Optimization Methods and Software

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/goms20

General framework for binary classification on top samples

L. Adam, V. Mácha, V. Šmídl & T. Pevný

To cite this article: L. Adam, V. Mácha, V. Šmídl & T. Pevný (2022) General framework for binary classification on top samples, Optimization Methods and Software, 37:5, 1636-1667, DOI: 10.1080/10556788.2021.1965601

To link to this article: https://doi.org/10.1080/10556788.2021.1965601

OPTIMIZATION METHODS & SOFTWARE
2022, VOL. 37, NO. 5, 1636–1667
https://doi.org/10.1080/10556788.2021.1965601

Taylor & Francis
Taylor & Francis Group

General framework for binary classification on top samples

L. Adam [a], V. Mácha[b], V. Šmídl [a,c] and T. Pevný[a]

[a]Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic; [b]Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czech Republic; [c]The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague, Czech Republic

ABSTRACT

Many binary classification problems minimize misclassification above (or below) a threshold. We show that instances of ranking problems, accuracy at the top, or hypothesis testing may be written in this form. We propose a general framework to handle these classes of problems and show which formulations (both known and newly proposed) fall into this framework. We provide a theoretical analysis of this framework and mention selected possible pitfalls the formulations may encounter. We show the convergence of the stochastic gradient descent for selected formulations even though the gradient estimate is inherently biased. We suggest several numerical improvements, including the implicit derivative and stochastic gradient descent. We provide an extensive numerical study.

ARTICLE HISTORY
Received 26 October 2020
Accepted 12 May 2021

KEYWORDS
General framework; classification; ranking; accuracy at the top; Neyman–Pearson; Pat&Mat

AMS CLASSIFICATIONS
90C20; 49M05; 65K10; 49K10

1. Introduction

Many binary classification problems focus on separating the dataset by a linear hyperplane $\boldsymbol{w}^\top \boldsymbol{x} - t$. A sample $\boldsymbol{x}$ is deemed to be positive or relevant (depending on the application) if its score $\boldsymbol{w}^\top \boldsymbol{x}$ 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

CONTACT L. Adam lukas.adam.cr@gmail.com, adam@utia.cas.cz

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 [9], 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 [8] maximizing the area under the ROC curve, the Infinite Push [2] or the p-norm push [19] 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 [16], where the authors performed the limit $p \to \infty$ in p-norm push and obtained the linear complexity in the number of samples. Moreover, since the l_∞-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 (τ-quantile) was formally defined in [5] and maximizes the number of relevant samples in the top τ-fraction of ranked samples. When the threshold equals the top τ-quantile of all scores, this problem falls into our framework. The early approaches aim at solving approximations, for example, [11] 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. Boyd *et al.* [5] presented an SVM-like formulation that 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. Kar *et al.* [12] derived upper approximations, their error bounds and solved these approximations. Grill and Pevný [9] proposed the projected gradient descent method where after each gradient step, the quantile is recomputed. Eban *et al.* [7] suggested new formulations for various criteria and argued that they keep desired properties such as convexity. Tasche [22] showed that accuracy at the top is maximized by thresholding the posterior probability of the relevant class. The closest approach to our framework is [14,15], where the authors considered multi-class classification problems, and their goal was to optimize the performance on the top few classes and [18], where the authors implicitly removed some variables and derived an efficient algorithm.

The paper is organized as follows. In Section 2, we introduce the unified framework and then show how the three applications above fall into it. To each application, we state several numerical formulations, some of them are known, and some are new. In Section 3, we perform a theoretical analysis of the unified framework. First, we focus on convexity, which ensures that no local minima are present. Second, we show which formulations from the previous section are differentiable. Finally, we present a simple example that shows the differences between these formulations and highlights the major problem that some of the

formulations may have a global minimum at $\boldsymbol{w} = 0$. Since the weights form the normal of the separating hyperplane, this solution does not provide any information. We further build on this example and analyse the case when zero weights are the global minimum. We show that the higher the threshold, the more likely a problem is to have the global minimum at zero. Based on this result, we compare the thresholds of individual formulations and show how stable they are. Section 4 argues that the threshold t depends on the weights $\boldsymbol{w}$ and therefore, we follow the implicit programming technique and remove the threshold constraint. We show how to compute derivatives for the reduced problem. We propose a more complex stochastic gradient descent method which is convergent. Section 5 presents the numerical results. We show both a good performance of our method and verify the theoretical results numerically. To keep the brevity of the paper, we postpone multiple results to the Appendix. The reader is welcome to refer to our codes online.[1]

2. Framework for minimizing misclassification 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 $\boldsymbol{x}$, we consider the linear classifier $f(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{x} - t$, where $\boldsymbol{w}$ 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 $z = \boldsymbol{w}^\top \boldsymbol{x}$. For example, *TopPush* from [16] sets the threshold t to the largest score z^- corresponding to negative samples and [9] sets it to the quantile of all scores.

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

$$\begin{aligned} \mathrm{tp}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x} \in \mathcal{X}^+} [\boldsymbol{w}^\top \boldsymbol{x} - t \geq 0], \quad \mathrm{fn}(\boldsymbol{w}, t) = \sum_{\boldsymbol{x} \in \mathcal{X}^+} [\boldsymbol{w}^\top \boldsymbol{x} - t < 0], \\ \mathrm{tn}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x} \in \mathcal{X}^-} [\boldsymbol{w}^\top \boldsymbol{x} - t < 0], \quad \mathrm{fp}(\boldsymbol{w}, t) = \sum_{\boldsymbol{x} \in \mathcal{X}^-} [\boldsymbol{w}^\top \boldsymbol{x} - t \geq 0]. \end{aligned} \tag{1}$$

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:

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

As the 0-1 loss in (1) is discontinuous, problem (2) is difficult to handle. The usual approach is to employ a surrogate function such as the hinge loss function defined by

$$l_{\text{hinge}}(z) = \max\{0, 1 + z\}. \tag{3}$$

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) may be approximated by their surrogate counterparts

$$\begin{aligned} \overline{\mathrm{tp}}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x}\in\mathcal{X}^+} l(\boldsymbol{w}^\top \boldsymbol{x} - t), \quad \overline{\mathrm{fn}}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x}\in\mathcal{X}^+} l(t - \boldsymbol{w}^\top \boldsymbol{x}), \\ \overline{\mathrm{tn}}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x}\in\mathcal{X}^-} l(t - \boldsymbol{w}^\top \boldsymbol{x}), \quad \overline{\mathrm{fp}}(\boldsymbol{w}, t) &= \sum_{\boldsymbol{x}\in\mathcal{X}^-} l(\boldsymbol{w}^\top \boldsymbol{x} - t). \end{aligned} \tag{4}$$

Since $l(\cdot) \geq [\cdot]$, the surrogate counts (4) provide upper approximations of the true counts (1). Replacing the counts in (2) by their surrogate counterparts and adding a regularization results in

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad \text{threshold } t \text{ is a function of } \{\boldsymbol{w}^\top \boldsymbol{x}_i\}_{i=1}^n. \end{aligned} \tag{5}$$

In the rest of this section, we list formulations that fall into the framework of (2) and (5).

2.1. Methods based on pushing positives to the top

The first category of formulations falling into our framework (2) and (5) are ranking methods which attempt to put as many positives (relevant samples) to the top as possible. Specifically, for each sample $\boldsymbol{x}$, they compute the score $z = \boldsymbol{w}^\top \boldsymbol{x}$ and then sort the vector $\boldsymbol{z}$ into $\boldsymbol{z}_{[\cdot]}$ with decreasing components $z_{[1]} \geq z_{[2]} \geq \cdots \geq z_{[n]}$. 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

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\,\mathrm{fn}\,(\boldsymbol{w}, t) \\ &\text{subject to} \quad t = z^-_{[1]}, \\ &\qquad\qquad \text{components of } \boldsymbol{z}^- \text{ equal to } z^- = \boldsymbol{w}^\top \boldsymbol{x}^- \text{ for } \boldsymbol{x}^- \in \mathcal{X}^-. \end{aligned} \tag{6}$$

As t is a function of the scores $z = \boldsymbol{w}^\top \boldsymbol{x}$, problem (6) is a special case of (2).

TopPush from [16] replaces the false-negatives in (6) by their surrogate and adds a regularization term to arrive at

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t = z^-_{[1]}, \\ &\qquad\qquad \text{components of } \boldsymbol{z}^- \text{ equal to } z^- = \boldsymbol{w}^\top \boldsymbol{x}^- \text{ for } \boldsymbol{x}^- \in \mathcal{X}^-. \end{aligned} \tag{7}$$

Note that this falls into the framework of (5).

As we will show in Section 3.4, *TopPush* is sensitive to outliers and mislabelled data. To robustify it, we follow the idea from [14] and propose to replace the largest negative score

by the mean of k largest negative scores. This results in

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\text{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t = \frac{1}{k}(z^-_{[1]} + \cdots + z^-_{[k]}), \\ &\qquad\qquad \text{components of } \boldsymbol{z}^- \text{ equal to } z^- = \boldsymbol{w}^\top \boldsymbol{x}^- \text{ for } \boldsymbol{x}^- \in \mathcal{X}^-. \end{aligned} \tag{8}$$

We used the mean of highest k negative scores instead of the value of the kth negative score to preserve convexity as shown in Section 3.2.

