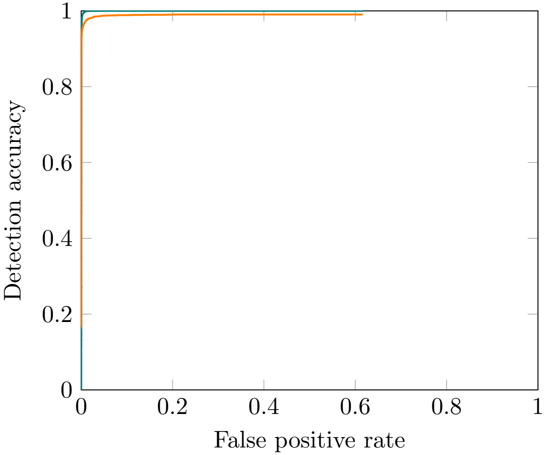
Towards dependable steganalysis

Tomáš Pevný[a,c], Andrew D. Ker[b]

[a]Cisco systems, Inc., Cognitive Research Team in Prague, CZ
[b]Department of Computer Science, University of Oxford, UK
[c]Department of Computers, CVUT in Prague, CZ

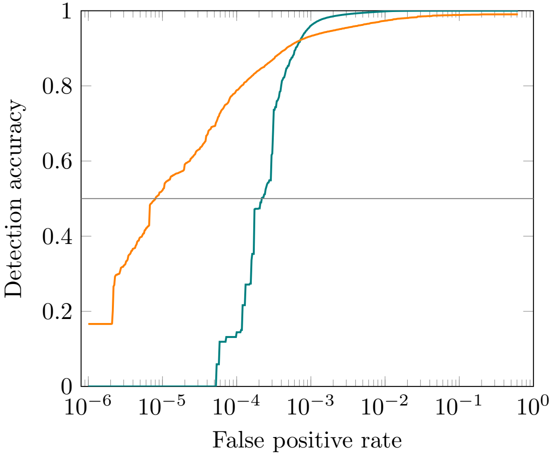
10th February 2015
SPIE/IS&T Electronic Imaging

Motivation



1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Detection accuracy
False positive rate

Motivation



Detection accuracy
False positive rate
0
0.2
0.4
0.6
0.8
1
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}

Millions of images

- In 2014, Yahoo! released 100 million CC Flickr images.
- Selected images with quality factor 80 and known camera, split into two sets:

Training & validation	449 395 cover	449 395 stego	from 4781 users
Testing	4 062 128 cover	407 417 stego	from 43026 users

- Stego images: nsF5 at 0.5 bits per nonzero coefficient.
- JRM features computed from every image.

Motivation

What is a good benchmark?

- Equal prior error rate?
- Emphasizing false positives?

Our error measure (FP-50)

False positive rate at 50% detection accuracy.

Motivation

What is a good benchmark?

- Equal prior error rate?
- Emphasizing false positives?

Our error measure (FP-50)

False positive rate at 50% detection accuracy.

Mathematical formulation

Exact optimization criterion

$$\arg\min_{f \in \mathscr{F}} \mathbb{E}_{x \sim \text{cover}} \left[I \left[f(x) > \text{median} \{ f(y) | y \sim \text{stego} \} \right] \right]$$

- $I(\cdot)$ is the indicator function
- $\mathscr{F}$ set of classifiers

Simplifications

- Restrict $\mathscr{F}$ to linear classifiers.
- $\arg\min_{f \in \mathscr{F}} \mathbb{E}_{x \sim \text{cover}} \left[I \left[f(x) > \mathbb{E}_{y \sim \text{stego}} [f(y)] \right] \right]$

Mathematical formulation

Exact optimization criterion

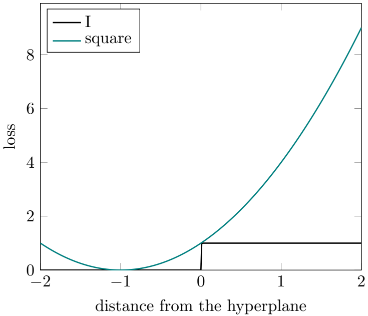
$$\arg\min_{f \in \mathscr{F}} \mathbb{E}_{x \sim \text{cover}} \left[I \left[f(x) > \text{median} \left\{ f(y) | y \sim \text{stego} \right\} \right] \right]$$

