LEARNING WITH CONSTRAINTS
Part I

Marco Gori
University of Siena (Italy)

ACDL 2018

Acknowledgments

Stefano Melacci, Marco Maggini, Claudio Saccà, Francesco Giannini
Giuseppe Marra - UNISI

Michelangelo Diligenti - UNISI and Google - Zurich

Luciano Serafini - FBK, Trento

Artur Gracez - City University, London

Michael Spranger - Sony

Luis Lamb - Institute of Informatics - UFRGS

Andrea Passerini - Univ. of Trento

Outline

Part I - Foundation

- Environment and constraints
- Bridging logic and real-valued constraints
- Representational issues
- Learning, inference, and reasoning with constraints

Outline

Part II

Neuro-symbolic models and case studies

- Supervised, unsupervised, semi-supervised learning
- Inference in formal logic
- Inference in the environment, and full inference
- Missing data and generation
- Recurrent neural networks
- Open issues

ENVIRONMENTS AND CONSTRAINTS

ACDL 2018

Supervised Learning

classic learning from examples

$x \in \mathcal{X}$

$\epsilon - \theta_j \parallel y_j(x) - f_j(x) \parallel_p \geq 0$

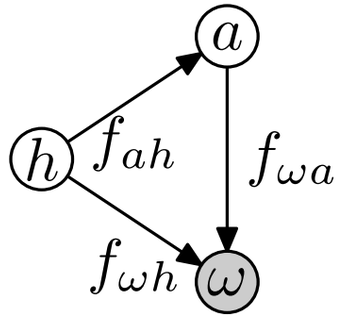
examples can be sets

$$\mathcal{E}_L = \{(\mathcal{X}_i, y_i) \in 2^{\mathcal{X}} \times \mathcal{Y}, i = 1, \ldots, m\}$$

as this becomes a box, the pair is a proposition

Enforcing Consistencies

$$f_{\omega h} : \mathcal{W} \to \mathscr{H} : h \to \omega(h),$$
$$f_{ah} : \mathcal{W} \to \mathcal{A} : h \to a(h),$$
$$f_{\omega a} : \mathcal{A} \to \mathcal{W} : a \to \omega(a),$$



$$f_{\omega h}(h) = f_{\omega a} \circ f_{ah}(h).$$

This functional equation is imposing the circulation of coherence. Since the functions are linear, this constraint can be converted to $w_{\omega h}h + b_{\omega h} = w_{\omega a}w_{ah}h + (w_{ah}b_{ah} + b_{\omega a})$. The equivalence $\forall h \in \mathbb{R}^+$ yields

$$w_{\omega a}w_{ah} - w_{\omega h} = 0,$$
$$w_{ah}b_{ah} + b_{\omega a} - b_{\omega h} = 0.$$

Diagnosis and Prognosis in Medicine

Pima Indian Diabetes Dataset

$$(MASS \geq 30) \wedge (PLASMA \geq 126) \Rightarrow positive$$

$$(MASS \leq 25) \wedge (PLASMA \leq 100) \Rightarrow negative$$

body mass index
blood glucose

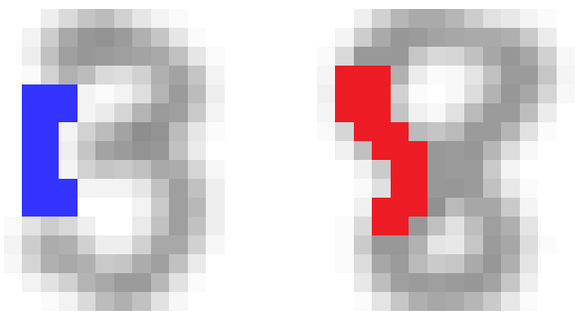
Wisconsin Breast Cancer Prognosis

$$(SIZE \geq 4) \wedge (NODES \geq 5) \quad \Rightarrow \quad recurrent$$

$$(SIZE \leq 1.9) \wedge (NODES = 0) \quad \Rightarrow \quad non\ recurrent$$

diameter of the tumor
number of metastasized lymph nodes

Handwritten Char Recognition



GRAYLEVEL >220 in blue region $\Rightarrow$ 3

GRAYLEVEL <160 in red region $\Rightarrow$ 8

blue region: a selection of (blue) coordinates out of the 256=16*16

red region: a selection of (red) coordinates out of the 256=16*16

Text Categorization

$$\text{graphic} \wedge \text{pixel} \wedge \text{bitmap} \Rightarrow \text{comp.graphics} \vee \text{comp.sys.ibm.pc.hardware}$$

keywords: input level

categories: decision level

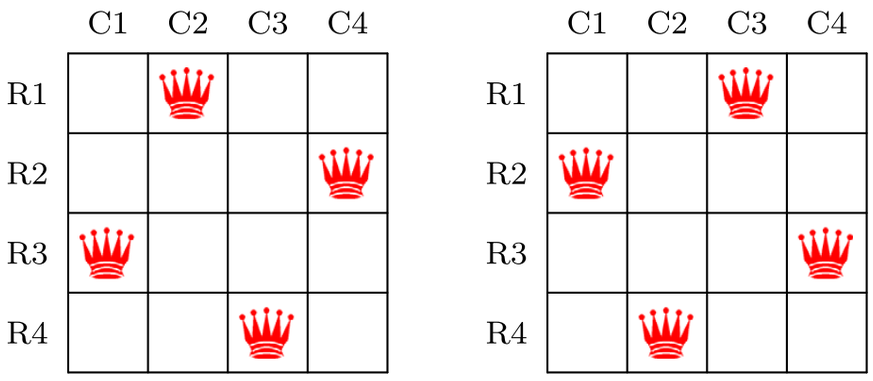
$$i. \quad \forall x \forall y \quad x \bowtie y \Leftrightarrow a(x) = a(y)$$ ← docs of the same author

$$ii. \quad \forall x \; c_1(x) \wedge c_2(x) \Rightarrow c_3(x)$$

$$iii \quad \forall x \; c_3(x) \Rightarrow c_4(x).$$

categories: decision level

N-Queens



$$(i)\ \forall i \in \mathbb{N}_n : \sum_{j \in \mathbb{N}_n} q_{i,j} = 1,$$

$$(ii)\ \forall j \in \mathbb{N}_n : \sum_{i \in \mathbb{N}_n} q_{i,j} = 1.$$

$$(i)\ \forall j = 1, \ldots, n : \ q_{1,j} + \sum_{k=1}^{n-j} q_{1+k,j+k} \leq 1,$$

