Probability, Agnosticism, and Guarantees in Inductive Learning Processes

Marco C. Campi

University of Brescia

Italy

(*joint work with Simone Garatti and Algo carè*)

Biblioteca Nazionale Braidense - 2019

“our” classroom at the Politecnico

Probability, Agnosticism, and Guarantees in Inductive Learning Processes

R nominal value 2G Ohm
standard deviation $= 0.1$G Ohm

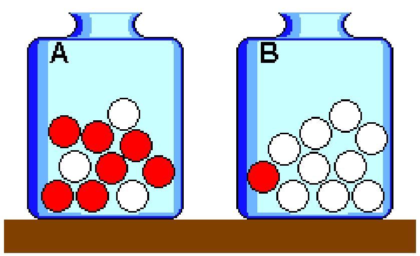
$$Y = R + N$$

- *construct bi-variate distribution for (R,Y)* $\longrightarrow$ *use Bayes rule*

- *what does prior on R describes?* $\longrightarrow$ *subjectivist interpretation of probability*

a misconception

- *subjectivism goes hand in hand with fully probabilistic models (and Bayesian updating)*



$$p = \frac{m}{50}, \quad m \in \{0, \ldots, 50\}$$

agnosticism

- *Every probability distribution is possible*

→ *practical reasons*

→ *philosophical reasons*

agnosticism

- *Every probability distribution is possible*

⟶ *practical reasons*

goal: predict whether the defibrillator shock will be effective

agnosticism

- *Every probability distribution is possible*

→ *practical reasons*

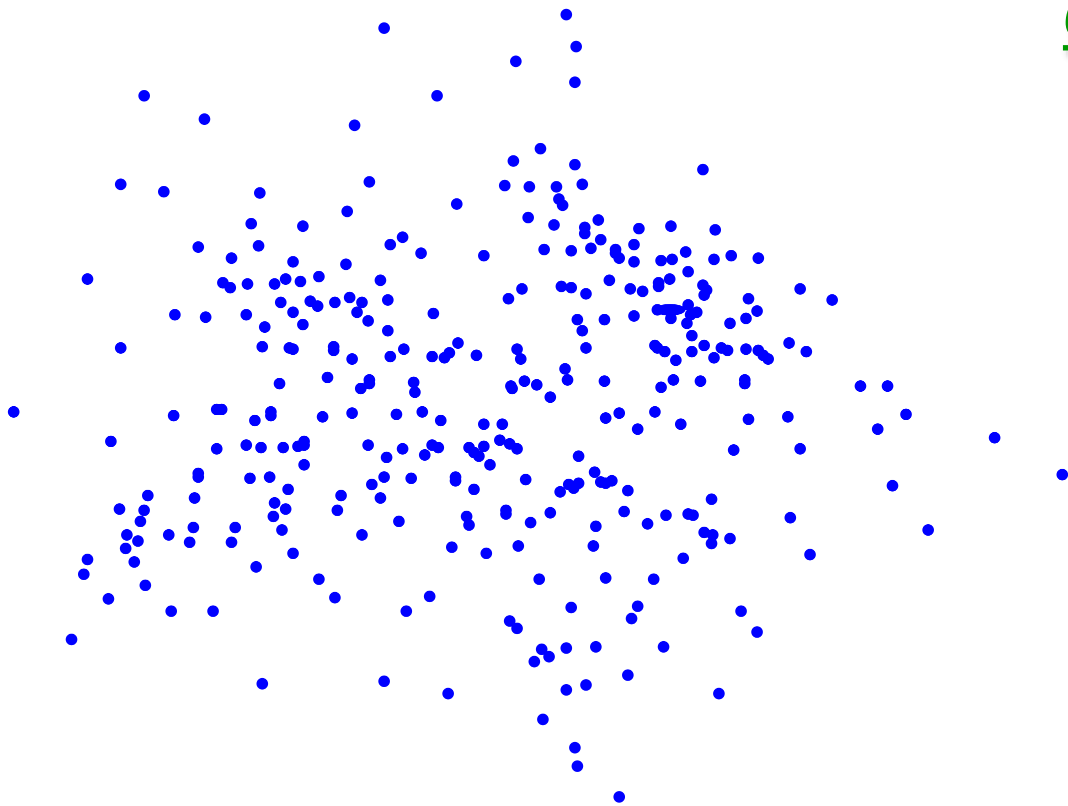
→ *philosophical reasons*

can knowledge be created out of lack of knowledge in the light of observations?

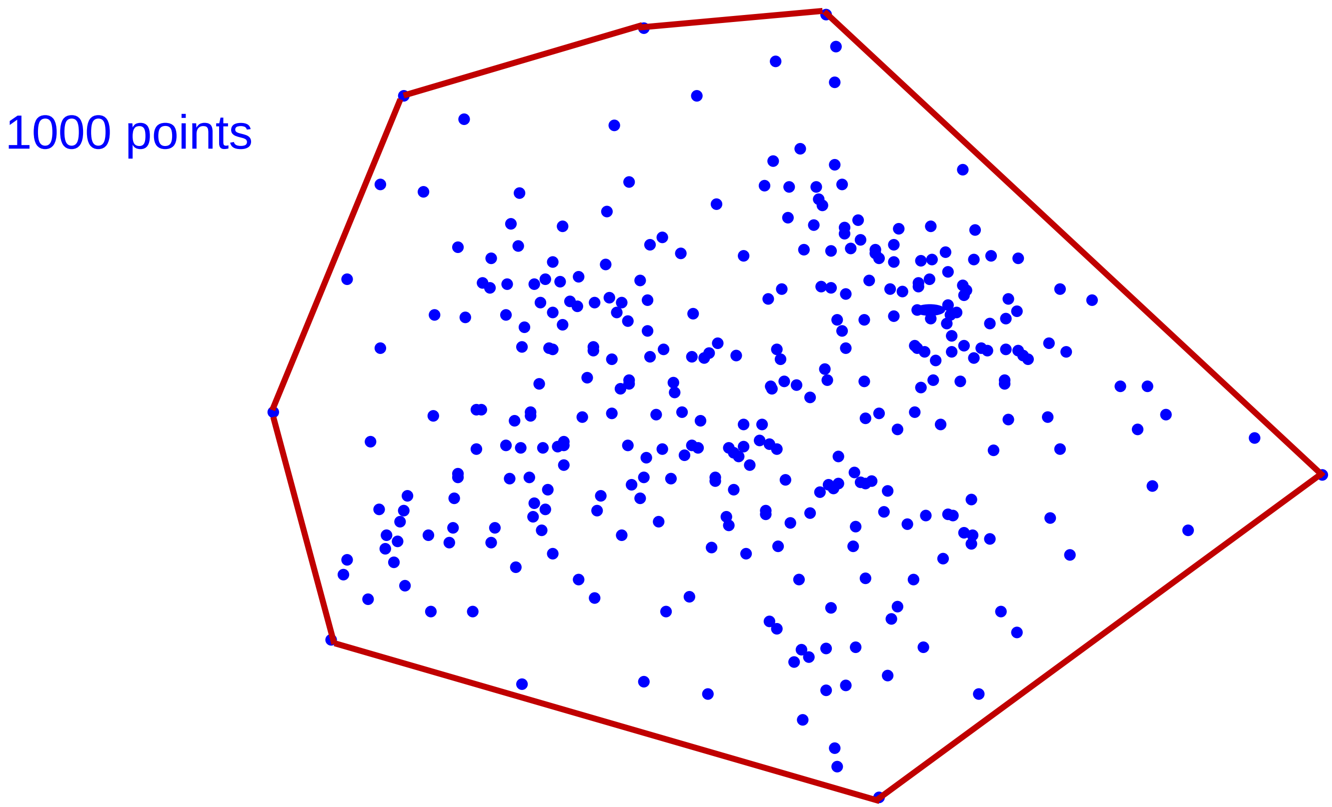
is it possible to derive mathematically rigorous and practically useful data-driven techniques within an agnostic setup?

