Classification via Multiple Hyperplanes: Loss functions, Overparametrization, and Interpolation

by

Yutong Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in The University of Michigan
2022

Doctoral Committee:

Professor Clayton D. Scott, Chair
Associate Professor Laura Balzano
Assistant Professor Qing Qu
Professor Ambuj Tewari

Yutong Wang

yutongw@umich.edu

ORCID iD: 0000-0001-7472-6750

To my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Professor Clay Scott. One of my greatest struggles especially early on is finishing what I started. Despite many projects left abandoned halfway, Clay always generously invested his time and energy in advising me. As I gradually overcame my struggle to complete projects, other weaknesses would come to the forefront. For instance, I would sometimes follow the path of least resistance and avoid the hard questions. Clay would then compassionately guide me toward focusing on these weak links. Because of his guidance, I felt seen, challenged, and transformed from a student to a researcher. I am truly grateful to have Clay as a teacher and a friend.

Next, I'd like to thank my committee members Professors Ambuj Tewari, Laura Balzano and Qing Qu. I am incredibly lucky to have the opportunity to continue building collaboration with them. Chapter 4 of this thesis in many ways stands on Ambuj's and Professor Peter Bartlett's prior work. Whenever I had the chance to meet and discuss with Ambuj, he was always deeply engaged and generously shared his insights, which continue to impact my research. While writing Chapter 6 of this thesis, I consulted with Laura regarding a related problem on Grassmannians. Our discussion became the basis of a new research collaboration with her student Kyle Gilman. Her vision and passion for impactful machine learning research has deeply impacted my perspective on both the application side, e.g., from our collaboration in applying machine learning to genomics, and the theory side (ongoing). I am grateful and excited to be advised by Qing and Professor Wei Hu in the upcoming year as a

postdoc. In addition to research guidance, Qing and Wei have given me an incredible amount of help on improving my other skills as researcher, such as mentorship and writing skills.

I'd like to thank all ECE staff, especially Shelly Feldkamp and Kristen Thornton. Throughout the program, they have gone out of their way to make students feel supported in the department.

This work would not have been possible without my teachers from my time as an undergrad and a masters student. Professors Ralf Spatzier and Mitya Boyarchenko first inspired me to study mathematics rigorously. Professor David Neuhoff and Dr. Matt Reyes for mentoring me on information theory research. Professor Lizhen Ji and Dr. Patrick Boland mentored me on an research project as an undergraduate that helped me decide to pursue research as a career. I'd like to thank especially my masters advisor Professor Brian Osserman at UC Davis who taught me the beautiful subject of algebraic geometry.

I thank all my friends and colleagues at Michigan who made the experience unforgettable. Working with Tasha Thong and Professor Justin Colacino earlier on in my PhD on an interdisciplinary project had been extremely motivating and inspiring for the latter half of my time here. I'd like to thank my friends and fellow students in the department Glen Chou, Yunus Emre, Kwesi Rutledge, David Hong, Aniket Deshmukh, Hojae Lee, Kyle Gilman, Alex Ritchie, Efrén Cruz Cortés, and Amanda Bower. I'd like to especially thank Haroon Raja, who is simultaneously an amazing mentor and a dear friend. They have shaped my experience as a PhD student here.

I thank my parents, Dejuan Kong and Yongcai Wang, for their boundless love, humor, and wisdom. They have raised me in a way so that I never give up pursuing what *truly makes me happy*. Completing this thesis has been a deeply fulfilling experience. This milestone achievement also belongs to my parents: it is the fruition of their hardwork as the amazing parents they are. This thesis is dedicated to them.

Finally, I thank my brilliant, beautiful wife, Rita Xiaochen Hu. It's hard to overstate how wonderful it is to be able to finish a day's work and then to spend time with her. When I am facing seemingly insurmountable difficulties, her unwavering belief in me is my greatest source of motivation. Her commitment to her own research, how she embodies it in her own life, is my greatest source of inspiration. Thank you, Rita, for *truly making me happy*.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Many well-established classification algorithms such as support vector machines (SVM) are originally proposed as large-margin classifiers from a single hyperplane. This dissertation is divided into two halves, each half studying classification from the perspective of using multiple hyperplanes.

The first half introduces a new framework for multiclass loss functions called the permutation-equivariant and relative margin-based (PERM) losses, inspired by multiclass classification with multiple hyperplanes. Using our framework, we establish statistical and optimization results on Weston-Watkins multiclass SVMs. Furthermore, we provide sufficient conditions for the classification-calibration of a general family of PERM losses. These sufficient conditions subsume all previously known and establish new classification-calibration results.

The second half focuses on hyperplane arrangement classifiers (HACs). When implemented as neural networks, we show that the HACs can be overparameterized yet still have small VC dimensions and further achieve minimax optimality (assuming the empirical risk minimization can be solved to optimality). By using an ensemble of randomly initialized HACs, we demonstrate for the first time an interpolating ensemble method that is consistent for a broad class of distributions in arbitrary dimensions. We discuss the significance of these results in the context of recent advances in the theory of overparameterized learning.

CHAPTER I

Introduction

> Mathematics is the art of giving the same name to different things
>
> *Henri Poincaré* [Ver12]

This introduction can be read as a guided-tour for the rest of the thesis. The focus will be on motivating each subject via simple examples.

In Section 1.1, we review binary classification with linear classifiers. The goal is to review the main concepts of the ± 1 *label encoding*, *margins*, *discriminants* and *margin losses*. In *binary* classification, these concepts are essentially standardized. However, in *multiclass* classification, there are several distinct notions of multiclass label encodings and margins.

A key ingredient behind the theory developed in Chapters II, III and IV is a novel label encoding for multiclass classification which we call the *multiplicative label encoding*. Let k denote the number of classes. The multiplicative label encoding is a set of $(k-1) \times (k-1)$ square matrices $\{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k\}$ which generalizes the well-known ± 1 label encoding for binary classification. In Section 1.3, we demonstrate the multiplicative label encoding for the case of ternary classification, i.e., when $k = 3$. Moreover, we give an overview of how this label encoding is used in the aforementioned chapters towards deriving our main results.

In Section 1.4, we review the concepts of *hyperplane arrangements* and their associated histogram classifiers which we refered to as simply *hyperplane arrangement classifiers* (HACs). We provide an intuitive introduction to a class of partially quantized neural networks which implement HACs, which is the focus of Chapter V. In Section 1.5, we discuss random *ensembles* of these hyperplane arrangement classifiers, which is the focus of Chapter VI.

1.1 Binary classification: two classes, one hyperplane

Let us consider one of the simplest non-trivial settings for classification: when $\mathcal{X} = \mathbb{R}^d$ is the Euclidean space and $\mathcal{Y} = \{\pm 1\}$ is binary. Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, our goal is to select a mapping $\mathbb{R}^d \to \{\pm 1\}$ that generalizes well to unseen data.

A classical approach is to use linear classifiers $x \mapsto \mathtt{sgn}(w^\top x)$. Although deceptively simple, linear classifiers have been continuously studied since the earliest days of machine learning research under various guises and names (perceptrons [Ros57], linear threshold functions [Blu+98], optimal margin separating hyperplanes [BGV92], support vector machines [CV95], halfspaces [Kal+08]). Most relevant to this thesis is the support vector machines. In the next few subsections, we recall definitions and facts from the theory of binary support vector machine with a view toward its multiclass extension. Some of these definitions seems unnecessarily complicated for the binary case, but will be beneficial when transitioning into the multiclass case.

1.1.1 Discriminants and margins

We begin by defining the *discriminant*, a quantity of relevance to most if not all binary classification algorithms (after replacing $w^\top \bullet : \mathbb{R}^d \to \mathbb{R}$ by a general function $f : \mathcal{X} \to \mathbb{R}$):

$$\mathtt{disc}_i := w^\top x_i \tag{1.1}$$

and the *margin*:

$$\texttt{marg}_i := y_i \cdot \texttt{disc}_i = y_i w^\top x_i. \tag{1.2}$$

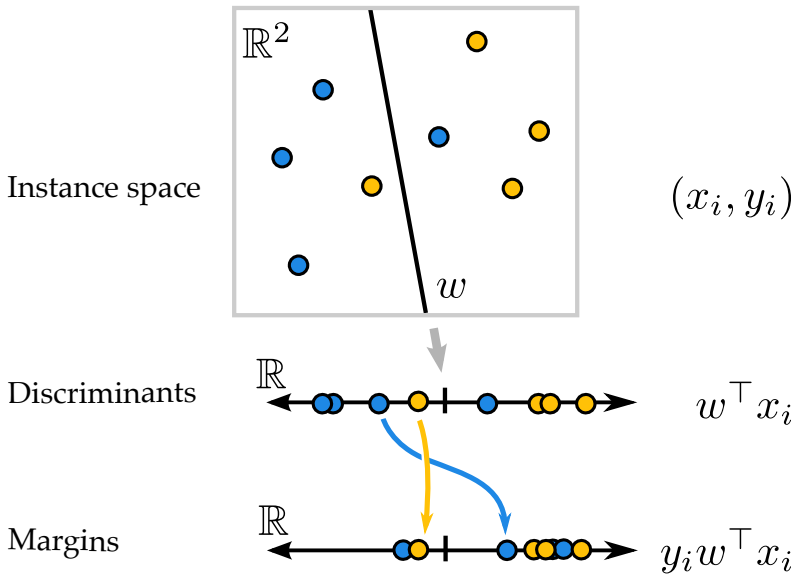

Figure 1.1: Discriminants and margins in binary classification. Here, yellow and blue points represent the "positive" and "negative" classes, respectively. For the yellow points, the margins and the discriminants are equal. For the blue points, the margins are the reflections across the origin of their respective discriminants.

The only difference between the discriminant and the margin is the y_i multiplier in front. By definition, the sign of the discriminant is the classifier's predicted label. The magnitude of the discriminant can be thought of intuitively as the "confidence" of the classifier. See Figure 1.1.

On the other hand, the margin *does* take into account the label. A large (positive) margin means that the classifier did a good job while a small (negative) margin means that the classifier did a poor job. The definition of the margin (Eqn. 1.2) can be stated in plain English as

"labels acting on discriminants by multiplication gives rise to margins". (1.3)

This definition seems unnecessary given the already clear mathematical definition Eqn. 1.2. However, when we transition from the binary case $\mathcal{Y} = \{\pm 1\}$ to the ternary case $\mathcal{Y} = \{1, 2, 3\}$ and beyond, defining the margin is not as straightforward anymore. This "supervised-label-as-an-action" perspective[1] will serve as the blueprint for our approach to the multiclass theory.

1.1.2 Margin-based loss functions

Finally, to formulate the hyperplane selection problem as an optimization, we recall *margin-based loss functions* [BJM06], i.e., nonnegative functions $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that converts margins into penalties for use in regularized empirical ψ-risk minimization[2]:

$$\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \psi(\underbrace{y_i w^\top x_i}_{=\texttt{marg}_i}). \tag{1.4}$$

Monotone non-increasing margin-based loss function formalizes the notion that "a large (positive) margin means that the classifier did a good job while a small (negative) margin means that the classifier did a poor job". See Figure 1.2.

Another way to visualize this is via a partition of the space of margins, i.e., the real line, into high and low penalty zones, namely the positive and the negative halves. This partition perspective extends easily into the multiclass case, with the margins being vector- instead of scalar-valued.

[1] This is an instance of "group action" in the mathematical subject of group theory. Here, the labels $\{\pm 1\}$ is a group with group action by multiplication on the set $\mathbb{R}$, where the discriminant lives in.

[2] Following Bartlett et al. [BJM06], we consider nonnegative losses throughout the thesis. No generality is gained by allowing the loss to be negative and while still lower bounded. In the full general case when the loss is allowed to tend to $-\infty$, empirical risk minimization may be $-\infty$.

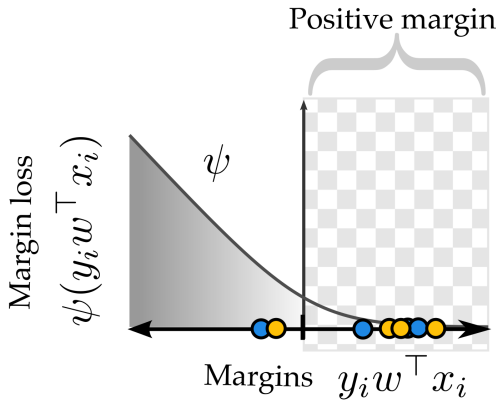

Figure 1.2: Loss function and the partition of the discriminant into high and low penalty zones (the latter denoted by the checkered region).

1.1.3 Consistency

Finally, we need to circle back to our original problem: finding a w such that $x \mapsto \texttt{sgn}(w^\top x)$ generalizes well. Unfortunately, if we require w to be a hyperplane in finite-dimensional Euclidean space, this is essentially impossible except in highly specialized settings[3]. Fortunately, if we allow w to be an element of a universal reproducing kernel Hilbert space $\mathcal{H}$ and x to be replaced by its kernel embedding in $\mathcal{H}$, then the problem has a solution. Namely, if ψ is *classification-calibrated* (defined in the next section), then there is a choice of hyperparameters $C = C_n$ such that solving the optimization Eqn. 1.4 results in asymptotically optimal choice of the "hyperplane" w for the classifier $x \mapsto \texttt{sgn}(w^\top x)$ [Ste05].

Before proceeding, we recall some definitions. Let $g : \mathcal{X} \to \{\pm 1\}$. The *01*-risk of g is defined as

$$R_{01}(g) := \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{I}\{Y \neq g(X)\}\right] \tag{1.5}$$

[3]Learning w^* that minimizes the number of misclassifications is well-known to be NP-hard. See Guruswami et al. [GR09]. On the other hand, with additional assumptions, polynomial time algorithms for finding such a w^* is known, e.g., see Blum et al. [Blu+98] and Diakonikolas et al. [DGT19] and the references therein.

where $\mathbb{I}$ is the indicator function. For a function $f : \mathcal{X} \to \mathbb{R}$ the ψ-risk is defined as

$$R_\psi(f) := \mathbb{E}_{(X,Y)\sim P}[\psi(Yf(X))]. \tag{1.6}$$

The ψ- and the *01-Bayes risk* are defined as $R^*_\psi := \inf_f R_\psi(f)$ and $R^*_{01} := \inf_f R_{01}(\mathtt{sgn} \circ f)$, respectively, where the infimum is taken over all Borel functions f. The following theorem relates when an algorithm that performs well respect to R_ψ can be converted to one that performs well respect to R_{01}:

Theorem I.1 ([BJM06])**.** *Let ψ be a margin-based loss function. Let $\mathcal{F}$ be the set of Borel functions $\mathcal{X} \to \mathbb{R}$. If ψ is classification calibrated then the following holds: For all sequence of function classes $\{\mathcal{F}_n\}_n$ such that $\mathcal{F}_n \subseteq \mathcal{F}$, $\bigcup_n \mathcal{F}_n = \mathcal{F}$, $\hat{f}_n \in \mathcal{F}_n$ and all data generating probability distribution P, we have*

$$R_\psi(\hat{f}_n) \xrightarrow{P} R^*_\psi \quad \textit{implies} \quad R_{01}(\mathtt{sgn} \circ \hat{f}_n) \xrightarrow{P} R^*_{01}.$$

Although the above theorem seems quite powerful, it does have limitations. For instance, if $\mathcal{F}$ is a smaller function space (such as linear functions on finite-dimensional Euclidean space), then all bets are off. As mentioned in Duchi et al. [DKR18], going beyond the "$\mathcal{F}$ = all Borel functions" setting is an important research direction. Partial progress have already been made by Duchi et al. [DKR18] and Zhang et al. [ZA20].

1.1.4 Classification-calibration

The previous section established the significance of the notion of classification-calibration in linking solution to the optimization Eqn. 1.4 and solution to the original problem (in a universal RKHS). In this section, we review the notion of a margin-loss being classification-calibrated.

Definition I.2. Let $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a margin-based loss function. Define $C_p^\psi(t) = p\psi(t) + (1-p)\psi(-t)$ for all $t \in \mathbb{R}$. We say that ψ is *classification-calibrated* if for all $p > (1/2, 1]$

$$\inf_{t \leq 0} C_p^\psi(t) > \inf_{t \in \mathbb{R}} C_p^\psi(t).$$

To understand the definition, consider the following scenario. Suppose that for a fixed sample x, we have $p := P_{Y=+1|X=x} > 1/2$. Then a discriminant function f should satisfy $f(x) > 0$ in order to minimize $\mathbb{E}[\mathbb{I}\{Y \neq \mathtt{sgn}(f(x))\}|X = x]$, sometimes called the *conditional 01-risk (at x)*. Analogously, the conditional ψ-risk (at x) is defined as $\mathbb{E}[\psi(Yf(x))|X = x]$ and is equal to C_p^ψ.

Consequently, the above definition can be interpreted as "if $\mathtt{sgn}(f(x))$ is suboptimal for the conditional 01-risk, then the discriminant $t := f(x)$ is also suboptimal for the conditional ψ-risk." The converse of this says that "optimality for the conditional ψ-risk implies optimality for the conditional 01-risk."

There is a remarkable characterization of *convex* classification-calibrated losses: ψ is differentiable at 0 and $\psi'(0) < 0$ [BJM06, Theorem 6]. Obtaining a similar characterization in the multiclass setting is the is the goal of Chapter 4 of this thesis.

Now, the hinge loss $\psi(t) := \max\{0, 1-t\}$ satisfies this characterization (in the binary case). It has the interesting property that for $p \in (0, 1/2)$, $\arg\min_{t \in \mathbb{R}} C_p^\psi(t) = \{-1\}$ and for $p \in (1/2, 1)$, $\arg\min_{t \in \mathbb{R}} C_p^\psi(t) = \{1\}$. Thus, outside $p \in \{0, 1/2, 1\}$ (which has measure zero), minimizers for C_p^ψ lies in $\{\pm 1\}$. See Figure 1.3.

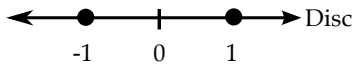

Figure 1.3: $\bigcup_{p \in (0,1/2) \cup (1/2,1)} \arg\min_t C_p^\psi(t) = \{\pm 1\}$ for the hinge loss.

Chapter 2 of this thesis studies the Weston-Watkins (WW) hinge loss [WW99], one of the various extension of the binary hinge loss proposed for multiclass SVMs. Unlike the binary hinge loss, the WW hinge loss is not classification-calibrated. Nevertheless, it has been shown to perform well in practice [DGI16]. The goal of Chapter 2 will

be to attempt to salvage the failure to be classification-calibrated by considering a different discrete loss.

1.2 Multiclass classification: Three classes, *three* hyperplanes?

Moving on to the multiclass case, for exposition we consider the simplest nontrival example of when $\mathcal{Y} = \{1, 2, 3\}$ is ternary. One way to define a hyperplane-based classifier $g : \mathbb{R}^d \to \{1, 2, 3\}$ is to use three hyperplanes $\mathbf{W} = [w_1, w_2, w_3] \in \mathbb{R}^{d \times 3}$ and define

$$g(x) = \arg\max_{j \in \{1,2,3\}} w_j^\top x.$$

Clearly, this generalizes to k classes, where we need k hyperplanes $w_1, \ldots, w_k$. Figure 1.4 shows a toy dataset with three classes that is completely interpolated by this linear classifier.

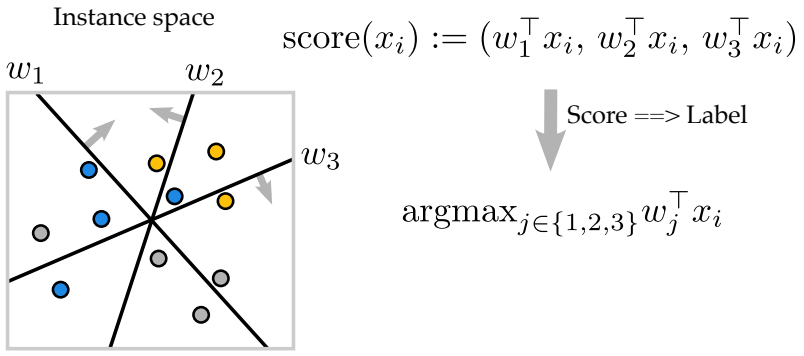

Figure 1.4: A toy dataset overlayed with three hyperplanes (lines) in $\mathbb{R}^2$. The black line represent the hyperplane itself, i.e., the set $\{x : w_i^\top x = 0\}$. The gray thick arrow represents the normal vector to the hyperplane, i.e., w_i.

However, the geometry of classifier is not apparent from the hyperplanes w_1, w_2, w_3

themselves, compared to the analogous "two classes, one hyperplane" picture. How can we think about this classifier intuitively?

1.3 Multiclass classification: Three classes, two hyperplanes

Instead of plotting w_1, w_2, w_3, the geometry becomes clear when we plot the difference of the hyperplanes $w_1 - w_2$, $w_1 - w_3$ and $w_2 - w3$ instead (Figure 1.5).

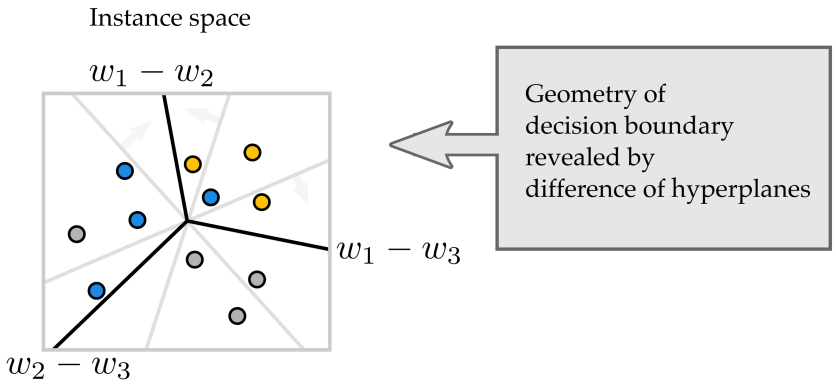

Figure 1.5: Difference of hyperplanes revealing the geometry of the decision regions of the multiclass linear classifier.

Doğan et al. [DGI16] introduced the term *relative margins*[4]. Figure 1.5 is indeed in line with the relative margins as defined by Doğan et al. [DGI16], which is defined in a case-wise manner for each $y_i \in \{1, 2, 3\}$. If $y_i = 1$, then

$$\texttt{marg}_i =: \begin{bmatrix} (w_1 - w_2)^\top x_i \\ (w_1 - w_3)^\top x_i \end{bmatrix}. \tag{1.7}$$

Pictorally, this correspond to Figure 1.7.

[4]We note that the term relative margin have been previously used by Jebara et al. [JS08] and Shivaswamy et al. [SJ10], but used in a completely different way.

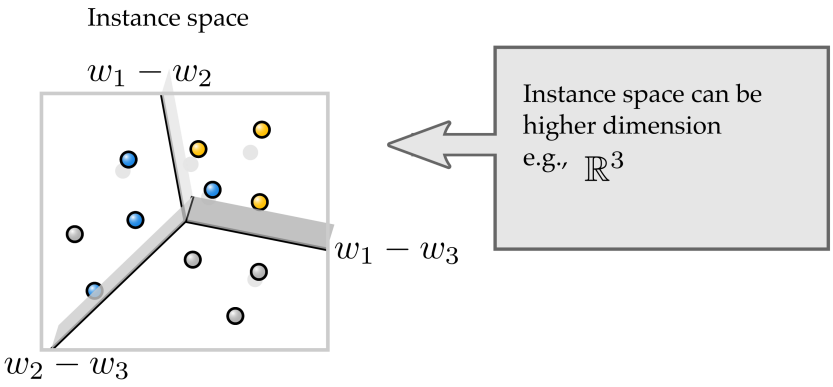

Figure 1.6: For $k = 3$, the discriminants are 2-dimensional. For disambiguating the instance space and the discriminant space, we take a toy dataset in 3-dimension space that is a "jittered" version of the dataset from Figure 1.5

If $y_i = 2$, then

$$\texttt{marg}_i =: \begin{bmatrix} (w_2 - w_1)^\top x_i \\ (w_2 - w_3)^\top x_i \end{bmatrix}. \tag{1.8}$$

And finally, if $y_i = 3$, then

$$\texttt{marg}_i =: \begin{bmatrix} (w_3 - w_2)^\top x_i \\ (w_3 - w_1)^\top x_i \end{bmatrix}. \tag{1.9}$$

However, we observe the following "issues" regarding the above definitions:

1. The margin in the binary case (Eqn. 1.2) is defined via a single equation. However, here in the multiclass case, the margins need to be defined case-wise depending on the label y_i.
2. There is no analog to the discriminant (Eqn. 1.1).

Let us attempt to propose a definition of the discriminant in the multiclass case as follows (see Figure 1.7)

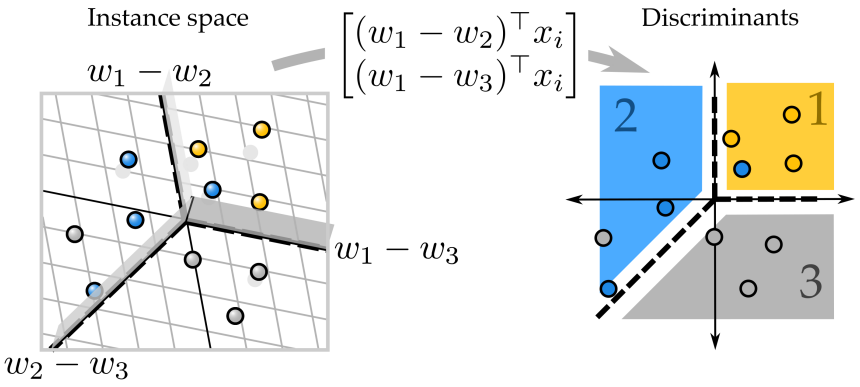

Figure 1.7: Discriminants when $k = 3$ for the toy dataset shown in Figure 1.6.

$$\texttt{disc}_i := \begin{bmatrix}(w_1 - w_2)^\top x_i \\ (w_1 - w_3)^\top x_i\end{bmatrix}. \tag{1.10}$$

Observe that the above definition of the discriminant (Eqn. 1.10) is equal to the margin when $y_i = 1$ (Eqn. 1.7). This is by design and is analogous to the binary case where one arbitrary class is chosen as the "positive" class.

Now, we recall the definition/slogan

"labels acting on discriminants by multiplication $\implies$ margins".

Thus, if the labels were to act on the 2-dimensional vector-valued discriminant in Eqn. 1.10 by multiplication (also see Figure 1.8 and 1.9), then they should be encoded as 2×2 matrices:

$$\boldsymbol{\rho}_1 = \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}, \quad \boldsymbol{\rho}_2 = \begin{bmatrix}-1 & 0 \\ -1 & 1\end{bmatrix}, \quad \boldsymbol{\rho}_3 = \begin{bmatrix}1 & -1 \\ 0 & -1\end{bmatrix}. \tag{1.11}$$

With this definition, it is straightforward to check that

$$\texttt{marg}_i = \boldsymbol{\rho}_{y_i} \texttt{disc}_i \tag{1.12}$$

in all cases $y_i \in \{1, 2, 3\}$. See Figure 1.9. Furthermore, note that by construction

$$y_i = \underset{j \in \{1,2,3\}}{\arg\min}\, w_j^\top x_i \quad \text{if and only if} \quad \texttt{marg}_i \in \mathbb{R}^2_{>0}.$$

The action of the matrices $\boldsymbol{\rho}_2$ and $\boldsymbol{\rho}_3$ can be visualized as linear *involutions* (i.e., $\boldsymbol{\rho}_j^2$ is the identity) on the discriminant space $\mathbb{R}^2$. Note that in binary classification, the negative class label -1 is a linear involution on $\mathbb{R}$, the binary discriminant space (see Figure 1.8). The analogous definition for Eqn. 1.12 when $k > 3$ is given in Chapter IV.

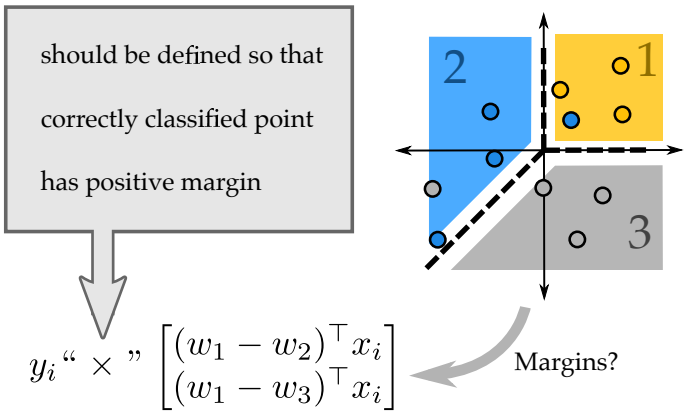

Figure 1.8: Defining “multiplication” by a label y_i.

1.3.1 Margin-based multiclass loss functions

Having defined multiclass margins, we now define ternary margin-based loss functions as multivariate-input, univariate-output function $\psi : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$. The regular-

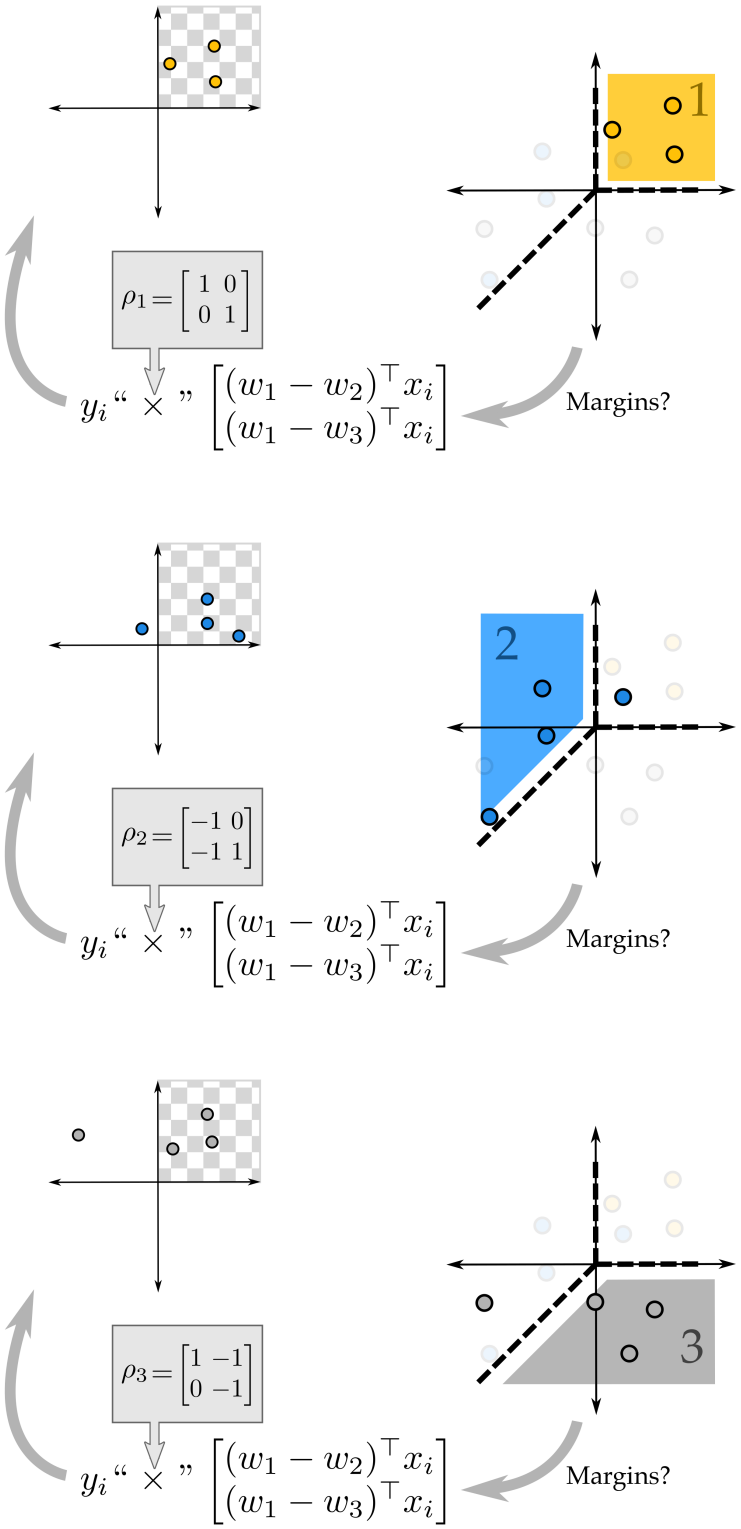

Figure 1.9: The matrices $\boldsymbol{\rho}_1$ (top), $\boldsymbol{\rho}_2$ (mid) and $\boldsymbol{\rho}_3$ (bottom).

ized[5] ψ-risk is defined as

$$\frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^{n}\psi\left(\boldsymbol{\rho}_{y_i}\begin{bmatrix}(w_1 - w_2)^\top x_i\\ (w_1 - w_3)^\top x_i\end{bmatrix}\right). \tag{1.13}$$

Like the binary case, we would like to formalizes the notion that "a large (positive) margin means that the classifier did a good job while a small (negative) margin means that the classifier did a poor job". However, the formalization should take into account the vector-valued nature of the multiclass margin. The analogous statement is "a margin inside the positive orthant means that the classifier did a good job while a margin outside the positive orthant means that the classifier did a poor job". Thus, the low-penalty zone in the space of margins, i.e., $\mathbb{R}^2$ in the ternary case, is the positive orthant. See Figure 1.10

1.3.2 Multiclass classification-calibration

Is there is a simple characterization of convex classification-calibrated *multiclass* losses extending the elegant result of Bartlett et al. [BJM06, Theorem 6]? Tewari et al. [TB07] developed the theoretical foundation towards such a characterization. However, ultimately, there remains a gap for a (relatively) simple to verify sufficient condition for classification-calibration for a general multiclass loss ψ.

In Chapter 4, we define a class of loss called *permutation equivariant and relative margin-based* losses, or PERM loss for short. For a PERM loss, we define a condition which we refer to as *total regularity*. One of the key property of total regularity is that the negative gradient ψ points into the positive orthant everywhere, i.e., $-\nabla\psi(z)$ is entrywise positive for all $z \in \mathbb{R}^{k-1}$. Our main result is that if ψ is totally regular,

[5] It is an open question what is the "right" norm for regularizing Eqn. (1.13). Amit et al. [Ami+07] proposes using the nuclear norm instead of the Frobenius norm. Lei et al. [Lei+19] proposes using p-Schatten norms. Tatsumi et al. [TT14] propose a multi-objective approach that departs from the framework of Eqn. (1.13) altogether. See Lee [Lee14] for an insightful perspective regarding whether it is even worthwhile to ponder which norm to use.

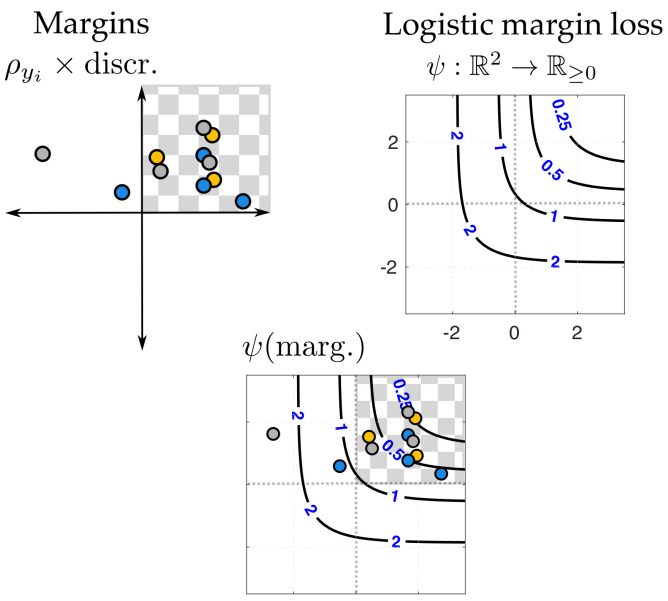

Figure 1.10: The low-penalty zone is in the positive orthant (denoted by the checkered region).

then ψ is classification-calibrated. In comparison to Bartlett et al. [BJM06, Theorem 6], our result is not as powerful, since the gradient condition is global. Furthermore, the regularity definition requires twice-differentiability. Nevertheless, our result is sufficiently general to significantly expand the known classification-calibration results regarding the recently proposed Fenchel-Young losses [BMN20; DKR18].

1.3.3 Weston-Watkins SVM: calibration

The Weston-Watkins (WW) SVM proposes solving Eqn. 1.13 with the "sum of hinge of margin components" extension of the binary hinge loss:

$$\psi(z) := \sum_{j=1}^{k-1} \max\{0, 1 - z_j\}.$$

Let us call the above the WW hinge loss. The WW hinge loss is not classification-calibrated but performs well in practice [DGI16]. Thus, a natural question is to

understand why. This requires us to go beyond classification-calibration. Instead of the 01 loss, we need to consider more general discrete losses. Instead of the classifier mapping to $\mathcal{Y} = \{1, \ldots, k\}$, we must allow making predictions in more "exotic" discrete spaces, which we'll denote by $\mathcal{D}$.

Some definitions are in order. Consider functions making predictions in $\mathcal{D}$, i.e., $g : \mathcal{X} \to \mathcal{D}$ and a *discrete loss* $\ell : \mathcal{Y} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$. The ℓ-risk of g is defined as

$$R_\ell(g) := \mathbb{E}_{(X,Y)\sim P}\left[\ell(Y, g(X))\right]. \tag{1.14}$$

Like in Theorem I.1, we are interested ψ satisfying the property that there exists a function $\texttt{pred} : \mathbb{R}^k \to \mathcal{D}$ such that

$$R_\psi(\hat{f}_n) \xrightarrow{P} R^*_\psi \quad \text{implies} \quad R_\ell(\texttt{pred} \circ \hat{f}_n) \xrightarrow{P} R^*_\ell.$$

This general theory behind calibration w.r.t arbitrary discrete losses were developed by Ramaswamy et al. [RA16] extending. Finocchiaro et al. [FFW19] developed a framework for calibration for discrete losses particularly using *polyhedral* surrogate losses, of which the WW hinge loss is one of. Ramaswamy et al. [RTA18] also characterized the calibration theory for the Crammer-Singer hinge loss, another hinge loss for the multiclass SVM that is not classification-calibrated. Using the label encoding $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k$, we completely characterize the optimizers of the conditional risk in Chapter 2

$$\sum_{j=1}^{k} p_j \psi_{\texttt{WW-Hinge}}(\boldsymbol{\rho}_j z)$$

for all $(p_1, \ldots, p_k) \in$ the k-dimensional probability simplex. These optimizers, outside of a measure zero subset of the k-simplex, form a finite discrete subset of the margin space. We show that this finite discrete subset corresponds to a combinatorial object known as the *ordered partitions* of $\{1, \ldots, k\}$, denoted $\mathcal{OP}_k$. See Figure 1.11.

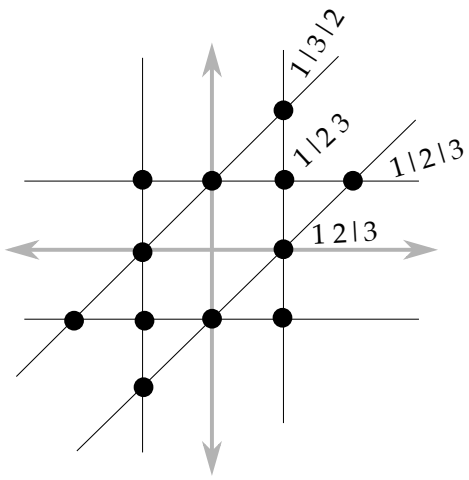

Figure 1.11: Finite discrete subset of minimizers of the conditional WW hinge-risk. Vertices corresponds to the 12 non-trivial order partitions. (The trivial order partition is the "everything in a single bin" partition).

Order partitions can be thought of the set of rankings on k objects, where ties are allowed. When $k = 3$, below are a couple of examples of an ordered partition:

$$\underbrace{1\,2}_{\text{ranked } 1^{\text{st}}} \mid \underbrace{3}_{2^{\text{nd}}} \qquad \underbrace{2}_{1^{\text{st}}} \mid \underbrace{1}_{2^{\text{nd}}} \mid \underbrace{3}_{3^{\text{rd}}} \quad \text{and} \quad \underbrace{2}_{1^{\text{st}}} \mid \underbrace{1\,3}_{2^{\text{nd}}}$$

We give an explicit formula for a discrete loss $\ell : \mathcal{Y} \times \mathcal{OP}_k \to \mathbb{R}_{\geq 0}$ we call the *ordered partition loss* for which the WW hinge loss is calibrated for. We use our calibration results and the formula for ℓ to give theoretical justification for the empirical finding of Doğan et al. [DGI16].

1.3.4 Weston-Watkins SVM: Optimization

In Chapter 3, we turn to the practical and the theoretical question of solving the WW-SVM. State-of-the-art techniques for solve the binary SVM and the Crammer-Singer SVM both employ the so-called *decomposition method*. The decomposition method breaks down the optimization into a series of subproblems. The overall solver's runtime is essentially proportion to how quickly the subproblem can be solved.

For the Crammer-Singer SVM, a $O(k \log k)$ subproblem solver is well-known in the literature [CS01; Duc+08; BFU14; Con16].

Using our novel reparametrization of the optimization objective Eqn. (1.13), we are able to derive the first known algorithm that solves the WW-SVM subproblem also in $O(k \log k)$ time.

1.4 Hyperplane arrangement classifiers and partially quantized neural networks

Previously in Section 1.2, we considered k-ary multiclass classifier obtained from a configuration of k hyperplanes. To derive a classifier, we used the "argmax" function to convert a vector-valued discriminants to a classifier. In this section, we consider a entirely different way to turn a vector-valued discriminants into a classifer based on hyperplane arrangements. Consider a toy dataset as in Figure 1.12 which will be our running example.

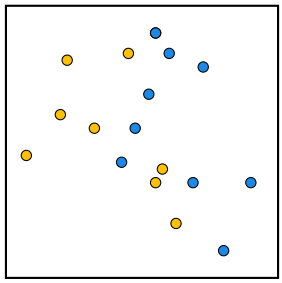

Figure 1.12: A toy binary classification dataset.

We will introduce *hyperplane arrangement classifiers* by demonstrating them in action on the toy dataset. A *hyperplane arrangement* is simply a set of k hyperplanes $\{(\mathbf{w}_i, b_i)\}_{i=1}^k$. See Figure 1.13 left panel. We sometimes say an *arrangement of k hyperplanes* if we need to specify the number of hyperplanes. In contrast to the previous sections, k here no longer denote the number of classes for classification. Instead, we use k to denote an arbitrary positive integer.

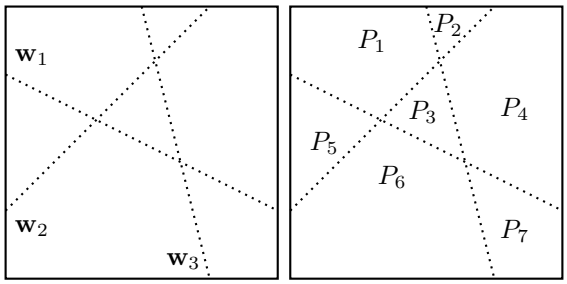

Figure 1.13: *Left*: An arrangement of three hyperplanes in $\mathbb{R}^2$. To lessen notational clutter, we hide the intercept/offset. *Right*: Regions or cells of the hyperplane arrangement.

Observe that in Figure 1.13 right panel, the input space has been partitioned into seven regions labeled by P_i. Each region can be assigned by a unique *sign pattern*, i.e., a vector of the form $\{\pm 1\}^k$. The sign pattern for a region P_i is simply $(\texttt{sgn}(w_1^\top x + b_1), \ldots, \texttt{sgn}(w_k^\top x + b_k))^\top \in \{\pm 1\}^k$ for any x in the interior of P_i. For notation simplicity, we drop the "1" in the sign patterns and write them as vectors of the form $\{\pm\}^k$ instead. See Figure 1.14 left panel.

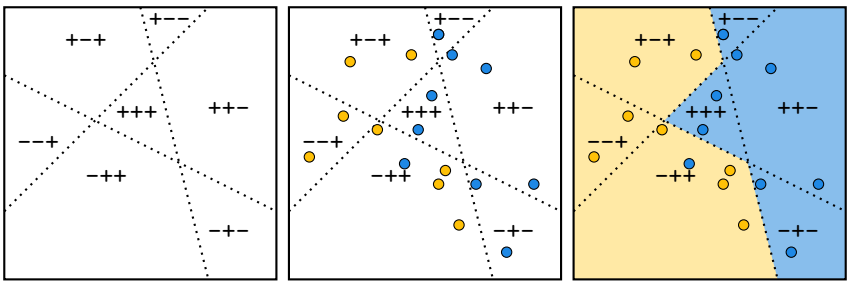

Figure 1.14: *Left*: Sign patterns of the hyperplane arrangement. *Mid*: The hyperplane arrangement overlaid with the toy dataset from Figure 1.12. *Right*: An example of a hyperplane arrangement classifier.

We define *hyperplane arrangement classifiers* (HACs) as functions that are piecewise constant functions over the regions of the arrangement. In other words, all points in a given region is assigned the same label. Consider the toy dataset visualized over the hyperplane arrangements in Figure 1.14 mid panel. Then Figure 1.14 right panel shows the decision region of one possible instance of a hyperplane arrangement clas-

sifier. In fact, the decision for each region is decided by a majority voter.

Another way to think about a hyperplane arrangement classifier involves using a boolean function/look-up table. A *k-look-up table*, or a k-LUT, is a function whose domain is a subset of $\{\pm\}^k$. Suppose that h is a k-look-up table so that all possible sign vectors are in the domain of h. Then a hyperplane arrangement classifier can be expressed as the composition of first mapping a point to its sign vector then applying h to the sign vector. See Figure 1.15.

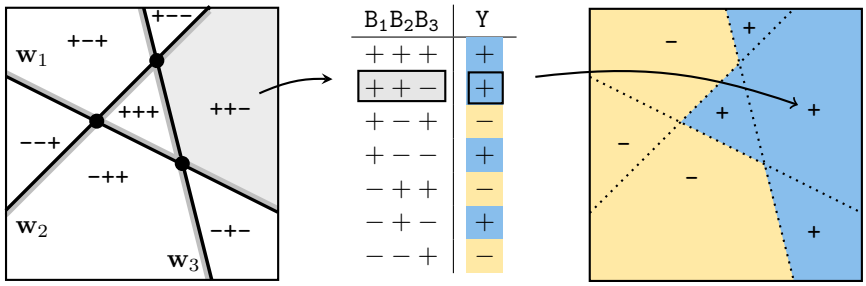

Figure 1.15: All points in gray region are mapped to a sign vector matching the second row of the look up table. The second row of the look up table is mapped to the "positive" class label.

Let $\mathbb{B} = \{\pm\}$. We can alternatively visualize the classifier in Figure 1.15 into the neural network-like architecture shown in Figure 1.16. Moreover, if the LUT is also implemented as neural network, then the entire architecture is a neural network as well.

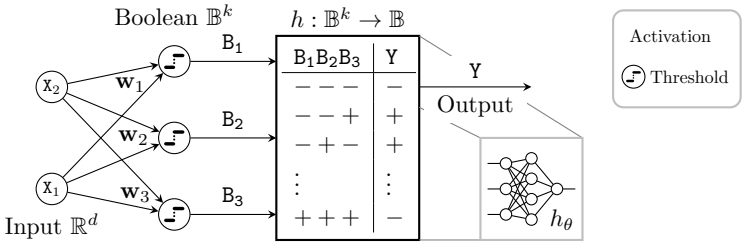

Figure 1.16: A neural network-like architecture for representing a HAC. The input vector $(\mathtt{X}_1, \mathtt{X}_2) \in \mathbb{R}^2$ is mapped to a sign vector $\mathtt{B}_1\mathtt{B}_2\mathtt{B}_3 \in \mathbb{B}^3$. The final output is $\mathtt{Y} = h(\mathtt{B}_1\mathtt{B}_2\mathtt{B}_3)$. The LUT can itself be implemented via a neural network h_θ.

Chapter V will analyze the VC dimension and the minimax theory of hyperplane arrangement classifiers. Moreover, we will see that the VC dimension does not depend on the size of the network used to implement the LUT. Hyperplane arrangement neural networks (HANNs) belong to the family of what is sometimes referred to as a *quantized* neural network. In practice, there's often a performance gap between quantized and non-quantized neural networks [Hub+16]. We benchmark hyperplane arrangement neural networks on 121 UCI datasets and show that its performance matches current state-of-the-art neural non-quantized networks tailored for unstructured datasets such as the UCI data [Kla+17; Wu+18].

1.5 Hyperplane arrangement and random partition kernels

In the previous section, we looked at hyperplane arrangement neural networks. In this section, we consider ensemble of random *hyperplane arrangement classifiers* such as the one shown in Figure 1.17. First, the hyperplanes are sampled randomly. Next, each region is assigned label according to majority rule over data points inside the region.

If we take the ensemble average of many random hyperplane arrangement classifiers as in Figure 1.17, the underlying data distribution becomes apparent in Figure 1.18. Subplot heading $n =$ the size of the ensemble.

Towards rigorously understanding the pattern in Figure 1.18, consider the following. Let `HA` denote a random k-hyperplane arrangement from some fixed distribution. Fix two points x_1, x_2 in the sample space and consider the probability

$$\Pr\{x_1 \text{ and } x_2 \text{ belong to the same region of } \texttt{HA}\}. \tag{1.15}$$

Intuitively, we can interpret the above quantity as a *similarity measure* of the two points. In fact, the above probability is precise the probability of `HA` having at least

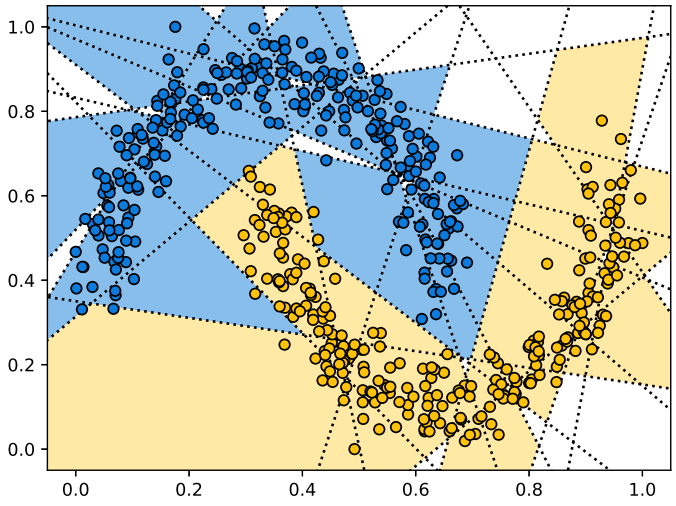

Figure 1.17: An random hyperplane arrangement classifier for the MOONS datasets. Each region is assigned the majority vote label. Regions without training data is shown in white.

one hyperplane intersecting the line segment between x_1 and x_2. See Figure 1.19.

Eqn. (1.15) defines what is known as a *random partition kernel* [DG14]. Let us denote this kernel as $k_{RP}(x_1, x_2)$. In general, it is difficult to write down an analytic formula for k_{RP}. The kernel depends on the distribution of the random hyperplane arrangements.

There is a special case where the kernel can be expressed analytically: when the data x_i and the hyperplane arrangements are restricted to the unit sphere. See Figure 1.20 for an geometric picture of spherical random hyperplane arrangements. In this case, the random partition kernel is given by $k_{RP}(x_1, x_2) = (1 - \angle(x_1, x_2)/\pi)^k$ where $\angle(x_1, x_2)$ is the angle between the two unit vectors and k is the number of hyperplanes.

If we project the moon dataset onto the unit sphere in $\mathbb{R}^3$ and take an ensemble of hyperplane arrangement classifiers[6], we get exactly the kernel smoothing classifier

[6]For technical reasons that will become clear in Chapter VI, we are able to prove this result when the hyperplane arrangement classifiers use a *weighted* majority vote rather than the vanilla majority

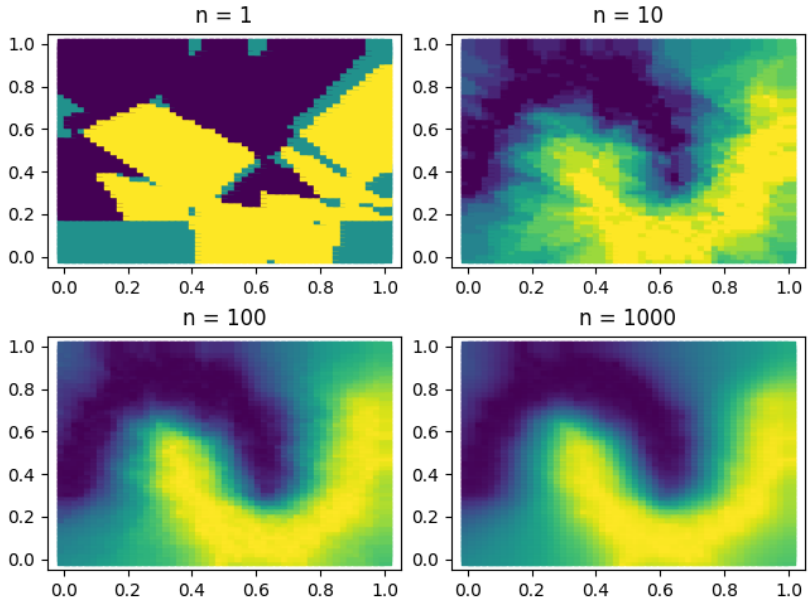

Figure 1.18: Ensemble of n random hyperplane arrangement classifiers.

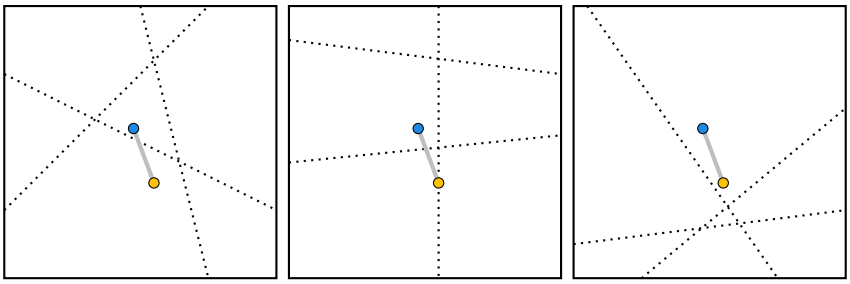

Figure 1.19: Three random hyperplane arrangements. In the left two panels, the points shown do not belong to the same region. In the right panel, the two points belong to the same region.

with the kernel k_{RP}:

$$x \mapsto \sum_{i=1}^{n} y_i k_{RP}(x, x_i).$$

This classifier is plotted in the bottom right panel of Figure 1.21. Note that the bottom right panel is the theoretically computed infinite ensemble. Observe that as $n \in \{1, 10, 100\}$ increases, the behavior of the finite ensemble approaches the classifier in the bottom right panel. In Chapter VI, we prove these facts rigorously. Further-

vote.

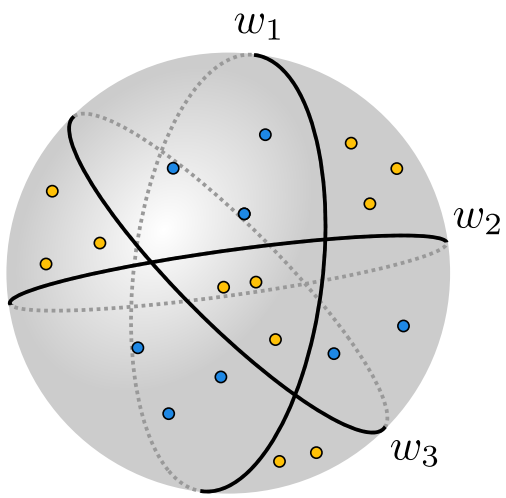

Figure 1.20: Spherical hyperplane arrangement.

more, we use the theoretical tools we develop to obtain the first demonstration of an interpolating ensemble method that is consistent for a broad class of distributions in arbitrary dimensions.

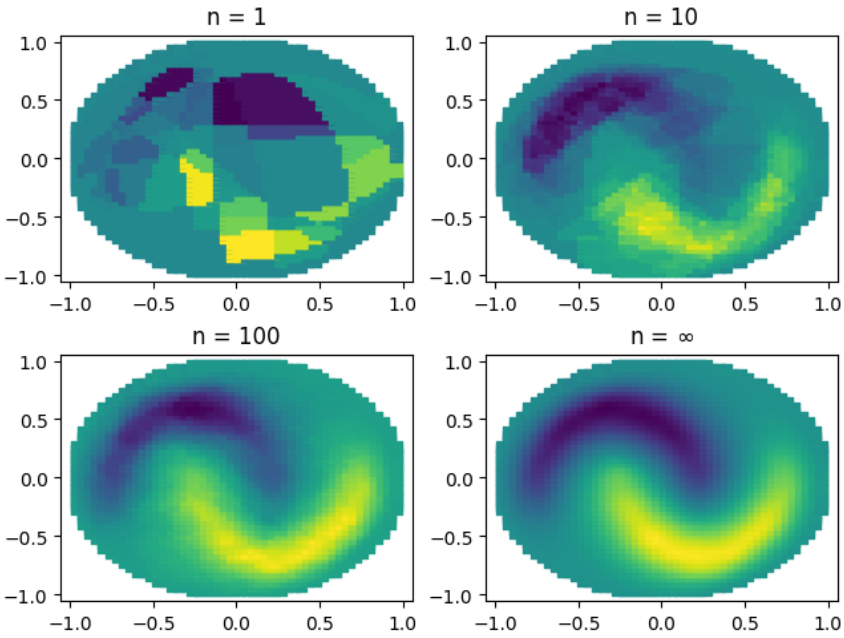

Figure 1.21: Ensemble classifier of hyperplane arrangement classifiers.

CHAPTER II

Weston-Watkins Hinge Loss and Ordered Partitions

Multiclass extensions of the support vector machine (SVM) have been formulated in a variety of ways. A recent empirical comparison of nine such formulations [DGI16] recommends the variant proposed by Weston and Watkins (WW), despite the fact that the WW-hinge loss is not calibrated with respect to the 0-1 loss. In this work we introduce a novel discrete loss function for multiclass classification, the *ordered partition loss*, and prove that the WW-hinge loss *is* calibrated with respect to this loss. We also argue that the ordered partition loss is *minimally emblematic* among discrete losses satisfying this property. Finally, we apply our theory to justify the empirical observation made by Doğan et al. [DGI16] that the WW-SVM can work well even under massive label noise, a challenging setting for multiclass SVMs.

2.1 Introduction

Classification is the task of assigning labels to instances, and a common approach is to minimize misclassification error corresponding to the 0-1 loss. However, the 0-1 loss is discrete and typically cannot be optimized efficiently. To address this, the 0-1 loss is often replaced by a surrogate loss during training. If the surrogate is *calibrated*

with respect to the 0-1 loss, then a classifier minimizing the expected surrogate loss will also minimize the expected 0-1 loss in the infinite sample limit.

For multiclass classification, several different multiclass extensions of the support vector machine (SVM) have been proposed, including the Weston-Watkins (WW) [WW99], Crammer-Singer (CS) [CS01], and Lee-Lin-Wahba (LLW) [LLW04] SVMs. The pertinent difference between these multiclass SVMs is the multiclass generalization of the hinge loss. Below, we refer to the hinge loss from WW-SVM as the WW hinge loss and so on. It is well-known that the LLW-hinge is calibrated with respect to the 0-1 loss, while the WW- and CS-hinge losses are not [Liu07; TB07].

Despite this result, the LLW-SVM is not more widely accepted than the WW-, CS-, and other SVMs. The first reason for this is that while the LLW-SVM is calibrated with respect to the 0-1 loss, this did not lead to superior performance empirically. In particular, Doğan et al. [DGI16] found that the LLW-SVM fails in low dimensional feature space even under the noiseless setting. On the other hand, Doğan et al. [DGI16] observed that the WW-SVM is the only multiclass SVM that succeeded in both the noiseless and noisy setting in their simulations. Indeed, Doğan et al. [DGI16] concluded that, among 9 different competing multiclass SVMs, the WW-SVM offers the best overall performance when considering accuracy and computation. The second reason is that the calibration framework is not limited to the 0-1 loss. There could be other discrete losses with respect to which a surrogate is calibrated, and which help to explain its performance. Indeed, Ramaswamy et al. [RTA18] recently showed that the CS-hinge loss is calibrated with respect to a discrete loss for classification with abstention.

In a vein similar to [RTA18], we show that the WW-hinge loss is calibrated with respect to a novel discrete loss that we call the *ordered partition* loss. Our results leverage the embedding framework for analyzing discrete losses and convex piecewise linear surrogates, introduced recently by Finocchiaro et al. [FFW19]. We also give

theoretical justification for the empirical performance of the WW-SVM observed by Doğan et al. [DGI16].

2.1.1 Related work

Cortes et al. [CV95] introduced the support vector machine for learning a binary classifier, using the hinge loss as a surrogate for the 0-1 loss. Steinwart [Ste02] showed that the binary SVM is *universally consistent*, a desirable property of a classification algorithm that ensures its convergence to the Bayes optimal classifier in the large sample limit. Steinwart [Ste05] later used calibration to give a more general proof of SVM consistency with respect to the 0-1 loss. Around that time, more general theories of when a loss is calibrated with respect to 0-1 loss, or "classification calibrated," began to emerge [Zha04a; BJM06; Ste07], and since then a proliferation of papers have extended these ideas to a variety of learning settings (see Bao et al. [BSS20] for a recent review).

Several natural extensions of the binary SVM exist, including the Weston-Watkins (WW) [WW99], Crammer-Singer (CS) [CS01], and Lee-Lin-Wahba (LLW) [LLW04] SVMs. Tewari et al. [TB07] extended the definition of calibration with respect to the 0-1 loss to the multiclass setting. Liu [Liu07] and Tewari et al. [TB07] analyzed these hinge losses and showed that WW and CS hinge losses are not calibrated with respect to the 0-1 loss while the LLW hinge loss is. Doğan et al. [DGI16] introduced a framework that unified existing multiclass SVMs, proved the 0-1 loss consistency of several multiclass SVMs when the kernel is allowed to change, and also conducted extensive experiments. Despite not being calibrated with respect to the 0-1 loss, Zhang [Zha04a] showed that the Crammer-Singer SVM is consistent given the "majority assumption", i.e., the most probable class has greater than 1/2 probability. When the majority assumption is violated, experiments conducted by Doğan et al. [DGI16] suggested that the CS-SVM fails, while the WW-SVM continues to perform well.

The LLW-hinge loss is calibrated with respect to the 0-1 loss while the WW-hinge loss is not [Liu07]. Nevertheless, the WW-SVM often outperforms the LLW-SVM in experiments [DGI16] which ostensibly undermines using calibration be as a justification for performance. To reconcile this, we refer the reader to the discussion in Doğan et al. [DGI16, Section 3.3] on *relative* and *absolute margin* losses. Doğan et al. [DGI16] argued that the poorer performance of losses based on absolute margin, including the LLW-hinge, is due to the issue of the absolute margin being incompatible the decision function. On the other hand, the CS and WW-hinge losses are relative margin based and do not suffer the same issue. We remark that Fathony et al. [Fat+16] proposed a relative margin hinge loss which is calibrated with respect to the 0-1 loss that outperforms the WW-hinge loss at the expense of greater computational complexity.

Ramaswamy et al. [RA16] extended the notion of calibration to an arbitrary discrete loss used in *general multiclass learning*. The general multiclass learning framework unifies several learning problems, including cost-sensitive classification [Sco12], classification with abstain option [RTA18], ranking [DMJ13], and partial label learning [Cid12]. Furthermore, Ramaswamy et al. [RA16] introduced the concept of *convex calibration dimension* which is defined for a discrete loss to be the minimum dimension required for the domain of a convex surrogate loss to be calibrated with respect to the given discrete loss. Ramaswamy et al. [RTA18] proved the consistency of CS-SVM with respect to the abstention loss where the cost of abstaining is $1/2$ by showing that the CS hinge is calibrated with respect to this abstention loss. They also proposed a new calibrated convex surrogate loss in dimension $\lceil \log_2 k \rceil$ for the abstention loss, implying that the CS hinge is suboptimal from the CC-dimension perspective.

Recently, several new multiclass hinge-like losses have been proposed, as well as frameworks for constructing convex losses. Doğan et al. [DGI16] used their framework to devise two new multiclass hinge losses, and using ideas from adversarial multiclass

classification, Fathony et al. [Fat+16] proposed a new multiclass hinge-like loss; all three are calibrated with respect to the 0-1 loss. Blondel et al. [BMN20] introduced a class of losses known as *Fenchel-Young losses* which contains non-smooth losses such as the CS hinge loss as well as smooth losses such as the logistic loss. Tan et al. [TZ20] proposed an approach for constructing hinge-like losses using generalized entropies. Finocchiaro et al. [FFW19] studied the calibration properties of *polyhedral* losses using the *embedding* framework that they developed. They analyzed several polyhedral losses in the literature including the CS hinge, the Lovász hinge [YB18], and the top-n loss [LHS17].

2.1.2 Our contributions

We introduce a novel discrete loss ℓ, the *ordered partition loss*. We show in theorem II.8 that the Weston-Watkins hinge loss L embeds the ordered partition loss ℓ. Our embedding result together with results of [FFW19] imply that L is calibrated with respect to ℓ (corollary II.9). To the best of our knowledge, this is the first calibration-theoretic result for the WW-hinge loss. We also introduce the notion of the *minimally emblematic* discrete loss that a polyhedral loss can embed and argue that the ordered partition loss is minimally emblematic for the WW-hinge loss. In section 2.5, we use properties of the ordered partition loss to give theoretical support for the empirical observations made by Doğan et al. [DGI16] on the success of WW-SVM in the massive label noise setting.

2.1.3 Notations

Let $k \geq 3$ be an integer which denotes the number of classes. For a positive integer n, we let $[n] = \{1, \ldots, n\}$. If $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$ and $i \in [k]$ is an index, then let $[v]_i := v_i$. Define $\max v = \max_{i \in [k]} v_i$ and $\arg\max v = \{i \in [k] : v_i = \max v\}$.

Let $\mathfrak{S}_k$ denote the set of permutations on $[k]$, i.e., elements of $\mathfrak{S}_k$ are bijections

$\sigma : [k] \to [k]$. Given $\sigma \in \mathfrak{S}_k$ and $v \in \mathbb{R}^k$, the vector $\sigma v \in \mathbb{R}^k$ is defined entrywise where the i-th entry is $[\sigma v]_i = v_{\sigma(i)}$. Equivalently, we view $\mathfrak{S}_k$ as the set of permutation matrices in $\mathbb{R}^{k \times k}$.

Let $\mathbb{R}_+$ denote the set of nonnegative reals. Denote $\Delta^k = \{(p_1, \ldots, p_k) \in \mathbb{R}_+^k : p_1 + \cdots + p_k = 1\}$ the probability simplex. For $p \in \Delta^k$, we write $Y \sim p$ to denote a discrete random variable $Y \in [k]$ whose probability mass function is p. Let $\langle \cdot, \cdot \rangle$ be the usual dot-product between vectors. Denote by $\mathbb{I}\{\texttt{input}\}$ the indicator function which returns 1 if `input` is true and 0 otherwise.

2.1.4 Background

Recall the *general multiclass learning* framework as described in [RA16]: $\mathcal{X}$ is a sample space and P is a joint distribution over $\mathcal{X} \times [k]$. A *multiclass classification loss* is a function $\ell : \mathcal{R} \to \mathbb{R}_+^k$ where $\mathcal{R}$ is called the *prediction space* and $[\ell(r)]_y \in \mathbb{R}_+$ is the penalty incurred for predicting $r \in \mathcal{R}$ when the label is $y \in [k]$. If $\mathcal{R}$ is finite, we refer to ℓ as a *discrete loss*. For example, a common setting for classification is $\mathcal{R} = [k]$ and ℓ is the 0-1 loss. The *ℓ-risk* of a *hypothesis* function $f : \mathcal{X} \to \mathcal{R}$ is

$$\mathrm{er}_P^\ell(f) := \mathbb{E}_{X,Y \sim P} \left\{ [\ell(f(X))]_Y \right\}. \tag{2.1}$$

The goal is to design *ℓ-consistent* algorithms, i.e., procedures that output a hypothesis f_n based on an input of n training samples sampled i.i.d from P such that $\mathrm{er}_P^\ell(f_n) \to \mathrm{er}_P^{\ell,*} = \inf_{f:\mathcal{X}\to\mathcal{R}} \mathrm{er}_P^\ell(f)$ as $n \to \infty$. Since ℓ is discrete, eq. (2.1) is difficult to directly minimize. To circumvent this difficulty, we consider a convex *surrogate loss* $L : \mathbb{R}^d \to \mathbb{R}^k$ for some positive integer d. The following property relates the surrogate loss L and the discrete loss ℓ.

Definition II.1 (Calibration)**.** For each $p \in \Delta^k$, define $\gamma_\ell(p) := \arg\min_{r \in \mathcal{R}} \langle p, \ell(r) \rangle$. We say that L *is calibrated with respect to* ℓ if there exists a function $\psi : \mathbb{R}^d \to \mathcal{R}$

such that for all $p \in \Delta^k$

$$\inf_{u \in \mathbb{R}^d : \psi(u) \notin \gamma_\ell(p)} \langle p, L(u) \rangle > \inf_{v \in \mathbb{R}^d} \langle p, L(v) \rangle.$$

By Ramaswamy et al. [RA16, Theorem 3], L being calibrated with respect to ℓ is equivalent to the following: there exists $\psi : \mathbb{R}^d \to \mathcal{R}$ such that for all joint distributions P on $\mathcal{X} \times [k]$ and all sequences of functions $g_n : \mathcal{X} \to \mathbb{R}^d$, we have

$$\mathrm{er}_P^L(g_n) \to \mathrm{er}_P^{L,*} \quad \text{implies} \quad \mathrm{er}_P^\ell(\psi \circ g_n) \to \mathrm{er}_P^{\ell,*}$$

where $\mathrm{er}_P^{L,*} = \inf_{g : \mathcal{X} \to \mathbb{R}^d} \mathrm{er}_P^L(g)$. Thus, the calibration property allows us to focus on finding L-consistent algorithms. In general it can be difficult to check that a given L is calibrated with respect to ℓ. Finocchiaro et al. [FFW19] introduced the following definition:

Definition II.2 (Finocchiaro et al. [FFW19])**.** The loss $L : \mathbb{R}^d \to \mathbb{R}^k$ *embeds* $\ell : \mathcal{R} \to \mathbb{R}^k$ if there exists an injection $\varphi : \mathcal{R} \to \mathbb{R}^d$ called an *embedding* such that

1. $L(\varphi(r)) = \ell(r)$ for all $r \in \mathcal{R}$
2. $r \in \arg\min_{r \in \mathcal{R}} \langle p, \ell(r) \rangle$ if and only if $\varphi(r) \in \arg\min_{v \in \mathbb{R}^d} \langle p, L(v) \rangle$.

The notion of embedding is important due to the following result from [FFW19, Theorem 3]:

Theorem II.3 (Finocchiaro et al. [FFW19])**.** *Let L be convex piecewise-linear and ℓ be discrete. If L embeds ℓ, then L is calibrated with respect to ℓ.*

Given L, ℓ and φ, Finocchiaro et al. [FFW19, Definition 6] provided an explicit construction for ψ with excess risk bound proved in [FFW19, Theorem 6].

In this work, we are interested in the case when L is the WW-hinge loss:

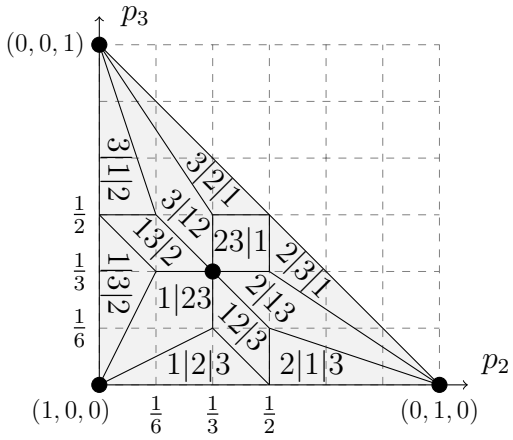

Figure 2.1: The gray triangle represents the probability simplex Δ^3, where $(p_1, p_2, p_3) \in \Delta^3$ is plotted as (p_2, p_3) in the plane. The interior of each polygonal region contains $p \in \Delta^3$ such that $\min_{\mathbf{S} \in \mathcal{OP}_k} \langle p, \ell(\mathbf{S}) \rangle$ has a unique minimizer. For the derivations, see Section 2.8. Ordered partitions are represented as follows:

$(\{1\}, \{2, 3\}) \mapsto 1|23,$

$(\{1\}, \{2\}, \{3\}) \mapsto 1|2|3,$

$\vdots$

$(\{3\}, \{2\}, \{1\}) \mapsto 3|2|1.$

Definition II.4. For $v \in \mathbb{R}^k$, define the *Weston-Watkins hinge loss* [WW99] $L(v) \in \mathbb{R}^k_+$ entrywise by

$$[L(v)]_y = \sum_{i \in [k] \,:\, i \neq y} h(v_y - v_i), \qquad y \in [k]$$

where $h : \mathbb{R} \to \mathbb{R}_+$ is the *hinge function* defined by $h(x) = \max\{0, 1 - x\}$.

By theorem II.3, to prove that L is calibrated with respect to ℓ, it suffices to show that L embeds ℓ. Going forward, L will refer to the WW-hinge loss. We now work toward showing that L embeds the ordered partition loss ℓ, which we introduce next.

2.2 The ordered partition loss

The prediction space $\mathcal{R}$ that we use is the set of ordered partitions, which we now define:

Definition II.5. An *ordered partition* on $[k]$ of length l is an ordered list $\mathbf{S} = (S_1, \ldots, S_l)$ of nonempty, pairwise disjoint subsets of $[k]$ such that $S_1 \cup \cdots \cup S_l = [k]$. Denote by $\mathcal{OP}_k$ the set of all ordered partitions on $[k]$ with length ≥ 2. We write the length of $\mathbf{S}$ as $l_{\mathbf{S}}$ to be precise when working with multiple ordered partitions.

Ordered partitions can be thought of as a complete ranking of k items where ties are allowed. They are widely studied in combinatorics [Man12; Gro62; IKZ08]. In the ranking literature, ordered partitions are called *bucket orders* [Fag+04] and the S_is are called the *buckets*. The first bucket S_1 contains the highest ranked items, and so on. There is only one ordered partition with $l_{\mathbf{S}} = 1$, namely the *trivial partition* $\mathbf{S} = ([k])$. Thus, $\mathcal{OP}_k$ is the set of nontrivial ordered partitions.

We now define the following discrete loss over the ordered partitions:

Definition II.6. The *ordered partition loss* $\ell : \mathcal{OP}_k \to \mathbb{R}_+^k$ is defined, for $i \in [k]$ and $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$, as $[\ell(\mathbf{S})]_i = |S_1| - 1 + \sum_{j=1}^{l_{\mathbf{S}}-1} |S_1 \cup \cdots \cup S_{j+1}| \cdot \mathbb{1}\{i \notin S_1 \cup \cdots \cup S_j\}$.

The intuition behind the ordered partition loss is that we want to rank the labels, where ties are allowed and each S_i is a set of labels that are tied. We want the correct label to be as high up the ranking as possible. The lower the true class is ranked, the larger the loss.

To build intuitions about ℓ, let $Y \sim p$ and consider the random variable $[\ell(\mathbf{S})]_Y$ whose expectation is

$$\mathbb{E}_{Y \sim p} \left\{ [\ell(\mathbf{S})]_Y \right\} = |S_1| - 1 + \sum_{j=1}^{l_{\mathbf{S}}-1} |S_1 \cup \cdots \cup S_{j+1}| \cdot \Pr_{Y \sim p} \left\{ Y \notin S_1 \cup \cdots \cup S_j \right\}. \quad (2.2)$$

Note that $\mathbb{E}_{Y \sim p} \left\{ [\ell(\mathbf{S})]_Y \right\} = \langle p, \ell(\mathbf{S}) \rangle$. In fig. 2.1, we visualize the decision rule for the Bayes optimal classifier in the $k = 3$ case by plotting the function $p \mapsto \arg\min_{\mathbf{S} \in \mathcal{OP}_k} \langle p, \ell(\mathbf{S}) \rangle$. When $l_{\mathbf{S}} = 2$, we have

$$\mathbb{E}_{Y \sim p} \left\{ [\ell(\mathbf{S})]_Y \right\} = |S_1| - 1 + k \Pr_{Y \sim p} \left\{ Y \notin S_1 \right\}. \quad (2.3)$$

Thus, we have a trade-off where adding elements to S_1 increases the $|S_1| - 1$ term but decreases the $k \Pr_{Y \sim p}\{Y \notin S_1\}$ term. More generally, when $l_{\mathbf{S}} \geq 2$, the ordered partition loss requires the predictor to associate each test instance x with a nested sequence of sets $S_1, S_1 \cup S_2, \cdots$ where these sets are designed to balance the probability of containing x's label with the size of the set. In the learning with partial labels settings [CST11; Cid12], for each training instance the learner observes a set of labels, one of which is the true label. The sets $S_1, S_1 \cup S_2, \ldots$ might be called *progressive partial labels* in the spirit of partial label learning [CST11; Cid12].

Next, we define the embedding that satisfies definition II.2 when L is the WW-hinge loss and ℓ is the ordered partition loss:

Definition II.7. The *embedding* $\varphi : \mathcal{OP}_k \to \mathbb{R}^k$ is defined as follows: Let $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$. Define $\varphi(\mathbf{S}) \in \mathbb{R}^k$ entrywise so that for all $i \in [l_{\mathbf{S}}]$ and all $j \in S_i$, we have $[\varphi(\mathbf{S})]_j = -(i-1)$.

With the discrete loss ℓ and the embedding map φ defined, we now proceed to the main results.

2.3 Main results

In this work, we establish that the WW-hinge loss embeds the ordered partition loss:

Theorem II.8. *The Weston-Watkins hinge loss $L : \mathbb{R}^k \to \mathbb{R}^k$ embeds the ordered partition loss $\ell : \mathcal{OP}_k \to \mathbb{R}^k$ with embedding φ as in definition II.7.*

In light of theorem II.3, theorem II.8 implies

Corollary II.9. *L is calibrated with respect to ℓ.*

In the remainder of this section, we develop the tools necessary to prove theorem II.8.

2.3.1 Vectorial representation of ordered partitions

First, we define the set $\mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$ whose elements serve as realizations of ordered partitions inside $\mathbb{R}^k$.

Definition II.10. Define the following sets:

$$\mathcal{C} := \{v \in \mathbb{R}^k : v_1 = 0, v_k \leq -1,\ v_i - v_{i+1} \in [0,1],\ \forall i \in [k-1]\}, \qquad \mathcal{C}_{\mathbb{Z}} := \mathcal{C} \cap \mathbb{Z}^k \quad (2.4)$$

and finally $\mathfrak{S}_k\mathcal{C}_{\mathbb{Z}} := \bigcup_{\sigma \in \mathfrak{S}_k} \sigma\mathcal{C}_{\mathbb{Z}}$ where $\sigma\mathcal{C}_{\mathbb{Z}} = \{\sigma v : v \in \mathcal{C}_{\mathbb{Z}}\}$.

A vector $v \in \mathbb{R}^k$ is *monotonic non-increasing* if $v_1 \geq v_2 \geq \cdots \geq v_k$. Note that vectors in $\mathcal{C}_{\mathbb{Z}}$ are nonconstant, integer-valued monotonic non-increasing such that consecutive entries decrease at most by 1. Furthermore, by construction, $\mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$ consists of all possible permutations of elements in $\mathcal{C}_{\mathbb{Z}}$. Therefore, the entries of an element $v \in \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$ take on every value in $0, -1, \ldots, -(l-1)$ for some integer $l \in \{2, \ldots, k\}$. Thus, $v \in \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$ can be thought of as vectorial representation of the ordered partition $\mathbf{S} = (S_1, \ldots, S_l)$ where $S_i = \{j : v_j = -(i-1)\}$ for each $i \in [l]$. In proposition II.13 below, we make this notion precise.

Lemma II.11. *The image of φ is contained in $\mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$.*

Proof. Let $\mathbf{S} \in \mathcal{OP}_k$. It suffices to prove that there exists some $\sigma \in \mathfrak{S}_k$ such that $\sigma\varphi(\mathbf{S}) \in \mathcal{C}_{\mathbb{Z}}$. Note that by definition, we have the set of unique values of $\varphi(\mathbf{S})$ is

$$\{[\varphi(\mathbf{S})]_j : j \in [k]\} = \{0, -1, -2, \ldots, -(l_{\mathbf{S}} - 1)\}.$$

Thus, let $\sigma \in \mathfrak{S}_k$ be such that $\sigma\varphi(\mathbf{S})$ is monotonic non-increasing. Then $\sigma\varphi(\mathbf{S}) \in \mathcal{C}_{\mathbb{Z}}$. □

Next, we define the inverse of φ.

Definition II.12. The *quasi-link map* $\tilde{\psi} : \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}} \to \mathcal{OP}_k$ is defined as follows: Given $v \in \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$, let $l = 1 - \min_{j\in[k]} v_j$. Define $S_i = \{j \in [k] : v_j = -(i-1)\}$ for each $i \in [l]$. Finally, define $\tilde{\psi}(v) = (S_1, \ldots, S_l)$.

The tilde in $\tilde{\psi}$ is to differentiate the quasi-link from ψ in definition II.1.

Proposition II.13. *The embedding map $\varphi : \mathcal{OP}_k \to \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$ given in definition II.7 is a bijection with inverse given by the quasi-link map $\tilde{\psi}$ from definition II.12.*

Proof. We first show that for all $\tilde{\psi}(\varphi(\mathbf{S})) = \mathbf{S}$ for all $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$. Observe that $S_i = \{j \in [k] : [\varphi(\mathbf{S})]_j = -(i-1)\}$ for all $i = 1, 2, \ldots, l$. This implies that $\tilde{\psi}(\varphi(\mathbf{S})) = \mathbf{S}$.

Next, we show that $\varphi(\tilde{\psi}(v)) = v$ for all $v \in \mathfrak{S}_k\mathcal{C}_{\mathbb{Z}}$. Let $\mathbf{S} = (S_1, \ldots, S_l) = \tilde{\psi}(v)$. Then $[\varphi(\mathbf{S})]_j = -(i-1)$ if and only if $j \in S_i$. By definition $S_i = \{j \in [k] : v_j = -(i-1)\}$. Hence, $[\varphi(\mathbf{S})]_j = -(i-1)$ if and only if $v_j = -(i-1)$ which implies that $\varphi(\mathbf{S}) = v$, as desired. □

In the next section, using φ, we prove a relationship between the inner risk functions of L and ℓ.

2.3.2 Inner risk functions

Define the *inner risk* functions $\underline{L} : \Delta^k \to \mathbb{R}_+$ and $\underline{\ell} : \Delta^k \to \mathbb{R}_+$ as follows:

$$\underline{L}(p) = \inf_{v\in\mathbb{R}^k} \langle p, L(v)\rangle, \quad \text{and} \quad \underline{\ell}(p) = \inf_{\mathbf{S}\in\mathcal{OP}_k} \langle p, \ell(\mathbf{S})\rangle. \tag{2.5}$$

Note that these functions appear in the second part of definition II.2, although here we have inf instead of min. Since $\mathcal{OP}_k$ is finite, the infimum in the definition of $\underline{\ell}$ is attained. Later, we will argue that the infimum in the definition of $\underline{L}$ is also attained.

We now state the main structural result regarding $\underline{L}$:

Theorem II.14. *For all $p \in \Delta^k$, we have*

$$\underline{L}(p) = \min_{v \in \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}} \langle p, L(v) \rangle.$$

Sketch of proof. Note that L is invariant under translation by any scalar multiple of the all ones vector. Thus, L has an extra degree of freedom. We introduce a loss function $\ell : \mathbb{R}^{k-1} \to \mathbb{R}^k$ called the *reduced WW-hinge loss*, which removes this extra degree freedom. Furthermore, there exists a mapping $\pi : \mathbb{R}^k \to \mathbb{R}^{k-1}$ such that $\langle p, L(v) \rangle = \langle p, \ell(\pi(v)) \rangle$ for all $p \in \Delta^k$ and $v \in \mathbb{R}^k$. Letting $z = \pi(v) \in \mathbb{R}^{k-1}$, we show that for a fixed p, the function $F_p(z) := \langle p, \ell(z) \rangle$ is convex and piecewise-linear and the minimization of which can be formulated as a linear program [BT97]. Furthermore, since F_p is nonnegative, the infimum $\inf_{z \in \mathbb{R}^{k-1}} F_p(z)$ is attained [BT97, Corollary 3.2], which implies that the infimum in the definition of $\underline{L}$ in eq. (2.5) is attained as well. The linear program is shown to be totally unimodular, which implies that an integral solution exists [Law01], i.e., $\min_{z \in \mathbb{R}^{k-1}} F_p(z) = F_p(z^*)$ for some $z^* \in \mathbb{Z}^{k-1}$. From z^*, we obtain an integral $v^* \in \mathbb{Z}^k$ such that $\underline{L}(p) = \langle p, L(v^*) \rangle$. Finally, we construct an element $v^\dagger \in \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}$ from v^* in such a way that the objective does not increase, i.e., $\langle p, L(v^*) \rangle \geq \langle p, L(v^\dagger) \rangle$, which implies that $\underline{L}(p) = \langle p, L(v^\dagger)$ by the optimality of v^*. □

The ordered partition loss ℓ and the WW-hinge loss L are related by the following:

Theorem II.15. *For all $p \in \Delta^k$ and all $\mathbf{S} \in \mathcal{OP}_k$, we have*

$$\langle p, \ell(\mathbf{S}) \rangle = \langle p, L(\varphi(\mathbf{S})) \rangle,$$

where φ is the embedding map as in definition II.7.

Sketch of proof. Let $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$ and $p \in \Delta^k$. Let $T \in \mathbb{R}^{k \times k}$ consist of ones on and below the main diagonal and zero everywhere else. Letting $D = T^{-1}$, we

have

$$\langle p, L(\varphi(\mathbf{S}))\rangle = \langle p, TDL(\varphi(\mathbf{S}))\rangle = \langle T'p, DL(\varphi(\mathbf{S}))\rangle.$$

Next, we observe that $[T'p]_i = p_i + \cdots + p_k$ for each $i \in [k]$. We then show through a lengthy calculation that for each $i \in [k]$

1. If $i = 1$, then $[T'p]_1 = 1$ and $[DL(\varphi(\mathbf{S}))]_1 = |S_1| - 1$.
2. If $i > 1$ and $i = |S_1 \cup \cdots \cup S_j| + 1$ for some $j \in [l]$, then $[T'p]_i = \Pr_{Y\sim p}\{Y \notin S_1 \cup \cdots \cup S_j\}$ and $[DL(\varphi(\mathbf{S}))]_i = |S_1 \cup \cdots \cup S_{j+1}|$.
3. For all other i, $[DL(\varphi(\mathbf{S}))]_i = 0$ (in which case the value of $[T'p]_i$ is irrelevant).

From this, we deduce that $\langle T'p, DL(\varphi(\mathbf{S}))\rangle$ is equal to eq. (2.2). □

Next, we show that the inner risks of L and ℓ from eq. (2.5) are in fact identical:

Corollary II.16. *For all $p \in \Delta^k$, we have $\underline{L}(p) = \underline{\ell}(p)$.*

Proof. Observe that

$$\underline{\ell}(p) \overset{(a)}{=} \min_{\mathbf{S}\in\mathcal{OP}_k} \langle p, \ell(\mathbf{S})\rangle \overset{(b)}{=} \min_{\mathbf{S}\in\mathcal{OP}_k} \langle p, L(\varphi(\mathbf{S}))\rangle \overset{(c)}{=} \min_{v\in\mathfrak{S}_k\mathcal{C}_\mathbb{Z}} \langle p, L(v)\rangle \overset{(d)}{=} \underline{L}(p)$$

where (a) follows from definition of $\underline{\ell}$, (b) from theorem II.15, (c) from the fact that $\varphi : \mathcal{OP}_k \to \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ is a bijection (proposition II.13), and (d) from theorem II.14. □

Having developed all the tools necessary, we turn toward the proof of our main result theorem II.8.

2.3.3 Proof of theorem II.8

We check that the two conditions in definition II.2 holds. The first condition is that $L(\varphi(\mathbf{S})) = \ell(\mathbf{S})$ for all $\mathbf{S} \in \mathcal{OP}_k$, which follows from theorem II.15. To see this,

note that for all $i \in [k]$ the i-th elementary basis vector $e_i \in \Delta^k$. Thus, we have

$$[L(\varphi(\mathbf{S}))]_i = \langle e_i, L(\varphi(\mathbf{S}))\rangle = \langle e_i, \ell(\mathbf{S})\rangle = [\ell(\mathbf{S})]_i$$

for all $i \in [k]$. This implies that $L(\varphi(\mathbf{S})) = \ell(\mathbf{S})$, which is the first condition of definition II.2.

Next, we check the second condition. Let $p \in \Delta^k$. Define $\gamma(p) := \arg\min_{\mathbf{S}\in\mathcal{OP}_k}\langle p, \ell(\mathbf{S})\rangle$, and $\Gamma(p) := \arg\min_{v\in\mathbb{R}^k}\langle p, L(v)\rangle$. Furthermore, by the definition of γ, $\mathbf{S} \in \gamma(p)$ if and only if $\langle p, \ell(\mathbf{S})\rangle = \underline{\ell}(p)$. Likewise, $\varphi(\mathbf{S}) \in \Gamma(p)$ if and only if $\langle p, L(\varphi(\mathbf{S}))\rangle = \underline{L}(p)$. By corollary II.16 and theorem II.15, we have $\langle p, \ell(\mathbf{S})\rangle = \underline{\ell}(p)$ if and only if $\langle p, L(\varphi(\mathbf{S}))\rangle = \underline{L}(p)$. Putting it all together, we get $\mathbf{S} \in \gamma(p)$ if and only if $\varphi(\mathbf{S}) \in \Gamma(p)$, which is the second condition of definition II.2.

2.4 Minimially emblematic losses

Going forward, let $L : \mathbb{R}^d \to \mathbb{R}^k_+$ be a generic surrogate loss. The WW-hinge loss is denoted by L^{WW} and the CS-hinge loss by L^{CS}. Likewise, let $\ell : \mathcal{R} \to \mathbb{R}^k_+$ be a generic discrete loss. The ordered partition loss is denoted by $\ell^{\mathcal{OP}}$ and the 0-1 loss by ℓ^{zo}.

We define a "dual" notion to the embedding dimension Finocchiaro et al. [FFW20, Definition 6]:

Definition II.17. Let $L : \mathbb{R}^d \to \mathbb{R}^k_+$ be a loss. Define the *embedding cardinality* of L as

$$\text{emb.card}(L) := \min\left\{n \in \{2, 3, \dots\} \mid \begin{array}{c}\text{there exists a discrete loss } \ell:[n]\to\mathbb{R}^k \\ \text{such that } L \text{ embeds } \ell\end{array}\right\}.$$

A discrete loss $\ell : \mathcal{R} \to \mathbb{R}^k$ is said to be *minimally emblematic* for L if $|\mathcal{R}| = \text{emb.card}(L)$ and L embeds ℓ.

Remark II.18. Intuitively, ℓ is minimally emblematic for L with embedding φ if $\varphi(\mathcal{R})$ captures all the *essential information* contained in the surrogate L in the most compact way. Let us say that a set of vectors $E \subseteq \mathbb{R}^k$ is an *emblem* of L if for all $p \in \Delta^k$, the set $E \cap \operatorname{argmin}_v \langle p, L(v) \rangle$ is nonempty. Then we can equivalently define ℓ with φ to be *minimally emblematic* for L if $\varphi(\mathcal{R})$ is an emblem of L of minimal cardinality. In other words, $\varphi(\mathcal{R})$ is a minimal set of minimizers of all possible L-inner risks.

For each $k \in \{3, \ldots, 15\}$, we showed by a computer search that for all $\mathbf{S} \in \mathcal{OP}_k$, there exists $p \in \Delta^k$ such that $\mathbf{S}$ is the *unique* minimizer of $\min_{\mathbf{T} \in \mathcal{OP}_k} \langle p, \ell(\mathbf{T}) \rangle$. A consequence of this is that

Proposition II.19. *For $k \in \{3, \ldots, 15\}$, $\operatorname{emb.card}(L^{WW}) = |\mathcal{OP}_k|$. In other words, the ordered partition loss is minimally emblematic for the WW-hinge loss.*

We conjecture this result holds for all $k \geq 3$.

2.5 The argmax link

Define $\gamma_\ell(p) := \arg\min_{r \in \mathcal{R}} \langle p, \ell(r) \rangle$ and $\Gamma_L(p) := \arg\min_{v \in \mathbb{R}^d} \langle p, L(v) \rangle$. For multiclass classification into k classes, most multiclass SVMs typically output a vector of scores $v \in \mathbb{R}^k$ which is converted to a class label by taking $\arg\max v$. In this section, we analyze the $\arg\max$ as a "link" function. Recall from section 2.1.3, $\arg\max$ is a set-valued function. Define

$$\Omega_L := \{p \in \Delta^k : |\arg\max p| = 1, \arg\max v = \arg\max p, \forall v \in \Gamma_L(p)\}.$$

When L is calibrated with respect to ℓ^{zo}, we have that $\Omega_L = \{p \in \Delta^k : |\arg\max p| = 1\}$. Hence, $\Delta^k \setminus \Omega_L$ has measure zero. For other L not necessarily calibrated with respect to ℓ^{zo}, it is desirable that Ω_L be as large as possible. Below, we will prove that $\Omega_{L^{CS}}$ is a proper subset of $\Omega_{L^{WW}}$.

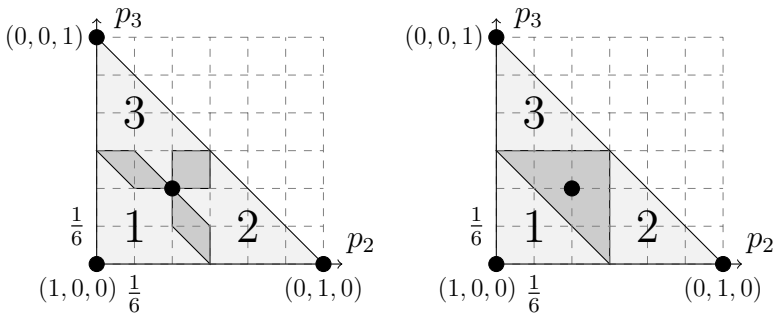

Figure 2.2: The gray triangle represents the probability simplex Δ^3, where $(p_1, p_2, p_3) \in \Delta^3$ is plotted as (p_2, p_3) in the plane. The light gray regions are $\Omega_{L^{WW}}$ (left) and $\Omega_{L^{CS}}$ (right). For the derivation, see Section 2.8.

Recall that $\mathcal{X}$ is a sample space and P is a distribution on $\mathcal{X} \times [k]$. For each $x \in \mathcal{X}$, define the *class conditional distribution* $\eta_P(x) \in \Delta^k$ by $[\eta_P(x)]_y = \Pr_{X,Y\sim P}(Y = y|X = x)$.

Proposition II.20. *Let P be a joint distribution on $\mathcal{X} \times [k]$ such that $\eta_P(x) \in \Omega_L$ for all x and $L : \mathbb{R}^d \to \mathbb{R}^k_+$ be a loss. Let $g^* : \mathcal{X} \to \mathbb{R}^k$ be such that $g^*(x) \in \Gamma_L(\eta_P(x))$ for all $x \in \mathcal{X}$. Then $\arg\max \circ g^*$ is Bayes optimal with respect to the 0-1 loss.*

Proof. By definition of Ω_L, we have $\arg\max \circ g^*(x) = \arg\max \eta_P(x)$ for all $x \in \mathcal{X}$. □

The following theorem asserts that for any $v \in \Gamma_{L^{WW}}(p)$, the $\arg\max v$ is contained in the top bucket S_1 for some $\mathbf{S} \in \gamma_{\ell^{\mathcal{OP}}}(p)$.

