Similarity Learning for Pointwise ROC Optimization

Robin Vogel[1,3] Stephan Clémençon[1] Aurélien Bellet[2]

[1] Telecom ParisTech, [2] Inria, [3] IDEMIA

ICML 2018

Outline

Biometric verification (1/2)

A biometric system uses:

- Two **measurements** X and X',
- A **similarity** S that quantifies the likeness of (X, X'),
- A **threshold** t that separates positive and negative pairs.

$S(X, X') > t$

Aim: $S(X, X') > t$ is a good indicator of $Z = +1$ with:

$$Z = \begin{cases} +1 & \text{if } (X, X') \text{ from the same person,} \\ -1 & \text{otherwise.} \end{cases}$$

Two types of errors:

$$\mathrm{TPR}_S(t) := \mathbb{P}\{S(X, X') > t \mid Z = +1\},$$
$$\mathrm{FPR}_S(t) := \mathbb{P}\{S(X, X') > t \mid Z = -1\}.$$

The set $\{(FPR_S(t), TPR_S(t)) \mid t \in \mathbb{R}\}$ is known as the **ROC curve**.

Biometric verification (2/2)

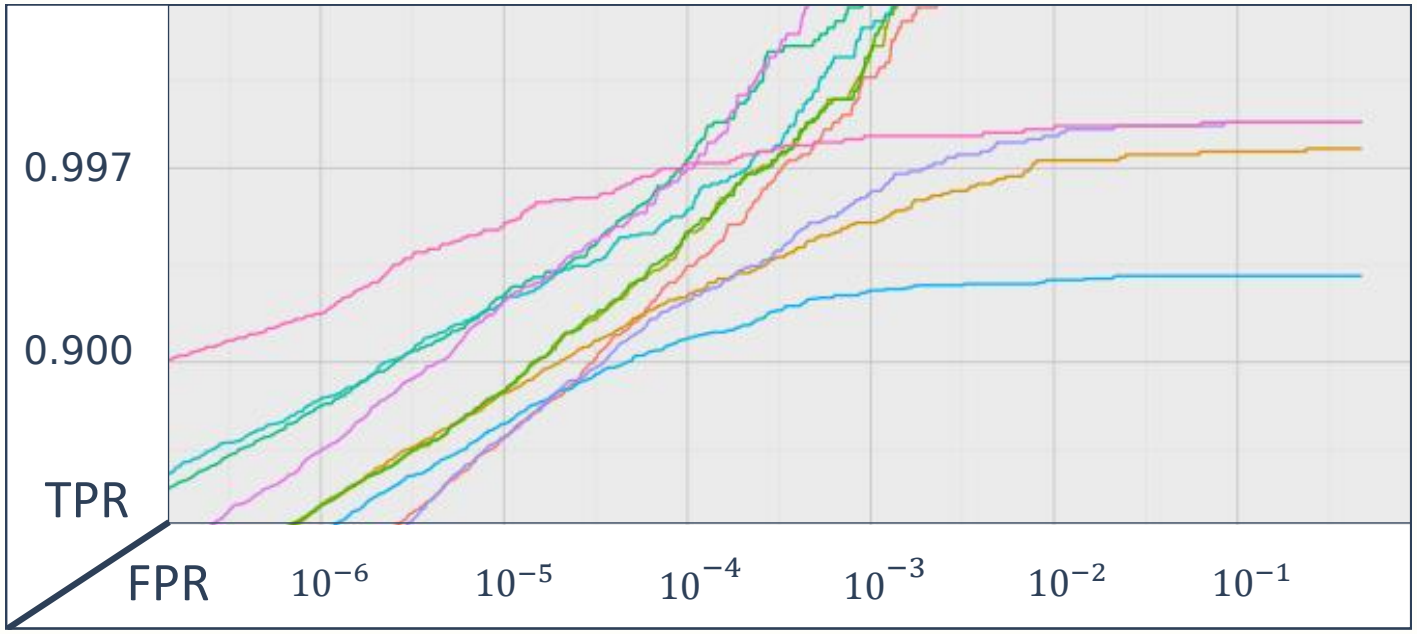


Figure: Example of a ROC curve (Source: NIST FRVT reports.)