2.2. Accuracy at the top

The previous category considers formulations which minimize the false-negatives below the highest-ranked negative. Accuracy at the Top [5] takes a different approach and minimizes false positives above the top τ-quantile defined by

$$t_1(\boldsymbol{w}) = \max\{t \mid \text{tp}(\boldsymbol{w}, t) + \text{fp}(\boldsymbol{w}, t) \geq n\tau\}. \tag{9}$$

Then the Accuracy at the Top problem is defined by

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^-}\text{fp}(\boldsymbol{w}, t) \\ &\text{subject to} \quad t \text{ is the top } \tau\text{-quantile: it solves (9).} \end{aligned} \tag{10}$$

Due to Lemma A.1 in the Appendix, the previous problem (10) is equivalent (up to a small theoretical issue) to

$$\begin{aligned} &\text{minimize} \quad \mu\ \text{fn}(\boldsymbol{w}, t) + (1-\mu)\text{fp}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t \text{ is the top } \tau\text{-quantile: it solves (9)} \end{aligned} \tag{11}$$

for any $\mu \in [0, 1]$. This problem with $\mu = 0$ equals to (10), with $\mu = 1$ it falls into our framework (2), while with $\mu = \frac{n^-}{n}$ it corresponds to the original definition from [5].

Apart from the quantile (9), there are two other possible choices of the threshold

$$t_2(\boldsymbol{w}) = \frac{1}{n\tau}\sum_{i=1}^{n\tau} z_{[i]}, \tag{12}$$

$$t_3(\boldsymbol{w}) \quad \text{solves} \quad \frac{1}{n}\sum_{i=1}^{n} l(\beta(z_i - t)) = \tau. \tag{13}$$

We again use the vector of scores $\boldsymbol{z}$ with components $z_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ and for the rest of the paper, we assume, for simplicity, that $n\tau$ is an integer. The quantile (9) is sometimes denoted as VaR (value at risk) and (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 (13) as surrogate top τ-quantile. We will investigate the relations between these three objects as well as their properties such as convexity, differentiability or stability in Section 3.

Paper [9] builds on the Accuracy at the Top problem (11), where it replaces $\mathrm{fn}(\boldsymbol{w}, t)$ and $\mathrm{fp}(\boldsymbol{w}, t)$ in the objective by their surrogate counterparts $\overline{\mathrm{fn}}(\boldsymbol{w}, t)$ and $\overline{\mathrm{fp}}(\boldsymbol{w}, t)$. This leads to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{1}{n^-}\overline{\mathrm{fp}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & t \text{ is the top } \tau\text{-quantile: it solves (9).} \end{aligned} \tag{14}$$

Based on the first author, we name this formulation *Grill*. The main purpose of (12) is to provide a convex approximation of the nonconvex quantile (9). Putting it into the constraint results in a convex approximation problem, which we call *TopMean*

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & t = \frac{1}{n\tau}(z_{[1]} + \cdots + z_{[n\tau]}), \\ & \text{components of } \boldsymbol{z} \text{ equal to } z = \boldsymbol{w}^\top \boldsymbol{x} \text{ for } \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{15}$$

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

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & t \text{ is the surrogate top } \tau\text{-quantile: it solves (13).} \end{aligned} \tag{16}$$

Note that *Grill* minimizes the convex combination of false positives and false-negatives while (15) and (16) minimize only the false negatives. The reason for this will be evident in Section 3.2 and amounts to preservation of convexity. Moreover, as will see later, problem (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).

2.3. Methods optimizing the Neyman–Pearson criterion

Another category falling into the framework of (2) and (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 $\boldsymbol{x}$ 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 τ, we may write this as

$$t_1^{\mathrm{NP}}(\boldsymbol{w}) = \max\{t \mid \mathrm{fp}(\boldsymbol{w}, t) \geq n^- \tau\}. \tag{17}$$

Then, we may write the Neyman–Pearson problem as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n^+}\mathrm{fn}(\boldsymbol{w}, t) \\ \text{subject to} \quad & t \text{ is Type I error at level } \tau\text{: it solves (17).} \end{aligned} \tag{18}$$

Since (18) differs from (11) only by counting only the false positives in (17) instead of counting all positives in (9), we can derive its approximations in exactly the same way as in

Section 2.2. We therefore provide only their brief description and start with approximations of (17)

$$t_2^{\mathrm{NP}}(\boldsymbol{w}) = \frac{1}{n^-\tau}\sum_{i=1}^{n^-\tau} z^-_{[i]}, \tag{19}$$

$$t_3^{\mathrm{NP}}(\boldsymbol{w}) \quad \text{solves} \quad \frac{1}{n}\sum_{i=1}^{n^-} l(\beta(z_i^- - t)) = \tau. \tag{20}$$

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

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{1}{n^-}\overline{\mathrm{fp}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t \text{ is the Neyman–Pearson threshold: it solves (17).} \end{aligned} \tag{21}$$

Similarly, the Neyman–Pearson alternative to *TopMean* reads

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t = \frac{1}{n^-\tau}(z^-_{[1]} + \cdots + z^-_{[n^-\tau]}), \\ &\qquad\qquad \text{components of } \boldsymbol{z}^- \text{ equal to } z^- = \boldsymbol{w}^\top \boldsymbol{x}^- \text{ for } \boldsymbol{x}^- \in \mathcal{X}. \end{aligned} \tag{22}$$

This problem already appeared in [23] under the name τ*-FPL*. Finally, *Pat*&*Mat-NP* reads

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t \text{ is the surrogate Neyman–Pearson threshold: it solves (20).} \end{aligned} \tag{23}$$

We may see (22) from two different viewpoints. First, τ*-FPL* provide convex approximations of *Grill-NP*. Second, τ*-FPL* has the same form as *TopPushK*. The only difference is that for τ*-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 that 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.

3. Theoretical analysis of the framework

In this section, we provide a theoretical analysis of the unified framework from Section 2. We consider the problem *formulations* purely 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 [4].
- *Differentiability* increases the speed of convergence.
- *Stability* is a general term, by which we mean that the global minimum is not at $\boldsymbol{w} = \boldsymbol{0}$. This actually happens for many formulations from Section 2 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 2.2. The results for methods from Section 2.3 are identical. For the same reason, we postpone the proofs to Appendix 1.

3.1. Threshold value comparison

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

Proposition 3.1 ([23]): *We always have*

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

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

3.2. 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 a faster convergence rates. We present the following two results.

Proposition 3.2: *Thresholds t_2 and t_3 are convex functions of the weights $\boldsymbol{w}$. The threshold function t_1 is nonconvex.*

Theorem 3.3: *If the threshold t is a convex function of the weights $\boldsymbol{w}$, then function $f(\boldsymbol{w}) = \overline{fn}(\boldsymbol{w}, t(\boldsymbol{w}))$ is convex.*

While the proof of Theorem 3.3 is simple, it points to the necessity of considering only false negatives in the objective of the problems in Section 2. In such a case, *TopPush*, *TopPushK*, *TopMean*, *τ-FPL*, *Pat&Mat* and *Pat&Mat-NP* are convex problems. At the same time, *Grill* and *Grill-NP* are not convex problems.

3.3. 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 3.4: *If the surrogate function l is differentiable, then threshold t_3 is a differentiable function of the weights $\boldsymbol{w}$ and its derivative equals to*

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

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 4.

3.4. Stability

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

3.4.1. Example of a degenerate behaviour

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(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 $\boldsymbol{w} = (1,0)$.

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

It can be shown that $\boldsymbol{w}_1 = (0,0)$ is even the global minimum for *TopPush* and *TopMean*. 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 as shown in Figure 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.

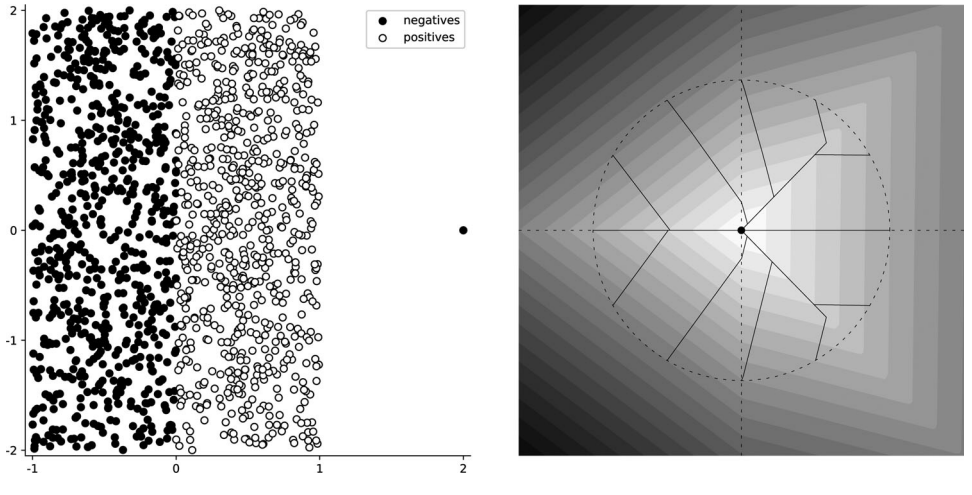

Figure 1. Left: distribution of positive (empty circle) and negative samples (full circles) for the example from Section 3.4.1. Right: contour plot for *TopPush* and its convergence to the zero vector from 12 initial points.

