# Nonparametric Estimation of the Precision-Recall Curve

**Stéphan Clémençon** STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
LTCI UMR Telecom ParisTech/CNRS No. 5141, 46 rue Barrault, 75634 Paris Cedex, France

**Nicolas Vayatis** NICOLAS.VAYATIS@CMLA.ENS-CACHAN.FR
CMLA UMR CNRS No. 8536 & UniverSud, 61 rue due Président Wilson, 94235 Cachan Cedex, France

## Abstract

The Precision-Recall (PR) curve is a widely used visual tool to evaluate the performance of scoring functions in regards to their capacities to discriminate between two populations. The purpose of this paper is to examine both theoretical and practical issues related to the statistical estimation of PR curves based on classification data. Consistency and asymptotic normality of the empirical counterpart of the PR curve in sup norm are rigorously established. Eventually, the issue of building confidence bands in the PR space is considered and a specific resampling procedure based on a smoothed and truncated version of the empirical distribution of the data is promoted. Arguments of theoretical and computational nature are presented to explain why such a bootstrap is preferable to a "naive" bootstrap in this setup.

## 1. Introduction

Although the ROC curve remains the golden standard in a wide variety of applications, ranging from anomaly detection in signal analysis to medical diagnosis, through credit-risk screening, for evaluating how a test statistic performs in regards to its capacity of discrimination between two populations, the *Precision-Recall* (PR) curve has recently received much attention in the machine-learning literature, through the development of statistical learning procedures tailored for the *bipartite ranking problem*. When the distribution of the pooled population is highly skewed, *i.e.* when the theoretical proportion of positive instances is (very) small, which is often the case in Information Retrieval (IR) applications, PR curves offer a scale-adapted graphical display that permit to visualize much more easily ranking performance, see (Manning & Schutze, 1999) or (Raghavan et al., 1989) for instance.

The goal of bipartite ranking consists of determining a test statistic with a PR curve "as high as possible" everywhere and assessing the performance of a given candidate from training data is thus of prime importance. However, although PR analysis is a more and more popular tool in IR applications, crucial questions related to its statistical estimation have not been addressed yet. It is hence the purpose of this paper to investigate the statistical properties of empirical counterparts of the PR curve from the asymptotic angle. Beyond consistency and asymptotic normality issues, we tackle the problem of building confidence bands for the PR curve of a given scoring function $\mathtt{s}(\mathtt{x})$ or parts of it in a fully data-driven fashion. Given the complexity of the asymptotic law of the (centered) empirical PR curve, we suggest to plot confidence bands in the PR space equipped with the sup norm by means of a resampling procedure, following in the footsteps of (Horvath et al., 2008) and (Bertail et al., 2008) where the bootstrap paradigm is applied to the construction of ROC confidence bands, see also (Macskassy & Provost, 2004) and (Macskassy et al., 2005) for a pointwise approach. The main novelty lies in the fact that the specific resampling method proposed here permits to remedy the possible computational difficulties caused by an extreme degree of asymmetry between the two populations forming the training data sample. Asymptotic validity of this bootstrap procedure is proved by means of the limit results previously established for the empirical PR curve.

The rest of the article is structured as follows. In Section 2, notations are first set out and key notions of PR analysis are recalled. Nonparametric estimation of the PR curve of a given scoring function based on clas-

Appearing in *Proceedings of the 26$^{\mathrm{th}}$ International Conference on Machine Learning*, Montreal, Canada, 2009. 

sification data is tackled in Section 3 from a functional perspective. The statistical procedure we propose for bootstrapping PR curves is presented in Section 4, together with the theoretical results establishing its asymptotic validity. Eventually, technical details are deferred to the Appendix.

## 2. Preliminaries

Here we briefly describe the issue of bipartite ranking and recall the related key concepts of PR analysis. We also set out the notations that shall be needed throughout the paper.

### 2.1. The bipartite framework

In the *bipartite ranking problem*, the matter is to order all the elements $X$ of a set $\mathcal{X}$ by degree of relevance, given a binary label information $Y$. Consider a system with binary random output $Y$, taking values in $\{-1, 1\}$, and a random input $X$, valued in a (generally high-dimensional) feature space $\mathcal{X}$. In the subsequent analysis, the distribution of the pair $(X, Y)$ is either described by the triplet $(p, G, H)$ where $p = \mathbb{P}(Y = +1) \in (0, 1)$ is the theoretical proportion of positive instances, and $G$ and $H$ respectively denote the conditional distribution of $X$ given $Y = +1$ and given $Y = -1$, or else by $(\mu, \eta)$ where $\mu$ denotes the marginal distribution of $X$ and $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, is the regression function. Here and throughout, we assume that $G$ and $H$ are equivalent. Equipped with these notation, we point out that $\mu = pG + (1-p)H$ and $dG/dH(X) = ((1-p)\eta(X))/(p(1-\eta(X))$. The probabilistic setup is the same as for standard binary classification but the goal is different. In the context of IR applications, one is concerned by ordering all the documents $x$ of the list $\mathcal{X}$ by degree of relevance for a particular query, rather than simply classifying them as relevant or not. This amounts to assign to each document $x$ in $\mathcal{X}$ a *score* $s(x)$ indicating its degree of relevance for this specific query. The challenge is thus to build a *scoring function* $s : \mathcal{X} \to \mathbb{R}$ from sampling data, so as to rank the observations $x$ by increasing order of their score $s(x)$ as accurately as possible: the higher the score $s(X)$ is, the more likely one should observe $Y = +1$. We denote by $\mathcal{S} = \{s : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ the set of all scoring functions.