Theorem II.21. *Let $p \in \Delta^k$ be such that $\max p > \frac{1}{k}$ and $v \in \Gamma_{L^{WW}}(p)$. Then there exists $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma_{\ell^{\mathcal{OP}}}(p)$ such that $\arg\max v \subseteq S_1$.*

Below, we consider two conditions on $p \in \Delta^k$ such that for *all* $\mathbf{S} \in \gamma_{\ell^{\mathcal{OP}}}(p)$, the top bucket $S_1 = \arg\max p$. By theorem II.21, for such $p \in \Delta^k$, we can recover $\arg\max p$ from any $v \in \Gamma_{L^{WW}}(p)$. The first condition covers $p \in \Delta^k$ such that the top class has a majority:

Proposition II.22. *Let $p \in \Delta^k$ satisfy the "majority condition": $\max p > 1/2$. Then for all $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma_{\ell^{\mathcal{OP}}}(p)$, we have $|S_1| = 1$ and $S_1 = \arg\max p$.*

While proposition II.22 does not guarantee that $\gamma_{\ell^{\mathcal{OP}}}(p)$ is a singleton, all $\mathbf{S} \in \gamma_{\ell^{\mathcal{OP}}}(p)$ have the same top bucket. The second condition covers $p \in \Delta^k$ whose top class may not have a majority, yet $\arg\max p$ can still be recovered from any $v \in \Gamma_{L^{WW}}(v)$ by taking $\arg\max v$:

Proposition II.23. *Fix a number α such that $1 > \alpha > \frac{1}{k}$. Let $p \in \Delta^k$ satisfy the "symmetric label noise (SLN) condition": there exists $j^* \in [k]$ so that $p_{j^*} = \alpha$ and $p_j = \frac{1-\alpha}{k-1}$ for all $j \neq j^*$. Then $(\{j^*\}, [k] \setminus \{j^*\})$ is the unique element of $\gamma_{\ell^{\mathcal{OP}}}(p)$.*

In particular, when $\alpha < 1/2$, p violates the majority condition. Under SLN, we have $\arg\max p = \{j^*\}$ since $\alpha - \frac{1-\alpha}{k-1} = \frac{(k-1)\alpha-1+\alpha}{k-1} = \frac{k\alpha-1}{k-1} > \frac{1-1}{k-1} = 0$. In light of theorem II.21, we have

Corollary II.24. *If $p \in \Delta^k$ satisfies the majority or the SLN condition, then $p \in \Omega_{L^{WW}}$.*

Thus, in two common regimes where for all $x \in \mathcal{X}$ the class conditional $\eta_P(x)$ satisfies the SLN or the majority condition, the Bayes optimal ordered partition has a top bucket consisting of a single element. When this occurs, the argmax link recovers the most probable class, i.e., the unique element from the top bucket. This supports the observation by Doğan et al. [DGI16] that the WW-SVM performs well under the SLN condition, even with significant label noise. For the CS-hinge loss, it is known that $\Omega_{L^{CS}} = \{p \in \Delta^k : p \text{ satisfies the majority condition}\}$ [Liu07, Lemma 4]. In particular, $\Omega_{L^{CS}}$ is a proper subset of $\Omega_{L^{WW}}$. For $k = 3$, we show in fig. 2.2 the regions $\Omega_{L^{WW}}$ and $\Omega_{L^{CS}}$. Our finding provides theoretical support for the finding of [DGI16] that WW outperforms CS.

2.6 Conclusion and future work

We proved that the Weston-Watkins hinge loss is calibrated with respect to the ordered partition loss, which we argue is minimally emblematic for the WW-hinge

loss. Furthermore, we showed the advantage of WW-hinge loss over the Crammer-Singer hinge loss when the popular "argmax" link is used. An interesting direction is to apply the ordered partition loss to other multiclass learning problems such as partial label and multilabel learning.

2.7 Omitted proofs

2.7.1 Additional notations

We introduce notations in addition to those already defined previously in section 2.1.3.

- L always denotes the WW-hinge loss (definition II.4) and ℓ always denotes the ordered partition loss (definition II.6). So far, we sometimes works with generic losses L and ℓ. However, below, we focus exclusively on the WW-hinge and the ordered partition loss. The exception is the last section section 2.8.2, where the explicit names L^{WW} and L^{CS} are used.
- All vectors are column vectors unless stated otherwise.
- $\mathbb{R}_+$ and $\mathbb{Z}_+$ denotes the set of non-negative reals and integers, respectively.
- Define $\mathbb{R}^k_\uparrow = \{v \in \mathbb{R}^k : v_1 \leq v_2 \leq \cdots \leq v_k\}$. Likewise, define $\mathbb{R}^k_\downarrow$.
- For a positive integer n, we let $[n] := \{1, \ldots, n\}$. By convention, $[0] = \emptyset$.
- Let $\mathbf{1}^k \in \mathbb{R}^k$ denote the vector all ones.
- For a number $t \in \mathbb{R}$, let $[t]_+ = \max\{0, t\}$. For a vector v, we denote by $[v]_+$ the vector resulting from applying $[\cdot]_+$ entrywise to v. The *hinge loss* $h : \mathbb{R} \to \mathbb{R}_+$ is defined by $h(x) = [1 - x]_+$.
- For a vector $v \in \mathbb{R}^k$, we use $[v]_i$ to denote the i-th entry of v in conjunction with the usual notation v_i.

- Given a vector $v \in \mathbb{R}^k$, we define

$$\max v := \max_{i \in [k]} v_i \quad \text{and} \quad \arg\max v := \{i \in [k] : v_i = \max v\}$$

  Define $\min v$ and $\arg\min v$ likewise.

- Probability simplex

$$\Delta^k = \{p = (p_1, \ldots, p_k) \in \mathbb{R}^k_+ : p_1 + \cdots + p_k = 1\}$$

  and *non-increasing* probability simplex

$$\Delta^k_\downarrow = \{p \in \Delta^k : p_1 \geq p_2 \geq \cdots \geq p_k\} = \Delta^k \cap \mathbb{R}^k_\downarrow.$$

- For $p \in \Delta^k$, we write $Y \sim p$ to denote a discrete random variable $Y \in [k]$ whose probability mass function is p.

- For each $i, j \in [k]$, $\sigma_{(i,j)} \in \mathbb{R}^{k \times k}$ is the permutation matrix that switches the i-th and j-th index. By convention, if $i = j$, then $\sigma_{(i,j)}$ is the identity. Also, for brevity, define $\sigma_i = \sigma_{(1,i)}$.

- According to the definition above, $\sigma_{(i,j)}$ acts on $\mathbb{R}^k$. However, we abuse notation and allow $\sigma_{(i,j)}$ to act on $[k]$ in the obvious way. In such cases, we write $\sigma_{(i,j)}(\ell)$ for $\ell \in [k]$.

2.7.2 Main results

Lemma II.25. *For all $v \in \mathbb{R}^k$ and $c \in \mathbb{R}$, we have $L(v) = L(v + c\mathbf{1}^k)$.*

Proof. For all $y \in [k]$, we have that

$$[L(v + c\mathbf{1})]_y = \sum_{i \in [k] \,:\, i \neq y} h(v_y + c - (v_i - c)) = \sum_{i \in [k] \,:\, i \neq y} h(v_y - v_i) = [L(v)]_y.$$

□

Lemma II.26. *For all $j \in [k]$, we have $L(\sigma_j v) = \sigma_j L(v)$.*

Proof. If $j = 1$, then the result is trivial. Hence, let $j > 1$. We prove

$$[L(\sigma_j v)]_y = [L(v)]_{\sigma_j(y)}$$

for the following three cases: $y \notin \{1, j\}$, $y = 1$ and $y = j$. Before we go through the cases, note that

$$[L(\sigma_j v)]_y = \sum_{i \in [k] : i \neq y} h([\sigma_j v]_y - [\sigma_j v]_i) = \sum_{i \in [k] : i \neq y} h(v_{\sigma_j(y)} - v_{\sigma_j(i)}).$$

Now, for the first case, suppose that $y \notin \{1, j\}$. Then $\sigma_j(y) = y$ and so

$$\begin{aligned}
[L(\sigma_j v)]_y &= \sum_{i \in [k] : i \neq y} h(v_y - v_{\sigma_j(i)}) \\
&= h(v_y - v_{\sigma_j(1)}) + h(v_y - v_{\sigma_j(j)}) + \sum_{i \in [k] \,:\, i \notin \{1, j, y\}} h(v_y - v_{\sigma_j(i)}) \\
&= h(v_y - v_j) + h(v_y - v_1) + \sum_{i \in [k] \,:\, i \notin \{1, j, y\}} h(v_y - v_i) \\
&= \sum_{i \in [k] \,:\, i \notin \{y\}} h(v_y - v_i) \\
&= [L(v)]_y = [L(v)]_{\sigma_j(y)}.
\end{aligned}$$

Next, suppose that $y = 1$. Thus, we have $\sigma_j(y) = \sigma_j(1) = j$. So

$$\begin{aligned}
[L(\sigma_j v)]_y = [L(\sigma_j v)]_1 &= \sum_{i \in [k] : i \neq 1} h(v_j - v_{\sigma_j(i)}) \\
&= \sum_{i \in [k] : i \neq j} h(v_j - v_i) \\
&= [L(v)]_j = [L(v)]_{\sigma_j(y)}.
\end{aligned}$$

Finally, if $y = j$, $\sigma_j(y) = 1$

$$\begin{aligned}[L(\sigma_j v)]_y = [L(\sigma_j v)]_j &= \sum_{i\in[k]:i\neq j} h(v_1 - v_{\sigma_j(i)}) \\ &= \sum_{i\in[k]:i\neq 1} h(v_j - v_i) \\ &= [L(v)]_1 = [L(v)]_{\sigma_j(j)} = [L(v)]_{\sigma_j(y)}.\end{aligned}$$

□

Lemma II.27. *Let $i, j \in \{2, \ldots, k\}$ be distinct. Then $\sigma_i\sigma_j\sigma_i = \sigma_{(i,j)}$.*

Proof. This is simply an exhaustive case-by-case proof over all inputs $y \in [k]$. First, let $y = 1$. Then $\sigma_{(i,j)}(1) = 1$ since $1 \notin \{i, j\}$. On the other hand $\sigma_i\sigma_j\sigma_i(1) = \sigma_i\sigma_j(i) = \sigma_i(i) = 1$. Now, let $y \in \{2, \ldots, k\}$. If $y \notin \{i, j\}$, then $\sigma_{(i,j)}(y) = y$ and $\sigma_i\sigma_j\sigma_i(y) = \sigma_i\sigma_j(y) = \sigma_i(y) = y$. If $y = i$, then $\sigma_{(i,j)}(i) = j$ and $\sigma_i\sigma_j\sigma_i(i) = \sigma_i\sigma_j(1) = \sigma_i(j) = j$. If $y = j$, then $\sigma_{(i,j)}(j) = i$ and $\sigma_i\sigma_j\sigma_i(j) = \sigma_i\sigma_j(j) = \sigma_i(1) = i$. □

Corollary II.28. *Every $\sigma \in \mathfrak{S}_k$ can be written as a product $\sigma = \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_l}$.*

Proof. We prove the equivalent statement that the set $\mathcal{S} := \{\sigma_i : i \in \{2, \ldots, k\}\}$ generates the group $\mathfrak{S}_k$. A standard result in group theory states that the set of transpositions $\mathcal{T}$ generates $\mathfrak{S}_k$. By lemma II.27, transpositions between labels in $\{2, \ldots, k\}$ can be generated by $\mathcal{S}$. Furthermore, $\sigma_i = \sigma_{(1,i)}$ by definition, so transposition between 1 and elements of $\{2, \ldots, k\}$ can be generated by $\mathcal{S}$ as well. Hence, all of $\mathcal{T}$ can be generated by $\mathcal{S}$. □

Corollary II.29. *For all $v \in \mathbb{R}^k$ and $\sigma \in \mathfrak{S}_k$, we have*

$$L(\sigma v) = \sigma L(v).$$

Proof. By corollary II.28, we may write $\sigma = \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_m}$. Hence,

$$\begin{aligned} L(\sigma v) &= L(\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_m}v) && (2.6)\\ &= \sigma_{i_1}L(\sigma_{i_2}\cdots\sigma_{i_m}v) && (2.7)\\ &\vdots \\ &= \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_m}L(v) && (2.8)\\ &= \sigma L(v), && (2.9) \end{aligned}$$

where for eq. (2.7) to eq. (2.8) we used lemma II.26. □

Lemma II.30. *Let $v \in \mathbb{R}^k$ and $j, j' \in [k]$ be distinct such that $v_j \geq v_{j'}$. Then $[L(v)]_j \leq [L(v)]_{j'}$. Furthermore, if $v_j > v_{j'}$, then $[L(v)]_j < [L(v)]_{j'}$.*

Proof. We have

$$\begin{aligned} &[L(v)]_j - [L(v)]_{j'} \\ &= \sum_{i\in[k]:i\neq j} h(v_j - v_i) \\ &\quad - \sum_{i\in[k]:i\neq j'} h(v_{j'} - v_i) \\ &= h(v_j - v_{j'}) + \sum_{i\in[k]:i\notin\{j,j'\}} h(v_j - v_i) \\ &\quad - h(v_{j'} - v_j) - \sum_{i\in[k]:i\notin\{j,j'\}} h(v_{j'} - v_i) \\ &= h(v_j - v_{j'}) - h(v_{j'} - v_j) \\ &\quad + \sum_{i\in[k]:i\notin\{j,j'\}} h(v_j - v_i) - h(v_{j'} - v_i). \end{aligned}$$

Since and h is monotonically non-increasing, we have

$$v_j - v_{j'} \geq 0 \geq v_{j'} - v_j \implies h(v_j - v_{j'}) - h(v_{j'} - v_j) \leq 0 \qquad (2.10)$$

For the same reason, we have $h(v_j - v_i) - h(v_{j'} - v_i) \leq 0$. Putting it all together, we have $[L(v)]_j - [L(v)]_{j'} \leq 0$, as desired.

For the "furthermore" part, note that under the assumption $v_j > v_{j'}$, all inequalities in eq. (2.10) becomes strict. □

For reasons that will become clear later, we define for each $n \in [k-1]$

$$\underline{L}^n(p) := \inf_{v \in \mathbb{R}^k : |\arg\max v| \geq n} \langle p, L(v) \rangle. \tag{2.11}$$

Since $\arg\max v$ is always nonempty, the condition that $|\arg\max v| \geq 1$ is always true. Thus, we have $\underline{L}^1 = \underline{L}$.

Lemma II.31. *For all $n \in [k-1]$, $p \in \Delta^k$ and $\sigma \in \mathfrak{S}_k$, we have $\underline{L}^n(p) = \underline{L}^n(\sigma p)$.*

Proof. Define $\mathcal{R}^{k,n} := \{v \in \mathbb{R}^k : |\arg\max v| \geq n\}$. Since $|\arg\max v| = |\arg\max \sigma v|$, we have $\sigma \mathcal{R}^{k,n} = \mathcal{R}^{k,n}$. Introducing the change of variables $u = \sigma v$, we have

$$\begin{aligned}
\underline{L}^n(p) &= \inf_{v \in \mathcal{R}^{k,n}} \langle p, L(v) \rangle \\
&= \inf_{\sigma' u \in \mathcal{R}^{k,n}} \langle p, L(\sigma' u) \rangle \quad \because \text{Definition of } u \\
&= \inf_{u \in \sigma \mathcal{R}^{k,n}} \langle p, L(\sigma' u) \rangle \quad \because \sigma^{-1} = \sigma' \\
&= \inf_{u \in \mathcal{R}^{k,n}} \langle p, L(\sigma' u) \rangle \quad \because \sigma \mathcal{R}^{k,n} = \mathcal{R}^{k,n} \\
&= \inf_{u \in \mathcal{R}^{k,n}} \langle p, \sigma' L(u) \rangle \quad \because \text{corollary II.29} \\
&= \inf_{u \in \mathcal{R}^{k,n}} \langle \sigma p, L(u) \rangle \\
&= \underline{L}^n(\sigma p).
\end{aligned}$$

□

Lemma II.32. *Let $p \in \mathbb{R}^k_\downarrow$, $q \in \mathbb{R}^k$ be arbitrary and $\sigma \in \mathfrak{S}_k$ be such that $\sigma q \in \mathbb{R}^k_\uparrow$. Then $\langle p, q \rangle \geq \langle p, \sigma q \rangle$.*

Proof. Consider the "bubble sort" algorithm applied to q:

1. Initialize $q^{(0)} = q$, $t \leftarrow 0$
2. While there exists $i \in [k-1]$ such that $q_i^{(t)} > q_{i+1}^{(t)}$, do
   (a) $q^{(t+1)} \leftarrow \sigma_{(i,i+1)} q^{(t)}$
   (b) $t \leftarrow t+1$
3. Output monotone non-decreasing vector $q^{(t)}$

We claim that at every step, we have $\langle p, q^{(t)} \rangle \geq \langle p, q^{(t+1)} \rangle$. Let $a = q_i^{(t)}$ and $b = q_{i+1}^{(t)}$ as in step 2 above. Let $c = p_i$ and $d = p_{i+1}$. Hence, we have $a > b$ and $c \geq d$. Observe that

$$\langle p, q^{(t)} \rangle - \langle p, q^{(t+1)} \rangle = ac + bd - (ad + bc) = (a-b)(c-d) \geq 0$$

which proves the claim. Thus, we have

$$\langle p, q \rangle = \langle p, q^{(0)} \rangle \geq \langle p, q^{(1)} \rangle \geq \cdots \geq \langle p, q^{(t)} \rangle.$$

By construction, there exists $\tau \in \mathfrak{S}_k$ such that $\tau q = q^{(t)}$. We must have $\tau q = \sigma q$ since both vectors are monotone non-increasing, although τ may not equal σ. □

Define the matrix $T \in \mathbb{R}^{k \times k}$

$$T_{ij} = \begin{cases} 1 & i \geq j \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

Also, define $D \in \mathbb{R}^{k \times k}$

$$D_{ij} = \begin{cases} 1 & : i = j \\ -1 & : i = j+1 \\ 0 & : \text{otherwise.} \end{cases}$$

In other words, D is the matrix with 1s on the main diagonal, -1s on the subdiagonal below the main diagonal, and 0 everywhere else. We have

$$[Dv]_i = \begin{cases} v_1 & : i = 1 \\ v_i - v_{i-1} & : i > 1. \end{cases}$$

Lemma II.33. $D^{-1} = T$.

Proof. Using Gaussian elimination for inverting a matrix, it is easy to see that $D'T'$ is the identity. □

Definition II.34. Define the following sets:

$$\mathcal{M} = \{v \in \mathbb{R}^k : v_1 = 0 \text{ and } 0 \leq v_i - v_{i+1}, \forall [k-1]\},$$

$$\mathcal{C} = \{v \in \mathbb{R}^k : v_1 = 0, v_k \leq -1, \text{ and } 0 \leq v_i - v_{i+1} \leq 1, \forall [k-1]\},$$

$\mathcal{M}_\mathbb{Z} = \mathcal{M} \cap \mathbb{Z}^k$ and $\mathcal{C}_\mathbb{Z} = \mathcal{C} \cap \mathbb{Z}^k$.

Lemma II.35. *We have the following equality of sets:*

$$\mathcal{M}_\mathbb{Z} = \{-Tc : c \in \mathbb{Z}_+^k, c_1 = 0\}$$
$$\mathcal{C}_\mathbb{Z} = \{-Ts : s \in \{0,1\}^k, s_1 = 0, \text{ and } \exists i \in \{2, \ldots, k\} : s_i = 1\}$$

Proof. If $v \in \mathcal{C}_\mathbb{Z}$, then we have $v_i \in \mathbb{Z}_+$ and $v_i - v_{i+1} \in [0,1]$. These two conditions together implies that $v_i - v_{i+1} \in \{0,1\}$ for all $i \in [k-1]$. Hence, $-Dv \in \{0,1\}^{k-1}$ with $[Dv]_1 = -v_1 = 0$. Let $-Dv = s$. Then lemma II.33 implies that $-Ts = TDv = v$. By construction, $s_1 = 0$. Furthermore, if $s_i = 0$ for all $i \in [k]$, then we would have $v = 0$ as well, which contradicts the fact that $v_k \leq -1$. Hence, there must exists $i \in \{2, \ldots, k\}$ such that $s_i = 1$. Clearly, all $v \in \mathcal{C}_\mathbb{Z}$ arise this way. The statement about $\mathcal{M}_\mathbb{Z}$ is similar. □

Lemma II.36. *Let $c \in \mathbb{Z}_+^k$ and define $s \in \{0,1\}^k$ entrywise where for each $i \in [k]$, $s_i = \mathbb{I}\{c_i \geq 1\}$. Then we have $[L(-Tc)]_y \geq [L(-Ts)]_y$ for all $y \in [k]$.*

Proof. By definition, we have

$$\begin{aligned}
&[L(-Tc)]_y - [L(-Ts)]_y \\
&= \sum_{i \in [k]: i \neq y} h([-Tc]_y - [-Tc]_i) - h([-Ts]_y - [-Ts]_i) \\
&= \sum_{i \in [k]: i \neq y} h([Tc]_i - [Tc]_y) - h([Ts]_i - [Ts]_y)
\end{aligned}$$

It suffices to show that $h([Tc]_i - [Tc]_y) - h([Ts]_i - [Ts]_y) \geq 0$ for all $i \in [k]$ such that $i \neq y$.

First, consider when $i > y$. We have

$$[Tc]_i - [Tc]_y = \sum_{j=y+1}^{i} c_j$$

Similarly, we have

$$[Ts]_i - [Ts]_y = \sum_{j=y+1}^{i} s_j = \sum_{j=y+1}^{i} \mathbb{I}\{c_j \geq 1\}.$$

From this, we see that

$$\begin{aligned}
[Ts]_i - [Ts]_y \geq 1 &\implies [Tc]_i - [Tc]_y \geq 1 \\
[Ts]_i - [Ts]_y = 0 &\implies [Tc]_i - [Tc]_y = 0.
\end{aligned}$$

For $i > y$, we have $h([Ts]_i - [Ts]_y) = h([Tc]_i - [Tc]_y)$.

Next, let $i < y$. We have

$$[Tc]_i - [Tc]_y = \sum_{j=i+1}^{y} -c_j.$$

Similarly, we have

$$[Ts]_i - [Ts]_y = \sum_{j=i+1}^{y} -\mathbb{I}\{c_j \geq 1\}.$$

Since $c_j \geq \mathbb{I}\{c_j \geq 1\}$, we have $[Ts]_i - [Ts]_y \geq [Tc]_i - [Tc]_y$ which implies that $h([Ts]_i - [Ts]_y) \leq h([Tc]_i - [Tc]_y)$. □

Definition II.37. Let $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$. Define the linear map $\pi : \mathbb{R}^k \to \mathbb{R}^{k-1}$

$$\pi(v) = (v_1 - v_2, v_1 - v_3, \ldots, v_1 - v_k).$$

We observe that for each $i \in [k-1]$, we have

$$[\pi v]_i = v_1 - v_{i+1}.$$

Definition II.38. Given $k \geq 2$, define the following $(k-1)$-by-$(k-1)$ square matrices $\rho_1, \rho_2, \ldots, \rho_k \in \mathbb{R}^{(k-1)\times(k-1)}$:

1. ρ_1 is the identity,
2. Let $z = (z_1, \ldots, z_{k-1}) \in \mathbb{R}^{k-1}$ be a vector. For each $i > 1$, define $\rho_i(z) \in \mathbb{R}^{k-1}$ entrywise for each $j \in [k-1]$ by

$$[\rho_i(z)]_j = \begin{cases} z_j - z_{i-1} & : j \neq i-1 \\ -z_{i-1} & : j = i-1. \end{cases} \tag{2.13}$$

Lemma II.39 (Commuting relations)**.** *For all $i \in [k]$, we have $\pi\sigma_i = \rho_i\pi$.*

Proof. If $i = 1$, then σ_i and ρ_i are both identity matrices and there is nothing to show. Otherwise, suppose that $i > 1$. Consider $v \in \mathbb{R}^k$. We first calculate $\pi\sigma_i v$. For

each $j \in [k-1]$, we have

$$[\pi\sigma_i v]_j = [\sigma_i v]_1 - [\sigma_i v]_{j+1} = v_i - v_{\sigma_i(j+1)} = \begin{cases} v_i - v_{j+1} & : i \neq j+1 \\ v_i - v_1 & : i = j+1. \end{cases} \tag{2.14}$$

Now, we compute $\rho_i \pi v$. Likewise, for each $j \in [k-1]$,

$$[\rho_i \pi v]_j = \begin{cases} [\pi v]_j - [\pi v]_{i-1} & : j \neq i-1 \\ -[\pi v]_{i-1} & : j = i-1. \end{cases}$$

Consider the two cases above separately: for $j \neq i-1$, we have

$$[\pi v]_j - [\pi v]_{i-1} = (v_1 - v_{j+1}) - (v_1 - v_i) = v_i - v_{j+1}.$$

On the other hand, for $i = j+1$, we have

$$-[\pi v]_{i-1} = -(v_1 - v_i) = v_i - v_1.$$

Thus, we have $[\pi\sigma_i v]_j = [\rho_i \pi v]_j$ for all j which implies that $\pi\sigma_i v = \rho_i \pi v$. Since v was arbitrary, we have $\pi\sigma_i = \rho_i \pi$. □

Definition II.40. The *reduced WW hinge function* $H : \mathbb{R}^{k-1} \to \mathbb{R}_{\geq 0}$ is defined as

$$H(z) = \sum_{i=1}^{k-1} h(z_i).$$

Definition II.41. For $z \in \mathbb{R}^{k-1}$, the *reduced WW hinge loss* $\ell(z) \in \mathbb{R}^k$ is defined entrywise for each $y \in [k]$ by

$$[\ell(z)]_y = H(\rho_y z).$$

Lemma II.42. *For all $v \in \mathbb{R}^k$, we have $\ell(\pi v) = L(v)$.*

Proof. We first check for all $y \in [k]$ that

$$\sum_{i \in [k] \,:\, i \neq y} h(v_y - v_i) = H(\pi \sigma_i v). \tag{2.15}$$

Unpacking the definition, we have $H(\pi\sigma_y v) = \sum_{i \in [k-1]} h([\pi\sigma_y v]_i)$. Now, if $y = 1$, then $[\pi v]_i = v_1 - v_{i+1}$ for all $i \in [k-1]$. Hence, eq. (2.15) holds. If $y > 1$. Then eq. (2.15) follows from the expression for $[\pi\sigma_y v]_i$ computed in eq. (4.13). Thus, we have proven eq. (2.15) for all $y \in [k]$. To conclude, we have

$$\begin{aligned}
[L(v)]_y &= \sum_{i \in [k] \,:\, i \neq y} h(v_y - v_i) && (2.16)\\
&= H(\pi \sigma_y v) && (2.17)\\
&= H(\rho_y \pi v) && (2.18)\\
&= [\ell(\pi v)]_y && (2.19)
\end{aligned}$$

where in eq. (2.18), we applied lemma IV.34. □

Lemma II.43. *Let $n \in [k-1]$. If $p \in \Delta^k_\downarrow$, then*

$$\underline{L}^n(p) = \min_{v \in \mathcal{C}_{\mathbb{Z}} \,:\, v_n = 0} \langle p, L(v) \rangle.$$

Proof. Define

$$\mathcal{N}^n = \{v \in \mathbb{R}^k : v_1 = \cdots = v_n = 0,\, v_i \leq 0,\, \forall i \in [k]\}.$$

We first claim that

$$\underline{L}^n(p) = \inf_{v \in \mathcal{N}^n} \langle p, L(v) \rangle. \tag{2.20}$$

Since $\mathcal{N}^n \subseteq \{v \in \mathbb{R}^k : |\arg\max v| \geq n\}$, the "$\leq$" part of eq. (2.20) is obvious. For the "$\geq$" part, let $v \in \mathbb{R}^k$ be such that $|\arg\max v| \geq n$. Then $w = v - \mathbf{1}^k \max_{i \in [k]} v_i$

is such that $w \in \mathcal{N}^n$. Furthermore, by lemma II.25, we have $\langle p, L(v)\rangle = \langle p, L(w)\rangle$. Thus, we have proven the claim.

Next, observe that if $v \in \mathcal{N}^n$, then

$$[\pi v]_i = v_1 - v_{i+1} \begin{cases} = 0 & : i \leq n-1 \\ \geq 0 & : i \geq n. \end{cases}$$

Therefore, we have

$$\pi(\mathcal{N}^n) = \{z \in \mathbb{R}^{k-1} : z \geq 0,\ z_i = 0, \forall i \in [n-1]\}$$

where $[0] = \emptyset$. Introducing the change of variable $z = \pi v \in \mathbb{R}^{k-1}$, we have

$$\begin{aligned} \inf_{v \in \mathcal{N}^n} \langle p, L(v)\rangle &= \inf_{v \in \mathcal{N}^n} \langle p, \ell(\pi v)\rangle \quad \because \text{lemma II.42} && (2.21) \\ &= \inf_{z \in \pi(\mathcal{N})} \langle p, \ell(z)\rangle && (2.22) \\ &= \inf_{\substack{z \in \mathbb{R}^{k-1} : z \geq 0 \\ z_i = 0,\, \forall i \in [n-1]}} \langle p, \ell(z)\rangle && (2.23) \end{aligned}$$

Below, let $\mathbf{1} := \mathbf{1}^{k-1}$. Unwinding the definition, we have

$$\langle p, \ell(z)\rangle = \sum_{i \in [k]} p_i H(\rho_i z) = \sum_{i \in [k]} p_i \mathbf{1}' \left[\mathbf{1} - \rho_i z\right]_+ .$$

Using slack variables $\xi_i \geq [\mathbf{1} - \rho_i z]_+$, we can rewrite eq. (2.23) as the following linear

program:

$$\min_{z\in\mathbb{R}^{k-1}} \min_{(\xi_1,\ldots,\xi_k):\,\xi_i\in\mathbb{R}^{k-1}} \quad \sum_i p_i \mathbf{1}'\xi_i \tag{2.24}$$

$$s.t. \quad \xi_i \geq \mathbf{1} - \rho_i z \tag{2.25}$$

$$\xi_i \geq 0, \quad \forall i \in [k] \tag{2.26}$$

$$z \geq 0, \tag{2.27}$$

$$z_i = 0,\ \forall i \in [n-1]. \tag{2.28}$$

By Bertsimas et al. [BT97, Corollary 3.2], for a linear programming minimization problem over a nonempty polyhedron, one of the following must be true: 1) the optimal cost is $-\infty$ or 2) a feasible minimum exists. Since eq. (2.24) is nonnegative and the feasible region is nonempty, a feasible minimum exists. Let

$$R = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} \in \mathbb{R}^{k(k-1)\times(k-1)}, \quad X = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_k \end{bmatrix} \in \mathbb{R}^{k(k-1)}, \quad p \otimes \mathbf{1} = \begin{bmatrix} p_1\mathbf{1} \\ p_2\mathbf{1} \\ \vdots \\ p_k\mathbf{1} \end{bmatrix} \in \mathbb{R}^{k(k-1)}.$$

We claim that

$$\underline{L}^n(p) = \min_{z\in\mathbb{R}_+^{k-1}:z_i=0\,\forall i\in[n-1]} \langle p, \ell(z)\rangle. \tag{2.29}$$

We first consider the case when $n = 1$ where we have $\underline{L}^1 = \underline{L}$. In this case, the linear

program eq. (2.24) can be rewritten as

$$\begin{aligned} \underline{L}(p) = \min_{z \in \mathbb{R}^{k-1}} \min_{X \in \mathbb{R}^{k(k-1)}} \quad & (p \otimes \mathbf{1})' X \\ s.t. \quad & X + Rz \geq \mathbf{1} \\ & X \geq 0 \\ & z \geq 0. \end{aligned}$$

For a positive integer m, let I_m denote the $m \times m$ identity matrix. Thus,

$$\min_{z \in \mathbb{R}^{k-1},\, X \in \mathbb{R}^{k(k-1)}} \quad (p \otimes \mathbf{1})' X \tag{2.30}$$

$$s.t. \quad \underbrace{\begin{bmatrix} R & I_{k(k-1)} \\ I_{k-1} & 0 \\ 0 & I_{k(k-1)} \end{bmatrix}}_{=:A} \begin{bmatrix} z \\ X \end{bmatrix} \geq \begin{bmatrix} \mathbf{1} \\ 0 \\ 0 \end{bmatrix}. \tag{2.31}$$

We prove that A is totally unimodular (TUM). The matrix R has the property that every row has at most one 1 and at most one -1, with all other entries being zeros. Hence, R is TUM by the Hoffman's sufficient condition Lawler [Law01]. Thus, (horizontally) concatenating R with an identity matrix, i.e., $R_0 := \begin{bmatrix} R & I_{k(k-1)} \end{bmatrix}$ results in another TUM matrix R_0. Finally, A is the (vertical) concatenation of R_0 with another identity matrix, i.e., $A = \begin{bmatrix} R_0 \\ I_{k(k-1)} \end{bmatrix}$. Hence, A is also TUM.

By a well-known result in combinatorial optimization Lawler [Law01], there exists an integral solution (X^*, z^*) to eq. (2.30). In particular, $z^* \in \mathbb{Z}_+^{k-1}$. Thus, we have proven that

$$\underline{L}(p) = \langle p, \ell(z^*) \rangle = \min_{z \in \mathbb{Z}_+^{k-1}} \langle p, \ell(z) \rangle.$$

This proves eq. (2.29) for the case when $n = 1$. For $n > 1$, we define the matrix

$J \in \mathbb{R}^{(n-1)\times(k-1)}$ to be the first $n-1$ rows of the $(k-1)$-by-$(k-1)$ identity matrix. In other words, for $i \in [n-1]$ and $j \in [k-1]$,

$$J_{ij} = \begin{cases} 1 & : i = j \\ 0 & : i \neq j \end{cases}.$$

Thus, we have

$$\begin{aligned} \underline{L}^n(p) = \min_{z\in\mathbb{R}^{k-1},\, X\in\mathbb{R}^{k(k-1)}} \quad & (p \otimes \mathbf{1})' X \\ s.t. \quad & \underbrace{\begin{bmatrix} R & I_{k(k-1)} \\ I_{k-1} & 0 \\ 0 & I_{k(k-1)} \\ -J & 0 \end{bmatrix}}_{=:B} \begin{bmatrix} z \\ X \end{bmatrix} \geq \begin{bmatrix} \mathbf{1} \\ 0 \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

The matrix B is formed by duplicating rows of A and multiplying the duplicated row by -1. Thus, B is also TUM. This proves eq. (2.29).

Below, let z^* be a solution to eq. (2.29). Define $v^* = \begin{pmatrix} 0 \\ -z^* \end{pmatrix}$. Furthermore, $\pi(v^*) = z^*$ and so

$$\begin{aligned} \underline{L}^n(p) &= \langle p, \ell(z^*) \rangle \\ &= \langle p, \ell(\pi(v^*)) \rangle \\ &= \langle p, L(v^*) \rangle. \end{aligned}$$

Pick $\sigma \in \mathfrak{S}_k$ such that $\sigma v^* \in \mathbb{R}^k_\downarrow$. First we note that $L(\sigma v^*) \in \mathbb{R}^k_\uparrow$ by lemma II.30. Next, by corollary II.29, $L(\sigma v^*) = \sigma L(v^*)$. Hence, by lemma II.32

$$\langle p, L(v^*) \rangle \geq \langle p, \sigma L(v^*) \rangle = \langle p, L(\sigma v^*) \rangle$$

which implies that σv^* is optimal. Also, we observe that $\sigma v^* \in \mathcal{M}_{\mathbb{Z}}$. By lemma II.35, we can write $\sigma v^* = -Tc$ for some $c \in \mathbb{Z}_+^k$. Note that since $z_1^* = \cdots = z_{n-1}^* = 0$, the vector v^* has at least n entries equal to 0. Since $v^* \leq 0$, we must have that $v_1 = \cdots = v_n^* = 0$. Thus, $c_1 = \cdots c_n = 0$ as well. Let $s \in \{0,1\}^k$ be as defined in lemma II.36. Then we have

$$\underline{L}^n(p) \geq \langle p, L(\sigma v^*)\rangle = \langle p, L(-Tc)\rangle \geq \langle p, L(-Ts)\rangle.$$

Hence, we have $\underline{L}(p) = \langle p, L(-Ts)\rangle$. Since $s_i = \mathbb{I}\{c_i \geq 1\}$, we have $s_1 = \cdots = s_n = 0$ which implies that $[-Ts]_1 = \cdots = [-Ts]_n = 0$. Consider the case when there exists some $i \in \{n+1, \ldots, k\}$ such that $s_i = 1$, then we have $-Ts \in \mathcal{C}_{\mathbb{Z}}$ which completes the proof of lemma II.43. Now, consider the case where there does not exists such i. Then we must have $s = 0$ and also $-Ts = 0$. Therefore, we have $\underline{L}^n(p) = \langle p, L(0)\rangle$. Define $\tilde{v} \in \mathbb{R}^k$ entrywise by

$$[\tilde{v}]_i = \begin{cases} 0 & : i \neq k \\ -1 & : i = k \end{cases}$$

Noting that $k \in \arg\min_{i\in[k]} p_i$ by the assumption that $p \in \Delta_{\downarrow}^k$. By lemma II.44 below, we get that $\langle p, L(\tilde{v})\rangle \leq \langle p, L(0)\rangle$ which implies that $\langle p, L(\tilde{v})\rangle = \underline{L}^n(v)$. Clearly, $\tilde{v} \in \mathcal{C}_{\mathbb{Z}}$ and $\tilde{v}_n = 0$, which implies that $\tilde{v}$ is feasible for the optimization in lemma II.43. □

Lemma II.44. *Let $p \in \Delta^k$ and $i^* \in \arg\min_{i\in[k]} p_i$. Consider the vector $\tilde{v} \in \mathbb{R}^k$ defined by*

$$[\tilde{v}]_i = \begin{cases} 0 & : i \neq i^* \\ -1 & : i = i^* \end{cases}$$

Then

1. *$p_{i^*} \leq \frac{1}{k}$*
2. *$p_i = \frac{1}{k}$ for all i if and only if $p_{i^*} = \frac{1}{k}$*

3. $\langle p, L(0)\rangle \geq \langle p, L(\tilde{v})\rangle$ *with equality if and only if* $p_{i^*} = \frac{1}{k}$.

Proof. If $p_{i^*} > \frac{1}{k}$, then we would have $\sum_i p_i \geq k p_{i^*} > 1$, a contradiction. This proves that $p_{i^*} \leq \frac{1}{k}$. For the second item, the "only if" direction is obvious. For the "if" direction, note that if $p_i > \frac{1}{k}$ for any i, then we again obtain $\sum_i p_i > 1$, a contradiction. For the third item, first observe that

$$[L(0)]_i = \sum_{j\in[k]:j\neq i} h(0) = k - 1.$$

Thus, $L(0) = (k-1)\mathbf{1}^k$ and $\langle p, L(0)\rangle = k - 1$. Next, we only $L(\tilde{v})$. For $i \neq i^*$, we have

$$[L(\tilde{v})]_i = \sum_{j\in[k]:j\neq i} h(\tilde{v}_i - \tilde{v}_j) = h(1) + \sum_{j\in[k]:j\neq i, j\neq i^*} h(0) = k - 2.$$

When $i = i^*$, we have

$$[L(\tilde{v})]_{i^*} = \sum_{j\in[k]:j\neq i^*} h(\tilde{v}_{i^*} - \tilde{v}_j) = \sum_{j\in[k]:j\neq i^*} h(-1) = 2(k-1) = k - 2 + k.$$

From this, we deduce that

$$\langle p, L(\tilde{v})\rangle = k - 2 + k p_{i^*}.$$

Therefore, we have $p_{i^*} \leq \frac{1}{k}$ and so

$$\langle p, L(\tilde{v})\rangle = k - 2 + k p_{i^*} \leq k - 2 + 1 = k - 1 = \langle p, L(0)\rangle.$$

Note if if $p_{i^*} < \frac{1}{k}$, then the inequality above is strict. □

2.7.2.1 Proof of theorem II.14

Proof of theorem II.14. Recall that $\underline{L}(p) = \min_{v \in \mathbb{R}^k} \langle p, L(v) \rangle$. Since $\mathbb{R}^k \supseteq \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}$, we immediately have $\underline{L}(p) \leq \min_{v \in \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}} \langle p, L(v) \rangle$. Below, we focus on the other inequality.

Pick $\sigma \in \mathfrak{S}_k$ such that $\sigma p \in \Delta^k_{\downarrow}$. By lemma II.43 where $n = 1$, we have

$$\underline{L}(\sigma p) = \min_{v \in \mathcal{C}_{\mathbb{Z}}} \langle \sigma p, L(v) \rangle.$$

Now, by corollary II.29, we have

$$\langle \sigma p, L(v) \rangle = \langle p, \sigma' L(v) \rangle = \langle p, L(\sigma' v) \rangle.$$

Thus,

$$\begin{aligned}
\underline{L}(p) &= \underline{L}(\sigma p) \quad \because \text{lemma II.31} \\
&= \min_{v \in \mathcal{C}_{\mathbb{Z}}} \langle p, L(\sigma' v) \rangle \\
&= \min_{v \in \sigma' \mathcal{C}_{\mathbb{Z}}} \langle p, L(v) \rangle \quad \because \text{change of variables} \\
&\geq \min_{v \in \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}} \langle p, L(v) \rangle
\end{aligned}$$

where for the last equality, we used the fact that $\sigma' \mathcal{C}_{\mathbb{Z}} \subseteq \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}$. □

Lemma II.45. *Let $s \in \{0,1\}^k$ be such that $s_1 = 0$. Then*

$$[DL(-Ts)]_y = \begin{cases} \min\{i \in [k] : s_i = 1\} - 2 & : y = 1 \\ \min\{i \in [k] : s_i = 1,\, i > y\} - 1 & : s_y = 1,\, y > 1 \\ 0 & : s_y = 0,\, y > 1 \end{cases} \tag{2.32}$$

Proof. By the definition of T, we have

$$[Ts]_j = \sum_{i=1}^{j} s_i. \tag{2.33}$$

First, consider the case when $y = 1$. Then by eq. (2.33) we have $[-Ts]_1 = 0$. Furthermore,

$$\begin{aligned}
[DL(-Ts)]_1 &= [L(-Ts)]_1 \\
&= \sum_{i\in[k]:i\neq 1} h([-Ts]_1 - [-Ts]_i) \\
&= \sum_{i\in[k]:i\neq 1} h([Ts]_i)
\end{aligned}$$

Note that by eq. (2.33), we have $[Ts]_i \geq 1$ if $i \geq \min\{j : s_j = 1\}$ and $[Ts]_i = 0$ otherwise. Hence, we get

$$\begin{aligned}
[DL(-Ts)]_1 &= \sum_{i\in[k]:1<i<\min\{j:s_j=1\}} h([Ts]_i) \\
&= \sum_{i\in[k]:1<i<\min\{j:s_j=1\}} 1 \\
&= \min\{j \in [k] : s_j = 1\} - 2.
\end{aligned}$$

This proves the first case of eq. (2.32). Below, let $y > 1$. We have

$$[DL(-Ts)]_y \tag{2.34}$$

$$= \sum_{i\in[k]:i\neq y} h([-Ts]_y - [-Ts]_i) - \sum_{i\in[k]:i\neq y-1} h([-Ts]_{y-1} - [-Ts]_i) \tag{2.35}$$

$$= \sum_{i\in[k]:i\neq y} h([Ts]_i - [Ts]_y) - \sum_{i\in[k]:i\neq y-1} h([Ts]_i - [Ts]_{y-1}) \tag{2.36}$$

$$= \sum_{i\in[k]:i<y-1} h([Ts]_i - [Ts]_y) - h([Ts]_i - [Ts]_{y-1}) \tag{2.37}$$

$$+ h([Ts]_{y-1} - [Ts]_y) - h([Ts]_y - [Ts]_{y-1}) \tag{2.38}$$

$$+ \sum_{i\in[k]:i>y} h([Ts]_i - [Ts]_y) - h([Ts]_i - [Ts]_{y-1}) \tag{2.39}$$

If $s_y = 0$, then $[Ts]_y = [Ts]_{y-1}$ and so we have $[DL(-Ts)]_y = 0$. This proves the last case of eq. (2.32).

Below, assume the setting of the second case, i.e., $y > 1$ and $s_y = 1$. We first evaluate eq. (2.37). Since $i < y - 1$, we have

$$([Ts]_i - [Ts]_y) - ([Ts]_i - [Ts]_{y-1}) = [Ts]_{y-1} - [Ts]_y = -1$$

and

$$([Ts]_i - [Ts]_{y-1}) \leq 0.$$

The two preceding facts together imply that

$$h([Ts]_i - [Ts]_y) - h([Ts]_i - [Ts]_{y-1}) = 1$$

and so

$$\sum_{i\in[k]:i<y-1} h([Ts]_i - [Ts]_y) - h([Ts]_i - [Ts]_{y-1}) = y - 2.$$

Next, we evaluate eq. (2.38)

$$h([Ts]_{y-1} - [Ts]_y) - h([Ts]_y - [Ts]_{y-1}) = h(-1) - h(1) = 2.$$

Finally, we evaluate eq. (2.39). Since $i > y$, we have

$$[Ts]_i - [Ts]_y = \sum_{j=y+1}^{i} s_i.$$

From this, we see that

$$[Ts]_i - [Ts]_y \begin{cases} = 0 & : i < \min\{j \in [k] : j > y,\ s_j = 1\} \\ \geq 1 & : \text{otherwise.} \end{cases}$$

Hence,

$$h([Ts]_i - [Ts]_y) \begin{cases} = 1 & : i < \min\{j \in [k] : j > y,\ s_j = 1\} \\ = 0 & : \text{otherwise.} \end{cases}$$

On the other hand, $[Ts]_i - [Ts]_{y-1} = \sum_{j=y}^{i} s_i \geq s_y = 1$ and so $h([Ts]_i - [Ts]_{y-1}) = 0$. Therefore,

$$\sum_{i \in [k]: i > y} h([Ts]_i - [Ts]_y) - h([Ts]_i - [Ts]_{y-1})$$
$$= \min\{j \in [k] : j > y,\ s_j = 1\} - y - 1$$

Putting it all together, we have

$$[DL(-Ts)]_y = y - 2 + 2 + \min\{j \in [k] : j > y,\ s_j = 1\} - y - 1$$
$$= \min\{j \in [k] : j > y,\ s_j = 1\} - 1.$$

□

2.7.2.2 Proof of theorem II.15

Proof of theorem II.15. Let $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$. Pick σ such that $\sigma\varphi(\mathbf{S})$ is monotonic non-increasing. Hence, we have

$$\sigma\varphi(\mathbf{S}) = -[\underbrace{0, \ldots, 0}_{|S_1|\text{-times}}, \underbrace{1, \ldots, 1}_{|S_2|\text{-times}}, \ldots, \underbrace{l-1, \ldots, l-1}_{|S_l|\text{-times}}].$$

For each $i = 1, \ldots, l-1$, define $c_i(\mathbf{S}) = |S_1| + \cdots + |S_i|$.

Note that

$$S_1 \cup \cdots \cup S_i = \{j \in [k] : 0 \geq [\varphi(\mathbf{S})]_j \geq -(i-1)\} \tag{2.40}$$

$$= \{\sigma(1), \sigma(2), \ldots, \sigma(c_i(\mathbf{S}))\}. \tag{2.41}$$

Also, note that by definition, $c_i(\mathbf{S})$ is precisely the index in $[k-1]$ such that

$$\begin{cases} [\sigma\varphi(\mathbf{S})]_{c_i(\mathbf{S})} = -(i-1) \\ [\sigma\varphi(\mathbf{S})]_{c_i(\mathbf{S})+1} = -i. \end{cases}$$

Motivated by this, we define $\zeta(\mathbf{S}) \in \{0,1\}^k$ where

$$[\zeta(\mathbf{S})]_j = \begin{cases} 1 & : j = c_i(\mathbf{S}) + 1 \text{ for some } i = 1, \ldots, l-1 \\ 0 & : \text{otherwise.} \end{cases}$$

Then

$$\sigma\varphi(\mathbf{S}) = -T\zeta(\mathbf{S}). \tag{2.42}$$

Next, note that

$$\begin{aligned}
\langle p, L(\varphi(\mathbf{S}))\rangle &= \langle p, L(\sigma'\sigma\varphi(\mathbf{S}))\rangle && (2.43)\\
&= \langle p, \sigma' L(\sigma\varphi(\mathbf{S}))\rangle && (2.44)\\
&= \langle \sigma p, L(\sigma\varphi(\mathbf{S}))\rangle && (2.45)\\
&= \langle T'(\sigma p), DL(\sigma\varphi(\mathbf{S}))\rangle && (2.46)\\
&= \langle T'(\sigma p), DL(-T\zeta(\mathbf{S}))\rangle && (2.47)
\end{aligned}$$

where eq. (2.43) is by $\sigma' = \sigma^{-1}$, eq. (2.44) is by corollary II.29, eq. (2.45) is a basic property of the dot product, eq. (2.46) is by lemma II.33, eq. (2.47) is by eq. (2.42).

We first calculate $DL(-T\zeta(\mathbf{S}))$ by applying eq. (2.32) from lemma II.45 to $s = \zeta(\mathbf{S})$. For the case $y = 1$ of eq. (2.32), we have

$$\begin{aligned}
[DL(-T\zeta(\mathbf{S}))]_1 &= \min\{j \in [k-1] : [\zeta(\mathbf{S})]_j = 1\} - 2\\
&= c_1(\mathbf{S}) + 1 - 2\\
&= |S_1| - 1.
\end{aligned}$$

By definition, for $y > 1$, we note that $[\zeta(\mathbf{S})]_y = 1$ if and only if $y = c_i(\mathbf{S}) + 1$ for some $i \in \{1, \ldots, l-1\}$. Thus,

$$\begin{aligned}
[DL(-T\zeta(\mathbf{S}))]_{c_i(\mathbf{S})+1} &= \min\{j \in [k] : [\zeta(\mathbf{S})]_j = 1,\, j > c_i(\mathbf{S}) + 1\} - 1\\
&= (c_{i+1}(\mathbf{S}) + 1) - 1 = c_{i+1}(\mathbf{S}).
\end{aligned}$$

We summarize the above as follows:

$$[DL(-T\zeta(\mathbf{S}))]_y = \begin{cases} |S_1| - 1 & : y = 1 \\ c_{i+1}(\mathbf{S}) & : y = c_i(\mathbf{S}) + 1 \text{ for some } i \in [l-1] \\ 0 & : \text{otherwise.} \end{cases}$$

Next, we calculate $T'(\sigma p)$. Note that

$$\begin{aligned} [T'(\sigma p)]_y &= p_{\sigma(y)} + p_{\sigma(y+1)} + \cdots + p_{\sigma(k)} \\ &= 1 - \left(p_{\sigma(1)} + \cdots + p_{\sigma(y-1)}\right). \end{aligned}$$

In particular, $[T'(\sigma p)]_1 = 1$. Hence,

$$\begin{aligned} &\langle p, L(\varphi(\mathbf{S}))\rangle \\ &= \langle T'(\sigma p), DL(-T\zeta(\mathbf{S}))\rangle \\ &= [T'(\sigma p)]_1(|S_1| - 1) \\ &\qquad + \sum_{i=1}^{l-1} \left([T'(\sigma p)]_{c_i(\mathbf{S})+1}\right) c_{i+1}(\mathbf{S}) \\ &= |S_1| - 1 \\ &\qquad + \sum_{i=1}^{l-1} \left(1 - \left(p_{\sigma(1)} + \cdots + p_{\sigma(c_i(\mathbf{S}))}\right)\right) c_{i+1}(\mathbf{S}). \end{aligned}$$

Recall from eq. (2.41)

$$\{\sigma(1), \sigma(2), \ldots, \sigma(c_i(\mathbf{S}))\} = S_1 \cup \cdots \cup S_i.$$

Hence,

$$\left(1 - \left(p_{\sigma(1)} + \cdots + p_{\sigma(c_i(\mathbf{S}))}\right)\right) = \Pr_{Y \sim p}(Y \notin S_1 \cup \cdots \cup S_i).$$

Putting it all together, we have

$$\begin{aligned}\langle p, L(\varphi(\mathbf{S}))\rangle &= |S_1| - 1 + \sum_{i=1}^{l_{\mathbf{S}}-1} |S_1 \cup \cdots \cup S_{i+1}| \Pr_{Y\sim p}(Y \notin S_1 \cup \cdots \cup S_i) \\ &= \mathbb{E}_{Y\sim p}\left[[\ell(\mathbf{S})]_Y\right] \\ &= \langle p, \ell(\mathbf{S})\rangle\end{aligned}$$

This concludes the proof of theorem II.15. □

2.7.3 Minimally emblematic losses

We first introduce some basic properties of hyperplane arrangements that will be needed later.

Definition II.46. A *hyperplane* in $\mathbb{R}^d$ is a subset $H \subseteq \mathbb{R}^d$ of the form $H = \{v \in \mathbb{R}^k : b - \langle a, v\rangle = 0\}$ for some (column) vector $a \in \mathbb{R}^k$ and $b \in \mathbb{R}$.

Definition II.47. Define the following:

1. A *hyperplane arrangement* is a set of hyperplanes $\{H_n\}_{n\in I}$ indexed by a finite set I. Let the hyperplanes be written as $H_n = \{v \in \mathbb{R}^k : b^{(n)} - \langle a^{(n)}, v\rangle = 0\}$ for each $n \in I$.

2. Define $\mathfrak{s} : \mathbb{R}^k \to \{-1, 0, 1\}^I$ entrywise by

$$[\mathfrak{s}(v)]_n = \operatorname{sgn}\left(b^{(n)} - \langle a^{(n)}, v\rangle\right), \qquad \text{where } \forall t \in \mathbb{R},\ \operatorname{sgn}(t) = \begin{cases} 1 & : t > 0 \\ 0 & : t = 0 \\ -1 & : t < 0 \end{cases}.$$

3. Define the set $\Theta := \mathfrak{s}(\mathbb{R}^k) \subseteq \{-1, 0, 1\}^I$.

4. For each $\theta \in \Theta$, define

$$\tilde{P}_\theta := \mathfrak{s}^{-1}(\theta) = \{v \in \mathbb{R}^k : \mathfrak{s}(v) = \theta\} \quad \text{and } P_\theta := \text{cl}(\tilde{P}_\theta)$$

where cl denotes the closure of a set in $\mathbb{R}^k$ with the Euclidean topology.

Definition II.48. An *affine subspace* of $\mathbb{R}^k$ is a set of the form $W + v$ where $W \subseteq \mathbb{R}^k$ is a linear subspace and $v \in \mathbb{R}^k$ is a vector. Let C be a convex set. The *affine hull* $\text{Aff}(C)$ of C is defined as the smallest affine subspace containing C. The *relative interior* of C, denoted $\text{relint}(C)$, is defined as the subset of $v \in C$ such that for all $\epsilon > 0$ sufficiently small, we have that

$$\text{Aff}(C) \cap \{w \in \mathbb{R}^k : \|w - v\| < \epsilon\} \subseteq C.$$

In other words, $\text{relint}(C)$ is an open subset of $\text{Aff}(C)$. Here $\| \bullet \|$ is the Euclidean 2-norm on $\mathbb{R}^k$.

The following result is "folklore". Since we cannot find its proof, we prove it here.

Lemma II.49. *Let $\{H_n\}_{n \in I}$ be an arrangement of hyperplanes. Adopt all notations from definition II.47. The following are true:*

1. *For all $\theta \in \Theta$,* $\tilde{P}_\theta = \left\{ v \in \mathbb{R}^k : \begin{cases} \theta_n(b^{(n)} - \langle a^{(n)}, v \rangle) > 0 & : \theta_n \neq 0 \\ b^{(n)} - \langle a^{(n)}, v \rangle = 0 & : \theta_n = 0 \end{cases}, \forall n \in I \right\},$

2. *For all $\theta \in \Theta$,* $P_\theta = \left\{ v \in \mathbb{R}^k : \begin{cases} \theta_n(b^{(n)} - \langle a^{(n)}, v \rangle) \geq 0 & : \theta_n \neq 0 \\ b^{(n)} - \langle a^{(n)}, v \rangle = 0 & : \theta_n = 0 \end{cases}, \forall n \in I \right\},$

3. *For all $\theta \in \Theta$,* $\text{relint}(P_\theta) = \tilde{P}_\theta$,

4. $\bigsqcup_{\theta \in \Theta} \text{relint}(P_\theta) = \mathbb{R}^k$ *as a disjoint union.*

Proof. First, we note that item 1 follows directly from definition.

For item 2, let Q_θ denote the set on the right hand side of the identity. We want to show that $P_\theta = Q_\theta$. Recall that $P_\theta = \text{cl}(\tilde{P}_\theta)$ is by definition the smallest closed set containing $\tilde{P}_\theta$. Clearly, Q_θ is a closed set. Furthermore, by item 1, we have $\tilde{P}_\theta \subseteq Q_\theta$. Thus, we have the $P_\theta \subseteq Q_\theta$.

Conversely, let $v \in Q_\theta$ and $w \in \tilde{P}_\theta$. Then by item 1, we have that $(1-\lambda)w+\lambda v \in \tilde{P}_\theta$ for all $\lambda \in [0, 1)$. Now, $\lim_{\lambda \to 1}(1-\lambda)w + \lambda v = v$. Since $\text{cl}(\tilde{P}_\theta)$ is closed, it contains all limits. Hence $v \in \text{cl}(\tilde{P}_\theta) = P_\theta$, as desired. This proves that $Q_\theta \subseteq P_\theta$, as desired.

Next, we prove item 3. From the first paragraph of Ben-Tal et al. [BN20, Section 1.1.6.D], we have $\text{relint}(\tilde{P}_\theta) \subseteq \tilde{P}_\theta \subseteq \text{cl}(\tilde{P}_\theta)$. By Ben-Tal et al. [BN20, Theorem 1.1.1 (iv)], we have $\text{relint}(\tilde{P}_\theta) = \text{relint}(\text{cl}(\tilde{P}_\theta))$. By definition $P_\theta = \text{cl}(\tilde{P}_\theta)$. Putting it all together, we get $\text{relint}(P_\theta) \subseteq \tilde{P}_\theta$.

For the other inclusion, let $v \in \tilde{P}_\theta$. Let

$$W = \{v \in \mathbb{R}^k : b^{(n)} - \langle a^{(n)}, v\rangle = 0, \forall n \in I \text{ such that } \theta_n = 0\}.$$

Then by item 2, W is an affine subspace containing P_θ. Thus, by definition of the affine hull, we have $W \supseteq \text{Aff}(P_\theta)$. Furthermore, by item 1, we have, for all $\epsilon > 0$ sufficiently small, that $W \cap \{w \in \mathbb{R}^k : \|w - v\| < \epsilon\} \subseteq P_\theta$. This proves that $v \in \text{relint}(P_\theta)$ and so $\tilde{P}_\theta \subseteq \text{relint}(P_\theta)$.

Finally, we prove item 4

$$\bigsqcup_{\theta \in \Theta} \text{relint}(P_\theta) = \bigsqcup_{\theta \in \Theta} \tilde{P}_\theta = \bigsqcup_{\theta \in \mathfrak{s}(\mathbb{R}^k)} \mathfrak{s}^{-1}(\theta) = \mathbb{R}^k,$$

where for the middle equality, we recall that $\Theta = \mathfrak{s}(\mathbb{R}^k)$ by definition. □

2.7.3.1 Semiordered hyperplane arrangement

Below, we apply the results of lemma II.49 to the "semiorder hyperplane arrangement", which is closely connected to the WW-hinge loss.

Definition II.50. The *semiorder hyperplane arrangement* is the hyperplane arrangement in $\mathbb{R}^k$ indexed by the finite set $I = \{(i,j) \in [k] \times [k] : i \neq j\}$ with the (i,j)-th hyperplane given by $H_{(i,j)} = \{v \in \mathbb{R}^k : 1 - (v_i - v_j) = 0\}$.

Lemma II.51. *Let $L : \mathbb{R}^k \to \mathbb{R}^k_+$ be the WW-hinge loss and $\mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ be as in definition II.10. Let $\{H_{(i,j)}\}_{(i,j)\in I}$ be the semiorder hyperplane arrangement as in definition II.50. Adopt all notations from definition II.47. Then we have for all $\theta \in \Theta$ that*

1. *the restriction of L to P_θ, denoted $L|_{P_\theta}$, is an affine function,*
2. *$P_\theta \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ is nonempty.*

Proof. For the first item, fix some $i \in [k]$ and note that

$$[L(v)]_i = \sum_{j\in[k]:j\neq i} \max\{0, 1 - (v_i - v_j)\}.$$

Fix $(i,j) \in I$ where I is as in definition II.50. Then by lemma II.49 item 2, for all $v \in P_\theta$, we have

$$\max\{0, 1 - (v_i - v_j)\} = \begin{cases} 1 - (v_i - v_j) & : \theta_{(i,j)} = 1 \\ 0 & : \text{otherwise.} \end{cases}$$

In either case, $\max\{0, 1 - (v_i - v_j)\}$ is affine over P_θ.

Next, we prove the second item. Define $H_0 = \{v \in \mathbb{R}^k : \sum_{i\in[k]} v_i = 0\}$. Then $H_0 \cap P_\theta$ is nonempty for all $\theta \in \Theta$. To see this, first note that P_θ is nonempty by

construction. Furthermore, if $v \in P_\theta$ then $v + c\mathbf{1}^k \in P_\theta$ as well for any $c \in \mathbb{R}$. Thus, $v + (-(1/k)\sum_{i\in[k]} v_i)\mathbf{1}^k \in H_0 \cap P_\theta$.

Lemma II.52. *$H_0 \cap P_\theta$ does not contain any line.*

Proof. Suppose that this is false, i.e., $\mathfrak{l} \subseteq H_0 \cap P_\theta$ where $\mathfrak{l} \subseteq \mathbb{R}^k$ is a line. In particular, $\mathfrak{l} \subseteq H_0$. This means that $\mathfrak{l} = \{cw : c \in \mathbb{R}\}$ where $w \in H_0$ is a nonzero vector. Thus, there exists $i \neq j$ such that $w_i > 0$ and $w_j < 0$. Recall from definition II.47 that $[\mathfrak{s}(cw)]_{(i,j)} = \operatorname{sgn}(1 - c(w_i - w_j))$. Thus, as c ranges over $\mathbb{R}$, we have that $[\mathfrak{s}(cw)]_{(i,j)}$ takes on all three values in $\{-1, 0, 1\}$. However, by lemma II.49 item 2, $[\mathfrak{s}(cw)]_{(i,j)}$ can only take on at most two distinct values in $\{-1, 0, 1\}$. □

Before proceeding, we recall a definition:

Definition II.53. A *polyhedron* P in $\mathbb{R}^k$ is a set of the form $P = \{x \in \mathbb{R}^k : \langle a^{(n)}, x\rangle \leq b^{(n)}, \forall n \in [m]\}$ where m is a positive integer, $a^{(n)} \in \mathbb{R}^k$ and $b^{(n)} \in \mathbb{R}$ for all $n \in [m]$. For each $n \in [m]$, the tuple $(a^{(n)}, b^{(n)})$ is called a *constraint* of P. A point $x \in P$ is a *basic feasible solution* (BFS) if there exists $n_1, \ldots, n_k \in [m]$ such that

1. $\langle a^{(n_i)}, x\rangle = b^{(n_i)}$ for all $i \in [k]$, and
2. $\mathcal{A} := \{a^{(n_1)}, \ldots, a^{(n_k)}\}$ is a basis for $\mathbb{R}^k$.

By Bertsimas et al. [BT97, Theorem 2.6] and [BT97, Theorem 2.3], a polyhedron which does not contain any line always have a BFS. Earlier, we proved that $H_0 \cap P_\theta$ does not contain any line. Hence, $H_0 \cap P_\theta$ contains a BFS. For the remainder of this proof, let $x \in \mathbb{R}^k$ be such a BFS with associated basis $\mathcal{A} = \{a^{(n_1)}, \ldots, a^{(n_k)}\}$ as in definition II.53.

Let $e^i \in \mathbb{R}^k$ be the i-th elementary basis vector in $\mathbb{R}^k$. By definition of $P_\theta \cap H_0$, we have

$$\mathcal{A} \subseteq \{e^i - e^j : (i,j) \in I\} \cup \{\mathbf{1}^k\}$$

where we recall that I is as in definition II.50. Observe that $\langle \mathbf{1}^k, e^i - e^j \rangle = 0$ for all $(i,j) \in I$. Hence, we must have that $\mathbf{1}^k \in \mathcal{A}$, since otherwise $\mathcal{A}$ cannot span $\mathbb{R}^k$. This implies that we necessarily have $\mathbf{1}^k \in \mathcal{A}$. Without the loss of generality, let $a^{(n_k)} = \mathbf{1}^k$. Since $\mathcal{A}$ is linearly independent, we have

$$\mathcal{B} := \mathcal{A} \setminus \{a^{(n_k)}\} = \{a^{(n_1)}, \ldots, a^{(n_{k-1})}\} \subseteq \{e^i - e^j : (i,j) \in I\}.$$

Now, for each $i \in [k-1]$, let $(t_i, h_i) \in I$ be such that $a^{(n_i)} = e^{t_i} - e^{h_i}$. By the definition of P_θ, we have $\langle a^{(n_i)}, x \rangle = x_{t_i} - x_{h_i} = \pm 1$. Note that this implies that x is not a scalar multiple of $\mathbf{1}^k$.

Next, consider the directed graph G with vertices $V(G) = [k]$ and edges are $E(G) = \{(t_i, h_i) : i \in [k-1]\}$. Since $\mathcal{B}$ is linearly independent, we observe that if $(t_i, h_i) \in E(G)$, then $(h_i, t_i) \notin E(G)$. Let G^u be the undirected graph obtained from G by forgetting the edge orientations. By the preceding observation, we have $|E(G^u)| = k - 1$. An undirected edge is denoted as $\{\alpha, \beta\} \in E(G^u)$.

Observe that if $\{\alpha, \beta\} \in E(G^u)$, then $x_\alpha - x_\beta = \pm 1$.

Lemma II.54. *G^u is a tree, i.e., a connected graph without cycles.*

Proof. Note that G^u does not contain any cycles. To see this, note that if G^u had a cycle, then $\mathcal{A}$ cannot be linearly independent. Thus, G^u is a disjoint union of trees $\{T_1, \ldots, T_f\}$ where f is a positive integer. Since each T_i is a tree, we have

$|E(T_i)| = |V(T_i)| - 1$. On the other hand, we have

$$\begin{aligned} k - 1 &= |E(G^u)| \\ &= |E(T_1)| + \cdots + |E(T_f)| \\ &= |V(T_i)| + \cdots + |V(T_f)| - f \\ &= |V(G^u)| - f \\ &= k - f \end{aligned}$$

which implies that $f = 1$. In other words, G^u is a tree to begin with. □

Although we know that G^u is a tree, we only need the fact that G^u is connected.

Let $\alpha, \beta \in V(G^u)$. A *path* of length l from α to β is a sequence $\phi_1, \ldots, \phi_l \in V(G^u)$ such that

1. $\phi_1 = \alpha$ and $\phi_l = \beta$
2. $\{\phi_i, \phi_{i+1}\} \in E(G^u)$ for all $i \in [m-1]$.

The fact that G^u is connected implies that there exists a path between any two vertices $\alpha, \beta \in V(G^u)$. Define $\overline{x} := \max x$ and $\underline{x} := \min x$.

Lemma II.55. *For all $\beta \in [k]$, we have $\overline{x} - x_\beta \in \mathbb{Z}$.*

Proof. Let $\alpha \in \arg\max x$ and consider a path $\phi_1, \ldots, \phi_l \in V(G^u)$ from α to β. Observe that $x_\alpha - x_\beta = \sum_{i \in [l-1]} x_{\phi_i} - x_{\phi_{i+1}}$. Since $\{\phi_i, \phi_{i+1}\} \in E(G^u)$, we have $x_{\phi_i} - x_{\phi_{i+1}} = \pm 1$. This proves that $x_\alpha - x_\beta \in \mathbb{Z}$. □

Let $D := \overline{x} - \underline{x}$. Since $x_\beta \geq \underline{x}$, we have $0 \leq \overline{x} - x_\beta \leq D$. Apply lemma II.55 with $\beta \in \arg\min x$, we get $\overline{x} - \underline{x} = D \in \mathbb{Z}$. In summarize, we have proven that

$$\{x_\beta - \overline{x} : \beta \in [k]\} \subseteq \{-D, -D+1, \ldots, -1, 0\}. \tag{2.48}$$

Below, we will show that the inclusion in eq. (2.48) is in fact an equality.

Next, let $\overline{\varrho} \in \arg\max x$ and $\underline{\varrho} \in \arg\min x$. Let $\phi_1, \ldots, \phi_l \in V(G^u)$ be a path between $\overline{\varrho}$ and $\underline{\varrho}$. Note that by definition we have

1. $x_{\phi_1} = \overline{x}$ and $x_{\phi_l} = \underline{x}$,

2. $x_{\phi_i} - x_{\phi_{i+1}} = \pm 1$ for all $i \in [l-1]$.

Consider the sequence of numbers

$$S := (\underbrace{x_{\phi_1} - \overline{x}}_{=-D}, x_{\phi_2} - \overline{x}, \ldots, x_{\phi_{l-1}} - \overline{x}, \underbrace{x_{\phi_l} - \overline{x}}_{=0}).$$

Notice that the difference between consecutive entries of S is ± 1. Thus, the sequence S takes on every value in $\{-D, -D+1, \ldots, -1, 0\}$ at least once. This proves that eq. (2.48) holds with equality, i.e.,

$$\{x_\beta - \overline{x} : \beta \in [k]\} = \{-D, -D+1, \ldots, -1, 0\}. \tag{2.49}$$

Now, let $\sigma \in \mathfrak{S}_k$ be the element such that σx is monotonic non-increasing. Earlier, we argued that x is not a scalar multiple of $\mathbf{1}^k$. Thus, eq. (2.49) implies that $\sigma x - \overline{x}\mathbf{1}^k \in \mathcal{C}_{\mathbb{Z}}$. Consequently, we have $x - \overline{x}\mathbf{1}^k \in \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}$. Since $x \in P_\theta$, we have $x - \overline{x}\mathbf{1}^k \in P_\theta$ as well. This proves that $P_\theta \cap \mathfrak{S}_k \mathcal{C}_{\mathbb{Z}}$ is nonempty, which concludes the proof of lemma II.49. □

2.7.3.2 Proof of proposition II.19

Proof of proposition II.19. Let $m = |\mathcal{OP}_k|$. Index the elements of $\mathcal{OP}_k$ by $[m]$, i.e.,

$$\mathcal{OP}_k = \{\mathbf{S}^1, \ldots, \mathbf{S}^m\}.$$

For each $i \in [m]$, let $p^{(i)} \in \Delta^k$ be such that $\{\mathbf{S}^i\} = \arg\min_{\mathbf{S} \in \mathcal{OP}_k} \langle p, \ell(\mathbf{S}) \rangle$. The existence of such $p^{(i)}$s was confirmed by computer search for $k \in \{3, \ldots, 15\}$. Equivalently, $\mathbf{S}^i$ is the unique element of $\mathcal{OP}_k$ such that

$$\langle p^{(i)}, \ell(\mathbf{S}^i) \rangle = \underline{\ell}(p^{(i)}) = \underline{L}(p^{(i)}) \tag{2.50}$$

where the second equality is by corollary II.16.

Next, suppose L embeds another discrete loss $\lambda : \mathcal{R} \to \mathbb{R}^k_+$ with embedding map $\chi : \mathcal{R} \to \mathbb{R}^k$. Our goal is to show that $|\mathcal{R}| \geq |\mathcal{OP}_k|$. To this end, let $\mathcal{R} = \{r^1, \ldots, r^n\}$. Since L embeds λ via χ, we have by definition that $\underline{L}(p) = \underline{\lambda}(p) = \min_{r \in \mathcal{R}} \langle p, L(\chi(r)) \rangle$. In particular, for a fixed $i \in [m]$, there exists $\iota(i) \in [n]$ such that $\underline{L}(p^{(i)}) = \langle p^{(i)}, L(\chi(r^{\iota(i)})) \rangle$. Note that this defines a mapping

$$\iota : [m] \to [n]. \tag{2.51}$$

Let $v^{(i)} := \chi(r^{\iota(i)})$. Combined with eq. (2.50), we have

$$\langle p^{(i)}, L(v^{(i)}) \rangle = \underline{L}(p^{(i)}) = \underline{\ell}(p^{(i)}). \tag{2.52}$$

Consider $\{P_\theta\}_{\theta \in \Theta}$ as in lemma II.51. For each $v \in \mathbb{R}^k$, let $\theta(v) \in \Theta$ be the unique element such that $v \in \operatorname{relint}\left(P_{\theta(v)}\right)$. The existence and uniqueness of $\theta(v)$ is guaranteed by lemma II.49 item 4.

By eq. (2.52), we have $v^{(i)} \in \arg\min_{v \in \mathbb{R}^k} \langle p^{(i)}, L(v) \rangle$. By lemma II.51, the function $v \mapsto \langle p^{(i)}, L(v) \rangle$ is affine over the domain $P_{\theta(v^{(i)})}$. Furthermore, it is minimized at $v^{(i)} \in \operatorname{relint}(P_{\theta(v)})$. Thus, by Ben-Tal et al. [BN20, Lemma 1.2.2], the function $v \mapsto \langle p^{(i)}, L(v) \rangle$ is constant over the domain $v \in P_{\theta(v^{(i)})}$. Since $v^{(i)} \in P_{\theta(v^{(i)})}$ and

$\langle p^{(i)}, L(v^{(i)})\rangle = \underline{L}(p^{(i)})$ by eq. (2.52), we have

$$\langle p^{(i)}, L(v)\rangle = \underline{L}(p^{(i)}),\ \forall v \in P_{\theta(v^{(i)})} \tag{2.53}$$

Next, recall that $P_\theta \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ is nonempty for all $\theta \in \Theta$. In particular, $P_{\theta(v^{(i)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ is nonempty. By proposition II.13, we have $\mathfrak{S}_k\mathcal{C}_\mathbb{Z} = \varphi(\mathcal{OP}_k)$. All elements of $P_{\theta(v^{(i)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$ are of the form $\varphi(\mathbf{S})$ for some $\mathbf{S} \in \mathcal{OP}_k$. Fix such an $\mathbf{S}$ so that $\varphi(\mathbf{S}) \in P_{\theta(v^{(i)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z}$. Now,

$$\langle p^{(i)}, L(\varphi(\mathbf{S}))\rangle \overset{\text{eq. (2.53)}}{=} \underline{L}(p^{(i)}) \overset{\text{eq. (2.52)}}{=} \underline{\ell}(p^{(i)}).$$

Recall from right before eq. (2.50), we have that $\mathbf{S}^i$ is the unique element of $\mathcal{OP}_k$ such that $\langle p^{(i)}, L(\varphi(\mathbf{S}^{(i)}))\rangle = \underline{\ell}(p^{(i)})$. This proves that $\mathbf{S} = \mathbf{S}^i$. Thus, we have shown that

$$P_{\theta(v^{(i)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z} = \{\varphi(\mathbf{S}^i)\}. \tag{2.54}$$

Finally, we are now ready to prove that $n = |\mathcal{R}| \geq |\mathcal{OP}_k| = m$. It suffices to show that the mapping $\iota : [m] \to [n]$ defined at eq. (2.51) is injective. Suppose that there exists distinct $i, j \in [m]$ such that $\iota(i) = \iota(j)$. Then

$$\begin{aligned}
& r^{\iota(i)} = r^{\iota(j)} \\
\implies & v^{(i)} = v^{(j)} \quad \because \text{definition of } v^{(i)} := \chi(r^{\iota(i)}) \\
\implies & \theta(v^{(i)}) = \theta(v^{(j)}) \\
\implies & P_{\theta(v^{(i)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z} = P_{\theta(v^{(j)})} \cap \mathfrak{S}_k\mathcal{C}_\mathbb{Z} \\
\implies & \{\varphi(\mathbf{S}^i)\} = \{\varphi(\mathbf{S}^j)\} \quad \because \text{eq. (2.54)} \\
\implies & \varphi(\mathbf{S}^i) = \varphi(\mathbf{S}^j) \\
\implies & \mathbf{S}^i = \mathbf{S}^j \quad \because \varphi \text{ is a bijection}
\end{aligned}$$

which contradicts $i \neq j$. Thus, we have that $\iota : [m] \to [n]$ is injective which implies that $n \geq m$. □

2.7.4 The argmax link

Definition II.56. For $\sigma \in \mathfrak{S}_k$ and $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$, define $\sigma(\mathbf{S}) \in \mathcal{OP}_k$ by

$$\sigma(\mathbf{S}) = (\sigma(S_1), \ldots, \sigma(S_l))$$

where $\sigma(S_i) = \{\sigma(j) : j \in S_i\}$ for each $i \in [l]$.

Lemma II.57. *For $\sigma \in \mathfrak{S}_k$ and $\mathbf{S} = (S_1, \ldots, S_l) \in \mathcal{OP}_k$, we have*

$$\sigma' \varphi(\mathbf{S}) = \varphi(\sigma(\mathbf{S})).$$

Proof. By definition, we have

$$[\varphi(\sigma(\mathbf{S}))]_j = -(i-1), \forall j \in \sigma(S_i).$$

Since $j \in \sigma(S_i) \iff \sigma^{-1}(j) \in S_i$, we have

$$[\varphi(\sigma(\mathbf{S}))]_j = -(i-1), \forall j \in [k] : \sigma^{-1}(j) \in S_i.$$

Introduce the change of variable $m = \sigma^{-1}(j)$, we have

$$[\varphi(\sigma(\mathbf{S}))]_{\sigma(m)} = -(i-1), \forall m \in S_i.$$

On the other hand, we have

$$[\sigma' \varphi(\mathbf{S})]_{\sigma(m)} = [\varphi(S)]_{\sigma' \sigma(m)} = [\varphi(S)]_m = -(i-1), \forall m \in S_i.$$

This proves that $\sigma'\varphi(\mathbf{S}) = \varphi(\sigma\mathbf{S})$. □

Corollary II.58. *For all $\mathbf{S} \in \mathcal{OP}_k$ and $\sigma \in \mathfrak{S}_k$, we have $\sigma\ell(\mathbf{S}) = \ell(\sigma'\mathbf{S})$.*

Proof. Since Δ^k spans $\mathbb{R}^k$, it suffices to check that $\langle p, \sigma\ell(\mathbf{S})\rangle = \langle p, \ell(\sigma'\mathbf{S})\rangle$ for all $p \in \Delta^k$. To this end, we have

$$\begin{aligned}
\langle p, \ell(\sigma'\mathbf{S})\rangle &= \langle p, L(\varphi(\sigma'\mathbf{S}))\rangle \quad \because \text{theorem II.15} \\
&= \langle p, L(\sigma\varphi(\mathbf{S}))\rangle \quad \because \text{lemma II.57} \\
&= \langle p, \sigma L(\varphi(\mathbf{S}))\rangle \quad \because \text{corollary II.29} \\
&= \langle \sigma' p, L(\varphi(\mathbf{S}))\rangle \\
&= \langle \sigma' p, \ell(\mathbf{S})\rangle \quad \because \text{theorem II.15} \\
&= \langle p, \sigma\ell(\mathbf{S})\rangle
\end{aligned}$$

as desired. □

For $p \in \Delta^k$, define

$$\gamma(p) := \operatorname*{arg\,min}_{\mathbf{S} \in \mathcal{OP}_k} \langle p, \ell(\mathbf{S})\rangle, \tag{2.55}$$

$$\Gamma(p) := \operatorname*{arg\,min}_{v \in \mathbb{R}^k} \langle p, L(v)\rangle. \tag{2.56}$$

Lemma II.59. *Let $p \in \Delta^k_\downarrow$, $v \in \Gamma(p)$, and σ be such that $\sigma v \in \mathbb{R}^k_\downarrow$. Then $\sigma p = p$ and $\sigma v \in \Gamma(p)$.*

Proof. Let $i \in [k-1]$ be such that $v_i < v_{i+1}$. We first prove that $p_i = p_{i+1}$. Let

$\tau = \sigma_{(i,i+1)}$. Since τ is a transposition, we have $\tau' = \tau$. Now,

$$\begin{aligned}
0 &\leq \langle p, L(\tau v)\rangle - \langle p, L(v)\rangle \quad \because \text{Optimality of } v \\
&= \langle p, \tau L(v)\rangle - \langle p, L(v)\rangle \quad \because \text{corollary II.29} \\
&= \langle \tau p, L(v)\rangle - \langle p, L(v)\rangle \quad \because \tau' = \tau. \\
&= (p_{i+1} - p_i)[L(v)]_i + (p_i - p_{i+1})[L(v)]_{i+1} \\
&= (p_{i+1} - p_i)([L(v)]_i - [L(v)]_{i+1})
\end{aligned}$$

By lemma II.30, we have $[L(v)]_i - [L(v)]_{i+1} > 0$. By assumption, we have $p_i \geq p_{i+1}$. If we have $p_i > p_{i+1}$, then

$$\underbrace{(p_{i+1} - p_i)}_{<0}\underbrace{([L(v)]_i - [L(v)]_{i+1})}_{>0} < 0$$

which is a contradiction. Hence, we must have $p_i = p_{i+1}$. Repeating the proof with the update $v \leftarrow \tau v$, we obtain a composition of transpositions

$$\sigma := \sigma_{(i_1,i_1+1)}\sigma_{(i_2,i_2+1)} \cdots \sigma_{(i_m,i_m+1)}$$

such that $\sigma v \in \mathbb{R}^k_\downarrow$ and $\sigma p = p$. Finally,

$$\underline{L}(p) = \langle p, L(v)\rangle = \langle p, \sigma'\sigma L(v)\rangle = \langle \sigma p, L(\sigma v)\rangle = \langle p, L(\sigma v)\rangle$$

implies that $\sigma v \in \Gamma(p)$. □

Lemma II.60. *Let $\sigma \in \mathfrak{S}_k$ and $v \in \mathbb{R}^k$. Then* $\arg\max \sigma v = \sigma^{-1}(\arg\max v)$.

Proof. Let $M = \max v = \max \sigma v$.

$$\begin{aligned}\arg\max \sigma v &= \{j \in [k] : [\sigma v]_j = M\} \\ &= \{j \in [k] : [v]_{\sigma(j)} = M\}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\sigma^{-1}(\arg\max v) &= \{j \in [k] : \sigma(j) \in \arg\max v\} \\ &= \{j \in [k] : [v]_{\sigma(j)} = M\} \\ &= \arg\max \sigma v\end{aligned}$$

as desired. □

Lemma II.61. *Let $p \in \Delta^k_\downarrow$ be such that $\max p > \frac{1}{k}$. Let $v \in \Gamma(p)$, then there exists $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma(p)$ such that $\arg\max v \subseteq S_1$.*

Proof. Recall by definition, $v \in \Gamma(p)$ if and only if $\underline{L}(p) = \langle p, L(v)$. We first claim that v is not a scalar multiple of the all ones vector. Suppose it is, then $\underline{L}(p) = \langle p, L(v)\rangle = \langle p, L(0)\rangle$ by lemma II.25, which implies that $0 \in \Gamma(p)$. Now, by lemma II.44, we have $0 \notin \Gamma(p)$ since $\min p < \frac{1}{k}$ by the assumption that $\max p > \frac{1}{k}$. This is a contradiction. Hence, the claim is proved.

Next, let $n = |\arg\max v|$. By our claim that v is non-constant, we have that $n \in [k-1]$. Let $\sigma \in \mathfrak{S}_k$ be such that $\sigma v \in \mathbb{R}^k_\downarrow$. Thus, by construction, we have $\arg\max v = [n]$. Hence, we have, by lemma II.60,

$$[n] = \arg\max \sigma v = \sigma^{-1}(\arg\max v)$$

or, equivalently, $\arg\max v = \sigma([n])$. Since $n = |\arg\max v| \in [k-1]$, v is feasible for

the right hand side of eq. (2.11). Thus, we have

$$\underline{L}(p) = \underline{L}^n(p).$$

By lemma II.43

$$\underline{L}^n(p) = \min_{w \in \mathcal{C}_{\mathbb{Z}} \,:\, w_n = 0} \langle p, L(w) \rangle. \tag{2.57}$$

Let w^* be a minimizer of the above optimization. Since $w^* \in \mathcal{C}_{\mathbb{Z}}$, consider $\mathbf{S} = (S_1, \ldots, S_l) := \tilde{\psi}(w^*)$. Hence, by the definition of $\tilde{\psi}$, we have that $S_1 = \arg\max w^*$. Note that

$$\begin{aligned}
\underline{L}(p) = \underline{L}^n(p) &= \langle p, L(w^*) \rangle \\
&= \langle p, L(\varphi(\mathbf{S})) \rangle \quad \because \text{proposition II.13} \\
&= \langle p, \ell(\mathbf{S}) \rangle \quad \because \text{theorem II.15} \\
&= \langle \sigma p, \ell(\mathbf{S}) \rangle \quad \because \sigma p = p \text{ by lemma II.59} \\
&= \langle p, \sigma' \ell(\mathbf{S}) \rangle \\
&= \langle p, \ell(\sigma \mathbf{S}) \rangle \quad \because \text{corollary II.58.}
\end{aligned}$$

Putting it all together, we have

$$\langle p, \ell(\sigma \mathbf{S}) \rangle = \underline{L}(p) = \underline{\ell}(p)$$

where the second equality follows from corollary II.16. This proves that $\sigma \mathbf{S} \in \gamma(p)$. Note that since w^* is feasible for the optimization on the right hand side of eq. (2.57), we have $\arg\max w^* = \{i \in [k] : w_i^* = 0\} \supseteq [n]$. Furthermore, recall that $S_1 = \arg\max w^*$. Putting it all together, we have $\sigma(S_1) \supseteq \sigma([n]) = \arg\max v$. Thus, $\sigma(\mathbf{S})$ satisfies the desired conditions. □

Lemma II.62. *For all $p \in \Delta^k$ and $\sigma \in \mathfrak{S}_k$, we have*

$$\mathbf{S} \in \gamma(\sigma p) \iff \sigma\mathbf{S} \in \gamma(p), \tag{2.58}$$

$$v \in \Gamma(\sigma p) \iff \sigma' v \in \Gamma(p). \tag{2.59}$$

Proof. We first prove eq. (2.58). Let $\mathbf{S} \in \gamma(\sigma p)$. Then

$$\begin{aligned}
\underline{\ell}(\sigma p) &= \langle \sigma p, \ell(\mathbf{S}) \rangle \\
&= \langle p, \sigma' \ell(\mathbf{S}) \rangle \\
&= \langle p, \ell(\sigma \mathbf{S}) \rangle \quad \because \text{corollary II.58} \\
&\geq \underline{\ell}(p).
\end{aligned}$$

By the same argument, we have $\underline{\ell}(p) \geq \underline{\ell}(\sigma p)$. Thus, $\underline{\ell}(p) = \underline{\ell}(\sigma p)$ and $\sigma\mathbf{S} \in \gamma(p)$. This proves the $\implies$ direction eq. (2.58). To prove the other direction, we first write $p = \sigma'\sigma p$ and note that

$$\sigma\mathbf{S} \in \gamma(\sigma'\sigma p) \implies \sigma'\sigma\mathbf{S} \in \gamma(\sigma p) \iff \mathbf{S} \in \gamma(\sigma p).$$

Next, we prove eq. (2.59). By lemma II.31, we have $\underline{L}(\sigma p) = \underline{L}(p)$. Let $v \in \Gamma(\sigma p)$, then

$$\begin{aligned}
\underline{L}(p) = \underline{L}(\sigma p) &= \langle \sigma p, L(v) \rangle \\
&= \langle p, \sigma' L(v) \rangle \\
&= \langle p, L(\sigma' v) \rangle \quad \because \text{corollary II.29.}
\end{aligned}$$

Thus, $\sigma' v \in \Gamma(p)$. This proves the $\implies$ direction of eq. (2.59). For the other direction,

$$\sigma' v \in \Gamma(\sigma' \sigma p) \implies \sigma\sigma' v \in \Gamma(\sigma p) \iff v \in \Gamma(\sigma p).$$

□

2.7.4.1 Proof of theorem II.21

Proof of theorem II.21. Let $\sigma \in \mathfrak{S}_k$ be such that $\sigma p \in \Delta^k_\downarrow$. By lemma II.62, we have $\sigma v \in \Gamma(\sigma p)$. Then by lemma II.61, there exists $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma(\sigma p)$ such that $S_1 \supseteq \arg\max \sigma v = \sigma^{-1}(\arg\max v)$, where the equality is due to lemma II.60. Applying σ, to both side, we have $\sigma S_1 \supseteq \arg\max v$. By lemma II.62, we have $\sigma \mathbf{S} \in \gamma(p)$. Hence, we are done. □

Lemma II.63. *Let $p \in \Delta^k_\downarrow$ be such that $\arg\max p = \{1\}$ and $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma(p)$. Then $1 \in S_1$.*

Proof. Let $v = \varphi(\mathbf{S})$. Since $\mathbf{S}$ is nontrivial, we have $\max v > \min v$. By construction, we have $\arg\max v = S_1$. Hence, if $1 \notin S_1$, then there exists some $j \in \{2, \ldots, k\}$ such that $v_j > v_1$. Then lemma II.30 implies that $[L(v)]_1 > [L(v)]_j$ and so

$$\langle p, L(v)\rangle - \langle p, \sigma_j L(v)\rangle = (p_1 - p_j)([L(v)]_1 - [L(v)]_j) > 0.$$

But $\underline{\ell}(p) = \langle p, \ell(\mathbf{S})\rangle = \langle p, L(v)\rangle$ and

$$\langle p, \sigma_j L(v)\rangle = \langle p, L(\sigma_j v)\rangle = \langle p, L(\sigma_j \varphi(\mathbf{S})\rangle = \langle p, L(\varphi(\sigma_j \mathbf{S}))\rangle = \langle p, \ell(\sigma_j \mathbf{S})\rangle$$

Thus, we have

$$\langle p, \ell(\mathbf{S})\rangle - \langle p, \ell(\sigma_j \mathbf{S})\rangle > 0$$

which contradicts that $\mathbf{S} \in \gamma(p)$. □

Definition II.64. A $\mathfrak{S}_k$*-invariant property* is a boolean function

$$\mathcal{B} : \Delta^k \to \{\texttt{true}, \texttt{false}\} \tag{2.60}$$

such that $\mathcal{B}(p) \implies \mathcal{B}(\sigma p)$ for all $\sigma \in \mathfrak{S}_k$ and $p \in \Delta^k$. Here, " $\implies$ " denotes logical implication.

Lemma II.65. *Let $\mathcal{B}$ and $\mathcal{C}$ be $\mathfrak{S}_k$-invariant properties. Suppose that for all $p \in \Delta^k_\downarrow$, $\mathcal{B}(p)$ implies $\mathcal{C}(p)$. Then for all $p \in \Delta^k$, we have $\mathcal{B}(p)$ implies $\mathcal{C}(p)$.*

Proof. Let $p \in \Delta^k$ be arbitrary. Pick σ such that $\sigma p \in \Delta^k_\downarrow$. Then

$$\mathcal{B}(p) \implies \mathcal{B}(\sigma p) \implies \mathcal{C}(\sigma p) \implies \mathcal{C}(p)$$

where for the first and last implications we used the $\mathfrak{S}_k$-invariance property of $\mathcal{B}$ and $\mathcal{C}$, and for the implication in the middle we used the assumption in the lemma. □

Lemma II.66. *Let $p \in \Delta^k$. Consider the statement $\mathcal{B}_1(p)$ which returns* `true` *if and only if*

$$\text{for all } \mathbf{S} \in \gamma(p),\ |S_1| = 1 \text{ and } S_1 = \arg\max p. \tag{2.61}$$

Then $\mathcal{B}_1$ is a $\mathfrak{S}_k$-invariant property.

Proof. Let $p \in \Delta^k$ and $\sigma \in \mathfrak{S}_k$. Suppose $\mathcal{B}_1(p)$ is true. We need to show that $\mathcal{B}_1(\sigma' p)$ is true. Let $\mathbf{S} \in \gamma(\sigma p)$. By lemma II.62, we have $\sigma\mathbf{S} \in \gamma(p)$. Since $\mathcal{B}_1(p)$ is true, we have $|\sigma(S_1)| = 1$ and $\sigma(S_1) = \arg\max p$. Thus, we immediately get that $|S_1| = 1$. By lemma II.60, we have $S_1 = \sigma^{-1}(\arg\max p) = \arg\max \sigma p$. The two preceding facts is equivalent to $\mathcal{B}_1(p)$ being true, by definition. □

2.7.4.2 Proof of proposition II.22

Proof of proposition II.22. By lemma II.66 and lemma II.65, we may assume $p \in \Delta^k_\downarrow$. lemma II.63 implies that $1 \in S_1$. If $|S_1| = 1$, then $S_1 = \{1\}$ and the result is proven. Below, suppose $|S_1| > 1$. We define

$$S_1' = \{1\}, \quad S_1'' = S_1 \setminus \{1\}.$$

Define

$$\mathbf{S}' = (S_1', S_1'', S_2, \ldots, S_l) \in \mathcal{OP}_k.$$

We claim that $\langle p, \ell(\mathbf{S}')\rangle < \langle p, \ell(\mathbf{S})\rangle$. Given the claim, we would have a contradiction that $\mathbf{S} \in \gamma(p)$ and so $|S_1| = 1$ must be true. Let $Y \sim p$ and define

$$\beta := \sum_{j=1}^{l-1} |S_1 \cup \cdots \cup S_{j+1}| \Pr(Y \notin S_1 \cup \cdots \cup S_j)$$

Observe that

$$\begin{aligned}
\langle p, \ell(\mathbf{S}')\rangle &= |S_1'| - 1 + |S_1' \cup S_1''| \Pr(Y \notin S_1') + \beta \\
&= |S_1| \Pr(Y \neq 1) + \beta \\
&< \frac{1}{2}|S_1| + \beta.
\end{aligned}$$

On the other hand, we have

$$\langle p, \ell(\mathbf{S})\rangle = |S_1| - 1 + \beta.$$

Hence, we have

$$\begin{aligned}
\langle p, \ell(\mathbf{S})\rangle - \langle p, \ell(\mathbf{S}')\rangle &= |S_1| - 1 - |S_1| \Pr(Y \neq 1) \\
&> |S_1| - 1 - \frac{1}{2}|S_1| \\
&= \frac{1}{2}|S_1| - 1 \\
&\geq \frac{2}{2} - 1 \\
&= 0.
\end{aligned}$$

which proves the claim. □

2.7.4.3 Proof of proposition II.23

Proof of proposition II.23. Since $\arg\max p = \{j^*\}$, we have $(\{j^*\}, [k] \setminus \{j^*\}) = (\arg\max p, [k] \setminus \arg\max p)$. We check that the statement below defines a $\mathfrak{S}_k$-invariant property:

"p satisfies $(\arg\max p, [k] \setminus \arg\max p)$ is the unique element of $\gamma(p)$." (2.62)

Let p satisfy eq. (2.62). By lemma II.62, we have $\sigma^{-1}(\arg\max p, [k] \setminus \arg\max p)$ is the unique element of $\gamma(\sigma p)$. By definition,

$$\sigma^{-1}(\arg\max p, [k] \setminus \arg\max p) = (\sigma^{-1} \arg\max p, \sigma^{-1}([k] \setminus \arg\max p)).$$

By lemma II.60, we have $\sigma^{-1} \arg\max p = \arg\max \sigma p$. Thus, we have

$$\sigma^{-1}(\arg\max p, [k] \setminus \arg\max p) = (\arg\max \sigma p, [k] \setminus \arg\max \sigma^{-1} p)$$

is the unique element of $\gamma(\sigma p)$. In other words, σp satisfies eq. (2.62), as desired.