Table 1. Comparison of formulations on the very simple problem from Section 3.4.1.

		$\boldsymbol{w}_1 = (0,0)$		$\boldsymbol{w}_2 = (1,0)$	
Name	Label	t	f	t	f
TopPush	(7)	0	1	2	2.5
TopPushK	(8)	0	1	$\frac{2}{k}$	$0.5 + \frac{2}{k}$
Grill	(14)	0	2	$1 - 2\tau$	$1.5 + 2\tau(1-\tau)$
TopMean	(15)	0	1	$1 - \tau$	$1.5 - \tau$
Pat&Mat	(16)	$\frac{1}{\beta}(1-\tau)$	$1 + \frac{1}{\beta}(1-\tau)$	$\frac{1}{\beta}(1-\tau)$	$0.5 + \frac{1}{\beta}(1-\tau)$

Notes: Two formulations have the global minimum (denoted by grey colour) at $\boldsymbol{w}_1 = (0,0)$ which does not generate any separating hyperplane. The optimal separating hyperplane is generated by $\boldsymbol{w}_2 = (1,0)$.

3.4.2. 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 $\boldsymbol{w} = \boldsymbol{0}$. This is highly undesirable since $\boldsymbol{w}$ is the normal vector to the separating hyperplane and the zero vector provides no information. In this section, we analyse when this situation happens. The first result states that if the threshold $t(\boldsymbol{w})$ is above a certain value, then zero has a better objective that $\boldsymbol{w}$. If this happens for all $\boldsymbol{w}$, then zero is the global minimum.

Theorem 3.5: *Consider any of these formulations:* TopPush, TopPushK, TopMean *or* τ-FPL. *Fix any* $\boldsymbol{w}$ *and denote the corresponding threshold* $t(\boldsymbol{w})$. *If we have*

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

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

$$\begin{aligned}
z^-_{[1]} \geq \frac{1}{n^+} \sum_{i=1}^{n^+} z^+_i \implies f(\boldsymbol{0}) \leq f(\boldsymbol{w}) \quad \textit{for} \text{ TopPush}, \\
\frac{1}{k} \sum_{i=1}^{k} z^-_{[i]} \geq \frac{1}{n^+} \sum_{i=1}^{n^+} z^+_i \implies f(\boldsymbol{0}) \leq f(\boldsymbol{w}) \quad \textit{for} \text{ TopPushK}, \\
\frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} z^-_{[i]} \geq \frac{1}{n^+} \sum_{i=1}^{n^+} z^+_i \implies f(\boldsymbol{0}) \leq f(\boldsymbol{w}) \quad \textit{for } \tau\text{–FPL}.
\end{aligned} \tag{24}$$

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

Corollary 3.6: *Consider the TopPush formulation. If the positive samples lie in the convex hull of negative samples, then* $\boldsymbol{w} = \boldsymbol{0}$ *is the global minimum.*

Corollary 3.7: *Consider the TopMean formulation. If* $n^+ \geq n\tau$*, then* $\boldsymbol{w} = \boldsymbol{0}$ *is the global minimum.*

The proof of Theorem 3.5 employs the fact that all formulations in the theorem statement have only false negatives in the objective. If $\boldsymbol{w}_0 = \boldsymbol{0}$, then $\boldsymbol{w}_0^\top \boldsymbol{x} = 0$ for all samples $\boldsymbol{x}$, the threshold equals to $t = 0$ and the objective equals to one. If the threshold is large for $\boldsymbol{w}$, many positives are below the threshold, and the false negatives have the average surrogate value larger than one. In such a case, $\boldsymbol{w}_0 = \boldsymbol{0}$ 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 $\boldsymbol{w}^\top \boldsymbol{x}$ 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 3.8: *Consider the Pat&Mat or Pat&Mat-NP formulation with the hinge surrogate and no regularization. Assume that for some* $\boldsymbol{w}$*, we have*

$$\frac{1}{n^+} \sum_{\boldsymbol{x}^+ \in \mathcal{X}^+} \boldsymbol{w}^\top \boldsymbol{x}^+ > \frac{1}{n^-} \sum_{\boldsymbol{x}^- \in \mathcal{X}^-} \boldsymbol{w}^\top \boldsymbol{x}^-. \tag{25}$$

Then there is a scaling parameter β_0 *from* (13) *such that* $f(\boldsymbol{w}) < f(\boldsymbol{0})$ *for all* $\beta \in (0, \beta_0)$*.*

These theorem shed some light on the behaviour of the formulations. Theorem 3.5 states that the stability of τ*-FPL* requires

$$\frac{1}{n^- \tau} \sum_{i=1}^{n^- \tau} z_{[i]}^- < \frac{1}{n^+} \sum_{i=1}^{n^+} z_i^+, \tag{26}$$

while Theorem 3.8 states that the stability of *Pat&Mat-NP* is ensured by

$$\frac{1}{n^-} \sum_{i=1}^{n^-} z_{[i]}^- < \frac{1}{n^+} \sum_{i=1}^{n^+} z_i^+. \tag{27}$$

The right-hand sides of (26) and (27) are the same, while the left-hand side of (27) is always smaller than the left-hand side of (26). This implies that if τ*-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 z may be interpreted as performance. Then formula (26) 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 τ*-FPL* fails. On the other hand, formula (27) 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.

3.5. Method comparison

We provide a summary of the obtained results in Table 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 $\boldsymbol{w} = \mathbf{0}$ being the global minimum.

A similar comparison is performed in Figure 2. Methods in green and grey are convex, while formulations in white are non-convex. Based on Theorem 3.5, four formulations in grey are vulnerable to have the global minimum at $\boldsymbol{w} = 0$. 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 usually holds but not always, the precise formulation is provided in Appendix A.5. This complies with Corollaries 3.6 and 3.7 which state that *TopPush* and *TopMean* are most vulnerable. At the same time, it says that τ*-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 τ*-FPL* due to Theorem 3.5, it is more stable due to the discussion after Theorem 3.8.

Table 2. Summary of the formulations from Section 2.

Name	Source	Definition	Hyperpars	Convex	Differentiable	Stable
TopPush	[16]	(7)	λ	✓	✗	✗
TopPushK	ours	(8)	λ, k	✓	✗	✗
Grill	[9]	(14)	λ	✗	✗	✓
Pat&Mat	ours	(16)	β, λ	✓	✓	✓
TopMean	—	(15)	λ	✓	✗	✗
Grill-NP	—	(21)	λ	✗	✗	✓
Pat&Mat-NP	ours	(23)	β, λ	✓	✓	✓
τ-FPL	[23]	(22)	λ	✓	✗	✗

Notes: 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 $\boldsymbol{w} = \mathbf{0}$).

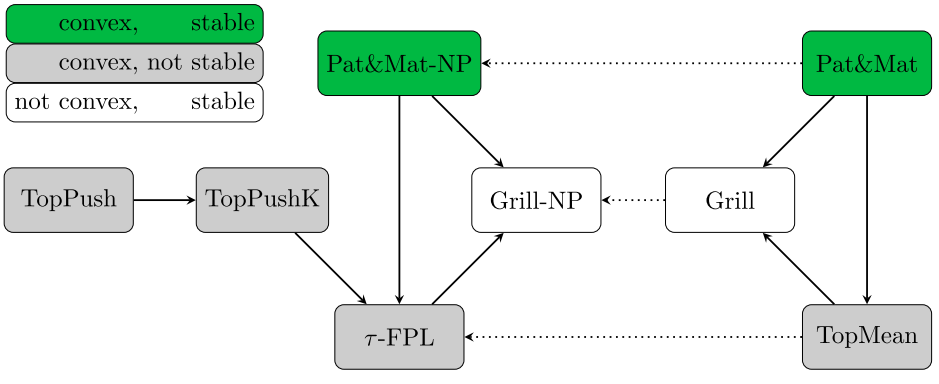

Figure 2. Summary of the formulations from Section 2. Methods in green and grey are convex, while formulations in white are non-convex. Methods in grey are vulnerable to have the global minimum at $\boldsymbol{w} = 0$. Full (dotted) arrow pointing from one formulation to another show that the latter formulation has always (usually) smaller threshold (Colour online).

4. Convergence of stochastic gradient descent

The previous section analysed the formulations from Section 2 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.

4.1. Stochastic gradient descent: basic

Many optimization algorithms for solving the formulations from Section 2 use primal-dual or purely dual formulations. [7] introduced dual variables and used alternating optimization to the resulting min-max problem. Li *et al.* [16] and Zhang *et al.* [23] dualized the problem and solved it with the steepest gradient ascent. Mácha *et al.* [17] followed the same path but added kernels to handle non-linearity. We follow the ideas of [18] and [1] and solve the problems directly in their primal formulations. Therefore, even though we use the same formulation for *TopPush* as [16] or for τ*-FPL* as [23], our solution process is different. However, due to convexity, both algorithms should converge to the same point.

