Some theoretical aspects of support vector machines and related kernel-based learning methods

Ingo Steinwart
Modeling, Algorithms, and Informatics Group, CCS-3
Los Alamos National Laboratory
ingo@lanl.gov
LA-UR 05-5392

Overview

- **The "Learning" Problem**
  The problem formulation, Some examples and results
- **Support Vector Machines**
  RKHS, SVMs
- **Kernels Revisited**
  Universal kernels, Examples and Counter-examples, Some geometric observations
- **Asymptotic Properties of SVMs**
  Consistency, Sparseness, Rates

The "Learning" Problem: Description

- **Informal description:**
  - ⋆ Given: a finite sequence of observations $T := \big((x_1, y_1), \ldots, (x_n, y_n)\big)$
  - ⋆ Goal: predict label y for new, unseen x.

- **Mathematical Model:**
  - ⋆ P is an *unknown* probability measure on $X \times Y$, $Y \subset \mathbb{R}$.
  - ⋆ $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$ sampled from P^n.
  - ⋆ $L : Y \times \mathbb{R} \to [0, \infty]$ *loss function* that measures cost $L(y, t)$ of predicting y by t.

- **Goal:**
  Find a function $f_T : X \to \mathbb{R}$ with small *risk*

$$\mathcal{R}_{L,P}(f_T) := \int_{X \times Y} L\big(y, f_T(x)\big)\, dP(x, y) \ .$$

Learning: Consistency and "No-Free-Lunch Theorem"

- **Bayes risk:**

$$\mathcal{R}^*_{L,P} := \inf\{\, \mathcal{R}_{L,P}(f) \mid f : X \to Y \text{ measurable} \,\}\,.$$

- **Learning method:**
  A *learning method* assigns to every training set T a decision function $f_T : X \to \mathbb{R}$.

- **Consistency:**
  A learning method is called *universally consistent* if

$$\mathcal{R}_{L,P}(f_T) \to \mathcal{R}^*_{L,P} \qquad \text{in probability} \tag{1}$$

  for $n \to \infty$ and *every* probability measure P on $X \times Y$.

- **Good news:**
  Many learning methods are universally consistent, see later.

- **Question:**
  Does there exist a learning method and a convergence rate in (3) that holds for all P?

- **Bad news (Devroye, 1982):**
  No! (if $|Y| \geq 2$, $|X| = \infty$, and L "non-trivial")

The "Learning" Problem: Examples I

- **Binary Classification:**
  - ⋆ **Labels:** $Y := \{-1, 1\}$.
  - ⋆ **Loss:** $\operatorname{sign} t$ predicts y, i.e.

$$L(y,t) := \begin{cases} 1 & \text{if } y \operatorname{sign} t \leq 0 \\ 0 & \text{else.} \end{cases}$$

  - ⋆ **Risk:**

$$\mathcal{R}_{L,P}(f) = P\big((x,y) : \operatorname{sign}(f(x)) \neq y\big)$$

  - ⋆ **Consistency:**
    E.g. k-Nearest Neighbour (Stone, 1977), but also many others ....
  - ⋆ **Rates:**
    No: If e.g. $P_X \ll \lambda^d$ and $\eta \in C^\infty$, where $\eta(x) := P(y = 1|x)$.
    Yes: e.g. by Yang 1999, Tsybakov 2004, ...

The "Learning" Problem: Examples II

- **Least Squares Regression:**
  - ⋆ **Labels:** continuous values, i.e. $Y := [-M, M]$ or $Y := \mathbb{R}$.
  - ⋆ **Least squares loss:** $L(y,t) := (y-t)^2$.
  - ⋆ **Risk:**

$$\mathcal{R}_{L,P}(f) = \int_{X\times Y} \big(y - f(x)\big)^2 \, dP(x,y)$$

  - ⋆ **Consistency:**
    Many methods including kernel estimators and (regularized) risk minimizers.
  - ⋆ **Rates:**
    Possible if e.g. the minimizer $f^*_{L,P}$ of $\mathcal{R}_{L,P}(.)$ is smooth in a Sobolev sense.

