Learning-based decision-making via sample compression: theoretical results and algorithms

speaker: **Simone Garatti**

(Politecnico di Milano, Italy – email: simone.garatti@polimi.it)

Many thanks to all collaborators!

Marco C. Campi

Many thanks to all collaborators!

Marco C. Campi

Algo Carè

Federico Ramponi

Kostas Margellos

Alessandro Falsone

Maria Prandini

Dario Paccagnan

Many thanks to all collaborators!

Marco C. Campi

Algo Carè

Federico Ramponi

Kostas Margellos

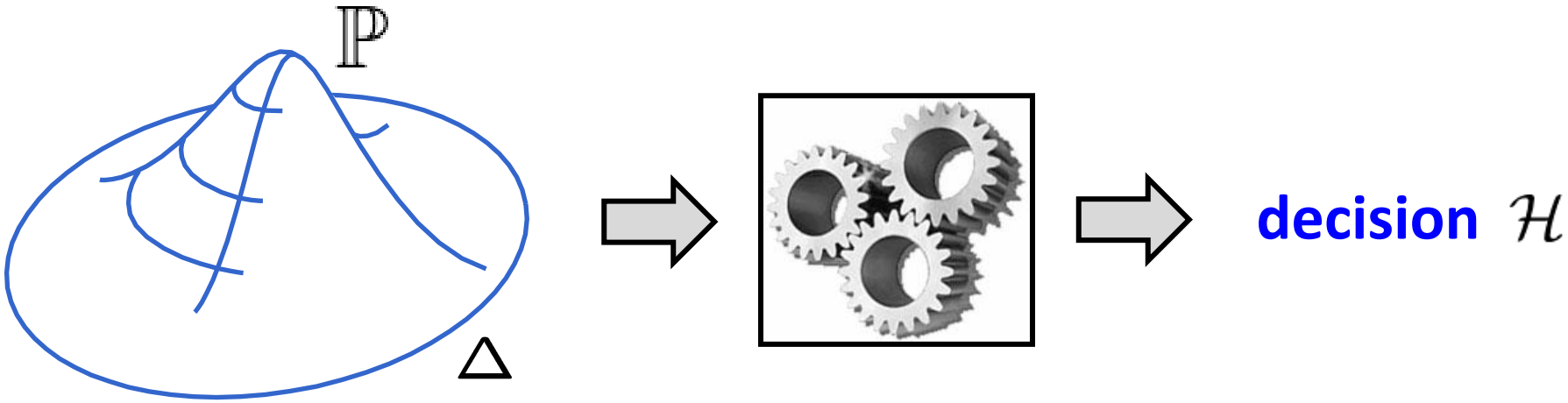
Alessandro Falsone

Maria Prandini

Dario Paccagnan

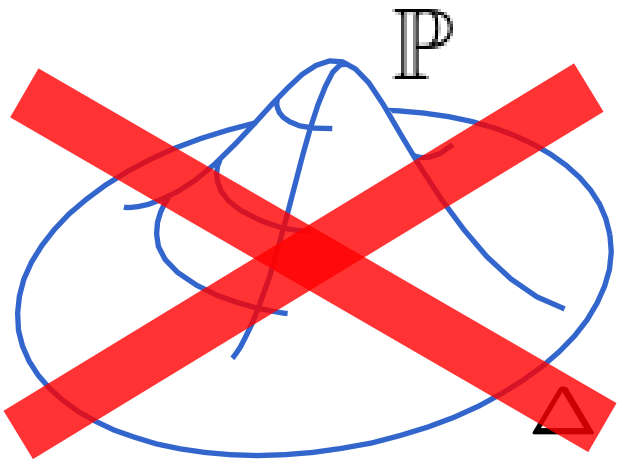
Learning-based decision-making

δ = uncertain element ⇨ exercise caution



Learning-based decision-making

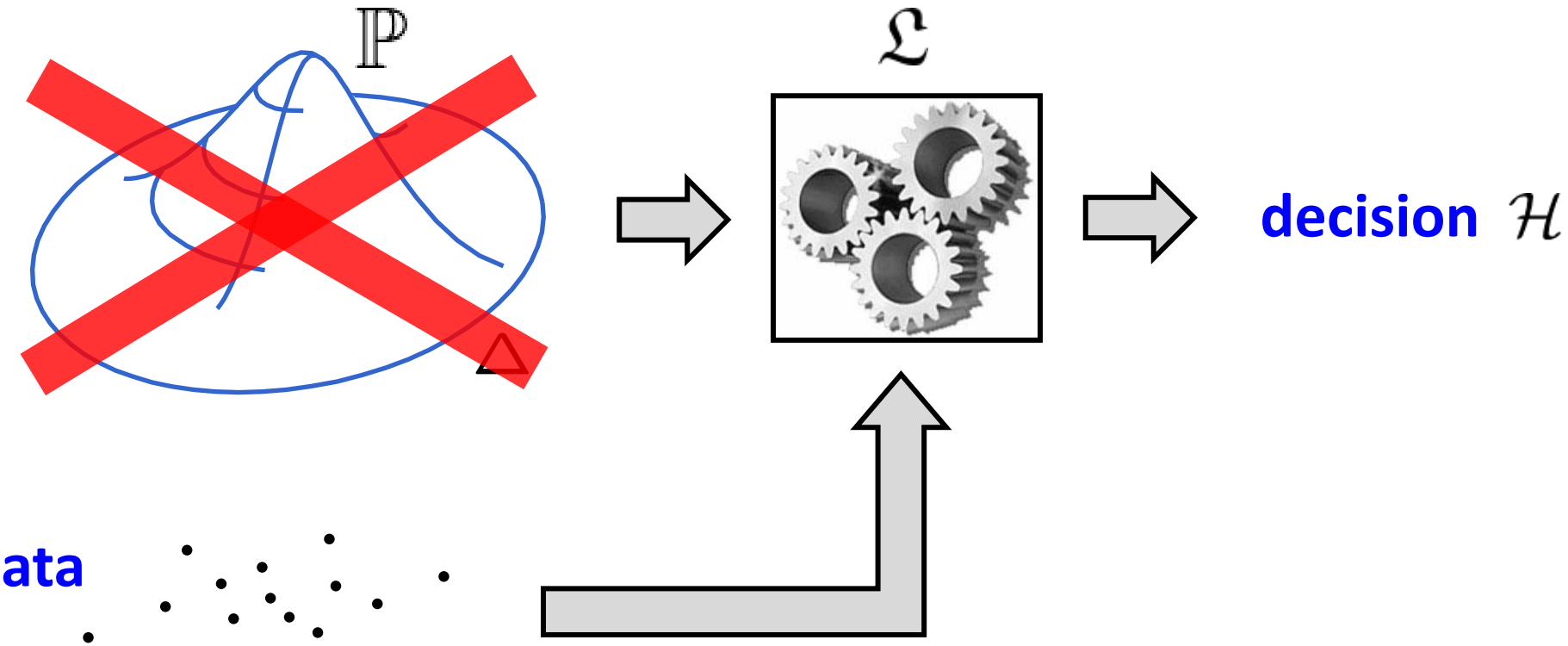
δ = uncertain element ⇨ exercise caution



