# A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization

Robin Vogel[1,3], Stéphan Clémençon[1] and Aurélien Bellet[2]   [1] Télécom ParisTech, [2] Inria, [3] IDEMIA

## Motivation

Biometric identification aims to check the claimed identity of an individual by matching his biometric information (e.g., a photo taken at an airport) with another measurement (e.g., a passport photo). Given a similarity function and a threshold, the pair is considered matching if its score is above the threshold.

Performance criteria are hence related to the ROC curve associated with the similarity function, i.e., the relation between the false positive rate and the true positive rate. In biometrics applications, the verification system is typically set to keep the proportion of people falsely considered a match below a predefined acceptable threshold.

The performance criterion we consider in this work is hence *pointwise* ROC *optimization*.

## Preliminaries

The random variable $Y$ denotes thes output label with values in the discrete set $\{1,\dots,K\}$ with $K \geq 1$, and $X$ is the input random variable, taking its values in a feature space $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$. The distribution of the random pair $(X,Y)$ is written $P$.

The objective of similarity learning is to learn, from a training sample $\mathcal{D}_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ composed of $n \geq 1$ independent copies of $(X,Y)$, a (measurable) similarity function $S : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ such that given two independent pairs $(X,Y)$ and $(X',Y')$ drawn from $P$, the larger the similarity $S(X,X')$ between two observations, the more likely they are to share the same label. The class $\mathcal{S}^*$ of optimal similarity rules naturally corresponds to the set of strictly increasing transforms of the pairwise posterior probability $\eta(x,x') = \mathbb{P}\{Y = Y' \mid (X,X') = (x,x')\}$.

Pointwise ROC optimization for a false positive rate $\alpha \in (0,1)$ can be written

$$\max_{S\in\mathcal{S}} R^+(S) \quad \text{subject to} \quad R^-(S) \leq \alpha, \qquad (1)$$

where $R^+(S) = \mathbb{E}[S(X,X') \mid Y = Y']$, $R^-(S) = \mathbb{E}[S(X,X') \mid Y \neq Y']$ and $\mathcal{S}$ is a proposal class of functions. Neymann Pearson's lemma implies that an optimal solution is the indicator function of the set $\mathcal{R}^*_\alpha = \{(x,x') \in \mathcal{X}^2 : \eta(x,x') \geq Q^*_\alpha\}$ where $Q^*_\alpha$ is the conditional quantile of $\eta(X,X')$ given $Y \neq Y'$ at level $1-\alpha$.

## Generalization

We investigate the generalization ability of solutions obtained by solving the empirical version of eq. (1). Natural estimates for the positive risk $R^+(S)$ and the negative risk $R^-(S)$ computed on $\mathcal{D}_n$ are given by:

$$\hat{R}^+_n(S) = \frac{1}{n_+}\sum_{1\leq i<j\leq n} S(X_i,X_j)\cdot\mathbb{I}\{Y_i = Y_j\},$$

$$\hat{R}^-_n(S) = \frac{1}{n_-}\sum_{1\leq i<j\leq n} S(X_i,X_j)\cdot\mathbb{I}\{Y_i \neq Y_j\},$$

where $n_+ = \sum_{1\leq i<j\leq n}\mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$. Using these estimates, one can derive the following empirical problem:

$$\max_{S\in\mathcal{S}} R^+_n(S) \quad \text{subject to} \quad R^-_n(S) \leq \alpha + \Phi, \qquad (2)$$

where we replaced the target level $\alpha$ by $\alpha+\Phi$, where $\Phi$ is some tolerance parameter that should be of the same order as the maximal deviation $\sup_{S\in\mathcal{S}}|\hat{R}^-_n(S) - R^-(S)|$.

Our first result describes the generalization capacities of solutions of the constrained optimization problem eq. (2) under specific conditions for the class $\mathcal{S}$ of similarity functions and a suitable choice of the tolerance parameter $\Phi$.

