Surrogate regret bounds for generalized classification performance metrics

Wojciech Kotłowski
Krzysztof Dembczyński

Poznań University of Technology

PL-SIGML, Częstochowa, 14.04.2016

Motivation

Kaggle Higgs Boson Machine Learning Challenge

Higgs challenge

Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

Mon 12 May 2014 – Mon 15 Sep 2014 (18 months ago)

Dashboard

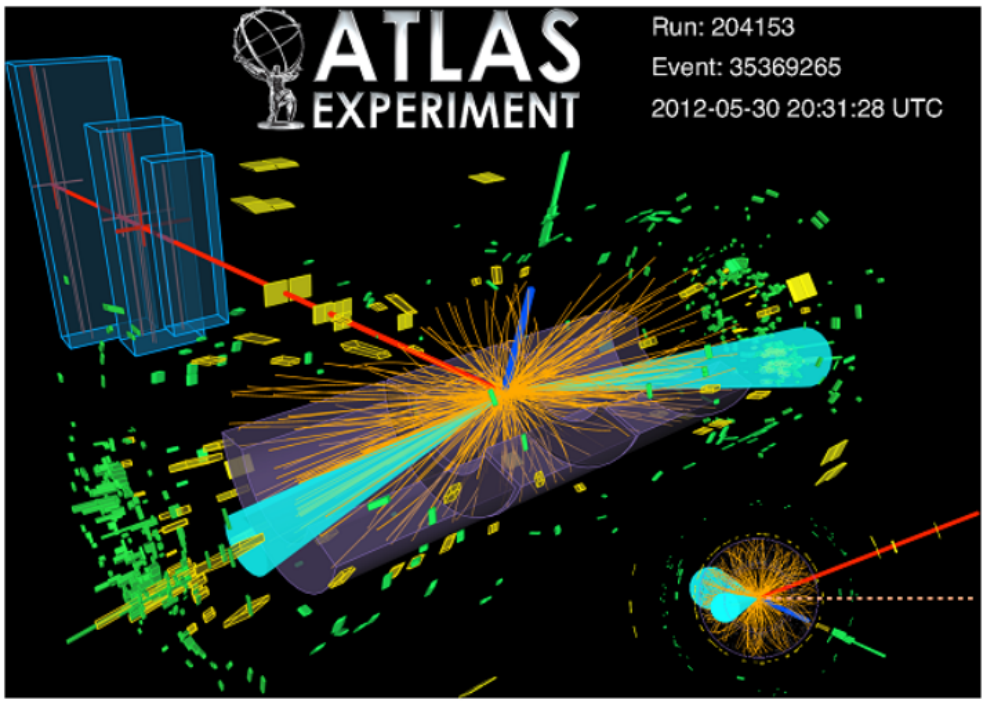
- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - About the Sponsors
  - Timeline
  - Winners
- Forum
- Leaderboard
  - Public
  - Private

Private Leaderboard

1. Gábor Melis
2. Tim Salimans
3. nhlx5haze
4. ChoKo Team
5. cheng chen
6. quantify
7. Stanislav Semenov & Co (HSE Yandex)

Competition Details » Get the Data » Make a submission

Use the ATLAS experiment to identify the Higgs boson



Data set and rules

- Classes: "signal" ($h \to \tau^+\tau^-$) and "background".

# events	# features	% signal weight
250 000	30	0.17

Data set and rules

- Classes: "signal" ($h \to \tau^+\tau^-$) and "background".

# events	# features	% signal weight
250 000	30	0.17

- Evaluation: *Approximate Median Significance* (AMS):

$$\mathrm{AMS} = \sqrt{2(s + b + 10)\log\left(1 + \frac{s}{b + 10}\right) - s}$$

s, b – weight of signal/background events classified as signal.

How to optimize AMS?

Research Problem

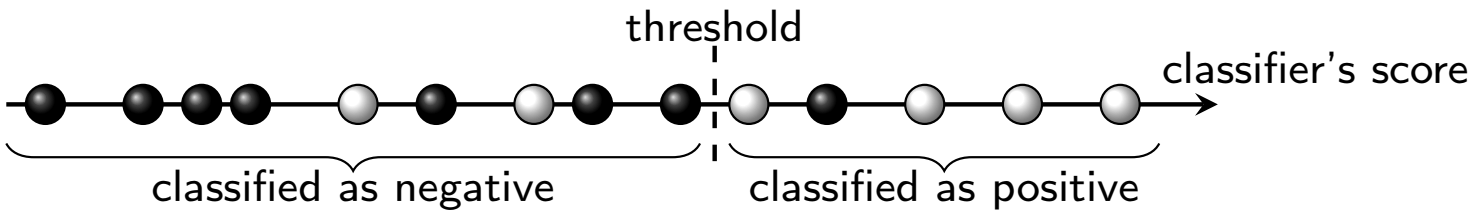
How to optimize a global function of true/false positives/negatives, not decomposable into individual losses over the observations?

How to optimize AMS?

Research Problem

How to optimize a global function of true/false positives/negatives, not decomposable into individual losses over the observations?

Most popular approach:
Sort classifier's scores and threshold to maximize AMS.



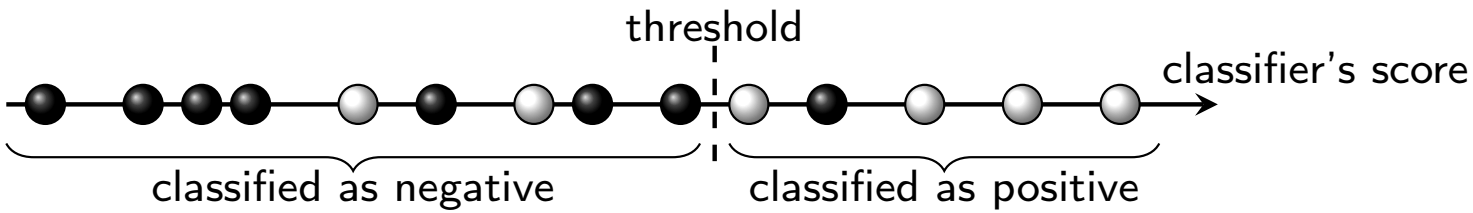
AMS not used while training, only for tuning the threshold.

How to optimize AMS?

Research Problem

How to optimize a global function of true/false positives/negatives, not decomposable into individual losses over the observations?

Most popular approach:
Sort classifier's scores and threshold to maximize AMS.



AMS not used while training, only for tuning the threshold.

Is this approach theoretically justified?

Optimization of generalized performance metrics

- Wojciech Kotłowski and Krzysztof Dembczyński. Surrogate regret bounds for generalized classification performance metrics. In *ACML*, volume 45 of *JMLR W&C Proc.*, pages 301–316, 2015 [Best Paper Award]
- Wojciech Kotłowski. Consistent optimization of AMS by logistic loss minimization. In *NIPS HEPML Workshop*, volume 42 of *JMLR W&C Proc.*, pages 99–108, 2015

Generalized performance metrics for binary classification

Given a binary classifier
$h\colon X \to \{-1, 1\}$, define:

$$\Psi(h) = \Psi\Big(\mathrm{FP}(h), \mathrm{FN}(h)\Big),$$

	predicted $\hat{y} = h(x)$		
true y	-1	$+1$	total
-1	TN	FP	$1 - P$
$+1$	FN	TP	P

where:

$$\mathrm{FP}(h) = \Pr(h(x) = 1 \wedge y = -1),$$
$$\mathrm{FN}(h) = \Pr(h(x) = -1 \wedge y = 1).$$

Goal: maximize $\Psi(h)$.

We assume Ψ is non-increasing in FP and FN.

Linear-fractional performance metric

Definition

$$\Psi(\mathrm{FP}, \mathrm{FN}) = \frac{a_0 + a_1 \mathrm{FP} + a_2 \mathrm{FN}}{b_0 + b_1 \mathrm{FP} + b_2 \mathrm{FN}},$$

Linear-fractional performance metric

Definition

$$\Psi(\mathrm{FP}, \mathrm{FN}) = \frac{a_0 + a_1 \mathrm{FP} + a_2 \mathrm{FN}}{b_0 + b_1 \mathrm{FP} + b_2 \mathrm{FN}},$$

Examples

Accuracy	$\mathrm{Acc} = 1 - \mathrm{FN} - \mathrm{FP}$
F_β-measure	$F_\beta = \frac{(1+\beta^2)(P-\mathrm{FN})}{(1+\beta^2)P-\mathrm{FN}+\mathrm{FP}}$
Jaccard similarity	$J = \frac{P-\mathrm{FN}}{P+\mathrm{FP}}$
AM measure	$\mathrm{AM} = 1 - \frac{1}{2P}\mathrm{FN} - \frac{1}{2(1-P)}\mathrm{FP}$
Weighted accuracy	$\mathrm{WA} = 1 - w_-\mathrm{FP} - w_+\mathrm{FN}$

Convex performance metrics

Definition

$\Psi(\mathrm{FP}, \mathrm{FN})$ is jointly convex in FP and FN.