Furthermore, "p satisfies the symmetric noise condition." is obviously $\mathfrak{S}_k$-invariant. Hence, by lemma II.66 and lemma II.65, we may assume $p \in \Delta^k_{\downarrow}$. Pick $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma(p)$. lemma II.63 implies that $1 \in S_1$. By definition II.5 of $\mathcal{OP}_k$, we have $l \geq 2$. We first show that $l = 2$ by contradiction. Suppose that $l > 2$. Define $\mathbf{S}' = (S'_1, \ldots, S'_{l-1})$ where

$$S'_1 := S_1, \quad S'_2 := S_2 \cup S_3, \quad S'_j := S_{j+1}, \forall j \in \{3, \cdots, l-1\}.$$

Let $Y \sim p$ and

$$\beta := \sum_{j=3}^{l-1} |S_1 \cup \cdots \cup S_{j+1}| \Pr(Y \notin S_1 \cup \cdots \cup S_j).$$

Then we have

$$\begin{aligned}\langle p, \ell(\mathbf{S})\rangle &= |S_1| - 1 + |S_1 \cup S_2| \Pr(Y \notin S_1) \\ &\quad + |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1 \cup S_2) + \beta\end{aligned}$$

and

$$\begin{aligned}\langle p, \ell(\mathbf{S}')\rangle &= |S_1'| - 1 + |S_1' \cup S_2'| \Pr(Y \notin S_1') \\ &\quad + \sum_{j=2}^{l-2} |S_1' \cup \cdots \cup S_{j+1}'| \Pr(Y \notin S_1' \cup \cdots \cup S_j') \\ &= |S_1| - 1 + |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1) \\ &\quad + \sum_{j=2}^{l-2} |S_1 \cup \cdots \cup S_{j+2}| \Pr(Y \notin S_1 \cup \cdots \cup S_{j+1}) \\ &= |S_1| - 1 + |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1) \\ &\quad + \sum_{j=3}^{l-1} |S_1 \cup \cdots \cup S_{j+1}| \Pr(Y \notin S_1 \cup \cdots \cup S_j) \\ &= |S_1| - 1 + |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1) + \beta\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}\langle p, \ell(\mathbf{S})\rangle - \langle p, \ell(\mathbf{S}')\rangle &= |S_1 \cup S_2| \Pr(Y \notin S_1) \\ &\quad + |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1 \cup S_2) \\ &\quad - |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1) \\ &= |S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1 \cup S_2) \\ &\quad - |S_3| \Pr(Y \notin S_1).\end{aligned}$$

Define $s_i := |S_i|$ for each $i \in [l]$. Then

$$|S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1 \cup S_2) = (s_1 + s_2 + s_3)(k - s_1 - s_2)\frac{1-\alpha}{k-1}$$

and

$$|S_3| \Pr(Y \notin S_1) = s_3(k - s_1)\frac{1-\alpha}{k-1}.$$

Now, we have

$$\begin{aligned}
&(s_1 + s_2 + s_3)(k - s_1 - s_2) - s_3(k - s_1) \\
&= ((s_1 + s_2) + s_3)((k - s_1) - s_2) - s_3(k - s_1) \\
&= (s_1 + s_2)(k - s_1) - s_2(s_1 + s_2) - s_2 s_3 \\
&= (s_1 + s_2)k - (s_1 + s_2)^2 - s_2 s_3 \\
&\geq (s_1 + s_2)(s_1 + s_2 + s_3) - (s_1 + s_2)^2 - s_2 s_3 \\
&= s_1 s_3
\end{aligned}$$

where for the inequality, we used the fact that $k \geq s_1 + s_2 + s_3$. Finally, we now get a contradiction of the optimality of $\mathbf{S}$:

$$|S_1 \cup S_2 \cup S_3| \Pr(Y \notin S_1 \cup S_2) - |S_3| \Pr(Y \notin S_1) \geq s_1 s_3 \frac{1-\alpha}{k-1} > 0$$

implies

$$\langle p, \ell(\mathbf{S}) \rangle - \langle p, \ell(\mathbf{S}') \rangle > 0.$$

This proves the claim that if $\mathbf{S} = (S_1, \ldots, S_l) \in \gamma(p)$, then $l = 2$ and so $\mathbf{S} = (S_1, [k] \setminus S_1)$. Next, we show that $S_1 = \{1\}$. We already have shown that $1 \in S_1$. We

calculate

$$\begin{aligned}\langle p, \ell((S_1, [k] \setminus S_1))\rangle &= |S_1| - 1 + k \Pr(Y \notin S_1) \\ &= |S_1| - 1 + k(k - |S_1|)\left(\frac{1-\alpha}{k-1}\right) \\ &= |S_1|\left(1 - k\left(\frac{1-\alpha}{k-1}\right)\right) + C\end{aligned}$$

where $C = -1 + k^2\left(\frac{1-\alpha}{k-1}\right)$ does not depend on $|S_1|$. To prove that $|S_1| = 1$, by minimality of $\mathbf{S}$ it suffices to show that

$$1 - k\left(\frac{1-\alpha}{k-1}\right) > 0.$$

To see this, note that

$$\begin{aligned}1 > k\left(\frac{1-\alpha}{k-1}\right) &\iff \frac{1}{k} > \frac{1-\alpha}{k-1} \\ &\iff \frac{k-1}{k} = 1 - \frac{1}{k} > 1 - \alpha \\ &\iff \alpha > \frac{1}{k}\end{aligned}$$

where the last line is part of our assumption in the lemma statement. □

2.8 Derivation of the figures

We discuss how figs. 2.1 and 2.2 are obtained.

2.8.1 fig. 2.1

When $k = 3$, there are 12 nontrivial ordered partitions. Below, we represent $\mathcal{OP}_3$ vectorially in $\mathbb{R}^3$ using proposition II.13:

```
OPk = [-2 -2 -1 -1 -1 -1  0 -1  0  0  0  0 ;
        0 -1  0  0  0 -1  0 -2 -1 -1 -1 -2 ;
       -1  0  0 -1 -2  0 -1  0  0 -1 -2 -1 ]
```

Every column of the matrix `OPk` is a nontrivial ordered partition, e.g., the first column $\begin{bmatrix} -2 \\ 0 \\ -1 \end{bmatrix} \mapsto 2|3|1$. Consider the following matrix whose columns are $\ell(\mathbf{S}) = L^{WW}(\varphi(\mathbf{S})) \in \mathbb{R}^3_+$ where ℓ is the ordered partition loss and $\mathbf{S} \in \mathcal{OP}_3$.

```
ell = [ 5  5  4  3  2  3  1  2  1  0  0  0 ;
        0  2  1  0  0  3  1  5  4  3  2  5 ;
        2  0  1  3  5  0  4  0  1  3  5  2 ]
```

For example, the first column of `ell` is the result of applying $L^{WW} : \mathbb{R}^k \to \mathbb{R}^k_+$ to the first column of `OPk` , *i.e.*, $\begin{bmatrix} 5 \\ 0 \\ 2 \end{bmatrix} = L^{WW}\left(\begin{bmatrix} -2 \\ 0 \\ -1 \end{bmatrix}\right) = \ell^{\mathcal{OP}}(2|3|1)$. Finally, to get the region in fig. 2.3 labelled by "2|3|1", we plot the (p_2, p_3) coordinates of the following polytope:

$$\mathrm{Reg}(2|3|1) := \{p \in \Delta^3 : \langle p, \ell(2|3|1) - \ell(\mathbf{S})\rangle \leq 0,\ \forall \mathbf{S} \in \mathcal{OP}_3,\ \mathbf{S} \neq 2|3|1\}.$$

Repeat this procedure for all of $\mathcal{OP}_3$, we obtain fig. 2.3.

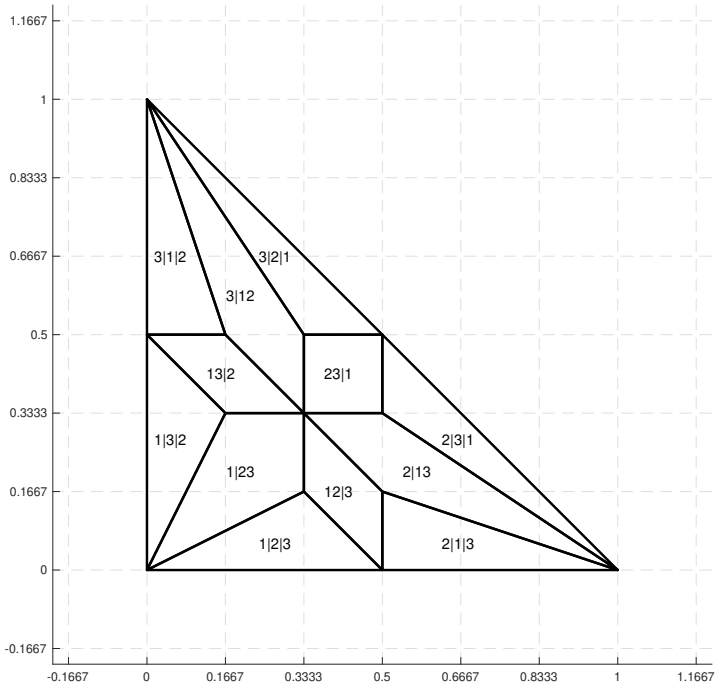

Figure 2.3: Each polygonal region is the polytope $\mathrm{Reg}(\mathbf{S})$ projected onto its last two coordinates overall $\mathbf{S} \in \mathcal{OP}_3$.

2.8.2 fig. 2.2

For the left panel of fig. 2.2, we compute $\Omega_{L^{ww}}$

$$\Omega_{L^{ww}} := \{p \in \Delta^k : |\arg\max p| = 1,\ \arg\max v = \arg\max p,\ \forall v \in \Gamma_{L^{ww}}(p)\}.$$

Thus, the region in light gray in the left panel of fig. 2.2 is the union of the polygons of fig. 2.1 labelled by an ordered partition whose the top bucket has 2 elements. This characterize $\Omega_{L^{ww}}$ up to a set of Lebesgue measure zero.

For the right panel, consider $v \in \Gamma_{L^{cs}}(p)$. Liu [Liu07, Lemma 4] states that if $\max p < 1/2$, then $v = (0, 0, 0)$. Furthermore, if $\max p > 1/2$, then $\arg\max v =$

$\arg\max p$. This characterize Ω_{LCS} up to a set of Lebesgue measure zero.

CHAPTER III

An Exact Solver for the Weston-Watkins SVM Subproblem

Recent empirical evidence suggests that the Weston-Watkins support vector machine is among the best performing multiclass extensions of the binary SVM. Current state-of-the-art solvers repeatedly solve a particular subproblem approximately using an iterative strategy. In this work, we propose an algorithm that solves the subproblem exactly using a novel reparametrization of the Weston-Watkins dual problem. For linear WW-SVMs, our solver shows significant speed-up over the state-of-the-art solver when the number of classes is large. Our exact subproblem solver also allows us to prove linear convergence of the overall solver.

3.1 Introduction

Support vector machines (SVMs) [BGV92; CV95] are a powerful class of algorithms for classification. In the large scale studies by Fernández-Delgado et al. [Fer+14] and by Klambauer et al. [Kla+17], SVMs are shown to be among the best performing classifers.

The original formulation of the SVM handles only binary classification. Subsequently, several variants of multiclass SVMs have been proposed [LLW04; CS01;

WW99]. However, as pointed out by Doğan et al. [DGI16], no variant has been considered canonical.

The empirical study of Doğan et al. [DGI16] compared nine prominent variants of multiclass SVMs and demonstrated that the Weston-Watkins (WW) and Crammer-Singer (CS) SVMs performed the best with the WW-SVM holding a slight edge in terms of both efficiency and accuracy. This work focuses on the computational issues of solving the WW-SVM optimization efficiently.

SVMs are typically formulated as quadratic programs. State-of-the-art solvers such as LIBSVM [CL11] and LIBLINEAR [Fan+08] apply block coordinate descent to the associated dual problem, which entails repeatedly solving many small subproblems. For the binary case, these subproblems are easy to solve exactly.

The situation in the multiclass case is more complex, where the form of the subproblem depends on the variant of the multiclass SVM. For the CS-SVM, the subproblem can be solved exactly in $O(k \log k)$ time where k is the number of classes [CS01; Duc+08; BFU14; Con16]. However, for the WW-SVM, only iterative algorithms that approximate the subproblem minimizer have been proposed, and these lack runtime guarantees [Kee+08; IHG08].

In this work, we propose an algorithm called *Walrus*[1] that finds the exact solution of the Weston-Watkins subproblem in $O(k \log k)$ time. We implement Walrus in C++ inside the LIBLINEAR framework, yielding a new solver for the *linear* WW-SVM. For datasets with large number of classes, we demonstrate significant speed-up over the state-of-the-art linear solver Shark [IHG08]. We also rigorously prove the linear convergence of block coordinate descent for solving the dual problem of linear WW-SVM, confirming an assertion of Keerthi et al. [Kee+08].

[1] WW-subproblem analytic log-linear runtime solver

3.1.1 Related works

Existing literature on solving the optimization from SVMs largely fall into two categories: linear and kernel SVM solvers. The seminal work of Platt [Pla98] introduced the sequential minimal optimization (SMO) for solving kernel SVMs. Subsequently, many SMO-type algorithms were introduced which achieve faster convergence with theoretical guarantees [Kee+01; FCL05; SHS11; TAD21].

SMO can be thought of as a form of *(block) coordinate descent* where where the dual problem of the SVM optimization is decomposed into small subproblems. As such, SMO-type algorithms are also referred to as *decomposition methods*. For binary SVMs, the smallest subproblems are 1-dimensional and thus easy to solve exactly. However, for multiclass SVMs with k classes, the smallest subproblems are k-dimensional. Obtaining exact solutions for the subproblems is nontrivial.

Many works have studied the convergence properties of decomposition focusing on asymptotics [LS04], rates [CFL06; LS09], binary SVM without offsets [SHS11], and multiclass SVMs [HL02]. Another line of research focuses on primal convergence instead of the dual [Hus+06; LS07; Lis+07; BPS18].

Although kernel SVMs include linear SVMs as a special case, solvers specialized for linear SVMs can scale to larger data sets. Thus, linear SVM solvers are often developed separately. Hsieh et al. [Hsi+08] proposed using coordinate descent (CD) to solve the linear SVM dual problem and established linear convergence. Analogously, Keerthi et al. [Kee+08] proposed block coordinate descent (BCD) for multiclass SVMs. Coordinate descent on the dual problem is now used by the current state-of-the-art linear SVM solvers LIBLINEAR [Fan+08], liquidSVM [ST17], and Shark [IHG08].

There are other approaches to solving linear SVMs, e.g., using the cutting plane method [Joa06], and stochastic subgradient descent on the primal optimization [Sha+11]. However, these approaches do not converge as fast as CD on the dual

problem [Hsi+08].

For the CS-SVM introduced by Crammer et al. [CS01], an exact solver for the subproblem is well-known and there is a line of research on improving the solver's efficiency [CS01; Duc+08; BFU14; Con16]. For solving the kernel CS-SVM dual problem, convergence of an SMO-type algorithm was proven in [Lin02]. For solving the linear CS-SVM dual problem, linear convergence of coordinate descent was proven by Lee et al. [LC19]. Linear CS-SVMs with ℓ_1-regularizer have been studied by Babichev et al. [BOB19]

The Weston-Watkins SVM was introduced by Bredensteiner et al. [BB99], Weston et al. [WW99], and Vapnik [Vap98]. Empirical results from Doğan et al. [DGI16] suggest that the WW-SVM is the best performing multiclass SVMs among nine prominent variants. The WW-SVM loss function has also been successfully used in natural language processing by [SS21].

Hsu et al. [HL02] gave an SMO-type algorithm for solving the WW-SVM, although without convergence guarantees. Keerthi et al. [Kee+08] proposed using coordinate descent on the linear WW-SVM dual problem with an iterative subproblem solver. Furthermore, they asserted that the algorithm converges linearly, although no proof was given. The software Shark [IHG08] features a solver for the linear WW-SVM where the subproblem is approximately minimized by a greedy coordinate descent-type algorithm. MSVMpack [DL15] is a solver for multiclass SVMs which uses the Frank-Wolfe algorithm. The experiments of [BG16] showed that MSVMpack did not scale to larger number of classes for the WW-SVM. To our knowledge, an exact solver for the subproblem has not previously been developed.

3.1.2 Notations

Let n be a positive integer. Define $[n] := \{1, \dots, n\}$. All vectors are assumed to be column vectors unless stated otherwise. If $v \in \mathbb{R}^n$ is a vector and $i \in [n]$, we use

the notation $[v]_i$ to denote the i-th component of v. Let $\mathbb{1}_n$ and $\mathbb{0}_n \in \mathbb{R}^n$ denote the vectors of all ones and zeros, respectively. When the dimension n can be inferred from the context, we drop the subscript and simply write $\mathbb{1}$ and $\mathbb{0}$.

Let m be a positive integer. Matrices $\mathbf{w} \in \mathbb{R}^{m \times n}$ are denoted by boldface font. The (j, i)-th entry of $\mathbf{w}$ is denoted by w_{ji}. The columns of $\mathbf{w}$ are denoted by the same symbol $w_1, \ldots, w_n$ using regular font with a single subscript, i.e., $[w_i]_j = w_{ji}$. A column of $\mathbf{w}$ is sometimes referred to as a *block*. We will also use boldface Greek letter to denote matrices, e.g., $\boldsymbol{\alpha} \in \mathbb{R}^{m \times n}$ with columns $\alpha_1, \ldots, \alpha_n$.

The 2-norm of a vector v is denoted by $\|v\|$. The Frobenius norm of a matrix $\mathbf{w}$ is denoted by $\|\mathbf{w}\|_F$. The $m \times m$ identity and all-ones matrices are denoted by $\mathbf{I}_m$ and $\mathbf{O}_m$, respectively. When m is clear from the context, we drop the subscript and simply write $\mathbf{I}$ and $\mathbf{O}$.

For referencing, section numbers from our supplementary materials will be prefixed with an "A", e.g., Section 3.8.4.

3.2 Weston-Watkins linear SVM

Throughout this work, let $k \geq 2$ be an integer denoting the number of classes. Let $\{(x_i, y_i)\}_{i \in [n]}$ be a training dataset of size n where the instances $x_i \in \mathbb{R}^d$ and labels $y_i \in [k]$. The Weston-Watkins linear SVM [2] solves the optimization

$$\min_{\mathbf{w} \in \mathbb{R}^{d \times k}} \frac{1}{2}\|\mathbf{w}\|_F^2 + C \sum_{i=1}^{n} \sum_{\substack{j \in [k]: \\ j \neq y_i}} \texttt{hinge}(w'_{y_i} x_i - w'_j x_i) \tag{P}$$

where $\texttt{hinge}(t) = \max\{0, 1 - t\}$ and $C > 0$ is a hyperparameter.

Note that if an instance x_i is the zero vector, then for any $\mathbf{w} \in \mathbb{R}^{d \times k}$ we have $\texttt{hinge}(w'_{y_i} x_i - w'_j x_i) = 1$. Thus, we can simply ignore such an instance. Below, we

[2] Similar to other works on multiclass linear SVMs [HL02; Kee+08], the formulation eq. (P) does not use *offsets*. For discussions, see Section 3.7.

assume that $\|x_i\| > 0$ for all $i \in [n]$.

3.2.1 Dual of the linear SVM

In this section, we recall the dual of eq. (P). Derivation of all results here can be found in Hsu et al. [HL02] and Keerthi et al. [Kee+08].

We begin by defining the function $f : \mathbb{R}^{k \times n} \to \mathbb{R}$

$$f(\boldsymbol{\alpha}) := \frac{1}{2} \sum_{i,s \in [n]} x_s' x_i \alpha_i' \alpha_s - \sum_{i \in [n]} \sum_{\substack{j \in [k]: \\ j \neq y_i}} \alpha_{ij}$$

and the set

$$\begin{aligned} \mathcal{F} := \Big\{ \boldsymbol{\alpha} \in \mathbb{R}^{k \times n} \mid \\ & 0 \leq \alpha_{ij} \leq C, \, \forall i \in [n], \, j \in [k], \, j \neq y_i, \\ & \alpha_{iy_i} = - \sum_{j \in [k] \setminus \{y_i\}} \alpha_{ij}, \, \forall i \in [n] \Big\}. \end{aligned}$$

The dual problem

$$\min_{\boldsymbol{\alpha} \in \mathcal{F}} \; f(\boldsymbol{\alpha}). \tag{D1}$$

The primal and dual variables $\mathbf{w}$ and $\boldsymbol{\alpha}$ are related via

$$\mathbf{w} = - \sum_{i \in [n]} x_i \alpha_i'. \tag{3.1}$$

State-of-the-art solver Shark [IHG08] uses coordinate descent on the dual problem eq. (D1). It is also possible to solve the primal problem eq. (P) using stochastic gradient descent (SGD) as in Pegasos [Sha+11]. However, the empirical results of Hsieh et al. [Hsi+08] show that CD on the dual problem converges faster than SGD on the primal problem. Hence, we focus on the dual problem.

3.2.2 Solving the dual with block coordinate descent

Block coordinate descent (BCD) is an iterative algorithm for solving the dual problem eq. (D1) by repeatedly improving a candidate solution $\boldsymbol{\alpha} \in \mathcal{F}$. Given an $i \in [n]$, an *inner iteration* performs the update $\boldsymbol{\alpha} \leftarrow \widetilde{\boldsymbol{\alpha}}$ where $\widetilde{\boldsymbol{\alpha}}$ is a minimizer of the *i-th subproblem*:

$$\min_{\widehat{\boldsymbol{\alpha}} \in \mathcal{F}} \; f(\widehat{\boldsymbol{\alpha}}) \text{ such that } \; \widehat{\alpha}_s = \alpha_s, \, \forall s \in [n] \setminus \{i\}. \tag{S1}$$

An *outer iteration* performs the inner iteration once for each $i \in [n]$ possibly in a random order. By running several outer iterations, an (approximate) minimizer of eq. (D1) is putatively obtained.

Later, we will see that it is useful to keep track of $\mathbf{w}$ so that eq. (3.1) holds throughout the BCD algorithm. Suppose that $\boldsymbol{\alpha}$ and $\mathbf{w}$ satisfy eq. (3.1). Then $\mathbf{w}$ must be updated via

$$\mathbf{w} \leftarrow \mathbf{w} - x_i(\widetilde{\alpha}_i - \alpha_i)' \tag{3.2}$$

prior to updating $\boldsymbol{\alpha} \leftarrow \widetilde{\boldsymbol{\alpha}}$.

3.3 Reparametrization of the dual problem

In this section, we introduce a new way to parametrize the dual optimization eq. (D1) which allows us to derive an algorithm for finding the exact minimizer of eq. (S1).

Define the matrix $\boldsymbol{\pi} := \begin{bmatrix} \mathbb{1} & -\mathbf{I} \end{bmatrix} \in \mathbb{R}^{(k-1)\times k}$. For each $y \in [k]$, let $\tau_y \in \mathbb{R}^{k\times k}$ be the permutation matrix which switches the 1st and the yth indices. In other words,

given a vector $v \in \mathbb{R}^k$, we have

$$[\tau_y(v)]_j = \begin{cases} v_1 & : j = y \\ v_y & : j = 1 \\ v_j & : j \notin \{1, y\}. \end{cases}$$

Define the function $g : \mathbb{R}^{(k-1)\times n} \to \mathbb{R}$

$$g(\boldsymbol{\beta}) := \frac{1}{2} \sum_{i,s \in [n]} x_s' x_i \beta_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_s} \boldsymbol{\pi}' \beta_s - \sum_{i \in [n]} \mathbb{1}' \beta_i$$

and the set

$$\mathcal{G} := \Big\{ \boldsymbol{\beta} \in \mathbb{R}^{(k-1)\times n} \mid \\ 0 \leq \beta_{ij} \leq C, \, \forall i \in [n], \, j \in [k-1] \Big\}.$$

Consider the following optimization:

$$\min_{\boldsymbol{\beta} \in \mathcal{G}} \; g(\boldsymbol{\beta}). \tag{D2}$$

Up to a change of variables, the optimization eq. (D2) is equivalent to the dual of the linear WW-SVM eq. (D1). In other words, eq. (D2) is a reparametrization of eq. (D1). Below, we make this notion precise.

Definition III.1. Define a map $\Psi : \mathcal{G} \to \mathbb{R}^{k\times n}$ as follows: Given $\boldsymbol{\beta} \in \mathcal{G}$, construct an element $\Psi(\boldsymbol{\beta}) := \boldsymbol{\alpha} \in \mathbb{R}^{k\times n}$ whose i-th block is

$$\alpha_i = -\tau_{y_i} \boldsymbol{\pi}' \beta_i. \tag{3.3}$$

The map Ψ will serve as the change of variables map, where $\boldsymbol{\pi}$ reduces the dual

variable's dimension from k for α_i to $k-1$ for β_i. Furthermore, τ_{y_i} eliminates the dependency on y_i in the constraints. The following proposition shows that Ψ links the two optimization problems eq. (D1) and eq. (D2).

Proposition III.2. *The image of Ψ is $\mathcal{F}$, i.e., $\Psi(\mathcal{G}) = \mathcal{F}$. Furthermore, $\Psi : \mathcal{G} \to \mathcal{F}$ is a bijection and*

$$f(\Psi(\boldsymbol{\beta})) = g(\boldsymbol{\beta}).$$

Sketch of proof. Define another map $\Xi : \mathcal{F} \to \mathbb{R}^{(k-1)\times n}$ as follows: For each $\boldsymbol{\alpha} \in \mathcal{F}$, define $\boldsymbol{\beta} := \Xi(\boldsymbol{\alpha})$ block-wise by

$$\beta_i := \mathtt{proj}_{2:k}(\tau_{y_i}\alpha_i) \in \mathbb{R}^{k-1}$$

where

$$\mathtt{proj}_{2:k} = \begin{bmatrix} \mathbb{0} & \mathbf{I}_{k-1} \end{bmatrix} \in \mathbb{R}^{(k-1)\times k}.$$

Then the range of Ξ is in $\mathcal{G}$. Furthermore, Ξ and Ψ are inverses of each other. This proves that Ψ is a bijection. □

3.3.1 Reparametrized subproblem

Since the map Ψ respects the block-structure of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the result below follows immediately from proposition III.2:

Corollary III.3. *Let $\boldsymbol{\beta} \in \mathcal{G}$ and $i \in [n]$. Let $\boldsymbol{\alpha} = \Psi(\boldsymbol{\beta})$. Consider*

$$\min_{\widehat{\boldsymbol{\beta}} \in \mathcal{G}} \; g(\widehat{\boldsymbol{\beta}}) \text{ such that } \widehat{\beta}_s = \beta_s, \forall s \in [n] \setminus \{i\}. \tag{S2}$$

Let $\widetilde{\boldsymbol{\beta}} \in \mathcal{F}$ be arbitrary. Then $\widetilde{\boldsymbol{\beta}}$ is a minimizer of eq. (S2) *if and only if $\widetilde{\boldsymbol{\alpha}} := \Psi(\widetilde{\boldsymbol{\beta}})$ is a minimizer of eq.* (S1)*.*

Below, we focus on solving eq. (D2) with BCD, i.e., repeatedly performing the up-

date $\boldsymbol{\beta} \leftarrow \widetilde{\boldsymbol{\beta}}$ where $\widetilde{\boldsymbol{\beta}}$ is a minimizer of eq. (S2) over different $i \in [n]$. By corollary III.3, this is equivalent to solving eq. (D1) with BCD, up to the change of variables Ψ.

The reason we focus on solving eq. (D2) with BCD is because the subproblem can be cast in a simple form that makes an exact solver more apparent. To this end, we first show that the subproblem eq. (S2) is a quadratic program of a particular form. Define the matrix $\boldsymbol{\Theta} := \mathbf{I}_{k-1} + \mathbf{O}_{k-1}$.

Theorem III.4. *Let $v \in \mathbb{R}^{k-1}$ be arbitrary and $C > 0$. Consider the optimization*

$$\begin{aligned} \min_{b \in \mathbb{R}^{k-1}} \quad & \frac{1}{2} b' \boldsymbol{\Theta} b - v' b \\ s.t. \quad & 0 \leq b \leq C. \end{aligned} \tag{3.4}$$

Then algorithm 2, `solve_subproblem`(v, C)*, computes the unique minimizer of eq.* (3.4) *in $O(k \log k)$ time.*

We defer further discussion of theorem III.4 and algorithm 2 to the next section. The quadratic program eq. (3.4) is the *generic* form of the subproblem eq. (S2), as the following result shows:

Proposition III.5. *In the situation of corollary III.3, let $\widetilde{\beta}_i$ be the i-th block of the minimizer $\widetilde{\boldsymbol{\beta}}$ of eq.* (S2)*. Then $\widetilde{\beta}_i$ is the unique minimizer of eq.* (3.4) *with*

$$v := (\mathbb{1} - \boldsymbol{\pi} \tau_{y_i} \mathbf{w}' x_i) / \|x_i\|_2^2 + \boldsymbol{\Theta} \beta_i$$

and $\mathbf{w}$ *as in eq.* (3.1)*.*

3.3.2 BCD for the reparametrized dual problem

As mentioned in section 3.2.2, it is useful to keep track of $\mathbf{w}$ so that eq. (3.1) holds throughout the BCD algorithm. In proposition III.5, we see that $\mathbf{w}$ is used to

compute v. The update formula eq. (3.2) for $\mathbf{w}$ in terms of $\widetilde{\boldsymbol{\alpha}}$ can be cast in terms of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ by using eq. (3.3):

$$\mathbf{w} \leftarrow \mathbf{w} - x_i(\widetilde{\alpha}_i - \alpha_i)' = \mathbf{w} + x_i(\widetilde{\beta}_i - \beta_i)'\boldsymbol{\pi}\tau_{y_i}.$$

We now have all the ingredients to state the reparametrized block coordinate descent pseudocode in algorithm 1.

Algorithm 1 Block coordinate descent on eq. (D2)

```
β ← 0_{(k−1)×n}
w ← 0_{d×k}
while not converged do
    for i ← 1 to n do
        v ← (𝟙 − π τ_{y_i} w′ x_i)/‖x_i‖_2^2 + Θ β_i
        β̃_i ← solve_subproblem(v, C) (algorithm 2)
        w ← w + x_i(β̃_i − β_i)′ π τ_{y_i}
        β_i ← β̃_i
    end for
end while
```

Multiplying a vector by the matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ both only takes $O(k)$ time. Multiplying a vector by τ_{y_i} takes $O(1)$ time since τ_{t_i} simply swaps two entries of the vector. Hence, the speed bottlenecks of algorithm 1 are computing $\mathbf{w}'x_i$ and $x_i(\widetilde{\beta}_i - \beta_i)'$, both taking $O(dk)$ time and running `solve_subproblem`(v, C), which takes $O(k \log k)$ time. Overall, a single inner iteration of algorithm 1 takes $O(dk + k \log k)$ time. If x_i is s-sparse (only s entries are nonzero), then the iteration takes $O(sk + k \log k)$ time.

3.3.3 Linear convergence

Similar to the binary case [Hsi+08], BCD converges *linearly*, i.e., it produces an ϵ-accurate solution in $O(\log(1/\epsilon))$ outer iterations:

Theorem III.6. *algorithm 1 has global linear convergence. More precisely, let $\boldsymbol{\beta}^t$ be $\boldsymbol{\beta}$ at the end of the t-th iteration of the outer loop of algorithm 1. Let $g^* = \min_{\boldsymbol{\beta} \in \mathcal{G}} g(\boldsymbol{\beta})$.*

Then there exists $\Delta \in (0,1)$ *such that*

$$g(\boldsymbol{\beta}^{t+1}) - g^* \leq \Delta(g(\boldsymbol{\beta}^t) - g^*), \qquad \forall t = 0, 1, 2 \dots \tag{3.5}$$

where Δ *depends on the data* $\{(x_i, y_i)\}_{i \in [n]}$*,* k *and* C*.*

Luo et al. [LT92] proved asymptotic[3] linear convergence for cyclic coordinate descent for a certain class of minimization problems where the subproblem in each coordinate is *exactly* minimized. Furthermore, Luo et al. [LT92] claim that the same result holds if the subproblem is *approximately* minimized, but did not give a precise statement (e.g., approximation in which sense).

Keerthi et al. [Kee+08] asserted without proof that the results of Luo et al. [LT92] can be applied to BCD for WW-SVM. Possibly, no proof was given since no solver, exact nor approximate with approximation guarantees, was known at the time. theorem III.6 settles this issue, which we prove in Section 3.8.3 by extending the analysis of Luo et al. [LT92] and Wang et al. [WL14] to the multiclass case.

3.4 Sketch of proof of theorem III.4

Throughout this section, let $v \in \mathbb{R}^{k-1}$ and $C > 0$ be fixed. We first note that eq. (3.4) is a minimization of a strictly convex function over a compact domain, and hence has unique minimizer $\widetilde{b} \in \mathbb{R}^{k-1}$. Furthermore, it is the unique point satisfying the KKT conditions, which we present below. Our goal is to sketch the argument that algorithm 2 outputs the minimizer upon termination. The full proof can be found in Section 3.8.4.

[3] Asymptotic in the sense that eq. (3.5) is only guaranteed after $t > t_0$ for some unknown t_0.

3.4.1 Intuition

We first study the structure of the minimizer $\widetilde{b}$ in and of itself. The KKT conditions for a point $b \in \mathbb{R}^{k-1}$ to be optimal for eq. (3.4) are as follows:

$$\begin{aligned} &\forall i \in [k-1], \exists \lambda_i, \mu_i \in \mathbb{R} \text{ satisfying} \\ [(\mathbf{I}+\mathbf{O})b]_i + \lambda_i - \mu_i = v_i \quad &\text{stationarity} \\ C \geq b_i \geq 0 \quad &\text{primal feasibility} \\ \lambda_i \geq 0, \text{ and } \mu_i \geq 0 \quad &\text{dual feasibility} \\ \lambda_i(C - b_i) = 0, \text{ and } \mu_i b_i = 0 \quad &\text{complementary slackness} \end{aligned} \tag{KKT}$$

Below, let $\max_{i\in[k-1]} v_i =: v_{\max}$, and $\langle 1\rangle, \ldots, \langle k-1\rangle$ be an argsort of v, i.e., $v_{\langle 1\rangle} \geq \cdots \geq v_{\langle k-1\rangle}$.

Definition III.7. The *clipping map* $\texttt{clip}_C : \mathbb{R}^{k-1} \to [0, C]^{k-1}$ is the function defined as follows: for $w \in \mathbb{R}^{k-1}$, $[\texttt{clip}_C(w)]_i := \max\{0, \min\{C, w_i\}\}$.

Using the KKT conditions, we check that $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$ for some (unknown) $\widetilde{\gamma} \in \mathbb{R}$ and that $\widetilde{\gamma} = \mathbb{1}'\widetilde{b}$.

Proof. Let $\widetilde{\gamma} \in \mathbb{R}$ be such that $\mathbf{O}\widetilde{b} = \widetilde{\gamma}\mathbb{1}$. The stationarity condition can be rewritten as $\widetilde{b}_i + \lambda_i - \mu_i = v_i - \widetilde{\gamma}$. Thus, by complementary slackness and dual feasibility, we have

$$\widetilde{b}_i \begin{cases} \leq v_i - \widetilde{\gamma} & : \widetilde{b}_i = C \\ = v_i - \widetilde{\gamma} & : \widetilde{b}_i \in (0, C) \\ \geq v_i - \widetilde{\gamma} & : \widetilde{b}_i = 0 \end{cases}$$

Note that this is precisely $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$. □

For $\gamma \in \mathbb{R}$, let $b^\gamma := \texttt{clip}_C(v - \gamma\mathbb{1}) \in \mathbb{R}^{k-1}$. Thus, the $(k-1)$-dimensional vector $\widetilde{b}$ can be recovered from the scalar $\widetilde{\gamma}$ via $b^{\widetilde{\gamma}}$, reducing the search space from $\mathbb{R}^{k-1}$ to

$\mathbb{R}$.

However, the search space $\mathbb{R}$ is still a continuum. We show that the search space for $\widetilde{\gamma}$ can be further reduced to a finite set of candidates. To this end, let us define

$$I_{\mathtt{u}}^{\gamma} := \{i \in [k-1] : b_i^{\gamma} = C\}$$
$$I_{\mathtt{m}}^{\gamma} := \{i \in [k-1] : b_i^{\gamma} \in (0, C)\}.$$

Note that $I_{\mathtt{u}}^{\gamma}$ and $I_{\mathtt{m}}^{\gamma}$ are determined by their cardinalities, denoted $n_{\mathtt{u}}^{\gamma}$ and $n_{\mathtt{m}}^{\gamma}$, respectively. This is because

$$I_{\mathtt{u}}^{\gamma} = \{\langle 1\rangle, \langle 2\rangle, \ldots, \langle n_{\mathtt{u}}^{\gamma}\rangle\}$$
$$I_{\mathtt{m}}^{\gamma} = \{\langle n_{\mathtt{u}}^{\gamma}+1\rangle, \langle n_{\mathtt{u}}^{\gamma}+2\rangle, \ldots, \langle n_{\mathtt{u}}^{\gamma}+n_{\mathtt{m}}^{\gamma}\rangle\}.$$

Let $[\![k]\!] := \{0\} \cup [k-1]$. By definition, $n_{\mathtt{m}}^{\gamma}, n_{\mathtt{u}}^{\gamma} \in [\![k]\!]$. For $(n_{\mathtt{m}}, n_{\mathtt{u}}) \in [\![k]\!]^2$, define $S^{(n_{\mathtt{m}},n_{\mathtt{u}})}$, $\widehat{\gamma}^{(n_{\mathtt{m}},n_{\mathtt{u}})} \in \mathbb{R}$ by

$$S^{(n_{\mathtt{m}},n_{\mathtt{u}})} := \sum_{i=n_{\mathtt{u}}+1}^{n_{\mathtt{u}}+n_{\mathtt{m}}} v_{\langle i\rangle}, \tag{3.6}$$
$$\widehat{\gamma}^{(n_{\mathtt{m}},n_{\mathtt{u}})} := \left(C \cdot n_{\mathtt{u}} + S^{(n_{\mathtt{m}},n_{\mathtt{u}})}\right)/(n_{\mathtt{m}}+1). \tag{3.7}$$

Furthermore, define $\widehat{b}^{(n_{\mathtt{m}},n_{\mathtt{u}})} \in \mathbb{R}^{k-1}$ such that, for $i \in [k-1]$, the $\langle i\rangle$-th entry is

$$\widehat{b}_{\langle i\rangle}^{(n_{\mathtt{m}},n_{\mathtt{u}})} := \begin{cases} C & : i \le n_{\mathtt{u}} \\ v_{\langle i\rangle} - \gamma^{(n_{\mathtt{m}},n_{\mathtt{u}})} & : n_{\mathtt{u}} < i \le n_{\mathtt{u}} + n_{\mathtt{m}} \\ 0 & : n_{\mathtt{u}} + n_{\mathtt{m}} < i. \end{cases}$$

Using the KKT conditions, we check that

$$\widetilde{b} = \widehat{b}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} = \mathtt{clip}_C(v - \widehat{\gamma}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})}\mathbb{1}).$$

Proof. It suffices to prove that $\widetilde{\gamma} = \widehat{\gamma}^{(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})}$. To this end, let $i \in [k-1]$. If $i \in I_{\mathfrak{m}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = v_i - \widetilde{\gamma}$. If $i \in I_{\mathfrak{u}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = C$. Otherwise, $\widetilde{b}_i = 0$. Thus

$$\widetilde{\gamma} = \mathbb{1}'\widetilde{b} = C \cdot n_{\mathfrak{u}}^{\widetilde{\gamma}} + S^{(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})} - \widetilde{\gamma} \cdot n_{\mathfrak{m}}^{\widetilde{\gamma}} \tag{3.8}$$

Solving for $\widetilde{\gamma}$, we have

$$\widetilde{\gamma} = \left(C \cdot n_{\mathfrak{u}}^{\widetilde{\gamma}} + S^{(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})} \right) / (n_{\mathfrak{m}}^{\widetilde{\gamma}} + 1) = \widehat{\gamma}^{(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})},$$

as desired. □

Now, since $(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}}) \in \llbracket k \rrbracket^2$, to find $\widetilde{b}$ we can simply check for each $(n_{\mathfrak{m}}, n_{\mathfrak{u}}) \in \llbracket k \rrbracket^2$ if $\widehat{b}^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})}$ satisfies the KKT conditions. However, this naive approach leads to an $O(k^2)$ runtime.

To improve upon the naive approach, define

$$\mathfrak{R} := \{(n_{\mathfrak{m}}^{\gamma}, n_{\mathfrak{u}}^{\gamma}) : \gamma \in \mathbb{R}\}. \tag{3.9}$$

Since $(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}}) \in \mathfrak{R}$, to find $\widetilde{b}$ it suffices to search through $(n_{\mathfrak{m}}, n_{\mathfrak{u}}) \in \mathfrak{R}$ instead of $\llbracket k \rrbracket^2$. Towards enumerating all elements of $\mathfrak{R}$, a key result is that the function $\gamma \mapsto (I_{\mathfrak{m}}^{\gamma}, I_{\mathfrak{u}}^{\gamma})$ is locally constant outside of the set of discontinuities:

$$\texttt{disc} := \{v_i : i \in [k-1]\} \cup \{v_i - C : i \in [k-1]\}.$$

Proof. Let $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \mathbb{R}$ satisfy the following: 1) $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4$, 2) $\gamma_1, \gamma_4 \in$ `disc`, and 3) $\gamma \notin$ `disc` for all $\gamma \in (\gamma_1, \gamma_4)$. Assume for the sake of contradiction that $(I_{\mathfrak{m}}^{\gamma_2}, I_{\mathfrak{u}}^{\gamma_2}) \neq (I_{\mathfrak{m}}^{\gamma_3}, I_{\mathfrak{u}}^{\gamma_3})$. Then $I_{\mathfrak{m}}^{\gamma_2} \neq I_{\mathfrak{m}}^{\gamma_3}$ or $I_{\mathfrak{u}}^{\gamma_2} \neq I_{\mathfrak{u}}^{\gamma_3}$. Consider the case $I_{\mathfrak{m}}^{\gamma_2} \neq I_{\mathfrak{m}}^{\gamma_3}$. Then at least one of the sets $I_{\mathfrak{m}}^{\gamma_2} \setminus I_{\mathfrak{m}}^{\gamma_3}$ and $I_{\mathfrak{m}}^{\gamma_3} \setminus I_{\mathfrak{m}}^{\gamma_2}$ is nonempty. Consider the case when $I_{\mathfrak{m}}^{\gamma_2} \setminus I_{\mathfrak{m}}^{\gamma_3}$ is nonempty. Then there exists $i \in [k-1]$ such that $v_i - \gamma_2 \in (0, C)$

but $v_i - \gamma_3 \notin (0, C)$. This implies that there exists some $\gamma' \in (\gamma_2, \gamma_3)$ such that $v_i - \gamma' \in \{0, C\}$, or equivalently, $\gamma' \in \{v_i, v_i - C\}$. Hence, $\gamma' \in \texttt{disc}$, which is a contradiction. For the other cases not considered, similar arguments lead to the same contradiction. □

Thus, as we sweep γ from $+\infty$ to $-\infty$, we observe finitely many distinct tuples of sets $(I_{\mathfrak{m}}^{\gamma}, I_{\mathfrak{u}}^{\gamma})$ and their cardinalities $(n_{\mathfrak{m}}^{\gamma}, n_{\mathfrak{u}}^{\gamma})$. Using the index $t = 0, 1, 2 \ldots$, we keep track of these data in the variables $(I_{\mathfrak{m}}^{t}, I_{\mathfrak{u}}^{t})$ and $(n_{\mathfrak{m}}^{t}, n_{\mathfrak{u}}^{t})$. For this proof sketch, we make the assumption that $|\texttt{disc}| = 2(k-1)$, i.e., no elements are repeated.

By construction, the maximal element of $\texttt{disc}$ is $v_{\max}$. When $\gamma > v_{\max}$, we check that $n_{\mathfrak{m}}^{\gamma} = n_{\mathfrak{u}}^{\gamma} = \emptyset$. Thus, we put $I_{\mathfrak{m}}^{0} = I_{\mathfrak{u}}^{0} = \emptyset$ and $(n_{\mathfrak{m}}^{0}, n_{\mathfrak{u}}^{0}) = (0, 0)$.

Now, suppose γ has swept across $t-1$ points of discontinuity and that $I_{\mathfrak{m}}^{t-1}, I_{\mathfrak{u}}^{t-1}, n_{\mathfrak{m}}^{t-1}, n_{\mathfrak{u}}^{t-1}$ have all been defined. Suppose that γ crossed a single new point of discontinuity $\gamma' \in \texttt{disc}$. In other words, $\gamma'' < \gamma < \gamma'$ where γ'' is the largest element of $\texttt{disc}$ such that $\gamma'' < \gamma'$.

By the assumption that no elements of $\texttt{disc}$ are repeated, exactly one of the two following possibilities is true:

$$\text{there exists } i \in [k-1] \text{ such that } \gamma' = v_i, \tag{Entry}$$

$$\text{there exists } i \in [k-1] \text{ such that } \gamma' = v_i - C. \tag{Exit}$$

Under the eq. (Entry) case, the index i gets added to $I_{\mathfrak{m}}^{t-1}$ while $I_{\mathfrak{u}}^{t-1}$ remains unchanged. Hence, we have the updates

$$I_{\mathfrak{m}}^{t} := I_{\mathfrak{m}}^{\gamma} = I_{\mathfrak{m}}^{t-1} \cup \{i\}, \quad I_{\mathfrak{u}}^{t} := I_{\mathfrak{u}}^{\gamma} = I_{\mathfrak{u}}^{t-1} \tag{3.10}$$

$$n_{\mathfrak{m}}^{t} := n_{\mathfrak{m}}^{\gamma} = n_{\mathfrak{m}}^{t-1} + 1, \quad n_{\mathfrak{u}}^{t} := n_{\mathfrak{u}}^{\gamma} = n_{\mathfrak{u}}^{t-1}. \tag{3.11}$$

Under the eq. (Exit) case, the index i moves from $I_{\mathfrak{m}}^{t-1}$ to $I_{\mathfrak{u}}^{t-1}$. Hence, we have the

updates

$$I^t_{\mathfrak{m}} := I^\gamma_{\mathfrak{m}} = I^{t-1}_{\mathfrak{m}} \setminus \{i\}, \quad I^t_{\mathfrak{u}} := I^\gamma_{\mathfrak{u}} = I^{t-1}_{\mathfrak{u}} \cup \{i\} \tag{3.12}$$

$$n^t_{\mathfrak{m}} := n^\gamma_{\mathfrak{m}} = n^{t-1}_{\mathfrak{m}} - 1, \quad n^t_{\mathfrak{u}} := n^\gamma_{\mathfrak{u}} = n^{t-1}_{\mathfrak{u}} + 1. \tag{3.13}$$

Thus, $\{(n^t_{\mathfrak{m}}, n^t_{\mathfrak{u}})\}_{t=0}^{2(k-1)} = \Re$. The case when `disc` has repeated elements requires more careful analysis which is done in the full proof. Now, we have all the ingredients for understanding algorithm 2 and its subroutines.

3.4.2 A walk through of the solver

If $v_{\max} \leq 0$, then $\widetilde{b} = \mathbb{0}$ satisfies the KKT conditions. algorithm 2-line 3 handles this exceptional case. Below, we assume $v_{\max} > 0$.

Algorithm 2 `solve_subproblem`(v, C)

1: **Input:** $v \in \mathbb{R}^{k-1}$
2: Let $\langle 1\rangle, \ldots, \langle k-1\rangle$ sort v, i.e., $v_{\langle 1\rangle} \geq \cdots \geq v_{\langle k-1\rangle}$.
3: **if** $v_{\langle 1\rangle} \leq 0$ **then HALT and output:** $\mathbb{0} \in \mathbb{R}^{k-1}$.
4: $n^0_{\mathfrak{u}} := 0$, $n^0_{\mathfrak{m}} := 0$, $S^0 := 0$
5: $(\delta_1, \ldots, \delta_\ell) \leftarrow$ `get_up_dn_seq`() (Subroutine 3)
6: **for** $t = 1, \ldots, \ell$ **do**
7: $(n^t_{\mathfrak{m}},\, n^t_{\mathfrak{u}},\, S^t) \leftarrow$ `update_vars`() (Subroutine 4).
8: $\widehat{\gamma}^t := (C \cdot n^t_{\mathfrak{u}} + S^t)/(n^t_{\mathfrak{m}} + 1)$
9: **if** `KKT_cond`() (Subroutine 5) returns true **then**
10: **HALT and output**: $\widehat{b}^t \in \mathbb{R}^{k-1}$ where

$$\widehat{b}^t_{\langle i\rangle} := \begin{cases} C & : i \leq n^t_{\mathfrak{u}} \\ v_{\langle i\rangle} - \gamma^t & : n^t_{\mathfrak{u}} < i \leq n^t_{\mathfrak{u}} + n^t_{\mathfrak{m}} \\ 0 & : n^t_{\mathfrak{u}} + n^t_{\mathfrak{m}} < i. \end{cases}$$

11: **end if**
12: **end for**

algorithm 2-line 4 initializes the state variables $n^t_{\mathfrak{m}}$ and $n^t_{\mathfrak{u}}$ as discussed in the last section. The variable S^t is also initialized and will be updated to maintain

$S^t = S^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$ where the latter is defined at eq. (3.6).

algorithm 2-line 5 calls Subroutine 3 to construct the `vals` ordered set, which is similar to the set of discontinuities `disc`, but different in three ways: 1) `vals` consists of tuples (γ', δ') where $\gamma' \in$ `disc` and $\delta' \in \{\mathtt{up}, \mathtt{dn}\}$ is a decision variable indicating whether γ' satisfies the eq. (Entry) or the eq. (Exit) condition, 2) `vals` is sorted so that the γ's are in descending order, and 3) only positive values of `disc` are needed. The justification for the third difference is because we prove that algorithm 2 always halts before reaching the negative values of `disc`. Subroutine 3 returns the list of symbols $(\delta_1, \ldots, \delta_\ell)$ consistent with the ordering.

Subroutine 3 `get_up_dn_seq` *Note: all variables from algorithm 2 are assumed to be visible here.*

1: `vals` $\leftarrow \{(v_i, \mathtt{dn}) : v_i > 0, i = 1, \ldots, k-1\} \cup \{(v_i - C, \mathtt{up}) : v_i > C, i = 1, \ldots, k-1\}$ as a multiset, where elements may be repeated.
2: Order the set `vals` $= \{(\gamma_1, \delta_1), \ldots, (\gamma_\ell, \delta_\ell)\}$ such that $\gamma_1 \geq \cdots \geq \gamma_\ell$, $\ell = |\mathtt{vals}|$, and for all $j_1, j_2 \in [\ell]$ such that $j_1 < j_2$ and $\gamma_{j_1} = \gamma_{j_2}$, we have $\delta_{j_1} = \mathtt{dn}$ implies $\delta_{j_2} = \mathtt{dn}$.
Note that by construction, for each $t \in [\ell]$, there exists $i \in [k-1]$ such that $\gamma_t = v_i$ or $\gamma_t = v_i - C$.
3: **Output:** sequence $(\delta_1, \ldots, \delta_\ell)$ whose elements are retrieved in order from left to right.

In the "for" loop, algorithm 2-line 7 calls Subroutine 4 which updates the variables $n_{\mathtt{m}}^t, n_{\mathtt{u}}^t$ using eq. (3.11) or eq. (3.13), depending on δ_t. The variable S^t is updated accordingly so that $S^t = S^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$.

Subroutine 4 `update_vars` *Note: all variables from algorithm 2 are assumed to be visible here.*

1: **if** $\delta_t = \mathtt{up}$ **then**
2: $\quad n_{\mathtt{u}}^t := n_{\mathtt{u}}^{t-1} + 1, \quad n_{\mathtt{m}}^t := n_{\mathtt{m}}^{t-1} - 1$
3: $\quad S^t := S^{t-1} - v_{\langle n_{\mathtt{u}}^{t-1} \rangle}$
4: **else**
5: $\quad n_{\mathtt{m}}^t := n_{\mathtt{m}}^{t-1} + 1, \quad n_{\mathtt{u}}^t := n_{\mathtt{u}}^{t-1}.$
6: $\quad S^t := S^{t-1} + v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \rangle}$
7: **end if**
8: **Output:** $(n_{\mathtt{m}}^t,\ n_{\mathtt{u}}^t,\ S^t)$

We skip to algorithm 2-line 9 which constructs the putative solution $\widehat{b}^t$. Observe that $\widehat{b}^t = \widehat{b}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$ where the latter is defined in the previous section.

Going back one line, algorithm 2-line 8 calls Subroutine 5 which checks if the putative solution $\widehat{b}^t$ satisfies the KKT conditions. We note that this can be done *before* the putative solution is constructed.

Subroutine 5 `KKT_cond` *Note: all variables from algorithm 2 are assumed to be visible here.*

`kkt_cond` $\leftarrow$ *true*
if $n_{\mathtt{u}}^t > 0$ **then**
    `kkt_cond` $\leftarrow$ `kkt_cond` $\wedge \left(C + \widehat{\gamma}^t \leq v_{\langle n_{\mathtt{u}}^t \rangle}\right)$
    Note: $\wedge$ denotes the logical "and".
end if
if $n_{\mathtt{m}}^t > 0$ **then**
    `kkt_cond` $\leftarrow$ `kkt_cond` $\wedge \left(v_{\langle n_{\mathtt{u}}^t + 1 \rangle} \leq C + \widehat{\gamma}^t\right)$
    `kkt_cond` $\leftarrow$ `kkt_cond` $\wedge \left(\widehat{\gamma}^t \leq v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \rangle}\right)$
end if
if $n_{\mathfrak{d}}^t := k - 1 - n_{\mathtt{u}}^t - n_{\mathtt{m}}^t > 0$ **then**
    `kkt_cond` $\leftarrow$ `kkt_cond` $\wedge \left(v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t + 1 \rangle} \leq \widehat{\gamma}^t\right)$
end if
Output: `kkt_cond`

For the runtime analysis, we note that Subroutines 5 and 4 both use $O(1)$ FLOPs without dependency on k. The main "for" loop of algorithm 2 (line 6 through 11) has $O(\ell)$ runtime where $\ell \leq 2(k-1)$. Thus, the bottlenecks are algorithm 2-line 2 and 5 which sort lists of length at most $k-1$ and $2(k-1)$, respectively. Thus, both lines run in $O(k \log k)$ time.

3.5 Experiments

LIBLINEAR is one of the state-of-the-art solver for linear SVMs [Fan+08]. However, as of the latest version 2.42, the linear Weston-Watkins SVM is not supported. We implemented our linear WW-SVM subproblem solver, *Walrus* (algorithm 2), along with the BCD algorithm 1 as an extension to LIBLINEAR. The solver and

Table 3.1: Data sets used. Variables k, n and d are, respectively, the number of classes, training samples, and features.

DATA SET	k	n	d
DNA	3	2,000	180
SATIMAGE	6	4,435	36
MNIST	10	60,000	780
NEWS20	20	15,935	62,061
LETTER	26	15,000	16
RCV1	53	15,564	47,236
SECTOR	105	6,412	55,197
ALOI	1,000	81,000	128

code for generating the figures are available[4].

We compare our implementation to *Shark* [IHG08], which solves the dual subproblem eq. (S1) using a form of greedy coordinate descent. For comparisons, we reimplemented Shark's solver also as a LIBLINEAR extension. When clear from the context, we use the terms "Walrus" and "Shark" when referring to either the subproblem solver or the overall BCD algorithm.

We perform benchmark experiments on 8 datasets from "LIBSVM Data: Classification (Multi-class)[5]" spanning a range of k from 3 to 1000. See table 3.1.

In all of our experiments, Walrus and Shark perform identically in terms of testing accuracy. We report the accuracies in Section 3.9.3. Below, we will only discuss runtime.

For measuring the runtime, we start the timer after the data sets have been loaded into memory and before the state variables $\boldsymbol{\beta}$ and $\mathbf{w}$ have been allocated. The primal objective is the value of eq. (P) at the current $\mathbf{w}$ and the dual objective is -1 times the value of eq. (D2) at the current $\boldsymbol{\beta}$. The duality gap is the primal minus the dual objective. The objective values and duality gaps are measured after each *outer* iteration, during which the timer is paused.

[4]See Section 3.9.
[5]See Section 3.9.2.

For solving the subproblem, Walrus is guaranteed to return the minimizer in $O(k \log k)$ time. On the other hand, to the best of our knowledge, Shark does not have such guarantee. Furthermore, Shark uses a doubly-nested for loop, each of which has length $O(k)$, yielding a worst-case runtime of $O(k^2)$. For these reasons, we hypothesize that Walrus scales better with larger k.

As exploratory analysis, we ran Walrus and Shark on the SATIMAGE and SECTOR data sets[6], which has 6 and 105 classes, respectively. The results, shown in fig. 3.1, support our hypothesis: Walrus and Shark are equally fast for SATIMAGE while Walrus is faster for SECTOR.

We test our hypothesis on a larger scale by running Walrus and Shark on the datasets in table 3.1 over the grid of hyperparameters $C \in \{2^{-6}, 2^{-5}, \ldots, 2^2, 2^3\}$. The results are shown in fig. 3.2 where each dot represents a triplet (DATA SET, C, δ) where δ is a quantity we refer to as the *duality gap decay*. The Y-axis shows the comparative metric of runtime $\texttt{ET}^{\delta}_{\texttt{Walrus}}/\texttt{ET}^{\delta}_{\texttt{Shark}}$ to be defined next.

Consider a single run of Walrus on a fixed data set with a given hyperparameter C. Let $\texttt{DG}^{t}_{\texttt{Walrus}}$ denote the duality gap achieved by Walrus at the end of the t-th outer iteration. Let $\delta \in (0, 1)$. Define $\texttt{ET}^{\delta}_{\texttt{Walrus}}$ to be the elapsed time at the end of the t-th iteration where t is minimal such that $\texttt{DG}^{t}_{\texttt{Walrus}} \leq \delta \cdot \texttt{DG}^{1}_{\texttt{Walrus}}$. Define $\texttt{DG}^{t}_{\texttt{Shark}}$ and $\texttt{ET}^{\delta}_{\texttt{Shark}}$ similarly. In all experiments $\texttt{DG}^{1}_{\texttt{Walrus}}/\texttt{DG}^{1}_{\texttt{Shark}} \in [0.99999, 1.00001]$. Thus, the ratio $\texttt{ET}^{\delta}_{\texttt{Walrus}}/\texttt{ET}^{\delta}_{\texttt{Shark}}$ measures how much faster Shark is relative to Walrus.

From fig. 3.2, it is evident that in general Walrus converges faster on data sets with larger number of classes. Not only does Walrus beat Shark for large k, but it also seems to not do much worse for small k. In fact Walrus seems to be at least as fast as Shark for all datasets except SATIMAGE.

The absolute amount of time saved by Walrus is often more significant on datasets with larger number of classes. To illustrate this, we let $C = 1$ and compare the times

[6]The regularizers are set to the corresponding values from Table 5 of the supplementary material of Doğan et al. [DGI16] chosen by cross-validation.

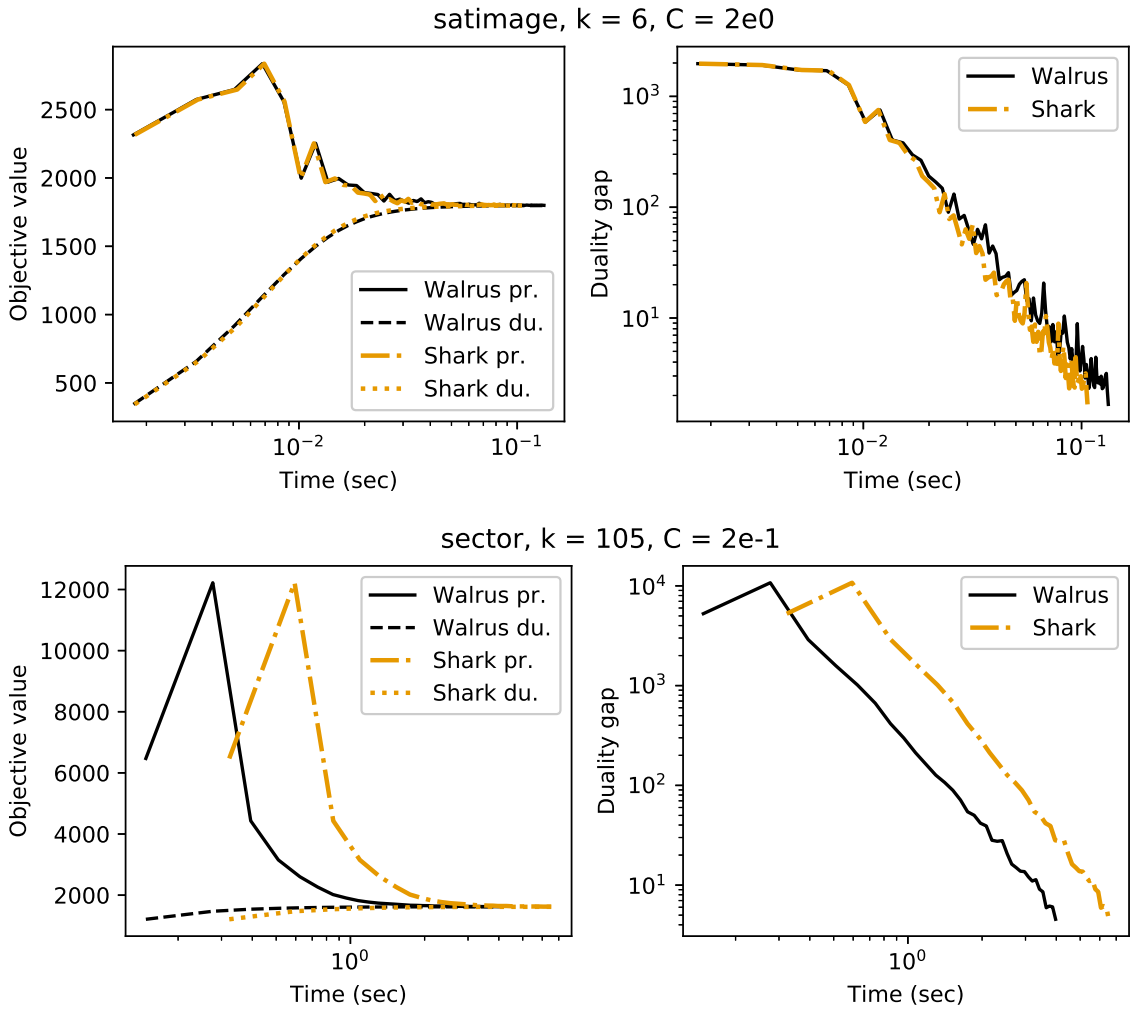

Figure 3.1: Runtime comparison of Walrus and Shark. Abbreviations: pr. = primal and du. = dual. The X-axes show time elapsed.

for the duality gap to decay by a factor of 0.01. On the data set SATIMAGE with $k = 6$, Walrus and Shark take 0.0476 and 0.0408 seconds, respectively. On the data set ALOI with $k = 1000$, Walrus and Shark take 188 and 393 seconds, respectively.

We remark that fig. 3.2 also suggests that Walrus tends to be faster during early iterations but can be slower at late stages of the optimization. To explain this phenomenon, we note that Shark solves the subproblem using an iterative descent algorithm and is set to stop when the KKT violations fall below a hard-coded threshold. When close to optimality, Shark takes fewer descent steps, and hence less time, to reach the stopping condition on the subproblems. On the other hand, Walrus takes the same amount of time regardless of proximity to optimality.

For the purpose of grid search, a high degree of optimality is not needed. In Section 3.9.3, we provide empirical evidence that stopping early versus late does not change the result of grid search-based hyperparameter tuning. Specifically, table 3.7 shows that running the solvers until $\delta \approx 0.01$ or until $\delta \approx 0.001$ does not change the cross-validation outcomes.

Finally, the optimization eq. (3.4) is a convex quadratic program and hence can be solved using general-purpose solvers [VL04]. However, we find that Walrus, being specifically tailored to the optimization eq. (3.4), is orders of magnitude faster. See tables 3.9 and 3.10 in the Appendix.

3.6 Discussions and future works

We presented an algorithm called Walrus for exactly solving the WW-subproblem which scales with the number of classes. We implemented Walrus in the LIBLINEAR framework and demonstrated empirically that BCD using Walrus is significantly faster than state-of-the-art linear WW-SVM solver Shark on datasets with a large number of classes, and comparable to Shark for small number of classes.

One possible direction for future research is whether Walrus can improve *kernel*

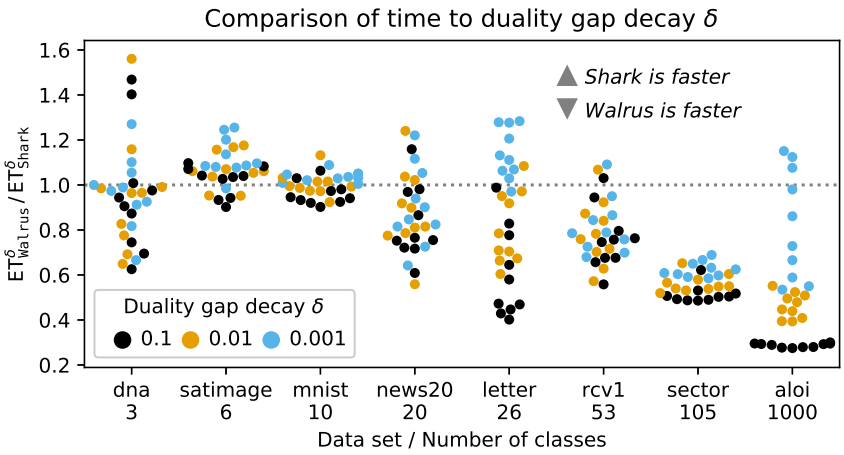

Figure 3.2: X-coordinates jittered for better visualization.

WW-SVM solver. Another direction is lower-bounding time complexity of solving the WW-subproblem eq. (3.4).

3.7 Regarding offsets

In this section, we review the literature on SVMs in particular with regard to offsets.

For binary kernel SVMs, Steinwart et al. [SHS11] demonstrates that kernel SVMs without offset achieve comparable classification accuracy as kernel SVMs with offset. Furthermore, they propose algorithms that solve kernel SVMs without offset that are significantly faster than solvers for kernel SVMs with offset.

For binary linear SVMs, Hsieh et al. [Hsi+08] introduced coordinate descent for the dual problem associated to linear SVMs without offsets, or with the bias term included in the $\mathbf{w}$ term. Chiu et al. [CLL20] studied whether the method of Hsieh et al. [Hsi+08] can be extended to allow offsets, but found evidence that the answer is negative. For multiclass linear SVMs, Keerthi et al. [Kee+08] studied block coordinate descent for the CS-SVM and WW-SVM, both without offsets. We are not aware of a multiclass analogue to Chiu et al. [CLL20] although the situation should be similar.

The previous paragraph discussed coordinate descent in relation to the offset. Including the offset presents challenges to primal methods as well. In Section 6 of Shalev-Shwartz et al. [Sha+11], the authors argue that including an unregularized offset term in the primal objective leads to slower convergence guarantee. Furthermore, Shalev-Shwartz et al. [Sha+11] observed that including an unregularized offset did not significantly change the classification accuracy.

The original Crammer-Singer (CS) SVM was proposed without offsets [CS01]. In Section VI of [HL02], the authors show the CS-SVM with offsets do *not* perform better than CS-SVM without offsets. Furthermore, CS-SVM with offsets requires twice as many iterations to converge than without.

3.8 Omitted proofs

3.8.1 Proof of proposition III.2

Below, let $i \in [n]$ be arbitrary. First, we note that $-\boldsymbol{\pi}' = \begin{bmatrix} -\mathbb{1}' \\ \mathbf{I}_{k-1} \end{bmatrix}$ and so

$$\boldsymbol{\pi}'\beta_i = \begin{bmatrix} -\mathbb{1}'\beta_i \\ \beta_i \end{bmatrix}. \tag{3.14}$$

Now, let $j \in [k]$, we have by eq. (3.3) that

$$[\alpha_i]_j = [-\tau_{y_i}\boldsymbol{\pi}'\beta_i]_j = [-\boldsymbol{\pi}'\beta_i]_{\tau_{y_i}(j)}. \tag{3.15}$$

Note that if $j \neq y_i$, then $\tau_{y_i}(j) \neq 1$ and so $[\alpha_i]_j = [-\boldsymbol{\pi}'\beta_i]_{\tau_{y_i}(j)} = [\beta_i]_{\tau_{y_i}(j)-1} \in [0, C]$. On the other hand, if $j = y_i$, then $\tau_{y_i}(y_i) = 1$ and $[\alpha_i]_{y_i} = [-\boldsymbol{\pi}'\beta_i]_1 = -\mathbb{1}'\beta_i = -\sum_{t\in[k-1]}[\beta_i]_t = -\sum_{t\in[k]:t\neq y_i}[\beta_i]_{\tau_{y_i}(t)-1} = -\sum_{t\in[k]:t\neq y_i}[\alpha_i]_t$. Thus, $\boldsymbol{\alpha} \in \mathcal{F}$. This proves that $\Psi(\mathcal{G}) \subseteq \mathcal{F}$.

Next, let us define another map $\Xi : \mathcal{F} \to \mathbb{R}^{(k-1)\times n}$ as follows: For each $\boldsymbol{\alpha} \in \mathcal{F}$, define $\boldsymbol{\beta} := \Xi(\boldsymbol{\alpha})$ block-wise by

$$\beta_i := \mathtt{proj}_{2:k}(\tau_{y_i}\alpha_i) \in \mathbb{R}^{k-1}$$

where

$$\mathtt{proj}_{2:k} = \begin{bmatrix} \mathbb{0} & \mathbf{I}_{k-1} \end{bmatrix} \in \mathbb{R}^{(k-1)\times k}.$$

By construction, we have for each $j \in [k-1]$ that $[\beta_i]_j = [\tau_{y_i}\alpha_i]_{j+1} = [\tau_{y_i}\alpha_i]_{j+1} = [\alpha_i]_{\tau_{y_i}(j+1)}$ Since $j+1 \neq 1$ for any $j \in [k-1]$, we have that $\tau_{y_i}(j+1) \neq y_i$ for any $j \in [k-1]$. Thus, $[\beta_i]_j = [\alpha_i]_{\tau_{y_i}(j+1)} \in [0, C]$. This proves that $\Xi(\mathcal{F}) \subseteq \mathcal{G}$.

Next, we prove that for all $\boldsymbol{\alpha} \in \mathcal{F}$ and $\boldsymbol{\beta} \in \mathcal{G}$, we have $\Xi(\Psi(\boldsymbol{\beta})) = \boldsymbol{\beta}$ and $\Psi(\Xi(\boldsymbol{\alpha})) = \boldsymbol{\alpha}$.

By construction, the i-th block of $\Xi(\Psi(\boldsymbol{\beta}))$ is given by

$$\begin{aligned}
\mathtt{proj}_{2:k}(\tau_{y_i}(-\tau_{y_i}\boldsymbol{\pi}'\beta_i)) &= -\mathtt{proj}_{2:k}(\tau_{y_i}\tau_{y_i}\boldsymbol{\pi}'\beta_i) \\
&= -\mathtt{proj}_{2:k}(\boldsymbol{\pi}'\beta_i) \\
&= -\begin{bmatrix} \mathbb{0} & \mathbf{I}_{k-1} \end{bmatrix} \begin{bmatrix} \mathbb{1}' \\ -\mathbf{I}_{k-1} \end{bmatrix} \beta_i \\
&= \mathbf{I}_{k-1}\beta_i = \beta_i.
\end{aligned}$$

For the second equality, we used the fact that $\tau_y^2 = \mathbf{I}$ for all $y \in [k]$. Thus, $\Xi(\Psi(\boldsymbol{\beta})) = \boldsymbol{\beta}$.

Next, note that the i-th block of $\Psi(\Xi(\boldsymbol{\alpha}))$ is, by construciton,

$$-\tau_{y_i}\boldsymbol{\pi}'\mathtt{proj}_{2:k}(\tau_{y_i}\alpha_i) = -\tau_{y_i}\pi' \begin{bmatrix} \mathbb{0} & \mathbf{I}_{k-1} \end{bmatrix} \tau_{y_i}\alpha_i = -\tau_{y_i} \begin{bmatrix} \mathbb{0} & \pi' \end{bmatrix} \tau_{y_i}\alpha_i \tag{3.16}$$

Recall that $\pi' = \begin{bmatrix} \mathbb{1}' \\ -\mathbf{I}_{k-1} \end{bmatrix}$ and so $\begin{bmatrix} \mathbb{0} & \pi' \end{bmatrix} = \begin{bmatrix} 0 & \mathbb{1}' \\ \mathbb{0} & -\mathbf{I}_{k-1} \end{bmatrix}$. Therefore,

$$\left[\begin{bmatrix} \mathbb{0} & \pi' \end{bmatrix} \tau_{y_i} \alpha_i \right]_1 = \sum_{j=2}^{k} [\tau_{y_i} \alpha_i] = \sum_{j \in [k]: j \neq y_i} [\alpha_i]_j = -[\alpha_i]_{y_i} = -[\tau_{y_i} \alpha_i]_1$$

and, for $j = 2, \ldots, k$,

$$\left[\begin{bmatrix} \mathbb{0} & \pi' \end{bmatrix} \tau_{y_i} \alpha_i \right]_j = -[\tau_{y_i} \alpha_i]_j.$$

Hence, we have just shown that $\begin{bmatrix} \mathbb{0} & \pi' \end{bmatrix} \tau_{y_i} \alpha_i = -\tau_{y_i} \alpha_i$. Continuing from eq. (3.16), we have

$$-\tau_{y_i} \boldsymbol{\pi}' \mathtt{proj}_{2:k}(\tau_{y_i} \alpha_i) = -\tau_{y_i}(-\tau_{y_i} \alpha_i) = \tau_{y_i} \tau_{y_i} \alpha_i = \alpha_i.$$

This proves that $\Psi(\Xi(\boldsymbol{\alpha})) = \boldsymbol{\alpha}$. Thus, we have shown that Ψ and Ξ are inverses of one another. This proves that Ψ is a bijection.

Finally, we prove that

$$f(\Psi(\boldsymbol{\beta})) = g(\boldsymbol{\beta}).$$

Recall that

$$f(\boldsymbol{\alpha}) := \frac{1}{2} \sum_{i,s \in [n]} x_s' x_i \alpha_i' \alpha_s - \sum_{i \in [k]} \sum_{\substack{j \in [k]: \\ j \neq y_i}} \alpha_{ij}$$

Thus,

$$\alpha_i' \alpha_s = (-\tau_{y_i} \boldsymbol{\pi}' \beta_i)'(-\tau_{y_s} \boldsymbol{\pi}' \beta_s) = \beta_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_s}' \boldsymbol{\pi}' \beta_s$$

On the other hand, eq. (3.3) implies that $\tau_{y_i} \alpha_i = -\boldsymbol{\pi}' \beta_i$. Hence

$$\sum_{j \in [k] \setminus \{y_i\}} \alpha_{ij} = \sum_{j \in [k]: j \neq 1} [\alpha_i]_{\tau_{y_i}(j)} = \sum_{j \in [k]: j \neq 1} [\tau_{y_i} \alpha_i]_j = \sum_{j \in [k]: j \neq 1} [-\boldsymbol{\pi}' \beta_i]_j = \sum_{j \in [k-1]} [\beta_i]_j = \mathbb{1}' \beta_i.$$

Thus,

$$f(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{i,s\in[n]} x_s' x_i \alpha_i' \alpha_s - \sum_{i\in[k]}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \alpha_{ij} = \frac{1}{2}\sum_{i,s\in[n]} x_s' x_i \beta_i' \boldsymbol{\pi}\tau_{y_i}\tau_{y_s}'\boldsymbol{\pi}'\beta_s - \sum_{i\in[k]} \mathbb{1}'\beta_i = g(\boldsymbol{\beta})$$

as desired. Finally, we note that $\tau_y = \tau_y'$ for all $y \in [k]$. This concludes the proof of proposition III.2. □

3.8.2 Proof of proposition III.5

We prove the following lemma which essentially unpacks the succinct proposition III.5:

Lemma III.8. *Recall the situation of corollary III.3: Let $\boldsymbol{\beta} \in \mathcal{G}$ and $i \in [n]$. Let $\boldsymbol{\alpha} = \Psi(\boldsymbol{\beta})$. Consider*

$$\min_{\widehat{\boldsymbol{\beta}}\in\mathcal{G}} g(\widehat{\boldsymbol{\beta}}) \text{ such that } \widehat{\beta}_s = \beta_s,\ \forall s \in [n]\setminus\{i\}. \tag{3.17}$$

Let $\mathbf{w}$ be as in eq. (3.1), i.e., $\mathbf{w} = -\sum_{i\in[n]} x_i\alpha_i'$. Then a solution to eq. (3.17) is given by $[\beta_1, \ldots, \beta_{i-1}, \widetilde{\beta}_i, \beta_{i+1}, \ldots, \beta_n]$ where $\widetilde{\beta}_i$ is a minimizer of

$$\min_{\widehat{\beta}_i\in\mathbb{R}^{k-1}} \frac{1}{2}\widehat{\beta}_i'\boldsymbol{\Theta}\widehat{\beta}_i - \widehat{\beta}_i'\left((\mathbb{1} - \boldsymbol{\pi}\tau_{y_i}\mathbf{w}'x_i)/\|x_i\|_2^2 + \boldsymbol{\Theta}\beta_i\right) \text{ such that } 0 \le \widehat{\beta}_i \le C.$$

Furthermore, the above optimization has a unique minimizer which is equal to the minimizer of eq. (3.4) where

$$v := (\mathbb{1} - \boldsymbol{\rho}_{y_i}\boldsymbol{\pi}\mathbf{w}'x_i + \boldsymbol{\Theta}\beta_i\|x_i\|_2^2)/\|x_i\|_2^2$$

and $\mathbf{w}$ is as in eq. (3.1).

Proof. First, we prove a simple identity:

$$\boldsymbol{\pi}\boldsymbol{\pi}' = \begin{bmatrix} \mathbb{1} & -\mathbf{I}_{k-1} \end{bmatrix} \begin{bmatrix} \mathbb{1}' \\ -\mathbf{I}_{k-1} \end{bmatrix} = \mathbf{I} + \mathbf{O} = \boldsymbol{\Theta}. \tag{3.18}$$

Next, recall that by definition, we have

$$g(\boldsymbol{\beta}) := \left(\frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \tau_{y_t} \tau_{y_s} \boldsymbol{\pi}' \beta_s \right) - \left(\sum_{s \in [n]} \mathbb{1}' \beta_s \right).$$

Let us group the terms of $g(\boldsymbol{\beta})$ that depends on β_i:

$$\begin{aligned}
g(\boldsymbol{\beta}) &= \frac{1}{2} x_i' x_i \beta_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_i} \boldsymbol{\pi}' \beta_i \\
&\quad + \frac{1}{2} \sum_{s \in [n]: s \neq i} x_s' x_i \beta_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_s} \boldsymbol{\pi}' \beta_s \\
&\quad + \frac{1}{2} \sum_{t \in [n]: t \neq i} x_i' x_t \beta_t' \boldsymbol{\pi} \tau_{y_t} \tau_{y_i} \boldsymbol{\pi}' \beta_i \\
&\quad + \frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \tau_{y_t} \tau_{y_s} \boldsymbol{\pi}' \beta_s - \sum_{s \in [n]} \mathbb{1}' \beta_s \\
&= \frac{1}{2} x_i' x_i \beta_i' \boldsymbol{\Theta} \beta_i \qquad \because \tau_{y_i}^2 = \mathbf{I} \text{ and } eq.\ (3.18) \\
&\quad + \sum_{s \in [n]: s \neq i} x_s' x_i \beta_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_s} \boldsymbol{\pi}' \beta_s \\
&\quad - \mathbb{1}' \beta_i \\
&\quad + \underbrace{\frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \tau_{y_t} \tau_{y_s} \boldsymbol{\pi}' \beta_s - \sum_{s \in [n]: s \neq i} \mathbb{1}' \beta_s}_{=: C_i}
\end{aligned}$$

where C_i is a scalar quantity which does not depend on β_i. Thus, plugging in $\widehat{\boldsymbol{\beta}}$, we have

$$g(\widehat{\boldsymbol{\beta}}) = \frac{1}{2} \|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i + \sum_{s \in [n]: s \neq i} x_s' x_i \widehat{\beta}_i' \boldsymbol{\pi} \tau_{y_i} \tau_{y_s} \boldsymbol{\pi}' \beta_s - \mathbb{1}' \widehat{\beta}_i + C_i. \tag{3.19}$$

Furthermore,

$$\begin{aligned}
\sum_{s\in[n]:s\neq i} x_s' x_i \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i}\tau_{y_s}\boldsymbol{\pi}'\beta_s &= \sum_{s\in[n]:s\neq i} \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i}\tau_{y_s}\boldsymbol{\pi}'\beta_s x_s' x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i} \left(\sum_{s\in[n]:s\neq i} \tau_{y_s}\boldsymbol{\pi}'\beta_s x_s' \right) x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i} \left(-\tau_{y_i}\boldsymbol{\pi}'\beta_i x_i' + \sum_{s\in[n]} \tau_{y_s}\boldsymbol{\pi}'\beta_s x_s' \right) x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i} \left(-\tau_{y_i}\boldsymbol{\pi}'\beta_i x_i' - \sum_{s\in[n]} \alpha_s x_s' \right) x_i \qquad \because eq.\ (3.3) \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\tau_{y_i} \left(-\tau_{y_i}\boldsymbol{\pi}'\beta_i x_i' + \mathbf{w}' \right) x_i \qquad \because eq.\ (3.1) \\
&= \widehat{\beta}_i' \left(-\boldsymbol{\pi}\tau_{y_i}\tau_{y_i}\boldsymbol{\pi}'\beta_i \|x_i\|_2^2 + \boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i \right) \\
&= \widehat{\beta}_i' \left(\boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i - \boldsymbol{\pi}\boldsymbol{\pi}'\beta_i \|x_i\|_2^2 \right) \qquad \because \tau_{y_i}^2 = \mathbf{I} \\
&= \widehat{\beta}_i' \left(\boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i - \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) \qquad \because eq.\ (3.18)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
g(\widehat{\boldsymbol{\beta}}) &= \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta}\widehat{\beta}_i + \widehat{\beta}_i' \left(\boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i - \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 - \mathbb{1} \right) + C_i \\
&= \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta}\widehat{\beta}_i - \widehat{\beta}_i' \left(\mathbb{1} - \boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i + \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) + C_i
\end{aligned}$$

Thus, eq. (3.17) is equivalent to

$$\begin{aligned}
\min_{\widehat{\boldsymbol{\beta}}\in\mathcal{G}} \quad & \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta}\widehat{\beta}_i - \widehat{\beta}_i' \left(\mathbb{1} - \boldsymbol{\pi}\tau_{y_i}\mathbf{w}' x_i + \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) + C_i \\
s.t. \quad & \widehat{\beta}_s = \beta_s,\ \forall s \in [n] \setminus \{i\}.
\end{aligned}$$

Dropping the constant C_i and dividing through by $\|x_i\|_2^2$ does not change the mini-

mizers. Hence, eq. (3.17) has the same set of minimizers as

$$\begin{aligned} \min_{\widehat{\beta}\in\mathcal{G}} \quad & \frac{1}{2}\widehat{\beta}_i'\mathbf{\Theta}\widehat{\beta}_i - \widehat{\beta}_i'\left((\mathbb{1} - \boldsymbol{\pi}\tau_{y_i}\mathbf{w}'x_i)/\|x_i\|_2^2 + \mathbf{\Theta}\beta_i\right) \\ s.t. \quad & \widehat{\beta}_s = \beta_s,\ \forall s \in [n] \setminus \{i\}. \end{aligned}$$

Due to the equality constraints, the only free variable is $\widehat{\beta}_i$. Note that the above optimization, when restricted to $\widehat{\beta}_i$, is equivalent to the optimization eq. (3.4) with

$$v := (\mathbb{1} - \boldsymbol{\pi}\tau_{y_i}\mathbf{w}'x_i)/\|x_i\|_2^2 + \mathbf{\Theta}\beta_i$$

and $\mathbf{w}$ is as in eq. (3.1). The uniqueness of the minimizer is guaranteed by theorem III.4. □

3.8.3 Proof of theorem III.6: global linear convergence

Wang et al. [WL14] established the global linear convergence of the so-called *feasible descent method* when applied to a certain class of problems. As an application, they prove global linear convergence for coordinate descent for solving the dual problem of the binary SVM with the hinge loss. Wang et al. [WL14] considered optimization problems of the following form:

$$\min_{x\in\mathcal{X}} f(x) := g(\mathbf{E}x) + b'x \tag{3.20}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a function such that ∇f is Lipschitz continuous, $\mathcal{X} \subseteq \mathbb{R}^n$ is a polyhedral set, $\arg\min_{x\in\mathcal{X}} f(x)$ is nonempty, $g : \mathbb{R}^m \to \mathbb{R}$ is a strongly convex function such that ∇g is Lipschitz continuous, and $\mathbf{E} \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$ are fixed matrix and vector, respectively.

Below, let $\mathcal{P}_{\mathcal{X}} : \mathbb{R}^n \to \mathcal{X}$ denote the orthogonal projection on $\mathcal{X}$.

Definition III.9. In the context of eq. (3.20), an iterative algorithm that produces

a sequence $\{x^0, x^1, x^2, \dots\} \subseteq \mathcal{X}$ is a *feasible descent method* if there exists a sequence $\{\epsilon^0, \epsilon^1, \epsilon^2, \dots\} \subseteq \mathbb{R}^n$ such that for all $t \geq 0$

$$x^{t+1} = \mathcal{P}_{\mathcal{X}}\left(x^t - \nabla f(x^t) + \epsilon^t\right) \tag{3.21}$$

$$\|\epsilon^t\| \leq B\|x^t - x^{t+1}\| \tag{3.22}$$

$$f(x^t) - f(x^{t+1}) \geq \Gamma\|x^t - x^{t+1}\|^2 \tag{3.23}$$

where $B, \Gamma > 0$.

One of the main result of [WL14] is

Theorem III.10 (Theorem 8 from [WL14])**.** *Suppose an optimization problem* $\min_{x\in\mathcal{X}} f(x)$ *is of the form eq.* (3.20) *and* $\{x^0, x^1, x^2, \dots\} \subseteq \mathcal{X}$ *is a sequence generated by a feasible descent method. Let* $f^* := \min_{x\in\mathcal{X}} f(x)$*. Then there exists* $\Delta \in (0, 1)$ *such that*

$$f(x^{t+1}) - f^* \leq \Delta(f(x^t) - f^*), \quad \forall t \geq 0.$$

Now, we begin verifying that the WW-SVM dual optimization and the BCD algorithm for WW-SVM satisfies the requirements of theorem III.10.

Given $\boldsymbol{\beta} \in \mathbb{R}^{(k-1)\times n}$, define its vectorization

$$\operatorname{vec}(\boldsymbol{\beta}) = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \in \mathbb{R}^{(k-1)n}.$$

Define the matrix $\mathbf{P}_{is} = \boldsymbol{\pi}\tau_{y_i}x_i'x_s\tau_{y_s}\boldsymbol{\pi}' \in \mathbb{R}^{(k-1)\times(k-1)}$, and $\mathbf{Q} \in \mathbb{R}^{(k-1)n\times(k-1)n}$ by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1n} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{n1} & \mathbf{P}_{n2} & \cdots & \mathbf{P}_{nn} \end{bmatrix}.$$

Let

$$\mathbf{E} = \begin{bmatrix} x_1\tau_{y_1}\boldsymbol{\pi}' \\ x_2\tau_{y_2}\boldsymbol{\pi}' \\ \vdots \\ x_n\tau_{y_n}\boldsymbol{\pi}' \end{bmatrix}.$$

We observe that $\mathbf{Q} = \mathbf{E}'\mathbf{E}$. Thus, $\mathbf{Q}$ is symmetric and positive semi-definite. Let $\|\mathbf{Q}\|_{op}$ be the operator norm of $\mathbf{Q}$.

Proposition III.11. *The optimization eq.* (D2) *is of the form eq.* (3.20)*. More precisely, the optimization eq.* (D2) *can be expressed as*

$$\min_{\boldsymbol{\beta}\in\mathcal{G}} \quad g(\boldsymbol{\beta}) = \varphi(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \tag{3.24}$$

where the feasible set $\mathcal{G}$ is a nonempty polyhedral set (i.e., defined by a system of linear inequalities, hence convex), φ is strongly convex, and ∇g is Lipschitz continuous with Lipschitz constant $L := \|\mathbf{Q}\|_{op}$. Furthermore, eq. (3.24) *has at least one minimizer.*

Proof. Observe

$$\begin{aligned} g(\boldsymbol{\beta}) &= \frac{1}{2}\sum_{i,s\in[n]} x_s' x_i \beta_i' \boldsymbol{\pi}\tau_{y_i}\tau_{y_s}\boldsymbol{\pi}'\beta_s - \sum_{i\in[n]} \mathbb{1}'\beta_i \\ &= \frac{1}{2}\mathrm{vec}(\boldsymbol{\beta})'\mathbf{Q}\mathrm{vec}(\boldsymbol{\beta}) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \\ &= \frac{1}{2}(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta}))'(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \\ &= \varphi(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \end{aligned}$$

where $\varphi(\bullet) = \frac{1}{2}\| \bullet \|^2$. Note that $\mathrm{vec}(\nabla g(\boldsymbol{\beta})) = \mathbf{Q}\mathrm{vec}(\boldsymbol{\beta}) - \mathbb{1}$. Hence, the Lipschitz constant of g is $\|\mathbf{Q}\|_{op}$. For the "Furthermore" part, note that the above calculation shows that eq. (3.24) is a quadratic program where the second order term is positive semi-definite and the constraint set is convex. Hence, eq. (3.24) has at least one minimizer. □

Let $B = [0, C]^{k-1}$. Let $\boldsymbol{\beta}^t$ be $\boldsymbol{\beta}$ at the end of the t-iteration of the outer loop of algorithm 1. Define

$$\boldsymbol{\beta}^{t,i} := [\beta_1^{t+1}, \cdots, \beta_i^{t+1}, \beta_{i+1}^t, \cdots, \beta_n^t].$$

By construction, we have

$$\beta_i^{t+1} = \arg\min_{\beta\in B} g\left([\beta_1^{t+1}, \cdots, \beta_{i-1}^{t+1}, \beta, \beta_{i+1}^t, \cdots, \beta_n^t]\right) \tag{3.25}$$

For each $i = 1, \ldots, n$, let

$$\nabla_i g(\boldsymbol{\beta}) = \left[\frac{\partial g}{\partial \beta_{1i}}(\boldsymbol{\beta}), \frac{\partial g}{\partial \beta_{2i}}(\boldsymbol{\beta}), \ldots, \frac{\partial g}{\partial \beta_{(k-1)i}}(\boldsymbol{\beta})\right]'.$$

By Lemma 24 [WL14], we have

$$\beta_i^{t+1} = \mathcal{P}_B(\beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i}))$$

where $\mathcal{P}_B$ denotes orthogonal projection on to B. Now, define $\boldsymbol{\epsilon}^t \in \mathbb{R}^{(k-1)\times n}$ such that

$$\epsilon_i^t = \beta_i^{t+1} - \beta_i^t - \nabla_i g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^t).$$

Proposition III.12. *The BCD algorithm for the WW-SVM is a feasible descent method. More precisely, the sequence* $\{\boldsymbol{\beta}^0, \boldsymbol{\beta}^1, \dots\}$ *satisfies the following conditions:*

$$\boldsymbol{\beta}^{t+1} = \mathcal{P}_{\mathcal{G}}\left(\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t\right) \tag{3.26}$$

$$\|\boldsymbol{\epsilon}^t\| \leq (1 + \sqrt{n}L)\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t+1}\| \tag{3.27}$$

$$g(\boldsymbol{\beta}^t) - g(\boldsymbol{\beta}^{t+1}) \geq \Gamma\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t+1}\|^2 \tag{3.28}$$

where L *is as in proposition III.11,* $\Gamma := \min_{i\in[n]} \frac{\|x_i\|^2}{2}$, $\mathcal{G}$ *is the feasible set of eq.* (D2)*, and* $\mathcal{P}_{\mathcal{G}}$ *is the orthogonal projection onto* $\mathcal{G}$*.*

The proof of proposition III.12 essentially generalizes Proposition 3.4 of [LT93] to the higher dimensional setting:

Proof. Recall that $\mathcal{G} = B^{\times n} := B \times \cdots \times B$. Note that the i-th block of $\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t$ is

$$\beta_i^t - \nabla_i g(\boldsymbol{\beta}^t) + \epsilon_i^t = \beta_i^t - \nabla_i g(\boldsymbol{\beta}^t) + (\beta_i^{t+1} - \beta_i^t - \nabla_i g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^t)) = \beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i}).$$

Thus, the i-th block of $\mathcal{P}_{\mathcal{G}}(\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t)$ is

$$\mathcal{P}_B(\beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i})) = \beta_i^{t+1}.$$

This is precisely the identity eq. (3.26).

Next, we have

$$\begin{aligned}\|\epsilon_i^t\| &\leq \|\beta_i^{t+1} - \beta_i^t\| + \|\nabla_i g(\boldsymbol{\beta}^{t,i}) - \nabla_i g(\boldsymbol{\beta}^t)\| \\ &\leq \|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t,i} - \boldsymbol{\beta}^t\| \\ &\leq \|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|.\end{aligned}$$

From this, we get that

$$\begin{aligned}\|\boldsymbol{\epsilon}^t\| &= \sqrt{\sum_{i=1}^n \|\epsilon_i^t\|^2} \\ &\leq \sqrt{\sum_{i=1}^n (\|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|)^2} \\ &\leq \sqrt{\sum_{i=1}^n \|\beta_i^{t+1} - \beta_i^t\|^2} + \sqrt{\sum_{i=1}^n L^2\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2} \\ &= \|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\| + \sqrt{n}L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\| \\ &= (1 + \sqrt{n}L)\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|.\end{aligned}$$

Thus, we conclude that $\|\boldsymbol{\epsilon}^t\| \leq (1 + \sqrt{n}L)\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|$ which is eq. (3.27).

Finally, we show that

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \geq \Gamma\|\beta_i^{t+1} - \beta_i^t\|^2$$

where $\Gamma := \min_{i\in[n]} \frac{\|x_i\|^2}{2}$.

Lemma III.13. *Let $\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n \in \mathbb{R}^{k-1}$ be arbitrary. Then there exist $v \in \mathbb{R}^{k-1}$ and $C \in \mathbb{R}$ which depend only on $\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n$, but not on β, such that*

$$g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) = \frac{1}{2}\|x_i\|^2\beta'\beta - v'\beta - C.$$

In particular, we have

$$\nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) = \|x_i\|^2 \beta - v.$$

Proof. The result follows immediately from the identity eq. (3.19). □

Lemma III.14. *Let $\beta_1, \cdots, \beta_{i-1}, \beta, \eta, \beta_{i+1}, \cdots, \beta_n \in \mathbb{R}^{k-1}$ be arbitrary. Then we have*

$$\begin{aligned} & g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\ & \qquad + \nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) \\ & = \frac{\|x_i\|^2}{2}\|\eta - \beta\|^2 \end{aligned}$$

Proof. Let v, C be as in lemma III.13. We have

$$\begin{aligned} & g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\ & = \frac{\|x_i\|^2}{2}\|\eta\|^2 - v'\eta - \frac{\|x_i\|^2}{2}\|\beta\|^2 + v'\beta \\ & = \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + v'(\beta - \eta) \end{aligned}$$

and

$$\begin{aligned} & \nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) \\ & = (\|x_i\|^2 \beta - v)'(\beta - \eta) \\ & = \|x_i\|^2(\|\beta\|^2 - \beta'\eta) - v'(\beta - \eta). \end{aligned}$$

Thus,

$$
\begin{aligned}
&g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\
&\qquad + \nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) \\
&= \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + v'(\beta - \eta) + \|x_i\|^2(\|\beta\|^2 - \beta'\eta) - v'(\beta - \eta) \\
&= \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + \|x_i\|^2(\|\beta\|^2 - \beta'\eta) \\
&= \|x_i\|^2\left(\frac{1}{2}(\|\eta\|^2 - \|\beta\|^2) + (\|\beta\|^2 - \beta'\eta)\right) \\
&= \|x_i\|^2\left(\frac{1}{2}(\|\eta\|^2 + \|\beta\|^2) - \beta'\eta\right) \\
&= \frac{\|x_i\|^2}{2}\|\eta - \beta\|^2
\end{aligned}
$$

as desired. □

Applying lemma III.14, we have

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \geq \frac{\|x_i\|^2}{2}\|\beta_i^{t+1} - \beta_i^t\|^2.$$

Since eq. (3.25) is true, we have by Lemma 24 of [WL14] that

$$\nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^t - \beta_i^{t+1}) \geq 0$$

Equivalently, $\nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \leq 0$. Thus, we deduce that

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) \geq \frac{\|x_i\|^2}{2}\|\beta_i^{t+1} - \beta_i^t\|^2 \geq \Gamma\|\beta_i^{t+1} - \beta_i^t\|^2$$

Summing the above identity over $i \in [n]$, we have

$$g(\boldsymbol{\beta}^{t,0}) - g(\boldsymbol{\beta}^{t,n}) = \sum_{i=1}^{n} g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) \geq \Gamma \sum_{i=1}^{n} \|\beta_i^{t+1} - \beta_i^t\|^2 = \Gamma\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2$$

Table 3.2: Variables used in Section 3.8.4

VARIABLE(S)	DEFINED IN	NOTA BENE
t	ALGORITHM 2	ITERATION INDEX
$\ell, \texttt{vals}, \delta_t, \gamma_t$	SUBROUTINE 3	$t \in [\ell]$ IS AN ITERATION INDEX
$\texttt{up}, \texttt{dn}$	SUBROUTINE 3	SYMBOLS
$\widetilde{b}, \widetilde{\gamma}, v_{\max}$	LEMMA III.16	
$\langle 1 \rangle, \ldots, \langle k-1 \rangle$	ALGORITHM 2	
$n_{\mathtt{m}}^t, n_{\mathtt{u}}^t, S^t, \widehat{\gamma}^t, \widehat{b}^t$	ALGORITHM 2	$t \in [\ell]$ IS AN ITERATION INDEX
$\llbracket k \rrbracket$, $I_{\mathtt{u}}^{\gamma}, I_{\mathtt{m}}^{\gamma}, n_{\mathtt{u}}^{\gamma}, n_{\mathtt{m}}^{\gamma}$	DEFINITION III.17	$\gamma \in \mathbb{R}$ IS A REAL NUMBER
$S^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$, $\widehat{\gamma}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$, $\widehat{b}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$	DEFINITION III.20	$(n_{\mathtt{m}}, n_{\mathtt{u}}) \in \llbracket k \rrbracket^2$
$\texttt{vals}^+$	DEFINITION III.25	
$\mathtt{u}(j), \mathtt{d}(j)$	DEFINITION III.26	$j \in [k-1]$ IS AN INTEGER
$\texttt{crit}_1$, $\texttt{crit}_2$	DEFINITION III.27	
$\texttt{KKT\_cond}()$	SUBROUTINE 5	

Since $(\boldsymbol{\beta}^{t,0}) = \boldsymbol{\beta}^t$ and $\boldsymbol{\beta}^{t,n} = \boldsymbol{\beta}^{t+1}$, we conclude that $g(\boldsymbol{\beta}^t) - g(\boldsymbol{\beta}^{t+1}) \geq \Gamma \|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2$. □

To conclude the proof of theorem III.6, we note that proposition III.12 and proposition III.11 together imply that the requirements of Theorem 8 from [WL14] (restated as theorem III.10 here) are satisfied for the BCD algorithm for WW-SVM. Hence, we are done. □

3.8.4 Proof of theorem III.4

The goal of this section is to prove theorem III.4. The time complexity analysis has been carried out at the end of section 3.4 of the main article. Below, we focus on the part of the theorem on the correctness of the output. Throughout this section, $k \geq 2$, $C > 0$ and $v \in \mathbb{R}^{k-1}$ are assumed to be fixed. Additional variables used are summarized in table 3.2.

