Multiclass Learning with Margin: Exponential Rates with No Bias-Variance Trade-Off

Stefano Vigogna[1], Giacomo Meanti[2],
Ernesto De Vito[2] and Lorenzo Rosasco[2,3,4]

[1] *RoMaDS, Università di Roma Tor Vergata*
[2] *MaLGa, Università degli Studi di Genova*
[3] *Center for Brains, Minds and Machines, MIT*
[4] *Istituto Italiano di Tecnologia, Genova*

Summary of Our Contributions

Exponential rates in classification under margin conditions:

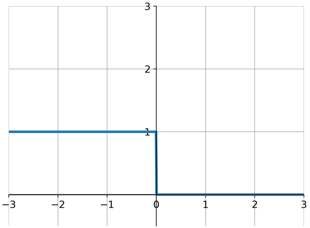
	Square loss	Margin losses e.g. logistic, exponential
Binary	*Audibert and Tsybakov 2007, Koltchinskii and Beznosova 2005*	*Nitanda and Suzuki, 2019*
Multiclass	*Cabannes et al. 2021*	**This work**

Classification

0-1 Loss

find $c : \mathcal{X} \to \mathcal{Y}$, $\#\mathcal{Y} = T$

$$c_* = \arg\min_c \mathcal{R}(c) = \arg\min_c \mathbb{E}\mathbb{1}\{c(X) \neq Y\}$$

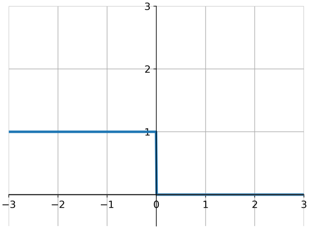


Classification

0-1 Loss

find $c : \mathcal{X} \to \mathcal{Y}$, $\#\mathcal{Y} = T$

$$c_* = \arg\min_c \mathcal{R}(c) = \arg\min_c \mathbb{E}\mathbb{1}\{c(X) \neq Y\}$$



Surrogate Losses

find $f : \mathcal{X} \to \mathbb{R}^{T-1}$

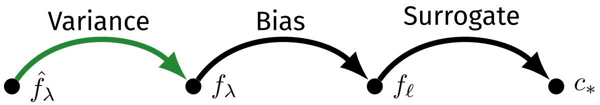
$$f_\ell = \arg\min_f \mathcal{R}_\ell(f) = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$D : \mathbb{R}^{T-1} \to \mathcal{Y}$ decoding operator