**Theorem 1.** *Suppose that $\mathcal{S}$ is a VC-major class of functions with finite VC-dimension $V < +\infty$ and that $S(x,x') \leq 1$ for all $S \in \mathcal{S}$ and any $(x,x') \in \mathcal{X}^2$. Assume also that there exists a constant $\kappa \in (0,1)$ such that $\kappa \leq \mathbb{P}\{Y = Y'\} \leq 1-\kappa$. For all $\delta \in (0,1)$ and $n > 1$, set:*

$$\Phi_{n,\delta} = 2C\kappa^{-1}\sqrt{\frac{V}{n}} + 2\kappa^{-1}(1+\kappa^{-1})\sqrt{\frac{\log(3/\delta)}{n-1}},$$

*where $C$ is a known universal constant. Consider a solution $\hat{S}_n$ of the contrained minimization problem eq.* (2) *with $\Phi = \Phi_{n,\delta/2}$. Then, for any $\delta \in (0,1)$, we have simultaneously with probability at least $1-\delta$: $\forall n \geq 1 + 4\kappa^{-2}\log(3/\delta)$,*

$$R^+(\hat{S}_n) \geq \sup_{S\in\mathcal{S}:\ R^-(S)\leq\alpha} R^+(S) - \Phi_{n,\delta/2}$$

$$\textit{and} \quad R^-(\hat{S}_n) \leq \alpha + \Phi_{n,\delta/2}.$$

It is established by combining an uniform bound over $\mathcal{S}$ on the variations of $\hat{R}^+_n$ around its mean, see [1], with the derivations of [2] on pointwise ROC optimization for bipartite ranking.

## Fast Rates

Except for a minor condition, the generalization bound stated in theorem 1 holds whatever the probability distribution of $(X,Y)$. This section introduce situations where rates faster than $O(1/\sqrt{n})$ can be achieved by solutions of eq. (2). It relies on the following noise assumption, that resembles the so-called Mammen-Tsybakov noise condition:

**Noise assumption (NA).** *There exist a constant $c$ and $a \in [0,1]$ such that, almost surely,*

$$\mathbb{E}_{X'}\left[|\eta(X,X') - Q^*_\alpha|^{-a}\right] \leq c.$$

By means of a variant of the Bernstein inequality for $U$-statistics, we can establish fast rate bounds under the preceding condition on the data distribution, which write:

**Theorem 2.** *Suppose that the assumptions of Theorem 1 are satisfied, that condition NA holds true and that the optimal similarity rule $S^*_\alpha(x,x') = \mathbb{I}\{(x,x') \in \mathcal{R}^*_\alpha\}$ belongs to $\mathcal{S}$. Fix $\delta > 0$. Then, there exists a constant $C'$, depending on $\delta$, $\kappa$, $Q^*_\alpha$, $a$, $c$ and $V$ such that, with probability at least $1-\delta$,*

$$R^+(S^*_\alpha) - R^+(\hat{S}_n) \leq C' n^{-(2+a)/4},$$

$$\textit{and} \quad R^-(\hat{S}_n) \leq \alpha + 2\Phi_{n,\delta/2}.$$

## Scalability

In the large-scale setting, solving eq. (2) can be computationally costly due to the very large number of training pairs. In the setting where we have a large number of classes, the number of negative pairs is dramatically higher than that of positive pairs. A natural strategy is to drastically subsample the negative pairs, while keeping all positive pairs.
For that matter, we study the equivalent of eq. (2) when replacing $\hat{R}^-_n(S)$ by the following approximation:

$$\bar{R}^-_B(S) := \frac{1}{B}\sum_{(i,j)\in\mathcal{P}_B} S(X_i,X_j),$$

where $\mathcal{P}_B$ is a set of cardinality $B$ built by sampling with replacement in the set of negative training pairs $\Lambda_P = \{(i,j) \mid i,j \in \{1,\dots,n\};\, Y_i \neq Y_j\}$. In [3], we show that the results of theorem 1 still hold, with a different $\Phi$ of the order $O(\sqrt{\log n/B})$). Remarkably, this implies that it is sufficient to sample $B = O(n)$ pairs to get an order of $O(\sqrt{\log(n)/n})$ learning rate in theorem 1.

