathrm{diam}(\mathcal{K}_{i(t)})^2. \tag{4.32}
$$

Then, we have that

$$
G_{t+1} \leq (1 - \frac{\gamma_t}{n})G_t + \frac{\gamma_t^2 C(1+\delta)}{2n}C,
$$

and so we get

$$
\mathbb{E}\big[g(z^{(t+1)}) - g(z^*)\big] \leq \frac{2n}{t+2n}\left(\mathbb{E}\big[g(z^{(0)}) - g(z^*)\big] + (1+\delta)\sum_{i=1}^n L_i \mathrm{diam}(\mathcal{K}_i)^2\right).
$$

In order to obtain the final bound only in terms of $\sum_{i=1}^n L_i \mathrm{diam}(\mathcal{K}_i)^2$, we can reuse the techniques from Lacoste-Julien et al. (2013), such as a single batch generalized Frank-Wolfe step, or use line search instead of constant step-sizes. Using these techniques, we can manage to set

$$
\mathbb{E}\big[g(z^{(0)}) - g(z^*)\big] \leq n \max_i \mathrm{diam}(\mathcal{K}_i)^2 \frac{\max_i L_i}{2},
$$

so that we obtain

$$
\mathbb{E}\big[g(z^{(t+1)}) - g(z^*)\big] \leq (2+\delta)\frac{2n^2}{t+2n} \max_i L_i \max_i \mathrm{diam}(\mathcal{K}_i)^2.
$$

E .2 Application to Our Setting, Proof of Theorem 5.1.

In our setting, we have that $\mathrm{diam}(\mathcal{K}_i) = \mathrm{diam}(\mathcal{M})$ and $L_i \leq \frac{R^2}{\lambda n^2}$ (R is the maximal norm of features). Hence, the bound simplifies to

$$
\mathbb{E}\big[g(z^{(t+1)}) - g(z^*)\big] \leq \frac{2(2+\delta)}{t+2n}\frac{R^2 \mathrm{diam}(\mathcal{M})^2}{\lambda}.
$$

Which means that in order to get $\mathbb{E}\big[g(z^{(t+1)}) - g(z^*)\big] \leq \varepsilon$ one needs

$$
t \geq \frac{2(2+\delta)R^2\,\mathrm{diam}(\mathcal{M})^2}{\lambda\varepsilon} + 2n = O\left(n + \frac{R^2\,\mathrm{diam}(\mathcal{M})^2}{\lambda\varepsilon}\right)
$$

iterations.

Algorithm 4: Generalized Block-Coordinate Frank-Wolfe (GBCFW)

Let $w^{(0)} \coloneqq w_i^{(0)} \coloneqq \bar{w}^{(0)} \coloneqq 0$;
for $t = 0$ ***to*** T **do**
  Pick i at random in $\{1, \dots, n\}$;
  $(\mu_i^\star, \nu_i^\star) = \mathcal{O}^{\varepsilon_t}(g_{w^{(t)}}(x_i))$ (solve oracle with precision ε_t);
  $w_s \coloneqq \Phi_n(\mu_i^\star - \varphi_n)/(\lambda n)$;
  $\gamma \coloneqq \frac{2n}{t+2n}$ (or line-search);
  $w_i^{(t+1)} \coloneqq (1-\gamma) w_i^{(t)} + \gamma w_s$;
  $w^{(t+1)} \coloneqq w^{(t)} + w_i^{(t+1)} - w_i^{(t)}$;
  (Optional averaging: $\bar{w}^{(t+1)} \coloneqq \frac{t}{t+2}\bar{w}^{(t)} + \frac{2}{t+2} w^{(t+1)}$);
end

F Solving the Oracle with Saddle Point Mirror Prox

Let $\mathcal{X} \subset \mathbb{R}^k$, $\mathcal{Y} \subset \mathbb{R}^k$ be compact and convex sets. Let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function such that $F(\cdot, y)$ is convex and $F(x, \cdot)$ is concave. We are interested in computing

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y).$$

By Sion's minimax theorem there exists a pair $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ such that

$$F(x^\star, y^\star) = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} F(x, y).$$

We assume that

$$\begin{aligned}
\|\nabla_x F(x, y) - \nabla_x F(x', y)\|_{\mathcal{X}}^* &\leq \beta_{1,1} \|x - x'\|_{\mathcal{X}} \\
\|\nabla_x F(x, y) - \nabla_x F(x, y')\|_{\mathcal{X}}^* &\leq \beta_{1,2} \|y - y'\|_{\mathcal{Y}} \\
\|\nabla_y F(x, y) - \nabla_y F(x', y)\|_{\mathcal{Y}}^* &\leq \beta_{2,1} \|x - x'\|_{\mathcal{X}} \\
\|\nabla_y F(x, y) - \nabla_y F(x, y')\|_{\mathcal{Y}}^* &\leq \beta_{2,2} \|y - y'\|_{\mathcal{Y}},
\end{aligned}$$

where $\|\cdot\|_{\mathcal{X}}^*, \|\cdot\|_{\mathcal{Y}}^*$ denote the dual norms of $\|\cdot\|_{\mathcal{X}}, \|\cdot\|_{\mathcal{Y}}$, respectively. We are interested in finding an algorithm that produces $(\widehat{x}, \widehat{y})$ that has small *duality gap* $g(\widehat{x}, \widehat{y})$ defined as

$$g(\widehat{x}, \widehat{y}) \coloneqq \max_{y \in \mathcal{Y}} F(\widehat{x}, y) - \min_{x \in \mathcal{X}} F(x, \widehat{y}).$$

F .1 Saddle Point Mirror Prox (SP-MP)

Define $H_{\mathcal{X}} : \mathcal{D}_{\mathcal{X}} \to \mathbb{R}$ and $H_{\mathcal{Y}} : \mathcal{D}_{\mathcal{Y}} \to \mathbb{R}$, which are 1-strongly concave w.r.t a norm $\|\cdot\|_{\mathcal{X}}$ on $\mathcal{X} \cap \mathcal{D}_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ on $\mathcal{Y} \cap \mathcal{D}_{\mathcal{Y}}$, respectively. Denote $R_{\mathcal{X}} = \sup_{x \in \mathcal{X}} H_{\mathcal{X}}(x) - \min_{x \in \mathcal{X}} H_{\mathcal{X}}(x)$ and $R_{\mathcal{Y}}$ similarily for $H_{\mathcal{Y}}$. Define $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$ and $H : \mathcal{D} \coloneqq \mathcal{D}_{\mathcal{X}} \times \mathcal{D}_{\mathcal{Y}} \to \mathbb{R}$ defined as $H(z) = \frac{1}{R_{\mathcal{X}}^2} H_{\mathcal{X}}(x) + \frac{1}{R_{\mathcal{Y}}^2} H_{\mathcal{Y}}(y)$, where $z = (x, y)$. The saddle point mirror prox (SP-MP) algorithm is defined as follows.

Start with $z^{(1)} = (x^{(1)}, y^{(1)}) = \arg\max_{z \in \mathcal{Z}} H(z)$. Then at every iteration k:

$$(u^{(k+1)}, v^{(k+1)}) \coloneqq \operatorname*{arg\,min}_{z \in \mathcal{Z} \cap \mathcal{D}} \eta(\nabla_x F(x^{(k)}, y^{(k)}), -\nabla_y F(x^{(k)}, y^{(k)}))^\top z + D_{-H}(z, z^{(k+1)})$$
$$(x^{(k+1)}, y^{(k+1)}) \coloneqq \operatorname*{arg\,min}_{z \in \mathcal{Z} \cap \mathcal{D}} \eta(\nabla_x F(u^{(k+1)}, v^{(k+1)}), -\nabla_y F(u^{(k+1)}, v^{(k+1)}))^\top z + D_{-H}(z, z^{(k+1)})$$

The following Theorem F .1 by Nemirovski (2004) studies the convergence of SP-MP.

Theorem F .1 (Nemirovski (2004))**.** *Let* $L = \max(\beta_{11} R_{\mathcal{X}}^2, \beta_{22} R_{\mathcal{Y}}^2, \beta_{12} R_{\mathcal{X}} R_{\mathcal{Y}}, \beta_{21} R_{\mathcal{X}} R_{\mathcal{Y}})$*. Then, the algorithm saddle point mirror prox (presented at the beginning of the section) runned with* $\eta = \frac{1}{2L}$ *satisfies*

$$g(\bar{u}^K, \bar{v}^K) \leq \frac{4L}{K},$$

where $\bar{u}^K \coloneqq \frac{1}{K} \sum_{k=1}^K u^{(k)}$ *and* $\bar{v}^K \coloneqq \frac{1}{K} \sum_{k=1}^K v^{(k)}$*.*

In our setting, we have that $\mathcal{X} = \mathcal{Y} = \mathcal{M}$ and

$$F(\nu, \mu) = \nu^\top A \mu + v^\top \mu. \tag{4.33}$$

The gradients have the following form:

$$\nabla_\nu F(\nu, \mu) = A\mu, \quad \text{and} \quad \nabla_\mu F(\nu, \mu) = A^\top \nu + v.$$

F .2 Max-Min Oracle for Sequences (special case of Example 2 .1)

Consider unary potentials and binary potentials between adjacent variables. The embeddings can be written as

$$\varphi(y) = (\varphi_u(y), \varphi_p(y)) = ((\varphi_m(y_m))_{m=1}^M, \varphi_{m,m+1}(y_{m,m+1})_{m=1}^{M-1}) \in \mathbb{R}^{RM + R^2(M-1)},$$

where $\varphi_m(y_m) = e_{y_m} \in \mathbb{R}^R$ and $\varphi_{m,m+1}(y_{m,m+1}) = e_{y_{m,m+1}} \in \mathbb{R}^{R^2}$ are vectors of the canonical basis. Here, φ_u and φ_p stand for unary and pair-wise embeddings. If the loss decomposes coordinate-wise as $L(y, y') = \frac{1}{M} \sum_{i=1}^M L_m(y_m, y'_m)$ as detailed in Example 2 .1, the loss decomposition reads

$$A = \left(\begin{array}{ccc|c} L_1/M & \cdots & 0_{R\times R} & \\ \vdots & \ddots & \vdots & \mathcal{O}_{MR \times (M-1)R^2} \\ 0_{R\times R} & \cdots & L_M/M & \\ \hline \multicolumn{3}{c|}{\mathcal{O}_{(M-1)R^2 \times MR}} & \mathcal{O}_{(M-1)R^2 \times (M-1)R^2} \end{array}\right), \qquad a = 0.$$

The bilinear function (4.33) takes the following form:

$$F(\nu, \mu) = \sum_{m=1}^M \nu_m^\top L_m \mu_m + \sum_{m=1}^M v_m^\top \mu_m + \sum_{p=1}^{M-1} v_p^\top \mu_p.$$

Note that as A is low-rank, the dependence on ν is only on the unary embeddings, which means that the minimization over ν is over a simpler domain that decomposes as $\mathcal{Q} = \Pi_{m=1}^M \Delta_R$.

We consider the entropies $H_{\mathcal{Q}} : \mathcal{Q} \to \mathbb{R}$ and $H_{\mathcal{M}} : \mathcal{M} \to \mathbb{R}$ defined as:

$$H_{\mathcal{Q}}(\nu) \coloneqq \sum_{m=1}^{M} H_S(\nu_m), \qquad H_{\mathcal{M}}(\mu) \coloneqq \left\{ \begin{array}{lll} \max_{q \in \Delta_{|\mathcal{Y}|}} & H_S(q) & \\ \text{s.t.} & \mathbb{E}_{y \sim q} \varphi_m(y_m) = \mu_m, & 1 \leq m \leq M \\ & \mathbb{E}_{y \sim q} \varphi_p(y_p) = \mu_p, & 1 \leq p \leq M-1 \end{array} \right. ,$$

where for $q \in \Delta_k$, we define the *Shannon entropy* $H_S : \Delta_k \to \mathbb{R}$ as $H_S(q) = -\sum_{j=1}^{k} q_j \log q_j$.

In order to apply SP-MP we need to compute two projections in $\mathcal{Q}$ and $\mathcal{M}$ with respect to the corresponding entropies described above. The update on $\nu \in \mathcal{Q}$ takes the form

$$\arg\min_{\nu \in \mathcal{Q}} \ \eta \sum_{m=1}^{M} \nu_m^\top L_m \mu_m^{(t)} + D_{-H_{\mathcal{Q}}}(\nu, \nu^{(t)}). \tag{4.34}$$

As the entropy $H_{\mathcal{Q}}$ is separable, the projection (4.34) is separable and can be computed with the softmax operator. The update on $\mu \in \mathcal{M}$ takes the form

$$\arg\min_{\mu \in \mathcal{M}} \ -\eta \sum_{m=1}^{M} \mu_m^\top (L_m^\top \nu_m^{(t)} + v_m) - \eta \sum_{p=1}^{M-1} \mu_p^\top v_p + D_{-H_{\mathcal{M}}}(\mu, \mu^{(t)}). \tag{4.35}$$

Projection (4.35) can be computed using marginal inference using the sum-product algorithm.

Norm $\|\cdot\|_{\mathcal{Q}}$ and constants $R_{\mathcal{Q}}, \sigma_{\mathcal{Q}}$. We choose the norm as the L_1-norm $\|\nu\|_{\mathcal{Q}} \coloneqq \|\nu\|_1 = \sum_{m=1}^{M} \|\nu_m\|_1$. From Pinsker's inequality, we know that $H(\nu_m)$ is 1-strongly convex with respect to $\|\cdot\|_1$ in Δ_R. Hence, we have that $H_{\mathcal{Q}}(\nu)$ is 1-strongly respect with respect to $\|\cdot\|_1$ in $\mathcal{Q}$. Moreover, using that $\min_{q \in \mathcal{Q}} H_{\mathcal{Q}}(\nu) = 0$, we have that

$$R_{\mathcal{Q}}^2 \coloneqq \max_{\nu \in \mathcal{Q}} H_{\mathcal{Q}}(\nu) = \max_{\nu \in \Pi_{m=1}^{M} \Delta_R} \sum_{m=1}^{M} H(\nu_m) = \sum_{m=1}^{M} \max_{\nu_m \in \Delta_R} H(\nu_m) = \sum_{m=1}^{M} \log R = M \log R.$$

Norm $\|\cdot\|_{\mathcal{M}}$ and constants $R_{\mathcal{M}}, \sigma_{\mathcal{M}}$. If we choose the L_2-norm $\|\mu\|_{\mathcal{M}} \coloneqq \|\mu\|_2$, the strong-convexity constant of $H_{\mathcal{M}} : \mathcal{M} \to \mathbb{R}$ defined in Section F .2 with respect to $\|\cdot\|_2$ is

$$\sigma_{\mathcal{M}} = \operatorname{diam}(\mathcal{M})^{-2}.$$

In order to see this, note that the strong-convexity parameter $\sigma_{\mathcal{M}}$ of $H_{\mathcal{M}}$ is equal to the inverse of the smoothness parameter of the partition function $A(v) = \log\left(\sum_{y \in \mathcal{Y}} \exp(\langle \varphi(y), v \rangle)\right)$, which corresponds to the maximal dual norm $\|\cdot\|_*$ of the covariance operator $\Sigma(v) = \mathbb{E}_{y \sim q_v} \varphi(y)\varphi(y)^\top - \mathbb{E}_{y \sim q_v} \varphi(y)\, \mathbb{E}_{y \sim q_v} \varphi(y)^\top$, where $q_v(y) = \exp\langle v, \varphi(y) \rangle / \sum_{y' \in \mathcal{Y}} \exp(\langle \varphi(y'), v \rangle))$ If we consider $\|\cdot\|_2$, it follows directly that $\sigma_{\mathcal{M}}^{-1} = \sup_v \|\Sigma(v)\|_2 \leq \operatorname{diam}(\mathcal{M})^2$. Finally, using that $\min_{\mu \in \mathcal{M}} H_{\mathcal{M}}(\mu) = 0$, we have that

$$R_{\mathcal{M}}^2 \coloneqq \max_{\mu \in \mathcal{M}} H_{\mathcal{M}}(\mu) = \sum_{m=1}^{M} \max_{\mu_m \in \Delta_R} H(\mu_m) + (\leq 0) \leq \sum_{m=1}^{M} \log R.$$

Computation of the smoothness constants ($\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}$).

- $\beta_{11} = 0$ as $\nabla_m F_x(q, \mu)$ is constant in q for all $m \in [M]$.
- We have that $\|L_m(\mu_m - \mu'_m)\|_\infty \leq \|L_m\|_\infty \|\mu_m - \mu'_m\|_1$. Hence, $\beta_{12} = \max_{m\in[M]} \|L_m\|_\infty$.
- We have that $\nabla_m F_y(q, \mu)$ and $\nabla_c F_y(q, \mu)$ are constant in μ for all $m \in [M]$ and $c \in C$, so $\beta_{12} = 0$.
- We have that $\|L_m^\top(q_m - q'_m)\|_2 \leq \|L_m^\top\|_2 \|q_m - q_m\|_2$. Hence, $\beta_{22} = \max_{m\in[M]} \|L_m^\top\|_2$.