Support Vector Machines: Kernels and RKHS

- **Kernels:**

  $k : X \times X \to \mathbb{R}$ is a *kernel*

  $:\Leftrightarrow$ there exist a Hilbert space H (*feature space*) and a *feature map* $\Phi : X \to H$ with

  $$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \qquad \text{for all } x, x' \in X.$$

  $\Leftrightarrow$ all quadratic kernel matrices are symmetric and positive semi-definite.

- **RKHS:**

  The RKHS of k is the "smallest" feature space of k consisting of functions.

  ⋆ **"Construction":** Equip

  $$\left\{ \sum_{i=1}^{n} \alpha_i k(x_i, .) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in X \right\}$$

  with the dot product

  $$\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, .), \sum_{j=1}^{m} \beta_j k(\hat{x}_j, .) \right\rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, \hat{x}_j)$$

  and take "the" completion.

  ⋆ **Feature map:** $x \mapsto k(x, .), \qquad x \in X.$

  ⋆ **Reproducing property:** $f(x) = \langle f, k(x, .) \rangle, \qquad f \in H, x \in X.$

Support Vector Machines: Examples of Kernels

- **Polynomial Kernels:**
  For $a \geq 0$ and $m \in \mathbb{N}$ let

$$k(x, x') := (\langle x, x' \rangle + a)^m, \qquad x, x' \in \mathbb{R}^d \ .$$

- **Gaussian RBF kernels:**
  For $\sigma > 0$ let

$$k(x, x') := \exp(-\sigma^2 \|x - x'\|_2^2), \qquad x, x' \in \mathbb{R}^d \ .$$

  The parameter σ is called *width*.

Support Vector Machines: Definition

- **Support vector machines (SVMs)** solve the problem

$$f_{T,\lambda} = \arg\min_{f \in H} \ \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \ , \tag{2}$$

  where

  - H is a RKHS,
  - $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$ is a training set,
  - $\lambda > 0$ is a *free* regularization parameter,
  - $L : Y \times \mathbb{R} \to [0, \infty)$ is a *convex* loss, e.g. hinge or least squares.

- **Representer Theorem:**
  The solution is of the form $f_{T,\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, .)$.
  Minimization actually takes place in the RKHS of k over $\{x_1, \ldots, x_n\}$.

- **Questions:**

  - Universally consistent?
  - Learning rates?
  - Additional properties?

Kernels Revisited: Universal Kernels I

- **Universal kernels:**
  A continuous kernel on compact X is *universal* if its RKHS H is dense in $C(X)$.
- **Examples:**
  - ⋆ Gaussian RBFs are universal.
  - ⋆ Polynomial kernels are not universal if $|X| = \infty$.
- **Approximation properties of universal kernels:**
  If k is a universal kernel then

$$\inf_{f \in H} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P} \,.$$

  Consequence:
  SVMs with universal kernel have the *potential* to be universally consistent.
- **Approximation properties of polynomial kernels:**
  If k is a polynomial kernel and $|X| = \infty$ then there is a P with

$$\inf_{f \in H} \mathcal{R}_{L,P}(f) \neq \mathcal{R}^*_{L,P} \,.$$

  Consequence:
  SVMs with polynomial kernel cannot be universally consistent.

Kernels Revisited: Universal Kernels II

- **Observation:**
  - ⋆ Let $x_1, \ldots, x_n \in X$ and k be a universal kernel on X.
  - ⋆ For all $y_1, \ldots, y_n = \pm 1$ there exists an $f \in H$ with

$$f(x_i) = y_i, \qquad i = 1, \ldots n.$$

  - ⋆ Reproducing property gives $\langle f, \Phi(x_i) \rangle = f(x_i)$.
- **Consequence I:**
  Every training set can be perfectly classified.
- **Consequence II:**
  RKHSs (or their balls) of universal kernels have *infinite* VC-dimension.
- **Consequence III:**
  Standard VC-dimension type arguments cannot explain SVMs with Gaussian kernels.

