



Czech Technical University in Prague  
Faculty of Nuclear Sciences and  
Physical Engineering

# General Framework for Classification at the Top

*Dissertation*



**Author:**  
**Academic year:**

Ing. Václav Mácha  
2021/2022



## Poděkování:

Thanks thanks thanks thanks thanks thanks thanks thanks thanks thanks thanks  
thanks thanks thanks thanks thanks thanks thanks thanks thanks thanks thanks  
thanks thanks

## Čestné prohlášení:

Prohlašuji na tomto místě, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškerou použitou literaturu.

V Praze dne 1. prosince 2021

.....  
Ing. Václav Mácha



Název:	Title title title title title title
Autor:	Ing. Václav Mácha
Obor:	Matematické inženýrství
Druh práce:	Disertační práce
Školitel:	doc. Ing Václav Šmídl, Ph.D.
Školitel specialista:	Mgr. Lukáš Adam, Ph.D.
Abstrakt:	Abstract abstract
Klíčová slova:	Keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords



Title:	Title title title title title title
Abstract:	Abstract abstract
Keywords:	Keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords keywords





# Contents

---

<b>1</b>	<b>Linear Classification at the Top</b>	<b>1</b>
1.1	Framework for Minimizing Missclassification Above a Threshold . . . . .	2
1.1.1	Methods based on pushing positives to the top . . . . .	3
1.1.2	Accuracy at the Top . . . . .	4
1.1.3	Methods optimizing the Neyman-Pearson criterion . . . . .	5
1.2	Theoretical Analysis of the Framework . . . . .	6
1.2.1	Threshold value comparison . . . . .	7
1.2.2	Convexity . . . . .	7
1.2.3	Differentiability . . . . .	7
1.2.4	Stability . . . . .	8
1.2.5	Method comparison . . . . .	11
1.3	Convergence of stochastic gradient descent . . . . .	12
1.3.1	Stochastic gradient descent: Basic . . . . .	12
1.3.2	Stochastic gradient descent: Convergent for <i>Pat&amp;Mat</i> and <i>Pat&amp;Mat-NP</i> . . . . .	13
1.4	Numerical experiments . . . . .	14
1.4.1	Implementational details and Hyperparameter choice . . . . .	14
1.4.2	Dataset description and Performance criteria . . . . .	15
1.4.3	Numerical results . . . . .	15
1.5	Conclusion . . . . .	18
<b>2</b>	<b>Non-linear Classification at the Top</b>	<b>21</b>
2.1	Introduction . . . . .	21
2.2	Derivation of dual problems . . . . .	22
2.2.1	<i>TopPushK</i> . . . . .	23
2.2.2	<i>Accuracy at the Top</i> . . . . .	24
2.3	New method for solving dual problems . . . . .	25
2.3.1	Adding kernels . . . . .	25
2.3.2	Update of dual optimization variables . . . . .	26
2.3.3	Algorithm summary and Complexity analysis . . . . .	26
2.3.4	Computing $\Delta$ for <i>TopPushK</i> with truncated quadratic loss . . . . .	27
2.4	Numerical experiments . . . . .	28
2.4.1	Performance criteria . . . . .	29
2.4.2	Hyperparameter choice . . . . .	29
2.4.3	Dataset description . . . . .	29
2.4.4	Experiments . . . . .	30
2.5	Conclusion . . . . .	32

<b>Appendices</b>	<b>35</b>
<b>A Linear case</b>	<b>37</b>
A.1 Additional results and proofs . . . . .	37
A.1.1 Equivalence of (1.10) and (1.11) . . . . .	37
A.1.2 Results related to convexity . . . . .	38
A.1.3 Results related to differentiability . . . . .	38
A.1.4 Results related to stability . . . . .	39
A.1.5 Results related to threshold comparison . . . . .	41
A.2 Computation for Section 1.2.4 . . . . .	41
A.3 Computing the threshold for <i>Pat&amp;Mat</i> . . . . .	42
A.4 Proof of Theorem 1.9 . . . . .	43
A.4.1 General result . . . . .	43
A.4.2 Proof of Theorem 1.9 . . . . .	45
A.4.3 Auxiliary results . . . . .	48
<b>B Non-linear case</b>	<b>51</b>
B.1 Convex conjugate . . . . .	51
B.2 Proofs for dual problems . . . . .	51
B.3 Computing $\Delta^*$ with truncated quadratic surrogate . . . . .	54
B.3.1 <i>TopPushK</i> . . . . .	54
B.3.2 <i>Pat&amp;Mat</i> . . . . .	56
B.4 Computing $\Delta^*$ with hinge loss function . . . . .	58
B.4.1 <i>TopPushK</i> . . . . .	59
B.4.2 <i>Pat&amp;Mat</i> . . . . .	60
<b>Bibliography</b>	<b>63</b>

# Linear Classification at the Top

---

Many binary classification problems focus on separating the dataset by a linear hyperplane  $\mathbf{w}^\top \mathbf{x} - t$ . A sample  $\mathbf{x}$  is deemed to be positive or relevant (depending on the application) if its score  $\mathbf{w}^\top \mathbf{x}$  is above a threshold  $t$ . Multiple problem categories belong to this framework:

- *Ranking problems* select the most relevant samples and rank them. To each sample, a numerical score is assigned, and the ranking is performed based on this score. Often, only scores above a threshold are considered.
- *Accuracy at the Top* is similar to ranking problems. However, instead of ranking the most relevant samples, it only maximizes the accuracy (equivalently minimizes the misclassification) in these top samples. The prime examples of both categories include search engines or problems where identified samples undergo expensive post-processing such as human evaluation.
- *Hypothesis testing* states a null and an alternative hypothesis. The Neyman-Pearson problem minimizes the Type II error (the null hypothesis is false but it fails to be rejected) while keeping the Type I error (the null hypothesis is true but is rejected) small. If the null hypothesis states that a sample has the positive label, then Type II error happens when a positive sample is below the threshold and thus minimizing the Type II error amounts to minimizing the positives below the threshold.

Examples of this type can be found in search engines, where the user is interested only in the first few queries. These queries need to be of high quality. Other examples include cybersecurity [1], where a low false-negative rate is crucial as a high number of false alarms would result in the software being uninstalled, or drug development, where potentially useful drugs need to be preselected and manually investigated. All these three applications may be written (possibly after a reformulation) in a similar form as a minimization of the false-negatives (misclassified positives) above a threshold. They only differ in the way they define the threshold. Despite this striking similarity, they are usually considered separately in the literature. The main goal of this paper is to provide a unified framework for these three applications and perform its theoretical and numerical analysis.

The goal of the ranking problems is to rank the relevant samples higher than the non-relevant ones. A prototypical example is the RankBoost [2] maximizing the area under the ROC curve, the Infinite Push [3] or the  $p$ -norm push [4] which concentrate on the high-ranked negatives and push them down. Since all these papers include pairwise comparisons of all samples, they can be used only for small datasets. This was alleviated in [5], where the authors performed the limit  $p \rightarrow \infty$  in  $p$ -norm push and obtained the

linear complexity in the number of samples. Moreover, since the  $l_\infty$ -norm is equal to the maximum, this method falls into our framework with the threshold equal to the largest score computed from negative samples.

Accuracy at the Top ( $\tau$ -quantile) was formally defined in [6] and maximizes the number of relevant samples in the top  $\tau$ -fraction of ranked samples. When the threshold equals the top  $\tau$ -quantile of all scores, this problem falls into our framework. The early approaches aim at solving approximations, for example, [7] optimizes a convex upper bound on the number of errors among the top samples. Due to the presence of exponentially many constraints, the method is computationally expensive. [6] presented an SVM-like formulation which fixes the index of the quantile and solves  $n$  problems. While this removes the necessity to handle the (difficult) quantile constraint, the algorithm is computationally infeasible for a large number of samples. [8] derived upper approximations, their error bounds and solved these approximations. [1] proposed the projected gradient descent method where after each gradient step, the quantile is recomputed. [9] suggested new formulations for various criteria and argued that they keep desired properties such as convexity. [10] showed that accuracy at the top is maximized by thresholding the posterior probability of the relevant class. The closest approach to our framework is [11, 12], where the authors considered multi-class classification problems, and their goal was to optimize the performance on the top few classes and [13], where the authors implicitly removed some variables and derived an efficient algorithm.

## 1.1 Framework for Minimizing Missclassification Above a Threshold

Many important binary classification problems minimize the number of misclassified samples below (or above) certain threshold. Since these problems are usually considered separately, in this section, we provide a unified framework for their handling and present several classification problems falling into this framework.

For samples  $\mathbf{x}$ , we consider the linear classifier  $f(\mathbf{w}) = \mathbf{w}^\top \mathbf{x} - t$ , where  $\mathbf{w}$  is the normal vector to the separating hyperplane and  $t$  is a threshold. The most well-known example is the support vector machines, where  $t$  is an optimization variable. In many cases the threshold  $t$  is computed from the scores  $s = \mathbf{w}^\top \mathbf{x}$ . For example, *TopPush* from [5] sets the threshold  $t$  to the largest score  $s^-$  corresponding to negative samples and [1] sets it to the quantile of all scores.

To be able to determine the missclassification above and below the threshold  $t$ , we define the true-positive, false-negative, true-negative and false-positive counts by

$$\begin{aligned} \text{tp}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^+} [\mathbf{w}^\top \mathbf{x} - t \geq 0], & \text{fn}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^+} [\mathbf{w}^\top \mathbf{x} - t < 0], \\ \text{tn}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^-} [\mathbf{w}^\top \mathbf{x} - t < 0], & \text{fp}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^-} [\mathbf{w}^\top \mathbf{x} - t \geq 0]. \end{aligned} \tag{1.1}$$

Here  $[\cdot]$  is the 0-1 loss (Iverson bracket, characteristic function) which is equal to 1 if the argument is true and to 0 otherwise. Moreover,  $\mathcal{X}/\mathcal{X}^+/\mathcal{X}^-$  denotes the sets of all/positive/negative samples and by  $n/n^+/n^-$  their respective sizes.

Since the misclassified samples below the threshold are the false-negatives, we arrive

at the following problem

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \text{fn}(\mathbf{w}, t) \\ & \text{subject to} && \text{threshold } t \text{ is a function of } \{\mathbf{w}^\top \mathbf{x}_i\}_{i=1}^n. \end{aligned} \quad (1.2)$$

As the 0-1 loss in (1.1) is discontinuous, problem (1.2) is difficult to handle. The usual approach is to employ a surrogate function such as the hinge loss function defined by

$$l_{\text{hinge}}(s) = \max\{0, 1 + s\}. \quad (1.3)$$

In the text below, the symbol  $l$  denotes any convex non-negative non-decreasing function with  $l(0) = 1$ . Using the surrogate function, the counts (1.1) may be approximated by their surrogate counterparts

$$\begin{aligned} \overline{\text{tp}}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^+} l(\mathbf{w}^\top \mathbf{x} - t), & \overline{\text{fn}}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^+} l(t - \mathbf{w}^\top \mathbf{x}), \\ \overline{\text{tn}}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^-} l(t - \mathbf{w}^\top \mathbf{x}), & \overline{\text{fp}}(\mathbf{w}, t) &= \sum_{\mathbf{x} \in \mathcal{X}^-} l(\mathbf{w}^\top \mathbf{x} - t). \end{aligned} \quad (1.4)$$

Since  $l(\cdot) \geq [\cdot]$ , the surrogate counts (1.4) provide upper approximations of the true counts (1.1). Replacing the counts in (1.2) by their surrogate counterparts and adding a regularization results in

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && \text{threshold } t \text{ is a function of } \{\mathbf{w}^\top \mathbf{x}_i\}_{i=1}^n. \end{aligned} \quad (1.5)$$

In the rest of this section, we list formulations which fall into the framework of (1.2) and (1.5).

### 1.1.1 Methods based on pushing positives to the top

The first category of formulations falling into our framework (1.2) and (1.5) are ranking methods which attempt to put as many positives (relevant samples) to the top as possible. Specifically, for each sample  $\mathbf{x}$ , they compute the score  $s = \mathbf{w}^\top \mathbf{x}$  and then sort the vector  $\mathbf{s}$  into  $\mathbf{s}_{[.]}$  with decreasing components  $s_{[1]} \geq s_{[2]} \geq \dots \geq s_{[n]}$ . The number of positives on top equals to the number of positives above the highest negative. This amounts to maximizing true-positives or, equivalently, minimizing false-negatives, which may be written as

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \text{fn}(\mathbf{w}, t) \\ & \text{subject to} && t = s_{[1]}^-, \\ & && \text{components of } \mathbf{s}^- \text{ equal to } s^- = \mathbf{w}^\top \mathbf{x}^- \text{ for } \mathbf{x}^- \in \mathcal{X}^-. \end{aligned} \quad (1.6)$$

As  $t$  is a function of the scores  $s = \mathbf{w}^\top \mathbf{x}$ , problem (1.6) is a special case of (1.2).

*TopPush* from [5] replaces the false-negatives in (1.6) by their surrogate and adds a regularization term to arrive at

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && t = s_{[1]}^-, \\ & && \text{components of } \mathbf{s}^- \text{ equal to } s^- = \mathbf{w}^\top \mathbf{x}^- \text{ for } \mathbf{x}^- \in \mathcal{X}^-. \end{aligned} \quad (1.7)$$

Note that this falls into the framework of (1.5).

As we will show in Section 1.2.4, *TopPush* is sensitive to outliers and mislabelled data. To robustify it, we follow the idea from [11] and propose to replace the largest negative score by the mean of  $k$  largest negative scores. This results in

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & t = \frac{1}{k} (s_{[1]}^- + \dots + s_{[k]}^-), \\ & \text{components of } \mathbf{s}^- \text{ equal to } s^- = \mathbf{w}^\top \mathbf{x}^- \text{ for } \mathbf{x}^- \in \mathcal{X}^-. \end{aligned} \quad (1.8)$$

We used the mean of highest  $k$  negative scores instead of the value of the  $k$ -th negative score to preserve convexity as shown in Section 1.2.2.

### 1.1.2 Accuracy at the Top

The previous category considers formulations which minimize the false-negatives below the highest-ranked negative. Accuracy at the Top [6] takes a different approach and minimizes false-positives above the top  $\tau$ -quantile defined by

$$t_1(\mathbf{w}) = \max\{t \mid \text{tp}(\mathbf{w}, t) + \text{fp}(\mathbf{w}, t) \geq n\tau\}. \quad (1.9)$$

Then the Accuracy at the Top problem is defined by

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^-} \text{fp}(\mathbf{w}, t) \\ \text{subject to} \quad & t \text{ is the top } \tau\text{-quantile: it solves (1.9)}. \end{aligned} \quad (1.10)$$

Due to Lemma A.1 in the Appendix, the previous problem (1.10) is equivalent (up to a small theoretical issue) to

$$\begin{aligned} \text{minimize} \quad & \mu \text{fn}(\mathbf{w}, t) + (1 - \mu) \text{fp}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & t \text{ is the top } \tau\text{-quantile: it solves (1.9)} \end{aligned} \quad (1.11)$$

for any  $\mu \in [0, 1]$ . This problem with  $\mu = 0$  equals to (1.10), with  $\mu = 1$  it falls into our framework (1.2), while with  $\mu = \frac{n^-}{n}$  it corresponds to the original definition from [6].

Apart from the quantile (1.9), there are two other possible choices of the threshold

$$t_2(\mathbf{w}) = \frac{1}{n\tau} \sum_{i=1}^{n\tau} s_{[i]}, \quad (1.12)$$

$$t_3(\mathbf{w}) \text{ solves } \frac{1}{n} \sum_{i=1}^n l(\beta(s_i - t)) = \tau. \quad (1.13)$$

We again use the vector of scores  $\mathbf{s}$  with components  $s_i = \mathbf{w}^\top \mathbf{x}_i$  and for the rest of the paper we assume, for simplicity, that  $n\tau$  is an integer. The quantile (1.9) is sometimes denoted as VaR (value at risk) and (1.12) as CVaR (conditional value of risk). It is known is that the latter is the tightest convex approximation of the former. We will sometimes denote (1.13) as surrogate top  $\tau$ -quantile. We will investigate the relations between these three objects as well as their properties such as convexity, differentiability or stability in Section 1.2.

Paper [1] builds on the Accuracy at the Top problem (1.11), where it replaces  $\text{fn}(\mathbf{w}, t)$  and  $\text{fp}(\mathbf{w}, t)$  in the objective by their surrogate counterparts  $\overline{\text{fn}}(\mathbf{w}, t)$  and  $\overline{\text{fp}}(\mathbf{w}, t)$ . This leads to

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{1}{n^-} \overline{\text{fp}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && t \text{ is the top } \tau\text{-quantile: it solves (1.9)}. \end{aligned} \quad (1.14)$$

Based on the first author, we name this formulation *Grill*. The main purpose of (1.12) is to provide a convex approximation of the non-convex quantile (1.9). Putting it into the constraint results in a convex approximation problem, which we call *TopMeanK*

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && t = \frac{1}{n\tau} (s_{[1]} + \dots + s_{[n\tau]}), \\ & && \text{components of } \mathbf{s} \text{ equal to } s = \mathbf{w}^\top \mathbf{x} \text{ for } \mathbf{x} \in \mathcal{X}. \end{aligned} \quad (1.15)$$

Similarly, we can use the surrogate top quantile in the constraint to arrive at

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && t \text{ is the surrogate top } \tau\text{-quantile: it solves (1.13)}. \end{aligned} \quad (1.16)$$

Note that *Grill* minimizes the convex combination of false-positives and false-negatives while (1.15) and (1.16) minimize only the false-negatives. The reason for this will be evident in Section 1.2.2 and amounts to preservation of convexity. Moreover, as will see later, problem (1.16) provides a good approximation to the Accuracy at the Top problem, it is easily solvable due to convexity and requires almost no tuning, we named it *Pat&Mat* (Precision At the Top & Mostly Automated Tuning).

### 1.1.3 Methods optimizing the Neyman-Pearson criterion

Another category falling into the framework of (1.2) and (1.5) is the Neyman-Pearson problem which is closely related to hypothesis testing, where null  $H_0$  and alternative  $H_1$  hypotheses are given. Type I error occurs when  $H_0$  is true but is rejected, and type II error happens when  $H_0$  is false, but it fails to be rejected. The standard technique is to minimize Type II error while a bound for Type I error is given.

In the Neyman-Pearson problem, the null hypothesis  $H_0$  states that a sample  $\mathbf{x}$  has the negative label. Then Type I error corresponds to false-positives while Type II error to false-negatives. If the bound on Type I error equals  $\tau$ , we may write this as

$$t_1^{\text{NP}}(\mathbf{w}) = \max \{ t \mid \text{fp}(\mathbf{w}, t) \geq n^- \tau \}. \quad (1.17)$$

Then, we may write the Neyman-Pearson problem as

$$\begin{aligned} & \text{minimize} && \frac{1}{n^+} \text{fn}(\mathbf{w}, t) \\ & \text{subject to} && t \text{ is Type I error at level } \tau: \text{ it solves (1.17)}. \end{aligned} \quad (1.18)$$

Since (1.18) differs from (1.11) only by counting only the false-positives in (1.17) instead of counting all positives in (1.9), we can derive its approximations in exactly the same

## 1.2 Theoretical Analysis of the Framework

way as in Section 1.1.2. We therefore provide only their brief description and start with approximations of (1.17)

$$t_2^{\text{NP}}(\mathbf{w}) = \frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} s_{[i]}^-, \quad (1.19)$$

$$t_3^{\text{NP}}(\mathbf{w}) \quad \text{solves} \quad \frac{1}{n} \sum_{i=1}^{n^-} l(\beta(s_i^- - t)) = \tau. \quad (1.20)$$

Replacing the true counts by their surrogates results in the Neyman-Pearson variant *Grill-NP*

$$\begin{aligned} & \text{minimize} \quad \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{1}{n^-} \overline{\text{fp}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad t \text{ is the Neyman-Pearson threshold: it solves (1.17).} \end{aligned} \quad (1.21)$$

Similarly, the Neyman-Pearson alternative to *TopMeanK* reads

$$\begin{aligned} & \text{minimize} \quad \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad t = \frac{1}{n^- \tau} (s_{[1]}^- + \dots + s_{[n^- \tau]}^-), \\ & \quad \text{components of } \mathbf{s}^- \text{ equal to } s^- = \mathbf{w}^\top \mathbf{x}^- \text{ for } \mathbf{x}^- \in \mathcal{X}. \end{aligned} \quad (1.22)$$

This problem already appeared in [14] under the name  $\tau$ -FPL. Finally, *Pat&Mat-NP* reads

$$\begin{aligned} & \text{minimize} \quad \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad t \text{ is the surrogate Neyman-Pearson threshold: it solves (1.20).} \end{aligned} \quad (1.23)$$

We may see (1.22) from two different viewpoints. First,  $\tau$ -FPL provide convex approximations of *Grill-NP*. Second,  $\tau$ -FPL has the same form as *TopPushK*. The only difference is that for  $\tau$ -FPL we have  $k = n^- \tau$  while for *TopPushK* the value of  $k$  is small. Thus, even though we started from two different problems, we arrived at two approximations which differ only in the value of one parameter. This shows a close relation of the ranking problem and the Neyman-Pearson problem and the need for a unified theory to handle these problems.

## 1.2 Theoretical Analysis of the Framework

In this section, we provide a theoretical analysis of the unified framework from Section 1.1. We consider purely the problem *formulations* and not individual *algorithms* which specify how to solve these formulations. We focus mainly on the following desirable properties:

- *Convexity* implies a guaranteed convergence for many optimization algorithms or their better convergence rates [15].
- *Differentiability* increases the speed of convergence.
- *Stability* is a general term, by which we mean that the global minimum is not at  $\mathbf{w} = \mathbf{0}$ . This actually happens for many formulations from Section 1.1 and results in the situation where the separating hyperplane is degenerate and does not actually exist.



For a nicer flow of text, we show the results only for formulations from Section 1.1.2. The results for methods from Section 1.1.3 are identical. For the same reason, we postpone the proofs to Appendix A.1.

### 1.2.1 Threshold value comparison

We start with the following proposition, which compares the threshold approximation quality.

#### Proposition 1.1: [14]

We always have

$$t_1(\mathbf{w}) \leq t_2(\mathbf{w}) \leq t_3(\mathbf{w}).$$

Whenever the objective contains only false-negatives, a lower threshold  $t$  means a lower objective function. Therefore, a lower threshold is preferred.

### 1.2.2 Convexity

Convexity is one of the most important properties in numerical optimization. It ensures that the optimization problem has neither stationary points nor local minima. All points of interest are global minima. Moreover, it allows for faster convergence rates. We present the following two results.

#### Proposition 1.2

Thresholds  $t_2$  and  $t_3$  are convex functions of the weights  $\mathbf{w}$ . The threshold function  $t_1$  is non-convex.