Finally, the constant L appearing in Theorem F .1 reads

$$L = \max_{m\in[M]} \|L_m\|_2 \operatorname{diam}(\mathcal{M})^2 M \log R.$$

F .3 Max-Min Oracle for Ranking and Matching of Example 2 .2

We represent the permutation $\sigma \in \mathcal{S}_M$ using the corresponding permutation matrix $\varphi(\sigma) = P_\sigma \in \mathbb{R}^{M\times M}$. The loss decomposition is

$$L(\sigma, \sigma') = \frac{1}{M} \sum_{m=1}^{M} 1(\sigma(j) \neq \sigma'(j)) = 1 - \frac{\langle P_\sigma, P_{\sigma'} \rangle}{M} = 1 - \frac{\langle \varphi(\sigma), \varphi(\sigma') \rangle}{M},$$

i.e., $A = -Id/M$ and $a = 1$. The marginal polytope $\mathcal{M}$ corresponds to the *Birkhoff polytope* or equivalently, the polytope of *doubly stochastic matrices*

$$\mathcal{M} = \operatorname{hull}\{P_\sigma \mid \sigma \in \mathcal{S}_M\} = \{P \in \mathbb{R}^{M\times M} \mid P1 = 1, P^T 1 = 1, 0 \leq P_{ij} \leq 1, i, j \in [M]\}.$$

The max-min oracle corresponds to the following saddle-point problem:

$$\arg\max_{P\in\mathcal{M}} \min_{Q\in\mathcal{M}} \langle S, P\rangle - \langle Q, P\rangle/m. \tag{4.36}$$

We have three natural options for the entropy, namely, the constrained Shannon entropy (which is the one used in the factor graph Example 2 .1), the entropy of marginals and the quadratic entropy.

Constrained Shannon Entropy. In this case,

$$H(Q) \coloneqq \max_{p\in\Delta_{\mathcal{S}_M}} -\sum_{\sigma\in\mathcal{S}_M} p(\sigma)\log p(\sigma) \quad \text{s.t.} \quad \sum_{\sigma\in\mathcal{S}_M} p(\sigma) P_\sigma = Q.$$

The projection corresponds to marginal inference, which is in general $\#P$-complete as we have to compute the permanent (Valiant, 1979). As noted by Petterson et al. (2009), it can be 'efficiently' computed exactly up to $M = 30$ with complexity $O(M2^M)$ using an algorithm by Ryser (1963). Note that this is way faster than enumeration which is of the order of $M! \sim M^M$.

Entropy of Marginals. We can define the entropy defined in the marginals as

$$H(Q) = -\sum_{i,j=1}^{M} Q_{ij} \log Q_{ij}. \tag{4.37}$$

The projection can be computed up to precision δ using the *Sinkhorn-Knopp algorithm* with complexity $O(M^2/\delta)$. Moreover, this can be easily implemented efficiently in `C++` as the algorithm corresponds to an alternating normalization between rows and columns.

Quadratic Entropy. We can use the following quadratic entropy

$$H(Q) \coloneqq -\|Q\|_F^2 = -\sum_{i,j=1}^{M} Q_{ij}^2.$$

The projection has essentially the same complexity as the entropy on marginals described above **?** and it provides sparse solutions. The algorithm consists in minimizing an unconstrained smooth and non-strongly convex function. The computation of the gradient requires M euclidean projections to the simplex Δ_M. Each projection can be performed exactly in worst-case $O(M \log M)$ using the algorithm by Michelot (1986) and in expected $O(M)$ using the randomized pivot algorithm of Duchi et al. (2008). The resulting computational complexity is of $O(M^2/\delta)$. Note that even though the complexity is the same as for the entropic regularization, the implementation is more involved and difficult to speed up.

In our experiments we focus on the entropy on marginals (4.37). We now compute the constants.

Norm $\|\cdot\|_{\mathcal{M}}$ and constants $R_{\mathcal{M}}, \sigma_{\mathcal{M}}$. If we consider $\|\cdot\|_{\mathcal{M}} = \|\cdot\|_1$, we have that $\sigma_{\mathcal{M}} = 1$ and $R_{\mathcal{M}}^2 = M$.

Computation of the smoothness constants ($\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}$). In this case we obtain $\beta_{11} = \beta_{12} = 0$ and $\beta_{21} = \beta_{22} = 1$. Hence

$$L = M.$$

G Generalization Bounds for $\mathrm{M^4N}$ solved via GBCFW and Approximate Oracle

Proof of Theorem 5.2 Denote by $\widehat{g}_{n,T}$ the result of Algorithm 1 where the oracle is approximated via Algorithm 2. In the same setting of Section E, by applying Theorem C .1, bounding L, B as in the proof of Theorem 3 .4 and applying the comparison inequality in Theorem 3 .3, we have that the following holds with probability $1 - \delta$

$$\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star) \leq 2(\mathcal{R}_n^\lambda(\widehat{g}_{n,T}) - \mathcal{R}_n^\lambda(g_n^\lambda)) + M\|\varphi(f^\star)\|_{\mathcal{G}} \sqrt{\frac{\log(1/\delta)}{n}},$$

when λ is chosen as $\lambda = \kappa L \log^{1/2}(1/\delta) n^{-1/2}$ and M defined as in Theorem 3 .4. Denote by $\varepsilon_{\text{opt}} = \mathcal{R}_n^\lambda(\widehat{g}_{n,T}) - \mathcal{R}_n^\lambda(g_n^\lambda)$. The result is obtained by optimizing until $\varepsilon_{\text{opt}} = O\Big(\|\varphi(f^\star)\|_{\mathcal{G}} \sqrt{\frac{\log(1/\delta)}{n}}\Big)$, we have that

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}(g^\star) \leq O\Big(\|\varphi(f^\star)\|_{\mathcal{G}} \sqrt{\frac{\log(1/\delta)}{n}}\Big).$$

According to Theorem 5.1 and F .1 this is obtained with a number of steps for Algorithm 1 of $T = O(n)$ and Algorithm 2 in the order of $O(\sqrt{n})$, for a total computational complexity of $O(n\sqrt{n})$. □

5 Max-Margin is Dead, Long Live Max-Margin!

Abstract

The foundational concept of Max-Margin in machine learning is ill-posed for output spaces with more than two labels such as in structured prediction. In this paper, we show that the Max-Margin loss can only be consistent to the classification task under highly restrictive assumptions on the discrete loss measuring the error between outputs. These conditions are satisfied by distances defined in tree graphs, for which we prove consistency, thus being the first losses shown to be consistent for Max-Margin beyond the binary setting. We finally address these limitations by correcting the concept of Max-Margin and introducing the Restricted-Max-Margin, where the maximization of the loss-augmented scores is maintained, but performed over a subset of the original domain. The resulting loss is also a generalization of the binary support vector machine and it is consistent under milder conditions on the discrete loss.

1 Introduction

One of the first binary classification methods learned in a machine learning course is the support vector machine (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) and it is introduced using the principle of maximum margin: assuming the data are linearly separable, the classification hyperplane must maximize the separation to the observed examples. Having this intuition in mind, the same principle has been used to extend this notion to larger output spaces $\mathcal{Y}$, such as multi-class classification (Crammer and Singer, 2001) and structured prediction (Taskar et al., 2004; Tsochantaridis et al., 2005), where the separation to the observed examples is controlled by a discrete loss $L(y, y')$ measuring the error between outputs y and y'. The resulting method generalizes the binary SVM and corresponds to minimizing the so-called Max loss

$$S_{\mathrm{M}}(v, y) = \max_{y' \in \mathcal{Y}} \; L(y, y') + v_{y'} - v_y, \tag{5.1}$$

where $v \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector with coordinate v_y encoding the score for output y. Unfortunately, this method may not be *consistent*, i.e., minimizing the Max loss (5.1) may not lead to a minimization of the discrete loss L of interest. In particular, it is known that the Max loss is only consistent for the 0-1 loss under the dominant label condition, i.e., when for every input there exists an output element with probability larger than $1/2$ (Liu, 2007), which is always satisfied in the binary case. However, far less is known for other tasks. Indeed, the Max loss is widely used for structured output spaces where the discrete loss L defining the task is different than the 0-1 loss, under the name of Structural SVM (SSVM) (Taskar et al., 2005b; Caetano et al., 2009; Smith, 2011) or Max-Margin Markov

Networks ($\mathrm{M}^3\mathrm{N}$) (Taskar et al., 2005b). In this general setting, the following questions remain unanswered:

> (i) *Does it exist a necessary condition on L for consistency to hold? Does it exist a space of losses for which consistency holds? Can we generalize the consistency result under the dominant label assumption beyond the 0-1 loss?*
>
> (ii) *Can we correct the Max loss to make it consistent by maintaining the additive and maximization structure of the Max loss?*

We answer these questions in this paper. In particular, we make the following contributions:

- We prove that the Max loss can only be consistent under a restrictive necessary condition on the structure of the loss L, indeed, the loss L has to be a distance and satisfy the triangle inequality as an *equality* for several groups of outputs. As a positive result, we show that a distance defined in a tree graph, such as the absolute deviation loss used in ordinal regression, satisfies this condition and it is consistent, thus providing the first set of losses for which consistency holds beyond the binary setting. We also extend the existing *partial* consistency result of the 0-1 loss by extending the result under the dominant label condition to all losses that are distances.
- We introduce the Restricted-Max loss, where the loss-augmented scores defining the Max loss are maximized over a restricted subset of the simplex. The resulting loss also generalizes the binary SVM and it is consistent under milder assumptions on L. Moreover, we show the connections between these losses and the Max-Min loss (Fathony et al., 2016; Duchi et al., 2018; Nowak-Vila et al., 2020), where consistency always holds independently of the discrete loss L.

2 Max-Margin and Main Results

In this section we introduce the concept of Max-Margin learning and its consistency from its origins in binary classification to the structured output setting. This is followed by the presentation of the main results of this paper and its implications are discussed.

2 .1 Max-Margin Learning

Binary output. Let $(x_1, y_1), \dots, (x_n, y_n)$ be n examples of input-output pairs sampled from an unknown distribution ρ defined in $\mathcal{X} \times \mathcal{Y}$. Let us first assume that the input space $\mathcal{X}$ is a vector space and $\mathcal{Y} = \{-1, 1\}$ represents binary labels. The goal is to construct a binary-valued function $f : \mathcal{X} \to \mathcal{Y}$ minimizing the expected classification error

$$\mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho} L(f(x), y), \tag{5.2}$$

where $L(y, y') = 1(y \neq y')$ is the binary 0-1 loss. The concept of max-margin was initially defined in this setting to construct a predictor of the form $\operatorname{sign}(g(x))$ where $g(x) =$

$w^\top x + b$ is an affine function defining a hyperplane with maximum separation to the examples assuming linearly separable data (Boser et al., 1992). In this setting, an example x_i is correctly classified if $(w^\top x_i + b_i)y_i > 0$ and misclassified otherwise. The max-margin hyperplane can be found by minimizing $\|w\|_2^2$ under the constraint $(w^\top x_i + b)y_i \geq 1$ for all n examples. When the data are not linearly separable, some examples are allowed to be misclassified by introducing some positive *slack variables* ξ_i and solving the optimization problem known as the *support vector machine (SVM)* (Cortes and Vapnik, 1995):

$$\min_{w,b,\xi} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\|w\|_2^2 \quad \text{s.t.} \quad (w^\top x_i + b)y_i \geq 1 - \xi_i, \quad \xi_i \geq 0,\ \forall i = 1, \ldots, n,$$

where $\lambda > 0$ is a parameter used to balance the first term with the second. We can re-write the constraints as $\xi_i \geq 1 - y_i g(x_i)$ for non-negative ξ_i's and extend the affine hypothesis space to a generic functional space $\mathcal{G}$ with associated norm $\|\cdot\|_{\mathcal{G}}$ to allow for non-linear predictors, such as reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950). Then, the problem above can be written as a convex regularized empirical risk minimization (ERM) (Vapnik, 1992) problem

$$\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} S_{\mathrm{M}}(g(x_i), y_i) + \frac{\lambda}{2}\|g\|_{\mathcal{G}}^2, \tag{5.3}$$

where $S_{\mathrm{M}}(v, y) = \max(1 - yv, 0)$ is the binary Max loss (also called SVM or hinge loss), and now λ can be interpreted as the regularization parameter. An important property of the classification method is that the estimated predictor solving (5.3) over all measurable functions converges to the predictor $f^\star$ minimizing the expected classification error (5.2) in the infinite data regime ($n \to \infty$ and $\lambda \to 0$) (Vapnik, 1995). More concretely, the minimizer $g^\star$ of the expected risk $\mathbb{E}_{(x,y)\sim\rho} S(g(x), y)$ must satisfy $f^\star = \operatorname{sign}(g^\star)$. This property is called *Fisher consistency* (Bartlett et al., 2006) (or simply consistency) and can be studied in terms of the conditional expectation $q(x) := \rho(1|x)$, as $f^\star(x)$ and $g^\star(x)$ can be characterized in terms of this quantity [1]. Note that in the rest of the paper we will drop the dependence in x from the function q: a statement $P(q)$ for all $q \in [0, 1]$ must then be read as $P(q(x))$ for all $x \in \mathcal{X}$. Let $v_{\mathrm{M}}^\star(q) \subseteq \mathbb{R}$ and $y^\star(q) \subseteq \mathcal{Y}$ be the minimizers of the conditional risks $\mathbb{E}_{y'\sim q} S_{\mathrm{M}}(v, y')$ and $\mathbb{E}_{y'\sim q} L(y, y')$, respectively. Then, Fisher consistency is equivalent to say that if $v \in v_{\mathrm{M}}^\star(q)$, then $\operatorname{sign}(v) \in y^\star(q)$ for all $q \in [0, 1]$ (Devroye et al., 1996). This property is satisfied as (see also left Figure 5.1):

$$y^\star(q) = \begin{cases} \{1\} & q \in (1/2, 1] \\ \{-1, 1\} & q = 1/2 \\ \{-1\} & q \in [0, 1/2), \end{cases}, \quad v_{\mathrm{M}}^\star(q) = \begin{cases} [1, \infty) & q = 1 \\ \{1\} & q \in (1, 1/2) \\ [-1, 1] & q = 1/2 \\ \{-1\} & q \in (0, 1/2) \\ (-\infty, -1] & q = 0. \end{cases}.$$

Structured prediction. In the structured prediction setting, we have $k = |\mathcal{Y}|$ possible outputs and the goal is to estimate a discrete-valued function f minimizing (5.2) where now $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a generic non-negative discrete loss function between output pairs defining the task at hand. We construct predictors of the form $\arg\max_{y\in\mathcal{Y}} g_y(x)$, where $g : \mathcal{X} \to \mathbb{R}^k$ is a vector-valued function assigning scores to each of the k possible

[1] This is because $f^\star$ and $g^\star$ are minimizers over all measurable functions of an expectation over $\mathcal{X}$.

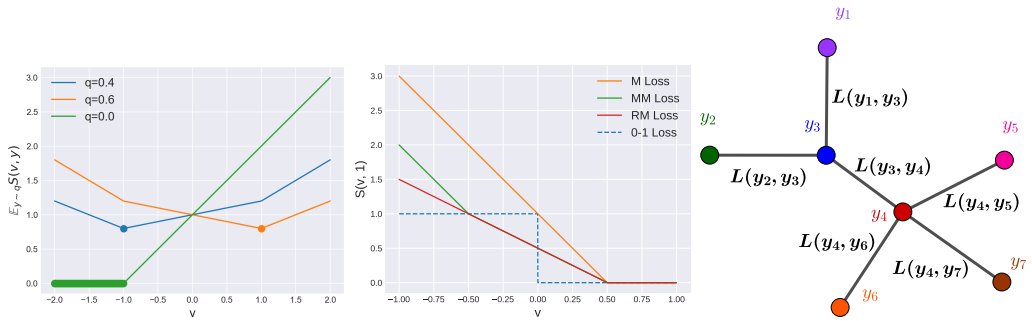


Figure 5.1: **Left:** Plots of the conditional risks of the Max loss for $q = 0.4, q = 0.6$ and $q = 0$, respectively. The conditional risk in the set of minimizers $v^\star_{\mathrm{M}}(q)$ is plotted with a thick point / line. **Middle:** Plot of $S_{\mathrm{M}}(v, 1), S_{\mathrm{MM}}(v, 1)$ and $S_{\mathrm{RM}}(v, 1)$ in the binary setting with $v = v_1 = -v_2$. In this case, $S_{\mathrm{M}} = 2S_{\mathrm{RM}}$ (so both losses generalize the binary SVM up to a factor of 2) and S_{M} is the only one upper-bounding the 0-1 loss. Moreover, the three losses are consistent with the 0-1 loss. **Right:** Distance defined in a tree: the distance/loss between two nodes is the sum of the distances between adjacent nodes of the path between them. For every triplet of outputs $y, y', y'' \in \mathcal{Y}$, or they are aligned in a path, or there exists $z \in \mathcal{Z}$ belonging to the shortest path between all pairs.

