On the Path to an Ideal ROC Curve: Considering Cost Asymmetry in Learning Classifiers

Francis R. Bach, David Heckerman, Eric Horvitz

Technical Report MSR-TR-2004-124

November 18, 2004

Abstract

Receiver Operating Characteristic (ROC) curves are a standard way to display the performance of a set of binary classifiers for all feasible ratios of the costs associated with false positives and false negatives. For linear classifiers, the set of classifiers is typically obtained by training once, holding constant the estimated slope and then varying the intercept to obtain a parameterized set of classifiers whose performances can be plotted in the ROC plane. In this paper, we consider the alternative of varying the asymmetry of the cost function used for training. We show that the ROC curve obtained by varying the intercept and the asymmetry—and hence the slope—always outperforms the ROC curve obtained by varying only the intercept. In addition, we present a path-following algorithm for the support vector machine (SVM) that can compute efficiently the entire ROC curve, that has the same computational properties as training a single classifier. Finally, we provide a theoretical analysis of the relationship between the asymmetric cost model assumed when training a classifier and the cost model assumed in applying the classifier. In particular, we show that the mismatch between the step function used for testing and its convex upper bounds usually used for training leads to a provable and quantifiable difference around extreme asymmetries.

1 Introduction

Receiver Operating Characteristic (ROC) analysis has seen increasing attention in the recent statistics and machine-learning literature (Pepe, 2000, Provost and Fawcett, 2001, Flach, 2003). The ROC is a representation of choice for displaying the performance of a classifier when the costs assigned by end users to false positives and false negatives are not known at the time of training. For example, when training a classifer for identifying cases of undesirable unsolicited email, end users may have different preferences about the likelihood of a false negative and false positive. The ROC curve for such a classifier reveals the ratio of false negatives and positives at different probability thresholds for classifying an email message as unsolicited or normal email.

In this paper, we consider linear binary classification of points in an Euclidean space—noting that it can be extended in a straightforward manner to non-linear classification problems by using Mercer kernels (Schölkopf and Smola, 2002). That is, given data $x \in \mathbb{R}^d$, $d \geqslant 1$, we consider classifiers of the form $f(x) = \text{sign}(w^\top x + b)$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are referred to as the *slope* and the *intercept*. To date, ROC curves have been usually constructed by training once, holding constant the estimated slope and varying the intercept to obtain the curve. In this paper, we show that, while the latter procedure appears to be the most practical thing to do, it may lead to classifiers with poor performance in some parts of the ROC curve.

The crux of our approach is that we allow the asymmetry of the cost function to vary—i.e., we vary the ratio of the cost of a false positive and the cost of a false negative. For each value of the ratio, we obtain a different slope and intercept, each optimized for this ratio. In a naive implementation, varying the asymmetry would require a retraining of the classifier for each point of the ROC curve, which would be computationally expensive. In Section 3.1, we present an algorithm that can compute the solution of an SVM (Schölkopf and Smola, 2002) for all possible costs of false positives and false negatives, with the same computational complexity as obtaining the solution for only one cost function. The algorithm extends to asymmetric costs the algorithm of Hastie et al. (2005) and is based on path-following techniques that take advantage of the piecewise linearity of the path of optimal solutions. In Section 3.2, we show how the path-following algorithm can be used to obtain the best possible ROC curve (in expectation). In particular, by allowing both the asymmetry and the intercept to vary, we can obtain provably better ROC curves than by methods that simply vary the intercept.

In Section 4, we provide a theoretical analysis of the link between the asymmetry of costs assumed in training a classifier and the asymmetry desired in its application. In particular, we show that—even in the population (*i.e.*, infinite sample) case—the use of a training loss function which is a convex upper bound on the true or testing loss function (a step function) creates classifiers with sub-optimal accuracy. We quantify this problem around extreme asymmetries for several classical convex-upper-bound loss functions—the square loss and the *erf loss*, an approximation of the logistic loss based on normal cumulative distribution functions (also referred to as the "error function", and usually abbreviated as erf). The analysis is carried through for Gaussian and mixture of Gaussian class-conditional distributions (see Section 4 for more details). As we shall see, the consequences of the potential mismatch between the cost functions assumed in testing versus training underscore the value of using the algorithm that we introduce in Section 4.3. Even when costs are known (i.e., when only one point on the ROC curve is needed), the classifier resulting from our approach which builds the entire ROC curve is never less accurate and can be more accurate than one trained with the known costs using a convex-upper-bound loss function.

2 Problem overview

Given data $x \in \mathbb{R}^d$ and labels $y \in \{-1, 1\}$, we consider linear classifiers of the form $f(x) = \operatorname{sign}(w^\top x + b)$, where w is the *slope* of the classifier and b the *intercept*. A classifier is determined by the parameters $(w, b) \in \mathbb{R}^{d+1}$. In Section 2.1, we introduce notation and definitions; in Section 2.2, we lay out the necessary concepts of ROC analysis. In Section 2.3, we describe how these classifiers and ROC curves are typically obtained from data.

2.1 Asymmetric cost and loss functions

Positive (resp. negative) examples are those for which $y = 1$ (resp. $y = -1$). The two types of misclassification, false positives and false negatives, are assigned two different costs, and the total *expected* cost is equal to

$$R(C_+, C_-, w, b) = C_+ P\{w^\top x + b < 0,\ y = 1\} + C_- P\{w^\top x + b \geqslant 0,\ y = -1\}$$

If we let $\phi_{0-1}(u) = 1_{u<0}$ be the *0-1 loss*, we can write the expected cost as

$$R(C_+, C_-, w, b) = C_+ E\{1_{y=1}\phi_{0-1}(w^\top x + b)\} + C_- E\{1_{y=-1}\phi_{0-1}(-w^\top x - b)\}$$

where E denotes the expectation with respect to the joint distribution of (x, y). The expected cost defined using the 0-1 loss is the cost that end users are usually interested in during the use of the

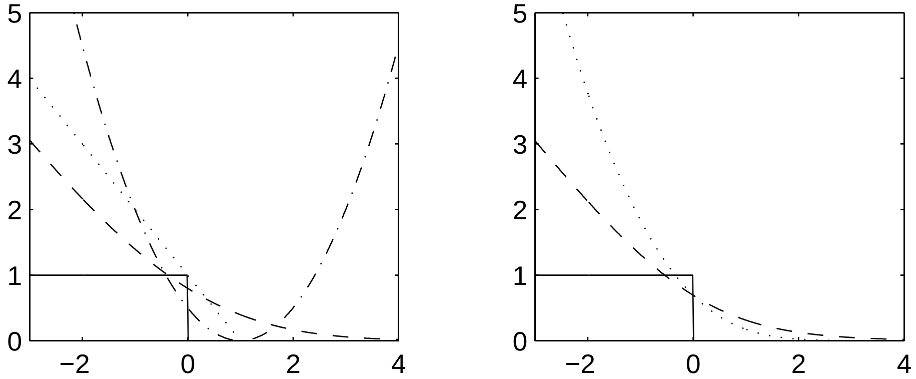

Figure 1: Loss functions: (left) plain: 0-1 loss, dotted: hinge loss, dashed: erf loss, dash-dotted: square loss. (Right) plain: 0-1 loss, dotted: probit loss, dashed: logistic loss.

classifier, while the other cost functions that we define below are used solely for training purposes. The convexity of these cost functions makes learning algorithms convergent without local minima, and leads to attractive asymptotic properties (Bartlett et al., 2004).

A traditional set-up for learning linear classifiers from labeled data is to consider a convex upper bound ϕ on the 0-1 loss ϕ_{0-1}, and use the expected ϕ-cost :

$$R_\phi(C_+, C_-, w, b) = C_+ E\{1_{y=1}\phi(w^\top x + b)\} + C_- E\{1_{y=-1}\phi(-w^\top x - b)\}$$

We refer to the ratio $C_+/(C_- + C_+)$ as the *asymmetry*. We shall use *training asymmetry* to refer to the asymmetry used for training a classifier using a ϕ-cost, and the *testing asymmetry* to refer to the asymmetric cost characterizing the testing situation (reflecting end user preferences) with the actual cost based on the 0-1 loss. In Section 4, we will show that these may be different in the general case.

We shall consider several common loss functions. Some of the loss functions (square loss, hinge loss) lead to attractive computational properties, while others (square loss, erf loss) are more amenable to theoretical manipulations (see Figure 1 for the plot of the loss functions, as they are commonly used and defined below[1]):

- **square loss** : $\phi_{sq}(u) = \frac{1}{2}(u-1)^2$; the classifier is equivalent to linear regression,

- **hinge loss** : $\phi_{hi}(u) = \max\{1-u, 0\}$; the classifier is the support vector machine (Schölkopf and Smola, 2002),

- **erf loss** : $\phi_{erf}(u) = 2\left[\frac{u}{2}\psi\left(\frac{u}{2}\right) - \frac{u}{2} + \psi'\left(\frac{u}{2}\right)\right]$, where ψ is the cumulative distribution of the standard normal distribution, i.e. : $\psi(v) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{v} e^{-t^2/2}dt$, and $\psi'(v) = \frac{1}{\sqrt{2\pi}}e^{-v^2/2}$. The erf loss provides a good approximation of the *logistic loss* $\log(1+e^{-u})$ as well as its derivative, and is amenable to closed-form computations for Gaussians and mixture of Gaussians (see Section 4 for more details). Note that the erf loss is different from the *probit loss* $-\log\psi(u)$, which leads to probit regression (Hastie et al., 2001).