The decision variables in (5) are the normal vector of the separating hyperplane $\boldsymbol{w}$ and the threshold t. To apply an efficient optimization method, we need to compute gradients. The simplest idea [9] is to compute the gradient only with respect to $\boldsymbol{w}$ and then recompute t. A more sophisticated way is based on the chain rule. For each $\boldsymbol{w}$, the threshold t can be computed uniquely. We stress this dependence by writing $t(\boldsymbol{w})$ instead of t. By doing so, we effectively remove the threshold t from the decision variables and $\boldsymbol{w}$ remains the only decision variable. Note that the convexity is preserved. Then we can compute the derivative via the chain rule

$$\begin{aligned} f(\boldsymbol{w}) &= \frac{1}{n^+} \sum_{\boldsymbol{x}\in\mathcal{X}^+} l(t(\boldsymbol{w}) - \boldsymbol{w}^\top \boldsymbol{x}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2, \\ \nabla f(\boldsymbol{w}) &= \frac{1}{n^+} \sum_{\boldsymbol{x}\in\mathcal{X}^+} l'(t(\boldsymbol{w}) - \boldsymbol{w}^\top \boldsymbol{x})(\nabla t(\boldsymbol{w}) - \boldsymbol{x}) + \lambda \boldsymbol{w}. \end{aligned} \tag{28}$$

The only remaining part is the computation of $\nabla t(\boldsymbol{w})$. It is simple for $\nabla t_1(\boldsymbol{w})$ and $\nabla t_2(\boldsymbol{w})$ and Theorem 3.4 shows the computation for $\nabla t_3(\boldsymbol{w})$. Appendix 3 provides an efficient computation method for $t_3(\boldsymbol{w})$.

Having derivative (28), deriving the stochastic gradient is simple. It partitions the dataset into minibatches and provides an update of the weights $\boldsymbol{w}$ based only on a minibatch, namely by replacing the mean over the whole dataset in (28) by a mean over the minibatch.

4.2. Stochastic gradient descent: convergent for Pat&Mat and Pat&Mat-NP

For the convergence proof, we need differentiability which is due to Theorem 3.4 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 [20].

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

$$z_i^k = \begin{cases} \boldsymbol{x}_i^\top \boldsymbol{w}^k & \text{for all } i \in I^k, \\ z_i^{k-1} & \text{for all } i \notin I^k. \end{cases} \tag{29}$$

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

$$\sum_{i \in X} l(\beta(z_i^k - t^k)) = n\tau. \tag{30}$$

This is an approximation of the surrogate quantile $t(\boldsymbol{w}^k)$ from (13). The only difference from the true value $t(\boldsymbol{w}^k)$ is that we use delayed scores. Then we introduce artificial variable

$$\boldsymbol{a}^k = \sum_{i \in I^k} l'(\beta(z_i^k - t^k))\boldsymbol{x}_i. \tag{31}$$

Finally, we approximate the derivative $\nabla f(\boldsymbol{w}^k)$ from (28) by

$$g(\boldsymbol{w}^k) = \frac{1}{n_+^k} \sum_{i \in I_+^k} l'(t^k - z_i^k)(\nabla t^k - \boldsymbol{x}_i), \tag{32}$$

where ∇t^k is an approximation of $\nabla t(\boldsymbol{w}^k)$ from Theorem 3.4 defined by

$$\nabla t^k = \frac{\boldsymbol{a}^k + \boldsymbol{a}^{k-1} + \cdots + \boldsymbol{a}^{k-s+1}}{\sum_{i \in X} l'(\beta(z_i^k - t^k))}. \tag{33}$$

A perhaps more straightforward possibility would be to consider only $\boldsymbol{a}^k$ in the numerator of (33). However, choice (33) 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 4.1. Note that a proper initialization for the first s iterations is needed. We finish the theoretical part by the convergence proof.

Theorem 4.1: *Consider the Pat&Mat or Pat&Mat-NP formulation, stepsizes* $\alpha^k = \frac{\alpha^0}{k+1}$ *and piecewise disjoint minibatches* $I^1, \ldots, I^s$ *which cycle periodically* $I^{k+s} = I^k$. *If* l *is the smoothened (Huberized) hinge function, then Algorithm* 4.1 *converges to the global minimum of* (16).

5. Numerical experiments

In this section, we present numerical results.

Algorithm 4.1 Stochastic gradient descent for maximizing accuracy at the top

Input: Dataset X, Minibatches $I^1, \ldots, I^s$, Stepsize α^k

Initialize weights $\boldsymbol{w}^0$
for $k = 0, 1, \ldots$ **do**
  Select a minibatch I^k
  Compute z_i^k for all $i \in I^k$ according to (29)
  Compute t^k according to (30)
  Compute $\boldsymbol{a}^k$ according to (31)
  Compute ∇t^k according to (33)
  Compute $g(\boldsymbol{w}^k)$ according to (32)
  Set $\boldsymbol{w}^{k+1} \leftarrow \boldsymbol{w}^k - \alpha^k g(\boldsymbol{w}^k)$
end for

5.1. Implementational details and hyperparameter choice

We recall that all methods fall into the framework of either (2) or (5). Since the threshold t depends on the weights $\boldsymbol{w}$, we can consider the decision variable to be only $\boldsymbol{w}$. Then to apply a method, we implemented the following iterative procedure. At iteration j, we have the weights $\boldsymbol{w}^j$ to which we compute the threshold $t^j = t(\boldsymbol{w}^j)$. Then according to (28), we compute the gradient of the objective and apply the ADAM descent scheme [13]. All methods were run for 10000 iterations using the stochastic gradient descent. The minibatch size was 512 except for the Ionosphere and Spambase datasets where the full gradient was used. All methods used the hinge surrogate (3). The initial point is generated randomly.

We run the methods for the following hyperparameters:

$$\begin{aligned} &\beta \in \{0.0001,\ 0.001,\ 0.01,\ 0.1,\ 1,\ 10\}, \\ &\lambda \in \{0,\ 0.00001,\ 0.0001,\ 0.001,\ 0.01,\ 0.1\}, \\ &k \in \{1, 3, 5, 10, 15, 20\}. \end{aligned} \tag{34}$$

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 τ*-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.

5.2. Dataset description and performance criteria

For the numerical results, we considered ten datasets summarized in Table 3. They can be downloaded from the UCI repository. Ionosphere [21] and Spambase are small, Hepmass [3] contains a large number of samples while Gisette [10] contains a large number of features. We also considered six visual recognition datasets: MNIST, FashionMNIST, CIFAR10, CIFAR20, CIFAR100 and SVHN2. MNIST and FashionMNIST are greyscale datasets of digits and fashion items, respectively. CIFAR100 is a dataset of coloured images

Table 3. Structure of the used datasets.

		Training		Validation		Testing	
	m	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	5,250,000	50.0%	1,750,000	50.0%	3,500,000	50.0%
MNIST	$28 \times 28 \times 1$	44,999	11.2%	15,001	11.2%	10,000	11.4%
FashionMNIST	$28 \times 28 \times 1$	45,000	10.0%	15,000	10.0%	10,000	10.0%
CIFAR10	$32 \times 32 \times 3$	37,500	10.0%	12,500	10.0%	10,000	10.0%
CIFAR20	$32 \times 32 \times 3$	37,500	5.0%	12,500	5.0%	10,000	5.0%
CIFAR100	$32 \times 32 \times 3$	37,500	1.0%	12,500	1.0%	10,000	1.0%
SVHN2	$32 \times 32 \times 3$	54,944	18.9%	18,313	18.9%	26,032	19.6%

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

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 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 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 $\boldsymbol{w}^\top \boldsymbol{x} - t$, it selects the threshold t so that the desired true-negative rate is satisfied and then computes the true-positive rate for this threshold.

5.3. Numerical results

Figure 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 the two more complicated datasets. We selected four representative methods: *Pat&Mat* and *Pat&Mat-NP* as our methods and *TopPush* and τ*-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.

While Figure 3 gave a glimpse of the behaviour of methods, Figures 4 and 5 provide a statistically more sound comparison. It employs the Nemenyi post hoc test for the Friedman test recommended in [6]. This test compares if the mean ranks of multiple methods are significantly different.

We consider 14 methods (we count different values of τ as different methods) as depicted in this table. For each dataset mentioned in Section 5.2 and each method, we evaluated the 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

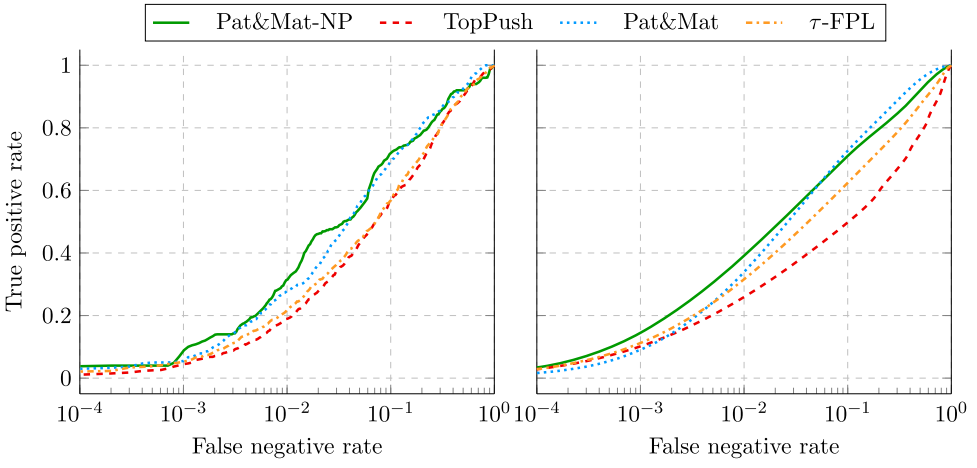

Figure 3. ROC curves (with logarithmic *x* axis) on CIFAR100 (left) and Hepmass (right).