Convex performance metrics

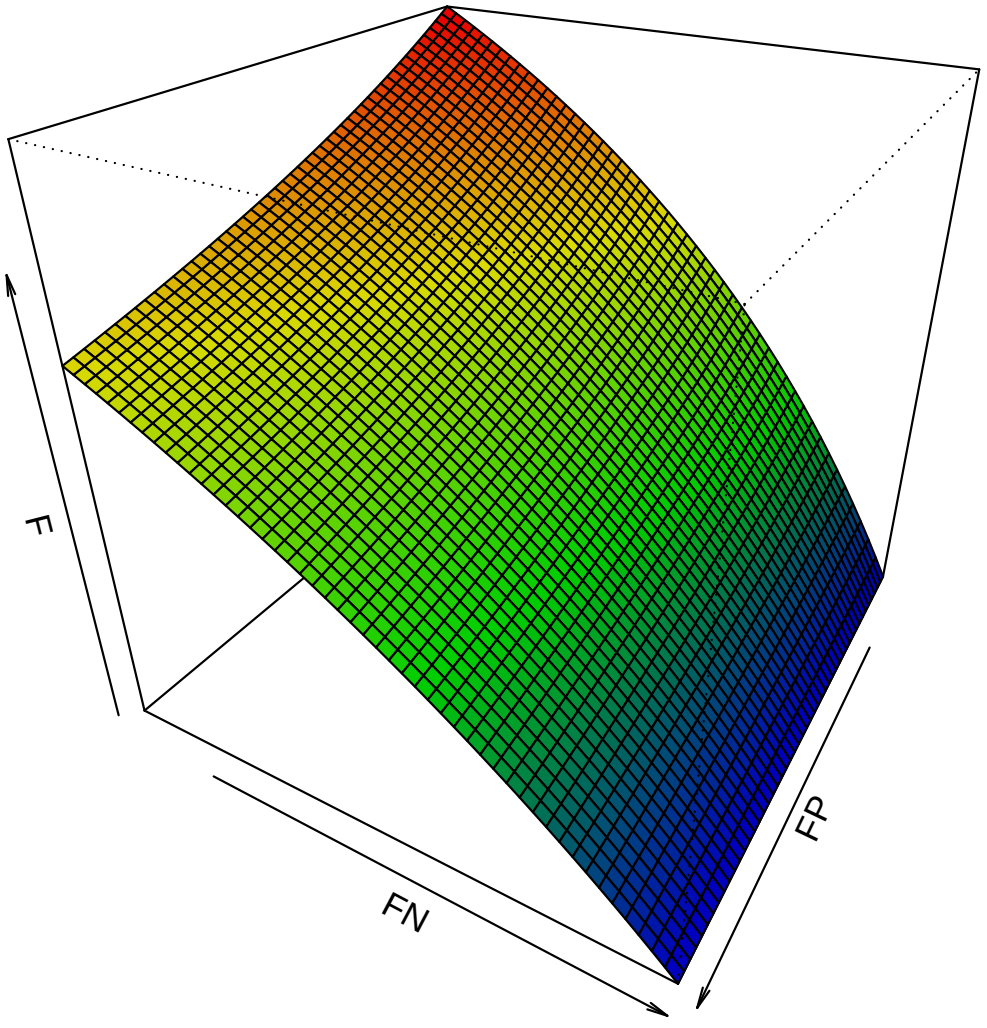
Definition

$\Psi(\mathrm{FP}, \mathrm{FN})$ is jointly convex in FP and FN.

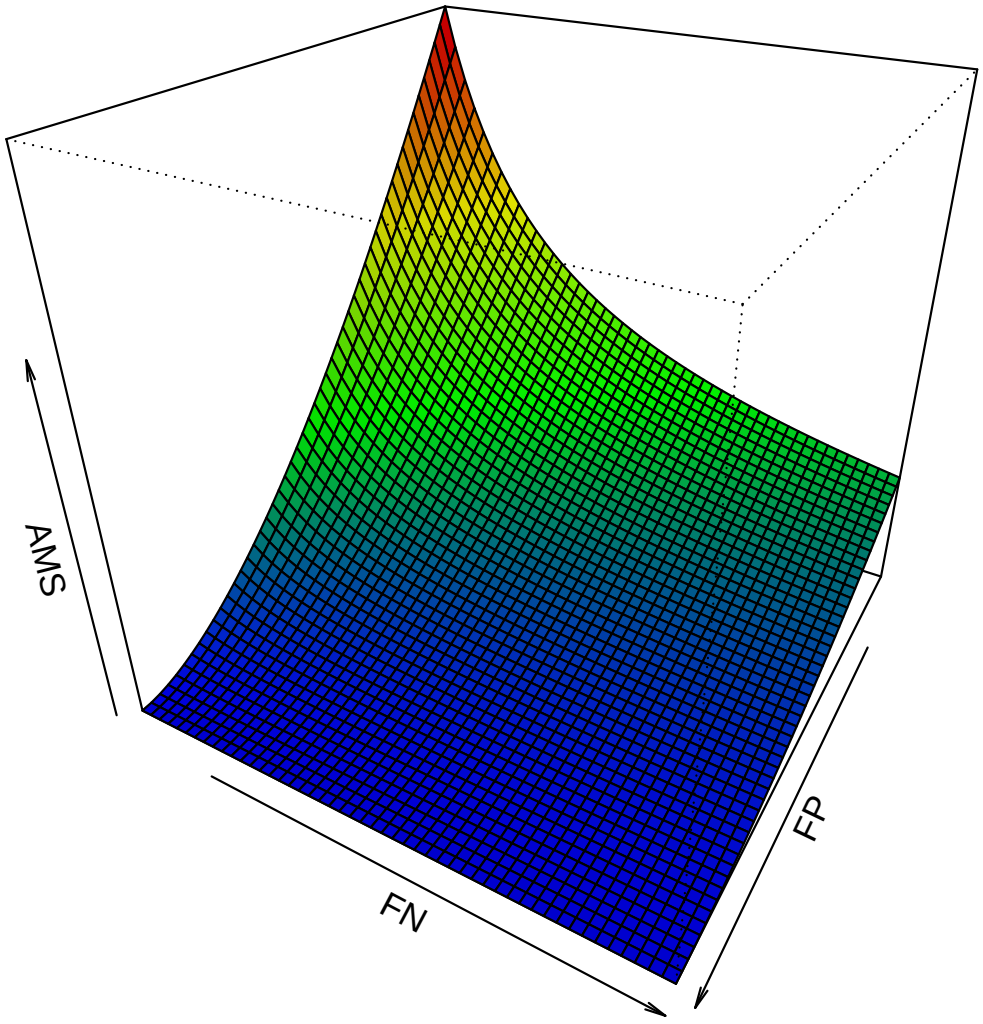
Example: AMS^2 score

$$\mathrm{AMS}^2(\mathrm{TP}, \mathrm{FP}) = 2\left((\mathrm{TP}+\mathrm{FP})\log\left(1+\frac{\mathrm{TP}}{\mathrm{FP}}\right)-\mathrm{TP}\right).$$