outputs. The maximum margin principle from binary classification is generalized as follows. For every example (x_i, y_i), the method minimizes the squared norm $\|g\|^2_{\mathcal{G}}$ under the constraints

$$g_{y_i}(x_i) \geq \underbrace{L(y, y_i) + g_y(x_i)}_{\text{loss-augmented scores}}, \tag{5.4}$$

for all possible outputs y. By writing the above constraint as $g_{y_i}(x_i) - g_y(x_i) \geq L(y, y_i)$, we observe that this generalizes the condition $y_i g(x_i) \geq 1$ from binary classification when $g = g_1 = -g_{-1}$ so that the argmax corresponds to the sign and L is the binary 0-1 loss. As in the binary case, introducing slack variables and turning it into a regularized ERM problem of the form (5.3) we obtain the Max loss $S_{\mathrm{M}}(v, y) = \max_{y' \in \mathcal{Y}} L(y, y') + v_{y'} - v_y$, which is constructed as a maximization of the loss-augmented scores defined in (5.4). To ease notation, the dependence of S on the loss L is deduced from the context. The Max loss (5.1) is known as the Crammer-Singer SVM (Crammer and Singer, 2001) when L is the 0-1 loss, and it is also widely used in structured prediction settings with exponentially large output spaces under the name of Structural SVM (Joachims, 2006) or Max-Margin Markov Networks ($\mathrm{M}^3\mathrm{N}$) (Taskar et al., 2004) by using losses between structured outputs such as sequences, permutations, graphs, etc (BakIr et al., 2007). An interesting property of this loss is that it upper-bounds the discrete loss as $L(\arg\max_{y' \in \mathcal{Y}} v_{y'}, y) \leq S_{\mathrm{M}}(v, y)$, for all $v \in \mathbb{R}^k$ and $y \in \mathcal{Y}$, which can guide us to think that minimizing $\mathbb{E}_{(x,y)\sim\rho} S_{\mathrm{M}}(g(x), y)$ leads to minimizing $\mathbb{E}_{(x,y)\sim\rho} L(\arg\max_{y' \in \mathcal{Y}} g_{y'}(x), y)$. Unfortunately this intuition is misleading, as this bound is in general far from tight. Analogously to the binary case, let $q : \mathcal{X} \to \Delta$ be the conditional distribution where Δ is the simplex over $\mathcal{Y}$ (we again drop the dependence on x, since each statement must be read as holding for every $x \in \mathcal{X}$). Moreover, we define for every q the set of minimizers of the conditional risks as

$$y^\star(q) = \arg\min_{y \in \mathcal{Y}} L_y^\top q \subseteq \mathcal{Y}, \quad v^\star_{\mathrm{M}}(q) = \arg\min_{v \in \mathbb{R}^k} S_{\mathrm{M}}(v)^\top q \subseteq \mathbb{R}^k,$$

where $L_y = (L(y, y'))_{y'} \in \mathbb{R}^k$ and $S_{\mathrm{M}}(v) = (S_{\mathrm{M}}(v, y'))_{y' \in \mathcal{Y}} \in \mathbb{R}^k$. We say that S_{M} is Fisher consistent to L if for all $q \in \Delta$

$$v \in v^\star_{\mathrm{M}}(q) \implies \arg\max_{y \in \mathcal{Y}} v_y \in y^\star(q). \tag{5.5}$$

Related works on consistency of Max-Margin. The Max loss is only consistent to the 0-1 loss under the dominant label assumption $\max_{y \in \mathcal{Y}} q_y \geq 1/2$ (Liu, 2007). Ramaswamy et al. (2018) show that it is consistent to the "abstain" loss, but in this case the loss appearing in the definition (5.1) is not the same as the classification loss L. There exist several generalizations of the binary SVM to larger output spaces other than Max-Margin (Dogan et al., 2016) such as Weston-Watkins (WW-SVM) (Weston and Watkins, 1999), Lee-Lin-Wahba (LLW-SVM) (Lee et al., 2004), Simplex-Coding (SC-SVM) (Mroueh et al., 2012), with the last two being consistent and defined as sums. However, the only loss with a max-structure is (5.1), which makes it computationally feasible to work in structured spaces of exponential size such as sequences or permutations. The Max-Min loss (Fathony et al., 2016; Duchi et al., 2018; Nowak-Vila et al., 2020) (defined below in Eq. (5.8)) is always consistent, it has a max-min structure and can be used in structured prediction settings. However, it does not correspond to the SVM in the binary setting, so it cannot be considered a generalization of the binary SVM.

2 .2 Main Results

We assume that L is symmetric and that $L(y, y') = 0$ if and only if $y = y'$. Symmetry of L is assumed for the sake of exposition, but it is only required for the results on the Max loss.

Main Results on Max-Margin. The following Theorem 2 .1 is our main negative result.

Theorem 2 .1 (Necessary condition for consistency S_{M})**.** *Let $k = |\mathcal{Y}| > 2$. If the Max loss is consistent to L, then L is a distance and for every three outputs $y_1, y_2, y_3 \in \mathcal{Y}$, there exists $z \in \mathcal{Y}$ for which the following three identities hold:*

$$\begin{aligned} L(y_1, y_2) &= L(y_1, z) + L(z, y_2) \\ L(y_1, y_3) &= L(y_1, z) + L(z, y_3) \\ L(y_2, y_3) &= L(y_2, z) + L(z, y_3) \end{aligned}$$

If $z = y_2$ in Theorem 2 .1, then the only informative condition is $L(y_1, y_3) = L(y_1, y_2) + L(y_2, y_3)$ as L is assumed to be symmetric, which means that the outputs y_1, y_2, y_3 are 'aligned' in the output space (analogously for $z = y_2, y_3$). On the other hand, if $z \neq y_1, y_2, y_3$, then the three equations are informative and all distances between the pairs can be decomposed into distances to z. The following discrete losses do not satisfy the above necessary condition (see Appendix B):

- Losses which are not distances (such as the squared discrete loss $(y - y')^2$).
- Losses with full rank loss matrix with existing $q \in \mathrm{int}(\Delta)$ for which all outputs are optimal, i.e., $y^\star(q) = \mathcal{Y}$ (such as the 0-1 loss).
- Hamming losses $L(y, y') = \frac{1}{M} \sum_{m=1}^M L_m(y_m, y'_m)$ with $y, y' \in \Pi_{m=1}^M \mathcal{Y}_m$ where L_m does not satisfy the necessary conditions for some $m = 1, \ldots, M$.

- Hamming loss on permutations $L(\sigma, \sigma') = \frac{1}{M}\sum_{m=1}^{M} 1(\sigma(m) \neq \sigma'(m))$ with σ, σ' permutations of size M, used for graph matching (Petterson et al., 2009; Caetano et al., 2009).

It is an open question whether the necessary condition of Theorem 2 .1 is also sufficient. The following Theorem 2 .2 shows that distances defined in a tree, which always satisfy this condition (see right Figure 5.1), are indeed consistent.

Theorem 2 .2 (Sufficient condition for consistency S_{M})**.** *If L is a distance defined in a tree, then the Max loss is consistent to L.*

An important example of these losses is the *absolute deviation* loss used in ordinal regression,

$$L(y, y') = |\gamma_y - \gamma_{y'}|, \qquad \gamma \in \mathbb{R}^k,$$

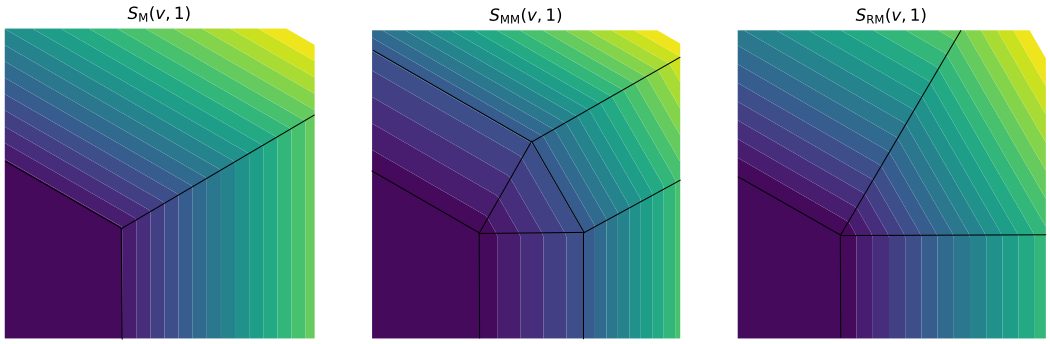


Figure 5.2: From left to right: plots of $S_{\mathrm{M}}(v, 1), S_{\mathrm{MM}}(v, 1)$ and $S_{\mathrm{M}}(v, 1)$ in the three-label setting with $v^\top 1 = 0$ for the 0-1 loss (note that $S(v + c1, y) = S(v, y)$ for all three losses). The Max-Min and the Restricted-Max loss coincide in the bottom-left region, but the Max-Min loss has three extra activated faces in the top-right region of the plot which are in general unnecessary for consistency, while the Restricted-Max uses just the necessary ones. The Max loss in the left plot uses just three faces, being insufficient for consistency.

for which the associated tree is a chain. Note that these losses are not the only ones satisfying the necessary condition given by Theorem 2 .1. Indeed, the Hamming loss with $M = 2$, $\mathcal{Y}_1 = \mathcal{Y}_2 = \{-1, 1\}$ and L_1, L_2 the 0-1 loss is *not* a distance in a tree, satisfies the necessary condition, and consistency can be proven to hold (see Appendix B). The following Proposition 2 .3 gives a much milder sufficient condition to ensure *partial* consistency under the dominant label assumption, generalizing thus the well-known results from Liu (2007).

Proposition 2 .3 (Sufficient condition for *partial* consistency S_{M})**.** *If L is a distance, then the Max loss is consistent to L under the dominant label assumption, i.e.,* $\max_{y \in \mathcal{Y}} q_y \geq 1/2$.

In other words, if the learning task is defined by a distance and it is close to deterministic, then the Max loss is consistent to the task.

Beyond Max-Margin. To overcome the limitations imposed by the maximum margin, but retaining the maximization structure of the loss, we propose a novel generalization

of the binary SVM to structured prediction by restricting the maximization of the loss-augmented scores in (5.1). First, note that the Max loss can be written as a maximization over the simplex Δ over $\mathcal{Y}$ as

$$S_{\mathrm{M}}(v, y) = \max_{q \in \Delta} \; L_y^\top q + v^\top q - v_y. \tag{5.6}$$

We restrict the maximization to the so-called *prediction set* $\Delta(y) = \{q \in \Delta \mid y \in y^\star(q)\}$, defined as the set of probabilities for which y is optimal. In binary classification the sets are $\Delta(-1) = [0, 1/2]$ and $\Delta(1) = [1/2, 1]$. The resulting *Restricted-Max loss* reads

$$S_{\mathrm{RM}}(v, y) = \max_{q \in \Delta(y)} \; L_y^\top q + v^\top q - v_y. \tag{5.7}$$

This loss satisfies $2S_{\mathrm{RM}} = S_{\mathrm{M}}$ in the binary setting (see middle Figure 5.1), thus, it corresponds to the binary SVM up to a scaling with a factor of two. The following Theorem 2 .4 states that consistency of S_{M} implies consistency of S_{RM} and provides a sufficient condition for consistency of S_{RM}.

Theorem 2 .4 (Sufficient condition for consistency S_{RM})**.** *The Restricted-Max loss is consistent to L whenever the Max loss is consistent. Moreover, if L satisfies $q_y > 0$ for every y optimal for $q \in \Delta$, i.e., $q \in \Delta(y)$, then the Restricted-Max loss is also consistent to L.*

In other words, if the output y is optimal for q, then the probability of this label has to be strictly greater than zero $q_y > 0$. The 0-1 loss, which does not satisfy the necessary condition of Theorem 2 .1, satisfies the sufficient condition for the Restricted-Max, as $\min_{q \in \Delta(y)} q_y = 1/k$ for all $y \in \mathcal{Y}$. However, there are still losses for which (5.7) is not consistent to, such as the squared discrete loss $(z - y)^2$ (see Appendix B). The remaining inconsistencies can be resolved by going beyond the maximization structure into a max-min structure. The resulting loss is the so-called *Max-Min loss* (Fathony et al., 2016; Duchi et al., 2018; Nowak-Vila et al., 2020) defined as

$$S_{\mathrm{MM}}(v, y) = \max_{q \in \Delta} \min_{z \in \mathcal{Y}} L_z^\top q + v^\top q - v_y. \tag{5.8}$$

It is known (Nowak-Vila et al., 2020) that the loss (5.8) is *always* consistent to L. As shown in Figure 5.1 (middle), this loss does not correspond to the SVM in the binary setting because it has two symmetric kinks instead of one. See also Figure 5.2 to compare the shape of the different losses for $k = 3$. Hence, while the structure of the loss gets more computationally involved from the Max loss (5.6) to the Max-Min loss (5.8), passing by the Restricted-Max loss (5.7), the consistency properties of these losses improve from one to the next.

3 Background and Preliminary Results

3 .1 Background on Polyhedral Losses

Fisher consistency. Let us now consider a generic loss $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ and let's generalize the argmax computing the prediction from the scores in the previous section to a generic *decoding* function $d : \mathbb{R}^k \to \mathcal{Y}$. The set of minimizers $v^\star(q) \subseteq \mathbb{R}^k$ of the conditional risk $S(v)^\top q$ is also defined as before. We say that S is *Fisher consistent* to L (Tewari and Bartlett, 2007) under the decoding $d : \mathbb{R}^k \to \mathcal{Y}$ if

$$v \in v^\star(q) \implies d(v) \in y^\star(q), \tag{5.9}$$

for all $q \in \Delta$. If the decoding is not specified, it means that there exists a decoding satisfying this property. An important quantity throughout the paper is the *Bayes risk*, a *concave* function defined as the minimum of the conditional expected loss respectively for L and S:

$$H_L(q) = \min_{y \in \mathcal{Y}} L_y^\top q, \qquad H_S(q) = \min_{v \in \mathbb{R}^k} S(v)^\top q.$$

Embedding of discrete losses. Fisher consistency in Eq. (5.9) states that every minimizer $v \in v^\star(q)$ can be assigned to a solution of the discrete task using the decoding d. For the analysis of this paper, it will be useful to work using the concept of embeddability between losses (Finocchiaro et al., 2019), a stronger notion than Fisher consistency.

Definition 3 .1 (Embeddability)**.** *S embeds L if there exists an embedding $\psi : \mathcal{Y} \to \mathbb{R}^k$ such that: (i) $y \in y^\star(q) \iff \psi(y) \in v^\star(q),\ \forall q \in \Delta$, and (ii) $S(\psi(y)) = L_y,\ \forall y \in \mathcal{Y}$.*

Condition (i) states that every solution of the discrete problem corresponds to a solution of the problem in S and vice versa. In particular, this rules out many smooth plug-in classifiers such as the squared loss or logistic regression, because they predict the vector of probabilities q which cannot be recovered from the discrete predictor $y^\star(q)$. It is known (Finocchiaro et al., 2019) that the existence of an embedding ψ satisfying (i) implies the existence of a decoding d satisfying Eq. (5.9), so it is already a sufficient condition for Fisher consistency. Note that both Eq. (5.9) and condition (i) are assumptions on the predictors $y^\star$ and $v^\star$, but there exist many possible losses L for which $y^\star(q)$ is the set of minimizers of the conditional risk $L_y^\top q$, and analogously for $v^\star$ and S. Condition (ii) restricts the relationship between pairs of losses by assuming that L can be recovered from S using the embedding ψ. The following Proposition 3 .2 shows that S embedding L is equivalent to having the same Bayes risks.

Proposition 3 .2 ((Finocchiaro et al., 2019))**.** *S embeds L if and only if $H_L = H_S$.*

Moreover, it is known that any discrete loss L is embedded by at least one loss S (Theorem 2 by Finocchiaro et al. (2019)), which corresponds precisely to the Max-Min loss S_{MM} defined in Eq. (5.8). Indeed, S_{MM} and L have the same Bayes risk as

$$H_{\mathrm{MM}}(q) = \min_{v \in \mathbb{R}^k} \big(\max_{p \in \Delta} \min_{y \in \mathcal{Y}} L_y^\top p + v^\top p \big) - v^\top q = \min_{v \in \mathbb{R}^k} (-H_L)^*(v) - v^\top q = H_L(q),$$

where $h^*(u) = \sup_{s \in \mathbb{R}^k} u^\top s - h(s)$ is the Fenchel conjugate of h. It can be checked (Nowak-Vila et al., 2020) that the embedding is $\psi(y) = -L_y$ and it is always Fisher consistent to L under the argmax decoding.

3 .2 Preliminary Results

Relationship between losses and Bayes risks. What makes the Max-Min loss simple to analyze is its Fenchel-Young structure (Blondel et al., 2020), i.e., it can be written in the form $S(v, y) = \Omega^*(v) - v_y$, for a certain convex Ω defined in the simplex. We extend this notion by allowing the convex function Ω to depend on the label y as

$$S(v, y) = (\Omega^y)^*(v) - v_y. \tag{5.10}$$

The losses $S_{\mathrm{M}}, S_{\mathrm{RM}}$ and S_{MM} can be written in this form with

$$\Omega^y_{\mathrm{MM}}(q) = -\min_{y' \in \mathcal{Y}} L_{y'}^\top q + i_\Delta(q), \quad \Omega^y_{\mathrm{M}}(q) = -L_y^\top q + i_\Delta(q), \quad \Omega^y_{\mathrm{RM}}(q) = -L_y^\top q + i_{\Delta(y)}(q),$$

where $i_U(u) = 0$ if $u \in U$ and ∞ otherwise. The first equation is the only one independent of y and we remove its dependence by simply writing Ω_{MM}. The following Proposition 3 .3 relates the functions $\Omega_{\mathrm{MM}}, \Omega_{\mathrm{M}}^y$ and Ω_{RM}^y.