⇨  ⇨ **decision** $\mathcal{H}$

Learning-based decision-making

δ = uncertain element ⇨ exercise caution



$\delta_1, \delta_2, \ldots, \delta_N$ (scenarios, i.i.d. sample)

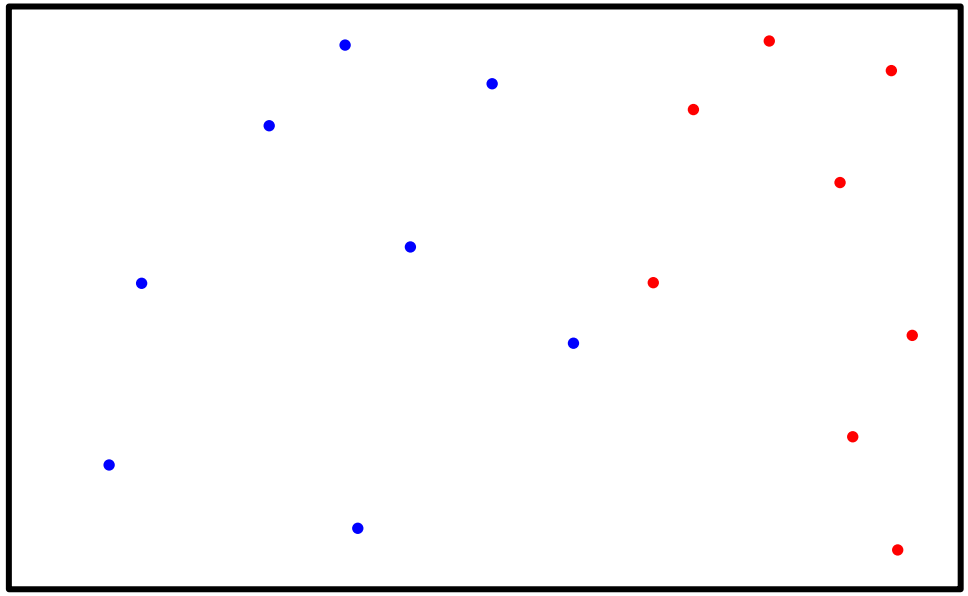
Example: classification

data

$\delta_i = (x_i, y_i)$

$x_i \in \mathbb{R}^d$

$y_i \in \{\text{red}, \text{blue}\}$



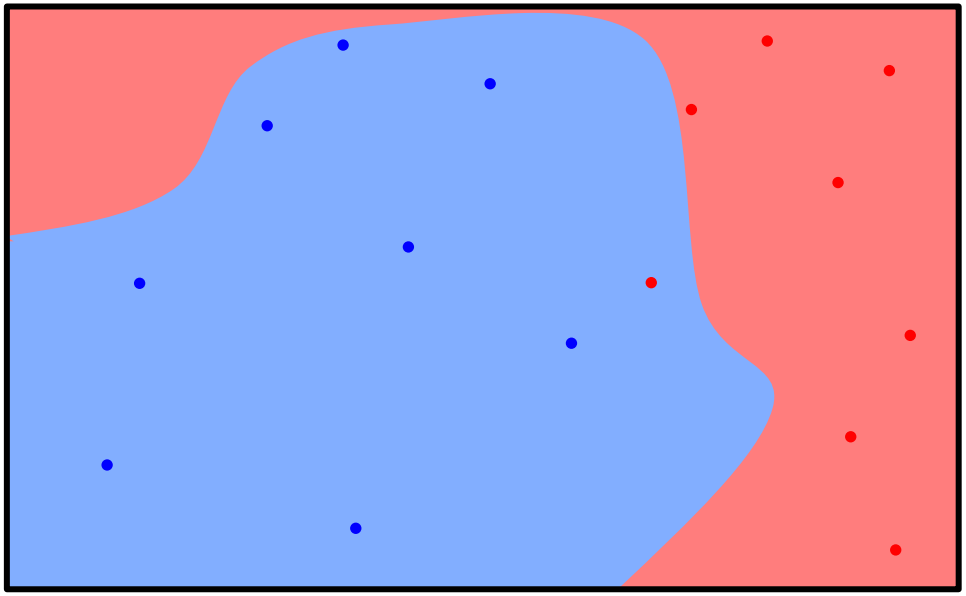
Example: classification

data

$\delta_i = (x_i, y_i)$

$x_i \in \mathbb{R}^d$

$y_i \in \{\text{red}, \text{blue}\}$



$\mathcal{H}$ = SVM classifier

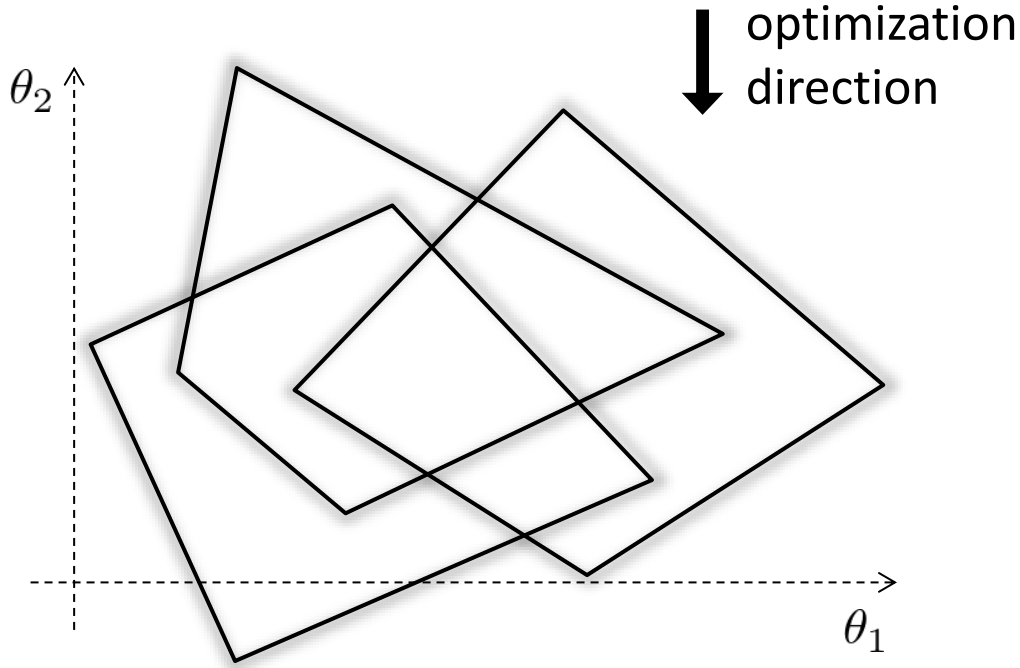
$$\min_{\substack{w \in \Phi, b \in \mathbb{R} \\ \xi_i \geq 0, i=1,\dots,N}} \quad \|w\|^2 + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad 1 - y_i(\langle w, \phi(x_i) \rangle - b) \leq \xi_i, \quad i = 1, \dots, N$$

