JMLR: Workshop and Conference Proceedings 42:99-108, 2015 HEPML 2014

Consistent optimization of AMS by logistic loss minimization

Wojciech Kotłowski WKOTLOWSKI@CS.PUT.POZNAN.PL
Poznań University of Technology, Poland

Editor: Glen Cowan, Cécile Germain, Isabelle Guyon, Balàzs Kégl, David Rousseau

Abstract

In this paper, we theoretically justify an approach popular among participants of the Higgs Boson Machine Learning Challenge to optimize approximate median significance (AMS). The approach is based on the following two-stage procedure. First, a real-valued function f is learned by minimizing a surrogate loss for binary classification, such as logistic loss, on the training sample. Then, given f, a threshold $\hat{\theta}$ is tuned on a separate validation sample, by direct optimization of AMS. We show that the regret of the resulting classifier (obtained from thresholding f on $\hat{\theta}$) measured with respect to the squared AMS, is upperbounded by the regret of f measured with respect to the logistic loss. Hence, we prove that minimizing logistic surrogate is a consistent method of optimizing AMS.

Keywords: Approximate median significance (AMS), Higgs Boson Machine Learning Challenge, Kaggle, logistic loss, regret bound, statistical consistency.

1. Introduction

This paper concerns a problem of learning a classifier to optimize approximate median significance (AMS), which was the goal of the Higgs Boson Machine Learning Challenge (HiggsML), hosted by Kaggle website (see Adam-Bourdarios et al. (2014) for details on this contest and description of the problem).

In particular, we are interested in an approach to optimize AMS, based on the following two-stage procedure. First, a real-valued function f is learned by minimizing a surrogate loss for binary classification, such as logistic loss function, on the training sample. In the second stage, given f, a threshold is tuned on a separate "validation" sample, by direct optimization of AMS with respect to a classifier obtained from f by classifying all observations with value of f above the threshold as positive class (signal event), and all observations below the threshold as negative class (background event).

This approach became very popular among HiggsML challenge participants, mainly due to the fact that its first stage, learning a classifier, does not exploit the task evaluation metric (AMS) in any way and thus can employ without modifications any standard classification tools such as logistic regression, LogitBoost, Stochastic Gradient Boosting, Random Forest, etc. (see, e.g., Hastie et al. (2009)). Despite its simplicity, this approach proved to be very effective in achieving high leaderboard score in HiggsML. [1]

[1] See the HiggsML forum https://www.kaggle.com/c/higgs-boson/forums for discussions and presentation of the top score solutions.

The intuition behind this approach is clear: minimization of logistic loss results in estimation of conditional probabilities of signal and background event, and the AMS is assumed to be maximized by classifying the events most likely to be signal as signal events.

This paper formalizes this intuition by showing that the approach described above constitutes a consistent method of optimizing AMS. More specifically, we use the notion of *regret* with respect to some evaluation metric, which is a difference between the performance of a given classifier and the performance of the optimal classifier with respect to this metric. Given a function f, and a classifier $h_{f,\hat{\theta}}$ obtained from f by thresholding f at $\hat{\theta}$, we give a bound on the regret of $h_{f,\hat{\theta}}$ measured with respect to the squared AMS by the regret of f measured with respect to the logistic loss, given that the threshold $\hat{\theta}$ is tuned by optimization of AMS among all classifiers of the form $h_{f,\theta}$ for any threshold value θ. Thus, the goal of this paper is to *theoretically explain the procedure of optimizing AMS by logistic loss minimizing*.

To our knowledge, this is the first regret bound of this form applicable to a non-decomposable performance measure such as AMS. We also discuss generalization of our approach to different performance measures and surrogate loss functions.

Related work. The issue of consistent optimization of performance measures which are functions of true positive and true negative rates has received increasing attention recently in machine learning community (Narasimhan et al., 2014; Natarajan et al., 2014; Zhao et al., 2013). However, these works are mainly concerned with *statistical consistency* also known as *calibration*, which determines whether convergence to the minimizer of a surrogate loss implies convergence to the minimizer of the task performance measure as sample size goes to infinity. Here we give a much stronger result which bounds the regret with respect to squared AMS by the regret with respect to logistic loss. Our result is valid for all finite sample sizes and informs about the rates of convergence.