Asymptotic Properties of SVMs: Consistency I

- **Universal consistency:**
  SVMs are universally (L-risk) consistent if RKHS is universal and $\lambda_n \longrightarrow 0$ "slowly".
- **Examples:**
  - ⋆ **Classification:**
    - ∗ T. Zhang '04, S. '02 & '05:
      e.g. L hinge loss, H Gaussian RBF then $n\lambda_n^{1+\varepsilon} \longrightarrow \infty$ is ok.
    - ∗ Bartlett, Jordan & McAufliffe '03 together with S. '05:
      - ○ A *convex* loss of the form $L(y,t) = \varphi(yt)$ is ok if and only if $\varphi'(0) < 0$.
      - ○ Condition on λ_n depends on both H and L.
  - ⋆ **Regression:**
    - ∗ Cucker & Smale '02:
      $Y = [-M, M]$, L least squares, H e.g. Gaussian RBF then $n\lambda_n^{2+\varepsilon} \longrightarrow \infty$ is ok.
    - ∗ Christmann & S. '04:
      $Y = \mathbb{R}$, P finite variance, H e.g. Gaussian RBF then $n\lambda_n^4 \longrightarrow \infty$ is ok.

Asymptotic Properties of SVMs: Consistency II

- **Basic ideas of the proof:**
  - ⋆ **Infinite sample SVM:**

$$f_{P,\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f)$$

  - ⋆ **Approximation error function**

$$a(\lambda) := \min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$$

  - ⋆ **Decomposition:**

$$\mathcal{R}_{L,P}(f_{T,\lambda}) - \mathcal{R}^*_{L,P} \;\leq\; \mathcal{R}_{L,P}(f_{T,\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda}) + a(\lambda)$$

  - ⋆ If H universal then $\lim_{\lambda \to 0} a(\lambda) \to 0$.
  - ⋆ $\mathcal{R}_{L,P}(f_{T,\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda})$ estimated by either
    - ∗ *Stability argument* ensuring $\|f_{T,\lambda} - f_{P,\lambda}\| \leq \varepsilon$ with high probability.
    - ∗ *Uniform bound* ensuring

$$\sup_{f \in \lambda^{-1/2} B_H} \left| \mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f) \right| \leq \varepsilon$$

      with high probability.

      Tools: Hoeffding's inequality & covering numbers, or scale-sensitive VC-dimensions.

Asymptotic Properties of SVMs: Sparseness I

- **Representer theorem:**

$$f_{T,\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, .) \, .$$

- **Support Vectors:**
  A sample x_i of T is called a *support vector* of the above representation of $f_{T,\lambda}$ if $\alpha_i \neq 0$.
- **Observations:**
  - ⋆ The number of support vectors may depend on the representation.
  - ⋆ Both training and testing time depend on the number of support vectors.
- **Question:**
  How many support vectors does (a representation of) $f_{T,\lambda}$ have?

Asymptotic Properties of SVMs: Sparseness II

- **Minimal Representation:**
  For $f \in H$ the *minimal number of support vectors* is

$$\#SV(f) := \inf\left\{n \in \mathbb{N} : \exists \alpha_1, \ldots, \alpha_n \neq 0 \text{ and } x_1, \ldots, x_n \in X \text{ with } f = \sum_{i=1}^{n} \alpha_i k(x_i, .)\right\}.$$

  A representation is called *minimal* if it has $\#SV(f)$ support vectors.

- **Observations:**
  - ⋆ We always have $\#SV(f) \leq \dim H$.
  - ⋆ k a polynomial kernel then $\#SV(f_{T,\lambda_T})/|T| \to 0$ for $n = |T| \to \infty$.
  - ⋆ If $\dim H = \infty$ there exists $f \in H$ with $\#SV(f) = \infty$.
  - ⋆ The representer theorem ensures $\#SV(f_{T,\lambda}) \leq |T|$.

- **Assumptions:**
  - ⋆ P is a distribution with no discrete component, i.e. $P_X(\{x\}) = 0$ for all $x \in X$.
  - ⋆ k is universal, e.g. a Gaussian RBF kernel.

  $\rightsquigarrow$ The representation found by the representer theorem is P-a.s. minimal and unique.