#### Theorem 1.3

If the threshold  $t$  is a convex function of the weights  $\mathbf{w}$ , then function  $f(\mathbf{w}) = \overline{\text{fn}}(\mathbf{w}, t(\mathbf{w}))$  is convex.

While the proof of Theorem 1.3 is simple, it points to the necessity of considering only false-negatives in the objective of the problems in Section 1.1. In such a case, *TopPush*, *TopPushK*, *TopMeanK*,  $\tau$ -FPL, *Pat&Mat* and *Pat&Mat-NP* are convex problems. At the same time, *Grill* and *Grill-NP* are not convex problems.

### 1.2.3 Differentiability

Similarly to convexity, differentiability allows for faster convergence rate and in some algorithms, better termination criteria. The next theorem shows which formulations are differentiable.

#### Theorem 1.4

If the surrogate function  $l$  is differentiable, then threshold  $t_3$  is a differentiable func-

## 1.2 Theoretical Analysis of the Framework

tion of the weights  $\mathbf{w}$  and its derivative equals to

$$\nabla t_3(\mathbf{w}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}} l'(\beta(\mathbf{w}^\top \mathbf{x} - t_3(\mathbf{w}))) \mathbf{x}}{\sum_{\mathbf{x} \in \mathcal{X}} l'(\beta(\mathbf{w}^\top \mathbf{x} - t_3(\mathbf{w})))}.$$

The threshold functions  $t_1$  and  $t_2$  are non-differentiable.

This theorem shows that the objective functions of *Pat&Mat* and *Pat&Mat-NP* are differentiable. This allows us to prove the convergence of the stochastic gradient descent for these two formulations in Section 1.3.

### 1.2.4 Stability

We first provide a simple example and show that many formulations from the previous section are degenerate for it. Then we analyze general conditions under which this degenerate behaviour happens.

#### Example of a Degenerate Behavior

We consider  $n$  negative samples uniformly distributed in  $[-1, 0] \times [-1, 1]$ ,  $n$  positive samples uniformly distributed in  $[0, 1] \times [-1, 1]$  and one negative sample at  $(2, 0)$ , see Figure 1.1 (left). We consider the hinge loss and no regularization. If  $n$  is large, the point at  $(2, 0)$  is an outlier and the dataset is separable and the separating hyperplane has the normal vector  $\mathbf{w} = (1, 0)$ .

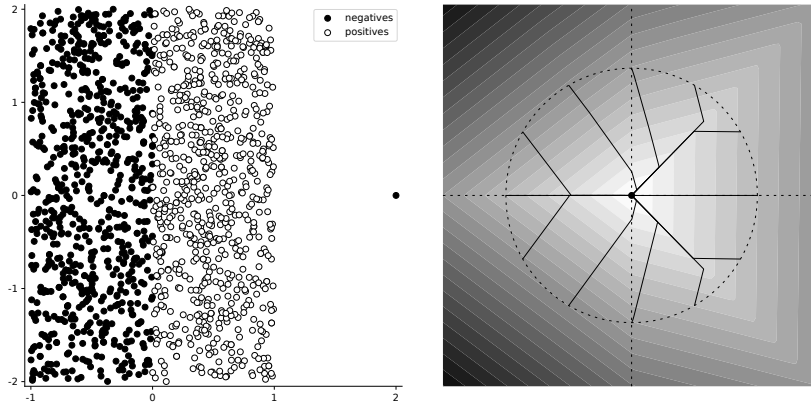


Figure 1.1: Left: distribution of positive (empty circle) and negative samples (full circles) for the example from Section 1.2.4. Right: contour plot for *TopPush* and its convergence to the zero vector from 12 initial points.

Table 1.1 shows the threshold  $t$  and the objective value  $f$  for two points  $\mathbf{w}_1 = (0, 0)$  and  $\mathbf{w}_2 = (1, 0)$ . These two points are both important:  $\mathbf{w}_1$  does not generate any separating hyperplane, while  $\mathbf{w}_2$  generates the optimal separating hyperplane. We show the precise computation in Appendix A.2. Since the dataset is perfectly separable by  $\mathbf{w}_2$ , we expect that  $\mathbf{w}_2$  provides a lower objective than  $\mathbf{w}_1$ . By shading the better objective in Table 1.1 by grey, we see that this did not happen for *TopPush* and *TopMeanK*.

It can be shown that  $\mathbf{w}_1 = (0, 0)$  is even the global minimum for *TopPush* and *TopMeanK*. This raises the question of whether some tricks, such as early stopping or excluding a small ball around zero, cannot overcome this difficulty. The answer is negative

Table 1.1: Comparison of formulations on the very simple problem from Section 1.2.4. Two formulations have the global minimum (denoted by grey color) at  $\mathbf{w}_1 = (0, 0)$  which does not generate any separating hyperplane. The optimal separating hyperplane is generated by  $\mathbf{w}_2 = (1, 0)$ .

Name	Label	$\mathbf{w}_1 = (0, 0)$		$\mathbf{w}_2 = (1, 0)$	
		$t$	$f$	$t$	$f$
<i>TopPush</i>	(1.7)	0	1	2	2.5
<i>TopPushK</i>	(1.8)	0	1	$\frac{2}{k}$	$0.5 + \frac{2}{k}$
<i>Grill</i>	(1.14)	0	2	$1 - 2\tau$	$1.5 + 2\tau(1 - \tau)$
<i>TopMeanK</i>	(1.15)	0	1	$1 - \tau$	$1.5 - \tau$
<i>Pat&amp;Mat</i>	(1.16)	$\frac{1}{\beta}(1 - \tau)$	$1 + \frac{1}{\beta}(1 - \tau)$	$\frac{1}{\beta}(1 - \tau)$	$0.5 + \frac{1}{\beta}(1 - \tau)$

as shown in Figure 1.1 (right). Here, we run *TopPush* from several starting points, and it always converges to zero from one of the three possible directions; all of them far from the normal vector to the separating hyperplane.

### Stability and Global minimum at zero

The convexity derived in the previous section guarantees that there are no local minima. However, as we showed in the example above, the global minimum may be at  $\mathbf{w} = \mathbf{0}$ . This is highly undesirable since  $\mathbf{w}$  is the normal vector to the separating hyperplane and the zero vector provides no information. In this section, we analyze when this situation happens. The first result states that if the threshold  $t(\mathbf{w})$  is above a certain value, then zero has a better objective than  $\mathbf{w}$ . If this happens for all  $\mathbf{w}$ , then zero is the global minimum.

#### Theorem 1.5

Consider any of these formulations: *TopPush*, *TopPushK*, *TopMeanK* or  $\tau$ -FPL. Fix any  $\mathbf{w}$  and denote the corresponding threshold  $t(\mathbf{w})$ . If we have

$$t(\mathbf{w}) \geq \frac{1}{n^+} \sum_{\mathbf{x}^+ \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x}^+, \quad (1.24)$$

then  $f(\mathbf{0}) \leq f(\mathbf{w})$ . Specifically, denote the scores  $s^+ = \mathbf{w}^\top \mathbf{x}^+$  for  $\mathbf{x}^+ \in \mathcal{X}^+$  and  $s^- = \mathbf{w}^\top \mathbf{x}^-$  for  $\mathbf{x}^- \in \mathcal{X}^-$  and the ordered variants with decreasing components of  $\mathbf{s}^-$  by  $\mathbf{s}^-_{[i]}$ . Then

$$\begin{aligned} s^-_{[1]} &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \textit{TopPush}, \\ \frac{1}{k} \sum_{i=1}^k s^-_{[i]} &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \textit{TopPushK}, \\ \frac{1}{n-\tau} \sum_{i=1}^{n-\tau} s^-_{[i]} &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \tau\text{-FPL}. \end{aligned} \quad (1.25)$$

## 1.2 Theoretical Analysis of the Framework

We can use this result immediately to deduce that some formulations have the global minimum at  $\mathbf{w} = \mathbf{0}$ . More specifically, *TopPush* fails if there are outliers, and *TopMeanK* fails whenever there are many positive samples.

### Corollary 1.6

Consider the *TopPush* formulation. If the positive samples lie in the convex hull of negative samples, then  $\mathbf{w} = \mathbf{0}$  is the global minimum.

### Corollary 1.7

Consider the *TopMeanK* formulation. If  $n^+ \geq n\tau$ , then  $\mathbf{w} = \mathbf{0}$  is the global minimum.

The proof of Theorem 1.5 employs the fact that all formulations in the theorem statement have only false-negatives in the objective. If  $\mathbf{w}_0 = \mathbf{0}$ , then  $\mathbf{w}_0^\top \mathbf{x} = 0$  for all samples  $\mathbf{x}$ , the threshold equals to  $t = 0$  and the objective equals to one. If the threshold is large for  $\mathbf{w}$ , many positives are below the threshold, and the false-negatives have the average surrogate value larger than one. In such a case,  $\mathbf{w}_0 = \mathbf{0}$  becomes the global minimum. There are two fixes to this situation:

- Include false-positives to the objective. This approach is taken by *Grill* and *Grill-NP* and necessarily results in the loss of convexity.
- Move the threshold away from zero even when all scores  $\mathbf{w}^\top \mathbf{x}$  are zero. This approach is taken by our formulations *Pat&Mat* and *Pat&Mat-NP* and keeps convexity.

The next theorem shows the advantage of the second approach.

### Theorem 1.8

Consider the *Pat&Mat* or *Pat&Mat-NP* formulation with the hinge surrogate and no regularization. Assume that for some  $\mathbf{w}$  we have

$$\frac{1}{n^+} \sum_{\mathbf{x}^+ \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x}^+ > \frac{1}{n^-} \sum_{\mathbf{x}^- \in \mathcal{X}^-} \mathbf{w}^\top \mathbf{x}^-. \quad (1.26)$$

Then there is a scaling parameter  $\beta_0$  from (1.13) such that  $f(\mathbf{w}) < f(\mathbf{0})$  for all  $\beta \in (0, \beta_0)$ .

These theorem shed some light on the behaviour of the formulations. Theorem 1.5 states that the stability of  $\tau$ -FPL requires

$$\frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} s_{[i]}^- < \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+, \quad (1.27)$$

while Theorem 1.8 states that the stability of *Pat&Mat-NP* is ensured by

$$\frac{1}{n^-} \sum_{i=1}^{n^-} s_{[i]}^- < \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+. \quad (1.28)$$

The right-hand sides of (1.27) and (1.28) are the same, while the left-hand side of (1.28) is always smaller than the left-hand side of (1.27). This implies that if  $\tau$ -FPL is stable, then *Pat&Mat-NP* is stable as well.

At the same time, there may be a huge difference in the stability of both formulations. Since the scores of positive samples should be above the scores of negative samples, the scores  $s$  may be interpreted as performance. Then formula (1.27) states that if the mean performance of a *small number of the best* negative samples is larger than the average performance of *all* positive samples, then  $\tau$ -FPL fails. On the other hand, formula (1.28) states that if the average performance of *all* positive samples is better than the average performance of *all* negative samples, then *Pat&Mat-NP* is stable. The former may well happen as accuracy at the top is interested in a good performance of only a small number of positive samples.

### 1.2.5 Method comparison

We provide a summary of the obtained results in Table 1.2. There we give basic characterizations of the formulations such as their definition label, their source, the hyperparameters, whether the formulation is differentiable and convex, and whether it has stability problems with  $\mathbf{w} = \mathbf{0}$  being the global minimum.

Table 1.2: Summary of the formulations from Section 1.1. The table shows their definition label, the source or the source they are based on, the hyperparameters, whether the formulation is differentiable, convex and stable (in the sense of having problems with  $\mathbf{w} = \mathbf{0}$ ).

Name	Source	Definition	Hyperpars	Convex	Differentiable	Stable
<i>TopPush</i>	[5]	(1.7)	$\lambda$	✓	✗	✗
<i>TopPushK</i>	ours	(1.8)	$\lambda, k$	✓	✗	✗
<i>Grill</i>	[1]	(1.14)	$\lambda$	✗	✗	✓
<i>Pat&amp;Mat</i>	ours	(1.16)	$\beta, \lambda$	✓	✓	✓
<i>TopMeanK</i>	-	(1.15)	$\lambda$	✓	✗	✗
<i>Grill-NP</i>	-	(1.21)	$\lambda$	✗	✗	✓
<i>Pat&amp;Mat-NP</i>	ours	(1.23)	$\beta, \lambda$	✓	✓	✓
$\tau$ -FPL	[14]	(1.22)	$\lambda$	✓	✗	✗

A similar comparison is performed in Figure 1.2. Methods in green and grey are convex, while formulations in white are non-convex. Based on Theorem 1.5, four formulations in grey are vulnerable to have the global minimum at  $\mathbf{w} = \mathbf{0}$ . This theorem states that the higher the threshold, the more vulnerable the formulation is. The full arrows depict this dependence. If it points from one formulation to another, the latter one has a smaller threshold and thus is less vulnerable to this undesired global minima. The dotted arrows indicate that this holds usually but not always, the precise formulation is provided in Appendix A.1.5. This complies with Corollaries 1.6 and 1.7 which state that *TopPush* and *TopMeanK* are most vulnerable. At the same time, it says that  $\tau$ -FPL is the best one from the grey-coloured formulations. Finally, even though *Pat&Mat-NP* has a worse approximation of the true threshold than  $\tau$ -FPL due to Theorem 1.5, it is more stable due to the discussion after Theorem 1.8.

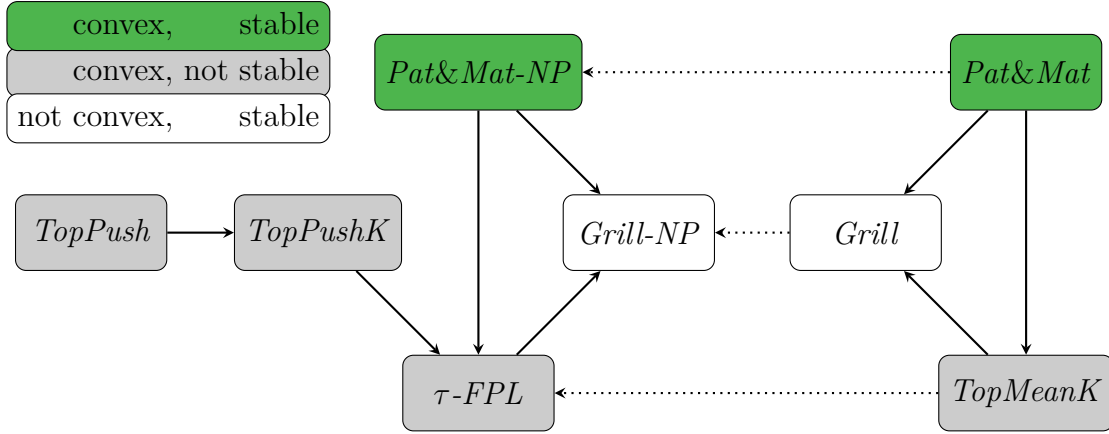


Figure 1.2: Summary of the formulations from Section 1.1. Methods in green and grey are convex, while formulations in white are non-convex. Methods in grey are vulnerable to have the global minimum at  $\mathbf{w} = 0$ . Full (dotted) arrow pointing from one formulation to another show that the latter formulation has always (usually) smaller threshold.

## 1.3 Convergence of stochastic gradient descent

The previous section analyzed the formulations from Section 1.1 but did not consider any optimization algorithms. In this section, we show a basic version of the stochastic gradient descent and then show its convergent version. Since due to considering the threshold, gradient computed on a minibatch is a biased estimate of the true gradient, we need to use variance reduction techniques, and the proof is rather complex.

### 1.3.1 Stochastic gradient descent: Basic

Many optimization algorithms for solving the formulations from Section 1.1 use primal-dual or purely dual formulations. [9] introduced dual variables and used alternating optimization to the resulting min-max problem. [5] and [14] dualized the problem and solved it with the steepest gradient ascent. [16] followed the same path but added kernels to handle non-linearity. We follow the ideas of [13] and [17] and solve the problems directly in their primal formulations. Therefore, even though we use the same formulation for *TopPush* as [5] or for  $\tau$ -FPL as [14], our solution process is different. However, due to convexity, both algorithms should converge to the same point.

The decision variables in (1.5) are the normal vector of the separating hyperplane  $\mathbf{w}$  and the threshold  $t$ . To apply an efficient optimization method, we need to compute gradients. The simplest idea [1] is to compute the gradient only with respect to  $\mathbf{w}$  and then recompute  $t$ . A more sophisticated way is based on the chain rule. For each  $\mathbf{w}$ , the threshold  $t$  can be computed uniquely. We stress this dependence by writing  $t(\mathbf{w})$  instead of  $t$ . By doing so, we effectively remove the threshold  $t$  from the decision variables and  $\mathbf{w}$  remains the only decision variable. Note that the convexity is preserved. Then we can compute the derivative via the chain rule

$$\begin{aligned} f(\mathbf{w}) &= \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l(t(\mathbf{w}) - \mathbf{w}^\top \mathbf{x}) + \frac{\lambda}{2} \|\mathbf{w}\|^2, \\ \nabla f(\mathbf{w}) &= \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l'(t(\mathbf{w}) - \mathbf{w}^\top \mathbf{x})(\nabla t(\mathbf{w}) - \mathbf{x}) + \lambda \mathbf{w}. \end{aligned} \tag{1.29}$$

The only remaining part is the computation of  $\nabla t(\mathbf{w})$ . It is simple for  $\nabla t_1(\mathbf{w})$  and  $\nabla t_2(\mathbf{w})$  and Theorem 1.4 shows the computation for  $\nabla t_3(\mathbf{w})$ . Appendix A.3 provides an efficient computation method for  $t_3(\mathbf{w})$ .

Having derivative (1.29), deriving the stochastic gradient is simple. It partitions the dataset into minibatches and provides an update of the weights  $\mathbf{w}$  based only on a minibatch, namely by replacing the mean over the whole dataset in (1.29) by a mean over the minibatch.

### 1.3.2 Stochastic gradient descent: Convergent for *Pat&Mat* and *Pat&Mat-NP*

For the convergence proof, we need differentiability which is due to Theorem 1.4 possessed only by *Pat&Mat* and *Pat&Mat-NP*. Therefore, we consider only these two formulations and for simplicity, show it only for *Pat&Mat*. We apply a variance reduction technique based on delayed values similar to SAG [18].

At iteration  $k$  we have the decision variable  $\mathbf{w}^k$  and the active minibatch  $I^k$ . First, we update the score vector  $\mathbf{s}^k$  only on the active minibatch by setting

$$s_i^k = \begin{cases} \mathbf{x}_i^\top \mathbf{w}^k & \text{for all } i \in I^k, \\ s_i^{k-1} & \text{for all } i \notin I^k. \end{cases} \quad (1.30)$$

We keep scores from previous minibatches intact. We use Appendix A.3 to compute the surrogate quantile  $t^k$  as the unique solution of

$$\sum_{i \in X} l(\beta(s_i^k - t^k)) = n\tau. \quad (1.31)$$

This is an approximation of the surrogate quantile  $t(\mathbf{w}^k)$  from (1.13). The only difference from the true value  $t(\mathbf{w}^k)$  is that we use delayed scores. Then we introduce artificial variable

$$\mathbf{a}^k = \sum_{i \in I^k} l'(\beta(s_i^k - t^k)) \mathbf{x}_i. \quad (1.32)$$

Finally, we approximate the derivative  $\nabla f(\mathbf{w}^k)$  from (1.29) by

$$g(\mathbf{w}^k) = \frac{1}{n_+^k} \sum_{i \in I_+^k} l'(t^k - s_i^k) (\nabla t^k - \mathbf{x}_i), \quad (1.33)$$

where  $\nabla t^k$  is an approximation of  $\nabla t(\mathbf{w}^k)$  from Theorem 1.4 defined by

$$\nabla t^k = \frac{\mathbf{a}^k + \mathbf{a}^{k-1} + \dots + \mathbf{a}^{k-m+1}}{\sum_{i \in X} l'(\beta(s_i^k - t^k))}. \quad (1.34)$$

A perhaps more straightforward possibility would be to consider only  $\mathbf{a}^k$  in the numerator of (1.34). However, choice (1.34) enables us to prove the convergence and it adds stability to the algorithm for small minibatches.

The whole procedure does not perform any vector operations outside of the current minibatch  $I^k$ . We summarize it in Algorithm 1.1. Note that a proper initialization for the first  $m$  iterations is needed. We finish the theoretical part by the convergence proof.

## 1.4 Numerical experiments

---

**Algorithm 1.1** Stochastic gradient descent for maximizing accuracy at the top

---

**Require:** Dataset  $X$ , Minibatches  $I^1, \dots, I^m$ , Stepsize  $\alpha^k$

```
1: Initialize weights  $\mathbf{w}^0$ 
2: for  $k = 0, 1, \dots$  do
3:   Select a minibatch  $I^k$ 
4:   Compute  $s_i^k$  for all  $i \in I^k$  according to (1.30)
5:   Compute  $t^k$  according to (1.31)
6:   Compute  $\mathbf{a}^k$  according to (1.32)
7:   Compute  $\nabla t^k$  according to (1.34)
8:   Compute  $g(\mathbf{w}^k)$  according to (1.33)
9:   Set  $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \alpha^k g(\mathbf{w}^k)$ 
10: end for
```

---

### Theorem 1.9

Consider the *Pat&Mat* or *Pat&Mat-NP* formulation, stepsizes  $\alpha^k = \frac{\alpha^0}{k+1}$  and piecewise disjoint minibatches  $I^1, \dots, I^m$  which cycle periodically  $I^{k+m} = I^k$ . If  $l$  is the smoothened (Huberized) hinge function, then Algorithm 1.1 converges to the global minimum of (1.16).

## 1.4 Numerical experiments

In this section, we present numerical results.

### 1.4.1 Implementational details and Hyperparameter choice

We recall that all methods fall into the framework of either (1.2) or (1.5). Since the threshold  $t$  depends on the weights  $\mathbf{w}$ , we can consider the decision variable to be only  $\mathbf{w}$ . Then to apply a method, we implemented the following iterative procedure. At iteration  $j$ , we have the weights  $\mathbf{w}^j$  to which we compute the threshold  $t^j = t(\mathbf{w}^j)$ . Then according to (1.29), we compute the gradient of the objective and apply the ADAM descent scheme [19]. All methods were run for 10000 iterations using the stochastic gradient descent. The minibatch size was 512 except for the sigillito1989classification and Spambase datasets where the full gradient was used. All methods used the hinge surrogate (1.3). The initial point is generated randomly.

We run the methods for the following hyperparameters

$$\begin{aligned}\beta &\in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}, \\ \lambda &\in \{0, 0.00001, 0.0001, 0.001, 0.01, 0.1\}, \\ k &\in \{1, 3, 5, 10, 15, 20\}.\end{aligned}\tag{1.35}$$

For *TopPushK*, *Pat&Mat* and *Pat&Mat-NP* we fixed  $\lambda = 0.001$  to have six hyperparameters for all methods. For all datasets, we choose the hyperparameter which minimized the criterion on the validation set. The results are computed on the testing set which was not used during training the methods.