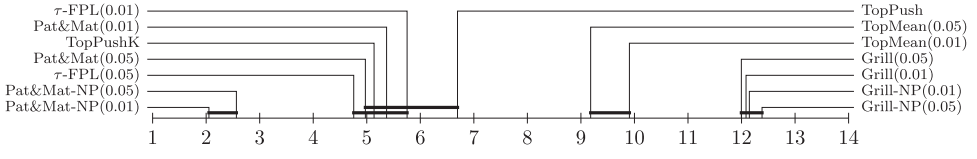

Figure 4. Critical difference (CD) diagrams (level of importance 0.05) of the Nemenyi post hoc test for the Friedman test. Each diagram shows the mean rank of each method, with rank 1 being the best. Black wide horizontal lines group together methods with the mean ranks that are not significantly different. The critical difference diagrams were computed for mean rank averages over all datasets of the tpr@fpr ($\tau = 0.01$) metric.

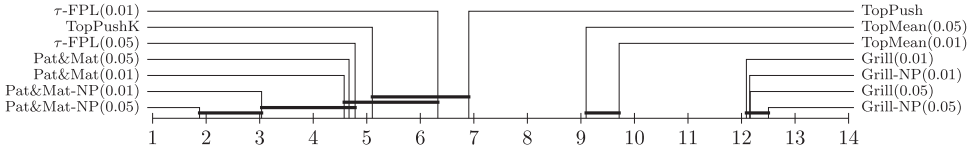

Figure 5. Critical difference (CD) diagrams (level of importance 0.05) of the Nemenyi post hoc test for the Friedman test. Each diagram shows the mean rank of each method, with rank 1 being the best. Black wide horizontal lines group together methods with the mean ranks that are not significantly different. The critical difference diagrams were computed for mean rank averages over all datasets of the tpr@fpr ($\tau = 0.05$) metric.

their critical difference, their performance is not deemed to be significantly different. Black wide horizontal lines group such methods.

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 4 and 5.

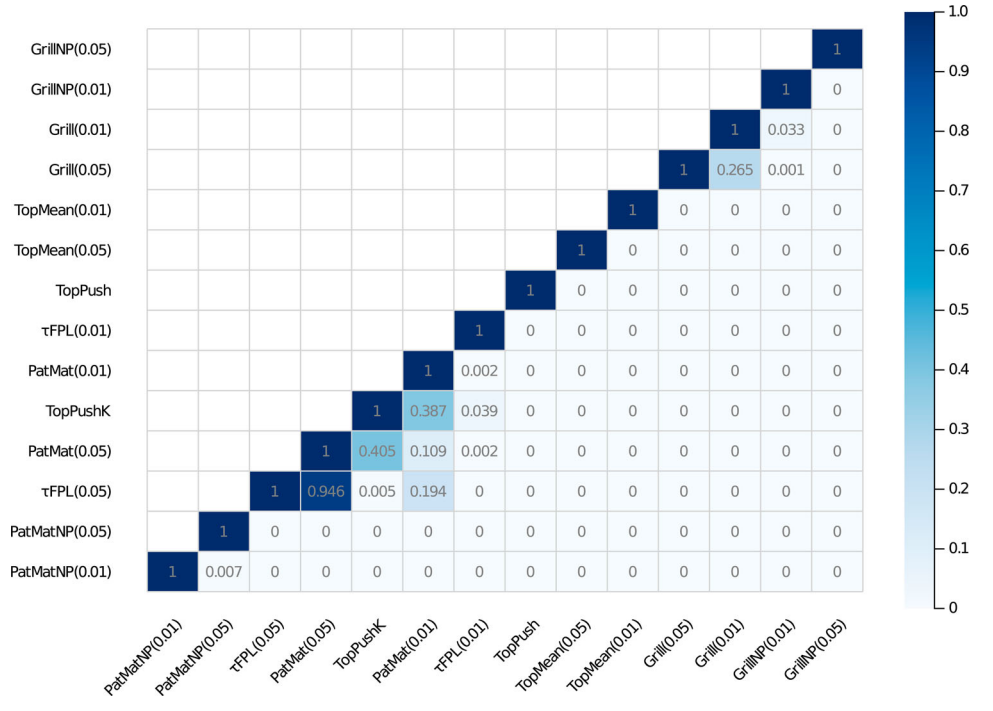

Figure 6. The *p*-value for the pairwise Wilcoxon signed-rank test, where the null hypothesis is that the mean tpr@fpr(0.01) of both methods is the same. The methods are sorted by mean rank (left = better).

- 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 3.3 and the discussion after that.
- *TopMean* (ranks 9.2 and 9.9) does not perform well either. Since the thresholds τ are small, then $\boldsymbol{w} = 0$ is the global minimum as proved in Corollary 3.7.
- *Pat&Mat-NP* (rank 2.1 and 2.6) seems to outperform other methods.
- *Pat&Mat* (ranks 5.0 and 5.4), τ*-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 5) outperforms *Pat&Mat-NP* at level 0.01 (rank 3.0 in Figure 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 5 but not for Figure 4.

Figure 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 tpr@fpr of both methods is the same. Even though Figure 4 employs a comparison of mean ranks and Figure 6 a pairwise comparison of 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 4. Necessary hyperparameter choice for the solution to have a better objective than zero.

	Ionosphere	Hepmass	FashionMNIST	CIFAR100
TopPush	✓	✗	✓	✗
TopPushK	✓	✗	✓	✗
Grill$\tau = 0.01$	✗	✗	✗	✗
$\tau = 0.0$	✗	✗	✗	✗
Pat&Mat$\tau = 0.01$	✓	$\beta \leq 0.1$	$\beta \leq 1$	$\beta \leq 1$
$\tau = 0.05$	✓	$\beta \leq 1$	✓	✓
TopMean$\tau = 0.01$	✗	✗	✗	✗
$\tau = 0.05$	✗	✗	✗	✗
Grill-NP$\tau = 0.01$	✗	✗	✗	✗
$\tau = 0.05$	✗	✗	✗	✗
Pat&Mat-NP$\tau = 0.01$	✓	$\beta \leq 1$	✓	$\beta \leq 1$
$\tau = 0.05$	✓	✓	✓	$\beta \leq 1$
τ-FPL$\tau = 0.01$	✓	✗	✓	✗
$\tau = 0.05$	✓	✓	✓	$\lambda \leq 0.001$

Notes: ✓ means that the solution was better than zero for all hyperparameters while ✗ means that it was worse for all hyperparameters.

Table 4 investigates the impact of $\boldsymbol{w} = 0$ 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 $\boldsymbol{w} = 0$. Thus, ✓ means that it is always better while ✗ means that the algorithm made no progress from the starting point $\boldsymbol{w} = 0$. The latter case implies that $\boldsymbol{w} = 0$ seems to be the global minimum. We make the following observations:

- *Pat&Mat* and *Pat&Mat-NP* are the only methods that succeeded at every dataset for some hyperparameter. Moreover, for each dataset, there was some β_0 such that these methods were successful if and only if $\beta \in (0, \beta_0)$. This is in agreement with Theorem 3.8.
- *TopMean* fails everywhere which agrees with Corollary 3.7.
- Figure 2 states that the methods from Section 2.2 has a higher threshold than their Neyman–Pearson variants from Section 2.3. This is documented in the table as the latter have a higher number of successes.

6. Conclusion

In this paper, we achieved the following results:

- We presented a unified framework for the three criteria from Section 2. These criteria include ranking, accuracy at the top and hypothesis testing.
- We showed that several known methods (*TopPush*, *Grill*, τ*-FPL*) fall into our framework and derived some completely new methods (*Pat&Mat*, *Pat&Mat-NP*).
- We performed a theoretical analysis of the methods. We showed that known methods suffer from certain disadvantages. While *TopPush* and τ*-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*.

Note

1. https://github.com/VaclavMacha/AccuracyAtTopPrimal_experiments.jl

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Grant Agency of the Czech Republic (18-21409S) and by the OP RDE (OP VVV) funded project Research Center for Informatics, reg. No: CZ02.1.01/0.0./0.0./16_019/0000765.

Notes on contributors

L. Adam received the MS and PhD degrees in mathematical optimization from the Charles University, Prague, Czech Republic, in 2011 and 2015, respectively. He is currently an Assistant Professor with the Artificial Intelligence Center, Faculty of Electrical Engineering at the Czech Technical University in Prague. He has published more than 25 papers in international journals and conference proceedings. His research interests include mathematical optimization, machine learning and loggerhead sea turtles.

V. Mácha received a master's degree in mathematical engineering from the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, in 2017. In the same year, he started PhD studies at the same faculty. He works as a junior researcher at the Artificial Intelligence Center, Czech Technical University in Prague. His research interests include classification problems, especially classification with low false alarm rates.

V. Šmídl received the PhD degree in electrical engineering from Trinity College Dublin, Dublin, Ireland, in 2004. Since 2004, he has been a Researcher with the Institute of Information Theory and Automation, Prague, Czech Republic. In October 2010, he joined the Regional Innovation Centre for Electrical Engineering, RICE. He published one research monograph, 30 journal papers, and over 100 conference papers. His research interests include advanced estimation and control techniques and their applications.