Recently, Mackey and Bryan (2014) proposed a classification cascade approach to optimize AMS. Their method, based on the theory of Fenchel's duality, iteratively alternates between solving a cost-sensitive binary classification problem and updating misclassification costs. In contrast, the method described here requires solving an ordinary binary classification problem just once.

Outline. The paper is organized as follows. In Section 2, we introduce basic concepts needed to state our main result presented in Section 3 and proved in Section 4. Section 5 discusses generalization of our results beyond AMS and logistic loss.

2. Problem Setting

Binary classifier. In binary classification, the goal is, given an input (feature vector) $x \in X$, to accurately predict the output (label) $y \in \{-1, 1\}$. We assume input-output pairs (x, y), which we call *observations*, are generated i.i.d. according to probability distribution $\Pr(x, y)$.[2] A *classifier* is a mapping $h \colon X \to \{-1, 1\}$. Given h, we define the following two

[2] The original HiggsML problem was to optimize the score on a finite (test) sample, which is a special case of a distribution. HiggML problem also involved observations' weights, but without loss of generality, they can be incorporated into the $\Pr(x, y)$.

quantities:

$$s(h) = \Pr(h(x) = 1, y = 1),$$
$$b(h) = \Pr(h(x) = 1, y = -1),$$

which can be interpreted as true positive and false positive rates of h.

AMS and regret. Given a classifier h, define its *approximate median significance* (AMS) score (Cowan et al., 2011) as $\mathrm{AMS}(h) = \mathrm{AMS}(s(h), b(h))$, where:[3]

$$\mathrm{AMS}(s,b) = \sqrt{2\left((s+b)\log\left(1+\frac{s}{b}\right) - s\right)}.$$

It is more convenient to deal with a *squared* AMS:

$$\mathrm{AMS}^2(s,b) = 2\left((s+b)\log\left(1+\frac{s}{b}\right) - s\right),$$

and this quantity is used throughout the paper. By calculating the derivatives, it is easy to verify that $\mathrm{AMS}^2(s,b)$ is increasing in s and decreasing in b:

$$\frac{\partial \mathrm{AMS}^2(s,b)}{\partial s} = \log\left(1+\frac{s}{b}\right) \geq 0, \qquad \frac{\partial \mathrm{AMS}^2(s,b)}{\partial b} = \log\left(1+\frac{s}{b}\right) - \frac{s}{b} \leq 0.$$

Moreover, $\mathrm{AMS}^2(s,b)$ is jointly convex with respect to (s,b). Indeed the Hessian matrix is given by:

$$\nabla^2 \mathrm{AMS}^2(s,b) \;=\; \begin{pmatrix} \frac{1}{b+s} & -\frac{s}{b(b+s)} \\ -\frac{s}{b(b+s)} & \frac{s^2}{b^2(b+s)} \end{pmatrix} \;=\; \frac{1}{b+s}\mathbf{u}\mathbf{u}^\top,$$

where $\mathbf{u} = \left(1, \frac{b}{s}\right)$. Hence $\nabla^2 \mathrm{AMS}^2(s,b)$ is positive semidefinite.

Let h^*_{AMS} be the classifier which maximizes the AMS^2 over all possible classifiers:

$$h^*_{\mathrm{AMS}} = \operatorname*{arg\,max}_{h \in \{-1,1\}^X} \mathrm{AMS}^2(h)$$

(if the maximizer is non-unique, we take any maximizer of AMS^2 as h^*_{AMS}). Given h, we define its *AMS regret* as the distance of h from the optimal classifier h^*_{AMS} measured by means of AMS^2:

$$R_{\mathrm{AMS}}(h) = \mathrm{AMS}^2(h^*_{\mathrm{AMS}}) - \mathrm{AMS}^2(h).$$

AMS regret is a better performance metric than the AMS itself, since it specifies how much worse is h comparing to the optimal h^*_{AMS}.