*TopPush* and  $\tau$ -FPL were originally implemented in the dual. However, to allow for the same framework and the stochastic gradient descent, we implemented it in the primal. These two approaches are equivalent.

### 1.4.2 Dataset description and Performance criteria

For the numerical results, we considered 10 datasets summarized in Table 1.3. They can be downloaded from the UCI repository. *sigillito1989classification* [20] and *Spambase* are small, *baldi2016parameterized* [21] contains a large number of samples while *guyon2005result* [22] contains a large number of features. We also considered six visual recognition datasets: MNIST, FashionMNIST, CIFAR10, CIFAR20, CIFAR100 and SVHN2. MNIST and FashionMNIST are grayscale datasets of digits and fashion items, respectively. CIFAR100 is a dataset of coloured images of items grouped into 100 classes. CIFAR10 and CIFAR20 merge these classes into 10 and 20 superclasses, respectively. SVHN2 contains coloured images of house numbers. As Table 1.3 shows, these datasets are imbalanced.

Each of the visual recognition datasets was converted into ten binary datasets by considering one of the classes  $\{0, \dots, 9\}$  as the positive class and the rest as the negative class. The experiments were repeated ten times for each dataset from different seeds, which influenced the starting point and minibatch creation. We use  $\text{tpr@fpr}$  as the evaluation criterion. This describes the true-positive rate at a prescribed true-negative rate, usually of 1% or 5%. For the linear classifier  $\mathbf{w}^\top \mathbf{x} - t$ , it selects the threshold  $t$  so that the desired true-negative rate is satisfied and then computes the true-positive rate for this threshold.

Table 1.3: Structure of the used datasets. The training, validation and testing sets show the number of features  $m$ , samples  $n$  and the fraction of positive samples  $\frac{n^+}{n}$ .

	$m$	Training		Validation		Testing	
		$n$	$\frac{n^+}{n}$	$n$	$\frac{n^+}{n}$	$n$	$\frac{n^+}{n}$
Ionosphere	34	175	36.0%	88	36.4%	88	35.2%
Spambase	57	2300	39.4%	1150	39.4%	1151	39.4%
Gisette	5000	1000	50.0%	1500	50.0%	500	50.0%
Hepmass	28	5250000	50.0%	1750000	50.0%	3500000	50.0%
MNIST	$28 \times 28 \times 1$	44999	11.2%	15001	11.2%	10000	11.4%
FashionMNIST	$28 \times 28 \times 1$	45000	10.0%	15000	10.0%	10000	10.0%
CIFAR10	$32 \times 32 \times 3$	37500	10.0%	12500	10.0%	10000	10.0%
CIFAR20	$32 \times 32 \times 3$	37500	5.0%	12500	5.0%	10000	5.0%
CIFAR100	$32 \times 32 \times 3$	37500	1.0%	12500	1.0%	10000	1.0%
SVHN2	$32 \times 32 \times 3$	54944	18.9%	18313	18.9%	26032	19.6%

### 1.4.3 Numerical results

Figure 1.3 presents the standard ROC (receiver operating characteristic) curves on selected datasets. Since all methods from this paper are supposed to work at low false-positive rates, the  $x$  axis is logarithmic. Both figures depict averages over ten runs with different seeds. The left column depicts CIFAR100 while the right one Hepmass. These are

## 1.4 Numerical experiments

the two more complicated datasets. We selected four representative methods: *Pat&Mat* and *Pat&Mat-NP* as our methods and *TopPush* and  $\tau$ -*FPL* as state-of-the-art methods. Even though all methods work well, *Pat&Mat-NP* seems to outperform the remaining methods on most levels of false-positive rate.

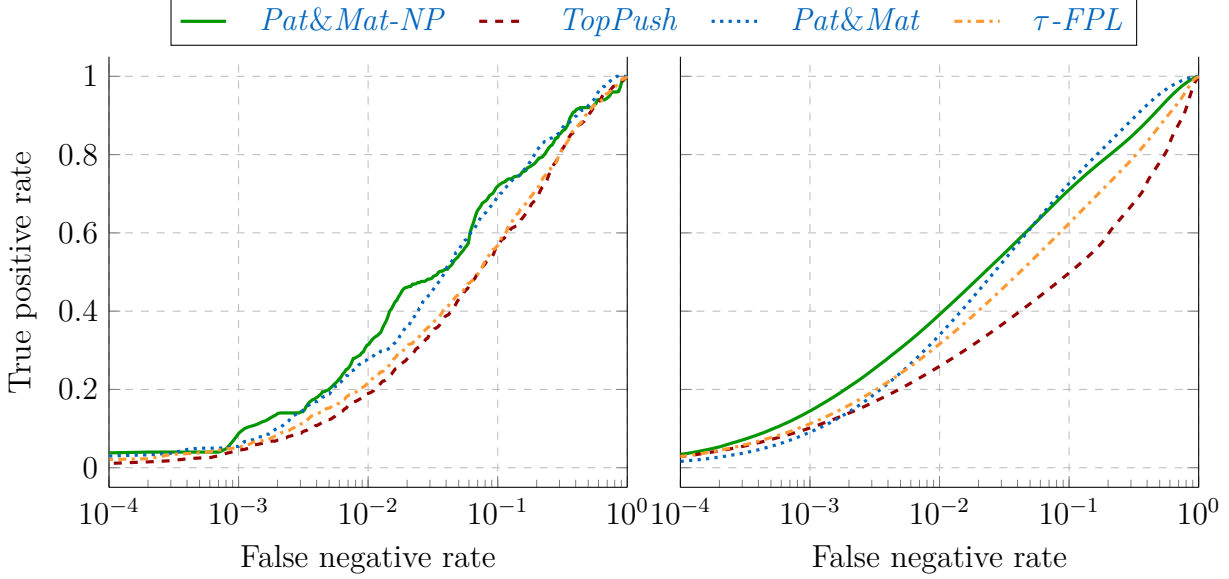


Figure 1.3: ROC curves (with logarithmic  $x$  axis) on CIFAR100 (left) and Hepmass (right).

While Figure 1.3 gave a glimpse of the behaviour of methods, Figures 1.4 and 1.5 provide a statistically more sound comparison. It employs the Nemenyi post hoc test for the Friedman test recommended in [23]. This test compares if the mean ranks of multiple methods are significantly different.

We consider 14 methods (we count different values of  $\tau$  as different methods) as depicted in this table. For each dataset mentioned in Section 1.4.2 and each method, we evaluated the  $\text{fpr@tpr}$  metric and ranked all methods. Rank 1 refers to the best performance for given criteria, while rank 14 is the worst. The  $x$ -axis shows the average rank over all datasets. The Nemenyi test computes the critical difference. If two methods are within their critical difference, their performance is not deemed to be significantly different. Black wide horizontal lines group such methods.

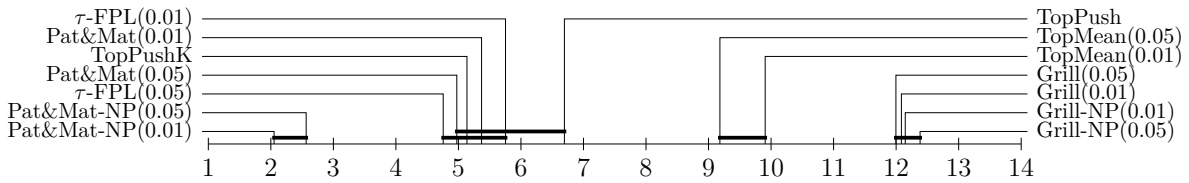


Figure 1.4: Critical difference (CD) diagrams (level of importance 0.05) of the Nemenyi post hoc test for the Friedman test. Each diagram shows the mean rank of each method, with rank 1 being the best. Black wide horizontal lines group together methods with the mean ranks that are not significantly different. The critical difference diagrams were computed for mean rank averages over all datasets of the  $\text{tpr@fpr}$  ( $\tau = 0.01$ ) metric.

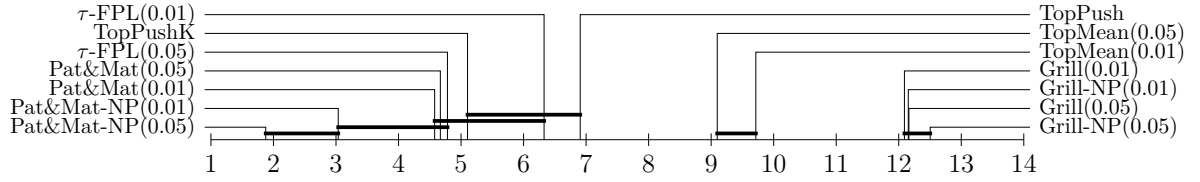


Figure 1.5: Critical difference (CD) diagrams (level of importance 0.05) of the Nemenyi post hoc test for the Friedman test. Each diagram shows the mean rank of each method, with rank 1 being the best. Black wide horizontal lines group together methods with the mean ranks that are not significantly different. The critical difference diagrams were computed for mean rank averages over all datasets of the  $\text{tpr@fpr}$  ( $\tau = 0.05$ ) metric.

From this figure and table, we make several observations:

- *TopPushK* (rank 5.1) provides a slight improvement over *TopPush* (rank 6.7) even though this improvement is not statistically significant as both methods are connected by the black line in both Figures 1.4 and 1.5.
- Neither *Grill* (ranks 12.0 and 12.1) nor *Grill-NP* (ranks 12.1 and 12.4) perform well. We believe this happened due to the lack of convexity as indicated in Theorem 1.3 and the discussion after that.
- *TopMeanK* (ranks 9.2 and 9.9) does not perform well either. Since the thresholds  $\tau$  are small, then  $\mathbf{w} = 0$  is the global minimum as proved in Corollary 1.7.
- *Pat&Mat-NP* (rank 2.1 and 2.6) seems to outperform other methods.
- *Pat&Mat* (ranks 5.0 and 5.4),  $\tau$ -FPL (ranks 4.8 and 5.8) and *TopPushK* (rank 5.1) perform similarly. Since they are connected, there is no statistical difference between their behaviours.
- *Pat&Mat-NP* at level 0.01 (rank 2.1) outperforms *Pat&Mat-NP* at level 0.05 (rank 2.6) for  $\tau = 0.01$ . *Pat&Mat-NP* at level 0.05 (rank 1.9 in Figure 1.5) outperforms *Pat&Mat-NP* at level 0.01 (rank 3.0 in Figure 1.5) for  $\tau = 0.05$ . This should be because these methods are optimized for the corresponding threshold. For  $\tau$ -FPL we observed this behaviour for Figure 1.5 but not for Figure 1.4.

Figure 1.6 provides a similar comparison. Both axes are sorted from the best (left) to the worst (right) average ranks. The numbers in the graph show the  $p$ -value for the pairwise Wilcoxon signed-rank test, where the null hypothesis is that the mean  $\text{tpr@fpr}$  of both methods is the same. Even though Figure 1.4 employs a comparison of mean ranks and Figure 1.6 a pairwise comparison of  $\text{fpr@tpr}$ , the results are almost similar. Methods grouped by the black line in the former figure usually show a large  $p$ -value in the latter figure.

Table 1.4 investigates the impact of  $\mathbf{w} = 0$  as a potential global minimum. Each method was optimized for six different values of hyperparameters. The table depicts the condition under which the final value has a lower objective than  $\mathbf{w} = 0$ . Thus, ✓ means that it is always better while ✗ means that the algorithm made no progress from the starting point  $\mathbf{w} = 0$ . The latter case implies that  $\mathbf{w} = 0$  seems to be the global minimum. We make the following observations:

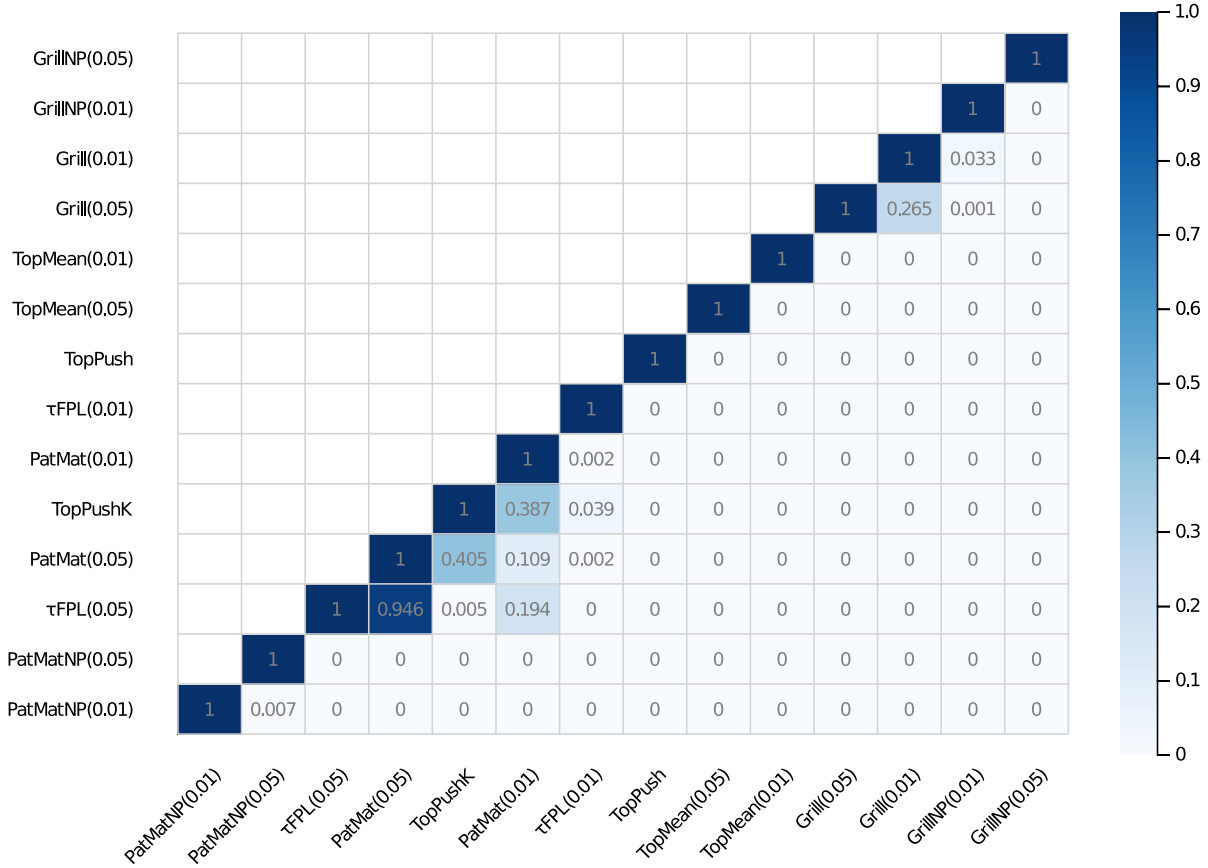


Figure 1.6: The  $p$ -value for the pairwise Wilcoxon signed-rank test, where the null hypothesis is that the mean  $\text{tpr@fpr}(0.01)$  of both methods is the same. The methods are sorted by mean rank (left = better).

- *Pat&Mat* and *Pat&Mat-NP* are the only methods which succeeded at every dataset for some hyperparameter. Moreover, for each dataset, there was some  $\beta_0$  such that these methods were successful if and only if  $\beta \in (0, \beta_0)$ . This is in agreement with Theorem 1.8.
- *TopMeanK* fails everywhere which agrees with Corollary 1.7.
- Figure 1.2 states that the methods from Section 1.1.2 has a higher threshold than their Neyman-Pearson variants from Section 1.1.3. This is documented in the table as the latter have a higher number of successes.

## 1.5 Conclusion

In this paper, we achieved the following results:

- We presented a unified framework for the three criteria from Section 1.1. These criteria include ranking, accuracy at the top and hypothesis testing.
- We showed that several known methods (*TopPush*, *Grill*,  $\tau$ -*FPL*) fall into our framework and derived some completely new methods (*Pat&Mat*, *Pat&Mat-NP*).

Table 1.4: Necessary hyperparameter choice for the solution to have a better objective than zero.  $\checkmark$  means that the solution was better than zero for all hyperparameters while  $\times$  means that it was worse for all hyperparameters.

	Ionosphere	Hepmass	FashionMNIST	CIFAR100
<i>TopPush</i>	$\checkmark$	$\times$	$\checkmark$	$\times$
<i>TopPushK</i>	$\checkmark$	$\times$	$\checkmark$	$\times$
<i>Grill</i> $\tau = 0.01$	$\times$	$\times$	$\times$	$\times$
$\tau = 0.05$	$\times$	$\times$	$\times$	$\times$
<i>Pat&amp;Mat</i> $\tau = 0.01$	$\checkmark$	$\beta \leq 0.1$	$\beta \leq 1$	$\beta \leq 1$
$\tau = 0.05$	$\checkmark$	$\beta \leq 1$	$\checkmark$	$\checkmark$
<i>TopMeanK</i> $\tau = 0.01$	$\times$	$\times$	$\times$	$\times$
$\tau = 0.05$	$\times$	$\times$	$\times$	$\times$
<i>Grill-NP</i> $\tau = 0.01$	$\times$	$\times$	$\times$	$\times$
$\tau = 0.05$	$\times$	$\times$	$\times$	$\times$
<i>Pat&amp;Mat-NP</i> $\tau = 0.01$	$\checkmark$	$\beta \leq 1$	$\checkmark$	$\beta \leq 1$
$\tau = 0.05$	$\checkmark$	$\checkmark$	$\checkmark$	$\beta \leq 1$
$\tau$ -FPL $\tau = 0.01$	$\checkmark$	$\times$	$\checkmark$	$\times$
$\tau = 0.05$	$\checkmark$	$\checkmark$	$\checkmark$	$\lambda \leq 0.001$

- We performed a theoretical analysis of the methods. We showed that known methods suffer from certain disadvantages. While *TopPush* and  $\tau$ -FPL are sensitive to outliers, *Grill* is non-convex. We proved the global convergence of the stochastic gradient descent for *Pat&Mat* and *Pat&Mat-NP*.
- We performed a numerical comparison and we showed a good performance of our method *Pat&Mat-NP*.



# Non-linear Classification at the Top

---

## 2.1 Introduction

The aim of classical linear binary classification is to separate positive and negative samples by a linear hyperplane. In many applications, it is desirable to separate only a certain number of samples. In such a case, the goal is not to maximize the performance on all samples but only the performance on the required samples with the highest relevance. Such classifiers have many applications. For example, in information retrieval systems, only the most relevant documents should be returned for a given query. Furthermore, they are useful in domains, where a large number of samples needs to be quickly screened and only a small subset of samples needs to be selected for further evaluation.

These problems can be generally written as pushing the positive samples above some decision threshold. The methods differ in the definition of the decision threshold. In our previous work [24], we introduced a general framework that unifies these methods. We showed that several problem classes, which were considered as separate problems so far, fit into the framework. As the most relevant we mention the following methods:

- *Ranking problems* focuses on ranking the positive samples higher than the negative ones. Many methods, such as *RankBoost* [2], *Infinite Push* [3] or *p-norm push* [4] employ a pairwise comparison of samples, which makes them infeasible for larger datasets. This was alleviated in *TopPush* [5] where the authors considered the limit  $p \rightarrow \infty$ . Since the  $l_\infty$  norm from *TopPush* is equal to the maximum, the decision threshold from our framework equals to the maximum of scores of negative samples. This was generalized into *TopPushK* [24] by considering the threshold to be the mean of  $K$  largest scores of negative samples.
- *Accuracy at the Top* [6] focuses on maximizing the number of positive samples above the top  $\tau$ -quantile of scores. There are many methods on how to solve accuracy at the top. In [6], the authors assume that the top quantile is one of the samples, construct  $n$  unconstrained optimization problems with fixed thresholds, solve them and select the best solution. This method is computationally expensive. In [1] the authors propose a fast projected gradient descent method. In our previous paper, we proposed a convex approximation of the accuracy at the top called *Pat&Mat*. This method is reasonably fast and guaranteed the existence of global optimum.

The deficiency of methods from this framework is that they usually cover only linear classifiers. However, as many problems are not linearly separable, nonlinear classifiers are needed. In this work, we show how to extend our framework into nonlinear classification problems. To do so, we use the fact that our framework is similar to the primal formulation

of support vector machines [25]. The classical way to incorporate nonlinearity into SVM is to derive the dual formulation [15] and to employ the kernels method [26]. In this work, we follow this approach, derive dual formulations for the considered problems and add nonlinear kernels to them. Moreover, as dual problems are generally expensive to solve, we derive a quick method to solve them. This is a modification of the coordinate-wise dual ascent from [27]. For a review of other approaches see [28, 29].

The paper is organized as follows: In Section 2.2 we recall the unified framework derived in [24] and two class of problems that falls into it. Moreover, for selected methods, we derive their dual formulations. Namely, we focus on *TopPush*, *TopPushK* and *Pat&Mat*. In Section 2.3, we show how to add nonlinear kernels into dual formulations, derive a new method for solving these dual problems and perform its complexity analysis. Since our method depends on the chosen problem and surrogate function, we provide a concrete form of the solution for *TopPushK* with the truncated quadratic loss. Solutions for other problems are provided in Appendix B.3 and B.4. Finally, in Section 2.4 we present the description of performance criteria, choice of hyperparameters and description of datasets. The rest of the section is focused on the results of numerical experiments. Here we compare all methods in terms of overall accuracy and accuracy at a given threshold. We also discuss the convergence and time consumption of all methods. All our codes are available online.<sup>1</sup>

## 2.2 Derivation of dual problems

Linear binary classification is a problem of finding a linear hyperplane that separates a group of positive samples from a group of negative samples and achieves the lowest possible error. For a sample  $\mathbf{x} \in \mathbb{R}^d$ , the prediction for a linear classifier amounts to

$$\mathbf{x} \text{ has } \begin{cases} \text{positive label} & \text{if } \mathbf{w}^\top \mathbf{x} \geq t, \\ \text{negative label} & \text{otherwise.} \end{cases}$$

Here,  $\mathbf{w} \in \mathbb{R}^d$  is the normal vector to the separating hyperplane and  $t \in \mathbb{R}$  is a decision threshold. The well-known example of such a classifier is a support vector machine [25] where the decision threshold  $t$  is a free variable. However, many important binary classification problems maximize the performance only for a certain amount of samples with the highest scores  $s = \mathbf{w}^\top \mathbf{x}$ . In these cases, the threshold  $t$  is not a free variable but a function of the scores. In our previous work [24], we formulated a general framework for maximizing performance above the threshold  $t$  as

$$\begin{aligned} \underset{\mathbf{w}, t}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} [t - \mathbf{w}^\top \mathbf{x}_i^+] \\ \text{subject to} \quad & \text{threshold } t \text{ is a function of } \{\mathbf{w}^\top \mathbf{x}_i\}_{i=1}^n, \end{aligned} \tag{2.1}$$

where  $C \in \mathbb{R}$  is a constant;  $[\cdot]$  is the 0 – 1 loss defined as 1 if the argument is true and 0 otherwise. To denote positive and negative samples, we use  $+$  and  $-$  symbols in the superscript, respectively. Note that  $[t - \mathbf{w}^\top \mathbf{x}_i^+]$  counts the number of positive samples  $\mathbf{x}_i^+$  whose score  $\mathbf{w}^\top \mathbf{x}_i^+$  is below the threshold  $t$ . Since the objective is to be minimized, the positive samples should lie above the threshold  $t$ .