convex hull

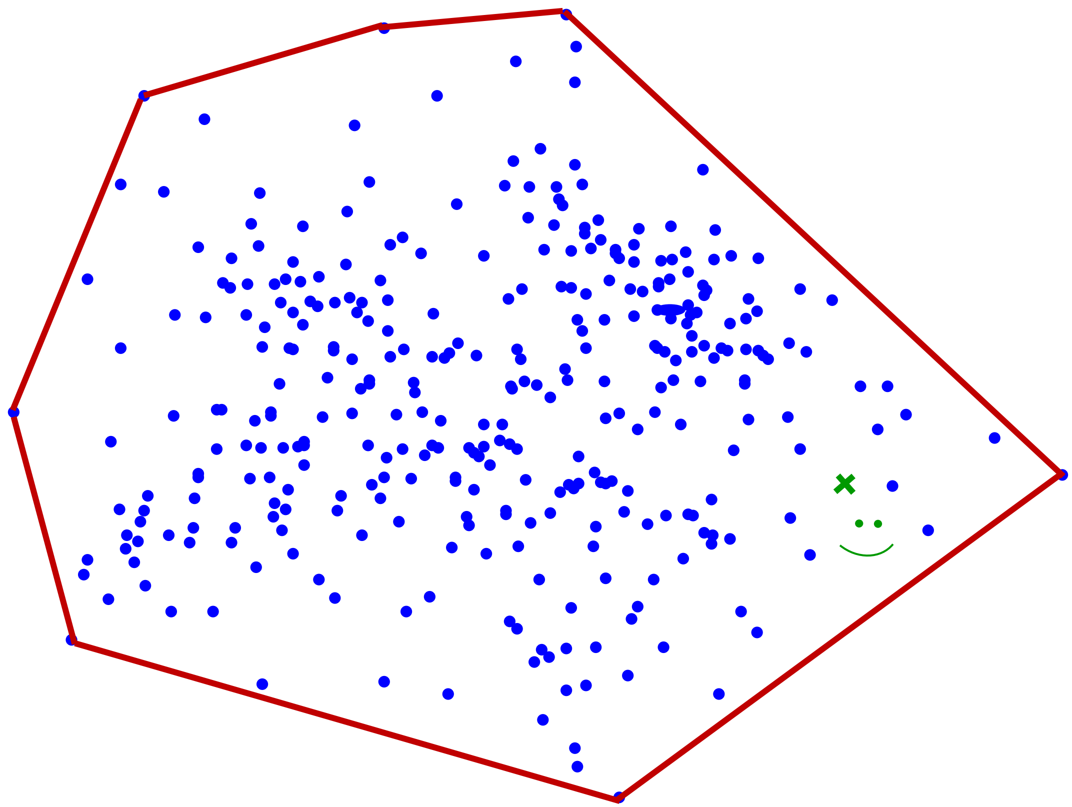
1000 points



Marco Fuhrman 60th birthday

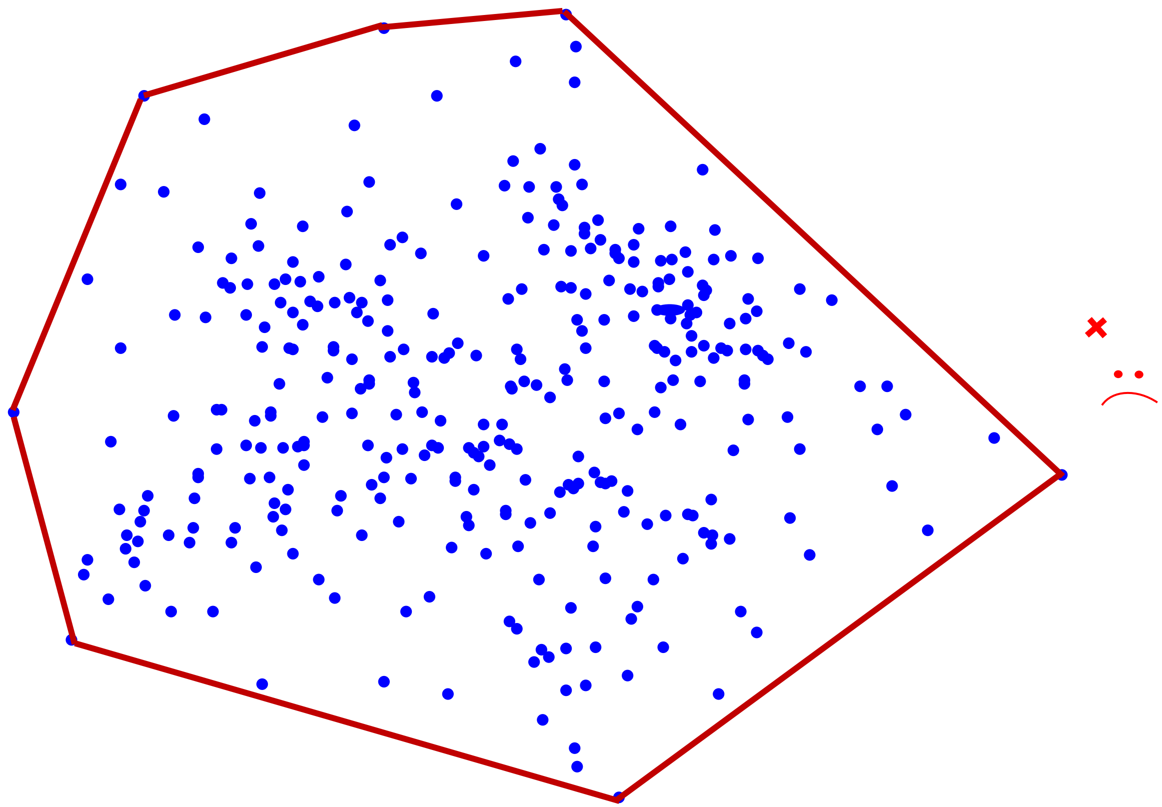

1000 points

1000 points

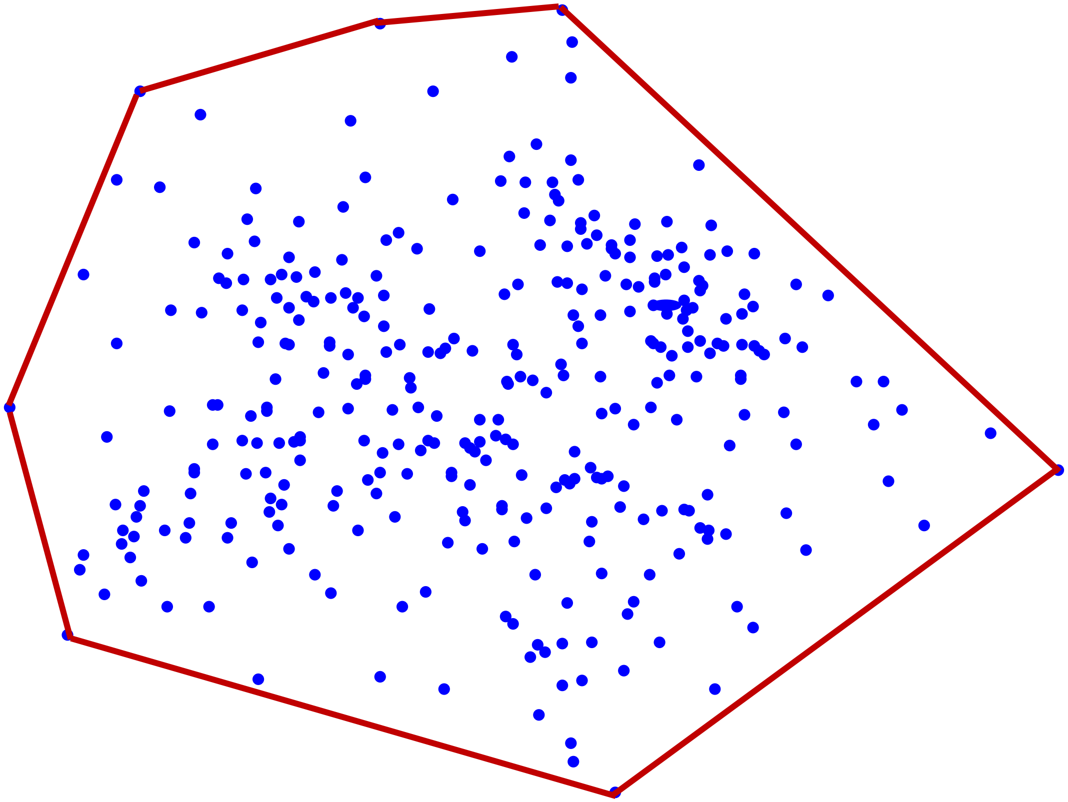


Marco Fuhrman 60th birthday

1000 points

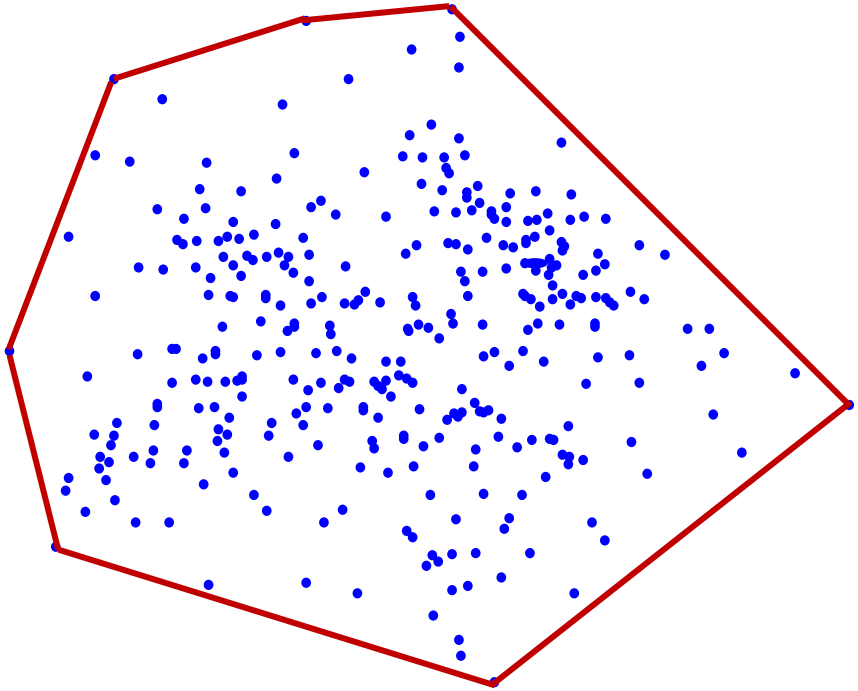


1000 points

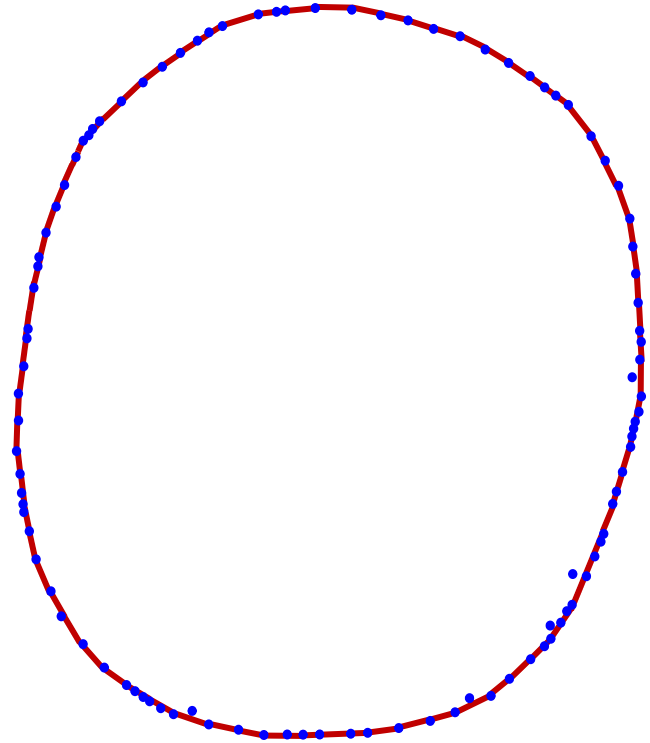


7 points at the boundary

Marco Fuhrman 60th birthday



7 points at the boundary



988 points at the boundary

Marco Fuhrman 60th birthday

complexity

$\cancel{p_1}, p_2, \cancel{p_3}, \cancel{p_4}, p_5, \ldots, \cancel{p_{1000}}$

"7 points at the boundary" $\Longrightarrow$ "if only these 7 points are maintained, "Opt" gives the same solution" complexity