Example: scenario robust optimization

data constraints

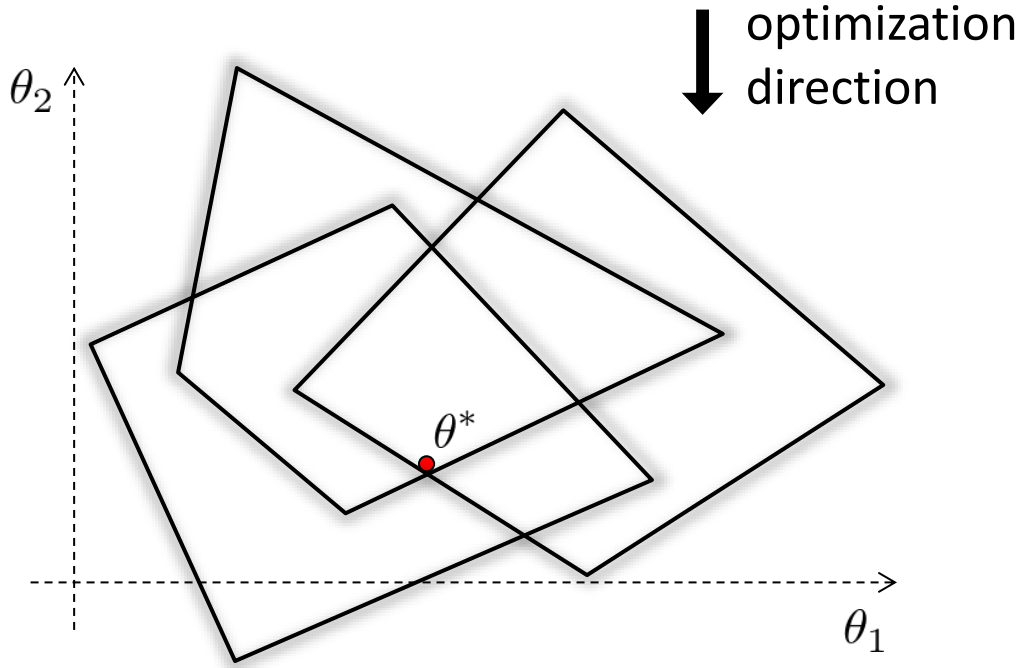
$$\delta_1 \rightarrow f(\theta, \delta_1) \leq 0$$
$$\delta_2 \rightarrow f(\theta, \delta_2) \leq 0$$
$$\vdots$$
$$\delta_N \rightarrow f(\theta, \delta_N) \leq 0$$



Example: scenario robust optimization

data → constraints

$$\delta_1 \rightarrow f(\theta, \delta_1) \leq 0$$
$$\delta_2 \rightarrow f(\theta, \delta_2) \leq 0$$
$$\vdots$$
$$\delta_N \rightarrow f(\theta, \delta_N) \leq 0$$



$$\mathcal{H} = \theta^* = \text{solution to} \begin{cases} \min\limits_{\theta \in \Theta} & c(\theta) \\ \text{s.t.} & f(\theta, \delta_i) \leq 0 \\ & i = 1, \ldots, N \end{cases}$$

Example: scenario robust optimization (cont'd)

rate of return of k financial assets in day i

$$\delta_i = (R_i^1, R_i^2, \ldots, R_i^k)$$

money split over the k financial assets

$$\theta = (\$^1, \$^2, \ldots, \$^k, L)$$

$$\min_{\substack{\$^i, L \\ \$^1 + \cdots + \$^k = 1}} \quad L$$

loss in day i

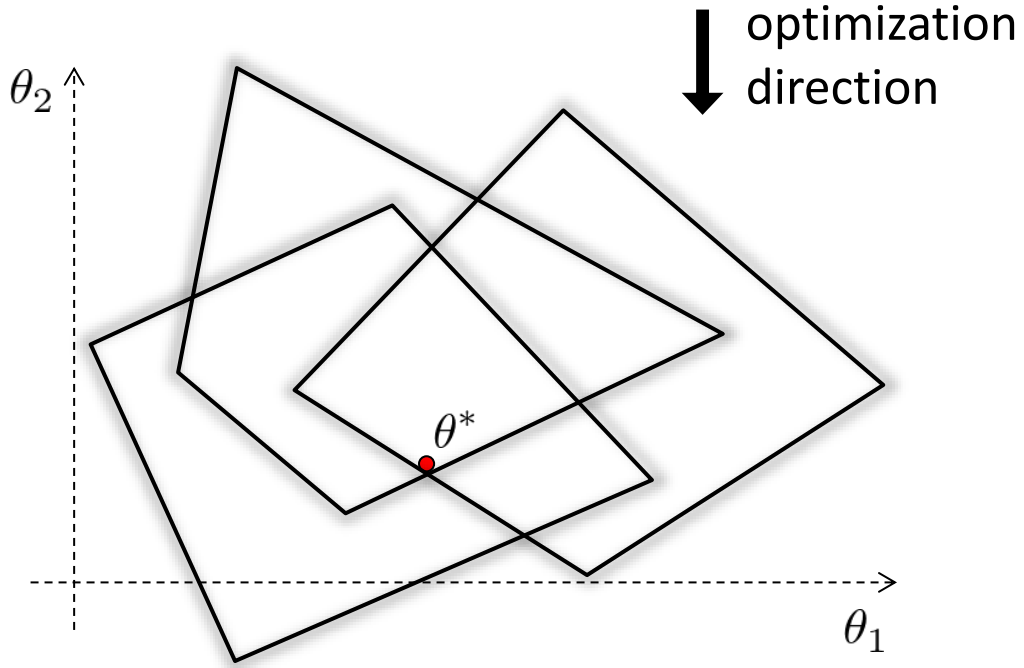
$$\text{s.t.} \; \underbrace{-(\$^1 R_i^1 + \$^2 R_i^2 + \cdots + \$^k R_i^k)}_{\text{loss in day } i} \leq L$$

$$i = 1, \ldots, N$$

Example: scenario optimization with relaxation

data → constraints

$$\delta_1 \rightarrow f(\theta, \delta_1) \leq 0$$
$$\delta_2 \rightarrow f(\theta, \delta_2) \leq 0$$
$$\vdots$$
$$\delta_N \rightarrow f(\theta, \delta_N) \leq 0$$