<sup>1</sup>[https://github.com/VaclavMacha/ClassificationOnTop\\_nonlinear.jl](https://github.com/VaclavMacha/ClassificationOnTop_nonlinear.jl)



Since the objective function in (2.1) is discontinuous due to the 0 – 1 loss, the optimization problem (2.1) is difficult. A typical approach to remove this unwanted feature is to approximate the 0 – 1 loss by a surrogate function such as the truncated quadratic or the hinge functions

$$l_{\text{quadratic}}(s) = (\max\{0, 1 + \vartheta s\})^2, \quad (2.2)$$

$$l_{\text{hinge}}(s) = \max\{0, 1 + \vartheta s\}, \quad (2.3)$$

where  $\vartheta > 0$  is a scaling parameter. In the following text, we use the symbol  $l$  to denote any convex non-negative non-decreasing function with  $l(0) = 1$ . Replacing the 0 – 1 loss in (2.1) by a surrogate function results in

$$\begin{aligned} \underset{\mathbf{w}, t}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l(t - \mathbf{w}^\top \mathbf{x}_i^+) \\ \text{subject to} \quad & t \text{ is a function of } \{\mathbf{w}^\top \mathbf{x}_i\}_{i=1}^n. \end{aligned} \quad (2.4)$$

Note that the objective function in (2.4) is continuous.

As we derived in [24], there are many problems belonging to the general framework (2.1). However, this framework handles only linear classification problems. As many problems are not linearly separable, this is often not sufficient. To generalize the framework to nonlinear classifiers, we realize that (2.4) is similar to the primal formulation of the SVM [25]. We will follow the standard way to incorporate nonlinearity into SVM by deriving the dual problem [15] and using the kernels methods [26].

In the remainder of this section, we recall two problem classes from [24] and their convex approximations, and for each of them, we derive its dual formulation. Namely, we will discuss *TopPushK* and *Accuracy at the Top* with its convex approximation *Pat&Mat*.

### 2.2.1 TopPushK

The first problem *TopPushK* is our modification of the *TopPush* method introduced in [5]. It selects the threshold  $t$  as the mean of the scores corresponding to  $K$  highest ranked negatives. By doing so, it enforces the positives to be ranked above negatives. Therefore, both *TopPush* ( $K = 1$ ) and *TopPushK* ( $K > 1$ ) fall into the category of ranking problems.

Writing this more formally, define vector  $\mathbf{s}$  of all scores as  $s_i = \mathbf{w}^\top \mathbf{x}_i$  for all  $i = 1, 2, \dots, n$  and its sorted version  $\mathbf{s}_{[ ]}$  with decreasing components, i.e.  $s_{[1]} \geq s_{[2]} \geq \dots \geq s_{[n]}$ . Then *TopPushK* reads

$$\begin{aligned} \underset{\mathbf{w}, t, \mathbf{s}^-}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l(t - \mathbf{w}^\top \mathbf{x}_i^+) \\ \text{subject to} \quad & t = \frac{1}{K} \sum_{j=1}^K s_{[j]}^-, \\ & s_j^- = \mathbf{w}^\top \mathbf{x}_j^-, \quad \forall j = 1, 2, \dots, n^-. \end{aligned} \quad (2.5)$$

It can be showed that this problem is convex and that for  $K = 1$  we get the original *TopPush*.

In the following theorem, we denote the positive semidefinite kernel matrix  $\mathbb{K}$  by

$$\mathbb{K} = \begin{pmatrix} \mathbb{X}^+ \\ -\mathbb{X}^- \end{pmatrix} \begin{pmatrix} \mathbb{X}^+ \\ -\mathbb{X}^- \end{pmatrix}^\top = \begin{pmatrix} \mathbb{X}^+ \mathbb{X}^{+\top} & -\mathbb{X}^+ \mathbb{X}^{-\top} \\ -\mathbb{X}^- \mathbb{X}^{+\top} & \mathbb{X}^- \mathbb{X}^{-\top} \end{pmatrix}. \quad (2.6)$$

## 2.2 Derivation of dual problems

and show the form of the *TopPushK* dual problem<sup>2</sup>.

### Theorem 2.1: *TopPushK* dual formulation

The dual problem corresponding to the problem (2.5) has the form

$$\underset{\alpha, \beta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - C \sum_{i=1}^{n^+} l^* \left( \frac{\alpha_i}{C} \right) \quad (2.7a)$$

$$\text{subject to} \quad \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (2.7b)$$

$$0 \leq \beta_j \leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \quad \forall j = 1, 2, \dots, n^-, \quad (2.7c)$$

where  $l^*$  is a conjugate of the surrogate loss function  $l$  from (2.5) and  $\mathbb{K}$  was defined in (2.6).

### 2.2.2 Accuracy at the Top

The second problem *Accuracy at the Top* was introduced in [6]. On the contrary to *TopPushK*, which focuses on the minimization of the number of positive samples below  $K$  highest ranked negatives, the *Accuracy at the Top* minimizes the number of positive samples below the top  $\tau$ -quantile from negative samples defined as

$$t = \max \left\{ t \mid \sum_{j=1}^{n^-} [\mathbf{w}^\top \mathbf{x}_j^- - t] \geq n\tau \right\}. \quad (2.8)$$

Then *Accuracy at the Top* is an optimization problem written as follows

$$\underset{\mathbf{w}, t}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l_1(t - \mathbf{w}^\top \mathbf{x}_i^+) \quad (2.9)$$

subject to  $t$  is the surrogate top  $\tau$ -quantile: it solves (2.8).

Since it is known that the quantile function (2.8) is non-convex, we derived its convex surrogate approximation

$$t \text{ solves } \sum_{j=1}^{n^-} l_2(\mathbf{w}^\top \mathbf{x}_j^- - t) \leq n\tau. \quad (2.10)$$

Replacing the true quantile (2.8) by its surrogate approximation (2.10) yields

$$\underset{\mathbf{w}, t}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l_1(t - \mathbf{w}^\top \mathbf{x}_i^+) \quad (2.11)$$

subject to  $t$  is the top  $\tau$ -quantile: it solves (2.10).

In [24], we called the convex problem (2.11) *Pat&Mat*. The following theorem shows its dual form.

<sup>2</sup>To keep the readability of the paper, we postpone all proofs to the Appendix

**Theorem 2.2: *Pat&Mat* dual formulation**

The dual problem corresponding to the problem (2.11) has the form

$$\underset{\alpha, \beta, \delta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - C \sum_{i=1}^{n^+} l_1^* \left( \frac{\alpha_i}{C} \right) - \delta \sum_{j=1}^{n^-} l_2^* \left( \frac{\beta_j}{\delta} \right) - \delta n \tau \quad (2.12a)$$

$$\text{subject to } \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (2.12b)$$

$$\delta \geq 0, \quad (2.12c)$$

where  $l_1^*$ ,  $l_2^*$  are conjugates of the surrogate loss functions  $l_1$ ,  $l_2$  from (2.11) and  $\mathbb{K}$  was defined in (2.6).

## 2.3 New method for solving dual problems

In the previous section, we derived the dual formulations for the *TopPushK* and *Pat&Mat* problems. These dual formulations allow us to incorporate nonlinearity using kernels [26] in the same way as in SVM. Since the dimension of (2.7) and (2.12) equals to the number of samples  $n$ , it is computationally expensive to use standard techniques such as the gradient descent. To handle this issue, the coordinate descent algorithm [30, 27] has been proposed in the context of SVMs. Since problems (2.7, 2.12) differ from original SVMs by additional constraints (2.7b, 2.12b), the key idea of our algorithm is to update two coordinates (instead of one) of  $\alpha$ ,  $\beta$  at every iteration. To summarize, we will solve the original tasks (2.7, 2.12) by an iterative procedure where in every iteration we need to find a solution of a one-dimensional quadratic optimization problem. As we will show later, these one-dimensional problems have a closed form solution, which means that every iteration is cheap.

### 2.3.1 Adding kernels

To add kernels, we realize first that from the proofs of Theorems 2.1 and 2.2 for any  $\mathbf{z} \in \mathbb{R}^d$  we have

$$\mathbf{w}^\top \mathbf{z} = \sum_{i=1}^{n^+} \alpha_i \mathbf{z}^\top \mathbf{x}_i^+ - \sum_{j=1}^{n^-} \beta_j \mathbf{z}^\top \mathbf{x}_j^- \quad (2.13)$$

Consider now any kernel function  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ . Using the standard trick, we replace the kernel matrix (2.6) by<sup>3</sup>

$$\mathbb{K} = \begin{pmatrix} k(\mathbb{X}^+, \mathbb{X}^+) & -k(\mathbb{X}^+, \mathbb{X}^-) \\ -k(\mathbb{X}^-, \mathbb{X}^+) & k(\mathbb{X}^-, \mathbb{X}^-) \end{pmatrix}, \quad (2.14)$$

<sup>3</sup>The first part of the objective of (2.7) and (2.12) amounts to

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \begin{pmatrix} \mathbb{X}^+ \mathbb{X}^{+\top} & -\mathbb{X}^+ \mathbb{X}^{-\top} \\ -\mathbb{X}^- \mathbb{X}^{+\top} & \mathbb{X}^- \mathbb{X}^{-\top} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ -\beta \end{pmatrix}^\top \begin{pmatrix} \mathbb{X}^+ \mathbb{X}^{+\top} & \mathbb{X}^+ \mathbb{X}^{-\top} \\ \mathbb{X}^- \mathbb{X}^{+\top} & \mathbb{X}^- \mathbb{X}^{-\top} \end{pmatrix} \begin{pmatrix} \alpha \\ -\beta \end{pmatrix},$$

from which (2.14) follows.

## 2.3 New method for solving dual problems

where  $k(\cdot, \cdot)$  is applied to all rows of both arguments. Then for a new sample  $\mathbf{z}$ , the prediction (2.13) is replaced by

$$\text{pred}(\mathbf{z}) = \sum_{i=1}^{n^+} \alpha_i k(\mathbf{z}, \mathbf{x}_i^+) - \sum_{j=1}^{n^-} \beta_j k(\mathbf{z}, \mathbf{x}_j^-), \quad (2.15)$$

where  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are the optimal solution of (2.7) or (2.12).

### 2.3.2 Update of dual optimization variables

Let us consider the kernel matrix  $\mathbb{K}$  as in (2.14) and define the score vector  $\mathbf{s}$  by

$$\mathbf{s} = \mathbb{K} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}. \quad (2.16)$$

There are three possible update rules which modify two coordinates of  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$  and which satisfy constraints (2.7b, 2.12b) and keep (2.16) satisfied. The first one updates two components of  $\boldsymbol{\alpha}$

$$\alpha_k \rightarrow \alpha_k + \Delta, \quad \alpha_l \rightarrow \alpha_l - \Delta, \quad \mathbf{s} \rightarrow \mathbf{s} + (\mathbb{K}_{\bullet k} - \mathbb{K}_{\bullet l})\Delta, \quad (2.17a)$$

where  $\mathbb{K}_{\bullet i}$  denotes  $i$ -th column of  $\mathbb{K}$ . Note that the update rule for  $\mathbf{s}$  does not use matrix multiplication but only vector addition. The second rule updates one component of  $\boldsymbol{\alpha}$  and one component of  $\boldsymbol{\beta}$

$$\alpha_k \rightarrow \alpha_k + \Delta, \quad \beta_l \rightarrow \beta_l + \Delta, \quad \mathbf{s} \rightarrow \mathbf{s} + (\mathbb{K}_{\bullet k} + \mathbb{K}_{\bullet l})\Delta, \quad (2.17b)$$

and the last one updates two components of  $\boldsymbol{\beta}$

$$\beta_k \rightarrow \beta_k + \Delta, \quad \beta_l \rightarrow \beta_l - \Delta, \quad \mathbf{s} \rightarrow \mathbf{s} + (\mathbb{K}_{\bullet k} - \mathbb{K}_{\bullet l})\Delta. \quad (2.17c)$$

These three update rules hold true for any surrogate function. However, the calculation of the optimal  $\Delta$  depends on the used problem formulation and surrogate function. In Subsection 2.3.4, we show the closed-form formula for  $\Delta$  for *TopPushK* problem (2.7) with truncated quadratic surrogate function (2.2). Computation of  $\Delta$  for the hinge surrogate or *Pat&Mat* is presented in Appendices B.3 and B.4.

### 2.3.3 Algorithm summary and Complexity analysis

We summarize the whole procedure in Algorithm 2.1. We will describe it only for *Pat&Mat* (right column) as for *TopPushK* (left column) it is almost identical. In step 1 we initialize  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$  and  $\delta$  to some feasible value and based on (2.16) compute  $\mathbf{s}$ . Each **repeat** loop in step 2 updates two coordinates as shown in (2.17). In step 3 we select a random index  $k$  and in the **for** loop in step 4 we compute the optimal  $(\Delta_l, \delta_l)$  for all possible combinations  $(k, l)$  as in (2.17). In step 7 we select the pair  $(\Delta_l, \delta_l)$  which maximizes the objective. Finally, based on (2.17) we update  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$ ,  $\mathbf{s}$  and  $\delta$  in steps 8 and 9.

Now we derive the computational complexity of each **repeat** loop from step 2. Since the computation of  $(\Delta_l, \delta_l)$  amounts to solving a quadratic optimization problem in one variable, there is a closed-form solution (see Section 2.3.4) and step 5 can be performed in  $O(1)$ . Since this is embedded in a **for** loop in step 4, the whole complexity of this loop is  $O(n)$ . Step 8 requires  $O(1)$  for the update of  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  while  $O(n)$  for the update of  $\mathbf{s}$ . Since the other steps are  $O(1)$ , the total complexity of the **repeat** loop is  $O(n)$ . This holds true only if the kernel matrix  $\mathbb{K}$  is precomputed. In the opposite case, all complexities must be multiplied by the cost of computation of components of  $\mathbb{K}$  which is  $O(d)$ . This complexity analysis is summarized in Table 2.1.

**Algorithm 2.1** Coordinate descent algorithm for *TopPushK* (left) and *Pat&Mat* (right).

1: set $(\alpha, \beta)$ feasible, set $\mathbf{s}$ based on (2.16)	1: set $(\alpha, \beta, \delta)$ feasible, set $\mathbf{s}$ based on (2.16)
2: <b>repeat</b>	2: <b>repeat</b>
3:   select random $k$ from $\{1, 2, \dots, n\}$	3:   select random $k$ from $\{1, 2, \dots, n\}$
4: <b>for</b> $l \in \{1, 2, \dots, n\}$ <b>do</b>	4: <b>for</b> $l \in \{1, 2, \dots, n\}$ <b>do</b>
5:     compute $\Delta_l$	5:     compute $(\Delta_l, \delta_l)$
6: <b>end for</b>	6: <b>end for</b>
7:   select best $\Delta_l$	7:   select best $(\Delta_l, \delta_l)$
8:   update $\alpha, \beta, \mathbf{s}$ according to (2.17)	8:   update $\alpha, \beta, \mathbf{s}$ according to (2.17)
9:	9:   set $\delta \leftarrow \delta_l$
10: <b>until</b> stopping criterion is satisfied	10: <b>until</b> stopping criterion is satisfied

Table 2.1: Computational complexity of one **repeat** loop (which updates two coordinates of  $\alpha$  or  $\beta$ ) from Algorithm 2.1.

Precomputed $\mathbb{K}$	Evaluation of $\Delta_l$	Update of $\mathbf{s}$	Total per iteration
✓	$O(1)$	$O(n)$	$O(n)$
✗	$O(d)$	$O(nd)$	$O(nd)$

### 2.3.4 Computing $\Delta$ for *TopPushK* with truncated quadratic loss

In this section, we show how to compute the stepsize  $\Delta$  from (2.17) for *TopPushK* (2.7) with the truncated quadratic surrogate function (2.2). Optimal  $\Delta$  for *Pat&Mat* can be found in a similar way as we show in Appendix B.3. In Appendix B.4 we present the computation of optimal  $\Delta$  for *TopPushK* and *Pat&Mat* with the hinge loss function (2.3).

Plugging the conjugate (B.2) of the truncated quadratic loss (2.2) into *TopPushK* dual formulation (2.7) yields

$$\underset{\alpha, \beta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{\vartheta} \sum_{i=1}^{n^+} \alpha_i - \frac{1}{4C\vartheta^2} \sum_{i=1}^{n^+} \alpha_i^2 \quad (2.18a)$$

$$\text{subject to} \quad \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (2.18b)$$

$$\alpha_i \geq 0, \quad \forall i = 1, 2, \dots, n^+, \quad (2.18c)$$

$$0 \leq \beta_j \leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \quad \forall j = 1, 2, \dots, n^-. \quad (2.18d)$$

This is a quadratic optimization problem. Moreover, for  $K = 1$  the upper bound in (2.18d) automatically follows from (2.18b) and the problem can be simplified. In the following theorem, we show that for each of update rules (2.17), problem (2.18) has a simple closed-form solution. For simplicity, we use  $\text{clip}_{[a, b]}(c)$  to denote clipping (projecting)  $c$  to the interval  $[a, b]$ .

**Theorem 2.3:** Update rule for  $\Delta^*$  for *TopPushK* with truncated quadratic loss

Consider problem (2.18). Then the optimal step  $\Delta^*$  equals to

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]}(\gamma), \quad (2.19)$$

where there are the following three cases (each corresponding to one update rule in (2.17)):

- For any  $1 \leq k, l \leq n^+$  we have

$$\begin{aligned} \Delta_{lb} &= -\alpha_k, \\ \Delta_{ub} &= \alpha_l, \\ \gamma &= -\frac{s_k - s_l + \frac{1}{2C\vartheta^2}(\alpha_k - \alpha_l)}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{C\vartheta^2}}. \end{aligned}$$

- For any  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n$  we define  $\hat{l} = l - n^+$  and  $\beta_{\max} = \max_{j \in \{1, 2, \dots, n^-\} \setminus \{\hat{l}\}} \beta_j$ . Then we have

$$\begin{aligned} \Delta_{lb} &= \begin{cases} \max\{-\alpha_k, -\beta_{\hat{l}}\} & K = 1, \\ \max\{-\alpha_k, -\beta_{\hat{l}}, K\beta_{\max} - \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} +\infty & K = 1, \\ \frac{1}{K-1} \left( \sum_{i=1}^{n^+} \alpha_i - K\beta_{\hat{l}} \right) & \text{otherwise,} \end{cases} \\ \gamma &= -\frac{s_k + s_l - \frac{1}{\vartheta} + \frac{1}{2C\vartheta^2} \alpha_k}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk} + \frac{1}{2C\vartheta^2}}. \end{aligned}$$

- For any  $n^+ + 1 \leq k, l \leq n$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$  and then have

$$\begin{aligned} \Delta_{lb} &= \begin{cases} -\beta_{\hat{k}} & K = 1, \\ \max\{-\beta_{\hat{k}}, \beta_{\hat{l}} - \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} \beta_{\hat{l}} & K = 1, \\ \min\{\beta_{\hat{l}}, \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i - \beta_{\hat{k}}\} & \text{otherwise,} \end{cases} \\ \gamma &= -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}}. \end{aligned}$$

## 2.4 Numerical experiments

In this section, we present numerical results. All codes were implemented in the Julia language [31] and are available online.<sup>4</sup>

<sup>4</sup>All codes are available at [https://github.com/VaclavMacha/ClassificationOnTop\\_new.jl](https://github.com/VaclavMacha/ClassificationOnTop_new.jl)

### 2.4.1 Performance criteria

For the evaluation of numerical experiments, we use precision and recall. For a threshold  $t$  they are defined by

$$\text{Precision} = \frac{\sum_{i=1}^{n^+} [\mathbf{w}^\top \mathbf{x}_i^+ - t]}{\sum_{i=1}^n [\mathbf{w}^\top \mathbf{x}_i - t]}, \quad \text{Recall} = \frac{1}{n^+} \sum_{i=1}^{n^+} [\mathbf{w}^\top \mathbf{x}_i^+ - t]. \quad (2.20)$$

We also use the Precision-Recall (PR) curve that are commonly used for unbalanced data [32] and precision at a certain level of recall which we denote by Precision@Recall.

### 2.4.2 Hyperparameter choice

In Section 2.3 we introduced Algorithm 2.1 for solving dual problems (2.7, 2.12). We let it run for 20000 `repeat` loops, which corresponds to 40000 updates of coordinates of  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ . We use the linear and Gaussian kernels defined by

$$k_{\text{lin}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}, \quad (2.21)$$

$$k_{\text{gauss}}(\mathbf{x}, \mathbf{y}) = \exp\{-\sigma \|\mathbf{x} - \mathbf{y}\|_2^2\} \quad (2.22)$$

and the truncated quadratic loss (2.2) with  $\vartheta = 1$  as a surrogate.

The classifiers were trained on the training set. We selected the optimal hyperparameter from

$$\tau \in \{0.01, 0.05, 0.1\}, \quad K \in \{5, 10\}, \quad C \in \{0.1, 1, 10\}, \quad \sigma \in \{0.01, 0.05\}$$

which gave the best performance on the validation set. All presented result are shown on the testing set which was not part of the training process.

Table 2.2: Summary of the used datasets. It shows which original labels  $y^+$  were selected as the positive class, the number of features  $d$ , samples  $n$ , and the fraction of positive samples  $\frac{n^+}{n}$ .

	$y^+$	$d$	Training		Validation		Testing	
			$n$	$\frac{n^+}{n}$	$n$	$\frac{n^+}{n}$	$n$	$\frac{n^+}{n}$
sigillito1989classification	—	34	176	64.2%	87	64.4%	88	63.6%
Spambase	—	57	2301	39.4%	1150	39.4%	1150	39.4%
WhiteWineQuality	7, 8, 9	11	2449	21.6%	1224	21.7%	1225	21.6%
RedWineQuality	7, 8	11	800	13.5%	400	13.8%	399	13.5%
Fashion-MNIST	0	784	50000	10.0%	10000	10.0%	10000	10.0%

### 2.4.3 Dataset description

For numerical experiments, we consider the FashionMNIST dataset [33] and four smaller datasets from the UCI repository [34]: sigillito1989classification [20], Spambase, WhiteWineQuality [35] and RedWineQuality [35]. Datasets that do not contain testing set were randomly divided into a training (50%), validation (25%) and testing (25%) sets. For datasets that contain a testing set, the training set was randomly divided into a training and a validation set, where the validation set has the same size as the testing set. FashionMNIST dataset was converted to binary classification tasks by selecting class with label 0 as the positive class and the rest as the negative class. All datasets are summarized in Table 2.2.