[1] Note that by rescaling, all of these loss functions can be made to be an upper bound on the 0-1 loss which is tight at zero.

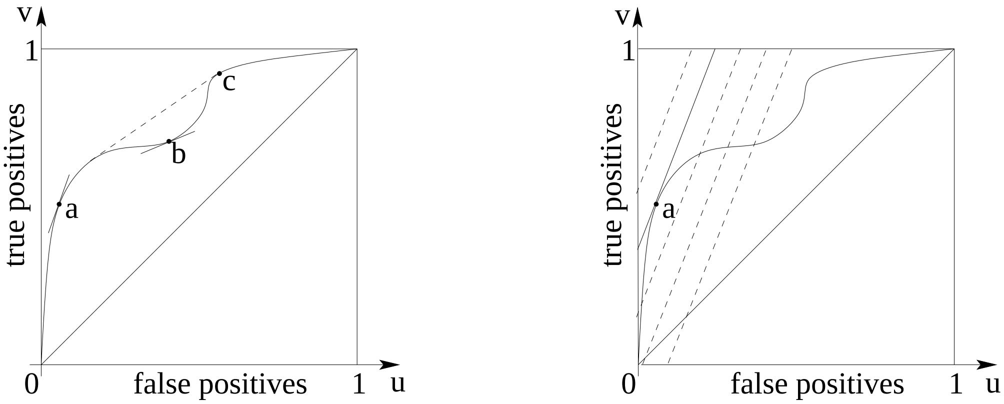

Figure 2: (Left) ROC curve: (plain) regular ROC curve; (dashed) convex envelope. The points a and c are ROC-consistent and the point b is ROC-inconsistent. (Right) ROC curve and dashed equi-cost lines: All lines have direction (p_+C+, p_-C_-), the plain line is optimal and the point "a" is the optimal classifier.

2.2 ROC analysis

The aim of ROC analysis is to display in a single graph the performance of classifiers for all possible costs of misclassification. In this paper, we consider sets of classifiers $f_\gamma(x)$, parameterized by a variable $\gamma \in \mathbb{R}$ (γ can either be the intercept or the training asymmetry).

For a classifier $f(x)$, we can define a point (u, v) in the "ROC plane," where u is the proportion of false positives $u = P(f(x) = 1|y = -1)$, and v is the proportion of true positives $v = P(f(x) = 1|y = 1)$.

When γ is varied, we obtain a curve in the ROC plane, the ROC curve (see Figure 2 for an example). Whether γ is the intercept or the training asymmetry, the ROC curve always passes through the point $(0, 0)$ and $(1, 1)$, which corresponds to classifying all points as negative (resp. positive).

The upper convex envelope of the curve is the set of optimal ROC points that can be achieved by the set of classifiers; indeed, if a point in the envelope is not one of the original points, it must lie in a segment between two points $(u(\gamma_0), v(\gamma_0))$ and $(u(\gamma_1), v(\gamma_1))$, and all points in a segment between two classifiers can always be attained by choosing randomly between the two classifiers (note that this classifier itself is not a linear classifier; this performance can only be achieved by a mixture of two linear classifiers).

Denoting $p_+ = P(y = 1)$ and $p_- = P(y = -1)$, the expected (C_+, C_-)-cost for a classifier (u, v) in the ROC space, is simply $p_+C_+(1 - v) + p_-C_-u$, and thus optimal classifiers for the (C_+, C_-)-cost can be found by looking at lines of slope that are normal to $(p_-C_-, -p_+C_+)$—and thus proportional to (p_+C_+, p_-C_-)—and which intersects the ROC curve and are as close as the point $(0, 1)$ as possible (see Figure 2).

A point $(u(\gamma), v(\gamma))$ is said to be *ROC-consistent* if it lies on the upper convex envelope; In this case, the tangent direction $(du/d\gamma, dv/d\gamma)$ defines a cost $(C_+(\gamma), C_-(\gamma))$ for which the classifier is optimal (for the testing cost, which is defined using the 0-1 loss), by having $(p_+C_+(\gamma), p_-C_-(\gamma))$ proportional to $(du/d\gamma, dv/d\gamma)$. This leads to an *optimal testing asymmetry*

$$\beta(\gamma) \triangleq \frac{C_+(\gamma)}{C_+(\gamma) + C_-(\gamma)} = \frac{1}{1 + \frac{p_+}{p_-}\frac{dv}{d\gamma}(\gamma)/\frac{du}{d\gamma}(\gamma)}.$$

If a point $(u(\gamma), v(\gamma))$ is ROC-inconsistent, then the quantity $\beta(\gamma)$ has no meaning, and such a classifier is generally useless, because, for all settings of the misclassification cost, that classifier can be outperformed by the other classifiers or a combination of classifiers.

In Section 4, we relate the optimal asymmetry of cost in the testing or eventual use of a classifer in the real world, to the asymmetry of cost used to train that classifier; in particular, we show that they differ and quantify this difference for extreme asymmetries (*i.e.*, close to the corner points $(0, 0)$ and $(1, 1)$). This analysis highlights the value of generating the entire ROC curve, even when only one point is needed, as we will present in Section 4.3.

2.3 Learning from data

Given n labelled data points (x_i, y_i), the *empirical cost* is equal to:

$$\widehat{R}(C_+, C_-, w, b) = \frac{C_+}{n}\#\{y_i(w^\top x_i + b) < 0, y_i = 1\} + \frac{C_-}{n}\#\{y_i(w^\top x_i + b) < 0, y_i = -1\}$$

while the *empirical ϕ-cost* is equal to

$$\widehat{R}_\phi(C_+, C_-, w, b) = \frac{C_+}{n}\sum_{i \in \mathcal{I}_+} \phi(y_i(w^\top x_i + b)) + \frac{C_-}{n}\sum_{i \in \mathcal{I}_-} \phi(y_i(w^\top x_i + b)),$$

where $\mathcal{I}_+ = \{i, y_i = 1\}$ and $\mathcal{I}_- = \{i, y_i = -1\}$. When learning a classifier from data, a classical setup is to minimize the sum of the *empirical ϕ-cost* and a regularization term $\frac{1}{2n}||w||^2$, *i.e.*, to minimize $\widehat{J}_\phi(C_+, C_-, w, b) = \widehat{R}_\phi(C_+, C_-, w, b) + \frac{1}{2n}||w^2||$.

Note that the objective function is no longer homogeneous in (C_+, C_-); the sum $C_+ + C_-$ is referred to as the total amount of regularization. Thus, two end-user–defined parameters are needed to train a linear classifier: the *total amount of regularization* $C_+ + C_- \in \mathbb{R}^+$, and the *asymmetry* $\frac{C_+}{C_+ + C_-} \in [0, 1]$. In Section 3.1, we show how the minimum of $\widehat{J}_\phi(C_+, C_-, w, b)$, with respect to w and b, can be computed efficiently for the hinge loss, for many values of (C_+, C_-), with a computational cost that is within a constant factor of the computational cost of obtaining a solution for one value of (C_+, C_-).

Building an ROC curve from data If a sufficiently large validation set is available, we can train on the training set and use the empirical distribution of the validation data to plot the ROC curve. If sufficient validation data is not available, then we can use 10 random half splits of the data, train a classifier on one half and use the other half to obtain the ROC scores. Then, for each value of the parameter γ that defines the ROC curve (either the intercept or the training asymmetry), we average the 10 scores. We can also use this approach to obtain confidence intervals (Flach, 2003).

3 Building ROC curves for the SVM

In this section, we will present an algorithm to compute ROC curves for the SVM that explores the two-dimensional space of cost parameters (C_+, C_-) efficiently. We first show how to obtain optimal solutions of the SVM without solving the optimization problems many times for each value of (C_+, C_-). This method generalizes the results of Hastie et al. (2005) to the case of asymmetric cost functions. We then describe how the space (C_+, C_-) can be appropriately explored and how ROC curves can be constructed.

3.1 Building paths of classifiers

Given n data points x_i, $i = 1, \ldots, n$ which belong to $\mathbb{R}^d$, and n labels $y_i \in \{-1, 1\}$, minimizing the regularized empirical hinge loss is equivalent to solve the following convex optimization problem Schölkopf and Smola (2002):

$$\min_{w,b,\xi} \sum_i C_i \xi_i + \frac{1}{2}||w||^2 \quad \text{such that} \quad \forall i,\ \xi_i \geqslant 0,\ \xi_i \geqslant 1 - y_i(w^\top x_i + b)$$

where $C_i = C_+$ if $y_i = 1$ and $C_i = C_-$ if $y_i = -1$.