[3] Comparing to the definition in (Adam-Bourdarios et al., 2014), we skip the regularization term b_{reg}. This comes without loss of generality, as b_{reg} can be incorporated into b and, since it affects all classifiers equally, will vanish in the definition of regret.

Logistic loss and logistic regret. Given a real number f, and a label y, we define the logistic loss $\ell_{\log}\colon \{-1,1\} \times \mathbb{R} \to \mathbb{R}_+$ as:

$$\ell_{\log}(y,f) = \log\left(1+e^{-yf}\right).$$

The logistic loss is a commonly used surrogate loss function for binary classification, employed in various learning methods, such as logistic regression, LogitBoost or Stochastic Gradient Boosting (see, e.g., Hastie et al. (2009)). It is convex in f, so minimizing logistic loss over the training sample becomes a convex optimization problem, which can be solved efficiently. Another advantage of logistic loss is that the sigmoid transform of f, $(1+e^{-f})^{-1}$, can be used to obtain probability estimates $\Pr(y|x)$.

Given a real-valued function $f\colon X \to \mathbb{R}$, its expected logistic loss $L_{\log}(f)$ is defined as:

$$L_{\log}(f) = \mathbb{E}_{(x,y)}[\ell_{\log}(y, f(x))].$$

Let:

$$f^*_{\log} = \operatorname*{arg\,min}_f L_{\log}(f),$$

be the minimizer of $L_{\log}(f)$ among all functions $f\colon X \to \mathbb{R}$ (as before, in case the minimizer is non-unique, we take any minimizer as $f^*_{\log}$). We define the logistic *regret* of f as:

$$R_{\log}(f) = L_{\log}(f) - L_{\log}(f^*_{\log}).$$

3. Main Result

Any real-valued function $f\colon X \to \mathbb{R}$ can be turned into a classifier $h_{f,\theta}\colon X \to \{-1,1\}$, by thresholding at some value θ:

$$h_{f,\theta}(x) = \operatorname{sgn}(f(x) - \theta),$$

where $\operatorname{sgn}(x)$ is the sign function, and we use the convention that $\operatorname{sgn}(0) = 1$.

The purpose of this paper is to address the following problem: given a function f with logistic regret $R_{\log}(f)$, and a threshold θ, what is the maximum AMS regret of $h_{f,\theta}$? In other words, can we bound $R_{\mathrm{AMS}}(h_{f,\theta})$ in terms of $R_{\log}(f)$? If f is close to $f^*_{\log}$ in terms of expected logistic loss, does it also imply that $h_{f,\theta}$ is close to h^*_{AMS} in terms of squared AMS? We give a positive answer to this question, which based on the following regret bound:

Lemma 1 *There exists a threshold θ^*, such that for any f,*

$$R_{\mathrm{AMS}}(h_{f,\theta^*}) \leq \frac{s(h^*_{\mathrm{AMS}})}{b(h^*_{\mathrm{AMS}})}\sqrt{\frac{1}{2}R_{\log}(f)}.$$

The proof is quite long and hence is postponed to Section 4. Interestingly, the proof goes by an intermediate bound of the AMS regret by a cost-sensitive classification regret, with misclassification costs proportional to the gradient coordinates of the AMS.

Lemma 1 has the following interpretation. If we are able to find a function f with small logistic regret, we are guaranteed that there exists a threshold θ^* such that h_{f,θ^*} has small AMS regret. Note that the same threshold θ^* will work for any f, and the right hand side of the bound is *independent* of θ^*. We are now ready to prove the main result of the paper:

Theorem 2 *Given a real-valued function f, let $\hat{\theta} = \arg\max_\theta \mathrm{AMS}(h_{f,\theta})$. Then:*

$$R_{\mathrm{AMS}}(h_{f,\hat{\theta}}) \leq \frac{s(h^*_{\mathrm{AMS}})}{b(h^*_{\mathrm{AMS}})} \sqrt{\frac{1}{2} R_{\log}(f)}.$$