### 2.4.4 Experiments

In Figure 2.1 we present the PR curves for all methods with two different kernels evaluated on the FashionMNIST dataset. The left column corresponds to the linear kernel (2.21) while the right one to the Gaussian kernel (2.22) with  $\sigma = 0.01$ . The nonlinear Gaussian kernel significantly outperforms the linear kernel. This will be confirmed later in Table 2.3 where we present a comparison from multiple datasets.

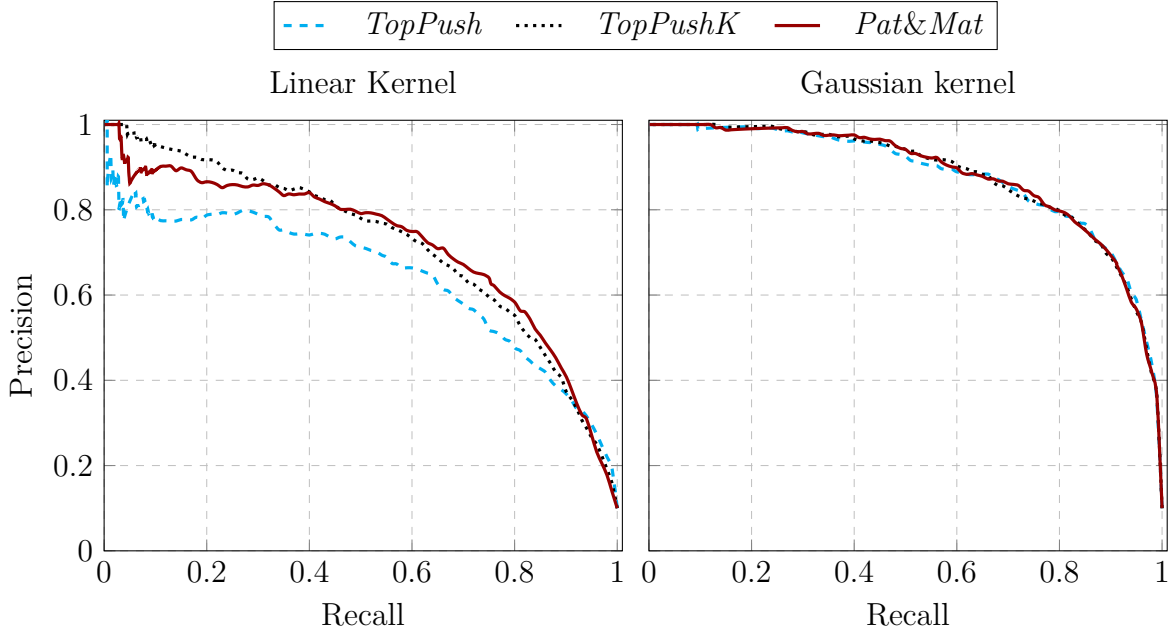


Figure 2.1: PR curves for all methods and FashionMNIST dataset. The left column corresponds to the linear kernel (2.21) and the right column corresponds to the Gaussian kernel (2.22).

For a better illustration of how the methods from Figure 2.1 work, we present density estimates of scores  $\mathbf{s}$  from (2.16). High scores predict positive labels while low scores predict negative labels. The rows of Figure 2.2 depict the linear (2.21) and the Gaussian kernels (2.22) with  $\sigma = 0.01$  while each column corresponds to one method. The black vertical lines depict the top 5%-quantile of all scores (on the testing set). Since a smaller overlap of scores of samples with positive and negative labels implies a better separation, we deduce the benefit of the Gaussian over the linear kernel.

In Table 2.3 we present the precision of all methods across all datasets from Table 2.2. For each dataset, we trained each method and computed precision at certain levels of recall. The depicted values are averages over all datasets. For each kernel and each level of recall, the best precision is highlighted in light green. Moreover, the best overall precision for each level of recall is depicted in dark green. We can make several observations from Table 2.3:

- All methods perform better with the Gaussian kernels than with the linear kernel.
- *TopPush* and *TopPushK* perform better for sufficiently small recall. This happened because they consider the threshold to be the maximal  $K$  negative scores and small recall corresponds to high threshold. However, for the same reason, *TopPush* is not robust.



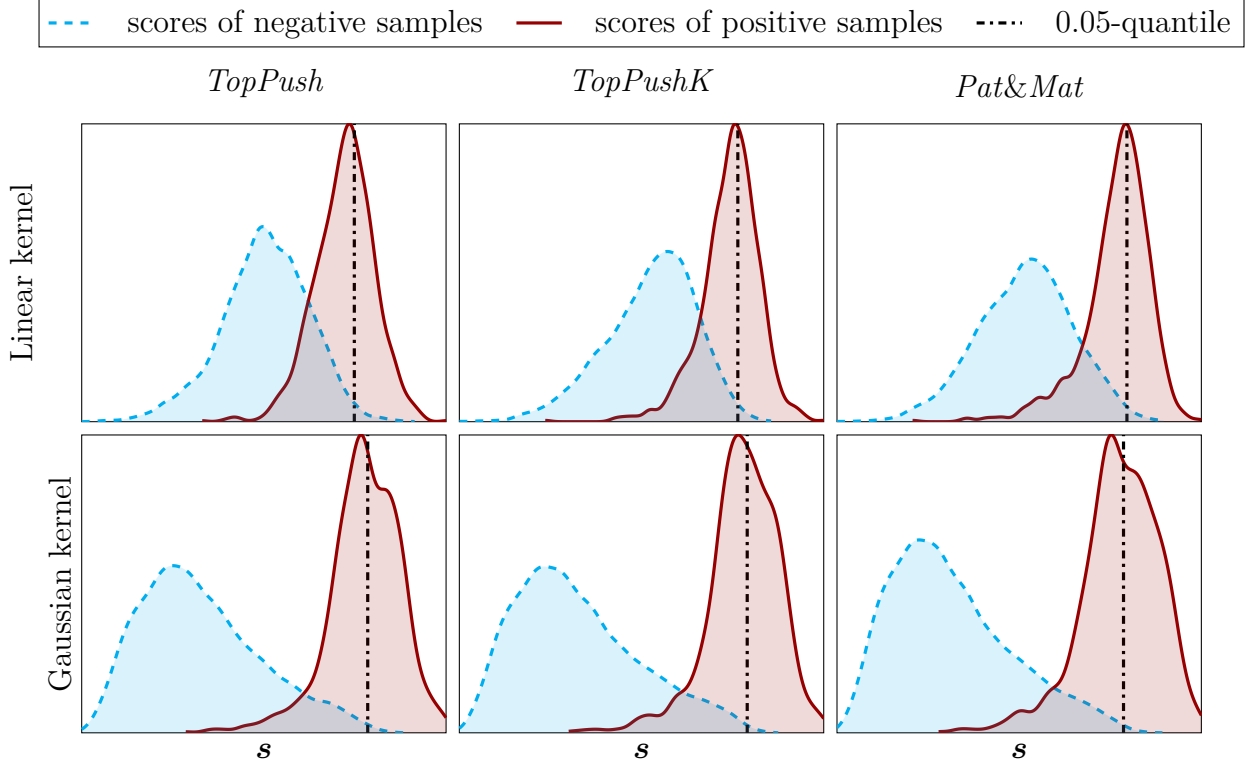


Figure 2.2: Density estimates for scores corresponding to samples with positive and negative labels for the FashionMNIST dataset.

- *Pat&Mat* is the best for all kernels if the recall is sufficiently large. The reason is again the form of the decision threshold.

In Figure 2.3, we investigate the convergence of methods. In each column, we show the convergence of primal and dual problems for one method. To solve the primal problem, we use the gradient method proposed in [24]. For the dual problem, we use our Algorithm 2.1. Since [24] considers only linear kernels, we present them. Moreover, since the computation of the objective is expensive, the results are presented for the sigillito1989classification dataset. We can see that *TopPush* and *TopPushK* converge to the same objective for primal and dual problems. This means that the problem was solved to optimality. However, there is a little gap between optimal solution of primal and dual problems for *Pat&Mat*.

Finally, Table 2.4 depicts the time comparison for all methods and all datasets. It shows the average time in milliseconds needed for one **repeat** loop in Algorithm 2.1. The time is relatively stable and for most of the datasets it is below one millisecond. Since we run all experiments for 20000 **repeat** loops, the evaluation of one method with one hyperparameter setting takes a few seconds for smaller datasets and approximately 7 minutes for FashionMNIST. The average time for one  $\Delta_i$  in step 5 in Algorithm 2.1 took between  $1.7 \cdot 10^{-7}$  and  $3.1 \cdot 10^{-7}$  seconds for each methods. It is almost the same for all datasets, which corresponds to the fact that the complexity of step 5 is independent of the size of the dataset. Note that in all experiments we used precomputed kernel matrix  $\mathbb{K}$  saved on the hard drive and not in memory.

## 2.5 Conclusion

Table 2.3: The precision of all methods averaged across all datasets from Table 2.2. Each column represents precision at a certain level of recall. Light green depicts the best method for the given kernel and dark green depicts the best overall method.

			Precision@Recall					
			0.05	0.1	0.2	0.4	0.6	0.8
Linear kernel	<i>TopPush</i>		79.83	64.27	65.55	61.85	57.89	51.83
	<i>TopPushK</i>	$K = 5$	73.96	65.41	64.82	60.28	56.94	50.52
		$K = 10$	60.63	61.97	59.69	56.89	54.40	49.83
	<i>Pat&amp;Mat</i>	$\tau = 0.01$	63.67	60.30	58.74	57.75	53.32	48.42
		$\tau = 0.05$	54.05	60.91	63.32	55.24	52.55	48.30
		$\tau = 0.1$	57.02	61.24	62.49	63.11	59.91	52.14
Gaussian kernel	<i>TopPush</i>		97.50	86.06	81.28	76.15	71.13	60.17
	<i>TopPushK</i>	$K = 5$	92.50	87.56	85.31	78.47	70.77	57.10
		$K = 10$	89.50	87.56	83.15	79.09	71.88	59.27
	<i>Pat&amp;Mat</i>	$\tau = 0.01$	89.65	89.11	86.75	80.77	75.44	65.95
		$\tau = 0.05$	80.77	81.28	85.74	82.92	74.91	65.04
		$\tau = 0.1$	81.30	84.14	82.58	83.12	77.82	66.50

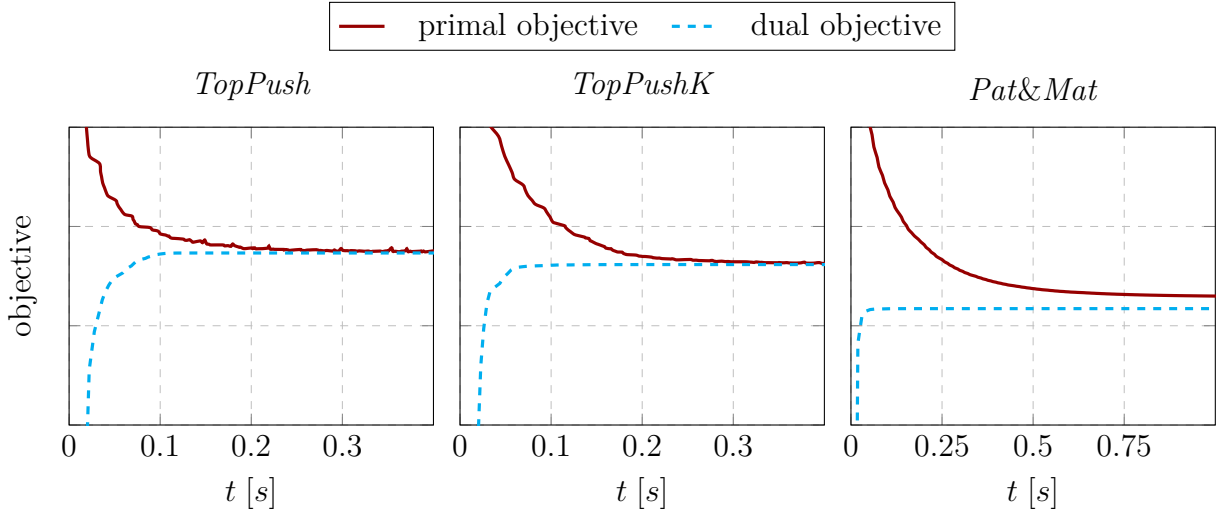


Figure 2.3: Convergence of the objectives for the primal (red line) and dual (blue line) problems for the sigillito1989classification dataset with linear kernel.

## 2.5 Conclusion

In this paper, we analyzed and extended the general framework for binary classification on top samples from [24] to nonlinear problems. Achieved results can be summarized as follows:

- We derived the dual formulations for *TopPush*, *TopPushK* and *Pat&Mat*.
- We proposed a new method for solving the dual problems. We performed its complexity analysis. For selected surrogate functions we also derived the exact formulas needed in the method.

Table 2.4: The average time with standard deviation (in milliseconds) for one **repeat** loop in Algorithm 2.1. The average time for one  $\Delta_l$  in step 5 in Algorithm 2.1 took between  $1.7 \cdot 10^{-7}$  and  $3.1 \cdot 10^{-7}$  seconds for each methods.

		<i>TopPush</i>	<i>TopPushK</i>	<i>Pat&amp;Mat</i>
One repeat loop [ms]	sigillito1989classification	$0.04 \pm 0.00$	$0.03 \pm 0.00$	$0.03 \pm 0.00$
	Spambase	$0.56 \pm 0.02$	$0.49 \pm 0.01$	$0.50 \pm 0.01$
	WhiteWineQuality	$0.62 \pm 0.03$	$0.53 \pm 0.01$	$0.54 \pm 0.01$
	RedWineQuality	$0.17 \pm 0.01$	$0.14 \pm 0.01$	$0.15 \pm 0.01$
	Fashion-MNIST	$17.16 \pm 0.74$	$15.95 \pm 0.14$	$15.54 \pm 0.80$

- We performed a numerical analysis of the proposed method. We showed its good convergence as well as improved performance of nonlinear kernels over the linear one.

Based on the numerical analysis from Section 2.4, we recommend using *TopPush* or *TopPushK* for problems where the resulting recall should be small. Otherwise, we recommend using *Pat&Mat* with an appropriately selected  $\tau$  parameter.



---

## Apendices



## Linear case

---

### A.1 Additional results and proofs

Here, we provide additional results and proofs of results mentioned in the main body. For convenience, we repeat the result statements.

#### A.1.1 Equivalence of (1.10) and (1.11)

To show this equivalence, we will start with an auxiliary lemma.

##### Lemma A.1

Denote by  $t$  the exact quantile from (1.9). Then for all  $\mu \in [0, 1]$  we have

$$\text{fp}(\mathbf{w}, t) = \mu \text{fp}(\mathbf{w}, t) + (1 - \mu) \text{fn}(\mathbf{w}, t) + (1 - \mu)(n\tau - n^+) + (1 - \mu)(q - 1), \quad (\text{A.1})$$

where  $q := \#\{\mathbf{x} \in \mathcal{X} \mid \mathbf{w}^\top \mathbf{x} = t\}$ .

##### *Proof:*

By the definition of the quantile we have

$$\text{tp}(\mathbf{w}, t) + \text{fp}(\mathbf{w}, t) = n\tau + q - 1.$$

This implies

$$\text{fp}(\mathbf{w}, t) = n\tau + q - 1 - \text{tp}(\mathbf{w}, t) = n\tau + q - 1 - n^+ + \text{fn}(\mathbf{w}, t).$$

From this relation we deduce

$$\begin{aligned} \text{fp}(\mathbf{w}, t) &= \mu \text{fp}(\mathbf{w}, t) + (1 - \mu) \text{fp}(\mathbf{w}, t) = \mu \text{fp}(\mathbf{w}, t) + (1 - \mu)(\text{fn}(\mathbf{w}, t) + n\tau - n^+ + q - 1) \\ &= \mu \text{fp}(\mathbf{w}, t) + (1 - \mu) \text{fn}(\mathbf{w}, t) + (1 - \mu)(n\tau - n^+) + (1 - \mu)(q - 1), \end{aligned}$$

which is precisely the lemma statement. ■

The right-hand side of (A.1) consists of three parts. The first one is a convex combination of false-positives and false-negatives. The second one is a constant term which has no impact on optimization. Finally, the third term  $(1 - \mu)(q - 1)$  equals the number of samples for which their classifier equals the quantile. However, this term is small in comparison with the true-positives and the false-negatives and can be neglected. Moreover, when the data are “truly” random such as when measurement errors are present, then  $q = 1$  and this term vanishes completely. This gives the (almost) equivalence of (1.10) and (1.11). Note that term  $q$  is ignored in many papers.

### A.1.2 Results related to convexity

#### Proposition 1.2

Thresholds  $t_2$  and  $t_3$  are convex functions of the weights  $\mathbf{w}$ . The threshold function  $t_1$  is non-convex.

#### *Proof Proposition 1.2 on page 7:*

It is easy to show that the quantile  $t_1$  is not convex. Due to [11], the mean of the  $k$  highest values of a vector is a convex function and therefore,  $t_2$  is a convex function. It remains to analyze  $t_3$ . It is defined via an implicit equation, where we consider for simplicity  $\beta = 1$ ,

$$g(\mathbf{w}, t) := \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} l(\mathbf{w}^\top \mathbf{x} - t) - \tau = 0.$$

Since  $l$  is convex, we immediately obtain that  $g$  is jointly convex in both variables.

To show the convexity, consider  $\mathbf{w}$ ,  $\hat{\mathbf{w}}$  and the corresponding  $t = t_3(\mathbf{w})$ ,  $\hat{t} = t_3(\hat{\mathbf{w}})$ . Note that this implies  $g(\mathbf{w}, t) = g(\hat{\mathbf{w}}, \hat{t}) = 0$ . Then for any  $\lambda \in [0, 1]$  we have

$$g(\lambda \mathbf{w} + (1 - \lambda) \hat{\mathbf{w}}, \lambda t + (1 - \lambda) \hat{t}) \leq \lambda g(\mathbf{w}, t) + (1 - \lambda) g(\hat{\mathbf{w}}, \hat{t}) = 0, \quad (\text{A.2})$$

where the inequality follows from the convexity of  $g$  and the equality from  $g(\mathbf{w}, t) = g(\hat{\mathbf{w}}, \hat{t}) = 0$ . From the definition of the surrogate quantile function  $t_3$  we have

$$g(\lambda \mathbf{w} + (1 - \lambda) \hat{\mathbf{w}}, t_3(\lambda \mathbf{w} + (1 - \lambda) \hat{\mathbf{w}})) = 0. \quad (\text{A.3})$$

Since  $g$  is non-increasing in the second variable, from (A.2) and (A.3) we deduce

$$t_3(\lambda \mathbf{w} + (1 - \lambda) \hat{\mathbf{w}}) \leq \lambda t + (1 - \lambda) \hat{t} = \lambda t_3(\mathbf{w}) + (1 - \lambda) t_3(\hat{\mathbf{w}}),$$

which implies that function  $\mathbf{w} \mapsto t_3(\mathbf{w})$  is convex. ■

#### Theorem 1.3

If the threshold  $t$  is a convex function of the weights  $\mathbf{w}$ , then function  $f(\mathbf{w}) = \bar{\text{fn}}(\mathbf{w}, t(\mathbf{w}))$  is convex.

#### *Proof of Theorem 1.3 on page 7:*

Due to the definition of the surrogate counts (1.4), the objective of (1.5) equals to

$$\frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l(t(\mathbf{w}) - \mathbf{w}^\top \mathbf{x}).$$

Here we write  $t(\mathbf{w})$  to stress the dependence of  $t$  on  $\mathbf{w}$ . Since  $\mathbf{w} \mapsto t(\mathbf{w})$  is a convex function, we also have that  $\mathbf{w} \mapsto t(\mathbf{w}) - \mathbf{w}^\top \mathbf{x}$  is a convex function. From its definition, the surrogate function  $l$  is convex and non-decreasing. Since a composition of a convex function with a non-decreasing convex function is a convex function, this finishes the proof. ■

### A.1.3 Results related to differentiability



**Theorem 1.4**

If the surrogate function  $l$  is differentiable, then threshold  $t_3$  is a differentiable function of the weights  $\mathbf{w}$  and its derivative equals to

$$\nabla t_3(\mathbf{w}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}} l'(\beta(\mathbf{w}^\top \mathbf{x} - t_3(\mathbf{w}))) \mathbf{x}}{\sum_{\mathbf{x} \in \mathcal{X}} l'(\beta(\mathbf{w}^\top \mathbf{x} - t_3(\mathbf{w})))}.$$

The threshold functions  $t_1$  and  $t_2$  are non-differentiable.

**Proof of Theorem 1.4 on page 7:**

The result for  $t_3$  follows directly from the implicit function theorem. The non-differentiability of  $t_1$  and  $t_2$  happens whenever the threshold value is achieved at two different scores. ■

**A.1.4 Results related to stability****Theorem 1.5**

Consider any of these formulations: *TopPush*, *TopPushK*, *TopMeanK* or  $\tau$ -FPL. Fix any  $\mathbf{w}$  and denote the corresponding threshold  $t(\mathbf{w})$ . If we have

$$t(\mathbf{w}) \geq \frac{1}{n^+} \sum_{\mathbf{x}^+ \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x}^+, \quad (1.24)$$

then  $f(\mathbf{0}) \leq f(\mathbf{w})$ . Specifically, denote the scores  $s^+ = \mathbf{w}^\top \mathbf{x}^+$  for  $\mathbf{x}^+ \in \mathcal{X}^+$  and  $s^- = \mathbf{w}^\top \mathbf{x}^-$  for  $\mathbf{x}^- \in \mathcal{X}^-$  and the ordered variants with decreasing components of  $\mathbf{s}^-$  by  $\mathbf{s}_{[i]}^-$ . Then

$$\begin{aligned} s_{[1]}^- &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \textit{TopPush}, \\ \frac{1}{k} \sum_{i=1}^k s_{[i]}^- &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \textit{TopPushK}, \\ \frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} s_{[i]}^- &\geq \frac{1}{n^+} \sum_{i=1}^{n^+} s_i^+ \implies f(\mathbf{0}) \leq f(\mathbf{w}) \text{ for } \tau\text{-FPL}. \end{aligned} \quad (1.25)$$

**Proof of Theorem 1.5 on page 9:**

Due to  $l(0) = 1$  and the convexity of  $l$  we have  $l(s) \geq 1 + cs$ , where  $c$  equals to the derivative of  $l$  at 0. Then we have

$$\begin{aligned} f(\mathbf{w}) &\geq \frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}, t) = \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l(t - \mathbf{w}^\top \mathbf{x}) \geq \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} (1 + c(t - \mathbf{w}^\top \mathbf{x})) \\ &= 1 + \frac{c}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} (t - \mathbf{w}^\top \mathbf{x}) = 1 + ct - \frac{c}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} \geq 1, \end{aligned}$$

where the last inequality follows from (1.24). Now we realize that for any formulation from the statement, the corresponding threshold for  $\mathbf{w} = \mathbf{0}$  equals to  $t = 0$ , and thus  $f(\mathbf{0}) = 1$ . But then  $f(\mathbf{0}) \leq f(\mathbf{w})$ . The second part of the result follows from the form of thresholds  $t(\mathbf{w})$ . ■

### Theorem 1.8

Consider the *Pat&Mat* or *Pat&Mat-NP* formulation with the hinge surrogate and no regularization. Assume that for some  $\mathbf{w}$  we have

$$\frac{1}{n^+} \sum_{\mathbf{x}^+ \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x}^+ > \frac{1}{n^-} \sum_{\mathbf{x}^- \in \mathcal{X}^-} \mathbf{w}^\top \mathbf{x}^-. \quad (1.26)$$

Then there is a scaling parameter  $\beta_0$  from (1.13) such that  $f(\mathbf{w}) < f(\mathbf{0})$  for all  $\beta \in (0, \beta_0)$ .