3.8.4.1 The clipping map

First, we recall the clipping map:

Definition III.15. The *clipping map* $\texttt{clip}_C : \mathbb{R}^{k-1} \to [0, C]^{k-1}$ is the function defined as follows: for $w \in \mathbb{R}^{k-1}$, $[\texttt{clip}_C(w)]_i := \max\{0, \min\{C, w_i\}\}$.

Lemma III.16. *Let* $v_{\max} = \max_{i \in [k-1]} v_i$. *The optimization eq.* (3.4) *has a unique global minimum* $\widetilde{b}$ *satisfying the following:*

1. $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$ *for some* $\widetilde{\gamma} \in \mathbb{R}$
2. $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i$. *In particular,* $\widetilde{\gamma} \geq 0$.
3. *If* $v_i \leq 0$, *then* $\widetilde{b}_i = 0$. *In particular, if* $v_{\max} \leq 0$, *then* $\widetilde{b} = \mathbb{0}$.
4. *If* $v_{\max} > 0$, *then* $0 < \widetilde{\gamma} < v_{\max}$.

Proof. We first prove part 1. The optimization eq. (3.4) is a minimization over a convex domain with strictly convex objective, and hence has a unique global minimum $\widetilde{b}$. For each $i \in [k-1]$, let $\lambda_i, \mu_i \in \mathbb{R}$ be the dual variables for the constraints $0 \geq b_i - C$ and $0 \geq -b_i$, respectively. The Lagrangian for the optimization eq. (3.4) is

$$\mathcal{L}(b, \lambda, \mu) = \frac{1}{2} b'(\mathbf{I} + \mathbf{O})b - v'b + (b - C)'\lambda + (-b)'\mu.$$

Thus, the stationarity (or gradient vanishing) condition is

$$0 = \nabla_b \mathcal{L}(b, \lambda, \mu) = (\mathbf{I} + \mathbf{O})b - v + \lambda - \mu.$$

The KKT conditions are as follows:

for all $i \in [k-1]$, the following holds:

$$[(\mathbf{I}+\mathbf{O})b]_i + \lambda_i - \mu_i = v_i \quad \text{stationarity} \tag{3.29}$$

$$C \geq b_i \geq 0 \quad \text{primal feasibility} \tag{3.30}$$

$$\lambda_i \geq 0 \quad \text{dual feasibility} \tag{3.31}$$

$$\mu_i \geq 0 \quad " \tag{3.32}$$

$$\lambda_i(C - b_i) = 0 \quad \text{complementary slackness} \tag{3.33}$$

$$\mu_i b_i = 0 \quad " \tag{3.34}$$

Equations (3.29) to (3.34) are satisfied if and only if $b = \widetilde{b}$ is the global minimum.

Let $\widetilde{\gamma} \in \mathbb{R}$ be such that $\widetilde{\gamma}\mathbb{1} = \mathbf{O}\widetilde{b}$. Note that by definition, part 2 holds. Furthermore, eq. (3.29) implies

$$\widetilde{b} = v - \widetilde{\gamma}\mathbb{1} - \lambda + \mu. \tag{3.35}$$

Below, fix some $i \in [k-1]$. Note that λ_i or μ_i cannot both be nonzero. Otherwise, eq. (3.33) and eq. (3.34) would imply that $C = \widetilde{b}_i = 0$, a contradiction. We claim the following:

1. If $v_i - \widetilde{\gamma} \in [0, C]$, then $\lambda_i = \mu_i = 0$ and $\widetilde{b}_i = v_i - \widetilde{\gamma}$.
2. If $v_i - \widetilde{\gamma} > C$, then $\widetilde{b}_i = C$.
3. $v_i - \widetilde{\gamma} < 0$, then $\widetilde{b}_i = 0$.

We prove the first claim. To this end, suppose $v_i - \widetilde{\gamma} \in [0, C]$. We will show $\lambda_i = \mu_i = 0$ by contradiction. Suppose $\lambda_i > 0$. Then we have $C = \widetilde{b}_i$ and $\mu_i = 0$. Now, eq. (3.35) implies that $C = \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i$. However, we now have $v_i - \widetilde{\gamma} - \lambda_i \leq C - \lambda_i < C$,

a contradiction. Thus, $\lambda_i = 0$. Similarly, assuming $\mu_i > 0$ implies

$$0 = \widetilde{b}_i = v_i - \lambda + \mu_i \geq 0 + \mu_i > 0,$$

a contradiction. This proves the first claim.

Next, we prove the second claim. Note that

$$C \geq \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i + \mu_i > C - \lambda_i + \mu_i \implies 0 > -\lambda_i + \mu_i \geq -\lambda_i.$$

In particular, we have $\lambda_i > 0$ which implies $C = \widetilde{b}_i$ by complementary slackness.

Finally, we prove the third claim. Note that

$$0 \leq \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i + \mu_i < -\lambda_i + \mu_i \leq \mu_i$$

Thus, $\mu_i > 0$ and so $0 = \widetilde{b}_i$ by complementary slackness. This proves that $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$, which concludes the proof of part 1.

For part 2, note that $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i$ holds by definition. The "in particular" portion follows immediately from $\widetilde{b} \geq 0$.

We prove part 3 by contradiction. Suppose there exists $i \in [k-1]$ such that $v_i \leq 0$ and $\widetilde{b}_i > 0$. Thus, by eq. (3.34), we have $\mu_i = 0$. By eq. (3.29), we have $b_i + \widetilde{\gamma} \leq b_i + \widetilde{\gamma} + \lambda_i = v_i \leq 0$. Thus, we have $-\widetilde{\gamma} \geq b_i > 0$, or equivalently, $\widetilde{\gamma} < 0$. However, this contradicts part 2. Thus, $\widetilde{b}_i = 0$ whenever $v_i \leq 0$. The " in particular" portion follows immediately from the observation that $v_{\max} \leq 0$ implies that $v_i \leq 0$ for all $i \in [k-1]$.

For part 4, we first prove that $\widetilde{\gamma} < v_{\max}$ by contradiction. Suppose that $\widetilde{\gamma} \geq v_{\max}$. Then we have $v - \widetilde{\gamma}\mathbb{1} \leq v - v_{\max}\mathbb{1} \leq \mathbb{0}$. Thus, by part 1, we have $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1}) = \mathbb{0}$. By part 2, we must have that $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i = 0$. However, $\widetilde{\gamma} \geq v_{\max} > 0$, which is a contradiction.

Finally, we prove that $\widetilde{\gamma} > 0$ again by contradiction. Suppose that $\widetilde{\gamma} = 0$. Then part 2 and the fact that $\widetilde{b} \geq \mathbb{0}$ implies that $\widetilde{b} = \mathbb{0}$. However, by part 1, we have $\widetilde{b} = \texttt{clip}_C(v)$. Now, let i^* be such that $v_{i^*} = v_{\max}$. This implies that $\widetilde{b}_{i^*} = \texttt{clip}_C(v_{\max}) > 0$, a contradiction. □

3.8.4.2 Recovering $\widetilde{\gamma}$ from discrete data

Definition III.17. For $\gamma \in \mathbb{R}$, let $b^\gamma := \texttt{clip}_C(v - \gamma\mathbb{1}) \in \mathbb{R}^{k-1}$. Define

$$I_{\mathfrak{u}}^\gamma := \{i \in [k-1] : b_i^\gamma = C\}$$
$$I_{\mathfrak{m}}^\gamma := \{i \in [k-1] : b_i^\gamma \in (0, C)\}$$
$$n_{\mathfrak{u}}^\gamma := |I_{\mathfrak{u}}^\gamma|, \quad \text{and} \quad n_{\mathfrak{m}}^\gamma := |n_{\mathfrak{m}}^\gamma|.$$

Let $[\![k]\!] := \{0\} \cup [k-1]$. Note that by definition, $n_{\mathfrak{m}}^\gamma, n_{\mathfrak{u}}^\gamma \in [\![k]\!]$.

Note that $I_{\mathfrak{u}}^\gamma$ and $I_{\mathfrak{m}}^\gamma$ are determined by their cardinalities. This is because

$$I_{\mathfrak{u}}^\gamma = \{\langle 1\rangle, \langle 2\rangle, \ldots, \langle n_{\mathfrak{u}}^\gamma\rangle\}$$
$$I_{\mathfrak{m}}^\gamma = \{\langle n_{\mathfrak{u}}^\gamma + 1\rangle, \langle n_{\mathfrak{u}}^\gamma + 2\rangle, \ldots, \langle n_{\mathfrak{u}}^\gamma + n_{\mathfrak{m}}^\gamma\rangle\}.$$

Definition III.18. Define

$$\texttt{disc}^+ := \{v_i : i \in [k-1],\, v_i > 0\} \cup \{v_i - C : i \in [k-1], v_i - C > 0\} \cup \{0\}.$$

Note that $\texttt{disc}^+$ is slightly different from $\texttt{disc}$ as defined in the main text.

Lemma III.19. *Let $\gamma', \gamma'' \in \texttt{disc}^+$ be such that $\gamma \notin \texttt{disc}^+$ for all $\gamma \in (\gamma', \gamma'')$. The*

functions

$$(\gamma', \gamma'') \ni \gamma \mapsto I_{\mathfrak{m}}^{\gamma}$$
$$(\gamma', \gamma'') \ni \gamma \mapsto I_{\mathfrak{u}}^{\gamma}$$

are constant.

Proof. We first prove $I_{\mathfrak{m}}^{\lambda} = I_{\mathfrak{m}}^{\rho}$. Let $\lambda, \rho \in (\gamma', \gamma'')$ be such that $\lambda < \rho$. Assume for the sake of contradiction that $I_{\mathfrak{m}}^{\lambda} \neq I_{\mathfrak{m}}^{\rho}$. Then either 1) $i \in [k-1]$ such that $v_i - \lambda \in (0, C)$ but $v_i - \rho \notin (0, C)$ or 2) $i \in [k-1]$ such that $v_i - \lambda \notin (0, C)$ but $v_i - \rho \in (0, C)$. This implies that there exists some $\gamma \in (\lambda, \rho)$ such that $v_i - \gamma \in \{0, C\}$, or equivalently, $\gamma \in \{v_i, v_i - C\}$. Hence, $\gamma \in \mathtt{disc}^+$, which is a contradiction. Thus, for all $\lambda, \rho \in (\gamma', \gamma'')$, we have $I_{\mathfrak{m}}^{\lambda} = I_{\mathfrak{m}}^{\rho}$.

Next, we prove $I_{\mathfrak{u}}^{\lambda} = I_{\mathfrak{u}}^{\rho}$. Let $\lambda, \rho \in (\gamma', \gamma'')$ be such that $\lambda < \rho$. Assume for the sake of contradiction that $I_{\mathfrak{u}}^{\lambda} \neq I_{\mathfrak{u}}^{\rho}$. Then either 1) $i \in [k-1]$ such that $v_i - \lambda \geq C$ but $v_i - \rho < C$ or 2) $i \in [k-1]$ such that $v_i - \lambda < C$ but $v_i - \rho \geq C$. This implies that there exists some $\gamma \in (\lambda, \rho)$ such that $v_i - \gamma = C$, or equivalently, $\gamma = v_i = C$. Hence, $\gamma \in \mathtt{disc}^+$, which is a contradiction. Thus, for all $\lambda, \rho \in (\gamma', \gamma'')$, we have $I_{\mathfrak{u}}^{\lambda} = I_{\mathfrak{u}}^{\rho}$. □

Definition III.20. For $(n_{\mathfrak{m}}, n_{\mathfrak{u}}) \in \llbracket k \rrbracket^2$, define $S^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})}, \widehat{\gamma}^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})} \in \mathbb{R}$ by

$$S^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})} := \sum_{i=n_{\mathfrak{u}}+1}^{n_{\mathfrak{u}}+n_{\mathfrak{m}}} v_{\langle i \rangle},$$
$$\widehat{\gamma}^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})} := \left(C \cdot n_{\mathfrak{u}} + S^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})}\right) / (n_{\mathfrak{m}} + 1).$$

Furthermore, define $\widehat{b}^{(n_{\mathfrak{m}},n_{\mathfrak{u}})} \in \mathbb{R}^{k-1}$ such that, for $i \in [k-1]$, the $\langle i \rangle$-th entry is

$$\widehat{b}^{(n_{\mathfrak{m}},n_{\mathfrak{u}})}_{\langle i \rangle} := \begin{cases} C & : i \leq n_{\mathfrak{u}} \\ v_{\langle i \rangle} - \gamma^{(n_{\mathfrak{m}},n_{\mathfrak{u}})} & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ 0 & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases}$$

Below, recall ℓ as defined on Subroutine 3-line 2.

Lemma III.21. *Let $t \in [\ell]$. Let $n^t_{\mathfrak{m}}$, $n^t_{\mathfrak{u}}$, and $\widehat{b}^t$ be as in the for loop of algorithm 2. Then $\widehat{\gamma}^{(n^t_{\mathfrak{m}},n^t_{\mathfrak{u}})} = \widehat{\gamma}^t$ and $\widehat{b}^{(n^t_{\mathfrak{m}},n^t_{\mathfrak{u}})} = \widehat{b}^t$.*

Proof. It suffices to show that $S^t = S^{(n^t_{\mathfrak{m}},n^t_{\mathfrak{u}})}$ where the former is defined as in algorithm 2 and the latter is defined as in definition III.20. In other words, it suffices to show that

$$S^t = \sum_{j \in [k-1] \,:\, n^t_{\mathfrak{u}} < j \leq n^t_{\mathfrak{u}} + n^t_{\mathfrak{m}}} v_{\langle j \rangle}. \tag{3.36}$$

We prove eq. (3.36) by induction. The base case $t = 0$ follows immediately due to the initialization in algorithm 2-line 4.

Now, suppose that eq. (3.36) holds for S^{t-1}:

$$S^{t-1} = \sum_{j \in [k-1] \,:\, n^{t-1}_{\mathfrak{u}} < j \leq n^{t-1}_{\mathfrak{u}} + n^{t-1}_{\mathfrak{m}}} v_{\langle j \rangle}. \tag{3.37}$$

Consider the first case that $\delta_t = \texttt{up}$. Then we have $n^t_{\mathfrak{u}} + n^t_{\mathfrak{m}} = n^{t-1}_{\mathfrak{u}} + n^{t-1}_{\mathfrak{m}}$ and $n^t_{\mathfrak{u}} = n^{t-1}_{\mathfrak{u}} + 1$. Thus, we have

$$\begin{aligned} S^t &= S^{t-1} - v_{\langle n^{t-1}_{\mathfrak{u}} \rangle} \quad \because \text{Subroutine 4-line 3,} \\ &= \sum_{j \in [k-1] \,:\, n^{t-1}_{\mathfrak{u}} + 1 < j \leq n^{t-1}_{\mathfrak{u}} + n^{t-1}_{\mathfrak{m}}} v_{\langle j \rangle} \quad \because \textit{eq. (3.37)} \\ &= \sum_{j \in [k-1] \,:\, n^t_{\mathfrak{u}} < j \leq n^t_{\mathfrak{u}} + n^t_{\mathfrak{m}}} v_{\langle j \rangle} \end{aligned}$$

which is exactly the desired identity in eq. (3.36).

Consider the second case that $\delta_t = \texttt{dn}$. Then we have $n_{\texttt{u}}^t + n_{\texttt{m}}^t = n_{\texttt{u}}^{t-1} + n_{\texttt{m}}^{t-1} + 1$ and $n_{\texttt{u}}^t = n_{\texttt{u}}^{t-1}$. Thus, we have

$$\begin{aligned} S^t &= S^{t-1} + v_{\langle n_{\texttt{u}}^t + n_{\texttt{m}}^t \rangle} \quad \because \text{Subroutine 4-line 6,} \\ &= \sum_{j \in [k-1] \,:\, n_{\texttt{u}}^{t-1} + 1 < j \leq n_{\texttt{u}}^{t-1} + n_{\texttt{m}}^{t-1} + 1} v_{\langle j \rangle} \quad \because \textit{eq.}\ (3.37) \\ &= \sum_{j \in [k-1] \,:\, n_{\texttt{u}}^{t} < j \leq n_{\texttt{u}}^{t} + n_{\texttt{m}}^{t}} v_{\langle j \rangle} \end{aligned}$$

which, again, is exactly the desired identity in eq. (3.36). □

Lemma III.22. *Let $\widetilde{\gamma}$ be as in lemma III.16. Then we have*

$$\widetilde{b} = \widehat{b}^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})} = \texttt{clip}_C(v - \widehat{\gamma}^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})} \mathbb{1}).$$

Proof. It suffices to prove that $\widetilde{\gamma} = \widehat{\gamma}^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})}$. To this end, let $i \in [k-1]$. If $i \in I_{\texttt{m}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = v_i - \widetilde{\gamma}$. If $i \in I_{\texttt{u}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = C$. Otherwise, $\widetilde{b}_i = 0$. Thus

$$\widetilde{\gamma} = \mathbb{1}' \widetilde{b} = C \cdot n_{\texttt{u}}^{\widetilde{\gamma}} + S^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})} - \widetilde{\gamma} \cdot n_{\texttt{m}}^{\widetilde{\gamma}}$$

Solving for $\widetilde{\gamma}$, we have

$$\widetilde{\gamma} = \left(C \cdot n_{\texttt{u}}^{\widetilde{\gamma}} + S^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})} \right) / (n_{\texttt{m}}^{\widetilde{\gamma}} + 1) = \widehat{\gamma}^{(n_{\texttt{m}}^{\widetilde{\gamma}}, n_{\texttt{u}}^{\widetilde{\gamma}})},$$

as desired. □

3.8.4.3 Checking the KKT conditions

Lemma III.23. *Let* $(n_{\mathfrak{m}}, n_{\mathfrak{u}}) \in [\![k]\!]^2$. *To simplify notation, let* $b := \widehat{b}^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})}$, $\gamma := \widehat{\gamma}^{(n_{\mathfrak{m}}, n_{\mathfrak{u}})}$. *We have* $\mathbf{O}b = \gamma \mathbb{1}$ *and for all* $i \in [k-1]$ *that*

$$[(\mathbf{I}+\mathbf{O})b]_{\langle i \rangle} = \begin{cases} C+\gamma & : i \leq n_{\mathfrak{u}} \\ v_{\langle i \rangle} & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ \gamma & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases} \tag{3.38}$$

Furthermore, b *satisfies the KKT conditions Equations* (3.29) *to* (3.34) *if and only if, for all* $i \in [k-1]$,

$$v_{\langle i \rangle} \begin{cases} \geq C+\gamma & : i \leq n_{\mathfrak{u}} \\ \in [\gamma, C+\gamma] & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ \leq \gamma & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases} \tag{3.39}$$

Proof. First, we prove $\mathbf{O}b = \gamma\mathbb{1}$ which is equivalent to $[\mathbf{O}b]_j = \gamma$ for all $j \in [k-1]$. This is a straightforward calculation:

$$\begin{aligned} [\mathbf{O}b]_j = \mathbb{1}'b &= \sum_{i \in [k-1]} b_{\langle i \rangle} \\ &= \sum_{i \in [k-1] \,:\, i \leq n_{\mathfrak{u}}} b_{\langle i \rangle} + \sum_{i \in [k-1] \,:\, n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}}} b_{\langle i \rangle} + \sum_{i \in [k-1] \,:\, n_{\mathfrak{u}} + n_{\mathfrak{m}} < i} b_{\langle i \rangle} \\ &= \sum_{i \in [k-1] \,:\, i \leq n_{\mathfrak{u}}} C + \sum_{i \in [k-1] \,:\, n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}}} v_{\langle i \rangle} - \gamma \\ &= C \cdot n_{\mathfrak{u}} + S^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)} - n_{\mathfrak{m}}\gamma \\ &= \gamma. \end{aligned}$$

Since $[(\mathbf{I}+\mathbf{O})b]_i = [\mathbf{I}b]_i + [\mathbf{O}b]_i$, the identity eq. (3.38) now follows immediately.

Next, we prove the "Furthermore" part. First, we prove the "only if" direction. By assumption, we have $b = \widetilde{b}$ and so $\gamma = \widetilde{\gamma}$. Furthermore, from lemma III.16 we have $\widetilde{b} = \mathtt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$ and so $b = \mathtt{clip}_C(v - \gamma\mathbb{1})$. To proceed, recall that by construction, we have

$$b_{\langle i \rangle} = \begin{cases} C & : i \leq n_{\mathfrak{u}} \\ v - \gamma & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ 0 & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i \end{cases}$$

Thus, if $i \leq n_{\mathfrak{u}}$, then $C = b_{\langle i \rangle} = [\mathtt{clip}_C(v - \gamma\mathbb{1})]_{\langle i \rangle}$ implies that $v_{\langle i \rangle} - \gamma \geq C$. If $n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}}$, then $b_{\langle i \rangle} = v_{\langle i \rangle} - \gamma$. Since $b_j \in [0, C]$ for all $j \in [k-1]$, we have in particular that $v_{\langle i \rangle} - \gamma \in [0, C]$. Finally, if $n_{\mathfrak{u}} + n_{\mathfrak{m}} < i$, then $0 = b_{\langle i \rangle} = [\mathtt{clip}_C(v - \gamma\mathbb{1})]_{\langle i \rangle}$ implies that $v - \gamma \leq 0$. In summary,

$$v_{\langle i \rangle} - \gamma \begin{cases} \geq C & : i \leq n_{\mathfrak{u}} \\ \in [0, C] & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ \leq 0 & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases}$$

Note that the above identity immediately implies eq. (3.39).

Next, we prove the "if" direction. Using eq. (3.38) and eq. (3.39), we have

$$[(\mathbf{I} + \mathbf{O})b]_{\langle i \rangle} - v_{\langle i \rangle} \begin{cases} \leq 0 & : i \leq n_{\mathfrak{u}} \\ = 0 & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ \geq 0 & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases}$$

For each $i \in [k-1]$, define $\lambda_i, \mu_i \in \mathbb{R}$ where

$$\lambda_{\langle i \rangle} = \begin{cases} -([(\mathbf{I}+\mathbf{O})b]_{\langle i \rangle} - v_{\langle i \rangle}) & : i \leq n_{\mathfrak{u}} \\ 0 & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ 0 & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i \end{cases}$$

and

$$\mu_{\langle i \rangle} = \begin{cases} 0 & : i \leq n_{\mathfrak{u}} \\ 0 & : n_{\mathfrak{u}} < i \leq n_{\mathfrak{u}} + n_{\mathfrak{m}} \\ [(\mathbf{I}+\mathbf{O})b]_{\langle i \rangle} - v_{\langle i \rangle} & : n_{\mathfrak{u}} + n_{\mathfrak{m}} < i. \end{cases}$$

It is straightforward to verify that all of Equations (3.29) to (3.34) are satisfied for all $i \in [k-1]$, i.e., the KKT conditions hold at b. □

Recall that we use indices with angle brackets $\langle 1 \rangle, \langle 2 \rangle, \ldots, \langle k-1 \rangle$ to denote a fixed permutation of $[k-1]$ such that

$$v_{\langle 1 \rangle} \geq v_{\langle 2 \rangle} \geq \cdots \geq v_{\langle k-1 \rangle}.$$

Corollary III.24. *Let $t \in [\ell]$ and $\widetilde{b}$ be the unique global minimum of the optimization eq. (3.4). Then $\widehat{b}^t = \widetilde{b}$ if and only if* `KKT_cond()` *returns true during the t-th iteration of algorithm 2.*

Proof. First, by lemma III.16 we have $\widehat{b}^t = \widetilde{b}$ if and only if $\widehat{b}^t$ satisfies the KKT conditions Equations (3.29) to (3.34). From lemma III.21, we have $\widehat{b}^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)} = \widehat{b}^t$ and $\widehat{\gamma}^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)} = \widehat{\gamma}^t$. To simplify notation, let $\gamma = \widehat{\gamma}^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)}$. By lemma III.23, $\widehat{b}^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)}$ satisfies

the KKT conditions Equations (3.29) to (3.34) if and only if the following are true:

$$v_{\langle i \rangle} \begin{cases} \geq C + \gamma & : i \leq n_{\mathtt{u}}^t \\ \in [\gamma, C + \gamma] & : n_{\mathtt{u}}^t < i \leq n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \\ \leq \gamma & : n_{\mathtt{u}}^t + n_{\mathtt{m}}^t < i. \end{cases}$$

Since $v_{\langle 1 \rangle} \geq v_{\langle 2 \rangle} \geq \cdots$, the above system of inequalities holds for all $i \in [k-1]$ if and only if

$$\begin{cases} C + \gamma \leq v_{\langle n_{\mathtt{u}}^t \rangle} & : \text{if } n_{\mathtt{u}}^t > 0. \\ \gamma \leq v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \rangle} \text{ and } v_{\langle n_{\mathtt{u}}^t + 1 \rangle} \leq C + \gamma & : \text{if } n_{\mathtt{m}}^t > 0, \\ v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t + 1 \rangle} \leq \gamma & : \text{if } n_{\mathtt{u}}^t + n_{\mathtt{m}}^t < k - 1. \end{cases}$$

Note that the above system holds if and only if `KKT_cond`() returns true. □

3.8.4.4 The variables $n_{\mathtt{m}}^t$ and $n_{\mathtt{u}}^t$

Definition III.25. Define the set $\mathtt{vals}^+ = \{(v_j, \mathtt{dn}, j) : v_j > 0, j = 1, \ldots, k-1\} \cup \{(v_j - C, \mathtt{up}, j) : v_j > C, j = 1, \ldots, k-1\}$. Sort the set $\mathtt{vals}^+ = \{(\gamma_1, \delta_1, j_1), \ldots, (\gamma_\ell, \delta_\ell, j_\ell)\}$ so that the ordering of $\{(\gamma_1, \delta_1), \ldots, (\gamma_\ell, \delta_\ell)\}$ is identical to `vals` from Subroutine 3-line 2.

Definition III.26. Define

$$\mathtt{u}(j) := \max\{\tau \in [\ell] : v_{\langle j \rangle} - C = \gamma_\tau\}, \quad \text{and} \quad \mathtt{d}(j) := \max\{\tau \in [\ell] : v_{\langle j \rangle} = \gamma_\tau\}, \tag{3.40}$$

where $\max \emptyset = \ell + 1$.

Definition III.27. Define the following sets

$$\texttt{crit}_1(v) = \{\tau \in [\ell] : \gamma_\tau > \gamma_{\tau+1}\}$$
$$\texttt{crit}_2(v) = \{\tau \in [\ell] : \gamma_\tau = \gamma_{\tau+1},\, \delta_\tau = \texttt{up},\, \delta_{\tau+1} = \texttt{dn}\}$$

where $\gamma_{\ell+1} = 0$.

Later, we will show that algorithm 2 will halt and output the global optimizer $\widetilde{b}$ on or before the t-th iteration where $t \in \texttt{crit}_1(v) \cup \texttt{crit}_2(v)$.

Lemma III.28. *Suppose that $t \in \texttt{crit}_1(v)$. Then*

$$\#\{j \in [k-1] : \texttt{d}(j) \le t\} = \#\{\tau \in [t] : \delta_\tau = \texttt{dn}\},$$

and

$$\#\{j \in [k-1] : \texttt{u}(j) \le t\} = \#\{\tau \in [t] : \delta_\tau = \texttt{up}\}.$$

Proof. First, we observe that

$$\#\{\tau \in [t] : \delta_\tau = \texttt{up}\} = \#\{(\gamma, \delta, j') \in \texttt{vals}^+ : \delta = \texttt{up},\, \gamma \ge \gamma_t\}$$

Next, note that $j \mapsto (\gamma_{\texttt{d}(j)}, \texttt{up}, \langle j \rangle)$ is a bijection from $\{j \in [k-1] : \texttt{d}(j) \le t\}$ to $\{(\gamma, \delta, j') \in \texttt{vals}^+ : \delta = \texttt{up},\, \gamma \ge \gamma_t\}$. To see this, we view the permutation $\langle 1 \rangle, \langle 2 \rangle, \ldots$ viewed as a bijective mapping $\langle \cdot \rangle : [k-1] \to [k-1]$ given by $j \mapsto \langle j \rangle$. Denote by $\rangle \cdot \langle$ the inverse of $\langle \cdot \rangle$. Then the (two-sided) inverse to $j \mapsto (\gamma_{\texttt{d}(j)}, \texttt{up}, \langle j \rangle)$ is clearly given by $(\gamma, \texttt{up}, j') \mapsto \rangle j' \langle$. This proves the first identity of the lemma.

The proof of the second identity is completely analogous. □

Lemma III.29. *The functions* $\texttt{u}$ *and* $\texttt{d} : [k-1] \to [\ell+1]$ *are non-decreasing. Furthermore, for all $j \in [k-1]$, we have $\texttt{u}(j) < \texttt{d}(j)$.*

Proof. Let $j', j'' \in [k-1]$ be such that $j' < j''$. By the sorting, we have $v_{\langle j' \rangle} \geq v_{\langle j'' \rangle}$. Now, suppose that $\mathtt{d}(j') > \mathtt{d}(j'')$, then by construction we have $\gamma_{\mathtt{d}(j')} < \gamma_{\mathtt{d}(j'')}$. On the other hand, we have

$$\gamma_{\mathtt{d}(j')} = v_{\langle j' \rangle} \geq v_{\langle j'' \rangle} = \gamma_{\mathtt{d}(j'')}$$

which is a contradiction.

For the "Furthermore" part, suppose the contrary that $\mathtt{u}(j) \geq \mathtt{d}(j)$. Then we have $\gamma_{\mathtt{u}(j)} \leq \gamma_{\mathtt{d}(j)}$. However, by definition, we have $\gamma_{\mathtt{u}(j)} = v_{\langle j \rangle} > v_{\langle j \rangle} - C = \gamma_{\mathtt{d}(j)}$. This is a contradiction. □

Lemma III.30. *Let $t \in \mathtt{crit}_1(v)$. Then $n^t_{\mathtt{u}} = \#\{j \in [k-1] : \mathtt{u}(j) \leq t\}$. Furthermore, $[n^t_{\mathtt{u}}] = \{j \in [k-1] : \mathtt{u}(j) \leq t\}$. Equivalently, for each $j \in [k-1]$, we have $j \leq n^t_{\mathtt{u}}$ if and only if $\mathtt{u}(j) \leq t$.*

Proof. First, we note that

$$\begin{aligned} n^t_{\mathtt{u}} &= \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\} \quad \because \text{Subroutine 4-line 2} \\ &= \#\{j \in [k-1] : \mathtt{u}(j) \leq t\} \quad \because \text{lemma III.28} \end{aligned}$$

This proves the first part. For the "Furthermore" part, let $N := \#\{j \in [k-1] : \mathtt{u}(j) \leq t\}$. Since $\mathtt{u}$ is monotonic non-decreasing (lemma III.29), we have $\{j \in [k-1] : \mathtt{u}(j) \leq t\} = [N]$. Since $N = n^t_{\mathtt{u}}$ by the first part, we are done. □

Lemma III.31. *Let $\hat{t}, \check{t} \in \mathtt{crit}_1(v)$ be such that there exists $t \in [\ell]$ where*

$$n^t_{\mathtt{m}} = \#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\} - \#\{j \in [k-1] : \mathtt{u}(j) \leq \hat{t}\}. \tag{3.41}$$

Then $\mathtt{d}(j) \leq \check{t}$ and $\hat{t} < \mathtt{u}(j)$ if and only if $n^{\hat{t}}_{\mathtt{u}} < j \leq n^{\hat{t}}_{\mathtt{u}} + n^t_{\mathtt{m}}$.

Proof. By lemma III.30 and eq. (3.41), we have $\#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\} = n^{\hat{t}}_{\mathtt{u}} + n^t_{\mathtt{m}}$. By lemma III.29, $\mathtt{d}$ is monotonic non-decreasing and so $[n^{\hat{t}}_{\mathtt{u}} + n^t_{\mathtt{m}}] = \{j \in [k-1] :$

$\mathtt{d}(j) \le \check{t}\,\}$. Now,

$$\begin{aligned}
&\{j \in [k-1] : \mathtt{d}(j) \le \check{t},\, \hat{t} < \mathtt{u}(j)\} \\
=&\{j \in [k-1] : \mathtt{d}(j) \le \check{t}\,\} \cap \{j \in [k-1] : \hat{t} < \mathtt{u}(j)\} \\
=&\{j \in [k-1] : \mathtt{d}(j) \le \check{t}\,\} \setminus \{j \in [k-1] : \mathtt{u}(j) \le \hat{t}\,\} \\
=&[n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t}] \setminus [n_{\mathtt{u}}^{\hat{t}}],
\end{aligned}$$

where in the last equality, we used lemma III.30. □

Corollary III.32. *Let $t \in \mathtt{crit}_1(v)$. Then $\mathtt{d}(j) \le t$ and $t < \mathtt{u}(j)$ if and only if $n_{\mathtt{u}}^t < j \le n_{\mathtt{u}}^t + n_{\mathtt{m}}^t$.*

Proof. We apply lemma III.31 with $t = \hat{t} = \check{t}$, which requires checking that

$$n_{\mathtt{m}}^t = \#\{j \in [k-1] : \mathtt{d}(j) \le t\} - \#\{j \in [k-1] : \mathtt{u}(j) \le t\}.$$

This is true because from Subroutine 4-line 2 and 5, we have

$$n_{\mathtt{m}}^t = \#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} - \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\}.$$

Applying lemma III.28, we are done. □

Lemma III.33. *Let $t \in \mathtt{crit}_1(v)$. Let $\varepsilon > 0$ be such that for all $\tau, \tau' \in \mathtt{crit}_1(v)$ where $\tau' < \tau$, we have $\gamma_{\tau'} - \varepsilon > \gamma_\tau$. Then $(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) = (n_{\mathtt{m}}^{\gamma_t - \varepsilon}, n_{\mathtt{u}}^{\gamma_t - \varepsilon})$.*

Proof. We claim that

$$v_{\langle j \rangle} - \gamma_t + \varepsilon \begin{cases} < 0 & : t < \mathtt{d}(j) \\ \in (0, C) & : \mathtt{d}(j) \le t < \mathtt{u}(j) \\ > C & : \mathtt{u}(j) \le t. \end{cases} \tag{3.42}$$

To prove the $t < \mathtt{d}(j)$ case of eq. (3.42), we have

$$\begin{aligned} v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{d}(j)} - \gamma_t + \varepsilon \quad \because \textit{eq.}\ (3.40) \\ &< -\varepsilon + \varepsilon = 0 \quad \because\ t < \mathtt{d}(j) \text{ implies that } \gamma_t - \varepsilon > \gamma_{\mathtt{d}(j)}. \end{aligned}$$

To prove the $\mathtt{d}(j) \leq t < \mathtt{u}(j)$ case of eq. (3.42), we note that

$$\begin{aligned} v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{d}(j)} - \gamma_t + \varepsilon \quad \textit{eq.}\ (3.40) \\ &\geq \varepsilon > 0 \quad \because \mathtt{d}(j) \leq t \text{ implies } \gamma_{\mathtt{d}(j)} \geq \gamma_t. \end{aligned}$$

For the other inequality,

$$\begin{aligned} v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{u}(j)} + C - \gamma_t + \varepsilon \quad \because \textit{eq.}\ (3.40) \\ &< -\varepsilon + C + \varepsilon = C \quad \because t < \mathtt{u}(j) \text{ implies } \gamma_t - \varepsilon > \gamma_{\mathtt{u}(j)}. \end{aligned}$$

Finally, we prove the $\mathtt{u}(j) \leq t$ case of eq. (3.42). Note that

$$\begin{aligned} v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{u}(j)} + C - \gamma_t + \varepsilon \quad \because \textit{eq.}\ (3.40) \\ &\geq C + \varepsilon > C \quad \because \mathtt{u}(j) \leq t \text{ implies that } \gamma_{\mathtt{u}(j)} \geq \gamma_t. \end{aligned}$$

Thus, we have proven eq. (3.42). By lemma III.30 and corollary III.32, eq. (3.42) can be rewritten as

$$v_{\langle j \rangle} - \gamma_t + \varepsilon \begin{cases} < 0 & : n_{\mathtt{u}}^t + n_{\mathtt{m}}^t < j, \\ \in (0, C) & : n_{\mathtt{u}}^t < j \leq n_{\mathtt{u}}^t + n_{\mathtt{m}}^t, \\ > C & : j \leq n_{\mathtt{u}}^t. \end{cases} \tag{3.43}$$

Thus, we have $I_{\mathtt{u}}^{\gamma_t - \varepsilon} = \{\langle 1 \rangle, \ldots, \langle n_{\mathtt{u}}^t \rangle\}$ and $I_{\mathtt{m}}^{\gamma_t - \varepsilon} = \{\langle n_{\mathtt{u}}^t + 1 \rangle, \ldots, \langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \rangle\}$. By the

definitions of $n_{\mathtt{u}}^{\gamma_t-\varepsilon}$ and $n_{\mathtt{m}}^{\gamma_t-\varepsilon}$, we are done. □

Lemma III.34. *Let $t \in \mathtt{crit}_2(v)$. Then $(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) = (n_{\mathtt{m}}^{\gamma_t}, n_{\mathtt{u}}^{\gamma_t})$.*

Proof. Let $\hat{t} \in \mathtt{crit}_1(v)$ be such that $\gamma_{\hat{t}} = \gamma_t$, and $\check{t} = \max\{\tau \in \mathtt{crit}_1(v) : \gamma_\tau > \gamma_t\}$. We claim that

$$v_{\langle j\rangle} - \gamma_{\hat{t}} \begin{cases} \leq 0 & : \check{t} < \mathtt{d}(j), \\ \in (0, C) & : \mathtt{d}(j) \leq \check{t},\ \hat{t} < \mathtt{u}(j), \\ \geq C & : \mathtt{u}(j) \leq \hat{t}. \end{cases} \tag{3.44}$$

Note that by definition, we have $\gamma_{\check{t}} > \gamma_{\hat{t}}$, which implies that $\check{t} < \hat{t}$.

Consider the first case of eq. (3.44) that $\check{t} < \mathtt{d}(j)$. We have by construction that $v_{\langle j\rangle} = \gamma_{\mathtt{d}(j)}$ and so $v_{\langle j\rangle} - \gamma_t = \gamma_{\mathtt{d}(j)} - \gamma_{\check{t}} \leq 0$.

Next, consider the case when $\mathtt{d}(j) \leq \check{t}$ and $\hat{t} < \mathtt{u}(j)$. Thus,

$$\begin{aligned} v_{\langle j\rangle} - \gamma_{\hat{t}} > v_{\langle j\rangle} - \gamma_{\check{t}} \quad & \because \gamma_{\check{t}} > \gamma_t \\ = \gamma_{\mathtt{d}(j)} - \gamma_{\check{t}} \quad & \because \text{definition of } \mathtt{d}(j) \\ \geq 0 \quad & \because \mathtt{d}(j) \leq \check{t} \implies \gamma_{\mathtt{d}(j)} \geq \gamma_{\check{t}}. \end{aligned}$$

On the other hand

$$\begin{aligned} v_{\langle j\rangle} - \gamma_{\hat{t}} = \gamma_{\mathtt{u}(j)} + C - \gamma_{\hat{t}} \quad & \because \text{definition of } \mathtt{u}(j) \\ < C \quad & \because \hat{t} < \mathtt{u}(j) \implies \gamma_{\hat{t}} > \gamma_{\mathtt{u}(j)} \end{aligned}$$

Thus, we've shown that in the second case, we have $v_{\langle j\rangle} - \gamma_{\hat{t}} \in (0, C)$.

We consider the final case that $\mathtt{u}(j) \leq \hat{t}$. We have

$$\begin{aligned} v_{\langle j\rangle} - \gamma_{\hat{t}} = \gamma_{\mathtt{u}(j)} + C - \gamma_{\hat{t}} \quad & \because \text{definition of } t \\ \geq C \quad & \because \mathtt{u}(j) \leq \hat{t} \implies \gamma_{\mathtt{u}(j)} \geq \gamma_{\hat{t}}. \end{aligned}$$

Thus, we have proven eq. (3.44).

Next, we claim that $t, \hat{t}, \check{t}$ satisfy the condition eq. (3.41) of lemma III.31, i.e.,

$$n_{\mathtt{m}}^{t} = \#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\,\} - \#\{j \in [k-1] : \mathtt{u}(j) \leq \hat{t}\,\}.$$

To this end, we first recall that

$$n_{\mathtt{m}}^{t} = \#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} - \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\}.$$

By assumption on t, for all τ such that $\check{t} < \tau \leq t$, we have $\delta_\tau = \mathtt{up}$. Thus,

$$\#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} = \#\{\tau \in [\check{t}\,] : \delta_\tau = \mathtt{dn}\} = \#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\,\}$$

where for the last equality, we used lemma III.28. Similarly, for all τ such that $t < \tau \leq \hat{t}$, we have $\delta_\tau = \mathtt{dn}$. Thus, we get that analogous result

$$n_{\mathtt{u}}^{t} = \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\} = \#\{\tau \in [\hat{t}\,] : \delta_\tau = \mathtt{up}\} = \#\{j \in [k-1] : \mathtt{u}(j) \leq \hat{t}\,\} = n_{\mathtt{u}}^{\hat{t}}. \tag{3.45}$$

Thus, we have verified the condition eq. (3.41) of lemma III.31. Now, applying lemma III.30 and lemma III.31, we get

$$v_{\langle j \rangle} - \gamma_{\hat{t}} \begin{cases} \leq 0 & : n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t} < j, \\ \in (0, C) & : n_{\mathtt{u}}^{\hat{t}} < j \leq n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t} \\ \geq C & : j \leq n_{\mathtt{u}}^{\hat{t}}. \end{cases} \tag{3.46}$$

By eq. (3.45) and that $\gamma_t = \gamma_{\hat{t}}$, the above reduces to

$$v_{\langle j \rangle} - \gamma_t \begin{cases} \leq 0 & : n_{\mathfrak{u}}^t + n_{\mathfrak{m}}^t < j, \\ \in (0, C) & : n_{\mathfrak{u}}^t < j \leq n_{\mathfrak{u}}^t + n_{\mathfrak{m}}^t \\ \geq C & : j \leq n_{\mathfrak{u}}^t. \end{cases} \tag{3.47}$$

Thus, $I_{\mathfrak{u}}^{\gamma_t} = \{\langle 1 \rangle, \ldots, \langle n_{\mathfrak{u}}^t \rangle\}$ and $I_{\mathfrak{m}}^{\gamma_t} = \{\langle n_{\mathfrak{u}}^t + 1 \rangle, \ldots, \langle n_{\mathfrak{u}}^t + n_{\mathfrak{m}}^t \rangle\}$. By the definitions of $n_{\mathfrak{u}}^{\gamma_t}$ and $n_{\mathfrak{m}}^{\gamma_t}$, we are done. □

3.8.4.5 Putting it all together

If $v_{\max} \leq 0$, then algorithm 2 returns $\mathbb{0}$.

Otherwise, by lemma III.16, we have $\widetilde{\gamma} \in (0, v_{\max})$.

Lemma III.35. *Let $t \in [\ell]$ be such that $(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t) = (n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})$. Then during the t-th loop of algorithm 2 we have $\widetilde{b} = \widehat{b}^t$ and* `KKT_cond()` *returns true. Consequently, algorithm 2 returns the optimizer $\widetilde{b}$ on or before the t-th iteration.*

Proof. We have

$$\begin{aligned} \widetilde{b} &= \widehat{b}^{(n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})} \quad \because \text{ lemma III.22} \\ &= \widehat{b}^{(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t)} \quad \because \text{Assumption} \\ &= \widehat{b}^t \quad \because \text{ lemma III.21.} \end{aligned}$$

Thus, by corollary III.24 `KKT_cond()` returns true on the t-th iteration. This means that algorithm 2 halts on or before iteration t. Let $\tau \in [\ell]$ be the iteration where algorithm 2 halts and outputs $\widehat{b}^\tau$. Then $\tau \leq t$. Furthermore, by corollary III.24, $\widehat{b}^\tau = \widetilde{b}$, which proves the "Consequently" part of the lemma. □

By lemma III.35, it suffices to show that $(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t) = (n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})$ for some $t \in [\ell]$.

We first consider the case when $\widetilde{\gamma} \neq \gamma_t$ for any $t \in \texttt{crit}_1(v)$. Thus, there exists $t \in \texttt{crit}_1(v)$ such that $\gamma_{t+1} < \widetilde{\gamma} < \gamma_t$, where we recall that $\gamma_{\ell+1} := 0$.

Now, we return to the proof of theorem III.4.

$$\begin{aligned}(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t) &= (n_{\mathfrak{m}}^{\gamma_t-\varepsilon}, n_{\mathfrak{u}}^{\gamma_t-\varepsilon}) \quad \because \text{lemma III.33} \\ &= (n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}}) \quad \because \text{lemma III.19, and that both } \widetilde{\gamma} \text{ and } \gamma_i - \varepsilon \in (\gamma_{t+1}, \gamma_t).\end{aligned}$$

Thus, lemma III.35 implies the result of theorem III.4 under the assumption that $\widetilde{\gamma} \neq \gamma_t$ for any $t \in \texttt{crit}_1(v)$.

Next, we consider when $\widetilde{\gamma} = \gamma_t$ for some $t \in \texttt{crit}_1(v)$. There are three possibilities:

1. There does not exist $j \in [k-1]$ such that $v_{\langle j \rangle} = \gamma_t$,
2. There does not exist $j \in [k-1]$ such that $v_{\langle j \rangle} - C = \gamma_t$,
3. There exist $j_1, j_2 \in [k-1]$ such that $v_{\langle j_1 \rangle} = \gamma_t$ and $v_{\langle j_2 \rangle} - C = \gamma_t$.

First, we consider case 1. We claim that

$$(n_{\mathfrak{m}}^{\gamma_t}, n_{\mathfrak{u}}^{\gamma_t}) = (n_{\mathfrak{m}}^{\gamma_t-\varepsilon'}, n_{\mathfrak{u}}^{\gamma_t-\varepsilon'}) \quad \text{for all } \varepsilon' > 0 \text{ sufficiently small.} \tag{3.48}$$

We first note that $n_{\mathfrak{u}}^{\gamma_t} = n_{\mathfrak{u}}^{\gamma_t-\varepsilon'}$ for all $\varepsilon' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned}i \in I_{\mathfrak{u}}^{\gamma_t} \iff v_i - \gamma_t \geq C &\iff v_i - \gamma_t + \varepsilon' \geq C, \forall \epsilon' > 0, \text{ sufficiently small} \\ &\iff i \in I_{\mathfrak{u}}^{\gamma_t-\varepsilon'}, \forall \epsilon' > 0, \text{ sufficiently small.}\end{aligned}$$

Next, we show that $n_{\mathfrak{m}}^{\gamma_t} = n_{\mathfrak{m}}^{\gamma_t-\varepsilon'}$ for all $\varepsilon' > 0$ sufficiently small. To see this, let

$i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathfrak{m}}^{\gamma_t} &\iff v_i - \gamma_t \in (0, C) \overset{\dagger}{\iff} v_i - \gamma_t + \varepsilon' \in (0, C), \forall \epsilon' > 0, \text{ sufficiently small} \\ &\iff i \in I_{\mathfrak{m}}^{\gamma_t - \varepsilon'}, \forall \epsilon' > 0, \text{ sufficiently small} \end{aligned}$$

where at " $\overset{\dagger}{\iff}$ ", we used the fact that $v_i - \gamma_t \neq 0$ for any $i \in [k-1]$. Thus, we have proven eq. (3.48). Taking $\varepsilon' > 0$ so small so that both eq. (3.48) and the condition in lemma III.33 hold, we have

$$(n_{\mathfrak{m}}^t, n_{\mathfrak{u}}^t) = (n_{\mathfrak{m}}^{\gamma_t - \varepsilon'}, n_{\mathfrak{u}}^{\gamma_t - \varepsilon'}) = (n_{\mathfrak{m}}^{\gamma_t}, n_{\mathfrak{u}}^{\gamma_t}) = (n_{\mathfrak{m}}^{\tilde{\gamma}}, n_{\mathfrak{u}}^{\tilde{\gamma}}).$$

This proves theorem III.4 under case 1.

Next, we consider case 2. We claim that

$$(n_{\mathfrak{m}}^{\gamma_t}, n_{\mathfrak{u}}^{\gamma_t}) = (n_{\mathfrak{m}}^{\gamma_t + \varepsilon''}, n_{\mathfrak{u}}^{\gamma_t + \varepsilon''}) \quad \text{for all } \varepsilon'' > 0 \text{ sufficiently small.} \tag{3.49}$$

We first note that $n_{\mathfrak{u}}^{\gamma_t} = n_{\mathfrak{u}}^{\gamma_t - \varepsilon''}$ for all $\varepsilon'' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathfrak{u}}^{\gamma_t} &\iff v_i - \gamma_t \geq C \overset{\ddagger}{\iff} v_i - \gamma_t - \varepsilon'' \geq C, \forall \epsilon'' > 0, \text{ sufficiently small} \\ &\iff i \in I_{\mathfrak{u}}^{\gamma_t + \varepsilon''}, \forall \epsilon'' > 0, \text{ sufficiently small.} \end{aligned}$$

where at " $\overset{\ddagger}{\iff}$ ", we used the fact that $v_i - \gamma_t \neq C$ for any $i \in [k-1]$. Next, we show that $n_{\mathfrak{m}}^{\gamma_t} = n_{\mathfrak{m}}^{\gamma_t - \varepsilon''}$ for all $\varepsilon'' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathfrak{m}}^{\gamma_t} &\iff v_i - \gamma_t \in (0, C) \overset{\ddagger}{\iff} v_i - \gamma_t - \varepsilon'' \in (0, C), \forall \epsilon'' > 0, \text{ sufficiently small} \\ &\iff i \in I_{\mathfrak{m}}^{\gamma_t + \varepsilon''}, \forall \epsilon'' > 0, \text{ sufficiently small} \end{aligned}$$

where again at " $\overset{\ddagger}{\iff}$ ", we used the fact that $v_i - \gamma_t \neq C$ for any $i \in [k-1]$. Thus, we have proven eq. (3.49). Since $\widetilde{\gamma} = \gamma_t \in (0, v_{\max})$ and $\gamma_1 = v_{\max}$, we have in particular that $\gamma_t < \gamma_1$. Thus, there exists $\tau \in \texttt{crit}_1(v)$ such that $\tau < t$ and $\gamma_t < \gamma_\tau$. Furthermore, we can choose τ such that for all $\gamma \in (\gamma_t, \gamma_\tau)$, $\gamma \notin \texttt{crit}_1(v)$. Let $\varepsilon'' > 0$ be so small that $\gamma_t + \varepsilon'', \gamma_\tau - \varepsilon'' \in (\gamma_t, \gamma_\tau)$, and furthermore both eq. (3.49) and the condition in lemma III.33 hold. We have

$$
\begin{aligned}
(n_{\mathfrak{m}}^{\tau}, n_{\mathfrak{u}}^{\tau}) &= (n_{\mathfrak{m}}^{\gamma_\tau - \varepsilon''}, n_{\mathfrak{u}}^{\gamma_\tau - \varepsilon''}) \quad \because \text{ lemma III.33} \\
&= (n_{\mathfrak{m}}^{\gamma_t + \varepsilon''}, n_{\mathfrak{u}}^{\gamma_t + \varepsilon''}) \quad \because \text{ lemma III.19 and } \gamma_t + \varepsilon'', \gamma_\tau - \varepsilon'' \in (\gamma_t, \gamma_\tau) \\
&= (n_{\mathfrak{m}}^{\gamma_t}, n_{\mathfrak{u}}^{\gamma_t}) \quad \because \text{ eq. (3.49)} \\
&= (n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}}) \quad \because \text{Assumption.}
\end{aligned}
$$

This proves theorem III.4 under case 2.

Finally, we consider the last case. Under the assumptions, we have $t \in \texttt{crit}_2(v)$. Then lemma III.34 $(n_{\mathfrak{m}}^{t}, n_{\mathfrak{u}}^{t}) = (n_{\mathfrak{m}}^{\gamma_t}, n_{\mathfrak{u}}^{\gamma_t}) = (n_{\mathfrak{m}}^{\widetilde{\gamma}}, n_{\mathfrak{u}}^{\widetilde{\gamma}})$. Thus, we have proven theorem III.4 under case 3. □

3.9 Experiments

The Walrus solver is available at:

`https://github.com/YutongWangUMich/liblinear`

The actual implementation is in the file `linear.cpp` in the class `Solver_MCSVM_WW`.

All code for downloading the datasets used, generating the train/test split, running the experiments and generating the figures are included. See the `README.md` file for more information.

All experiments are run on a single machine with the following specifications:

Processor: `Intel(R) Core(TM) i7-6850K CPU @ 3.60GH`

Memory: `31GiB System memory`

3.9.1 On Sharks linear WW-SVM solver

Shark's linear WW-SVM solver is publicly available in the GitHub repository `https://github.com/Shark-ML`. Specifically, the C++ code is in `Algorithms/QP/QpMcLinear.h` in the class `QpMcLinearWW`. Our reimplementation follows their implementation with two major differences. In our implementations, neither Shark nor Walrus use the shrinking heuristic. Furthermore, we use a stopping criterion based on duality gap, following [SHS11].

We also remark that Shark solves the following variant of the WW-SVM which is equivalent to ours after a change of variables. Let $0 < A \in \mathbb{R}$ be a hyperparameter.

$$\min_{\mathbf{u}\in\mathbb{R}^{d\times k}} F_A(\mathbf{u}) := \frac{1}{2}\|\mathbf{u}\|_F^2 + A\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}\left((u'_{y_i}x_i - u'_j x_i)/2\right). \tag{3.50}$$

Recall the formulation eq. (P) that we consider in this work, which we repeat here:

$$\min_{\mathbf{w}\in\mathbb{R}^{d\times k}} G_C(\mathbf{w}) := \frac{1}{2}\|\mathbf{w}\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}(w'_{y_i}x_i - w'_j x_i). \tag{3.51}$$

The formulation eq. (3.50) is used by Weston et al. [WW99], while the formulation eq. (3.51) is used by Vapnik [Vap98]. These two formulations are equivalent under

the change of variables $\mathbf{w} = \mathbf{u}/2$ and $A = 4C$. To see this, note that

$$\begin{aligned}
G_C(\mathbf{w}) &= G_C(\mathbf{u}/2) \\
&= \frac{1}{2}\|\mathbf{u}/2\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2) \\
&= \frac{1}{8}\|\mathbf{u}\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2) \\
&= \frac{1}{4}\left(\frac{1}{2}\|\mathbf{u}\|_F^2 + 4C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2)\right) \\
&= \frac{1}{4}F_{4C}(\mathbf{u}) = \frac{1}{4}F_A(\mathbf{u}).
\end{aligned}$$

Thus, we have proven

Proposition III.36. *Let $C > 0$ and $\mathbf{u} \in \mathbb{R}^{d\times k}$. Then $\mathbf{u}$ is a minimizer of F_{4C} if and only if $\mathbf{u}/2$ is a minimizer of G_C.*

In our experiments, we use the above proposition to rescale the variant formulation to the standard formulation.

3.9.2 Data sets

The data sets used are downloaded from the "LIBSVM Data: Classification (Multi-class)" repository:

`https://csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html`

We use the scaled version of a data set whenever available. For testing accuracy, we use the testing set provided whenever available. The data set ALOI did not have an accompanying test set. Thus, we manually created a test set using methods described in the next paragraph. See table 3.3 for a summary.

The original, unsplit ALOI dataset has $k = 1000$ classes, where each class has 108

Table 3.3: Data sets used from the "LIBSVM Data: Classification (Multi-class)" repository. Variables k, n and d are, respectively, the number of classes, training samples, and features. The SCALED column indicates whether a scaled version of the dataset is available on the repository. The TEST SET PROVIDED column indicates whether a test set of the dataset is provided on the repository.

DATA SET	k	n	d	SCALED	TEST SET AVAILABLE
DNA	3	2,000	180	YES	YES
SATIMAGE	6	4,435	36	YES	YES
MNIST	10	60,000	780	YES	YES
NEWS20	20	15,935	62,061	YES	YES
LETTER	26	15,000	16	YES	YES
RCV1	53	15,564	47,236	NO	YES
SECTOR	105	6,412	55,197	YES	YES
ALOI	1,000	81,000	128	YES	NO

instances. For creating the test set, we split instances from each class such that first 81 elements are training instances while the last 27 elements are testing instances. This results in a "75% train /25% test" split with training and testing set consisting of 81,000 and 27,000 samples, respectively.

3.9.3 Classification accuracy results

For both algorithms, we use the same stopping criterion: after the first iteration t such that $\texttt{DG}^t_\bullet < \delta \cdot \texttt{DG}^1_\bullet$. The results are reported in table 3.5 and table 3.6 where $\delta = 0.009$ and $\delta = 0.0009$, respectively. The highest testing accuracies are in **bold**.

Note that going from table 3.5 to table 3.6, the stopping criterion becomes more stringent. The choice of hyperparameters achieving the highest testing accuracy are essentially unchanged. Thus, for hyperparameter tuning, it suffices to use the more lenient stopping criterion with the larger δ.

Table 3.4: Accuracies under the stopping criterion $\mathtt{DG}^t_\bullet < \delta \cdot \mathtt{DG}^1_\bullet$ with $\delta = 0.09$

$\log_2(C)$	-6	-5	-4	-3	-2	-1	0	1	2	3
DATA SET										
DNA	94.60	**94.69**	94.52	94.44	93.59	92.92	93.09	92.83	92.50	92.83
SATIMAGE	81.95	82.45	82.85	83.75	83.55	**84.10**	83.95	83.30	83.95	84.00
MNIST	92.01	**92.16**	91.97	92.15	91.92	91.76	91.62	91.66	91.70	91.58
NEWS20	82.24	83.17	84.20	84.85	**85.45**	85.15	85.07	84.30	84.40	83.90
LETTER	69.62	**71.46**	70.92	69.82	69.72	70.74	70.50	70.74	71.00	69.22
RCV1	87.23	87.93	88.46	**88.79**	88.78	88.68	88.51	88.29	88.19	88.09
SECTOR	93.08	93.33	93.64	93.92	**94.20**	94.17	**94.20**	94.08	94.14	94.14
ALOI	86.81	87.49	88.22	88.99	89.53	89.71	**89.84**	89.53	89.06	88.21

Table 3.5: Accuracies under the stopping criterion $\texttt{DG}^t_\bullet < \delta \cdot \texttt{DG}^1_\bullet$ with $\delta = 0.009$

$\log_2(C)$	-6	-5	-4	-3	-2	-1	0	1	2	3
DATA SET										
DNA	94.77	94.77	**94.94**	94.69	93.59	93.09	92.24	92.24	92.16	92.16
SATIMAGE	82.35	82.50	82.95	83.55	83.55	84.10	**84.35**	84.20	84.05	84.25
MNIST	92.34	92.28	**92.41**	92.37	92.26	92.13	92.12	91.98	91.94	91.70
NEWS20	82.29	83.35	84.15	85.02	**85.45**	85.30	84.97	84.40	84.12	84.07
LETTER	69.98	71.02	**71.74**	71.52	71.36	71.46	71.20	71.56	71.44	70.74
RCV1	87.24	87.96	88.46	88.76	**88.80**	88.70	88.48	88.25	88.15	88.03
SECTOR	93.14	93.36	93.64	93.95	94.04	**94.08**	94.04	**94.08**	93.98	93.92
ALOI	86.30	87.21	88.20	89.00	89.34	89.63	89.99	**90.18**	89.78	89.80

Table 3.6: Accuracies under the stopping criterion $\texttt{DG}_\bullet^t < \delta \cdot \texttt{DG}_\bullet^1$ with $\delta = 0.0009$.

$\log_2(C)$	-6	-5	-4	-3	-2	-1	0	1	2	3
DATA SET										
DNA	94.77	94.69	**95.11**	94.77	93.76	93.34	92.41	92.24	92.24	92.24
SATIMAGE	82.35	82.65	83.20	83.65	83.80	84.10	**84.20**	84.10	84.15	84.10
MNIST	92.28	92.38	**92.43**	92.24	92.21	92.13	92.16	91.92	91.79	91.65
NEWS20	82.27	83.45	84.00	85.00	**85.40**	85.22	85.02	84.52	84.10	83.97
LETTER	70.04	71.28	**71.70**	71.66	71.48	71.30	71.26	71.30	71.02	71.22
RCV1	87.23	87.98	88.46	88.76	**88.79**	88.69	88.48	88.25	88.12	88.02
SECTOR	93.20	93.39	93.64	93.92	94.01	**94.04**	**94.04**	94.01	93.95	93.83
ALOI	86.17	87.01	87.99	88.66	89.04	89.46	89.64	**89.70**	89.69	89.51

Table 3.7: Accuracies under the stopping criterion $\texttt{DG}^t_\bullet < \delta \cdot \texttt{DG}^1_\bullet$ with $\delta = 0.09$ (first row in each cell), $= 0.009$ (second row) and $= 0.0009$ (third row). For datasets: DNA, SATIMAGE, MNIST, NEWS20.

$\log_2(C)$	-6	-5	-4	-3	-2	-1	0	1	2	3
DATA SET										
DNA ($\delta = 0.09$)	94.60	**94.69**	94.52	94.44	93.59	92.92	93.09	92.83	92.50	92.83
$\delta = 0.009$	94.77	94.77	**94.94**	94.69	93.59	93.09	92.24	92.24	92.16	92.16
$\delta = 0.0009$	94.77	94.69	**95.11**	94.77	93.76	93.34	92.41	92.24	92.24	92.24
SATIMAGE	81.95	82.45	82.85	83.75	83.55	**84.10**	83.95	83.30	83.95	84.00
	82.35	82.50	82.95	83.55	83.55	84.10	**84.35**	84.20	84.05	84.25
	82.35	82.65	83.20	83.65	83.80	84.10	**84.20**	84.10	84.15	84.10
MNIST	92.01	**92.16**	91.97	92.15	91.92	91.76	91.62	91.66	91.70	91.58
	92.34	92.28	**92.41**	92.37	92.26	92.13	92.12	91.98	91.94	91.70
	92.28	92.38	**92.43**	92.24	92.21	92.13	92.16	91.92	91.79	91.65
NEWS20	82.24	83.17	84.20	84.85	**85.45**	85.15	85.07	84.30	84.40	83.90
	82.29	83.35	84.15	85.02	**85.45**	85.30	84.97	84.40	84.12	84.07
	82.27	83.45	84.00	85.00	**85.40**	85.22	85.02	84.52	84.10	83.97

Table 3.8: Accuracies under the stopping criterion $\texttt{DG}^t_\bullet < \delta \cdot \texttt{DG}^1_\bullet$ with $\delta = 0.09$ (first row in each cell), $= 0.009$ (second row) and $= 0.0009$ (third row). For datasets: LETTER, RCV1, SECTOR, ALOI.

$\log_2(C)$	-6	-5	-4	-3	-2	-1	0	1	2	3
DATA SET										
LETTER	69.62	**71.46**	70.92	69.82	69.72	70.74	70.50	70.74	71.00	69.22
	69.98	71.02	**71.74**	71.52	71.36	71.46	71.20	71.56	71.44	70.74
	70.04	71.28	**71.70**	71.66	71.48	71.30	71.26	71.30	71.02	71.22
RCV1	87.23	87.93	88.46	**88.79**	88.78	88.68	88.51	88.29	88.19	88.09
	87.24	87.96	88.46	88.76	**88.80**	88.70	88.48	88.25	88.15	88.03
	87.23	87.98	88.46	88.76	**88.79**	88.69	88.48	88.25	88.12	88.02
SECTOR	93.08	93.33	93.64	93.92	**94.20**	94.17	**94.20**	94.08	94.14	94.14
	93.14	93.36	93.64	93.95	94.04	**94.08**	94.04	**94.08**	93.98	93.92
	93.20	93.39	93.64	93.92	94.01	**94.04**	**94.04**	94.01	93.95	93.83
ALOI	86.81	87.49	88.22	88.99	89.53	89.71	**89.84**	89.53	89.06	88.21
	86.30	87.21	88.20	89.00	89.34	89.63	89.99	**90.18**	89.78	89.80
	86.17	87.01	87.99	88.66	89.04	89.46	89.64	**89.70**	89.69	89.51

3.9.4 Comparison with convex program solvers

For solving eq. (3.4), we compare the speed of Walrus (algorithm 2) versus the general-purpose, commercial convex program (CP) solver MOSEK. We generate random instances of the subproblem eq. (3.4) by randomly sampling v. The runtime results of Walrus and the CP solver are shown in table 3.9 and table 3.10, where each entry is the average over 10 random instances.

Table 3.9: Runtime in seconds for solving random instances of the problem eq. (3.4). The parameter $C = 1$ is fixed while k varies.

$\log_2(k-1)$	2	4	6	8	10	12
Walrus	0.0009	0.0001	0.0001	0.0001	0.0002	0.0005
CP solver	0.1052	0.0708	0.0705	0.1082	0.5721	12.6057

Table 3.10: Runtime in seconds for solving random instances of the problem eq. (3.4). The parameter $k = 2^8 + 1$ is fixed while C varies.

$\log_{10}(C)$	-3	-2	-1	0	1	2	3
Walrus	0.0004	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
CP Solver	0.1177	0.1044	0.1046	0.1005	0.1050	0.1127	0.1206

As shown here, the analytic solver Walrus is faster than the general-purpose commercial solver by orders of magnitude.

CHAPTER IV

Permutation Equivariant Relative Margin Losses for Multiclass Classification

We introduce the permutation equivariant and relative margin-based (PERM) loss for k-ary multiclass classification, and the *multiplicative label encoding*, which generalizes the $\{\pm 1\}$ binary label encoding. By using these tools in conjunction, we can formulate multiclass classification in a way that directly generalize discriminant-based binary classification and prove an extension of the seminal classification-calibration (CC) result of Bartlett et al. [BJM06] to the multiclass setting. PERM losses include the Gamma-Phi [Bei+14], and Fenchel-Young loss families [BMN20]. Using our theoretical framework, we prove sufficient conditions for CC of these two previous families. We demonstrate that the only previously-known sufficient condition for Gamma-Phi loss proposed by Pires et al. [PS16] turns out to be insufficient. Thus, our work establishes the first sufficient condition for general Gamma-Phi losses. For the Fenchel-Young losses, our result recover all known CC sufficient conditions [NBR19; Blo19]. Moreover, we establish CC for Fenchel-Young losses not satisfying previously known sufficient conditions. While this work mainly concerns CC, we believe our framework will be useful for other problems in multiclass classfication.

4.1 Introduction

Multiclass classification into $k \geq 2$ categories is one of the most common tasks in machine learning. Labelled training instances $\texttt{Train}_n := \{(x_i, y_i)\}_{i=1}^n$ are drawn from a joint distribution P over $\mathcal{X} \times [k]$ where $[k] := \{1, \ldots, k\}$ and $\mathcal{X}$ is a space of unlabelled instances. The goal is to select a classifier $g : \mathcal{X} \to [k]$ that makes as few mistakes as possible on test instances. In other words, g should have low *01*-risk, defined as

$$R_{01}(g) := \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{I}\{Y \neq g(X)\}\right] \tag{4.1}$$

where $\mathbb{I}$ is the indicator function. However, directly minimizing the 01 risk is difficult. To address this, many classification algorithms often minimize a *surrogate risk* based a *surrogate loss* $\mathcal{L} : [k] \times \mathbb{R}^k \to \mathbb{R}_+$. Discrete "hard label" classifiers g are replaced by continuous "soft label" classifiers $f : \mathcal{X} \to \mathbb{R}^k$. Instead of minimizing the (empirical) 01-risk, surrogate-based approach seeks to minimize the $\mathcal{L}$-risk, defined as

$$R_{\mathcal{L}}(f) := \mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(Y, f(X))]. \tag{4.2}$$

The *$\mathcal{L}$-Bayes risk* is defined as $R^*_{\mathcal{L}} := \inf_f R_{\mathcal{L}}(f)$ where the infimum is over all Borel functions f. Note that $R^*_{\mathcal{L}}$ is the optimal achievable $\mathcal{L}$-risk for any classifier. However, note that the original goal is to minimize the 01-risk, i.e., to approach the *01-Bayes risk* $R^*_{01} := \inf_g R_{01}(g)$ where the infimum is over all Borel functions g.

The theory of classification-calibration of loss functions is concerned with the following question. A score function $f : \mathcal{X} \to \mathbb{R}^k$ has components $(f_1, \ldots, f_k)$ representing the score assigned to each of the k classes, where higher score implies greater preference for the corresponding class. The final predicted label for an instance x is $\arg\max_{j=1,\ldots,k} f_j(x)$. Suppose that $\{f^{(n)}\}$ is a sequence of score functions $\mathcal{X} \to \mathbb{R}^k$, obtained from, say, empirical $\mathcal{L}$-risk minimization on $\texttt{Train}_n$. In this setup, a natural question is: if $R_{\mathcal{L}}(f^{(n)}) \to R^*_{\mathcal{L}}$ as $n \to \infty$, then does $R_{01}(\arg\max f^{(n)}) \to R^*_{01}$ as well?

If this is the case for all P, then we say that $\mathcal{L}$ *is classification-calibrated*.

For the binary case when $k = 2$, the classification-calibration of $\mathcal{L}$ is relatively well-understood with easy to check sufficient conditions available [Zha04b; BJM06]. In contrast, the theory of multiclass losses is less developed than its binary counterpart. Zhang [Zha04a] and Tewari et al. [TB07] derived an abstract definition of multiclass classification-calibration and provided sufficient conditions in special cases. Nevertheless, new multiclass loss functions continue to be introduced and developed. Two prominent families of loss functions are the Gamma-Phi and the Fenchel-Young losses, introduced by Beijbom et al. [Bei+14] and Blondel et al. [BMN20], respectively. Gamma-Phi losses have been successfully applied in multiclass boosting algorithsm [SV19]. Fenchel-Young losses are defined using a procedure of constructing multiclass loss functions from generalized entropies [DKR18]. The multinomial logistic/cross entropy loss is perhaps the most well-known example and is constructed using the Shannon entropy.

Many works have analyzed sufficient conditions for classification-calibration (CC) of Gamma-Phi [Zha04a; PS16] and Fenchel-Young losses [NBR19; Blo19]. However, several important theoretical gaps remain. Towards addressing these gaps, we introduce the *permutation-equivariant and relative margin-based* (PERM) multiclass classification loss family subsuming both the Gamma-Phi and Fenchel-Young losses. We prove sufficient condition for PERM losses to be CC and apply this theory to Fenchel-Young losses, expanding previously known sufficient conditions. We also establish the first sufficient conditions for a general subfamily of Gamma-Phi loss.

A key ingredient in the analysis of PERM losses is a novel label encoding for multiclass classification which we call *multiplicative label encoding*. Multiplicative label encoding generalizes the $\{\pm 1\}$ encoding in binary classification. Taken together, our framework consisting of PERM losses and multiplicative label encoding provides a natural formulation of margin-based multiclass classification. In the next section,

we discussed our results in greater details.

4.1.1 Our contributions

Establishing the first sufficient condition of CC for Gamma-Phi losses. Prior to our work, Pires et al. [PS16] proposed the only existing sufficient condition of CC for Gamma-Phi loss. However, we show that their proposed conditions turns out to be insufficient. Namely, we construct a Gamma-Phi loss in Section 4.9.1 that is not classification-calibrated but satisfies the conditions of Pires et al. [PS16]. Thus, our Theorem IV.15 establishes the first sufficient condition of CC for Gamma-Phi loss.

Expanding previous sufficient conditions of CC for Fenchel-Young losses. The key ingredient of a Fenchel-Young loss is the so-called *negenetropy*. The recent line of work on Fenchel-Young losses [NBR19; Blo19] proved sufficient conditions for CC under the assumption that the negentropy is *strongly convex*. Our Theorem IV.22 shows that CC holds for the more general *strictly convex* negentropy. Moreover, we exhibit in Section 4.8.2 a large class of negentropies that are strictly convex but not strongly convex. Thus, calibration-classification of the Fenchel-Young losses corresponding to these aforementioned negentropies are guaranteed by our sufficient conditions, and not by any previous ones.

PERM loss and multiplicative label encoding — bridging the gap between binary and multiclass margin loss. Binary classification commonly uses a *discriminant function*, i.e., a function $f : \mathcal{X} \to \mathbb{R}$ mapping an instance x to a real number $f(x) \in \mathbb{R}$ called the *discriminant*. The sign of $f(x)$ is used to classify x as either the positive or the negative class. The *margin* is the multiplication of the training label $y \in \{\pm 1\}$ and the discriminant $f(x)$. A *margin-based loss* is characterized by a function $\psi : \mathbb{R} \to \mathbb{R}$. The learner incurs a penalty of $\psi(yf(x))$ for outputing $f(x)$ given training label y.

To generalize the above to the multiclass case, we view the multiclass discriminant of an instance as a $(k-1)$-dimensional vector $f(x) \in \mathbb{R}^{k-1}$, where k is the number of classes. Our multiplicative label encoding is a set of k square matrices $\{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k\}$, where each matrix is of the size $(k-1)\times(k-1)$. As in the binary case, a (vector-valued) margin is the multiplication of the training label, encoded as $\boldsymbol{\rho}_y$, and the discriminant $f(x) \in \mathbb{R}^{k-1}$. A *PERM loss* is characterized by a function $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$. The learner incurs a penalty of $\psi(\boldsymbol{\rho}_y f(x))$ for outputing $f(x)$ given training label y.

A seminal result of Bartlett et al. [BJM06] in the binary case shows that a convex margin loss is classification-calibrated if and only if ψ is differentiable at 0 and has negative derivative at there. Our sufficient condition for classification-calibration result (Theorem IV.27) can be viewed as a multiclass partial extension of the seminal result of Bartlett et al. [BJM06] in the binary case. While this work focuses on classification-calibration, we foresee that the PERM loss and the multiplicative label encoding framework could have many implications for the theory and practice of multiclass classification. For instance, in Chapters II and III, the multiplicative label encoding plays an important role in establishing new theoretical results regarding the Weston-Watkins support vector machine.

4.1.2 Related works

Gamma-Phi losses. Gamma-Phi losses were introduced and studied in a series of papers [SV19; SV11; Bei+14]. They have been shown to perform well in boosting [SV19]. Classification-calibration have been shown for special instances of Gamma-Phi, namely for the *coherence loss* [Zha+09] and the *pairwise-comparison* loss [Zha04a][1].

Fenchel-Young losses. Fenchel-Young losses were developed by Duchi et al.

[1]However, the pairwise-comparison loss proof by Zhang [Zha04a] is incomplete (see Remark IV.16). Our sufficient condition (Theorem IV.15) for Gamma-Phi losses subsumes that of pairwise comparison loss ([Zha04a, Theorem 6]).

[DKR18] and Blondel et al. [BMN20]. While our focus is on multiclass classification, Fenchel-Young losses apply more generally to other learning problems such as label proportion estimation and dependency parsing [BMN20]. Furthermore, Mensch et al. [MBP19] applies a Fenchel-Young loss in the "infinite-dimensional" problem of distribution learning using a strictly convex negentropy known as the Sinkhorn negentropy[2].

Multiclass frameworks — label encodings, margins and losses. The simplex encoding [BG16; Mro+12] is a $(k-1)$-vector-valued encoding of the labels, which have been proposed to analyze specific losses used in multiclass support vector machines. [SV19] applied simplex encoding in the context of multiclass boosting with Gamma Phi loss. The error-correcting output codes [DB94] encodes the k-ary labels as a bit-string of zeros and ones. In contrast, our work encodes the labels as matrices that directly generalizes the binary case. As the result, in our framework, the label encoding, the margin, and the margin loss are clearly separated and can thus be studied in isolation.

There are other frameworks for analyzing multiclass loss functions in the literature. Williamson et al. [WVR16] proposes the family of losses known as the *composite multiclass losses*. However, neither frameworks provided a sufficient condition for classification-calibration[3]. Tan et al. [TZ22] proved a sufficent condition for non-differentiable multiclass hinge losses to be classification-calibrated. In contrast, our sufficient conditions are for differentiable losses.

Zou et al. [ZZH08] proposed definitions of multiclass margin vectors and margin

[2]To the best of our knowledge, this negentropy is only known to be strictly convex as strong convexity was not discussed [MBP19; Fey+19]. However, we note that our result on Fenchel-Young loss is specifically for multiclass classification and thus does *not* apply to their setting. Expanding our analysis to their setting is an interesting direction of future work.

[3]Somewhat confusingly, Williamson et al. [WVR16] re-defines the term "classification-calibrated" to be different from the definitions of Zhang [Zha04a] and Tewari et al. [TB07]. Williamson et al. [WVR16] establishes a sufficient condition (characterization in fact) for their redefined version of classification-calibration, which is no longer necessarily related to the 01-consistency of surrogate risk minimization.

losses. They proved classification-calibration guarantees that required certain sum-to-zero constraints. However, in practice, enforcing these sum-to-zero constraints leads to significantly slower computations, while simply dropping the constraint result in poor model accuracy [DGI16]. Doğan et al. [DGI16] developed a framework using relative-margins to unify the analysis of several variants of multiclass support vector machines. Our work builds upon the relative-margins notion by introducing a natural label encoding that allowed losses defined over relative margins to be analyzed.

Multiclass losses and overparametrized learning. Loss functions for multiclass classification have recently been studied in the context of learning in overparametrized settings where models can interpolate the training data. While the cross-entropy/multinomial logistic loss is the *de facto* choice in training neural networks, recent works have questioned this convention and pushed forward understanding of alternative losses such as the squared loss [HB20; Mut+21].

Calibration beyond classification. While this paper is concerned with classification-calibration, we remark that there are many works calibration for other learning tasks. Steinwart [Ste07] introduced the extension of loss calibration-theory to cost-sensitive classification, regression and unsupervised learning tasks such as density estimation. Ramaswamy et al. [RA16] developed theory for multiclass classification with abstain option and, more generally, losses defined over finite sets i.e., discrete losses. Finocchiaro et al. [FFW19] showed that there exists polyhedral losses that are calibrated with respect to arbitrary discrete losses.

4.1.3 Notations

Denote by $k \geq 2$ the number of classes and by $\Delta^k = \{\mathbf{p} \in \mathbb{R}^k_{\geq 0} : \sum_{j=1}^k p_j = 1\}$ the k-probability simplex. Let $\Delta^k_{\texttt{desc}} = \{\mathbf{p} \in \Delta^k : p_1 \geq \cdots \geq p_k\}$.

Operations on vectors. Let the square bracket with subscript $[\cdot]_j$ be the projection of a vector onto its j-th component, i.e., $[\mathbf{v}]_j := v_j$ where $\mathbf{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$. Given

Mathematical object	Notation	Example
Vector	Bold lower case	$\mathbf{v}, \mathbf{w}$
Entries of vector	Normal font lower case	v_1
Special vector	Blackboard font	
All zeros/ones vector in $\mathbb{R}^n$		$\mathbb{0}_n, \mathbb{1}_n$
i-th elem. basis vector in $\mathbb{R}^n$		$\mathbb{e}_i^n$
Matrix	Bold upper case	$\mathbf{A}$
j-th Column		$[\mathbf{A}]_{:j}$
$n \times n$ Identity		$\mathbf{Id}_n$
Permutations	Lower case sigma or tau	σ, τ
Transpositions	Lower case tau & subscripts	$\tau_{(i,j)}$

Table 4.1: Symbols used throughout this work.

two vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^k$, we write $\mathbf{w} \geq \mathbf{v}$ if $w_j \geq v_j$ for all $j \in [k]$. Likewise, we write $\mathbf{w} > \mathbf{v}$ if $w_j > v_j$ for all $j \in [k]$.

Permutations. A bijection from $[k]$ to itself is called a *permutation* (on $[k]$). Denote by $\mathtt{Sym}(k)$ the set of all permutations on $[k]$. We often write $\sigma\sigma'$ instead of $\sigma \circ \sigma'$ for the compositions of two permutations $\sigma, \sigma' \in \mathtt{Sym}(k)$. For $i, j \in [k]$, let $\tau_{(i,j)} \in \mathtt{Sym}(k)$ denote the *transposition* which swaps i and j, leaving all other elements unchanged. More precisely, $\tau_{(i,j)}(i) = j$, $\tau_{(i,j)}(j) = i$ and $\tau_{(i,j)}(y) = y$ for $y \in [k] \setminus \{i, j\}$. Define the notational shorthand $\tau_i := \tau_{(1,i)}$, the transposition that swaps 1 and i.

Permutation matrices. For each $\sigma \in \mathtt{Sym}(k)$, let $\mathbf{S}_\sigma$ denote the permutation matrix corresponding to σ. In other words, if $\mathbf{v} \in \mathbb{R}^k$ is a vector, then $[\mathbf{S}_\sigma \mathbf{v}]_j = [\mathbf{v}]_{\sigma(j)} = v_{\sigma(j)}$. Note that if $\sigma, \sigma' \in \mathtt{Sym}(k)$, then $\mathbf{S}_{\sigma\sigma'} = \mathbf{S}_\sigma \mathbf{S}_{\sigma'}$. Define the notational shorthand $\mathbf{T}_{(i,j)} := \mathbf{S}_{\tau_{(i,j)}}$ the matrix corresponding to the transposition of i and j. Likewise, define $\mathbf{T}_i := \mathbf{T}_{(1,i)}$.

Topology. Let S be a subset of a topological space. Let $\mathtt{int}(S)$ and $\mathtt{bdry}(S)$ denote the interior and the boundary of the set S, respectively. See Table 4.1 for the full list of symbols.

4.2 Permutation equivariant and margin based (PERM) losses

In this section, we state the definitions used throughout the rest of this work.

Definition IV.1 (Loss functions)**.** Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k$ be a vector-valued function, whose component functions are denoted $\mathcal{L}_1, \ldots, \mathcal{L}_k$ where $\mathcal{L}_y : \mathbb{R}^k \to \mathbb{R}$ for each $y \in [k]$. We say that $\mathcal{L}$ is a *k-ary multiclass loss function* if for all $\mathbf{v} \in \mathbb{R}^k$ and all $y, y' \in [k]$, $v_y \leq v_{y'}$ implies $\mathcal{L}_y(\mathbf{v}) \geq \mathcal{L}_{y'}(\mathbf{v})$.

To understand the definition, suppose that $f : \mathcal{X} \to \mathbb{R}^k$ is a score function and $(x, y) \in \mathcal{X} \times [k]$ is a training data instance. Let $\mathbf{v} := f(x)$ be the k-dimensional score assigned to the instance x. The quantity $\mathcal{L}_y(\mathbf{v})$ is the loss incurred at the training data instance (x, y). In the loss function literature, a standard approach is to analyze $\mathcal{L}$ via only the vector $\mathbf{v}$ of class scores, while the score function f and the unlabelled instance x are "abstract away". We take this approach as well.

Definition IV.2 (PERM loss)**.** Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k$ be a loss function. We say that $\mathcal{L}$ is

1. *permutation equivariant* if $\mathcal{L}(\mathbf{S}_\sigma(\mathbf{v})) = \mathbf{S}_\sigma(\mathcal{L}(\mathbf{v}))$ for all $\mathbf{v} \in \mathbb{R}^k$ and $\sigma \in \mathtt{Sym}(k)$,
2. *relative margin-based* if there exists a vector-valued function $\ell : \mathbb{R}^{k-1} \to \mathbb{R}^k$, whose component functions are denoted $\ell_1, \ldots, \ell_k$ where $\ell_y : \mathbb{R}^{k-1} \to \mathbb{R}$, such that
$$\mathcal{L}_y(\mathbf{v}) = \ell_y(v_1 - v_2, \ldots, v_1 - v_k)$$
for all $y \in [k]$,
3. *PERM* if $\mathcal{L}$ is both permutation equivariant and relative margin-based.

The function ℓ is called the *reduced form* associated to $\mathcal{L}$.

The term "relative margin" was introduced by Doğan et al. [DGI16] in the context

of multiclass SVMs [4] to distinguish with another type of margin called "absolute margin". Here, the adjective "relative" refers to the situation when the loss $\mathcal{L}$ only depends on the set of differences of the scores $v_i - v_j$ where $i, j \in [k]$ such that $i \neq j$. If $\mathcal{L}$ depends on the quantity v_j as well, then $\mathcal{L}$ is said to be absolute margin-based. Below, we only consider relative margin-based losses.

Proposition IV.3 (Template of a PERM loss)**.** *Let $\mathcal{L}$ be a multiclass loss function. Then $\mathcal{L}$ is PERM if and only if there exists a symmetric function*[5] *$\psi : \mathbb{R}^{k-1} \to \mathbb{R}$ such that*

$$\mathcal{L}_1(\mathbf{v}) = \psi(v_1 - v_2,\, v_1 - v_3, \ldots, v_1 - v_k),\ \textit{and} \tag{4.3}$$

$$\mathcal{L}_y(\mathbf{v}) = \psi(v_y - v_1,\, \ldots,\, v_y - v_{y-1},\, v_y - v_{y+1},\, \ldots,\, v_y - v_k),\ \textit{for } y \in \{2, \ldots, k\} \tag{4.4}$$

for all $y \in [k]$. Below, we often denote the $(k-1)$-dimensional vector $(v_1 - v_2, v_1 - v_3, \ldots, v_1 - v_k)$ as $\mathbf{z}$.

Proposition IV.3 states that there is an one-to-one correspondence between PERM losses and symmetric template functions. Thus, we can refer to a PERM loss and its template interchangeably without ambiguity. We will prove Proposition IV.3 in Section 4.5 where we will state a more detailed result relating $\mathcal{L}$, its reduced form, and its template in Proposition IV.33.

Remark IV.4. The notation ψ is chosen intentionally to match that of Bartlett et al. [BJM06]. Recall from Bartlett et al. [BJM06] that a (binary) margin based loss is a function $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that the loss incurred by a function $f : \mathcal{X} \to \mathbb{R}$ on a sample (x, y) is $\psi(yf(x))$. In Proposition IV.33 below, we show that the template leads to a multiclass analog of this scenario in the binary case.

[4] Unfortunately, the term "relative margin" conflicts with another unrelated definition of the same name proposed by Jebara et al. [JS08] in the context of binary support vector machines.

[5] a function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric if $f \circ \sigma = f$ for all $\sigma \in \mathtt{Sym}(n)$.

Next, we recall the Gamma-Phi and Fenchel-Young losses.

Example IV.5 (Gamma-Phi losses)**.** Let $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be functions. Introduced and studied in a series of papers [SV19; SV11; Bei+14], the *Gamma-Phi* loss associated to γ and ϕ is the PERM loss $\mathcal{L}$ whose y-th component is given by

$$\mathcal{L}_y(\mathbf{v}) := \gamma\left(\sum_{y' \in [k] : y' \neq y} \phi(v_y - v_{y'})\right). \tag{4.5}$$

Thus, $\psi(\mathbf{z}) := \gamma\left(\sum_{j \in [k-1]} \phi(z_j)\right)$. When $\gamma(\bullet) := \log(1 + \bullet)$ and $\phi(\bullet) := \exp(-\bullet)$, we recover the multinomial logistic/cross entropy loss. When $\gamma(\bullet) = T \log(1 + \bullet)$ and $\phi(\bullet) = \exp((1 - \bullet)/T)$ where $T > 0$ is a hyperparameter, we recover the coherence loss [Zha+09]. When γ is the identity and ϕ is a decreasing function, we recover the pairwise comparison loss [Zha04a, Section 4.1].

Example IV.6 (Fenchel-Young losses)**.** Let $\Omega : \Delta^k \to \mathbb{R}$ be a continuous function and $\mu \in \mathbb{R}_{\geq 0}$. Define $\mathbf{c}_y := \mu(\mathbb{1}_k - \mathbb{e}_y^k)$. Introduced by Blondel et al. [BMN20], the *Fenchel-Young* loss associated to Ω and μ is the loss function $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k$ whose y-th component is given by

$$\mathcal{L}_y(\mathbf{v}) := \max_{\mathbf{p} \in \Delta^k} -\Omega(\mathbf{p}) + \Omega(\mathbb{e}_y^k) + \langle \mathbf{v} + \mathbf{c}_y, \mathbf{p} - \mathbb{e}_y^k \rangle. \tag{4.6}$$

The reason for the name is that the above is actually a convex conjugate, also known as the Fenchel conjugate. See Definition IV.18. Later in Proposition IV.86, we show that the Fenchel-Young loss is a PERM loss with template

$$\psi(\mathbf{z}) = \max_{\tilde{\mathbf{p}} \in \tilde{\Delta}^k} -\tilde{\Omega}(\mathbf{p}) + \mu \mathbb{1}^\top \tilde{\mathbf{p}} - \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle \tag{4.7}$$

where $\tilde{\Delta}^k$ is defined in Eqn. (4.9) and $\tilde{\Omega}$ in Eqn. 4.10.

Remark IV.7. Blondel et al. [BMN20] allow the vector $\mathbf{c}_y \in \mathbb{R}^k$ to be arbitrary, in

which case the resulting loss is known as *cost-sensitive Fenchel-Young* loss. However, known calibration results are limited to the case in Example IV.6 above where $\mathbf{c}_y$ has the special form [Blo19; NBR19].

4.2.1 Classification-calibration and Consistency

In this section, we review fundamental definitions in the theory of classification-calibration and recall the key result Theorem IV.12.

Definition IV.8. The *conditional risk* of $\mathcal{L}$ is the function $C_{\mathbf{p}}^{\mathcal{L}} : \mathbb{R}^k \to \mathbb{R}$ defined by

$$C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}) = \sum_{y \in [k]} p_y \mathcal{L}_y(\mathbf{v}).$$

The *conditional Bayes risk* is defined as $C_{\mathbf{p}}^{\mathcal{L},*} := \inf_{\mathbf{v} \in \mathbb{R}^k} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v})$. When there is no ambiguity about the loss function, we drop the superscript $\mathcal{L}$ and simply write $C_{\mathbf{p}}(\mathbf{v})$ and $C_{\mathbf{p}}^*$.

This terminology was used in Bartlett et al. [BJM06]. It was also called *inner $\mathcal{L}$-risk* by Steinwart [Ste07].

The following is from Zhang [Zha04a, Definition 1].

Definition IV.9. A loss $\mathcal{L}$ has the *infinite-sample consistency* (ISC) property if for all $\mathbf{p} \in \Delta^k$ and y such that $p_y < \max_j p_j$, we have $C_{\mathbf{p}}^{\mathcal{L},*} < \inf \left\{ C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_y = \max \mathbf{v} \right\}$.

As explained in Zhang [Zha04a], the name "infinite-sample consistency" is chosen precisely because the property in Definition IV.9 implies that "$\mathcal{L}$-surrogate risk minimization is 01-consistent". See Theorem IV.12 below. Next, we review the closely related concept of multiclass classification-calibration as developed in Tewari et al. [TB07].

Definition IV.10 (Range and its convex hull)**.** Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function. Denote by $\mathcal{R}(f) := \{f(x) : x \in \mathbb{R}^m\}$ the *range* of f. Define $\mathcal{S}(f) := \operatorname{conv}(\mathcal{R}(f))$ to be the convex hull of the range of f.

The following is from Tewari et al. [TB07, Definition 1].

Definition IV.11. A set $S \subseteq \mathbb{R}^k_+$ is *classification-calibrated* if there exists a function[6] $\theta : \mathbb{R}^k \to [k]$ such that

$$\inf\{\langle \mathbf{p}, \boldsymbol{\zeta} \rangle : \boldsymbol{\zeta} \in S \,:\, p_{\theta(\boldsymbol{\zeta})} < \max \mathbf{p}\} > \inf_{\boldsymbol{\zeta} \in S} \langle \mathbf{p}, \boldsymbol{\zeta} \rangle \tag{4.8}$$

for all $\mathbf{p} \in \Delta^k$.

Intuitively, Definition IV.11 says that the lowest achievable conditional risk when predicting the wrong label (Eqn. (4.8) LHS) is still strictly larger than the conditional Bayes risk (Eqn. (4.8) RHS). Formally, the importance of Definitions IV.9 and IV.11 is manifested by the following theorem, which paraphrases Zhang [Zha04a, Theorem 3] and one implication[7] of Tewari et al. [TB07, Theorem 2] when $\mathcal{L}$ is a permutation equivariant loss. Define $\underline{\arg\max} : \mathbb{R}^k \to [k]$ by $\underline{\arg\max}(v) = \min\{i \in [k] : v_i = \max_{j \in [k]} v_j\}$. When $\mathcal{L}$ is permutation equivariant and classification-calibrated, we can assume that θ from Definition IV.11 is $\underline{\arg\max}$. See [TB07, Lemma 4].

Theorem IV.12 ([Zha04a; TB07])**.** *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_+$ be a permutation equivariant loss function. Let $\mathcal{F}$ be the set of Borel functions $\mathcal{X} \to \mathbb{R}^k$. If either $\mathcal{S}(\mathcal{L})$ is classification-calibrated or $\mathcal{L}$ has the ISC property, then* ***$\mathcal{L}$-surrogate risk minimization is 01-consistent****, namely: For all sequence of function classes $\{\mathcal{F}_n\}_n$ such that $\mathcal{F}_n \subseteq \mathcal{F}$, $\bigcup_n \mathcal{F}_n = \mathcal{F}$, $\hat{f}_n \in \mathcal{F}_n$ and all data generating probability distribu-*

[6] The function θ is called a *calibrated link* for S.

[7] Tewari et al. [TB07, Theorem 2] says the other implication is true as well: $\mathcal{L}$-surrogate risk minimization being 01-consistent implies that $\mathcal{S}(\mathcal{L})$ is classification-calibrated. However, we do not need the implication in this direction. It is nevertheless a curious question if there exists $\mathcal{L}$ having the ISC property when $\mathcal{S}(\mathcal{L})$ is not classification-calibrated.

tions P

$$R_{\mathcal{L}}(\hat{f}_n) \xrightarrow{P} R^*_{\mathcal{L}} \quad \textit{implies} \quad R_{01}(\underline{\arg\max} \circ \hat{f}_n) \xrightarrow{P} R^*_{01}.$$

4.3 Sufficient conditions for classification-calibration

In this section is divided into three subsections, each containing a sufficient condition of classification-calibration of the loss family in the subsection's title.

4.3.1 Gamma-Phi loss

In this section, we consider the Gamma-Phi loss as in Example IV.5.

Definition IV.13 (Conditions on γ)**.** Let $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a function with the property that $\sup_{x \in [0,\infty)} \gamma(x) = +\infty$. We say that γ satisfies condition (G1) if γ is strictly increasing, i.e., $\gamma(x) < \gamma(\tilde{x})$ if $x < \tilde{x}$, and condition (G2) if γ is continuously differentiable and $\frac{d\gamma}{dx}(x) > 0$ for all $x \geq 0$.

Note that condition (G2) implies condition (G1), but the converse is not true.

Definition IV.14 (Condition on ϕ)**.** Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a function with the property that $\sup_{x \in \mathbb{R}} \phi(x) = 0$. We say that ϕ satisfies condition (F) if ϕ is differentiable where $\frac{d\phi}{dx}(x) \leq 0$ for all $x \in \mathbb{R}$, and $\frac{d\phi}{dx}(0) < 0$.

Theorem IV.15. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where γ satisfies Definition IV.13 condition (G2), and ϕ satisfies Definition IV.14 condition (F). Then $\mathcal{L}$ has the ISC property.*

In light of Theorem IV.12, if $\mathcal{L}$ satisfies the conditions of Theorem IV.15, then $\mathcal{L}$-surrogate risk minimization is 01-consistent. As stated in the introduction, Theorem IV.15 establishes the first sufficient condition of CC for Gamma-Phi loss. The only previously proposed sufficient condition by Pires et al. [PS16] turns out to be insufficient. See Section 4.9.1.

Remark IV.16. Both the coherence loss and pairwise comparison loss (see Example IV.5) clearly satisfy the conditions of Theorem IV.15. On the other hand, the ISC property for both of these two losses have not been established previously. For the coherence loss, Zhang et al. [Zha+09] proves only a restricted form of ISC, i.e., when $\mathbf{p} > 0$ entrywise in Definition IV.9. For the pairwise comparison loss, the proof of the sufficient condition for ISC, i.e., Zhang [Zha04a, Theorem 6], explicitly omits the edge case where the minimizers in Definition IV.9 occur at infinity and asserts that the extension to handle this edge case is trivial. In Section 4.9, we handle this edge case which turn out to be rather involved. Moreover, we significantly generalize the result of Zhang [Zha04a] to cover Gamma-Phi losses, a much larger family encompassing pairwise comparison losses.