Asymptotic Properties of SVMs: Sparseness III

- **Lower bound for hinge loss:**
  If $\lambda_n \to 0$ and $n\lambda_n^2 \to \infty$ then

$$\liminf_{n\to\infty} \frac{\#SV(f_{T,\lambda_n})}{n} \;\geq\; 2\mathcal{R}_P^*,$$

  and under some conditions this is sharp (see S. '03).

- **Lower bound squared hinge loss:**
  Write $\eta(x) := P(y = 1|x)$. If $\lambda_n \to 0$ and $n\lambda_n^3 \to \infty$ we have

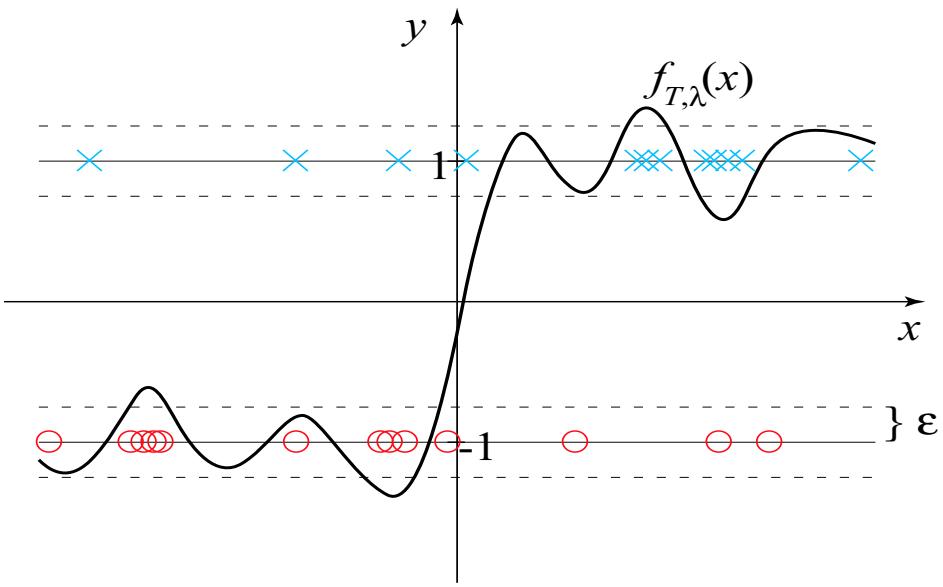
$$\liminf_{n\to\infty} \frac{\#SV(f_{T,\lambda_n})}{n} \;\geq\; P_X\Big(x \in X : 0 < \eta(x) < 1\Big).$$

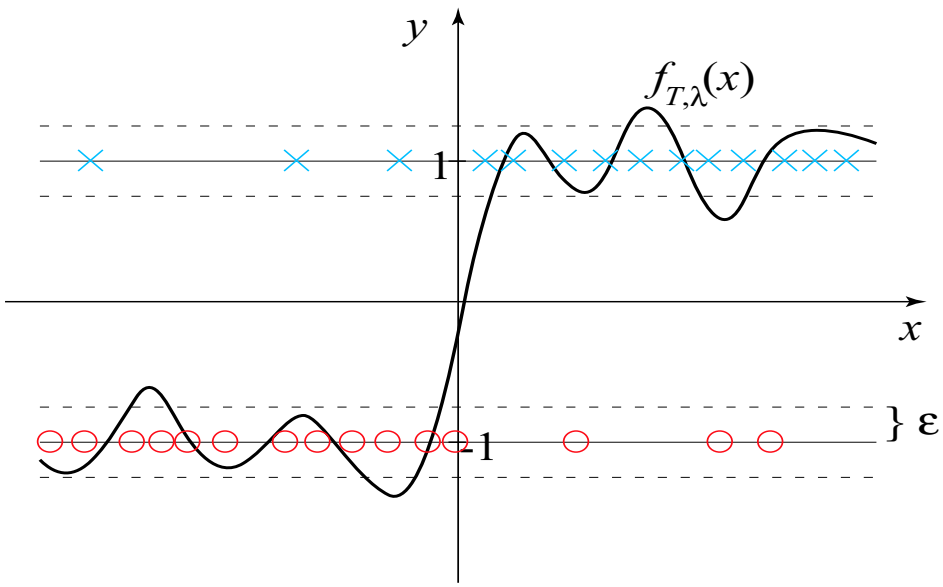
  But: $f_{T,\lambda} \to 2\eta(x) - 1$ for the squared hinge $\rightsquigarrow$ estimate of $\eta(x)$.