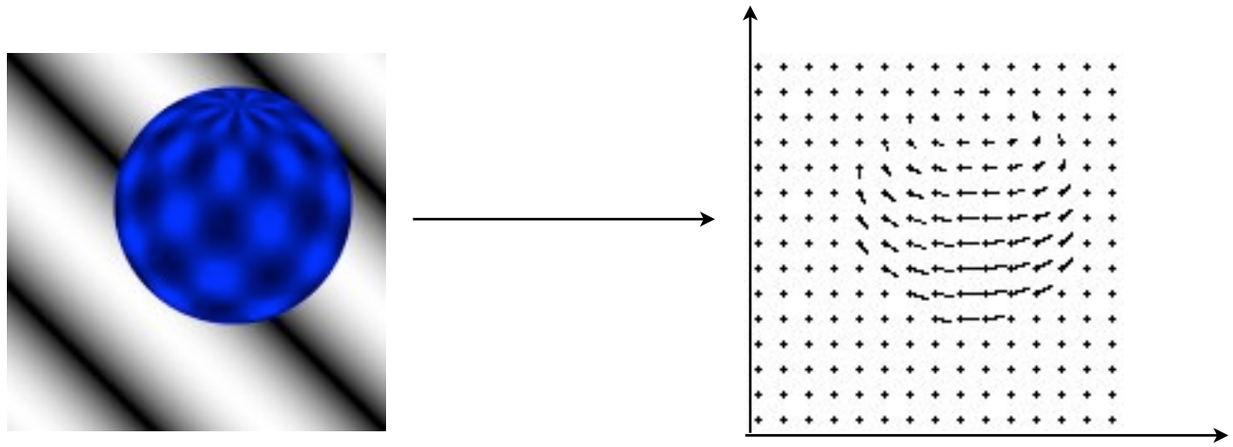
$$(ii)\ \forall i = 2, \ldots, n : \ q_{i,1} + \sum_{k=1}^{n-i} q_{i+k,1+k} \leq 1.$$

$$(i)\ \forall j = 1, \ldots, n : \ q_{1,j} + \sum_{k=1}^{n-j} q_{1+k,j-k} \leq 1,$$

$$(ii)\ \forall j = 2 \ldots, n : \ q_{i,n} + \sum_{k=1}^{n-i} q_{i+k,n-k} \leq 1.$$

$$q^\star = \arg\min_{q \in \mathcal{Q}} -\frac{1}{2} \sum_{(i,j) \in \mathscr{P}} (q_{i,j} - \overline{q}_{i,j})^2$$

Optical Flow in Computer Vision

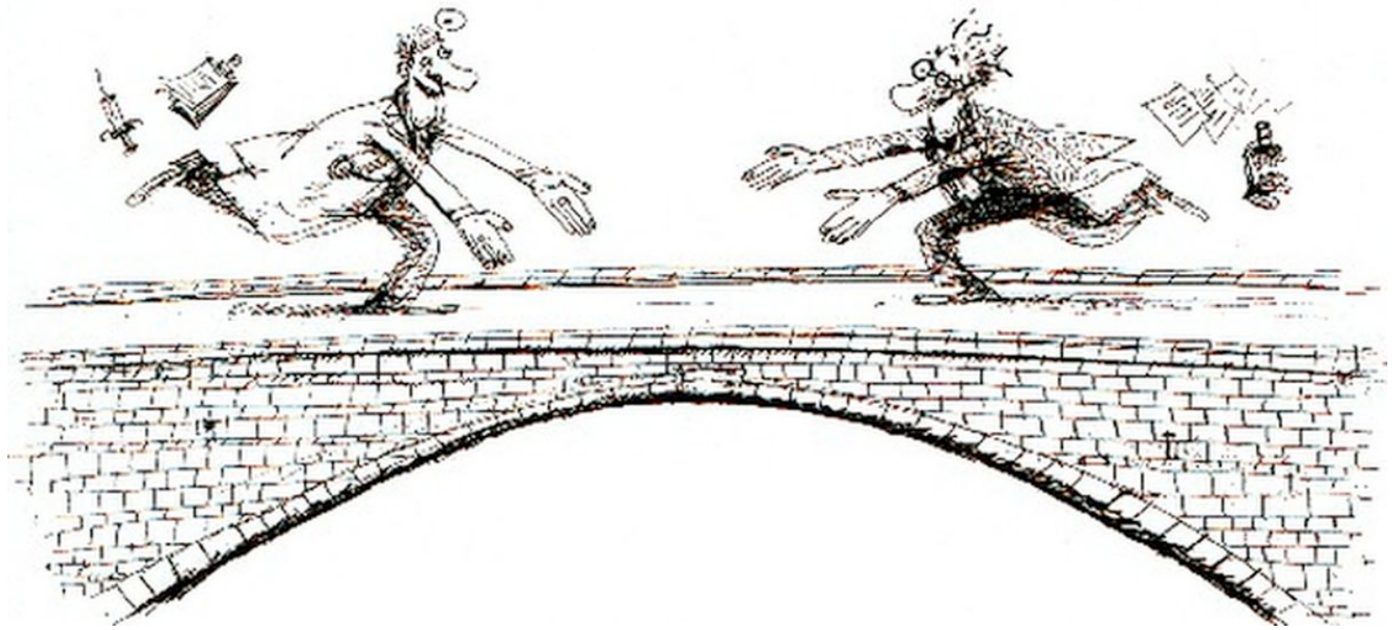


$E(x, y, t)$ *is constant*

$$u = \dot{x}, \quad v = \dot{y}$$

$$\forall t \; \forall x \; \forall y \quad \frac{\partial E}{\partial x} u + \frac{\partial E}{\partial y} v + \frac{\partial E}{\partial t} = 0$$

BRIDGING LOGIC AND REAL-VALUED CONSTRAINTS



learning relations and logic

"There are finer fish in the sea that have ever been caught," Irish proverb

Two Schools of Thought

(Formal) Logic

Optimization, statistics

Any break through the wall?