- $I(\cdot)$ is the indicator function
- $\mathscr{F}$ set of classifiers

Simplifications

- Restrict $\mathscr{F}$ to linear classifiers.
- $\arg\min_{f \in \mathscr{F}} \mathbb{E}_{x \sim \text{cover}} \left[I \left[f(x) > \mathbb{E}_{y \sim \text{stego}} \left[f(y) \right] \right] \right]$

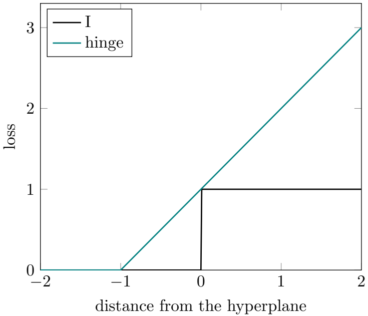
Approximation by square loss



optimization criterion

$$\arg\min_{w} \sum_{x\,\text{cover}} \left(w^{\mathrm{T}}(x-\bar{y})\right)^2 + \lambda \|w\|^2$$

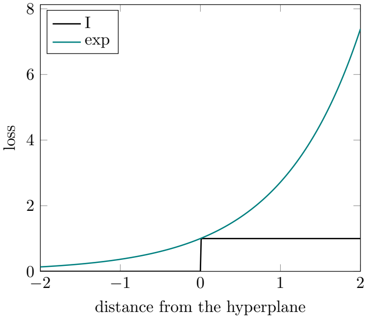
Approximation by hinge loss



optimization criterion

$$\arg\min_{w} \sum_{x\,\text{cover}} \max\left\{0, w^{\mathrm{T}}(x - \bar{y} - 1)\right\} + \lambda \|w\|^2$$

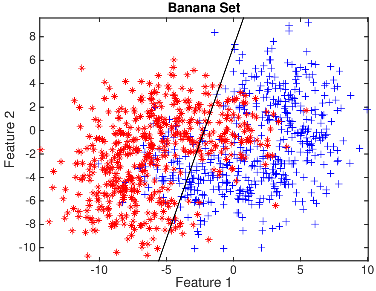
Approximation by exponential loss



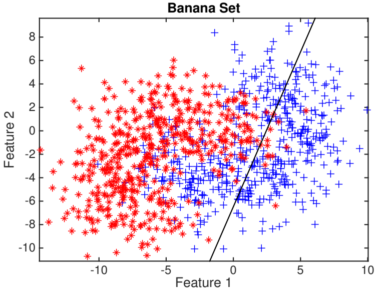
optimization criterion

$$\arg\min_{w} \sum_{x\,\text{cover}} e^{\left(w^{\mathrm{T}}(x-\bar{y})\right)} + \lambda \|w\|^2$$

Toy example



Fisher linear discriminant



Optimizing exponential loss

Linear classifiers on JRM features

- 22510 features
- 2 x 40 000 training images
- 2 x 250 000 validation images

FP-50	FLD	weighted SVM*	Square loss	Exponential loss
training set	$1.11 \cdot 10^{-4}$	$2.18 \cdot 10^{-5}$	$1.45 \cdot 10^{-5}$	0
validation set	$2.52 \cdot 10^{-4}$	$1.99 \cdot 10^{-4}$	$5.61 \cdot 10^{-4}$	$9.87 \cdot 10^{-4}$

$$^{*}\ \arg\min_{w} \eta \mathbb{E}_{x\sim\mathrm{cover}} \max\{0, w^{\mathrm{T}}x\} + (1-\eta)\mathbb{E}_{y\sim\mathrm{stego}} \max\{0, -w^{\mathrm{T}}y\} + \lambda \|w\|^2$$

Optimizing an ensemble

Ensembles based on random subspaces à la Kodovský:

- L base learners,
- Each trained on random d_{sub} features, and all data.