$\cancel{p_1}, p_2, \cancel{p_3}, \cancel{p_4}, p_5, \ldots, \cancel{p_{1000}}$

"7 points at the boundary" $\Longrightarrow$ "if only these 7 points are maintained, "Opt" gives the same solution"

complexity of representation

intuitively: high complexity $\Longrightarrow$ high risk

wait-&-judge scenario optimization

Theorem (with S. Garatti)

$Pr\{\text{risk} \leq \epsilon(\text{complexity})\} \geq 1 - \beta$

where

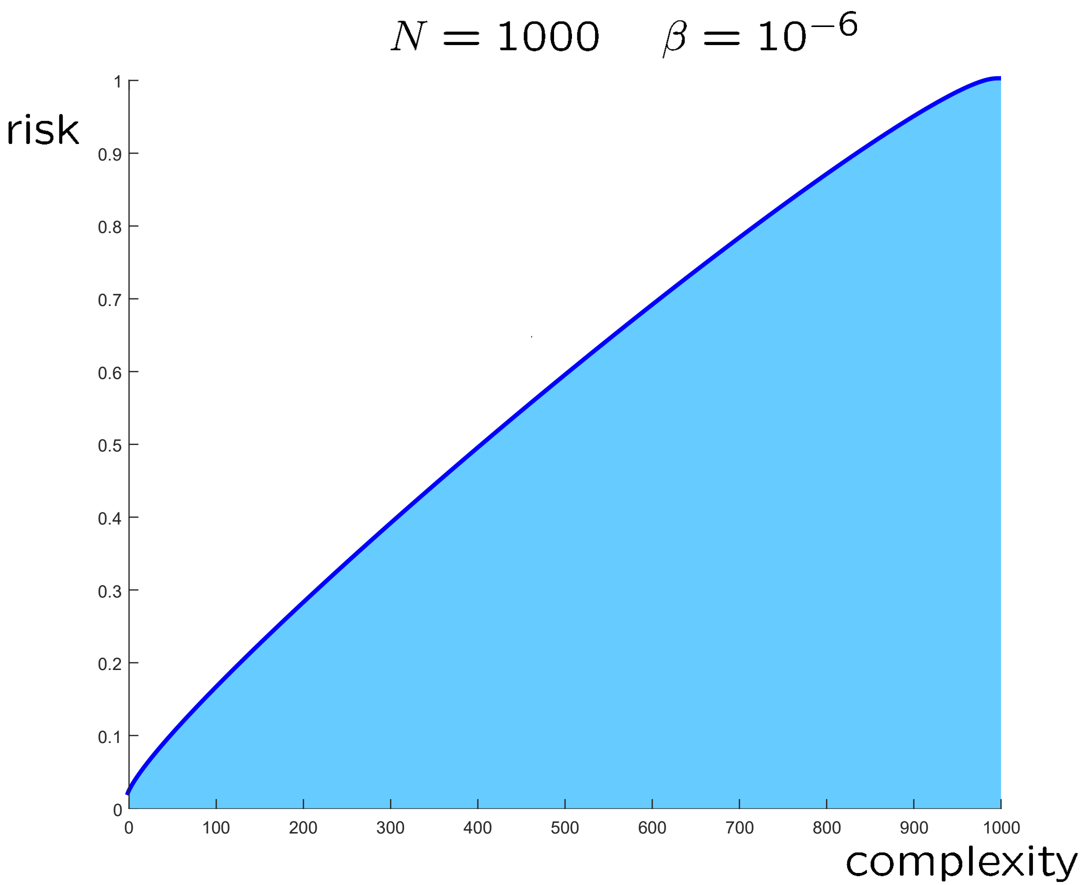
$$\beta = \inf_{\xi(\cdot) \in \mathsf{P}_N} \quad \xi(1)$$

subject to: $\displaystyle \frac{1}{k!}\frac{d^k}{dt^k}\xi(t) \geq \binom{N}{k} t^{N-k} \cdot 1\{t \in [0, 1-\epsilon(k))\} \quad t \in [0,1],$