Logic by Real Numbers

$$\forall x \quad a(x) \wedge b(x) \Rightarrow c(x)$$

p-norm

$$\neg(a(x) \wedge b(x)) \vee c(x))$$
$$\neg\neg(\neg(a(x) \wedge b(x)) \wedge c(x))$$
$$\neg(a(x) \wedge b(x) \wedge \neg c(x))$$
$$1 - f_a(x) \cdot f_b(x) \cdot (1 - f_c(x)) = 1$$
$$f_a(x) f_b(x)(1 - f_c(x)) = 0$$

Logic by Real Numbers (con't)

$$\forall x \quad a(x) \wedge b(x) \Rightarrow c(x)$$

$$\neg(a(x) \wedge b(x) \wedge \neg c(x))$$

Gödel T-norm

$$1 - min\{f_a(x), f_b(x), 1 - f_c(x)\} = 1$$

$$min\{f_a(x), f_b(x), 1 - f_c(x)\} = 0$$

Tricky Issues

$1 \Rightarrow 2 \qquad f_1(x_1)(1 - f_2(x_2)) = 0$

$2 \Rightarrow 1 \qquad f_2(x_2)(1 - f_1(x_1)) = 0$

$2 \Leftrightarrow 1 \qquad f_1(x_1) + f_2(x_2) - 2f_1(x_1)f_2(x_2) = 0$

$$f_1^2(x_1) + f_2^2(x_2) - 2f_1(x_1)f_2(x_2)$$
$$= (f_1(x_1) - f_2(x_2))^2 = 0$$

$$f_1(x_1) = f_2(x_2)$$

Taxonomy

Gnecco et al, Neural Computation 2015

Definition 1 *(types of constraints). Let $\mathcal{X}$ denote a subset of the perceptual space $\mathbb{R}^d$, $\mathcal{F}$ a space of functions $f : \mathcal{X} \to \mathbb{R}^n$, $\mathcal{X}_i$ open subsets of $\mathcal{X}$, $\phi_i : \mathcal{X}_i \times \mathbb{R}^n \to \mathbb{R}$ and $\check{\phi}_i : \mathcal{X}_i \times \mathbb{R}^n \to \mathbb{R}$ continuous functions, $\Phi_i : \mathcal{F} \to \mathbb{R}$ and $\check{\Phi}_i : \mathcal{F} \to \mathbb{R}$ continuous functionals, and m_H, m_I, $\check{m}_H$, and $\check{m}_I$ positive integers. We consider the following types of constraints:*

i. *Holonomic (ho) bilateral (bi):*

$$\forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \ \phi_i(x, f(x)) = 0, \ i = 1, \ldots, m_H.$$

ii. *Holonomic (ho) unilateral (un):*

$$\forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \ \check{\phi}_i(x, f(x)) \geq 0, \ i = 1, \ldots, \check{m}_H.$$

iii. *Isoperimetric (is) bilateral (bi):*

$$\Phi_i(f) = 0, \ i = 1, \ldots, m_I.$$

iv. *Isoperimetric (is) unilateral (un):*

$$\check{\Phi}_i(f) \geq 0, \ i = 1, \ldots, \check{m}_I.$$

v,vi. *Pointwise (pw) bilateral (bi) and pointwise (pw) unilateral (un): as constraints i and ii, respectively, with each $\mathcal{X}_i$ made up of finitely many points (in this case, the continuity of ϕ_i—respectively, of $\check{\phi}_i$—is required with respect to the second vector argument).*

Taxonomy

Examples of Constraints

Number of Constraint	Linguistic Description	Real-Valued Representation	Classification	Typical interpretation
i	*i*th supervised pair for classification	$y_\kappa \cdot f(x_\kappa) - 1 \geq 0$	(pw,un)	(sf)
ii	Probabilistic normalization	$\forall x \in \mathcal{X}:$	(ho,bi)	(hr)
	for classification	$f_1(x) + f_2(x) + f_3(x) = 1;$		
		$\forall x \in \mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathcal{X}:$	(ho,un)	(hr)
		$f_j(x) \geq 0 \ (j = 1, 2, 3)$		
iii	Probabilistic normalization	$\int_{\mathcal{X}} f(x)dx = 1;$	(is,bi)	(hr)
	of a density function	$\forall x \in \mathcal{X}: \quad f(x) \geq 0$	(ho,un)	(hr)
iv	Coherence constraint	$\forall x = (x_1, x_2) \in \mathcal{X}:$		
	(2 classes)	$f_1(x_1) \cdot f_2(x_2) \geq 0$	(ho,un)	(sf)
v	Asset allocation	$\forall x \in \mathcal{X}:$	(ho,bi)	(hr)
	Cash, bond, and stock in USD	$f_c^d(x) + f_b^d(x) + f_s^d(x) = t_d(x);$		
	Cash, bond, and stock in euro	$f_c^e(x) + f_b^e(x) + f_s^e(x) = t_e(x);$		
	Overall investment in USD and euro	$t_d(x) + c \cdot t_e(x) = T$		
vi	Optical flow	$\frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t} = 0$	(ho,bi)	(sf)
vii	Wernicke's aphasia (*i*th) rule:	$\forall x \in \mathcal{X}_i: \quad y_{we}^i \cdot f_{we}(x) - 1 \geq 0$	(ho,un)	(sf)
	if $P1 > 3$ and $P2 \leq 4$ and $P5 > 2$			
	and $N5 \leq 22$ and $V0 \leq 62$			
	and $V1 > 38$, then W			
viii	Document classification:	$f_{na}(x) \cdot f_{nn}(x) \cdot (1 - f_{ml}(x)) \leq \epsilon$	(ho,un)	(sf)
	$\forall x : na(x) \wedge nn(x) \Rightarrow ml(x)$	($\epsilon > 0$ and $\epsilon \simeq 0$)		