### Proof of Theorem 1.8 on page 10:

Define first

$$s_{\min} = \min_{\mathbf{x} \in \mathcal{X}} \mathbf{w}^\top \mathbf{x}, \quad \bar{s} = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{w}^\top \mathbf{x}, \quad s_{\max} = \max_{\mathbf{x} \in \mathcal{X}} \mathbf{w}^\top \mathbf{x}.$$

Then we have the following chain of relations

$$\begin{aligned} \bar{s} &= \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{w}^\top \mathbf{x} = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} + \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^-} \mathbf{w}^\top \mathbf{x} < \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} + \frac{n^-}{nn^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} \\ &= \frac{1}{n} \left( 1 + \frac{n^-}{n^+} \right) \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} = \frac{1}{n} \frac{n^+ + n^-}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} = \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x}. \end{aligned} \quad (\text{A.4})$$

The only inequality follows from (1.26) and the last equality follows from  $n^+ + n^- = n$ .

Due to (1.26) we have  $s_{\min} < \bar{s} < s_{\max}$ . Then we can define

$$\beta_0 = \min \left\{ \frac{\tau}{\bar{s} - s_{\min}}, \frac{1 - \tau}{s_{\max} - \bar{s}}, \tau \right\},$$

observe that  $\beta_0 > 0$ , fix any  $\beta \in (0, \beta_0)$  and define

$$t = \frac{1 - \tau}{\beta_0} + \bar{s}.$$

Then we obtain

$$1 + \beta(\mathbf{w}^\top \mathbf{x} - t) \geq 1 + \beta(s_{\min} - t) = 1 + \beta s_{\min} - 1 + \tau - \beta \bar{s} = \beta(s_{\min} - \bar{s}) + \tau \geq 0. \quad (\text{A.5})$$

Here, the first equality follows from the definition of  $t$  and the last inequality from the definition of  $\beta_0$ . Moreover, we have

$$\begin{aligned} \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} l(\beta(\mathbf{w}^\top \mathbf{x} - t)) &= \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} \max\{1 + \beta(\mathbf{w}^\top \mathbf{x} - t), 0\} = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} (1 + \beta(\mathbf{w}^\top \mathbf{x} - t)) \\ &= 1 - \beta t + \frac{\beta}{n} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{w}^\top \mathbf{x} = 1 - \beta t + \beta \bar{s} = \tau, \end{aligned}$$

where the second equality employs (A.5), the third one the definition of  $\bar{s}$  and the last one the definition of  $t$ . But this means that  $t$  is the threshold corresponding to  $\mathbf{w}$ .

Similarly to (A.5) we get

$$1 + t - \mathbf{w}^\top \mathbf{x} \geq 1 + t - s_{\max} = 1 + \frac{1 - \tau}{\beta} + \bar{s} - s_{\max} \geq \frac{1 - \tau}{\beta} + \bar{s} - s_{\max} \geq 0, \quad (\text{A.6})$$

where the last inequality follows from the definition of  $\beta_0$ . Then for the objective we have

$$\begin{aligned} f(\mathbf{w}) &= \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l(t - \mathbf{w}^\top \mathbf{x}) = \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \max\{1 + t - \mathbf{w}^\top \mathbf{x}, 0\} = \\ &= \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} (1 + t - \mathbf{w}^\top \mathbf{x}) = 1 + t - \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \mathbf{w}^\top \mathbf{x} < 1 + t - \bar{s} \\ &= 1 + \frac{1 - \tau}{\beta} + \bar{s} - \bar{s} = 1 + \frac{1 - \tau}{\beta} = f(\mathbf{0}), \end{aligned}$$

where the third equality follows from (A.6), the only inequality from (A.4) and the last equality from Appendix A.2. Thus, we finished the proof for *Pat&Mat*. The proof for *Pat&Mat-NP* can be performed in an identical way by replacing in the definition of  $\bar{s}$  the mean with respect to all samples by the mean with respect to all negative samples. ■

### A.1.5 Results related to threshold comparison

#### Lemma A.7

Define vector  $\mathbf{s}^+$  with components  $s^+ = \mathbf{w}^\top \mathbf{x}^+$  for  $\mathbf{x}^+ \in \mathcal{X}^+$  and similarly define vector  $\mathbf{s}^-$  with components  $s^- = \mathbf{w}^\top \mathbf{x}^-$  for  $\mathbf{x}^- \in \mathcal{X}^-$ . Denote by  $\mathbf{s}_{[\cdot]}^+$  and  $\mathbf{s}_{[\cdot]}^-$  the sorted versions of  $\mathbf{s}^+$  and  $\mathbf{s}^-$ , respectively. Then we have the following statements:

$$\begin{aligned} s_{[n+\tau]}^+ > s_{[n-\tau]}^- &\implies \text{Grill has larger threshold than Grill-NP,} \\ \frac{1}{n^+ \tau} \sum_{i=1}^{n^+ \tau} s_{[i]}^+ > \frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} s_{[i]}^- &\implies \text{TopMeanK has larger threshold than } \tau\text{-FPL.} \end{aligned}$$

#### Proof:

Since  $\mathbf{s}^+$  and  $\mathbf{s}^-$  are computed on disjunctive indices, we have

$$s_{[n\tau]} \geq \min\{s_{[n+\tau]}^+, s_{[n-\tau]}^-\}.$$

Since  $s_{[n\tau]}$  is the threshold for *Grill* and  $s_{[n-\tau]}^-$  is the threshold for *Grill-NP*, the first statement follows. The second part can be shown in a similar way. ■

Since the goal of the presented formulations is to push  $s^+$  above  $s^-$ , we may expect that the conditions in Lemma A.7 hold true.

## A.2 Computation for Section 1.2.4

We derive the results presented in Section 1.2.4 more properly. We recall that we have  $n$  negative samples randomly distributed in  $[-1, 0] \times [-1, 1]$ ,  $n$  positive samples randomly distributed in  $[0, 1] \times [-1, 1]$  and one negative sample at  $(2, 0)$ . We assume that  $n$  is large and the outlier may be ignored for the computation of thresholds which require a large number of points. Since the computation is simple for other formulations, we show it only for *Pat&Mat*.

For  $\mathbf{w}_0 = (0, 0)$ , we get

$$\tau = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} l(\beta(\mathbf{w}_0^\top \mathbf{x} - t)) = l(0 - \beta t) = 1 - \beta t,$$

### A.3 Computing the threshold for *Pat&Mat*

which implies  $t = \frac{1}{\beta}(1 - \tau)$  and consequently

$$\frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}_0, t) = \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} l(t - 0) = l(t) = 1 + t.$$

This finishes the computation for  $\mathbf{w}_0$ .

For  $\mathbf{w}_1 = (1, 0)$  the computation goes similar. Then  $\mathbf{w}_1^\top \mathbf{x}^+$  has the uniform distribution on  $[0, 1]$  while  $\mathbf{w}_1^\top \mathbf{x}$  has the uniform distribution on  $[-1, 1]$ . If  $\beta \leq \tau$ , then

$$\begin{aligned} \tau &= \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} l(\mathbf{w}_1^\top \mathbf{x} - t) \approx \frac{1}{2} \int_{-1}^1 l(s - t) \, ds = \frac{1}{2} \int_{-1}^1 \max\{0, 1 + \beta(s - t)\} \, ds \\ &= \frac{1}{2} \int_{-1}^1 (1 + \beta(s - t)) \, ds = 1 - \beta t + \frac{\beta}{2} \int_{-1}^1 s \, ds = 1 - \beta t, \end{aligned} \quad (\text{A.7})$$

and thus again  $t = \frac{1}{\beta}(1 - \tau)$ . Note that

$$1 + \beta(s - t) \geq 1 + \beta(-1 - t) = 1 - \beta - 1 + \tau = -\beta + \tau \geq 0$$

and we could have ignored the max operator in (A.7). Finally, we have

$$\frac{1}{n^+} \overline{\text{fn}}(\mathbf{w}_1, t) \approx \int_0^1 l(t - s) \, ds = \int_0^1 (1 + t - s) \, ds = 0.5 + t.$$

### A.3 Computing the threshold for *Pat&Mat*

We show how to efficiently compute the threshold (1.13) for *Pat&Mat* and the hinge surrogate (1.3). As always define the scores  $s_i = \mathbf{w}^\top \mathbf{x}_i$  and consider function

$$h(t) = \sum_{i=1}^n l(\beta(s_i - t)) - n\tau. \quad (\text{A.8})$$

Then solving (1.31) is equivalent to looking for  $\hat{t}$  such that  $h(\hat{t}) = 0$ . We have the following properties of  $h$ .

#### Lemma A.8

Function  $h$  is continuous and strictly decreasing (until it hits the global minimum) with  $h(t) \rightarrow \infty$  as  $t \rightarrow -\infty$  and  $h(t) \rightarrow -n\tau$  as  $t \rightarrow \infty$ . Thus, there is a unique solution to the equation  $h(t) = 0$ .

For sorted data, the following lemma gives advice on how to solve equation  $h(t) = 0$ .

#### Lemma A.9

Let  $s_1 \leq s_2 \leq \dots \leq s_n$  be sorted. Define  $\gamma = \frac{1}{\beta}$ . Then

$$h(s_i + \gamma) = h(s_{i+1} + \gamma) + (n - i)\beta(s_{i+1} - s_i) \quad (\text{A.9})$$

for all  $i = n - 1, \dots, 1$  with the initial condition  $h(s_n + \gamma) = -n\tau$ .

### Proof:

Observe first that

$$h(s_j + \gamma) = \sum_{i=1}^n l(\beta(s_i - s_j - \gamma)) - n\tau = \sum_{i=1}^n \max(0, \beta(s_i - s_j)) - n\tau = \sum_{i=j+1}^n \beta(s_i - s_j) - n\tau.$$

From here, we obtain  $h(s_n + \gamma) = -n\tau$ . Moreover, we have

$$\begin{aligned} h(s_j + \gamma) &= \sum_{i=j+1}^n \beta(s_i - s_j) - n\tau = \sum_{i=j+2}^n \beta(s_i - s_j) + \beta(s_{j+1} - s_j) - n\tau \\ &= \sum_{i=j+2}^n \beta(s_i - s_{j+1}) + \sum_{i=j+2}^n \beta(s_{j+1} - s_j) + \beta(s_{j+1} - s_j) - n\tau \\ &= \sum_{i=j+2}^n \beta(s_i - s_{j+1}) + (n-j)\beta(s_{j+1} - s_j) - n\tau \\ &= h(s_{j+1} + \gamma) + (n-j)\beta(s_{j+1} - s_j), \end{aligned}$$

which finishes the proof. ■

Thus, to solve  $h(t) = 0$  with the hinge surrogate, we start with  $t_n = s_n + \gamma$  and  $h(t_n) = -n\tau$ . Then we start decreasing  $t$  according to (A.9) until we find some  $t_i = s_i + \gamma$  such that  $h(t_i) > 0$ . The desired  $t$  then lies between  $t_i$  and  $t_{i+1}$ . Since  $h$  is a piecewise linear function with

$$h(t) = h(t_i) + \frac{t - t_i}{t_{i+1} - t_i} (h(t_{i+1}) - h(t_i))$$

for  $t \in [t_i, t_{i+1}]$ , the precise value of  $\hat{t}$  can be computed by a simple interpolation

$$\hat{t} = t_i - h(t_i) \frac{t_{i+1} - t_i}{h(t_{i+1}) - h(t_i)} = t_i - h(t_i) \frac{t_{i+1} - t_i}{-(n-i)\beta(t_{i+1} - t_i)} = t_i + \frac{h(t_i)}{\beta(n-i)}.$$

## A.4 Proof of Theorem 1.9

The proof is divided into three parts. In Section A.4.1, we prove a general statement for convergence of stochastic gradient descent with a convex objective. In Section A.4.2 we apply it to Theorem 1.9. The proof is based on auxiliary results from Section A.4.3.

### A.4.1 General result

Consider a differentiable objective function  $f$  and the optimization method

$$w^{k+1} = w^k - \alpha^k g(w^k), \tag{A.10}$$

where  $\alpha^k > 0$  is a stepsize and  $g(w^k)$  is an approximation of the gradient  $\nabla f(w^k)$ . Assume the following:

- (A1)  $f$  is differentiable, convex and attains a global minimum;
- (A2)  $\|g(w^k)\| \leq B$  for all  $k$ ;
- (A3) the stepsize is non-increasing and satisfies  $\sum_{k=0}^{\infty} \alpha^k = \infty$ ;
- (A4) the stepsize satisfies  $\sum_{k=0}^{\infty} (\alpha^k)^2 < \infty$ ;

## A.4 Proof of Theorem 1.9

(A5) the stepsize satisfies  $\sum_{k=0}^{\infty} \|\alpha^{k+1} - \alpha^k\| < \infty$ .

Assumptions (A3)-(A5) are satisfied for example for  $\alpha^k = \alpha^0 \frac{1}{k+1}$ . We start with the general result.

### Theorem A.10

Assume that (A1)-(A4) is satisfied. If there exists some  $C$  such that for some global minimum of  $w^*$  of  $f$  we have

$$\sum_{k=0}^{\infty} \alpha^k \langle g(w^k) - \nabla f(w^k), w^* - w^k \rangle \leq C, \quad (\text{A.11})$$

then the sequence  $\{w^k\}$  generated by (A.10) is bounded and  $f(w^k) \rightarrow f(w^*)$ . Thus, all its convergent subsequences converge to some global minimum of  $f$ .

### Proof:

Note first that the convexity from (A1) implies

$$\langle \nabla f(w^k), w^* - w^k \rangle \leq f(w^*) - f(w^k). \quad (\text{A.12})$$

Then we have

$$\begin{aligned} \|w^{k+1} - w^*\|^2 &= \|w^k - \alpha^k g(w^k) - w^*\|^2 \\ &= \|w^k - w^*\|^2 + 2\alpha^k \langle g(w^k), w^* - w^k \rangle + (\alpha^k)^2 \|g(w^k)\|^2 \\ &\leq \|w^k - w^*\|^2 + 2\alpha^k \langle g(w^k) - \nabla f(w^k), w^* - w^k \rangle + 2\alpha^k (f(w^*) - f(w^k)) + (\alpha^k)^2 B^2, \end{aligned}$$

where the inequality follows from (A.12) and assumption (A2). Summing this expression for all  $k$  and using (A.11) leads to

$$\limsup_k \|w^k - w^*\|^2 \leq \|w^0 - w^*\|^2 + 2C + 2 \sum_{k=0}^{\infty} \alpha^k (f(w^*) - f(w^k)) + \sum_{k=0}^{\infty} (\alpha^k)^2 B^2.$$

Using assumption (A4) results in the existence of some  $\hat{C}$  such that

$$\limsup_k \|w^k - w^*\|^2 + 2 \sum_{k=0}^{\infty} \alpha^k (f(w^k) - f(w^*)) \leq 2\hat{C}. \quad (\text{A.13})$$

Since  $\alpha^k > 0$  and  $f(w^k) \geq f(w^*)$  as  $w^*$  is a global minimum of  $f$ , we infer that sequence  $\{w^k\}$  is bounded and (A.13) implies

$$\sum_{k=0}^{\infty} \alpha^k (f(w^k) - f(w^*)) \leq \hat{C}.$$

Since  $f(w^k) - f(w^*) \geq 0$ , due to assumption (A3) we obtain  $\lim f(w^k) \rightarrow f(w^*)$ , which implies the theorem statement. ■

### A.4.2 Proof of Theorem 1.9

For the proof, we will consider a general surrogate which satisfies:

- (S1)  $l(s) \geq 0$  for all  $s \in \mathbb{R}$ ,  $l(0) = 1$  and  $l(s) \rightarrow 0$  as  $s \rightarrow -\infty$ ;
- (S2)  $l$  is convex and strictly increasing function on  $(s_0, \infty)$ , where  $s_0 := \sup\{s \mid l(s) = 0\}$ ;
- (S3)  $\frac{l'}{l}$  is a decreasing function on  $(s_0, \infty)$ ;
- (S4)  $l'$  is a bounded function;
- (S5)  $l'$  is a Lipschitz continuous function with modulus  $L$ .

All these requirements are satisfied for the surrogate logistic or by the Huber loss, which is the hinge surrogate which is smoothened on an  $\varepsilon$ -neighborhood of zero.

#### Theorem 1.9

Consider the *Pat&Mat* or *Pat&Mat-NP* formulation, stepsizes  $\alpha^k = \frac{\alpha^0}{k+1}$  and piecewise disjoint minibatches  $I^1, \dots, I^m$  which cycle periodically  $I^{k+m} = I^k$ . If  $l$  is the smoothened (Huberized) hinge function, then Algorithm 1.1 converges to the global minimum of (1.16).

#### Proof of Theorem 1.9 on page 14:

We intend to apply Theorem A.10 and thus, we need to verify its assumptions. Assumption (A1) is satisfied as  $f$  is convex due to Theorem 1.3. Assumption (A2) follows directly from Lemma A.13. Assumptions (A3), (A4) and (A5) are imposed directly in the statement of this theorem. It remains to verify (A.11).

For simplicity, we will do so only for  $\beta = 1$  and for  $s = 2$  minibatches of the same size. However, the proof would be identical for other values. This implies that there are some  $I^k$  and  $I^{k+1}$  which are pairwise disjoint, they cover all samples and that  $I^k = I^{k+2}$  for all  $k$ . The assumptions imply that the number of positive samples in each minibatch equal to  $n_+^k = \frac{1}{2}n_+$ , where  $n_+$  is the total number of positive samples.

First we estimate the difference between  $s_i^k$  defined in (1.30) and  $x_i^\top w^k$ . For any  $i \in I^k$  due to the construction (1.30) we have

$$\begin{aligned} s_i^k &= x_i^\top w^k, \\ s_i^{k-1} &= s_i^{k-2} = x_i^\top w^{k-2} = x_i^\top (w^k + \alpha^{k-2}g(w^{k-2}) + \alpha^{k-1}g(w^{k-1})) \\ &= x_i^\top w^k + \alpha^{k-2}x_i^\top g(w^{k-2}) + \alpha^{k-1}x_i^\top g(w^{k-1}). \end{aligned} \quad (\text{A.14})$$

Similarly, for  $i \notin I^k$  we have

$$s_i^k = s_i^{k-1} = x_i^\top w^{k-1} = x_i^\top (w^k + \alpha^{k-1}g(w^{k-1})) = x_i^\top w^k + \alpha^{k-1}x_i^\top g(w^{k-1}). \quad (\text{A.15})$$

Recall that we already verified (A1)-(A5). Combining (A2) with (A.14) and (A.15) yields the existence of some  $C_2$  such that for all  $i \in X$  we have

$$\begin{aligned} \|s_i^k - x_i^\top w^k\| &\leq C_2 \alpha^{k-1}, \\ \|s_i^{k-1} - x_i^\top w^k\| &\leq C_2 (\alpha^{k-1} + \alpha^{k-2}). \end{aligned} \quad (\text{A.16})$$

This also immediately implies

$$\begin{aligned}\|t^k - t(w^k)\| &\leq C_2 \alpha^{k-1}, \\ \|t^{k-1} - t(w^k)\| &\leq C_2 (\alpha^{k-1} + \alpha^{k-2}).\end{aligned}\tag{A.17}$$

Since  $l'$  is Lipschitz continuous with modulus  $L$  according to (S5), due to (A.16) and (A.17) we get

$$\|l'(t^k - s_i^k) - l'(t(w^k) - x_i^\top w^k)\| \leq L \|t^k - s_i^k - t(w^k) + x_i^\top w^k\| \leq 2C_2 L \alpha^{k-1}.\tag{A.18}$$

In an identical way we can show

$$\begin{aligned}\|l'(t^{k-1} - s_i^{k-1}) - l'(t(w^k) - x_i^\top w^k)\| &\leq 2C_2 L (\alpha^{k-1} + \alpha^{k-2}), \\ \|l'(s_i^k - t^k) - l'(x_i^\top w^k - t(w^k))\| &\leq 2C_2 L \alpha^{k-1}, \\ \|l'(s_i^{k-1} - t^{k-1}) - l'(x_i^\top w^k - t(w^k))\| &\leq 2C_2 L (\alpha^{k-1} + \alpha^{k-2}).\end{aligned}\tag{A.19}$$

Now we need to estimate the distance between  $\nabla t(w^k)$  and  $\nabla t^k$ . We have

$$\begin{aligned}\nabla t^k &= \frac{\sum_{i \in I^k} l'(s_i^k - t^k) x_i + \sum_{i \in I^{k-1}} l'(s_i^{k-1} - t^{k-1}) x_i}{\sum_{i \in X} l'(s_i^k - t^k)}, \\ \nabla t(w^k) &= \frac{\sum_{i \in I^k} l'(x_i^\top w^k - t(w^k)) x_i + \sum_{i \in I^{k-1}} l'(x_i^\top w^k - t(w^k)) x_i}{\sum_{i \in X} l'(x_i^\top w^k - t(w^k))}.\end{aligned}\tag{A.20}$$

The first equality in (A.20) follows from (1.34) and (1.32) while the second equality in (A.20) follows from Theorem 1.4 and  $X = I^k \cup I^{k-1}$ . From Lemma A.12 we deduce that the denominators in (A.20) are bounded away from zero uniformly in  $k$ . Assumption (A4) implies  $\alpha^k \rightarrow 0$ . This allows us to use Lemma A.14 which together with (A.19) implies that there is some  $C_3$  such that for all sufficiently large  $k$  we have

$$\|\nabla t^k - \nabla t(w^k)\| \leq C_3 (\alpha^{k-1} + \alpha^{k-2}).\tag{A.21}$$

Using the assumptions above, we can simplify the terms for  $g(w^k)$  and  $\nabla f(w^k)$  to

$$\begin{aligned}g(w^k) &= \frac{2}{n_+} \sum_{i \in I_+^k} l'(t^k - s_i^k) (\nabla t^k - x_i), \\ g(w^{k+1}) &= \frac{2}{n_+} \sum_{i \in I_+^{k+1}} l'(t^{k+1} - s_i^{k+1}) (\nabla t^{k+1} - x_i), \\ \nabla f(w^k) &= \frac{1}{n_+} \sum_{i \in \mathcal{X}_+} l'(t(w^k) - x_i^\top w^k) (\nabla t(w^k) - x_i), \\ \nabla f(w^{k+1}) &= \frac{1}{n_+} \sum_{i \in \mathcal{X}_+} l'(t(w^{k+1}) - x_i^\top w^{k+1}) (\nabla t(w^{k+1}) - x_i).\end{aligned}$$

Due to the assumptions, we have  $\mathcal{X}_+ = I_+^k \cup I_+^{k+1}$  and  $\emptyset = I_+^k \cap I_+^{k+1}$ , which allows us to



write

$$n^+(g(w^k) + g(w^{k+1}) - \nabla f(w^k) - \nabla f(w^{k+1})) \quad (\text{A.22a})$$

$$= \sum_{i \in I_+^k} l'(t^k - s_i^k)(\nabla t^k - x_i) - \sum_{i \in I_+^k} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i) \quad (\text{A.22b})$$

$$+ \sum_{i \in I_+^k} l'(t^k - s_i^k)(\nabla t^k - x_i) - \sum_{i \in I_+^k} l'(t(w^{k+1}) - x_i^\top w^{k+1})(\nabla t(w^{k+1}) - x_i) \quad (\text{A.22c})$$

$$+ \sum_{i \in I_+^{k+1}} l'(t^{k+1} - s_i^{k+1})(\nabla t^{k+1} - x_i) - \sum_{i \in I_+^{k+1}} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i) \quad (\text{A.22d})$$

$$+ \sum_{i \in I_+^{k+1}} l'(t^{k+1} - s_i^{k+1})(\nabla t^{k+1} - x_i) - \sum_{i \in I_+^{k+1}} l'(t(w^{k+1}) - x_i^\top w^{k+1})(\nabla t(w^{k+1}) - x_i). \quad (\text{A.22e})$$

Then relations (A.21) and (A.18) applied to Lemma A.15 imply

$$\left\| \sum_{i \in I_+^k} l'(t^k - s_i^k)(\nabla t^k - x_i) - \sum_{i \in I_+^k} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i) \right\| \leq C_4(\alpha^{k-1} + \alpha^{k-2})$$

for some  $C_4$ , which gives a bound for (A.22b). Bound for (A.22e) is obtained by increasing  $k$  by one. Bounds for (A.22c) and (A.22d) can be find similarly using (A.19). Altogether, we showed

$$\|g(w^k) + g(w^{k+1}) - \nabla f(w^k) - \nabla f(w^{k+1})\| \leq C_1(\alpha^{k-2} + \alpha^{k-1} + \alpha^k + \alpha^{k+1}) \quad (\text{A.23})$$

for some  $C_1$ .