$k = 0, 1, \ldots, N.$

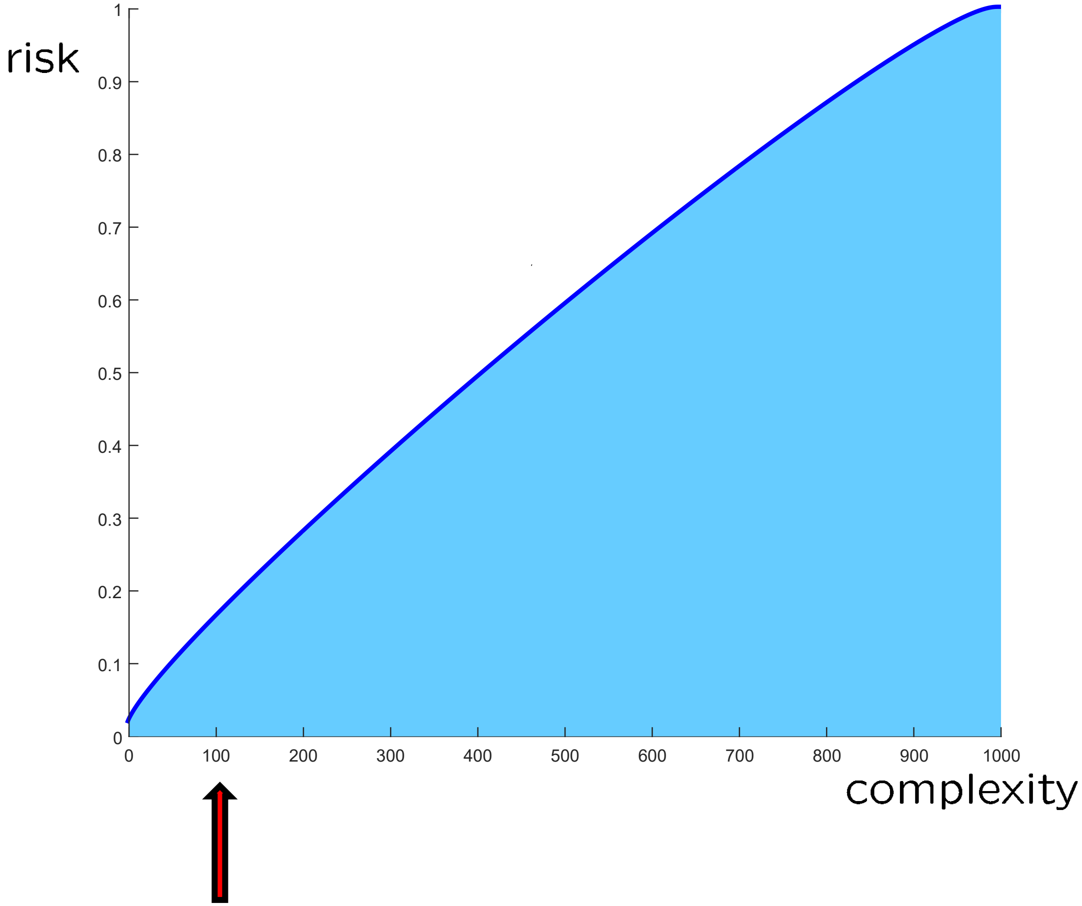
[M.C. Campi, S. Garatti, Mathematical Programming, 2016]