### 2.2. True PR curves

A standard way of measuring the ranking performance of a given scoring function $s$ consists of plotting its precision-recall (PR) curve:

$$t \in \mathcal{D}_s \mapsto \left(\mathbb{P}\{s(X) \geq t \mid Y = +1\},\ \mathbb{P}\{Y = +1 \mid s(X) \geq t\}\right), \tag{1}$$

where $\mathcal{D}_s = \{t \in \mathbb{R} :\ \mu(\{x\ :\ s(x) \geq t\}) > 0\}$. The quantity $\mathrm{prec}_s(t) \overset{\text{def}}{=} \mathbb{P}\{Y = +1 \mid s(X) \geq t\}$ represents the *precision* of the test based on the thresholded diagnostic statistic $s(X)$ for discriminating between the composite hypotheses $\mathcal{H}_0 : Y = -1$ and $\mathcal{H}_1 : Y = +1$ (*i.e.* $\mathcal{H}_0 : X \sim H$ and $\mathcal{H}_1 : X \sim G$), while its power $\beta_s(t) \overset{\text{def}}{=} \mathbb{P}\{s(X) \geq t \mid Y = +1\}$ is generally termed *recall* or *true positive rate*. We also consider the *false positive rate* $\alpha_s(t) \overset{\text{def}}{=} \mathbb{P}\{s(X) \geq t \mid Y = -1\}$ and write:

$$\forall t \in \mathcal{D}_s,\ \ \mathrm{prec}_s(t) = \frac{p\beta_s(t)}{p\beta_s(t) + (1-p)\alpha_s(t)}.$$

The PR space is the set of all achievable points $\{(\alpha_s(t), \beta_s(t));\ \ (s, t) \in \mathcal{S} \times \mathbb{R}\}$. For notational convenience, we introduce the conditional cumulative distribution functions (cdf) of $s(X)$ denoted by $G_s(t) = 1-\beta_s(t)$ and $H_s(t) = 1-\alpha_s(t)$. We consider the subset $\mathcal{S}_0$ of scoring functions $s(x)$ such that the equivalent distributions $H_s$ and $G_s$ are continuous. In the case where $s \in \mathcal{S}_0$, the PR curve matches with the graph of the *càd-làg* (right-continuous, left-limit) mapping:

$$\mathrm{PR}_s : \beta \in (0, 1) \mapsto \frac{p\beta}{p\beta + (1-p)\alpha(s, \beta)}, \tag{2}$$

where $\alpha(s, \beta) \overset{\text{def}}{=} 1 - H_s \circ G_s^{-1}(1-\beta)$ for all $\beta \in (0, 1)$, $F^{-1}(\alpha) = \inf\{x \in \mathbb{R} / F(x) \geq \alpha\}$ denotes the generalized inverse of any cdf $F$. We point out that the curve $\mathrm{PR}_s$ is continuous as soon as $G_s$ is strictly increasing and, in the case where $H_s = G_s$ (*i.e.* the scoring function $s(x)$ has no capacity to discriminate between the two populations), we have $\mathrm{PR}_s \equiv p$. Notice additionally that $\lim_{\beta \to 1} \mathrm{PR}_s(\beta) = p$ (the distributions $G_s$ and $H_s$ have the same support) and, as $\beta \to 0$, $\mathrm{PR}_s(\beta)$ has a finite limit $pl(s)/(pl(s) + 1 - p)$, when the (possibly infinite) limit $l(s) = \lim_{t \to \infty} dG_s/dH_s(t)$ exists, the latter being determined by the right tail behavior of the distributions $H_s$ and $G_s$: in particular, if $(1 - G_s(t))/(1-H_s(t)) \to \infty$ as $t \to \infty$, $l(s) = \infty$ and $\mathrm{PR}_s$ has limit value $1 \overset{\text{def}}{=} \mathrm{PR}_s(0)$ as $\beta \to 0$. In addition, $\mathrm{PR}_s$ and the mapping $\beta \in (0, 1) \mapsto \alpha(s, \beta)/\beta$ vary in the opposite direction: $\mathrm{PR}_s$ is thus non increasing as soon as the mapping $\beta \in (0, 1) \mapsto \alpha(s, \beta)$ is convex (*i.e.* the likelihood ratio $dG_s/dH_s(s(X))$ is monotone).

Given the goal pursued in the bipartite ranking, it would be desirable that the scoring function is such that the positive instances tend to have higher scores than the negative ones, or, more formally, that $G_s$ is *stochastically larger* than $H_s$: $\forall t \in \mathcal{D}_s$, $1 - H_s(t) \leq$

$1 - G_s(t)$. This boils down to assume that s(x)'s PR curve is above the horizontal line $\beta \equiv p$ everywhere in the PR space. This is naturally not sufficient in practice and, actually, one would like to select a scoring function $s$ so that $G_s$ is "as stochastically larger" than $H_s$ as possible, provided a rigorous meaning can be given to such an attempt of quantification. The concept of PR curve permits to formalize this precisely.

**A partial order on $\mathcal{S}_0$.** As a matter of fact, the PR curve indeed induces a partial order on the set of scoring functions $\mathcal{S}_0$. Let $(s_1, s_2) \in \mathcal{S}_0^2$, we will say that the scoring function $s_1$ is more accurate than $s_2$ if and only if its PR curve is above the one of $s_2$ everywhere, *i.e.* $\forall \beta \in (0,1)$, $PR_{s_1}(\beta) \geq PR_{s_2}(\beta)$: for any fixed recall, $s_1$ yields a better precision. In regards to this functional performance criterion, the class of optimal scoring functions is the set of increasing transforms of the regression function: $\mathcal{S}^* = \{\psi \circ \eta,\ \psi \text{ strictly increasing}\}$. Indeed, it follows from Neyman-Pearson's lemma that the test based on the likelihood ratio statistic $\phi(X) = dG/dH(X)$, or equivalently based on $\eta(X) = (1-p)\phi(X)/(p+(1-p)\phi(X))$, is *uniformly more powerful* among all unbiased tests: for a given power (recall) $\beta \in (0,1)$, it has minimum type I error, $\alpha(\eta, \beta)$ namely, and thus maximum precision: $\forall \beta \in (0,1)$,

$$PR^*(\beta) \stackrel{\text{def}}{=} PR_\eta(\beta) \geq PR_s(\beta),$$

for all $s \in \mathcal{S}$. We set $G^* = G_\eta$ and $H^* = H_\eta$ and suppose that these distributions are absolutely continuous with respect to Lebesgue measure. We recall from (Clémençon & Vayatis, 2008) that:

$$\phi(X) = \frac{1-p}{p} \cdot \frac{\eta(X)}{1-\eta(X)} = \frac{dG^*}{dH^*}(\eta(X)).$$

Thus, $\lim_{\beta \to 0} PR^*(\beta)=1$ as soon as the essential supremum of $\eta(X)$ is equal to 1, or equivalently the one of the likelihood ratio $\phi(X)$ is equal to $+\infty$. We point out in addition that $PR^*$ is non increasing, since the likelihood ratio $dG^*/dH^*(\eta(X))$ is monotone.