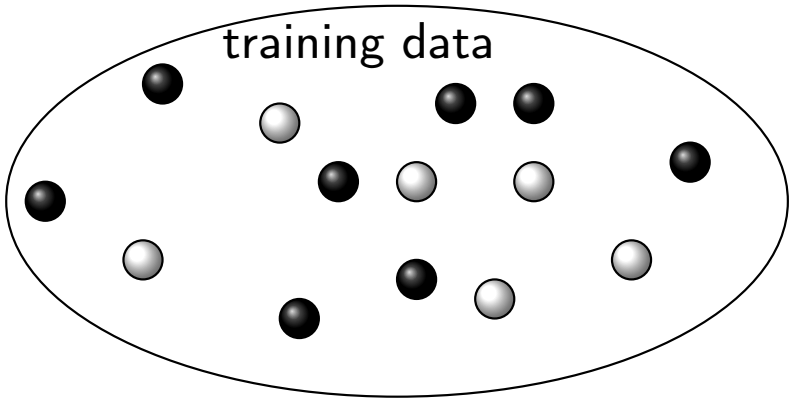
Example - F_1-measure



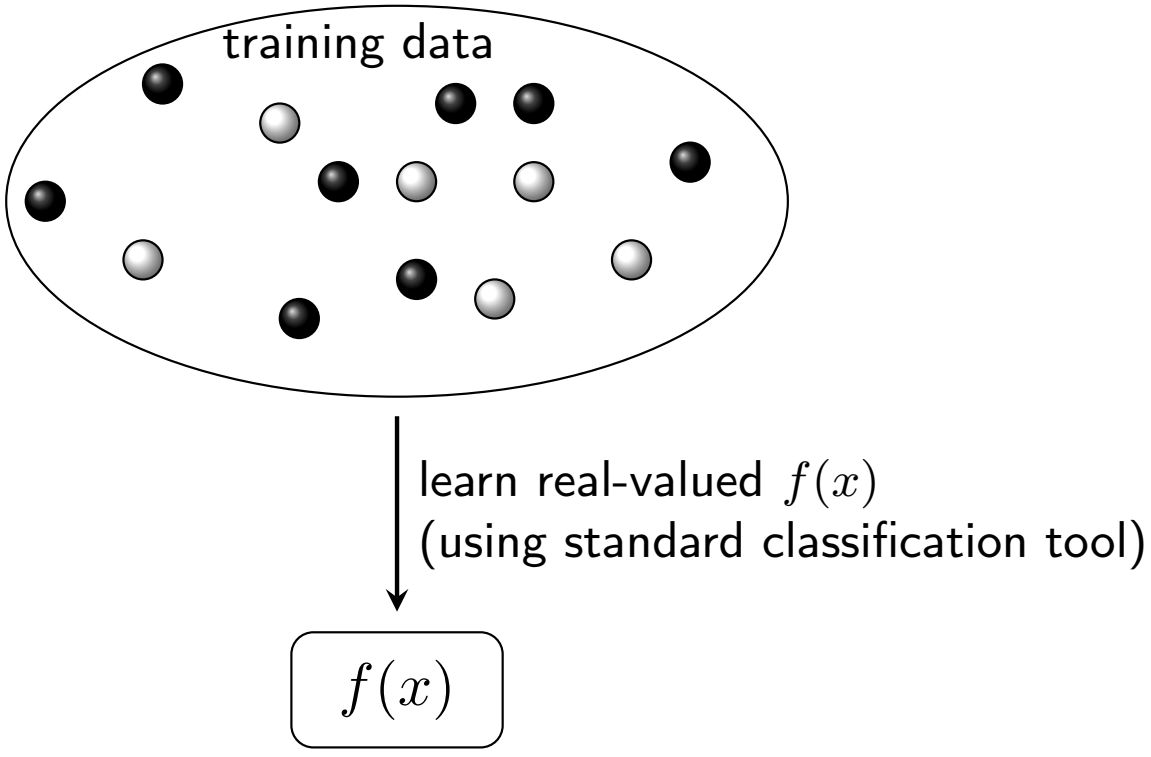
Example - AMS^2 score



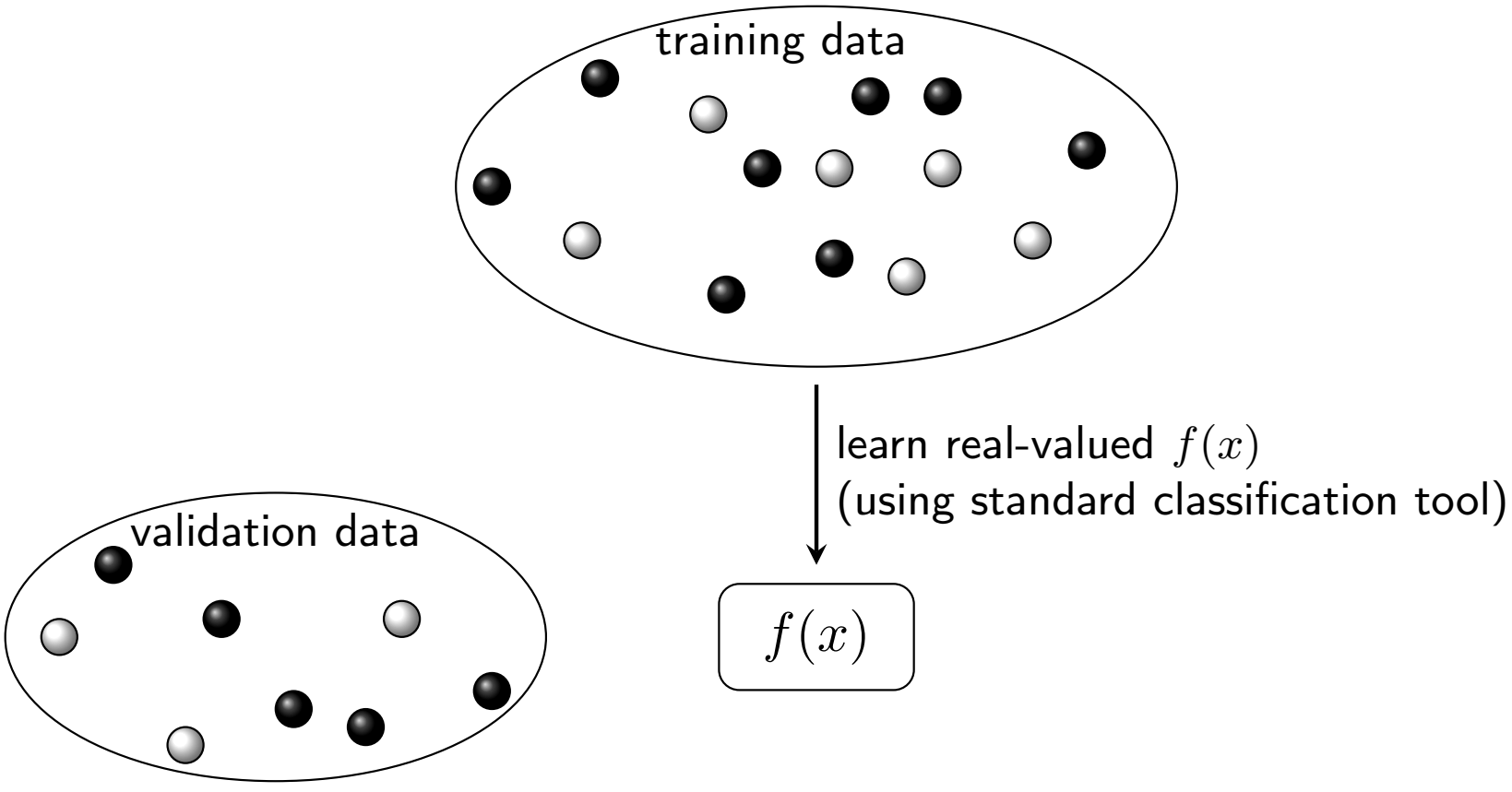
A simple approach to optimization of $\Psi(h)$



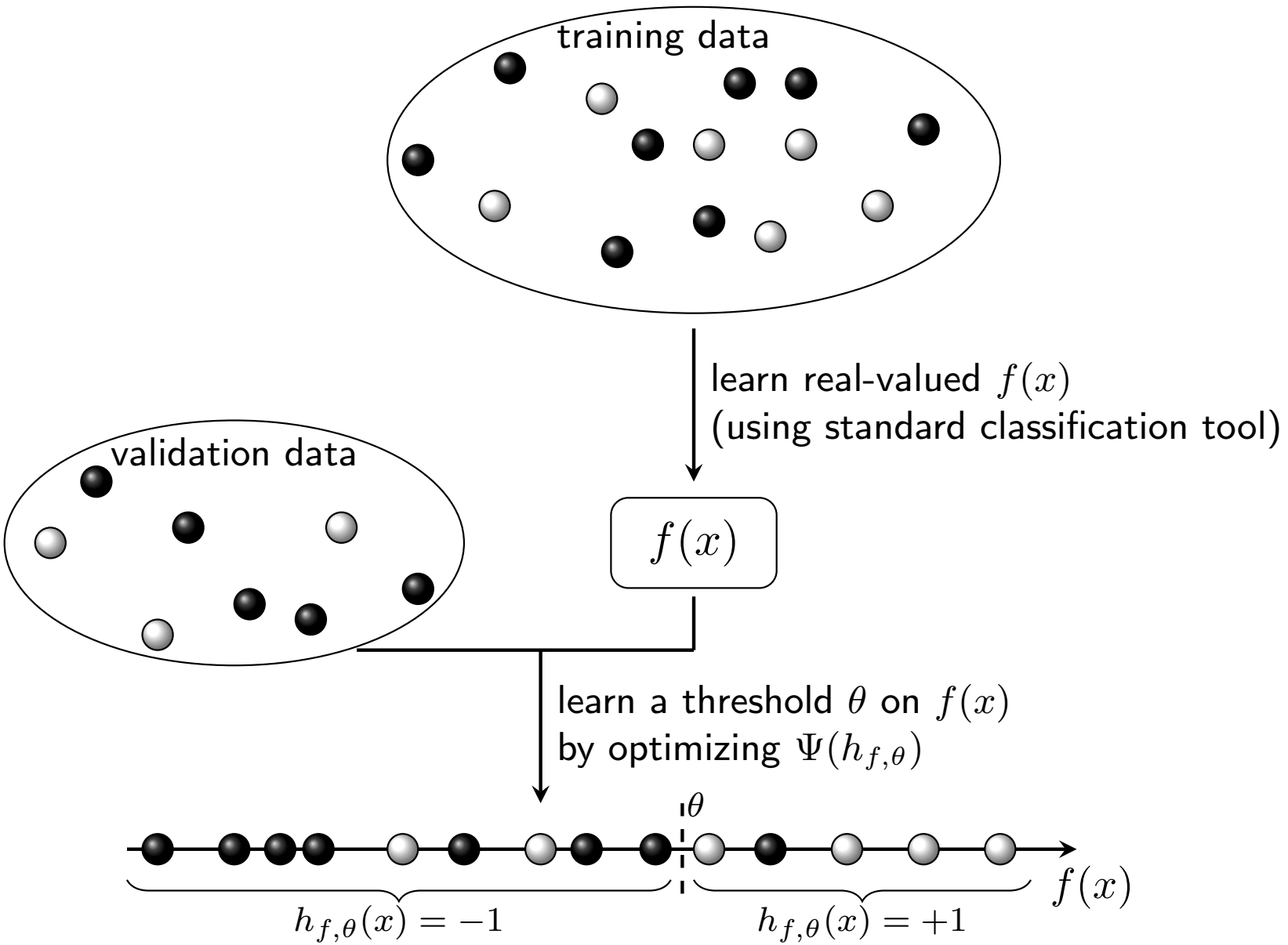
A simple approach to optimization of $\Psi(h)$