wait-&-judge scenario optimization



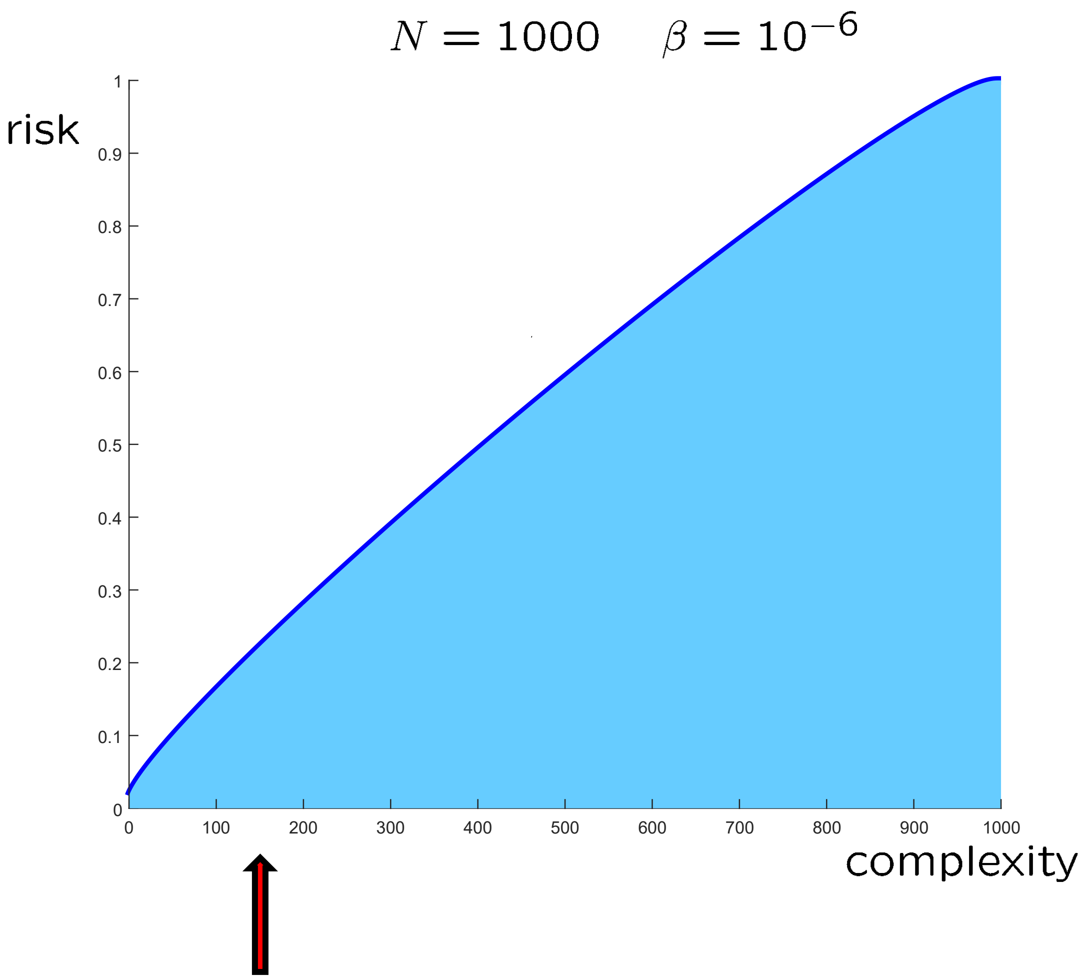
wait-&-judge scenario optimization

$N = 1000 \quad \beta = 10^{-6}$



complexity $= 100 \;\rightarrow\;$ risk ≤ 0.16

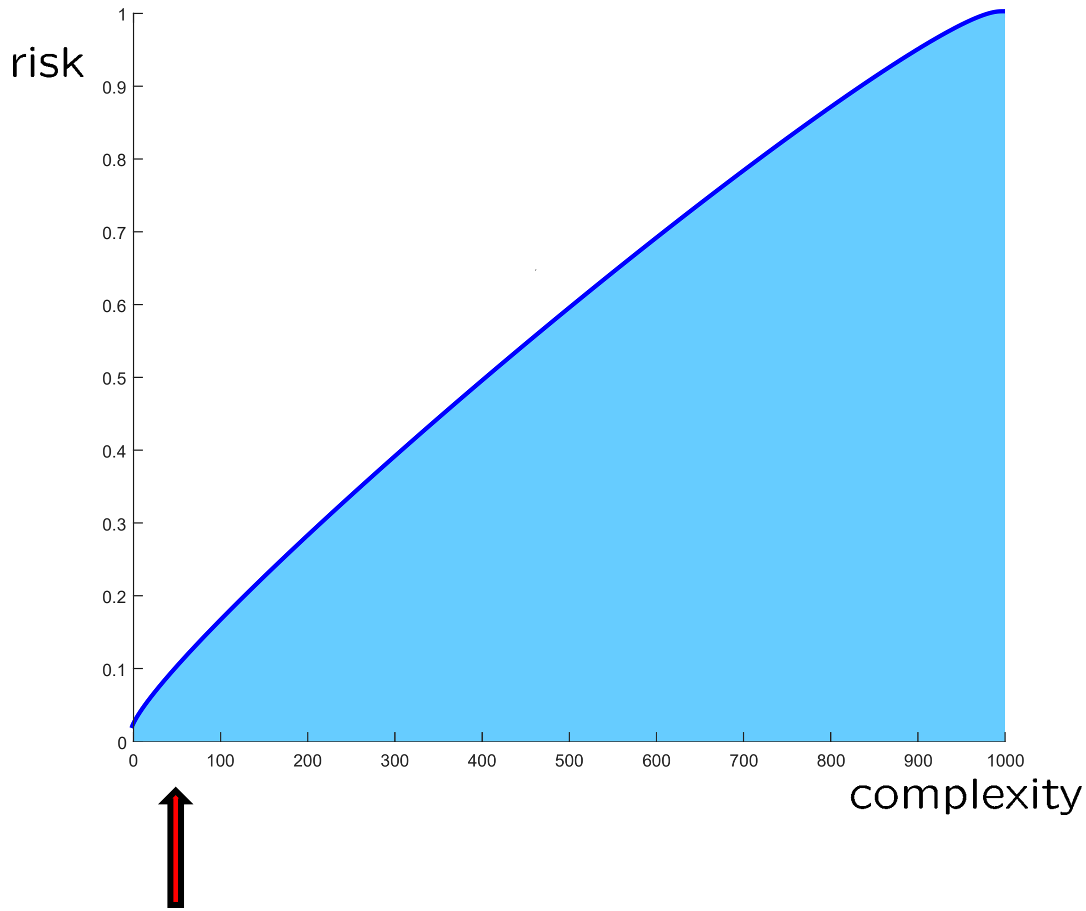
wait-&-judge scenario optimization



complexity $= 150 \;\rightarrow\;$ risk ≤ 0.22

wait-&-judge scenario optimization

$N = 1000 \quad \beta = 10^{-6}$



complexity $= 50 \rightarrow$ risk ≤ 0.099

wait-&-judge scenario optimization