Optimality conditions and dual problems The Lagrangian of the problem is (with dual variables $\alpha, \beta \geqslant 0$):

$$L(w, b, \alpha, \beta) = \sum_i C_i \xi_i + \frac{1}{2}||w||^2 + \sum_i \alpha_i (1 - y_i(w^\top x_i + b)) - \xi_i) - \sum_i \beta_i \xi_i$$

The derivatives with respect to the primal variables are

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i x_i\ , \quad \frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i\ , \quad \frac{\partial L}{\partial \xi_i} = C_i - \alpha_i - \beta_i\ ,$$

The slackness conditions are

$$\forall i,\ \alpha_i(1 - y_i(w^\top x_i + b)) = 0 \ \text{ and } \ \beta_i \xi_i = 0$$

and the dual problem is the following (where $\widetilde{K} = \operatorname{Diag}(y) K \operatorname{Diag}(y)$) :

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{2}\alpha^\top \widetilde{K} \alpha + 1^\top \alpha \quad \text{such that} \quad \alpha^\top y = 0, \quad \forall i,\ 0 \leqslant \alpha_i \leqslant C_i$$

Piecewise affine solutions For an optimal set of primal-dual variables (w, b, α), we can separate data points in three disjoint sets:

$$\begin{array}{lllll}
\text{Margin :} & \mathcal{M} & = \{i, & \alpha_i \in [0, C_i], & y_i(w^\top x_i + b) = 1 \quad \} \\
\text{Left of margin :} & \mathcal{L} & = \{i, & \alpha_i = C_i, & y_i(w^\top x_i + b) < 1 \quad \} \\
\text{Right of margin :} & \mathcal{R} & = \{i, & \alpha_i = 0, & y_i(w^\top x_i + b) > 1 \quad \}
\end{array}$$

These sets are usually referred to as *active sets* Boyd and Vandenberghe (2003). If the sets $\mathcal{M}$, $\mathcal{L}$ and $\mathcal{R}$ are known, then α, b are optimal if and only if :

$$\begin{aligned}
&\forall i \in \mathcal{L}, \alpha_i = C_i \\
&\forall i \in \mathcal{R}, \alpha_i = 0 \\
&\forall i \in \mathcal{M}, (\widetilde{K}\alpha)_i + b y_i = 1 \\
&\alpha^\top y = 0
\end{aligned}$$

This is a linear system with as many equations as unknowns (i.e., $n + 1$). The real unknowns (not clamped to C_i or 0) are $\alpha_{\mathcal{M}}$ and b, and the smaller system is

$$\begin{aligned}
\widetilde{K}_{\mathcal{M},\mathcal{M}} \alpha_{\mathcal{M}} + b y_{\mathcal{M}} &= 1_{\mathcal{M}} - \widetilde{K}_{\mathcal{M},\mathcal{L}} C_{\mathcal{L}} \\
y_{\mathcal{M}}^\top \alpha_{\mathcal{M}} &= -y_{\mathcal{L}}^\top C_{\mathcal{L}}
\end{aligned}$$

whose solution is affine in $C_{\mathcal{L}}$ and thus in C.

Consequently, for known active sets, the solution is affine in C, which implies that the optimal variables (w, α, b) are piecewise affine continuous functions of the vector C. In our situation, C depends linearly on C_+ and C_-, and thus the path is piecewise affine in (C_+, C_-).

Following a path The active sets (and thus the linear system) remain the same as long as all constraints defining the active sets are satisfied, i.e., (a) $y_i(w^\top x_i + b) - 1$ is positive for all $i \in \mathcal{R}$ and negative for all $i \in \mathcal{L}$, and (b) for each $i \in \mathcal{M}$, α_i remains between 0 and C_i. This defines a set of linear inequalities in (C_+, C_-). The facets of the polytope defined by these inequalities can always be found in linear time in n, if efficient convex hull algorithms are used Avis et al. (1997). However, when we only follow a straight line in the (C_+, C_-)-space, the polytope is then a segment and its extremities are trivial to find (also in linear time $O(n)$).

Following Hastie et al. (2005), if a solution is known for one value of (C_+, C_-), we can follow the path along a line, by monitoring which constraints are violated at the boundary of the polytope that defines the allowed domain of (C_+, C_-) for the given active sets.

Numerical issues However, several numerical issues have to be solved before the previous approach can be made efficient and stable. Some of the issues directly follow the known issues of the simplex method for linear programming (which is itself an active set method) Maros (2002).

- *Path initialization*: In order to easily find a point of entry into the path, we look at situations when all points are in $\mathcal{L}$. In order to verify $\alpha^\top y = 0$, this implies that $\sum_{i \in \mathcal{I}_+} C_i = \sum_{i \in \mathcal{I}_-} C_i$, i.e., this means $C_+ n_+ = C_- n_-$ (where $n_+ = |\mathcal{I}_+|$ and $n_- = |\mathcal{I}_-|$), which we now assume. This is valid as long as $\forall i, y_i(w^\top x_i + b) \leqslant 1$, i.e.:

  $$\begin{aligned} \forall i \in \mathcal{I}_+ \quad &, \quad b \leqslant 1 - C_+ \left((\widetilde{K}\delta_+)_i + \frac{n_+}{n_-} (\widetilde{K}\delta_-)_i \right) \\ \forall i \in \mathcal{I}_- \quad &, \quad b \geqslant -1 + C_+ \left((\widetilde{K}\delta_+)_i + \frac{n_+}{n_-} (\widetilde{K}\delta_-)_i \right) \end{aligned}$$

  Let us define the following two maxima: $m_+ = \max_{i \in \mathcal{I}_+} \left((\widetilde{K}\delta_+)_i + \frac{n_+}{n_-} (\widetilde{K}\delta_-)_i \right)$ (attained at i_+) and $m_- = \max_{i \in \mathcal{I}_-} \left((\widetilde{K}\delta_+)_i + \frac{n_+}{n_-} (\widetilde{K}\delta_-)_i \right)$ (attained at i_-). The previous conditions are equivalent to

  $$-1 + C_+ m_- \leqslant b \leqslant 1 - C_+ m_+$$

  Thus, all points are in $\mathcal{L}$ as long as $C_+ \leqslant 2/(m_- + m_+)$, and when this is verified, b is undetermined, between $-1 + C_+ m_-$ and $1 - C_+ m_+$. At the boundary point $C_+ = 2/(m_- + m_+)$, then both i_+ and i_- are going from $\mathcal{L}$ to $\mathcal{M}$.

  Since we vary both C_+ and C_- we are able to avoid to solve a quadratic program to enter the path, as is done in Hastie et al. (2005) when the datasets are not perfectly balanced.

- *Switching between active sets*: As in Hastie et al. (2005), indices can go from $\mathcal{L}$ to $\mathcal{M}$, $\mathcal{R}$ to $\mathcal{M}$, or $\mathcal{M}$ to $\mathcal{R}$ or $\mathcal{L}$. Empirically, when we follow a line in the plane (C_+, C_-), almost all points go from $\mathcal{L}$ to $\mathcal{R}$ through $\mathcal{M}$ (or from $\mathcal{R}$ to $\mathcal{M}$ through $\mathcal{M}$), with a few points going back and forth.

  Note that specific care has to be taken when $\mathcal{M}$ becomes empty.

- *Efficient implementation of linear system*: The use of Cholesky updating and downdating is necessary for stability and speed Golub and Van Loan (1996).

- *Computational complexity* Same as Hastie et al. (2005), i.e., same as obtaining the solution for one SVM. As shown by Hastie et al. (2005), if the appropriate online linear algebra tools are used, the complexity of obtaining one path of classifiers across one line is the same as obtaining the solution for one SVM using classical techniques such as sequential minimal optimization (Platt, 1998).

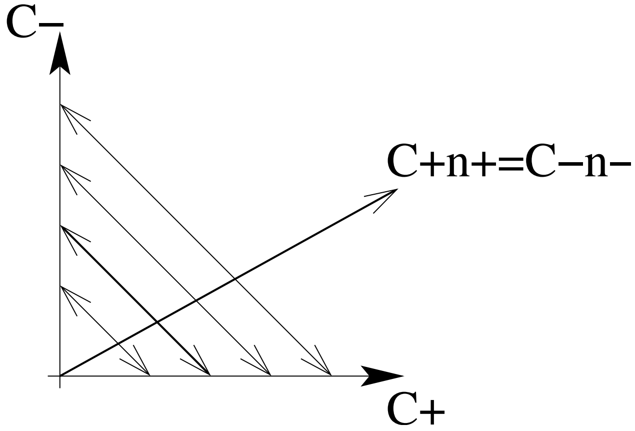

Figure 3: Lines in the (C_+, C_-)-space. The line $C_+n_+ = C_-n_-$ is always followed first; then several lines with constant $C_+ + C_-$ are followed in parallel, around the optimal line for the validation data (bold curve).