Remark IV.17. The multiclass savage loss [SV19] is a Gamma-Phi loss with $\gamma(x) = (x/(1+x))^2$ and $\phi(x) = \exp(-2x)$ which does not satisfy the condition of Theorem IV.15. More precisely, the condition $\sup_{x\in[0,\infty)} \gamma(x) = +\infty$ fails. While the binary savage loss is classification-calibrated [MV08], to the best of our knowledge it is unknown whether the multiclass savage loss has the ISC property.

4.3.2 Fenchel-Young loss

In this section, we consider the Fenchel-Young loss as in Example IV.6. Define the *reduced k-probability simplex* as

$$\tilde{\Delta}^k := \{\tilde{\mathbf{p}} := (p_2, \ldots, p_k) \in [0,1]^k : \sum_{i=2}^{k} p_i \leq 1\}. \tag{4.9}$$

In other words, $\tilde{\Delta}^k$ is simply Δ^k without the first coordinate. To every function $\Omega : \Delta^k \to \mathbb{R}$ with domain on the k-simplex, we define a corresponding function

$\tilde{\Omega} : \tilde{\Delta}^k \to \mathbb{R}$ called the *reduced form* of Ω,

$$\tilde{\Omega}(\tilde{\mathbf{p}}) := \Omega\left(1 - \sum_{i=2}^{k} p_i, p_2, \ldots, p_k\right), \quad \forall \tilde{\mathbf{p}} = (p_2, \ldots, p_k)^\top \in \tilde{\Delta}^k. \tag{4.10}$$

Clearly, Eqn. 4.10 gives a one-to-one correspondence between functions $\Omega : \Delta^k \to \mathbb{R}$ on the simplex Δ^k and functions $\tilde{\Omega} : \tilde{\Delta}^k \to \mathbb{R}$ on the reduced simplex $\tilde{\Delta}^k$.

Next, we briefly review the theory of convex analysis and Legendre transformation following Rockafellar [Roc70, Section 26]

Definition IV.18. Let $D \subseteq \mathbb{R}^n$ be a closed convex set. Let $f : D \to \mathbb{R}$ be a function. Define $D^* := \{y \in \mathbb{R}^n : \sup_{x \in D} \langle y, x \rangle - f(x) < \infty\}$. The *convex conjugate* of a function $f : D \to \mathbb{R}$ is the function $f^* : D^* \to \mathbb{R}$ given by

$$f^*(y) = \sup_{x \in D} \langle y, x \rangle - f(x).$$

Definition IV.19. Let $D \subseteq \mathbb{R}^n$ be a closed convex set. A convex function $f : D \to \mathbb{R}$ is said to be of *Legendre type* if

1. $C := \mathrm{int}(D)$ is an open convex subset of $\mathbb{R}^n$,
2. f is strictly convex and differentiable on C,
3. for all sequences $\{x_i\} \subseteq C$ such that $\lim_{i \to \infty} x_i \in \mathtt{bdry}(D)$ we have $\lim_{i \to \infty} \|\nabla_f(x_i)\| = +\infty$.

For example, when $D = \tilde{\Delta}^k$ and $f = -H$ is the negative Shannon entropy, then $f : D \to \mathbb{R}_{\leq 0}$ is of Legendre type. See paragraph immediately following Blondel et al. [BMN20, Definition 3].

Definition IV.20 (Regular negentropy)**.** A function $\Omega : \Delta^k \to \mathbb{R}$ is a *negentropy* if :

1. Ω is closed (maps closed sets to closed sets) and convex,
2. Ω is symmetric, i.e., $\Omega(\sigma(\mathbf{p})) = \Omega(\mathbf{p})$ for all $\mathbf{p} \in \Delta^k$ and $\sigma \in \mathtt{Sym}(k)$,
3. $-\Omega(\mathbf{p}) \geq 0$ for all $\mathbf{p} \in \Delta^k$ and $\Omega(\mathbf{e}_i^k) = 0$ for all $i \in [k]$.

If in addition the reduced form $\tilde{\Omega}$ is of Lengedre type and twice differentiable, then Ω is a *regular* negentropy.

The term "negentropy" was previously used by Mensch et al. [MBP19], although the origin of the term is unclear. To the best of our knowledge, the definition of a *regular* negentropy is new. Since Eqn. 4.10 is a one-to-one correspondence between functions on the simplex Δ^k and functions on the reduced simplex $\tilde{\Delta}^k$, we sometimes refer to a negentropy by its reduced form $\tilde{\Omega}$.

For $n \in \{2, \dots, k\}$, define $\texttt{inj}_{n-1}^{k-1} : \tilde{\Delta}^n \to \tilde{\Delta}^k$ by padding zeros, i.e.,

$$\texttt{inj}_{n-1}^{k-1}(\tilde{\mathbf{q}}) = (q_2, \dots, q_n, 0, \dots, 0) \in \tilde{\Delta}^k, \quad \forall \mathbf{q} = (q_2, \dots, q_n) \in \tilde{\Delta}^n.$$

Definition IV.21 (Totally regular negentropy)**.** Let $\tilde{\Omega} : \tilde{\Delta}^k \to \mathbb{R}$ be a negentropy and $n \in \{2, \dots, k\}$. The *n-ary retracted negentropy* of $\tilde{\Omega}$, which we denote by $\tilde{\Omega}^{(n)} : \tilde{\Delta}^n \to \mathbb{R}$, is defined as

$$\tilde{\Omega}^{(n)}(\mathbf{q}) := \tilde{\Omega}(\texttt{inj}_{n-1}^{k-1}(\mathbf{q})), \quad \forall \mathbf{q} \in \tilde{\Delta}^n.$$

We say that Ω is a *totally regular* negentropy if $\Omega^{(n)}$ is a regular negentropy for each $n \in \{2, \dots, k\}$.

To the best of our knowledge, the definition of a totally regular negentropy is new. The next result establishes it as a sufficient condition for classification-calibration:

Theorem IV.22. *Let Ω be a totally regular negentropy, $\mu \in \mathbb{R}_+$ be fixed, and $\mathcal{L}$ be the Fenchel-Young loss associated to Ω and the μ. Then $\mathcal{S}(\mathcal{L})$ is classification-calibrated.*

In light of Theorem IV.12, if Ω satisfies the conditions of Theorem IV.22, then $\mathcal{L}$-surrogate risk minimization is 01-consistent. As stated in the abstract, Theorem IV.22 recovers all known classification-calibration sufficient conditions [NBR19;

Blo19]. Moreover, Theorem IV.22 establish classification-calibration for Fenchel-Young losses not satisfying previously known sufficient conditions. See Section 4.8.2.

4.3.3 Regular PERM losses

In this section, we define *regular* and *totally regular* PERM losses, which generalizes Fenchel-Young losses with regular and totally regular negentropies, respectively.

Definition IV.23. A function $f : \mathbb{R}^n \to \mathbb{R}$ is

1. *coercive* if for all $c \in \mathbb{R}$, the *c-sublevel set* $\{\mathbf{v} \in \mathbb{R}^n : f(\mathbf{v}) \leq c\}$ is bounded,
2. *semi-coercive* if for all $c \in \mathbb{R}$ there exists $b \in \mathbb{R}$ such that

$$\{\mathbf{v} \in \mathbb{R}^n : f(\mathbf{v}) \leq c\} \subseteq \{\mathbf{v} \in \mathbb{R}^n : \min \mathbf{v} \geq b\}.$$

The definition of a coercive function is well-known. However, semi-coercivity appears to be a novel concept. Intuitively, a function is semi-coercive if, for all $c \in \mathbb{R}$, its c-sublevel set is contained in a translate of the positive orthant.

Definition IV.24 (Regular PERM loss)**.** Let $\mathcal{L}$ be a PERM loss with template ψ. We say that $\mathcal{L}$ is *regular* if ψ is nonnegative, twice differentiable, strictly convex, semi-coercive, the partial derivative $\frac{\partial \psi}{\partial z_1} : \mathbb{R}^{k-1} \to \mathbb{R}$ is semi-bounded, and the gradient $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ is entrywise negative for all $\mathbf{z} \in \mathbb{R}^{k-1}$.

We note that the condition $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ in Definition IV.24 is reminiscient of a condition in Bartlett et al. [BJM06, Theorem 6], which shows that in the binary case a convex margin loss ψ is classification-calibrated if and only if ψ is differentiable at 0 and $\psi'(0) < 0$. However, Definition IV.24 requires the strict negativity of the gradient for all of $\mathbb{R}^{k-1}$, whereas the derivative of ψ is only required to be negative at 0 in Bartlett et al. [BJM06, Theorem 6]. Compared to the binary case, the multiclass case ostensibly requires much stronger assumption to establish classification-calibration. Future work will investigate whether this can be weakened.

Proposition IV.25 (Retraction of a PERM loss)**.** *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_{\geq 0}$ be a regular PERM loss with template $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$. Define a function called the* retraction *of ψ by $\texttt{ret}[\psi] : \mathbb{R}^{k-2} \to \mathbb{R}$ by $\texttt{ret}[\psi](\mathbf{w}) := \lim_{\lambda \to \infty} \psi\left(\begin{bmatrix}\mathbf{w}^\top & \lambda\end{bmatrix}^\top\right)$ for all $\mathbf{w} \in \mathbb{R}^{k-1}$. Then $\texttt{ret}[\psi]$ is a well-defined (the limit exists in $\mathbb{R}$) symmetric function.*

Proof. The condition that $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ implies that the function $\lambda \mapsto \psi\left(\begin{bmatrix}\mathbf{w}^\top & \lambda\end{bmatrix}^\top\right)$ is (strictly) decreasing as a function of λ. Thus, the limit exists for all $\mathbf{w}$. The symmetry of $\texttt{ret}[\psi]$ follows immediately from the symmetry of ψ. □

Definition IV.26 (Totally regular PERM loss)**.** Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_{\geq 0}$ be a regular PERM loss with template $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$. For each $n \in \{2, \ldots, k\}$, define the symmetric functions $\psi^{(n)} : \mathbb{R}^{n-1} \to \mathbb{R}$ by $\psi^{(n)} := \underbrace{\texttt{ret} \circ \cdots \circ \texttt{ret}}_{(k-n)\text{-times}}[\psi]$. Let $\mathcal{L}^{(n)}$ be the PERM loss associated to $\psi^{(n)}$ (see Proposition IV.3). Below, we refer to $\mathcal{L}^{(n)}$ as the *n-ary retracted loss* associated to $\mathcal{L}$. We say that $\mathcal{L}$ is a *totally regular* PERM loss if $\mathcal{L}^{(n)}$ is regular for each $n \in \{2, \ldots, k\}$.

The n-ary retracted loss captures the behavior of ψ when the last $k - n$ inputs to the function approach $+\infty$. We now state our main theorem:

Theorem IV.27. *If $\mathcal{L}$ is totally regular, then $\mathcal{S}(\mathcal{L})$ is classification-calibrated.*

We will see in Section 4.9 that Theorem IV.27 implies sufficient condition for classificaiton-calibration of Fenchel-Young loss (Theorem IV.22). On the other hand, proof of the analogous result for Gamma-Phi loss, i.e., Theorem IV.15, requires a different set of techniques introduced in the following section.

4.4 Conditional risks of permutation equivariant losses

In this section, we study some of the basic properties of the conditional risk (Definition IV.8) of permutation equivariant losses Definition IV.2 part 1.

Lemma IV.28. *Let $\mathcal{L}$ be a permutation-equivariant loss. Let $\sigma \in \mathtt{Sym}(k)$, $\mathbf{v} \in \mathbb{R}^k$ and $\mathbf{p} \in \Delta^k$ be arbitrary. Then $C_{\mathbf{p}}(\mathbf{v}) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v}))$. Furthermore, we have $C^*_{\mathbf{p}} = C^*_{\sigma(\mathbf{p})}$.*

Proof. For the first assertion, we have

$$C_{\mathbf{p}}(\mathbf{v}) = \sum_{y \in [k]} p_y \mathcal{L}_y(\mathbf{v}) = \sum_{y \in [k]} p_{\sigma(y)} \mathcal{L}_{\sigma(y)}(\mathbf{v}) = \sum_{y \in [k]} [\sigma(\mathbf{p})]_y \mathcal{L}_y(\sigma(\mathbf{v})) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v})).$$

For the "Furthermore" part, note that $\sigma : \mathbb{R}^k \to \mathbb{R}^k$ is a bijection. Hence,

$$\begin{aligned} C^*_{\mathbf{p}} &= \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k\} \\ &= \inf\{C_{\mathbf{p}}(\sigma^{-1}(\mathbf{v})) : \mathbf{v} \in \mathbb{R}^k\} \\ &= \inf\{C_{\sigma(\mathbf{p})}(\sigma(\sigma^{-1}(\mathbf{v}))) : \mathbf{v} \in \mathbb{R}^k\}. \end{aligned}$$

The right hand side is equal to $\inf\{C_{\sigma(\mathbf{p})}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k\} = C^*_{\sigma(\mathbf{p})}$. □

Lemma IV.29. *Suppose that $\mathcal{L}$ is permutation equivariant. Let $\mathbf{p} \in \Delta^k$, $y, y' \in [k]$ and $\mathbf{v} \in \mathbb{R}^k$. Let $\tau \in \mathtt{Sym}(k)$ be the* transposition *of y and y', i.e., $\tau(y) = y'$, $\tau(y') = y$ and $\tau(j) = j$ for all $j \in [k] \setminus \{y, y'\}$. Then $C_{\mathbf{p}}(\mathbf{v}) - C_{\mathbf{p}}(\tau(\mathbf{v})) = (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v}))$.*

Proof. This is a straightforward computation:

$$\begin{aligned}
&C_{\mathbf{p}}(\mathbf{v}) - C_{\mathbf{p}}(\tau(\mathbf{v})) \\
&= \left(\sum_{j\in[k]} p_j \mathcal{L}_j(\mathbf{v})\right) - \left(\sum_{j\in[k]} p_j \mathcal{L}_j(\tau(\mathbf{v}))\right) \\
&= \left(\sum_{j\in[k]} p_j \mathcal{L}_j(\mathbf{v})\right) - \left(\sum_{j\in[k]} p_j \mathcal{L}_{\tau(j)}(\mathbf{v})\right) \quad \because \text{Definition IV.2} \\
&= (p_y \mathcal{L}_y(\mathbf{v}) + p_{y'} \mathcal{L}_{y'}(\mathbf{v})) - (p_y \mathcal{L}_{y'}(\mathbf{v}) + p_{y'} \mathcal{L}_y(\mathbf{v})) \\
&= p_y(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v})) + p_{y'}(\mathcal{L}_{y'}(\mathbf{v}) - \mathcal{L}_y(\mathbf{v})) \\
&= (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v})),
\end{aligned}$$

as desired. □

Proposition IV.30. *Let* $\mathbf{p} \in \Delta^k_{\texttt{desc}}$. *Let* $\mathbf{v} \in \mathbb{R}^k$ *be arbitrary. Let* $\sigma \in \texttt{Sym}(k)$ *be such that* $v_{\sigma(1)} \geq v_{\sigma(2)} \geq \cdots \geq v_{\sigma(k)}$. *Then* $C_{\mathbf{p}}(\mathbf{v}) \geq C_{\mathbf{p}}(\sigma(\mathbf{v}))$.

Proof. This proof is essentially Lemma S3.8 from Wang et al. [WS20] Supplemental Materials. First, we note that if $\tilde{\sigma} \in \texttt{Sym}(k)$ is another permutation such that $v_{\tilde{\sigma}(1)} \geq v_{\tilde{\sigma}(2)} \geq \cdots \geq v_{\tilde{\sigma}(k)}$, then $\tilde{\sigma}(\mathbf{v}) = \sigma(\mathbf{v})$. Thus, it suffices to prove the result while assuming that the permutation σ that sorts $\mathbf{v}$ is given by the *bubble sort* algorithm:

L1. Initialize the iteration index $t \leftarrow 0$ and $\mathbf{v}^0 := \mathbf{v}$,

L2. While there exists $i \in [k]$ such that $v^t_i < v^t_{i+1}$, do

   (a) Let $\tau^t \in \texttt{Sym}(k)$ be the permutation that swaps i and $i+1$, leaving other indices unchanged.

   (b) $\mathbf{v}^{t+1} \leftarrow \tau^t(\mathbf{v}^t)$

   (c) $t \leftarrow t + 1$

L3. Output $\mathbf{v}^T$, where $T \leftarrow t$ is the final iteration index.

Let $\langle \cdot, \cdot \rangle$ be the ordinary dot product on $\mathbb{R}^k$. Note that $C_{\mathbf{p}}(\mathbf{v}) = \langle \mathbf{p}, \mathcal{L}(\mathbf{v}) \rangle$.

Furthermore, at termination, there exists $\sigma \in \mathtt{Sym}(k)$ such that $\mathbf{v}^T = \sigma(\mathbf{v})$ is sorted as in the statement of Proposition IV.30. We claim that at every intermediate step $t \in \{0, \ldots, T\}$, we have $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle \geq \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^{t+1}) \rangle$. This would prove Proposition IV.30, since $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^0) \rangle = C_{\mathbf{p}}(\mathbf{v})$ and $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^T) \rangle = C_{\mathbf{p}}(\sigma(\mathbf{v}))$.

Now, towards proving our claim, let t be an intermediate iteration of the above "bubble sort" algorithm, and let $i \in [k]$ be as in L2. Then we have

$$\begin{aligned}
&\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle - \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^{t+1}) \rangle \\
&= \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle - \langle \mathbf{p}, \mathcal{L}(\tau^t(\mathbf{v}^t)) \rangle \qquad \because \text{Definition on L2.(b)} \\
&= (p_i - p_{i+1})(\mathcal{L}_i(\mathbf{v}^t) - \mathcal{L}_{i+1}(\mathbf{v}^t)) \geq 0, \qquad \text{Lemma IV.29}
\end{aligned}$$

as desired. □

4.5 Multiplicative label encoding

The goal of this section is to prove Proposition IV.3. In the following definition, we introduce the *multiplicative label code*, a set of matrices $\{\boldsymbol{\rho}_1^{(k)}, \ldots, \boldsymbol{\rho}_k^{(k)}\}$ generalizing of the familiar $\{\pm 1\}$ label in binary classification to the k-ary multiclass classification.

Definition IV.31 (Multiplicative label code)**.** For $k \geq 2$ and $i \in [k]$, define matrices $\boldsymbol{\rho}_i^{(k)} \in \mathbb{R}^{(k-1)\times(k-1)}$ as follows: For $i = 1$, $\boldsymbol{\rho}_1^{(k)}$ is the identity. For $i \in \{2, \ldots, k\}$, define $\boldsymbol{\rho}_i^{(k)}$ column-wise by

$$[\boldsymbol{\rho}_i^{(k)}]_{:j} := \begin{cases} \mathbb{e}_j^{(k)} & : j \neq i-1 \\ -\mathbb{1}_k & : j = i-1, \end{cases} \qquad \text{for each } j \in \{1, \ldots, k-1\}.$$

When there is no ambiguity, we write $\boldsymbol{\rho}_i$ to denote $\boldsymbol{\rho}_i^{(k)}$.

Note that for $i \in \{2, \ldots, k\}$, the matrix $\boldsymbol{\rho}_i^{(k)}$ acts on a vector $\mathbf{z} = (z_1, \ldots, z_{k-1})^\top \in$

$\mathbb{R}^{k-1}$ by

$$\forall j \in [k-1],\ \left[\boldsymbol{\rho}_i^{(k)}\mathbf{z}\right]_j := \begin{cases} z_j - z_{i-1} & : j \neq i-1 \\ -z_{i-1} & : j = i-1. \end{cases} \tag{4.11}$$

Definition IV.32. Define the linear map $\mathbf{M}^{(k)} : \mathbb{R}^k \to \mathbb{R}^{k-1}$

$$\mathbf{M}^{(k)}(\mathbf{v}) = (v_1 - v_2, v_1 - v_3, \ldots, v_1 - v_k)^\top, \text{ where } \mathbf{v} = (v_1, \ldots, v_k)^\top \in \mathbb{R}^k.$$

Observe that $[\mathbf{M}^{(k)}(\mathbf{v})]_i = v_1 - v_{i+1}$ for $i \in [k-1]$. When there is no ambiguity, we write $\mathbf{M} = \mathbf{M}^{(k)}$. Note that $\mathbf{M}^{(k)} = \begin{bmatrix} \mathbb{1}_{k-1} & -\mathbf{Id}_{k-1} \end{bmatrix}$ as a matrix.

Proposition IV.33. *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k$ be a PERM loss with template ψ and reduced form ℓ. Then $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$ is symmetric and for all $y \in [k]$ and $\mathbf{v} \in \mathbb{R}^k$, we have*

$$[\mathcal{L}(\mathbf{v})]_y = \ell_y(\mathbf{M}\mathbf{v}) = \psi(\boldsymbol{\rho}_y\mathbf{M}\mathbf{v}). \tag{4.12}$$

Conversely, given a symmetric function $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$, define $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k$ by $[\mathcal{L}(\mathbf{v})]_y = \psi(\boldsymbol{\rho}_y\mathbf{M}\mathbf{v})$. Then $\mathcal{L}$ is a PERM loss whose template is ψ.

The proof of Proposition IV.33 will be given at the end of this section after developing the necessary machinery.

Lemma IV.34. *For all $i \in [k]$, we have $\mathbf{M}\mathbf{T}_i = \boldsymbol{\rho}_i\mathbf{M}$. In particular, for all $i > 1$ and $j \in [k-1]$, we have*

$$[\boldsymbol{\rho}_i\mathbf{M}\mathbf{v}]_j = \begin{cases} v_i - v_{j+1} & : i \neq j+1 \\ v_i - v_1 & : i = j+1. \end{cases}$$

Proof. If $i = 1$, then $\mathbf{T}_i$ and $\boldsymbol{\rho}_i$ are both identity matrices and there is nothing to show. Otherwise, suppose that $i > 1$. Consider $\mathbf{v} \in \mathbb{R}^k$. We first calculate $\mathbf{M}\mathbf{T}_i\mathbf{v}$.

For each $j \in [k-1]$, we have

$$[\mathbf{M}\mathbf{T}_i\mathbf{v}]_j = [\mathbf{T}_i\mathbf{v}]_1 - [\mathbf{T}_i\mathbf{v}]_{j+1} = v_i - v_{\tau_i(j+1)} = \begin{cases} v_i - v_{j+1} & : i \neq j+1 \\ v_i - v_1 & : i = j+1. \end{cases} \tag{4.13}$$

Next, we compute $\boldsymbol{\rho}_i\mathbf{M}\mathbf{v}$. Using Eqn. 4.11, we have for each $j \in [k-1]$ that

$$[\boldsymbol{\rho}_i\mathbf{M}\mathbf{v}]_j = \begin{cases} [\mathbf{M}\mathbf{v}]_j - [\mathbf{M}\mathbf{v}]_{i-1} & : j \neq i-1 \\ -[\mathbf{M}\mathbf{v}]_{i-1} & : j = i-1. \end{cases}$$

For the $j \neq i-1$ case, we have

$$[\mathbf{M}\mathbf{v}]_j - [\mathbf{M}\mathbf{v}]_{i-1} = (v_1 - v_{j+1}) - (v_1 - v_i) = v_i - v_{j+1}.$$

For the $i = j+1$ case, we have

$$-[\mathbf{M}\mathbf{v}]_{i-1} = -(v_1 - v_i) = v_i - v_1.$$

Thus, $[\mathbf{M}\mathbf{T}_i\mathbf{v}]_j = [\boldsymbol{\rho}_i\mathbf{M}\mathbf{v}]_j$ for all j, which implies that $\mathbf{M}\mathbf{T}_i\mathbf{v} = \boldsymbol{\rho}_i\mathbf{M}\mathbf{v}$. Since $\mathbf{v}$ was arbitrary, we have $\mathbf{M}\mathbf{T}_i = \boldsymbol{\rho}_i\mathbf{M}$. □

Let $\mathbf{M}^\dagger$ denote the Moore-Penrose inverse of $\mathbf{M}$. Since $\mathbf{M}$ is surjective, $\mathbf{M}\mathbf{M}^\dagger$ is the identity. Define a mapping $\boldsymbol{\Pi} : \mathtt{Sym}(k) \to \mathbb{R}^{(k-1)\times(k-1)}$ by

$$\boldsymbol{\Pi}(\sigma) := \mathbf{M}\mathbf{S}_\sigma\mathbf{M}^\dagger.$$

Lemma IV.35. *For all $i \in [k]$, we have $\boldsymbol{\Pi}(\tau_i) = \boldsymbol{\rho}_i$.*

Proof. Lemma IV.34 says that $\boldsymbol{\rho}_i\mathbf{M} = \mathbf{M}\mathbf{T}_i$ which implies that $\boldsymbol{\rho}_i = \boldsymbol{\rho}_i\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}\mathbf{T}_i\mathbf{M}^\dagger = \boldsymbol{\Pi}(\tau_i)$. □

Lemma IV.36. *For all $\sigma, \sigma' \in \mathtt{Sym}(k)$, we have $\mathbf{\Pi}(\sigma\sigma') = \mathbf{\Pi}(\sigma)\mathbf{\Pi}(\sigma')$.*

Proof. Unwinding the definition of $\mathbf{\Pi}$, it suffices to show $\mathbf{M}\mathbf{S}_\sigma\mathbf{S}_{\sigma'}\mathbf{M}^\dagger = \mathbf{M}\mathbf{S}_\sigma\mathbf{M}^\dagger\mathbf{M}\mathbf{S}_{\sigma'}\mathbf{M}^\dagger$. First, we observe that $\ker \mathbf{M} = \{\mathbf{v} \in \mathbb{R}^k : v_1 = v_2 = \cdots = v_k\}$ which follows directly from the definition of $\mathbf{M}$. Next, let $\mathcal{R} \subseteq \mathbb{R}^k$ denote the subspace $\mathcal{R} := \{\mathbf{v} \in \mathbb{R}^k : v_1 + \cdots + v_k = 0\}$. Then we have $\{\mathbf{v} \in \mathbb{R}^k : v_1 = v_2 = \cdots = v_k\} = \mathcal{R}^\perp$ is the subspace of vectors orthogonal to $\mathcal{R}$. A fundamental result in linear algebra states that $\operatorname{ran} \mathbf{M}^\dagger = \operatorname{ran} \mathbf{M}^\top = (\ker \mathbf{M})^\perp$. Thus, $\operatorname{ran} \mathbf{M}^\dagger = \mathcal{R}$. Taken together, if we let $\mathbf{P} := \mathbf{M}^\dagger\mathbf{M} \in \mathbb{R}^{k\times k}$, then $\mathbf{P}$ is a projection matrix on $\mathcal{R}$. Thus, $\mathbf{P}(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v} \in \mathcal{R}$. Since $\mathbf{S}_{\sigma'}$ is a permutation matrix, we have $\mathbf{S}_{\sigma'}(\mathcal{R}) \subseteq \mathcal{R}$. Thus, $\operatorname{ran}(\mathbf{S}_{\sigma'}\mathbf{M}^\dagger) \subseteq \mathcal{R}$ which implies that $\mathbf{P}\mathbf{S}_{\sigma'}\mathbf{M}^\dagger = \mathbf{S}_{\sigma'}\mathbf{M}^\dagger$. This proves $\mathbf{M}\mathbf{S}_\sigma\mathbf{S}_{\sigma'}\mathbf{M}^\dagger = \mathbf{M}\mathbf{S}_\sigma\mathbf{P}\mathbf{S}_{\sigma'}\mathbf{M}^\dagger = \mathbf{M}\mathbf{S}_\sigma\mathbf{M}^\dagger\mathbf{M}\mathbf{S}_{\sigma'}\mathbf{M}^\dagger$ as desired.

□

Lemma IV.37. *For all $i \in [k]$, $\boldsymbol{\rho}_i^2$ is the identity.*

Proof. Using Lemma IV.35 and Lemma IV.36, we have $\boldsymbol{\rho}_i^2 = \mathbf{\Pi}(\tau_i)\mathbf{\Pi}(\tau_i) = \mathbf{\Pi}(\tau_i^2)$. Since τ_i is a transposition, τ_i^2 is the identity. Thus, $\boldsymbol{\rho}_i^2$ is also the identity. □

Lemma IV.38. *Let $i_1, i_2 \in \{2, \ldots, k\}$ be distinct. Then $\tau_{i_1}\tau_{i_2}\tau_{i_1} = \tau_{(i_1,i_2)}$ and $\mathbf{T}_{i_1}\mathbf{T}_{i_2}\mathbf{T}_{i_1} = \mathbf{T}_{(i_1,i_2)}$.*

Proof. This is simply an exhaustive case-by-case proof over all inputs $j \in [k]$. First, let $j = 1$. Then $\tau_{(i_1,i_2)}(1) = 1$ since $1 \notin \{i_1, i_2\}$. On the other hand $\tau_{i_1}\tau_{i_2}\tau_{i_1}(1) = \tau_{i_1}\tau_{i_2}(i_1) = \tau_{i_1}(i_1) = 1$. Now, let $j \in \{2, \ldots, k\}$. If $j \notin \{i_1, i_2\}$, then $\tau_{(i_1,i_2)}(j) = j$ and $\tau_{i_1}\tau_{i_2}\tau_{i_1}(j) = \tau_{i_1}\tau_{i_2}(j) = \tau_{i_1}(j) = j$. If $j = i_1$, then $\tau_{(i_1,i_2)}(i_1) = i_2$ and $\tau_{i_1}\tau_{i_2}\tau_{i_1}(i_1) = \tau_{i_1}\tau_{i_2}(1) = \tau_{i_1}(i_2) = i_2$. If $j = i_2$, then $\tau_{(i_1,i_2)}(i_2) = i_1$ and $\tau_{i_1}\tau_{i_2}\tau_{i_1}(i_2) = \tau_{i_1}\tau_{i_2}(i_2) = \tau_{i_1}(1) = i_1$. □

Corollary IV.39. *Every $\sigma \in \mathtt{Sym}(k)$ can be written as a product $\sigma = \tau_{i_1}\tau_{i_2}\cdots\tau_{i_l}$.*

Proof. We prove the equivalent statement that the set $\mathcal{S} := \{\tau_i : i \in \{2, \ldots, k\}\}$ generates the group $\mathtt{Sym}(k)$. A standard result in group theory states that the set of transpositions $\mathcal{T}$ generates $\mathtt{Sym}(k)$. By Lemma IV.38, transpositions between labels in $\{2, \ldots, k\}$ can be generated by $\mathcal{S}$. Furthermore, $\tau_i = \tau_{(1,i)}$ by definition, so transposition between 1 and elements of $\{2, \ldots, k\}$ can be generated by $\mathcal{S}$ as well. Hence, all of $\mathcal{T}$ can be generated by $\mathcal{S}$. □

Lemma IV.40. *Let $i_1, i_2 \in \{2, \ldots, k\}$ be distinct. Then $\mathbf{T}_{(i_1-1,i_2-1)} = \boldsymbol{\rho}_{i_1}\boldsymbol{\rho}_{i_2}\boldsymbol{\rho}_{i_1}$.*

Proof. First, we note that

$$\begin{aligned}
\boldsymbol{\rho}_{i_1}\boldsymbol{\rho}_{i_2}\boldsymbol{\rho}_{i_1} &= \mathbf{\Pi}(\tau_{i_1})\mathbf{\Pi}(\tau_{i_2})\mathbf{\Pi}(\tau_{i_1}) \quad \because \text{Lemma IV.35} \\
&= \mathbf{\Pi}(\tau_{i_1}\tau_{i_2}\tau_{i_1}) \quad \because \text{Lemma IV.36} \\
&= \mathbf{\Pi}(\tau_{(i_1,i_2)}) \quad \because \text{Lemma IV.38}
\end{aligned}$$

Now, let $\mathbf{v} \in \mathbb{R}^k$ be arbitrary. Then, by definition, for all $j \in [k-1]$, we have

$$[\mathbf{M}\mathbf{T}_{(i_1,i_2)}\mathbf{v}]_j = v_1 - v_{\tau_{(i_1,i_2)}(j+1)}.$$

On the other hand,

$$[\mathbf{T}_{(i_1-1,i_2-1)}\mathbf{M}\mathbf{v}]_j = [\mathbf{M}\mathbf{v}]_{\tau_{(i_1-1,i_2-1)}(j)} = v_1 - v_{\tau_{(i_1-1,i_2-1)}(j)+1}.$$

Since $\tau_{(i_1,i_2)}(j+1) = \tau_{(i_1-1,i_2-1)}(j) + 1$ for all $j \in [k-1]$, we have

$$\mathbf{T}_{(i_1-1,i_2-1)}\mathbf{M}\mathbf{v} = \mathbf{M}\mathbf{T}_{(i_1,i_2)}\mathbf{v}$$

which proves that $\mathbf{T}_{(i_1-1,i_2-1)}\mathbf{M} = \mathbf{M}\mathbf{T}_{(i_1,i_2)}$. To conclude, we have

$$\mathbf{T}_{(i_1-1,i_2-1)} \overset{(1)}{=} \mathbf{T}_{(i_1-1,i_2-1)}\mathbf{M}\mathbf{M}^\dagger \overset{(2)}{=} \mathbf{M}\mathbf{T}_{(i_1,i_2)}\mathbf{M}^\dagger \overset{(3)}{=} \mathbf{\Pi}(\tau_{(i_1,i_2)})$$

where (1) follows from $\mathbf{M}\mathbf{M}^\dagger$ being the identity, (2) follows from multiplying both sides of $\mathbf{T}_{(i_1-1,i_2-1)}\mathbf{M} = \mathbf{M}\mathbf{T}_{(i_1,i_2)}$ on the right by $\mathbf{M}^\dagger$, and (3) follows from the definition of $\mathbf{\Pi}$. □

Corollary IV.41. *For all $i, j \in [k-1]$, we have $\mathbf{T}_{(i,j)}\mathbf{M} = \mathbf{M}\mathbf{T}_{(i+1,j+1)}$.*

Proof. By Lemma IV.40, we have

$$\begin{aligned}
\mathbf{T}_{(i,j)}\mathbf{M} &= \boldsymbol{\rho}_{i+1}\boldsymbol{\rho}_{j+1}\boldsymbol{\rho}_{i+1}\mathbf{M} && \because \text{Lemma IV.40} \\
&= \mathbf{M}\mathbf{T}_{i+1}\mathbf{T}_{j+1}\mathbf{T}_{i+1} && \because \text{Lemma IV.34} \\
&= \mathbf{M}\mathbf{T}_{(i+1,j+1)} && \because \text{Lemma IV.34}
\end{aligned}$$

□

Lemma IV.42. *Let $i, j \in [k]$. Then*

$$\boldsymbol{\rho}_{\tau_i(j)} = \begin{cases} \mathbf{T}_{(j-1,i-1)}\boldsymbol{\rho}_j\boldsymbol{\rho}_i & : i, j \in \{2, \ldots, k\} \\ \boldsymbol{\rho}_j\boldsymbol{\rho}_i & : \textit{otherwise.} \end{cases} \tag{4.14}$$

Proof. Suppose that $i = j$, then $\tau_i(j) = \tau_i(i) = 1$. Hence, the left hand side is the identity by definition. For the right hand side, we observe that $\mathbf{T}_{(i-1,j-1)}$ reduces to the identity element. Furthermore, $\boldsymbol{\rho}_j\boldsymbol{\rho}_i = \boldsymbol{\rho}_i^2$ is also the identity by Lemma IV.37. Thus, below, we may assume that $i \neq j$.

Consider the $i, j \in \{2, \ldots, k\}$ case first. In this case, we must have $\tau_i(j) = j$, thus $\boldsymbol{\rho}_{\tau_i(j)} = \boldsymbol{\rho}_j$. For the right hand side of eq. (4.14), we have

$$\begin{aligned}
\mathbf{T}_{(j-1,i-1)}\boldsymbol{\rho}_j\boldsymbol{\rho}_i &= (\boldsymbol{\rho}_j\boldsymbol{\rho}_i\boldsymbol{\rho}_j)\boldsymbol{\rho}_j\boldsymbol{\rho}_i && \because \text{Lemma IV.40} \\
&= \boldsymbol{\rho}_j\boldsymbol{\rho}_i\boldsymbol{\rho}_i && \because \text{Lemma IV.37} \\
&= \boldsymbol{\rho}_j && \because \text{Lemma IV.37.}
\end{aligned}$$

Next, consider the case when $j = 1$ and $i > 1$. Then $\tau_i(j) = \tau_i(1) = i$. So $\boldsymbol{\rho}_{\tau_i(j)} = \boldsymbol{\rho}_i$. The right hand side of eq. (4.14), we have $\boldsymbol{\rho}_j \boldsymbol{\rho}_i = \boldsymbol{\rho}_i$ since $\boldsymbol{\rho}_j = \boldsymbol{\rho}_1$ is the identity.

Finally, consider the case when $i = 1$ and $j > 1$. Then $\tau_1(j) = j$ and so $\boldsymbol{\rho}_{\tau_i(j)} = \boldsymbol{\rho}_j$. Similar to the previous case, the right hand side of eq. (4.14) is also $\boldsymbol{\rho}_j$. □

Proposition IV.43. *Let $\sigma \in \mathtt{Sym}(k-1)$ and define $\sigma' \in \mathtt{Sym}(k)$ by $\sigma'(1) = 1$ and $\sigma'(i) = \sigma(i-1)$ for $i \in \{2, \ldots, k\}$. Then we have $\mathbf{S}_\sigma \mathbf{M} = \mathbf{M}\mathbf{S}_{\sigma'}$.*

Proof. Let $\sigma = \tau_{(i_1,j_1)}\tau_{(i_2,j_2)} \cdots \tau_{(i_n,j_n)}$. By Corollary IV.41, we have

$$\begin{aligned}
\mathbf{S}_\sigma \mathbf{M} &= \mathbf{T}_{(i_1,j_1)} \mathbf{T}_{(i_2,j_2)} \cdots \mathbf{T}_{(i_n,j_n)} \mathbf{M} \\
&= \mathbf{M} \mathbf{T}_{(i_1+1,j_1+1)} \mathbf{T}_{(i_2+1,j_2+1)} \cdots \mathbf{T}_{(i_n+1,j_n+1)} \\
&= \mathbf{M}\mathbf{S}_{\sigma'}
\end{aligned}$$

where the last equality follows from the observation that $\sigma' = \tau_{(i_1+1,j_1+1)}\tau_{(i_2+1,j_2+1)} \cdots \tau_{(i_n+1,j_n+1)}$. □

Proof of Proposition IV.33. We first check that $\psi : \mathbb{R}^{k-1} \to \mathbb{R}$ is symmetric. Let $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\sigma \in \mathtt{Sym}(k-1)$ be fixed and arbitrary. Our goal is to show that $\psi(\mathbf{S}_\sigma \mathbf{z}) = \psi(\mathbf{z})$.

To this end, first pick $\mathbf{v} \in \mathbb{R}^k$ such that $\mathbf{M}\mathbf{v} = \mathbf{z}$. Define $\sigma \in \mathtt{Sym}(k)$ as in Proposition IV.43. Recall that by definition $\psi(\mathbf{z}) = \ell_1(\mathbf{z})$. Thus, it suffices to show that $\ell_1(\mathbf{S}_\sigma \mathbf{z}) = \ell_1(\mathbf{z})$. This is just a straight forward computation:

$$\begin{aligned}
[\ell(\mathbf{S}_\sigma \mathbf{z})]_1 = [\ell(\mathbf{M}\mathbf{S}_{\sigma'}\mathbf{v})]_1 = [\mathcal{L}(\mathbf{S}_{\sigma'}\mathbf{v})]_1 = [\mathbf{S}_{\sigma'}\mathcal{L}(\mathbf{v})]_1 \\
= [\mathcal{L}(\mathbf{v})]_{\sigma'(1)} = [\mathcal{L}(\mathbf{v})]_1 = [\ell(\mathbf{M}\mathbf{v})]_1 = [\ell(\mathbf{z})]_1.
\end{aligned}$$

This proves that ψ is symmetric.

Next, we prove Eqn. 4.12, i.e., $[\mathcal{L}(\mathbf{v})]_y = \psi(\boldsymbol{\rho}_y \mathbf{M}\mathbf{v})$ for all $y \in [k]$. Again, this is a straight forward computation: $[\mathcal{L}(\mathbf{v})]_y = [\mathbf{T}_y \mathcal{L}(\mathbf{v})]_1 = [\mathcal{L}(\mathbf{T}_y \mathbf{v})]_1 = [\ell(\mathbf{M}\mathbf{T}_y \mathbf{v})]_1 = [\ell(\boldsymbol{\rho}_y \mathbf{M}\mathbf{v})]_1 = \psi(\boldsymbol{\rho}_y \mathbf{M}\mathbf{v})$. This proves Eqn. 4.12.

For the "conversely" part, we note that $\mathcal{L}$ is margin-based by construction. It remains to check that $\mathcal{L}$ is permutation equivariant. Let $\mathbf{v} \in \mathbb{R}^k$ be arbitrary. We claim that $\mathcal{L}(\mathbf{T}_i \mathbf{v}) = \mathbf{T}_i \mathcal{L}(\mathbf{v})$ for all $i \in [k]$. To see this, we have

$$
\begin{aligned}
[\mathbf{T}_i \mathcal{L}(\mathbf{v})]_y &= [\mathcal{L}(\mathbf{v})]_{\tau_i(y)} && \because \text{Definition of } \mathbf{T}_i \\
&= \psi(\boldsymbol{\rho}_{\tau_i(y)} \mathbf{M}\mathbf{v}) && \because \text{Definition of } \mathcal{L} \\
&= \begin{cases} \psi(\mathbf{T}_{(y-1,i-1)} \boldsymbol{\rho}_y \boldsymbol{\rho}_i \mathbf{M}\mathbf{v}) & : i, y \in \{2, \dots, k\} \\ \psi(\boldsymbol{\rho}_y \boldsymbol{\rho}_i \mathbf{M}\mathbf{v}) & : \text{otherwise} \end{cases} && \because \text{Lemma IV.42} \\
&= \psi(\boldsymbol{\rho}_y \boldsymbol{\rho}_i \mathbf{M}\mathbf{v}) && \because \psi \text{ is symmetric} \\
&= \psi(\boldsymbol{\rho}_y \mathbf{M}\mathbf{T}_i \mathbf{v}) && \because \text{Lemma IV.34} \\
&= [\mathcal{L}(\mathbf{T}_i \mathbf{v})]_y && \because \text{Definition of } \mathcal{L}
\end{aligned}
$$

This proves that $\mathbf{T}_i \mathcal{L}(\mathbf{v}) = \mathcal{L}(\mathbf{T}_i \mathbf{v})$.

Now, for an arbitrary $\sigma \in \mathtt{Sym}(k)$, write $\sigma = \tau_{i_1} \cdots \tau_{i_l}$ as in Corollary IV.39. Then we have

$$
\mathcal{L}(\mathbf{S}_\sigma \mathbf{v}) = \mathcal{L}(\mathbf{T}_{i_1} \cdots \mathbf{T}_{i_l} \mathbf{v}) = \mathbf{T}_{i_1} \mathcal{L}(\mathbf{T}_{i_2} \cdots \mathbf{T}_{i_l} \mathbf{v}) = \cdots = \mathbf{T}_{i_1} \cdots \mathbf{T}_{i_l} \mathcal{L}(\mathbf{v}) = \mathbf{S}_\sigma \mathcal{L}(\mathbf{v}).
$$

This proves that $\mathcal{L}$ is permutation equivariant. □

4.6 Regular PERM losses

In this section, we will prove several key properties of regular PERM losses which was introduced in Definition IV.24. Recall that a regular PERM loss has a template

ψ such that ψ is nonnegative, twice differentiable, strictly convex, semi-coercive, the partial derivative $\frac{\partial \psi}{\partial z_1} : \mathbb{R}^{k-1} \to \mathbb{R}$ is semi-bounded, and the gradient $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ is entrywise negative for all $\mathbf{z} \in \mathbb{R}^{k-1}$. Section 4.6.1 focuses on consequences of the semi-coercivity condition. Sections 4.6.2 and 4.6.3 focus on consequences of the other aforementioned conditions. Finally, Section 4.7 presents the proof of one of our main result Theorem IV.27

4.6.1 Semi-coercive functions

Lemma IV.44. *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_+$ be a* nonnegative *PERM loss whose ψ is semi-coercive. Let ℓ be the reduced form of $\mathcal{L}$. Then, for all $\boldsymbol{\zeta} \in \mathbb{R}^k$, the set $\{\mathbf{z} \in \mathbb{R}^{k-1} : \ell(\mathbf{z}) \leq \boldsymbol{\zeta}\}$ is bounded.*

Proof. Observe that

$$\{\mathbf{z} \in \mathbb{R}^{k-1} : \ell(\mathbf{z}) \leq \boldsymbol{\zeta}\} = \bigcap_{i \in [k]} \{\mathbf{z} \in \mathbb{R}^{k-1} : \ell_i(\mathbf{z}) \leq \zeta_i\} \tag{4.15}$$

$$= \bigcap_{i \in [k]} \{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\boldsymbol{\rho}_i \mathbf{z}) \leq \zeta_i\} \quad \because \text{Eqn. (4.12)} \tag{4.16}$$

$$= \bigcap_{i \in [k]} \boldsymbol{\rho}_i \left(\{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\mathbf{z}) \leq \zeta_i\}\right) \tag{4.17}$$

where for the last equality, we used the fact that $\boldsymbol{\rho}_i = \boldsymbol{\rho}_i^{-1}$ (Lemma IV.37) and that

$$\{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\boldsymbol{\rho}_i \mathbf{z}) \leq \zeta_i\} = \boldsymbol{\rho}_i^{-1} \left(\{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\mathbf{z}) \leq \zeta_i\}\right).$$

By assumption, there exists $b_i \in \mathbb{R}$ such that $\{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\mathbf{z}) \leq \zeta_i\} \subseteq \{\mathbf{z} \in \mathbb{R}^{k-1} : \min \mathbf{z} \geq b_i\}$. Putting it all together, we have

$$\{\mathbf{z} \in \mathbb{R}^{k-1} : \ell(\mathbf{z}) \leq \boldsymbol{\zeta}\} \subseteq \bigcap_{i \in [k]} \boldsymbol{\rho}_i(\{\mathbf{z} \in \mathbb{R}^{k-1} : \min \mathbf{z} \geq b_i\}) =: B \tag{4.18}$$

Thus, it suffices to show that B is bounded. Below, we prove this.

Recall that the infinity-norm is defined as $\|\mathbf{u}\|_\infty = \max\{|u_i| : i \in [k-1]\}$. Moreover,

$$\|\mathbf{u}\|_\infty = \max\{|\max \mathbf{u}|, |\min \mathbf{u}|\}. \tag{4.19}$$

Since the empty set is bounded, we may assume that B is nonempty. Define $M_1 = \max\{|b_i| : i \in [k]\}$ and $M_2 = \max\{|b_i + b_j| : i \in [k], j \in [k]\}$. Finally, define $M = \max\{M_1, M_2\}$. To show that B is bounded, if suffices to prove that $\|\mathbf{u}\|_\infty \leq M$ for an arbitrary $\mathbf{u} \in B$. Below, fix such a $\mathbf{u} \in B$.

First, we note that $\min \mathbf{u} \geq b_1$. To see this, recall that $\boldsymbol{\rho}_1$ is the identity. So from Eqn. 4.18 we have $\mathbf{u} \in B \subseteq \{\mathbf{z} \in \mathbb{R}^{k-1} : \min \mathbf{z} \geq b_1\}$.

Let $j \in \arg\min \mathbf{u}$. From Eqn. 4.18, we have $\mathbf{u} \in \boldsymbol{\rho}_{j+1}(\{\mathbf{z} \in \mathbb{R}^{k-1} : \min \mathbf{z} \geq b_{j+1}\})$. Thus, $\boldsymbol{\rho}_{j+1}\mathbf{u} \in \{\mathbf{z} \in \mathbb{R}^{k-1} : \min \mathbf{z} \geq b_{j+1}\}$ and in particular, $[\boldsymbol{\rho}_{j+1}\mathbf{u}]_j \geq b_{j+1}$. Moreover, by Equation (4.11), we have $[\boldsymbol{\rho}_{j+1}\mathbf{u}]_j = -u_j = -\min \mathbf{u}$, and thus $\min \mathbf{u} \leq -b_{j+1}$. Note that we now have $\min \mathbf{u} \in [b_1, -b_{j+1}]$ and, in particular, $|\min \mathbf{u}| \leq M_1$.

Next, let $i \in \arg\max \mathbf{u}$ (and j be as above). First consider the case when $i = j$. Then $\mathbf{u}$ is a constant vector and $\|\mathbf{u}\|_\infty = |\min \mathbf{u}|$. Thus, in this case, we have shown that $\|\mathbf{u}\|_\infty \leq M_1 \leq M$.

Next, consider the case when $i \neq j$. Then we have $[\boldsymbol{\rho}_{i+1}\mathbf{u}]_j = u_j - u_i = (\min \mathbf{u}) - (\max \mathbf{u})$ by Equation (4.11). By similar argument as in the preceding paragraph, we have $[\boldsymbol{\rho}_{i+1}\mathbf{u}]_j \geq b_{i+1}$. Thus, $\max \mathbf{u} \leq \min \mathbf{u} - b_{i+1} \leq -(b_{j+1} + b_{i+1})$. Furthermore, $\max \mathbf{u} \geq \min \mathbf{u} \geq b_1$. Thus, we've shown that $\max \mathbf{u} \in [b_1, -(b_{j+1} + b_{i+1})]$. This implies that $|\max \mathbf{u}| \leq \max\{M_1, M_2\} = M$. Since $|\min \mathbf{u}| \leq M$, by Equation (4.19), we get $\|\mathbf{u}\|_\infty \leq M$. □

Proposition IV.45. *If ψ is semi-coercive, then $C^{\mathcal{L}}_{\mathbf{p}}$ is coercive for all $\mathbf{p} \in \mathtt{int}(\Delta^k)$.*

Proof. Let $C = \{\mathbf{z} \in \mathbb{R}^{k-1} : \langle \mathbf{p}, \ell(\mathbf{z}) \rangle \leq c\}$. Observe that for all $\mathbf{z} \in C$ we have

$$c \geq \langle \mathbf{p}, \ell(\mathbf{z}) \rangle = \sum_{i \in [k]} p_i \psi(\boldsymbol{\rho}_i \mathbf{z}) \geq p_i \psi(\boldsymbol{\rho}_i \mathbf{z})$$

for all $i \in [k]$. Thus, $C \subseteq \bigcap_{i \in [k]} \{\mathbf{z} \in \mathbb{R}^{k-1} : \psi(\boldsymbol{\rho}_i \mathbf{z}) \leq c/p_i\}$. By Lemma IV.44, the right hand side is a bounded set. Hence, C is also bounded. □

4.6.2 The link function

In this section, we study the set of minimizers of the conditional risk of a PERM loss $\mathcal{L}$, i.e., the set $\arg\min_{\mathbf{z} \in \mathbb{R}^{k-1}} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{z})$. When $\mathcal{L}$ is the multinomial cross entropy (Example IV.5), this argmin is a singleton set for all $\mathbf{p} \in \mathtt{int}(\Delta^k)$ and the mapping from $\mathbf{p}$ to this unique minimizer recovers the logit function.

For a general loss $\mathcal{L}$, this mapping is sometimes referred to as the *link function* [NBR19; WVR16]. See Definition IV.48 below. This section will study the properties of the link function, culminating in a sufficient condition for when the link function is a bijection (Proposition IV.52).

Proposition IV.46. *Let $\mathcal{L}$ be a PERM loss with template ψ. If ψ is convex, then $C^{\mathcal{L}}_{\mathbf{p}}$ is convex for all $\mathbf{p} \in \Delta^k$. Furthermore, if ψ is strictly convex, then $C^{\mathcal{L}}_{\mathbf{p}}$ is strictly convex for all $\mathbf{p} \in \mathtt{int}(\Delta^k)$.*

Proof. Recall that $C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{z}) = \sum_{i \in [k]} p_i \psi(\boldsymbol{\rho}_i \mathbf{z})$ where $\boldsymbol{\rho}_i$ is an invertible matrix by Lemma IV.37. Thus, if ψ is (strictly) convex, then $\mathbf{z} \mapsto \psi(\boldsymbol{\rho}_i \mathbf{z})$ is (strictly) convex for each $i \in [k]$. For each $\mathbf{p} \in \Delta^k$, $C^{\mathcal{L}}_{\mathbf{p}}$ is a convex combination of convex function and is thus convex. Furthermore, if $\mathbf{p} \in \mathtt{int}(\Delta^k)$, then $C^{\mathcal{L}}_{\mathbf{p}}$ is a convex combination of strictly convex function and is thus strictly convex. See Boyd et al. [BV04, Section 3.2.1] for instance.

□

An easy consequence of the above result is the following:

Corollary IV.47. *Let* $\mathbf{p} \in \mathtt{int}(\Delta^k)$ *be arbitrary and* $\mathcal{L}$ *be a nonnegative PERM loss whose* ψ *is semi-coercive. If* ψ *is convex, then the infimum* $\inf_{\mathbf{z}\in\mathbb{R}^{k-1}} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{z})$ *is attained. Furthermore, if* ψ *is strictly convex, then the infimum is attained by a* unique *minimizer, i.e.,* $\arg\min_{\mathbf{z}\in\mathbb{R}^{k-1}} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{z})$ *is a singleton set.*

Proof. By Proposition IV.45, $C_{\mathbf{p}}^{\mathcal{L}}$ is coercive. By Proposition IV.46, $C_{\mathbf{p}}^{\mathcal{L}}$ is strictly convex. By the Extreme Value Theorem, a continuous and coercive functions have at least one global minimum. Furthermore, a strictly convex functions have at most one global minimum. See Boyd et al. [BV04, Section 4.2] for instance. □

In view of Corollary IV.47, we define:

Definition IV.48. Let $\mathcal{L}$ be a PERM loss whose template ψ is nonnegative, strictly convex and semi-coercive. Define the *link function* $\mathtt{lnk}^{\mathcal{L}} : \mathtt{int}(\Delta^k) \to \mathbb{R}^{k-1}$ by letting $\mathtt{lnk}^{\mathcal{L}}(\mathbf{p})$ be the unique element of $\arg\min_{\mathbf{z}\in\mathbb{R}^{k-1}} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{z})$.

In this section, we give a sufficient condition on $\mathcal{L}$ for $\mathtt{lnk}^{\mathcal{L}}$ of Definition IV.48 to be a *bijection*. We will need the concept of an *M-matrix*, which is reviewed in Section 4.11.1.

Lemma IV.49. *Let* $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}_+^k$ *be a regular PERM loss. For all* $\mathbf{z} \in \mathbb{R}^{k-1}$*, the* $(k-1) \times (k-1)$ *matrix*

$$\mathbf{A}(\mathbf{z}) := \begin{bmatrix} \nabla_{\ell_2}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_k}(\mathbf{z}) \end{bmatrix}$$

is a non-singular M-matrix. Thus, by Theorem IV.113, $\mathbf{A}(\mathbf{z})$ *is a monotone matrix. Furthermore,* $\mathbf{A}(\mathbf{z})$ *is strictly monotone.*

Note that we use the convention that the gradient of a multivariate-input univariate-output function is a row vector. Conversely, the gradient of a univariate-input multivariate-output function is a column vector. See Section 4.11.2.

Proof. First, we compute the Jacobian of $\ell = (\ell_1, \ldots, \ell_k) : \mathbb{R}^{k-1} \to \mathbb{R}^k$. For each $i \in [k]$, we have $\ell_i(\mathbf{z}) = \psi(\boldsymbol{\rho}_i \mathbf{z})$ (Eqn. 4.12). Thus, by the chain rule and Eqn. (4.12), we have

$$\nabla_{\ell_i}(\mathbf{z}) = \nabla_\psi(\boldsymbol{\rho}_i \mathbf{z})\boldsymbol{\rho}_i. \tag{4.20}$$

Next, fix $i \in \{2, \ldots, k\}$ and $\mathbf{z} \in \mathbb{R}^k$. Let $\mathbf{w} := \nabla_\psi(\boldsymbol{\rho}_i \mathbf{z})$. Then by assumption, we have $\mathbf{w} < 0$. Note that $\mathbf{w}$ is a row vector. Now, $\nabla_{\ell_i}(\mathbf{z}) = \nabla_\psi(\boldsymbol{\rho}_i \mathbf{z})\boldsymbol{\rho}_i = \mathbf{w}\boldsymbol{\rho}_i$. Thus, for each $j \in [k-1]$, we have $[\mathbf{w}\boldsymbol{\rho}_i]_j = \mathbf{w}[\boldsymbol{\rho}_i]_{:j}$. Recall that

$$[\boldsymbol{\rho}_i]_{:j} = \begin{cases} \mathbb{e}_j^{k-1} & : j \neq i-1 \\ -\mathbb{1}_{k-1} & : j = i-1. \end{cases}$$

Thus, we have

$$[\mathbf{w}\boldsymbol{\rho}_i]_j = \begin{cases} w_j & : j \neq i-1 \\ -\sum_{l \in [k-1]} w_l & : j = i-1. \end{cases}$$

In particular, $[\mathbf{w}\boldsymbol{\rho}_i]_j \leq 0$ for all $j \neq i-1$ which proves that $\mathbf{A}(\mathbf{z})$ is a Z-matrix. Furthermore, note that the fact $\mathbf{w} < 0$ and $[\mathbf{w}\boldsymbol{\rho}_i]_{i-1} = -\sum_{l \in [k-1]} w_l$ implies that the diagonals of $\mathbf{A}(\mathbf{z})$ are positive. Observe that $\mathbf{w}\boldsymbol{\rho}_i$ has the property that

$$\begin{aligned} |[\mathbf{w}\boldsymbol{\rho}_i]_{i-1}| &= -\sum_{l \in [k-1]} w_l \\ &> -\sum_{l \in [k-1] : l \neq i-1} w_l \\ &= \sum_{l \in [k-1] : l \neq i-1} |[\mathbf{w}\boldsymbol{\rho}_i]_l|. \end{aligned}$$

This proves that $\mathbf{A}(\mathbf{z})$ is strictly diagonally dominant. By Corollary IV.112, we have that $\mathbf{A}(\mathbf{z})$ is a non-singular M-matrix. For the "Furthermore" part, we can apply Lemma IV.114 since the diagonal elements of $\mathbf{A}(\mathbf{z})$ are positive. □

Lemma IV.50. *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_+$ be a regular PERM loss. Let $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\mathbf{p} \in \Delta^k$. Then $\mathbf{z}$ minimizes $C^{\mathcal{L}}_{\mathbf{p}}$ if and only if*

$$-p_1 \nabla_\psi(\mathbf{z}) = \begin{bmatrix} p_2 & \cdots & p_k \end{bmatrix} \mathbf{A}(\mathbf{z}) \tag{4.21}$$

where $\mathbf{A}(\mathbf{z})$ is defined as in Lemma IV.49. Furthermore, if $\mathbf{z}$ minimizes $C^{\mathcal{L}}_{\mathbf{p}}$, then $\mathbf{p} \in \mathtt{int}(\Delta^k)$.

Proof. Proposition IV.46 asserts that $C^{\mathcal{L}}_{\mathbf{p}}$ is convex. For a differentiable convex function, recall that the gradient-vanishing condition is necessary and sufficient for optimality [BV04]. Thus, $\mathbf{z}$ minimizes $C^{\mathcal{L}}_{\mathbf{p}}$ if and only if

$$0 = \nabla_{C^{\mathcal{L}}_{\mathbf{p}}}(\mathbf{z}) = \sum_{j \in [k]} p_j \nabla_{\ell_j}(\mathbf{z}) = p_1 \nabla_\psi(\mathbf{z}) + \begin{bmatrix} p_2 & \cdots & p_k \end{bmatrix} \mathbf{A}(\mathbf{z}). \tag{4.22}$$

Rearranging Equation (4.22), we get

$$-p_1 \nabla_\psi(\mathbf{z}) = \begin{bmatrix} p_2 & \cdots & p_k \end{bmatrix} \mathbf{A}(\mathbf{z}). \tag{4.23}$$

For the "Furthermore" part, first note that Lemma IV.49 says $\mathbf{A}(\mathbf{z})$ is a non-singular M-matrix. If $p_1 = 0$, then Equation (4.23) reduces to

$$0 = \begin{bmatrix} p_2 & \cdots & p_k \end{bmatrix} \mathbf{A}(\mathbf{z}) \tag{4.24}$$

Since $\mathbf{A}(\mathbf{z})$ is non-singular, we have $p_2 = \cdots = p_k = 0$ which contradicts that $\mathbf{p} \in \Delta^k$. Thus, $p_1 > 0$ and so $-p_1 \nabla_\psi(\mathbf{z}) > 0$. From Lemma IV.49, we have that $\mathbf{A}(\mathbf{z})$ is strictly monotone. Thus, Equation 4.23 implies that $p_i > 0$ for each $i = 2, \ldots, k$ as well. □

The "Furthermore" part of Lemma IV.50 immediately implies the following.

Corollary IV.51. *If $\mathbf{p} \in \Delta^k \setminus \mathtt{int}(\Delta^k)$, then $\arg\min_{\mathbf{z} \in \mathbb{R}^{k-1}} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{z}) = \emptyset$.*

Proposition IV.52. *Let $\mathcal{L}$ be a regular PERM loss. Recall the mapping $\texttt{lnk}^{\mathcal{L}} : \texttt{int}(\Delta^k) \to \mathbb{R}^{k-1}$ from Definition IV.48. Then $\texttt{lnk}^{\mathcal{L}}$ is a bijection.*

Proof. First, we prove that $\texttt{lnk}$ is injective. Suppose that $\mathbf{p}, \mathbf{q} \in \texttt{int}(\Delta^k)$ are such that $\texttt{lnk}^{\mathcal{L}}(\mathbf{p}) = \texttt{lnk}^{\mathcal{L}}(\mathbf{q}) =: \mathbf{z}$. Then by Lemma IV.50, we have that

$$-\nabla_\psi(\mathbf{z})\mathbf{A}(\mathbf{z})^{-1} = p_1^{-1}\begin{bmatrix} p_2 & \cdots & p_k \end{bmatrix} = q_1^{-1}\begin{bmatrix} q_2 & \cdots & q_k \end{bmatrix}. \tag{4.25}$$

Thus, $(1-p_1)/p_1 = (p_2 + \cdots + p_k)/p_1 = (q_2 + \cdots + q_k)/q_1 = (1-q_1)/q_1$ implies that $p_1 = q_1$. Therefore, Equation 4.25 implies that $p_i = q_i$ for each $i = 2, \ldots, k$ as well. This proves that $\texttt{lnk}$ is injective.

Next, we prove that $\texttt{lnk}$ is surjective. Pick $\mathbf{z} \in \mathbb{R}^{k-1}$. From Lemma IV.49, we have that $\mathbf{A}(\mathbf{z})$ is non-singular and strictly monotone. Since $\mathbf{A}(\mathbf{z})$ is non-singular, there exists $\mathbf{v} \in \mathbb{R}^{k-1}$ such that $-\nabla_\psi(\mathbf{z}) = \mathbf{v}\mathbf{A}(\mathbf{z})$. Furthermore, since $-\nabla_\psi(\mathbf{z}) > 0$ and $\mathbf{A}(\mathbf{z})$ is strictly monotone, we have $\mathbf{v} > 0$. Define $p_1, \ldots, p_k$ by $p_1 := (1 + v_1 + \cdots + v_{k-1})^{-1}$ and $p_i := v_{i-1}p_1$ for each $i = 2, \ldots, k$. Clearly, we have $\mathbf{p} = (p_1, \ldots, p_k) > 0$. Furthermore,

$$p_1 + p_2 + \cdots + p_k = p_1(1 + v_1 + \cdots + v_{k-1}) = 1.$$

Thus, we have $\mathbf{p} \in \texttt{int}(\Delta^k)$. By construction, $\mathbf{z}$ and $\mathbf{p}$ satisfy Equation 4.21. This proves that $\texttt{lnk}^{\mathcal{L}}(\mathbf{p}) = \mathbf{z}$. □

Remark IV.53. Before proceeding, we remark that Proposition IV.52 gives theoretical support to the conjectural observation in Nowak-Vila et al. [NBR19, Remark 3.1] regarding the injectivity of the *link function*.

4.6.3 Geometry of the loss surface

Recall from Definition IV.10 and Theorem IV.12 that the classification-calibration of the *set* $\mathcal{S}(\mathcal{L})$ implies the classification-calibration of the *loss* $\mathcal{L}$. In general, the set $\mathcal{S}(\mathcal{L})$ may be difficult to compute. In this section, we study the geometry of the set

$\mathcal{R}(\mathcal{L})$ when $\mathcal{L}$ is a regular PERM loss which enables us to compute the convex hull $\mathcal{S}(\mathcal{L})$ of $\mathcal{R}(\mathcal{L})$. One of the main tools is the mapping defined below:

Corollary IV.54. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Then $\ell : \mathbb{R}^{k-1} \to \mathbb{R}^k$ is injective.*

Proof. Suppose that $\mathbf{z}, \mathbf{w} \in \mathbb{R}^{k-1}$ are such that $\ell(\mathbf{z}) = \ell(\mathbf{w})$. By Proposition IV.52, there exists $\mathbf{p} \in \texttt{int}(\Delta^k)$ such that $\texttt{lnk}^{\mathcal{L}}(\mathbf{p}) = \mathbf{z}$. Now, $\langle \mathbf{p}, \ell(\mathbf{z}) \rangle = \langle \mathbf{p}, \ell(\mathbf{w}) \rangle$ implies that both $\mathbf{z}, \mathbf{w}$ minimize $C^{\mathcal{L}}_{\mathbf{p}}$. By Corollary IV.47, we have $\mathbf{z} = \mathbf{w}$ and so ℓ is injective. □

Definition IV.55. Given a PERM loss $\mathcal{L}$ with reduced form ℓ, we define two functions F and G mapping from $\mathbb{R}^{k-1} \times \mathbb{R}$ to $\mathbb{R}^k$ by

$$F(\mathbf{z}, \lambda) = \ell(\mathbf{z}) + \lambda \mathbb{1}, \quad \text{and} \quad G(\mathbf{z}, t) = \ell(\mathbf{z}) + t\mathbb{e}_k^{(k)}.$$

Below, we will study the properties of the two functions from Definition IV.55.

4.6.3.1 Properties of the F function

Lemma IV.56. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ and F be as in Definition IV.55. Then F is injective, i.e., if $\ell(\mathbf{z}) + \lambda \mathbb{1} = \ell(\mathbf{w}) + \mu \mathbb{1}$ for some $\mathbf{z}, \mathbf{w} \in \mathbb{R}^{k-1}$ and $\lambda, \mu \in \mathbb{R}$. Then $\mathbf{z} = \mathbf{w}$ and $\lambda = \mu$.*

Proof. If $\mathbf{z} = \mathbf{w}$, then $\ell(\mathbf{z}) = \ell(\mathbf{w})$ and so $\lambda = \mu$. Thus, $\mathbf{z} = \mathbf{w}$ implies $\lambda = \mu$.

If $\lambda = \mu$, then we have $\ell(\mathbf{z}) = \ell(\mathbf{w})$. By Corollary IV.54, we have $\mathbf{z} = \mathbf{w}$. Therefore, $\lambda = \mu$ implies $\mathbf{z} = \mathbf{w}$.

It remains to show that $\mathbf{z} \neq \mathbf{w}$ and $\lambda \neq \mu$ leads to a contradiction. Without loss of generality, suppose that $\lambda > \mu$. Then we have $\ell(\mathbf{z}) + (\lambda - \mu)\mathbb{1} = \ell(\mathbf{w})$. Thus, for all $\mathbf{p} \in \Delta^k$, we have

$$\langle \mathbf{p}, \ell(\mathbf{w}) \rangle = \langle \mathbf{p}, \ell(\mathbf{z}) + (\lambda - \mu)\mathbb{1} \rangle > \langle \mathbf{p}, \ell(\mathbf{z}) \rangle.$$

Thus, $\mathbf{w}$ is never the minimizer of $C^{\mathcal{L}}_{\mathbf{p}}$ for any $\mathbf{p} \in \Delta^k$. But this contradicts since Proposition IV.52 implies that `lnk` is surjective. □

Lemma IV.57. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let F be as in Definition IV.55. Then for all $(\mathbf{z}, \lambda) \in \mathbb{R}^{k-1} \times \mathbb{R}$, $\nabla_F(\mathbf{z}, \lambda)$ is non-singular.*

Proof. Let $\mathbf{v} \in \mathbb{R}^k$ be arbitrary. It suffices to check that if $\mathbf{v}^\top \nabla_F(\mathbf{z}, \lambda) = 0$ then $\mathbf{v} = 0$. Note that

$$\nabla_F(\mathbf{z}, \lambda) = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{1} \end{bmatrix} \quad \text{where } \nabla_\ell(\mathbf{z}) \in \mathbb{R}^{k \times (k-1)} \text{ and } \mathbb{1} \in \mathbb{R}^k.$$

Hence, $\mathbf{v}^\top \nabla_F(\mathbf{z}, \lambda) = 0$ implies $\mathbf{v}^\top \nabla_\ell(\mathbf{z}) = 0$ and $\mathbf{v}^\top \mathbb{1} = v_1 + \cdots + v_k = 0$. Replacing $\mathbf{v}$ by $-\mathbf{v}$ if necessary, we can assume that $v_1 \geq 0$. The equation $\mathbf{v}^\top \nabla_\ell(\mathbf{z}) = 0$ can be rewritten as

$$-v_1 \nabla_\psi(\mathbf{z}) = \begin{bmatrix} v_2 & \cdots & v_k \end{bmatrix} \mathbf{A}(\mathbf{z}) \tag{4.26}$$

Since $\mathbf{A}(\mathbf{z})$ is monotone and $-v_1 \nabla_\psi(\mathbf{z}) \geq 0$, we get that $v_i \geq 0$ for each $i = 2, \ldots, k$. Now, $\mathbf{v}^\top \mathbb{1} = v_1 + \cdots + v_k = 0$ implies that $\mathbf{v} = 0$, as desired. □

Now, by applying the inverse function theorem (Theorem IV.116), we immediately have the following.

Corollary IV.58. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let F be as in Definition IV.55. For all $(\mathbf{z}, \lambda) \in \mathbb{R}^{k-1} \times \mathbb{R}$, there exist open neighborhoods $U \ni (\mathbf{z}, \lambda)$ and $V \ni F(\mathbf{z}, \lambda)$ such that $F|_U : U \to V$ is a diffeomorphism.*

Proposition IV.59. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let F be as in Definition IV.55. The map F is a bijection.*

Proof. Lemma IV.56 shows that F is injective. To show that F is surjective, we prove that $\mathcal{R}(F)$ is both open and closed as a subset of $\mathbb{R}^k$. This would imply that $\mathcal{R}(F) = \mathbb{R}^k$ since the only subets of $\mathbb{R}^k$ that are both open and closed are $\emptyset$ and $\mathbb{R}^k$.

Now, Lemma IV.58 shows that $\mathcal{R}(F)$ is an open subset of $\mathbb{R}^k$. It remains to prove that $\mathcal{R}(F)$ is closed. To this end, consider a sequence $\{(\mathbf{z}^{(i)}, \lambda^{(i)})\}_{i=1}^{\infty}$ such that $F(\mathbf{z}^{(i)}, \lambda^{(i)}) = \ell(\mathbf{z}^{(i)}) + \lambda^{(i)}\mathbb{1}$ converges to $\boldsymbol{\zeta} \in \mathbb{R}^k$. Our goal is to show that $\boldsymbol{\zeta} \in \mathcal{R}(F)$.

To this end, first pick $\epsilon > 0$. For the rest of this proof, $\mathbb{1} := \mathbb{1}_k$ denotes the k-dimensional vector of all ones.

There exists M such that

$$\boldsymbol{\zeta} - \epsilon\mathbb{1} \leq \ell(\mathbf{z}^{(i)}) + \lambda^{(i)}\mathbb{1} \leq \boldsymbol{\zeta} + \epsilon\mathbb{1}$$

for all $i \geq M$. Before proceeding, we prove a helper lemma.

Lemma IV.60 (Helper lemma)**.** *Let $\mathcal{L}$ be a PERM loss with reduced form ℓ and template ψ such that $\nabla_\psi \leq 0$. Then for all $\mathbf{z} \in \mathbb{R}^{k-1}$, we have that $\min(\ell(\mathbf{z})) \leq \psi(\mathbb{0}_{k-1}) =: C$, where $\mathbb{0}_{k-1}$ is the $(k-1)$-dimensional all-zeros vector.*

Proof of helper lemma. Let $\mathbf{v} \in \mathbb{R}^k$ be such that $\mathbf{M}(\mathbf{v}) = \mathbf{z}$. For instance, we can take $\mathbf{v} = [0 \ -\mathbf{z}^\top]^\top$. Let $y \in \arg\max \mathbf{v}$. Then by Definition IV.1, we have that $\min(\ell(\mathbf{z})) = \min(\mathcal{L}(\mathbf{v})) = [\mathcal{L}(\mathbf{v})]_y$. Next by Eqn. 4.12, we have

$$[\mathcal{L}(\mathbf{v})]_y = \ell_y(\mathbf{z}).$$

Let $\mathbf{w} := \sigma_{(1,y)}(\mathbf{v})$. Note that by construction we have $1 \in \arg\max \mathbf{w}$. Recall that $\sigma_{(1,y)} \in \mathtt{Sym}(k)$ is the transposition that swaps 1 and y. By permutation-equivariance, we have

$$[\mathcal{L}(\mathbf{v})]_y = [\mathcal{L}(\mathbf{v})]_{\sigma_{(1,y)}(1)} = [\mathcal{L}(\sigma_{(1,y)}(\mathbf{v}))]_1 = [\mathcal{L}(\mathbf{w})]_1.$$

By Eqn. 4.12, we have

$$[\mathcal{L}(\mathbf{w})]_1 = \psi(\mathbf{M}(\mathbf{w})).$$

Since $1 \in \arg\max \mathbf{w}$, we have that $\mathbf{M}(\mathbf{w}) \in \mathbb{R}^{k-1}_{\geq 0}$ belongs to the non-negative or-

thtant. In other words, $\mathbf{M}(\mathbf{w}) \geq \mathbb{0}_{k-1}$. Furthermore, since $\nabla_\psi \leq 0$, we have that $\psi(\mathbf{M}(\mathbf{w})) \leq \psi(\mathbb{0}_{k-1})$, as desired. □

We now return to the proof of the proposition. Let C be as in the helper lemma. Then we have

$$\min(\ell(\mathbf{z}^{(i)}) + \lambda^{(i)}\mathbb{1}) = \min(\ell(\mathbf{z}^{(i)})) + \lambda^{(i)} \leq C + \lambda^{(i)}.$$

Thus, we have

$$\min(\boldsymbol{\zeta}) - \epsilon \leq C + \lambda^{(i)}$$

and so $-\lambda^{(i)} \leq C + \epsilon - \min(\boldsymbol{\zeta}) =: D$. From this, we get that

$$\ell(\mathbf{z}^{(i)}) \leq \boldsymbol{\zeta} + \epsilon\mathbb{1} - \lambda^{(i)}\mathbb{1} \leq \boldsymbol{\zeta} + (\epsilon + D)\mathbb{1}$$

Thus, $\mathbf{z}^{(i)} \in \{\mathbf{z} \in \mathbb{R}^{k-1} : \ell(\mathbf{z}) \leq \boldsymbol{\zeta} + (\epsilon + D)\mathbb{1}\}$ which is a bounded set by Lemma IV.44. By passing to a subsequence, we may assume that $\mathbf{z}^{(i)}$ converges to some $\mathbf{z}^* \in \mathbb{R}^{k-1}$. Thus, we have $\lambda^{(i)}\mathbb{1}$ converges to $\ell(\mathbf{z}^*) + \boldsymbol{\zeta}$, which implies in particular that $\lambda^{(i)}$ converges to some λ^*. Putting it all together, we have shown that $F(\mathbf{z}^{(i)}, \lambda^{(i)})$ converges to $\boldsymbol{\zeta} = F(\mathbf{z}^*, \lambda^*)$ and so $\mathcal{R}(F)$ is closed. □

Corollary IV.61. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let F be as in Definition IV.55. The map F is a diffeomorphism, i.e., F is a differentiable bijection with a differentiable inverse. In particular, F is a homeomorphism.*

Proof. Lee [Lee13, Proposition 4.6 (f)] states that every bijective local diffeomorphism is a (global) diffeomorphism. Thus, the result follows in view of the facts that F is a bijection (Proposition IV.59) and that F is a local diffeomorphism (Corollary IV.58). □

Proposition IV.62. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let F be as*

in Definition IV.55. Consider $\mathbf{v}, \mathbf{x} \in \mathbb{R}^k$ *and* $t \in \mathbb{R}$*. Let* $\alpha(t)$ *and* $\beta(t)$ *be such that* $t\mathbf{v} + \mathbf{x} = F(\alpha(t), \beta(t)) = \ell(\alpha(t)) + \beta(t)\mathbb{1}$*. Then for all* $t \in \mathbb{R}$*, we have*

1. α *and* β *are differentiable.*

2. *If* $\mathbf{v} > 0$*, then* $\nabla_\beta(t) > 0$*.*

3. β *is concave, i.e.,* $\nabla^2_\beta(t) \leq 0$*.*

Proof. To prove the first part, first note that we have $(\alpha(t), \beta(t)) = F^{-1}(t\mathbf{v} + \mathbf{x})$. Hence, α and β are differentiable.

Next, we prove the second part. Pick $i \in [k]$. The i-th coordinate of $t\mathbf{v} + \mathbf{x} = \ell(\alpha(t)) + \beta(t)\mathbb{1}$ is

$$v_i t + x_i = \ell_i(\alpha(t)) + \beta(t). \tag{4.27}$$

Differentiating (4.27) on both sides with respect to t, we get

$$v_i = \nabla_{\ell_i}(\alpha(t))\nabla_\alpha(t) + \nabla_\beta(t). \tag{4.28}$$

We claim that $\nabla_{\ell_i}(\alpha(t))\nabla_\alpha(t) \leq 0$ for some i.

Lemma IV.63. *Let* $\mathcal{L}$ *be a regular PERM loss with reduced form* ℓ*. Let* $\mathbf{z}, \mathbf{w} \in \mathbb{R}^{k-1}$*. Then there exists* $i \in [k]$ *such that* $\nabla_{\ell_i}(\mathbf{z}) \cdot \mathbf{w} \leq 0$*.*

Proof. Suppose for the sake of contradiction that $\nabla_{\ell_i}(\mathbf{z}) \cdot \mathbf{w} > 0$ for all $i \in [k]$. Then

$$0 < \begin{bmatrix} \nabla_\psi(\mathbf{z}) \cdot \mathbf{w} \\ \nabla_{\ell_2}(\mathbf{z}) \cdot \mathbf{w} \\ \vdots \\ \nabla_{\ell_k}(\mathbf{z}) \cdot \mathbf{w} \end{bmatrix} = \begin{bmatrix} \nabla_\psi(\mathbf{z}) \\ \nabla_{\ell_2}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_k}(\mathbf{z}) \end{bmatrix} \mathbf{w} = \begin{bmatrix} \nabla_\psi(\mathbf{z}) \\ \mathbf{A}(\mathbf{z}) \end{bmatrix} \mathbf{w}.$$

In other words, we have $\mathbf{A}(\mathbf{z})\mathbf{w} > 0$ and $\nabla_\psi(\mathbf{z}) \cdot \mathbf{w} > 0$. Since $\mathbf{A}(\mathbf{z})$ is strictly monotone, we have $\mathbf{w} > 0$. But $\nabla_\psi(\mathbf{z}) < 0$ by assumption that $\mathcal{L}$ is a regular PERM

loss. Hence, $\nabla_\psi(\mathbf{z}) \cdot \mathbf{w} < 0$, a contradiction. $\square$

Applying Lemma IV.63 with $\mathbf{w} = \nabla_\alpha(t)$, we get the desired claim. Now, pick $i \in [k]$ such that $\nabla_{\ell_i}(\alpha(t)) \cdot \nabla_\alpha(t) \leq 0$. Thus, from Eqn. (4.28) we have

$$\nabla_\beta(t) = v_i - \nabla_{\ell_i}(\alpha(t)) \cdot \nabla_\alpha(t) > 0.$$

This proves the second part of Lemma IV.63.

Finally, we prove the last part of Lemma IV.63. Note that (4.28) can be rewritten as follows:

$$v_i = \langle \nabla_{\ell_i}(\alpha(t))^\top, \nabla_\alpha(t) \rangle + \nabla_\beta(t). \tag{4.29}$$

Differentiating (4.29) with respect to t, we get

$$0 = \left\langle \left(\nabla^2_{\ell_i}(\alpha(t))\nabla_\alpha(t)\right)^\top, \nabla_\alpha(t) \right\rangle + \left\langle \nabla_{\ell_i}(\alpha(t))^\top, \nabla^2_\alpha(t) \right\rangle + \nabla^2_\beta(t) \tag{4.30}$$

$$= \nabla_\alpha(t)^\top \nabla^2_{\ell_i}(\alpha(t))\nabla_\alpha(t) + \nabla_{\ell_i}(\alpha(t))\nabla^2_\alpha(t) + \nabla^2_\beta(t). \tag{4.31}$$

Thus, we have

$$-\nabla^2_\beta(t) = \nabla_\alpha(t)^\top \nabla^2_{\ell_i}(\alpha(t))\nabla_\alpha(t) + \nabla_{\ell_i}(\alpha(t))\nabla^2_\alpha(t).$$

Since ℓ_i is convex, we have $\nabla_\alpha(t)^\top \nabla^2_{\ell_i}(\alpha(t))\nabla_\alpha(t) \geq 0$. Next, we claim that for some choice of $i \in [k]$, we have have $\nabla_{\ell_i}(\alpha(t))\nabla^2_\alpha(t) \geq 0$. The claim follows immediately from applying Lemma IV.63 with $\mathbf{z} = \alpha(t)$ and $\mathbf{v} = -\nabla^2_\alpha(t)$. Below, pick $i \in [k]$ such that $\nabla_{\ell_i}(\alpha(t))\nabla^2_\alpha(t) \geq 0$. Then for such a choice of i, we have

$$-\nabla^2_\beta(t) = \nabla_\alpha(t)^\top \nabla^2_{\ell_i}(\alpha(t))\nabla_\alpha(t) + \nabla_{\ell_i}(\alpha(t))\nabla^2_\alpha(t) \geq 0,$$

or equivalently, $\nabla^2_\beta(t) \leq 0$. This proves that β is concave. $\square$

4.6.3.2 Properties of the G function

Lemma IV.64. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let G be as in Definition IV.55. Then for all $(\mathbf{z}, t) \in \mathbb{R}^{k-1} \times \mathbb{R}$, the Jacobian matrix $\nabla_G(\mathbf{z}, t)$ is non-singular.*

Proof. Let $\mathbf{v} \in \mathbb{R}^k$ be arbitrary. It suffices to check that $\mathbf{v}^\top \nabla_G(\mathbf{z}, \lambda) = 0$ implies $\mathbf{v} = 0$. Note that

$$\nabla_G(\mathbf{z}, \lambda) = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{e}_k^{(k)} \end{bmatrix} \quad \text{where } \nabla_\ell(\mathbf{z}) \in \mathbb{R}^{k \times (k-1)} \text{ and } \mathbb{e}_k^{(k)} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^\top.$$

Thus, $\mathbf{v}^\top \nabla_G(\mathbf{z}, \lambda) = 0$ implies $\mathbf{v}^\top \mathbb{e}_k^{(k)} = v_k = 0$ and

$$\mathbf{v}^\top \nabla_\ell(\mathbf{z}) = \begin{bmatrix} v_1 & \cdots & v_{k-1} & 0 \end{bmatrix} \begin{bmatrix} \nabla_{\ell_1}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_k}(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_{k-1} \end{bmatrix} \underbrace{\begin{bmatrix} \nabla_{\ell_1}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_{k-1}}(\mathbf{z}) \end{bmatrix}}_{\dagger} = 0.$$

Thus, it only remains to show that the matrix marked by † is non-singular. To this end, we first recall from Lemma IV.49 that

$$\mathbf{A}(\mathbf{z}) := \begin{bmatrix} \nabla_{\ell_2}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_k}(\mathbf{z}) \end{bmatrix}.$$

Let $j \in \{2, \ldots, k\}$, then we have $\nabla_{\ell_j}(\boldsymbol{\rho}_k \mathbf{z})$. By the chain rule (Theorem IV.115), we have $\nabla_{\ell_j(\boldsymbol{\rho}_k \bullet)}(\mathbf{z}) = \nabla_{\ell_j}(\boldsymbol{\rho}_k \mathbf{z}) \boldsymbol{\rho}_k$. Below, we will use the $\bullet$ notation to denote the placeholder for the input of a function. For instance, $\ell_j(\boldsymbol{\rho}_k \bullet)$ denotes the function

$\mathbf{z} \mapsto \ell_j(\boldsymbol{\rho}_k \bullet)$. Now, note that

$$\begin{aligned}
\ell_j(\boldsymbol{\rho}_k \bullet) &= \psi(\boldsymbol{\rho}_j \boldsymbol{\rho}_k \bullet) && \because \text{Eqn. 4.12} \\
&= \psi(\mathbf{T}_{(j-1,k-1)} \boldsymbol{\rho}_{\sigma_k(j)} \bullet) && \because \text{Lemma IV.42} \\
&= \psi(\boldsymbol{\rho}_{\sigma_k(j)} \bullet) && \because \psi \text{ is symmetric (Proposition IV.33)} \\
&= \begin{cases} \psi(\boldsymbol{\rho}_j \bullet) & : j \neq k \\ \psi(\boldsymbol{\rho}_1 \bullet) & : j = k \end{cases} && \because \text{Definition of } \sigma_k \text{ (Section 4.1.3)} \\
&= \begin{cases} \ell_j(\bullet) & : j \neq k \\ \ell_1(\bullet) & : j = k \end{cases} && \because \text{Eqn. 4.12.}
\end{aligned}$$

The above calculation shows that

$$\nabla_{\ell_j(\boldsymbol{\rho}_k \bullet)}(\mathbf{z}) = \begin{cases} \nabla_{\ell_j}(\mathbf{z}) & : j \neq k \\ \nabla_{\ell_1}(\mathbf{z}) & : j = k \end{cases}.$$

Combined with the result earlier that $\nabla_{\ell_j(\boldsymbol{\rho}_k \bullet)}(\mathbf{z}) = \nabla_{\ell_j}(\boldsymbol{\rho}_k \mathbf{z}) \boldsymbol{\rho}_k$, we have

$$\nabla_{\ell_j}(\boldsymbol{\rho}_k \mathbf{z}) \boldsymbol{\rho}_k = \begin{cases} \nabla_{\ell_j}(\mathbf{z}) & : j \neq k \\ \nabla_{\ell_1}(\mathbf{z}) & : j = k. \end{cases}$$

Multiplying both side by $\boldsymbol{\rho}_k$, we get

$$\nabla_{\ell_j}(\boldsymbol{\rho}_k \mathbf{z}) = \begin{cases} \nabla_{\ell_j}(\mathbf{z}) \boldsymbol{\rho}_k & : j \neq k \\ \nabla_{\ell_1}(\mathbf{z}) \boldsymbol{\rho}_k & : j = k. \end{cases}$$

Thus, we have

$$\mathbf{A}(\boldsymbol{\rho}_k \mathbf{z}) = \begin{bmatrix} \nabla_{\ell_2}(\boldsymbol{\rho}_k \mathbf{z}) \\ \vdots \\ \nabla_{\ell_k}(\boldsymbol{\rho}_k \mathbf{z}) \end{bmatrix} = \begin{bmatrix} \nabla_{\ell_2}(\mathbf{z})\boldsymbol{\rho}_k \\ \vdots \\ \nabla_{\ell_{k-1}}(\mathbf{z})\boldsymbol{\rho}_k \\ \nabla_{\ell_1}(\mathbf{z})\boldsymbol{\rho}_k \end{bmatrix} = \underbrace{\begin{bmatrix} \nabla_{\ell_2}(\mathbf{z}) \\ \vdots \\ \nabla_{\ell_{k-1}}(\mathbf{z}) \\ \nabla_{\ell_1}(\mathbf{z}) \end{bmatrix}}_{\ddagger} \boldsymbol{\rho}_k.$$

By Lemma IV.49, the left-hand side is nonsingular. Lemma IV.37 says that $\boldsymbol{\rho}_k$ is its own inverse and hence nonsingular. Finally, the matrix marked by ‡ is clearly obtainable by permuting the rows of the matrix marked by † from earlier. Thus, the matrix marked by † is nonsingular. □

While the next lemma is not about the G function *per se*, the proof of the lemma uses the G function heavily. Define $\texttt{prj}^{(k)}$ to be the projection $\mathbb{R}^k \to \mathbb{R}^{k-1}$ that drops the last coordinate, i.e., $\texttt{prj}^{(k)}([v_1, \ldots, v_k]^\top) = [v_1, \ldots, v_{k-1}]^\top$. When k is clear from context, we drop the superscript (k) from $\texttt{prj}^{(k)}$.

Lemma IV.65. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\mathbf{w} \in \mathbb{R}^{k-2}$. Suppose that $\texttt{prj}(\ell^{(k)}(\mathbf{z})) = \ell^{(k-1)}(\mathbf{w})$. Then there exists $\mathbf{z}^* \in \mathbb{R}^{k-1}$ and $t^* \in \mathbb{R}$ such that $t^* > 0$ and $\texttt{prj}(\ell^{(k)}(\mathbf{z}^*) + t^* \mathbb{1}) = \ell^{(k-1)}(\mathbf{w})$.*

Proof. Before proceeding, first let F and G be as in Definitions IV.55. Let $\tilde{\mathbf{z}}(\cdot, \cdot) : \mathbb{R}^{k-1} \times \mathbb{R} \to \mathbb{R}^{k-1}$ denote the first $k-1$ component functions of $F^{-1} \circ G$. Likewise, let $\tilde{\lambda}(\cdot, \cdot) : \mathbb{R}^{k-1} \times \mathbb{R} \to \mathbb{R}$ denote the last component function of $F^{-1} \circ G$. In other words, we have

$$F^{-1} \circ G(\mathbf{z}, t) = \begin{bmatrix} \tilde{\mathbf{z}}(\mathbf{z}, t) \\ \tilde{\lambda}(\mathbf{z}, t) \end{bmatrix} \tag{4.32}$$

for all $\mathbf{z} \in \mathbb{R}^{k-1}$ and $t \in \mathbb{R}$. (Note that the fact that F has an inverse was proven in Corollary IV.61.)

Similar to earlier, we will use the $\bullet$ notation to denote the placeholder for the input of a function with the understanding that $\bullet$ represents an input *tuple*, e.g., $(\mathbf{z}, t)$.

Now, by the chain rule (Theorem IV.115), we have $\nabla_{F^{-1}\circ G}(\bullet) = \nabla_{F^{-1}}(G(\bullet))\nabla_G(\bullet)$. By the inverse function theorem (Theorem IV.116 and Corollary IV.117), we have $\nabla_{F^{-1}}(\bullet) = \nabla_F(F^{-1}(\bullet))^{-1}$. Putting these two equations together, we get

$$\nabla_{F^{-1}\circ G}(\bullet) = \nabla_F(F^{-1} \circ G(\bullet))^{-1}\nabla_G(\bullet).$$

Now, $G(\mathbf{z}, 0) = \ell(\mathbf{z}) = F(\mathbf{z}, 0)$. Thus we have $F^{-1}(G(\mathbf{z}, 0)) = (\mathbf{z}, 0)$ and

$$\nabla_{F^{-1}\circ G}(\mathbf{z}, 0) = \nabla_F(F^{-1} \circ G(\mathbf{z}, 0))^{-1}\nabla_G(\mathbf{z}, 0) = \nabla_F(\mathbf{z}, 0)^{-1}\nabla_G(\mathbf{z}, 0).$$

Furthermore, recall from the proofs of Lemmas IV.57 and IV.64 that the Jacobians of F and G are given by

$$\nabla_F(\mathbf{z}, \lambda) = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{1} \end{bmatrix} \quad \text{and} \quad \nabla_G(\mathbf{z}, \lambda) = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{e}_k^{(k)} \end{bmatrix}.$$

Next, recall $\tilde{\mathbf{z}}(\mathbf{z}, t)$ and $\tilde{\lambda}(\mathbf{z}, t)$ as defined at the beginning of this proof.

Thus, we have

$$\begin{bmatrix} \frac{\partial}{\partial \mathbf{z}^\top}\tilde{\mathbf{z}}(\mathbf{z}, t) \\ \frac{\partial}{\partial \mathbf{z}^\top}\tilde{\lambda}(\mathbf{z}, t) \end{bmatrix} = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{1} \end{bmatrix}^{-1} \nabla_\ell(\mathbf{z}) = \begin{bmatrix} \mathbf{Id}_{k-1} \\ \mathbb{0}_{k-1}^\top \end{bmatrix}.$$

Putting it all together, we have

$$\begin{bmatrix} \frac{\partial}{\partial t}\tilde{\mathbf{z}}(\mathbf{z}, t) \\ \frac{\partial}{\partial t}\tilde{\lambda}(\mathbf{z}, t) \end{bmatrix} = \begin{bmatrix} \nabla_\ell(\mathbf{z}) & \mathbb{1} \end{bmatrix}^{-1} \nabla_\ell(\mathbf{z}) = \begin{bmatrix} \mathbf{Id}_{k-1} & \frac{\partial}{\partial t}\tilde{\mathbf{z}}(\mathbf{z}, t) \\ \mathbb{0}_{k-1}^\top & \frac{\partial}{\partial t}\tilde{\lambda}(\mathbf{z}, t) \end{bmatrix}$$

Since the above matrix is non-singular, we must have that $\frac{\partial}{\partial t}\tilde{\lambda}(\mathbf{z}, t) \neq 0$. On the other hand, the fact that $(\tilde{\mathbf{z}}(\mathbf{z}, t), \tilde{\lambda}(\mathbf{z}, t)) = F^{-1}(G(\mathbf{z}, 0)) = (\mathbf{z}, 0)$ implies that $\tilde{\lambda}(\mathbf{z}, 0) = 0$.

We claim that there exists $t^* \in \mathbb{R}$ such that $\tilde{\lambda}(\mathbf{z}, t^*) > 0$. To prove the claim, suppose that it is not true. Then $\tilde{\lambda}(\mathbf{z}, t) \leq 0$ for all $t \in \mathbb{R}$. Earlier, we saw that

$\tilde{\lambda}(\mathbf{z}, 0) = 0$. Thus, viewed as a function of t, we have that $\tilde{\lambda}(\mathbf{z}, t)$ attains a maximum at $t = 0$. This implies that $\frac{\partial}{\partial t}\tilde{\lambda}(\mathbf{z}, 0) = 0$, a contradiction.

Now, fix a $t^* \in \mathbb{R}$ such that $\tilde{\lambda}(\mathbf{z}, t^*) > 0$. Then we have

$$\ell(\mathbf{z}) + t^* \mathbb{e}_k^{(k)} = G(\mathbf{z}, t^*) = F(\tilde{\mathbf{z}}(\mathbf{z}, t^*), \tilde{\lambda}(\mathbf{z}, t^*)) = \ell(\tilde{\mathbf{z}}(\mathbf{z}, t^*)) + \tilde{\lambda}(\mathbf{z}, t^*)\mathbb{1}.$$

Note that we have $\mathtt{prj}(\ell(\tilde{\mathbf{z}}(\mathbf{z}, t^*)) + \tilde{\lambda}(\mathbf{z}, t^*)\mathbb{1}) = \mathtt{prj}(\ell^{(k)}(\mathbf{z}) + t^* \mathbb{e}_k^{(k)}) = \ell^{(k-1)}(\mathbf{w})$ as desired. □

Definition IV.66. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function. Define the following sets:

1. $\mathcal{S}^{\times}(f) := \{\boldsymbol{\zeta} + \lambda\mathbb{1} : \boldsymbol{\zeta} \in \mathcal{R}(f), \lambda \in [0, \infty)\}$
2. $\mathcal{S}^{\circ}(f) := \{\boldsymbol{\zeta} + \lambda\mathbb{1} : \boldsymbol{\zeta} \in \mathcal{R}(f), \lambda \in (0, \infty)\}$

When $f = \ell$ is a PERM loss, the above two sets are closely related to $\mathcal{S}(\ell)$ (Definition IV.10), as the following lemma and Proposition IV.70 show. The reason we define them is because they are convenient alternative characterizations.