Hence, the performance of a scoring function $s \in S_0$ depends on the closeness between $PR_s(\beta)$ and $PR^*(\beta)$ for all $\beta \in (0,1)$. It can be thus naturally quantified by measuring the deviation between its PR curve and $PR^*$ in sup norm.

**Remark 1** (PR ANALYSIS *vs.* ROC ANALYSIS) Recall that the ROC curve of a scoring function $s \in \mathcal{S}$ is the *PP*-plot $t \in \mathbb{R} \mapsto (\alpha_s(t),\ \beta_s(t))$. When $s \in \mathcal{S}_0$, it is actually the graph of the function $\alpha \in (0,1) \mapsto ROC(s,\alpha) = 1 - G_s \circ H_s(1-\alpha)$ and one may show that it is concave iff the likelihood ratio $dG_s/dH_s(s(X))$ is monotone ($PR_s$ is then non increasing). As for PR curves, all such curves are dominated by $ROC^* = ROC(s^*, .)$, $s^* \in \mathcal{S}^*$, in the ROC space. The major difference lies in the fact that the rate of positive instances $p$ is not involved in the definition of the ROC curve and when the latter is very small for instance, a graphical display of $ROC(s, .)$ may yield a wrong judgement about $s(x)$' discrimination capacity, see (Davis & Goadrich, 2006).

### 2.3. Empirical PR curve estimates.

In practice, learning strategies for selecting a good scoring function are based on training data $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$ and should thus rely on accurate empirical estimates of the true PR curves. Let $s \in \mathcal{S}$. Denote the empirical versions of $H_s(t)$ and $G_s(t)$ by

$$\widehat{H}_s(t) = \frac{1}{n_-} \sum_{1 \leq j \leq n,\ Y_j = -1} K(t - s(X_j)),$$

$$\widehat{G}_s(t) = \frac{1}{n_+} \sum_{1 \leq j \leq n,\ Y_j = +1} K(t - s(X_j)),$$

where $n_+ = \sum_{j=1}^n \mathbb{I}\{Y_i = +1\} = n - n_-$ is the number of positive instances among the sample (distributed as the binomial $Bin(n, p)$) and $K(u)$ denotes the step function $\mathbb{I}\{u \geq 0\}$. As discussed in (Davis & Goadrich, 2006), if empirical estimation of the PR curve of $s(x)$ is based on the definition given in (1), one faces the question of how to interpolate the points $(\widehat{\beta}_{s,i}, \widehat{prec}_{s,i})$, $1 \leq i \leq n$, where: $\forall i \in \{1, \ldots, n\}$,

$$\widehat{\alpha}_{s,i} = 1 - \widehat{H}_s(s(X_i)),\ \widehat{\beta}_{s,i} = 1 - \widehat{G}_s(s(X_i)),$$

and $\widehat{prec}_{s,i} = n_+\widehat{\beta}_{s,i}/(n_+\widehat{\beta}_{s,i} + n_-\widehat{\alpha}_{s,i})$. We point out that, in the situation where $H_s$ and $G_s$ are continuous distributions, given the representation (2), it is natural to consider as statistical version of the curve $PR_s$ the graph $\{(\beta, \widehat{PR}_s(\beta);\ \beta \in (0,1)\}$, where:

$$\forall \beta \in (0,1),\ \widehat{PR}_s(\beta) = \frac{n_+\beta}{n_+\beta + n_-\widehat{\alpha}(s,\beta)}, \quad (3)$$

with $\widehat{\alpha}(s,\beta) = 1 - \widehat{H}_s \circ \widehat{G}_s(1-\beta)$.

**Smoothed PR curve estimates.** In order to obtain a smoother estimate of a supposedly regular curve $PR_s$, one may consider smoothed versions $\widetilde{G}_s(x)$ and $\widetilde{H}_s(x)$ of the class cdf's. A typical choice consists of picking, instead of the step function $\mathbb{I}\{u \geq 0\}$, a function $K(u)$ of the form $\int_{v \geq 0} \mathbf{K}_h(u - v)dv$, with $\mathbf{K}_h(u) = h^{-1}\mathbf{K}(h^{-1} \cdot u)$ where $\mathbf{K} \geq 0$ is a regularizing Parzen-Rosenblatt kernel (*i.e.* a bounded square integrable function such that $\int K(v)dv = 1$) and $h > 0$ is the smoothing bandwidth.

## 3. Consistency and asymptotic law

Throughout this section, the scoring function $s \in \mathcal{S}_0$ is fixed. Let $Z = s(X)$ and denote by $h_s(x)$ and $g_s(x)$ the densities of the class distributions $H_s$ and $G_s$, by $\mathcal{P}$ the joint distribution of $(Z, Y)$ on $\mathbb{R} \times \{-1, +1\}$ and by $\widehat{\mathcal{P}}_n$ its empirical version based on the sample $\mathcal{D}_n = \{(Z_i, Y_i)\}_{1 \le i \le n}$ where $Z_i = s(X_i)$ for all $i \in \{1, \dots, n\}$. Equipped with these notations, we have $\mathcal{P}(dz, y) = p\mathbb{I}\{y = +1\}H_s(dz) + (1-p)\mathbb{I}\{y = -1\}G_s(dz)$ and $\widehat{\mathcal{P}}_n(dz, y) = (n_+/n)\mathbb{I}\{y = +1\}\widehat{H}_s(dz) + (1 - n/n_+)\mathbb{I}\{y = -1\}\widehat{G}_s(dz)$.