A simple approach to optimization of $\Psi(h)$

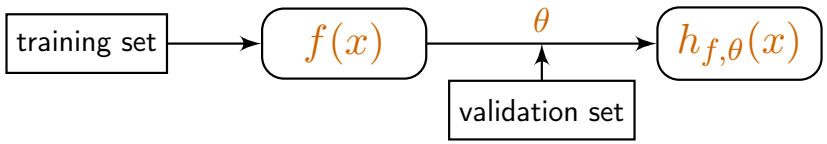


A simple approach to optimization of $\Psi(h)$



Our results

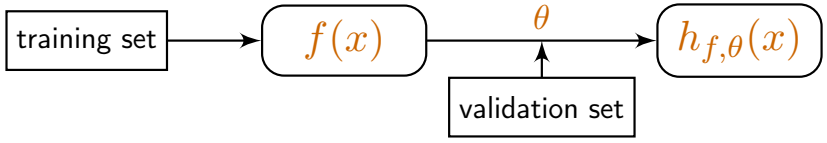
Algorithm



1. Learn f minimizing a surrogate loss on the training sample.
2. Given f, tune a threshold θ on f on a the validation sample by direct optimization of Ψ.

Our results

Algorithm



1. Learn f minimizing a surrogate loss on the training sample.
2. Given f, tune a threshold θ on f on a the validation sample by direct optimization of Ψ.

Our results (informally)

Assumptions:

- the surrogate loss is strongly proper composite (e.g., logistic, exponential, squared-error loss),
- Ψ is linear-fractional or jointly convex,

Claim:

- If f is close to the minimizer of the surrogate loss, then $h_{f,\theta}$ is close to the maximizer of Ψ.

Ψ-regret and ℓ-regret

- Ψ-regret of a classifier $h\colon X \to \{-1, 1\}$:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*) - \Psi(h) \qquad \text{where} \quad h^* = \operatorname*{argmax}_h \Psi(h).$$

Measures suboptimality.

Ψ-regret and ℓ-regret

- Ψ-regret of a classifier $h\colon X \to \{-1, 1\}$:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*) - \Psi(h) \qquad \text{where} \quad h^* = \operatorname*{argmax}_h \Psi(h).$$

  Measures suboptimality.

- Surrogate loss $\ell(y, f(x))$ of a real-valued function $f\colon X \to \mathbb{R}$.
  Used in training: logistic loss, squared loss, hinge loss, ...

Ψ-regret and ℓ-regret

- Ψ-regret of a classifier $h\colon X \to \{-1, 1\}$:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*) - \Psi(h) \qquad \text{where} \quad h^* = \operatorname*{argmax}_h \Psi(h).$$

  Measures suboptimality.

- Surrogate loss $\ell(y, f(x))$ of a real-valued function $f\colon X \to \mathbb{R}$.
  Used in training: logistic loss, squared loss, hinge loss, ...

- Expected loss (ℓ-risk) of f:

$$\mathrm{Risk}_\ell(f) = \mathbb{E}_{(x,y)}\left[\ell(y, f(x))\right].$$

Ψ-regret and ℓ-regret

- Ψ-regret of a classifier $h\colon X \to \{-1, 1\}$:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*) - \Psi(h) \qquad \text{where} \quad h^* = \operatorname*{argmax}_h \Psi(h).$$

  Measures suboptimality.

- Surrogate loss $\ell(y, f(x))$ of a real-valued function $f\colon X \to \mathbb{R}$.
  Used in training: logistic loss, squared loss, hinge loss, ...

- Expected loss (ℓ-risk) of f:

$$\mathrm{Risk}_\ell(f) = \mathbb{E}_{(x,y)}\left[\ell(y, f(x))\right].$$

- ℓ-regret of f:

$$\mathrm{Reg}_\ell(f) = \mathrm{Risk}_\ell(f) - \mathrm{Risk}_\ell(f^*) \qquad \text{where} \quad f^* = \operatorname*{argmin}_f \mathrm{Risk}_\ell(f).$$

Ψ-regret and ℓ-regret

- Ψ-regret of a classifier $h\colon X \to \{-1, 1\}$:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*) - \Psi(h) \qquad \text{where} \quad h^* = \operatorname*{argmax}_h \Psi(h).$$

  Measures suboptimality.

- Surrogate loss $\ell(y, f(x))$ of a real-valued function $f\colon X \to \mathbb{R}$.
  Used in tra... ...

- Expected lo...

$$\mathrm{Risk}_\ell(f) = \mathbb{E}_{(x,y)}\left[\ell(y, f(x))\right].$$

Relate Ψ-regret of $h_{f,\theta}$ to ℓ-regret of f

- ℓ-regret of f:

$$\mathrm{Reg}_\ell(f) = \mathrm{Risk}_\ell(f) - \mathrm{Risk}_\ell(f^*) \qquad \text{where} \quad f^* = \operatorname*{argmin}_f \mathrm{Risk}_\ell(f).$$

Examples of surrogate losses

Logistic loss

$$\ell(y, \widehat{y}) = \log\left(1 + e^{-y\widehat{y}}\right).$$

Risk minimizer $f^*(x)$:

$$f^*(x) = \log \frac{\eta(x)}{1 - \eta(x)}, \qquad \text{where} \quad \eta(x) = \Pr(y = 1|x).$$

Invertible function of conditional probability $\eta(x)$.

Examples of surrogate losses

Logistic loss

$$\ell(y, \widehat{y}) = \log\left(1 + e^{-y\widehat{y}}\right).$$

Risk minimizer $f^*(x)$:

$$f^*(x) = \log \frac{\eta(x)}{1 - \eta(x)}, \qquad \text{where } \eta(x) = \Pr(y = 1|x).$$

Invertible function of conditional probability $\eta(x)$.