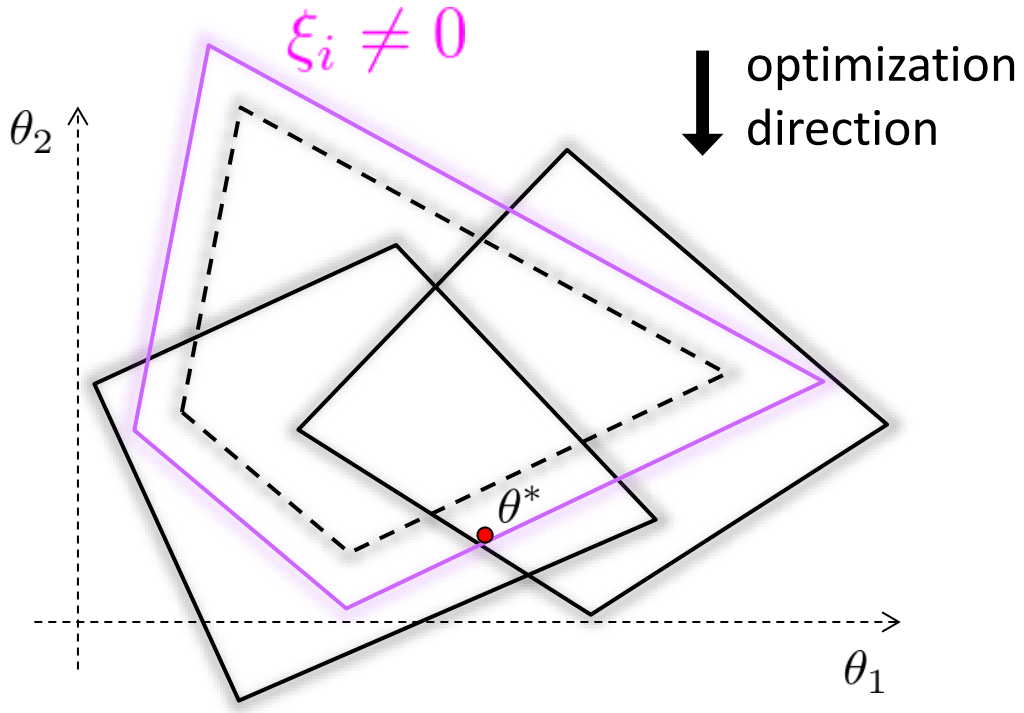


$$\mathcal{H} = \theta^* = \text{solution to} \begin{cases} \min\limits_{\theta \in \Theta} & c(\theta) \\ \text{s.t.} & f(\theta, \delta_i) \leq 0 \\ & i = 1, \ldots, N \end{cases}$$

Example: scenario optimization with relaxation

data → constraints

$$\delta_1 \rightarrow f(\theta, \delta_1) \leq \xi_1$$
$$\delta_2 \rightarrow f(\theta, \delta_2) \leq \xi_2$$
$$\vdots$$
$$\delta_N \rightarrow f(\theta, \delta_N) \leq \xi_N$$



$\mathcal{H}$ = θ^* = solution to

$$\begin{cases} \min\limits_{\theta \in \Theta} & c(\theta) + \rho \sum\limits_{i=1}^{N} \xi_i \\ \text{s.t.} & f(\theta, \delta_i) \leq \xi_i \\ & i = 1, \ldots, N \end{cases}$$

A lesson from supervised learning

Which is the learning-based scheme for the problem at hand?

Difficult to say a-priori without incurring in over-conservatism … a blend of approximate knowledge and heuristics, often in various attempts (hyperparameters tuning)

No limits in exploration, but some guidance is needed…

⇨ accurate and rigorous certification of the actual performance of the explored decisions

… is it possible?

dependable utilization of it

final decision selection

The risk of a decision

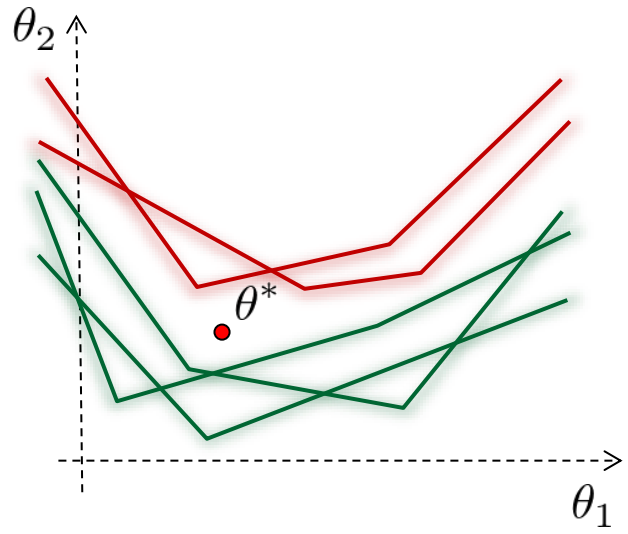
$$\mathrm{R}(\mathcal{H}) = \mathbb{P}_{\delta}\{\, \mathcal{H} \text{ is } \textbf{inappropriate} \text{ for a new } \delta \,\}$$

↓

interaction between decision and environment

For instance, in scenario optimization:

$$\mathbb{P}_{\delta}\{\, \text{a new constraint is violated by } \theta^* \,\}$$



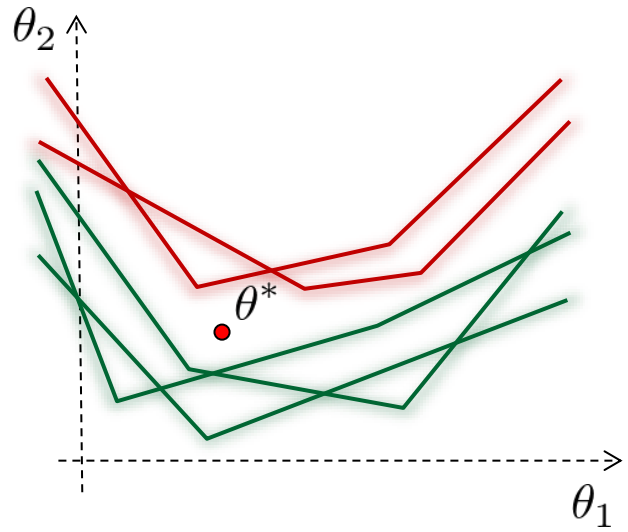
The risk of a decision

$$\mathrm{R}(\mathcal{H}) = \mathbb{P}_{\delta}\{\,\mathcal{H} \text{ is } \textbf{inappropriate} \text{ for a new } \delta\,\}$$

interaction between decision and environment

For instance, in scenario optimization:

$$\mathbb{P}_{\delta}\{\,\text{a new constraint is violated by } \theta^*\,\}$$



Issue: $\mathbb{P}$ is not available…

⇨ is it possible to assess $\mathrm{R}(\mathcal{H})$ from data ?