The purpose of this section is to investigate the asymptotic properties of the random function $\beta \in (0,1) \mapsto \widehat{\mathrm{PR}}(\beta)$ as an estimator of the curve $\mathrm{PR}_s$, as the sample size $n$ tends to infinity. The next theorem reveals it is strongly consistent and establishes a strong approximation result for the PR *fluctuation process*:

$$R_n(\beta) = \sqrt{n}\left(\widehat{\mathrm{PR}}_s(\beta) - \mathrm{PR}_s(\beta)\right), \quad \beta \in (0,1). \quad (4)$$

Although it is not directly useful from a practical perspective, this limit result plays a crucial role in understanding the asymptotic behavior of the empirical PR curve and of its bootstrap counterpart, as will be seen in the next section. The technical assumptions listed below are required. Let $\epsilon \in (0, 1/2)$ be fixed.

**$\mathbf{H_1}$** The slope of the function $\beta \mapsto \alpha(s, \beta)$ is bounded on $[\epsilon, 1-\epsilon]$:

$$\sup_{\beta \in [\epsilon, 1-\epsilon]} \frac{h_s(G_s^{-1}(\beta))}{g_s(G_s^{-1}(\beta))} < \infty. \quad (5)$$

**$\mathbf{H_2}$** the density $g_s$ is differentiable and

$$\forall \beta \in (\epsilon, 1-\epsilon), \quad g_s(G_s^{-1}(\beta)) > 0, \quad (6)$$

and there exists $\gamma > 0$ such that

$$\sup_{\beta \in (\epsilon, 1-\epsilon)} \frac{d \log(g_s \circ G_s^{-1}(\beta))}{d\beta} \le \gamma < \infty. \quad (7)$$

**Theorem 1** *(Strong approximation) Suppose that assumptions $\mathbf{H}_1 - \mathbf{H}_2$ are fulfilled. Then,*

*(i) the empirical* PR *curve is strongly consistent, uniformly over $[\epsilon, 1-\epsilon]$:*

$$\sup_{\beta \in [\epsilon, 1-\epsilon]} |\widehat{\mathrm{PR}}_s(\beta) - \mathrm{PR}_s(\beta)| \to 0 \text{ a.s. as } n \to \infty,$$

*(ii) there exist two independent sequences of brownian bridges $\{B_1^{(n)}(\beta)\}_{\beta \in (0,1)}$ and $\{B_2^{(n)}(\beta)\}_{\beta \in (0,1)}$ and a normal r.v. $W$, independent from those sequences, such that we almost-surely have, uniformly over $[\epsilon, 1-\epsilon]$: as $n \to \infty$,*

$$R_n(\beta) = Z^{(n)}(\beta) + o\left(\frac{(\log\log n)^{\rho_1(\gamma)} \log^{\rho_2(\gamma)} n}{\sqrt{n}}\right), \quad (8)$$

*where, with $a_p = \sqrt{1-p}/p$,*

$$Z^{(n)}(\beta) = \frac{\mathrm{PR}_s(\beta)^2}{\beta}\{\frac{a_p}{p}\alpha(s,\beta)W + a_p^2\sqrt{p}\frac{h_s(G_s(1-\beta))}{g_s(G_s(1-\beta))}B_1^{(n)}(\beta) + a_p B_2^{(n)}(\alpha(s,\beta))\},$$

*and*

$$\begin{cases} \rho_1(\gamma) = 0, & \rho_2(\gamma) = 1, & \text{if } \gamma < 1 \\ \rho_1(\gamma) = 0, & \rho_2(\gamma) = 2, & \text{if } \gamma = 1 \\ \rho_1(\gamma) = \gamma, & \rho_2(\gamma) = \gamma - 1 + \varepsilon,\ \varepsilon > 0, & \text{if } \gamma > 1. \end{cases}$$

The proof is deferred to the appendix section. The strong approximation result stated in (ii) implies that the PR fluctuation process $\{R_n(\beta)\}_{\beta \in [\epsilon, 1-\epsilon]}$ converges weakly, in the space $(\mathbb{D}([\epsilon, 1-\epsilon]), \|.\|_\infty)$ of *càd-làg* functions equipped with the sup norm, to the law of the gaussian stochastic process $\{Z^{(1)}(\beta)\}_{\beta \in [\epsilon, 1-\epsilon]}$.

**Remark 2** (On assumptions) Hypothesis $\mathbf{H}_2$ is a standard assumption for obtaining a strong approximation for the quantile process $\{G_s^{-1}(\beta)\}$ involved in $\mathrm{PR}_s$'s definition, see (Csorgo & Revesz, 1981).

**Remark 3** (On pointwise limit results) Freezing the precision $\beta \in [\epsilon, 1-\epsilon]$, the asymptotic behavior of the recall estimator $\widehat{\mathrm{PR}}_s(\beta)$ is described by the following pointwise limit result, obtained as a direct consequence of Theorem 1's part (ii): as the sample size $n$ tends to infinity, we have the convergence in distribution

$$\sqrt{n}\left(\widehat{\mathrm{PR}}_s(\beta) - \mathrm{PR}_s(\beta)\right) \Rightarrow \mathcal{N}\left(0, \sigma_s^2(\beta)\right),$$

where the asymptotic variance is given by

$$\begin{aligned} \sigma_s^2(\beta) &= \frac{(1-p)\mathrm{PR}_s(\beta)^4 \alpha(s,\beta)^2}{\beta^2 p^3} \\ &+ \frac{(1-p)^2 \beta(1-\beta) h_s^2(G_s(1-\beta))}{g_s^2(G_s(1-\beta)) p^3} \\ &+ \frac{(1-p)\alpha(s,\beta)(1-\alpha(s,\beta))}{p^2}. \end{aligned} \quad (9)$$

## 4. Confidence bands for PR curves

Beyond consistency of the empirical PR curve in sup norm and the asymptotic normality of the PR fluctuation process, we now tackle the question of constructing confidence bands for the true PR curve. Truth

should be said, the complex, though explicit, form for the limiting gaussian law of the fluctuation process $R_n$ given in Theorem 1 can hardly be used for this purpose: indeed, simulating an empirical counterpart of the limit process would require to estimate in particular the densities of the distributions $H_s(dz)$ and $G_s(dz)$. Such computational difficulties strongly advocate for using the bootstrap approach introduced by (Efron, 1979). The latter suggests to consider, as an estimate of the law of the fluctuation process $\{R_n(\beta)\}_{\beta\in[\epsilon,1-\epsilon]}$, the conditional law given the data sample $\mathcal{D}_n = \{(Z_i, Y_i)\}_{1\le i\le n}$ of the *bootstrapped* PR *fluctuation process*

$$R_n^* = \left\{\sqrt{n}(PR_s^*(\beta) - \widehat{PR}_s(\beta))\right\}_{\beta\in(0,1)},$$

where $PR_s^*$ is the PR curve corresponding to a sample $\mathcal{D}_n^* = \{(Z_i^*, Y_i^*)\}_{1\le i\le n}$ of i.i.d. random pairs, with a common distribution $\widetilde{\mathcal{P}}_n$ close to $\mathcal{P}_n$.