3.2 Constructing the ROC curve

Given the tools of Section 3.1, we can learn paths of linear classifiers from data. In this section, we present an algorithm to build ROC curves from the paths. We do this by exploring relevant parts of the (C_+, C_-) space, selecting the best classifiers among the ones that are visited.

We assume that we have two separate datasets, one for training and one for testing. This approach generalizes to cross-validation settings in a straightforward manner.

Exploration phase In order to start the path-following algorithm, we need to start at $C_+ = C_- = 0$ and follow the line $C_+n_+ = C_-n_-$. We follow this line up to a large upper bound on $C_+ + C_-$. For all classifiers along that line, we compute a misclassification cost on the testing set, with given asymmetry (C_+^0, C_-^0) (as given by the user, and usually, but not necessarily, close to a point of interest in the ROC space). We then compute the best performing pair (C_+^1, C_-^1) and we select pairs of the form (rC_+^1, rC_-^1), where r belongs to a set R of the type $R = \{1, 10, 1/10, 100, 1/100, \dots\}$. The set R provides further explorations for the total amount of regularization $C_+ + C_-$.

Then, for each r, we follow the paths of direction $(1, -1)$ and $(-1, 1)$ starting from the point (rC_+^1, rC_-^1). Those paths have a fixed total amount of regularization but vary in asymmetry. In Figure 3, we show all of lines that are followed in the (C_+, C_-) space.

Selection phase After the exploration phase, we have $|R|+1$ different lines in the (C_+, C_-) space: the line $C_-n_- = C_+n_+$, and the $|R|$ lines $C_+ + C_- = r(C_+^1 + C_-^1), r \in R$. For each of these lines, we know the optimal solution (w, b) for any cost settings on that line. The line $C_-n_- = C_+n_+$ is used for computational purposes (*i.e.*, to enter the path), so we do not use it for testing purposes.

From R lines in the (C_+, C_-)-plane, we build the three ROC curves shown in Figure 4, for a finite sample problem and for an infinite sample problem (for the infinite sample, the solution of the SVM was obtained by working directly with densities):

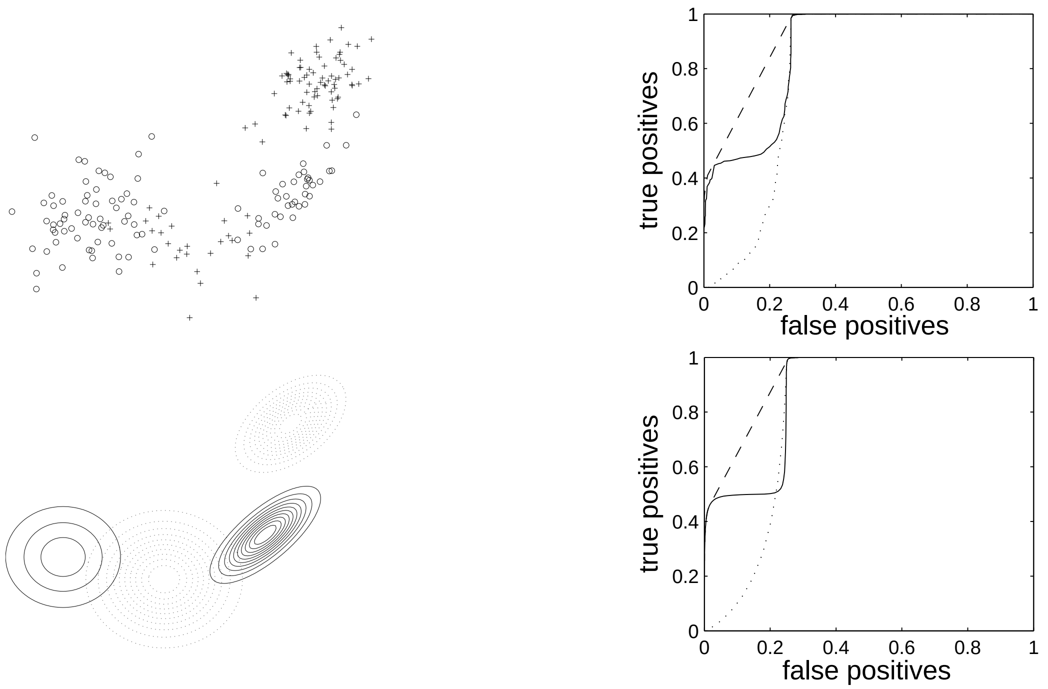

Figure 4: Two examples of ROC curves for bimodal class conditional densities, varying intercept (dotted), varying asymmetry (plain) and varying both (dashed). (Top) obtained from 10 random splits, using the data shown on the left side (one class is plotted as circles, the other one as crosses), (Bottom) obtained from population densities (one class with plain density contours, the other one with dotted contours).

- *Varying intercept*: we extract the slope w corresponding to the best setting $(C_+^1 + C_-^1)$, and vary the intercept b from $-\infty$ to ∞. This is the traditional method for building an ROC curve for an SVM.
- *Varying asymmetry*: we only consider the line $C_+ + C_- = C_+^1 + C_-^1$ in the (C_+, C_-)-plane; the classifiers that are used are the optimal solutions of the convex optimization problem. Note that for each value of the asymmetry, we obtain a different value of the slope and the intercept.
- *Varying intercept and asymmetry*: for each of the points on the R lines in the (C_+, C_-)-plane, we discard the intercept b and keep the slope w obtained from the optimization problem; we then use all possible intercept values; this leads to R two-dimensional surfaces in the ROC plane. We then compute the convex envelope of these, to obtain a single curve.

Since all classifiers obtained by varying only the intercept (resp. the asymmetry) are included in the set used for varying both the intercept and the asymmetry, the third ROC curve always outperforms the first two curves (*i.e.*, it is always closer to the top left corner). This is illustrated in Figure 4.

Intuitively, the ROC curve obtained by varying the asymmetry should be better than the ROC generated by varying the intercept because, for each point, the slope of the classifier is optimized. Empirically, this is generally true, but is not always the case, as displayed in Figure 4. This is not a small sample effect, as the infinite sample simulation shows. Another troubling fact is that the ROC curve obtained by varying asymmetry, exhibits strong concavities, *i.e.*, there are many ROC-inconsistent points: for those points, the solution of the SVM with the corresponding asymmetry is far from being the best linear classifier when performance is measured with the same asymmetry but with the exact 0-1 loss. In addition, even for ROC-consistent points, the training asymmetry and the testing asymmetry differ. In the next section, we analyze why they may differ and characterize their relationships in some situations.

4 Training vs. testing asymmetry

We observed in Section 3.2 that the training cost asymmetry can differ from the testing asymmetry. In this section, we analyze their relationships more closely for the population (*i.e.*, infinite sample) case. Although a small sample effect might alter some of the results presented in this section, we argue that most of the discrepancies come from using a convex surrogate to the $0-1$ loss.

The *Bayes optimal* classifier for a given asymmetry (C_+, C_-), is the (usually non-linear) classifier with minimal expected cost. A direct consequence of results in Bartlett et al. (2004) is that, if the Bayes optimal classifier is linear, then using a convex surrogate has no effect, *i.e.*, using the expected ϕ-cost will lead to the minimum expected cost. Thus, if the Bayes optimal classifier is linear, then, in the population case (infinite sample), there should be no difference. However, when the Bayes optimal classifier is not linear, then we might expect to obtain a difference, and we demonstrate that we do have one and quantify it for several situations.

Since we are using population densities, we can get rid of the regularization term and thus only the asymmetry will have an influence on the results, *i.e.*, we can restrict ourselves to $C_+ + C_- = 1$. We let $\gamma = C_+/(C_+ + C_-) = C_+$ denote the training asymmetry. For a given training asymmetry γ and a loss function ϕ, in Section 2.2, we defined the *optimal testing asymmetry* $\beta(\gamma)$ for the training asymmetry γ. In this section, we will refer to the $\beta(\gamma)$ simply as the *testing asymmetry.*

Although a difference might be seen empirically for all possible asymmetries, we analyze the relationship between the testing cost asymmetry and training asymmetry in cases of extreme asymmetry, *i.e.*, in the ROC framework, close to the corner points $(0,0)$ and $(1,1)$. We prove that, depending on the class conditional densities, there are two possible different regimes for extreme asymmetries: either the optimal testing asymmetry is more extreme, or it is less extreme. We also provide, under certain conditions, a simple test that can determine the regime given class conditional densities.

In this section, we choose class conditional densities that are either Gaussian or a mixture of Gaussians, because (a) any density can be approximated as well as desired by mixtures of Gaussians (Hastie et al., 2001), and (b) for the square loss and the erf loss, they enable closed-form calculations that lead to Taylor expansions.