Pointwise ROC optimization (**pROC**):

$$\text{Maximize TPR s.t. FPR} \leq \alpha,$$

with α some target, for example $\alpha = 10^{-3}$.

Objective:

Procedure to solve **pROC** from data, and theoretical guarantees.

Related work

[Clémençon and Vayatis, 2008]
$\rightarrow$ guarantees for ERM for **pROC** for **bipartite ranking**.

- **Bipartite ranking:** Rank $X_1, X_2, \ldots$ by relevance,
- **Similarity ranking:** Rank $(X_1, X_2), (X_1, X_3), \ldots$ by similarity.

Their analysis does not generalize to **similarity ranking**.
$\rightarrow$ Ours requires results on U-statistics, see [Lee, 1990] and [Clémençon et al., 2008].

Instead of a **constrained approach**, **asymmetric weighting** of positive and negative risks is common in metric learning.
$\rightarrow$ Theoretical guarantees exist, see [Cao et al., 2016].

However, we have no indication on the asymmetry factor γ that guarantees FPR $\leq \alpha$.

Our problem (1/3)

We introduce the following problem:

$$\max_{G \in \mathcal{G}} R^+(G) \text{ s.t. } R^-(G) \leq \alpha, \quad (1)$$

with $R^+(G) = \mathbb{E}\left[G(X, X') \mid Z = +1\right]$ and $R^-(G) = \mathbb{E}\left[G(X, X') \mid Z = -1\right]$.

Remark: When $G(X, X') = \mathbb{I}\{S(X, X') > t\}$, **eq. (1)** is **pROC**.

Our data $\mathcal{D}_n$ follows a **standard classification model**, i.e.

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

is composed of n i.i.d copies of $(X, Y) \sim P$ where $Y \in \{1, \ldots, K\}$.

Given $(X, Y) \perp (X', Y') \sim P$, it follows that:

$$Z = \begin{cases} +1 & Y = Y', \\ -1 & Y \neq Y'. \end{cases}$$

Our problem (2/3)

Theoretical problem:

$$\max_{G \in \mathcal{G}} R^+(G) \text{ s.t. } R^-(G) \leq \alpha, \quad (1)$$

with $R^+(G) = \mathbb{E}\left[G(X, X') \mid Z = +1\right]$ and $R^-(G) = \mathbb{E}\left[G(X, X') \mid Z = -1\right]$.

We can estimate the quantities $R^+(G)$ and $R^-(G)$ by:

$$R_n^+(G) := \frac{1}{n_+} \sum_{i<j} G(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\},$$

$$R_n^-(G) := \frac{1}{n_-} \sum_{i<j} G(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\},$$

where $n_+ = \sum_{i<j} \mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$.

We can now derive an **empirical problem** !

Our problem (3/3)

Theoretical problem:

$$\max_{G \in \mathcal{G}} R^+(G) \text{ s.t. } R^-(G) \leq \alpha, \quad (1)$$

with $R^+(G) = \mathbb{E}\left[G(X, X') \mid Z = +1\right]$ and $R^-(G) = \mathbb{E}\left[G(X, X') \mid Z = -1\right]$.

Empirical problem:

$$\max_{G \in \mathcal{G}} R_n^+(G) \text{ s.t. } R_n^-(G) \leq \alpha + \Phi. \quad (2)$$

with $\Phi > 0$.

Why a tolerance parameter Φ ?

- Tolerate variations of $R_n^-(G)$ around its expectation $R^-(G)$.
- Φ will depend on $\mathcal{G}$'s complexity and n.

Let G^* and $\hat{G}_n$ be respectively the solutions of eq. (1) and eq. (2).