suppose: no concentrated mass

Theorem (with S. Garatti)

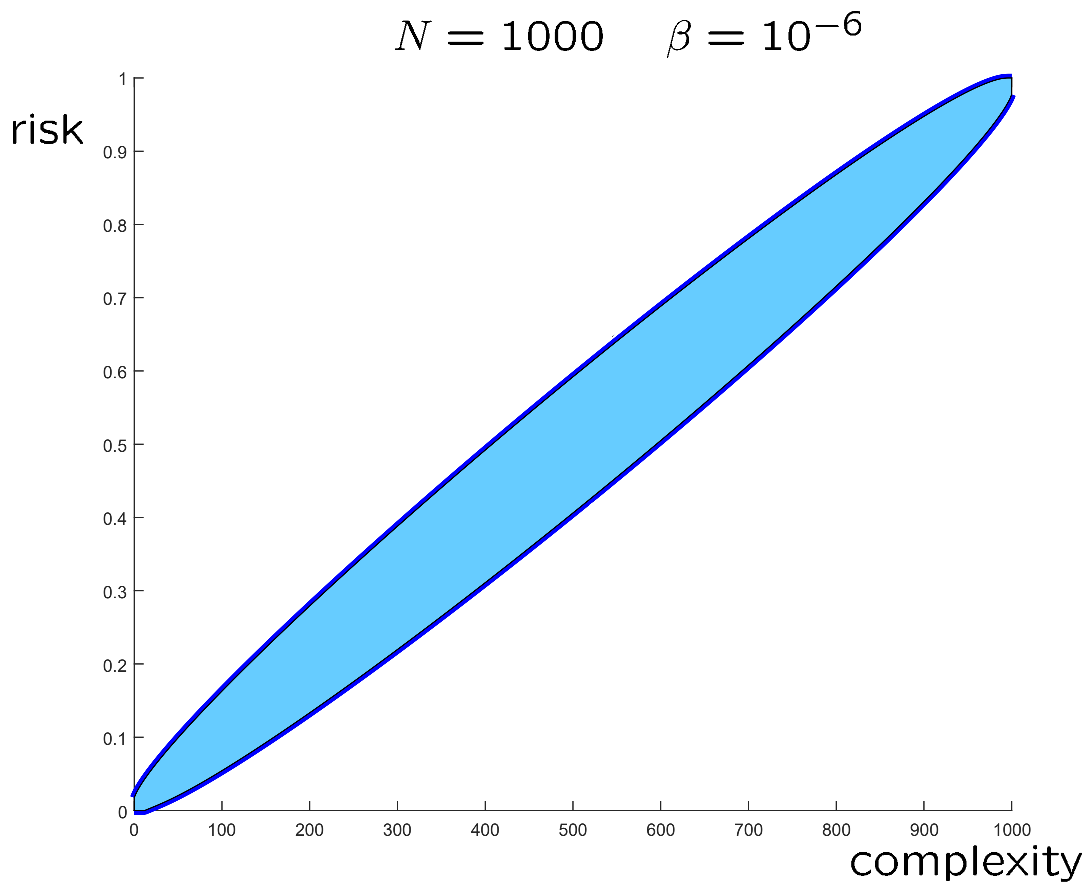
$Pr\{\underline{\epsilon}(\text{complexity}) \leq \text{risk} \leq \overline{\epsilon}(\text{complexity})\}$
$\geq 1 - \beta$

where, with the position $t = 1-\epsilon$, $\underline{\epsilon}$ and $\overline{\epsilon}$ are given by:

$$\binom{N}{k} t^{N-k} - \frac{\beta}{2N} \sum_{i=k}^{N-1} \binom{i}{k} t^{i-k} - \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{k} t^{i-k} = 0.$$