4.1 Optimal solutions for extreme cost asymmetries

We assume that the class conditional densities are mixtures of Gaussian, *i.e.*, the density of positive (resp. negative) examples is a mixture of k_+ Gaussians, with means μ_+^i and covariance matrix Σ_+^i, and mixing weights π_+^i, $i \in \{1, \ldots, m_+\}$ (resp. k_- Gaussians, with means μ_-^i and covariance matrix Σ_-^i, and mixing weights π_-^i, $i \in \{1, \ldots, m_-\}$). We denote M_+ (resp. M_-) the $d \times k_+$ (resp. $d \times k_-$) the matrix of means.

We denote p_+ and p_- as the marginal class densities, $p_+ = P(y = 1)$, $p_- = P(y = -1)$. We assume that all mixing weights $\pi_\pm^i$ are strictly positives and that all covariance matrices $\Sigma_\pm^i$ have full rank.

In the following sections, we provide Taylor expansions of various quantities around the null training asymmetry $\gamma = 0$. They trivially extend around the reverse asymmetry $\gamma = 1$. We start with an expansion of the unique global minimum (w, b) of the ϕ-cost with asymmetry γ. For the square loss, (w, b) can be obtained in closed form for any class conditional densities so the expansion is easy to obtain, while for the erf loss, an asymptotic analysis of the optimality conditions has to be carried through, and is only valid for mixtures of Gaussians (see Appendix A for a proof).

Proposition 1 (square loss) *Under previous assumptions, we have the following expansions:*

$$\begin{aligned} w &= 2\frac{p_+}{p_-}\gamma\Sigma_-^{-1}(\mu_+ - \mu_-) + O(\gamma^2) \\ b &= -1 + \frac{p_+}{p_-}\gamma[2 - 2\mu_-^\top(\mu_+ - \mu_-)] + O(\gamma^2) \end{aligned}$$

where $m = \mu_+ - \mu_-$, and $\Sigma_\pm$ and $\mu_\pm$ are the class conditional means and variances. For mixtures, of Gaussians, we have $\Sigma_\pm = \sum_i \pi_\pm^i \Sigma_\pm^i + M_\pm(\mathrm{diag}(\pi_\pm) - \pi_\pm\pi_\pm^\top)M_\pm^\top$ and $\mu_\pm = \sum_i \pi_\pm^i \mu_\pm^i$.

Proposition 2 (erf loss) *Under previous assumptions, we have the following expansions (see the proof in the Appendix):*

$$\begin{aligned} w &= (2\log(1/\gamma))^{-1/2}\,\widetilde{\Sigma}_-^{-1}(\tilde{\mu}_+ - \tilde{\mu}_-) + o\left(\log(1/\gamma)^{-1/2}\right) \\ b &= -\,(2\log(1/\gamma))^{1/2} + o\left(\log(1/\gamma)^{1/2}\right) \end{aligned}$$

where $\tilde{m} = \tilde{\mu}_+ - \tilde{\mu}_-$, and $\widetilde{\Sigma}_\pm$ and $\tilde{\mu}_\pm$ are convex combinations of the mixture means and covariances, i.e., *there exists strictly positive weights $\tilde{\pi}_\pm^i$, that sum to one for each sign, such that $\widetilde{\Sigma}_\pm = \sum_i \tilde{\pi}_\pm^i \Sigma_\pm^i$ and $\tilde{\mu}_\pm = \sum_i \tilde{\pi}_\pm^i \mu_\pm^i$. The weights $\tilde{\pi}_+$ are equal to π_+, while the weights $\tilde{\pi}_-$ are the unique solution of a strictly convex optimization problem (see Appendix A for details).*

Note that when there is only one mixture component (Gaussian densities), then $\tilde{\pi}_\pm^1 = 1$.

4.2 Expansion of testing asymmetries

Using the expansions of Proposition 1 and 2, we can readily derive an expansion of the ROC coordinates for small γ, as well as the testing asymmetry $\beta(\gamma)$. We have (see Appendix B for a proof):

Proposition 3 (square loss) *Under previous assumptions, we have the following expansion:*

$$\log\left(\frac{p_-}{p+}(\beta(\gamma)^{-1}–1)\right) = \frac{p_-^2}{8p_+^2\gamma^2}\left(\frac{1}{m^\top\Sigma_-^{-1}\Sigma_-^{i_-}\Sigma_-^{-1}m} - \frac{1}{m^\top\Sigma_-^{-1}\Sigma_+^{i_+}\Sigma_-^{-1}m}\right) + o(1/\gamma^2) \qquad (1)$$

where i_- (resp. i_+) is one of the negative (resp. positive) mixture component.

Proposition 4 (erf loss) *Under previous assumptions, we have the following expansion:*

$$\log\left(\frac{p_-}{p+}(\beta(\gamma)^{-1}–1)\right) = 2\log(1/\gamma)\left(\frac{1}{\tilde{m}^\top\widetilde{\Sigma}_-^{-1}\Sigma_-^{i_-}\widetilde{\Sigma}_-^{-1}\tilde{m}} - \frac{1}{\tilde{m}^\top\widetilde{\Sigma}_-^{-1}\Sigma_+^{i_+}\widetilde{\Sigma}_-^{-1}\tilde{m}}\right) + o(\log(1/\gamma)) \qquad (2)$$

where i_- (resp. i_+) is one of the negative (resp. positive) mixture component.

The rest of the analysis is identical for both losses and thus, for simplicity, we focus primarily on the square loss. For the square loss, we have two different regimes, depending on the sign of $m^\top\Sigma_-^{-1}\Sigma_-^{i_-}\Sigma_-^{-1}m - m^\top\Sigma_-^{-1}\Sigma_+^{i_+}\Sigma_-^{-1}m$:

- if $m^\top\Sigma_-^{-1}\Sigma_-^{i_-}\Sigma_-^{-1}m > m^\top\Sigma_-^{-1}\Sigma_+^{i_+}\Sigma_-^{-1}m$, then from the expansion in Eq. (1) and Eq. (2), we see that the testing asymmetry tends to 1 exponentially fast. Because this is an expansion around the null training asymmetry, the ROC curve must be starting on the bottom right part of the main diagonal and the points close to $\gamma = 0$ are not ROC-consistent, *i.e.*, the classifiers with training asymmetry too close to zero are useless as they are too extreme.

- if $m^\top \Sigma_-^{-1} \Sigma_-^{i-} \Sigma_-^{-1} m < m^\top \Sigma_-^{-1} \Sigma_+^{i+} \Sigma_-^{-1} m$, then from the expansion in Eq. (1) and Eq. (2), we see that the testing asymmetry tends to 0 exponentially fast, in particular, the derivative $d\beta/d\gamma$ is null at $\gamma = 0$, meaning, that the testing asymmetry is significantly smaller than the training asymmetry, *i.e.*, less extreme.

- if $m^\top \Sigma_-^{-1} \Sigma_-^{i-} \Sigma_-^{-1} m = m^\top \Sigma_-^{-1} \Sigma_+^{i+} \Sigma_-^{-1} m$, then the asymptotic expansion does not provide any information relating to the behavior of the testing asymmetry. We are currently investigating higher-order expansions in order to study the behavior of this limiting case. Note that when the two class conditional densities are Gaussians with identical covariance (a case where the Bayes optimal classifier with symmetric cost is indeed linear), we are in the present case.

The strength of the effects we have described above depends on the norm of $m = \mu_+ - \mu_-$: if m is large, *i.e.*, the classification problem is simple, then those effects are less strong, while when m is small, they are stronger. In Figure 5, we provide several examples for the square loss, with the two regimes and different strengths. It is worth noting, that, although the theoretical results obtained in this section are asymptotic expansions around the corners (*i.e.*, extreme asymmetries), the effects also remain valid far from the corners.

We thus must test to identify which regime we are in, namely testing for the sign of $m^\top \Sigma_-^{-1} \Sigma_-^{i-} \Sigma_-^{-1} m - m^\top \Sigma_-^{-1} \Sigma_+^{i+} \Sigma_-^{-1} m$. This test requires knowledge of the class conditional densities; it can currently always be performed in closed form for the square loss and mixture of Gaussian, while for the erf loss, it requires to solve a convex optimization problem.

4.3 Building the entire ROC curve for a single point

As shown empirically in Section 3.2, and demonstrated theoretically in this section, training and testing asymmetries differ; and this difference suggests that even when the user is interested in only one cost asymmetry, the training procedure should explore more cost asymmetries, i.e. build the ROC curve as presented in Section 3.2 and chose the best classifier as follows: a given asymmetry in cost for the test case leads to a unique slope in the ROC space, and the optimal point for this asymmetry is the point on the ROC curve whose tangent has the corresponding slope and which is closest to the upper-left corner.

We compare in Figure 1, for various datasets and linear classifiers, the performance with cost asymmetry γ of training a classifier with cost asymmetry γ to the performance of training with all cost asymmetries. Using all asymmetries always perform better than assuming a single asymmetry—we simply have more classifiers to choose from. In Figure 1, we report only the performance for the cost asymmetries which show the greatest differences, showing that in some cases, it is very significant, and that a simple change in the training procedure may lead to substantial gains.