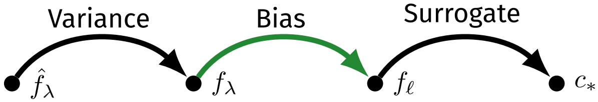
$c = Df$ plug-in classifier

UniGe | MaLGa

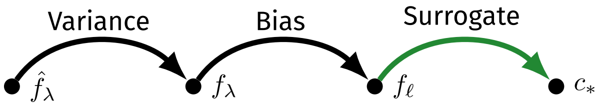
Bias-Variance in Classification



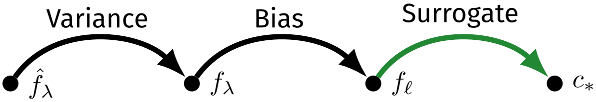
Bias-Variance in Classification

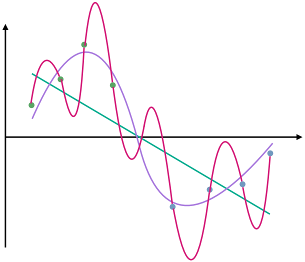


Bias-Variance in Classification



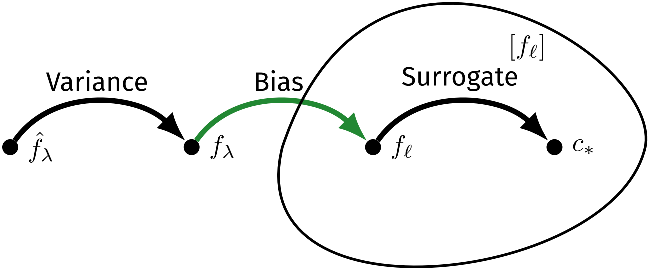
Bias-Variance in Classification

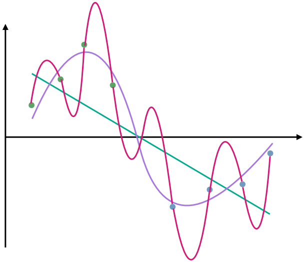




UniGe | MaLGa

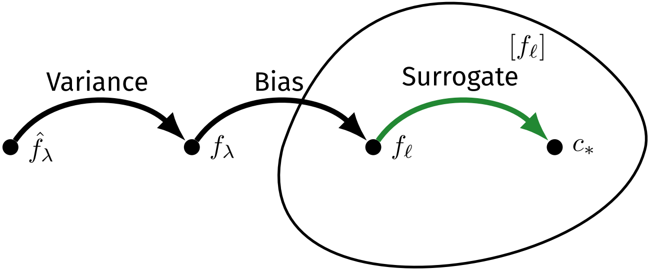
Bias-Variance in Classification

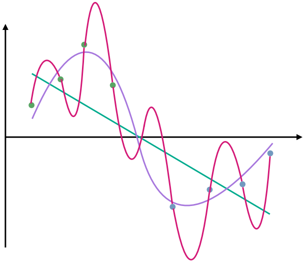




$$[f_\ell] = \{f : \mathcal{X} \to \mathbb{R}^{T-1} : Df = Df_\ell \text{ almost surely}\}$$

Bias-Variance in Classification





$[f_\ell] = \{f : \mathcal{X} \to \mathbb{R}^{T-1} : Df = Df_\ell$ almost surely$\}$

Fisher consistent: $Df_\ell = c_*$ almost surely

Hard-Margin Condition

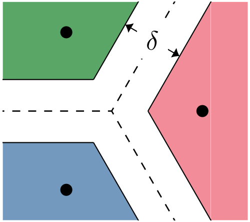
$$\min_{y \neq c_*(X)} \rho(c_*(X) \mid X) - \rho(y \mid X) \geq \delta \quad \text{almost surely}$$

Proposition (Vigogna, Meanti, De Vito, Rosasco)

The hard-margin condition holds if and only if η is confident:

$$\mathrm{dist}(\eta(X), \mathcal{B}) \geq \delta \quad \text{almost surely}$$

where $\mathcal{B}$ is the decision boundary



binary case ($T = 2$): $|\eta(X)| \geq \delta$ almost surely [Mammen and Tsybakov 1999]

Exponential Convergence for Margin Losses

$$\ell(w,y) = \phi(\langle w, y\rangle) \qquad \phi : \mathbb{R} \to [0,\infty)$$

assume

(i) $\|f_\lambda - f_\ell\|_\infty \underset{\lambda}{\to} 0$

(ii) $\mathbb{P}\{\|\widehat{f}_\lambda - f_\lambda\|_\infty > \epsilon\} \lesssim \exp(-n\epsilon^2/b^2)$

(iii) ℓ is Fisher consistent

(iv) ϕ is twice differentiable, decreasing and convex

Theorem (Vigogna, Meanti, De Vito, Rosasco)

If the hard-margin condition holds, then for $\lambda \leq \lambda_*$

$$\mathbb{E}|\mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}_*| \lesssim \exp(-n\, m(\delta)^2 \lambda / b^2)$$

$T = 2$: [Nitanda and Suzuki, 2019]

Exponential Convergence for Margin Losses

$$\ell(w,y) = \phi(\langle w, y\rangle) \qquad \phi : \mathbb{R} \to [0,\infty)$$

assume

(i) $\|f_\lambda - f_\ell\|_\infty \underset{\lambda}{\to} 0$

(ii) $\mathbb{P}\{\|\widehat{f}_\lambda - f_\lambda\|_\infty > \epsilon\} \lesssim \exp(-n\epsilon^2/b^2)$

(iii) ℓ is Fisher consistent

(iv) ϕ is twice differentiable, decreasing and convex

Theorem (Vigogna, Meanti, De Vito, Rosasco)

If the hard-margin condition holds, then for $\lambda \leq \lambda_*$

$$\mathbb{E}|\mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}_*| \lesssim \exp(-n\, m(\delta)^2 \lambda / b^2)$$

hard margin $\overset{\text{(iv)}}{\Rightarrow}$ f_ℓ is confident $\overset{\text{(i)}}{\Rightarrow}$ f_λ is confident $\overset{\text{(ii)}}{\Rightarrow}$ exponential convergence

$T = 2$: [Nitanda and Suzuki, 2019]

UniGe | MaLGa