T. Pevný received the master's degree in computer science from the School of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, in 2003, and the PhD degree in computer science from the State University of New York, Binghamton, in 2008. From 2008 to 2009, he held a post-doctoral position at Gipsa Laboratory, Grenoble, France. He is with the Artificial Intelligence Center, Czech Technical University in Prague. His research interests include applications of non-parametric statistics (machine learning, density modeling) with a focus on computer security, steganography, and steganalysis.

ORCID

L. Adam http://orcid.org/0000-0001-8748-4308
V. Šmídl http://orcid.org/0000-0003-3027-6174

References

[1] L. Adam and M. Branda, *Machine learning approach to chance-constrained problems: An algorithm based on the stochastic gradient descent*, preprint (2019). Available at arXiv:1905.10986.

[2] S. 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*, SIAM, Philadelphia, PA, 2011, pp. 839–850.
[3] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, *Parameterized neural networks for high-energy physics*, Eur. Phys. J. C 76(5) (2016), pp. 235.
[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, 2004.
[5] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic, *Accuracy at the top*, In *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, 2012, pp. 953–961.
[6] J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, J. Mach. Learn. Res. 7 (2006), pp. 1–30.
[7] E. Eban, M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan, Scalable learning of non-decomposable objectives, In *Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2017, pp. 832–840.
[8] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer, *An efficient boosting algorithm for combining preferences*, J. Mach. Learn. Res. 4 (2003), pp. 933–969.
[9] M. Grill and T. Pevný, *Learning combination of anomaly detectors for security domain*, Comput. Netw. 107 (2016), pp. 55–63.
[10] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, *Result analysis of the NIPS 2003 feature selection challenge*, In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2005, pp. 545–555.
[11] T. Joachims, *A support vector method for multivariate performance measures*, In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, New York, NY, USA, ACM, 2005, pp. 377–384.
[12] P. Kar, H. Narasimhan, and P. Jain, *Surrogate functions for maximizing precision at the top*, In *International Conference on Machine Learning*, Lille, France, 2015, pp. 189–198.
[13] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, preprint (2014). Available at arXiv:1412.6980.
[14] M. Lapin, M. Hein, and B. Schiele, *Top-k multiclass SVM*, In *Advances in Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 325–333.
[15] M. Lapin, M. Hein, and B. Schiele, *Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification*, IEEE. Trans. Pattern Anal. Mach. Intell. 40(7) (2018), pp. 1533–1554.
[16] N. Li, R. Jin, and Z.-H. Zhou, *Top rank optimization in linear time*, In *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 1*, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 1502–1510.
[17] V. Mácha, L. Adam, and V. Šmídl, *Nonlinear classifiers for ranking problems based on kernelized SVM*, preprint (2020). Available at arXiv:2002.11436.
[18] A. Mackey, X. Luo, and E. Eban, *Constrained classification and ranking via quantiles*, preprint (2018). Available at arXiv:1803.00067.
[19] C. Rudin, *The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list*, J. Mach. Learn. Res. 10 (2009), pp. 2233–2271.
[20] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Math. Program. 162(1–2) (2017), pp. 83–112.
[21] V.G. Sigillito, S.P. Wing, L.V. Hutton, and K.B. Baker, *Classification of radar returns from the ionosphere using neural networks*, Johns Hopkins APL Tech. Dig. 10(3) (1989), pp. 262–266.
[22] D. Tasche, *A plug-in approach to maximising precision at the top and recall at the top*, preprint (2018). Available at arXiv:1804.03077.
[23] A. Zhang, N. Li, J. Pu, J. Wang, J. Yan, and H. Zha, *tau-FPL: Tolerance-constrained learning in linear time*, preprint (2018). Available at arXiv:1801.04701.

Appendices

Appendix 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 Equivalence of (10) and (11)

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

Lemma A.1: *Denote by t the exact quantile from (9). Then for all $\mu \in [0,1]$, we have*

$$\mathrm{fp}(\boldsymbol{w},t) = \mu\mathrm{fp}(\boldsymbol{w},t) + (1-\mu)\mathrm{fn}(\boldsymbol{w},t) + (1-\mu)(n\tau - n^+) + (1-\mu)(q-1), \tag{A1}$$

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

Proof: By the definition of the quantile, we have

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

This implies

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

From this relation, we deduce

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

which is precisely the lemma statement. ■

The right-hand side of (A1) consists of three parts. The first one is a convex combination of false positives and false negatives. The second one is a constant term that 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 (10) and (11). Note that the term q is ignored in many papers.

A.2 Results related to convexity

Proposition 3.2: *Thresholds t_2 and t_3 are convex functions of the weights. The threshold function t_1 is non-convex.*

Proof: It is easy to show that the quantile t_1 is not convex. Due to [14], 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 analyse t_3. It is defined via an implicit equation, where we consider for simplicity $\beta = 1$,

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

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

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

$$g(\lambda\boldsymbol{w} + (1-\lambda)\hat{\boldsymbol{w}},\ \lambda t + (1-\lambda)\hat{t}) \leq \lambda g(\boldsymbol{w}, t) + (1-\lambda) g(\hat{\boldsymbol{w}}, \hat{t}) = 0, \tag{A2}$$

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

$$g(\lambda\boldsymbol{w} + (1-\lambda)\hat{\boldsymbol{w}}, t_3(\lambda\boldsymbol{w} + (1-\lambda)\hat{\boldsymbol{w}})) = 0. \tag{A3}$$

Since g is nonincreasing in the second variable, from (A2) and (A3), we deduce

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

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

Theorem 3.3: *If the threshold t is a convex function of the weights* $\boldsymbol{w}$*, then function* $f(\boldsymbol{w}) = \overline{fn}(\boldsymbol{w}, t(\boldsymbol{w}))$ *is convex.*

Proof: Due to the definition of the surrogate counts (4), the objective of (5) equals to

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

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

A.3 Results related to differentiability

Theorem 3.4: *If the surrogate function l is differentiable, then threshold* t_3 *is a differentiable function of the weights* $\boldsymbol{w}$ *and its derivative equals to*

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

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

Proof: The result for t_3 follows directly from the implicit function theorem. The nondifferentiability of t_1 and t_2 happens whenever the threshold value is achieved at two different scores. ■

A.4 Results related to stability

Theorem 3.5: *Consider any of these formulations:* TopPush, TopPushK, TopMean *or* τ*-FPL. Fix any* $\boldsymbol{w}$ *and denote the corresponding threshold* $t(\boldsymbol{w})$*. If we have*

$$t(\boldsymbol{w}) \geq \frac{1}{n^+} \sum_{\boldsymbol{x}^+ \in \mathcal{X}^+} \boldsymbol{w}^\top \boldsymbol{x}^+, \tag{A4}$$

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

$$z^-_{[1]} \geq \frac{1}{n^+} \sum_{i=1}^{n^+} z^+_i \implies f(\mathbf{0}) \leq f(\boldsymbol{w}) \quad \text{for TopPush,}$$

$$\frac{1}{k} \sum_{i=1}^{k} z^-_{[i]} \geq \frac{1}{n^+} \sum_{i=1}^{n^+} z^+_i \implies f(\mathbf{0}) \leq f(\boldsymbol{w}) \quad \text{for TopPushK,}$$

$$\frac{1}{n^-\tau}\sum_{i=1}^{n^-\tau} z^-_{[i]} \geq \frac{1}{n^+}\sum_{i=1}^{n^+} z^+_i \implies f(\mathbf{0}) \leq f(\boldsymbol{w}) \quad \text{for } \tau\text{−FPL}.$$

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

$$f(\boldsymbol{w}) \geq \frac{1}{n^+}\overline{\mathrm{fn}}(\boldsymbol{w}, t) = \frac{1}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+} l(t - \boldsymbol{w}^\top\boldsymbol{x}) \geq \frac{1}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}(1 + c(t - \boldsymbol{w}^\top\boldsymbol{x}))$$
$$= 1 + \frac{c}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}(t - \boldsymbol{w}^\top\boldsymbol{x}) = 1 + ct - \frac{c}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x} \geq 1,$$

where the last inequality follows from (A4). Now, we realize that for any formulation from the statement, the corresponding threshold for $\boldsymbol{w} = 0$ equals to $t = 0$, and thus $f(\mathbf{0}) = 1$. But then $f(\mathbf{0}) \leq f(\boldsymbol{w})$. The second part of the result follows from the form of thresholds $t(\boldsymbol{w})$. ■

Theorem 3.8: *Consider the Pat&Mat or Pat&Mat-NP formulation with the hinge surrogate and no regularization. Assume that for some* $\boldsymbol{w}$*, we have*

$$\frac{1}{n^+}\sum_{\boldsymbol{x}^+\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x}^+ > \frac{1}{n^-}\sum_{\boldsymbol{x}^-\in\mathcal{X}^-}\boldsymbol{w}^\top\boldsymbol{x}^-. \tag{A5}$$

Then there is a scaling parameter β_0 *from* (13) *such that* $f(\boldsymbol{w}) < f(\mathbf{0})$ *for all* $\beta \in (0, \beta_0)$.