Two thresholds:

- base learner threshold: optimize equal prior accuracy
  - Neyman-Pearson criterion (identical FP rate)
- voting threshold: majority vote
  - arbitrary threshold

Optimizing an ensemble

Ensembles based on random subspaces à la Kodovský:

- L base learners,
- Each trained on random d_{sub} features, and all data.

Two thresholds:

- base learner threshold: optimize equal prior accuracy
  - Neyman-Pearson criterion (identical FP rate)
- voting threshold: majority vote
  - arbitrary threshold

Optimizing an ensemble

Ensembles based on random subspaces à la Kodovský:

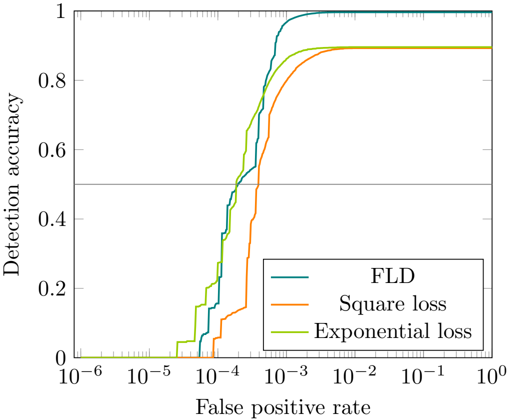
- L base learners,
- Each trained on random d_{sub} features, and all data.

Two thresholds:

- base learner threshold: optimize equal prior accuracy
  - Neyman-Pearson criterion (identical FP rate)
- voting threshold: majority vote
  - arbitrary threshold

ROC of ensembles

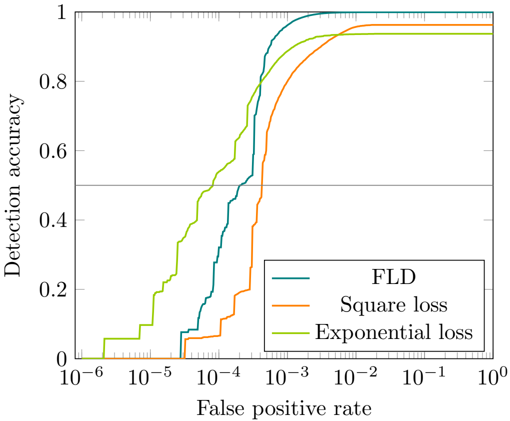
- 2 x 40 000 training images
- 2 x 250 000 validation images



$L = 300,\ d_{sub} = 1000$

ROC of ensembles

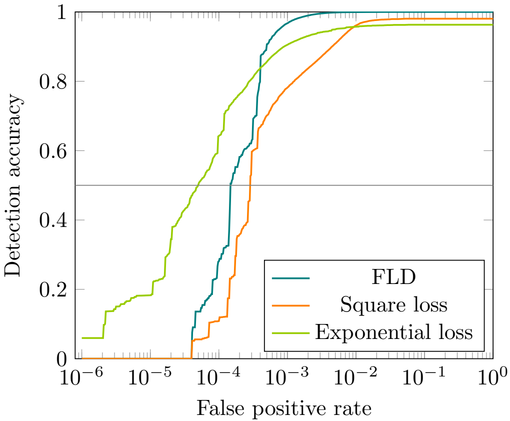
- 2 x 40 000 training images
- 2 x 250 000 validation images



$L = 300,\ d_{sub} = 500$

ROC of ensembles

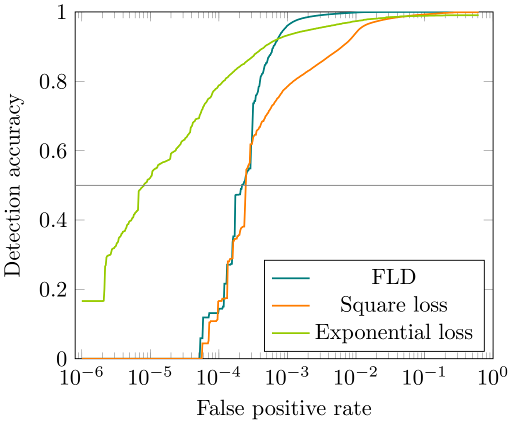
- 2 x 40 000 training images
- 2 x 250 000 validation images