Proof The result follows immediately from Lemma 1 by noticing that solving $\max_\theta \mathrm{AMS}(h_{f,\theta})$ is equivalent to solving $\min_\theta R_{\mathrm{AMS}}(h_{f,\theta})$, and that $\min_\theta R_{\mathrm{AMS}}(h_{f,\theta}) \leq R_{\mathrm{AMS}}(h_{f,\theta^*})$. ■

Theorem 2 motivates the following procedure for AMS maximization:

1. Find f with small logistic regret, e.g. by employing a learning algorithm minimizing logistic loss on the training sample.

2. Given f, solve $\hat{\theta} = \arg\max_\theta \mathrm{AMS}(h_{f,\theta})$.

Theorem 2 states that the AMS regret of the classifier obtained by this procedure is upperbounded by the logistic regret of the underlying real-valued function.

We now discuss how to approach step 2 of the procedure in practice. In principle, this step requires maximizing AMS defined by means of an unknown distribution $\Pr(x, y)$. However, it is sufficient to optimize θ on the empirical counterpart of AMS calculated on a separate validation sample. Indeed, step 2 involves optimization within a class of threshold functions (since f is fixed), which has VC-dimension equal to 2 (Devroye et al., 1996). By convexity of AMS^2,

$$\mathrm{AMS}^2(s,b) - \mathrm{AMS}^2(\hat{s},\hat{b}) \leq \left(\frac{\partial \mathrm{AMS}^2(s,b)}{\partial s}, \frac{\partial \mathrm{AMS}^2(s,b)}{\partial b}\right)^\top (s - \hat{s}, b - \hat{b}) \tag{1}$$

(see, e.g. Boyd and Vandenberghe (2004)), where $\hat{s}$ and $\hat{b}$ are empirical counterparts of s and b. Since $\hat{s}$ and $\hat{b}$ are empirical means of some quantities, we employ the results of VC theory and state that the deviations of empirical means $\hat{s}$ and $\hat{b}$ from their expectations s and b, respectively, can be upperbounded with high probability *uniformly* over the class of all threshold functions by $O(1/\sqrt{m})$, where m is the validation sample size. This and (1) implies[4] uniform convergence on AMS^2. This in turn means, that $\mathrm{AMS}^2(s,b)$ of the empirical maximizer is $O(1/\sqrt{m})$ close to the $\max_\theta \mathrm{AMS}^2(h_{f,\theta})$. Hence, step 2 can be performed within $O(1/\sqrt{m})$ accuracy on a validation sample independent from the training sample.

4. Proof of Lemma 1

The proof consists of two steps. First, we bound the AMS regret of any classifier h by its cost-sensitive classification regret (introduced below). Next, we show that there exists a threshold θ^*, such that for any f, the cost-sensitive classification regret of h_{f,θ^*} is upperbounded by the logistic regret of f.

[4] In the HiggsML problem, the gradients of AMS^2 are bounded due to regularization term b_r.

Bounding AMS regret by cost-sensitive classification regret. Given a real number $c \in (0,1)$, define a *cost-sensitive classification loss* $\ell_c \colon \{-1,1\} \times \{-1,1\} \to \mathbb{R}_+$ as:

$$\ell_c(y,h) = c\mathbb{1}[y=-1]\mathbb{1}[h=1] + (1-c)\mathbb{1}[y=1]\mathbb{1}[h=-1],$$

where $\mathbb{1}[A]$ is the indicator function equal to 1 if predicate A is true, and 0 otherwise. The cost-sensitive loss assigns different costs of misclassification for positive and negative labels. Given classifier h, the expected cost-sensitive loss of h is:

$$L_c(h) = \mathbb{E}_{(x,y)}[\ell_c(y,h(x))] = cb(h) + (1-c)(\Pr(y=1) - s(h)),$$