5 Conclusion

We have presented an efficient algorithm to build ROC curves by varying the training cost asymmetries for SVMs. The algorithm is based on the piecewise linearity of the path of solutions when the cost of false positives and false negatives vary. We have also provided a theoretical analysis of the relationship between the potentially different cost asymmetries assumed in training and testing classifiers, showing that they differ under certain circumstances. We have characterized key relationships, and have worked to highlight the potential value of building the entire ROC curve even when a single point may be needed. Such an approach can lead to a significant improvement of performance with little added computational cost. Finally, we note that, although we have focused in this paper

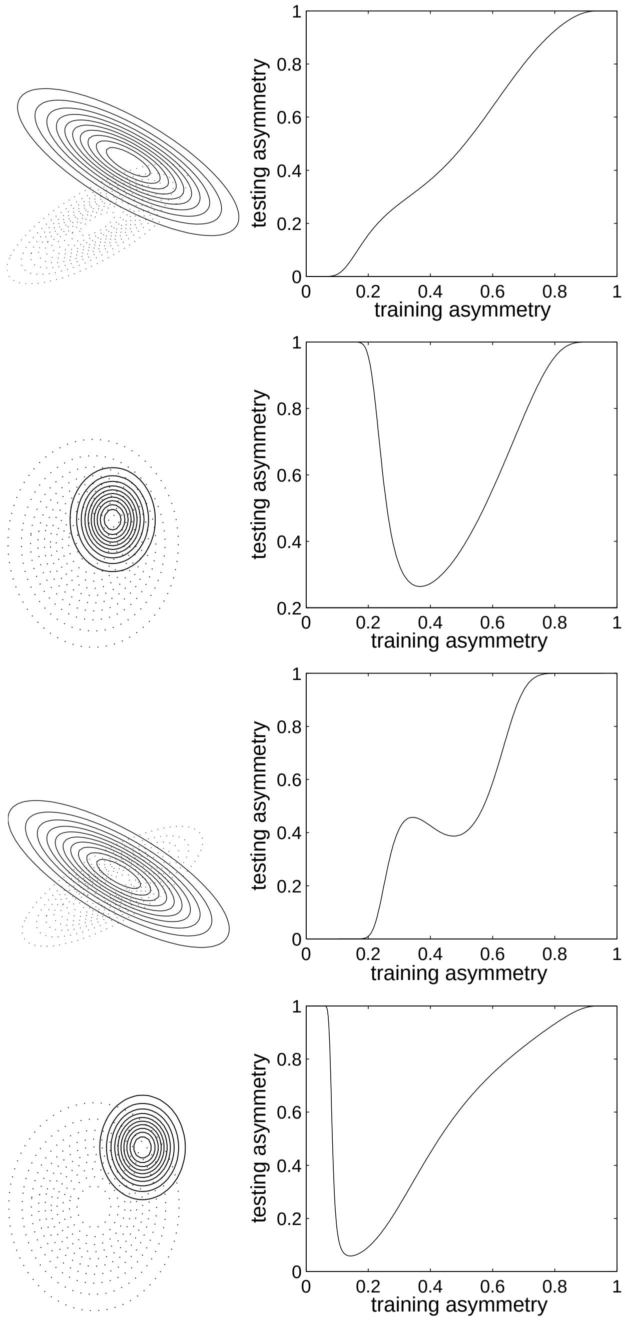

Figure 5: Training asymmetry vs. testing asymmetry, **square loss**: (Left) Gaussian class conditional densities, (right) testing asymmetry vs. training asymmetry; from top to bottom, the values of $(m^\top \Sigma_-^{-1} \Sigma_-^{i_-} \Sigma_-^{-1} m)^{-1} - (m^\top \Sigma_-^{-1} \Sigma_+^{i_+} \Sigma_-^{-1} m)^{-1}$ are 0.12, -6, 3, -0.96.

Dataset	γ	one asym.	all asym.
Pima	0.68	41 ± 0.4	22 ± 1
Breast	0.99	0.9 ± 0.03	0.09 ± 0.04
Ionosphere	0.82	10 ± 0.5	4 ± 0.8
Liver	0.32	27 ± 1.8	23.8 ± 0.02
Ringnorm	0.94	6.3 ± 0.06	4.3 ± 0.1
Twonorm	0.16	15 ± 0.2	1.2 ± 0.2
Adult	0.70	12.8 ± 0.8	11.5 ± 0.3

Table 1: Training with the testing asymmetry γ vs. training with all cost asymmetries: we report validation costs obtained from 10 half-random splits (premultiplied by 100). Only the asymmetry with the largest difference is reported. Given an asymmetry γ we use the cost settings $C_+ = 2\gamma$, $C_- = 2(1-\gamma)$ (which leads to the misclassification error if $\gamma = 1/2$).

on the single kernel learning problem, our approach can be readily extended to the multiple kernel learning setting (Bach et al., 2005) with appropriate numerical path following techniques.

Acknowledgments

We thank John Platt for sharing his experiences and insights with considerations of cost functions in training and testing support vector machines.

References

D. Avis, D. Bremner, and R. Seidel. How good are convex hull algorithms ? In *Computational Geometry: Theory and Applications*, volume 7, 1997.

F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *NIPS 17*, 2005.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Large margin classifiers: convex loss, low noise, and convergence rates. In *NIPS 16*, 2004.

N. Bleistein and R. A. Handelsman. *Asymptotic Expansions of Integrals.* Dover, 1986.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2003.

P. A. Flach. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *ICML*, 2003.

G. H. Golub and C. F. Van Loan. *Matrix Computations.* J.Hopkins Univ. Press, 1996.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *JMLR*, 5:1391–1415, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, 2001.

I. Maros. *Computational Techniques of the Simplex Method.* Klüwer Academic Publishers, 2002.

M. S. Pepe. Receiver operating characteristic methodology. *J. Am. Stat. Assoc.*, 95(449):308–311, 2000.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42 (3):203–231, 2001.

B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, 2002.

A Proof of expansion of optimal solutions

In this appendix, we give the proof of the expansion of optimal solutions for extreme asymmetries, for the square and erf loss. We perform the expansion using the variable $\rho = \frac{C_+ p_+}{C_- p_-} = \frac{\gamma p_+}{(1-\gamma)p_-} = \frac{\gamma p_+}{p_-} + O(\gamma^2)$ around zero.

A.1 Square loss. Proof of Proposition 1.

In this case, the classifier is simply a linear regression and (w, b) can be obtained in closed form as the solution of the following linear system (obtained from the normal equations):

$$\begin{aligned} b &= \frac{\rho - 1}{\rho + 1} - w^\top \frac{\rho\mu_+ + \mu_-}{\rho + 1} && (3) \\ \left(\rho\Sigma_+ + \Sigma_- + \frac{\rho}{\rho+1}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^\top\right) w &= \frac{2\rho}{\rho+1}(\mu_+ - \mu_-), && (4) \end{aligned}$$

where Σ_+ and Σ_- are the class conditional means and covariance matrices.

The first two terms of the Taylor expansions around $\rho = 0$ (i.e., around $\gamma = 0$) are straightforward to obtain:

$$\begin{aligned} w &= 2\rho\Sigma_-^{-1}m - 2\rho^2\left[\Sigma_-^{-1}m + \Sigma_-^{-1}(\Sigma_+ + mm^\top)\Sigma_-^{-1}m\right] + O(\rho^3) \\ b &= -1 + \rho\left[2 - 2\mu_-^\top\Sigma_-^{-1}m\right] + O(\rho^2). \end{aligned}$$

A.2 Erf loss. Proof of Proposition 2.

We begin by proving Proposition 2 in the Gaussian case, where the proof is straightforward.

A.2.1 Expectation of the erf loss and its derivatives for Gaussian densities

A short calculation shows that, when expectations are taken with respect to a normal distribution with mean μ and variance Σ, we have:

$$\begin{aligned} E\phi_{erf}(w^\top x + b) &= (-w^\top\mu - b)\psi\left(\frac{-w^\top\mu - b}{(1 + w^\top\Sigma w)^{1/2}}\right) + (1 + w^\top\Sigma w)^{1/2}\psi'\left(\frac{-w^\top\mu - b}{(1 + w^\top\Sigma w)^{1/2}}\right) \\ E\frac{\partial\phi_{erf}(w^\top x + b)}{\partial w} &= -\mu\psi\left(\frac{-w^\top\mu - b}{(1 + w^\top\Sigma w)^{1/2}}\right) + \frac{\Sigma w}{(1 + w^\top\Sigma w)^{1/2}}\psi'\left(\frac{-w^\top\mu - b}{(1 + w^\top\Sigma w)^{1/2}}\right) \\ E\frac{\partial\phi_{erf}(w^\top x + b)}{\partial b} &= -\psi\left(\frac{-w^\top\mu - b}{(1 + w^\top\Sigma w)^{1/2}}\right) \end{aligned}$$

A.2.2 Gaussian case

Let's define $t_- = \dfrac{w^\top \mu_- + b}{(1+w^\top \Sigma_- w)^{1/2}}$ and $t_+ = \dfrac{-w^\top \mu_+ - b}{(1+w^\top \Sigma_+ w)^{1/2}}$.