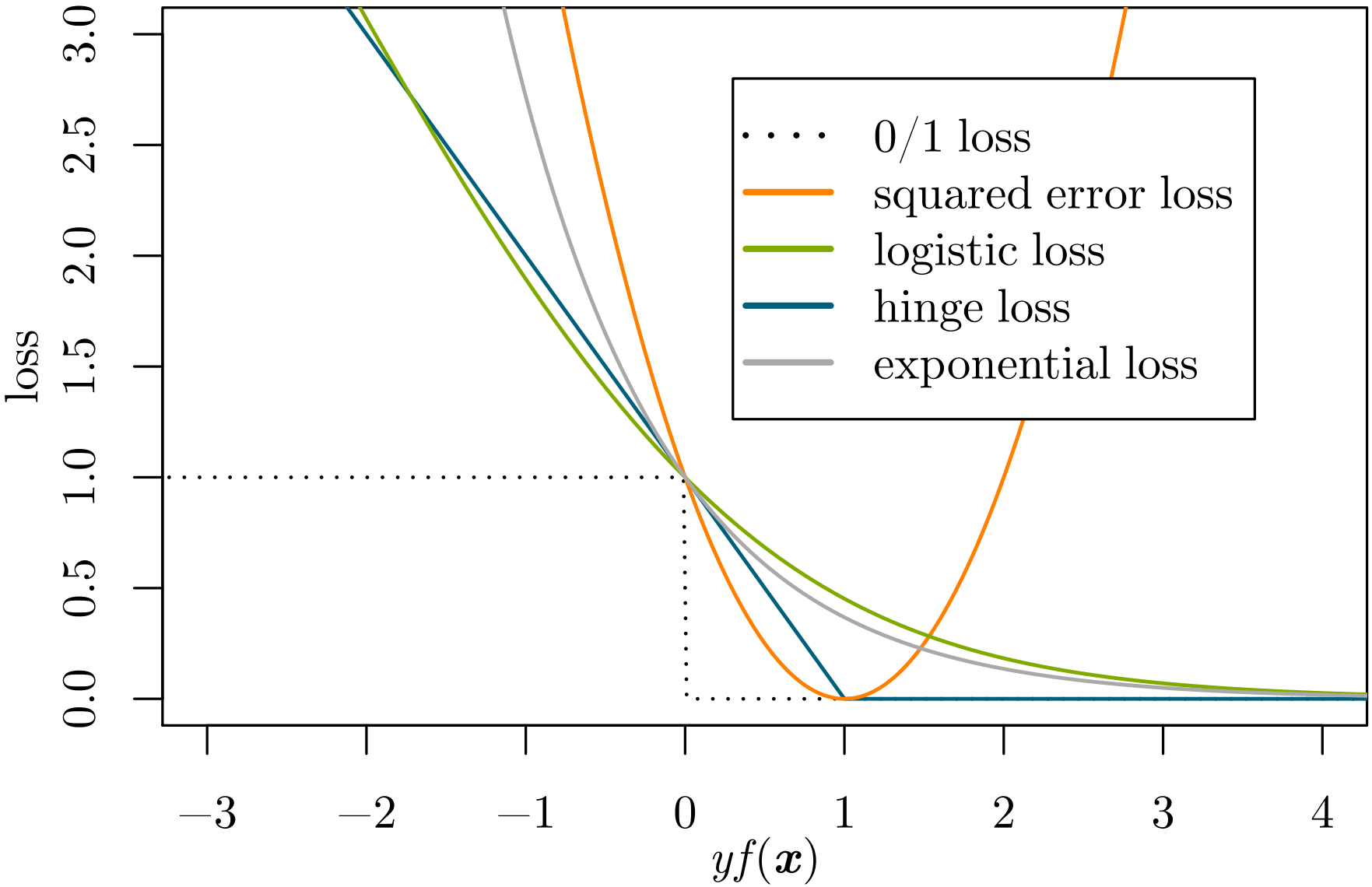
Hinge loss

$$\ell(y, \widehat{y}) = (1 - y\widehat{y})_+ .$$

Its risk minimizer $f^*(x)$ is non-invertible:

$$f^*(x) = \operatorname{sgn}(\eta(x) - 1/2).$$

Examples of surrogate losses



Examples of surrogate losses

loss	$f^*(\eta) = \psi(\eta)$	$\eta(f^*) = \psi^{-1}(f^*)$
squared error	$2\eta - 1$	$\frac{1+f^*}{2}$
logistic	$\log \frac{\eta}{1-\eta}$	$\frac{1}{1+e^{-f^*}}$
exponential	$\frac{1}{2} \log \frac{\eta}{1-\eta}$	$\frac{1}{1+e^{-2f^*}}$
hinge	$\operatorname{sgn}(\eta - 1/2)$	doesn't exist

Proper composite losses

$\ell(y, f)$ is proper composite if there exists a strictly increasing link function ψ, such that:

$$f^*(x) = \psi(\eta(x)), \qquad \text{where} \quad \eta(x) = \Pr(y = 1|x).$$

Minimizing proper composite losses implies probability estimation.

Strongly proper composite losses [Agarwal, 2014]

$\ell(y, f)$ is λ-strongly proper composite if it is proper composite and for any f, x, and distribution $y|x$:

$$\mathbb{E}_{y|x}\left[\ell(y, f(x)) - \ell(y, f^*(x))\right] \geq \frac{\lambda}{2}\left(\eta(x) - \psi^{-1}(f(x))\right)^2.$$

Technical condition.

Strongly proper composite losses: Examples

loss	$f^*(\eta) = \psi(\eta)$	$\eta(f^*) = \psi^{-1}(f^*)$	λ
squared error	$2\eta - 1$	$\frac{1+f^*}{2}$	8
logistic	$\log \frac{\eta}{1-\eta}$	$\frac{1}{1+e^{-f^*}}$	4
exponential	$\frac{1}{2} \log \frac{\eta}{1-\eta}$	$\frac{1}{1+e^{-2f^*}}$	4

Main result

Theorem for linear fractional measures

If:

- $\Psi(\mathrm{FP}, \mathrm{FN})$ is linear-fractional, non-increasing in FP and FN,
- ℓ is λ-strongly proper composite,

Then, there exists a threshold θ^*, such that for any real-valued function f,

$$\mathrm{Reg}_{\Psi}(h_{f,\theta^*}) \leq C_{\Psi} \sqrt{\tfrac{2}{\lambda}} \sqrt{\mathrm{Reg}_{\ell}(f)}.$$

Main result

Theorem for linear fractional measures

If:

- $\Psi(\mathrm{FP}, \mathrm{FN})$ is linear-fractional, non-increasing in FP and FN,
- ℓ is λ-strongly proper composite,

Then, there exists a threshold θ^*, such that for any real-valued function f,

$$\mathrm{Reg}_\Psi(h_{f,\theta^*}) \leq C_\Psi \sqrt{\tfrac{2}{\lambda}} \sqrt{\mathrm{Reg}_\ell(f)}.$$

metric	C_Ψ
F_β-measure	$\frac{1+\beta^2}{\beta^2 P}$
Jaccard similarity	$\frac{J^*+1}{P}$
AM measure	$\frac{1}{2P(1-P)}$

Main result

Theorem for linear fractional measures

If:

- $\Psi(\mathrm{FP}, \mathrm{FN})$ is linear-fractional, non-increasing in FP and FN,
- ℓ is λ-strongly proper composite,

Then, the alued function f

Similar theorem for convex performance metrics (such as AMS[2])

$$\mathrm{Reg}_\Psi(h_{f,\theta^*}) \leq C_\Psi \sqrt{\tfrac{2}{\lambda}} \sqrt{\mathrm{Reg}_\ell(f)}.$$

metric	C_Ψ
F_β-measure	$\frac{1+\beta^2}{\beta^2 P}$
Jaccard similarity	$\frac{J^*+1}{P}$
AM measure	$\frac{1}{2P(1-P)}$

Explanation of the theorem

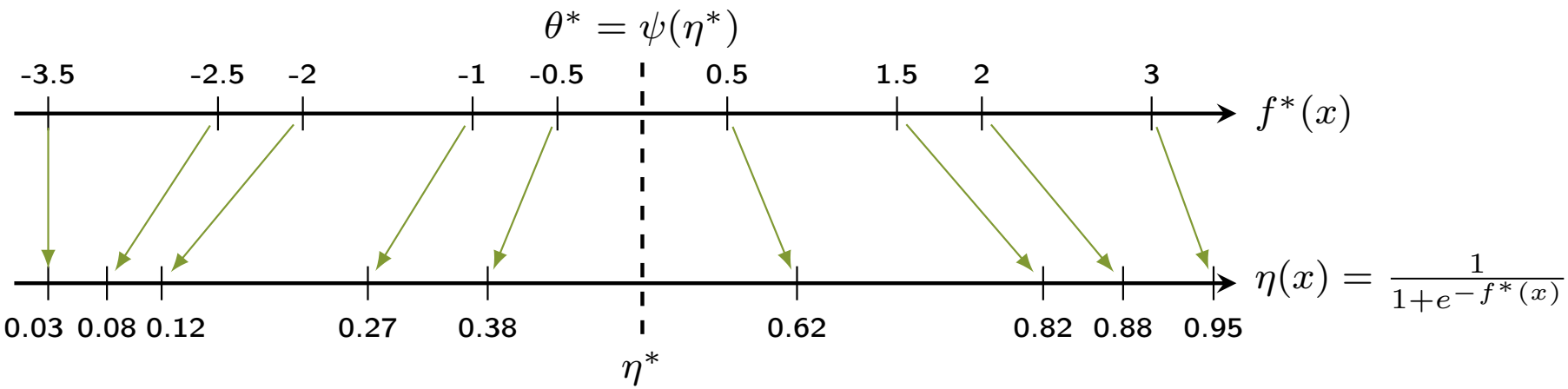
- The maximizer of Ψ is $h^*(x) = \operatorname{sgn}(\eta(x) - \eta^*)$ for some η^*.

Explanation of the theorem

- The maximizer of Ψ is $h^*(x) = \operatorname{sgn}(\eta(x) - \eta^*)$ for some η^*.
- The minimizer of ℓ is $f^*(x) = \psi(\eta(x))$ for invertible ψ.