Proposition 3 .3. *The following holds:* $\Omega_{\mathrm{MM}} = \max_{y\in\mathcal{Y}} \Omega_{\mathrm{M}}^y$ *and* $(\Omega_{\mathrm{MM}})^* = \max_{y\in\mathcal{Y}} (\Omega_{\mathrm{RM}}^y)^*$.

Proof. The first identity is trivial from the definition. For the second identity, note that by construction the prediction sets necessarily cover the simplex as $\Delta = \cup_{y'\in\mathcal{Y}} \Delta(y')$. Hence,

$$(\Omega_{\mathrm{MM}})^*(v) = \max_{y\in\mathcal{Y}} \big(\max_{q\in\Delta(y)} \min_{y'\in\mathcal{Y}} L_{y'}^\top q + v^\top q \big) = \max_{y\in\mathcal{Y}} \big(\max_{q\in\Delta(y)} L_y^\top q + v^\top q \big) = \max_{y\in\mathcal{Y}} (\Omega_{\mathrm{RM}}^y)^*(v),$$

where we have used the fact that $\min_{y'\in\mathcal{Y}} L_{y'}^\top q = L_y^\top q$ whenever $q \in \Delta(y)$ by construction. □

Moreover, it can be readily seen that $S_{\mathrm{RM}}(v, y) \leq S_{\mathrm{MM}}(v, y) \leq S_{\mathrm{M}}(v, y)$, for all $v \in \mathbb{R}^k$ and $y \in \mathcal{Y}$. See Figure 5.1 (left and middle) and Figure 5.2 for the shape of these losses when L is the 0-1 loss for two and three dimensions, respectively. Our main quantity of interest is the Bayes risk H_S, because by comparing it with H_L we are able to tell whether L is embedded by S using Proposition 3 .2, thus proving consistency. The following Proposition 3 .4 gives the form of the Bayes risks.

Proposition 3 .4 (Bayes risks)**.** *For all $q \in \Delta$, the Bayes risks read*

$$H_{\mathrm{MM}}(q) = H_L(q), \qquad H_{\mathrm{M}}(q) = \max_{Q\in U(q,q)} \langle L, Q\rangle_{\mathrm{F}}, \qquad H_{\mathrm{RM}}(q) = \max_{Q\in U(q,q)\cap\mathcal{C}_L} \langle L, Q\rangle_{\mathrm{F}},$$

where $U(q, q) = \{Q \in \mathbb{R}^{k\times k} \mid Q1 = q, Q^\top 1 = q, Q \succeq 0\}$ and $\mathcal{C}_L = \{Q \in \mathbb{R}^{k\times k} \mid (1L_y^\top - L)Q_y \preceq 0,\ \forall y \in \mathcal{Y}\}$. Moreover, we have that $H_{\mathrm{RM}} \leq H_{\mathrm{MM}} \leq H_{\mathrm{M}}$, and there exists L for which $H_L \neq H_{\mathrm{M}}$ and/or $H_{\mathrm{RM}} \neq H_L$.

The proof can be found in Appendix A . As a corollary, the only one of these losses always embedding L is the Max-Min loss. Our sufficient conditions for consistency will correspond to the conditions on L for which the Bayes risks H_{M} and/or H_{RM} are equal (or proportional) to H_L, thus implying consistency. In the next section, we use the expressions of the Bayes risks given by Proposition 3 .4 as a basis to prove the consistency results.

4 Fisher Consistency Analysis

4 .1 Analysis of Max Loss

In this section we want to provide a necessary condition for consistency of the Max loss. Note that it is not enough to provide a condition for which $H_S \neq H_L$, as S consistent to L does not imply S embedding L. The following Proposition 4 .1 gives a necessary condition for consistency in terms of the extreme points of the prediction sets and the Bayes risk of S.

Proposition 4 .1. *Assume S is Fisher consistent to L. Then, for any extreme point $q \in \Delta$ of a prediction set $\Delta(y),\ y \in \mathcal{Y}$, we necessarily have $\{q\} = \partial(-H_S)^*(v)$ for some $v \in \mathbb{R}^k$.*

The set $\partial h(x)$ denotes the subgradient of the function h. This result is proven in Appendix B . To use this necessary condition, we first have to compute the Fenchel conjugate $(-H_{\mathrm{M}})^*$ of the Max loss. This is given by the following Proposition 4 .2.

Proposition 4 .2. *If L is symmetric, then $(-H_{\mathrm{M}})^*(v) = \max_{y,y' \in \mathcal{Y}} L(y,y') + \frac{v_y + v_{y'}}{2}$ for all $v \in \mathbb{R}^k$.*

As a corollary, we obtain the specific form of the images of the sub-gradient mapping $\partial(-H_{\mathrm{M}})^*$ at the differentiable points, i.e., when the sub-gradient is a singleton.

Corollary 4 .3. *The 0-dimensional images of $\partial(-H_{\mathrm{M}})^*$ are of the form $q = \frac{1}{2}(e_y + e_{y'}), y, y' \in \mathcal{Y}$.*

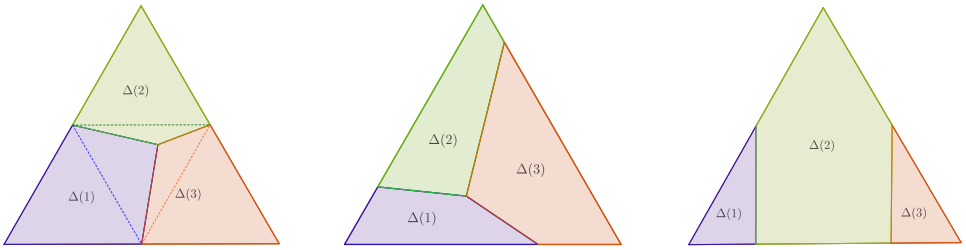


Figure 5.3: **Left:** Prediction sets $\Delta(y)$ for a symmetric loss. The points given by Corollary 4 .3 are the 3 extremes of the simplex and the middle points in the faces of the simplex. The interior point is not of the form $\frac{1}{2}(e_y + e_{y'})$ for any $y, y' \in \{1,2,3\}$. **Middle:** The sufficient condition from Proposition 4 .6 is satisfied for any point in the simplex. **Right:** The sufficient condition from Proposition 4 .6 is not satisfied as the prediction set $\Delta(2)$ (in green) intersects the line $q_2 = 0$.

Note that the points $\{q\} \in \operatorname{Im} \partial(-H_{\mathrm{M}})^*$ are independent of the discrete loss L. In Figure 5.3 (left), we show that this is not the case for the extreme points of the prediction sets $\Delta(y)$'s of a generic L. In this example, the points $\frac{1}{2}(e_y + e_{y'})$'s are extreme points of the prediction sets because L is symmetric, but the extreme point in the interior is not of this form. By moving this point to $\frac{1}{2}(e_1 + e_3)$, we enforce $L(1,2) + L(2,3) = L(1,3)$, which corresponds precisely to the necessary condition given by Theorem 2 .1 when $z = y_2$. The generic necessary condition is obtained by extending this argument to consider all possibilities in larger output spaces. The full proof can be found in Appendix B .

Consistency of distances defined in trees. This result is proved by showing $H_{\mathrm{M}} = 2H_L$ ($H_S \propto H_L$ also implies consistency (Finocchiaro et al., 2019)) whenever L is a distance defined in a tree, and it is done by proving equality of the Fenchel conjugates $(-H_{\mathrm{M}})^* = (-2H_L)^*$. The proof is based on the analysis of the extreme points of a certain polytope defined in terms of L, and can be found in Appendix B .

Partial positive results on consistency. In this section, we generalize the well-known result by Liu (2007) which states that the Max loss is consistent to the 0-1 loss under the dominant label condition. More specifically, we provide a generalization of this result to all distance losses, i.e., symmetric and satisfying the triangle inequality.

Proposition 4 .4. *If L is a distance, then we have $H_{\mathrm{M}} \leq 2H_L$ and $H_{\mathrm{M}}(q) = 2H_L(q)$ (S embeds $2L$, thus consistent), for all $q \in \Delta$ such that $\max_{y\in\mathcal{Y}} q_y \geq 1/2$.*

Some intuition can be obtained for $k = 3$. The interior of the set delimited by the dashed lines in the left image from Figure 5.3 shows precisely the points where it is satisfied $\max_{y\in\mathcal{Y}}\ q_y \leq 1/2$. If L is a distance, then the interior extreme point can only move inside this region, and hence, it remains consistent at the exterior of this set, which is precisely where the dominant label condition is satisfied. The proof of this result can be found in Appendix B .

4 .2 Analysis of Restricted-Max Loss

In this section we sketch the proof of the consistency of the Restricted-Max loss (Theorem 2 .4). We prove that S_{RM} embeds L by showing equal Bayes risks using the expressions from Proposition 3 .4. The proof is done in two steps. In the first one (Proposition 4 .5), we show that under a condition on the extreme points of the prediction sets both Bayes risks are equal, while in the second we obtain that this condition is satisfied whenever the hypothesis from Theorem 2 .4 holds.

Proposition 4 .5. *If $qq^\top \in \mathcal{C}_L = \{Q \in \mathbb{R}^{k\times k} \mid (1L_y^\top - L)Q_y \preceq 0,\ \forall y \in \mathcal{Y}\}$ for all extreme points $q \in \Delta$ of the prediction sets $\Delta(y)$'s, then $H_{\mathrm{RM}} = H_L$.*

Proof. By Proposition 3 .4, we already know that $H_{\mathrm{RM}} \leq H_L$, so we only need to show that $H_{\mathrm{RM}} \geq H_L$. If q is an extreme point of the prediction set $\Delta(y)$, then $H_{\mathrm{RM}}(q) \geq H_L(q)$ by taking the matrix $qq^\top \in U(q,q) \cap \mathcal{C}_L$, which satisfies $\langle L, qq^\top\rangle_{\mathrm{F}} = \sum_{y'} L_{y'}^\top q_{y'} u \geq \sum_{y'} q_{y'} L_y^\top q = L_y^\top q = H_L(q)$, where we have used that $L_{y'}^\top q \geq L_y^\top q$ for all $y' \in \mathcal{Y}$ as $q \in \Delta(y)$. If q is not an extreme point, then it can be written as a convex combination of extreme points q_i of $\Delta(y)$ as $q = \sum_{i=1}^m \alpha_i q_i$ with $\alpha^\top 1 = 1, \alpha \succeq 0$. Then, it is straightforward to show that the matrix $Q = \sum_{i=1}^m \alpha_i q_i q_i^\top$ is in $U(q,q) \cap \mathcal{C}_L$, and satisfies $\langle L, Q\rangle_{\mathrm{F}} \geq L_y^\top q = H_L(q)$. More details in Appendix C . □

Proposition 4 .6. *Assume that for every $q \in \Delta$, if y is optimal for q (i.e., $q \in \Delta(y)$), then $q_y > 0$. In this case, all extreme points of the prediction sets satisfy $qq^\top \in \mathcal{C}_L$.*

The proof of this result and the fact that consistency of S_{M} implies consistency of S_{RM} can be found in Appendix C . In Figure 5.3 (middle and right), we show geometrically in dimension $k = 3$ what this condition means.

4 .3 Argmax Decoding

The positive consistency results of both S_{M} and S_{RM} do not specify whether the argmax decoding $d(v) = \arg\max_{y\in\mathcal{Y}}\ v_y$ is consistent, but just that there exists a decoding for which consistency holds. From Nowak-Vila et al. (2020), we know that the set of minimizers of the Max-Min loss S_{MM} is

$$v^\star_{\mathrm{MM}}(q) = -\operatorname{hull}(L_y)_{y\in y^\star(q)} + \mathcal{N}_\Delta(q), \quad \forall q \in \Delta, \tag{5.11}$$

where hull denotes the convex hull and $\mathcal{N}_\Delta(q) = \{u \in \mathbb{R}^k \mid u^\top(p-q) \leq 0, \forall p \in \Delta\}$ is the normal cone of the simplex at the point q. In this case, the argmax decoding is consistent

because $\arg\max_y v_y \in y^\star(q)$ whenever $v \in v^\star_{\mathrm{MM}}(q)$ if Eq. (5.11) is satisfied. The following Theorem 4 .7 shows that the same holds true for the other losses whenever they embed L (or $2L$).

Theorem 4 .7. *S_{M} and S_{RM} are consistent to L under the argmax decoding whenever $H_{\mathrm{M}} = 2H_L$ (S_{M} embeds $2L$) and $H_{\mathrm{RM}} = H_L$ (S_{RM} embeds L), respectively.*

5 Limitations of the Approach and Computational Considerations

Fisher consistency does not take into account the hypothesis space. Fisher consistency ensures that the minimizer $g^\star$ of the expected risk $\mathbb{E}_{(x,y)\sim\rho} S(g(x), y)$ over all *measurable functions* provides the Bayes discrete predictor $f^\star$ minimizing the expected classification risk (5.2) through the decoding as $f^\star = d \circ g^\star$. Although this is an important sanity check to produce reliable machine learning, in practice the minimizer $g^\star$ is approximated in a specific hypothesis space $\mathcal{G}$, and Fisher consistency is not always the best property to consider. Another notion that can be more adapted to these settings is $\mathcal{H}$-consistency, where the hypothesis space of scoring functions is fixed (Long and Servedio, 2013).

Comparing the complexity of subgradient computation between losses. One needs to maximize a linear function over the prediction sets $\Delta(y)$ to compute a sub-gradient of the Restricted-Max loss. It can be efficiently performed if a tractable maximization oracle is available on the polytopes $\Delta(y)$'s. If it does not exist, note that these sets can be written as an intersection of the cone $\{u \in \mathbb{R}^k \mid (L_y - L_{y'})^\top u \leq 0, \forall y' \in \mathcal{Y}\}$ with the simplex. This problem can be converted into a bi-linear max-min problem of the form $\max_{q\in\Delta} \min_{\lambda\succeq 0}$ by constructing the Lagrangian. This has to be compared with the bi-linear max-min problem for the Max-Min loss, which is of the form $\max_{q\in\Delta} \min_{p\in\Delta}$ (Nowak-Vila et al., 2020). Both problems can be solved using efficient saddle-point algorithms such as Saddle-Point Mirror-Prox (Nemirovski, 2004). It is interesting to study whether computing the subgradient of S_{RM} is easier than computing the one of S_{MM}, as then S_{RM} will be more attractive for the practitioner than S_{MM}. However, this question is out of the scope of this paper and we leave it for future work.

6 Conclusion and Future Directions

In this work, we have analyzed the consistency properties of the well-known Max loss for general classification tasks. We show that a restrictive condition on the task, which is not satisfied by most of the used losses in practice, is necessary for Fisher consistency. We also show consistency for the absolute deviation loss used in ordinal regression, and more generally to any distance defined in a tree, showing that the necessary condition is meaningful. Moreover, we have overcome this limitation and introduced a novel generalization of the binary SVM loss called Restricted-Max loss, which maintains the maximization over the loss-augmented scores and it is consistent under milder conditions on the task at hand. Several questions remain unanswered, such as whether the proposed necessary condition for the Max loss is sufficient or not, and whether there exists

tractable linear maximization oracles over the prediction sets for specific structured prediction problems, which would make the Restricted-Max loss more attractive than the Max-Min loss.

Appendices

Outline. The supplementary material is organized as follows. In Appendix A , we prove general results on embeddings of losses, we compute the Bayes risks for each of the losses and we provide an algebraic characterization of the extreme points of the prediction sets. In Appendix B and Appendix C , we provide the main results of the Max loss and the Restricted-Max loss, respectively.

A Preliminary Results

A .1 Results on Embeddability of Losses

Proposition A .1. *Let $\psi : \mathcal{Y} \to \mathbb{R}^k$ be an embedding of the output space. If $H_S = H_L$ and $S(\psi(y)) = L_y$ for all $y \in \mathcal{Y}$, then S embeds L with embedding ψ.*

Proof. To prove that S embeds L with embedding $\psi(y) = -L_y$, we need to show that

$$y \in y^\star(q) \iff \varphi(y) \in v^\star(q).$$

If $y \in y^\star(q)$, then

$$H_L(q) = L_y^\top q = S(\varphi(y))^\top q = H(q) = \min_{v \in \mathbb{R}^k} S(v)^\top q.$$

Thus, $S(\varphi(y))^\top q = \min_{v \in \mathbb{R}^k} S(v)^\top q$ implies that necessarily $\varphi(y) \in v^\star(q)$. Similarly, if $\varphi(y) \in v^\star(q)$, then $\min_z L_z^\top q = L_y^\top q$ which implies $y \in y^\star(q)$. □

A .2 Bayes risk identities

The following Lemma A .2 provides an identity which will be useful to provide the forms of the Bayes risk for S_{M} and S_{RM}.