REPRESENTATIONAL ISSUES

"the simplest solution" compatible with the constraints

- Intuition
- representational issues
- dealing with logic constraints

New Protocols for Learning!

Beyond induction and black-box perception
Abstract representations of reality

A breakthrough in the communication protocol of statistical machine learning:
we need a **language** to express properties

A New Communication Protocol

data + constraints

$\forall x \quad \Phi(x, f(x)) = 0$ from constraints to

$\sum_{\kappa \in U} \phi^2(x_\kappa, f(x_\kappa))$ loss functions

A New Communication Protocol

data + constraints

- Supervised
- Unsupervised
- Semi-supervised

learning problem

?

learning of constraints

cognitive laws

$\phi_i(x, f(x)) = 0$

tasks

$f(x)$

perceptual space

x

The New Role of Learning Data

$$\texttt{hair}(\texttt{x}) \Rightarrow \texttt{mammal}(\texttt{x})$$
$$\texttt{mammal}(\texttt{x}) \wedge \texttt{hoofs}(\texttt{x}) \Rightarrow \texttt{ungulate}(\texttt{x})$$
$$\texttt{ungulate}(\texttt{x}) \wedge \texttt{white}(\texttt{x}) \wedge \texttt{blackstripes}(\texttt{x}) \Rightarrow \texttt{zebra}(\texttt{x}).$$

$$f_{\texttt{hair}}(x)(1 - f_{\texttt{mammal}}(x)) = 0$$
$$f_{\texttt{mammal}}(x) f_{\texttt{hoofs}}(x)(1 - f_{\texttt{ungulate}}(x)) = 0$$
$$f_{\texttt{ungulate}}(x) f_{\texttt{white}}(x) f_{\texttt{blackstripes}}(x)(1 - f_{\texttt{zebra}}(x)) = 0.$$

penalty functions

perceptual space

x

cognitive laws

$\phi_i(x, f(x)) = 0$

?

tasks

$f(x)$

perceptual space

x

The Marriage of Parsimony Principle and Constraints

Constraints turn out to be loss functions

keep these loss functions as small as possible

$$f_{\texttt{hair}}(x)(1 - f_{\texttt{mammal}}(x)) = 0$$
$$f_{\texttt{mammal}}(x) f_{\texttt{hoofs}}(x)(1 - f_{\texttt{ungulate}}(x)) = 0$$
$$f_{\texttt{ungulate}}(x) f_{\texttt{white}}(x) f_{\texttt{blackstripes}}(x)(1 - f_{\texttt{zebra}}(x)) = 0.$$

penalty functions

perceptual space

x

Parsimony Principle

$\| f \|$

f

$f_{\texttt{hair}}$

$f_{\texttt{hoofs}}$

$f_{\texttt{mammal}}$

$f_{\texttt{ungulate}}$

$f_{\texttt{white}}$

$f_{\texttt{blackstripes}}$

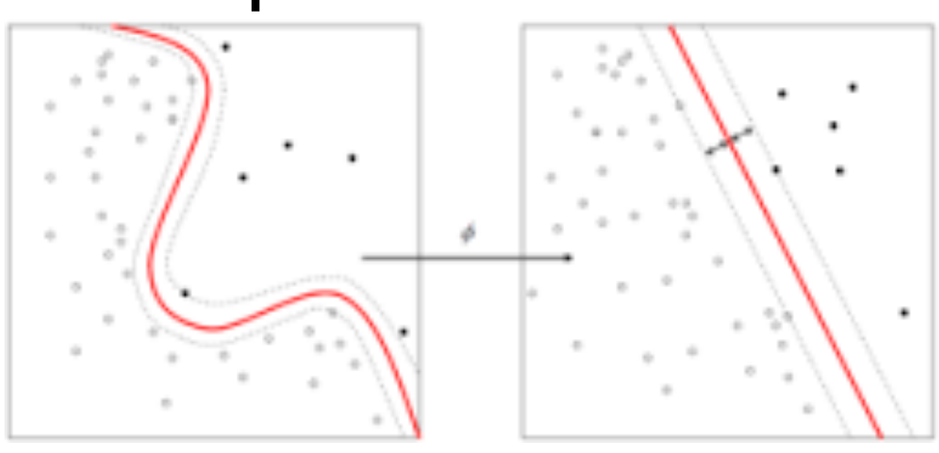
$f_{\texttt{zebra}}$

How to represent the tasks?

$$f\ ?$$

Primal space

Dual Space

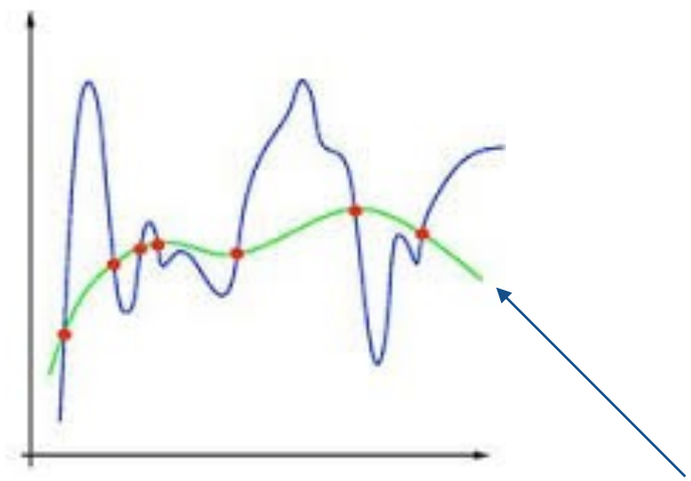


Kernel Machines

. . .