## Experiment on Fast Rates

This section illustrates the faster rates of generalization presented in theorem 2. For that matter, we introduce distributions that satisfy the noise assumption (NA) with different $a$'s and show the difference in the generalization speed.

We put ourselves in a simple scenario where $K = 2$, $\mathcal{X} = [0,1]$, $X \sim \mathcal{U}[0,1]$, $\mathbb{P}\{Y = 1\} = 1/2$ and $Q^*_\alpha = 1/2$. It then suffices to define $\mu_1$ as the density of $X$ conditioned upon $Y = 1$ to have fully defined the distribution of the pair $(X,Y)$. We introduce a family of $\mu_1$'s parameterized by $a$, with two examples on fig. 1.

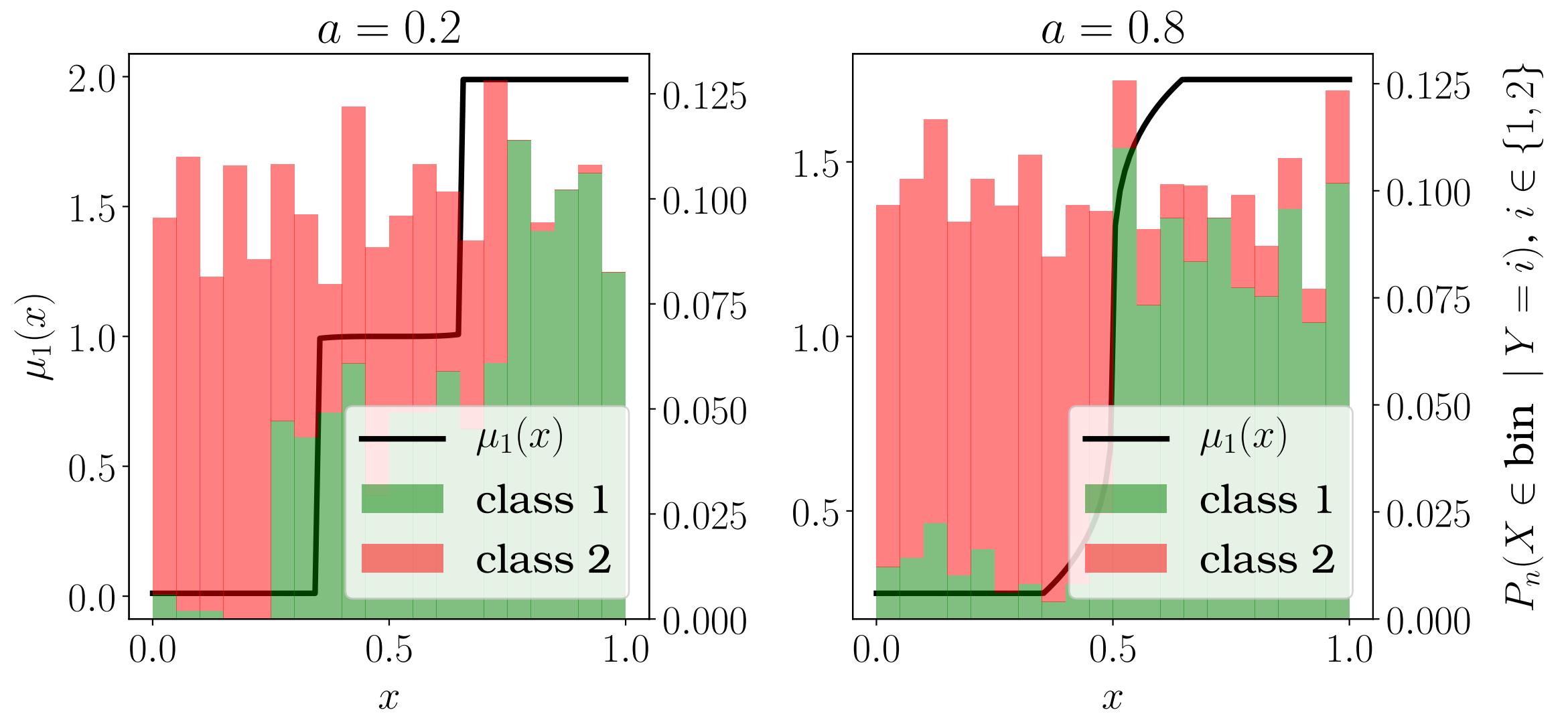

**Figure 1:** Example distributions and $\mu_1$'s for two values of $a$.

For some $a$, the empirical 90-quantile of $R^+(S^*_\alpha) - R^+(\hat{S}_n)$ is computed for different values of $n$ and its logarithm is fitted to $C_a \times \log(n) + D_a$ to get the empirical generalization speed $C_a$. The downward trend when $a$ increases illustrates the fast rates in practice, see fig. 2.