The barrier to implementing the bootstrap in this setup is twofold. Firstly, due to a possible extreme asymmetry between the positive and negative populations within the pooled sample $\mathcal{D}_n$, which situation precisely justifies the use of PR analysis for evaluating the performance of a scoring function, resampling the data in a naive fashion, by drawing with replacement among $\mathcal{D}_n$, may yield a bootstrap sample containing no positive instances with overwhelming probability. Secondly, the PR fluctuation process is a functional of the the quantile process $\{\widehat{G}_s^{-1}(\beta)\}_{\beta\in[0,1]}$. It is well-known that the *naive bootstrap* (*i.e.* resampling from the raw empirical distribution) generally provides bad approximations of the distribution of empirical quantiles in practice: the rate of convergence of the bootstrap distribution estimate for a given quantile is of order $O_{\mathbb{P}}(n^{-1/4})$ (Falk & Reiss, 1989), whereas the rate of the gaussian approximation is $n^{-1/2}$. The bootstrap procedure we describe next permits to overcome both difficulties.

### 4.1. The PR bootstrap algorithm

Here we describe an algorithm for building a confidence band at level $1-\delta \in (0,1)$ in the PR space from sampling data $\mathcal{D}_n = \{(Z_i, Y_i);\ 1 \le i \le n\}$. Let $\widetilde{p} \in (0,1)$. It is performed in four steps described in the pseudocode Algorithm 1.

Similarly to what has been suggested by (Bertail et al., 2008) and (Horvath et al., 2008) for ROC curves, the algorithm above implements a *smoothed version* of the bootstrap method. In a similar fashion to what is recommended for ROC curves, we propose to implement a *smoothed version* of the bootstrap algorithm in order to improve the approximation rate of $\sup_{\beta\in[\epsilon,1-\epsilon]} |R_n(\beta)|$'s distribution. Here this boils down to resampling the data from a smoothed version of the empirical distribution $\mathcal{P}_n$.

**Algorithm 1** Precision-Recall bootstrap

1. Based on $\mathcal{D}_n$, compute the empirical class cdf estimates $\widehat{G}_s$ and $\widehat{H}_s$, as well as their smoothed versions $\widetilde{G}_s$ and $\widetilde{H}_s$. Plot the PR curve estimate: $\forall \beta \in (0,1)$,

$$\widehat{PR}_s(\beta) = \frac{\widehat{p}_n \beta}{\widehat{p}_n \beta + (1-\widehat{p}_n)(1 - \widehat{H}_s \circ \widehat{G}_s^{-1}(1-\beta))}.$$

2. From the smooth distribution estimate

$$\widetilde{\mathcal{P}}_n(dz, dy) = \widetilde{p}\mathbb{I}\{y = +1\}\widetilde{G}_s(dz) + (1-\widetilde{p})\mathbb{I}\{y = -1\}\widetilde{H}_s(dz),$$

draw an i.i.d. bootstrap sample $\mathcal{D}_n^* = \{(Z_i^*, Y_i^*)\}_{1\le i\le n}$ conditioned on $\mathcal{D}_n$.

3. Based on $\mathcal{D}_n^*$, compute the bootstrap versions of the empirical class cdf estimates:

$$G_s^*(z) = \frac{1}{n_+^*} \sum_{1\le i\le n,\ Y_i^*=+1} \mathbb{I}\{Z_i^* \le z\},$$
$$H_s^*(z) = \frac{1}{n_-^*} \sum_{1\le i\le n,\ Y_i^*=-1} \mathbb{I}\{Z_i^* \le z\},$$

where $n_+^* = \sum_{i=1}^n \mathbb{I}\{Y_i^* = +1\} = n - n_-^*$. Set $p_n^* = n_+^*/n$ and plot the bootstrap PR curve

$$PR_s^*(\beta) = \frac{p_n^* \beta}{p_n^* \beta + (1-p_n^*)(1 - H_s^* \circ G_s^{*-1}(1-\beta))}.$$

4. Get the *bootstrap confidence bands at level* $1-\delta$ defined by the ball of center $\widehat{PR}$ and radius $r(\delta)/\sqrt{n}$ in $\mathbb{D}([\epsilon, 1-\epsilon])$, with $r(\delta)$ defined by

$$\mathbb{E}^*\left[\gamma_n \cdot \mathbb{I}\left\{\sup_{\beta\in[\epsilon,1-\epsilon]} |R_n^*(\beta)| \le r(\delta)\right\}\right] = 1-\delta,$$

where the *importance function* is given by

$$\gamma_n = \left(\frac{\widehat{p}_n}{\widetilde{p}}\right)^{n_+^*} \cdot \left(\frac{1-\widehat{p}_n}{1-\widetilde{p}}\right)^{n-n_+^*},$$

denoting by $\mathbb{E}^*[.]$ the conditional expectation given the original data $\mathcal{D}_n$.

The heuristics underlying the gain in accuracy resulting from the smoothing stage is as follows: when based on a smooth distribution, the bootstrap PR curve and the original one share similar regularity properties: in particular, an asymptotic result analogous to the one given in part (ii) of Theorem 1 holds for the bootstrap curve, with a limit law close to the one of $Z^{(1)}$, provided the instrumental distribution is close to $\mathcal{P}$.