where $s(h)$ and $b(h)$ are true positive and false positive rates defined before. Let $h_c^* = \arg\min_h L_c(h)$ be the minimizer of the expected cost-sensitive loss among all classifiers. Define the cost-sensitive classification regret as:

$$R_c(h) = L_c(h) - L_c(h_c^*).$$

Any convex and differentiable function $g(x)$ satisfies $g(x) \geq g(y) + \nabla g(y)^\top (x-y)$ for any x, y in its convex domain (Boyd and Vandenberghe, 2004). Applying this inequality to $\mathrm{AMS}^2(s,b)$ jointly convex in (s,b), we have for any $s, b, s^*, b^* \in [0,1]$:

$$\mathrm{AMS}^2(s,b) \geq \mathrm{AMS}^2(s^*,b^*) + \left(\frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial s^*}, \frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial b^*}\right)^\top (s-s^*, b-b^*). \tag{2}$$

Given classifier h, we set $s = s(h), b = b(h), s^* = s(h_{\mathrm{AMS}}^*), b^* = b(h_{\mathrm{AMS}}^*)$, and:

$$C := \frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial s^*} - \frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial b^*}, \qquad c := -\frac{1}{C}\frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial b^*}.$$

Since $\mathrm{AMS}^2(s,b)$ is increasing in s and decreasing in b, both $\frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial s^*}$ and $-\frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial b^*}$ are positive, which implies $C > 0$ and $0 < c < 1$. In this notation, (2) boils down to:

$$\begin{aligned} R_{\mathrm{AMS}}(h) = \mathrm{AMS}^2(h_{\mathrm{AMS}}^*) - \mathrm{AMS}^2(h) &\leq C\Big(c(b(h) - b(h_{\mathrm{AMS}}^*)) + (1-c)(s(h_{\mathrm{AMS}}^*) - s(h))\Big) \\ &= C\big(L_c(h) - L_c(h_{\mathrm{AMS}}^*)\big) \\ &\leq C\big(L_c(h) - L_c(h_c^*)\big) = CR_c(h), \end{aligned}$$

where the last inequality follows from the definition of h_c^*. Thus, the AMS regret is upper-bounded by the cost-sensitive classification regret with costs proportional to the gradient coordinates of $\mathrm{AMS}^2(s^*,b^*)$ at optimum h_{AMS}^*.[5]

Bounding cost-sensitive classification regret by logistic regret. We first give a bound on cost-sensitive classification regret by means of logistic regret *conditioned* at a given x. This part relies on the techniques used by Bartlett et al. (2006). Then, the final bound is obtained by taking expectation with respect to x, and applying Jensen's inequality.

Given a label $h \in \{-1,1\}$, and $\eta \in [0,1]$, define *conditional* cost-sensitive classification loss as:

$$\ell_c(\eta,h) = c(1-\eta)\mathbb{1}[h=1] + (1-c)\eta\mathbb{1}[h=-1].$$

[5] Note that the gradient at optimum does not vanish, as the optimum is with respect to h, not (s,b).

The reason this quantity is called "conditional loss" becomes clear if we note that for any classifier h, $L_c(h) = \mathbb{E}_x[\ell_c(\eta(x), h(x))]$, where $\eta(x) = \Pr(y = 1|x)$. In other words, $\ell_c(\eta(x), h(x))$ is the loss of h conditioned on x.

Given η, let $h_c^* = \arg\min_{h \in \{-1,1\}} \ell_c(\eta, h)$. It can be easily verified that:

$$h_c^* = \operatorname{sgn}(\eta - c),$$

and:

$$\ell_c(\eta, h_c^*) = \min\{c(1-\eta), (1-c)\eta\}.$$

The conditional regret of h is defined as $r_c(\eta, h) = \ell_c(\eta, h) - \ell_c(\eta, h_c^*)$. If $h = h_c^*$ then the obviously $r_c(\eta, h) = 0$. On the other hand, if $h \neq h_c^*$, then:

$$\begin{aligned} r_c(\eta, h) &= c(1-\eta)\mathbb{1}[h = 1] + (1-c)\eta\mathbb{1}[h = -1] - \min\{c(1-\eta), (1-c)\eta\} \\ &= \max\{c(1-\eta), (1-c)\eta\} - \min\{c(1-\eta), (1-c)\eta\} \\ &= |c(1-\eta) - (1-c)\eta| = |\eta - c|. \end{aligned}$$