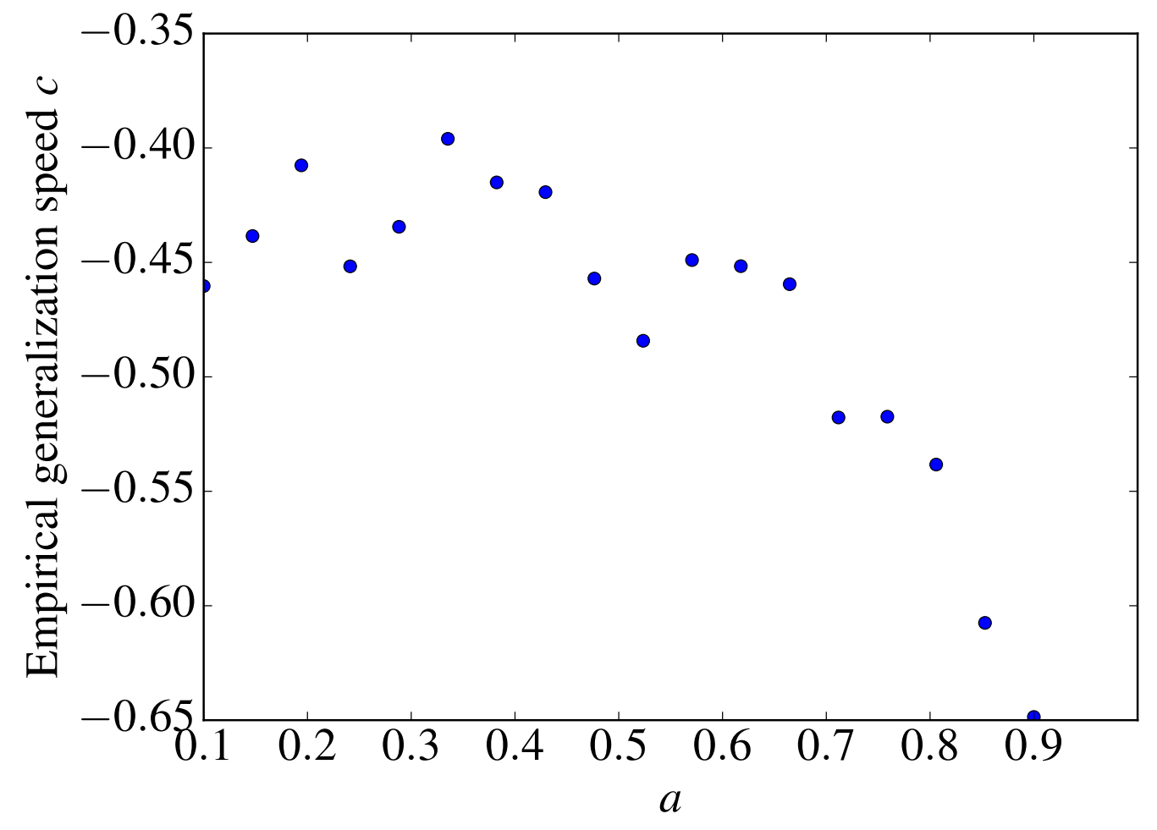

**Figure 2:** Generalization speed for different values of $a$.

## References

[1] Stephan Clémençon, Gàbor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[2] Stephan Clémençon and Nicolas Vayatis. Overlaying classifiers: A practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010.

[3] Robin Vogel, Stephan Clémençon, and Aurélien Bellet. A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization. *ICML*, 2018.