*Importance bootstrap resampling* is also implemented through Algorithm 1, see subsection 5.4.6 in (Shao & Tu, 1995). The principle consists of sampling novel data from a distribution, different from the raw distribution, or its smoothed version $\widehat{\mathcal{P}}_n(dz, dy) = \widehat{p}_n \mathbb{I}\{y = +1\}\widetilde{G}_s(dz) + (1-\widehat{p}_n)\mathbb{I}\{y = -1\}\widetilde{H}_s(dz)$ in our case, under which positive instances are observed with much larger probability, making the situation where the bootstrap sample is formed of observations with negative labels solely less frequent, or even rare. Here the procedure simply consists of replacing the empirical rate of positive instances $\widehat{p}_n = n_+/n$ in $\widehat{\mathcal{P}}_n(dz, dy)$'s formula by a constant $\widetilde{p}$ (equal to $1/2$ say), yielding the instrumental sampling distribution $\widetilde{\mathcal{P}}_n(dz, dy)$. The resulting biased bootstrap statistics based on pairs $(Z^*, Y^*)$ sampled from $\widetilde{\mathcal{P}}_n(dz, dy)$ will be then corrected by a multiplicative factor, the *importance function* $\gamma_n$.

Before turning to the theoretical properties of this algorithm, a few remarks are in order.

**Remark 4** (Monte-Carlo simulation) From a practical perspective, the true smoothed bootstrap distribution must be approximated too, through a Monte-Carlo approximation scheme. A computationally convenient way of performing such a smoothed resampling consists of drawing $B \sim n$ bootstrap samples, of size $n$, with replacement in the original data and then adding to each drawn data an independent centered gaussian r.v. of variance $h^2$. This is equivalent to drawing bootstrap data from a smooth estimate $\widetilde{\mathcal{P}}_n(dz, dy)$ computed using a gaussian kernel $K_h(u) = (2\pi h^2)^{-1/2}\exp(-u^2/(2h^2))$, see Silverman & Young (1987).

**Remark 5** (Tuning parameters) Implementation of Algorithm 1 requires to select two tuning parameters essentially: the constant $\widetilde{p}$ involved in the importance sampling (IS) step and the bandwidth $h$ related to the smoothing stage. When using a gaussian regularizing kernel $K_h$ (or the trick recalled in Remark 4 above), one should classically pick $h = h_n \sim n^{-1/5}$ in order to minimize the mean square error. Concerning the IS parameter $\widetilde{p}$, a practical strategy consists in choosing it so as to minimize the variance of the resulting Monte-Carlo estimate. Owing to space limitation, here we refer to (Bucklew, 2003) for further details on variance reduction techniques.

### 4.2. Asymptotic validity - Convergence rate

Here we establish the asymptotic validity of the bootstrap distribution estimate output by Algorithm 1. In order to formulate the result rigorously, we introduce further notation. Let

$$H_{n,\epsilon}(t) = \mathbb{P}\left\{\sup_{\beta\in[\epsilon,1-\epsilon]} |R_n(\beta)| \leq t\right\}$$

be the the root's cumulative distribution function and

$$H_{n,\epsilon}^{boot}(t) = \mathbb{E}^*\left[\gamma_n \cdot \mathbb{I}\left\{\sup_{\beta\in[\epsilon,1-\epsilon]} R_n^*(\beta)| \leq t\right\}\right]$$

its bootstrap counterpart, produced by Algorithm 1. In the subsequent analysis, the kernel $K$ used in the smoothing step is supposed to be either gaussian or else of the form $\mathbb{I}_{\{u\in[-1,+1]\}}$.

**Theorem 2** *(Asymptotic accuracy) Assume that the assumptions of Theorem 1 are satisfied. Suppose in addition that smoothed versions of the cdfs $\widetilde{G}$ and $\widetilde{H}$ are computed at step 1 using a scaled kernel $\mathbf{K}_{h_n}(u)$ with $h_n \downarrow 0$ as $n \to \infty$ in a way that $nh_n^3 \to \infty$ and $nh_n^5 \log^2 n \to 0$. Then, we have as $n \to \infty$:*

$$\sup_{t\in\mathbb{R}_+} |H_{n,\epsilon}(t) - H_{n,\epsilon}^{boot}(t)| = o_{\mathbb{P}}\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right) ,$$

*where the notation $U_n = o_{\mathbb{P}}(a_n)$ designates a random variable such that $U_n/a_n$ converges to 0 in probability.*

Picking the bandwidth $h_n$ of order $1/(\log^{2+\eta} n^{1/5})$ with $\eta > 0$ thus leads to an approximation error of order $n^{-2/5}$, up to log factors, for the bootstrap distribution estimate. This rate is slower than the one of the gaussian approximation given in part (ii) of Theorem 1, the PR bootstrap algorithm is however very appealing from a computational angle. The construction of gaussian confidence bands from estimates of the class densities $g_s(z)$ and $h_s(z)$ and simulated brownian bridges is indeed very challenging to implement.

In addition, we point out that the gain acquired through smoothing is significant in terms of convergence rate. In absence of it, it may be easily shown that the pointwise rate of approximation would have been then of order $O(n^{-1/4})$, due to the order of magnitude of the fluctuations of the bootstrap quantile $G_s^{*-1}(1-\beta)$ around its expected value $\widehat{G}_s^{-1}(1-\beta)$ given the data $\mathcal{D}_n$, see (Falk & Reiss, 1989). Eventually, we underline that, in the pointwise setup (*i.e.* for a fixed precision $\beta \in [\epsilon, 1-\epsilon]$), the rate of convergence may

be improved by bootstrapping a *studentized* version of the deviation $R_n(\beta)$. Given the complexity of the asymptotic variance $\sigma_s^2(\beta)$, see Eq. (9), a bootstrap version of the square root of $R_n(\beta)$'s variance $\widehat{\sigma}_s^2(\beta)$ should naturally be prefered to a plug-in estimate as renormalization factor. Precisely, if the standardization $\widehat{\sigma}_s^2(\beta)$ is computed by means of a smoothed bootstrap with a bandwidth of order $n^{-1/3}$ (different thus from the one used in the resampling step of Algorithm 1), it may be established that $R_n(\beta)/\widehat{\sigma}_s(\beta)$'s distribution is approximated by the resulting bootstrap distribution at the rate $n^{-2/3}$, faster then than by the gaussian approximation provided by the Central Limit Theorem. Due to space limitations, details are omitted here.