The optimality conditions for (w,b) are the following (obtained by zeroing derivatives with respect to b and w):

$$p_+ C_+ \left(-\mu_+ \psi(t_+) + \frac{\Sigma_+ w}{(1+w^\top \Sigma_+ w)^{1/2}} \psi'(t_+) \right) + p_- C_- \left(\mu_- \psi(t_-) + \frac{\Sigma_- w}{(1+w^\top \Sigma_- w)^{1/2}} \psi'(t_-) \right) = 0$$

$$-p_+ C_+ \psi(t_+) + p_- C_- \psi(t_-) = 0$$

which can be rewritten as (with $\rho = \dfrac{C_+ p_+}{C_- p_-}$):

$$\rho \left(-\mu_+ \psi(t_+) + \frac{\Sigma_+ w}{(1+w^\top \Sigma_+ w)^{1/2}} \psi'(t_+) \right) + \left(\mu_- \psi(t_-) + \frac{\Sigma_- w}{(1+w^\top \Sigma_- w)^{1/2}} \psi'(t_-) \right) = 0 \quad (5)$$

$$\rho \psi(t_+) = \psi(t_-) \quad (6)$$

Eq. (6) implies that as ρ tends to zero, $\psi(t_-)$ tends to zero, and thus t_- tends to $-\infty$. This in turn implies that $b/(1+||w||)$ tends to $-\infty$, which in turn implies that $\psi(t_+)$ tends to 1. It is well known that as z tends to $-\infty$, we have $\psi(z) \approx \frac{\psi'(z)}{-z}$ (see e.g., Bleistein and Handelsman (1986)). Thus, if we divide Eq. (5) by $\psi(t_+)$, we get:

$$-\rho \mu_+ + \rho \frac{\Sigma_+ w}{(1+w^\top \Sigma_+ w)^{1/2}} \frac{\psi'(t_+)}{\psi(t_+)} + \mu_- \rho + \frac{\Sigma_- w}{(1+w^\top \Sigma_- w)^{1/2}} \frac{\psi'(t_-)}{\psi(t_-)} \frac{\psi(t_-)}{\psi(t_+)} = 0, \text{ i.e.,}$$

$$\frac{\Sigma_+ w}{(1+w^\top \Sigma_+ w)^{1/2}} \frac{\psi'(t_+)}{\psi(t_+)} + \frac{\Sigma_- w}{(1+w^\top \Sigma_- w)^{1/2}} (-t_-) = \mu_+ - \mu_-$$

The first term in the left hand side is going to zero while the second term goes to infinity. Thus w tends to zero, and thus we have the expansion

$$w \approx \frac{1}{-t_-} \Sigma_-^{-1} (\mu_+ - \mu_-)$$

Since $\psi(t_-) \approx \rho$, we have $t_- \approx= -\sqrt{2\log(1/\rho)}$ (other classical result on the erf function). This finally leads to:

$$\begin{array}{rcl} b(\rho) & \approx & -(2\log(1/\rho))^{1/2} \\ w(\rho) & \approx & (2\log(1/\rho))^{-1/2} \Sigma_-^{-1} (\mu_+ - \mu_-) \end{array}$$

A.2.3 General case

We assume that each $\pi^i_\pm$ is strictly positive and that each covariance matrix $\Sigma^i_\pm$ is positive definite. We define $t^i_- = \dfrac{w^\top \mu^i_- + b}{(1+w^\top \Sigma^i_- w)^{1/2}}$ and $t^i_+ = \dfrac{-w^\top \mu^i_+ - b}{(1+w^\top \Sigma^i_+ w)^{1/2}}$.

Without loss of generality, we can assume that $\mu_+ = \sum_i \pi^i_+ \mu^i_+ = 0$. The optimality conditions for (w,b) are the following (obtained by zeroing derivatives with respect to b and w):

$$\rho \left(\sum_i \pi^i_+ \left\{ -\mu^i_+ \psi(t^i_+) + \frac{\Sigma^i_+ w}{(1+w^\top \Sigma^i_+ w)^{1/2}} \psi'(t^i_+) \right\} \right)$$

$$+ \left(\sum_i \pi^i_- \left\{ \mu^i_- \psi(t^i_-) + \frac{\Sigma^i_- w}{(1 + w^\top \Sigma_- w)^{1/2}} \psi'(t^i_-) \right\} \right) = 0 \qquad (7)$$

$$-\rho \sum_i \pi^i_+ \psi(t^i_+) + \sum_i \pi^i_- \psi(t^i_-) = 0 \qquad (8)$$

From Eq. (8), we obtain that $\psi(t^i_-)$ tends to zero for for all i, and this implies that $b/(1+||w||)$ tends to $-\infty$, and in turn that $\psi(t^i_+)$ tends to 1 for all i. We can now divide Eq. (7) by $\sum_i \pi^i_- \psi(t^i_-) = \rho \sum_i \pi^i_+ \psi(t^i_+)$, to obtain:

$$\begin{aligned} \frac{\sum_i \pi^i_- \psi(t^i_-) \mu^i_-}{\sum_i \pi^i_- \psi(t^i_-)} - \frac{\sum_i \pi^i_+ \psi(t^i_+) \mu^i_+}{\sum_i \pi^i_+ \psi(t^i_+)} &= -\frac{1}{\sum_i \pi^i_+ \psi(t^i_+)} \sum_i \pi^i_+ \psi(t^i_+) \left\{ \frac{\Sigma^i_+ w}{(1 + w^\top \Sigma^i_+ w)^{1/2}} \frac{\psi'(t^i_+)}{\psi(t^i_+)} \right\} \\ &\quad -\frac{1}{\sum_i \pi^i_- \psi(t^i_-)} \sum_i \pi^i_- \psi(t^i_-) \left\{ \frac{\Sigma^i_- w}{(1 + w^\top \Sigma^i_- w)^{1/2}} \frac{\psi'(t^i_-)}{\psi(t^i_-)} \right\} \end{aligned}$$

As in the Gaussian case, the first term of the right hand is negligible compared to the second term when ρ goes to zero. Moreover, for all i, we have $\psi'(t^i_-)/\psi(t^i_-) \approx -t^i_-$ and we thus get:

$$\frac{\sum_i \pi^i_- \psi(t^i_-) \mu^i_-}{\sum_i \pi^i_- \psi(t^i_-)} - \frac{\sum_i \pi^i_+ \psi(t^i_+) \mu^i_+}{\sum_i \pi^i_+ \psi(t^i_+)} = \frac{-1}{\sum_i \pi^i_- \psi(t^i_-)} \sum_i \pi^i_- \psi(t^i_-) \left\{ \frac{-t^i_-}{(1 + w^\top \Sigma^i_- w)^{1/2}} \right\} \Sigma^i_- w \qquad (9)$$

Since $-t^i_-$ goes to $+\infty$ faster than $(2\log(1/\rho))^{1/2}$, this implies that w goes to zero at least as fast as $(2\log(1/\rho))^{-1/2}$, which in turn implies that $b \approx t^i_-$ for all i and bw is bounded as ρ tends to zero.

Let $\tilde{\pi}_i^+ = \frac{\pi^i_+ \psi(t^i_+)}{\sum_j \pi^j_+ \psi(t^j_+)}$ and $\tilde{\pi}_i^- = \frac{\pi^i_- \psi(t^i_-)}{\sum_j \pi^j_- \psi(t^j_-)}$. The quantities bw, $\tilde{\pi}_+$ and $\tilde{\pi}_-$ are functions of ρ. As ρ tends to zero, they all remain bounded. We now proceed to prove that all points of accumulation of those quantities as ρ tends to zero satisfy a set of equations with an unique solution. This will imply that those quantities converge as ρ tends to zero.