Why not relying on a test dataset…

- holding out some data for testing rather than designing… waste of information, questionable!
- scenarios (data) are often limited resources (collecting data can be time-consuming or burdensome, involving a monetary cost)
- in the context of this talk, testing is not necessary…

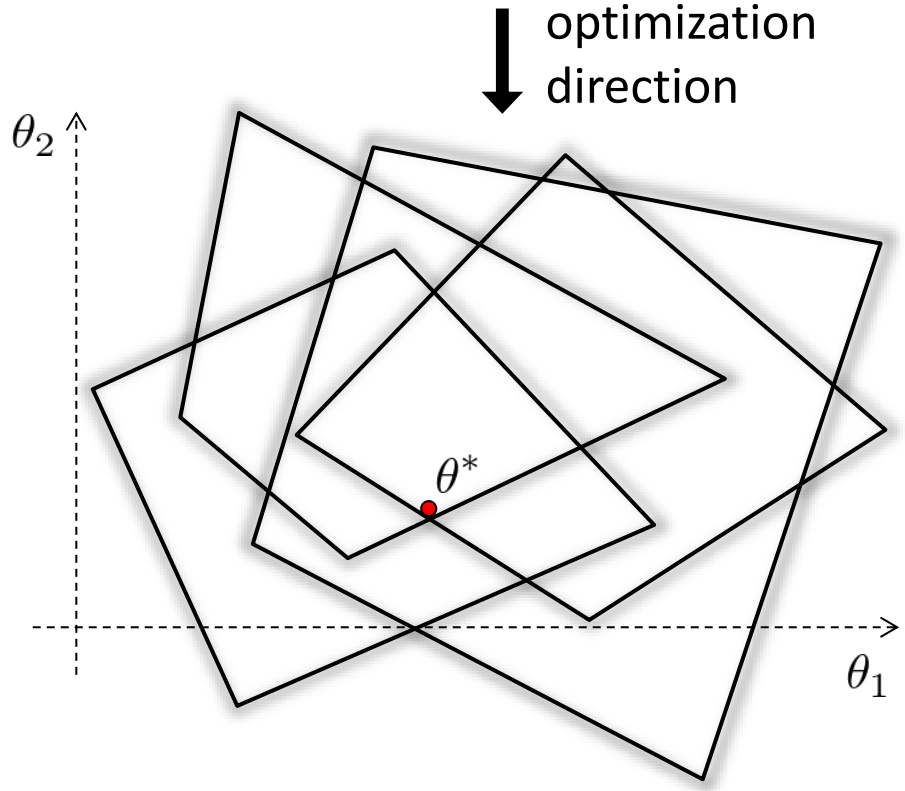
… data can well play a double role!

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Preference

$$\kappa(\delta_1, \ldots, \delta_N) \subseteq S \subseteq (\delta_1, \ldots, \delta_N)$$

$$\Downarrow$$

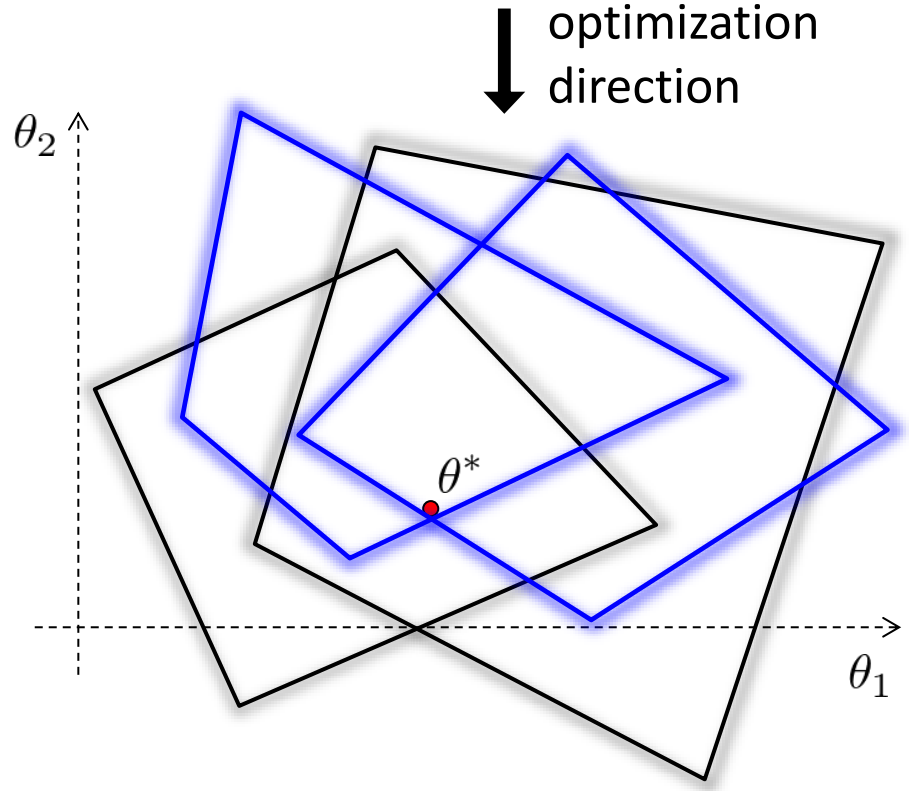
$$\kappa(S) = \kappa(\delta_1, \ldots, \delta_N)$$

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Preference

$$\kappa(\delta_1, \ldots, \delta_N) \subseteq S \subseteq (\delta_1, \ldots, \delta_N)$$

$$\Downarrow$$

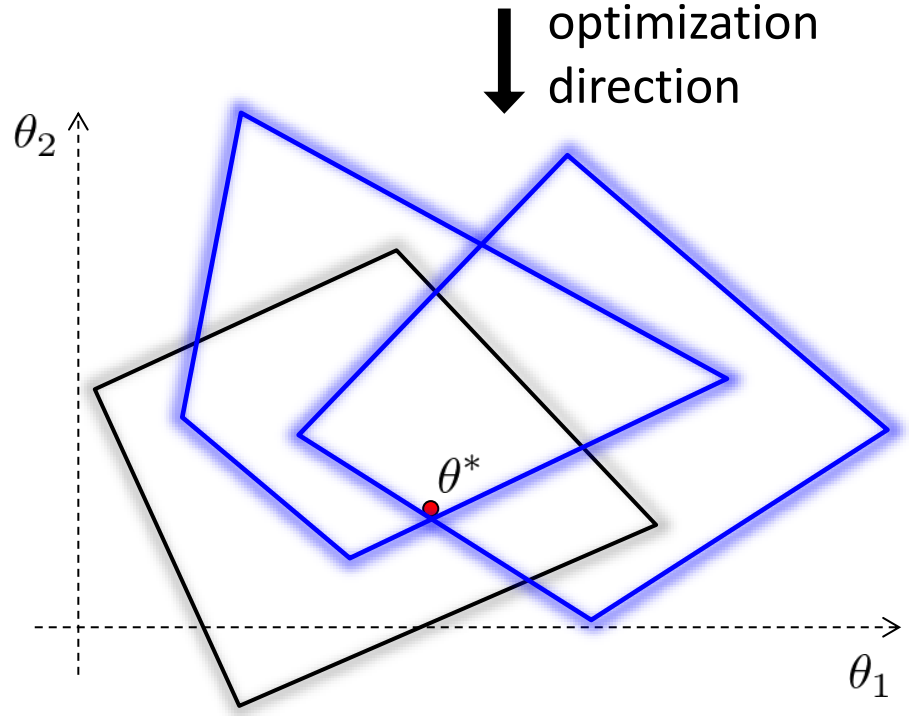
$$\kappa(S) = \kappa(\delta_1, \ldots, \delta_N)$$