[S. Garatti, M.C. Campi, Mathematical Programming, 2019]

wait-&-judge scenario optimization



how general is all this?

consistent decision-making

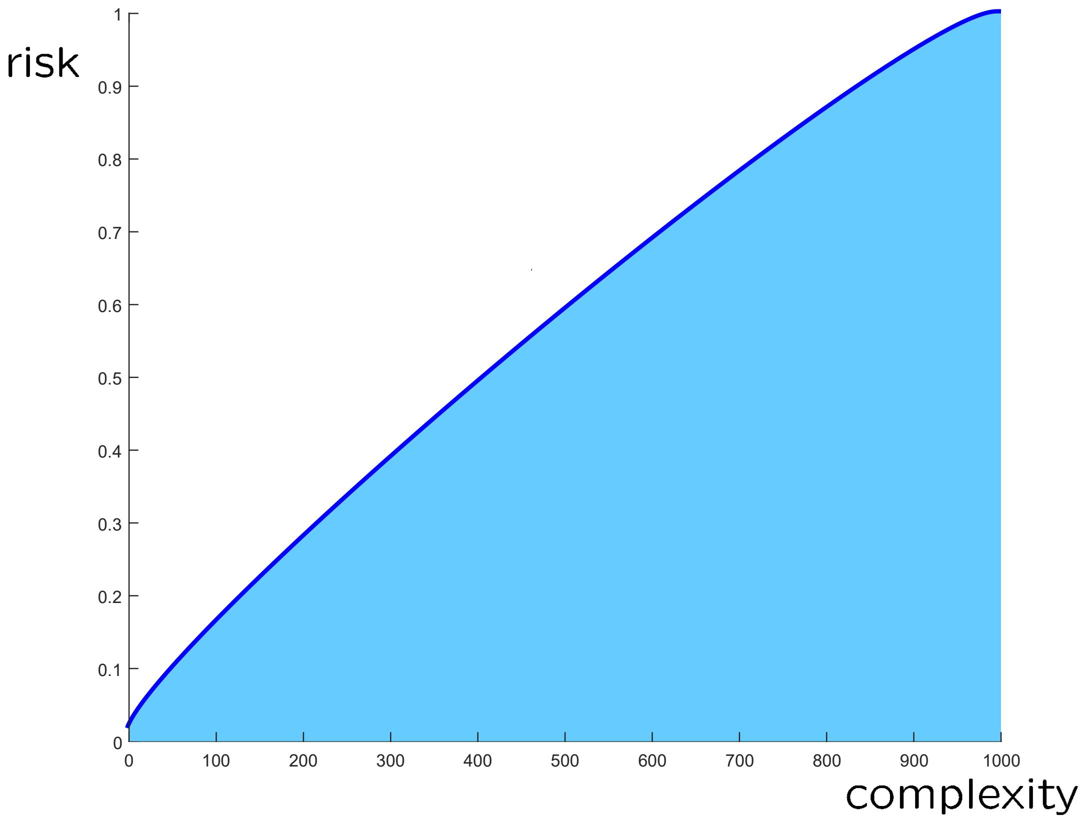
- δ = observation
- $\mathcal{D}$ = set of decisions
- each δ has associated a subset $\mathcal{D}_\delta \subseteq \mathcal{D}$, the set of decisions that are *appropriate* for δ

(a) **permutation invariance**: $P(\delta_1, \delta_2, \ldots, \delta_n) = P(\delta_{i_1}, \delta_{i_2}, \ldots, \delta_{i_n})$;

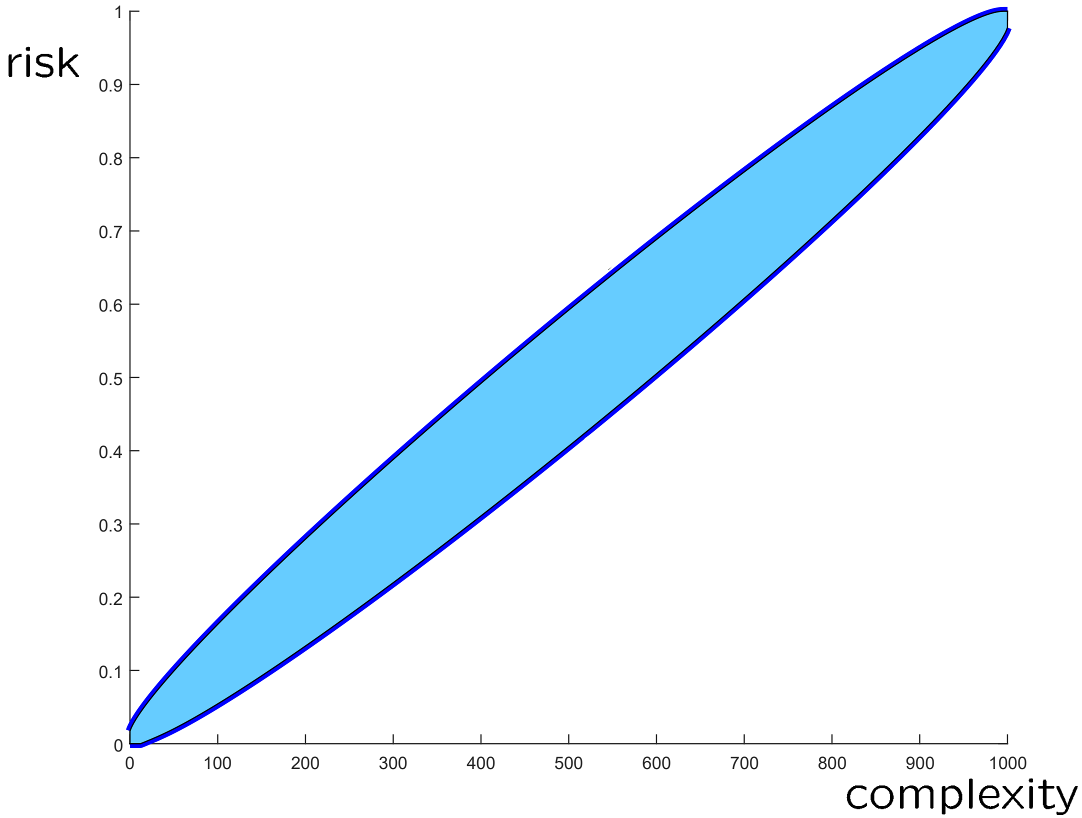
(b) **stability in the case of confirmation**: if $P(\delta_1, \delta_2, \ldots, \delta_n)$ is appropriate for m new observations $\delta_{n+1}, \ldots, \delta_{n+m}$, then
$P(\delta_1, \delta_2, \ldots, \delta_n, \delta_{n+1}, \ldots, \delta_{n+m}) = P(\delta_1, \delta_2, \ldots, \delta_n)$;