Equation for $\tilde{\pi}_+$ Since t^i_+ tends to $+\infty$, $\tilde{\pi}^i_+$ tends to π^i_+, and $\sum_i \tilde{\pi}^i_+ \mu^i_+$ tends to zero, since we have assumed that $\sum_i \pi^i_+ \mu^i_+ = 0$

Equation for $\tilde{\pi}_-$ and bw Let θ and ξ be points of accumulation of bw and $\tilde{\pi}_-$ as ρ tends to zero (i.e., θ and ξ are limits of sequences $b(\rho_k)w(\rho_k)$ and $\tilde{\pi}_-(\rho_k)$ as k tends to ∞, with $\rho_k \to 0$). From Eq. (9), we get:

$$\sum_i \xi_i \mu^i_- = \left(\sum_i \xi_i \Sigma^i_- \right) \theta. \qquad (10)$$

We can now expand

$$\begin{aligned} (t^i_-)^2 - (t^j_-)^2 &= \frac{(w^\top \mu^i_- + b)^2}{1 + w^\top \Sigma^i_- w} - \frac{(w^\top \mu^j_- + b)^2}{1 + w^\top \Sigma^j_- w} \\ &\approx 2bw^\top(\mu^i_- - \mu^j_-) - b^2 w^\top (\Sigma^i_- - \Sigma^j_-) w \\ &\to (2\theta^\top \mu^i_- - \theta^\top \Sigma^i_- \theta) - (2\theta^\top \mu^j_- - \theta^\top \Sigma^j_- \theta) \text{ as } \rho_k \to 0. \end{aligned}$$

We thus have

$$\frac{\tilde{\pi}^i_-}{\tilde{\pi}^j_-} = \frac{\pi^i_- \psi(t^i_-)}{\pi^j_- \psi(t^j_-)} \approx \frac{\pi^i_-}{\pi^j_-} \frac{\psi'(t^i_-)}{\psi'(t^j_-)} \frac{t^j_-}{t^i_-} \to \frac{\pi^i_- \exp(-\theta^\top \mu^i_- + \frac{1}{2}\theta^\top \Sigma^i_- \theta)}{\pi^j_- \exp(-\theta^\top \mu^j_- + \frac{1}{2}\theta^\top \Sigma^j_- \theta)} \text{ as } \rho_k \to 0,$$

which implies that

$$\xi = G\left\{(\log \pi_-^i - \theta^\top \mu_-^i + \frac{1}{2}\theta^\top \Sigma_-^i \theta)_i\right\} \tag{11}$$

where G is the "softmax" function from $\mathbb{R}^{k_-}$ to $\mathbb{R}^{k_-}$ defined as $G_j(x) = \dfrac{e^{x_j}}{\sum_k e^{x_k}}$.

Unique solution of Eq. (10) and Eq. (11) We now prove that Eq. (10) and Eq. (11) together have an unique solution, obtained as the optimum solution of a strictly convex problem. From Eq. (10), we can write θ as a function of ξ as:

$$\theta(\xi) = \left(\sum_i \xi_i \Sigma_-^i\right)^{-1} \sum_i \xi_i \mu_-^i \tag{12}$$

We now show that the equation

$$\xi = G\left\{(\log \pi_-^i - \theta(\xi)^\top \mu_-^i + \frac{1}{2}\theta(\xi)^\top \Sigma_-^i \theta(\xi))_i\right\} \tag{13}$$

has a unique solution on the simplex $\{\xi, \sum_i \xi_i = 1, \xi_i > 0, \forall i\}$. Let us define the following function defined on the positive orthant $\{\xi, \xi_i > 0, \forall i\}$:

$$H(\xi) = \sum_i \xi_i \log \xi_i - \sum_i \xi_i \left\{\log \pi_-^i - \theta(\xi)^\top \mu_-^i + \frac{1}{2}\theta(\xi)^\top \Sigma_-^i \theta(\xi)\right\}).$$

Short calculations show that:

$$\begin{aligned}
\frac{\partial \theta}{\partial \xi_i} &= \left(\sum_k \xi_k \Sigma_-^k\right)^{-1} \left(\mu_-^i - \Sigma_-^i \theta\right) \\
\frac{\partial\left\{-\theta(\xi)^\top \mu_-^i + \frac{1}{2}\theta(\xi)^\top \Sigma_-^i \theta(\xi)\right\}}{\partial \xi_j} &= -\left(\mu_-^i - \Sigma_-^i \theta\right) \left(\sum_k \xi_k \Sigma_-^k\right)^{-1} \left(\mu_-^j - \Sigma_-^j \theta\right) \\
\frac{\partial H}{\partial \xi_i} &= \log \xi_i + 1 - \left(\log \pi_-^i - \theta(\xi)^\top \mu_-^i + \frac{1}{2}\theta(\xi)^\top \Sigma_-^i \theta(\xi)\right) \\
\frac{\partial^2 H}{\partial \xi_i \partial \xi_j} &= \delta_{ij} \frac{1}{\xi_i} + \left(\mu_-^i - \Sigma_-^i \theta\right) \left(\sum_k \xi_k \Sigma_-^k\right)^{-1} \left(\mu_-^j - \Sigma_-^j \theta\right).
\end{aligned}$$

The last equation shows that the function H is strictly convex in the positive orthant. Thus, minimizing $H(\xi)$ subject to $\sum_i \xi_i = 1$ has an unique solution. Optimality conditions are derived by writing down the Lagrangian:

$$\mathcal{L}(\xi, \alpha) = H(\xi) + \alpha(\sum_i \xi_i - 1),$$

which leads to the following optimality conditions:

$$\forall i,\ \frac{\partial H}{\partial \xi_i} + \alpha = 0 \tag{14}$$

$$\sum_i \xi_i = 1 \tag{15}$$

The last two equations are exactly equivalent to Eq. (13). We have thus proved that the system defining θ and ξ (Eq. (10) and Eq. (11)) has an unique solution obtained as the solution of a convex optimization problem.

Asymptotic equivalent It is easy to show that $\psi(b)/\rho$ is bounded as ρ tends to zero, which implies that $b \approx -(2\log(1/\rho))^{1/2}$. From the fact that bw tends to a limit θ, we immediately obtain that $w \approx \theta/b$, which completes the proof of Proposition 2.

B Proof of expansion of testing asymmetries

For the two losses we considered (square and erf), the expansions of w and b around $\gamma = 0$ lead to

$$\frac{w(\gamma)}{b(\gamma)} \approx -c(\gamma)a$$

where $c(\gamma) = 2\frac{p_+}{p_-}\gamma$, $a = \Sigma_-^{-1}(\mu_+ - \mu_-)$ for the square loss and $c(\gamma) = (2\log(1/\gamma))^{-1}$, $a = \widetilde{\Sigma}_-^{-1}(\tilde{\mu}_+ - \tilde{\mu}_-)$ for the erf loss.

The proportion of false positives $u(\rho)$ and true positives $v(\rho)$ can be obtained as:

$$\begin{aligned} u(\rho) &= P(w^\top x + b \geqslant 0 | y = -1) = \sum_i \pi_-^i \psi\left(\frac{w^\top \mu_-^i + b}{(w^\top \Sigma_-^i w)^{1/2}}\right) = \psi(t_u^i(\gamma)) \\ v(\rho) &= P(w^\top x + b \geqslant 0 | y = 1) = \sum_i \pi_+^i \psi\left(\frac{w^\top \mu_+^i + b}{(w^\top \Sigma_+^i w)^{1/2}}\right) = \psi(t_v^i(\gamma)) \end{aligned}$$

and we have the expansions

$$\begin{aligned} t_u^i(\gamma) \triangleq \frac{w^\top \mu_-^i + b}{(w^\top \Sigma_-^i w)^{1/2}} &\approx \frac{-1}{c(\gamma)(a^\top \Sigma_-^i a)^{1/2}} \\ t_v^i(\gamma) \triangleq \frac{w^\top \mu_+^i + b}{(w^\top \Sigma_+^i w)^{1/2}} &\approx \frac{-1}{c(\gamma)(a^\top \Sigma_+^i a)^{1/2}} \\ \frac{du}{d\gamma} &= \sum_i \pi_-^i \frac{dt_u^i}{dc}\frac{dc}{d\gamma}\psi'(t_u^i(\gamma)) \approx \frac{dc}{d\gamma}\sum_i \pi_-^i \frac{1}{\sqrt{2\pi}} \frac{\exp(-(a^\top \Sigma_-^i a)^{-1}/c(\gamma)^2)}{c(\gamma)^2 (a^\top \Sigma_-^i a)^{-1/2}} \\ \frac{dv}{d\gamma} &= \sum_i \pi_+^i \frac{dt_v^i}{dc}\frac{dc}{d\gamma}\psi'(t_v^i(\gamma)) \approx \frac{dc}{d\gamma}\sum_i \pi_+^i \frac{1}{\sqrt{2\pi}} \frac{\exp(-(a^\top \Sigma_+^i a)^{-1}/c(\gamma)^2)}{c(\gamma)^2 (a^\top \Sigma_+^i a)^{-1/2}} \end{aligned}$$

The expansions of $\frac{du}{d\gamma}$ and $\frac{dv}{d\gamma}$ are each dominated by a single term, corresponding to indices i_- and i_+ that respectively minimize $a^\top \Sigma_-^i a$ and $a^\top \Sigma_+^i a$ (we assume for simplicity that all values of $a^\top \Sigma_\pm^i a$ are distinct). We then obtain

$$\log\left(\frac{dv}{d\gamma} / \frac{du}{d\gamma}\right) \approx \frac{1}{c(\gamma)^2}\left(\frac{1}{a^\top \Sigma_-^{i_-} a} - \frac{1}{a^\top \Sigma_-^{i_+} a}\right)$$

Proposition 3 and 4 follows from $\frac{dv}{d\gamma} / \frac{du}{d\gamma} = \frac{p_-}{p_+}(\beta(\gamma)^{-1} - 1)$.