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Preference

$$\kappa(\delta_1, \ldots, \delta_N) \subseteq S \subseteq (\delta_1, \ldots, \delta_N)$$

$$\Downarrow$$

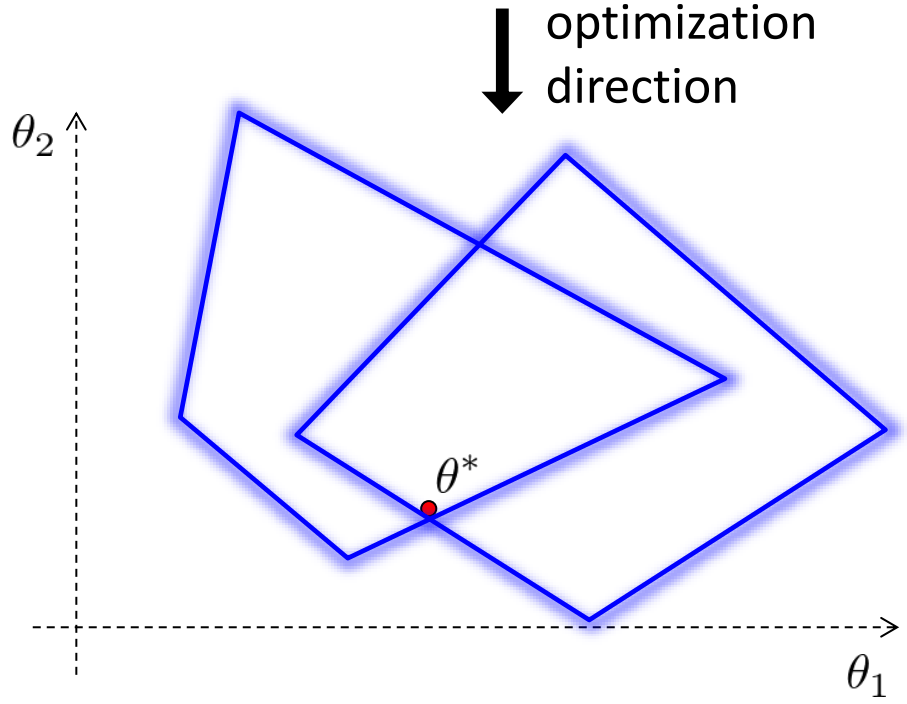
$$\kappa(S) = \kappa(\delta_1, \ldots, \delta_N)$$