Lemma A .2. *Let $\mathcal{C}_y \subseteq \Delta$, $\Omega^y(q) = -L_y^\top q + i_{\mathcal{C}_y}(q)$ and $S(v, y) = (\Omega^y)^*(v) - v_y$ for every $y \in \mathcal{Y}$. Then,*

$$H_S(q) = \max_{\substack{\sum_y q_y \nu_y = q \\ \nu_y \in \mathcal{C}_y}} \sum_y q_y L_y^\top \nu_y. \tag{12}$$

Proof. Recall the definition of the Bayes risk $H(q) = \min_{v \in \mathbb{R}^k} S(v)^\top q$. Using the structural assumption on S, we can re-write it as

$$H(q) = \min_{v \in \mathbb{R}^k} \sum_{y \in \mathcal{Y}} q_y (\Omega^y)^*(v) - v^\top q = -\max_{v \in \mathbb{R}^k} v^\top q - \sum_{y \in \mathcal{Y}} q_y (\Omega^y)^*(v) = -\Big(\sum_{y \in \mathcal{Y}} q_y (\Omega^y)^* \Big)^*(q).$$

Recall that if the functions h_i are convex, then the conjugate of the sum is the infimum convolution of the individual conjugates (Rockafellar, 1997) as

$$\Big(\sum_i h_i\Big)^*(t) = \min_{\sum_i x_i = t} \sum_i h_i^*(x_i).$$

If we apply this property to the functions $h_i = q_i(\Omega^y)^*$, we obtain:

$$\begin{aligned}
-\Big(\sum_{y\in\mathcal{Y}} q_y(\Omega^y)^*\Big)^*(q) &= -\min_{\sum_{y\in\mathcal{Y}}\nu_y = q} \sum_{y\in\mathcal{Y}} (q_y(\Omega^y)^*)^*(\nu_y) \\
&= -\min_{\sum_{y\in\mathcal{Y}}\nu_y = q} \sum_{y\in\mathcal{Y}} q_y \Omega^y(\nu_y/q_y) \qquad (ah)^*(x) = ah^*(x/a) \\
&= -\min_{\substack{\sum_{y\in\mathcal{Y}}\nu_y = q \\ \nu_y/q_y\in\mathcal{C}_y,\ \forall y\in\mathcal{Y}}} -\sum_{y\in\mathcal{Y}} L_y^\top \nu_y \qquad \Omega^y(q) = -L_y^\top q + i_{\mathcal{C}_y}(q) \\
&= \max_{\substack{\sum_{y\in\mathcal{Y}}\nu_y = q \\ \nu_y/q_y\in\mathcal{C}_y,\ \forall y\in\mathcal{Y}}} \sum_{y\in\mathcal{Y}} L_y^\top \nu_y \qquad \text{redefine } \nu_y \text{ as } \nu_y/q_y \\
&= \max_{\substack{\sum_{y\in\mathcal{Y}} q_y\nu_y = q \\ \nu_y\in\mathcal{C}_y,\ \forall y\in\mathcal{Y}}} \sum_{y\in\mathcal{Y}} q_y L_y^\top \nu_y.
\end{aligned}$$

□

The following Proposition A .3 provides us with the first part of Proposition 3 .4.

Proposition A .3 (Bayes risks). *For all $q \in \Delta$, the Bayes risks read*

$$\begin{aligned}
H_{\mathrm{MM}}(q) &= \min_{y\in\mathcal{Y}} L_y^\top q = H_L(q) \\
H_{\mathrm{M}}(q) &= \max_{Q\in U(q,q)} \langle L, Q\rangle_{\mathrm{F}} \\
H_{\mathrm{RM}}(q) &= \max_{Q\in U(q,q)\cap\mathcal{C}_L} \langle L, Q\rangle_{\mathrm{F}},
\end{aligned}$$

where

$$U(q,q) = \{Q \in \mathbb{R}_+^{k\times k} \,|\, Q1 = q, Q^\top 1 = q\}, \quad \textit{and} \quad \mathcal{C}_L = \{Q \in \mathbb{R}^{k\times k} \,|\, (1L_y^\top - L)Q_y \preceq 0,\ \forall y \in \mathcal{Y}\}.$$

Proof. The first identity is trivial and has already been derived in the main body of the paper. We use the above Lemma A .2 to obtain the identities corresponding to S_{M} and S_{RM}.

1. *Bayes risk of Max loss:* In this case $\mathcal{C}_y = \Delta$. If we define $\Gamma \in \mathbb{R}^{k\times k}$ as the matrix whose rows are ν_y, the maximization reads

$$\max_{\substack{\Gamma^\top q = q \\ \Gamma 1 = 1 \\ \Gamma \succeq 0}} \sum_{y\in\mathcal{Y}} q_y L_y^\top \Gamma_y$$

If we now define $Q \in \mathbb{R}^{k\times k}$ as $Q = \mathrm{diag}(q)\Gamma$, i.e., $Q_y = q_y\Gamma_y$, the objective can be re-written as a matrix scalar product as

$$\sum_{y\in\mathcal{Y}} q_y L_y^\top \Gamma_y = \sum_{y\in\mathcal{Y}} L_y^\top (q_y\Gamma_y) = \sum_{y\in\mathcal{Y}} L_y^\top Q_y = \langle L, Q\rangle_{\mathrm{F}}.$$

Whenever $q_y > 0$, the change of variables $Q_y = q_y \Gamma_y$ is invertible and the constraints satisfy

$$\begin{aligned}(Q^\top 1)_y = q_y &\iff (\Gamma^\top q)_y = q_y \\ (Q1)_y = q_y &\iff (\Gamma 1)_y = 1 \\ Q_y \succeq 0 &\iff \Gamma_y \succeq 0.\end{aligned}$$

On the other hand, if $q_y = 0$ then $Q_y = 0$ but the objective is not affected as it is independent of Γ_y.

2. *Bayes risk of Restricted-Max loss:* In this case $\mathcal{C}_y = \Delta(y) = \{q \in \Delta \mid (L_y - L_{y'})^\top q \preceq 0, \forall y' \in \mathcal{Y}\}$. The maximization now reads

$$\max_{\substack{\Gamma^\top q = q \\ \Gamma 1 = 1 \\ \Gamma \succeq 0 \\ (1L_y^\top - L)\Gamma_y \preceq 0, \forall y}} \quad \sum_{y \in \mathcal{Y}} q_y L_y^\top \Gamma_y.$$

The result follows as $(1L_y^\top - L)\Gamma_y \preceq 0$ if and only if $(1L_y^\top - L)Q_y \preceq 0$ whenever $q_y > 0$.

□

A .3 Extreme points of a polytope

We will need to analyse the extreme points of the polytope $\mathcal{P} = \{(q, u) \in \mathbb{R}^{k+1} \mid q \in \Delta, L_y^\top q \geq u, \forall y \in \mathcal{Y}\} \subseteq \mathbb{R}^{k+1}$ in the proof of the sufficient condition for consistency of Max loss in Appendix B .3.

Algebraic characterization of extreme points of a polyhedron. The following Proposition A .4 provides us with an algebraic characterization of the extreme points of a polyhedron $\mathcal{Q} = \{x \in \mathbb{R}^n \mid Ax \succeq b\}$.

Proposition A .4 (Theorem 3.17 of Andreasson et al. (2020))**.** *Let $x \in \mathcal{Q} = \{x \in \mathbb{R}^n \mid Ax \succeq b\}$, where $A \in \mathbb{R}^{m \times n}$ has $\operatorname{rank}(A) = n$ and $b \in \mathbb{R}^m$. Let $I \subseteq [m]$ be a set of indexes for which the subsystem is an equality, i.e., $A_I x_I = b_I$ with $Ax_I \succeq b$. Then x_I is an extreme point of $\mathcal{Q}$ if and only if $\operatorname{rank}(A_S) = n$.*

Let $\mathcal{P} \subseteq \mathbb{R}^{k+1}$ be the polyhedron defined as

$$\mathcal{P} = \{(q, u) \in \mathbb{R}^{k+1} \mid q \in \Delta, L_y^\top q \geq u, \forall y \in \mathcal{Y}\} \subseteq \mathbb{R}^{k+1}.$$

The polyhedron $\mathcal{P}$ can be written as $\mathcal{P} = \{x = (q, u) \in \mathbb{R}^{k+1} \mid Ax \succeq b\}$ where

$$\begin{array}{c} S \\ \\ \\ T \\ \\ \\ \end{array} \left\{ \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \right. \underbrace{\left(\begin{array}{c|c} & \vdots \\ \mathbf{L} & -1 \\ & \vdots \\ \hline & \vdots \\ \mathbf{Id} & 0 \\ & \vdots \\ \hline \cdots \ 1 \ \cdots & 0 \\ \cdots \ -1 \ \cdots & 0 \end{array} \right)}_{A} \left(\begin{array}{c} q \\ \hline u \end{array} \right) \succeq \underbrace{\left(\begin{array}{c} \vdots \\ 0 \\ \vdots \\ \hline \vdots \\ 0 \\ \vdots \\ \hline 1 \\ -1 \end{array} \right)}_{b}, \tag{13}$$

with $A \in \mathbb{R}^{(2k+2)\times(k+1)}$ and $b \in \mathbb{R}^{k+1}$. Note that $\mathrm{rank}(A) = k+1$. Given $x = (q, u)$, define $S, T \subseteq \mathcal{Y}$ as the subsets of outputs such that

$$y \in S \iff L_y^\top q = u, \quad y \in T \iff q_y = 0, \tag{14}$$

i.e., S and T correspond to the indexes of the first and second block of the matrix A for which the inequality holds as an equality, respectively. More concretely, if I are the indices of A for which $A_I x_I = b_I$, we have that $I = S \cup (k+T) \cup \{2k+1\} \cup \{2k+1\}$, because the last two inequalities must be an equality as $q \in \Delta$. Moreover, the sets S have the following properties:

- We necessarily have $|S| \geq 1$: if $S = \emptyset$, then $\mathrm{rank}(A_I) = k$ and so the rank is not maximal, thus $x = (q, u)$ cannot be an extreme point.
- We necessarily have $|S| + |T| \geq k$ (using the fact that $\mathrm{rank}(A) = k+1$).

B Results on Max Loss

B .1 Bayes Risk of Max loss for Symmetric Losses

The following Proposition B .1 gives another expression of the Bayes risk of S_{M} and its Fenchel conjugate assuming the loss L is symmetric.

Proposition B .1. *Let $H_{\mathrm{M}}(q) = \max_{Q \in U(q,q)} \langle L, Q\rangle_{\mathrm{F}}$ and assume L symmetric. Then, the following identities hold:*

$$H_{\mathrm{M}}(q) = \min_{\frac{1}{2}(a_y + a_{y'}) \geq L(y,y')} a^\top q, \quad \forall q \in \Delta,$$
$$(-H_{\mathrm{M}})^*(v) = \max_{y,y' \in \mathcal{Y}} L(y,y') + \frac{v_y + v_{y'}}{2}, \quad \forall v \in \mathbb{R}^k.$$

Proof. The first part corresponds to the dual of the maximization problem defining the Bayes risk H_{M} when L is symmetric:

$$\begin{aligned} -\min_{Q \in U(q,q)} -\langle L, Q\rangle_{\mathrm{F}} &= \min_{Q \succeq 0} \max_{a,b \in \mathbb{R}^k} a^\top(Q1 - q) + b^\top(Q^\top 1 - q) - \langle L, Q\rangle_{\mathrm{F}} \\ &= -\max_{a,b \in \mathbb{R}^k} -a^\top q - b^\top q + \min_{Q \succeq 0} a^\top Q 1 + b^\top Q^\top 1 - \langle L, Q\rangle_{\mathrm{F}} \end{aligned}$$

We can now re-write the minimization objective as a matrix scalar product with Q as $a^\top Q 1 = \mathrm{Tr}(a^\top Q 1) = \mathrm{Tr}(\mathrm{Q1a}^\top) = \langle Q, a1^\top\rangle_{\mathrm{F}}$ and analogously $b^\top Q 1 = \langle Q, 1b^\top\rangle_{\mathrm{F}}$. Hence, the objective of the minimum becomes $\langle a1^\top + 1b^\top - L, Q\rangle_{\mathrm{F}}$, which gives

$$\min_{Q \succeq 0} \langle a1^\top + 1b^\top - L, Q\rangle_{\mathrm{F}} = \begin{cases} 0 & \text{if } a1^\top + 1b^\top - L \succeq 0 \\ -\infty & \text{otherwise} \end{cases}.$$

We obtain the following minimization problem in $a, b \in \mathbb{R}^k$

$$-\max_{a1^\top + 1b^\top \succeq L} -(a+b)^\top q = \min_{a1^\top + 1b^\top \succeq L} (a+b)^\top q.$$

Using that L is symmetric, we can add the constraint $a = b$. In order to see this, let $(a^\star, b^\star)$ be a solution of the linear problem. If L is symmetric, then $(b^\star, a^\star)$ is also a solution, which implies that $\frac{1}{2}(a^\star + b^\star, a^\star + b^\star)$ too. Hence, we can assume $a = b$ and we obtain the desired result.

For the second part, note that if L is symmetric, the matrix Q can be assumed also symmetric. To see this, let $Q^\star = \arg\max_{Q \in U(q,q)} \langle L, Q \rangle_{\mathrm{F}}$. Then if L symmetric $(Q^\star)^\top$ is also a solution, which means that $\frac{1}{2}(Q^\star + (Q^\star)^\top)$ too, which is symmetric. Hence, we can write

$$\begin{aligned} H_{\mathrm{M}}(q) &= \max_{\substack{Q = Q^\top \\ Q1 = q \\ Q \succeq 0}} \langle L, Q \rangle_{\mathrm{F}} \\ &= \min_{v \in \mathbb{R}^k} \max_{\substack{Q = Q^\top \\ Q \in \mathrm{Prob}(\mathcal{Y} \times \mathcal{Y})}} \langle L, Q \rangle_{\mathrm{F}} - v^\top (Q1 - q) \\ &= \min_{v \in \mathbb{R}^k} \Big\{ \underbrace{\max_{\substack{Q = Q^\top \\ Q \in \mathrm{Prob}(\mathcal{Y} \times \mathcal{Y})}} \langle L + v1^\top, Q \rangle_{\mathrm{F}}}_{(-H_{\mathrm{M}})^*(v)} \Big\} - v^\top q, \end{aligned}$$

where at the last step we have used that $q \in \Delta$ and so $1^\top Q 1 = 1$, which together with $Q \succeq 0$ implies $Q \in \mathrm{Prob}(\mathcal{Y} \times \mathcal{Y})$. The extreme points of the problem domain $\{Q \in \mathrm{Prob}(\mathcal{Y} \times \mathcal{Y})\}$ where the maximization of the linear objective is achieved are precisely the points $\{\frac{1}{2}(e_y + e_{y'})\}_{y,y' \in \mathcal{Y}}$. □

B .2 Necessary Conditions for Consistency

Recall that S is consistent to L if there exists a decoding $d : \mathbb{R}^k \to \mathcal{Y}$ such that if $v \in v^\star(q)$, then necessarily $d(v) \in y^\star(q)$ for all $q \in \Delta$. A necessary condition for this to hold is that every level set of $v^\star$ must be included in a level set of $y^\star$, which are precisely the prediction sets.

Lemma B .2. *If S is consistent to L, then for every $v \in \mathbb{R}^k$ there must exist a $y \in \mathcal{Y}$ such that*

$$(v^\star)^{-1}(v) \subseteq (y^\star)^{-1}(y) = \Delta(y). \tag{15}$$

Proof. If (15) does not hold, then there exists $q_1, q_2 \in (v^\star)^{-1}(v)$ with $y^\star(q_1) \cap y^\star(q_2) = \emptyset$. However, Fisher consistency means that $v \in v^\star(q_1)$ implies $d(v) \in y^\star(q_1)$ and $v \in v^\star(q_2)$ implies $d(v) \in y^\star(q_2)$, which is not possible because $y^\star(q_1) \cap y^\star(q_2) = \emptyset$. □

The following Corollary B .4 re-writes (15) in terms of the Bayes risk H_S.

Proposition B .3. *The level sets of $v^\star$ are the image of $-\partial(-H_S)^* : \mathbb{R}^k \to 2^\Delta$, i.e.,*

$$\mathrm{Im}((v^\star)^{-1}) = \mathrm{Im}(-\partial(-H_S)^*).$$

Proof. First of all, note that $-\partial(-H_S)^* = (\partial H_{\mathrm{M}})^{-1}$ (Rockafellar, 1997). We have

$$H_S(q) = \min_{v \in \mathbb{R}^k} S(v)^\top q + i_\Delta(q), \quad \partial H_S(q) = S(\bar{v}) + \langle 1 \rangle, \ \bar{v} \in v^\star(q).$$

Let's now prove the two inclusions.

($\subseteq$): Let $Q \in \mathrm{Im}((v^\star)^{-1})$. This means that there exists $V \in \mathbb{R}^k$ such that $V = \arg\min_{v\in\mathbb{R}^k} S(v)^\top q$ for all $q \in Q$. If we define $T = S(V) + \langle 1\rangle$, then $T = \partial H_{\mathrm{M}}(q)$ for all $q \in Q$.

($\supseteq$): Let $Q \in \mathrm{Im}((\partial H_{\mathrm{M}})^{-1})$. This means that there exists T such that $T = \partial H_{\mathrm{M}}(q)$ for all $q \in Q$. For every $q \in Q$, the set T can be written as $T = S(v^\star(q)) + \langle 1\rangle$. To show that $v^\star(q) = v^\star(q')$ for all $q, q' \in Q$, we need to show that if $S(v) = S(v') + c1, v \in v^\star(q), v' \in v^\star(q')$ for some $q, q' \in Q$, then necessarily $c = 0$. This is because $S(v(q))^\top q' \geq S(v(q'))^\top q' \implies c \geq 0$ and $S(v(q))^\top q \leq S(v(q'))^\top q \implies c \leq 0$.