- Bartlett & Tewari '04:
  Exact relation between estimating (parts of) $\eta(x)$ and sparseness. Roughly:
  You can either ensure (partial) sparseness or an (partial) estimate of η.

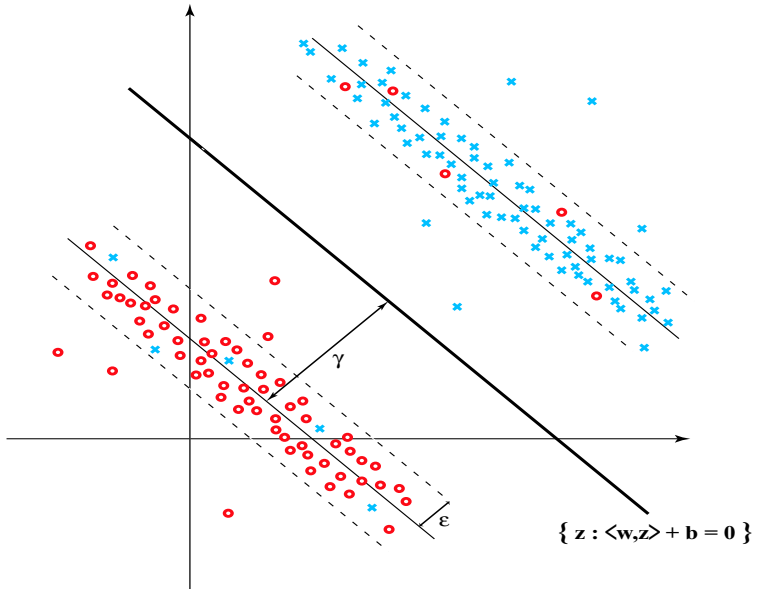
- **Basic idea of proofs:**
  - $\star$ $f_{P,\lambda} \to f_{L,P}^*$ in probability.
  - $\star$ $\|f_{T,\lambda} - f_{P,\lambda}\| \to 0$ with high probability.

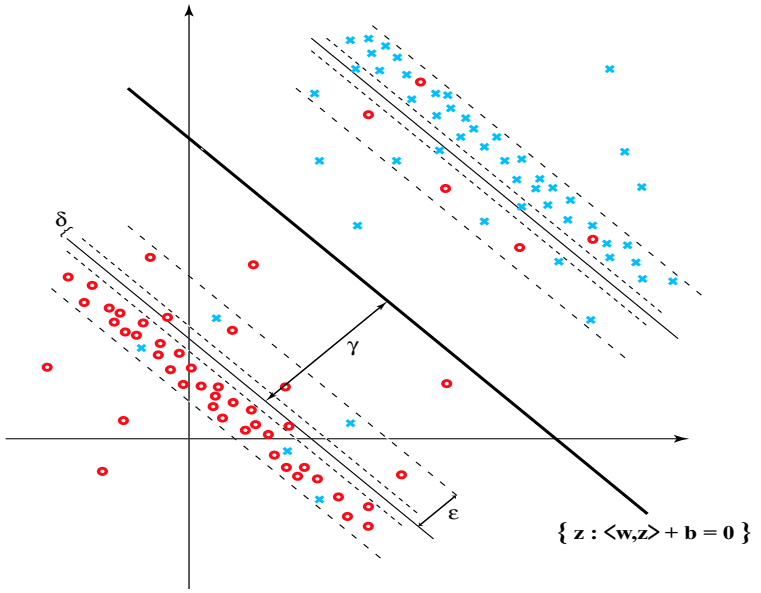
Asymptotic Properties of SVMs: Sparseness IV





Asymptotic behaviour of an SVM with hinge loss for a noisy problem.