Proof: Define first

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

Then we have the following chain of relations:

$$\bar{z} = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}}\boldsymbol{w}^\top\boldsymbol{x} = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x} + \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}^-}\boldsymbol{w}^\top\boldsymbol{x} < \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x} + \frac{n^-}{nn^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x}$$
$$= \frac{1}{n}\left(1 + \frac{n^-}{n^+}\right)\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x} = \frac{1}{n}\frac{n^+ + n^-}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x} = \frac{1}{n^+}\sum_{\boldsymbol{x}\in\mathcal{X}^+}\boldsymbol{w}^\top\boldsymbol{x}. \tag{A6}$$

The only inequality follows from (A5) and the last equality follows from $n^+ + n^- = n$.

Due to (A5) we have $z_{\min} < \bar{z} < z_{\max}$. Then we can define

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

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

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

Then we obtain

$$1 + \beta(\boldsymbol{w}^\top\boldsymbol{x} - t) \geq 1 + \beta(z_{\min} - t) = 1 + \beta z_{\min} - 1 + \tau - \beta\bar{z} = \beta(z_{\min} - \bar{z}) + \tau \geq 0. \tag{A7}$$

Here, the first equality follows from the definition of t and the last inequality from the definition of β_0. Moreover, we have

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

$$= 1 - \beta t + \frac{\beta}{n} \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{w}^\top \boldsymbol{x} = 1 - \beta t + \beta \bar{z} = \tau,$$

where the second equality employs (A7), the third one the definition of $\bar{z}$ and the last one the definition of t. But this means that t is the threshold corresponding to $\boldsymbol{w}$.

Similarly to (A7), we get

$$1 + t - \boldsymbol{w}^\top \boldsymbol{x} \geq 1 + t - z_{\max} = 1 + \frac{1-\tau}{\beta} + \bar{z} - z_{\max} \geq \frac{1-\tau}{\beta} + \bar{z} - z_{\max} \geq 0, \tag{A8}$$

where the last inequality follows from the definition of β_0. Then for the objective, we have

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

where we the third equality follows from (A8), the only inequality from (A6) and the last equality from Appendix A2. 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 $\bar{z}$ the mean with respect to all samples by the mean with respect to all negative samples. ■

A.5 Results related to threshold comparison

Lemma A.2: *Define vector $\boldsymbol{z}^+$ with components $z^+ = \boldsymbol{w}^\top \boldsymbol{x}^+$ for $\boldsymbol{x}^+ \in \mathcal{X}^+$ and similarly define vector $\boldsymbol{z}^-$ with components $z^- = \boldsymbol{w}^\top \boldsymbol{x}^-$ for $\boldsymbol{x}^- \in \mathcal{X}^-$. Denote by $\boldsymbol{z}^+_{[\cdot]}$ and $\boldsymbol{z}^-_{[\cdot]}$ the sorted versions of $\boldsymbol{z}^+$ and $\boldsymbol{z}^-$, respectively. Then we have the following statements:*

$$z^+_{[n^+\tau]} > z^-_{[n^-\tau]} \implies \text{Grill has larger threshold than Grill} - \text{NP},$$

$$\frac{1}{n^+\tau} \sum_{i=1}^{n^+\tau} z^+_{[i]} > \frac{1}{n^-\tau} \sum_{i=1}^{n^-\tau} z^-_{[i]} \implies \text{TopMean has larger threshold than } \tau\text{−FPL}.$$

Proof: Since $\boldsymbol{z}^+$ and $\boldsymbol{z}^-$ are computed on disjunctive indices, we have

$$z_{[n\tau]} \geq \min\{z^+_{[n^+\tau]}, z^-_{[n^-\tau]}\}.$$

Since $z_{[n\tau]}$ is the threshold for *Grill* and $z^-_{[n^-\tau]}$ 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 z^+ above z^-, we may expect that the conditions in Lemma A.2 hold true.

Appendix 2. Computation for Section 3.4.1

We derive the results presented in Section 3.4.1 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 that require a large number of points. Since the computation is simple for other formulations, we show it only for *Pat&Mat*.

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

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

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

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

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

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

$$\begin{aligned} \tau &= \frac{1}{n} \sum_{\boldsymbol{x} \in \mathcal{X}} l(\boldsymbol{w}_1^\top \boldsymbol{x} - t) \approx \frac{1}{2} \int_{-1}^{1} l(z - t)\, \mathrm{d}z = \frac{1}{2} \int_{-1}^{1} \max\{0, 1 + \beta(z - t)\}\, \mathrm{d}z \\ &= \frac{1}{2} \int_{-1}^{1} (1 + \beta(z - t))\, \mathrm{d}z = 1 - \beta t + \frac{\beta}{2} \int_{-1}^{1} z\, \mathrm{d}z = 1 - \beta t, \end{aligned} \tag{A9}$$

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

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

and we could have ignored the max operator in (A9). Finally, we have

$$\frac{1}{n^+} \overline{\mathrm{fn}}(\boldsymbol{w}_1, t) \approx \int_0^1 l(t - z)\, \mathrm{d}z = \int_0^1 (1 + t - z)\, \mathrm{d}z = 0.5 + t.$$

Appendix 3. Computing the threshold for Pat&Mat

We show how to efficiently compute the threshold (13) for *Pat&Mat* and the hinge surrogate (3). As always define the scores $z_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ and consider function

$$h(t) = \sum_{i=1}^{n} l(\beta(z_i - t)) - n\tau. \tag{A10}$$

Then solving (30) is equivalent to looking for $\hat{t}$ such that $h(\hat{t}) = 0$. We have the following properties of h.

Lemma A.3: *Function h is continuous and strictly decreasing (until it hits the global minimum) with $h(t) \to \infty$ as $t \to -\infty$ and $h(t) \to -n\tau$ as $t \to \infty$. 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.4: *Let $z_1 \leq z_2 \leq \cdots \leq z_n$ be sorted. Define $\gamma = \frac{1}{\beta}$. Then*

$$h(z_i + \gamma) = h(z_{i+1} + \gamma) + (n - i)\beta(z_{i+1} - z_i) \tag{A11}$$

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

Proof: Observe first that

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

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

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

which finishes the proof. ■

Thus, to solve $h(t)=0$ with the hinge surrogate, we start with $t_n=z_n+\gamma$ and $h(t_n)=-n\tau$. Then we start decreasing t according to (A11) until we find some $t_i=z_i+\gamma$ 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

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

Appendix 4. Proof of Theorem 4.1

The proof is divided into three parts. In Section A.6, we prove a general statement for convergence of stochastic gradient descent with a convex objective. In Section A.7 we apply it to Theorem 4.1. The proof is based on auxiliary results from Section A.8.

A.6 General result

Consider a differentiable objective function f and the optimization method

$$w^{k+1}=w^k-\alpha^k g(w^k), \tag{A12}$$

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) $\|g(w^k)\|\le B$ for all k;
(A3) the stepsize is non-increasing and satisfies $\sum_{k=0}^{\infty}\alpha^k=\infty$;
(A4) the stepsize satisfies $\sum_{k=0}^{\infty}(\alpha^k)^2<\infty$;
(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.6: *Assume that* (A1)–(A4) *is satisfied. If there exists some C such that for some global minimum of w^* of f, we have*

$$\sum_{k=0}^{\infty} \alpha^k \langle g(w^k) - \nabla f(w^k), w^* - w^k \rangle \leq C, \tag{A13}$$

then the sequence $\{w^k\}$ generated by (A12) *is bounded and $f(w^k) \to f(w^*)$. Thus, all its convergent subsequences converge to some global minimum of f.*

Proof: Note first that the convexity from (A1) implies

$$\langle \nabla f(w^k), w^* - w^k \rangle \leq f(w^*) - f(w^k). \tag{A14}$$

Then we have

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

where the inequality follows from (A14) and assumption (A2). Summing this expression for all k and using (A13) 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

$$\limsup_k \|w^k - w^*\|^2 + 2\sum_{k=0}^{\infty} \alpha^k (f(w^k) - f(w^*)) \leq 2\hat{C}. \tag{A15}$$

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 (A15) 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 $\lim f(w^k) \to f(w^*)$, which implies the theorem statement. ■

A.7 Proof of Theorem 4.1

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

(S1) $l(z) \geq 0$ for all $z \in \mathbb{R}$, $l(0) = 1$ and $l(z) \to 0$ as $z \to -\infty$;
(S2) l is convex and strictly increasing function on (z_0, ∞), where $z_0 := \sup\{z \mid l(z) = 0\}$;
(S3) $\frac{l'}{l}$ is a decreasing function on (z_0, ∞);
(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 that is smoothened on an ε-neighbourhood of zero.

Theorem 4.1: *Consider the Pat&Mat or Pat&Mat-NP formulation, stepsizes* $\alpha^k = \frac{\alpha^0}{k+1}$ *and piecewise disjoint minibatches* $I^1, \dots, I^s$ *which cycle periodically* $I^{k+s} = I^k$. *If* l *is the smoothened (Huberized) hinge function, then Algorithm* 4.1 *converges to the global minimum of* (16).

Proof: We intend to apply Theorem A.6 and thus, we need to verify its assumptions. Assumption (A1) is satisfied as f is convex due to Theorem 3.3. Assumption (A2) follows directly from Lemma A.6. Assumptions (A3), (A4) and (A5) are imposed directly in the statement of this theorem. It remains to verify (A13).

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 is equal to $n_+^k = \frac{1}{2} n_+$, where n_+ is the total number of positive samples.