Lemma IV.67. *Let $\mathcal{L}$ be a regular PERM loss. Then we have the following:*

1. *$\mathcal{R}(\ell)$ is closed.*
2. *$\mathcal{S}^{\times}(\ell)$ is closed and $\mathtt{bdry}(\mathcal{S}^{\times}(\ell)) = \mathcal{R}(\ell)$.*
3. *$\mathcal{S}^{\circ}(\ell) = \mathtt{int}(\mathcal{S}^{\times}(\ell))$ and $\mathtt{bdry}(\mathcal{S}^{\circ}(\ell)) = \mathcal{R}(\ell)$.*

Proof. For part one, define the set $C = \mathbb{R}^{k-1} \times \{0\}$ which is a closed subset of $\mathbb{R}^{k-1} \times \mathbb{R}$. Now, note that $\mathcal{R}(\ell) = F(C)$. Since F is a homeomorphism, we have $F(C)$ is closed as well.

Next, note that $D = \mathbb{R}^{k-1} \times [0, \infty)$ is a closed subset of $\mathbb{R}^{k-1} \times \mathbb{R}$. Furthermore, $\mathcal{S}^{\times}(\ell) = F(D)$ by construction. Thus, $\mathcal{S}^{\times}(\ell)$ is closed. Next, we have

$$\mathtt{bdry}(\mathcal{S}^{\times}(\ell)) = \mathtt{bdry}(F(D)) = F(\mathtt{bdry}(D)) = F(C) = \mathcal{R}(\ell),$$

where the second equality from the left follows from F being a homeomorphism (Corollary IV.61).

Finally, let $E = \mathbb{R}^{k-1} \times (0, \infty)$. Then similar to the above, we have

$$\texttt{int}(\mathcal{S}^{\times}(\ell)) = \texttt{int}(F(D)) = F(\texttt{int}(D)) = F(E) = \mathcal{S}^{\circ}(\ell),$$

as desired. To conclude, note that $\texttt{bdry}(\mathcal{S}^{\circ}(\ell)) = \texttt{bdry}(F(E)) = F(\texttt{bdry}(E)) = F(C) = \mathcal{R}(\ell)$. □

Proposition IV.68. *Let $\mathcal{L}$ be a regular PERM loss. Then $\mathcal{S}^{\times}(\ell)$ is convex.*

Proof. Let $\boldsymbol{\zeta}, \boldsymbol{\xi} \in \mathcal{S}^{\times}(f)$. Write $\boldsymbol{\zeta} = \ell(\mathbf{z}) + \lambda \mathbb{1}$ and $\boldsymbol{\xi} = \ell(\mathbf{w}) + \mu \mathbb{1}$, where $\mathbf{z}, \mathbf{w} \in \mathbb{R}^{k-1}$ and $\lambda, \mu \in [0, \infty)$. Let $\mathbf{v} = \boldsymbol{\xi} - \boldsymbol{\zeta} \in \mathbb{R}^k$ and $\mathbf{x} = \boldsymbol{\zeta}$. Take α and β as defined in Proposition IV.62, i.e., we have for all $t \in \mathbb{R}$ that

$$t\mathbf{v} + \mathbf{x} = F(\alpha(t), \beta(t)) = \ell(\alpha(t)) + \beta(t)\mathbb{1}. \tag{4.33}$$

Plugging in $t = 0$ into (4.33), we get $\boldsymbol{\zeta} = \ell(\alpha(0)) + \beta(0)$. Thus, $\alpha(0) = \mathbf{z}$ and $\beta(0) = \lambda$. Likewise, plugging in $t = 1$, we get $\alpha(1) = \mathbf{w}$ and $\beta(1) = \mu$. In particular, we have $\beta(0) \geq 0$ and $\beta(1) \geq 0$. By Proposition IV.62, β is concave. Thus, $\beta(t) \geq 0$ for all $t \in [0, 1]$. In other words,

$$t\mathbf{v} + \mathbf{w} = t\boldsymbol{\xi} + (1 - t)\boldsymbol{\zeta} = \ell(\alpha(t)) + \beta(t)\mathbb{1} \in \mathcal{S}^{\times}(\ell)$$

for all $t \in [0, 1]$. This proves that $\mathcal{S}^{\times}(\ell)$ is convex. □

We will need the following result from [BS13, Theorem 9].

Theorem IV.69 (Beltagy et al. [BS13])**.** *Let C be a nonempty closed convex subset of $\mathbb{R}^n$. If C contains no hyperplane, then $C = \text{conv}(\texttt{bdry}(C))$.*

Proposition IV.70. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Then $\mathcal{S}(\ell) = \mathcal{S}^{\times}(\ell)$ and $\texttt{bdry}(\mathcal{S}(\ell)) = \mathcal{R}(\ell)$.*

Proof. Clearly, $\mathcal{S}^{\times}(\ell)$ is nonempty. Furthermore, by Lemma IV.67 and Lemma IV.68, $\mathcal{S}^{\times}(\ell)$ is closed and convex. Next, note that $\mathcal{S}^{\times}(\ell)$ lies in the nonnegative quadrant $[0,\infty)^k$. Since no hyperplane lies entirely inside the nonnegative quadrant, $\mathcal{S}^{\times}(\ell)$ cannot contain any hyperplane. Hence, we have verified that $\mathcal{S}^{\times}(\ell)$ satisfies the condition of Theorem IV.69. To finish the proof, we have

$$\begin{aligned}
\mathcal{S}(\ell) &= \text{conv}(\mathcal{R}(\ell)) \quad \because \text{Definition of } \mathcal{S}(\ell) && (4.34)\\
&= \text{conv}(\texttt{bdry}(\mathcal{S}^{\times}(\ell))) \quad \because \text{Lemma IV.68} && (4.35)\\
&= \mathcal{S}^{\times}(\ell) \quad \because \text{Theorem IV.69} && (4.36)
\end{aligned}$$

This proves the first part. For the second part, note that by Lemma IV.68, we have $\texttt{bdry}(\mathcal{S}(\ell)) = \texttt{bdry}(\mathcal{S}^{\times}(\ell)) = \mathcal{R}(\ell)$. □

Before we move on, we summarize the important results on $\mathcal{S}(\ell)$ below:

Corollary IV.71. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Recall from Definition IV.66*

$$\mathcal{S}^{\circ}(\ell) := \{\boldsymbol{\zeta} + \lambda\mathbb{1} : \boldsymbol{\zeta} \in \mathcal{R}(\ell), \lambda \in (0,\infty)\}.$$

Then $\mathcal{S}(\ell)$ is a closed and convex set with the following properties:

1. $\mathcal{S}(\ell) = \{\boldsymbol{\zeta} + \lambda\mathbb{1} : \boldsymbol{\zeta} \in \mathcal{R}(\ell), \lambda \in [0,\infty)\}$
2. $\texttt{int}(\mathcal{S}(\ell)) = \mathcal{S}^{\circ}(\ell)$ *(see Definitions IV.10 and IV.66)*
3. $\texttt{bdry}(\mathcal{S}(\ell)) = \texttt{bdry}(\mathcal{S}^{\circ}(\ell)) = \mathcal{R}(\ell)$.

Before we proceed, we state one more result about the set $\mathcal{S}^{\circ}(\ell)$ which will be useful later.

Lemma IV.72. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Then $\mathcal{S}^\circ(\ell) = \{\boldsymbol{\zeta} \in \mathbb{R}^k : \exists \mathbf{z} \in \mathbb{R}^{k-1}$ such that $\boldsymbol{\zeta} > \ell(\mathbf{z})\}$.*

Proof. Recall that by definition we have $\mathcal{S}^\circ(\ell) = \{\ell(\mathbf{z}) + \lambda\mathbb{1} : \mathbf{z} \in \mathbb{R}^{k-1}, \lambda \in (0, \infty)\}$. Thus, the "$\subseteq$" direction is immediate. For the other inclusion, take $\boldsymbol{\zeta} = \ell(\mathbf{z}) + \mathbf{v}$ where $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\mathbf{v} > 0$. Let $\mathbf{x} = \ell(\mathbf{z})$. Take α and β as defined in Proposition IV.62, i.e., we have for all $t \in \mathbb{R}$ that

$$t\mathbf{v} + \mathbf{x} = F(\alpha(t), \beta(t)) = \ell(\alpha(t)) + \beta(t)\mathbb{1}. \tag{4.37}$$

Plugging in $t = 0$ into (4.37), we get $\mathbf{x} = \ell(\mathbf{z}) = \ell(\alpha(0)) + \beta(0)$. Thus, $\alpha(0) = \mathbf{z}$ and $\beta(0) = 0$. Recall that $\mathbf{v} = \boldsymbol{\zeta} - \ell(\mathbf{z}) \geq 0$ by assumption. Hence, by Proposition IV.62, β is strictly increasing. Hence, $\beta(1) > \beta(0) = 0$. Now, plugging in $t = 1$ into (4.37), we get $\mathbf{v} + \mathbf{x} = \mathbf{v} + \ell(\mathbf{z}) = \boldsymbol{\zeta} = \ell(\alpha(1)) + \beta(1)\mathbb{1}$. This shows that $\boldsymbol{\zeta} \in \mathcal{S}^\circ(\ell)$, as desired. □

Remark IV.73. From basic topology, we know that $\mathcal{S}(\ell) = \texttt{int}(\mathcal{S}(\ell)) \cup \texttt{bdry}(\mathcal{S}(\ell))$. Hence, a consequence of Lemma IV.72 and Lemma IV.67 is that $\mathcal{S}(\ell)$ is precisely the *superprediction set* of ℓ (see Williamson et al. [WVR16, Definition 15] and Kalnishkan et al. [KV08]):

$$\mathcal{S}(\ell) = \{\boldsymbol{\zeta} \in \mathbb{R}^k : \exists \mathbf{z} \in \mathbb{R}^{k-1} \text{ such that } \boldsymbol{\zeta} \geq \ell(\mathbf{z})\}.$$

Recall [TB07, Definition 5]:

Definition IV.74 (Tewari et al. [TB07])**.** Let $S \subseteq \mathbb{R}^k_+$ be a set and $\boldsymbol{\zeta} \in \mathbb{R}^k_+$. Define the set

$$\mathcal{N}(\boldsymbol{\zeta}; S) := \{\mathbf{p} \in \Delta^k : \langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0, \forall \boldsymbol{\xi} \in S\}.$$

We say that S is *admissible* if for all $\boldsymbol{\zeta} \in \texttt{bdry}(S)$ and $\mathbf{p} \in \mathcal{N}(\boldsymbol{\zeta}; S)$ we have $\arg\min(\boldsymbol{\zeta}) \subseteq \arg\max(\mathbf{p})$.

Proposition IV.75 (Tewari et al. [TB07])**.** *Let $S \subseteq \mathbb{R}^k_+$ be a symmetric set. If $|\mathcal{N}(\boldsymbol{\zeta}; S)| = 1$ for all $\boldsymbol{\zeta} \in \texttt{bdry}(S)$, then S is admissible.*

Lemma IV.76. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Let $\boldsymbol{\zeta} \in \mathcal{R}(\ell)$. Then we have $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell)) = \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell)) = \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}^\circ(\ell))$.*

Proof. We first prove that $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell)) = \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))$. Since $\mathcal{S}(\ell) \supseteq \mathcal{R}(\ell)$, we immediately have $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell)) \supseteq \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))$. For the other inclusion, we first note that $\mathcal{S}(\ell) = \mathcal{S}^\times(\ell)$ by Proposition IV.70. Thus, every $\boldsymbol{\xi} \in \mathcal{S}(\ell)$ can be written as $\boldsymbol{\xi} = \boldsymbol{\alpha} + \beta\mathbb{1}$ for some $\boldsymbol{\alpha} \in \mathcal{R}(\ell)$ and $\beta \geq 0$. Now, let $\mathbf{p} \in \mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell))$ and let $\boldsymbol{\xi} \in \mathcal{S}(\ell)$ be decomposed as in the preceding sentence. Then

$$\langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle = \langle \boldsymbol{\alpha} + \beta\mathbb{1} - \boldsymbol{\zeta}, \mathbf{p} \rangle = \langle \boldsymbol{\alpha} - \boldsymbol{\zeta}, \mathbf{p} \rangle + \beta \langle \mathbb{1}, \mathbf{p} \rangle \geq 0$$

where the last inequality holds since (1) $\langle \boldsymbol{\alpha} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0$ because $\mathbf{p} \in \mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell))$, and (2) $\beta \geq 0$. Hence, such a $\mathbf{p}$ satisfies $\langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0$, $\forall \boldsymbol{\xi} \in \mathcal{S}(\ell)$ as well which implies that $\mathbf{p} \in \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))$, as desired.

Next, we prove $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell)) = \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}^\circ(\ell))$. Again, since $\mathcal{S}(\ell) \supseteq \mathcal{S}^\circ(\ell)$, we immediately have $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}^\circ(\ell)) \supseteq \mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))$. For the other inclusion, we first note that $\texttt{cl}(\mathcal{S}^\circ(\ell)) = \mathcal{S}(\ell)$. Suppose $\mathbf{p} \in \Delta^k$ is such that $\langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0$ for al $\boldsymbol{\xi} \in \mathcal{S}^\circ(\ell)$. Then by continuity, we must have that $\langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0$ for all $\boldsymbol{\xi} \in \texttt{cl}(\mathcal{S}^\circ(\ell)) = \mathcal{S}(\ell)$. □

Proposition IV.77. *Let $\mathcal{L}$ be a regular PERM loss with reduced form ℓ. Then $\mathcal{S}(\ell)$ and $\mathcal{S}^\circ(\ell)$ are both admissible.*

Proof. By Proposition IV.75, it suffices to check the following two claims hold:

1. for all $\boldsymbol{\zeta} \in \texttt{bdry}(\mathcal{S}(\ell))$ we have $|\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))| = 1$, and
2. for all $\boldsymbol{\zeta} \in \texttt{bdry}(\mathcal{S}^\circ(\ell))$ we have $|\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}^\circ(\ell))| = 1$.

By Corollary IV.71, we have $\texttt{bdry}(\mathcal{S}(\ell)) = \texttt{bdry}(\mathcal{S}^\circ(\ell)) = \mathcal{R}(\ell)$. Hence, by Lemma IV.76, to show both claims it suffices to show that $|\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell))| = 1$ for all $\boldsymbol{\zeta} \in \mathcal{R}(\ell)$.

Note that here we can replace $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}(\ell))$ and $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{S}^\circ(\ell))$ by $\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell))$ because of Lemma IV.76. Below, fix $\boldsymbol{\zeta} = \ell(\mathbf{z}) \in \mathcal{R}(\ell)$ where $\mathbf{z} \in \mathbb{R}^{k-1}$. Then

$$\begin{aligned}
\mathcal{N}(\boldsymbol{\zeta}; \mathcal{R}(\ell)) &= \{\mathbf{p} \in \Delta^k : \langle \boldsymbol{\xi} - \boldsymbol{\zeta}, \mathbf{p} \rangle \geq 0,\ \forall \boldsymbol{\xi} \in \mathcal{R}(\ell)\} \\
&= \{\mathbf{p} \in \Delta^k : \langle \ell(\mathbf{w}) - \ell(\mathbf{z}), \mathbf{p} \rangle \geq 0,\ \forall \mathbf{w} \in \mathbb{R}^{k-1}\} \quad \because \text{Corollary IV.54} \\
&= \left\{\mathbf{p} \in \Delta^k : \mathbf{z} \in \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^{k-1}} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{w})\right\} \quad \because \text{minimizer exists by} \\
&= \left\{\mathbf{p} \in \mathtt{int}(\Delta^k) : \mathbf{z} \in \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^{k-1}} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{w})\right\} \quad \because \text{Corollary IV.51} \\
&= \{\mathbf{p} \in \mathtt{int}(\Delta^k) : \mathbf{z} = \mathtt{lnk}^{\mathcal{L}}(\mathbf{p})\} \quad \because \text{Definition IV.48 and Corollary IV.47}
\end{aligned}$$

By Proposition IV.52, $\mathtt{lnk}^{\mathcal{L}}$ is an injection. Thus, $|\left\{\mathbf{p} \in \mathtt{int}(\Delta^k) : \mathbf{z} = \mathtt{lnk}^{\mathcal{L}}(\mathbf{p})\right\}| = 1$. □

4.7 Proof of Theorem IV.27

Throughout this section, assume that we are in the following **situation**:

1. $\mathcal{L}$ is totally regular. For each $n \in \{2, \ldots, k\}$,
2. $\mathcal{L}^{(n)}$ is the n-ary retracted loss of $\mathcal{L}$,
3. $\psi^{(n)}$ is the template of $\mathcal{L}^n$ (Proposition IV.3)
4. $\ell^{(n)}$ is the reduced form of $\mathcal{L}^n$ (Definition IV.2)

For the reader's convenience, we restate Theorem IV.27, whose proof is the goal of this section:

Theorem (IV.27, restated)**.** *If $\mathcal{L}$ is totally regular, then $\mathcal{S}(\mathcal{L})$ is classification-calibrated.*

Unpacking the definition of totally regular PERM (Definition IV.26), we have that the n-ary retracted loss is a PERM loss for each $n \in \{2, \ldots, k\}$.

Lemma IV.78. *For all* $\mathbf{z} \in \mathbb{R}^{k-1}$ *and all* $j \in [k-1]$*, we have* $\mathtt{prj}\left(\boldsymbol{\rho}_j^{(k)}\mathbf{z}\right) = \boldsymbol{\rho}_j^{(k-1)}\mathtt{prj}(\mathbf{z})$.

Proof. If $j = 1$, then $\boldsymbol{\rho}_1^{(k-1)}$ and $\boldsymbol{\rho}_1^{(k)}$ are both identity matrices. Thus, below we assume that $j > 1$. For each $i \in [k-2]$, we have

$$\left[\mathtt{prj}\left(\boldsymbol{\rho}_j^{(k)}\mathbf{z}\right)\right]_i = [\boldsymbol{\rho}_j^{(k)}\mathbf{z}]_i = \begin{cases} z_i - z_{j-1} & : i \neq j-1 \\ -z_{j-1} & : i = j-1. \end{cases}$$

On the other hand, let $\mathbf{w} = \mathtt{prj}(\mathbf{z})$. Then

$$\left[\boldsymbol{\rho}_j^{(k-1)}\mathbf{w}\right]_i = \begin{cases} w_i - w_{j-1} & : i \neq j-1 \\ -w_{j-1} & : i = j-1. \end{cases}$$

Note that $w_i = z_i$ and $w_{j-1} = z_{j-1}$ since $i, j-1 \in [k-2]$. □

Lemma IV.79. *Assume that we are in the situation stated at the beginning of Section 4.7. Let* $\mathbf{z} \in \mathbb{R}^{k-1}$ *and* $\mathbf{x} \in [0,\infty)^k$*. Define* $\tilde{\mathbf{z}} := \mathtt{prj}(\mathbf{z}) \in \mathbb{R}^{k-2}$ *and* $\tilde{\mathbf{x}} := \mathtt{prj}(\mathbf{x}) \in [0,\infty)^{k-2}$*. Then we have*

$$\lim_{\lambda\to+\infty} \mathtt{prj}\left(\ell^{(k)}\left(\mathbf{z} + \lambda\mathbb{e}_{k-1}^{(k-1)}\right) + \mathbf{x}\right) = \ell^{(k-1)}(\tilde{\mathbf{z}}) + \tilde{\mathbf{x}} \tag{4.38}$$

and

$$\mathtt{prj}\left(\ell^{(k)}(\mathbf{z}) + \mathbf{x}\right) > \ell^{(k-1)}(\tilde{\mathbf{z}}). \tag{4.39}$$

Proof. For brevity, let $\mathbf{u} = \mathbb{e}_{k-1}^{(k-1)}$. First, we claim that $\boldsymbol{\rho}_j^{(k)}\mathbf{u} = \mathbf{u}$ for each $j \in [k-1]$. To see this, first note that if $j = 1$, then $\boldsymbol{\rho}_1^{(k)}$ is the identity. For $k-1 \geq j > 1$, we recall that

$$\forall i \in [k-1],\ [\boldsymbol{\rho}_j^{(k-1)}\mathbf{u}]_i = \begin{cases} u_i - u_{j-1} & : i \neq j-1 \\ -u_{j-1} & : i = j-1. \end{cases}$$

Since $j-1 < k-1$, we have $u_{j-1} = 0$ and so

$$\forall i \in [k-1],\ [\boldsymbol{\rho}_j^{(k)}\mathbf{u}]_i = \begin{cases} u_i & : i \neq j-1 \\ u_{j-1} = 0 & : i = j-1. \end{cases}$$

This shows that $[\boldsymbol{\rho}_j^{(k)}\mathbf{u}]_i = u_i$ for all $i \in [k-1]$, which proves our claim.

Next, still assuming $j \in [k-1]$, we have

$$\ell_j^{(k)}(\mathbf{z}+\lambda\mathbf{u}) = \psi^{(k)}(\boldsymbol{\rho}_j^{(k)}(\mathbf{z}+\lambda\mathbf{u})) = \psi^{(k)}(\boldsymbol{\rho}_j^{(k)}\mathbf{z}+\lambda\mathbf{u}).$$

Hence, we have

$$\begin{aligned}
\lim_{\lambda\to+\infty} \ell_j^{(k)}(\mathbf{z}+\lambda\mathbf{u}) &= \lim_{\lambda\to+\infty} \psi^{(k)}(\boldsymbol{\rho}_j^{(k)}\mathbf{z}+\lambda\mathbf{u}) \quad \because \text{ Proposition IV.33.} \\
&= \psi^{(k-1)}(\mathtt{prj}(\boldsymbol{\rho}_j^{(k)}\mathbf{z})) \quad \because \text{ Proposition IV.25} \\
&= \psi^{(k-1)}\left(\boldsymbol{\rho}_j^{(k-1)}(\mathtt{prj}(\mathbf{z}))\right) \quad \because \text{ Lemma IV.78} \\
&= \psi^{(k-1)}\left(\boldsymbol{\rho}_j^{(k-1)}\tilde{\mathbf{z}}\right) \quad \because \text{ Definition of } \tilde{\mathbf{z}} \\
&= \ell_j^{(k-1)}(\tilde{\mathbf{z}}) \quad \because \text{ Proposition IV.33.}
\end{aligned}$$

Thus,

$$\lim_{\lambda\to+\infty} \mathtt{prj}\left(\ell^{(k)}(\mathbf{z}+\lambda\mathbf{u})+\mathbf{x}\right) = \ell^{(k-1)}(\tilde{\mathbf{z}})+\tilde{\mathbf{x}}.$$

Next, for every $j \in [k-1]$, we note that the function

$$g_j(\lambda) := \ell_j^{(k)}(\mathbf{z}+\lambda\mathbf{u}) = \psi^{(k)}(\boldsymbol{\rho}_j^{(k)}\mathbf{z}+\lambda\mathbf{u})$$

is strictly decreasing. To see this, by the chain rule, we have

$$\nabla_{g_j}(\lambda) = \nabla_{\psi^{(k)}}(\boldsymbol{\rho}_j^{(k)}\mathbf{z}+\lambda\mathbf{u})\mathbf{u} < 0.$$

Thus, $\ell_j^{(k)}(\mathbf{z}) = g_j(0) > \lim_{\lambda \to +\infty} g_j(\lambda) = \ell_j^{(k-1)}(\tilde{\mathbf{z}})$, which proves that

$$\texttt{prj}\left(\ell^{(k)}(\mathbf{z}) + \mathbf{x}\right) > \ell^{(k-1)}(\tilde{\mathbf{z}}) + \tilde{\mathbf{x}} \geq \ell^{(k-1)}(\tilde{\mathbf{z}})$$

as desired. □

Lemma IV.80. *Assume that we are in the situation stated at the beginning of Section 4.7. Then* $\texttt{prj}(\mathcal{S}(\ell^{(k)})) \subseteq \mathcal{S}^\circ(\ell^{(k-1)})$ *and* $\texttt{cl}[\texttt{prj}(\mathcal{S}(\ell^{(k)}))] = \mathcal{S}(\ell^{(k-1)})$.

Proof. Let $C := \texttt{prj}(\mathcal{S}(\ell^{(k)}))$ and take $\boldsymbol{\zeta} \in C$. We first prove $C \subseteq \mathcal{S}^\circ(\ell^{(k-1)})$. By the characterization of $\mathcal{S}(\ell^{(k)})$ from Corollary IV.71 item 1, there exists $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\mathbf{x} \in [0, \infty)^k$ such that $\boldsymbol{\zeta} = \texttt{prj}(\ell(\mathbf{z}) + \mathbf{x})$. Applying Eqn. (4.39) from Lemma IV.79, we get

$$\boldsymbol{\zeta} = \texttt{prj}\left(\ell(\mathbf{z}) + \mathbf{x}\right) > \ell^{(k-1)}(\tilde{\mathbf{z}})$$

where we recall that $\tilde{\mathbf{z}} = \texttt{prj}(\mathbf{z})$. In particular, by the characterization of $\mathcal{S}^\circ(\ell^{(k-1)})$ from Lemma IV.72, we have that $\boldsymbol{\zeta} \in \mathcal{S}^\circ(\ell^{(k-1)})$. This proves that $C \subseteq \mathcal{S}^\circ(\ell^{(k-1)})$.

Next, we prove $\texttt{cl}[C] = \mathcal{S}(\ell^{(k-1)})$. We first show that $\texttt{cl}[C] \supseteq \mathcal{S}(\ell^{(k-1)})$ by proving that every point $\mathcal{S}(\ell^{(k-1)})$ is a limit point of C.

Let $\boldsymbol{\zeta} \in \mathcal{S}(\ell^{(k-1)})$. By the characterization of $\mathcal{S}(\ell^{(k-1)})$ as in Corollary IV.71, there exists $\tilde{\mathbf{z}} \in \mathbb{R}^{k-2}$ and $\tilde{\mathbf{x}} \in [0, \infty)^{k-1}$ such that $\boldsymbol{\zeta} = \ell^{(k-1)}(\tilde{\mathbf{z}}) + \tilde{\mathbf{x}}$. Now, pick $\mathbf{z} \in \mathbb{R}^{k-1}$ and $\mathbf{x} \in [0, \infty)^k$ such that $\tilde{\mathbf{z}} = \texttt{prj}(\mathbf{z})$ and $\tilde{\mathbf{x}} = \texttt{prj}(\mathbf{x})$. Applying Lemma IV.79 Eqn. (4.38), we get that $\boldsymbol{\zeta}$ is a limit point of S, which proves the desired claim. This proves that $\texttt{cl}(C) \supseteq \mathcal{S}(\ell^{(k-1)})$. By the first part, we know that $C \subseteq \mathcal{S}^\circ(\ell^{(k-1)})$. By Corollary IV.71, $\mathcal{S}^\circ(\ell^{(k-1)}) = \texttt{int}(\mathcal{S}(\ell^{(k-1)})) \subseteq \mathcal{S}(\ell^{(k-1)})$. Putting it all together, we have

$$C \subseteq \mathcal{S}^\circ(\ell^{(k-1)}) \subseteq \mathcal{S}(\ell^{(k-1)}) \subseteq \texttt{cl}(C).$$

From Corollary IV.71, we have that $\mathcal{S}(\ell^{(k-1)})$ is closed. Since by definition $\texttt{cl}(C)$ is the smallest closed set containing C, we get that $\mathcal{S}(\ell^{(k-1)}) = \texttt{cl}(C)$, as desired. □

Theorem IV.81 (Blackwell et al. [BG79])**.** *Let $C \subseteq \mathbb{R}^n$ be a convex set. Then* $\mathtt{int}(C) = \mathtt{int}(\mathtt{cl}(C))$.

Proposition IV.82. *Assume that we are in the situation stated at the beginning of Section 4.7. Then we have* $\mathtt{prj}(\mathcal{S}(\ell^{(k)})) = \mathcal{S}^\circ(\ell^{(k-1)})$

Proof. For brevity, let $C := \mathtt{prj}(\mathcal{S}(\ell^{(k)}))$. By Corollary IV.71, $\mathcal{S}(\ell^{(k-1)})$ is convex. Since convexity is preserved under projection, we have that C is convex as well. Now,

$$\begin{aligned}
\mathtt{int}(C) &= \mathtt{int}(\mathtt{cl}(C)) \quad \because \text{ Theorem IV.81} && (4.40)\\
&= \mathtt{int}(\mathcal{S}(\ell^{(k-1)})) \quad \because \text{ Lemma IV.80} && (4.41)\\
&= \mathcal{S}^\circ(\ell^{(k-1)}) \quad \because \text{ Lemma IV.67} && (4.42)\\
&\supseteq C \quad \because \text{ Lemma IV.80} && (4.43)
\end{aligned}$$

Since $C \supseteq \mathtt{int}(C)$ by definition, we conclude that $C = \mathcal{S}^\circ(\ell^{(k-1)})$. □

Proposition IV.83. *Assume that we are in the situation stated at the beginning of Section 4.7. Then we have* $\mathtt{prj}(\mathcal{S}^\circ(\ell^{(k)})) = \mathcal{S}^\circ(\ell^{(k-1)})$.

Proof. By the preceding Proposition IV.82, we have $\mathtt{prj}(\mathcal{S}(\ell^{(k)})) = \mathcal{S}^\circ(\ell^{(k-1)})$. Since $\mathcal{S}^\circ(\ell^{(k)}) \subseteq \mathcal{S}(\ell^{(k)})$ we have $\mathtt{prj}(\mathcal{S}^\circ(\ell^{(k)})) \subseteq \mathtt{prj}(\mathcal{S}(\ell^{(k)}))$. Thus, to prove the result we only have to show $\mathtt{prj}(\mathcal{S}^\circ(\ell^{(k)})) \supseteq \mathcal{S}^\circ(\ell^{(k-1)})$.

To this end, let $\ell^{(k-1)}(\mathbf{w}) \in \mathcal{S}^\circ(\ell^{(k-1)})$ and $\mathbf{z} \in \mathcal{S}(\ell^{(k)})$ be such that $\mathtt{prj}(\ell^{(k)}(\mathbf{z})) = \ell^{(k-1)}(\mathbf{w})$. By Lemma IV.65, there exist $\mathbf{z}^* \in \mathbb{R}^{k-1}$ and $t^* \in \mathbb{R}$ such that $t^* > 0$ and $\mathtt{prj}(\ell^{(k)}(\mathbf{z}^*) + t^*\mathbb{1}) = \ell^{(k-1)}(\mathbf{w})$. Since $\ell^{(k)}(\mathbf{z}^*) + t^*\mathbb{1} \in \mathcal{S}^\circ(\ell^{(k)})$, we get that $\ell^{(k-1)}(\mathbf{w}) \in \mathtt{prj}_n^k(\mathcal{S}^\circ(\ell^{(k)}))$ as desired. □

Below, let $\mathtt{prj}^{(n)}$ denote the n-fold iterated composition of $\mathtt{prj}$. In other words, $\mathtt{prj}^{(n)} := \mathtt{prj} \circ \cdots \circ \mathtt{prj}$ repeated n times.

Proposition IV.84. *Assume that we are in the situation stated at the beginning of Section 4.7. Then we have* $\mathtt{prj}^{(n)}(\mathcal{S}(\ell^{(k)})) = \mathcal{S}^\circ(\ell^{(k-n)})$ *for each* $n \in \{1, \ldots, k-2\}$.

Proof. We prove by induction. The case when $n = 1$ is simply Lemma IV.80. Now, suppose that the result holds for n where $1 < n < k - 2$. Then

$$\begin{aligned}
\mathtt{prj}^{(n+1)}(\mathcal{S}(\ell^{(k)})) &= \mathtt{prj}(\mathtt{prj}^{(n)}(\mathcal{S}(\ell^{(k)}))) \\
&= \mathtt{prj}(\mathcal{S}^{\circ}(\ell^{(k-n)})) \qquad \because \text{Induction hypothesis} \\
&= \mathcal{S}^{\circ}(\ell^{(k-n-1)}) \qquad \because \text{ Proposition IV.84} \\
&= \mathcal{S}^{\circ}(\ell^{(k-(n+1))}).
\end{aligned}$$

This completes the induction step and the desired result follows. □

We recall the following from [TB07, Theorem 7]:

Theorem IV.85 (Tewari et al. [TB07])**.** *Let $S \subseteq \mathbb{R}^k_+$ be a symmetric convex set. Then S is classification calibrated if and only if S is admissible and $\mathtt{prj}^{(n)}(S)$ is admissible for all $n \in \{1, \ldots, k-2\}$.*

Proof of Theorem IV.27. Assume that we are in the situation stated at the beginning of Section 4.7. Let $S = \mathcal{S}(\ell^{(k)})$. By Theorem IV.85, it suffices to prove that S is admissible and $\mathtt{prj}^{(n)}(S)$ is admissible for all $n \in \{1, \ldots, k-2\}$. From Proposition IV.77, we have that $S = \mathcal{S}(\ell^{(k)})$ is admissible. For each $n \in \{1, \ldots, k-2\}$, we have by Proposition IV.84 that $\mathtt{prj}^{(n)}(\mathcal{S}(\ell^{(k)})) = \mathcal{S}^{\circ}(\ell^{(k-n)})$. Again by Proposition IV.77, $\mathcal{S}^{\circ}(\ell^{(k-2)})$ is admissible, which proves the theorem in view of Theorem IV.85. □

4.8 Classification-Calibration of Fenchel-Young losses

The goal of this section is two fold. The first subsection presents the proof of Theorem IV.22. The second subsection shows the existence of a totally regular negentropy that is strictly convex, but not strongly convex.

4.8.1 Proof of Theorem IV.22

Before proceeding with the proof, we establish two key results.

Proposition IV.86. *Let Ω be a negentropy (Definition IV.20) and $\mu \in \mathbb{R}$. Then the Fenchel-Young loss $\mathcal{L}$ associated to Ω and μ is a PERM loss that is closed, convex, and non-negative. The template ψ of $\mathcal{L}$ is semi-coercive and is given by*

$$\psi(\mathbf{z}) = \max_{\tilde{\mathbf{p}} \in \tilde{\Delta}^k} -\tilde{\Omega}(\mathbf{p}) + \mu \mathbb{1}^\top \tilde{\mathbf{p}} - \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle.$$

Furthermore, if Ω is a regular negentropy, then $\mathcal{L}$ is a regular PERM loss.

Proof. In this proof, all elementary basis vectors are implicitly assumed to be k-dimensional, i.e., we write $\mathbb{e}_y$ instead of $\mathbb{e}_y^k$. First, recall that the Fenchel conjugate of a closed convex function is again closed convex [Roc70]. Next, we show that $\mathcal{L}$ is permutation equivariant:

$$\begin{aligned}
[\sigma_j \mathcal{L}(\mathbf{v})]_y &= [\mathcal{L}(\mathbf{v})]_{\sigma_j(y)} \\
&= \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_{\sigma_j(y)}) - \Omega(\mathbf{p}) + \langle \mathbf{v} + \mathbf{c}_{\sigma_j(y)}, \mathbf{p} - \mathbb{e}_{\sigma_j(y)} \rangle \\
&= \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_y) - \Omega(\mathbf{p}) + \langle \sigma_j(\mathbf{v} + \mathbf{c}_{\sigma_j(y)}), \sigma_j(\mathbf{p} - \mathbb{e}_{\sigma_j(y)}) \rangle \\
&= \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_y) - \Omega(\sigma_j(\mathbf{p})) + \langle \sigma_j(\mathbf{v}) + \mathbf{c}_y, \sigma_j(\mathbf{p}) - \mathbb{e}_y) \rangle \\
&= \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_y) - \Omega(\mathbf{p}) + \langle \sigma_j(\mathbf{v}) + \mathbf{c}_y, \mathbf{p} - \mathbb{e}_y) \rangle \\
&= [\mathcal{L}(\sigma_j(\mathbf{v}))]_y.
\end{aligned}$$

This shows that $\sigma \mathcal{L} = \mathcal{L} \sigma$.

Next, we show that $\mathcal{L}$ is margin-based. Recall that

$$[\mathcal{L}(\mathbf{v})]_y = \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_y) - \Omega(\mathbf{p}) + \langle \mathbf{c}_y, \mathbf{p} - \mathbb{e}_y \rangle + \langle \mathbf{v}, \mathbf{p} - \mathbb{e}_y \rangle$$

Since the only term that depends on $\mathbf{v}$ is the last summand $\langle \mathbf{v}, \mathbf{p} - \mathbb{e}_y \rangle$, which we show to only depend on $\mathbf{M}(\mathbf{v})$. First, we observe that

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{p} \rangle &= p_1 v_1 + \cdots + p_k v_k \\
&= (1 - (p_2 + \cdots + p_k)) v_1 + p_2 v_2 + \cdots + p_k v_k \\
&= v_1 - (p_2(v_1 - v_2) + \cdots + p_k(v_1 - v_k)) \\
&= v_1 - \langle \tilde{\mathbf{p}}, \mathbf{M}(\mathbf{v}) \rangle
\end{aligned}$$

where we write $\tilde{\mathbf{p}}$ to denote the vector $(p_2, \ldots, p_k)^\top$. Thus,

$$\langle \mathbf{v}, \mathbf{p} - \mathbb{e}_y \rangle = \langle \mathbf{v}, \mathbf{p} \rangle - v_y = v_1 - v_y - \langle \tilde{\mathbf{p}}, \mathbf{M}(\mathbf{v}) \rangle$$

From this, we deduced that

$$\langle \mathbf{v}, \mathbf{p} - \mathbb{e}_y \rangle = \begin{cases} [\mathbf{M}(\mathbf{v})]_{y-1} - \langle \tilde{\mathbf{p}}, \mathbf{M}(\mathbf{v}) \rangle & : y > 1 \\ -\langle \tilde{\mathbf{p}}, \mathbf{M}(\mathbf{v}) \rangle & : y = 1. \end{cases}$$

This shows that $\mathcal{L}$ is margin-based.

Furthermore,

$$\mathcal{L}_1(\mathbf{v}) = \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_1) - \Omega(\mathbf{p}) + \langle \mathbf{c}_1, \mathbf{p} - \mathbb{e}_1 \rangle - \langle \tilde{\mathbf{p}}, \mathbf{M}(\mathbf{v}) \rangle.$$

Thus,

$$\psi(\mathbf{z}) = \max_{\mathbf{p} \in \Delta^k} \Omega(\mathbb{e}_1) - \Omega(\mathbf{p}) + \langle \mathbf{c}_1, \mathbf{p} - \mathbb{e}_1 \rangle - \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle.$$

Since $\mathbb{e}_1 \in \Delta^k$ and $[\mathbb{e}_1]_i = 0$ for $i \in \{2, \ldots, k\}$, we have by construction that

$$\psi(\mathbf{z}) \geq \Omega(\mathbb{e}_1) - \Omega(\mathbb{e}_1) + \langle \mathbf{c}_1, \mathbb{e}_1 - \mathbb{e}_1 \rangle - \langle \mathbb{0}, \mathbf{z} \rangle = 0.$$

When $\mathbf{c_1} = \mu(\mathbb{1} - \mathbb{e}_\mathbf{1})$ and $\Omega(\mathbb{e}_1) = 0$, we have

$$\psi(\mathbf{z}) = \max_{\tilde{\mathbf{p}} \in \tilde{\Delta}^k} -\tilde{\Omega}(\tilde{\mathbf{p}}) + \mu \mathbb{1}^\top \tilde{\mathbf{p}} - \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle. \tag{4.44}$$

Finally, we prove that ψ is semi-coercive. Let $c \in \mathbb{R}$ and $\mathbf{z} \in \mathbb{R}^{k-1}$ be such that $c \geq \psi(\mathbf{z})$. Let $j \in \arg\min \mathbf{z}$. Then since $\mathbb{e}_j = \mathbb{e}_j^{k-1} \in \tilde{\Delta}^k$, we have

$$c \geq \psi(\mathbf{z}) = \sup_{\mathbf{p} \in \tilde{\Delta}^k} -\tilde{\Omega}(\mathbf{p}) - \langle \mathbf{p}, \mathbf{z} \rangle \geq -\tilde{\Omega}(\mathbb{e}_j) + \mu - \langle \mathbb{e}_j, \mathbf{z} \rangle \geq -z_j = -\min \mathbf{z}.$$

Thus, we have $c \geq \psi(\mathbf{z})$ implies that $\min \mathbf{z} \geq -c$.

Next, we prove the "Furthermore" part. By the first part, it remains to show that ψ is strictly convex, twice differentiable and $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ for all $\mathbf{z} \in \mathbb{R}^{k-1}$. Define $\Upsilon(\tilde{\mathbf{p}}) := \tilde{\Omega}(\tilde{\mathbf{p}}) - \mu \mathbb{1}^\top \tilde{\mathbf{p}}$. Then $\Upsilon : \tilde{\Delta}^k \to \mathbb{R}$ is also of Legendre type. Note that

$$\begin{aligned}
\psi(\mathbf{z}) &= \max_{\tilde{\mathbf{p}} \in \tilde{\Delta}^k} \langle \tilde{\mathbf{p}}, -\mathbf{z} \rangle - \tilde{\Omega}(\tilde{\mathbf{p}}) + \mu \mathbb{1}^\top \tilde{\mathbf{p}} \quad \because \text{Eqn. (4.44)} \\
&= \max_{\tilde{\mathbf{p}} \in \tilde{\Delta}^k} \langle \tilde{\mathbf{p}}, -\mathbf{z} \rangle - \Upsilon(\tilde{\mathbf{p}}) \\
&= \Upsilon^*(-\mathbf{z}) \quad \because \text{definition of Fenchel conjugate}
\end{aligned}$$

We recall the following fundamental theorem regarding convex conjugates [Roc70].

Theorem IV.87 (Rockafellar [Roc70])**.** *If (C, f) is a convex function of Legendre type, then (C^*, f^*) is a convex function of Legendre type. The map $\nabla_f : C \to C^*$ is a homeomorphism and $\nabla_{f^*} = (\nabla_f)^{-1}$.*

By Theorem IV.87, we have

1. The function Υ^*, and hence ψ, is of Legendre type. In particular, ψ is strictly convex.

2. The derivative $\nabla_\Upsilon : \mathtt{int}(\tilde{\Delta}^k) \to \mathbb{R}^{k-1}$ is a bijection and the derivative of Υ^*

satisfies $\nabla_{\Upsilon^*} = (\nabla_\Upsilon)^{-1} : \mathbb{R}^{k-1} \to \mathtt{int}(\tilde{\Delta}^k)$.

It follows that if Υ is twice differentiable, then so is Υ^*. Finally, by the chain rule, we have $\nabla_\psi(\mathbf{z}) = -\nabla_{\Upsilon^*}(\mathbf{z})$. Since $\nabla_{\Upsilon^*}(\mathbf{z}) \in \mathtt{int}(\tilde{\Delta}^k)$ for all $\mathbf{z}$, we have in particular that $\nabla_{\Upsilon^*}(\mathbf{z}) > \mathbb{0}$. Thus, $\nabla_\psi(\mathbf{z}) < \mathbb{0}$ for all $\mathbf{z}$. □

Theorem IV.88. *Let $\mathcal{L}$ be the Fenchel-Young loss corresponding to Ω and μ. Let $n \in \{2, \dots, k\}$. Let $\Omega^{(n)}$ be the n-ary retracted negentropy of Ω (Definition IV.21). Then the n-ary retracted loss $\mathcal{L}^{(n)}$ of $\mathcal{L}$ (Definition IV.26) is the Fenchel-Young loss corresponding to the n-ary retracted negentropy $\Omega^{(n)}$ and μ.*

Definition IV.21. Let $\tilde{\Omega} : \tilde{\Delta}^k \to \mathbb{R}$ be a negentropy and $n \in \{2, \dots, k\}$. The *n-ary retracted negentropy* of $\tilde{\Omega}$, which we denote by $\tilde{\Omega}^{(n)} : \tilde{\Delta}^n \to \mathbb{R}$, is defined as

$$\tilde{\Omega}^{(n)}(\mathbf{q}) := \tilde{\Omega}(\mathtt{inj}_{n-1}^{k-1}(\mathbf{q})), \quad \forall \mathbf{q} \in \tilde{\Delta}^n.$$

Proof. Fix $\mathbf{z} \in \mathbb{R}^{k-1}$ and let $\mathbf{w} = \mathtt{prj}(\mathbf{z}) \in \mathbb{R}^{k-2}$. Let $\mathbf{u} := \mathbb{e}_{k-1}^{(k-1)}$. Let $\tilde{H}^{(k-1)} := -\tilde{\Omega}^{(k-1)}$. To simplify the notation, elements of $\tilde{\Delta}^k$ will be denoted as $\mathbf{p}$ instead of $\tilde{\mathbf{p}}$ (same for $\tilde{\Delta}^{k-1}$). Our goal is to show that

$$\lim_{\lambda \to +\infty} \max_{\mathbf{p} \in \tilde{\Delta}^k} \tilde{H}^{(k)}(\mathbf{p}) + \mu \mathbb{1}^\top \mathbf{p} - \langle \mathbf{z} + \lambda \mathbf{u}, \mathbf{p} \rangle = \max_{\mathbf{q} \in \tilde{\Delta}^{k-1}} \tilde{H}^{(k-1)}(\mathbf{q}) + \mu \mathbb{1}^\top \mathbf{q} - \langle \mathbf{w}, \mathbf{q} \rangle.$$

Note that the left-hand side is $\mathtt{prj}[\psi](\mathbf{w}) := \lim_{\lambda \to \infty} \psi \left(\begin{bmatrix} \mathbf{w}^\top & \lambda \end{bmatrix}^\top \right)$ as in Proposition IV.25 where ψ is the template of the Fenchel-Young loss of $\Omega^{(k)}$ defined as in Eqn. 4.7. Moreover, note that the right-hand side is $\psi'(\mathbf{w})$ where ψ' is the template corresponding to the Fenchel-Young loss correspondin to $\Omega^{(k-1)}$. A priori, it is not immediately obvious why $\mathtt{prj}[\psi] = \psi'$. This proof will confirm that this equality indeed hold.

For brevity, we define $g(\lambda) := \max_{\mathbf{p} \in \tilde{\Delta}^k} \tilde{H}^{(k)}(\mathbf{p}) + \mu \mathbb{1}^\top \mathbf{p} - \langle \mathbf{z} + \lambda \mathbf{u}, \mathbf{p} \rangle$. For all $\lambda \in \mathbb{R}$, let $\mathbf{p}_\lambda^*$ be an arbitrary element of $\arg\max_{\mathbf{p} \in \tilde{\Delta}^k} \tilde{H}^{(k)}(\mathbf{p}) - \langle \mathbf{z} + \lambda \mathbf{u}, \mathbf{p} \rangle$. Note

that $g : \mathbb{R} \to \mathbb{R}_+$ is monotone non-increasing: For $\lambda, \nu \in \mathbb{R}$ such that $\lambda \leq \nu$, we have

$$\begin{aligned} g(\nu) &= \tilde{H}^{(k)}(\mathbf{p}^*_\nu) + \mu \mathbb{1}^\top \mathbf{p}^*_\nu - \langle \mathbf{z} + \nu \mathbf{u}, \mathbf{p}^*_\nu \rangle \\ &\leq \tilde{H}^{(k)}(\mathbf{p}^*_\nu) + \mu \mathbb{1}^\top \mathbf{p}^*_\nu - \langle \mathbf{z} + \lambda \mathbf{u}, \mathbf{p}^*_\nu \rangle \quad \because \nu \geq \lambda \\ &= g(\lambda). \end{aligned}$$

Next, let $\mathbf{q}^* \in \arg\max_{\mathbf{q} \in \tilde{\Delta}^{k-1}} \tilde{H}^{(k-1)}(\mathbf{q}) + \mu \mathbb{1}^\top \mathbf{q} - \langle \mathbf{w}, \mathbf{q} \rangle$ and let $\mathbf{r}^* = \mathtt{inj}^{k-1}_{n-1}(\mathbf{q}^*) \in \tilde{\Delta}^k$. Then we observe that

$$\begin{aligned} g(\lambda) &\geq \tilde{H}^{(k)}(\mathbf{r}^*) + \mu \mathbb{1}^\top \mathbf{r}^* - \langle \mathbf{z} + \lambda \mathbf{u}, \mathbf{r}^* \rangle \\ &= \tilde{H}^{(k)}(\mathbf{r}^*) + \mu \mathbb{1}^\top \mathbf{r}^* - \langle \mathbf{z}, \mathbf{r}^* \rangle \quad \because \langle \mathbf{u}, \mathbf{r}^* \rangle = 0 \\ &= \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{z}, \mathbf{r}^* \rangle \quad \because \mathbb{1}^\top \mathbf{q}^* = \mathbb{1}^\top \mathbf{r}^* \\ &= \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{w}, \mathbf{q}^* \rangle. \end{aligned}$$

Thus, we have that for all $\lambda \in \mathbb{R}$,

$$g(\lambda) \geq \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{w}, \mathbf{q}^* \rangle. \tag{4.45}$$

Now, take a sequence $\{\lambda_t\}_t$ such that $\lim_{t \to \infty} \lambda_t = +\infty$ and $\mathbf{p}^*_{\lambda_t}$ converges to some $\overline{\mathbf{p}} = [\overline{p}_1, \ldots, \overline{p}_k]^\top \in \tilde{\Delta}^k$ as $t \to \infty$. Such a sequence exists because $\tilde{\Delta}^k$ is compact.

We claim that $\overline{p}_k = 0$. Suppose that this is false. Then for all t sufficiently large, there exists an $\epsilon > 0$ such that

$$[\mathbf{p}^*_{\lambda_t}]_k \geq \epsilon.$$

Now, we have

$$\begin{aligned}
g(\lambda_t) &= \tilde{H}^{(k)}(\mathbf{p}^*_{\lambda_t}) + \mu \mathbb{1}^\top \mathbf{p}^*_{\lambda_t} - \langle \mathbf{z} + \lambda_t \mathbf{u}, \mathbf{p}^*_{\lambda_t} \rangle \\
&= \tilde{H}^{(k)}(\mathbf{p}^*_{\lambda_t}) + \mu \mathbb{1}^\top \mathbf{p}^*_{\lambda_t} - \langle \mathbf{z}, \mathbf{p}^*_{\lambda_t} \rangle - \lambda_t \langle \mathbf{u}, \mathbf{p}^*_{\lambda_t} \rangle \\
&\leq g(0) - \lambda_t \langle \mathbf{u}, \mathbf{p}^*_{\lambda_t} \rangle \quad \because \text{definition of } g \\
&\leq g(0) - \lambda_t [\mathbf{p}^*_{\lambda_t}]_k \quad \because \text{definition of } \mathbf{u} \\
&\leq g(0) - \lambda_t \epsilon
\end{aligned}$$

Thus, we have $\lim_{t\to\infty} g(\lambda_t) = -\infty$, which contradicts (4.45). This proves the claim.

Define $\overline{\mathbf{q}} := \mathtt{prj}(\overline{\mathbf{p}})$. Note that the claim we just proved implies that $\mathtt{inj}(\overline{\mathbf{q}}) = \overline{\mathbf{p}}$. Then, we have

$$\begin{aligned}
&\lim_{t\to\infty} g(\lambda_t) \\
&= \lim_{t\to\infty} \tilde{H}^{(k)}(\mathbf{p}^*_{\lambda_t}) + \mu \mathbb{1}^\top \mathbf{p}^*_{\lambda_t} - \langle \mathbf{z}, \mathbf{p}^*_{\lambda_t} \rangle \\
&= \tilde{H}^{(k)}(\overline{\mathbf{p}}) + \mu \mathbb{1}^\top \overline{\mathbf{p}} - \langle \mathbf{z}, \overline{\mathbf{p}} \rangle \quad \because \text{continuity} \\
&= \tilde{H}^{(k)}(\mathtt{inj}(\overline{\mathbf{q}})) + \mu \mathbb{1}^\top \mathtt{inj}(\overline{\mathbf{q}}) - \langle \mathbf{z}, \mathtt{inj}(\overline{\mathbf{q}}) \rangle \\
&= \tilde{H}^{(k)}(\mathtt{inj}(\overline{\mathbf{q}})) + \mu \mathbb{1}^\top \overline{\mathbf{q}} - \langle \mathbf{w}, \overline{\mathbf{q}} \rangle \\
&= \tilde{H}^{(k-1)}(\overline{\mathbf{q}}) + \mu \mathbb{1}^\top \overline{\mathbf{q}} - \langle \mathbf{w}, \overline{\mathbf{q}} \rangle \quad \because \tilde{H}^{(k-1)} \text{ is a nested } \Delta\text{-family} \\
&\leq \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{w}, \mathbf{q}^* \rangle \quad \because \text{Definition of } \mathbf{q}^*.
\end{aligned}$$

From Eqn. (4.45), we get the other inequality

$$\lim_{t\to\infty} g(\lambda_t) \geq \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{w}, \mathbf{q}^* \rangle.$$

Thus, we conclude that $\lim_{t\to\infty} g(\lambda_t) = \tilde{H}^{(k-1)}(\mathbf{q}^*) + \mu \mathbb{1}^\top \mathbf{q}^* - \langle \mathbf{w}, \mathbf{q}^* \rangle$ as desired. This proves the theorem for the case when $n = k-1$. (Note that the case $n = k$ is vacuous and thus trivial). Applying the theorem to $\mathcal{L}^{(k-1)}$, we get the $n = k-2$ case and so

on. □

For the reader's convenience, we restate Theorem IV.22.

Theorem (IV.22, restated)**.** *Let Ω be a totally regular negentropy, $\mu \in \mathbb{R}_+$ be fixed, and $\mathcal{L}$ be the Fenchel-Young loss associated to Ω and the μ. Then $\mathcal{S}(\mathcal{L})$ is classification-calibrated.*

Proof. By Theorem IV.27, it suffices to show that $\mathcal{L}$ is totally regular. For each $n \in \{2, \dots, k\}$, let $\mathcal{L}^{(n)}$ be the n-ary retracted loss of $\mathcal{L}$. Our goal is show is to show that $\mathcal{L}^{(n)}$ is regular. Let $\Omega^{(n)}$ be the n-ary retracted negentropy of Ω (Definition IV.21). By Theorem IV.88, $\mathcal{L}^{(n)}$ is the Fenchel-Young loss corresponding to $\Omega^{(n)}$ and μ. By assumption, $\Omega^{(n)}$ is a regular negentropy. Thus, by Proposition IV.86, $\mathcal{L}^{(n)}$ is regular, as desired. □

4.8.2 Totally regular negentropy that is not strongly convex

In this section, we show that there exists totally regular entropies that are not strongly convex. See Example IV.94. Thus, the associated Fenchel-Young loss is calibrated by Theorem IV.22. Moreover, this calibration result is outside of the purview of previously established results [Blo19; NBR19] which requires strong convexity

Proposition IV.89. *Let $f : D \to \mathbb{R}_{\geq 0}$ be of Legendre type and $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be convex, differentiable and strictly increasing. Let $C = \operatorname{int}(D)$. Suppose that D is compact and there exists $\mathbf{x}^* \in C$ such that $\inf_{\mathbf{x} \in D} f(\mathbf{x}) = f(\mathbf{x}^*)$. Then $g \circ f : D \to \mathbb{R}$ is of Legendre type.*

Proof. We check that the items of Definition IV.19 hold. Item 1 clearly holds since f and $g \circ f$ have the same domain.

Now for Item 2, note that $g \circ f$ is differentiable by the Chain Rule. Thus it remains to show that $g \circ f$ is strictly convex. For all $\mathbf{x}, y \in D$ such that $\mathbf{x} \neq y$ and $\lambda \in (0, 1)$,

we have

$$f(\lambda \mathbf{x} + (1-\lambda)y) < \lambda f(\mathbf{x}) + (1-\lambda) f(y).$$

This is due to f being strictly convex. Next, since g is strictly increasing, we have

$$g(f(\lambda \mathbf{x} + (1-\lambda)y)) < g(\lambda f(\mathbf{x}) + (1-\lambda) f(y))$$

By the convexity of g, we have $g(\lambda f(\mathbf{x}) + (1-\lambda) f(y)) \leq \lambda g(f(\mathbf{x})) + (1-\lambda) g(f(y)))$ which shows that $g \circ f$ is strictly convex.

For Item 3, we check that $\lim_{i\to\infty} \|\nabla_{g\circ f}(\mathbf{x}^i)\| = +\infty$ for all sequences $\{\mathbf{x}^i\} \subseteq C$ such that $\lim_{i\to\infty} \mathbf{x}^i \in \partial D$. To this end, we first prove the claim that there exists $\epsilon > 0$ such that for all sequences $\{\mathbf{x}^i\} \subseteq C$ with $\lim_{i\to\infty} \mathbf{x}^i \in \partial D$ we have $\lim_{i\to\infty} f(\mathbf{x}^i) \geq \epsilon$. Since f is convex on D, we know that f is continuous on D. This is Rockafellar [Roc70, Corollary 10.1.1]. In particular, f is continuous on $\partial D = D \setminus C$ as well. Since ∂D is compact, we have $\inf_{\mathbf{x}\in\partial D} f(\mathbf{x}) = f(\mathbf{x}^\dagger)$ for some $\mathbf{x}^\dagger \in \partial D$. Since $\mathbf{x}^\dagger \neq \mathbf{x}^*$, we must have $f(\mathbf{x}^\dagger) \neq f(\mathbf{x}^*)$ by the strict convexity of f. In particular, $f(\mathbf{x}^\dagger) > f(\mathbf{x}^*)$. Now, letting $\epsilon = f(\mathbf{x}^\dagger)$, the claim follows.

Next we prove that $g'(\epsilon) > 0$. Since g is increasing, we have $g \geq 0$. We proceed by considering the two cases $g'(0) > 0$ and $g'(0) = 0$ separately. In the first case, the convexity of g implies that g' is non-decreasing and so $g'(\epsilon) > 0$ holds. In the second case, if $g'(\epsilon) = 0$, then we must have $g'(t) = 0$ for all $t \in [0, \epsilon]$. But this implies that g is constant on $[0, \epsilon]$ which contradicts that g is strictly increasing. Thus, $g'(\epsilon) > 0$.

Finally, by the Chain Rule, we have $\nabla_{g\circ f}(\mathbf{x}) = \nabla_g(f(\mathbf{x}))\nabla_f(\mathbf{x}) = g'(f(\mathbf{x}))\nabla_f(\mathbf{x})$. Thus,

$$\lim_{i\to\infty} \nabla_{g\circ f}(\mathbf{x}^i) = \lim_{i\to\infty} g'(f(\mathbf{x}^i)) \lim_{i\to\infty} \nabla_f(\mathbf{x}^i) \geq g'(\epsilon) \lim_{i\to\infty} \nabla_f(\mathbf{x}^i).$$

Since $g'(\epsilon) > 0$ and does not depend on i, we have $\lim_{i\to\infty} \|\nabla_{g\circ f}(\mathbf{x}^i)\| = +\infty$, as desired. □

Lemma IV.90. *Let f and g be as in Proposition IV.89. If $g'(0) = 0$ and $\mathbf{x}^* \in \operatorname{int}(D)$ is such that $\inf_{\mathbf{x}\in D} f(\mathbf{x}) = f(\mathbf{x}^*) = 0$, then the Hessian of $g \circ f$ vanishes at $\mathbf{x}^*$, i.e., $\nabla^2_{g\circ f}(\mathbf{x}^*) = 0$.*

Proof. First, we have by the Chain Rule that $\nabla_{g\circ f}(\mathbf{x}) = g'(f(\mathbf{x}))\nabla_f(\mathbf{x})$ and

$$\nabla^2_{g\circ f}(\mathbf{x}) = g''(f(\mathbf{x}))\nabla_f(\mathbf{x})^\top \nabla_f(\mathbf{x}) + g'(f(\mathbf{x}))\nabla^2_f(\mathbf{x}).$$

Note that $\nabla_f(\mathbf{x})$ is a row vector by our convention. By assumption, we have $\nabla_f(\mathbf{x}^*) = 0$ and $g'(f(\mathbf{x})) = g'(0) = 0$. Thus, in light of the formula for $\nabla^2_{g\circ f}(\mathbf{x})$ derived above, we are done. □

Corollary IV.91. *Let f and g be as in Lemma IV.90. Then $g \circ f$ is not α-strongly convex for any $\alpha > 0$.*

Proposition IV.92. *Let $\Omega : \Delta^k \to \mathbb{R}$ be a regular negentropy and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be as in Proposition IV.89. Furthermore, suppose that g is twice differentiable. Let $a \in \mathbb{R}$ be a negative number such that $a \leq \Omega(\mathbf{p})$ for all $\mathbf{p} \in \Delta^k$. Define $\Theta : \Delta^k \to \mathbb{R}$ by*

$$\Theta(\mathbf{p}) := g(\Omega(\mathbf{p}) - a) - g(-a), \quad \forall \mathbf{p} \in \Delta^k.$$

Then Ω is a regular negentropy.

Proof. We first check that Θ is a negentropy. Clearly, Θ is symmetric (item 2 of Definition IV.20). Below, let $\mathbf{p} \in \Delta^k$ be arbitrary and let $\tilde{\mathbf{p}} = (p_2, \ldots, p_k)^\top \in \tilde{\Delta}^k$.

By assumption on a, we have $0 \leq \Omega(\mathbf{p}) - a \leq -a$. Therefore, by g being monotone, we have $g(\Omega(\mathbf{p}) - \Omega(\mathbf{u})) \leq g(-a)$. This proves that $\Theta(\mathbf{p}) \leq 0$. Since $\Omega(\mathbb{e}^k_i) = 0$, we have $\Theta(\mathbb{e}^k_i) = 0$ as well. This proves item 3 of Definition IV.20.

Next, since $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous and strictly increasing, g is a homeomorphism. In particular, g is closed. Since Ω and g are both closed, it follows that Θ is also closed.

It is easy to see that $\tilde{\Theta}(\tilde{\mathbf{p}}) := g(\tilde{\Omega}(\tilde{\mathbf{p}}) - a) - g(-a)$. Thus, $\tilde{\Theta}$ is twice differentiable in the interior of $\tilde{\Delta}^k$. Furthermore, by Proposition IV.89, we get that $\tilde{\Theta}$ is of Legendre type. In particular, $\tilde{\Theta}$ is strictly convex, and so Θ is convex. This proves item 1 of Definition IV.20. Thus, we have prove that Θ is a regular negentropy. □

Proposition IV.93. *Let $\Omega : \Delta^k \to \mathbb{R}$ be a totally regular negentropy and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be as in Proposition IV.89. Furthermore, suppose that g is twice differentiable. Let $\mathbf{u} := (1/k)\mathbb{1}_k \in \Delta^k$ be uniform probability vector. Define $\Theta : \Delta^k \to \mathbb{R}$ by*

$$\Theta(\mathbf{p}) := g(\Omega(\mathbf{p}) - \Omega(\mathbf{u})) - g(-\Omega(\mathbf{u})), \quad \forall \mathbf{p} \in \Delta^k.$$

Then Θ is a totally regular negentropy. Furthermore, if $g'(0) = 0$, then Θ is not strongly-convex.

Proof. By Definition IV.21, we must show that $\Theta^{(n)}$ is a regular negentropy for each $n \in \{2, \ldots, k\}$. Let $a = \Omega(\mathbf{u})$. Note that since Ω is symmetric and convex, we must have that $a \leq \Omega(\mathbf{p})$ for all $\mathbf{p} \in \Delta^k$. Furthermore, it is easy to see that $\tilde{\Theta}^{(n)}(\tilde{\mathbf{q}}) := g(\tilde{\Omega}^{(n)}(\tilde{\mathbf{q}}) - a) - g(-a)$ for all $\mathbf{q} \in \tilde{\Delta}^n$. Now, apply Proposition IV.92 to $\Theta^{(n)}$ and $a = \Omega(\mathbf{u})$, we get the desired result.

The "Furthermore" part follows immediately from Corollary . □

Example IV.94. For a concrete example, take $\Omega = -H$ to be the negative Shannon entropy and $g(x) = x^2$ the square function.

4.9 Gamma-Phi loss

Definition IV.95 (Convergence in extended reals)**.** Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ and $\overline{\mathbb{R}}_{\geq 0} = \mathbb{R}_{\geq 0} \cup \{+\infty\}$. A sequence $\{z^t\}_t \subseteq \mathbb{R}$ has a limit in $\overline{\mathbb{R}}$ if one of the following holds: 1) $\{z^t\}$ has a limit in the usual sense, 2) for all $c \in \mathbb{R}$, we have $z^t \geq c$ (resp. $z^t \leq c$) for all $t \gg 0$ in which case we say $\lim_t z^t = +\infty$ (resp. $\lim_t z^t = -\infty$).

We have the following elementary properties regarding convergence in the extended reals:

Proposition IV.96. *Let* $\{z^t\}$ *and* $\{\tilde{z}^t\}$ *be sequences in* $\mathbb{R}$ *that has a limit in* $\overline{\mathbb{R}}$*. Then* $z^t + \tilde{z}^t$ *has a limit in* $\overline{\mathbb{R}}$ *equal to* $\lim_t z^t + \lim_t \tilde{z}^t$ *if any of the following holds:*

1. *at least one of* $\lim_t z^t$ *or* $\lim_t \tilde{z}^t$ *is finite, i.e.,* $\in \mathbb{R}$*,*
2. $\{z^t\}_t$ *and* $\{\tilde{z}^t\}_t$ *are both* $\subseteq [0, \infty)$*,*
3. $\{z^t\}_t$ *and* $\{\tilde{z}^t\}_t$ *are both* $\subseteq (-\infty, 0]$*.*

Definition IV.97. A function $f : \mathbb{R} \to \mathbb{R}$ is *monotone non-increasing* (resp. non-decreasing) if $f(x) \geq f(y)$ for all $x, y \in \mathbb{R}$ such that $x \leq y$ (resp. $x \geq y$).

Lemma IV.98. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be continuous and monotone non-increasing. Suppose that* $\{z^t\}_t \subseteq \mathbb{R}$ *has a limit* $z^* \in \overline{\mathbb{R}}$*. Then* $f(z^t)$ *has a limit* $\in \overline{\mathbb{R}}$ *and*

$$\lim_t f(z^t) = \begin{cases} f(z^*) & : z^* \in \mathbb{R} \\ \inf_{x \in \mathbb{R}} f(x) & : z^* = +\infty \\ \sup_{x \in \mathbb{R}} f(x) & : z^* = -\infty. \end{cases} \tag{4.46}$$

Thus, the statement $\lim_t f(z^t) = f(\lim_t z^t)$ *is correct. When* f *is monotone non-*decreasing*, Equation* (4.46) *holds with the* inf *and* sup *swapped.*

Proof. If $z^* \in \mathbb{R}$, then the result is simply the definition of continuity. Next, suppose that $z^* = +\infty$. Our goal is to show that $\lim_t f(z^t)$ exists and converges to $I := \inf_{x \in \mathbb{R}} f(x)$.

Consider the case that $I = -\infty$. Then for any $U \in \mathbb{R}$, there exists $u \in \mathbb{R}$ such that $f(u) \leq U$. Since $z^* = +\infty$, $z_t \geq u$ for all $t \gg 0$ sufficiently large, and in which case $f(z_t) \leq f(u) \leq U$. Since $U \in \mathbb{R}$ is arbitrary, we have that $\lim_t f(z^t) = -\infty$ (Definition IV.95).

Now, consider the case that $I \in \mathbb{R}$. Then by definition $f(z^t) \geq I$ for all t. Furthermore, for any $\epsilon > 0$, there exists u such that $f(u) \leq I + \epsilon$. Again, since

$z^* = +\infty$, $z_t \geq u$ for all $t \gg 0$ sufficiently large, in which case $f(z_t) \leq f(u) \leq I + \epsilon$. Since $\epsilon > 0$ is arbitrary, this proves that $\lim_t f(z^t) = I$. The proof for the case when $z^* = -\infty$ is completely analogous. Furthermore, when f is monotone non-decreasing, the roles of inf and sup are clearly swapped. □

Definition IV.99. A sequence of vectors $\{\mathbf{v}^t\}_t \in \mathbb{R}^k$ is *totally convergent* if for all $y, j \in [k]$, both sequences of real numbers $\{v_y^t\}$ and $\{v_y^t - v_j^t\}$ have limits in $\overline{\mathbb{R}}$.

Lemma IV.100. *Every sequence $\{\mathbf{v}^t\}_t \in \mathbb{R}^k$ has a subsequence that is totally convergent.*

Proof. Every sequence of real numbers has a convergent subsequence with limit in $\mathbb{R} \cup \{\pm\infty\}$. By repeatedly passing to convergent subsequences, first for all $j \in [k]$, then for all pairs $j, j' \in [k]$ with $j < j'$, we get the desired result. □

Lemma IV.101. *Let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ be a totally convergent sequence and $\mathbf{p} \in \Delta^k$. Then the limit $\lim_t C_{\mathbf{p}}(\mathbf{v}^t)$ exists and is $\in [0, +\infty]$. If $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ is another totally convergent sequence such that $\lim_t v_y^t - v_j^t = \lim_t \tilde{v}_y^t - \tilde{v}_j^t$ for all $y, j \in [k]$, then $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = \lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t)$.*

Proof. Define $a_y^t := \sum_{j \in [k]: j \neq y} \phi(v_y^t - v_j^t)$ and $\tilde{a}_y^t := \sum_{j \in [k]: j \neq y} \phi(\tilde{v}_y^t - \tilde{v}_j^t)$. We proceed stepwise as follows:

Step 1: $\lim_t \phi(v_y^t - v_j^t) = \lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t)$ as elements of $[0, +\infty]$,
Step 2: $\lim_t a_y^t = \lim_t \tilde{a}_y^t$ as elements of $[0, +\infty]$,
Step 3: $\lim_t \gamma(a_y^t) = \lim_t \gamma(\tilde{a}_y^t)$ as elements of $[0, +\infty]$
Step 4: $\lim_t \sum_{y \in [k]} p_y \gamma(a_y^t) = \lim_t \sum_{y \in [k]} p_y \gamma(\tilde{a}_y^t)$

Proof of Step 1. From Lemma IV.98 and that fact that ϕ is monotone and continuous, we get that $\lim_t \phi(v_y^t - v_j^t) = \phi(\lim_t v_y^t - v_j^t)$ and $\lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t) = \phi(\lim_t \tilde{v}_y^t - \tilde{v}_j^t)$. Note that Lemma IV.98 also guarantees that these limits exist. Non-negativity of the limit values follows from the non-negativity of ϕ.

Step 2. From Proposition IV.96 and the non-negativity of ϕ, we have

$$\lim a_y^t = \sum_{j\in[k]\setminus\{y\}} \lim_t \phi(v_y^t - v_j^t) = \sum_{j\in[k]\setminus\{y\}} \lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t) = \lim \tilde{a}_y^t$$

where the equality in the middle follows from Step 1. Note that Proposition IV.96 also guarantees that these limits exist.

Step 3. This follows from Step 2, Lemma IV.98 and the non-negativity of γ on $[0,\infty)$.

Step 4. This follows from Step 3 and Proposition IV.96. □

Corollary IV.102. *Let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ be a totally convergent sequence and $S \subseteq [k]$ be a set such that $\lim_t v_y^t \in \mathbb{R}$ for all $y \in S$. Define $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ by $\tilde{v}_j^t := v_j^t$ if $j \notin S$ and $\tilde{v}_j^t := \lim_t v_j^t$ if $j \in S$. Then $\lim_t C_p(\mathbf{v}^t) = \lim_t C_p(\tilde{\mathbf{v}}^t)$ as elements of $[0,+\infty]$.*

Proof. Note that $\{\mathbf{v}^t\}_t$ and $\{\tilde{\mathbf{v}}^t\}_t$ satisfy the conditions of Lemma IV.101. □

Proposition IV.103. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where γ satisfies Definition IV.13 condition (G2) and ϕ satisfies Definition IV.14 condition (F). Let $\mathbf{p} \in \Delta^k$ and $y, y' \in [k]$ be such that $p_{y'} > p_y$. Suppose $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ is a sequence where $\liminf_t v_y^t - v_{y'}^t > 0$ and $\lim_t C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}^t) < +\infty$ exists. Then $\lim_t C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}^t) > C_{\mathbf{p}}^{\mathcal{L}*}$.*

Proof. Suppose that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$. We show that this leads to a contradiction. Since $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) < +\infty$ and $p_{y'} > p_y \geq 0$, we have $\limsup_t p_{y'}\gamma\left(\sum_{j\in[k]\setminus\{y'\}} \phi(v_{y'}^t - v_j^t)\right) < \infty$. By our assumptions on γ from Theorem IV.15, we have

$$M := \limsup_t \sum_{j\in[k]\setminus\{y'\}} \phi(v_{y'}^t - v_j^t) < \infty.$$

By assumption, there exists $\epsilon > 0$ such that $v_y^t \geq v_{y'}^t + \epsilon$ for all $t \gg 0$. Below, we assume t is in this sufficiently large regime. Hence, for all $j \in [k]$ we have $v_y^t - v_j^t >$

$v^t_{y'} - v^t_j$ and consequently $\phi(v^t_y - v^t_j) \leq \phi(v^t_{y'} - v^t_j)$. Furthermore, $v^t_y - v^t_{y'} \geq \epsilon > 0 > -\epsilon \geq v^t_{y'} - v^t_y$ and so $\phi(v^t_y - v^t_{y'}) \leq \phi(\epsilon) < \phi(-\epsilon) \leq \phi(v^t_{y'} - v^t_y)$. Let $a^t = \sum_{j \in [k] \setminus \{y'\}} \phi(v^t_{y'} - v^t_j)$ and $b^t = \sum_{j \in [k] \setminus \{y\}} \phi(v^t_y - v^t_j)$. Furthermore, define $\tilde{a}^t := \phi(-\epsilon) + \sum_{j \in [k] \setminus \{y, y'\}} \phi(v^t_{y'} - v^t_j)$ and $\tilde{b}^t := \phi(\epsilon) + \sum_{j \in [k] \setminus \{y, y'\}} \phi(v^t_y - v^t_j)$. Observe that

$$\tilde{a}^t - \tilde{b}^t = \phi(-\epsilon) - \phi(\epsilon) + \sum_{j \in [k] \setminus \{y, y'\}} \phi(v^t_{y'} - v^t_j) - \phi(v^t_y - v^t_j) \geq \phi(-\epsilon) - \phi(\epsilon).$$

In summary, we have $0 \leq b^t \leq \tilde{b}^t \leq \tilde{a}^t \leq a^t \leq M < \infty$. Let $\tau \in \mathtt{Sym}(k)$ be the permutation that swaps y and y'. By Lemma IV.29, we have

$$C_p(\mathbf{v}^t) - C_p(\tau(\mathbf{v}^t)) = (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}^t) - \mathcal{L}_{y'}(\mathbf{v}^t)) = (p_y - p_{y'})(\gamma(a^t) - \gamma(b^t)).$$

By the Fundamental Theorem of Calculus, we have

$$\begin{aligned} \gamma(a^t) - \gamma(b^t) = \int_{b^t}^{a^t} \gamma'(x)dx \geq \int_{\tilde{b}^t}^{\tilde{a}^t} \gamma'(x)dx &\geq (\tilde{a}^t - \tilde{b}^t) \inf_{x \in [\tilde{b}^t, \tilde{a}^t]} \gamma'(x) \\ &\geq (\phi(-\epsilon) - \phi(\epsilon)) \inf_{x \in [0, M]} \gamma'(x). \end{aligned}$$

By our assumption on γ, we have $\delta := \inf_{x \in [0,M]} \gamma'(x) > 0$. Thus,

$$\lim_{t \to \infty} C_{\mathbf{p}}(\mathbf{v}^t) - C_{\mathbf{p}}(\tau(\mathbf{v}^t)) \geq (p_y - p_{y'})(\phi(-\epsilon) - \phi(\epsilon))\delta > 0$$

where the right hand side is a positive quantity independent of t. Therefore,

$$\lim_{t \to \infty} C_{\mathbf{p}}(\mathbf{v}^t) > \lim_{t \to \infty} C_{\mathbf{p}}(\tau(\mathbf{v}^t)).$$

This contradicts that $\lim_{t \to \infty} C_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$. □

Before proceeding, we adopt the notation $\{v^t\}_t \equiv \alpha$ to denote that $v^t = \alpha$ for all

t, where $\{v^t\}_t \subseteq \mathbb{R}$ is a sequence of real numbers and $c \in \mathbb{R}$ is a constant.

Proposition IV.104. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where γ satisfies Definition IV.13 (G1) and ϕ satisfies Definition IV.14 (F). Let $\mathbf{p} \in \Delta^k_{\texttt{desc}}$ and $z \in [k]$ be such that $C^*_{\mathbf{p}} = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_z = \max \mathbf{v}\}$. Then there exists a sequence $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^{k-1}$ satisfying the following properties:*

1. $\lim_{t\to\infty} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$
2. *there exists an index $\ell \in [k]$ and a vector $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_\ell) \in \mathbb{R}^\ell$ such that for each $j \in \{1, \ldots, \ell\}$ we have $\{v^t_j\} \equiv \alpha_j$ and $\lim_t v^t_j = -\infty$ for $j > \ell$. In addition, $\alpha_1 = 0$.*
3. *Let $\mathbf{q} := (\sum_{j=1}^{\ell} p_j)^{-1} (p_1, \ldots, p_\ell) \in \Delta^\ell_{\texttt{desc}}$. Then $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$.*

Furthermore, suppose γ satisfies Definition IV.13 (G2), $z > 1$, and $p_{z-1} > p_z$. Then $\{\mathbf{v}^t\}$ can be chosen to further satisfy $\alpha_j = 0$ for all $j \in [z]$.

Proof. Let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^{k-1}$ be a sequence such that $\lim_{t\to\infty} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$ and $v^t_z = \max \mathbf{v}^t$ for all $t \in \mathbb{N}$. Throughout, t denotes the index of the sequence where "for all t" means "for all $t \in \mathbb{N}$". We will refine the sequence $\mathbf{v}^t$ until all properties P1-5 below are met in addition. Properties marked by $(\star)$ are only guaranteed when γ satisfies Definition IV.13 (G2), i.e., the condition in the "Furthermore" part of the result.

Properties

I. $\max \mathbf{v}^t = 0$ for all t

II. $\{\mathbf{v}^t\}_t$ is totally convergent and $\{v^t_j\}_t$ has a limit in $[-\infty, 0]$ for each $j \in [k]$

III. $(\star)$ the sequence $\{v^t_j\}_t \equiv 0$ for each $j \in [z]$

IV. there exists an $\ell \in [k]$ such that for each $j \in [\ell]$, we have $\{v^t_j\} \equiv \alpha_j$ where $\alpha_j \in (-\infty, 0]$ and for each $j \in [k] \setminus [\ell] := \{\ell + 1, \ldots, k\}$, we have $\lim_t v^t_j = -\infty$.

V. $(\star)$ $\ell \geq z$.

Properties I and II. To begin, note that $C_p(\mathbf{v}) = C_p(\mathbf{v} - c\mathbf{1})$ for any $c \in \mathbb{R}$ and any

$\mathbf{v} \in \mathbb{R}^k$. Replacing each $\mathbf{v}^t$ by $\mathbf{v}^t - (\max \mathbf{v}^t)\mathbf{1}$ for all t, we may assume $v_z^t = \max \mathbf{v}^t = 0$ for all t. In particular, $v_j^t \in (-\infty, 0]$ for all $j \in [k]$ and t. Passing to a subsequence if necessary, we may assume that $\{\mathbf{v}^t\}_t$ is totally convergent (Lemma IV.100).

Property III. ($\star$) By Property I, we already have $v_z^t = \max \mathbf{v}^t = 0$. By the assumption that $p_{z-1} > p_z$ and that $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$, we have $p_j > p_z$ for each $j \in [z-1]$. Furthermore, by Property 2, $v_z^t - v_j^t = -v_j^t$ has a limit in $[0, \infty]$. By Proposition IV.103, $\lim_t -v_j^t \notin (0, \infty]$. Thus $\lim_t -v_j^t = 0$. Now, define the sequence $\{\tilde{\mathbf{v}}^t\}$ by

$$\tilde{v}_j^t := \begin{cases} 0 & : j \in \{1, \ldots, z-1\} \\ v_j^t & : j \in \{z, \ldots, k\} \end{cases}$$

for all t. By Corollary IV.102, we have $\{\tilde{\mathbf{v}}^t\}_t$ is also totally convergent, $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$. Thus, $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = C^*_{\mathbf{p}}$. Replacing $\mathbf{v}^t$ by $\tilde{\mathbf{v}}^t$, we have that Property III holds.

Property IV. Let $\sigma^t \in \mathtt{Sym}(k)$ be the permutation that sorts $\mathbf{v}^t$ in non-increasing order as in Proposition IV.30, i.e., $v^t_{\sigma^t(1)} \geq \cdots \geq v^t_{\sigma^t(k)}$. By Proposition IV.30, $C_{\mathbf{p}}(\sigma^t(\mathbf{v}^t)) \leq C_{\mathbf{p}}(\mathbf{v}^t)$ and hence $\lim_t C_{\mathbf{p}}(\sigma^t(\mathbf{v}^t)) = C^*_{\mathbf{p}}$ as well. We now replace $\mathbf{v}^t$ by $\sigma^t(\mathbf{v}^t)$. Due to the sorting, the new $\mathbf{v}^t$ may no longer be totally convergent. However, passing to a subsequence if necessary, we can still assume that the new $\mathbf{v}^t$ is totally convergent. Note that Property III still holds after sorting. From Property I, we have that $\max \mathbf{v}^t = 0$ and so $\{v_1^t\}_t \equiv 0$.

By Property II, we have $\lim_t v_j^t \in [-\infty, 0]$. By the sorting in the preceding paragraph, we have that $\lim_t v_1^t \geq \cdots \geq \lim_t v_k^t$. Now, let $\ell \in [k]$ be the largest index such that $\lim_t v_\ell^t > -\infty$. Such an index exists because $\lim_t v_1^t = 0$. Let $\alpha_j := \lim_t v_j^t \in (-\infty, 0]$ for each $j \in \{1, \ldots, \ell\}$. Define $\tilde{\mathbf{v}}^t$ such that $\{\tilde{v}_j^t\}_t \equiv \alpha_j$ for $j \in \{1, \ldots, \ell\}$ and $\{\tilde{v}_j^t\}_t = \{v_j^t\}_t$ for $j > \ell$. Then by Corollary IV.102, we have $\tilde{\mathbf{v}}^t$ is totally convergent, and $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$. Replace v^t by $\tilde{v}^t$.

Property V. ($\star$) By Property III, $\{v_j^t\} \equiv 0$ for each $j \in [z]$. Hence, by the definition of ℓ, we have $\ell \geq z$.

We now proceed with the rest of the proof for Proposition IV.104. Consider the sequence $\{\mathbf{v}^t\}_t$ constructed as above. Then items 1 and 2 of Proposition IV.104, as well as the "Furthermore" part already hold. It only remains to check item 3 of Proposition IV.104. Below, we write $[k] \setminus [\ell] := \{\ell+1, \ldots, k\}$. Now, note that

$$\lim_t C_{\mathbf{p}}(\mathbf{v}^t) \tag{4.47}$$

$$= \sum_{y\in[k]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) \tag{4.48}$$

$$= \sum_{y\in[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) + \underbrace{\sum_{y\in[k]\setminus[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big)}_{=:A} \tag{4.49}$$

$$= (\underbrace{p_1 + \cdots + p_\ell}_{=:S}) \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) + A. \tag{4.50}$$

Now, we focus on $\lim_t v_y^t - v_j^t$ case by case:

$$\lim_t v_y^t - v_j^t = \begin{cases} \alpha_y - \alpha_j & : y \in [\ell],\, j \in [\ell] \\ \alpha_y - \lim_t v_j^t = +\infty & : y \in [\ell],\, j \in [k] \setminus [\ell] \\ \lim_t v_y^t - \alpha_j = -\infty & : y \in [k] \setminus [\ell],\, j \in [\ell]. \end{cases}$$

Note that we omitted the $y \in \{\ell+1, \ldots, k\}$, $j \in \{\ell+1, \ldots, k\}$ case in which we leave $\lim_t v_y^t - v_j^t$ as is without further simplification. Now,

$$\phi(\lim_t v_y^t - v_j^t) = \begin{cases} \phi(\alpha_y - \alpha_j) & : j \in [\ell] \\ \phi(+\infty) = 0 & : j \in \{\ell+1, \ldots, k\}. \end{cases}$$

Putting it all together, we have

$$\lim_t C_{\mathbf{p}}(v^t) \tag{4.51}$$

$$= S \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) + A \tag{4.52}$$

$$= S \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[\ell]:j\neq y} \phi(\alpha_y - \alpha_j) + \sum_{j\in[k]\setminus[\ell]:j\neq y} \phi(+\infty)\Big) + A \tag{4.53}$$

$$= S \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[\ell]:j\neq y} \phi(\alpha_y - \alpha_j)\Big) + A \tag{4.54}$$

$$= S \cdot C_{\mathbf{q}}(\boldsymbol{\alpha}) + A. \tag{4.55}$$

Now, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\ell) \in \mathbb{R}^\ell$ be arbitrary and define a sequence $\{w^t\} \subseteq \mathbb{R}^k$ by

$$w_j^t := \begin{cases} \beta_j & : j \in [\ell] \\ v_j^t & : j \in [k] \setminus [\ell]. \end{cases}$$

Then similar to the above, we have the decomposition

$$\lim_t C_{\mathbf{p}}(w_j^t) = \sum_{y\in[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t w_y^t - w_j^t)\Big) + \underbrace{\sum_{y\in[k]\setminus[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t w_y^t - w_j^t)\Big)}_{=:B}.$$

We claim that $A = B$ and $\lim_t C_{\mathbf{p}}(w^t) = S \cdot C_{\mathbf{q}}(\boldsymbol{\beta}) + A$. We first prove that $A = B$. To this end, observe that

$$\lim_t w_y^t - w_j^t = \begin{cases} \beta_y - \beta_j & : y \in [\ell],\, j \in [\ell] \\ \beta_y - \lim_t v_j^t = +\infty & : y \in [\ell],\, j \in [k] \setminus [\ell] \\ \lim_t v_y^t - \beta_j = -\infty & : y \in [k] \setminus [\ell],\, j \in [\ell] \\ \lim_t v_y^t - v_j^t & : y \in [k] \setminus [\ell],\, j \in [k] \setminus [\ell]. \end{cases}$$

In particular, for $y \in [k] \setminus [\ell], j \in [\ell]$, we have $\lim_t w_y^t - w_j^t = -\infty = \lim_t v_y^t - v_j^t$. Thus,

$$\begin{aligned} B &= \sum_{y\in[k]\setminus[\ell]} p_y \gamma\Big(\sum_{j\in[\ell]:j\neq y} \phi(\lim_t w_y^t - w_j^t) + \sum_{j\in[k]\setminus[\ell]:j\neq y} \phi(\lim_t w_y^t - w_j^t)\Big) \\ &= \sum_{y\in[k]\setminus[\ell]} p_y \gamma\Big(\sum_{j\in[\ell]:j\neq y} \phi(-\infty) + \sum_{j\in[k]\setminus[\ell]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) \\ &= A. \end{aligned}$$

Next, we have

$$\begin{aligned} \lim_t C_{\mathbf{p}}(w^t) &= \sum_{y\in[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t w_y^t - w_j^t)\Big) + A \\ &= S \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[\ell]:j\neq y} \phi(\beta_y - \beta_j) + \sum_{j\in[k]\setminus[\ell]:j\neq y} \phi(+\infty)\Big) + A \\ &= S \cdot C_{\mathbf{q}}(\boldsymbol{\beta}) + A. \end{aligned}$$

Since $\lim_t C_{\mathbf{p}}(w^t) \geq \lim_t C_{\mathbf{p}}(v^t) = C_{\mathbf{p}}^*$, we have $C_{\mathbf{q}}(\boldsymbol{\beta}) \geq C_{\mathbf{q}}(\boldsymbol{\alpha})$. Since β is arbitrary, this proves that $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C_{\mathbf{q}}^*$.

□

Lemma IV.105. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where γ satisfies Definition IV.13 (G1) and ϕ satisfies Definition IV.14 (F). Let $\{\mathbf{v}^t\}_t$ be any sequence satisfying items 1, 2 and 3 of Proposition IV.104. If $p_y = 0$ for each $y > \ell$, then $C_{\mathbf{q}}(\boldsymbol{\alpha}) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$.*

Proof. In Equation (4.51), we showed that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = S \cdot C_{\mathbf{q}}(\boldsymbol{\alpha}) + A$. Where S and A are defined on Equations (4.50) and (4.49) respectively. If $p_y = 0$ for all $y > \ell$, then clearly $S = 1$ and $A = 0$. □

Proposition IV.106. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where γ satisfies Definition IV.13 condition (G2), and ϕ satisfies Definition IV.14 condition*

(F). Suppose that $\mathbf{q} \in \Delta^\ell$ *and* $\boldsymbol{\alpha} \in \mathbb{R}^\ell$ *are such that* $\boldsymbol{\alpha}$ *is a minimizer of* $C_{\mathbf{q}}(\cdot)$ *and* $\alpha_1 = \alpha_2$*. Then* $q_1 = q_2$*.*

Proof. Recall that

$$C_{\mathbf{q}}(\mathbf{v}) = \sum_{y \in [\ell]} q_y \gamma \left(\sum_{j \in [k] \setminus \{y\}} \phi(v_y - v_j) \right).$$

For each y, define $\Gamma_y(\mathbf{v}) := \gamma' \left(\sum_{j \in [k] \setminus \{y\}} \phi(v_y - v_j) \right)$. Thus

$$\frac{\partial C_{\mathbf{q}}}{\partial v_y}(\mathbf{v}) = \left(q_y \Gamma_y(\mathbf{v}) \sum_{j \in [k] \setminus \{y\}} \phi'(v_y - v_j) \right) - \left(\sum_{j \in [k] \setminus \{y\}} q_j \Gamma_j(\mathbf{v}) \phi'(v_j - v_y) \right).$$

The vanishing of the first two partial derivatives $\begin{bmatrix} \frac{\partial C_{\mathbf{q}}}{\partial v_1}(\mathbf{v}) & \frac{\partial C_{\mathbf{q}}}{\partial v_2}(\mathbf{v}) \end{bmatrix} = 0$ can be cast in matrix form equivalently as follows:

$$\begin{bmatrix} q_1 \Gamma_1(\mathbf{v}) \\ q_2 \Gamma_2(\mathbf{v}) \\ q_3 \Gamma_3(\mathbf{v}) \\ \vdots \\ q_k \Gamma_k(\mathbf{v}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j \in [k] \setminus \{1\}} \phi'(v_1 - v_j) & -\phi'(v_1 - v_2) \\ -\phi'(v_2 - v_1) & \sum_{j \in [k] \setminus \{2\}} \phi'(v_2 - v_j) \\ -\phi'(v_3 - v_1) & -\phi'(v_3 - v_2) \\ \vdots & \vdots \\ -\phi'(v_k - v_1) & -\phi'(v_k - v_2) \end{bmatrix} = 0.$$

The above equation is satisfied at $\mathbf{v} = \boldsymbol{\alpha}$, which satisfies $\alpha_1 = \alpha_2$ by assumption.

$$\begin{bmatrix} q_1 \Gamma_1(\boldsymbol{\alpha}) \\ q_2 \Gamma_2(\boldsymbol{\alpha}) \\ q_3 \Gamma_3(\boldsymbol{\alpha}) \\ \vdots \\ q_k \Gamma_k(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j \in [k] \setminus \{1\}} \phi'(\alpha_1 - \alpha_j) & -\phi'(0) \\ -\phi'(0) & \sum_{j \in [k] \setminus \{2\}} \phi'(\alpha_2 - \alpha_j) \\ -\phi'(\alpha_3 - \alpha_1) & -\phi'(\alpha_3 - \alpha_1) \\ \vdots & \vdots \\ -\phi'(\alpha_k - \alpha_1) & -\phi'(\alpha_k - \alpha_1) \end{bmatrix} = 0.$$

Equivalently, we can rearrange the above equation as

$$\begin{bmatrix} q_1\Gamma_1(\boldsymbol{\alpha}) \\ q_2\Gamma_2(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j\in[k]\setminus\{1\}} \phi'(\alpha_1-\alpha_j) & -\phi'(0) \\ -\phi'(0) & \sum_{j\in[k]\setminus\{2\}} \phi'(\alpha_2-\alpha_j) \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} q_3\Gamma_3(\boldsymbol{\alpha}) \\ \vdots \\ q_k\Gamma_k(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \phi'(\alpha_3-\alpha_1) \\ \vdots \\ \phi'(\alpha_k-\alpha_1) \end{bmatrix}}_{=:d} \begin{bmatrix} 1 & 1 \end{bmatrix} = d\mathbf{1}^\top$$

Furthermore, note that

$$\begin{aligned} \sum_{j\in[k]\setminus\{1\}} \phi'(\alpha_1-\alpha_j) &= \phi'(\alpha_1-\alpha_2) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j) \\ &= \phi'(0) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j) \\ &= \phi'(0) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_2-\alpha_j) \\ &= \sum_{j\in[k]\setminus\{2\}} \phi'(\alpha_2-\alpha_j). \end{aligned}$$

Likewise, $\Gamma_1(\boldsymbol{\alpha}) = \gamma'(\phi(0) + \sum_{j\in[k]\setminus\{1,2\}} \phi(v_1-v_j)) = \Gamma_2(\boldsymbol{\alpha})$. Let $a := \phi'(0)$, $b := \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j)$, and $c := \Gamma_1(\boldsymbol{\alpha})$. Since $\gamma'(\cdot) > 0$, we have $c > 0$ and so

$$c\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}^\top \begin{bmatrix} a+b & -a \\ -a & a+b \end{bmatrix} = d\mathbf{1}^\top \implies \begin{bmatrix} a+b & -a \\ -a & a+b \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \frac{d}{c}\mathbf{1}.$$

Note that since $\phi' \le 0$ and $\phi'(0) \neq 0$, we have $a \in (-\infty, 0)$ and $b \in (-\infty, 0]$. First consider the case when $b < 0$. Then $\det\left(\begin{bmatrix} a+b & -a \\ -a & a+b \end{bmatrix}\right) = (a+b)^2 - a^2 > 0$ which

implies that

$$\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \frac{d}{c((a+b)^2 - a^2)} \begin{bmatrix} a+b & a \\ a & a+b \end{bmatrix} \mathbf{1} = \frac{d}{c((a+b)^2 - a^2)} \begin{bmatrix} 2a+b \\ 2a+b \end{bmatrix}.$$

And thus, when $b < 0$, we have $q_1 = q_2$. On the other hand, if $b = 0$, then

$$\begin{bmatrix} a & -a \\ -a & a \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \frac{d}{c}\mathbf{1} \implies a(q_1 - q_2) = a(q_2 - q_1) \implies q_1 - q_2 = 0.$$

Thus, in the case that $b = 0$, we have $q_1 = q_2$ as well. □

Lemma IV.107. *Suppose $\mathcal{L}$ does not satisfy the ISC property. Then there exists a probability vector $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$ and an index $z \in \{2, \dots, k\}$ satisfying 1) $p_{z-1} > p_z$ and 2) $C^*_{\mathbf{p}} = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_z = \max \mathbf{v}\}$.*

Proof. By Definition IV.9, there exists some $\mathbf{q} \in \Delta^k$ and $y \in [k]$ such that $q_y < \max_{j \in [k]} q_j$ and

$$C^*_{\mathbf{q}} = \inf\{C_{\mathbf{q}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_y = \max_{j \in [k]} v_j\}.$$

The above implies that there exists a sequence $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ such that $\lim_t C_{\mathbf{q}}(\mathbf{v}^t) = C^*_{\mathbf{q}}$ and $v^t_y = \max_{j \in [k]} v^t_j$ for all t. Let $\sigma \in \mathtt{Sym}(k)$ be such that $\sigma(\mathbf{q}) \in \Delta^k_{\mathtt{desc}}$. Let $\tilde{y} := \sigma^{-1}(y)$ and $z \in [k]$ be the smallest index such that $q_{\sigma(z)} = q_{\sigma(\tilde{y})}$ (note that $\sigma(\tilde{y}) = y$ by definition). Furthermore, we have that $z > 1$ since $q_{\sigma(1)} = \max \mathbf{q} > q_y = q_{\sigma(z)}$.