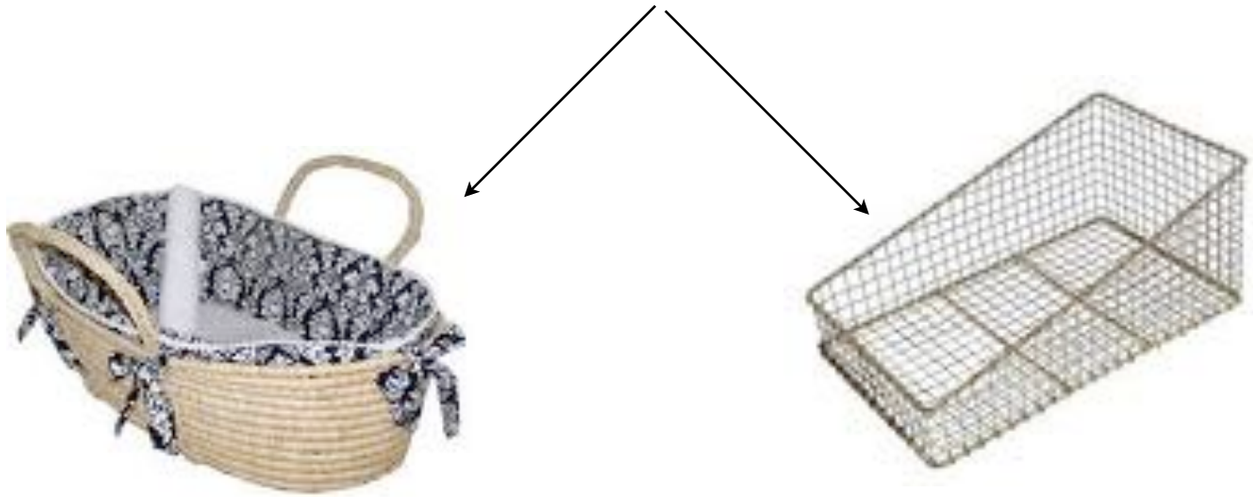
Parsimony Principle
smoothness of the tasks

$$\|f\|^2 := b_0 \int_{\mathcal{X}} f^2(x)dx + b_1 \int_{\mathcal{X}} \left(\frac{df}{dx}\right)^2 dx$$



Occam's razor, lex parsimoniae

Ambient Space



RKHS

$$\mathcal{X} \subset \mathbb{R}^d \quad f = [f_1, \ldots, f_n]' \quad f_j : \mathcal{X} \to \mathbb{R}$$

$$\forall j \in \mathbb{N}_n : \quad f_j \in W^{k,p}$$

Search in Sobolev spaces:
it is related to the topic of learning kernels!

Semi-norm in Sobolev Spaces

$$P = \sum_{|\alpha|<m} a_\alpha D_x^\alpha = \sum_{|\alpha|<m} a_\alpha \left(\frac{\partial}{\partial x_1} + \ldots + \frac{\partial}{\partial x_d}\right)^\alpha$$

∞ $\quad a_\alpha \in C^\infty$

under proper boundary conditions ...

$$P = \sum_{h=0}^{m} a_h \sum_{|\alpha|=h} \frac{h!}{\alpha!} \left(\frac{\partial}{\partial x}\right)^\alpha \qquad P^\star = \sum_{h=0}^{m} (-1)^h a_h \sum_{|\alpha|=h} \frac{h!}{\alpha!} \left(\frac{\partial}{\partial x}\right)^\alpha$$

Given P and $\gamma_i > 0, \ldots, i = 1, \ldots, n$

$$E(f) = \| f \|_{P,\gamma} = \sum_{j=1}^{n} \gamma_j < Pf_j, Pf_j > = \sum_{j=1}^{n} \gamma_j < f_j, P^\star P f_j > = \sum_{j=1}^{n} \gamma_j < f_j, L f_j >$$

Parsimony Principle

$\mathcal{F}_\phi$ admissible w.r.t the collection of constraints $\mathcal{C}_\phi$

$$f^\star = argmin_{f \in \mathcal{F}_\phi} \parallel f \parallel_{P,\gamma}$$

strictly (hard)

partially (soft)

inference in the environment!

check of a "new" constraint

$$\forall x \quad \phi(x, f^\star(x), Df^\star(x)) = 0 \ ?$$

Representation of the Solution
by variational calculus

hard constraints

$$\forall x \in \mathcal{X}_i \subset X : \phi_i(x, f(x)) = 0,\ i \in I\!N_m \qquad \frac{D(\phi_1, \ldots, \phi_m)}{D(f_1, \ldots, f_m)} \neq 0$$

$$\mathcal{L}(f) = \| f \|^2_{P,\gamma} + \sum_{i=1}^{m} \int_{\mathcal{X}} \lambda_i(x) \cdot \phi_i(x, f(x)) dx$$

Lagrangian approach

$$Lf(x) + \sum_{i=1} \lambda_i(x) \cdot \nabla_f \phi_i(x, f(x)) = 0$$

Euler-Lagrange equations

$$Lg = \delta$$

Green function

$$\omega_i(\cdot) = -\lambda_i(\cdot) \nabla_f \phi_i(\cdot, f^\star(\cdot))$$

reaction of the constraint

support constraints

$$f^\star(\cdot) = \sum_{i=1}^{m} g(\cdot) \otimes \omega_i(f^\star(\cdot))$$

Fredholm eq. (II kind)

"merging of two ideas ..."