We now estimate

$$\alpha^k \langle g(w^k) - \nabla f(w^k), w^* - w^k \rangle + \alpha^{k+1} \langle g(w^{k+1}) - \nabla f(w^{k+1}), w^* - w^{k+1} \rangle \quad (\text{A.24a})$$

$$= \langle g(w^k) - \nabla f(w^k), \alpha^k(w^* - w^k) \rangle + \langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) \rangle \quad (\text{A.24b})$$

$$= \langle g(w^k) - \nabla f(w^k) + g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^k(w^* - w^k) \rangle \quad (\text{A.24c})$$

$$+ \langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k) \rangle. \quad (\text{A.24d})$$

To estimate (A.24d), we make use of Lemma A.13 to obtain the existence of some  $C_5$  such that

$$\begin{aligned} & \langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k) \rangle \\ & \leq 2B \|\alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k)\| \\ & = 2B \|\alpha^{k+1}(w^* - w^k + \alpha^k g(w^k)) - \alpha^k(w^* - w^k)\| \\ & = 2B \|(\alpha^{k+1} - \alpha^k)w^* + (\alpha^k - \alpha^{k+1})w^k + \alpha^k \alpha^{k+1} g(w^k)\| \\ & \leq C_5 \|\alpha^{k+1} - \alpha^k\| + C_5(\alpha^k)^2 + C_5(\alpha^{k+1})^2. \end{aligned} \quad (\text{A.25})$$

In the last inequality we used the equality  $2ab \leq a^2 + b^2$ . To estimate (A.24c), we can apply (A.23) together with the boundedness of  $\{w^k\}$  to obtain the existence of some  $C_6$  such that

$$\begin{aligned} & \langle g(w^k) - \nabla f(w^k) + g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^k(w^* - w^k) \rangle \\ & \leq C_6(\alpha^{k-2})^2 + C_6(\alpha^{k-1})^2 + C_6(\alpha^k)^2 + C_6(\alpha^{k+1})^2. \end{aligned} \quad (\text{A.26})$$

Plugging (A.25) and (A.26) into (A.24) and summing the terms yields (A.11). Then the assumptions of Theorem A.10 are verified and the theorem statement follows.  $\blacksquare$

### A.4.3 Auxiliary results

#### Lemma A.12

Let  $l$  satisfy (S1)-(S3). Then there exists some  $\hat{C}$  such that for all  $k$  we have

$$\begin{aligned} \sum_{i \in X} l'(s_i^k - t^k) &\geq \hat{C} > 0, \\ \sum_{i \in X} l'(x_i^\top w^k - t(w^k)) &\geq \hat{C} > 0. \end{aligned}$$

#### Proof:

First, we will find an upper bound of  $s_i^k - t^k$ . Fix any index  $i_0$ . Since  $l$  is nonnegative due to (S1), equation (1.31) implies

$$n\tau = \sum_{i \in X} l(s_i^k - t^k) \geq l(s_{i_0}^k - t^k).$$

Moreover, as  $l$  is a strictly increasing function due to (S2) and  $n\tau > 0$ , this means

$$l^{-1}(n\tau) \geq s_{i_0}^k - t^k. \quad (\text{A.27})$$

Since  $i_0$  was an arbitrary index, it holds true for all indices. Then (S3) which leads to a further estimate

$$\begin{aligned} \sum_{i \in X} l'(s_i^k - t^k) &= \sum_{i \in X} l(s_i^k - t^k) \frac{l'(s_i^k - t^k)}{l(s_i^k - t^k)} \geq \sum_{i \in X} l(s_i^k - t^k) \frac{l'(l^{-1}(n\tau))}{l(l^{-1}(n\tau))} \\ &= n\tau \frac{l'(l^{-1}(n\tau))}{l(l^{-1}(n\tau))} = l'(l^{-1}(n\tau)), \end{aligned}$$

where the inequality follows from (A.27) and the following equality from (1.31). Due to (S2) we obtain that  $l'(l^{-1}(n\tau))$  is a positive number, which finishes the proof of the first part. The second part can be obtained in an identical way. ■

#### Lemma A.13

Let  $l$  satisfy (S1)-(S4). Then there exists some  $B$  such that for all  $k$  we have  $\|\nabla f(w^k)\| \leq B$  and  $\|g(w^k)\| \leq B$ .

#### Proof:

Due to (S4) the derivative  $l'$  is bounded by some  $\hat{B}$ . Then Theorem 1.4 and Lemma A.12 imply

$$\|\nabla t(w^k)\| \leq \frac{\hat{B} \sum_{i \in X} \|x_i\|}{\sum_{i \in X} l'(x_i^\top w - t(w))} \leq \frac{\hat{B}}{\hat{C}} \sum_{i \in X} \|x_i\|,$$

which is independent of  $k$ . Then (1.29) and again the boundedness of  $l'$  imply the existence of some  $B$  such that  $\|\nabla f(w^k)\| \leq B$  for all  $k$ . The proof for  $g(w^k)$  can be performed identically. ■

**Lemma A.14**

Consider uniformly bounded positive sequences  $c_1^k, c_2^k, d_1^k, d_2^k, \alpha^k$  and positive constants  $C_1, C_2$  such that for all  $k$  we have  $\|c_1^k - c_2^k\| \leq C_1 \alpha^k$ ,  $\|d_1^k - d_2^k\| \leq C_1 \alpha^k$ ,  $d_1^k \geq C_2$  and  $d_2^k \geq C_2$ . If  $\alpha^k \rightarrow 0$ , then there exists a constant  $C_3$  such that for all sufficiently large  $k$  we have

$$\left\| \frac{c_1^k}{d_1^k} - \frac{c_2^k}{d_2^k} \right\| \leq C_3 \alpha^k.$$

**Proof:**

Since  $d_1^k$  and  $d_2^k$  are bounded away from zero and since  $\alpha^k \rightarrow 0$ , we have

$$\left\| \frac{c_1^k}{d_1^k} - \frac{c_2^k}{d_2^k} \right\| \leq \max \left\{ \frac{c_1^k}{d_1^k} - \frac{c_1^k + C_1 \alpha^k}{d_1^k - C_1 \alpha^k}, \frac{c_1^k}{d_1^k} - \frac{c_1^k - C_1 \alpha^k}{d_1^k + C_1 \alpha^k} \right\}.$$

The first term can be estimated as

$$\left\| \frac{c_1^k}{d_1^k} - \frac{c_1^k + C_1 \alpha^k}{d_1^k - C_1 \alpha^k} \right\| = \left\| \frac{(c_1^k + d_1^k) C_1 \alpha^k}{d_1^k (d_1^k - C_1 \alpha^k)} \right\| \leq \frac{(c_1^k + d_1^k) C_1 \alpha^k}{C_2 |d_1^k - C_1 \alpha^k|}.$$

Since  $\alpha^k \rightarrow 0$  by assumption, for large  $k$  we have  $\|d_1^k - C_1 \alpha^k\| \geq \frac{1}{2} C_2$ . Since the sequences are uniformly bounded, the statement follows. ■

**Lemma A.15**

Consider scalars  $a_i, c_i$  and vectors  $b_i, d_i$ . If there is some  $\hat{C}$  such that  $\|a_i\| \leq \hat{C}$  and  $\|d_i\| \leq \hat{C}$ , then

$$\left\| \sum_{i=1}^n a_i b_i - \sum_{i=1}^n c_i d_i \right\| \leq \hat{C} \sum_{i=1}^n (\|a_i - c_i\| + \|b_i - d_i\|).$$

**Proof:**

It is simple to verify

$$\left\| \sum_{i=1}^n a_i b_i - \sum_{i=1}^n c_i d_i \right\| \leq \sum_{i=1}^n \|d_i\| \|a_i - c_i\| + \sum_{i=1}^n \|a_i\| \|b_i - d_i\|,$$

from which the statement follows. ■



## Non-linear case

---

### B.1 Convex conjugate

Recall that for a convex lower semi-continuous function  $f$ , its convex conjugate  $f^*$  is defined by

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \text{dom } f} \{\mathbf{y}^\top \mathbf{x} - f(\mathbf{x})\} = - \inf_{\mathbf{x} \in \text{dom } f} \{f(\mathbf{x}) - \mathbf{y}^\top \mathbf{x}\}.$$

For details, see [15, page 91]. For the hinge loss function, it is well-known [36] that

$$\begin{aligned} l_{\text{hinge}}(x) &= \max\{0, 1 + \vartheta x\}, \\ l_{\text{hinge}}^*(y) &= \begin{cases} -\frac{y}{\vartheta} & \text{if } y \in [0, \vartheta], \\ \infty & \text{otherwise.} \end{cases} \end{aligned} \quad (\text{B.1})$$

Similarly, for the truncated quadratic loss function, it is well-known [37] that

$$\begin{aligned} l_{\text{quad}}(x) &= (\max\{0, 1 + \vartheta x\})^2, \\ l_{\text{quad}}^*(y) &= \begin{cases} \frac{y^2}{4\vartheta^2} - \frac{y}{\vartheta} & \text{if } y \geq 0, \\ \infty & \text{otherwise.} \end{cases} \end{aligned} \quad (\text{B.2})$$

### B.2 Proofs for dual problems

#### Theorem 2.1: *TopPushK* dual formulation

The dual problem corresponding to the problem (2.5) has the form

$$\underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} - C \sum_{i=1}^{n^+} l^*\left(\frac{\alpha_i}{C}\right) \quad (\text{2.7a})$$

$$\text{subject to } \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (\text{2.7b})$$

$$0 \leq \beta_j \leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \quad \forall j = 1, 2, \dots, n^-, \quad (\text{2.7c})$$

where  $l^*$  is a conjugate of the surrogate loss function  $l$  from (2.5) and  $\mathbb{K}$  was defined

in (2.6).

### Proof of Theorem 2.1 on page 24:

Recall the primal formulation of the *TopPushK* problem (2.5)

$$\underset{\mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l \left( \frac{1}{K} \sum_{j=1}^K s_{[j]}^- - \mathbf{w}^\top \mathbf{x}_i^+ \right).$$

Using [38, Lemma 1] and the fact that the surrogate function  $l$  is non-decreasing, we can write

$$\begin{aligned} \sum_{i=1}^{n^+} l \left( \frac{1}{K} \sum_{j=1}^K s_{[j]}^- - \mathbf{w}^\top \mathbf{x}_i^+ \right) &= \sum_{i=1}^{n^+} l \left( \frac{1}{K} \min_t \left\{ Kt + \sum_{j=1}^{n^-} \max\{0, s_j^- - t\} \right\} - \mathbf{w}^\top \mathbf{x}_i^+ \right) \\ &= \min_t \sum_{i=1}^{n^+} l \left( t + \frac{1}{K} \sum_{j=1}^{n^-} \max\{0, \mathbf{w}^\top \mathbf{x}_j^- - t\} - \mathbf{w}^\top \mathbf{x}_i^+ \right). \end{aligned}$$

Using auxiliary variables  $\mathbf{y} \in \mathbb{R}^{n^+}$  and  $\mathbf{z} \in \mathbb{R}^{n^-}$ , we observe that (2.5) is equivalent to

$$\begin{aligned} \underset{\mathbf{w}, t, \mathbf{y}, \mathbf{z}}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l(y_i) \\ \text{subject to} \quad & y_i = t + \frac{1}{K} \sum_{j=1}^{n^-} z_j - \mathbf{w}^\top \mathbf{x}_i^+, \quad \forall i = 1, 2, \dots, n^+, \\ & z_j \geq \mathbf{w}^\top \mathbf{x}_j^- - t, \quad \forall j = 1, 2, \dots, n^-, \\ & z_j \geq 0, \quad \forall j = 1, 2, \dots, n^-. \end{aligned}$$

The dual objective function is  $g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \min_{\mathbf{w}, t, \mathbf{y}, \mathbf{z}} L(\mathbf{w}, t, \mathbf{y}, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ , where the Lagrange function  $L$  is defined by

$$\begin{aligned} L(\mathbf{w}, t, \mathbf{y}, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l(y_i) + \sum_{i=1}^{n^+} \alpha_i \left( t + \frac{1}{K} \sum_{j=1}^{n^-} z_j - \mathbf{w}^\top \mathbf{x}_i^+ - y_i \right) \\ &\quad + \sum_{j=1}^{n^-} \beta_j (\mathbf{w}^\top \mathbf{x}_j^- - t - z_j) - \sum_{j=1}^{n^-} \gamma_j z_j \end{aligned} \tag{B.3}$$

Since this function is separable in primal variables, it can be minimized with respect to each variable separately. Optimality conditions with respect to  $t$  and  $\mathbf{z}$  read

$$\begin{aligned} \sum_{i=1}^{n^+} \alpha_i - \sum_{j=1}^{n^-} \beta_j &= 0, \\ \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i - \beta_j - \gamma_j &= 0, \quad \forall j = 1, 2, \dots, n^-. \end{aligned}$$

From the first condition we deduce constraint (2.7b). Plugging the feasibility condition  $\gamma_j \geq 0$  into the second optimal condition and combining it with the feasibility conditions  $\beta_j \geq 0$  yields constraint (2.7c). By minimizing (B.3) with respect to  $\mathbf{w}$  we deduce

$$\mathbf{w} = \sum_{i=1}^{n^+} \alpha_i \mathbf{x}_i^+ - \sum_{j=1}^{n^-} \beta_j \mathbf{x}_j^- = \begin{pmatrix} \mathbb{X}^+ \\ -\mathbb{X}^- \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix},$$

where  $\mathbb{X}^+$ ,  $\mathbb{X}^-$  are matrices of positive and negative samples respectively (each row corresponds to one sample). Finally, minimization of (B.3) with respect to  $\mathbf{y}$  yields

$$C \min_{y_i} \left( l(y_i) - \frac{\alpha_i}{C} y_i \right) = -Cl^* \left( \frac{\alpha_i}{C} \right), \quad \forall i = 1, 2, \dots, n^+.$$

Plugging all these relations into (B.3) yields the objective function of (2.7), which finishes the proof. ■

### Theorem 2.2: *Pat&Mat* dual formulation

The dual problem corresponding to the problem (2.11) has the form

$$\underset{\alpha, \beta, \delta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - C \sum_{i=1}^{n^+} l_1^* \left( \frac{\alpha_i}{C} \right) - \delta \sum_{j=1}^{n^-} l_2^* \left( \frac{\beta_j}{\delta} \right) - \delta n\tau \quad (2.12a)$$

$$\text{subject to} \quad \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (2.12b)$$

$$\delta \geq 0, \quad (2.12c)$$

where  $l_1^*$ ,  $l_2^*$  are conjugates of the surrogate loss functions  $l_1$ ,  $l_2$  from (2.11) and  $\mathbb{K}$  was defined in (2.6).

### *Proof of Theorem 2.2 on page 25:*

Let us first realize that the primal *Pat&Mat* problem (2.11) is equivalent to

$$\begin{aligned} \underset{\mathbf{w}, t, \mathbf{y}, \mathbf{z}}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l_1(y_i) \\ \text{subject to} \quad & \sum_{j=1}^{n^-} l_2(z_j) \leq n\tau, \\ & y_i = t - \mathbf{w}^\top \mathbf{x}_i^+, \quad \forall i = 1, 2, \dots, n^+, \\ & z_j = \mathbf{w}^\top \mathbf{x}_j^- - t, \quad \forall j = 1, 2, \dots, n^-. \end{aligned}$$

Then the dual function is  $g(\alpha, \beta, \delta) = \min_{\mathbf{w}, t, \mathbf{y}, \mathbf{z}} L(\mathbf{w}, t, \mathbf{y}, \mathbf{z}; \alpha, \beta, \delta)$ , where the Lagrange function  $L$  is defined by

$$\begin{aligned} L(\mathbf{w}, t, \mathbf{y}, \mathbf{z}; \alpha, \beta, \delta) = & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n^+} l_1(y_i) + \sum_{i=1}^{n^+} \alpha_i (t - \mathbf{w}^\top \mathbf{x}_i^+ - y_i) \\ & + \sum_{j=1}^{n^-} \beta_j (\mathbf{w}^\top \mathbf{x}_j^- - t - z_j) + \delta \left( \sum_{j=1}^{n^-} l_2(z_j) - n\tau \right). \end{aligned} \quad (B.4)$$

Since this function is separable in primal variables, it can be minimized with respect to each variable separately. Optimality condition with respect to  $t$  read

$$\sum_{i=1}^{n^+} \alpha_i - \sum_{j=1}^{n^-} \beta_j = 0,$$

from which we deduce constraint (2.12b). By minimizing (B.4) with respect to  $\mathbf{w}$  we deduce

$$\mathbf{w} = \sum_{i=1}^{n^+} \alpha_i \mathbf{x}_i^+ - \sum_{j=1}^{n^-} \beta_j \mathbf{x}_j^- = \begin{pmatrix} \mathbb{X}^+ \\ -\mathbb{X}^- \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix},$$

where  $\mathbb{X}$ ,  $\mathbb{X}^+$  are matrices of all and positive samples respectively (each row corresponds to one sample). Finally, minimization of (B.4) with respect to  $\mathbf{y}$  and  $\mathbf{z}$  yields

$$\begin{aligned} C \min_{y_i} \left( l_1(y_i) - \frac{\alpha_i}{C} y_i \right) &= -C l_1^* \left( \frac{\alpha_i}{C} \right), \\ \delta \min_{z_j} \left( l_2(z_j) - \frac{\beta_j}{\delta} z_j \right) &= -\delta l_2^* \left( \frac{\beta_j}{\delta} \right). \end{aligned}$$

Plugging all these relations into (B.4) yields the objective function of (2.12), which finishes the proof.  $\blacksquare$

## B.3 Computing $\Delta^*$ with truncated quadratic surrogate

### B.3.1 *TopPushK*

**Theorem 2.3:** Update rule for  $\Delta^*$  for *TopPushK* with truncated quadratic loss

Consider problem (2.18). Then the optimal step  $\Delta^*$  equals to

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]}(\gamma), \quad (2.19)$$

where there are the following three cases (each corresponding to one update rule in (2.17)):

- For any  $1 \leq k, l \leq n^+$  we have

$$\begin{aligned} \Delta_{lb} &= -\alpha_k, \\ \Delta_{ub} &= \alpha_l, \\ \gamma &= -\frac{s_k - s_l + \frac{1}{2C\vartheta^2}(\alpha_k - \alpha_l)}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{C\vartheta^2}}. \end{aligned}$$

- For any  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n$  we define  $\hat{l} = l - n^+$  and  $\beta_{\max} = \max_{j \in \{1, 2, \dots, n^-\} \setminus \{\hat{l}\}} \beta_j$ . Then we have

$$\begin{aligned} \Delta_{lb} &= \begin{cases} \max\{-\alpha_k, -\beta_{\hat{l}}\} & K = 1, \\ \max\{-\alpha_k, -\beta_{\hat{l}}, K\beta_{\max} - \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} +\infty & K = 1, \\ \frac{1}{K-1} \left( \sum_{i=1}^{n^+} \alpha_i - K\beta_{\hat{l}} \right) & \text{otherwise,} \end{cases} \\ \gamma &= -\frac{s_k + s_l - \frac{1}{\vartheta} + \frac{1}{2C\vartheta^2} \alpha_k}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk} + \frac{1}{2C\vartheta^2}}. \end{aligned}$$



- For any  $n^+ + 1 \leq k, l \leq n$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$  and then have

$$\begin{aligned}\Delta_{lb} &= \begin{cases} -\beta_{\hat{k}} & K = 1, \\ \max\{-\beta_{\hat{k}}, \beta_{\hat{l}} - \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} \beta_{\hat{l}} & K = 1, \\ \min\{\beta_{\hat{l}}, \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i - \beta_{\hat{k}}\} & \text{otherwise,} \end{cases} \\ \gamma &= -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}}.\end{aligned}$$