Summarizing:

$$r_c(\eta, h) = \begin{cases} 0 & \text{if } \; h = h_c^*, \\ |\eta - c| & \text{if } \; h \neq h_c^*. \end{cases}$$

Given a real number f, and $\eta \in [0,1]$, define *conditional* logistic loss as:

$$\ell_{\log}(\eta, f) = (1-\eta)\log\left(1 + e^{f}\right) + \eta \log\left(1 + e^{-f}\right).$$

Let $f_{\log}^* = \arg\min_{f \in \mathbb{R}} \ell_{\log}(\eta, f)$. By differentiating $\ell_{\log}(\eta, f)$ with respect to f, and setting the derivative to 0, we get that:

$$f_{\log}^* = \log \frac{\eta}{1-\eta},$$

and $\ell_{\log}(\eta, f_{\log}^*) = -\eta\log\eta - (1-\eta)\log(1-\eta)$, which is the binary entropy of η. The *conditional logistic regret* of f is given by

$$r_{\log}(\eta, f) = \ell_{\log}(\eta, f) - \ell_{\log}(f_{\log}^*).$$

The conditional regret has a particularly simple form when f is re-expressed as a probability estimate η_f:

$$r_{\log}(\eta, f) = D(\eta \| \eta_f), \qquad \text{where} \quad \eta_f := \frac{1}{1 + e^{-f}},$$

and $D(\eta\|\eta_f) = \eta \log \frac{\eta}{\eta_f} + (1-\eta)\log\frac{1-\eta}{1-\eta_f}$ is the Kullback-Leibler divergence. By Pinsker's inequality,

$$D(\eta\|\eta_f) \geq 2(\eta - \eta_f)^2.$$

Given real number f, define $h_{f,\theta^*} = \operatorname{sgn}(f - \theta^*)$, where:

$$\theta^* = \log \frac{c}{1-c}.$$

We will now bound the conditional cost-sensitive classification regret $r_c(\eta, h_{f,\theta^*})$ in terms of conditional logistic regret $r_{\log}(\eta, f)$. First note that:

$$h_{f,\theta^*} = 1 \iff f \geq \theta^* = \log \frac{c}{1-c} \iff \frac{1}{1+e^{-f}} \geq c \iff \eta_f \geq c,$$

so that we can equivalently write $h_{f,\theta^*} = \operatorname{sgn}(\eta_f - c)$. Since $h_c^* = \operatorname{sgn}(\eta - c)$, then whenever $(\eta_f - c)(\eta - c) > 0$, it holds $h_{f,\theta^*} = h_c^*$, and $r_c(\eta, h_{f,\theta^*}) = 0$. On the other hand, when $(\eta_f - c)(\eta - c) \leq 0$, it holds[6] $r_c(\eta, h_{f,\theta^*}) \leq |\eta - c|$, whereas:

$$\begin{aligned} r_{\log}(\eta, f) &= D(\eta \| \eta_f) \\ \text{(Pinsker's inequality)} \quad &\geq 2(\eta - \eta_f)^2 \\ &= 2(\eta - c + c - \eta_f)^2 \\ &= 2(\eta - c)^2 + 4(\eta - c)(c - \eta_f) + 2(c - \eta_f)^2 \\ \text{(because } (\eta_f - c)(\eta - c) \leq 0)) \quad &\geq 2(\eta - c)^2 \\ &\geq 2r_c^2(\eta, h_{f,\theta^*}). \end{aligned}$$