## 5. Conclusion

In the paper, we established statistical properties of the Precision-Recall curve. We provided theoretical arguments in favor of using PR curve constructed from data as a proxy to the unknown expected PR curve resulting from the underlying distribution of the data. In particular, we showed consistency and established the asymptotic approximation rate under the supremum norm. We also proposed a practical algorithm for building confidence bands of the PR curve based on smoothed bootstrap and importance sampling. Eventually, the asymptotic validity of such a procedure is proved for a careful tuning of the regularization parameter involved in the smoothing kernel.

## Appendix - Technical Proofs

**Proof of Theorem 1.** In order to establish the result, we introduce $Z^+ = \{Z_n^+\}_{n\geq 1}$ and $Z^- = \{Z_n^-\}_{n\geq 1}$ two independent sequences of i.i.d. random variables with distributions $G_s$ and $H_s$ respectively, as well as $\{Y_n\}_{n\geq 1}$ a sequence of i.i.d. binary random variables, independent from $Z^+$ and $Z^-$ and such that $p = \mathbb{P}\{Y_1 = +1\} = 1 - \mathbb{P}\{Y_n = -1\}$. For all $n \geq 1$ and $z \in \mathbb{R}$, we set $\widehat{G}_{s,n}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Z_i^+ \leq t\}$, $\widehat{H}_{s,n}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Z_i^- \leq t\}$, as well as $\widehat{\alpha}_{m,q}(s,\beta) = 1 - \widehat{H}_{s,m} \circ \widehat{G}_{s,q}(1-\beta)$ for all $(m,q) \in \mathbb{N}^{*2}$. Set also $Z_n = Z_{n_+}^+ \cdot \mathbb{I}\{Y_n = +1\} + Z_{n_-}^- \cdot \mathbb{I}\{Y_n = -1\}$ for all $n \geq 1$, we point out that the collection $\{(Z_n, Y_n)\}_{n\geq 1}$ forms a sequence of i.i.d. copies of the random pair $(Z, Y)$ and, equipped with these notations, we have: $\forall \beta \in (0,1)$, $\widehat{\alpha}(s,\beta) = \widehat{\alpha}_{n_+,n_-}(s,\beta)$.

The proof is based on the following lemma, describing the asymptotic behavior of the stochastic process:

$$r_{m,n}(\beta) = \sqrt{n}\left(\widehat{\alpha}_{m,n-m}(s,\beta) - \alpha(s,\beta)\right),\ \beta \in (0,1).$$

**Lemma 3** *Under assumptions $H_1 - H_2$, if $m \to \infty$ in a way that $m/n \to p$ as $n \to \infty$, we have:*

*(i)* $\sup_{\beta\in[\epsilon,1-\epsilon]} |\widehat{\alpha}_{m,n-m}(s,\beta) - \alpha(s,\beta)| \underset{n\to\infty}{\to} 0$ *a.s.,*

*(ii) there exists a probability space on which one can define two independent sequences of brownian bridges $B_1^{(n)}$ and $B_2^{(n)}$ such that we have, with probability one, as $n \to \infty$,*

$$r_{m,n}(\beta) = z^{(n)}(\beta) + o\left(\frac{(\log\log n)^{\rho_1(\gamma)}(\log n)^{\rho_2(\gamma)}}{\sqrt{n}}\right),$$

*uniformly over $[\epsilon, 1-\epsilon]$, where: $\forall n \geq 1$,*

$$z^{(n)}(\beta) = p^{-1/2}\frac{h_s(G_s(1-\beta))}{g_s(G_s(1-\beta))}B_1^{(n)}(\beta) + (1-p)^{-1/2}B_2^{(n)}(\alpha(s,\beta)).$$

Proof of the lemma. This directly follows from Theorems 2.1 and 2.2 in (Hsieh & Turnbull, 1996), see also Theorem 1 in (Bertail et al., 2008), which results are based on standard results in strong approximation theory, refer to (Csorgo & Revesz, 1981). □

Now, for any $\beta \in (0,1)$ consider the mapping $F_\beta(u,v) = \beta u/(\beta u + (1-u)v)$ defined on $(0,1)^2$. Notice that $PR_s(\beta) = F_\beta(p, \alpha(s,\beta))$ and $\widehat{PR}_s(\beta) = F_\beta(n_+/n, \widehat{\alpha}(s,\beta))$. We also introduce $\widehat{PR}_{s,q}(\beta) = F_\beta(n_+/n, \widehat{\alpha}_{m,n-m}(s,\beta))$. For all $\beta \in [\epsilon, 1-\epsilon]$, the function $F_\beta$ is $\mathcal{C}^2$ on $[\epsilon,1-\epsilon]\times[0,1]$ and one may check that the norm of its Hessian matrix is bounded on $[\epsilon,1-\epsilon]\times[0,1]$, uniformly over $\beta \in [\epsilon, 1-\epsilon]$. Hence, combining a Taylor expansion of $F_\beta$ at the first order with Lemma 3 and the Law of Iterated Logarithm applied to the standardized binomial $\sqrt{n/(p(1-p))}(n_+/n - p)$, we obtain that

$$\sqrt{n}\left(\widehat{PR}_{s,m}(\beta) - PR_s(\beta)\right) = W_{n,m}(s,\beta) + o\left(\frac{(\log\log n)^{2\rho_1(\gamma)}(\log n)^{2\rho_2(\gamma)}}{n}\right),$$

as $m$ and $n$ tend to infinity so that $m/n \sim p$, where

$$W_{n,m}(s,\beta) = \sqrt{n}\left(\frac{n_+}{n} - p\right)\cdot\frac{\partial F_\beta}{\partial u}(p,\alpha(s,\beta)) + r_{m,n}(\beta)\cdot\frac{\partial F_\beta}{\partial v}(p,\alpha(s,\beta)).$$

One may then conclude the proof by using Lemma 3's second assertion, combined with Prohorov 's theorem (guaranteeing the existence of a version of $Y_n$ such that $(n_+ - np)/\sqrt{np(1-p)}$ almost surely converges to a

r.v. $Z \sim \mathcal{N}(0,1)$ independent from $B^{(1)}$ and $B^{(2)}$) and the fact that $n_+/n \to p$ a.s. as $n \to \infty$.