Lagrange Multipliers and Probability Density

hard constraints

$$\forall x \in \mathcal{X}_i \subset X : \phi_i(x, f(x)) = 0,\ i \in \mathbb{N}_m$$

$$\mathcal{L}(f) = \| f \|^2_{P,\gamma} + \sum_{i=1}^{m} \int_{\mathcal{X}} \lambda_i(x) \check{\phi}_i(x, f(x)) dx$$

soft constraints

$$\mathcal{L}(f) = \| f \|^2_{P,\gamma} + C \sum_{i=1}^{m} \int_{\mathcal{X}} p_i(x) \check{\phi}_i(x, f(x)) dx$$

concepts acquired in strongly skewed distributions

Representation (soft constraints)

$$E(f) = \| f \|^2_{P,\gamma} + C \cdot 1' < \Xi(x, f(x))) >$$

$$Lf^\star(x) + C \cdot 1' \nabla_f \Xi(x, f^\star(x)) = 0$$

$$\omega_i(f^\star(x)) = -C \cdot \nabla_f \Xi_i(x, f^\star(x))$$

regularization parameter

$$f^\star = -C \sum_{i=1}^{m} g(\cdot) \otimes \nabla_f \Xi_i(\cdot, f^\star(\cdot)) = \sum_{i=1}^{m} g \otimes \omega_i(f^\star)$$

Two Remarkable Examples

Optical flow *(Horn and Schunck (1981))*

$$\frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t} = 0$$

$$\int_X \int_X \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right] dxdy$$

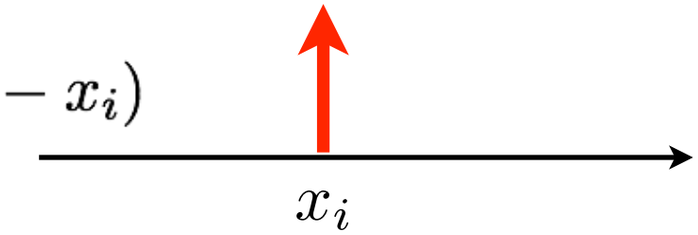
Learning from examples *(Poggio and Girosi (1989))*

$$E(f) = C \sum_{i=1}^{m} V(x_i, f(x_i)) + \frac{1}{2} < Pf, Pf >$$

Where Do Kernel Machines Come From?

$$Lf^\star + C\sum_{i=1}^{m} V_f'(x_i, f^\star(x_i))\delta(x - x_i) = 0$$

$$\omega_i(f^\star(x)) \quad = \quad -C \cdot V_f'(x_i, f^\star(x_i)) \cdot \delta(x - x_i)$$



reaction of the constraint

$$f^\star(x) = \sum_{i=1}^{m} \alpha_i g(x, x_i)$$ finite convolution

When Kernels Arise from Regularization Operators

$Lg = \delta$ Green function / "plain kernel"

$L = d^4/dx^4$ $g(x) = |x|^3$

$L = (\sigma^2 I - \nabla^2)^n$ Sobolev spline kernel

$$L = \sum_{\kappa=0}^{\infty} (-1)^\kappa \frac{\sigma^{2\kappa}}{\kappa! 2^\kappa} \nabla^{2\kappa}$$ Gaussian kernel

Polynomial kernels don't come from regularization operators!

Multi-Internvals

Back to kernels (under some hyp)!

$$\phi_i(x, f(x)) := max\{0, 1 - y_i f(x_i)\} \cdot c_{\mathcal{X}_i}(x)$$

$$\omega_i = -\lambda \nabla \phi(x, f(x)) \propto c_{\mathcal{X}_i}(x)$$

sign consistency uniform weight reaction

$g \otimes c_{\mathcal{X}_i}$ ← constraint reaction

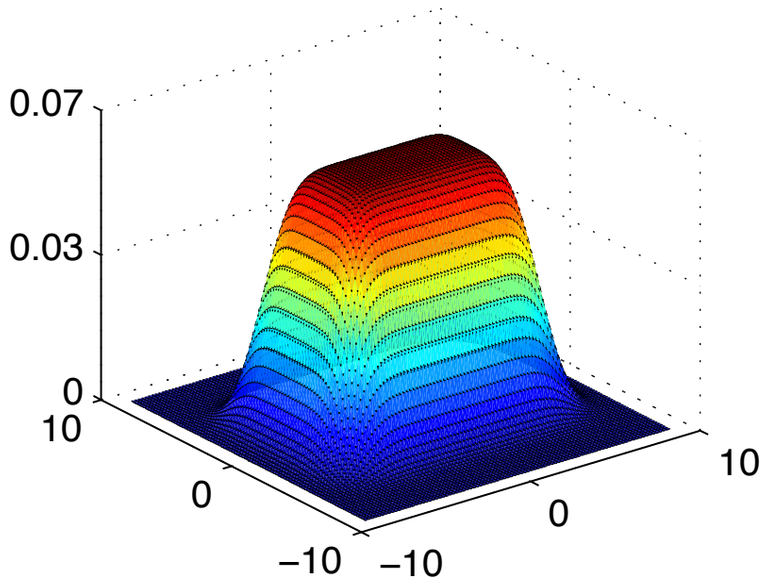
box kernels!

the case of soft-constraints ...