Taking both cases together, we get:

$$r_c(\eta, h_{f,\theta^*}) \leq \sqrt{r_{\log}(\eta, f)/2}.$$

Now, given any function f,

$$\begin{aligned} R_c(h_{f,\theta^*}) &= \mathbb{E}_x[r_c(\eta, h_{f,\theta^*})] \\ &\leq \mathbb{E}_x\left[\sqrt{r_{\log}(\eta, f)/2}\right] \\ &\leq \sqrt{\mathbb{E}_x[r_{\log}(\eta, f)]/2} \\ &= \sqrt{R_{\log}(f)/2}, \end{aligned}$$

where the last inequality is from Jensen's inequality applied to the concave function $x \mapsto \sqrt{x}$.

Finishing the proof. Combining the results from both parts, we get:

$$R_{\mathrm{AMS}}(h_{f,\theta^*}) \leq C R_c(h_{f,\theta^*}) \leq C\sqrt{R_{\log}(f)/2},$$

where $\theta^* = \log \frac{c}{1-c}$ is independent of f. Recalling that $C = \frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial s^*} - \frac{\partial \mathrm{AMS}^2(s^*,b^*)}{\partial b^*}$, we calculate:

$$C = \log\left(1 + \frac{s^*}{b^*}\right) - \left(\log\left(1 + \frac{s^*}{b^*}\right) - \frac{s^*}{b^*}\right) = \frac{s^*}{b^*},$$

where $s^* = s(h^*_{\mathrm{AMS}})$ and $b^* = b(h^*_{\mathrm{AMS}})$. This finished the proof. □

Note that the proof actually specifies the exact value of the universal threshold θ^*:

$$\theta^* = \log \frac{c}{1-c}, \qquad \text{where } c = 1 - \frac{b^*}{s^*} \log\left(1 + \frac{s^*}{b^*}\right).$$

[6] $r_c(\eta, h_{f,\theta^*}) = |\eta - c|$ if $(\eta_f - c)(\eta - c) < 0$, and can be either 0 or $|\eta - c|$ when $(\eta_f - c)(\eta - c) = 0$.

5. Generalization beyond AMS and logistic loss

Results of this paper can be generalized beyond AMS metric and logistic loss surrogate. The AMS can be replaced by any other evaluation metric, which enjoys the following two properties: 1) is increasing in s, and decreasing in b; 2) is jointly convex in s and b. These were the only two properties of the AMS used in the proof of Lemma 1. The logistic loss surrogate can be replaced by any other convex surrogate loss ℓ, such that the following property holds: There exists a threshold θ^* which is a function of the cost c, such that for all f,

$$R_c(h_{f,\theta^*}) \leq \lambda \sqrt{R_\ell(f)}, \tag{3}$$

for some positive constant λ. This property is satisfied by, e.g., squared error loss $\ell_{\mathrm{sq}}(y, f) = (y - f)^2$ with $\lambda = 1$, which can be verified by noticing that the logistic regret upperbounds the squared error regret by Pinsker's inequality. More generally, (3) is closely related to the properties of loss functions known as *strongly proper composite losses* (Agarwal, 2014).

Acknowledgments

The author was supported by the Foundation For Polish Science Homing Plus grant, co-financed by the European Regional Development Fund. The author would like to thank Krzysztof Dembczyński for interesting discussions and proofreading the paper.

References

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. Learning to discover: the Higgs boson machine learning challenge, 2014. URL http://higgsml.lal.in2p3.fr/documentation/.

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014. URL http://jmlr.org/papers/v15/agarwal14b.html.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C-Particles and Fields*, 71(2):1–19, 2011.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1st edition, 1996.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Lester Mackey and Jordan Bryan. Weighted classification cascades for optimizing discovery significance in the HiggsML challenge. *CoRR*, abs/1409.2655, 2014. URL `http://arxiv.org/abs/1409.2655`.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Neural Information Processing Systems (NIPS)*, 2014.

Nagarajan Natarajan, Oluwasanmi Koyejo, Pradeep K. Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. In *Neural Information Processing Systems (NIPS)*, 2014.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14:1033–1090, 2013.