**Proof of Theorem 2.3 on page 27:**

We will show that for each update rule (2.17) and for fixed  $\alpha, \beta$ , problem (2.18) can be rewritten as a quadratic one-dimensional problem

$$\begin{aligned}\underset{\Delta}{\text{maximize}} \quad & -\frac{1}{2}a(\alpha, \beta)\Delta^2 - b(\alpha, \beta)\Delta - c(\alpha, \beta) \\ \text{subject to} \quad & \Delta_{lb}(\alpha, \beta) \leq \Delta \leq \Delta_{ub}(\alpha, \beta).\end{aligned}$$

where  $a, b, c, \Delta_{lb}, \Delta_{ub}$  do not depend on  $\Delta$ . The optimal solution to this problem is

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]} \left( -\frac{b}{a} \right), \quad (\text{B.5})$$

which amounts to (2.19). Before discussing the three update rules (2.17), we realize that (2.18b) is always satisfied after the update. For the three updates we have:

- For update rule (2.17a) with  $1 \leq k, l \leq n^+$ , constraint (2.18d) is satisfied since no  $\beta_j$  was updated and the sum of all  $\alpha_i$  did not change. Constraint (2.18c) reads  $-\alpha_k \leq \Delta \leq \alpha_l$  and objective (2.18a) can be rewritten as

$$-\frac{1}{2} \left[ \mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{C\vartheta^2} \right] \Delta^2 - \left[ s_k - s_l + \frac{1}{2C\vartheta^2} (\alpha_k - \alpha_l) \right] \Delta + c(\alpha, \beta).$$

- For update rule (2.17b) with  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n^+ + n^-$  we define  $\hat{l} = l - n^+$ . Constraint (2.18c) reads  $\Delta \geq -\alpha_k$ . Denoting  $\beta_{\max} = \max_{j \in \{1, 2, \dots, n^-\} \setminus \{\hat{l}\}} \beta_j$ , then for any  $K \geq 2$  constraint (2.18d) reads

$$\begin{aligned}0 \leq \beta_{\hat{l}} + \Delta &\leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i + \frac{1}{K} \Delta, \\ 0 \leq \beta_{\max} &\leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i + \frac{1}{K} \Delta.\end{aligned} \quad (\text{B.6})$$

If  $K = 1$ , the upper bounds for  $\beta_j$  may be omitted as discussed in Section 2.3.4. Combining this with  $\Delta \geq -\alpha_k$  yields the lower and upper bound of  $\Delta$ . Using update rule (2.17b), objective (2.18a) can be rewritten as

$$-\frac{1}{2} \left[ \mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk} + \frac{1}{2C\vartheta^2} \right] \Delta^2 - \left[ s_k + s_l - \frac{1}{\vartheta} + \frac{1}{2C\vartheta^2} \alpha_k \right] \Delta + c(\alpha, \beta).$$

- For update rule (2.17c) with  $n^+ + 1 \leq k, l \leq n^+ + n^-$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$ . Since no  $\alpha_i$  was updated, constraint (2.18c) is always satisfied. Moreover, since we update only two coordinates of  $\beta$ , constraint (2.18d) for any  $K \geq 2$  reads

$$\begin{aligned} 0 \leq \beta_{\hat{k}} + \Delta &\leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \\ 0 \leq \beta_{\hat{l}} - \Delta &\leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \end{aligned} \tag{B.7}$$

As in the previous case, the upper bounds for  $\beta_j$  may be omitted for  $K = 1$ . Combining the previous results yields the lower and upper bound of  $\Delta$ . Using update rule (2.17c), objective (2.18a) can be rewritten as

$$-\frac{1}{2}[\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}]\Delta^2 - [s_k - s_l]\Delta + c(\alpha, \beta).$$

The proofs follows by plugging these cases into the solution (B.5). ■

#### B.3.2 *Pat&Mat*

Plugging the conjugate (B.2) of the truncated quadratic loss (2.2) into *Pat&Mat* dual formulation (2.12) yields

$$\begin{aligned} \underset{\alpha, \beta, \delta}{\text{maximize}} \quad & -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{\vartheta_1} \sum_{i=1}^{n^+} \alpha_i - \frac{1}{4C\vartheta_1^2} \sum_{i=1}^{n^+} \alpha_i^2 \\ & + \frac{1}{\vartheta_2} \sum_{j=1}^{n^-} \beta_j - \frac{1}{4\delta\vartheta_2^2} \sum_{j=1}^{n^-} \beta_j^2 - \delta n\tau \end{aligned} \tag{B.8a}$$

$$\text{subject to } \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \tag{B.8b}$$

$$\alpha_i \geq 0, \quad \forall i = 1, 2, \dots, n^+, \tag{B.8c}$$

$$\beta_j \geq 0, \quad \forall j = 1, 2, \dots, n^-, \tag{B.8d}$$

$$\delta \geq 0, \tag{B.8e}$$

The following theorem provides a formula for the optimal step  $\Delta^*$  for the update rule (2.17). Note that we do not perform a joint minimization in  $(\alpha_k, \beta_l, \delta)$  but perform a minimization with respect to  $(\alpha_k, \beta_l)$ , update these two values and then optimize the objective with respect to  $\delta$ .

#### **Theorem B.4: Update rule for $\Delta^*$ for *Pat&Mat* with truncated quadratic loss**

Consider problem (B.8). Then the optimal step  $\Delta^*$  equals to

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]}(\gamma), \tag{B.9}$$

where there are the following three cases (each corresponding to one update rule in (2.17)):

- If  $1 \leq k, l \leq n^+$ , then we have

$$\begin{aligned}\Delta_{lb} &= -\alpha_k, \\ \Delta_{ub} &= \alpha_l, \\ \gamma &= -\frac{s_k - s_l + \frac{\alpha_k - \alpha_l}{2C\vartheta_1^2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{C\vartheta_1^2}}, \\ \delta^* &= \delta.\end{aligned}$$

- If  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n$ , then defining  $\hat{l} = l - n^+$  we have

$$\begin{aligned}\Delta_{lb} &= \max\{-\alpha_k, -\beta_{\hat{l}}\}, \\ \Delta_{ub} &= +\infty, \\ \gamma &= -\frac{s_k + s_l - \frac{1}{\vartheta_1} + \frac{\alpha_k}{2C\vartheta_1^2} - \frac{1}{\vartheta_2} + \frac{\beta_{\hat{l}}}{2\delta\vartheta_2^2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk} + \frac{1}{2C\vartheta_1^2} + \frac{1}{2\delta\vartheta_2^2}}, \\ \delta^* &= \sqrt{\delta^2 + \frac{1}{4\vartheta_2 n\tau}(\Delta^{*2} + 2\Delta^*\beta_{\hat{l}})}.\end{aligned}$$

- If  $n^+ + 1 \leq k, l \leq n$ , then defining  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$  we have

$$\begin{aligned}\Delta_{lb} &= -\beta_{\hat{k}}, \\ \Delta_{ub} &= \beta_{\hat{l}}, \\ \gamma &= -\frac{s_k - s_l + \frac{\beta_{\hat{k}} - \beta_{\hat{l}}}{2\delta\vartheta_2^2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{\delta\vartheta_2^2}}, \\ \delta^* &= \sqrt{\delta^2 + \frac{1}{2\vartheta_2 n\tau}(\Delta^{*2} + \Delta^*(\beta_{\hat{k}} - \beta_{\hat{l}}))}.\end{aligned}$$

**Proof:**

In the beginning of this subsection we derived problem (B.8). As in the proof of Theorem 2.3, we show, that for each of update rules (2.17) and for fixed  $\alpha, \beta, \delta$ , this problem can be rewritten as a simple one-dimensional quadratic problem with bound constraints. In this case, however, we have to also consider the third primal variable  $\delta$ . For fixed  $\alpha$  and  $\beta$ , maximizing objective function (B.8a) with respect to  $\delta$  leads to the

$$\begin{aligned}\underset{\delta}{\text{maximize}} \quad & -(n\tau)\delta - \left(\frac{1}{4\vartheta_2^2} \sum_{j=1}^{n^-} \beta_j^2\right) \frac{1}{\delta}, \\ \text{subject to} \quad & \delta \geq 0.\end{aligned}$$

The solution of this problem equals to

$$\delta^* = \sqrt{\frac{1}{4\vartheta_2^2 n\tau} \sum_{j=1}^n \beta_j^2}. \quad (\text{B.10})$$

In the following list, we discuss each of update rules (2.17):

- For update rule (2.17a) and any  $1 \leq k, l \leq n^+$ , constraint (B.8d) is satisfied since no  $\beta_j$  was updated. Constraint (B.8c) reads  $-\alpha_k \leq \Delta \leq \alpha_l$  while objective (B.8a) can be rewritten as

$$-\frac{1}{2} \left[ \mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{C\vartheta_1^2} \right] \Delta^2 - \left[ s_k - s_l + \frac{1}{2C\vartheta_1^2} (\alpha_k - \alpha_l) \right] \Delta + c(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Since optimal  $\delta$  is given by (B.10) and no  $\beta_j$  was updated, the optimal  $\delta$  does not change.

- For update rule (2.17b) with  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n$  we define  $\hat{l} = l - n^+$ . In this case, constraints (B.8c, B.8d) can be written in a simple form  $\Delta \geq \max\{-\alpha_k, -\beta_{\hat{l}}\}$  and  $\Delta$  has no upper bound. Objective (B.8a) can be rewritten as

$$-\frac{1}{2} \left[ \mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk} + \frac{1}{2C\vartheta_1^2} + \frac{1}{2\delta\vartheta_2^2} \right] \Delta^2 \dots \\ - \left[ s_k + s_l - \frac{1}{\vartheta_1} - \frac{1}{\vartheta_2} + \frac{\alpha_k}{2C\vartheta_1^2} + \frac{\beta_{\hat{l}}}{2\delta\vartheta_2^2} \right] \Delta + c(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

We know that the optimal  $\delta^*$  is given by (B.10), then

$$\delta^* = \sqrt{\frac{1}{4\vartheta_2^2 n\tau} \left( \sum_{j \neq \hat{l}} \beta_j^2 + (\beta_{\hat{l}} + \Delta^*)^2 \right)} = \sqrt{\delta^2 + \frac{1}{4\vartheta_2^2 n\tau} (\Delta^{*2} + 2\Delta^* \beta_{\hat{l}})}.$$

- For update rule (2.17c) with  $n^+ + 1 \leq k, l \leq n^+ + n^-$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$ . Since no  $\alpha_i$  was updated, constraint (B.8c) is always satisfied. Constraint (B.8d) can be written in a simple form  $-\beta_{\hat{k}} \leq \Delta \leq \beta_{\hat{l}}$  and objective (B.8a) can be rewritten as

$$-\frac{1}{2} \left[ \mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk} + \frac{1}{2\delta\vartheta_2^2} \right] \Delta^2 - \left[ s_k - s_l + \frac{\beta_{\hat{k}} - \beta_{\hat{l}}}{\delta\vartheta_2^2} \right] \Delta + c(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

We know that the optimal  $\delta^*$  is given by (B.10), then

$$\delta^* = \sqrt{\frac{1}{4\vartheta_2^2 n\tau} \left( \sum_{j \notin \{\hat{l}, \hat{k}\}} \beta_j^2 + (\beta_{\hat{k}} + \Delta^*)^2 + (\beta_{\hat{l}} - \Delta^*)^2 \right)} = \sqrt{\delta^2 + \frac{1}{2\vartheta_2^2 n\tau} (\Delta^{*2} + \Delta^* (\beta_{\hat{k}} - \beta_{\hat{l}}))}.$$

The proofs follows by plugging these cases into the solution (B.9). ■

## B.4 Computing $\Delta^*$ with hinge loss function

In this section, we provide the results when the truncated quadratic surrogate is replaced by the hinge surrogate. Since the proofs are identical, we omit them.

### B.4.1 *TopPushK*

Plugging the conjugate (B.1) of the hinge loss (2.3) into *TopPushK* dual formulation (2.7) yields

$$\underset{\alpha, \beta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{\vartheta} \sum_{i=1}^{n^+} \alpha_i \quad (\text{B.11a})$$

$$\text{subject to } \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^{n^-} \beta_j, \quad (\text{B.11b})$$

$$0 \leq \alpha_i \leq C\vartheta, \quad \forall i = 1, 2, \dots, n^+, \quad (\text{B.11c})$$

$$0 \leq \beta_j \leq \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i, \quad \forall j = 1, 2, \dots, n^-. \quad (\text{B.11d})$$

This is a convex quadratic problem. Moreover, for  $K = 1$ , the upper limit in (B.11d) is always satisfied due to (B.11b) and the problem can be simplified. The following theorem provides a formula for optimal  $\Delta$  for each of update rules (2.17).

#### Theorem B.5: Update rule for $\Delta^*$ for *TopPushK* with hinge loss

Consider problem (B.11). Then

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]}(\gamma),$$

where there are the following cases:

- For any  $1 \leq k, l \leq n^+$  we have

$$\begin{aligned} \Delta_{lb} &= \min\{-\alpha_k, \alpha_l - C\vartheta\}, \\ \Delta_{ub} &= \max\{C\vartheta - \alpha_k, \alpha_l\}, \\ \gamma &= -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}}. \end{aligned}$$

- For any  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n$  we define  $\hat{l} = l - n^+$  and  $\beta_{\max} = \max_{j \in \{1, 2, \dots, n^-\} \setminus \{\hat{l}\}} \beta_j$ . Then we have

$$\begin{aligned} \Delta_{lb} &= \begin{cases} \min\{-\alpha_k, -\beta_{\hat{l}}\} & K = 1, \\ \min\{-\alpha_k, -\beta_{\hat{l}}, K\beta_{\max} - \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} C\vartheta - \alpha_k & K = 1, \\ \max\{C\vartheta - \alpha_k, \frac{1}{K-1}(\sum_{i=1}^{n^+} \alpha_i - K\beta_{\hat{l}})\} & \text{otherwise.} \end{cases} \\ \gamma &= -\frac{s_k + s_l - \frac{1}{\vartheta}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk}}. \end{aligned}$$

- For any  $n^+ + 1 \leq k, l \leq n^+ + n^-$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$  and then we

have

$$\begin{aligned}\Delta_{lb} &= \begin{cases} -\beta_{\hat{k}} & K = 1, \\ \min\{-\beta_{\hat{k}}, \beta_{\hat{l}} - \frac{1}{K} \sum_{i=1}^{n^+} \alpha_i\} & \text{otherwise,} \end{cases} \\ \Delta_{ub} &= \begin{cases} \beta_{\hat{l}} & K = 1, \\ \max\{\frac{1}{K} \sum_{i=1}^{n^+} \alpha_i - \beta_{\hat{k}}, \beta_{\hat{l}}\} & \text{otherwise.} \end{cases} \\ \gamma &= -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}}.\end{aligned}$$

### B.4.2 *Pat&Mat*

Plugging the conjugate (B.1) of the hinge loss (2.3) into *Pat&Mat* dual formulation (2.12) yields

$$\underset{\alpha, \beta, \delta}{\text{maximize}} \quad -\frac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^\top \mathbb{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{\vartheta_1} \sum_{i=1}^{n^+} \alpha_i + \frac{1}{\vartheta_2} \sum_{j=1}^n \beta_j - \delta n \tau \quad (\text{B.12a})$$

$$\text{subject to } \sum_{i=1}^{n^+} \alpha_i = \sum_{j=1}^n \beta_j, \quad (\text{B.12b})$$

$$0 \leq \alpha_i \leq C\vartheta_1, \quad \forall i = 1, 2, \dots, n^+, \quad (\text{B.12c})$$

$$0 \leq \beta_j \leq \delta\vartheta_2, \quad \forall j = 1, 2, \dots, n, \quad (\text{B.12d})$$

$$\delta \geq 0, \quad (\text{B.12e})$$

The following theorem provides a formula for optimal  $\Delta$  for each of update rules (2.17).

#### Theorem B.6: Update rule for $\Delta^*$ for *Pat&Mat* with hinge loss

Consider problem (B.12). Then

$$\Delta^* = \text{clip}_{[\Delta_{lb}, \Delta_{ub}]}(\gamma),$$

where there are the following cases:

- For any  $1 \leq k, l \leq n^+$  we have

$$\begin{aligned}\Delta_{lb} &= \min\{-\alpha_k, \alpha_l - C\vartheta_1\}, \\ \Delta_{ub} &= \max\{C\vartheta_1 - \alpha_k, \alpha_l\}, \\ \gamma &= -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}}, \\ \delta^* &= \delta.\end{aligned}$$

- For any  $1 \leq k \leq n^+$  and  $n^+ + 1 \leq l \leq n^+$  we define  $\hat{l} = l - n^+$  and  $\beta_{\max} = \max_{j \in \{1, 2, \dots, n\} \setminus \{\hat{l}\}} \beta_j$ . Then we have

$$\Delta_{lb} = \max\{-\alpha_k, -\beta_{\hat{l}}\}, \quad \Delta_{ub} = C\vartheta_1 - \alpha_k$$

and the optimal solution is one of the two following possibilities which maximizes the original objective:

1. If  $\beta_{\hat{l}} + \Delta^* \leq \beta_{\max}$ , then

$$\gamma = -\frac{s_k + s_l - \frac{1}{\vartheta_1} - \frac{1}{\vartheta_2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk}},$$

$$\delta^* = \frac{\beta_{\max}}{\vartheta_2}.$$

2. If  $\beta_{\hat{l}} + \Delta^* \geq \beta_{\max}$ , then

$$\gamma = -\frac{s_k + s_l - \frac{1}{\vartheta_1} - \frac{1-n\tau}{\vartheta_2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} + \mathbb{K}_{kl} + \mathbb{K}_{lk}},$$

$$\delta^* = \frac{\beta_{\hat{l}} + \Delta^*}{\vartheta_2}.$$

- For any  $n^+ + 1 \leq k, l \leq n$  we define  $\hat{k} = k - n^+$ ,  $\hat{l} = l - n^+$ . Then we have

$$\Delta_{lb} = -\beta_{\hat{k}}, \quad \Delta_{ub} = \beta_{\hat{l}},$$

and the optimal solution is one of the three following possibilities which maximizes the original objective:

1. If  $\beta_{\max} \geq \max\{\beta_{\hat{k}} + \Delta^*, \beta_{\hat{l}} - \Delta^*\}$ , then

$$\gamma = -\frac{s_k - s_l}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}},$$

$$\delta^* = \frac{\beta_{\max}}{\vartheta_2}.$$

2. If  $\beta_{\hat{k}} + \Delta^* \geq \max\{\beta_{\max}, \beta_{\hat{l}} - \Delta^*\}$ , then

$$\gamma = -\frac{s_k - s_l + \frac{n\tau}{\vartheta_2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}},$$

$$\delta^* = \frac{\beta_{\hat{k}} + \Delta^*}{\vartheta_2}.$$

3. If  $\beta_{\hat{l}} - \Delta^* \geq \max\{\beta_{\hat{k}} + \Delta^*, \beta_{\max}\}$ , then

$$\gamma = -\frac{s_k - s_l - \frac{n\tau}{\vartheta_2}}{\mathbb{K}_{kk} + \mathbb{K}_{ll} - \mathbb{K}_{kl} - \mathbb{K}_{lk}},$$

$$\delta^* = \frac{\beta_{\hat{l}} - \Delta^*}{\vartheta_2}.$$





# Bibliography

---

- [1] Martin Grill and Tomáš Pevný. Learning combination of anomaly detectors for security domain. *Computer Networks*, 107:55–63, 2016.
- [2] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.
- [3] Shivani Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 839–850. SIAM, 2011.
- [4] Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *J. Mach. Learn. Res.*, 10:2233–2271, December 2009.
- [5] Nan Li, Rong Jin, and Zhi-Hua Zhou. Top rank optimization in linear time. In *Advances in neural information processing systems*, NIPS’14, pages 1502–1510, Cambridge, MA, USA, 2014. MIT Press.
- [6] Stephen Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In *Advances in neural information processing systems*, pages 953–961, 2012.
- [7] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML ’05, pages 377–384, New York, NY, USA, 2005. ACM.
- [8] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Surrogate functions for maximizing precision at the top. In *International Conference on Machine Learning*, pages 189–198, 2015.
- [9] Elad ET Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. In *Artificial Intelligence and Statistics*, pages 832–840, 2017.
- [10] Dirk Tasche. A plug-in approach to maximising precision at the top and recall at the top. *arXiv preprint arXiv:1804.03077*, 2018.
- [11] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass svm. In *Advances in Neural Information Processing Systems*, pages 325–333, 2015.
- [12] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1533–1554, 2018.

- [13] Alan Mackey, Xiyang Luo, and Elad Eban. Constrained classification and ranking via quantiles. *arXiv preprint arXiv:1803.00067*, 2018.
- [14] Ao Zhang, Nan Li, Jian Pu, Jun Wang, Junchi Yan, and Hongyuan Zha. *tau-fpl*: Tolerance-constrained learning in linear time. *arXiv preprint arXiv:1801.04701*, 2018.
- [15] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [16] Václav Mácha, Lukáš Adam, and Václav Šmídl. Nonlinear classifiers for ranking problems based on kernelized svm. *arXiv preprint arXiv:2002.11436*, 2020.
- [17] Lukáš Adam and Martin Branda. Machine learning approach to chance-constrained problems: An algorithm based on the stochastic gradient descent. *arXiv preprint arXiv:1905.10986*, 2019.
- [18] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.
- [19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] Vincent G Sigillito, Simon P Wing, Larrie V Hutton, and Kile B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.
- [21] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5):235, 2016.
- [22] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.
- [23] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7:1–30, 2006.
- [24] Lukáš Adam, Václav Mácha, Václav Šmídl, and Tomáš Pevný. General framework for binary classification on top samples. *arXiv preprint arXiv:2002.10923*, 2019.
- [25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [26] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [27] 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*, pages 408–415. ACM, 2008.
- [28] Zeynep Batmaz, Ali Yurekli, Alper Bilge, and Cihan Kaleli. A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, 52(1):1–37, 2019.

- 
- [29] Tino Werner. A review on ranking problems in statistical learning. *arXiv preprint arXiv:1909.02998*, 2019.
  - [30] Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate descent method for large-scale l2-loss linear support vector machines. *Journal of Machine Learning Research*, 9(Jul):1369–1398, 2008.
  - [31] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.
  - [32] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.
  - [33] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
  - [34] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
  - [35] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
  - [36] Shai Shnlev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *31st International Conference on Machine Learning, ICML 2014*, volume 1, page 111, 2014.
  - [37] Takafumi Kanamori, Akiko Takeda, and Taiji Suzuki. Conjugate relation between loss functions and uncertainty sets in classification problems. *The Journal of Machine Learning Research*, 14(1):1461–1504, 2013.
  - [38] Wlodzimierz Ogryczak and Arie Tamir. Minimizing the sum of the k largest functions in linear time. *Information Processing Letters*, 85(3):117–122, 2003.