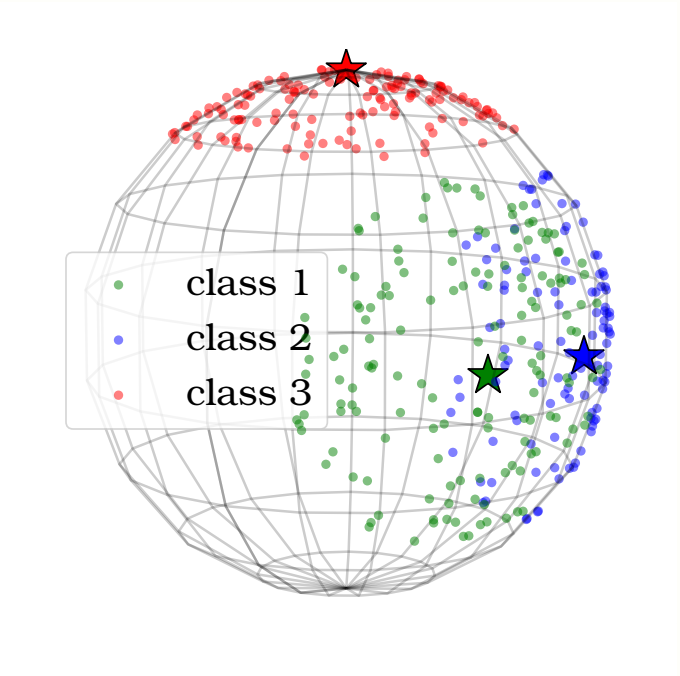
Illustration: **pROC** with linear $\mathcal{G}$

Setting:

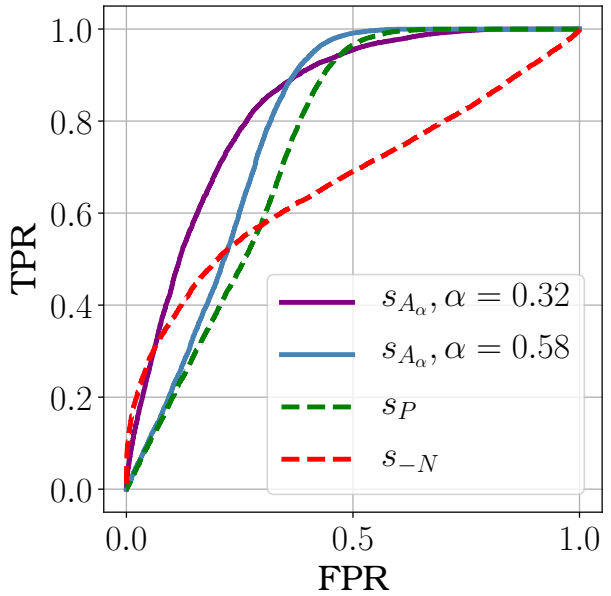
$$\mathcal{G} = \left\{ s_A : (x, x') \mapsto \frac{1}{2}\left(1 + x^\top A x'\right) \mid \|A\|_F^2 \leq 1 \right\},$$

where $\|A\|_F^2 = \sum_{i<j} a_{ij}^2$.

Easy analytical solutions of eq. (2) !



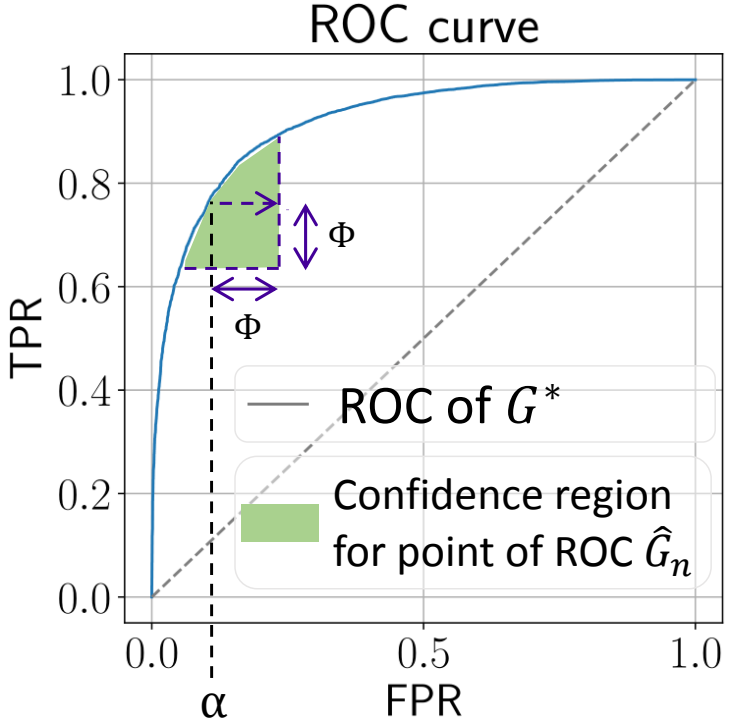
(a) Simulated data with simple structure.