□

Corollary B .4 (Necessary condition for consistency)**.** *If S is Fisher consistent to L, then for every $v \in \mathbb{R}^k$, there exists $y \in \mathcal{Y}$ such that*

$$-\partial(-H_S)^*(v) \subseteq \Delta(y). \tag{16}$$

Proof. This follows directly from Eq. (15) and Proposition B .3. □

Proposition B .5 (Weaker necessary condition for consistency)**.** *If S is consistent to L, then every extreme point of $\Delta(y)$ for some $y \in \mathcal{Y}$ must be a 0-dimensional image of $-\partial(-H_S)^*$.*

Proof. Let $\Delta_S(v) = -\partial(-H_S)^*(v)$. There exists a finite set $\mathcal{V} \subseteq \mathbb{R}^k$ such that $\bigcup_{v\in\mathcal{V}} \Delta_S(v) = \Delta(y)$. In particular, if q is an extreme point of $\Delta(y)$, then there exists $v \in \mathcal{V}$ such that $q \in \Delta_S(v)$. We need to show that q is also an extreme point of $\Delta_S(v)$. Indeed, if $\Delta_S(v) \subseteq \Delta(y)$ are polyhedrons and $q \in \Delta_S(v), \Delta(y)$ is an extreme point of $\Delta(y)$, then it is also necessarily an extreme point of $\Delta_S(v)$. □

Theorem B .6. *Let L be a symmetric loss with $k > 2$. If the Max loss is consistent to L, then L is a distance, and for every three outputs $y_1, y_2, y_3 \in \mathcal{Y}$, there exists $z \in \mathcal{Y}$ for which these the following three identities are satisfied:*

$$\begin{aligned} L(y_1, y_2) &= L(y_1, z) + L(z, y_2), \\ L(y_1, y_3) &= L(y_1, z) + L(z, y_3), \\ L(y_2, y_3) &= L(y_2, z) + L(z, y_3). \end{aligned}$$

Proof. From Proposition B .1 and Proposition B .5, we obtain that if the Max loss is consistent to L, then the extreme points of the prediction sets $\Delta(y)$'s have to be of the form $1/2(e_y + e_{y'})$. Hence, the projection of the sets $\Delta(y)$'s into a three-dimensional simplex can only be of the form depicted in Figure 4. The necessary condition follows directly from these possibilities (see caption of Figure 4). Moreover, note that if the three identities of the theorem hold, then L is a distance. To see that, note that the triangle inequality holds for any triplet $y_1, y_2, y_3 \in \mathcal{Y}$ as:

$$\begin{aligned} L(y_1, y_2) &= L(y_1, z) + L(z, y_2) \\ &= L(y_1, y_3) - L(z, y_3) + L(y_3, y_2) - L(y_3, z) \\ &= L(y_1, y_3) + L(y_3, y_1) - 2L(y_3, z) \\ &\leq L(y_1, y_3) + L(y_3, y_1). \end{aligned}$$

□

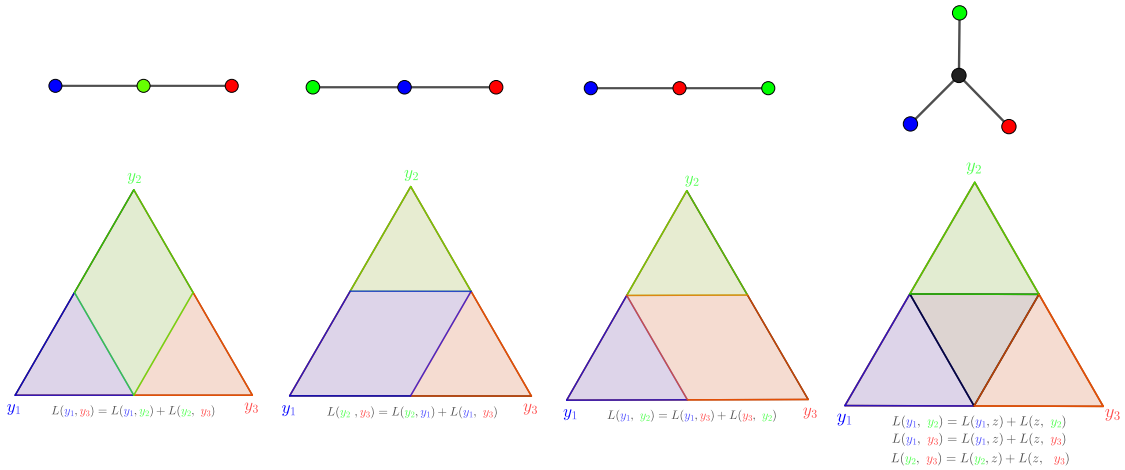


Figure 4: These are the only possible possibilities of the prediction sets in a three-dimensional face of the simplex. The equations associated to each configuration is written below the corresponding simplex and all together can be compactly written as the necessary condition given by the theorem. An edge from a corner of the simplex to the middle point of the opposite side is not possible as $L(y, y') = 0$ if and only if $y = y'$ by assumption.

Examples of losses not satisfying the necessary condition. We now show that the examples exposed in Section 2 .2 do not satisfy the necessary condition of Theorem B .6.

Lemma B .7. *Let L such with full rank loss matrix L and existing $q \in \operatorname{int}(\Delta)$ for which all outputs optimal, i.e., $y^\star(q) = \mathcal{Y}$. Then, the Max loss is* not *consistent to L.*

Proof. The point q, which is not of the form $1/2(e_y + e_{y'})$ for some $y, y' \in \mathcal{Y}$, is an extreme point of the polytope $\Delta(y) = \{q \in \Delta \mid L_z^\top q \geq L_y^\top q, \forall z \in \mathcal{Y}\}$ for every $y \in \mathcal{Y}$. This is because $q \in \Delta$ is the unique solution of $Lq = u$ with $u = L_y^\top q$ for all $y \in \mathcal{Y}$. Hence, by Proposition B .5, the Max loss is not consistent to L. □

Lemma B .8. *The Max loss is not consistent to the the Hamming loss on permutations $L(\sigma, \sigma') = \frac{1}{M}\sum_{m=1}^{M} 1(\sigma(m) \neq \sigma'(m))$ where σ, σ' permutations of size M.*

Proof. Take the transpositions $\sigma_1 = (3, 2), \sigma_2 = (2, 1), \sigma_3 = (3, 1)$. We have that $L(\sigma_i, \sigma_j) = 2/M$ for $i \neq j$ and $L(\sigma, \sigma') > \frac{2}{M}$ for all permutations $\sigma \neq \sigma'$. Hence, the necessary condition can't be satisfied. □

The Hamming loss with $M = k = 2$ is consistent and it is not defined in a tree. The Hamming loss $L(y, y') = \frac{1}{2}(1(y_1 \neq y_1') + 1(y_2 \neq y_2'))$ is consistent as it decomposes additively and each term is consistent as it is the binary 0-1 loss. However, it can't be described as the shortest path distance in a tree, but rather the shortest path distance in a cycle of size four with all weights equal to $1/2$.

B .3 Sufficient condition on the discrete loss L

Theorem B .9. *If L is a distance defined in a tree, then the Max loss S_{M} embeds L with embedding $\psi(y) = -L_y$ and it is consistent under the argmax decoding.*

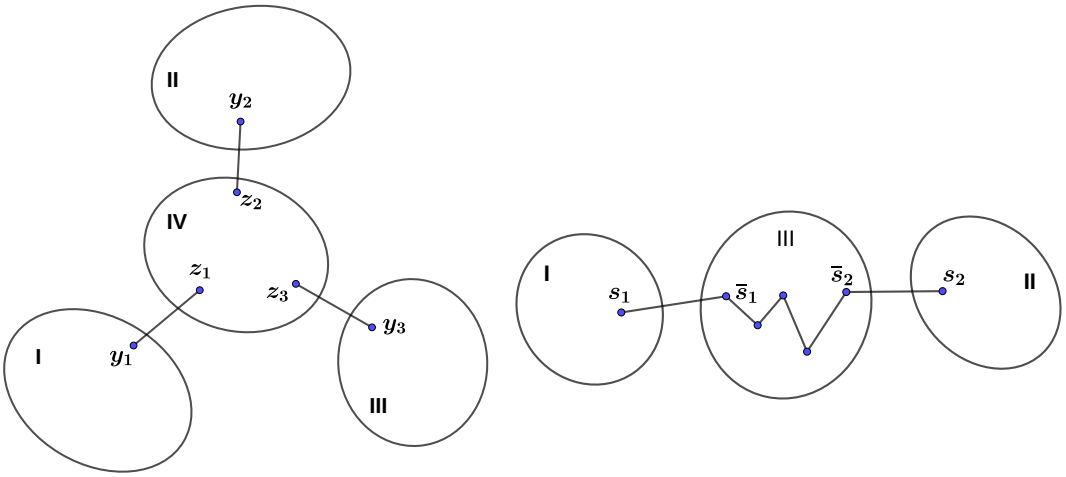


Figure 5: **Left:** The 4-partition of the output space where $y_1, y_2, y_3 \in S$. The different sets only communicate through the edges $z_i - y_i$ for $i = 1, 2, 3$ respectively. **Right:** The 3-partition of the output space where $s_1, s_2 \in S$ and the depicted path is the path between these two points.

Proof. We just have to prove that the extremes points of the polytope $\mathcal{P}$ defined as

$$\mathcal{P} = \{(q, u) \in \mathbb{R}^{k+1} \mid q \in \Delta, L_y^\top q \geq u, \forall y \in \mathcal{Y}\} \subseteq \mathbb{R}^{k+1}$$

are of the form $(\frac{1}{2}(e_y + e_{y'}), \frac{1}{2}L(y, y'))$, where $y, y' \in \mathcal{Y}$. Using this, we obtain

$$\begin{aligned}
(-H_{2L})^*(v) &= \max_{q \in \Delta} \min_{z \in \mathcal{Y}} 2L_z^\top q + v^\top q \\
&= \max_{(q,u) \in \mathcal{P}} 2u + v^\top q \\
&= \max_{(q,u) \in \operatorname{ext}(\mathcal{P})} 2u + v^\top q \\
&= \max_{y, y' \in \mathcal{Y}} L(y, y') + \frac{v_y + v_{y'}}{2} = (-H_{\mathrm{M}})^*(v),
\end{aligned}$$

for all $v \in \mathbb{R}^k$. This implies $H_{2L} = 2H_L = H_{\mathrm{M}}$. We will use the algebraic framework introduced in Appendix A .3.

Let $x \in \operatorname{ext}(\mathcal{P})$ and S are T the sets of indices for which $L_y^\top q = u$ and $q_{y'} = 0$ for $y \in S$ and $y' \in T$ (as defined in (14)).

If $|S| = 1$, then we necessarily. have $|T| = k - 1$ because $|T| \geq k - |S| = k - 1$ and $|T| = k$ is not possible because q is in the simplex. In this case, the extreme point is equal to $q = (e_y, 0)$, which is of the desired form.

First part of the proof. If $|S| \geq 2$, let's prove that the elements in S must be necessarily aligned, i.e., contained in a chain (always true for $|S| = 2$). If we denote by $\mathrm{SP}(s, s') \subseteq \mathcal{Y}$ the elements in the shortest path between s, s', this means that there exists $s_1, s_2 \in S$ such that $s \in \mathrm{SP}(s_1, s_2)$ for all $s \in S$.

If the elements in S are not aligned, then there exist pairwise different elements $y_1, y_2, y_3 \in S$ and $z_1, z_2, z_3 \in \mathcal{Y}$ (possibly repeated) such that the tree defining the loss L can be partitioned into four sets $\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}$ of the form depicted in the left Figure 5, where the edges $y_i - z_i$ belong to the tree for $i = 1, 2, 3$.

$$
\begin{aligned}
L_{z_1}^\top q &= \sum_{y\in \mathrm{I}} L(z_1, y)q_y + \sum_{\substack{y\notin \mathrm{I}\\ y\neq z_1}} L(z_1, y)q_y \\
&= \sum_{y\in \mathrm{I}} (L(z_1, y_1) + L(y_1, y))q_y + \sum_{\substack{y\notin \mathrm{I}\\ y\neq z_1}} (L(y_1, y) - L(z_1, y_1))q_y \\
&= \sum_{y\neq z_1} L(y_1, y)q_y + \Big(\sum_{y\in \mathrm{I}} q_i - \sum_{\substack{y\notin \mathrm{I}\\ i\neq z_1}} q_y\Big)L(z_1, y_1) \\
&= \sum_{y\in \mathcal{Y}} L(y_1, y)q_y + \Big(\sum_{y\in \mathrm{I}} q_i - \sum_{y\notin \mathrm{I}} q_y\Big)L(z_1, y_1) \\
&= u + \Big(\sum_{y\in \mathrm{I}} q_y - \sum_{y\notin \mathrm{I}} q_y\Big)L(z_1, y_1).
\end{aligned}
$$

If we repeat the same procedure for II and III, we obtain that

$$
\begin{cases}
L_{z_1}^\top q &= u + \Big(\sum_{y\in \mathrm{I}} q_y - \sum_{y\notin \mathrm{I}} q_y\Big)L(z_1, y_1) \\
L_{z_2}^\top q &= u + \Big(\sum_{y\in \mathrm{II}} q_y - \sum_{y\notin \mathrm{II}} q_y\Big)L(z_2, y_2) \\
L_{z_3}^\top q &= u + \Big(\sum_{y\in \mathrm{III}} q_y - \sum_{y\notin \mathrm{III}} q_y\Big)L(z_3, y_3)
\end{cases}
$$

However, note that $\sum_{y\in \mathrm{I}} q_y + \sum_{y\in \mathrm{II}} q_y + \sum_{y\in \mathrm{III}} q_y + \sum_{y\in \mathrm{IV}} q_y = 1$, which implies that

$$
\min_{\mathcal{A}\in\{\mathrm{I},\mathrm{II},\mathrm{III}\}} \Big(\sum_{y\in\mathcal{A}} q_y - \sum_{y\notin\mathcal{A}} q_y\Big) < 0.
$$

Hence, there exists $i \in \{1, 2, 3\}$ for which $L_{z_i}^\top q < u$, which leads to a contradiction as $L_y^\top q \geq u$ for all $y \in \mathcal{Y}$.

Second part of the proof. Now that we have that the elements in S must be aligned, let's proceed with the proof by analyzing separately particular cases:

- ($\boldsymbol{S \cap T = \emptyset}$): This means that $q_s > 0$ for all $s \in S$. Let $x = (\frac{1}{2}(e_{s_1} + e_{s_2}), \frac{1}{2}L(s_1, s_2))$, where $S \subseteq \mathrm{SP}(s_1, s_2)$. Then, it satisfies the equality constraints as $L_s q = 1/2(L(s_1, s)+L(s, s_2)) = 1/2L(s_1, s_2)$ because $s \in \mathrm{SP}(s_1, s_2)$ for all $s \in S$. Hence, it has to be equal to the unique solution of the linear system of equations.
- ($\boldsymbol{S \cap T \neq \emptyset}$): Let's separate into two more cases.
  - ($\boldsymbol{\exists r_1, r_2 \in [k] \setminus T}$ **such that** $\boldsymbol{S \subseteq \mathrm{SP}(r_1, r_2)}$): Let $x = (\frac{1}{2}(e_{r_1}+e_{r_2}), \frac{1}{2}L(r_1, r_2))$. Then, it satisfies the equality constraints as $L_s q = 1/2(L(r_1, s)+L(s, r_2)) = 1/2L(r_1, r_2)$ because $s \in \mathrm{SP}(r_1, r_2)$ for all $s \in S$. Hence, it has to be equal to the unique solution of the linear system of equations.
  - ($\boldsymbol{\nexists r_1, r_2 \in [k] \setminus T}$ **such that** $\boldsymbol{S \subseteq \mathrm{SP}(r_1, r_2)}$): We will show that this case is not possible. Consider the shortest path between s_1 and s_2 in S and the partition of the vertices of the tree into the sets $\mathrm{I}, \mathrm{II}, \mathrm{III}$ depicted in the right Figure 5. We know that

$$
\{\mathrm{I} \cap ([k] \setminus T) = \emptyset\} \vee \{\mathrm{II} \cap ([k] \setminus T) = \emptyset\}.
$$

If this is not true, then taking $r_1 \in \mathrm{I}$ and $r_2 \in \mathrm{II}$ we obtain $S \subseteq \mathrm{SP}(r_1, r_2)$. Assume that $\mathrm{I} \cap ([k] \setminus T) = \emptyset$. We have that $L(s_1, y) > L(\bar{s}_1, y)$ for all $y \in [k] \setminus T$, which means that

$$L_{s_1}^\top q > L_{\bar{s}_1}^\top q.$$

This is a contradiction because $L_y^\top q \geq u = L_{s_1}^\top q$ as $s_1 \in S$. The case $\mathrm{II} \cap ([k] \setminus T) = \emptyset$ can be done analogously.