**Proof of Theorem 2 (Sketch of).** The proof results from the strong approximation stated in Theorem 1, combined with a standard coupling argument. The main steps of the argument are as follows. Set $\widetilde{g}_s(z) = d\widetilde{G}_s(z)/dz$, $\widetilde{h}_s(z) = d\widetilde{H}_s(z)/dz$ and $\widetilde{\alpha}_s = 1 - \widetilde{H}_s \circ \widetilde{G}_s$. Let $\mathbb{P}^*$ the conditional probability given the data $\mathcal{D}_n$ and consider $\widehat{\mathbb{P}}^*$ the equivalent probability measure defined by: $\forall n \geq 1$, $d\widehat{\mathbb{P}}^*/d\mathbb{P}\,|_{\mathcal{F}_n} = \gamma_n$ where $\mathcal{F}_n$ is the $\sigma$-field generated by $\mathcal{D}_n$. Note first that, conditioned on $\mathcal{D}_n$, by applying Theorem 1's part (ii) with $\widehat{PR}_s$ as target curve (instead of $PR_s$), one gets that, with probability 1 under $\widehat{\mathbb{P}}^*$, $R_n^*(\beta)$ is equivalent to a stochastic process with same law as

$$Z^{(n)*}(\beta) = \frac{\widetilde{PR}_s(\beta)^2}{\beta}\left(\frac{a_n}{\sqrt{\widehat{p}_n}}\widetilde{\alpha}(s,\beta)W + a_n^2\sqrt{\widehat{p}_n}\frac{\widetilde{h}_s(\widetilde{G}_s(1-\beta))}{\widetilde{g}_s(\widetilde{G}_s(1-\beta))}B_1^{(n)}(\beta) + a_n B_2^{(n)}(\widetilde{\alpha}(s,\beta))\right),$$

with a remainder of order $o(\frac{(\log\log n)^{\rho_1(\gamma)}\log^{\rho_2(\gamma)} n}{\sqrt{n}})$, uniformly over $[\epsilon, 1-\epsilon]$. In the last display, we have set: $a_n = \sqrt{1-\widehat{p}_n}/\widehat{p}_n$. In addition, we a.s. have

$$\frac{\widetilde{h}_s(\widetilde{G}_s^{-1}(1-\beta))}{\widetilde{g}_s(\widetilde{G}_s^{-1}(1-\beta))} - \frac{h_s(G_s^{-1}(1-\beta))}{g_s(G_s^{-1}(1-\beta))} = O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right),$$

under the stipulated conditions, see (Giné & Guillou, 2002). In addition, from standard result on the modulus of continuity of the brownian bridge (Shorack & Wellner, 1986), we a.s. have: as $n \to \infty$,

$$\sup_{\alpha\in[0,1]}|B_2^{(n)}(\alpha_s(\beta)) - B_2^{(n)}(\widetilde{\alpha}_s(\beta))| = O\left(\frac{\log n}{n^{1/2}}\right).$$

Applying then the Law of Iterated Logarithm to $\widehat{p}_n$, it follows that almost surely, uniformly over $[\epsilon, 1-\epsilon]$,

$$Z^{(n)*}(\alpha) = Z^{(n)} + O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right).$$

Since this results holds uniformly in $\beta \in [\epsilon, 1-\epsilon]$, by virtue of the continuous mapping theorem applied to the function $\sup_{\beta\in[\epsilon,1-\epsilon]}(.)$, the result also holds in distribution up to same order.

## References


Bertail, P., Clémençon, S., & Vayatis, N. (2008). On bootstrapping the ROC curve. *In Proc. of Neur. Inf. Proc. Syst. 2008, Vancouver, Canada.*

Bucklew, J. (2003). Introduction to rare event simulation. *Springer.*

Clémençon, S., & Vayatis, N. (2008). Tree-structured ranking rules and approximation of the optimal ROC curve. *Proceedings of the 2008 conference on Algorithmic Learning Theory.* Lect. Notes Art. Int. 5254, pp. 22-37, Springer.

Csorgo, M., & Revesz, P. (1981). Strong approximations in probability and statistics. *Academic Press.*

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *In Proceedings of the 23 rd International Conference on Machine Learning, Vol. 148, pp. 233–240.*

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, *7*, 1–26.

Falk, M., & Reiss, R. (1989). Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *Annals of Probability*, *17*, 362–371.

Giné, E., & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Poincaré (B), Probabilités et Statistiques*, *38*, 907–921.

Horvath, L., Horvath, Z., & Zhou (2008). Confidence bands for ROC curves. *Journal of Statistical Planning and Inference*, *138*, 1894–1904.

Hsieh, F., & Turnbull, B. (1996). Nonparametric and semi-parametric statistical estimation of the ROC curve. *The Annals of Statistics*, *24*, 25–40.

Macskassy, S., & Provost, F. (2004). Confidence bands for ROC curves: methods and an empirical study. *In Proceedings of the first Workshop on ROC Analysis in Artif. Int. at Eur. Conf. on Artif. Int. 2004.*

Macskassy, S., Provost, F., & Rosset, S. (2005). Bootstrapping the ROC curve: an empirical evaluation. *In Proceedings of Int. Conf. Mach. Learn.-2005 Workshop on ROC Analysis in Machine Learning.*

Manning, C. M., & Schutze, H. (1999). Foundations of statistical natural language processing. *MIT Press.*

Raghavan, V., Bollmann, P., & Jung, G. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, *7*, 205–229.

Shao, G., & Tu, J. (1995). The jackknife and bootstrap. *Springer, NY.*

Shorack, G., & Wellner, J. (1986). Empirical processes with applications to statistics. *Wiley, NY.*

Silverman, B., & Young, G. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, *74*, 469–479.