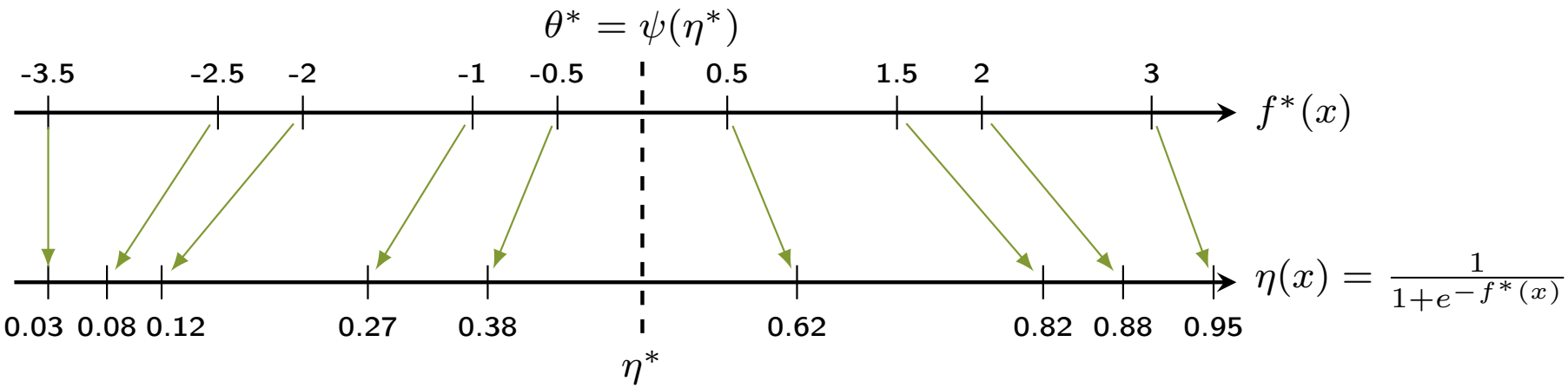
Explanation of the theorem

- The maximizer of Ψ is $h^*(x) = \operatorname{sgn}(\eta(x) - \eta^*)$ for some η^*.
- The minimizer of ℓ is $f^*(x) = \psi(\eta(x))$ for invertible ψ.
- Thresholding $f^*(x)$ at $\theta^* = \psi(\eta^*) \equiv$ Thresholding $\eta(x)$ at η^*.



Explanation of the theorem

- The maximizer of Ψ is $h^*(x) = \operatorname{sgn}(\eta(x) - \eta^*)$ for some η^*.
- The minimizer of ℓ is $f^*(x) = \psi(\eta(x))$ for invertible ψ.
- Thresholding $f^*(x)$ at $\theta^* = \psi(\eta^*) \equiv$ Thresholding $\eta(x)$ at η^*.



- Gradients of Ψ and λ measure local variations of Ψ and ℓ when f is not equal to f^*.

The optimal threshold θ^*

- $\Psi =$ classification accuracy $\implies$ $\theta^* = 0$ $(\eta^* = \frac{1}{2})$.

The optimal threshold θ^*

- $\Psi =$ classification accuracy $\implies$ $\theta^* = 0$ $(\eta^* = \frac{1}{2})$.
- $\Psi =$ weighted accuracy $\implies$ $\eta^* = \frac{w_-}{w_+ + w_-}$.

The optimal threshold θ^*

- Ψ = classification accuracy $\implies$ $\theta^* = 0$ ($\eta^* = \frac{1}{2}$).
- Ψ = weighted accuracy $\implies$ $\eta^* = \frac{w_-}{w_+ + w_-}$.
- More complex Ψ $\implies$ θ^* unknown (depends on Ψ^*)...

The optimal threshold θ^*

- Ψ = classification accuracy $\implies$ $\theta^* = 0$ ($\eta^* = \frac{1}{2}$).
- Ψ = weighted accuracy $\implies$ $\eta^* = \frac{w_-}{w_+ + w_-}$.
- More complex Ψ $\implies$ θ^* unknown (depends on Ψ^*)...

$\implies$ estimate θ^* from validation data

Tuning the threshold

Corollary

Given real-valued function f, validation sample of size m, let:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \widehat{\Psi}(h_{f,\theta}) \qquad (\widehat{\Psi} = \text{estimate of } \Psi)$$

Then, under the same assumptions and notation:

$$\mathrm{Reg}_{\Psi}(h_{f,\hat{\theta}}) \leq C_{\Psi} \sqrt{\tfrac{2}{\lambda}} \sqrt{\mathrm{Reg}_{\ell}(f)} + O\left(\tfrac{1}{\sqrt{m}}\right).$$

Tuning the threshold

Corollary

Given real-valued function f, validation sample of size m, let:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \widehat{\Psi}(h_{f,\theta}) \qquad (\widehat{\Psi} = \text{estimate of } \Psi)$$

Then, under the same assumptions and notation:

$$\mathrm{Reg}_{\Psi}(h_{f,\hat{\theta}}) \leq C_{\Psi} \sqrt{\tfrac{2}{\lambda}} \sqrt{\mathrm{Reg}_{\ell}(f)} + O\left(\tfrac{1}{\sqrt{m}}\right).$$

- Learning standard binary classifier and tuning the threshold afterwards is able to recover the maximizer of Ψ in the limit.

Multilabel classification

- A vector of m labels $\boldsymbol{y} = (y_1, \ldots, y_m)$ for each x.
- Multilabel classifier $\boldsymbol{h}(x) = (h_1(x), \ldots, h_m(x))$.
- False positive/negative rates for each label:

$$\mathrm{FP}_i(h_i) = \Pr(h_i = 1, y_i = -1),$$
$$\mathrm{FN}_i(h_i) = \Pr(h_i = -1, y_i = 1).$$

Multilabel classification

- A vector of m labels $\boldsymbol{y} = (y_1, \ldots, y_m)$ for each x.
- Multilabel classifier $\boldsymbol{h}(x) = (h_1(x), \ldots, h_m(x))$.
- False positive/negative rates for each label:

$$\mathrm{FP}_i(h_i) = \Pr(h_i = 1, y_i = -1),$$
$$\mathrm{FN}_i(h_i) = \Pr(h_i = -1, y_i = 1).$$

- How to use binary classification Ψ in the multilabel setting?
- We extend our bounds to cover micro- and macro-averaging.

Micro- and macro-averaging

Macro-averaging

Average outside Ψ:

$$\Psi_{\text{macro}}(\boldsymbol{h}) = \sum_{i=1}^{m} \Psi\big(\text{FP}_i(h_i), \text{FN}_i(h_i)\big).$$

Our bound suggests that a separate threshold needs to be tuned for each label.

Micro- and macro-averaging

Macro-averaging

Average outside Ψ:

$$\Psi_{\text{macro}}(\boldsymbol{h}) = \sum_{i=1}^{m} \Psi\big(\text{FP}_i(h_i), \text{FN}_i(h_i)\big).$$

Our bound suggests that a separate threshold needs to be tuned for each label.

Micro-averaging

Average inside Ψ:

$$\Psi_{\text{micro}}(\boldsymbol{h}) = \Psi\left(\frac{1}{m}\sum_{i=1}^{m} \text{FP}_i(h_i), \frac{1}{m}\sum_{i=1}^{m} \text{FN}_i(h_i)\right).$$

Our bound suggests that all labels share a single threshold.

Experiments

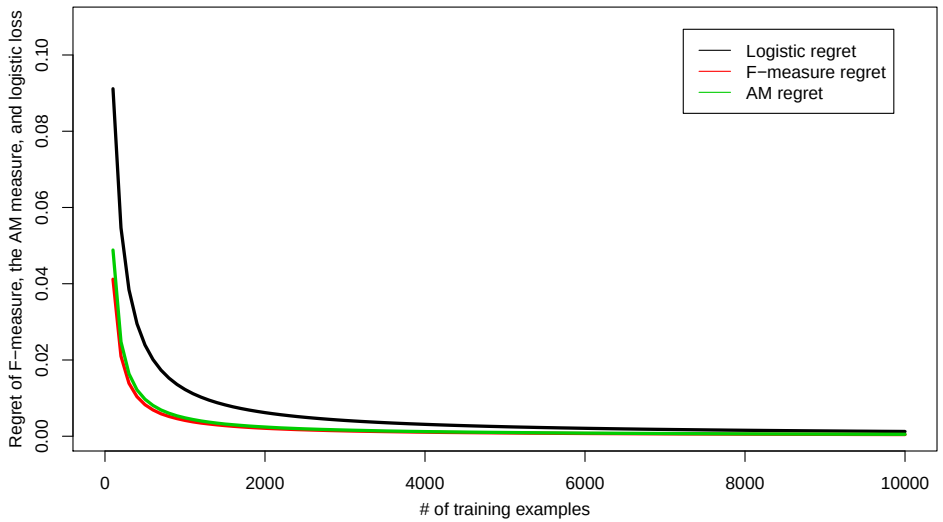
Experimental results

- Two synthetic and two benchmark data sets.
- Surrogates: Logistic loss (LR) and hinge loss (SVM).
- Performance metrics: F-measure and AM measure.
- Minimize ℓ (logistic or hinge) on training data + tune the threshold $\hat{\theta}$ on validation data (optimizing the metrics).