Asymptotic Properties of SVMs: Rates I

- **Consistency:**
  A learning method is *universally consistent* if

$$\mathcal{R}_{L,P}(f_T) \to \mathcal{R}^*_{L,P} \qquad \text{in probability} \tag{3}$$

  for $n \to \infty$ and *every* probability measure P on $X \times Y$.

- **"No-free-Lunch" (Devroye, 1982):**
  For classification, no uniform rates are possible if $|X| = \infty$.

- **Some insufficient restrictions:**
  - $\star$ P_X has a specified density with respect to the Lebesgue measure and P is noise-free.
  - $\star$ P_X has a specified density with respect to the Lebesgue measure and η is C^∞.
  - $\star$ P is noise-free and $X_1 \subset \mathbb{R}^d$, $d \geq 2$ is compact and convex.

- **Some sufficient restrictions:**
  - $\star$ P_X has a specified density with respect to the Lebesgue measure and η is smooth with bounds on derivatives (e.g. in Sobolev sense).
  - $\star$ Decision boundary is smooth.
  - $\star$ Bayes classifier is contained in some "small" hypothesis set.
  - $\star$ **Question:** Are smoothness assumptions realistic for classification?

Asymptotic Properties of SVMs: Rates II

- **Observation:**
  Write $\eta(x) := P(y = 1|x)$. Then:

$$|2\eta(x) - 1| \text{ is close to 1} \quad \longleftrightarrow \quad \text{“low noise”}$$
$$|2\eta(x) - 1| \text{ is close to 0} \quad \longleftrightarrow \quad \text{“high noise”} .$$

- **Tsybakov's noise exponent:**
  A distribution P *has noise exponent* $q \in [0, \infty]$ if there exists a constant $C > 0$ such that for all sufficiently small $t > 0$ we have

$$P_X\big(x \in X : |2\eta(x) - 1| \leq t\big) \;\leq\; Ct^q .$$

- **Interpretation:**
  The exponent q measure the amount of noise. Every P has $q = 0$. If $q = \infty$ then η is bounded away from the level $1/2$. The larger q is the “less noisy” is P.

- **Application to Empirical Risk Minimization:**
  - ⋆ General P, i.e. no noise assumption, then rate up to $n^{-\frac{1}{2}}$.
  - ⋆ No noise, i.e. $\mathcal{R}^*_P = 0$, then rate up to n^{-1}.
  - ⋆ Tsybakov '04: noise exponent q then rate up to $n^{-\frac{q+1}{q+2}}$.

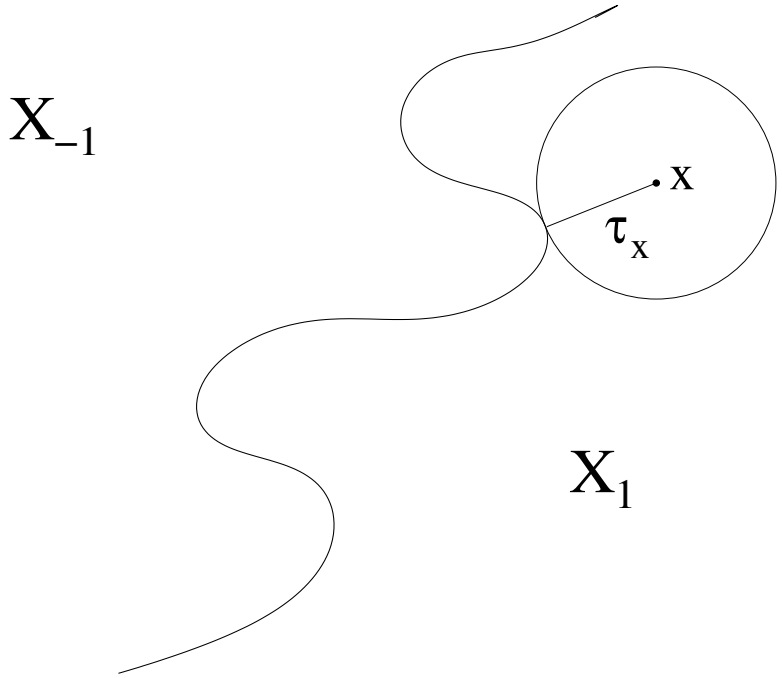
Asymptotic Properties of SVMs: Rates III