(b) ROC curves for different α's.

Outline

Universal bound



Theorem 1

Suppose that:

1. $\mathcal{G}$ *is a "nice" family of functions,*
2. $\mathbb{P}\{Y = Y'\} \geq \kappa > 0$ *"away from zero".*

Introducing:

$$\Phi_{n,\delta} := C_{\delta,\mathcal{G}} \cdot n^{-1/2},$$

we have that, w.p. $\geq 1 - \delta$, $\forall n \geq n_0$,

$$R^+(\hat{G}_n) \geq R^+(G^*) - \Phi_{n,\delta},$$
$$R^-(\hat{G}_n) \leq \alpha + \Phi_{n,\delta}.$$

Sketch of proof

- Show that: Uniform control over $\mathcal{G}$ of $R_n^+ - R^+$, $R_n^- - R^-$

  $\Longrightarrow$ bilateral control of the risks.

- Use results on U-statistics.

Fast rates (1/2)

Under a noise assumption (**NA**),

The regret in R^+ is bounded by a higher power of n with same $\Phi_{n,\delta}$.

Introduce:

- the **posterior probability** η:

$$\eta(x, x') = \mathbb{P}\{Z = +1 \mid (X, X') = (x, x')\}.$$

- the value Q^*_α of the $(1-\alpha)$-**quantile** of $\eta(X, X') \mid Z = -1$.

NA is inspired by [Mammen and Tsybakov, 1995] and means that:

For almost every $x \in \mathcal{X}$, the CDF of $\eta(x, X')$ is smooth around Q^*_α.

Some parameter a quantifies the smoothness.

Fast rates (2/2)

Theorem 2

Suppose that assumptions of theorem 1 and ***NA*** *are verified, we have that, w.p.* $\geq 1-\delta$, $\forall n \geq n_0$,

$$R^+(\hat{G}_n) \geq R^+(G^*) - C_0 \cdot n^{-(2+a)/4},$$
$$R^-(\hat{G}_n) \leq \alpha + \Phi_{n,\delta}.$$

where C_0 *depends of* $\delta, \kappa, Q^*_\alpha, a, c$ *and* $\mathcal{G}$.

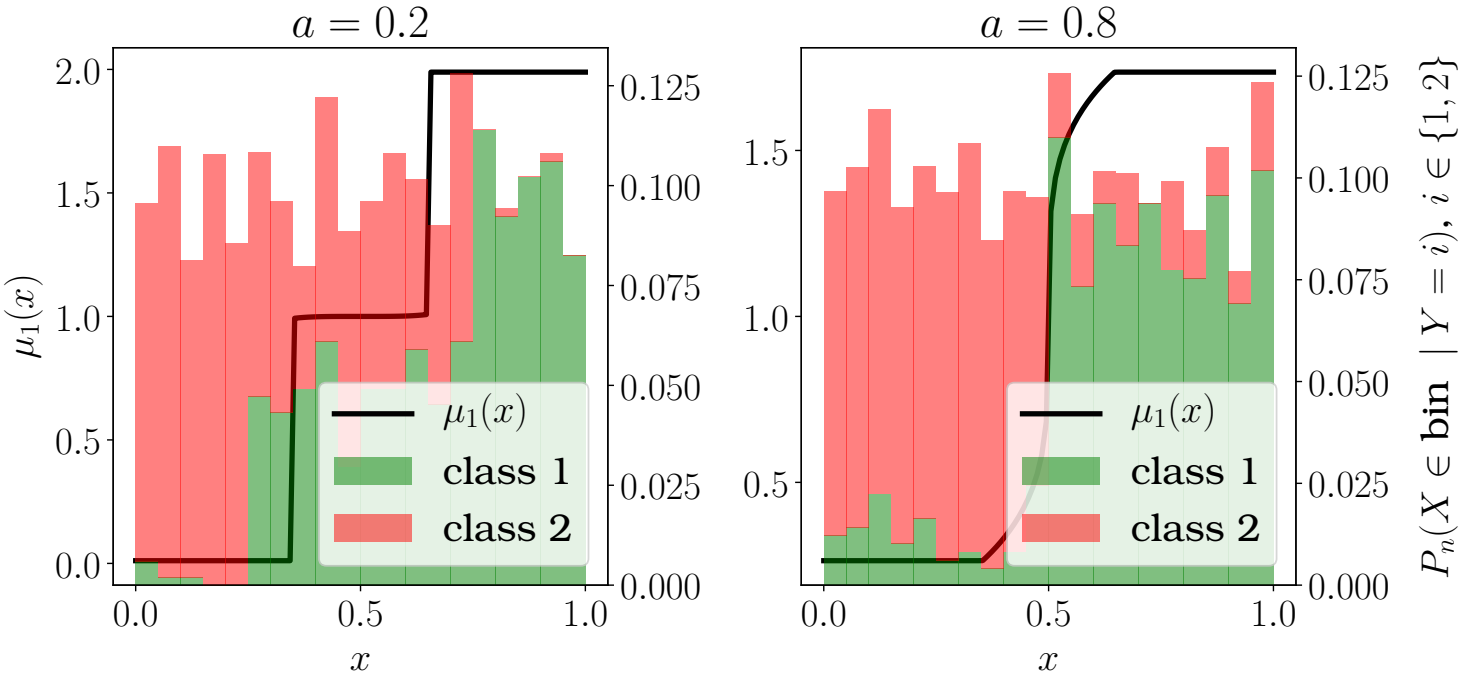
Remark: These fast rates are slower than fast classification rates.