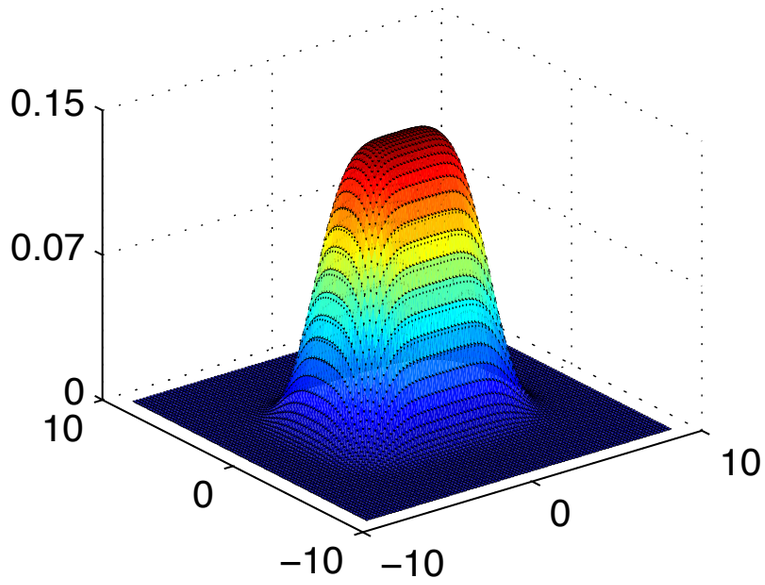
Box Kernels

$$g \otimes c\chi_i$$

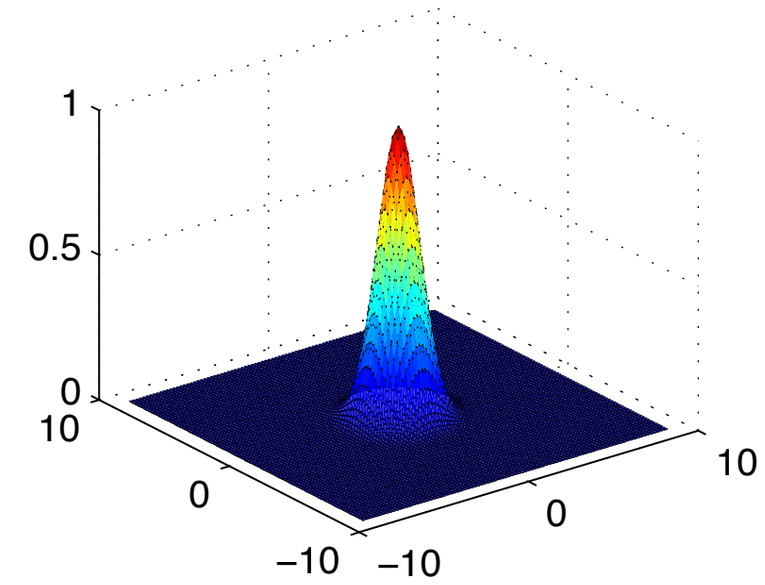
Gaussian (plain kernel) $\Longrightarrow$ $\prod_{i=1}^{d} \frac{(\sqrt{2\pi}\sigma)}{2}(erfc(\frac{x^i - b^i_j}{\sqrt{2}\sigma}) - erfc(\frac{x^i - a^i_j}{\sqrt{2}\sigma}))$



$[-6, -4] \times [6, 4]$



$[-3, -2] \times [3, 2]$



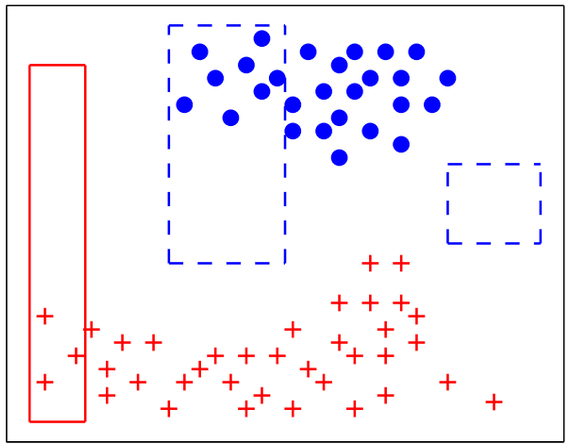
degeneration to the Gaussian

$(0, 0)$

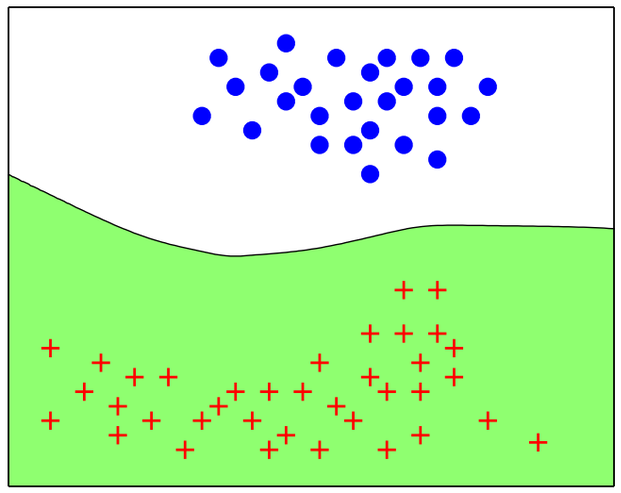
plain kernel (Gaussian) + multi-interval knowledge = box kernel

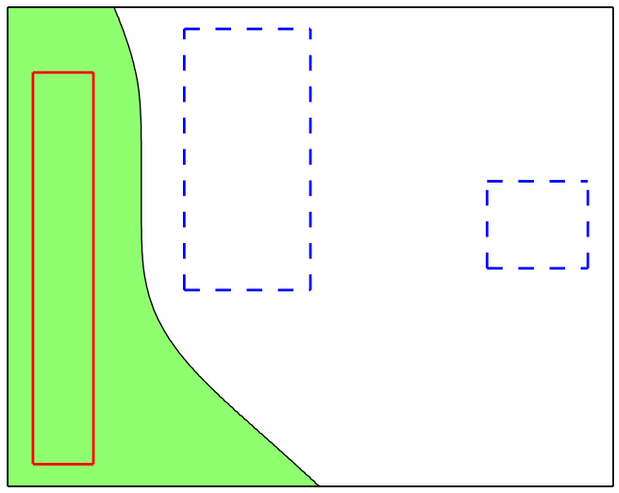
response to the “rectangular impulse”