(c) **responsiveness to contradiction**: if there is at least one observation δ_{n+i} for which $P(\delta_1, \delta_2, \ldots, \delta_n)$ is not appropriate,
then $P(\delta_1, \delta_2, \ldots, \delta_n, \delta_{n+1}, \ldots, \delta_{n+m}) \neq P(\delta_1, \delta_2, \ldots, \delta_n)$.

wait-&-judge scenario optimization



wait-&-judge scenario optimization

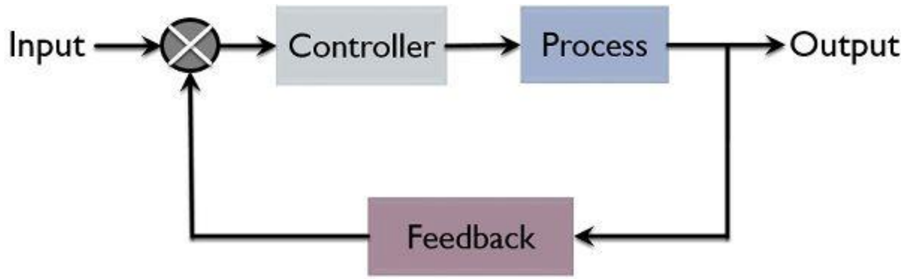


robust optimization

$$\min_{x \in \mathcal{X}} \quad c(x)$$

$$\text{subject to:} \quad x \in \mathcal{X}_{\delta_i}, \quad i = 1, \ldots, N$$

for example: Robust Control



relaxed schemes

$$\min_{x \in \mathcal{X}} \quad c(x)$$

$$\text{subject to:} \quad x \in \mathcal{X}_{\delta_i}\,, \qquad i = 1, \ldots, N$$

relaxed schemes

$$\min_{x \in \mathcal{X}} \quad c(x)$$

$$\text{subject to:} \quad f(x, \delta_i) \leq 0, \quad i = 1, \ldots, N$$

relaxed schemes

$$\min_{x \in \mathcal{X}, \xi_i \geq 0} \quad c(x) + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{subject to:} \quad f(x, \delta_i) \leq \xi_i, \quad i = 1, \ldots, N$$

relaxed schemes

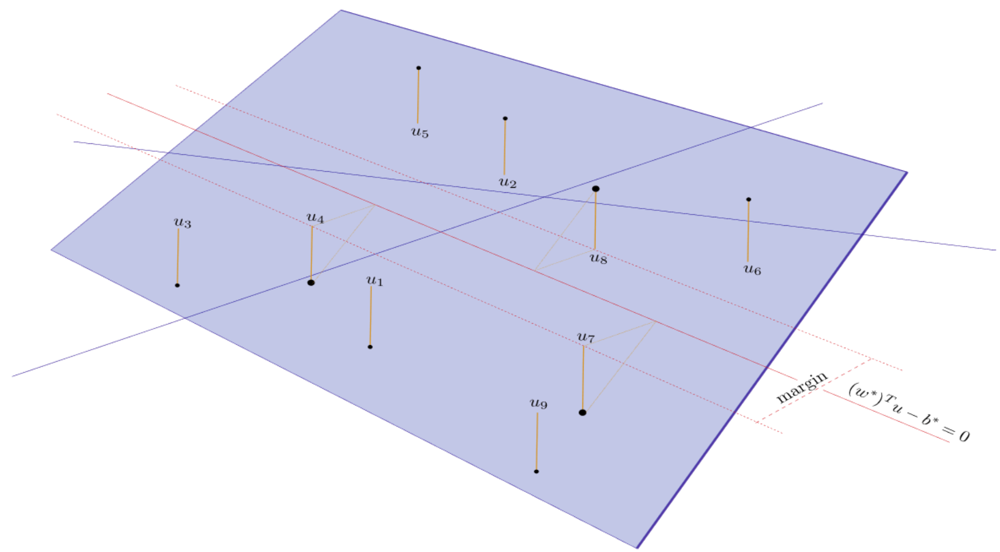
$$\min_{x \in \mathcal{X}, \xi_i \geq 0} \quad c(x) + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{subject to:} \quad f(x, \delta_i) \leq \xi_i, \quad i = 1, \ldots, N$$

for example, SVM:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi_i \geq 0} \|w\|^2 + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{subject to: } 1 - y_i(\langle w, u_i \rangle - b) \leq \xi_i$$



more …

- constraint discarding

- CVaR = Conditional Value at Risk

…

beyond Popper

- complexity is an impartial judge of reliability

and

- inductive procedures with continuous updates to observations are valid if complexity acts as a referee to assess model risk.(*)

(*) In "Conjectures and refutations: the growth of scientific knowledge", Karl Popper explicitly condemns the practice of adapting theories to observations. Speaking of the Marxist theory of history, he writes that its followers ``re-interpreted both the theory and the evidence in order to make them agree. [...] They thus gave a `conventionalist twist' to the theory; and by this stratagem they destroyed its much advertised claim of scientific status''

a beacon called "Marco"

Happy birthday, Marco!

Please, send me comments/remarks/observations: marco.campi@unibs.it

*A **monograph** that links philosophical aspects to mathematical results can be downloaded from here:*

https://marco-campi.unibs.it/pdf-pszip/inductive%20methods.pdf

*The monograph is not yet published. **I would greatly appreciate receiving your comments, even on partial aspects of the work**. Thank you!*