Sketch of proof

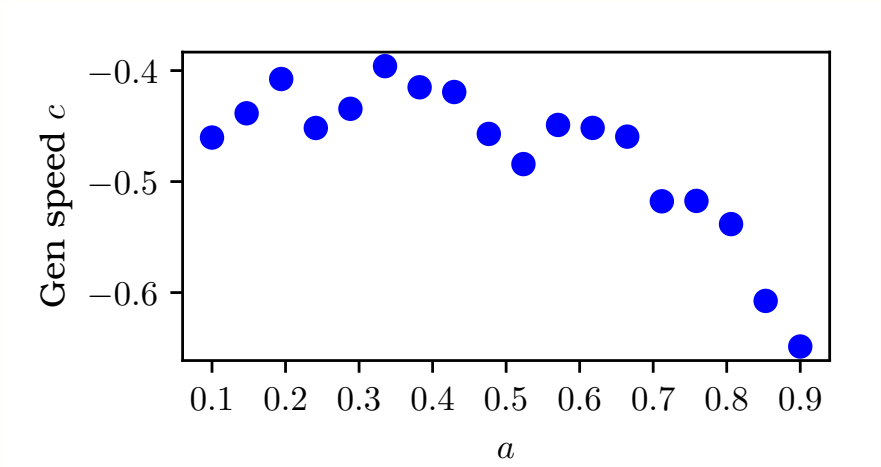
- The noise hypothesis implies a control of the variance of a linearization of the excess positive risk (epr) $R^+_n(G) - R^+_n(G^*)$.
- Apply a concentration inequality using the variance of the linearization of the epr.
- Difference between the epr and its linearization is negligible.

Illustrating the fast rates

We can choose P to satisfy **NA** with different a's:



And we can compute the generalization rates:



Outline

Tractability of eq. (2)

In biometric applications, K is high.

$\Longrightarrow$ R_n^- is an average of too many pairs.

$\Longrightarrow$ **We need to approximate R_n^-.**

Example: LFW dataset:

$$n_+ = 2 \cdot 10^5, \quad n_- = 9 \cdot 10^7.$$

We can:

1. Compute $R_n^-(G)$ with less observations,
2. Use an average of G on B negative pairs, uniformly selected.

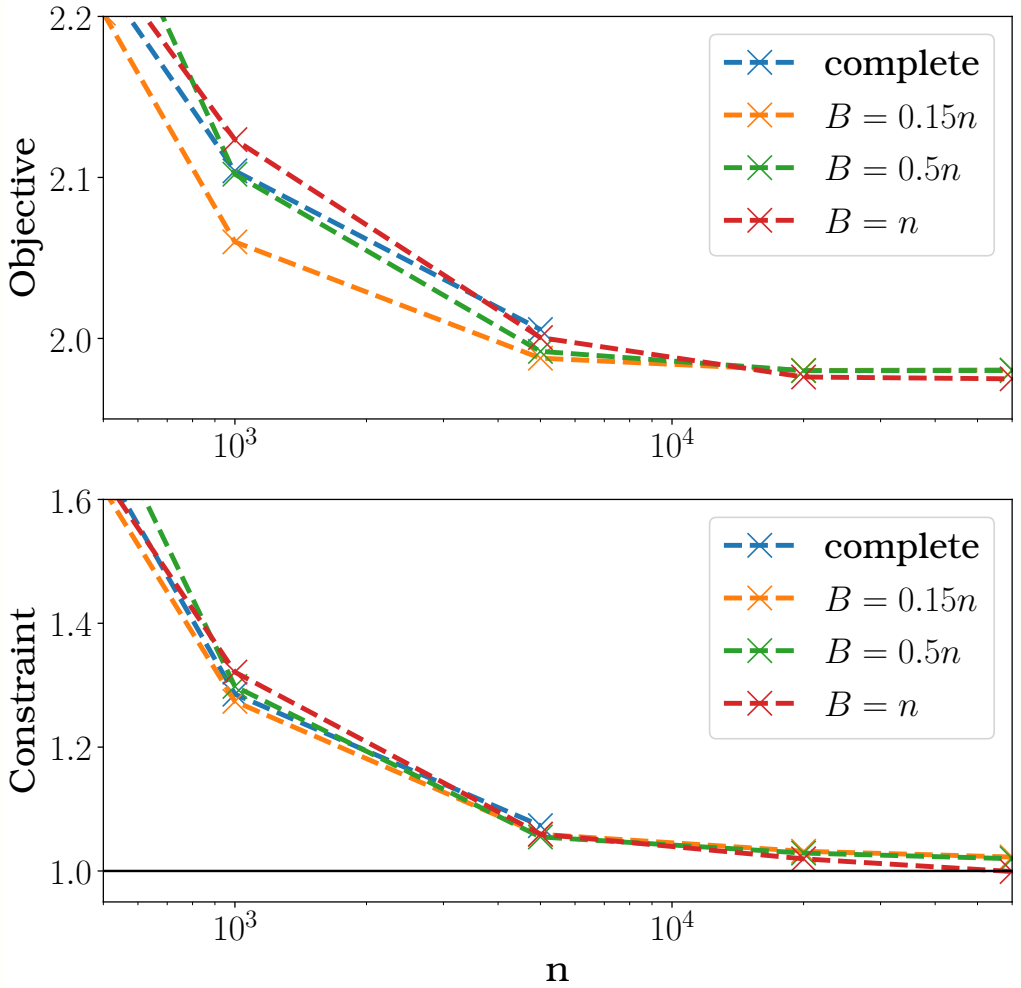
The 2nd proposition gives us an equivalent of theorem 1 with:

$$\Phi = O(B^{-1/2}) + O(n^{-1/2}).$$

It suffices to sample $B = O(n)$ to **have same order guarantees**.

Illustration: Scalability strategy

Illustrated on MMC algorithm with subsampling of negative pairs, see [Xing et al., 2002], and MNIST data. Results:



Subsampling negative pairs does not hinder learning.

Merci !

Come and see us at poster #74 !

References

Cao, Q., Guo, Z.-C., and Ying, Y. (2016).
Generalization Bounds for Metric and Similarity Learning.
Machine Learning, 102(1):115–132.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008).
Ranking and Empirical Minimization of U-Statistics.
The Annals of Statistics, 36(2):844–874.

Clémençon, S. and Vayatis, N. (2008).
Overlaying classifiers: a practical approach for optimal ranking.
In *NIPS 2008*, pages 313–320.

Lee, A. J. (1990).
U-statistics: Theory and practice.
Marcel Dekker, Inc., New York.

Mammen, E. and Tsybakov, A. B. (1995).
Asympotical minimax recovery of the sets with smooth boundaries.
The Annals of Statistics, 23(2):502–524.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J. (2002).
Distance Metric Learning with Application to Clustering with Side-Information.
In *NIPS*.