Let $\tau \in \mathtt{Sym}(k)$ be the permutation that swaps z and $\tilde{y}$ while leaving all other elements of $[k]$ unchanged. Note that if $z = \tilde{y}$, then τ is the trivial permutation, i.e., the identity map on $[k]$. Define $\mathbf{p} := \tau(\sigma(\mathbf{q}))$, and $\mathbf{w}^t := \tau(\sigma(\mathbf{v}^t))$. Observe that $\mathbf{p} = \tau(\sigma(\mathbf{q})) = \sigma(\mathbf{q})$ and thus $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$ as well. We claim that $p_{z-1} > p_z$. To see this, note that

$$p_{z-1} = [\tau(\sigma(\mathbf{q}))]_{z-1} = [\sigma(\mathbf{q})]_{\tau(z-1)} = [\sigma(\mathbf{q})]_{z-1} = q_{\sigma(z-1)} > q_{\sigma(z)} = q_y$$

and

$$p_z = [\tau(\sigma(\mathbf{q}))]_z = [\sigma(\mathbf{q})]_{\tau(z)} = [\sigma(\mathbf{q})]_{\tilde{y}} = q_{\sigma(\tilde{y})} = q_y.$$

By Lemma IV.28, we have

$$\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = \lim_t C_{\tau(\sigma(\mathbf{q}))}(\tau(\sigma(\mathbf{v}^t))) = \lim_t C_{\mathbf{q}}(\mathbf{v}^t) = C^*_{\mathbf{q}} = C^*_{\tau(\sigma(\mathbf{q}))} = C^*_{\mathbf{p}}.$$

Furthermore, we have $\max \mathbf{v}^t = \max \sigma(\mathbf{v}^t) = \max \mathbf{w}^t$ and so

$$w^t_z = [\mathbf{w}^t]_z = [\tau(\sigma(\mathbf{v}^t))]_z = [\sigma(\mathbf{v}^t)]_{\tau(z)} = v^t_{\sigma(\tau(z))} = v^t_{\sigma(\tilde{y})} = v^t_y = \max \mathbf{v}^t = \max \mathbf{w}^t.$$

In summary, we have an index $z \in [k]$ where $z > 1$ and a probability vector $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$ such that $p_{z-1} > p_z$. Furthermore, we have a sequence $\{\mathbf{w}^t\}_t$ such that $\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = C^*_{\mathbf{p}}$ and $w^t_z = \max \mathbf{w}^t$. This implies the desired condition in the statement of Lemma IV.107. □

Proof. (of Theorem IV.15). Let $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$ and $z \in \{2, \ldots, k\}$ be as in Lemma IV.107, which states that $\mathbf{p}$ and z satisfies the conditions of Proposition IV.104. Next, let $\ell \in [k]$, $\boldsymbol{\alpha} \in \mathbb{R}^\ell$, and $\mathbf{q} \in \Delta^\ell_{\mathtt{desc}}$ be as in Proposition IV.104, which satisfy $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$ and $q_z < q_{z-1} \leq q_1 = \max \mathbf{q}$. Let $\tau \in \mathtt{Sym}(\ell)$ be the permutation which swaps z and 2 leaving all elements in $[\ell] \setminus \{2, z\}$ unchanged. Then

$$C^*_{\tau(\mathbf{q})} = C^*_{\mathbf{q}} = C_{\mathbf{q}}(\boldsymbol{\alpha}) = C_{\tau(\mathbf{q})}(\tau(\boldsymbol{\alpha})).$$

Let $\tilde{\mathbf{q}} := \tau(\mathbf{q})$ and $\tilde{\boldsymbol{\alpha}} := \tau(\boldsymbol{\alpha})$. Then $[\tilde{\boldsymbol{\alpha}}]_1 = [\boldsymbol{\alpha}]_{\tau(1)} = \alpha_1 = 0$ and $[\tilde{\boldsymbol{\alpha}}]_2 = [\boldsymbol{\alpha}]_{\tau(2)} = \alpha_z = 0$. In particular, $\tilde{\alpha}_1 = \tilde{\alpha}_2$. Thus, by Proposition IV.106, we have $\tilde{q}_1 = \tilde{q}_2$. However, $\tilde{q}_1 = [\mathbf{q}]_{\tau(1)} = q_1$ and $\tilde{q}_2 = [\mathbf{q}]_{\tau(2)} = q_z$. Since $q_z < q_1$, we have a contradiction. □

4.9.1 A Gamma-Phi loss that is not ISC

In this section, we show an example of a Gamma Phi loss that satisfies the conditions of [PS16] and yet is not classification-calibrated. The paragraph before Pires et al. [PS16, Section 3.4.2] states that the Gamma-Phi loss is calibrated when γ is strictly increasing and ϕ satisfies the same condition as [Zha04a, Theorem 6], namely that ϕ is non-negative, non-increasing and $\phi'(0) < 0$. However, in the following example, we give a counterexample to the aforementioned statement.

Proposition IV.108. *Let $\mathcal{L}$ be the Gamma-Phi loss as in Example IV.5 where*

$$\gamma(x) = \begin{cases} 1 - (x-1)^2 & : x < 1 \\ 2(x-1)^2 + 1 & : x \geq 1 \end{cases}$$

and $\phi(x) = \exp(-x)$. Then $\mathcal{L}$ is not ISC.

For $r \in (\frac{1}{2}, 1)$, define $\mathbf{p} := [r, 1-r, 0, \ldots, 0] \in \Delta^k_{\texttt{desc}}$. Thus, for a generic $\mathbf{v} \in \mathbb{R}^k$, we have

$$C_{\mathbf{p}}(\mathbf{v}) = r\gamma\left(\sum_{j \in [k]\setminus\{1\}} \phi(v_1 - v_j)\right) + (1-r)\gamma\left(\sum_{j \in [k]\setminus\{2\}} \phi(v_2 - v_j)\right).$$

Consider the set $\texttt{SEQ}$ of all sequences $\{\mathbf{v}^t\}_t$ satisfying Proposition IV.104 all items 1, 2 and 3. For sequence $\{\mathbf{v}^t\}_t \in \texttt{SEQ}$, there exists an $\ell \in [k]$ as in Proposition IV.104 item 2 such that $\lim_t v^t_j = -\infty$ if and only if $j \in [k]$ satisfies $j > \ell$. Below,

$$\text{fix a sequence } \{\mathbf{v}^t\}_t \in \texttt{SEQ} \text{ such that } \ell \text{ is as small as possible.} \tag{4.56}$$

Furthermore, let $\boldsymbol{q} \in \Delta^\ell_{\texttt{desc}}$, $\boldsymbol{\alpha} \in \mathbb{R}^\ell$ be from Proposition IV.104 item 3. Recall that we have $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$ and that $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$. Furthermore, Proposition IV.104 asserts that $v^t_1 = 0$.

Claim. $\ell = 2$ and $\boldsymbol{\alpha} = [0, 0]$.

We first show that $\ell = 2$. To this end, we show that assuming $\ell = 1$ or $\ell \in \{3, \dots, k\}$ both lead to contradictions. First, assume that $\ell = 1$. Then we have $\lim_t v_2^t = \cdots = \lim_t v_k^t = -\infty$. Since γ is increasing and $\phi \geq 0$, we have for any $\mathbf{v} \in \mathbb{R}^k$ that

$$C_{\mathbf{p}}(\mathbf{v}) \geq r\gamma \left(\sum_{j \in [k] \setminus \{1\}} \phi(v_1 - v_j) \right) + (1 - r)\gamma \left(\phi(v_2 - v_1) \right).$$

Since $v_1^t = 0$ for all t, we have

$$\begin{aligned}
\lim_t C_{\mathbf{p}}(\mathbf{v}^t) &\geq \lim_t r\gamma \left(\sum_{j \in [k] \setminus \{1\}} \phi(-v_j) \right) + (1 - r)\gamma \left(\phi(v_2) \right) \\
&= r\gamma \left((k - 1)\phi(+\infty) \right) + (1 - r)\gamma \left(\phi(-\infty) \right) \\
&= r\gamma(0) + (1 - r)\gamma(+\infty) \\
&\geq +\infty. \qquad \because \gamma(+\infty) = +\infty \text{ (Definition IV.13)}
\end{aligned}$$

This is a contradiction since $C_{\mathbf{p}}(\mathbb{0}) = \gamma \left((k - 1)\phi(0) \right) < +\infty$.

Next, we assume that $\ell \in \{3, \dots, k\}$ and derive a contradiction. For a generic $\mathbf{w} \in \mathbb{R}^\ell$, recall that

$$C_{\mathbf{q}}(\mathbf{w}) = r\mathcal{L}_1(\mathbf{w}) + (1 - r)\mathcal{L}_2(\mathbf{w})$$

where for $y \in \{1, 2\}$, we have

$$\mathcal{L}_y(\mathbf{w}) = \gamma \left(\sum_{j \in [\ell] \setminus \{y\}} \phi(w_y - w_j) \right).$$

Let $\epsilon > 0$ and define $\boldsymbol{\beta} \in \mathbb{R}^\ell$ by

$$\beta_j = \begin{cases} \alpha_j & : j \neq 3 \\ \alpha_\ell - \epsilon & : j = \ell. \end{cases}$$

For $y \in \{1, 2\}$, since $\beta_\ell < \alpha_\ell$ and $\beta_j = \alpha_j$ for $j \in [k] \setminus \{\ell\}$, we have

$$\begin{cases} \beta_y - \beta_j = \alpha_y - \alpha_j & : j \neq \ell \\ \beta_y - \beta_\ell > \alpha_y - \alpha_\ell & : j = \ell \end{cases}$$

$$\implies \begin{cases} \phi(\beta_y - \beta_j) = \alpha_y - \alpha_j & : j \neq \ell \\ \phi(\beta_y - \beta_\ell) \leq \phi(\alpha_y - \alpha_\ell) & : j = \ell \end{cases}$$

$$\implies \mathcal{L}_y(\boldsymbol{\beta}) = \gamma \left(\sum_{j \in [\ell] \setminus \{y\}} \phi(\beta_y - \beta_j) \right) \leq \gamma \left(\sum_{j \in [\ell] \setminus \{y\}} \phi(\alpha_y - \alpha) \right) = \mathcal{L}_y(\boldsymbol{\alpha}).$$

Thus, $C_{\mathbf{q}}(\boldsymbol{\alpha}) \geq C_{\mathbf{q}}(\boldsymbol{\beta})$ and so $C_{\mathbf{q}}(\boldsymbol{\alpha}) \geq \lim_{\epsilon \to \infty} C_{\mathbf{q}}(\boldsymbol{\beta})$ as well. By Lemma IV.105 and that $p_y = 0$ for $y \geq 2$, we have $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{q}}(\boldsymbol{\alpha})$. Now, define $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ by

$$\tilde{v}_j^t := \begin{cases} v_j^t & : j \neq \ell \\ -t & : j = \ell. \end{cases}$$

By construction we have $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_{\epsilon \to \infty} C_{\mathbf{q}}(\boldsymbol{\beta})$ and $\{\tilde{\mathbf{v}}^t\}_t \in \texttt{SEQ}$. Furthermore, since $\lim_t \tilde{v}_\ell^t = -\infty$, we have a contradiction of the minimality of ℓ (Equation 4.56).

Below, we can assume that $\ell = 2$, where we have $\mathbf{q} = [r, 1 - r] \in \Delta^2_{\texttt{desc}}$ and so

$$C_{\mathbf{q}}(\boldsymbol{\alpha}) = r\gamma\left(\phi(-\alpha_2)\right) + (1 - r)\gamma\left(\phi(\alpha_2)\right) = \inf_{\mathbf{w} \in \mathbb{R}^2} C_{\mathbf{q}}(\mathbf{w}).$$

Consider the function

$$F(x) = r\gamma\left(\phi(x)\right) + (1-r)\gamma\left(\phi(-x)\right).$$

Then we have $C_{\mathbf{q}}(\boldsymbol{\alpha}) = \inf_x F(x)$. Now, let us compute the derivative of $F(x)$. Using the chain rule, we have

$$\frac{dF}{dx}(x) = r\frac{d\gamma}{dx}(\phi(x))\frac{d\phi}{dx}(x) - (1-r)\frac{d\gamma}{dx}(\phi(-x))\frac{d\phi}{dx}(-x).$$

Now, $\frac{d\phi}{dx}(x) = -\exp(-x)$ and

$$\frac{d\gamma}{dx}(x) = \begin{cases} -2(x-1) & : x < 1 \\ 4(x-1) & : x \geq 1. \end{cases}$$

If $x > 0$, then $\phi(x) < 1$ and $\phi(-x) > 1$. Thus, when $x > 0$, we have

$$\begin{aligned} \frac{dF}{dx}(x) &= r(-2(\exp(-x)-1))(-\exp(-x)) - (1-r)(4(\exp(x)-1))(-\exp(x)) \\ &= 2r(\exp(-x)-1)\exp(-x) + 4(1-r)(\exp(x)-1)\exp(x) \\ &=: G_+(x). \end{aligned}$$

If $x \leq 0$, then $\phi(x) \geq 1$ and $\phi(-x) \leq 1$. Thus, when $x \leq 0$, we have

$$\begin{aligned} \frac{dF}{dx}(x) &= r(4(\exp(-x)-1))(-\exp(-x)) - (1-r)(-2(\exp(x)-1))(-\exp(x)) \\ &= -4r(\exp(-x)-1)\exp(-x) - 2(1-r)(\exp(x)-1)\exp(x) \\ &=: G_-(x). \end{aligned}$$

Thus, by definition, we have

$$\frac{dF}{dx}(x) = \begin{cases} G_+(x) & : x > 0 \\ G_-(x) & : x < 0 \\ 0 & : x = 0. \end{cases}$$

Lemma IV.109. *If $r \in [\frac{1}{3}, \frac{2}{3}]$, then $\frac{dF}{dx}(x)$ vanishes only at $x = 0$.*

We now consider the zeros of both $G_+(x)$ and $G_-(x)$, i.e., $x \in \mathbb{R}$ where the functions vanish. Clearly, both functions vanish at $x = 0$. For $x \neq 0$, we compute

$$\begin{aligned} 0 = G_+(x) &= 2r(\exp(-x) - 1)\exp(-x) + 4(1-r)(\exp(x) - 1)\exp(x) \\ \iff \frac{r}{2(1-r)} &= -\frac{\exp(x)(\exp(x) - 1)}{\exp(-x)(\exp(-x) - 1)}. \end{aligned}$$

Simplifying the right hand side, we have

$$\begin{aligned} -\frac{\exp(x)(\exp(x) - 1)}{\exp(-x)(\exp(-x) - 1)} &= -\exp(2x)\frac{\exp(x) - 1}{\exp(-x) - 1} \\ &= -\exp(2x)\exp(x)\frac{1 - \exp(-x)}{\exp(-x) - 1} \\ &= \exp(3x). \end{aligned}$$

Thus,

$$0 = G_+(x) \iff \frac{1}{3}\ln\left(\frac{r}{2(1-r)}\right) = x.$$

Similarly, for the zeroes of $G_(x)$ we have

$$0 = G_-(x) \iff \frac{1}{3}\ln\left(\frac{2r}{(1-r)}\right) = x.$$

Thus, $G_+(x)$ has a zero on $x > 0$ if and only if

$$\frac{1}{3}\ln\left(\frac{r}{2(1-r)}\right) > 0 \iff \frac{r}{2(1-r)} > 1 \iff r > 2/3.$$

Similarly, $G_-(x)$ has a zero on $x < 0$ if and only if

$$\frac{1}{3}\ln\left(\frac{2r}{1-r}\right) < 0 \iff \frac{2r}{1-r} < 1 \iff r < 1/3.$$

Taken together, we see that if $r \in [\frac{1}{3}, \frac{2}{3}]$, then $\frac{dF}{dx}(x)$ only vanishes at $x = 0$. Thus, we conclude that when $r \in (\frac{1}{2}, \frac{2}{3}]$, we must have $\boldsymbol{\alpha} = [0, 0] \in \mathbb{R}^2$. Now, define another sequence $\{\mathbf{w}^t\}_{t=1}^{\infty} \subseteq \mathbb{R}^k$ where

$$w_j^t := \begin{cases} 0 & : j = 1 \\ 1/t & : j = 2 \\ -t & : j \in \{3, \dots, k\}. \end{cases}$$

Then we have $\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = C_{\mathbf{p}}^*$ and $\arg\max_{j \in [k]} w_j^t = 2$ for all t. Thus, we have demonstrated an example of $\mathbf{p}$ and $y \in [k]$ where

$$C_{\mathbf{p}}^* = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_y = \max \mathbf{v}\},$$

namely when $\mathbf{p} = [r, 1-r, 0, \dots, 0] \in \Delta_{\texttt{desc}}^k$ and $y = 2$. This shows that $\mathcal{L}$ is not ISC (Definition IV.9).

4.10 Discussion

We proved sufficient conditions for two families of losses: the Gamma-Phi and the Fenchel-Young losses. We also showed that previous attempts to prove sufficient condition for the Gamma-Phi loss did not account for behavior at infinity. As such,

we were able to construct a counterexample of a non-classification-calibrated loss that would previously have been deemed to be calibrated. Thus, our work augments the repertoire of the existing sufficient conditions.

Moving forward, there are many important open questions. Perhaps the most important question is whether classification-calibration results translates into the "real world" where classifiers are from a restricted candidate set $\mathcal{H}$ strictly smaller than the set of all Borel functions $\mathcal{F}$ from Theorem IV.12. Progress in this area have already been made by Duchi et al. [DKR18] where $\mathcal{H}$ are certain quantized functions, and by Zhang et al. [ZA20] where $\mathcal{H}$ are linear functions.

Another interesting future direction would be bounding the regret functions of PERM losses as Nowak-Vila et al. [NBR19] and Blondel [Blo19] have done for Fenchel-Young losses of strongly convex negentropy. Frongillo et al. [FW21] showed the polyhedral losses (such as the hinge loss) are, under certain conditions, optimal from the perspective of regret functions. Exploring the relationship between the quality of the regret function and the empirical performance, e.g., comparing polyhedral losses with Fenchel-Young losses, will be interesting.

Much progress have been made on implicit regularization of optimization algorithms [Sou+18; Ji+20] for binary classification. However, to the best of our knowledge, the only known result regarding implicit regularization in multiclass classification is for the cross entropy [Sou+18]. In the binary case, Ji et al. [Ji+20] showed that implicit regularization depends on the loss function in a subtle way. Therefore, we believe that understanding implicit regularization for multiclass losses besides the cross-entropy is an important future direction.

4.11 Mathematical Backgrounds

4.11.1 Non-singular M-matrix

We recall some definitions from linear algebra.

Definition IV.110. Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ be a matrix. We say that $\mathbf{A}$ is a

1. *Z-matrix* if $a_{ij} \leq 0$ whenever $i \neq j$.
2. *M-matrix* if $\mathbf{A}$ is a Z-matrix and all eigenvalues of $\mathbf{A}$ have nonnegative real parts.
3. *strictly diagonally dominant matrix* if $|a_{ii}| > \sum_{j\in[n]:j\neq i} |a_{ij}|$ for all $i \in [n]$.
4. *monotone matrix* if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{Ax} \geq 0$ implies $\mathbf{x} \geq 0$. If, in addition, $\mathbf{Ax} > 0$ implies $\mathbf{x} > 0$, then $\mathbf{A}$ is said to be *strictly monotone*.

The following result is known as the Levy–Desplanques theorem and the Gershgorin circle theorem.

Theorem IV.111. *Let* $\mathbf{A}$ *be a strictly diagonally dominant matrix. Then* $\mathbf{A}$ *is non-singular and all eigenvalues of* $\mathbf{A}$ *have nonnegative real parts.*

The above result immediately implies the following:

Corollary IV.112. *If* $\mathbf{A}$ *is a strictly diagonally dominant Z-matrix, then* $\mathbf{A}$ *is a non-singular M-matrix.*

Non-singular M-matrix has many equivalent characterizations. The one relevant to us is the following:

Theorem IV.113 ([Ple77])**.** *Let* $\mathbf{A}$ *be a Z-matrix. Then* $\mathbf{A}$ *is a non-singular M-matrix if and only* $\mathbf{A}$ *is a monotone matrix.*

Lemma IV.114. *Let* $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ *be a non-singular M-matrix. If the diagonals of* $\mathbf{A}$ *are positive, then* $\mathbf{A}$ *is strictly monotone.*

Proof. From Theorem IV.113, we have that $\mathbf{A}$ is monotone. Thus, $\mathbf{Ax} > 0$ implies $\mathbf{x} \geq 0$. We only have to check additionally that $\mathbf{x} > 0$. Since $\mathbf{A}$ is a Z-matrix, the off-diagonals are non-positive, i.e., $a_{ij} \leq 0$ for $i \neq j$. Let $i \in [n]$. We need to check that $x_i > 0$. To this end, note that

$$0 < [\mathbf{Ax}]_i = \sum_{j=1}^{n} a_{ij}x_j = a_{ii}x_i + \underbrace{\sum_{j \neq i} a_{ij}x_j}_{\leq 0} \leq a_{ii}x_i.$$

Since $a_{ii} > 0$, we get $x_i > 0$. □

4.11.2 Vector calculus

Given a differentiable function $f = (f_1, \ldots, f_m)' : \mathbb{R}^n \to \mathbb{R}^m$, the *Jacobian* of f at $\mathbf{x} \in \mathbb{R}^n$, denoted $\nabla_f(\mathbf{x})$ is the $m \times n$ matrix whose (i, j)-th entry is

$$[\nabla_f(\mathbf{x})]_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$$

where $i \in [m]$ and $j \in [n]$. Note that we can write the above as

$$\nabla_f(\mathbf{x}) = \begin{bmatrix} \nabla_{f_1}(\mathbf{x}) \\ \vdots \\ \nabla_{f_m}(\mathbf{x}) \end{bmatrix}.$$

If f is a linear map, i.e., $f(\mathbf{x}) = \mathbf{Ax}$ for some a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\nabla_f(\mathbf{x}) = \mathbf{A}$.

Theorem IV.115 (Chain rule)**.** *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ are differentiable, then $\nabla_{g \circ f}(\mathbf{x}) = \nabla_g(f(\mathbf{x})) \cdot \nabla_f(\mathbf{x})$.*

The follow is taken from Munkres [Mun18, Theorem 8.2].

Theorem IV.116 (Inverse function theorem)**.** *Let U be open in $\mathbb{R}^n$, $f : U \to \mathbb{R}^n$ be r-times continuously differentiable and $V = f(U)$. If f is one-to-one on U and if*

$\nabla_f(\mathbf{x})$ *is non-singular for all* $\mathbf{x} \in U$*, then* V *is open in* $\mathbb{R}^n$ *and the inverse function* $g : V \to U$ *is* r*-times continuously differentiable.*

The following is an immediate consequence of Theorems IV.116 and IV.115:

Corollary IV.117. *In the setting of Theorem IV.116, we have* $\nabla_{f^{-1}}(f(\mathbf{x})) = \nabla_f(\mathbf{x})^{-1}$.

CHAPTER V

VC Dimension of Partially Quantized Neural Networks in the Overparametrized Regime

Vapnik-Chervonenkis (VC) theory has so far been unable to explain the small generalization error of overparametrized neural networks. Indeed, existing applications of VC theory to large networks obtain upper bounds on VC dimension that are proportional to the number of weights, and for a large class of networks, these upper bound are known to be tight. In this work, we focus on a subclass of partially quantized networks that we refer to as *hyperplane arrangement neural networks* (HANNs). Using a sample compression analysis, we show that HANNs can have VC dimension significantly smaller than the number of weights, while being highly expressive. In particular, empirical risk minimization over HANNs in the overparametrized regime achieves the minimax rate for classification with Lipschitz posterior class probability. We further demonstrate the expressivity of HANNs empirically. On a panel of 121 UCI datasets, overparametrized HANNs match the performance of state-of-the-art full-precision models.

5.1 Introduction

Neural networks have become an indispensable tool for machine learning practitioners, owing to their impressive performance especially in vision and natural language processing [GBC16]. In practice, neural networks are often applied in the *overparametrized* regime and are capable of fitting even random labels [Zha+21]. Evidently, these overparametrized models perform well on real world data despite their ability to grossly overfit, a phenomenon that has been dubbed "the generalization puzzle" [NK19].

Toward solving this puzzle, several research directions have flourished and offer potential explanations, including implicit regularization [CB20], interpolation [CL21], and benign overfitting [Bar+20]. So far, VC theory has not been able to explain the puzzle, because existing bounds on the VC dimensions of neural networks are on the order of the number of weights [Maa94; Bar+19]. It remains unknown whether there exist neural network architectures capable of modeling rich set of classfiers with low VC dimension.

The focus of this work is on a class of neural networks with threshold activation that we refer to as *hyperplane arrangement neural networks* (HANNs). Using the theory of sample compression schemes [LW86], we show that HANNs can have VC dimension that is significantly smaller than the number of parameters. Furthermore, we apply this result to show that HANNs have high expressivity by proving that HANN classifiers achieve minimax-optimality when the data has Lipschitz posterior class probability in an overparametrized setting.

We benchmark the empirical performance of HANNs on a panel of 121 UCI datasets, following several recent neural network and neural tangent kernel works [Kla+17; Wu+18; Aro+19; Sha+20]. In particular, [Kla+17] showed that, using a properly chosen activation, overparametrized neural networks perform competitively compared to classical shallow methods on this panel of datasets. Our experiments

show that HANNs, a partially-quantized model, match the classification accuracy of the self-normalizing neural network [Kla+17] and the dendritic neural network [Wu+18], both of which are full-precision models.

5.2 Notations

The set of real numbers is denoted $\mathbb{R}$. The unit interval is denoted $[0, 1]$. For an integer $k \geq 1$, let $[k] = \{1, \ldots, k\}$. We use $\mathcal{X}$ to denote the feature space, which in this work will either be $\mathbb{R}^d$ or $[0, 1]^d$ where $d \geq 1$ is the ambient dimension/number of features.

Denote by $\mathbb{I}\{\texttt{input}\}$ the *indicator* function which returns 1 if `input` is true and 0 otherwise. The *sign* function is given by $\texttt{sgn}(t) = \mathbb{I}\{t \geq 0\} - \mathbb{I}\{t < 0\}$. For vector inputs, `sgn` applies entry-wise.

The set of labels for binary classification is denoted $\mathbb{B} := \{\pm 1\}$. Joint distributions on $\mathcal{X} \times \mathbb{B}$ are denoted by P, where $X, Y \sim P$ denotes a random instance-label pair distributed according to P. Let $f : \mathcal{X} \to \mathbb{B}$ be a binary classifier. The *risk* with respect to P is denoted by $R_P(f) := P(f(X) \neq Y)$. For an integer $n \geq 1$, the *empirical risk* is the random variable $\hat{R}_{P,n}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{f(X_i) \neq Y_i\}$, where $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ are i.i.d. The *Bayes risk* $\inf_{f:\mathcal{X}\to\mathbb{B}} R_P(f)$ with respect to P is denoted by R_P^*.

Let $f, g : \{1, 2, \ldots\} \to \mathbb{R}_{\geq 0}$ be nonnegative functions on the natural numbers. We write $f \asymp g$ if there exists $\alpha, \beta > 0$ such that for all $n = 1, 2, \ldots$ we have $\alpha g(n) \leq f(n) \leq \beta g(n)$.

5.3 Hyperplane arrangement neural networks

A hyperplane H in $\mathbb{R}^d$ is specified by its normal vector $w \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$. The mapping $x \mapsto \texttt{sgn}(w^\top x + b)$ indicates the side of H that x lies on, and hence

induces a partition of $\mathbb{R}^d$ into two halfspaces. A set of $k \geq 1$ hyperplanes is referred to as a *k-hyperplane arrangement*, and specified by a matrix of normal vectors and a vector of offsets:

$$\mathbf{W} = [w_1 \cdots w_k] \in \mathbb{R}^{d \times k} \quad \text{and} \quad b = [b_1, \ldots, b_k]^\top.$$

Let $q_{\mathbf{W},b}(x) := \mathtt{sgn}(\mathbf{W}^\top x + b)$ for all $x \in \mathbb{R}^d$. The vector $q_{\mathbf{W},b}(x) \in \mathbb{B}^k$ is called a *sign vector* and the set of all realizable sign vectors is denoted $\mathfrak{S}_{\mathbf{W},b} := \{q_{\mathbf{W},b}(x) : x \in \mathbb{R}^d\}$. Each sign vector $s \in \mathfrak{S}_{\mathbf{W},b}$ uniquely defines a set $\{x \in \mathbb{R}^d : q_{\mathbf{W},b}(x) = s\}$ known as a *cell* of the hyperplane arrangement. The set of all cells forms a partition of $\mathbb{R}^d$. For an example, see fig. 5.1-left.

A classical result in the theory of hyperplane arrangement due to [Buc43] gives the following tight upper bound on the number of distinct sign patterns/cells:

$$|\mathfrak{S}_{\mathbf{W},b}| \leq \binom{k}{\leq d} := \begin{cases} 2^k & : k < d, \\ \binom{k}{0} + \binom{k}{1} + \cdots + \binom{k}{d} & : k \geq d. \end{cases} \tag{5.1}$$

See [Fuk15] Theorem 10.1 for a simple proof. A *hyperplane arrangement classifier* assigns a binary label $y \in \mathbb{B}$ to a point $x \in \mathbb{R}^d$ solely based on the sign vector $q_{\mathbf{W},b}(x)$.

Definition V.1. Let $\mathbb{B}^{\mathcal{X}}$ be the set of all functions from $\mathcal{X}$ to $\mathbb{B}$. A *concept class* $\mathcal{C}$ over $\mathcal{X}$ is a subset of $\mathbb{B}^{\mathcal{X}}$. Fix r, k positive integers, $r \leq \min\{d, k\}$. Let $\mathtt{Bool}_k$ be the set of all Boolean functions $\mathbb{B}^k \to \mathbb{B}$. The *hyperplane arrangement classifier* class is the concept class, denoted $\mathtt{HAC}(d, r, k)$, over $\mathbb{R}^d$ defined by

$$\mathtt{HAC}(d, r, k) = \{h \circ q_{\mathbf{W},b} : h \in \mathtt{Bool}_k,\ q_{\mathbf{W},b}(x) := \mathtt{sgn}(\mathbf{W}^\top x + b), \\ \mathbf{W} \in \mathbb{R}^{d \times k},\ \mathrm{rank}(\mathbf{W}) \leq r,\ b \in \mathbb{R}^k\}.$$

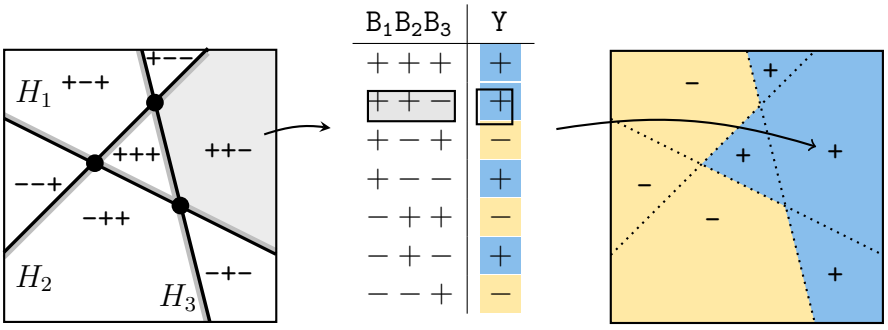

Figure 5.1: Left: An arrangement of 3 hyperplanes $\{H_1, H_2, H_3\}$ in $\mathbb{R}^2$. There are 7 sign patterns. Middle: An example of a lookup table (see Remark V.2). Right: the resulting classifier.

See fig. 5.2 for a graphical representation of $\texttt{HAC}(d, r, k)$. When the set of Boolean functions is realized by a neural network, we refer to the resulting classifier as a *hyperplane arrangement neural network* (HANN).

Remark V.2. Consider a fixed hyperplane arrangement $\mathbf{W}$, b and Boolean function $h \in \texttt{Bool}_k$. When performing prediction with the classifer $h \circ q_{\mathbf{W},b}$, the feature vector x is mapped to a sign vector to which h is applied. Thus, we do not need to know how h behaves outside of $\mathfrak{S}_{\mathbf{W},b}$. The restriction of h to $\mathfrak{S}_{\mathbf{W},b}$ is a *partially defined Boolean function* or a *lookup table*.

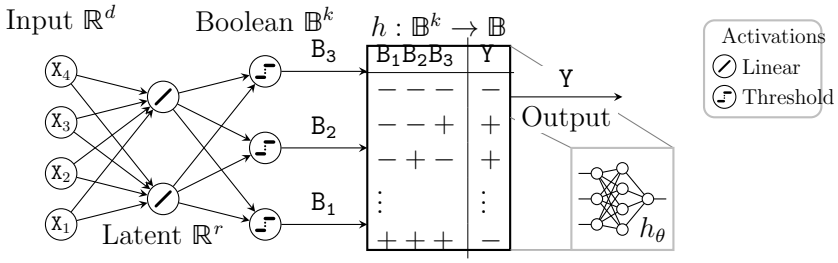

Figure 5.2: The $\texttt{HAC}(d, r, k)$ concept class as a neural network where $d = 4$, $r = 2$ and $k = 3$. The Boolean function h is realized as a neural network h_θ.

Remark V.3. The hidden layer of width r in fig. 5.2 allows the user to impose the restriction that the hyperplane arrangement classifier depends only on r relevant features, which can be either learned or defined by data preprocessing. When $r =$

d, no restriction is imposed. In this case, the input layer is directly connected to the Boolean layer. This is consistent with Definition V.1 where the rank constraint $\text{rank}(\mathbf{W}) \leq r$ becomes trivial.

Our next goal is to upper bound the VC dimension of $\texttt{HAC}(d, r, k)$.

Definition V.4 (VC-dimension)**.** Let $\mathcal{C} \subseteq \mathbb{B}^{\mathcal{X}}$ be a concept class over $\mathcal{X}$. A set $S := \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ is *shattered* by $\mathcal{C}$ if for all sequences $(y_1, \ldots, y_n) \in \mathbb{B}^n$, there exists $f \in \mathcal{C}$ such that $f(x_i) = y_i$ for all $i \in [n]$. The *VC-dimension* of $\mathcal{C}$ is defined as

$$\texttt{VC}(\mathcal{C}) = \sup\{|S| : S \subseteq \mathcal{X},\, S \text{ is shattered by } \mathcal{C}\}.$$

The VC-dimension has many far-reaching consequences in learning theory and, in particular, classification. One of these consequences is a sufficient (in fact also necessary) condition for *uniform convergence* in the sense of the following well-known theorem. See [SB14] Theorem 6.8.

Theorem V.5. *Let $\mathcal{C}$ be a concept class over $\mathcal{X}$. There exists a constant $C > 0$ such that for all joint distributions P on $\mathcal{X} \times \mathbb{B}$ and all $f \in \mathcal{C}$, we have $|\hat{R}_{P,n}(f) - R_P(f)| \leq C\sqrt{(\texttt{VC}(\mathcal{C}) + \log(1/\delta))/n}$ with probability at least $1 - \delta$ with respect to the draw of $(X_1, Y_1), \ldots, (X_n, Y_n)$.*

The above VC bound is useless in the overparametrized setting if $\texttt{VC}(\mathcal{C}) = \Theta(\#\text{ of weights}) \gg n$. We now present our main result: an upper bound on the VC dimension of $\texttt{HAC}(d, r, k)$.

Theorem V.6. *Let $d, r, k \geq 1$ be integers and $\texttt{HAC}(d, r, k)$ be defined as in Definition V.1. Then*

$$\texttt{VC}(\texttt{HAC}(d, r, k)) \leq 8 \cdot \left(k(d+1) + k(d+1)(1 + \lceil \log_2 k \rceil) + \binom{k}{\leq r} \right).$$

In the next section, we will prove this result using a sample compression scheme. Before proceeding, we comment on the significance of the result.

Remark V.7. Since $\binom{k}{\leq r} = O(k^r)$, we have $\texttt{VC}(\texttt{HAC}(d, r, k)) = O(k^r + dk \log k)$ which only involves the input dimension d and the width of the first two hidden layers r and k. For constant d and $r \geq 2$, this reduces to $\texttt{VC}(\texttt{HAC}(d, r, k)) = O(k^r)$. In particular, the number of weights used by an architecture to implement the Boolean function h does not affect the VC dimension at all and can be even infinitely wide.

For instance, [MB17] Lemma 2.1 states that a 1-hidden layer neural network with ReLU activation can model any k-input Boolean function if the hidden layer has width $\geq 2^k$. Note that this network uses $\geq k2^k$ weights, and $k2^k \gg k^r$ for fixed r and k large.

[BV19] study implementation of Boolean functions using threshold networks. A consequence of their Theorem 9.3 is that a 2-hidden layer network with widths $\geq c2^{k/2}/\sqrt{k}$ can implement all k input Boolean functions, where c is a constant not depending on k. This requires $\geq c^2 2^k/k$ weights which again is exponentially larger than k^r. Furthermore, this lower bound on the weights is also necessary as $k \to \infty$.

5.4 A sample compression scheme

In this section, we will construct a sample compression scheme for $\texttt{HAC}(d, r, k)$. As alluded to in the Related Work section, the size of a sample compression scheme upper bounds the VC-dimension of a concept class, which will be applied to prove theorem V.6. We first recall the definition of sample compression schemes with side information introduced in [LW86].

Definition V.8. Let $\mathcal{C}$ be a concept class. A length n sequence $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{B}\}_{i \in [n]}$ is $\mathcal{C}$-*labelled* if there exists $f \in \mathcal{C}$ such that $f(x_i) = y_i$ for all $i \in [n]$. Denote by $L_{\mathcal{C}}(n)$ the set of $\mathcal{C}$-labelled sequences of length at most n. Denote by $L_{\mathcal{C}}(\infty)$ the set of all

$\mathcal{C}$-labelled sequences of finite length. The concept class $\mathcal{C}$ over $\mathcal{X}$ has an *m-sample compression scheme with s-bits of side information* if there exists a pair of maps (ρ, κ) where

$$\kappa : L_{\mathcal{C}}(\infty) \to L_{\mathcal{C}}(m) \times \mathbb{B}^s, \quad \rho : L_{\mathcal{C}}(m) \times \mathbb{B}^s \to \mathbb{B}^{\mathcal{X}}$$

such that for all $\mathcal{C}$-labelled sequences $S := \{(x_i, y_i)\}_{i \in [n]}$, we have $\rho(\kappa(S))(x_i) = y_i$ for all $i \in [n]$. The *size* of the sample compression scheme is $\texttt{size}(\rho, \kappa) := m + s$.

Intuitively, κ and ρ can be thought of as the *compression* and the *reconstruction* maps, respectively. The compression map κ keeps m elements from the training set and s bits of additional information, which ρ uses to reconstruct a classifier that correctly labels the uncompressed training set.

The main result of this section is:

Theorem V.9. $\texttt{HAC}(d, r, k)$ *has a sample compression scheme (ρ, κ) of size*

$$\texttt{size}(\rho, \kappa) = k(d+1) + k(d+1)(1 + \lceil \log_2 k \rceil) + \binom{k}{\leq r}.$$

Both the hyperplane arrangement $(\mathbf{W}, b)$ and the Boolean function h contribute to the number of parameters/weights, which $\gg \texttt{size}(\rho, \kappa)$ for h in the examples of Remark V.7. The rest of this section will work toward the proof of theorem V.9. The following result states that a $\mathcal{C}$-labelled sequence can be labelled by a hyperplane arrangement classifier of a special form.

Proposition V.10. *Let $\{(x_i, y_i)\}_{i \in [n]}$ be $\texttt{HAC}(d, r, k)$-labelled. Then there exist $\mathbf{V} = [v_1 \cdots v_k] \in \mathbb{R}^{d \times k}, c \in \mathbb{R}^k$ and $h \in \texttt{Bool}_k$ such that for all $i \in [n]$, we have 1) $y_i = h(\texttt{sgn}(\mathbf{V}^\top x_i + c))$, 2) $\operatorname{rank}(\mathbf{V}) \leq r$ and 3) $|v_j^\top x_i + c_j| \geq 1$ for all $i \in [n], j \in [k]$.*

The proof, given in Section 5.8.1, is similar to showing the existence of a max-margin separating hyperplane for a linearly separable dataset.

Definition V.11. Let I be a finite set and let $a_i \in \mathbb{R}^n$ for each $i \in I$. Let $A = \{a_i\}_{i\in I}$. A *conical combination* of A is a linear combination $\sum_{i\in I} \lambda_i a_i$ where the weights $\lambda_i \in \mathbb{R}_{\geq 0}$ are nonnegative. The *conical hull* of A, denoted $\texttt{coni}(A)$, is the set of all conical combinations of A, i.e., $\texttt{coni}(\{a_i\}_{i\in I}) := \left\{\sum_{i\in I} \lambda_i a_i : \lambda_i \in \mathbb{R}_{\geq 0},\, \forall i \in I\right\}$.

The result below follows easily from the Carathédory's theorem for the conical hull [LP09]. For the sake of completeness, we included the proof in Section 5.8.2.

Proposition V.12. *Let $a_1, \ldots, a_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$. For each subset $I \subseteq [m]$, define $\mathcal{P}_I := \{x \in \mathbb{R}^n : a_i^\top x \leq b_i \, \forall i \in I\}$. Suppose that $\mathcal{P}_{[m]}$ is nonempty. Then 1) $\min_{x\in\mathcal{P}_I} \frac{1}{2}\|x\|^2$ has a unique minimizer, denoted by x_I^* below, and 2) there exists a subset $J \subseteq [m]$ such that $|J| = \min\{m, n\}$ and for all $I \subseteq [m]$ with $J \subseteq I$, we have $x_{[m]}^* = x_I^*$.*

Proof of theorem V.9. Let (x_i, y_i) be $\texttt{HAC}(d, r, k)$-realizable, and $\mathbf{V}, c$ and h be as in Proposition V.10. For each $i \in [n]$, define the Boolean vectors $s_i := \texttt{sgn}(\mathbf{V}^\top x_i + c) \in \{\pm 1\}^k$ and $s_{ij} = \texttt{sgn}(v_j^\top x_i + c_j)$ denote the j-th entry of s_i. Note that $s_{ij}(v_j^\top x_i + c_j) = |v_j^\top x_i + c_j| \geq 1$.

We first outline the steps of the proof:

1. Using a subset of the samples $\{(x_{i_\ell}, y_{i_\ell}) : \ell \in [d(k+1)]\}$ with additional $k(d+1)(1 + \lceil \log_2 k \rceil)$ bits of side information $\{(s_{i_\ell j_\ell}, j_\ell) : \ell \in [d(k+1)]\}$, we can reconstruct $\overline{\mathbf{W}}, \bar{b}$ such that $\texttt{sgn}(\overline{\mathbf{W}}^\top x_i + \bar{b}) = s_i$ for all $i \in [n]$.

2. Using an additional subset of samples $\{(x_{\iota_\ell}, y_{\iota_\ell}) : \ell = 1, \ldots, \binom{k}{\leq r}\}$ in conjunction with the $\overline{\mathbf{W}}, \bar{b}$ reconstructed in the previous step, we can find $g \in \texttt{Bool}_k$ such that $g(s_i) = h(s_i)$ for all i.

Now, consider the set

$$\mathcal{P} := \left\{(\mathbf{W}, b) \in \mathbb{R}^{d\times k} \times \mathbb{R}^k : s_{ij}(w_j^\top x_i + b_j) \geq 1,\, \forall i \in [n],\, j \in [k]\right\}.$$

Note that $\mathcal{P}$ is a convex polyhedron in $(d+1)k$-dimensional space. Let $(\overline{\mathbf{W}}, \bar{b})$ be the minimum norm element of $\mathcal{P}$. Note that $\mathtt{sgn}(\overline{\mathbf{W}}^\top x_i + \bar{b}) = \mathtt{sgn}(\mathbf{V}^\top x_i + c) = s_i$ by construction.

By Proposition V.12, there exists a set of tuples

$$\{(i_\ell, j_\ell)\}_{\ell=1,\ldots,(d+1)k}, \text{ where } (i_\ell, j_\ell) \in [n] \times [k]$$

such that $\overline{\mathbf{W}}, \bar{b}$ is also the minimum norm element of

$$\mathcal{P}' := \left\{(\mathbf{W}, b) \in \mathbb{R}^{d\times k} \times \mathbb{R}^k : s_{i_\ell j_\ell}(w_{j_\ell}^\top x_{i_\ell} + b_{j_\ell}) \geq 1,\, \ell = 1, \ldots, d(k+1)\right\}.$$

To encode the defining equations of $\mathcal{P}'$, we need to store

$$\text{samples } \{(x_{i_\ell}, y_{i_\ell})\}_{\ell=1}^{d(k+1)} \text{ and side information } \{(s_{i_\ell j_\ell}, j_\ell)\}_{\ell=1}^{d(k+1)}. \tag{5.2}$$

Note that each $s_{i_\ell j_\ell}$ requires 1 bit while each $j_\ell \in [k]$ requires $\lceil \log_2 k \rceil$ bits. In total, encoding $\mathcal{P}'$ requires storing $d(k+1)$ samples and $d(k+1)(1 + \lceil \log_2 k \rceil)$ of bits.

To reconstruct $g \in \mathtt{Bool}_k$ that agrees with h on all the samples, it suffices to know h when restricted to $\{s_i\}_{i=1}^n$. Since $\{s_i\}_{i=1}^n$ is a subset of $\mathfrak{S}_{\overline{\mathbf{W}}, \bar{b}}$, we have by eq. (5.1) that $|\{s_i\}_i^n| \leq \binom{k}{\leq r}$. Thus, $\{s_i\}_{i=1}^n$ has at most $\binom{k}{\leq r}$ unique elements. Let $\left\{s_{\iota_\ell} : \ell = 1, \ldots, \binom{k}{\leq r}\right\}$ be a set containing all such unique elements. Thus, we store

$$\text{samples } \{(x_{\iota_\ell}, y_{\iota_\ell}) : \ell = 1, \ldots, \tbinom{k}{\leq r}\}. \tag{5.3}$$

Using $\overline{\mathbf{W}}, \bar{b}$ as defined above, we have $s_{\iota_\ell} = \mathtt{sgn}(\overline{\mathbf{W}}^\top x_{\iota_\ell} + \bar{b})$. Now, simply choose g such that $g(s_{\iota_\ell}) = y_{\iota_\ell}$ for all $\ell = 1, \ldots, \binom{k}{\leq r}$.

To summarize, we formally define the compression and reconstruction functions (κ, ρ). Let κ take the full sample $\{(x_i, y_i)\}_{i=1}^n$ and output the subsample (and side

information) in eq. (5.2) and eq. (5.3). The reconstruction function ρ first constructs $\overline{\mathbf{W}}, \overline{b}$ using eq. (5.2). Next, ρ constructs g using $\overline{\mathbf{W}}, \overline{b}$ and the samples of eq. (5.3). $\square$

Now, the following result together with the sample compression scheme for $\texttt{HAC}(d, r, k)$ we constructed imply theorem V.6 from the previous section.

Theorem V.13 ([LW86])**.** *If $\mathcal{C}$ has sample compression scheme (ρ, κ), then $\texttt{VC}(\mathcal{C}) \leq 8 \cdot \texttt{size}(\rho, \kappa)$.*

Remark V.14. Note that the reconstruction function ρ is *not* permutation-invariant. Furthermore, the overall sample compression scheme ρ, κ is *not* stable in the sense of [HK21]. In general, sample compression schemes with permutation-invariant ρ [FW95] and *stable* sample compression schemes [HK21] enjoy tighter generalization bounds compared to ordinary sample compression schemes. We leave as an open question whether $\texttt{HAC}(d, r, k)$ has such specialized compression schemes.

5.5 Minimax-optimality for learning Lipschitz class

In this section, we show that empirical risk minimization (ERM) with respect to the 0-1 loss over $\texttt{HAC}(d, r, k)$, for properly chosen r and k, is minimax optimal for classification where the posterior class probability function is L-Lipschitz, for fixed $L > 0$. Furthermore, the choices for r and k is such that the associated HANN, the neural network realization of $\texttt{HAC}(d, r, k)$, is overparametrized for the Boolean function implementations discussed in Remark V.7.

Below, let $X \in [0,1]^d$ and $Y \in \mathbb{B}$ be the random variables corresponding to a sample and label jointly distributed according to P. Write $\eta_P(x) := P(Y = 1 | X = x)$ for the posterior class probability function.

Let $\Sigma(L, [0,1]^d)$ denote the class of L-Lipschitz functions $f : [0,1]^d \to \mathbb{R}$, i.e.,

$$|f(x) - f(x')| \leq L\|x - x'\|_2, \quad \forall x, x' \in [0,1]^d.$$

The following minimax lower bound result[1] concerns classification when η_P is L-Lipschitz:

Theorem V.15 ([AT07])**.** *There exists a constant $C > 0$ such that*

$$\inf_{\tilde{f}_n} \sup_{P\,:\,\eta_P \in \Sigma(L,[0,1]^d)} \mathbb{E}[R(\tilde{f}_n)] - R_P^* \geq C n^{-\frac{1}{d+2}}.$$

The infimum above is taken over all possible learning algorithms $\tilde{f}_n$, i.e., mappings from $(\mathcal{X} \times \mathbb{B})^n$ to Borel measurable functions $\mathcal{X} \to \mathbb{B}$. When $\hat{f}_n$ is an empirical risk minimizer (ERM) over $\texttt{HAC}(d, r, k)$ where $d = r$ for $k = n^{\frac{1}{d+2}}$, this minimax rate is achieved.

Theorem V.16. *Let $d \geq 1$ be fixed. Let $\hat{f}_n$ be an ERM over $\texttt{HAC}(d, d, k)$ where $k = k(n) \asymp n^{\frac{1}{d+1}}$. Then there exists a constant C' such that*

$$\sup_{P\,:\,\eta_P \in \Sigma(L,[0,1]^d)} \mathbb{E}[R(\hat{f}_n)] - R_P^* \leq C' n^{-\frac{1}{d+2}}.$$

Proof sketch (see Section 5.8.3 for full proof). We first show that the histogram classifier over the standard partition of $[0, 1]^d$ into smaller cubes is an element of $\mathcal{C} := \texttt{HAC}(d, d, k)$, thus reducing the problem to proving minimax-optimality of the histogram classifier. Previous work [Gyö+06] Theorem 4.3 established this for the histogram *regressor*. The analogous result for the histogram *classifier*, to the best of our knowledge, has not appeared in the literature and thus is included for completeness.

The neural network implementation of $\texttt{HAC}(d, d, k)$ where $k \asymp n^{1/(d+2)}$ in theorem V.16 can be overparametrized. Using either the 1- or the 2-hidden layer neural network implementations of Boolean functions as in Remark V.7, the resulting HANN is overparametrized and has number of weights either $\geq k2^k$ or $\geq c^2 2^k/k$ respectively. Both $k2^k$ and $c^2 2^k/k \gg n$ exponentially while $\texttt{VC}(\texttt{HAC}(d, d, k)) = o(n)$.

[1] The result we cite here is a special case of [AT07, Theorem 3.5], which gives minimax lower bound for when η_P has additional smoothness assumptions.

5.6 Empirical results

In this section, we discuss experimental results of using HANNs for classifying synthetic and real datasets. Our implementation uses TensorFlow [Aba+16] with the Larq [GT20] library for training neural networks with threshold activations. Note that theorem V.16 holds for ERM with respect to the 0-1 loss over HANNs, which is intractable in practice. Furthermore, our theory is for binary classification, while some of the datasets in the experiments are multiclass.

Synthetic datasets. We apply a HANN (model specification shown in fig. 5.3-top left) to the MOONS synthetic dataset with two classes with the hinge loss.

The heuristic for training networks with threshold activation can significantly affect the performance [Kim+19]. We consider two of the most popular heuristics: the straight-through-estimator (SteSign) and the SwishSign, introduced by [Hub+17] and [Dar+19], respectively. SwishSign reliably leads to higher validation accuracy (fig. 5.3-bottom left), consistent with the finding of [Dar+19]. Subsequently, we use SwishSign and plot a learned decision boundary in fig. 5.3-right.

By [MB17] Lemma 2.1, any Boolean function $\mathbb{B}^k \to \mathbb{B}$ can be implemented by a 1-hidden layer ReLU network with 2^k hidden nodes. Here, the width of the hidden layer is $2^{10} = 1024$. Thus, the architecture in fig. 5.3 can assign labels to the bold boundary cells arbitrarily without changing the training loss. Nevertheless, the optimization appears to be biased toward a topologically simpler classifier. This behavior is consistently reproducible. See fig. 5.7.

Real-world datasets. [Kla+17] introduced *self-normalizing neural networks* (SNN) which were shown to outperform other neural networks on a panel of 121 UCI datasets. Subsequently, [Wu+18] proposed the *dendritic neural network* architecture, which further improved classification performance on this panel of datasets. Following their works, we evaluate the performance of HANNs on the 121 UCI datasets.

A crucial hyperparameter for HANN is k, the number of hyperplanes used. We ran

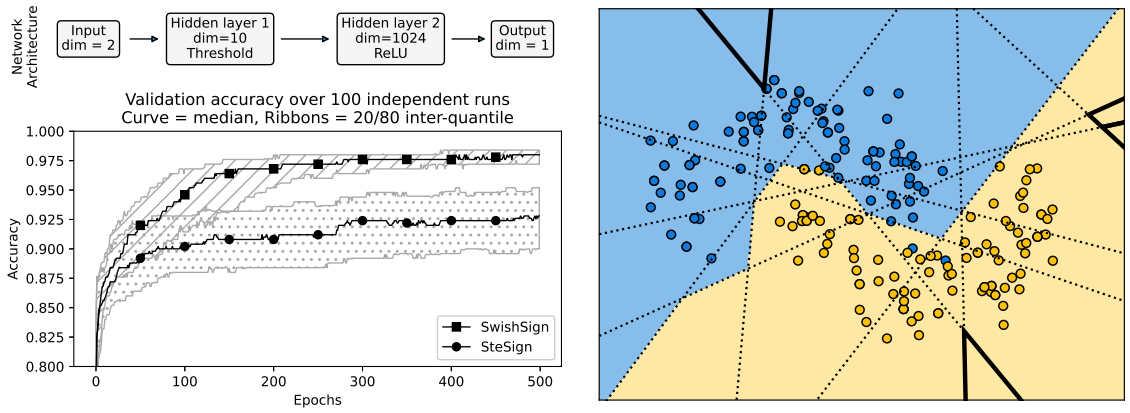

Figure 5.3: *Top left.* Architecture of HANN used for the MOONS dataset. *Bottom left.* Validation accuracies from 100 independent runs with random initialization and data generation. *Right.* Data points (circles) drawn from `make_moons` in `sklearn` colored by ground truth labels. The hyperplane arrangement is denoted by dotted lines. Coloring of the cells corresponds to the decision region of the trained classifier. A cell Δ is highlighted by bold boundaries if 1) no training data lies in Δ and 2) Δ does not touch the decision boundary.

the experiments with $k \in \{15, 100\}$ to test the hyperparameter's impact on accuracy. The Boolean function h is implemented as a 1-hidden layer residual network [He+16] of width 1000. The logistic loss is used.

We use the same train, validation, and test sets from the public code repository of [Kla+17]. The reported accuracies on the held-out test set are based on the best performing model according to the validation set. The models will be referred to as `HANN15` and `HANN100`, respectively. The results are shown in fig. 5.4. The accuracies of SNN and DENN are obtained from Table A1 in the supplemental materials of [Wu+18]. Full details for the training and accuracy tables can be found at the end of the chapter.

The `HANN15` model (top row of fig. 5.4) already achieves median accuracy within 1.5% of both SNN and DENN. With the larger `HANN100` model (bottom row), the gap is reduced to zero. The largest training set in this panel of datasets has size 77904. The `HANN15` and `HANN100` models use $\approx 10^4$ and 10^5 weights, respectively.

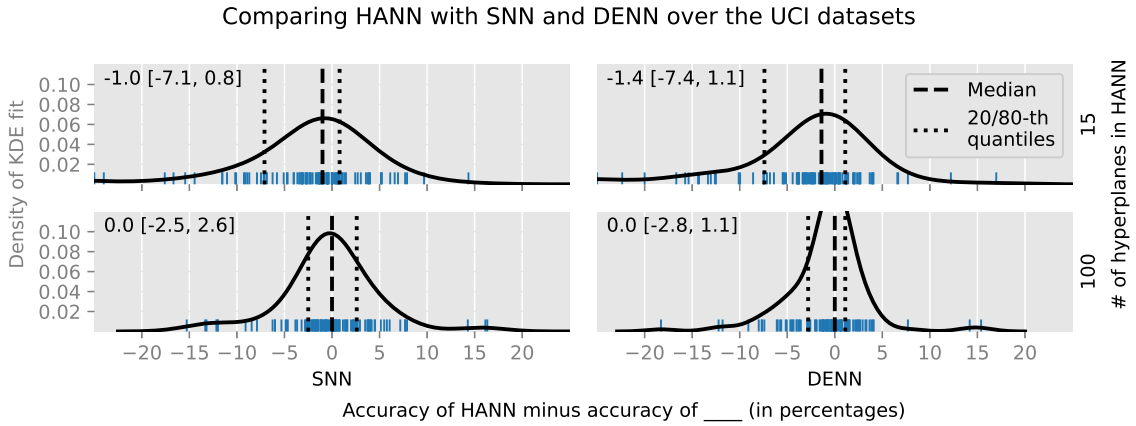

Figure 5.4: Each blue tick above the x-axis represents a single dataset, where the x-coordinate of the tick is the difference of the accuracy of HANN and either SNN (left) or DENN (right) on the dataset. The solid black curves are kernel density estimates for the blue ticks. The number of hyperplanes used by HANN is either 15 (top) or 100 (bottom). The quantities shown in the top-left corner of each subplot are the median, 20-th and 80-th quantiles of the differences, respectively, rounded to 1 decimal place.

By comparison, the average numbers of weights[2] used by SNN and DENN are both $\geq 5 * 10^5$. Thus, all three models considered here, namely HANN, SNN and DENN, are overparametrized for this panel of datasets.

5.7 Discussion

We have introduced an architecture for which the VC theorem can be used to prove minimax-optimality of ERM over HANNs in an overparametrized setting with Lipschitz posterior. To our knowledge, this is the first time VC theory has been used to analyze the performance of a neural network in the overparametrized regime. Furthermore, the same architecture leads to state-of-the-art performance over a benchmark collection of unstructured datasets.

[2]Details on these estimates are included in Section 5.10.

To the best of our knowledge, no existing theoretical bound for overparametrized NNs yields meaningful results. Yet there is immense interest in understanding what aspects of deep NNs can explain their performance, even if the bounds aren't yet small [BFT17; Ney+17; Jia+19]. Our work shows that the compressibility of the network, as reflected by the sample compression scheme, is a useful avenue, and one that has not previously been explored – ours is the first work applying sample compression to NNs. This seems likely to open the door to further analysis of quantized NNs.

5.8 Omitted proofs

5.8.1 Proof of Proposition V.10

By definition, there exists $h \in \texttt{Bool}_k$, $\mathbf{W} \in \mathbb{R}^{d\times k}$ of rank at most r, and $b \in \mathbb{R}^k$ such that $y_i = h(\texttt{sgn}(\mathbf{W}^\top x_i + b))$.

Now, let $j \in [k]$ be fixed. Since $|w_j^\top x_i + b_j| \geq 0$ for all $i \in [n]$, there exists a small perturbation $\tilde{c}_j$ of b_j such that $|w_j^\top x_i + \tilde{c}_j| > 0$ for all $i \in [n]$. Now, let $\lambda_j := \min_{i\in[n]} |w_j^\top x_i + \tilde{c}_j|$ which is positive. Define $v_j := w_j/\lambda_j$ and $c_j = \tilde{c}_j/\lambda_j$, we have $|v_j^\top x_i + c_j| \geq 1$ for all $i \in [n]$, as desired. Note that $\text{rank}(\mathbf{V}) = \text{rank}(\mathbf{W})$. □

5.8.2 Proof of Proposition V.12

Let $g_i(x) = a_i^\top x - b_i$ for each $i \in [m]$ and $f(x) = \frac{1}{2}\|x\|_2^2$. Then $\nabla f(x) = x$ and $\nabla g_i(x) = a_i$. By definition, x_I^* is a minimizer of

$$\min_{x\in\mathbb{R}^n} f(x) \text{ s.t. } g_i(x) \leq 0, \forall i \in I,$$

which is a convex optimization with strongly convex objective. Thus, the minimizer x_I^* is unique and furthermore is the unique element x of $\mathbb{R}^n$ satisfying the KKT

conditions:

$$x \in \mathcal{P}_I \text{ and } \exists \text{ a set of nonnegative weights } \{\lambda_i\}_{i\in I} \text{ such that } -x = \sum_{i\in I} \lambda_i a_i.$$

Thus, x^*_I can be equivalently characterized as the unique element of $x \in \mathbb{R}^n$ satisfying

$$x \in \mathcal{P}_I \text{ and } -x \in \texttt{coni}(\{a_i\}_{i\in I}). \tag{5.4}$$

In particular, $x^*_{[m]} \in \mathcal{P}_{[m]}$ and $-x^*_{[m]} \in \texttt{coni}(\{a_i\}_{i\in[m]})$. By the Carathédory's theorem for the conical hull [LP09], there exists $\underline{I} \subseteq [m]$ such that $|\underline{I}| = n$ and $-x^*_{[m]} \in \texttt{coni}(\{a_i\}_{i\in \underline{I}})$. Thus, for any $J \subseteq [m]$ such that $\underline{I} \subseteq J$, we have $-x^*_{[m]} \in \texttt{coni}(\{a_i\}_{i\in J})$. Furthermore, $J \subseteq [m]$ implies $\mathcal{P}_J \supseteq \mathcal{P}_{[m]}$. In particular, $x^*_{[m]} \in \mathcal{P}_J$. Putting it all together, we have $x^*_{[m]} \in \mathcal{P}_J$ and $-x^*_{[m]} \in \texttt{coni}(\{a_i\}_{i\in J})$. By the uniqueness, we have $x^*_J = x^*_{[m]}$. □

5.8.3 Proof of theorem V.16

In this proof, the constant C does not depending on n, and may change from line to line.

We fix a joint distribution P such that $\eta_P \in \Sigma(L, [0,1]^d)$ throughout the proof. Thus, the notation for risks will omit the P in their subscript, e.g., we write $\hat{R}_n(f)$ instead of $\hat{R}_{P,n}(f)$ and R^* instead of R^*_P. Below, let $\beta > \alpha > 0$ be constants such that $\alpha d n^{1/(d+2)} \le k \le \beta d n^{1/(d+2)}$. Let $\tilde{k} := \lceil k/d \rceil$.

Let $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{\tilde{k}^d}$ denote the hypercubes of side length $\ell = 1/\tilde{k}$ forming a partition of $[0,1]^d$. For each $i \in [\tilde{k}^d]$, let $\mathcal{R}_i^- := \{x \in \mathcal{R}_i : \eta_P(x) < 1/2\}$ and $\mathcal{R}_i^+ := \{x \in \mathcal{R}_i : \eta_P(x) \ge 1/2\}$.

Let $\tilde{f} : [0,1]^d \to \mathbb{B}$ be the classifier such that

$$\tilde{f}(x) = \begin{cases} +1 & : x \in \mathcal{R}_i,\ \int_{\mathcal{R}_i} \eta_P(x) dP(x) \geq \int_{\mathcal{R}_i} (1 - \eta_P(x)) dP(x) \\ -1 & : x \in \mathcal{R}_i,\ \int_{\mathcal{R}_i} \eta_P(x) dP(x) < \int_{\mathcal{R}_i} (1 - \eta_P(x)) dP(x). \end{cases}$$

In other words, $\tilde{f}$ classifies all $x \in \mathcal{R}_i$ as $+1$ if and only if $P(Y = 1 | X \in \mathcal{R}_i) \geq 1/2$. This is commonly referred to as the *histogram classifier* [Gyö+06]. It is easy to see that

$$P(\tilde{f}(X) \neq Y, X \in \mathcal{R}_i) = \min \left\{ \int_{\mathcal{R}_i} (1 - \eta_P(x)) dP(x), \int_{\mathcal{R}_i} \eta_P(x) dP(x) \right\}$$

For the remainder of this proof, we write "$\sum_i$" to mean "$\sum_{i \in [\tilde{k}^d]}$". Thus,

$$R(\tilde{f}) = \sum_i P(\tilde{f}(X) \neq Y, X \in \mathcal{R}_i) = \sum_i \min \left\{ \int_{\mathcal{R}_i} (1 - \eta_P(x)) dP(x), \int_{\mathcal{R}_i} \eta_P(x) dP(x) \right\}.$$

Next, we note that $\tilde{f} \in \texttt{HAC}(d, d, k)$. To see this, let $j \in [d]$. Take $H_{j1}, \ldots, H_{j(\tilde{k}-1)} \subseteq \mathbb{R}^d$ to be the hyperplanes perpendicular to the j-th coordinate where, for each $\ell \in [\tilde{k}]$, $H_{j\ell}$ intersects the j-th coordinate axis at $\ell/\tilde{k}$. Consider the hyperplane arrangement consisting of all $\{H_{j\ell}\}_{j \in [d], \ell \in [\tilde{k}-1]}$ and let $\{C_1, C_2, \ldots\}$ be its cells. Then $\{C_1 \cap [0,1]^d, C_2 \cap [0,1]^d, \ldots\} = \{\mathcal{R}_1, \ldots, \mathcal{R}_{\tilde{k}^d}\}$ is the partition of $[0,1]^d$ by $1/\tilde{k}$ side length hypercubes. See fig. 5.5.

Let $\mathbf{W}$ be the matrix of normal vectors and b be the vector of offsets representing this hyperplane arrangement, which requires $d(\tilde{k} - 1) = d(\lceil k/d \rceil - 1) \leq d(k/d) = k$ hyperplanes. Since $\tilde{f}$ is constant on $\mathcal{R}_i$, there exists a Boolean function $h \in \texttt{Bool}_k$ such that $h \circ q_{\mathbf{W},b}|_{[0,1]^d} = \tilde{f}$. From this, we conclude that $\tilde{f} \in \texttt{HAC}(d, d, k)$.

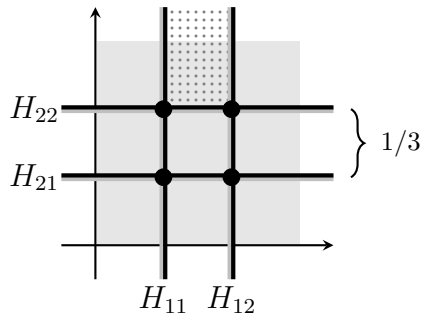

Figure 5.5:
Partition of $[0,1]^d$ into $1/\tilde{k}$ hypercubes via arrangement of $d(\tilde{k}-1)$ hyperplanes, where $d=2$ and $\tilde{k}=3$. Shaded region is $[0,1]^d$. Dotted region is a cell of the hyperplane arrangement.

Thus $\hat{R}_n(\hat{f}_n) - \hat{R}_n(\tilde{f}) \leq 0$ and so

$$R(\hat{f}_n) - R^* = R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) + \underbrace{\hat{R}_n(\hat{f}_n) - \hat{R}_n(\tilde{f})}_{\leq 0} + \hat{R}_n(\tilde{f}) - R(\tilde{f}) + R(\tilde{f}) - R^*$$

$$\leq \underbrace{R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)}_{\text{Term 1}} + \underbrace{\hat{R}_n(\tilde{f}) - R(\tilde{f})}_{\text{Term 2}} + \underbrace{R(\tilde{f}) - R^*}_{\text{Term 3}}.$$

We now bound Terms 1 and 2 using the uniform deviation bound. From theorem V.6, we know that there exists a constant C independent of n such that

$$\texttt{VC}(\texttt{HAC}(d,d,k)) \leq 8 \cdot \left(k(d+1) + k(d+1)(1 + \lceil \log_2(k) \rceil) + \binom{k}{\leq d} \right) \leq Ck^d.$$

Thus, by theorem V.5 with $\delta = 1/(2n)$ and a union bound, with probability at least $1 - 1/n$

$$\max\left\{ |\hat{R}_n(\hat{f}_n) - R(\hat{f}_n)|, |\hat{R}_n(\tilde{f}) - R(\tilde{f})| \right\} \leq C\sqrt{\frac{k^d + \log(n)}{n}} \tag{5.5}$$

for some $C > 0$.

Next, we focus on Term 3. Recall that

$$R^* = \int_{[0,1]^d} \min\{\eta_P(x), 1 - \eta_P(x)\} dP(x) = \sum_i \int_{\mathcal{R}_i} \min\{\eta_P(x), 1 - \eta_P(x)\} dP(x)$$

and that

$$R(\tilde{f}) = \sum_i \min\left\{\int_{\mathcal{R}_i} \eta_P(x)dP(x), \int_{\mathcal{R}_i} 1-\eta_P(x)dP(x)\right\}.$$

Fix some $i \in [k^d]$. Our goal now is to bound the difference between the i-th summands in the above expressions for $R(\tilde{f})$ and R^*:

$$\min\left\{\int_{\mathcal{R}_i} \eta_P(x)dP(x), \int_{\mathcal{R}_i} 1-\eta_P(x)dP(x)\right\} - \int_{\mathcal{R}_i} \min\{\eta_P(x), 1-\eta_P(x)\}dP(x). \quad (5.6)$$

First, consider the case that

$$\min\left\{\int_{\mathcal{R}_i} \eta_P(x)dP(x), \int_{\mathcal{R}_i} 1-\eta_P(x)dP(x)\right\} = \int_{\mathcal{R}_i} \eta_P(x)dP(x). \quad (5.7)$$

We claim that there must exist $x_0 \in \mathcal{R}_i$ such that $\eta_P(x_0) \leq 1/2$. Suppose $\eta_P(x) > 1/2$ for all $x \in \mathcal{R}_i$. Then $\eta_P(x) > 1/2 > 1-\eta_P(x)$. Since $\eta_P(x)$ is continuous, this would contradict eq. (5.7).

Continue assuming eq. (5.7), we further divide into two subcases: (1) $\eta_P(x) \leq 1/2$ for all $x \in \mathcal{R}_i$, and (2) there exists some $x_1 \in \mathcal{R}_i$ such that $\eta_P(x_1) > 1/2$.

Under subcase (1), $\min\{\eta_P(x), 1-\eta_P(x)\} = \eta_P(x)$ for all $x \in \mathcal{R}_i$ in which case *eq.* (5.6) $= 0$.

Under subcase (2), since $\eta_P(x_0) \leq 1/2 < \eta_P(x_1)$, we know by the intermediate

value theorem that there must exist $x' \in \mathcal{R}_i$ such that $\eta_P(x') = 1/2$. Now,

$$\begin{aligned}
\textit{eq. (5.6)} &= \int_{\mathcal{R}_i} (\eta_P(x) - \min\{\eta_P(x), 1-\eta_P(x)\}) dP(x) \\
&\leq \int_{\mathcal{R}_i^+} (\eta_P(x) - \min\{\eta_P(x), 1-\eta_P(x)\}) dP(x) \\
&\quad + \int_{\mathcal{R}_i^-} (\eta_P(x) - \min\{\eta_P(x), 1-\eta_P(x)\}) dP(x) \\
&= \int_{\mathcal{R}_i^+} (\eta_P(x) - (1-\eta_P(x)) dP(x) \qquad \because \text{Definition of } \mathcal{R}_i^{\pm} \\
&\quad + \int_{\mathcal{R}_i^-} (\eta_P(x) - \eta_P(x)) dP(x) \\
&= \int_{\mathcal{R}_i^+} (2\eta_P(x) - 1) dP(x) \\
&= 2 \int_{\mathcal{R}_i^+} (\eta_P(x) - \eta_P(x')) dP(x) \qquad \because 2\eta_P(x') = 1 \\
&\leq 2L \int_{\mathcal{R}_i^+} \|x - x'\|_2 dP(x) \\
&\leq 2L\sqrt{d} \Pr(\mathcal{R}_i)/\tilde{k} \qquad \because \|x-x'\|_2 \leq \sqrt{d}\|x-x'\|_1 \leq \sqrt{d}(1/\tilde{k}) \\
&\leq 2Ld^{3/2} \Pr(\mathcal{R}_i)/k \qquad \because 1/\tilde{k} = 1/\lceil k/d \rceil \leq 1/(k/d) = d/k.
\end{aligned}$$

Thus, under assumption eq. (5.7), we have proven that *eq.* (5.6) $\leq 2Ld^{3/2}/k$. For the other assumption, i.e., the minimum in eq. (5.7) is attained by $\int_{\mathcal{R}_i} 1-\eta_P(x) dP(x)$, a completely analogous argument again shows that *eq.* (5.6) $\leq 2Ld^{3/2}/k$.

Putting it all together, we have

$$R(\tilde{f}_n) - R^* \leq 2Ld^{3/2} \sum_i P(\mathcal{R}_i)/k = 2Ld^{3/2}/k. \tag{5.8}$$

We have shown that, with probability at least $1 - 1/n$,

$$R(\hat{f}_n) - R^* \leq C\sqrt{\frac{k^d + \log(n)}{n}} + \frac{2Ld^{3/2}}{k}.$$

Using $\alpha d n^{1/(d+2)} \leq k \leq \beta d n^{1/(d+2)}$, we have with probably at least $1 - 1/n$ that

$$\begin{aligned}
R(\hat{f}_n) - R^* &\leq C\sqrt{\frac{k^d + \log(n)}{n}} + \frac{2Ld^{3/2}}{k} \\
&\leq C\sqrt{\frac{(\beta d)^d n^{d/(d+2)} + \log(n)}{n}} + \frac{2Ld^{3/2}}{\alpha d n^{1/(d+2)}} \\
&\leq C\left(\sqrt{\frac{n^{d/(d+2)}}{n}} + n^{-1/(d+2)}\right) \qquad \because \log(n) = o(n^{1/d+2}) \\
&= C\left(\sqrt{n^{-2/(d+2)}} + n^{-1/(d+2)}\right) \\
&\leq C n^{-\frac{1}{d+2}}.
\end{aligned}$$

Taking expectation, we have $\mathbb{E}[R(\hat{f}_n)] - R^* \leq (1 - 1/n)Cn^{-\frac{1}{d+2}} + 1/n \cdot 1 \leq Cn^{-\frac{1}{d+2}}$.

□

5.9 Training details

Data preprocessing. The pooled training and validation data is centered and standardized using the `StandardScaler` function from sklearn. The transformation is also applied to the test data, using the centers and scaling from the pooled training and validation data:

```
scaler = StandardScaler().fit(X_train_valid)
X_train_valid = scaler.transform(X_train_valid)
X_test = scaler.transform(X_test)
```

If the feature dimension and training sample size are both > 50, then the data is dimension reduced to 50 principal component features:

```
if min(X_train_valid.shape) > 50:
```

```
    pca = PCA(n_components = 50).fit(X_train_valid)
    X_train_valid = pca.transform(X_train_valid)
    X_test = pca.transform(X_test)
```

Note that this is equivalent to freezing the weights between the Input and the Latent layer in fig. 5.2.

Validation and test accuracy. Every 10 epochs, the validation accuracy during the past 10 epochs are averaged. A smoothed validation accuracy is calculated as follows:

```
val_acc_sm = (1-sm_param)*val_acc_sm + sm_param*val_acc_av
## Variable description:
# sm_param = 0.1
# val_acc_av = average of the validation in the past 10 epochs
# val_acc_sm = smoothed validation accuracy
```

The predicted test labels is based on the snapshot of the model at the highest smoothed validation accuracy, at the end once max epochs is reached.

Heuristic for coarse gradient of the threshold function. We use the Swish-Sign from the Larq library [GT20].

```
# import larq as lq
qtz = lq.quantizers.SwishSign()
```

Dropout. During training, dropout is applied to the Boolean output of the threshold function, i.e, the variables $\mathtt{B}_1, \mathtt{B}_2, \ldots, \mathtt{B}_k$ in fig. 5.2. This improves generalization by preventing the training accuracy from reaching 100%.

```
# from tensorflow.keras.layers import Dense, Dropout
hyperplane_enc = Dense(n_hyperplanes, activation = qtz)(inputs)
hyperplane_enc = Dropout(dropout_rate)(hyperplane_enc)
```

Implementation of the Boolean function. For the Boolean function h, we use a 1-hidden layer residual network [He+16] with 1000 hidden nodes:

```
# from tensorflow.keras.layers import Dense, Add
```

```
# output_dim = num_classes
n_hidden = 1000
hidden = Dense(n_hidden, activation="relu")(hyperplane_enc)
out_hidden = Dense(output_dim, activation = "linear")(hidden)
out_skip = Dense(output_dim, activation = "linear")(hyperplane_enc)
outputs = Add()([out_skip,out_hidden])
```

Hyperparameters. `HANN15` is trained with a hyperparameter grid of size 3 where only the dropout rate is tuned. The hyperparameters are summarized in Table 5.2. The model with the highest smoothed validation accuracy is chosen.

The model `HANN15` is trained with the following hyperparameters:

Table 5.1: `HANN15` model and training hyperparameter grid

Optimizer	SGD
Learning rate	0.01
Dropout rate	$\{0.1, 0.25, 0.5\}$
Minibatch size	128
Boolean function	1-hidden layer resnet with 1000 hidden nodes
Epochs	100 for miniboone, 5000 for all others

For `HANN100`, we only used 1 set of hyperparameters.

Table 5.2: `HANN100` model and training hyperparameter

Optimizer	SGD
Learning rate	0.01
Dropout rate	0.5
Minibatch size	128
Boolean function	1-hidden layer resnet with 1000 hidden nodes
Epochs	100 for miniboone, 5000 for all others

5.10 Parameter counts

The widest part of `HANN15` and `HANN100` models are the weights mapping from $\mathbb{B}^k$ (k = number of hyperplanes) to $\mathbb{R}^{1000}$ (1000 = number of hidden layer of the boolean function) where $k \in \{15, 100\}$. Thus, the two HANN models use $\geq 15 \times 1000 \geq 10^4$ and $\geq 100 \times 1000 = 10^5$ weights, respectively.

The weight count estimates for the Self-normalized Neural Network (SNN) and Dendritic Neural Network (DENN) use the formula (# layers $-$ 1) $\times$ (# neurons per layer)^2.

For the Self-normalized Neural Network (SNN), average number of layers = 10.8, and the number of neurons per layers $\geq$ 256, found on page 7 and Table A4 of [Kla+17], respectively. The number of weights is $\geq (10-1) * (256^2) = 655,360$ weights.

The parameters for the dendritic neural network (DENN) is found in the public GitHub repository xiangwenliu/DENN of [Wu+18] which lists number of layers = 3 and number of neurons per layer = 512, found on line 41 and 52 of `train_uci.py`, respectively. The number of weights is $\geq (3-1) * (512^2) = 524,288$ weights.

5.11 Additional plots

Multiclass hinge versus cross-entropy loss. fig. 5.6 shows the accuracy differences when the Weston-Watkins hinge loss is used. Compared to the results shown in fig. 5.4, the performance for `HANN100` is slightly worse and the performance for `HANN15` is slightly better.

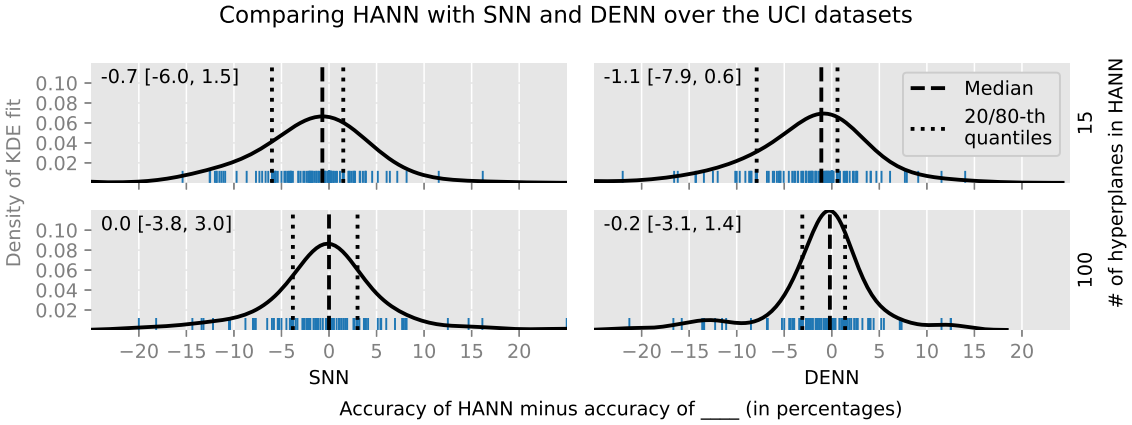

Figure 5.6: Each blue tick above the x-axis represents a single dataset, where the x-coordinate of the tick is the difference of the accuracy of HANN and either SNN (left) or DENN (right) on the dataset. The number of hyperplanes used by HANN is either 15 (top) or 100 (bottom). The quantities shown in the top-left corner of each subplot are the median, 20-th and 80-th quantiles of the differences, respectively, rounded to 1 decimal place.

Implicit bias for low complexity decision boundary. In fig. 5.7, we show additional results ran with the same setting for the MOONS synthetic dataset as in the Empirical Results section. From the perspective of the training loss, the label assignment in the bold-boundary regions is irrelevant. Nevertheless, the optimization consistently appears to be biased toward the geometrically simpler classifier, despite the capacity for fitting complex classifiers.

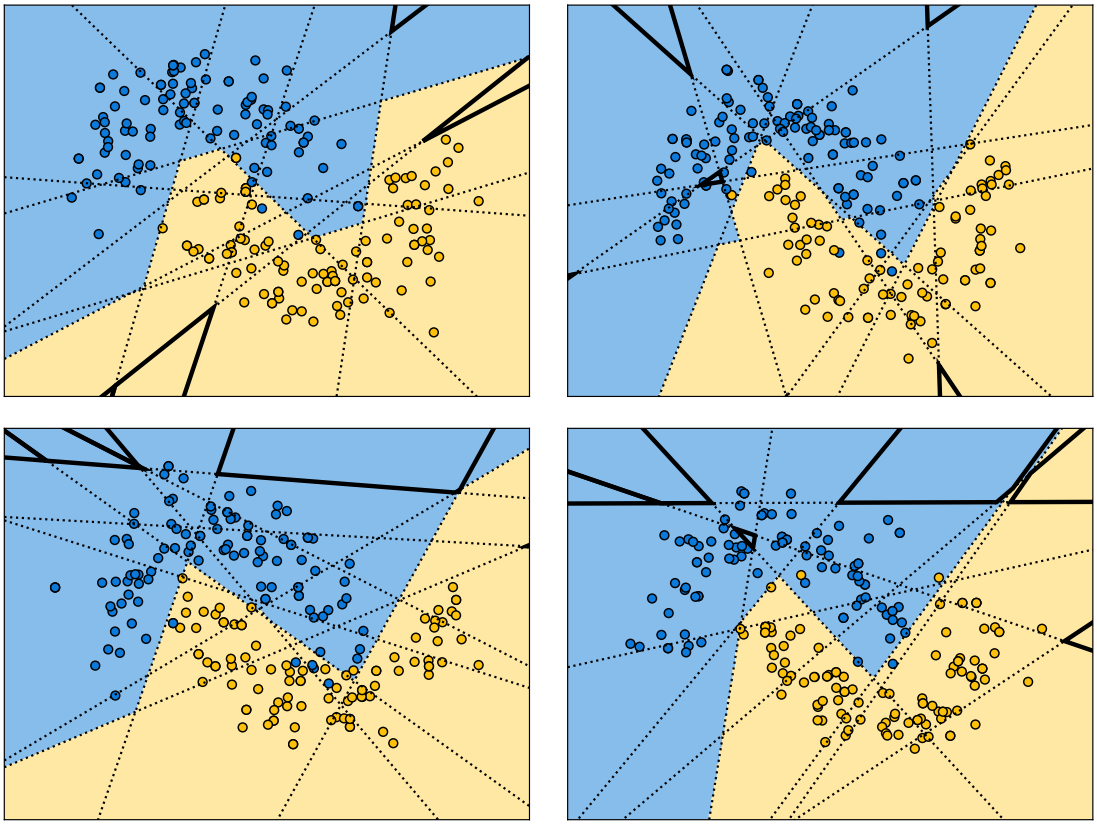

Figure 5.7: Four independent runs of HANN on the MOONS synthetic dataset. Data points (circles) drawn from `make_moons` in `sklearn` colored by ground truth labels. The hyperplane arrangement is denoted by dotted lines. Coloring of the cells corresponds to the decision region of the trained classifier. A cell $\mathcal{C}$ is highlighted by bold boundaries if 1) no training data lies in $\mathcal{C}$ and 2) $\mathcal{C}$ does not touch the decision boundary.

5.12 Table of accuracies

Table 5.3: The table of accuracies used to make fig. 5.4. The last column "HANN100trn" records the *training* accuracy at the epoch of the highest validation accuracy.

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
abalone	63.41	65.13	66.57	66.38	70.60
acute-inflammation	100.00	100.00	100.00	100.00	100.00
acute-nephritis	100.00	100.00	100.00	100.00	100.00
adult	84.32	85.04	84.76	84.80	86.35
annealing	47.00	74.00	76.00	75.00	96.81
arrhythmia	62.83	64.60	65.49	67.26	97.19
audiology-std	56.00	68.00	80.00	76.00	99.79
balance-scale	92.95	96.79	92.31	98.08	98.32
balloons	100.00	100.00	100.00	100.00	100.00
bank	88.50	88.05	89.03	89.65	95.62
blood	75.94	75.40	77.01	73.26	82.03
breast-cancer	70.42	63.38	71.83	69.01	86.63
breast-cancer-wisc	97.71	98.29	97.14	97.71	98.65
breast-cancer-wisc-diag	97.89	98.59	97.89	98.59	99.65
breast-cancer-wisc-prog	73.47	71.43	67.35	71.43	98.50

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
breast-tissue	61.54	80.77	73.08	65.38	95.78
car	98.84	100.00	98.38	98.84	99.32
cardiotocography-10clases	78.91	82.11	83.99	82.30	94.69
cardiotocography-3clases	90.58	93.97	91.53	94.35	97.09
chess-krvk	47.75	72.77	88.05	80.41	67.16
chess-krvkp	98.62	99.37	99.12	99.62	99.93
congressional-voting	61.47	57.80	61.47	57.80	66.54
conn-bench-sonar-mines-rocks	78.85	84.62	78.85	82.69	99.51
conn-bench-vowel-deterding	89.39	98.92	99.57	99.35	98.38
connect-4	78.96	86.39	88.07	86.46	86.83
contrac	52.72	49.73	51.90	54.89	62.95
credit-approval	81.98	79.65	84.30	82.56	98.47
cylinder-bands	69.53	73.44	72.66	78.12	99.54
dermatology	98.90	97.80	92.31	97.80	99.63
echocardiogram	84.85	87.88	81.82	87.88	94.62

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
ecoli	86.90	84.52	89.29	85.71	92.83
energy-y1	93.23	97.40	95.83	95.83	97.84
energy-y2	89.06	91.15	90.63	90.62	95.72
fertility	92.00	92.00	92.00	88.00	95.44
flags	39.58	50.00	45.83	52.08	94.96
glass	77.36	60.38	73.58	60.38	96.77
haberman-survival	72.37	65.79	73.68	65.79	82.34
hayes-roth	71.43	82.14	67.86	85.71	85.61
heart-cleveland	53.95	59.21	61.84	57.89	96.65
heart-hungarian	72.60	79.45	79.45	78.08	95.67
heart-switzerland	45.16	51.61	35.48	48.39	89.86
heart-va	36.00	30.00	36.00	32.00	94.34
hepatitis	82.05	82.05	76.92	79.49	100.00
hill-valley	66.83	68.81	52.48	54.62	72.69
horse-colic	80.88	83.82	80.88	82.35	97.65

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
ilpd-indian-liver	70.55	69.18	69.86	71.92	86.75
image-segmentation	87.76	90.57	91.14	90.57	99.31
ionosphere	89.77	87.50	88.64	96.59	98.84
iris	100.00	97.30	97.30	100.00	97.39
led-display	73.60	75.20	76.40	76.00	75.96
lenses	50.00	66.67	66.67	66.67	100.00
letter	81.82	96.86	97.26	96.20	96.47
libras	64.44	81.11	78.89	77.78	98.20
low-res-spect	86.47	90.23	85.71	90.23	99.81
lung-cancer	37.50	62.50	62.50	62.50	100.00
lymphography	89.19	94.59	91.89	94.59	99.67
magic	86.52	87.49	86.92	86.81	87.84
mammographic	81.25	80.00	82.50	80.83	87.10
miniboone	90.04	90.73	93.07	93.30	89.98
molec-biol-promoter	73.08	80.77	84.62	88.46	99.38

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
molec-biol-splice	79.05	78.04	90.09	85.45	98.35
monks-1	65.97	69.91	75.23	81.71	99.89
monks-2	66.20	66.44	59.26	65.05	98.42
monks-3	54.63	61.81	60.42	80.09	99.78
mushroom	100.00	100.00	100.00	100.00	99.98
musk-1	77.31	84.87	87.39	89.92	98.86
musk-2	97.21	98.61	98.91	99.27	99.83
nursery	99.75	99.91	99.78	100.00	99.56
oocytes-merluccius-nucleus-4d	86.27	83.14	82.35	83.92	94.29
oocytes-merluccius-states-2f	92.16	92.55	95.29	92.94	97.61
oocytes-trisopterus-nucleus-2f	81.14	82.02	79.82	82.46	96.13
oocytes-trisopterus-states-5b	93.86	96.05	93.42	94.74	99.26
optical	93.10	95.94	97.11	96.38	99.55
ozone	96.53	95.58	97.00	97.48	99.78
page-blocks	96.49	96.13	95.83	96.13	98.33

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
parkinsons	87.76	89.80	89.80	85.71	99.47
pendigits	94.40	97.11	97.06	97.37	99.79
pima	71.88	73.44	75.52	69.79	84.47
pittsburg-bridges-MATERIAL	88.46	92.31	88.46	92.31	99.33
pittsburg-bridges-REL-L	76.92	73.08	69.23	73.08	97.66
pittsburg-bridges-SPAN	60.87	69.57	69.57	73.91	95.09
pittsburg-bridges-T-OR-D	84.00	84.00	84.00	84.00	99.67
pittsburg-bridges-TYPE	65.38	65.38	65.38	57.69	96.39
planning	66.67	55.56	68.89	60.00	93.45
plant-margin	50.50	79.50	81.25	83.25	98.18
plant-shape	39.00	66.50	72.75	72.50	81.77
plant-texture	51.75	75.25	81.25	81.00	99.14
post-operative	40.91	63.64	72.73	68.18	95.64
primary-tumor	54.88	47.56	52.44	53.66	79.04
ringnorm	90.43	85.35	97.51	97.57	98.42

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
seeds	92.31	96.15	88.46	92.31	98.52
semeion	74.37	92.71	91.96	96.73	99.12
soybean	77.93	88.83	85.11	88.03	99.48
spambase	93.57	94.17	94.09	94.87	98.14
spect	62.90	63.44	63.98	62.37	90.87
spectf	91.98	91.98	49.73	89.30	99.66
statlog-australian-credit	65.12	63.37	59.88	61.05	78.57
statlog-german-credit	72.40	72.40	75.60	72.00	97.52
statlog-heart	85.07	91.04	92.54	92.54	93.88
statlog-image	95.15	96.88	95.49	97.75	99.00
statlog-landsat	87.55	89.25	91.00	89.90	96.39
statlog-shuttle	99.92	99.92	99.90	99.91	99.96
statlog-vehicle	78.67	77.25	80.09	81.04	95.04
steel-plates	73.61	76.49	78.35	77.53	94.86
synthetic-control	94.00	98.00	98.67	99.33	99.76

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
teaching	57.89	57.89	50.00	57.89	78.35
thyroid	98.37	98.25	98.16	98.22	99.65
tic-tac-toe	96.65	97.07	96.65	98.33	99.84
titanic	78.73	78.73	78.36	78.73	78.61
trains	100.00	50.00	NaN	NaN	100.00
twonorm	97.30	98.27	98.05	98.16	99.51
vertebral-column-2clases	88.31	85.71	83.12	85.71	93.85
vertebral-column-3clases	81.82	80.52	83.12	80.52	95.40
wall-following	92.45	94.79	90.98	91.86	98.16
waveform	85.84	84.00	84.80	83.92	95.14
waveform-noise	84.72	84.96	86.08	84.32	96.39
wine	97.73	100.00	97.73	100.00	99.80
wine-quality-red	62.50	65.00	63.00	63.50	90.25
wine-quality-white	54.82	61.03	63.73	62.25	81.79
yeast	59.03	60.65	63.07	58.22	68.48

Continued on next page

DSName	HANN15	HANN100	SNN	DENN	HANN100trn
zoo	96.00	96.00	92.00	100.00	99.68

CHAPTER VI

Consistent Interpolating Ensembles via the Manifold-Hilbert Kernel

Recent research in the theory of overparametrized learning has sought to establish generalization guarantees in the interpolating regime. Such results have been established for a few common classes of methods, but so far not for ensemble methods. We devise an ensemble classification method that simultaneously interpolates the training data, and is consistent for a broad class of data distributions. To this end, we define the *manifold-Hilbert kernel* for data distributed on a Riemannian manifold. We prove that kernel smoothing regression and classification using the manifold-Hilbert kernel are weakly consistent in the setting of Devroye et al. [DGK98]. For the sphere, we show that the manifold-Hilbert kernel can be realized as a weighted random partition kernel, which arises as an infinite ensemble of partition-based classifiers.

6.1 Introduction

Ensemble methods are among the most often applied learning algorithms, yet their theoretical properties have not been fully understood [BS16]. Based on empirical evidence, Wyner et al. [Wyn+17] conjectured that interpolation of the training data plays a key role in explaining the success of AdaBoost and random forests. However,

while a few classes of learning methods have been analyzed in the interpolating regime [Bel+19; Bar+20], ensembles have not.

Towards developing the theory of interpolating ensembles, we examine an ensemble classification method for data distributed on the sphere, and show that this classifier interpolates the training data and is consistent for a broad class of data distributions. To show this result, we develop two additional contributions that may be of independent interest. First, for data distributed on a Riemannian manifold M, we introduce the *manifold-Hilbert kernel* $K_M^{\mathcal{H}}$, a manifold extension of the *Hilbert kernel* [She68]. Under the same setting as Devroye et al. [DGK98], we prove that kernel smoothing regression with $K_M^{\mathcal{H}}$ is weakly consistent while interpolating the training data. Consequently, the classifier obtained by taking the sign of the kernel smoothing estimate has zero training error and is consistent.

Second, we introduce a class of kernels called weighted random partition kernels. These are kernels that can be realized as an infinite, weighted ensemble of partition-based histogram classifiers. Our main result is established by showing that when $M = \mathbb{S}^d$, the d-dimensional sphere, the manifold-Hilbert kernel is a weighted random partition kernel. In particular, we show that on the sphere, the manifold-Hilbert kernel is a weighted ensemble based on random hyperplane arrangements. This implies that the kernel smoothing classifier is a consistent, interpolating ensemble on $\mathbb{S}^d$. To our knowledge, this is the first demonstration of an interpolating ensemble method that is consistent for a broad class of distributions in arbitrary dimensions.

6.1.1 Problem statement

Consider the problem of binary classification on a Riemannian manifold M. Let (X, Y) be random variables jointly distributed on $M \times \{\pm 1\}$. Let $D^n := \{(X_i, Y_i)\}_{i=1}^n$ be the (random) training data consisting of n i.i.d copies of X, Y. A *classifier*, i.e., a mapping from D^n to a function $\widehat{f}(\bullet \| D^n) : M \to \{\pm 1\}$, has the **interpolating-**

consistent property if, when X has a continuous distribution, both of the following hold: 1) $\widehat{f}(X_i \| D^n) = Y_i$, for all $i \in \{1, \ldots, n\}$, and 2)

$$\Pr\{\widehat{f}(X\|D^n) \neq Y\} \to \inf_{f:M\to\{\pm 1\} \text{ measurable}} \Pr\{f(X) \neq Y\} \quad \text{in probability as } n \to \infty. \tag{6.1}$$

Our goal is to find an interpolating-consistent ensemble of *histogram classifiers*, to be defined below.

A *partition* on M, denoted by $\mathcal{P}$, is a set of subsets of M such that $P \cap P' = \emptyset$ for all $P, P' \in \mathcal{P}$ and $M = \bigcup_{P\in\mathcal{P}} P$. Given $x \in M$, let $\mathcal{P}[x]$ denote the unique element $P \in \mathcal{P}$ such that $x \in P$. The set of all partitions on a space M is denoted $\texttt{Part}(M)$. The *histogram classifier* with respect to D^n over $\mathcal{P}$ is the sign of the function $\widehat{h}(\bullet\|D^n, \mathcal{P}) : M \to \mathbb{R}$ given by

$$\widehat{h}(x\|D^n, \mathcal{P}) := \sum_{i=1}^{n} Y_i \cdot \mathbb{I}\{x \in \mathcal{P}[X_i]\}, \tag{6.2}$$

where $\mathbb{I}$ is the indicator function.

Definition VI.1. A *weighted random partition* (WRP) over M is a 3-tuple $(\Theta, \mathfrak{P}, \alpha)$ consisting of (i) *parameter space of partitions*: a set Θ where $\mathcal{P}_\theta \in \texttt{Part}(M)$ for each $\theta \in \Theta$, (ii) *random partitions*: a probability measure $\mathfrak{P}$ on Θ, and (iii) *weights*: a nonnegative function $\alpha : \Theta \to \mathbb{R}_{\geq 0}$.

Example VI.2 (Regular partition of the d-cube)**.** Let $M = [0,1]^d$ and $\Theta = \{1, 2 \ldots\} =: \mathbb{N}_+$. For each $n \in \mathbb{N}_+$, denote by $\mathcal{P}_n$ the regular partition of M into n^d d-cubes of side length $1/n$. For any probability mass function $\mathfrak{P}$ on $\mathbb{N}_+$ and weights $\alpha : \mathbb{N}_+ \to \mathbb{R}_{\geq 0}$, the 3-tuple $(\Theta, \mathfrak{P}, \alpha)$ is a WRP.