First, we estimate the difference between z_i^k defined in (29) and $x_i^\top w^k$. For any $i \in I^k$ due to the construction (29), we have

$$\begin{aligned} z_i^k &= x_i^\top w^k, \\ z_i^{k-1} = z_i^{k-2} &= x_i^\top w^{k-2} = x_i^\top (w^k + \alpha^{k-2} g(w^{k-2}) + \alpha^{k-1} g(w^{k-1})) \\ &= x_i^\top w^k + \alpha^{k-2} x_i^\top g(w^{k-2}) + \alpha^{k-1} x_i^\top g(w^{k-1}). \end{aligned} \tag{A16}$$

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

$$z_i^k = z_i^{k-1} = x_i^\top w^{k-1} = x_i^\top (w^k + \alpha^{k-1} g(w^{k-1})) = x_i^\top w^k + \alpha^{k-1} x_i^\top g(w^{k-1}). \tag{A17}$$

Recall that we already verified (A1)–(A5). Combining (A2) with (A16) and (A17) yields the existence of some C_2 such that for all $i \in X$ we have

$$\begin{aligned} |z_i^k - x_i^\top w^k| &\le C_2 \alpha^{k-1}, \\ |z_i^{k-1} - x_i^\top w^k| &\le C_2 (\alpha^{k-1} + \alpha^{k-2}). \end{aligned} \tag{A18}$$

This also immediately implies

$$\begin{aligned} |t^k - t(w^k)| &\le C_2 \alpha^{k-1}, \\ |t^{k-1} - t(w^k)| &\le C_2 (\alpha^{k-1} + \alpha^{k-2}). \end{aligned} \tag{A19}$$

Since l' is Lipschitz continuous with modulus L according to (S5), due to (A18) and (A19) we get

$$|l'(t^k - z_i^k) - l'(t(w^k) - x_i^\top w^k)| \le L |t^k - z_i^k - t(w^k) + x_i^\top w^k| \le 2 C_2 L \alpha^{k-1}. \tag{A20}$$

In an identical way, we can show

$$\begin{aligned} |l'(t^{k-1} - z_i^{k-1}) - l'(t(w^k) - x_i^\top w^k)| &\le 2 C_2 L (\alpha^{k-1} + \alpha^{k-2}), \\ |l'(z_i^k - t^k) - l'(x_i^\top w^k - t(w^k))| &\le 2 C_2 L \alpha^{k-1}, \\ |l'(z_i^{k-1} - t^{k-1}) - l'(x_i^\top w^k - t(w^k))| &\le 2 C_2 L (\alpha^{k-1} + \alpha^{k-2}). \end{aligned} \tag{A21}$$

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

$$\begin{aligned} \nabla t^k &= \frac{\sum_{i \in I^k} l'(z_i^k - t^k) x_i + \sum_{i \in I^{k-1}} l'(z_i^{k-1} - t^{k-1}) x_i}{\sum_{i \in X} l'(z_i^k - t^k)}, \\ \nabla t(w^k) &= \frac{\sum_{i \in I^k} l'(x_i^\top w^k - t(w^k)) x_i + \sum_{i \in I^{k-1}} l'(x_i^\top w^k - t(w^k)) x_i}{\sum_{i \in X} l'(x_i^\top w^k - t(w^k))}. \end{aligned} \tag{A22}$$

The first equality in (A22) follows from (33) and (31) while the second equality in (A22) follows from Theorem 3.4 and $X = I^k \cup I^{k-1}$. From Lemma A.5, we deduce that the denominators in (A22)

are bounded away from zero uniformly in k. Assumption (A4) implies $\alpha^k \to 0$. This allows us to use Lemma A.7 which together with (A21) implies that there is some C_3 such that for all sufficiently large k we have

$$\|\nabla t^k - \nabla t(w^k)\| \leq C_3(\alpha^{k-1} + \alpha^{k-2}). \tag{A23}$$

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

$$g(w^k) = \frac{2}{n_+} \sum_{i \in I_+^k} l'(t^k - z_i^k)(\nabla t^k - x_i),$$

$$g(w^{k+1}) = \frac{2}{n_+} \sum_{i \in I_+^{k+1}} l'(t^{k+1} - z_i^{k+1})(\nabla t^{k+1} - x_i),$$

$$\nabla f(w^k) = \frac{1}{n_+} \sum_{i \in X_+} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i),$$

$$\nabla f(w^{k+1}) = \frac{1}{n_+} \sum_{i \in X_+} l'(t(w^{k+1}) - x_i^\top w^{k+1})(\nabla t(w^{k+1}) - x_i).$$

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

$$n^+(g(w^k) + g(w^{k+1}) - \nabla f(w^k) - \nabla f(w^{k+1})) \tag{A24a}$$

$$= \sum_{i \in I_+^k} l'(t^k - z_i^k)(\nabla t^k - x_i) - \sum_{i \in I_+^k} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i) \tag{A24b}$$

$$+ \sum_{i \in I_+^k} l'(t^k - z_i^k)(\nabla t^k - x_i) - \sum_{i \in I_+^k} l'(t(w^{k+1}) - x_i^\top w^{k+1})(\nabla t(w^{k+1}) - x_i) \tag{A24c}$$

$$+ \sum_{i \in I_+^{k+1}} l'(t^{k+1} - z_i^{k+1})(\nabla t^{k+1} - x_i) - \sum_{i \in I_+^{k+1}} l'(t(w^k) - x_i^\top w^k)(\nabla t(w^k) - x_i) \tag{A24d}$$

$$+ \sum_{i \in I_+^{k+1}} l'(t^{k+1} - z_i^{k+1})(\nabla t^{k+1} - x_i) - \sum_{i \in I_+^{k+1}} l'(t(w^{k+1}) - x_i^\top w^{k+1})(\nabla t(w^{k+1}) - x_i). \tag{A24e}$$

Then relations (A23) and (A20) applied to Lemma A.8 imply

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

for some C_4, which gives a bound for (A24b). Bound for (A24e) is obtained by increasing k by one. Bounds for (A24c) and (A24d) can be find similarly using (A21). Altogether, we showed

$$\|g(w^k) + g(w^{k+1}) - \nabla f(w^k) - \nabla f(w^{k+1})\| \leq C_1(\alpha^{k-2} + \alpha^{k-1} + \alpha^k + \alpha^{k+1}) \tag{A25}$$

for some C_1.

We now estimate

$$\alpha^k \langle g(w^k) - \nabla f(w^k), w^* - w^k \rangle + \alpha^{k+1} \langle g(w^{k+1}) - \nabla f(w^{k+1}), w^* - w^{k+1} \rangle \tag{A26a}$$

$$= \langle g(w^k) - \nabla f(w^k), \alpha^k(w^* - w^k) \rangle + \langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) \rangle \tag{A26b}$$

$$= \langle g(w^k) - \nabla f(w^k) + g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^k(w^* - w^k) \rangle \tag{A26c}$$

$$+ \langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k) \rangle. \tag{A26d}$$

To estimate (A26d), we make use of Lemma A.6 to obtain the existence of some C_5 such that

$$\begin{aligned}
&\langle g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k)\rangle \\
&\quad \leq 2B\|\alpha^{k+1}(w^* - w^{k+1}) - \alpha^k(w^* - w^k)\| \\
&\quad = 2B\|\alpha^{k+1}(w^* - w^k + \alpha^k g(w^k)) - \alpha^k(w^* - w^k)\| \\
&\quad = 2B\|(\alpha^{k+1} - \alpha^k)w^* + (\alpha^k - \alpha^{k+1})w^k + \alpha^k\alpha^{k+1}g(w^k)\| \\
&\quad \leq C_5|\alpha^{k+1} - \alpha^k| + C_5(\alpha^k)^2 +_5 (\alpha^{k+1})^2. \qquad \text{(A27)}
\end{aligned}$$

In the last inequality, we used the equality $2ab \leq a^2 + b^2$. To estimate (A26c), we can apply (A25) together with the boundedness of $\{w^k\}$ to obtain the existence of some C_6 such that

$$\begin{aligned}
&\langle g(w^k) - \nabla f(w^k) + g(w^{k+1}) - \nabla f(w^{k+1}), \alpha^k(w^* - w^k)\rangle \\
&\quad \leq C_6(\alpha^{k-2})^2 + C_6(\alpha^{k-1})^2 + C_6(\alpha^k)^2 + C_6(\alpha^{k+1})^2. \qquad \text{(A28)}
\end{aligned}$$

Plugging (27) and (A28) into (A26) and summing the terms yields (A13). Then the assumptions of Theorem A.6 are verified and the theorem statement follows. ■

A.8 Auxiliary results

Lemma A.5: *Let l satisfy* (S1)–(S3). *Then there exists some $\hat{C}$ such that for all k we have*

$$\sum_{i\in X} l'(z_i^k - t^k) \geq \hat{C} > 0,$$

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

Proof: First, we will find an upper bound of $z_i^k - t^k$. Fix any index i_0. Since l is nonnegative due to (S1), Equation (30) implies

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

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

$$l^{-1}(n\tau) \geq z_{i_0}^k - t^k. \qquad \text{(A29)}$$

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

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

where the inequality follows from (A29) and the following equality from (30). 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.6: *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 3.4 and Lemma A.5 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 (28) 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.7: *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 $\alpha^k \to 0$, 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^k are bounded away from zero and since $\alpha^k \to 0$, 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 $\alpha^k \to 0$ 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.8: *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. ■