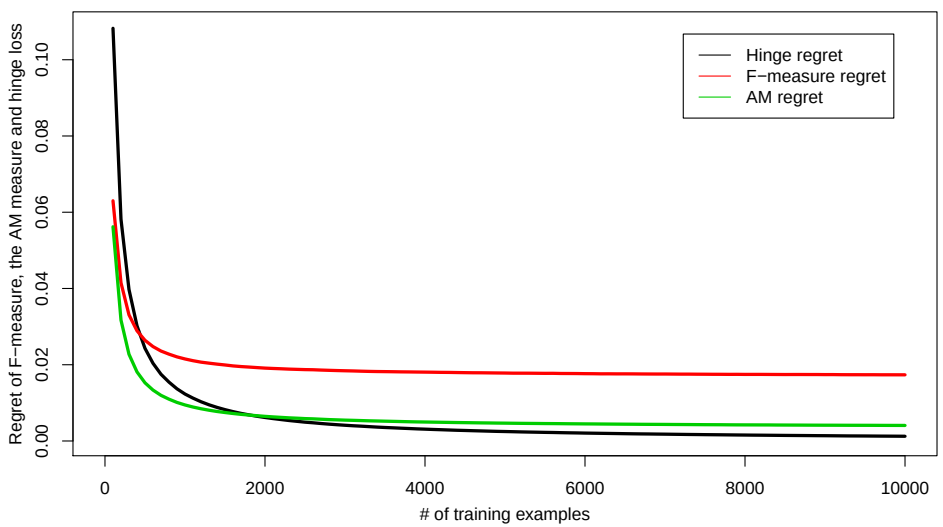
Synthetic experiment I

- $X = \{1, 2, \ldots, 25\}$ with $\Pr(x)$ uniform.
- For each $x \in X$, $\eta(x) \sim \text{Unif}[0, 1]$.

logistic loss surrogate
(converges as expected)



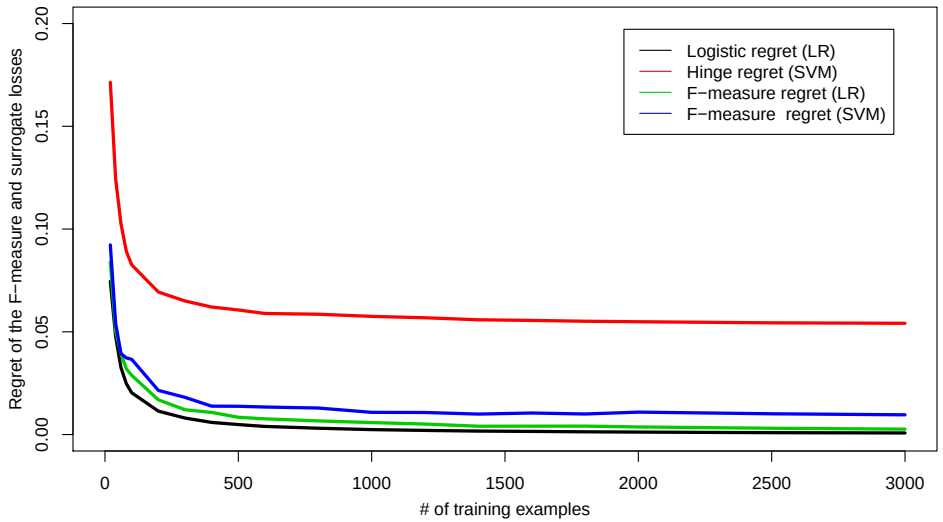
hinge loss surrogate
(does not converge)



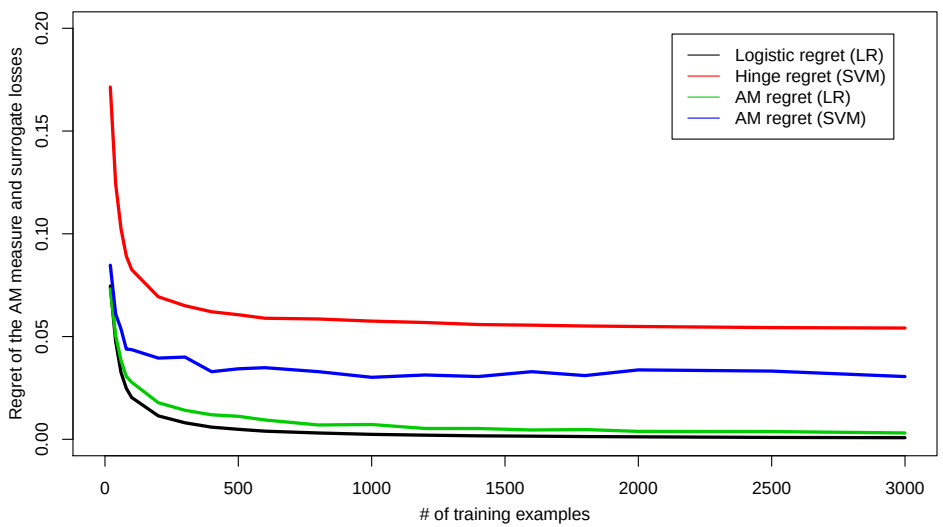
Synthetic experiment II

- $X = \mathbb{R}^2$, $x \sim N(0, I)$.
- Logistic model: $\eta(x) = (1 + \exp(-a_0 - a^\top x))^{-1}$.