Below, WRPs will be denoted with 2-letter names in the sans-serif font, e.g., "rp" for a generic WRP, and "ha" for the weighted hyperplane arrangement random partition (Definition VI.14). The *weighted random partition kernel* associated to

$\mathsf{rp} = (\Theta, \mathfrak{P}, \alpha)$ is defined as

$$K_M^{\mathsf{rp}} : M \times M \to \mathbb{R}_{\geq 0} \cup \{\infty\}, \quad K_M^{\mathsf{rp}}(x, z) := \mathbb{E}_{\theta \sim \mathfrak{P}}[\alpha(\theta)\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}]. \tag{6.3}$$

When $\alpha \equiv 1$, we recover the notion of unweighted random partition kernel introduced in [DG14]. Note that the kernel is symmetric since $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\} = \mathbb{I}\{z \in \mathcal{P}_\theta[x]\}$. If $K_M^{\mathsf{rp}} < \infty$, then K_M^{rp} is a positive definite (PD) kernel. When K_M^{rp} can evaluate to ∞, the definition of a PD kernel is not applicable since the positive definite property is defined only for to kernels taking finite values [BT11].

Let $\mathtt{sgn} : \mathbb{R} \cup \{\pm\infty\} \to \{\pm 1\}$ be the sign function. For a WRP, define the weighted infinite-ensemble

$$\widehat{u}(x \| D^n, K_M^{\mathsf{rp}}) := \sum_{i=1}^{n} Y_i \cdot K_M^{\mathsf{rp}}(x, X_i) = \mathbb{E}_{\theta \sim \mathfrak{P}}[\alpha(\theta)\widehat{h}(x \| D^n, \mathcal{P}_\theta)]. \tag{6.4}$$

Note that the equality on the right follows immediately from linearity of the expectation and the definition of $\widehat{h}(\bullet \| D^n, \mathcal{P}_\theta)$ in Equation (6.2).

Main problem. Find a WRP such that $\mathtt{sgn}(\widehat{u}(\bullet \| D^n, K_M^{\mathsf{rp}}))$ has the interpolating-consistent property.

6.1.2 Outline of approach and contributions

In the regression setting, we have (X, Y) jointly distributed on $M \times \mathbb{R}$. Let $m(x) := \mathbb{E}[Y|X = x]$. Recall from Belkin et al. [BRT19, Equation (7)] the definition of the *kernel smoothing estimator* with a so-called *singular*[1] kernel $K : M \times M \to [0, +\infty]$:

$$\widehat{m}(x \| D^n, K) := \begin{cases} Y_i & : \exists i \in [n] \text{ such that } x = X_i \\ \frac{\sum_{i=1}^n Y_i K(x, X_i)}{\sum_{j=1}^n K(x, X_j)} & : \sum_{j=1}^n K(x, X_j) > 0 \\ 0 & : \text{otherwise.} \end{cases} \tag{6.5}$$

[1]The "singular" modifier refers to the fact that $K(x, x) = +\infty$ for all $x \in M$.

We note that Equation (6.5) is referred as the *Nadaraya-Watson* estimate in [BRT19]. Now, we simply write $\widehat{m}_n(x)$ instead of $\widehat{m}(x\|D^n, K)$ when there is no ambiguity. Similarly, we write $\widehat{u}_n(x)$ instead of $\widehat{u}(x\|D^n, K)$ from earlier. Note that $\texttt{sgn}(\widehat{m}_n(x)) = \texttt{sgn}(\widehat{u}_n(x))$ if $\sum_{j=1}^n K(x, X_j) > 0$.

Observe that $\widehat{m}_n$ is interpolating by construction. Let μ_X denote the marginal distribution of X. The *L_1-error* of $\widehat{m}_n$ in approximating m is $J_n := \int_M |\widehat{m}_n(x) - m(x)|\mu_X(dx)$. For $M = \mathbb{R}^d$ and the *Hilbert kernel* defined by $K^{\mathcal{H}}_{\mathbb{R}^d}(x, z) := \|x - z\|^{-d}$, Devroye et al. [DGK98] proved L_1-consistency for *regression*: $J_n \to 0$ in probability when Y is bounded and X is continuously distributed.

Our contributions. Our primary contribution is to demonstrate an ensemble method with the consistent-interpolating property. Toward this end, in Section 6.3, we introduce the manifold-Hilbert kernel $K^{\mathcal{H}}_M$ on a Riemannian manifold M. When show that when M is complete, connected, and smooth, kernel smoothing regression with $K^{\mathcal{H}}_M$ has the same consistency guarantee (Theorem VI.4) as $K^{\mathcal{H}}_{\mathbb{R}^d}$ mentioned in the preceding paragraph. In Section 6.5, we consider the case when $M = \mathbb{S}^d$, and show that the manifold-Hilbert kernel $K^{\mathcal{H}}_{\mathbb{S}^d}$ is a weighted random partition kernel (Proposition VI.15).

Devroye et al. [DGK98, Section 7] observed that the L_1-consistency of $\widehat{m}_n$ for regression implies the consistency for classification of $\texttt{sgn} \circ \widehat{u}_n$. Furthermore, $\widehat{m}_n$ is interpolating for regression implies that $\texttt{sgn} \circ \widehat{u}_n$ is interpolating for classification. These observations together with our results demonstrate the existence of a weighted infinite-ensemble classifier with the interpolating-consistent property.

6.1.3 Related work

Kernel regression. Kernel smoothing regression, or simply kernel regression, is an interpolator when the kernel used is singular, a fact known to Shepard [She68] in 1968. Devroye et al. [DGK98] showed that kernel regression with the Hilbert kernel is

interpolating and weakly consistent for data with a density and bounded labels. Using singular kernels with compact support, Belkin et al. [BRT19] showed that minimax optimality can be achieved under additional distributional assumptions.

Random forests. Wyner et al. [Wyn+17] proposed that interpolation may be a key mechanism for the success of random forests and gave a compelling intuitive rationale. Belkin et al. [Bel+19] studied empirically the double descent phenomenon in random forests by considering the generalization performance past the interpolation threshold. The PERT variant of random forests, introduced by Cutler et al. [CZ01], provably interpolates in 1-dimension. Belkin et al. [BHM18] pose as an interesting question whether the result of Cutler et al. [CZ01] extends to higher dimension. Many work have established consistency of random forest and its variants under different settings [Bre04; BDL08; SBV15]. However, none of these work addressed interpolation.

Boosting. For classification under the noiseless setting (i.e., the Bayes error is zero), AdaBoost is interpolating and consistent (see Freund et al. [FS12, first paragraph of Chapter 12]). However, this setting is too restrictive and the result does not answer if consistency is possible when fitting the noise. Bartlett et al. [BT07] proved that AdaBoost with early stopping is universally consistent, however without the interpolation guarantee. To the best of our knowledge, whether AdaBoost or any other variant of boosting can be interpolating and consistent remains open.

Random partition kernels. Breiman [Bre00] and Geurts et al. [GEW06] studied infinite ensembles of simplified variants of random forest and connections to certain kernels. Davies et al. [DG14] formalized this connection and coined the term *random partition kernel*. Scornet [Sco16] further developed the theory of random forest kernels and obtained upper bounds on the rate of convergence. However, it is not clear if these variants of random forests are interpolating.

Previously defined (unweighted) random partition kernels are bounded, and thus

cannot be singular. On the other hand, the manifold-Hilbert kernel is always singular. To bridge between ensemble methods and theory on interpolating kernel smoothing regression, we propose *weighted* random partitions (Definition VI.1), whose associated kernel (Equation 6.3) can be singular.

Learning on Riemannian manifolds. Strong consistency of a kernel-based classification method on manifolds has been established by Loubes et al. [LP08]. However, the result requires the kernel to be bounded and thus the method is not guaranteed to be interpolating. See Feragen et al. [FH16] for a review of theoretical results regarding kernels on Riemannian manifolds.

Beyond kernel methods, other classical methods for Euclidean data have been extended to Riemannian manifolds, e.g., regression [Tho13], classification [YZ20], and dimensionality reduction and clustering [ZZ04][Mar+22]. To the best of our knowledge, no previous works have demonstrated an interpolating-consistent classifiers on manifolds other than $\mathbb{R}^d$.

In many applications, the data naturally belong to a Riemannian manifold. Spherical data arise from a range of disciplines in natural sciences. See the influential textbook by Mardia et al. [MJ00, Ch.1§4]. For applications of the Grassmanian manifold in computer vision, see Jayasumana et al. [Jay+15] and the references therein. Topological data analysis [Was18] presents another interesting setting of manifold-valued data in the form of *persistence diagrams* [Ani+16; LY18].

6.2 Background on Riemannian Manifolds

We give an intuitive overview of the necessary concepts and results on Riemannian manifolds. A longer, more precise version of this overview is in the Section 6.6.1.

A smooth d-dimensional manifold M is a topological space that is locally diffeomorphic[2] to open subsets of $\mathbb{R}^d$. For simplicity, suppose that M is embedded in $\mathbb{R}^N$

[2] A diffeomorphism is a smooth bijection whose inverse is also smooth.

for some $N \geq d$, e.g., $\mathbb{S}^d \subseteq \mathbb{R}^{d+1}$. Let $x \in M$ be a point. The *tangent space* at x, denoted T_xM, is the set of vectors that is tangent to M at x. Since linear combinations of tangent vectors are also tangent, the tangent space T_xM is a vector space. Tangent vectors can also be viewed as the time derivative of smooth curves. In particular, let $x \in M$. If $\epsilon > 0$ is an open set and $\gamma : (-\epsilon, \epsilon) \to M$ is a smooth curve such that $\gamma(0) = x$, then $\frac{d\gamma}{dt}(0) \in T_xM$.

A *Riemannian metric* on M is a choice of inner product $\langle \cdot, \cdot \rangle_x$ on T_xM for each x such that $\langle \cdot, \cdot \rangle_x$ varies smoothly with x. Naturally, $\|z\|_x := \sqrt{\langle z, z \rangle_x}$ defines a norm on T_xM. The length of a piecewise smooth curve $\gamma : [a, b] \to M$ is defined by $\texttt{len}(\gamma) := \int_a^b \|\dot{\gamma}(t)\|_{\gamma(t)} dt$. Define $\texttt{dist}_M(x, \xi) := \inf\{\texttt{len}(\gamma) : \gamma \text{ is a piecewise smooth curve from } x \text{ to } \xi\}$, which is a metric on M in the sense of metric spaces (see Sakai [Sak96, Proposition 1.1]). For $x \in M$ and $r \in (0, \infty)$, the open metric ball centered at x of radius r is denoted $\texttt{B}_x(r, M) := \{\xi \in M : \texttt{dist}_M(x, \xi) < r\}$.

A curve $\gamma : [a, b] \to M$ is a geodesic if γ is *locally* distance minimizing and has constant speed, i.e., $\|\frac{d\gamma}{dt}(\tau)\|_{\gamma(\tau)}$ is constant. Now, suppose $x \in M$ and $v \in T_xM$ are such that there exists a geodesic $\gamma : [0, 1] \to M$ where $\gamma(0) = x$ and $\frac{d\gamma}{dt}(0) = v$. Define $\exp_x(v) := \gamma(1)$, the element reached by traveling along γ at time $= 1$. See Figure 6.1 for the case when $M = \mathbb{S}^2$.

For a fixed $x \in M$, the above function $\exp_x$, the *exponential map*, can be defined on an open subset of T_xM containing the origin. The Hopf-Rinow theorem ([Do 92, Ch. 8, Theorem 2.8]) states that if M is connected and complete with respect to the metric $\texttt{dist}_M$, then $\exp_x$ can be defined on all of T_xM.

6.3 The Manifold-Hilbert kernel

Throughout the remainder of this work, we assume that M is a complete, connected, and smooth Riemannian manifold of dimension d.

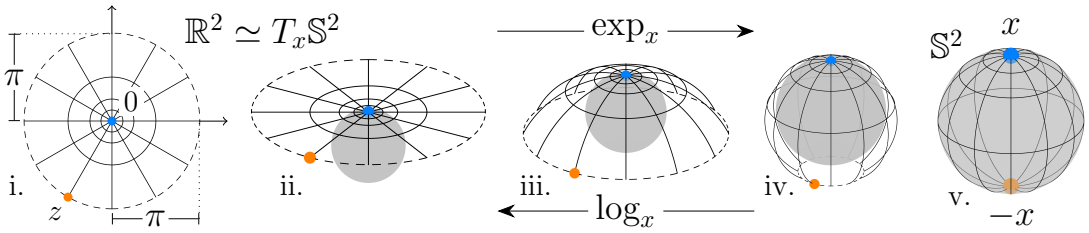

Figure 6.1: An illustration of the exponential map $\exp_x$ for the manifold $M = \mathbb{S}^2$, where x is the "northpole" (blue) and $-x$ the "southpole" (orange). The logarithm map $\log_x$, discussed in Section 6.4.1, is a right-inverse to $\exp_x$, i.e., $\exp_x \circ \log_x$ is the identity. *Panel i.* The tangent space $T_x\mathbb{S}^2$ visualized as $\mathbb{R}^2$. The dashed circle encloses a disc of radius π. *Panel ii.* The tangent space realized as the hyperplane tangent to sphere at x. *Panel iii-v.* Animation showing $\exp_x$ as a bijection from the open disc of radius π to $\mathbb{S}^2 \setminus \{-x\}$. The entire dashed circle in Panel i is mapped to $-x$ the southpole. Thus, $\log_x$ maps the southpole $-x$ to a point z on the dashed circle.

Definition VI.3. We define the *manifold-Hilbert kernel* $K_M^{\mathcal{H}} : M \times M \to [0, \infty]$ for each $x, \xi \in M$ by $K_M^{\mathcal{H}}(x, \xi) := \mathtt{dist}_M(x, \xi)^{-d}$ if $x \neq \xi$ and $K_M^{\mathcal{H}}(x, x) := \infty$ otherwise.

Let λ_M be the *Riemann–Lebesgue volume measure* of M. Integration with respect to this measure is denoted $\int_M f d\lambda_M$ for a function $f : M \to \mathbb{R}$. For details of the construction of λ_M, see Amann et al. [AE09, Proposition 1.5]. When $M = \mathbb{R}^d$, λ_M is the ordinary Lebesgue measure and $\int_{\mathbb{R}^d} f d\lambda_{\mathbb{R}^d}$ is the ordinary Lebesgue integral. For this case, we simply write λ instead of $\lambda_{\mathbb{R}^d}$.

We now state our first main result, a manifold theory extension of Devroye et al. [DGK98, Theorem 1].

Theorem VI.4. *Suppose that X has a density f_X with respect to λ_M and that Y is bounded. Let $P_{Y|X}$ be a conditional distribution of Y given X and $m_{Y|X}$ be its conditional expectation. Let $\widehat{m}_n(x) := \widehat{m}(x \| D^n, K_M^{\mathcal{H}})$. Then*

1. *at almost all $x \in M$ with $f_X(x) > 0$, we have $\widehat{m}_n(x) \to m_{Y|X}(x)$ in probability,*
2. *$J_n := \int_M |\widehat{m}_n(x) - m_{Y|X}(x)| f_X(x) d\lambda_M(x) \to 0$ in probability.*

In words, the kernel smoothing regression estimate $\widehat{m}_n$ based on the manifold-

Hilbert kernel is consistent and interpolates the training data, provided X has a density and Y is bounded. As a consequence, following the same logic as in Devroye et al. [DGK98], the associated classifier $\texttt{sgn} \circ \widehat{u}_n$ has the interpolating-consistent property. Before proving Theorem VI.4, we first review key concepts in probability theory on Riemannian manifolds.

6.3.1 Probability on Riemannian manifolds

Let $\mathcal{B}_M$ be the Borel σ-algebra of M, i.e., the smallest σ-algebra containing all open subsets of M. We recall the definition of M-valued random variables, following Pennec [Pen06, Definition 2]:

Definition VI.5. Let $(\Omega, \mathbb{P}, \mathcal{A})$ be a probability space with measure $\mathbb{P}$ and σ-algebra $\mathcal{A}$. A *M-valued random variable* X is a Borel-measurable function $\Omega \to M$, i.e., $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}_M$.

Definition VI.6 (Density)**.** A random variable X taking values in M has a *density* if there exists a nonnegative Borel-measurable function $f : M \to [0, \infty]$ such that for all Borel sets B in M, we have $\Pr(X \in B) = \int_B f d\lambda_M$. The function f is said to be a *probability density function* (PDF) of X.

Next, we recall the definition of conditional distributions, following Dudley [Dud18, Ch. 10 §2]:

Definition VI.7 (Conditional distribution[3])**.** Let (X, Y) be a random variable jointly distributed on $M \times \mathbb{R}$. Let $P_X(\cdot)$ be the probability measure corresponding to the marginal distribution of X. A *conditional distribution* for Y given X is a collection of probability measures $P_{Y|X}(\cdot|x)$ on $\mathbb{R}$ indexed by $x \in M$ satisfying the following:

1. For all Borel sets $A \subseteq \mathbb{R}$, the function $M \ni x \mapsto P_{Y|X}(A|x) \in [0, 1]$ is Borel-measurable.

[3] also known as *disintegration measures* according to Chang et al. [CP97].

2. For all $A \subseteq \mathbb{R}$ and $B \subseteq M$ Borel sets, $\Pr(Y \in A, X \in B) = \int_B P_{Y|X}(A|x)P_X(dx)$.

The *conditional expectation*[4] is defined as $m_{Y|X}(x) := \int_{\mathbb{R}} yP_{Y|X}(dy|x)$.

The existence of a conditional probability for a joint distribution (X, Y) is guaranteed by Dudley [Dud18, Theorem 10.2.2]. When (X, Y) has a joint density f_{XY} and marginal density f_X, the above definition gives the classical formula $P_{Y|X}(A|x) = \int_A f_{XY}(x, y)/f_X(x)dy$ when $\infty > f_X(x) > 0$. See the first example in Dudley [Dud18, Ch. 10 §2].

6.3.2 Lebesgue points on manifolds

Devroye et al. [DGK98] proved Theorem VI.4 when $M = \mathbb{R}^d$ and, moreover, that part 1 holds for the so-called *Lebesgue points*, whose definition we now recall.

Definition VI.8. Let $f : M \to \mathbb{R}$ be an absolutely integrable function and $x \in M$. We say that x is a *Lebesgue point* of f if $f(x) = \lim_{r\to 0} \frac{1}{\lambda_M(\mathtt{B}_x(r,M))} \int_{\mathtt{B}_x(r,M)} f d\lambda_M$.

For an integrable function, the following result states that almost all points are its Lebesgue points. For the proof, see Fukuoka [Fuk06, Remark 2.4].

Theorem VI.9 (Lebesgue differentation)**.** *Let $f : M \to \mathbb{R}$ be an absolutely integrable function. Then there exists a set $A \subseteq M$ such that $\lambda_M(A) = 0$ and every $x \in M \setminus A$ is a Lebesgue point of f.*

Next, for the reader's convenience, we restate Devroye et al. [DGK98, Theorem 1], emphasizing the connection to Lebesgue points.

Theorem VI.10 (Devroye et al. [DGK98])**.** *Let $M = \mathbb{R}^d$ be the flat Euclidean space. Then Theorem VI.4 holds. Moreover, Part 1 holds for all x that is a Lebesgue point to both f_X and $m_{Y|X} \cdot f_X$.*

The above result will be used in our proof of Theorem VI.4 below.

[4]More often, the conditional expectation is denoted $\mathbb{E}[Y|X = x]$. However, our notation is more convenient for function composition and compatible with that of [DGK98].

6.4 Proof of Theorem VI.4

The focal point of the first subsection is Lemma VI.11 which shows the Borel measurability of extensions of the so-called Riemannian logarithm. The second subsection contains two key results regarding densities of M-valued random variables transformed by the Riemannian logarithm. The final subsection proves Theorem VI.4 leveraging results from the preceding two subsections.

6.4.1 The Riemannian logarithm

Throughout, x is assumed to be an arbitrary point of M. Let $U_xM = \{v \in T_xM : \|v\|_x = 1\} \subseteq T_xM$ denote the set of unit tangent vectors. Define a function $\tau_x : U_xM \to (0, \infty]$ as follows[5]:

$$\tau_x(u) := \sup\{t > 0 : t = \mathtt{dist}_M(x, \exp_x(tu))\}.$$

The *tangent cut locus* is the set $\tilde{C}_x \subseteq T_xM$ defined by $\tilde{C}_x := \{\tau_x(u)u : u \in U_xM,\, \tau_x(u) < \infty\}$. Note that it is possible for $\tau_x(u) = \infty$ for all $u \in U_xM$ in which case $\tilde{C}_x$ is empty. The *cut locus* is the set $C_x := \exp_x(\tilde{C}_x) \subseteq M$.

The *tangent interior set* is $\tilde{I}_x := \{tu : 0 \leq t < \tau_x(u), u \in U_xM\}$ and the *interior set* is the set $I_x := \exp_x(\tilde{I}_x)$. Finally, define $\tilde{D}_x := \tilde{I}_x \cup \tilde{C}_x$. Note that for each $z = tu \in \tilde{I}_x$, we have

$$\|z\|_x = t = \mathtt{dist}_M(x, \exp_x(tu)) = \mathtt{dist}_M(x, \exp_x(z)). \tag{6.6}$$

Consider the example where $M = \mathbb{S}^2$ as in Figure 6.1. Then $\tau_x(u) = \pi$ for all $u \in U_xM$. Thus, the tangent interior set $\tilde{I}_x = \mathtt{B}_0(\pi, \mathbb{R}^2)$, the open disc of radius π centered at the origin.

[5]Positivity of τ_x is asserted at Sakai [Sak96, eq. (4.1)]

When restricted to $\tilde{I}_x$, the exponential map $\exp_x|_{\tilde{I}_x} : \tilde{I}_x \to I_x$ is a diffeomorphism. Its functional inverse, denoted by $\log_x|_{I_x}$, is called the *Riemannian Logarithm* [BZA20; Zim17]. In previous works, $\log_x|_{I_x}$ is only defined from I_x to $\tilde{I}_x$. The next result shows that the domain of $\log_x|_{I_x} : I_x \to \tilde{I}_x$ can be extended to $\log_x : M \to \tilde{D}_x$ while remaining Borel-measurable.

Lemma VI.11. *For all $x \in M$, there exists a Borel measurable map $\log_x : M \to T_xM$ such that $\log_x(M) \subseteq \tilde{D}_x$ and $\exp_x \circ \log_x$ is the identity on M. Furthermore, for all $x, \xi \in M$, we have $\texttt{dist}_M(x, \xi) = \|\log_x(\xi)\|_x$.*

Proof sketch. The full proof of the lemma is provided in Section 6.6.2. Below, we illustrate the idea of the proof using the example when $M = \mathbb{S}^2$ as in Figure 6.1.

Let $x \in \mathbb{S}^2$ be the "northpole" (the blue point). The tangent cut locus $\tilde{C}_x$ is the dashed circle in the left panel of Figure 6.1. The exponential map $\exp_x$ is one-to-one on $\tilde{D}_x$ except on the dashed circle, which all gets mapped to $-x$, the "southpole" (the orange point). A consequence of the measurable selection theorem[6] is that $\log_x$ can be extended to be a Borel-measurable right inverse of $\exp_x$ by selecting z point on $\tilde{C}_x$ such that $\log_x(-x) = z$. □

6.4.2 Random variable transforms

In the previous subsection, we showed that $\log_x : M \to T_xM$ is Borel-measurable. Now, recall that T_xM is equipped with the inner product $\langle \cdot, \cdot \rangle_x$, i.e., the Riemannian metric. Below, for each $x \in M$ choose an orthonormal basis on T_xM with respect to $\langle \cdot, \cdot \rangle$. Then T_xM is isomorphic as an inner product space to $\mathbb{R}^d$ with the usual dot product.

Our first result of this subsection is a "change-of-variables formula" for computing the densities of M-valued random variables after the $\log_x$ transform. Recall that λ_M

[6] Kuratowski–Ryll-Nardzewski measurable selection theorem (see [BR07, Theorem 6.9.3])

is the Riemann-Lebesgue measure on M and λ is the ordinary Lebesgue measure on $\mathbb{R}^d = T_xM$.

Proposition VI.12. *Let $x \in M$ be fixed. There exists a Borel measurable function $\nu_x : M \to \mathbb{R}$ with the following properties:*

(i) Let X be a random variable on M with density f_X and let $Z := \log_x(X)$. Then Z is a random variable on T_xM with density $f_Z(z) := f_X(\exp_x(z)) \cdot \nu_x(\exp_x(z))$.

(ii) Let $f : M \to \mathbb{R}$ be an absolutely integrable function such that x is a Lebesgue point of f. Define $f : T_xM \to \mathbb{R}$ by $h(z) := f(\exp_x(z)) \cdot \nu_x(\exp_x(z))$. Then $0 \in T_xM$ is a Lebesgue point for h.

Proof sketch. The full proof of the proposition is in Section 6.6.3. The function ν_x is the Jacobian of the change-of-variables formula for integrating $\int_{\tilde{B}} f_Z d\lambda$ where $\tilde{B} \subseteq T_xM$ is a Borel subset. See Lemma VI.22 for the exact definition of ν_x. Part (i) is a simple consequence of this change-of-variables formula, which says that $\int_{\tilde{B}} f_Z d\lambda = \int_{\exp_x(\tilde{B})} h d\lambda_M$.

For part (ii), the key observations are that (a) $\nu_x(\exp_x(0)) = \nu_x(0) = 1$ and (b) the volumes of $\mathtt{B}_x(r, M)$ and $\mathtt{B}_0(r, T_xM)$ are equal as $r \to 0$. More precisely, $\lim_{r\to 0} \frac{\lambda_M(\mathtt{B}_x(r,M))}{\lambda(\mathtt{B}_0(r,T_xM))} = 1$. From these two observations, it is straightforward to directly verify Definition VI.8. □

Proposition VI.13. *Let (X, Y) have a joint distribution on $M \times \mathbb{R}$ such that the marginal of X has a density f_X on M. Let $P_{Y|X}(\cdot|\cdot)$ be a conditional distribution for Y given X. Let $x \in M$. Define $Z := \log_x(X)$ and consider the joint distribution (Z, Y) on $T_pM \times \mathbb{R}$. Then $P_{Y|Z}(\cdot|\cdot) := P_{Y|X}(\cdot|\exp_x(\cdot))$ is a conditional distribution for Y given Z. Consequently, $m_{Y|X} \circ \exp_x = m_{Y|Z}$.*

Proof sketch. The full proof of the Proposition is in Section 6.6.4. The idea is the same as in the proof of Proposition VI.12, except that the probability density f_Z is replaced by an appropriate conditional probability density. □

6.4.3 Finishing up the Proof of Theorem VI.4

Fix $x \in M$ such that x is a Lebesgue point of f_X and $m_{Y|X} \cdot f_X$. Note that by Theorem VI.9, almost all $x \in M$ has this property. Next, let $Z = \log_x(X)$ and f_Z be as in Proposition VI.12-(i). Then

1. $f_Z = (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x)$, and
2. $(m_{Y|X} \circ \exp_x) \cdot f_Z = (m_{Y|X} \circ \exp_x) \cdot (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x)$.

Now, proposition VI.12-(ii) implies that 0 is a Lebesgue point of both f_Z and $(m_{Y|X} \circ \exp_x) \cdot f_Z$. Furthermore, by Proposition VI.13, we have $m_{Y|X} \circ \exp_x = m_{Y|Z}$. Thus, 0 is a Lebesgue point of f_Z and $m_{Y|Z} \cdot f_Z$.

Now, let $D_n := \{(X_i, Y_i)\}_{i \in [n]}$. Define $Z_i := \log_x(X_i)$, which are i.i.d copies of the random variable $Z := \log_x(X)$, and let $\tilde{D}_n := \{(Z_i, Y_i)\}_{i \in [n]}$. Then we have

$$\begin{aligned}\widehat{m}(x \| D^n, K_M^{\mathcal{H}}) &\overset{\text{(a)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \mathtt{dist}_M(x, X_i)^{-d}}{\sum_{j=1}^n \mathtt{dist}_M(x, X_j)^{-d}} \overset{\text{(b)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \|Z_i\|_x^{-d}}{\sum_{j=1}^n \|Z_j\|_x^{-d}} \\ &\overset{\text{(c)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \mathtt{dist}_{\mathbb{R}^d}(0, Z_i)^{-d}}{\sum_{j=1}^n \mathtt{dist}_{\mathbb{R}^d}(0, Z_j)^{-d}} \overset{\text{(d)}}{=} \widehat{m}(0 \| \tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}})\end{aligned}$$

where equations marked by (a) and (d) follow from Equation (6.5), (b) from Lemma VI.11, and (c) from the fact that the inner product space T_xM with $\langle \cdot, \cdot \rangle_x$ is isomorphic to $\mathbb{R}^d$ with the usual dot product. By Theorem VI.10, we have $\widehat{m}(0 \| \tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) \to m_{Y|Z}(0)$ in probability. In other words, for all $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr\{|\widehat{m}(0 \| \tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) - m_{Y|Z}(0)| > \epsilon\} = 0.$$

By Proposition VI.13, we have $m_{Y|Z}(0) = m_{Y|Z}(\exp_x(0)) = m_{Y|Z}(x)$. Therefore,

$$\left\{|\widehat{m}(0 \| \tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) - m_{Y|Z}(0)| > \epsilon\right\} = \left\{|\widehat{m}(x \| D^n, K_M^{\mathcal{H}}) - m_{Y|X}(x)| > \epsilon\right\}$$

as events. Thus, $\widehat{m}(x \| D^n, K_M^{\mathcal{H}}) \to m_{Y|X}(x)$ converges in probability, proving Theo-

rem VI.4 part 1. As noted in Devroye et al. [DGK98, §2], part 2 of Theorem VI.4 is an immediate consequence of part 1.

6.5 Application to the d-Sphere

The d-dimensional round sphere is $\mathbb{S}^d := \{x \in \mathbb{R}^{d+1} : x_1^2 + \cdots + x_{d+1}^2 = 1\}$. Here, a *round* sphere assumes that $\mathbb{S}^d$ has the *arc-length metric*:

$$\texttt{dist}_{\mathbb{S}^d}(x, z) = \angle(x, z) = \cos^{-1}(x^\top z) \in [0, \pi]. \tag{6.7}$$

Let $\mathcal{S}$ be a set and $\sigma : M \to \mathcal{S}$ be a function. The *partition induced* by σ is defined by $\{\sigma^{-1}(s) : s \in \texttt{Range}(\sigma)\}$. For example, when $M = \mathbb{S}^d$ and $W \in \mathbb{R}^{(d+1)\times h}$, then the function $\sigma_W : \mathbb{S}^d \to \{\pm 1\}^h$ defined by $\sigma_W(x) = \texttt{sgn}(W^\top x)$ induces a hyperplane arrangement partition.

Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denote the positive and non-negative integers.

Definition VI.14 (Random hyperplane arrangement partition)**.** Let $d \in \mathbb{N}$ and $M = \mathbb{S}^d$. Let $q < 0$ be a negative number, and let H be a random variable with probability mass function $p_H : \mathbb{N}_0 \to [0, 1]$ such that $p_H(h) > 0$ for all h. Define the following weighted random partition $\mathsf{ha} := (\Theta, \mathfrak{P}, \alpha)$:

1. The parameter space $\Theta = \bigsqcup_{h=0}^{\infty} \mathbb{R}^{(d+1)\times h}$ is the disjoint union of all $(d+1) \times h$ matrices. Element of Θ are matrices $\theta = W \in \mathbb{R}^{(d+1)\times h}$ where the number of columns $h \in \{0, 1, 2, \dots\}$ varies. By convention, if $h = 0$, the partition $\mathcal{P}_\theta = \mathcal{P}_W$ is the trivial partition $\{\mathbb{S}^d\}$. If $h > 0$, $\mathcal{P}_W$ is the partition induced by $x \mapsto \texttt{sgn}(W^\top x)$.
2. The probability $\mathfrak{P}$ is constructed by the procedure where we first sample $h \sim p_H(h)$, then sample the entries of $W \in \mathbb{R}^{d\times h}$ i.i.d according to $\texttt{Gaussian}(0, 1)$.
3. For $\theta \in \Theta$, define $\alpha(\theta) := \pi^q p_H(h)^{-1}(-1)^h \binom{q}{h}$, where $\binom{q}{h} := \frac{1}{h!}\prod_{j=0}^{h-1}(q - j)$.

Note that $(-1)^h \binom{q}{h} = \frac{1}{h!}\prod_{j=0}^{h-1}(q - j) > 0$ when $q < 0$.

Theorem VI.15. *Let* $\mathsf{ha} = (\Theta, \mathfrak{P}, \alpha)$ *be as in Definition VI.14. Then*

$$K^{\mathsf{ha}}_{\mathbb{S}^d}(x,z) = \begin{cases} \angle(x,z)^q & : \angle(x,z) \neq 0 \\ +\infty & : \textit{otherwise.} \end{cases}$$

When $q = -d$*, we have* $K^{\mathsf{ha}}_{\mathbb{S}^d} = K^{\mathcal{H}}_{\mathbb{S}^d}$ *where the right hand side is the manifold-Hilbert kernel.*

Proof of Theorem VI.15. Before proceeding, we have the following useful lemma:

Lemma VI.16. *Let* $\mathsf{rp} = (\Theta, \mathfrak{P}, \alpha)$ *be a WRP. Let* H *be a random variable. Let* $\theta \sim \mathfrak{P}$*. Suppose that for all* $x, z \in M$*, the random variables* $\alpha(\theta)$ *and* $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}$ *are conditionally independent given* H*. Then we have* $K^{\mathsf{rp}}_M(x,z) = \mathbb{E}_H\Big[\overline{\alpha}(H) \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H]\Big]$ *where* $\overline{\alpha}(h) := \mathbb{E}_{\theta\in\mathfrak{P}}\left[\alpha(\theta)|H = h\right]$ *for a realization* h *of* H*.*

The lemma follows immediately from the Definition of $K^{\mathsf{rp}}_M(x,z)$ in Equation 6.3 and the conditional independence assumption. Now, we proceed with the proof of Theorem VI.15.

Let $\phi := \angle(x,z)/\pi$. Let $H \sim p_H$ and $\theta \sim \mathfrak{P}$ be the random variables in Definition VI.14. Note that by construction, the following condition is satisfied: for all $x, z \in M$, the random variables $\alpha(\theta)$ and $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}$ are conditionally independent given H. In fact, $\alpha(\theta) = \pi^q p_H(h)^{-1}(-1)^h\binom{q}{h}$ is constant given $H = h$. Hence, applying Lemma VI.16, we have

$$\begin{aligned} K^{\mathsf{ha}}_{\mathbb{S}^d}(x,z) &= \mathbb{E}_H\Big[\overline{\alpha}(H) \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H]\Big] \\ &= \sum_{h=0}^{\infty} \pi^q(-1)^h\binom{q}{h} \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H = h] \\ &= \sum_{h=0}^{\infty} \pi^q(-1)^h\binom{q}{h} \cdot \Pr\{x \in \mathcal{P}_\theta[z]|H = h\}. \end{aligned}$$

Next, we claim that $\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} = (1-\phi)^h$. When $h = 0$, $x \in \mathcal{P}_\theta[z]$ is

always true since $\mathcal{P}_\theta = \{\mathbb{S}^d\}$ is the trivial partition. In this case, we have $\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} = 1 = (1-\phi)^0$. When $h > 0$, we recall a result of Pinelis [Pin19]:

Lemma VI.17. *Let $x, z \in \mathbb{S}^d$. Let $w \in \mathbb{R}^{d+1}$ be a random vector whose entries are sampled i.i.d according to* $\texttt{Gaussian}(0,1)$. *Then* $\Pr\{\texttt{sgn}(w^\top x) = \texttt{sgn}(w^\top z)\} = 1 - (\angle(x,z)/\pi)$.

Let $W = [w_1, \ldots, w_h]$ be as in Definition VI.14 where w_j denotes the j-th column of W. Then by construction, w_j is distributed identically as w in Lemma VI.17. Furthermore, w_j and $w_{j'}$ are independent for $j, j' \in [h]$ where $j \neq j'$. Thus, the claim follows from

$$\begin{aligned}\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} &\overset{\text{(a)}}{=} \Pr\{\texttt{sgn}(W^\top x) = \texttt{sgn}(W^\top z)|H = h\} \\ &\overset{\text{(b)}}{=} \prod_{j=1}^{h} \Pr\{\texttt{sgn}(w_j^\top x) = \texttt{sgn}(w_j^\top z)\} \overset{\text{(c)}}{=} \prod_{j=1}^{h} (1-\phi) = (1-\phi)^h.\end{aligned}$$

where equality (a) follows from Definition VI.14, (b) from $W \in \mathbb{R}^{(d+1)\times h}$ having i.i.d standard Gaussian entries given $H = h$, and (c) from Lemma VI.17. Putting it all together, we have

$$K^{\texttt{part}}_{\mathfrak{P},\alpha}(x,z) = \sum_{h=0}^{\infty} \pi^q (-1)^h \binom{q}{h} (1-\phi)^h = \pi^q \sum_{h=0}^{\infty} \binom{q}{h} (\phi - 1)^h = \angle(x,z)^q.$$

For the last step, we used the fact that for all $q \in \mathbb{R}$ the binomial series $(1+t)^q = \sum_{h=0}^{\infty} \binom{q}{h} t^h$ converges absolutely for $|t| < 1$ (when $\phi \in (0,1]$) and diverges to $+\infty$ for $t = -1$ (when $\phi = 0$). □

Corollary VI.18. *Let $q := -d$ and $K^{\mathsf{ha}}_{\mathbb{S}^d}$ be as in Theorem VI.15. The infinite-ensemble classifier $\texttt{sgn}(\widehat{u}(\bullet \| D^n, K^{\mathsf{ha}}_{\mathbb{S}^d}))$ (see Equation 6.4 for definition) has the interpolating-consistent property.*

Proof. As observed in Devroye et al. [DGK98, Section 7], for an arbitrary kernel K,

the L_1-consistency of $\widehat{m}(\bullet \| D^n, K)$ for regression implies the consistency for classification of $\mathtt{sgn}(\widehat{u}(\bullet \| D^n, K))$. Furthermore, $\widehat{m}(\bullet \| D^n, K)$ is interpolating for regression implies that $\mathtt{sgn}(\widehat{u}(\bullet \| D^n, K))$ is interpolating for classification. While the argument there is presented in the $\mathbb{R}^d$ case, the argument holds in the more general manifold case *mutatis mutandis*.

Thus, by Theorem VI.4, we have $\mathtt{sgn}(\widehat{u}(\bullet \| D^n, K^{\mathcal{H}}_{\mathbb{S}^d}))$ is consistent for classification, i.e., Equation (6.1) holds. It is also interpolating since $\widehat{m}(\bullet \| D^n, K)$ is interpolating. By Proposition VI.15, we have $K^{\mathsf{ha}}_{\mathbb{S}^d} = K^{\mathcal{H}}_{\mathbb{S}^d}$. Thus $\mathtt{sgn}(\widehat{u}(\bullet \| D^n, K^{\mathsf{ha}}_{\mathbb{S}^d}))$ is an ensemble method having the interpolating-consistent property. □

6.6 Discussion

We have shown that using the manifold-Hilbert kernel in kernel smoothing regression, also known as Nadaraya-Watson regression, results in a consistent estimator that interpolates the training data on a Riemannian manifold M. Furthermore, when $M = \mathbb{S}^d$ is the sphere, we showed that the manifold-Hilbert kernel is a weighted random partition kernel, where the random partitions are induced by random hyperplane arrangements. This demonstrates an ensemble method that has the interpolating-consistent property.

A limitation of this work is that the random hyperplane arrangement partition is data-independent. Thus, the resulting ensemble method considered in this work are easier to analyze than popular ensemble methods used in practice. Nevertheless, we believe our work offers one theoretical basis towards understanding generalization in the interpolation regime of ensembles of histogram classifiers over data-dependent partitions, e.g., decision trees à la CART [Bre+84].

6.6.1 Basics of Riemannian Manifolds

In this section, we review the main concepts from Riemannian manifold theory essential to this work. Our main references are Sakai [Sak96] and Do Carmo [Do 92]. Throughout, $d \in \mathbb{N}$ denotes the dimension. We use the word *smooth* to mean infinitely differentiable.

Manifolds. A smooth *manifold* M of dimension d is a Hausdorff, second countable topological space together with an *atlas*: a set $\texttt{Atlas} := \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in A}$ where 1). $\{U_\alpha\}_{\alpha \in A}$ is an open cover of M, 2). for each $\alpha \in A$, $\varphi_\alpha : U_\alpha \to \varphi_\alpha(U_\alpha) \subseteq \mathbb{R}^d$ is a homeomorphism onto its image, and 3). $\varphi_\alpha \circ \varphi_\beta^{-1} : \varphi_\beta(U_\alpha \cap U_\beta) \to \varphi_\alpha^{-1}(U_\alpha \cap U_\beta)$ is smooth for each pair $\alpha, \beta \in A$. An element (U, φ) of $\texttt{Atlas}$ is called a *chart*.

Smooth maps. A real-valued function $f : M \to \mathbb{R}$ is a *smooth function* if $f \circ \varphi^{-1}$ is smooth (in the elementary calculus sense) for all charts (U, φ). The set of all smooth functions is denoted $\texttt{Fn}(M)$, which forms an $\mathbb{R}$-vectorspace. Let N be another smooth manifold with atlas $\mathcal{B}$. A function $\Phi : M \to N$ is a *smooth map* if $g \circ \Phi \in \texttt{Fn}(M)$ for all $g \in \texttt{Fn}(N)$.

Tangent space. Let $x \in M$. A *derivation at* x is a linear function $v : \texttt{Fn}(M) \to \mathbb{R}$ satisfying the *product rule*: $v[fg] = f(x)v[f] + g(x)v[g]$ for all $f, g \in \texttt{Fn}(M)$. The *tangent space at* x, denoted T_xM, is the vector space of all derivations at x. Elements of T_xM are referred to as *tangent vectors* at x. For a given chart (U, φ) where $x \in U$, define a derivation at x, denoted $\partial_i x$, by $f \mapsto \frac{d(f \circ \varphi^{-1})}{dz_i}(\varphi(x))$ where $\frac{d}{dz_i}$ is the i-th partial derivative in ordinary calculus. It is a fact that $\{\partial_i x : i = 1, \ldots, d\}$ is a basis for T_xM.

Although the above definition of a tangent vector is abstract, it can be concretely interpreted in terms of derivative along a curve. Let $a < t_0 < b$ be real numbers. A *curve through* x is a smooth map $\gamma : (a, b) \to M$ such that $\gamma(t_0) = x$. Then $\texttt{Fn}(M) \ni f \mapsto \frac{d}{dt} f(\gamma(t))|_{t=t_0} \in \mathbb{R}$ defines a derivation at x. Oftentimes, this derivation is denoted $\dot{\gamma}(t_0) \in T_xM$

Riemannian metric. The *tangent bundle* is the set $TM := \bigcup_x T_xM$, which itself is a smooth manifold of dimension $2d$. A *vector field on* M is a smooth map $\mathsf{V} : M \to TM$ such that $\mathsf{V}(x) \in T_xM$ for all $x \in M$. The set of all vectors fields on M is denoted $\mathtt{Vf}(M)$.

A *Riemannian metric* on M is a choice of an inner product $\langle \cdot, \cdot \rangle_x$ (and thus, a norm $\| \cdot \|_x$) on T_xM for each $x \in M$ such that the function $M \to \mathbb{R}$ given by $x \mapsto \langle \mathsf{V}(x), \mathsf{U}(x) \rangle_x$ is smooth for all $\mathsf{V}, \mathsf{U} \in \mathtt{Vf}(M)$. As shorthands, when x is clear from context, we drop the subscripts and simply write $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ instead. Choosing an orthonormal basis for T_xM with respect to $\langle \cdot, \cdot \rangle_x$ for each x, we can identify T_xM with $\mathbb{R}^d$ with the ordinary dot inner product.

Let $x \in M$ and (U, φ) be a chart such that $x \in U$. Define $g_{ij}(x) = \langle \partial_i x, \partial_j x \rangle_x$. Denote by $G(x)$ the $d \times d$ positive definite matrix $[g_{ij}(x)]_{ij}$. Below, we will refer to the function $G : U \to \mathbb{R}^{d \times d}$ as the *coordinate representation* of the Riemannian metric. Define $g^{ij}(x) := [G(x)^{-1}]_{ij}$. The *Christoffel symbols* with respect to (U, φ) are defined by $\Gamma^k_{ij} := \frac{1}{2} \sum_{\ell=1}^d g^{k\ell}(\partial_i x g_{j\ell} + \partial_j x g_{i\ell} - \partial_\ell x g_{ij})$. Note that $g_{k\ell}$, $g^{k\ell}$, G, Γ^k_{ij}, and $\partial_i x g_{j\ell}$ are all functions with domain U.

Geodesics. Fix a chart (U, φ). Consider a smooth curve $\gamma : [a, b] \to U$. Let $\zeta_i(t) := [\varphi(\gamma(t))]_i$ be the i-th component functions. The curve γ is a *geodesic* if ζ is a solution to the following system of second order ordinary differential equations (ODEs): $\frac{d^2 \zeta_i}{dt^2} + \sum_{j,\ell=1}^d \Gamma^i_{j\ell} \circ \gamma \frac{d\zeta_j}{dt} \frac{d\zeta_\ell}{dt} = 0$ for all $i = 1, \ldots, d$ at all time $t \in [a, b]$.

Geodesics are minimizers of the so-called *energy functional* $E(\gamma) = \frac{1}{2} \int_a^b \|\dot{\gamma}(t)\|^2_{\gamma(t)} dt$. The above system of ODEs are the analog of the "first derivative test" for local minimizers of E. Thus, geodesics are defined independently of the choice of the chart.

Exponential map. For $x \in M$ and $v \in T_xM$, there exists $\epsilon > 0$ and a unique geodesic curve $\gamma_v : [-\epsilon, \epsilon] \to M$ such that $\gamma_v(0) = x$ and $\dot{\gamma}_v(0) = v$. This follows from the existence and uniqueness of the solution to an ODE given initial conditions where

the ODE is as discussed above. Note that although geodesics are previously defined in U where (U, φ) is a chart, they can be extended outside of U using additional charts.

Let $x \in M$ and $v \in T_xM$ be fixed and let $\gamma_v : [-\epsilon, \epsilon] \to M$ be as in the preceding paragraph. If $\|v\|_x \leq \epsilon$, then define $\exp_x(v) := \gamma_v(1)$. A fundamental fact is that $\exp_x$, known as the *exponential map at* x, can be defined on an open set of T_xM containing the origin.

Distance function. Let $x, \xi \in M$ and $a < b$ be real numbers. A *piecewise smooth curve* from x to ξ is a piecewise smooth map $\gamma : [a, b] \to M$ such that $\gamma(a) = x$ and $\gamma(b) = \xi$. Assume that M is connected. Then for all $x, \xi \in M$, there exists a piecewise smooth curve from x to ξ. The *length of* γ is defined as $\mathtt{len}(\gamma) := \int_a^b \|\dot{\gamma}(t)\|_{\gamma(t)} dt$. Define $\mathtt{dist}_M(x, \xi) := \inf\{\mathtt{len}(\gamma) : \gamma \text{ is a piecewise smooth curve from } x \text{ to } \xi\}$, which is a metric on M in the sense of metric spaces (see [Sak96, Proposition 1.1]). For $x \in M$ and $r \in (0, \infty)$, the open ball centered at x of radius r is denoted $\mathtt{B}_x(r, M) := \{z \in M : \mathtt{dist}_M(x, z) < r\}$.

Complete Riemannian manifolds. A Riemannian manifold is *complete* if it is a complete metric space under the metric $\mathtt{dist}_M$. The Hopf-Rinow theorem ([Do 92, Ch. 8, Theorem 2.8]) states that if M is connected and complete, then the exponential $\exp_x$ can be defined on the entire T_xM.

6.6.2 Proof of Lemma VI.11

This section uses definitions and notations introduced in Section 6.4.1. In particular, recall the cut locus C_x, the tangent cut locus $\tilde{C}_x$, the interior set I_x and the tangent interior set $\tilde{I}_x$. The proof of Lemma VI.11 is presented towards the end of the section. At this point, we compile some facts from various sources about the cut locus.

Lemma VI.19. *For all $x \in M$, we have*

1. *C_x is a closed subset of M (Hebda [Heb87, Proposition 1.2]).*
2. *$I_x \cap C_x = \emptyset$ and $I_x \cup C_x = M$ (Sakai [Sak96, Ch II, Lemma 4.4 (1)])*
3. *I_x is an open subset of M (immediate from 1 and 2 above)*
4. *$\exp_x : \tilde{I}_x \to I_x$ is a diffeomorphism ([Sak96, Ch II, Lemma 4.4 (2)])*
5. *$\lambda_M(C_x) = 0$, where λ_M is the Riemann-Lebesgue measure ([Sak96, Lemma 4.4 (3)])*
6. *τ_x is continuous and $\inf_{u \in U_x M} \tau_x(u) > 0$ ([Sak96, Ch II, Propositions 4.1 (2) and 4.13 (1)])*

While the following lemma is elementary, we provide a proof since we could not find one in the literature.

Lemma VI.20. *For all $x \in M$, the (topological) closure of $\tilde{I}_x$ in T_xM is $\tilde{D}_x$. Furthermore, for all $x \in M$, we have $\exp_x(\tilde{D}_x) = M$.*

Proof of Lemma VI.20. Take a convergent sequence $\{t_i u_i\}_{i \in \mathbb{N}} \subseteq \tilde{I}_x$ where $u_i \in U_x M$ and $0 \leq t_i < \tau_x(u_i)$. Let $v^* = \lim_i t_i u_i$. Our goal is to show that $v^* \in \tilde{D}_x = \tilde{I}_x \cup \tilde{C}_x$.

Since $U_x M$ is compact, we may assume that $u^* := \lim_i u_i$ exists after passing to a subsequence if necessary. Furthermore, $\|t_i u_i\|_x = t_i$ implies that $t^* := \lim_i t_i$ exists as well (i.e., $t^* < \infty$). Hence, $v^* = t^* u^*$.

Consider the case that $\tau_x(u^*) = \infty$. Then $0 \leq t^* < \tau_x(u^*)$ implies that $v^* = t^* u^* \in \tilde{I}_x$. For the other case that $t(u) < \infty$, we first note that $t_i u_i \in \tilde{I}_x$ implies that $t_i < \tau_x(u_i)$. Taking the limit of both sides, we have $t^* = \lim_i t_i \leq \lim_i \tau_x(u_i) = \tau_x(u^*)$. Note that the last limit can be exchanged since τ_x is continuous (Lemma VI.19 part 6). Thus, either $t^* < \tau_x(u^*)$ in which case $v^* \in \tilde{I}_x$, or $t^* = \tau_x(u^*)$ in which case $v^* = \tau_x(u^*) u^* \in \tilde{C}_x$.

For the "furthermore" part, note that

$$\exp_x(\tilde{D}_x) = \exp_x(\tilde{I}_x \cup \tilde{C}_x) = \exp_x(\tilde{I}_x) \cup \exp_x(\tilde{C}_x) = I_x \cup C_x = M$$

where the last equality is Lemma VI.19 part 2. □

Proof of Lemma VI.11. Denote by $\mathrm{cl}(T_xM)$ the set of closed subsets of T_xM. Define $\psi : M \to \mathrm{cl}(T_xM)$ by $\psi(\xi) := \{x \in \tilde{D}_x : \exp_x(x) = \xi\} = \exp_x^{-1}(\xi) \cap \tilde{D}_x$. Note that $\psi(\xi)$ is a closed set by Lemma VI.20.

We claim that ψ is *weakly-measurable*, i.e., for every open set $\tilde{U} \subseteq T_xM$, the subset of M defined by $\{\xi \in M : \psi(\xi) \cap \tilde{U} \neq \emptyset\}$ is Borel. To see this, note that

$$\begin{aligned}
&\{\xi \in M : \psi(\xi) \cap \tilde{U} \neq \emptyset\} \\
&= \{\xi \in M : \exp_x^{-1}(\xi) \cap \tilde{D}_x \cap \tilde{U} \neq \emptyset\} \\
&= \{\xi \in M : \exp_x(\tilde{D}_x \cap \tilde{U}) \ni \xi\} \\
&= \exp_x(\tilde{D}_x \cap \tilde{U}).
\end{aligned}$$

As inner product spaces, T_xM and $\mathbb{R}^d$ are isomorphic (see Section 6.6.1-Riemannian metric). Since, T_xM and $\mathbb{R}^d$ are homeomorphic as topological spaces, $\mathbb{R}^d$ being locally compact implies T_xM is locally compact as well. Thus, we can write $\tilde{U} = \bigcup_{i\in\mathbb{N}} \tilde{K}_i$ as a countable union of compact sets $\tilde{K}_i \subseteq T_xM$. Furthermore, $\tilde{D}_x \cap \tilde{U} = \bigcup_{i\in\mathbb{N}} \tilde{D}_x \cap \tilde{K}_i$ and so $\exp_x(\tilde{D}_x \cap \tilde{U}) = \bigcup_{i\in\mathbb{N}} \exp_x(\tilde{D}_x \cap \tilde{K}_i)$.

Since $\exp_x$ is continuous, $\exp_x(\tilde{D}_x \cap \tilde{K}_i)$ is a compact subset of M, and hence closed and bounded by the Hopf-Rinow theorem ([Do 92, Ch. 8, Theorem 2.8]). Thus, $\exp_x(\tilde{D}_x \cap \tilde{U}) = \bigcup_{i\in\mathbb{N}} \exp_x(\tilde{D}_x \cap \tilde{K}_i)$ is a countable union of closed sets, which is Borel. This proves the claim that ψ is weakly Borel measurable.

By the Kuratowski–Ryll-Nardzewski measurable selection theorem (see [BR07, Theorem 6.9.3]), there exists a Borel measurable function $M \to T_xM$, which we denote by $\log_x$, such that $\log_x(\xi) \in \psi(\xi) = \exp_x^{-1}(\xi)$ for all $\xi \in M$, as desired. By construction, $\log_x(\xi) \in \exp_x^{-1}(\xi)$ for all $\xi \in M$, and so $\exp_x(\log_x(\xi)) = \xi$ is immediate.

For the "furthermore" part, let $\xi \in M$ be arbitrary and let $z := \log_x(\xi) \in \tilde{D}_x$. Let $\{z_i\} \subseteq \tilde{I}_x$ be a sequence such that $\lim_i z_i = z$. By Equation (6.6), we

have $\mathtt{dist}_M(x, \exp_x(z_i)) = \|z_i\|_x$. By continuity of $\mathtt{dist}_M$ and $\exp_x$, we have $\mathtt{dist}_M(x, \xi) = \mathtt{dist}_M(x, \exp_x(z)) = \lim_i \mathtt{dist}_M(x, \exp_x(z_i))$. To conclude, we have $\lim_i \mathtt{dist}_M(x, \exp_x(z_i)) = \lim_i \|z_i\|_x = \|z\|_x = \|\log_x(\xi)\|_x$, as desired. □

6.6.3 Proof of Proposition VI.12

Recall from Section 6.6.1-Riemannian metric, given a chart (U, φ), one can define the matrix-valued function $G : U \to \mathbb{R}^{d\times d}$ referred to earlier as the coordinate representation of the Riemannian metric. Now, Lemma VI.19 part 3 states that I_x is an open neighborhood of x. Furthermore, $\tilde{I}_x$ is an open subset of T_xM, which is identified with $\mathbb{R}^d$ using an orthonormal basis (see Section 6.6.1-Riemannian metric). Hence, $\{(I_x, \log_x|_{I_x})\}_{x\in M}$ is an atlas of M (see Section 6.6.1-Manifolds).

Definition VI.21. The chart $(I_x, \log_x|_{I_x})$ is called a *normal coordinate system* at x. Let $G : I_x \to \mathbb{R}^{d\times d}$ be the coordinate representation of the Riemannian metric for this chart. To emphasize the dependency on x, we write $G_x := G$. Denote by $G_x^\perp : M \to \mathbb{R}^{d\times d}$ the zero extension of G_x to the rest of M, i.e., $G_x^\perp(\xi) = G_x(\xi)$ for $\xi \in I_x$ and $G_x^\perp(\xi)$ is the zero matrix for $\xi \notin I_x$.

The normal coordinate system has the property that $G_x(x) = G_x^\perp(x)$ is the identity matrix. This is the result of Sakai [Sak96, Ch. II §2 Exercise 4].

Lemma VI.22 (Change-of-Variables)**.** *Let $x \in M$ be fixed. Define the function $\nu_x : M \to \mathbb{R}$ by $\nu_x(\xi) = \sqrt{|\det G_x^\perp(\xi)|}$ where $G_x^\perp$ is as in Definition VI.21. Then ν_x is Borel-measurable. Furthermore, ν_x satisfies the following property: Let $f : M \to \mathbb{R}$ be an absolutely integrable function. Define the function*

$$h : T_xM \to \mathbb{R} \quad by \quad h(z) := f(\exp_x(z)) \cdot \nu_x(\exp_x(z)).$$

Then (i) $h(0) = f(x)$ *and* (ii) *for all Borel set $\tilde{B} \subseteq T_xM$ we have $\int_B f d\lambda_M = \int_{\tilde{B}} h d\lambda$ where $B := \exp_x(\tilde{B} \cap \tilde{I}_x)$.*

Proof of Lemma VI.22. We first show that ν_x is Borel-measurable. Recall that $G_x^\perp : M \to \mathbb{R}^{d\times d}$ is the zero extension of $G_x : I_x \to \mathbb{R}$, which is by definition smooth (see Section 6.6.1-Riemannian metric). In particular, $G_x : I_x \to \mathbb{R}$ is continuous and so $\sqrt{\det(G_x(\bullet))}$ is Borel-measurable. Now, note that $\sqrt{\det(G_x^\perp(\bullet))}$ is the zero extension of $\sqrt{\det(G_x(\bullet))}$ from I_x to M. Hence, $\sqrt{\det(G_x^\perp(\bullet))}$, which is ν_x by definition, is Borel-measurable.

Next, we prove the "Furthermore" part (i). Note that $\exp_x(0) = x$. Moreover, $G_x^\perp(x) = G_x(x)$ is the identity matrix as asserted after Definition VI.21 (see Sakai [Sak96, Ch. II §2 Exercise 4]). Thus, $h(0) = f(\exp_x(0))\sqrt{|\det G_x^\perp(\exp_x(0))|} = f(x)\sqrt{1} = f(x)$, as desired.

For the "Furthermore" part (ii), we first note that $\tilde{B} = (\tilde{B} \cap \tilde{I}_x) \cup (\tilde{B} \cap \tilde{C}_x)$ expresses $\tilde{B}$ as a disjoint union. Thus, $B = \exp_x(\tilde{B}) = \exp_x(\tilde{B} \cap \tilde{I}_x) \cup \exp(\tilde{B} \cap \tilde{C}_x)$ expresses B as a disjoint union as well. Moreover, $\exp(\tilde{B} \cap \tilde{C}_x) \subseteq \exp(\tilde{C}_x) = C_x$, which has λ_M-measure zero (Lemma VI.19 part 5).

Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition VI.3). Now, we directly compute to obtain the formula

$$\begin{aligned}
\int_{\tilde{B}} h d\lambda &= \int_{\tilde{B}\cap\tilde{I}_x} f \circ \exp_x \sqrt{|\det(G_x^\perp \circ \exp_x)|} d\lambda \\
&= \int_{\log_x(\exp_x(\tilde{B}\cap\tilde{I}_x))} f \circ \exp_x \sqrt{|\det(G_x^\perp \circ \exp_x)|} d\lambda \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_x)} f d\lambda_M \qquad \because \text{Amann et al. [AE09, Ch XII, Thm 1.10]} \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_x)} f d\lambda_M + \int_{\exp_x(\tilde{B}\cap\tilde{C}_x)} f d\lambda_M \\
&= \int_B f d\lambda_M,
\end{aligned}$$

as desired. □

Proposition VI.23. *Let $x \in M$ be fixed. Let X be a random variable on M with density f_X where the underlying probability space is $(\Omega, \mathbb{P}, \mathcal{A})$ (see Definition VI.5). Define $Z := \log_x(X)$. Then Z is a random variable on T_xM such that for all events $E \in \mathcal{A}$ and Borel sets $\tilde{B} \subseteq T_xM$ we have $\Pr(E \cap \{Z \in \tilde{B}\}) = \Pr(E \cap \{X \in \exp_x(\tilde{B} \cap \tilde{I}_x)\})$,*

Proof of Proposition VI.23. To start with, we have

$$\begin{aligned}
&\Pr(E \cap \{Z \in \tilde{B}\}) \\
&= \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{D}_x\}) \qquad \because \log_x(M) \subseteq \tilde{D}_x \\
&= \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{I}_x\}) + \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{C}_x\}) \qquad \because \tilde{D}_x = \tilde{I}_x \cup \tilde{C}_x,\ \emptyset = \tilde{I}_x \cap \tilde{C}_x \\
&= \Pr(E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{I}_x\}) + \Pr(E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{C}_x\}).
\end{aligned}$$

Since $\exp_x : \tilde{I}_x \to I_x$ is a diffeomorphism (Lemma VI.19-part 4) with inverse $\log_x$, we have

$$E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{I}_x\} = E \cap \{X \in \exp_x(\tilde{B} \cap \tilde{I}_x)\}$$

as sets. On the other hand,

$$E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{C}_x\} \subseteq \{X \in C_x\}.$$

Finally, $\Pr(X \in C_x) = \int_{C_x} f_X d\lambda_M = 0$ since C_x has λ_M-measure zero (Lemma VI.19-part 5). □

Proof of Proposition VI.12 part (i). Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition VI.3). Let $E = \Omega$ in

Proposition VI.23. Then we have

$$\begin{aligned}
&\Pr(Z \in \tilde{B}) \\
&= \Pr(X \in \exp_x(\tilde{B} \cap \tilde{I}_x)) \qquad \because \text{Part (i)} \\
&= \int_{\exp_x(\tilde{B} \cap \tilde{I}_x)} f_X d\lambda_M \qquad \because f_X \text{ is the density of } X \\
&= \int_{\tilde{B} \cap \tilde{I}_x} (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x) d\lambda \qquad \because \text{Lemma VI.22} \\
&= \int_{\tilde{B}} f_Z d\lambda \qquad \because \text{Definition of } f_Z
\end{aligned}$$

By assumption, f_X is Borel-measurable. By Lemma VI.22, ν_x is Borel-measurable. Since $\exp_x$ is continuous, we have that both $f_X \circ \exp_x$ and $\nu_x \circ \exp_x$ are Borel-measurable. This proves that f_Z is Borel-measurable. Hence, the integrand is Borel-measurable and a density function for Z. □

Proof of Proposition VI.12 part (ii). Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition VI.3). By Lemma VI.19 part 6, we have $\tau_x^* := \inf_{u \in U_x M} \tau_x(u) > 0$. Now, let $r \in (0, \tau_x^*)$. By the definition of r, we have $\mathtt{B}_x(r, M) \subseteq \tilde{I}_x$. Hence letting $z = \log_x(\xi)$ for $\xi \in \mathtt{B}_x(r, M)$, by Equation (6.6) we have

$$\mathtt{dist}_M(x, \xi) = \mathtt{dist}_M(x, \exp_x(z)) = \|z\|_x. \tag{6.8}$$

Thus,

$$\log_x(\mathtt{B}_x(r, M)) = \{z \in T_x M : \|z\|_x < r\} = \mathtt{B}_0(r, T_x M) \tag{6.9}$$

and

$$\mathtt{B}_x(r, M) = \exp_x(\mathtt{B}_0(r, T_x M)). \tag{6.10}$$

Thus, by Lemma VI.22, we have

$$\int_{\mathtt{B}_x(r,M)} f d\lambda_M = \int_{\mathtt{B}_0(r,T_xM)} h d\lambda. \tag{6.11}$$

Before proceeding, we need the following lemma:

Lemma VI.24. *For all $x \in M$, we have $\lim_{r\to 0} \frac{\lambda_M(\mathtt{B}_x(r,M))}{\lambda(\mathtt{B}_0(r,T_xM))} = 1$.*

Proof of Lemma VI.24. Let $\omega_d := \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ be the volume of the unit ball in $\mathbb{R}^d$ where Γ is the gamma function. Then $\lambda(\mathtt{B}_0(r,T_xM)) = \omega_d r^d$. Next, [Sak96, Ch II.5 Exercise 3] states that

$$\lim_{r\to 0} \frac{r^d\omega_d - \lambda_M(\mathtt{B}_x(r,M))}{r^{d+2}} = \frac{\omega_d}{6(d+2)} S_x$$

where $S_x \in \mathbb{R}$ is a constant that depends only on x (it is the scalar curvature of M at x). By simple algebra, the above yields

$$0 = \lim_{r\to 0} \frac{1}{r^2}\left(1 - \frac{\lambda_M(\mathtt{B}_x(r,M))}{\omega_d r^d} - \frac{S_x r^2}{6(d+2)}\right)$$

In particular, we have $\lim_{r\to 0} 1 - \frac{\lambda_M(\mathtt{B}_x(r,M))}{\omega_d r^d} = 0$, as desired. □

Now we continue with the proof of Proof of Proposition VI.12 part (ii). We observe that

$$\begin{aligned}
f(x) &= \lim_{r\to 0} \frac{\int_{\mathtt{B}_x(r,M)} f d\lambda_M}{\lambda_M(\mathtt{B}_x(r,M))} && \because x \text{ is a Lebesgue point of } f \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_x(M))} h d\lambda}{\lambda_M(\mathtt{B}_x(r,M))} && \because \text{definition of } h \text{ and equation (6.11)} \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_xM)} h d\lambda}{\lambda_M(\mathtt{B}_x(r,M))} \frac{\lambda_M(\mathtt{B}_x(r,M))}{\lambda(\mathtt{B}_0(r,T_xM))} && \because \text{Lemma VI.24} \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_xM)} h d\lambda}{\lambda(\mathtt{B}_0(r,T_xM))}.
\end{aligned}$$

Since $f(x) = h(0)$ (Lemma VI.22) , we've shown that

$$g(0) = \lim_{r \to 0} \frac{\int_{\mathtt{B}_0(r, T_x M)} h d\lambda}{\lambda(\mathtt{B}_0(r, T_x M))}.$$

Thus, 0 is a Lebesgue point of h, as desired. □

6.6.4 Proof of Proposition VI.13

Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition VI.3). Let $A \subseteq \mathbb{R}$ and $\tilde{B} \subseteq T_x M$ be Borel subsets. Then

$$\begin{aligned}
&\int_{\tilde{B}} P_{Y|Z}(A|z) f_Z(z) d\lambda(z) \\
&= \int_{\tilde{B}} P_{Y|X}(A|\exp_x(z)) f_Z(z) d\lambda(z) && \because \text{Definition of } P_{Y|Z=z} \\
&= \int_{\exp_x(\tilde{B} \cap \tilde{I}_p)} P_{Y|X}(A|x) f_X(x) d\lambda_M(x) && \because \text{Lemma VI.22 and Proposition VI.12 (ii)} \\
&= \Pr(Y \in A, X \in \exp_x(\tilde{B} \cap \tilde{I}_x)) \\
&= \Pr(Y \in A, Z \in \tilde{B}) && \because \text{Proposition VI.12 (i) with } E := \{Y \in A\}
\end{aligned}$$

This proves that $P_{Y|Z}(\cdot|\cdot)$ is a conditional probability for Y given Z.

CHAPTER VII

Future directions

This thesis explored several algorithms defined via multiple hyperplanes, namely, multiclass classification with linear discriminant functions, hyperplane arrangement classifiers, and ensemble of random hyperplane arrangements on the sphere. Below, we discuss questions inspired by this thesis research that do not fit neatly into any of the chapters.

Exact characterization of the Natarajan dimension of linear classifiers. It is well-known that the VC dimension of linear classifiers for binary classification in $\mathbb{R}^d$ is exactly $d + 1$. This is proven using Radon's theorem from convex geometry. For the k-ary multiclass case, the Natarajan dimension of linear classifier is upper bounded by $3kd \log(kd)$. See [SB14, Lemma 29.5]. When we substitute in $k = 2$ to the multiclass result, we get the upper bound $6d \log(2d)$. Given that the Natarajan dimension is the multiclass generalization of the VC dimension [SB14, Chapter 29], can this logarithmic term be removed from this upper bound? Is there an analogous Radon's theorem for analyzing the Natarajan dimension for the multiclass linear classifiers?

Multiclass hinge loss beyond SVMs. Fathony et al. [Fat+16] and Duchi et al. [DKR18] proposed a variant of multiclass hinge loss (hereinafter referred to as the

FD hinge loss) that is classification-calibrated. Frongillo et al. [FW21] proved the polyhedral surrogate loss functions (including the FD hinge loss) has a regret function that is superior to smooth surrogate (such as the cross-entropy). From this point of view, the FD hinge loss is superior to the cross-entropy. Given this, can the FD hinge loss be as competitive as the cross-entropy for training neural networks? If not, can we pinpoint the reason?

BIBLIOGRAPHY

[Aba+16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. "TensorFlow: A system for large-scale machine learning". In: *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016, pp. 265–283.

[AE09] Herbert Amann and Joachim Escher. *Analysis III*. Springer, 2009, pp. 389–455.

[Ami+07] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. "Uncovering shared structures in multiclass classification". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 17–24.

[Ani+16] Rushil Anirudh, Vinay Venkataraman, Karthikeyan Natesan Ramamurthy, and Pavan Turaga. "A Riemannian framework for statistical analysis of topological persistence diagrams". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 68–76.

[Aro+19] Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. "Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks". In: *International Conference on Learning Representations*. 2019.

[AT07] Jean-Yves Audibert and Alexandre B Tsybakov. "Fast learning rates for plug-in classifiers". In: *The Annals of Statistics* 35.2 (2007), pp. 608–633.

[Bar+19] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks". In: *Journal of Machine Learning Research* 20.63 (2019), pp. 1–17.

[Bar+20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[BB99] Erin J Bredensteiner and Kristin P Bennett. "Multicategory classification by support vector machines". In: *Computational Optimization*. Springer, 1999, pp. 53–79.

[BDL08] Gérard Biau, Luc Devroye, and Gäbor Lugosi. "Consistency of random forests and other averaging classifiers." In: *Journal of Machine Learning Research* 9.9 (2008).

[Bei+14] Oscar Beijbom, Mohammad Saberian, David Kriegman, and Nuno Vasconcelos. "Guess-averse loss functions for cost-sensitive multiclass boosting". In: *International Conference on Machine Learning*. 2014, pp. 586–594.

[Bel+19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. "Spectrally-normalized margin bounds for neural networks". In: *Proceedings of the*

31st International Conference on Neural Information Processing Systems. 2017, pp. 6241–6250.

[BFU14] Mathieu Blondel, Akinori Fujino, and Naonori Ueda. "Large-scale multiclass support vector machine training via Euclidean projection onto the simplex". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 1289–1294.

[BG16] Gerrit van den Burg and Patrick Groenen. "GenSVM: A generalized multiclass support vector machine". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7964–8005.

[BG79] David A Blackwell and Meyer A Girshick. *Theory of games and statistical decisions*. Courier Corporation, 1979.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. 1992, pp. 144–152.

[BHM18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 2300–2311.

[BJM06] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. "Convexity, classification, and risk bounds". In: *Journal of the American Statistical Association* 101.473 (2006), pp. 138–156.

[Blo19] Mathieu Blondel. "Structured prediction with projection oracles". In: *Advances in Neural Information Processing Systems*. 2019, pp. 12145–12156.

[Blu+98] Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. "A polynomial-time algorithm for learning noisy linear threshold functions". In: *Algorithmica* 22.1-2 (1998), pp. 35–52.

[BMN20] Mathieu Blondel, André FT Martins, and Vlad Niculae. "Learning with Fenchel-Young losses". In: *Journal of Machine Learning Research* 21.35 (2020), pp. 1–69.

[BN20] Aharon Ben-Tal and Arkadi Nemirovski. *Optimization III: Convex Analysis, Nonlinear Programming Theory, Standard Nonlinear Programming Algorithms*. Lecture notes, Georgia Institute of Technology. URL: `https://www2.isye.gatech.edu/~nemirovs/OPTIIILectureNotes2020.pdf` Last visited on 2020/06/08. 2020.

[BOB19] Dmitry Babichev, Dmitrii Ostrovskii, and Francis Bach. "Efficient primal-dual algorithms for large-scale multiclass classification". In: *arXiv preprint arXiv:1902.03755* (2019).

[BPS18] Amir Beck, Edouard Pauwels, and Shoham Sabach. "Primal and dual predicted decrease approximation methods". In: *Mathematical Programming* 167.1 (2018), pp. 37–73.

[BR07] Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*. Vol. 1. Springer, 2007.

[Bre+84] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[Bre00] Leo Breiman. *Some infinity theory for predictor ensembles*. Technical Report. Department of Statistics, 2000.

[Bre04] Leo Breiman. "Consistency for a simple model of random forests". In: (2004).

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. "Does data interpolation contradict statistical optimality?" In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1611–1619.

[BS13] M Beltagy and S Shenawy. "On the boundary of closed convex sets in E^n". In: *arXiv preprint arXiv:1301.0688* (2013).

[BS16] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25.2 (2016), pp. 197–227.

[BSS20] Han Bao, Clayton Scott, and Masashi Sugiyama. "Calibrated Surrogate Losses for Adversarially Robust Classification". In: *Conference on Learning Theory*. 2020.

[BT07] Peter L Bartlett and Mikhail Traskin. "AdaBoost is consistent". In: *Journal of Machine Learning Research* 8.Oct (2007), pp. 2347–2368.

[BT11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Vol. 6. Athena Scientific Belmont, MA, 1997.

[Buc43] Robert Creighton Buck. "Partition of space". In: *The American Mathematical Monthly* 50.9 (1943), pp. 541–544.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BV19] Pierre Baldi and Roman Vershynin. "The capacity of feedforward neural networks". In: *Neural networks* 116 (2019), pp. 288–311.

[BZA20] Thomas Bendokat, Ralf Zimmermann, and P-A Absil. "A Grassmann manifold handbook: Basic geometry and computational aspects". In: *arXiv preprint arXiv:2011.13699* (2020).

[CB20] Lenaic Chizat and Francis Bach. "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss". In: *Conference on Learning Theory*. PMLR. 2020, pp. 1305–1338.

[CFL06] Pai-Hsuen Chen, Rong-En Fan, and Chih-Jen Lin. "A study on SMO-type decomposition methods for support vector machines". In: *IEEE Trans. Neural Networks* 17.4 (2006), pp. 893–908.

[Cid12] Jesús Cid-Sueiro. "Proper losses for learning from partial labels". In: *Advances in neural information processing systems*. 2012, pp. 1565–1573.

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[CL21] Niladri S Chatterji and Philip M Long. "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime". In: *Journal of Machine Learning Research* 22.129 (2021), pp. 1–30.

[CLL20] Chi-Cheng Chiu, Pin-Yen Lin, and Chih-Jen Lin. "Two-variable Dual Coordinate Descent Methods for Linear SVM with/without the Bias Term". In: *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM. 2020, pp. 163–171.

[Con16] Laurent Condat. "Fast projection onto the simplex and the l_1 ball". In: *Mathematical Programming* 158.1-2 (2016), pp. 575–585.

[CP97] Joseph T Chang and David Pollard. "Conditioning as disintegration". In: *Statistica Neerlandica* 51.3 (1997), pp. 287–317.

[CS01] Koby Crammer and Yoram Singer. “On the algorithmic implementation of multiclass kernel-based vector machines”. In: *Journal of Machine Learning Research* 2.Dec (2001), pp. 265–292.

[CST11] Timothee Cour, Ben Sapp, and Ben Taskar. “Learning from partial labels”. In: *Journal of Machine Learning Research* 12.May (2011), pp. 1501–1536.

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (1995), pp. 273–297.

[CZ01] Adele Cutler and Guohua Zhao. “PERT-perfect random tree ensembles”. In: *Computing Science and Statistics* 33 (2001), pp. 490–497.

[Dar+19] Sajad Darabi, Mouloud Belbahri, Matthieu Courbariaux, and Vahid Partovi Nia. “Regularized Binary Network Training”. In: *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*. 2019. arXiv: `1812.11800 [cs.LG]`.

[DB94] Thomas G Dietterich and Ghulum Bakiri. “Solving multiclass learning problems via error-correcting output codes”. In: *Journal of artificial intelligence research* 2 (1994), pp. 263–286.

[DG14] Alex Davies and Zoubin Ghahramani. “The random forest kernel and other kernels for big data from random partitions”. In: *arXiv preprint arXiv:1402.4293* (2014).

[DGI16] Ürün Doğan, Tobias Glasmachers, and Christian Igel. “A unified view on multi-class support vector classification”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1550–1831.

[DGK98] Luc Devroye, László Györfi, and Adam Krzyzak. “The Hilbert kernel regression estimate”. In: *Journal of Multivariate Analysis* 65.2 (1998), pp. 209–227.

[DGT19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. "Distribution-independent pac learning of halfspaces with massart noise". In: *Advances in Neural Information Processing Systems*. 2019, pp. 4749–4760.

[DKR18] John Duchi, Khashayar Khosravi, and Feng Ruan. "Multiclass classification, information, divergence and surrogate risk". In: *The Annals of Statistics* 46.6B (2018), pp. 3246–3275.

[DL15] Emmanuel Didiot and Fabien Lauer. "Efficient optimization of multiclass support vector machines with MSVMpack". In: *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Springer, 2015, pp. 23–34.

[DMJ13] John C Duchi, Lester Mackey, and Michael I Jordan. "The asymptotics of ranking algorithms". In: *The Annals of Statistics* 41.5 (2013), pp. 2292–2323.

[Do 92] Manfredo Perdigao Do Carmo. *Riemannian Geometry*. Vol. 6. Springer, 1992.

[Duc+08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. "Efficient projections onto the l_1-ball for learning in high dimensions". In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 272–279.

[Dud18] Richard M Dudley. *Real Analysis and Probability*. CRC Press, 2018.

[Fag+04] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D Sivakumar, and Erik Vee. "Comparing and aggregating rankings with ties". In: *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2004, pp. 47–58.

[Fan+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: A Library for Large Linear Classification". In: *Journal of Machine Learning Research* 9 (2008), pp. 1871–1874.

[Fat+16] Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian Ziebart. "Adversarial multiclass classification: A risk minimization perspective". In: *Advances in Neural Information Processing Systems*. 2016, pp. 559–567.

[FCL05] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. "Working set selection using second order information for training support vector machines". In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 1889–1918.

[Fer+14] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. "Do we need hundreds of classifiers to solve real world classification problems?" In: *Journal of Machine Learning Research* 15.1 (2014), pp. 3133–3181.

[Fey+19] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. "Interpolating between optimal transport and mmd using sinkhorn divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 2681–2690.

[FFW19] Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. "An embedding framework for consistent polyhedral surrogates". In: *Advances in neural information processing systems*. 2019, pp. 10780–10790.

[FFW20] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. "Embedding Dimension of Polyhedral Losses". In: *Conference on Learning Theory*. 2020, pp. 1558–1585.

[FH16] Aasa Feragen and Søren Hauberg. "Open Problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive def-

inite geodesic exponential kernel?" In: *Conference on Learning Theory*. PMLR. 2016, pp. 1647–1650.

[FS12] Yoav Freund and Robert E Schapire. "Boosting: Foundations and Algorithms". In: *MIT Press* 1.6 (2012), p. 7.

[Fuk06] Ryuichi Fukuoka. "Mollifier smoothing of tensor fields on differentiable manifolds and applications to Riemannian Geometry". In: *arXiv preprint math/0608230* (2006).

[Fuk15] Komei Fukuda. *Lecture: Polyhedral Computation, Spring 2013*. 2015. URL: `http://www-oldurls.inf.ethz.ch/personal/fukudak/lect/pclect/notes2015/PolyComp2015.pdf`.

[FW21] Rafael Frongillo and Bo Waggoner. "Surrogate Regret Bounds for Polyhedral Losses". In: *Advances in Neural Information Processing Systems* 34 (2021).

[FW95] Sally Floyd and Manfred Warmuth. "Sample compression, learnability, and the Vapnik-Chervonenkis dimension". In: *Machine learning* 21.3 (1995), pp. 269–304.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. "Hardness of learning halfspaces with noise". In: *SIAM Journal on Computing* 39.2 (2009), pp. 742–765.

[Gro62] Oliver A Gross. "Preferential arrangements". In: *The American Mathematical Monthly* 69.1 (1962), pp. 4–8.

[GT20] Lukas Geiger and Plumerai Team. "Larq: An Open-Source Library for Training Binarized Neural Networks". In: *Journal of Open Source Software* 5.45 (Jan. 2020), p. 1746.

[Gyö+06] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[HB20] Like Hui and Mikhail Belkin. "Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks". In: *arXiv preprint arXiv:2006.07322* (2020).

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Heb87] James J Hebda. "Parallel translation of curvature along geodesics". In: *Transactions of the American Mathematical Society* 299.2 (1987), pp. 559–572.

[HK21] Steve Hanneke and Aryeh Kontorovich. "Stable Sample Compression Schemes: New Applications and an Optimal SVM Margin Bound". In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 697–721.

[HL02] Chih-Wei Hsu and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines". In: *IEEE Transactions on Neural Networks* 13.2 (2002), pp. 415–425.

[Hsi+08] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. "A dual coordinate descent method for large-scale linear SVM". In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 408–415.

[Hub+16] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. "Binarized neural networks". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 4114–4122.

[Hub+17] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. "Quantized neural networks: Training neural networks with low precision weights and activations". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6869–6898.

[Hus+06] Don Hush, Patrick Kelly, Clint Scovel, and Ingo Steinwart. "QP algorithms with guaranteed accuracy and run time for support vector machines". In: *Journal of Machine Learning Research* 7.May (2006), pp. 733–769.

[IHG08] Christian Igel, Verena Heidrich-Meisner, and Tobias Glasmachers. "Shark". In: *Journal of Machine Learning Research* 9.Jun (2008), pp. 993–996.

[IKZ08] Masao Ishikawa, Anisse Kasraoui, and Jiang Zeng. "Euler–Mahonian statistics on ordered set partitions". In: *SIAM Journal on Discrete Mathematics* 22.3 (2008), pp. 1105–1137.

[Jay+15] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. "Kernel methods on Riemannian manifolds with Gaussian RBF kernels". In: *IEEE transactions on pattern analysis and machine intelligence* 37.12 (2015), pp. 2464–2477.

[Ji+20] Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. "Gradient descent follows the regularization path for general losses". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2109–2136.

[Jia+19] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. "Fantastic Generalization Measures and Where to Find Them". In: *International Conference on Learning Representations*. 2019.

[Joa06] Thorsten Joachims. "Training linear SVMs in linear time". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, pp. 217–226.

[JS08] Tony Jebara and Pannagadatta Shivaswamy. "Relative margin machines". In: *Advances in Neural Information Processing Systems* 21 (2008).

[Kal+08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. "Agnostically learning halfspaces". In: *SIAM Journal on Computing* 37.6 (2008), pp. 1777–1805.

[Kee+01] S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. "Improvements to Platt's SMO algorithm for SVM classifier design". In: *Neural Computation* 13.3 (2001), pp. 637–649.

[Kee+08] S Sathiya Keerthi, Sellamanickam Sundararajan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. "A sequential dual method for large scale multi-class linear SVMs". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008, pp. 408–416.

[Kim+19] Hyungjun Kim, Kyungsu Kim, Jinseok Kim, and Jae-Joon Kim. "BinaryDuo: Reducing Gradient Mismatch in Binary Activation Network by Coupling Binary Activations". In: *International Conference on Learning Representations*. 2019.

[Kla+17] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. "Self-normalizing neural networks". In: *Advances in Neural Information Processing Systems*. 2017, pp. 971–980.

[KV08] Yuri Kalnishkan and Michael V Vyugin. "The weak aggregating algorithm and weak mixability". In: *Journal of Computer and System Sciences* 74.8 (2008), pp. 1228–1244.

[Law01] Eugene L Lawler. *Combinatorial optimization: networks and matroids*. Courier Corporation, 2001.

[LC19] Ching-Pei Lee and Kai-Wei Chang. "Distributed block-diagonal approximation methods for regularized empirical risk minimization". In: *Machine Learning* (2019), pp. 1–40.

[Lee13] John M Lee. "Smooth manifolds". In: *Introduction to Smooth Manifolds*. Springer, 2013, pp. 1–31.

[Lee14] Yoonkyung Lee. "Comments on: Support vector machines maximizing geometric margins for multi-class classification". In: *Top* 22.3 (2014), pp. 852–855.

[Lei+19] Yunwen Lei, Ürün Doğan, Ding-Xuan Zhou, and Marius Kloft. "Data-dependent generalization bounds for multi-class classification". In: *IEEE Transactions on Information Theory* 65.5 (2019), pp. 2995–3021.

[LHS17] Maksim Lapin, Matthias Hein, and Bernt Schiele. "Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification". In: *IEEE transactions on pattern analysis and machine intelligence* 40.7 (2017), pp. 1533–1554.

[Lin02] Chih-Jen Lin. "A formal analysis of stopping criteria of decomposition methods for support vector machines". In: *IEEE Transactions on Neural Networks* 13.5 (2002), pp. 1045–1052.

[Lis+07] Nikolas List, Don Hush, Clint Scovel, and Ingo Steinwart. "Gaps in support vector optimization". In: *International Conference on Computational Learning Theory*. Springer. 2007, pp. 336–348.

[Liu07] Yufeng Liu. "Fisher consistency of multicategory support vector machines". In: *Artificial intelligence and statistics*. 2007, pp. 291–298.

[LLW04] Yoonkyung Lee, Yi Lin, and Grace Wahba. "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data". In: *Journal of the American Statistical Association* 99.465 (2004), pp. 67–81.

[LP08] Jean-Michel Loubes and Bruno Pelletier. "A kernel-based classifier on a Riemannian manifold". In: *Statistics & Decisions* 26.1 (2008), pp. 35–51.

[LP09] László Lovász and Michael D Plummer. *Matching theory*. Vol. 367. American Mathematical Soc., 2009.

[LS04] Niko List and Hans Ulrich Simon. "A general convergence theorem for the decomposition method". In: *International Conference on Computational Learning Theory*. Springer. 2004, pp. 363–377.

[LS07] Nikolas List and Hans Ulrich Simon. "General polynomial time decomposition algorithms". In: *Journal of Machine Learning Research* 8.Feb (2007), pp. 303–321.

[LS09] Nikolas List and Hans Ulrich Simon. "SVM-optimization and steepest-descent line search". In: *Proceedings of the 22nd Annual Conference on Computational Learning Theory*. 2009.

[LT92] Zhi-Quan Luo and Paul Tseng. "On the convergence of the coordinate descent method for convex differentiable minimization". In: *Journal of Optimization Theory and Applications* 72.1 (1992), pp. 7–35.

[LT93] Zhi-Quan Luo and Paul Tseng. "Error bounds and convergence analysis of feasible descent methods: a general approach". In: *Annals of Operations Research* 46.1 (1993), pp. 157–178.

[LW86] Nick Littlestone and Manfred Warmuth. "Relating data compression and learnability". In: (1986).

[LY18] Tam Le and Makoto Yamada. "Persistence Fisher kernel: A Riemannian manifold kernel for persistence diagrams". In: *Advances in Neural Information Processing Systems* 31 (2018).

[Maa94] Wolfgang Maass. "Neural nets with superlinear VC-dimension". In: *Neural Computation* 6.5 (1994), pp. 877–884.

[Man12] Toufik Mansour. *Combinatorics of set partitions*. CRC Press, 2012.

[Mar+22] Kanti V Mardia, Henrik Wiechers, Benjamin Eltzner, and Stephan F Huckemann. "Principal component analysis and clustering on manifolds". In: *Journal of Multivariate Analysis* 188 (2022), p. 104862.

[MB17] Anirbit Mukherjee and Amitabh Basu. "Lower bounds over Boolean inputs for deep neural networks with ReLU gates". In: *arXiv preprint arXiv:1711.03073* (2017).

[MBP19] Arthur Mensch, Mathieu Blondel, and Gabriel Peyré. "Geometric losses for distributional learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4516–4525.

[MJ00] Kanti V Mardia and Peter E Jupp. *Directional statistics*. Vol. 2. Wiley Online Library, 2000.

[Mro+12] Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jeacques Slotine. "Multiclass learning with simplex coding". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2789–2797.

[Mun18] James R Munkres. *Analysis on manifolds*. CRC Press, 2018.

[Mut+21] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. "Classification vs regression in overparameterized regimes: Does the loss function matter?" In: *Journal of Machine Learning Research* 22.222 (2021), pp. 1–69.

[MV08] Hamed Masnadi-Shirazi and Nuno Vasconcelos. "On the design of loss functions for classification: theory, robustness to outliers, and savageboost". In: *Advances in neural information processing systems* 21 (2008).

[NBR19] Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. "A general theory for structured prediction with smooth convex surrogates". In: *arXiv preprint arXiv:1902.01958* (2019).

[Ney+17] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. "Exploring generalization in deep learning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 5949–5958.

[NK19] Vaishnavh Nagarajan and J Zico Kolter. "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems*. 2019.

[Pen06] Xavier Pennec. "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements". In: *Journal of Mathematical Imaging and Vision* 25.1 (2006), pp. 127–154.

[Pin19] Iosif Pinelis. *Probability of two points being divided by an high-dimensional hyperplane*. MathOverflow. URL:https://mathoverflow.net/q/323697 (version: 2019-02-21). 2019.

[Pla98] John Platt. *Sequential minimal optimization: A fast algorithm for training support vector machines*. Tech. rep. 1998.

[Ple77] Robert J Plemmons. "M-matrix characterizations. I. nonsingular M-matrices". In: *Linear Algebra and its Applications* 18.2 (1977), pp. 175–188.

[PS16] Bernardo Ávila Pires and Csaba Szepesvári. "Multiclass classification calibration functions". In: *arXiv preprint arXiv:1609.06385* (2016).

[RA16] Harish G Ramaswamy and Shivani Agarwal. "Convex calibration dimension for multiclass loss matrices". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 397–441.

[Roc70] R Tyrrell Rockafellar. *Convex analysis*. 28. Princeton university press, 1970.

[Ros57] Frank Rosenbaltt. "The Perceptron – A Perceiving and Recognizing Automaton". In: *Cornell Aeronautical Laboratory* (1957).

[RTA18] Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. "Consistent algorithms for multiclass classification with an abstain option". In: *Electronic Journal of Statistics* 12.1 (2018), pp. 530–554.

[Sak96] Takashi Sakai. *Riemannian Geometry*. Vol. 149. American Mathematical Society, 1996.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SBV15] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (2015), pp. 1716–1741.

[Sco12] Clayton Scott. "Calibrated asymmetric surrogate losses". In: *Electronic Journal of Statistics* 6 (2012), pp. 958–992.

[Sco16] Erwan Scornet. "Random forests and kernel methods". In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1485–1500.

[Sha+11] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. "Pegasos: Primal estimated sub-gradient solver for SVM". In: *Mathematical Programming* 127.1 (2011), pp. 3–30.

[Sha+20] Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. "Neural kernels without tangents". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8614–8623.

[She68] Donald Shepard. "A two-dimensional interpolation function for irregularly-spaced data". In: *Proceedings of the 1968 23rd ACM national conference*. 1968, pp. 517–524.

[SHS11] Ingo Steinwart, Don Hush, and Clint Scovel. "Training SVMs Without Offset". In: *Journal of Machine Learning Research* 12.1 (2011).

[SJ10] Pannagadatta K Shivaswamy and Tony Jebara. "Maximum Relative Margin and Data-Dependent Regularization." In: *Journal of Machine Learning Research* 11.2 (2010).

[Sou+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2822–2878.

[SS21] Timo Schick and Hinrich Schütze. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 2339–2352.

[ST17] Ingo Steinwart and Philipp Thomann. "liquidSVM: A fast and versatile SVM package". In: *arXiv preprint arXiv:1702.06899* (2017).

[Ste02] Ingo Steinwart. "Support vector machines are universally consistent". In: *Journal of Complexity* 18.3 (2002), pp. 768–791.

[Ste05] Ingo Steinwart. "Consistency of support vector machines and other regularized kernel classifiers". In: *IEEE Transactions on Information Theory* 51.1 (2005), pp. 128–142.

[Ste07] Ingo Steinwart. "How to compare different loss functions and their risks". In: *Constructive Approximation* 26.2 (2007), pp. 225–287.

[SV11] Mohammad Saberian and Nuno Vasconcelos. "Multiclass Boosting: Theory and Algorithms". In: *Advances in Neural Information Processing Systems* 24 (2011), pp. 2124–2132.

[SV19] Mohammad Saberian and Nuno Vasconcelos. "Multiclass Boosting: Margins, Codewords, Losses, and Algorithms." In: *Journal of Machine Learning Research* 20.137 (2019), pp. 1–68.

[TAD21] Alberto Torres-Barrán, Carlos M. Alaíz, and José R. Dorronsoro. "Faster SVM training via conjugate SMO". In: *Pattern Recognition* 111 (2021), p. 107644. ISSN: 0031-3203.

[TB07] Ambuj Tewari and Peter L Bartlett. "On the consistency of multiclass classification methods". In: *Journal of Machine Learning Research* 8.May (2007), pp. 1007–1025.

[Tho13] P Thomas Fletcher. "Geodesic regression and the theory of least squares on Riemannian manifolds". In: *International journal of computer vision* 105.2 (2013), pp. 171–185.

[TT14] Keiji Tatsumi and Tetsuzo Tanino. "Support vector machines maximizing geometric margins for multi-class classification". In: *Top* 22.3 (2014), pp. 815–840.

[TZ20] Zhiqiang Tan and Xinwei Zhang. “On loss functions and regret bounds for multi-category classification”. In: *arXiv preprint arXiv:2005.08155* (2020).

[TZ22] Zhiqiang Tan and Xinwei Zhang. “On loss functions and regret bounds for multi-category classification”. In: *IEEE Transactions on Information Theory* (2022).

[Vap98] Vladimir Vapnik. *Statistical learning theory*. 1998.

[Ver12] Ferdinand Verhulst. “Mathematics is the art of giving the same name to different things: An interview with Henri Poincaré”. In: *Nieuw archief voor wiskunde. Serie 5* 13.3 (2012), pp. 154–158.

[VL04] Costas Voglis and Isaac E Lagaris. “BOXCQP: An algorithm for bound constrained convex quadratic problems”. In: *Proceedings of the 1st International Conference: From Scientific Computing to Computational Engineering, IC-SCCE, Athens, Greece*. 2004.

[Was18] Larry Wasserman. “Topological data analysis”. In: *Annual Review of Statistics and Its Application* 5 (2018), pp. 501–532.

[WL14] Po-Wei Wang and Chih-Jen Lin. “Iteration complexity of feasible descent methods for convex optimization”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1523–1548.

[WS20] Yutong Wang and Clayton Scott. “Weston-Watkins Hinge Loss and Ordered Partitions”. In: *Advances in Neural Information Processing Systems*. 2020.

[Wu+18] Xundong Wu, Xiangwen Liu, Wei Li, and Qing Wu. “Improved expressivity through dendritic neural networks”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 8068–8079.

[WVR16] Robert C Williamson, Elodie Vernet, and Mark D Reid. "Composite multiclass losses". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7860–7911.

[WW99] Jason Weston and Chris Watkins. "Support Vector Machines for Multi-Class Pattern Recognition". In: *Proc. 7th European Symposium on Artificial Neural Networks, 1999*. 1999.

[Wyn+17] Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. "Explaining the success of AdaBoost and random forests as interpolating classifiers". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1558–1590.

[YB18] Jiaqian Yu and Matthew B Blaschko. "The Lovász hinge: A novel convex surrogate for submodular losses". In: *IEEE transactions on pattern analysis and machine intelligence* (2018).

[YZ20] Zhigang Yao and Zhenyue Zhang. "Principal boundary on Riemannian manifolds". In: *Journal of the American Statistical Association* 115.531 (2020), pp. 1435–1448.

[ZA20] Mingyuan Zhang and Shivani Agarwal. "Bayes Consistency vs. H-Consistency: The Interplay between Surrogate Loss Functions and the Scoring Function Class." In: *Advances in neural information processing systems*. 2020.

[Zha+09] Zhihua Zhang, Michael Jordan, Wu-Jun Li, and Dit-Yan Yeung. "Coherence functions for multicategory margin-based classification methods". In: *Artificial Intelligence and Statistics*. 2009, pp. 647–654.

[Zha+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3 (2021), pp. 107–115.

[Zha04a] Tong Zhang. “Statistical analysis of some multi-category large margin classification methods”. In: *Journal of Machine Learning Research* 5.Oct (2004), pp. 1225–1251.

[Zha04b] Tong Zhang. “Statistical behavior and consistency of classification methods based on convex risk minimization”. In: *The Annals of Statistics* (2004), pp. 56–85.

[Zim17] Ralf Zimmermann. “A matrix-algebraic algorithm for the Riemannian logarithm on the Stiefel manifold under the canonical metric”. In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017), pp. 322–342.

[ZZ04] Zhenyue Zhang and Hongyuan Zha. “Principal manifolds and nonlinear dimensionality reduction via tangent space alignment”. In: *SIAM journal on scientific computing* 26.1 (2004), pp. 313–338.

[ZZH08] Hui Zou, Ji Zhu, and Trevor Hastie. “New multicategory boosting algorithms based on multicategory fisher-consistent losses”. In: *The Annals of Applied Statistics* 2.4 (2008), p. 1290.