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Coherence

a new scenario for which $\mathcal{H}$ is inappropriate is added

⇓

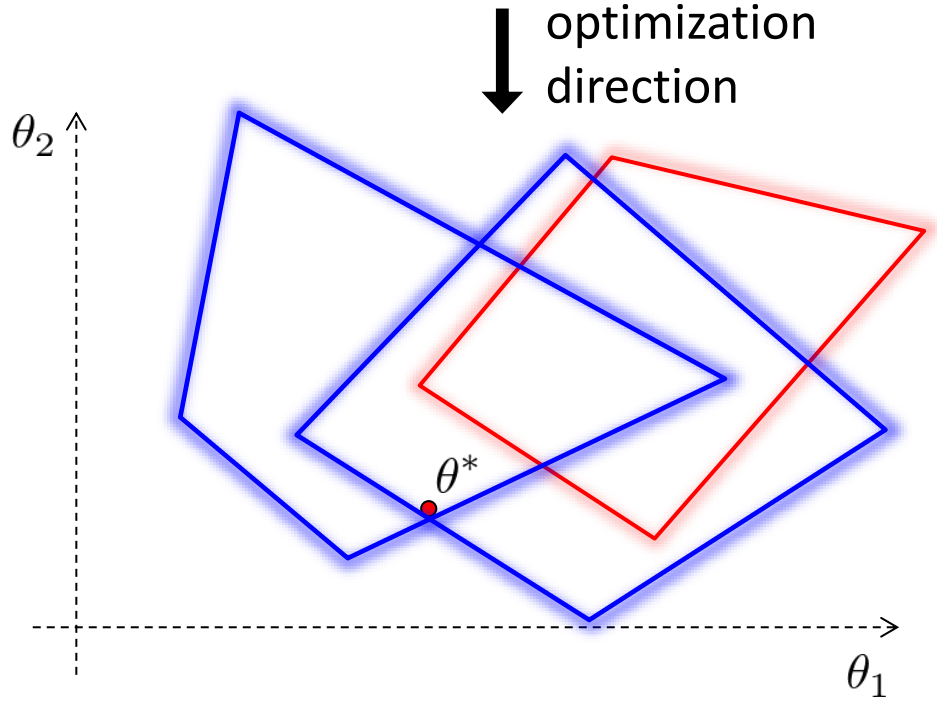
the compression must change

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Coherence

a new scenario for which $\mathcal{H}$ is inappropriate is added

⇓

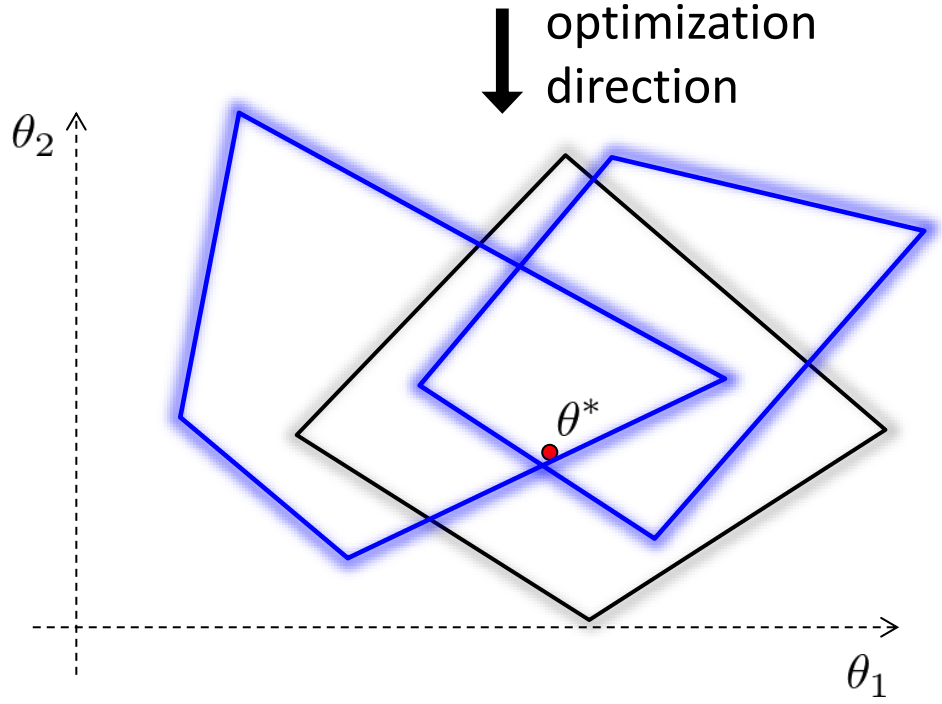
the compression must change

Risk assessment via sample compression

Compression function

$$\kappa(\delta_1, \ldots, \delta_N) = \delta_{i_1}, \ldots, \delta_{i_k}$$

map extracting a data subsample from the dataset



Coherence

a new scenario for which $\mathcal{H}$ is inappropriate is added

⇓

the compression must change

The main result in a nutshell

Risk: $$\mathsf{R}(\mathcal{H}) = \mathsf{R}\big(\mathcal{H}(\delta_1, \ldots, \delta_N)\big)$$

The main result in a nutshell

Risk: $\mathsf{R}(\mathcal{H}) = \mathsf{R}\big(\mathcal{H}(\delta_1, \ldots, \delta_N)\big)$

Complexity: $\pi = \big|\kappa(\delta_1, \ldots, \delta_N)\big|$

} random variables

size of compressed set

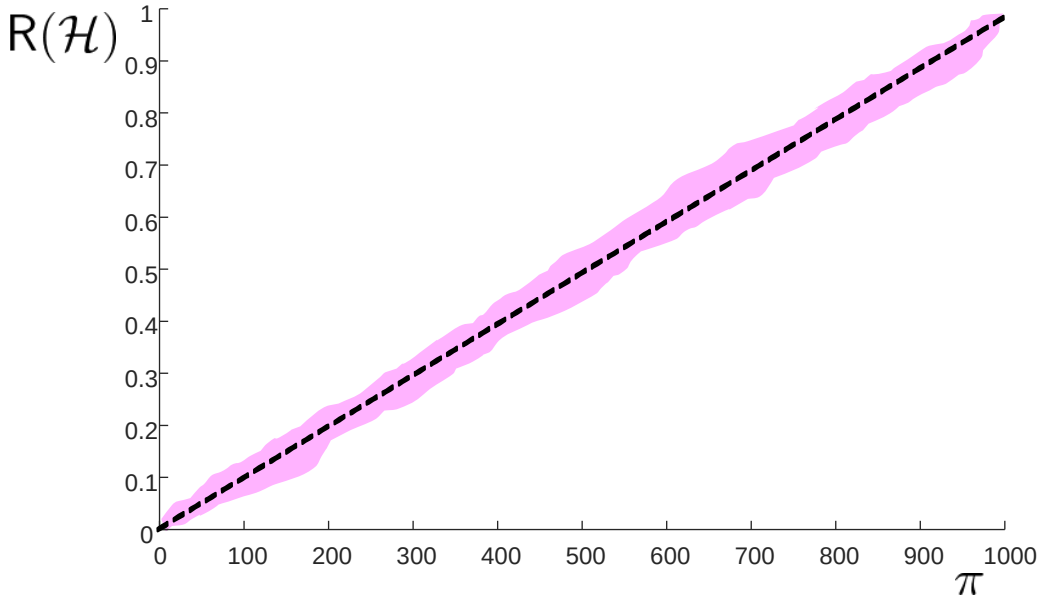
The main result in a nutshell

Risk: $\mathsf{R}(\mathcal{H}) = \mathsf{R}\big(\mathcal{H}(\delta_1, \ldots, \delta_N)\big)$

Complexity: $\pi = \big|\kappa(\delta_1, \ldots, \delta_N)\big|$

} random variables

Under preference and coherence, the joint distribution of risk and complexity is concentrated around/below $\mathsf{R}(\mathcal{H}) = \pi/N$



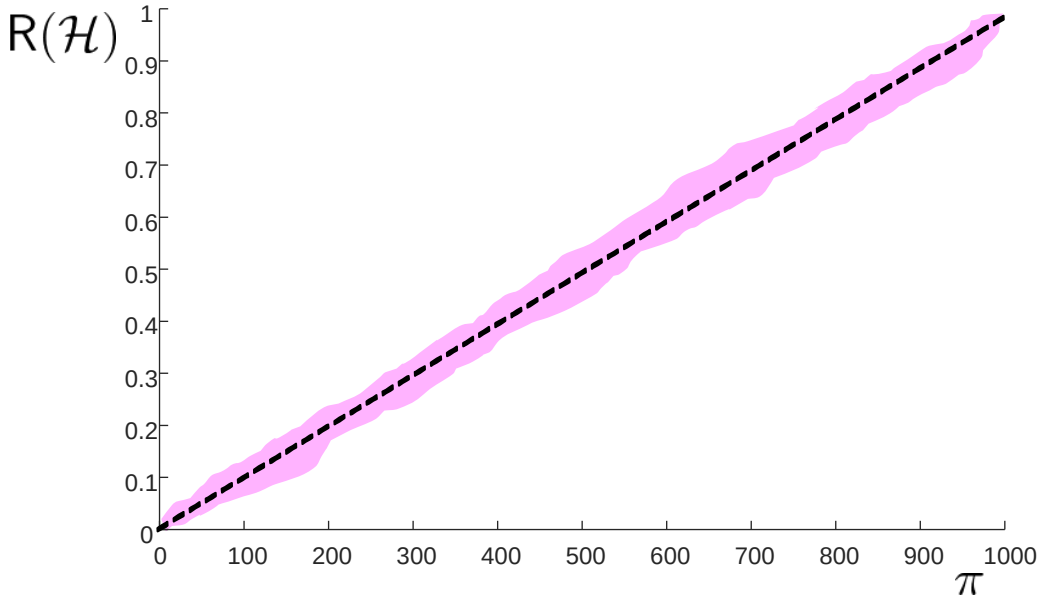
The main result in a nutshell

Risk: $\mathrm{R}(\mathcal{H}) = \mathrm{R}\big(\mathcal{H}(\delta_1, \ldots, \delta_N)\big)$

Complexity: $\pi = \big|\kappa(\delta_1, \ldots, \delta_N)\big|$

} random variables

Under preference and coherence, the joint distribution of risk and complexity is concentrated around/below $\mathrm{R}(\mathcal{H}) = \pi/N$