ACDL 2018

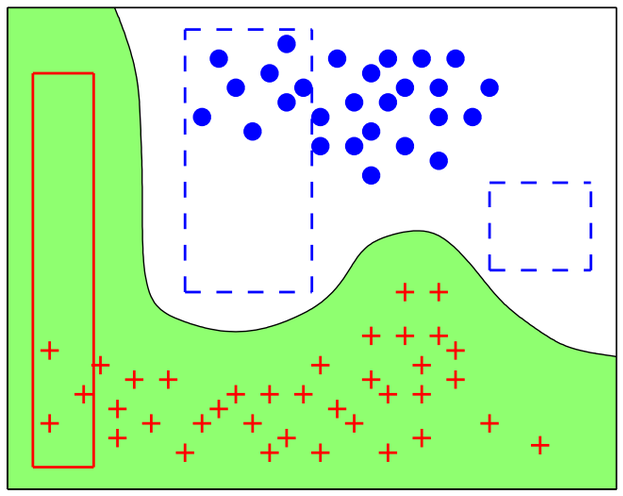


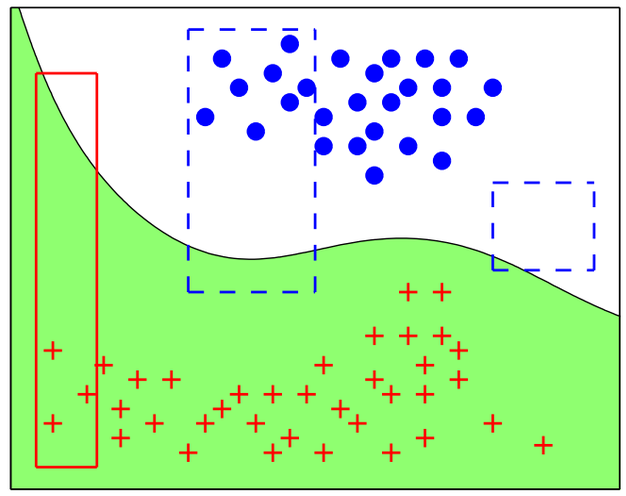
points only





boxes only





changing the regularization parameter

ACDL 2018

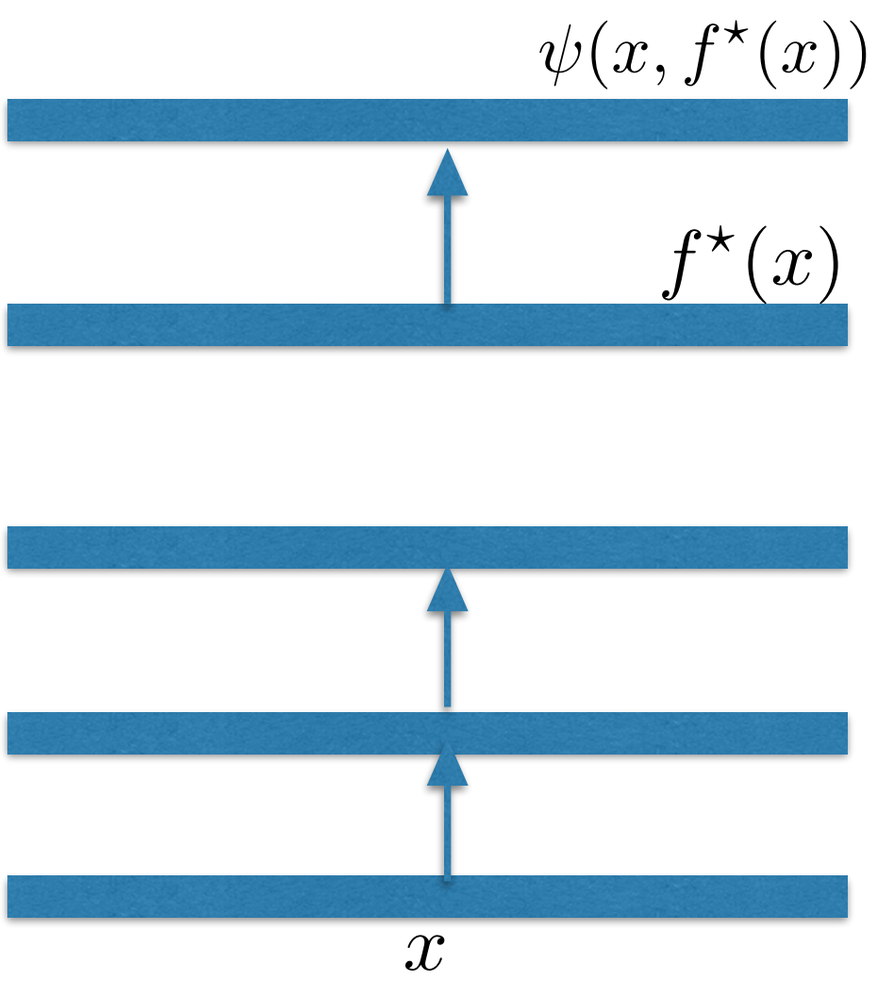
LEARNING, INFERENCE, AND REASONING

ACDL 2018

From parsimonious inference to induction

$f^\star = argmin_{f \in \mathcal{F}_\phi} \parallel f \parallel_{P,\gamma}$ learning and the active role

$\forall x \;\; \phi(x, f^\star(x), Df^\star(x)) = 0$? inference



inductive learning of new constraints by MMI clustering

a cyclic process: learning from and of constraints!