- **Distance to the decision boundary:**
  For simplicity assume $\{\eta = 1/2\} = \emptyset$. Then for $x \in X \subset \mathbb{R}^d$ we define

$$\tau_x := \begin{cases} d(x, X_1), & \text{if } x \in X_{-1}, \\ d(x, X_{-1}), & \text{if } x \in X_1, \end{cases} \tag{4}$$

  where $d(x, A)$ denotes the distance between x and A.

- **Interpretation:**
  τ_x measures the distance of x to the "decision boundary".

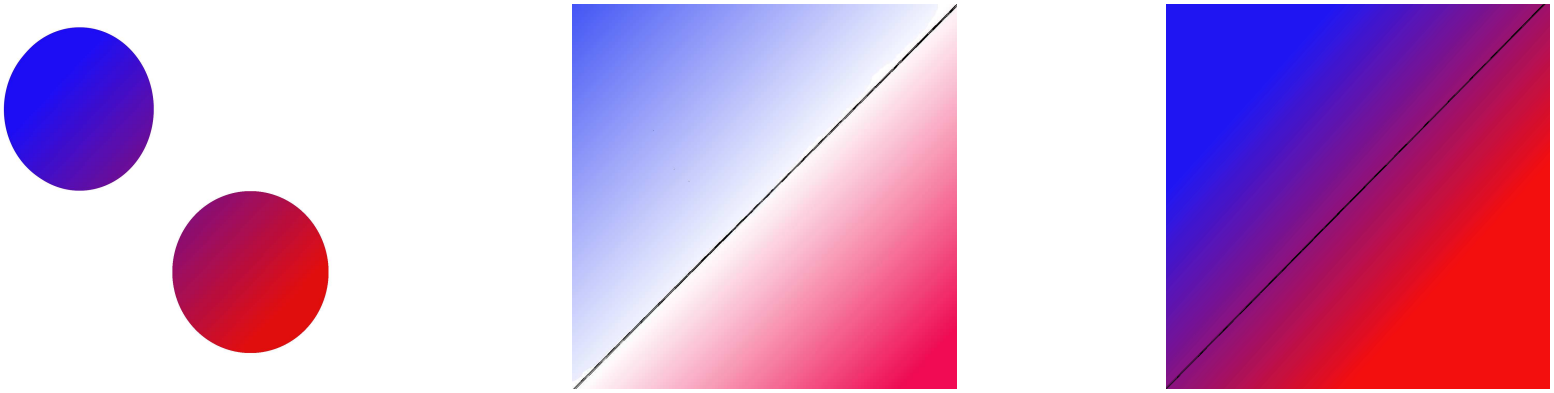


Asymptotic Properties of SVMs: Rates IV

- **Geometric noise exponent:**
  P has geometric noise exponent $\alpha \in (0, \infty]$ if $(x \mapsto \tau_x^{-1}) \in L_{\alpha d}(|2\eta - 1| dP_X)$, i.e.

$$\int \tau_x^{-\alpha d} \, |2\eta(x) - 1| dP_X(x) \;<\; \infty \,.$$

- **Interpretation:**
  The exponent α measures how much $|2\eta - 1| dP_X$ is concentrated around the decision boundary. In particular, $d(X_{-1}, X_1) > 0$ iff $\alpha = \infty$.



Left: X_{-1} and X_1 are strictly separated.
Middle: P_X is lowly concentrated around the decision boundary.
Right: $|2\eta - 1|$ is close to 0 around the decision boundary.