Third part of the proof. By Proposition A .1, to prove that S_{M} embeds $2L$ with embedding $\psi(y) = -L_y$, we only need to show that $S_{\mathrm{M}}(-L_y) = 2L_y$. For every $z \in \mathcal{Y}$, we have

$$\begin{aligned} S_{\mathrm{M}}(-L_y, z) &= \max_{y' \in \mathcal{Y}} \; L(z, y') + (-L_y)^\top e_{y'} - (-L_y)^\top e_z \\ &= \max_{y' \in \mathcal{Y}} \; \{L(z, y') - L(y, y')\} + L(y, z) \\ &= 2L(y, z), \end{aligned}$$

where at the last step we have used $L(z, y') - L(y, y') \leq L(y, y')$ as L is a distance and so the maximization is achieved at $y' = y$.

Finally, the argmax decoding is consistent as it is an inverse of the embedding $\psi(y) = -L_y$ as

$$d(\psi(y)) = \arg\max_{y' \in \mathcal{Y}} \; - L(y, y') = \arg\min_{y' \in \mathcal{Y}} \; L(y, y') = y.$$

□

B .4 Partial Consistency through dominant label condition

Lemma B .10. *Let $q \in \Delta$ such that $q_1 \geq 1/2 \geq p_y$ for all $y \neq 1$. If L is a distance, then $L_1^\top q \leq L_y^\top q$ for all $y \in \mathcal{Y}$.*

Proof.

$$\begin{aligned} L_z^\top q &= q_1 L_{z1} + \sum_{y \neq 1, z} L_{zy} q_y \\ &\geq \frac{1}{2} L_{z1} + \sum_{y \neq 1, z} L_{zy} q_y \\ &= \Big(\frac{1}{2} - \sum_{y \neq 1, z} q_y\Big) L_{z1} + \sum_{y \neq 1, z} (L_{z1} + L_{zy}) q_y \\ &\geq \Big(\frac{1}{2} - \sum_{y \neq 1, z} q_y\Big) L_{z1} + \sum_{y \neq 1, z} L_{1y} q_y \\ &\geq L_{z1} q_z + \sum_{y \neq 1, z} L_{1y} q_y \\ &= \sum_{y \neq 1} L_{1y} q_y = L_1^\top q. \end{aligned}$$

□

Lemma B .11. *If L is a distance, then $H_{\mathrm{M}} \leq 2H_L$.*

Proof. In particular, L is also symmetric. Recall that

$$H_L(q) = \min_{y \in \mathcal{Y}} L_y^\top q = \min_{\substack{a \succeq Lp, \\ p \in \Delta.}} a^\top q = \min_{a \in \mathcal{P}_L} a^\top q$$
$$\frac{1}{2} H_{\mathrm{M}}(q) = \min_{1a^\top + a1^\top \succeq L} a^\top q = \min_{a \in \mathcal{P}_L^{\mathrm{M}}} a^\top q,$$

where the second expression is given by Proposition A .3. To show that $2H_{\mathrm{M}} \leq H_L$ we will show that $\mathcal{P}_L \subseteq \mathcal{P}_L^{\mathrm{M}}$. If $a \in \mathcal{P}_L$, then there exists $p \in \Delta$ such that $a \succeq Lp$. Moreover, if L is a distance, it means that the triangle inequality $L(y, y') \leq L(y, z) + L(z, y')$, which is equivalent to $L \preceq 1L_z^\top + L_z 1^\top$ for all $z \in \mathcal{Z}$. This can also be written as

$$L \preceq 1p^\top L + Lp1^\top, \quad \forall p \in \Delta.$$

Finally, note that if $Lp \preceq a$, then $Lp1^\top \preceq a1^\top$, and the same holds for its transpose $1p^\top L \preceq 1a^\top$. Hence, we obtain that $L \preceq 1a^\top + a1^\top$, which is equivalent to $a \in \mathcal{P}_L^{\mathrm{M}}$. □

Proposition B .12. *Assume that L is a distance. Then $\frac{1}{2} H_{\mathrm{M}}(q) = H_L(q)$, for all $q \in \Delta$ satisfying $\|q\|_\infty \geq 1/2$. Moreover, under this condition on q, it is calibrated with the argmax decoding.*

Proof. Combining Lemma B .10 and Lemma B .11 gives

$$H_{\mathrm{M}}(q) \leq 2H_L(q) = 2L_y^\top q,$$

for all $q \in \Delta$ such that $p_y \geq 1/2 \geq p_z$ for all $z \neq y$. Hence, in order to prove the equality at these dominant label points, we just need to find a matrix $Q \in U(q, q)$ such that $\langle L, Q \rangle_{\mathrm{F}} = 2L_y^\top q$. We define the matrix Q as

$$Q_{ij} = \begin{cases} q_i & \text{if } (i = y) \wedge (i \neq j) \\ q_j & \text{if } (j = y) \wedge (j \neq i) \\ 2q_y - 1 & \text{if } i = j = y \\ 0 & \text{otherwise.} \end{cases}.$$

The matrix Q has q at the y-th row and y-th column and $2q_y - 1$ at the crossing point, instead of $2q_y$. The matrix is in $U(q, q)$ as the sum of the rows and columns gives q and it is non-negative because $2q_y - 1 \geq 0$ by assumption. Moreover, the objective satisfies

$$\begin{aligned} \langle L, Q \rangle_{\mathrm{F}} = \sum_{y \in \mathcal{Y}} L_{y'}^\top Q_{y'} &= L_y^\top Q_y + \sum_{y' \neq y} L_{y'} Q_{y'} \\ &= L_y^\top q + \sum_{y' \neq y} q_{y'} L_{y'} e_y \\ &= L_y^\top q + \sum_{y' \neq y} q_{y'} L(y, y') = 2L_y^\top q. \end{aligned}$$

The first part of the result follows. For the second part, we show that $v^\star_{\mathrm{M}}(q) = -L_y$ at the points q satisfying the assumption. Note that we have $H_{\mathrm{M}}(q) = 2L_y^\top q = S(v^\star(q))^\top q$, so we only need to show $S_{\mathrm{M}}(-L_y) = 2L_y$. For every $z \in \mathcal{Y}$, we have

$$\begin{aligned} S_{\mathrm{M}}(-L_y, z) &= \max_{y' \in \mathcal{Y}} \; L(z, y') + (-L_y)^\top e_{y'} - (-L_y)^\top e_z \\ &= \max_{y' \in \mathcal{Y}} \; \{L(z, y') - L(y, y')\} + L(y, z) \\ &= 2L(y, z), \end{aligned}$$

where at the last step we have used $L(z, y') - L(y, y') \leq L(y, y')$ as L is a distance and so the maximization is achieved at $y' = y$. □

C Proofs on Restricted-Max Loss

The following assumption **A1** will be key to prove our consistency results.
Assumption A1: If q is an extreme point of $\Delta(y')$ for some $y' \in \mathcal{Y}$, then

$$\{q \in \Delta(y)\} \vee \{q_y = 0\}, \quad \forall y \in \mathcal{Y}.$$

The following Lemma C .1 will be useful for the results below.

Lemma C .1. *If **A1** is satisfied, then $\Delta(y) \cap \Delta(y') \neq \emptyset$ for all $y, y' \in \mathcal{Y}$.*

Proof. If $\Delta(y) \cap \Delta(y') = \emptyset$ and **A1** is satisfied, then for every q extreme point of $\Delta(y')$ we have that $q_y = 0$. Hence, the prediction set $\Delta(y)$ is included in the non full-dimensional polyhedron $\Delta \cap \{e_y = 0\}$. As $\Delta = \cup_{y \in \mathcal{Y}} \Delta(y)$, this implies that the point $e_{y'} \in \Delta(y')$ must be necessarily included in another $\Delta(z)$, which can only be possible if $L(z, y) = 0$. However, by assumption $L(z, y) = 0$ if and only if $z = y$. □

The following Lemma C .2 shows that under **A1**, the Restricted-Max loss embeds the loss L, which in turn implies consistency.

Lemma C .2. *Assume **A1**. If q is an extreme point of $\Delta(y)$ for some $y \in \mathcal{Y}$, then*

$$qq^\top \in \operatorname*{arg\,max}_{Q \in U(q,q) \cap \mathcal{C}_L} \langle L, Q \rangle_{\mathrm{F}} \quad \textit{and} \quad \Omega_{\mathrm{MM}}(q) = \Omega_{\mathrm{RM}}(q).$$

Proof. The matrix $qq^\top$ belongs to $U(q, q)$ as $qq^\top 1 = q$ and $qq^\top \succeq 0$ and to $\mathcal{C}_L$ by assumption. Let $z \in y^\star(q)$. We have that

$$\begin{aligned} -\Omega_{\mathrm{RM}}(q) &= \max_{Q \in U(q,q) \cap \mathcal{C}_L} \langle L, Q \rangle_{\mathrm{F}} \\ &\geq \langle L, qq^\top \rangle_{\mathrm{F}} = \sum_y q_y L_y^\top q = \sum_y q_y L_z^\top q \\ &= L_z^\top (1^\top qq^\top) = L_z^\top q = -\Omega_{\mathrm{MM}}(q) \end{aligned}$$

We have shown that $\Omega_{\mathrm{RM}}(q) \leq \Omega_{\mathrm{MM}}(q)$. Combining with Proposition 3 .4 that states $\Omega_{\mathrm{RM}} \geq \Omega_{\mathrm{MM}}$, we obtain that $\Omega_{\mathrm{RM}}(q) = \Omega_{\mathrm{MM}}(q)$. □

Theorem C .3. *If **A1** is satisfied, the Restricted-Max loss embeds L with embedding $\psi(y) = -L_y$ and the loss is consistent to L under the argmax decoding.*

Proof. We split the proof into two parts.

First part: S_{RM} embeds L. Let $z \in y^\star(q)$, so that $q \in \Delta(z)$. We can write q as a convex combination of extreme points of the polytope $\Delta(z)$ as

$$q = \sum_{i=1}^{m} \alpha_i q_i,$$

where $\alpha \in \Delta_m$ and q_i is an extreme point of $\Delta(z)$. The matrix $Q = \sum_{i=1}^m \alpha_i q_i q_i^\top$ belongs to $U(q,q) \cap \mathcal{C}_L$ as:

- $Q \in U(q,q)$: We have $Q1 = \sum_{i=1}^m \alpha_i q_i q_i^\top 1 = \sum_{i=1}^m \alpha_i q_i = q$, the same holds for $Q^\top$ and $\sum_{i=1}^m \alpha_i q_i q_i^\top \succeq 0$.
- $Q \in \mathcal{C}_L$: For all $y \in \mathcal{Y}$, we have $(1L_y^\top - L)Q_y = \sum_{i=1}^m \alpha_i \underbrace{q_{i,y}(1L_y^\top - L)q_i}_{\preceq 0} \preceq 0$.

Moreover, we obtain:

$$\begin{aligned}
-\Omega_{\mathrm{RM}}(q) &= \max_{Q \in U(q,q) \cap \mathcal{C}_L} \langle L, Q \rangle_{\mathrm{F}} \\
&\geq \langle L, Q \rangle_{\mathrm{F}} = \sum_{i=1}^m \alpha_i \langle L, q_i q_i^\top \rangle_{\mathrm{F}} \\
&= -\sum_{i=1}^m \alpha_i \Omega_{\mathrm{RM}}(q_i) = -\sum_{i=1}^m \alpha_i \Omega_{\mathrm{MM}}(q_i) \\
&= \sum_{i=1}^m \alpha_i L_z^\top q_i = L_z^\top q = -\Omega_{\mathrm{MM}}(q)
\end{aligned}$$

We have shown $\Omega_{\mathrm{RM}} \leq \Omega_{\mathrm{MM}}$. Combining with Proposition 3 .4 that states $\Omega_{\mathrm{RM}} \geq \Omega_{\mathrm{MM}}$, we obtain $\Omega_{\mathrm{RM}} = \Omega_{\mathrm{MM}}$.

Second part: the embedding is $\psi(y) = -L_y$. By Proposition A .1, we only need to show that $\psi(z) = -L_z$, i.e.,

$$S_{\mathrm{RM}}(-L_z, y) = \sup_{q \in \Delta(y)} L_y^\top q + (-L_z)^\top q - (-L_z)^\top e_y = L(y,z),$$

which holds whenever

$$\max_{q \in \Delta(y)} (L_y - L_z)^\top q = 0, \tag{17}$$

for all $y, z \in \mathcal{Y}$. Note that by construction $(L_y - L_z)^\top q \leq 0$ for all $q \in \Delta(y)$. Moreover, by Lemma C .1 we have that $\Delta(z) \cap \Delta(y) \neq \emptyset$, so there exists $q \in \Delta(y)$ with $L_y^\top q = L_z^\top q$. Finally, the argmax decoding is consistent as it is an inverse of the embedding $\psi(y) = -L_y$ as

$$d(\psi(y)) = \arg\max_{y' \in \mathcal{Y}} \; -L(y,y') = \arg\min_{y' \in \mathcal{Y}} \; L(y,y') = y.$$

□

Proposition C .4. *Assume that $q \in \Delta(y) \implies q_y > 0$ for all $q \in \Delta$. Then **A1** is satisfied.*

Proof. We will prove that if the Assumption is not satisfied then it exists a vertex of $\Delta(y)$ for some $y \in \mathcal{Y}$ such that $S \cap T \neq \emptyset$. If the Assumption is not satisfied at vertex q, then $\{q \notin \Delta(y)\} \wedge \{q_y > 0\}$, which means in particular that $S \cup T \subsetneq [k]$. This necessarily means that $S \cap T \neq \emptyset$ because we must have $|S| + |T| \geq k$ to have maximal rank as q is a vertex. □

Proposition C .5. *Consistency of the Max loss implies consistency of Restricted-Max loss.*

Proof. From Proposition B .1 and Proposition B .5, we know that if the Max loss is consistent to L, then the extreme points of the prediction sets $\Delta(y)$'s have to be of the form $1/2(e_y + e_{y'})$. We will see that in this case **(A1)** is always satisfied. Indeed, if q is an extreme point of a prediction set, then is of the form $q = 1/2(e_y + e_{y'})$, which satisfies $\{q \in \Delta(z)\} \vee \{q_z = 0\}$ for all $z \in \mathcal{Y}$, because $q \in \Delta(z)$ if $z \in \{y, y'\}$ and $q_z = 0$ otherwise. □

Conclusion and Future Directions

The research presented in this thesis is towards a general framework for surrogate methods in the context of structured prediction. Both the structure of this thesis and the underlying theory clearly discloses the distinction between two major types of surrogate strategies: probabilistic estimators and non-probabilistic estimators. The first ones are generally smooth, have a quadratic-type calibration function, are consistent to any discrete loss over the output space, have a unique consistent decoding for each discrete loss and have a continuous surrogate Bayes predictor. The major type from the second group are polyhedral losses, which have a linear calibration function, are calibrated to a very limited number of discrete losses, have multiple consistent decodings and have a non-continuous surrogate Bayes predictor. The main contribution of this thesis regarding probabilistic estimators is a user-friendly quantitative analysis of the closeness to the Bayes risk using calibration functions. The contributions on non-probabilistic estimators are the necessary and sufficient conditions for calibration of the Max loss and the introduction of the Max-Min loss in the general structured setting with a full-stack analysis including the derivation of an efficient optimization algorithm accompanied with finite-sample generalization bounds to the Bayes predictor.

Future Research Directions

Several open questions and interesting research directions have been constantly showing up during the last three years of research. Unfortunately, many of the ideas have not been further investigated due to time constraint. While some of them have already been briefly discussed at some point throughout the thesis, we now provide a synthetized and more detailed inspection of these perspectives for future research by classifying them into four main categories.

Choosing the discrete loss for learning. The statistical learning setup for supervised learning introduced in Section 2 is given by a data distribution ρ over input-output pairs and a loss function L to minimize. However, the loss measuring the error between predictions and observations is not directly given by the learning problem as it is the case for the distribution, and thus has to be 'artificially' designed. For binary classification, it is quite natural to minimize the expected 0-1 loss, as the average number of prediction errors is a sensible measure of error. However, in structured output spaces there is usually a compromise between informativeness and computational tractability (such as small affine dimension). For instance, whereas the Hamming loss measuring the average number of errors in a sequence has logarithmic affine dimension in $|\mathcal{Y}|$, it is not as informative of the closeness of two sequences as the edit distance, for which computing the Bayes predictor from a distribution over sequences is computationally intractable.

Designing discrete losses from basic primitives on the learning problem accounting for a good balance between informativeness and computational tractability is indeed an important open question.