Convergence of F-measure

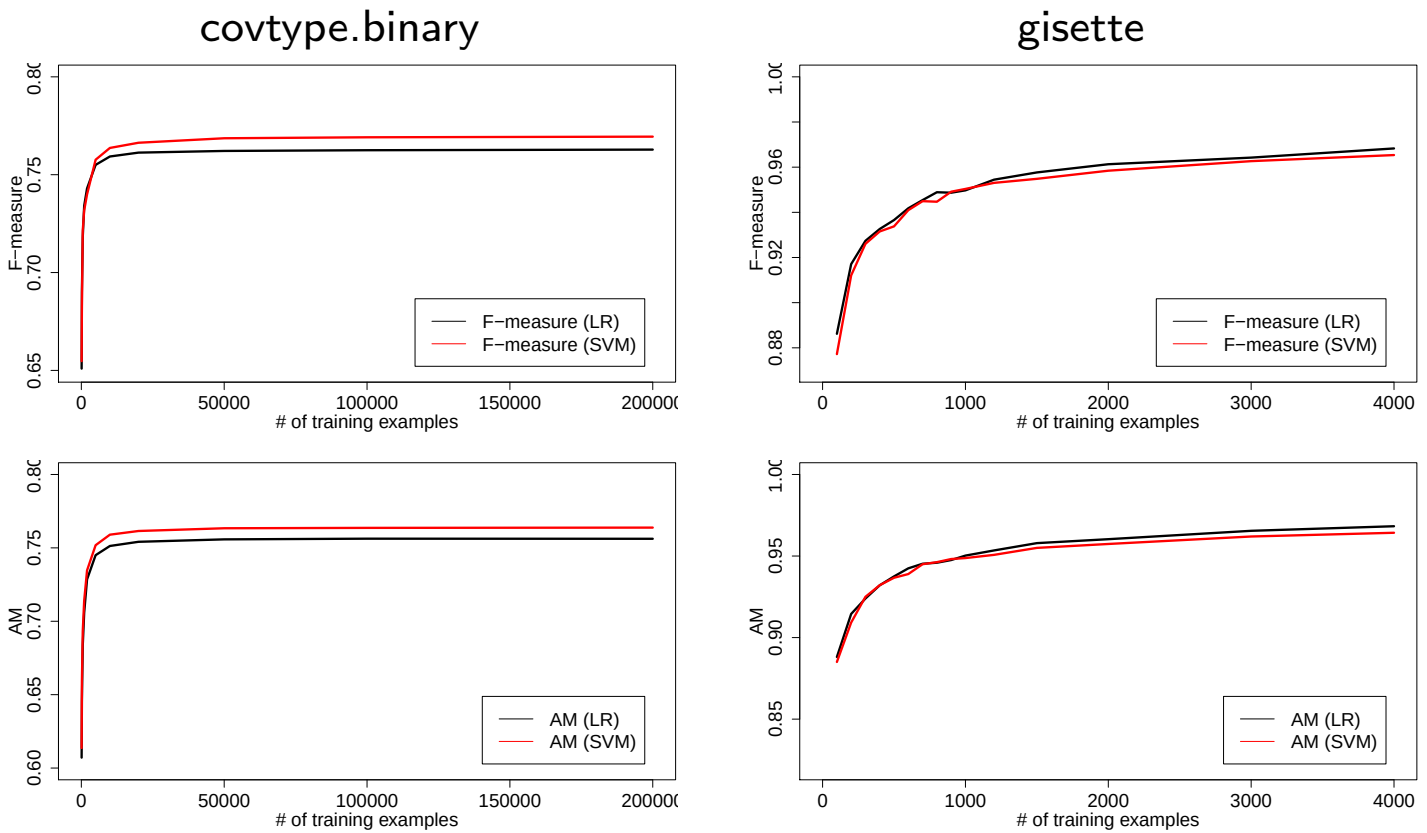


Convergence of AM measure



Benchmark data experiment – a bit of surprise

dataset	#examples	#features
covtype.binary	581,012	54
gisette	7,000	5,000



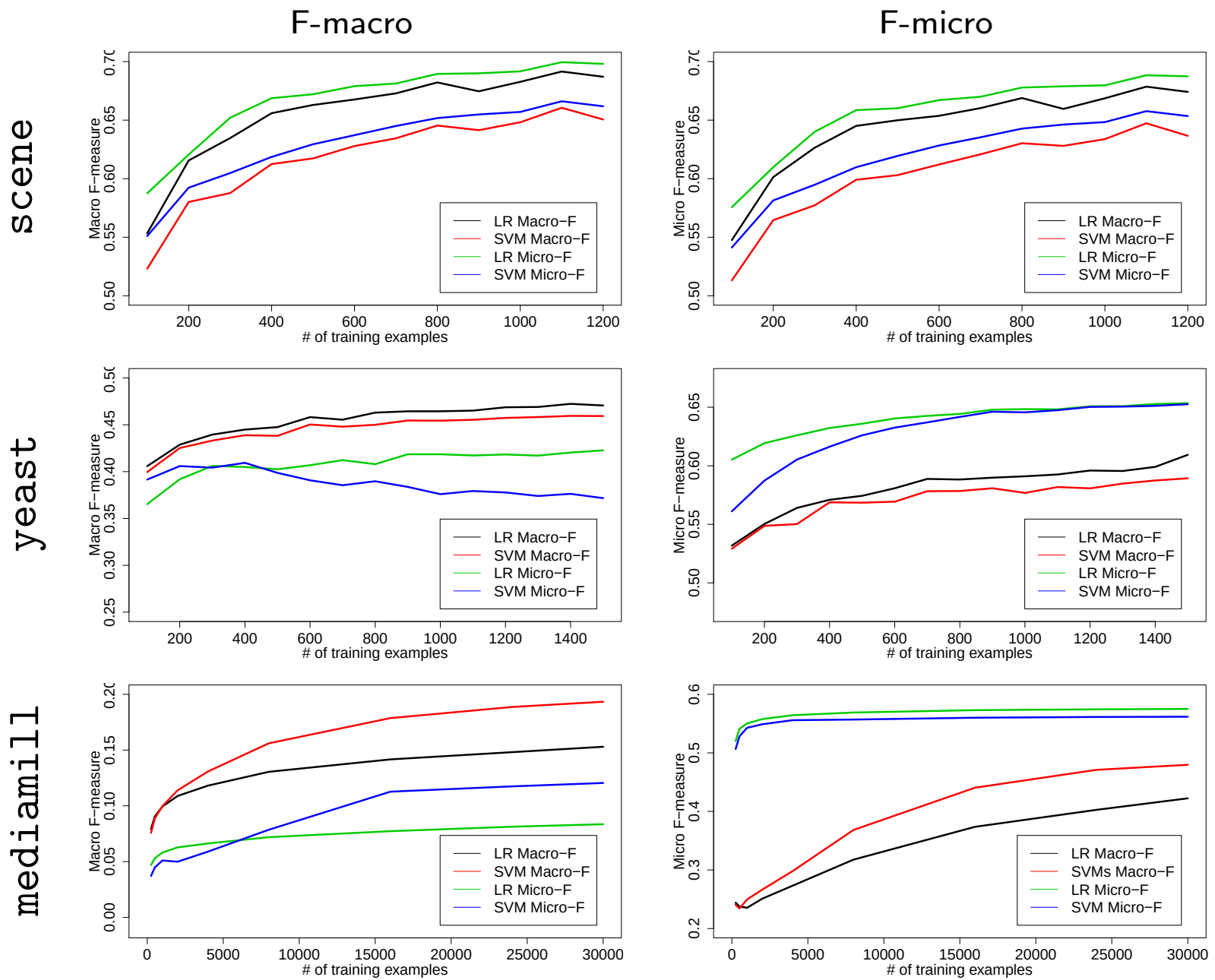
Logistic loss as expected, but also hinge loss surprisingly well.

Multilabel classification

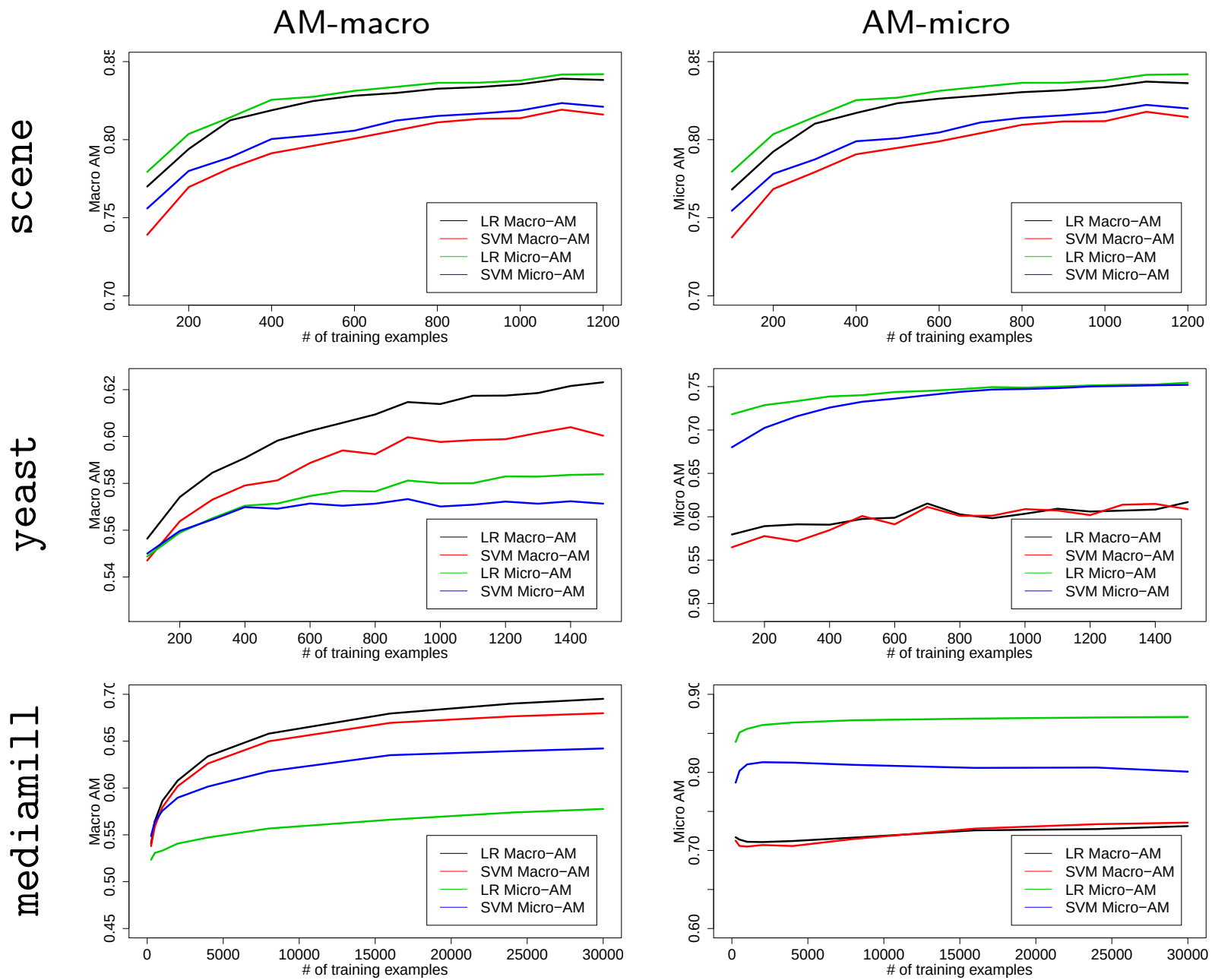
data set	# labels	# training examples	# test examples	#features
`scene`	6	1211	1169	294
`yeast`	14	1500	917	103
`mediamill`	101	30993	12914	120

- Surrogates: Logistic loss (LR) and hinge loss (SVM).
- Performance metrics: F-measure and AM measure.
- Macro-averaging (separate threshold for each label) and micro-averaging (single threshold for all labels).
- Cross-evaluation of algorithms tuned for micro-averaging in terms of macro-averaged metrics, and vice versa.

Multilabel classification: F measure



Multilabel classification: AM measure



Open problem

$$\Psi(h^*) - \Psi(h) \quad \leq \quad \text{const} \cdot \sqrt{\text{Risk}_\ell(f) - \text{Risk}_\ell(f^*)}$$

f^* is the minimizer over **all** functions

Open problem

$$\Psi(h^*) - \Psi(h) \quad \leq \quad \text{const} \cdot \sqrt{\text{Risk}_\ell(f) - \text{Risk}_\ell(f^*)}$$

f^* is the minimizer over **all** functions

- We often optimize within some class $\mathcal{F}$ (e.g., linear functions).
- If $f^* \notin \mathcal{F} \Rightarrow$ r.h.s. does not converge to 0.
- This can be beneficial non non-proper losses (e.g., hinge loss).
- Most often $f^* \notin \mathcal{F}$, and a theory for this is needed.

Summary

- Theoretical analysis of the two-step approach to optimize generalized performance metrics for classification.
- Regret bounds for linear-fractional and convex functions optimized by means of strongly proper composite surrogates.
- The theorem relates convergence to the global risk minimizer.
  - Can we say anything about convergence to the risk minimizer within some class of functions?
- Why does hinge loss perform so well if the risk minimizer is outside the family of classification functions.