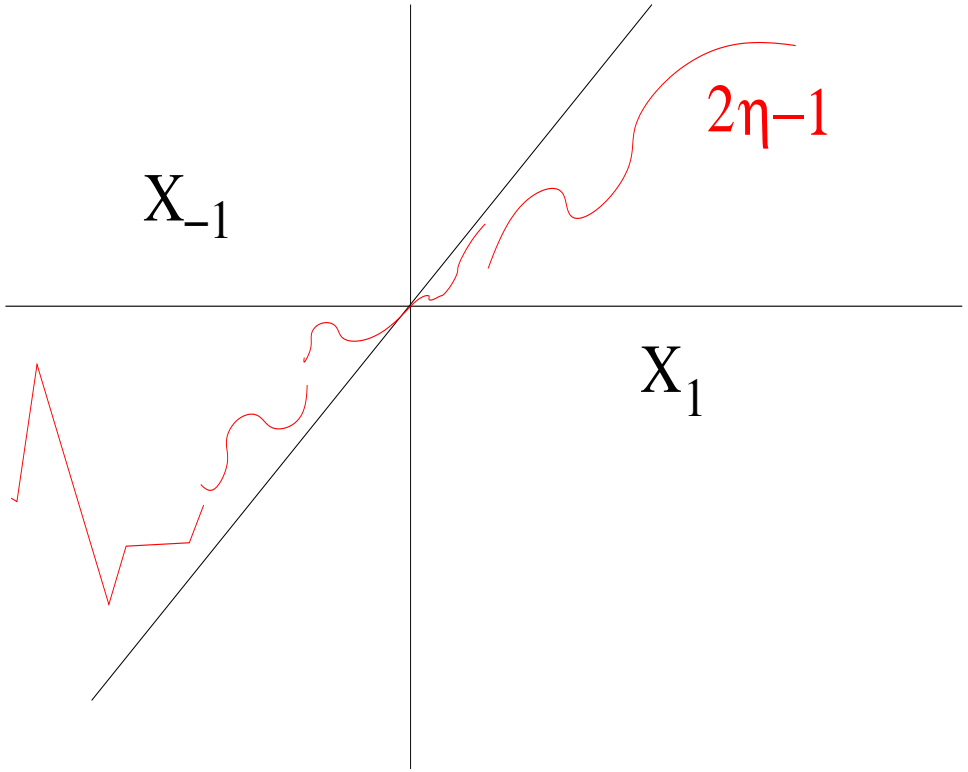
Asymptotic Properties of SVMs: Rates V

- **Example:**
  If P has noise exponent $q > 0$ and if there are $c > 0$ and $\gamma > 0$ with

$$|2\eta(x) - 1| \leq c\,\tau_x^{\gamma}, \qquad x \in X$$

  then P has geometric noise exponent α for all $\alpha < \gamma\frac{q+1}{d}$.

- **Illustration:**



Situation for $\gamma = 1$ if the decision boundary is at 0.

Asymptotic Properties of SVMs: Rates VI

- **Main Theorem:** (S. & Scovel '03)
  - ⋆ $X \subset \mathbb{R}^d$ compact.
  - ⋆ P has noise exponent $q \in [0, \infty]$.
  - ⋆ P has noise geometric noise exponent $\alpha \in (0, \infty]$.
  - ⋆ k is Gaussian RBF kernel with width σ and L is hinge loss.

  Then the SVM learns with rate up to

$$\begin{cases} n^{-\frac{\alpha}{2\alpha+1}} & \text{if } \alpha \leq \frac{1}{2} + \frac{1}{q} \\ n^{-\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}} & \text{otherwise .} \end{cases}$$

- **Discussion**
  - ⋆ Best values for λ and σ depend on unknown exponents q and α.
  - ⋆ Result can be stated as an oracle inequality and can be extended to other losses (S., Hush & Scovel '05).
  - ⋆ If $q \to \infty$ then $\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4} \to \frac{2\alpha}{2\alpha+3}$, i.e. we have rates faster than $n^{-1/2}$ if $\alpha > 3/2$.
  - ⋆ If $\alpha \to \infty$ then $\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4} \to \frac{q+1}{q+2}$, i.e. we have the behaviour of Tsybakov's result.