Towards a general theory of surrogate losses for discrete prediction. A general theory regarding a set of mathematical objects is generally understood as a classification of these objects and a characterization of the main properties of interest in each category. In the case of surrogate methods, we already discussed extensively the two main categories, namely, probabilistic estimators and non-probabilistic estimators:

- **Probabilistic estimators:** The first ones can be constructed in surrogate spaces as small as the affine dimension of the task loss. Moreover, they can be characterized with a Bregman representation (also known as proper composite representation (Vernet et al., 2011; Agarwal and Agarwal, 2015). An open question regarding these surrogates is how to make them task loss dependent. Indeed, most probabilistic estimators presented in this thesis do not contain full information about the discrete loss L, which is generally only or mostly contained in the decoding. Making the surrogate loss dependent on the task loss by maintaining its main properties is key to have good calibration properties when the approximation error is large (Lacoste-Julien et al., 2011). In particular, it would be interesting to develop a generalization of the assymetric calibration theory from binary classification (Scott, 2012) to the structured output case.
- **Non-probabilistic estimators:** Much less is known for non-probabilistic surrogates, for which there exist much more variety than in the probabilistic case. In this thesis we just investigated polyhedral losses in a surrogate space of dimension as small as the affine dimension of the task loss. However, one can construct non-probabilistic calibrated convex surrogates (smooth or non-smooth) of even smaller dimension depending on the task loss. As an example, there exist non-probabilistic convex surrogate losses in one dimension which are calibrated to the absolute deviation loss with $|\mathcal{Y}| > 1$. Given L, the minimum dimension of the surrogate space for which there exist convex calibrated surrogates for a given discrete loss is known as the *convex calibration dimension* (Ramaswamy and Agarwal, 2016). While there exist some recent results on one-dimenional surrogates by Finocchiaro et al. (2020) it still remains an open question how to characterize this quantity and systematically construct them. A proper understanding of the convex calibration dimension will shed light to the nature of non-probabilistic surrogate methods.

Which surrogate loss to choose? An important open question for which is still hard to give concrete answers is the problem of characterizing the learning settings for which a surrogate loss works better than another one. A classical sufficient condition for this question is to check whether the surrogate Bayes predictor $g^\star$ can be arbitrarly approximated with functions from the surrogate hypothesis class $\mathcal{G}$. Indeed, in this case one can make use of calibration properties to show superiority of one surrogate loss over another. However, when the $g^\star$'s cannot be approximated by predictors from the class, which is the most common setting in practice, it is not an easy task to understand when a surrogate method works better than another. Proper answers to this question will cer-

tainly come from a deeper understanding of $\mathcal{G}$-consistency presented in Section 2 of the introductory part.

Parts-based generalization bounds for structured prediction. As we have seen in this thesis, finite-sample generalization bounds for surrogate methods can be easily derived under the assumption that the surrogate Bayes predictor $g^\star$ belongs to hypothesis class by plugging-in learning bounds on the surrogate excess risk into the comparison inequality. Typically, these bounds are of the form $\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star) \leq \mathcal{O}(n^{-\gamma})$ for a positive $\gamma > 0$. However, as already discussed in Section 5 , when the structured output is made of parts, it is a common practice to tie the predictors of different parts under the assumption that the learning problem has some form of stationarity. In particular, some works such as Ciliberto et al. (2019); London et al. (2016) are able to derive generalization bounds on the excess risk of the form $\mathcal{O}(p^{-\beta}n^{-\gamma})$ with $\beta > 0$, i.e., decreasing both with the number of samples n and parts p. Indeed, if the same predictor is repeated across the output parts, having larger outputs in the dataset, (i.e., more parts p), corresponds to an increase of information for learning under sufficient correlation decay between the parts. Unfortunately, the works cited above use vacuous margin bounds for convex losses or are specific for the quadratic surrogate. It is an interesting research direction to derive parts dependent generalization bounds for widely used losses in structured prediction such as conditional random fields.

Bibliography

Abernethy, J. D. and Frongillo, R. M. (2012). A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pages 27–1.

Agarwal, A. and Agarwal, S. (2015). On consistent surrogate risk minimization and property elicitation. In *Conference on Learning Theory*, pages 4–22.

Andreasson, N., Patriksson, M., and Evgrafov, A. (2020). *An Introduction to Continuous Optimization: Foundations and Fundamental Algorithms*. Courier Dover Publications.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331.

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

Aurenhammer, F. (1987). Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96.

Bach, F. (2015). Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129.

BakIr, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., and Vishwanathan, S. (2007). *Predicting Structured Data*. MIT press.

Barahona, F. (1982). On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241.

Bartlett, P., Freund, Y., Lee, W. S., and Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1):85–113.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bauschke, H. H. and Borwein, J. M. (2001). Joint and separate convexity of the bregman distance. In *Studies in Computational Mathematics*, volume 8, pages 23–36. Elsevier.

Baxter, R. J. (2016). *Exactly solved models in statistical mechanics*. Elsevier.

Belanger, D. and McCallum, A. (2016). Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J. L., et al. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Blondel, M. (2019). Structured prediction with projection oracles. In *Advances in Neural Information Processing Systems*, pages 12145–12156.

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Burkard, R., Dell'Amico, M., and Martello, S. (2012). *Assignment problems, revised reprint*, volume 106. Siam.

Caetano, T. S., McAuley, J. J., Cheng, L., Le, Q. V., and Smola, A. J. (2009). Learning graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1048–1058.

Calauzenes, C., Usunier, N., and Gallinari, P. (2012). On the (non-) existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems*, pages 197–205.

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

Chatalbashev, V., Taskar, B., and Koller, D. (2005). Disulfide connectivity prediction via kernelized matching. RECOMB.

Chen, W., Liu, T.-Y., Lan, Y., Ma, Z.-M., and Li, H. (2009). Ranking measures and loss functions in learning to rank. In *Advances in Neural Information Processing Systems*, pages 315–323.

Ciliberto, C., Bach, F., and Rudi, A. (2019). Localized structured prediction. In *Advances in Neural Information Processing Systems*, pages 7299–7309.

Ciliberto, C., Rosasco, L., and Rudi, A. (2016). A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pages 4412–4420.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020). A general framework for consistent structured prediction with implicit loss embeddings. *J. Mach. Learn. Res.*, 21(98):1–67.

Cortes, C., Kuznetsov, V., Mohri, M., and Yang, S. (2016). Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems*, pages 2514–2522.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

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

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Daumé, H., Langford, J., and Marcu, D. (2009). Search-based structured prediction. *Machine learning*, 75(3):297–325.

Daumé III, H. and Marcu, D. (2005). Learning as search optimization: Approximate large margin methods for structured prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 169–176.

De Loera, J. A., Hemmecke, R., and Köppe, M. (2012). *Algebraic and Geometric Ideas in the Theory of Discrete Optimization*. SIAM.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer.

Deza, M. M. and Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of distances*, pages 1–583. Springer.

Djolonga, J. and Krause, A. (2014). From map to marginals: Variational inference in bayesian submodular models. In *NIPS*, pages 244–252.

Dogan, Ü., Glasmachers, T., and Igel, C. (2016). A unified view on multi-class support vector classification. *J. Mach. Learn. Res.*, 17(45):1–32.

Doppa, J. R., Fern, A., and Tadepalli, P. (2014). Hc-search: A learning framework for search-based structured prediction. *Journal of Artificial Intelligence Research*, 50:369–407.

Duchi, J., Khosravi, K., and Ruan, F. (2018). Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning*, pages 272–279.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. (2010). On the consistency of ranking algorithms. In *ICML*, pages 327–334.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eboli, T., Nowak-Vila, A., Sun, J., Bach, F., Ponce, J., and Rudi, A. (2020). Structured and localized image restoration. *arXiv preprint arXiv:2006.09261*.

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*, volume 425. Springer.

Fathony, R., Asif, K., Liu, A., Bashiri, M. A., Xing, W., Behpour, S., Zhang, X., and Ziebart, B. D. (2018a). Consistent robust adversarial prediction for general multiclass classification. *arXiv preprint arXiv:1812.07526*.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. (2016). Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems*, pages 559–567.

Fathony, R., Rezaei, A., Bashiri, M. A., Zhang, X., and Ziebart, B. (2018b). Distributionally robust graphical models. In *Advances in Neural Information Processing Systems*, pages 8353–8364.

Finley, T. and Joachims, T. (2008). Training structural SVMs when exact inference is intractable. In *Proceedings of the 25th International Conference on Machine Learning*, pages 304–311. ACM.

Finocchiaro, J., Frongillo, R., and Waggoner, B. (2019). An embedding framework for consistent polyhedral surrogates. In *Advances in Neural Information Processing Systems*.

Finocchiaro, J., Frongillo, R., and Waggoner, B. (2020). Embedding dimension of polyhedral losses. In *Conference on Learning Theory*, pages 1558–1585. PMLR.

Forsyth, D. A. and Ponce, J. (2012). *Computer vision: a modern approach*. Pearson,.

Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Frongillo, R. and Kash, I. A. (2015a). Vector-valued property elicitation. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of the Conference on Learning Theory*, pages 710–727.

Frongillo, R. and Kash, I. A. (2015b). Vector-valued property elicitation. In *Conference on Learning Theory*, pages 710–727.

Gao, W. and Zhou, Z.-H. (2011). On the consistency of multi-label learning. In *Proceedings of the Conference on Learning Theory*, pages 341–358.

Gidel, G., Jebara, T., and Lacoste-Julien, S. (2017). Frank-wolfe algorithms for saddle point problems. In *Artificial Intelligence and Statistics*, pages 362–371. PMLR.

Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, number CONF, pages 427–435.

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM.

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.

Keshet, J. and McAllester, D. A. (2011). Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing Systems*, pages 2205–2212.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Koltchinskii, V. and Beznosova, O. (2005). Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, pages 295–307. Springer.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer.

Korba, A., Garcia, A., and d'Alché Buc, F. (2018). A structured prediction approach for label ranking. In *Advances in Neural Information Processing Systems*, pages 8994–9004.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Kuhn, T. (1970). *The structure of scientific revolutions*. University of Chicago Press.

Lacoste-Julien, S., Huszár, F., and Ghahramani, Z. (2011). Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 416–424. JMLR Workshop and Conference Proceedings.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2013). Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proceedings of the International Conference on Machine Learning*, pages 53–61.

Lacoste-Julien, S., Taskar, B., Klein, D., and Jordan, M. (2006). Word alignment via quadratic assignment.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81.

Lin, H.-T. and Li, L. (2006). Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *International Conference on Algorithmic Learning Theory*, pages 319–333. Springer.

Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82.

Liu, Y. (2007). Fisher consistency of multicategory support vector machines. In *Artificial Intelligence and Statistics*, pages 291–298.

Liu, Y. and Shen, X. (2006). Multicategory ψ-learning. *Journal of the American Statistical Association*, 101(474):500–509.

London, B., Huang, B., and Getoor, L. (2016). Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859.

Long, P. and Servedio, R. (2013). Consistency versus realizable h-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809. PMLR.

Lugosi, G. and Vayatis, N. (2004). On the bayes-risk consistency of regularized boosting methods. *The Annals of statistics*, 32(1):30–55.

Mammen, E. and Tsybakov, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *The Annals of Statistics*, pages 502–524.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142.

Michelot, C. (1986). A finite algorithm for finding the projection of a point onto the canonical simplex of n. *Journal of Optimization Theory and Applications*, 50(1):195–200.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-J. (2012). Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems*, pages 2789–2797.

Nemirovski, A. (2004). Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Niculae, V., Martins, A., Blondel, M., and Cardie, C. (2018). Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pages 3799–3808. PMLR.

Nowak, A., Bach, F., and Rudi, A. (2019). Sharp analysis of learning with discrete losses. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1920–1929. PMLR.

Nowak-Vila, A., Bach, F., and Rudi, A. (2019). A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*.

Nowak-Vila, A., Bach, F., and Rudi, A. (2020). Consistent structured prediction with max-min margin markov networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Nowozin, S., Gehler, P. V., Jancsary, J., and Lampert, C. H. (2014). *Advanced Structured Prediction*. MIT Press.

Nowozin, S. and Lampert, C. H. (2011). Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. (2018). An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*.

Osokin, A., Bach, F., and Lacoste-Julien, S. (2017). On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pages 302–313.

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Pedregosa, F., Bach, F., and Gramfort, A. (2017). On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(1):1769–1803.

Petterson, J., Yu, J., McAuley, J. J., and Caetano, T. S. (2009). Exponential family graph matching and ranking. In *Advances in Neural Information Processing Systems*, pages 1455–1463.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018a). Exponential convergence of testing error for stochastic gradient methods. In *Proceedings of the Conference On Learning Theory*, pages 250–296.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018b). Exponential convergence of testing error for stochastic gradient methods. *Proceedings of the Conference on Learning Theory*.

Pires, B. A., Szepesvari, C., and Ghavamzadeh, M. (2013). Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, pages 1391–1399.

Pollard, D. (1984). Convergence of stochastic processes.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Ramaswamy, H. G. and Agarwal, S. (2016). Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(1):397–441.

Ramaswamy, H. G., Agarwal, S., and Tewari, A. (2013). Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, pages 1475–1483.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2015). Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554.

Ravikumar, P., Tewari, A., and Yang, E. (2011). On NDCG consistency of listwise ranking methods. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 618–626.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Robinson, J. (1951). An iterative method of solving a game. *Annals of mathematics*, pages 296–301.

Rockafellar, R. T. (1997). Convex analysis. princeton landmarks in mathematics.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Rudi, A., Carratino, L., and Rosasco, L. (2017). Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901.

Rudi, A., Ciliberto, C., Marconi, G. M., and Rosasco, L. (2018). Manifold structured prediction. *arXiv preprint arXiv:1806.09908*.

Russell, S. (1998). Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103.

Ryser, H. J. (1963). *Combinatorial mathematics*, volume 14. American Mathematical Soc.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992.

Settles, B. (2004). 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*.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599.

Shapiro, H. N. (1958). Note on a computation method in the theory of games. *Communications on Pure and Applied Mathematics*, 11(4):587–593.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Smith, N. A. (2011). Linguistic structure prediction. *Synthesis Lectures on Human Language Technologies*, 4(2):1–274.

Smola, A. J. and Schökopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the International Conference on Machine Learning*, ICML '00, pages 911–918.

Sontag, D. A. (2010). *Approximate inference in graphical models using LP relaxations*. PhD thesis, Massachusetts Institute of Technology.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pages 1545–1552.

Steinwart, I. (2002). Support vector machines are universally consistent. *Journal of Complexity*, 18(3):768–791.

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. (2005a). Learning structured prediction models: A large margin approach. In *Proceedings of the International Conference on Machine Learning*, pages 896–903.

Taskar, B., Klein, D., Collins, M., Koller, D., and Manning, C. D. (2004). Max-margin parsing. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Taskar, B., Lacoste-Julien, S., and Klein, D. (2005b). A discriminative matching approach to word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 73–80.

Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(May):1007–1025.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning*, page 104. ACM.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(Jul):2411–2414.

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

Valiant, L. G. (1979). The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Valizadegan, H., Jin, R., Zhang, R., and Mao, J. (2009). Learning to rank by optimizing ndcg measure. In *Advances in Neural Information Processing Systems*, pages 1883–1891.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Vapnik, V. and Lerner, A. Y. (1963). Recognition of patterns with help of generalized portraits. *Avtomat. i Telemekh*, 24(6):774–780.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

Vernet, E., Reid, M. D., and Williamson, R. C. (2011). Composite multiclass losses. In *Advances in Neural Information Processing Systems*, pages 1224–1232.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Volkovs, M. N., Larochelle, H., and Zemel, R. S. (2011). Loss-sensitive training of probabilistic conditional random fields. *arXiv preprint arXiv:1107.1805*.

Waegeman, W., Dembczyński, K., Jachnik, A., Cheng, W., and Hüllermeier, E. (2014). On the bayes-optimality of f-measure maximizers. *Journal of Machine Learning Research*, 15(1):3333–3388.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Conference on Learning Theory*, pages 25–54.

Weston, J. and Watkins, C. (1999). Support vector machines for multi-class pattern recognition. In *ESANN*, volume 99, pages 219–224.

Williamson, R., Vernet, E., Reid, M., et al. (2016). Composite multiclass losses.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, 51(7):2282–2312.

Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

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

Zhang, T. (2004b). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85.

RÉSUMÉ

La classification est la branche de l'apprentissage supervisé qui vise à estimer une fonction à valeurs discrètes à partir de données constituées de paires d'entrées et de sorties. Le cadre le plus classique et le plus étudié est celui de la classification binaire, où le prédicteur discret prend pour valeur zéro ou un. Cependant, la plupart des problèmes de classification qu'on retrouve en pratique sont definis sur de grands espaces de sortie structurés tels que des séquences, des grilles, des graphs, des permutations, etc. Il existe des différences fondamentales entre la prédiction structurée et la classification multiclasse ou binaire non structurée: la grandeur exponentielle de l'espace de sortie par rapport à la dimension naturelle des objets à prédire et la sensibilité des coûts de la tâche de classification. Cette thèse se concentre sur les méthodes de substitution pour la prédiction structurée, dans lesquelles le problème discret typiquement insoluble est abordé à l'aide d'un problème continu convexe qui, à son tour, peut être résolu à l'aide de techniques de régression.

MOTS CLÉS

Théorie d'aprentissage statistique, prediction structurée, méthodes de substitution, méthodes à noyaux.

ABSTRACT

Classification is the branch of supervised learning that aims at estimating a discrete valued mapping from data made of input-output pairs. The most classical and well studied setting is binary classification, where the discrete predictor takes zero or one as value. However, most of the practical classification settings deal with large structured output spaces such as sequences, grids, graphs, permutations, matchings, etc. There are many fundamental differences between structured prediction and vanilla binary or multi-class classification, such as the exponentially large size of the output space with respect to the natural dimension of the output objects and the cost-sensitive nature of the learning task. This thesis focuses on surrogate methods for structured prediction, whereby the typically intractable discrete problem is approached using a convex continuous surrogate problem which in turn can be addressed using techniques from regression.

KEYWORDS

Statistical learning theory, structured prediction, surrogate methods, kernel methods.