$L = 300,\ d_{sub} = 250$

ROC of ensembles

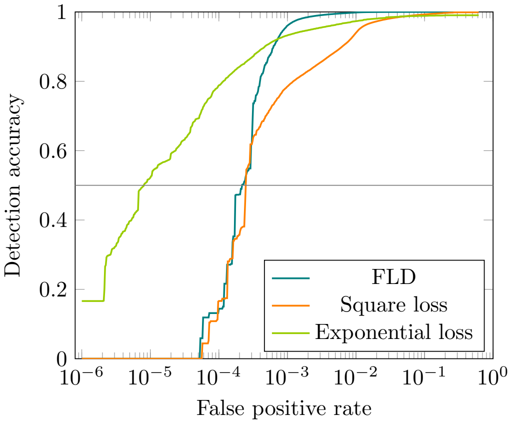
- 2 x 40 000 training images
- 2 x 250 000 validation images



$L = 300,\ d_{sub} = 100$

ROC of ensembles

- 4.5M image testing set:
- False negative rate 51.2%
- False positive rate $5.56 \cdot 10^{-5}$



$L = 300,\ d_{sub} = 100$

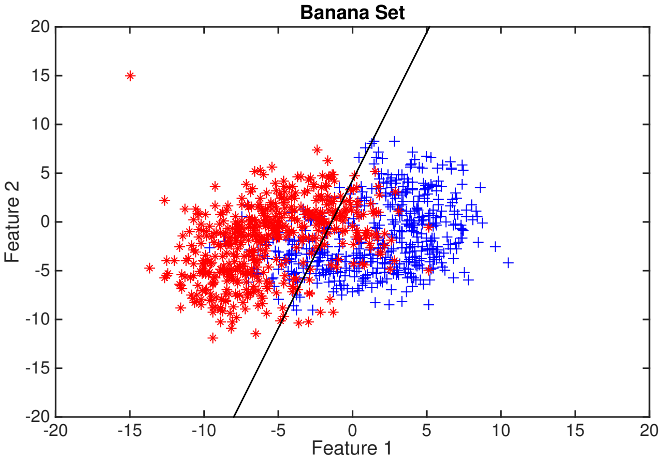
Errors on testing set

Base learner	Thresholds	False negative rate	False positive rate
FLD	Traditional	$1.33 \cdot 10^{-3}$	$9.07 \cdot 10^{-3}$
FLD	Proposed	$4.58 \cdot 10^{-1}$	$3.26 \cdot 10^{-4}$
Exponential loss	Proposed	$5.12 \cdot 10^{-1}$	$5.56 \cdot 10^{-5}$

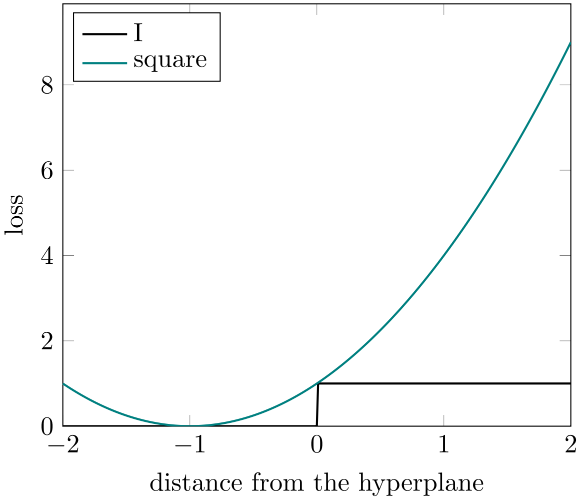
Summary

- Classifiers derived from the FP-50 measure.
  - Can derive same classifiers in two different ways.
- Various convex surrogates for step function:
  - Non-smooth loss is difficult to optimize.
  - Exponential loss encourages over-fitting.
  - Square loss (FLD) has a hidden weakness.
- Ensemble subdimension is an indirect regularizer.
- Ensemble thresholds need to be optimized differently.

Summary


Banana Set
Feature 2
Feature 1
20
15
10
5
0
-5
-10
-15
-20
-20
-15
-10
-5
0
5
10
15
20

Summary



Summary

- We detected lousy, very high-bit rate, steganography with 1 in 18000 false positive rate.