$\mathrm{R}(\mathcal{H})$ **can be accurately estimated from** π

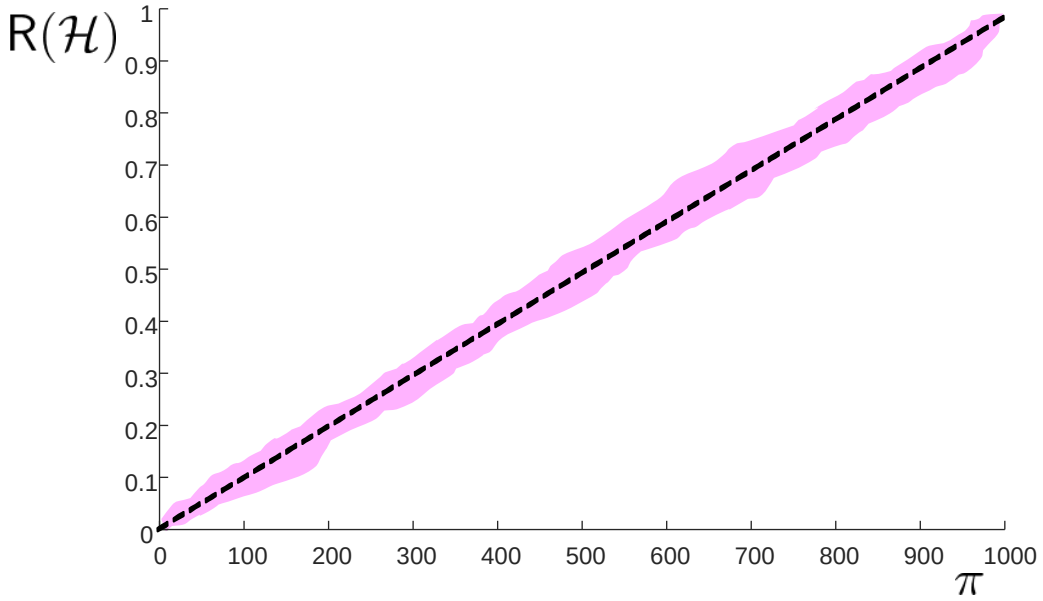
The main result in a nutshell

Risk: $\mathrm{R}(\mathcal{H}) = \mathrm{R}\big(\mathcal{H}(\delta_1, \ldots, \delta_N)\big)$

Complexity: $\pi = \big|\kappa(\delta_1, \ldots, \delta_N)\big|$

} random variables

Under preference and coherence, the joint distribution of risk and complexity is concentrated around/below $\mathrm{R}(\mathcal{H}) = \pi/N$



$\mathrm{R}(\mathcal{H})$ **can be accurately estimated from** π

observable!

Main result (formal statement)

Theorem (with M. Campi)

Assume preference and coherence

Choose $\beta \in (0,1)$ (confidence parameter)

Let $\epsilon^U(k)$ be the unique roots in (0,1) of polynomials

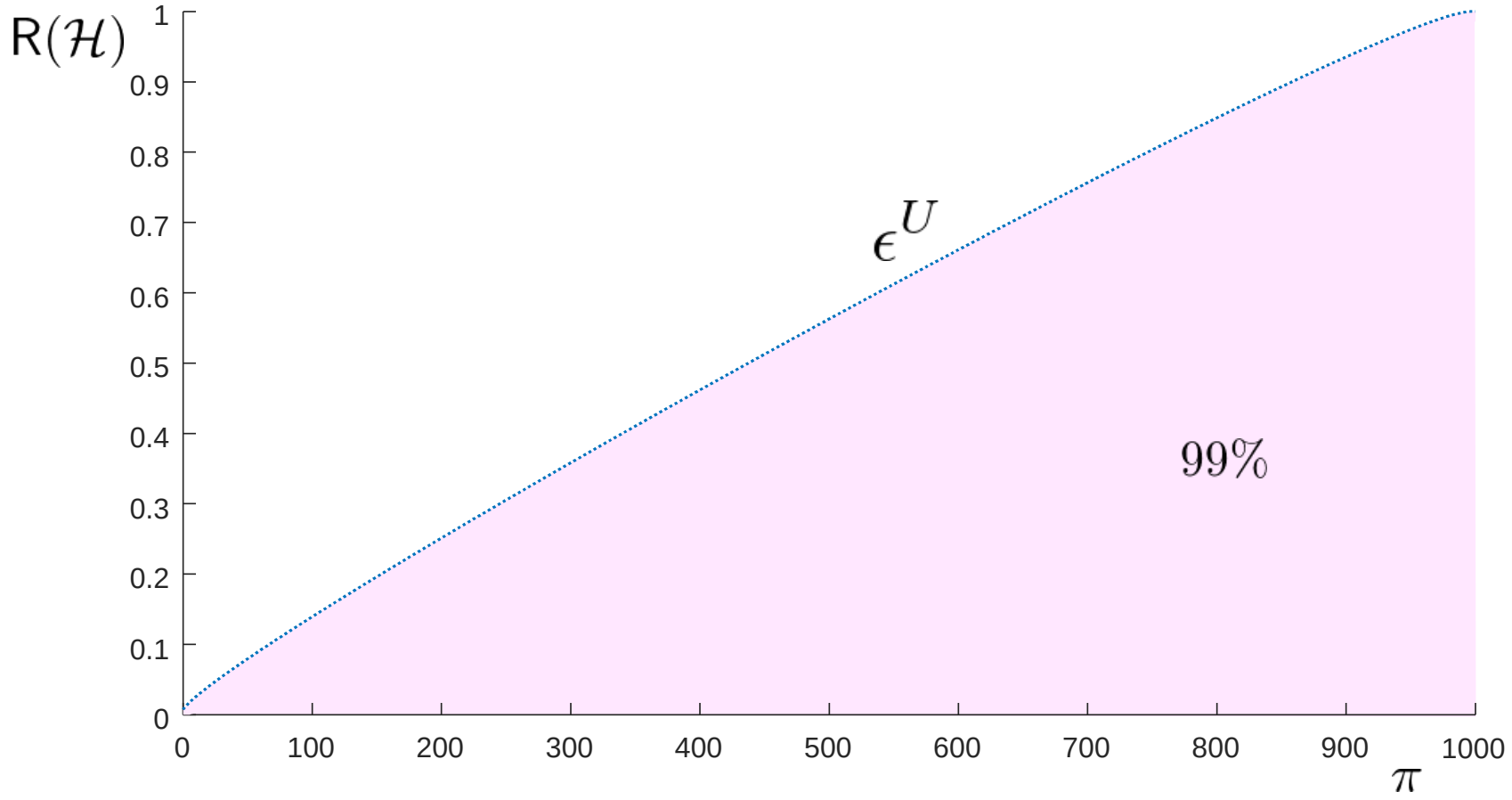
- $\binom{N}{k}(1-\epsilon)^{N-k} - \frac{\beta}{2N}\sum_{m=k}^{N-1}\binom{m}{k}(1-\epsilon)^{m-k}$

Then, irrespective of $\mathbb{P}$ (distribution-free),

$$\mathbb{P}^N\left\{\delta_1,\ldots,\delta_N:\ \mathsf{R}(\mathcal{H}) \le \epsilon^U(\pi)\right\} \ge 1-\beta$$

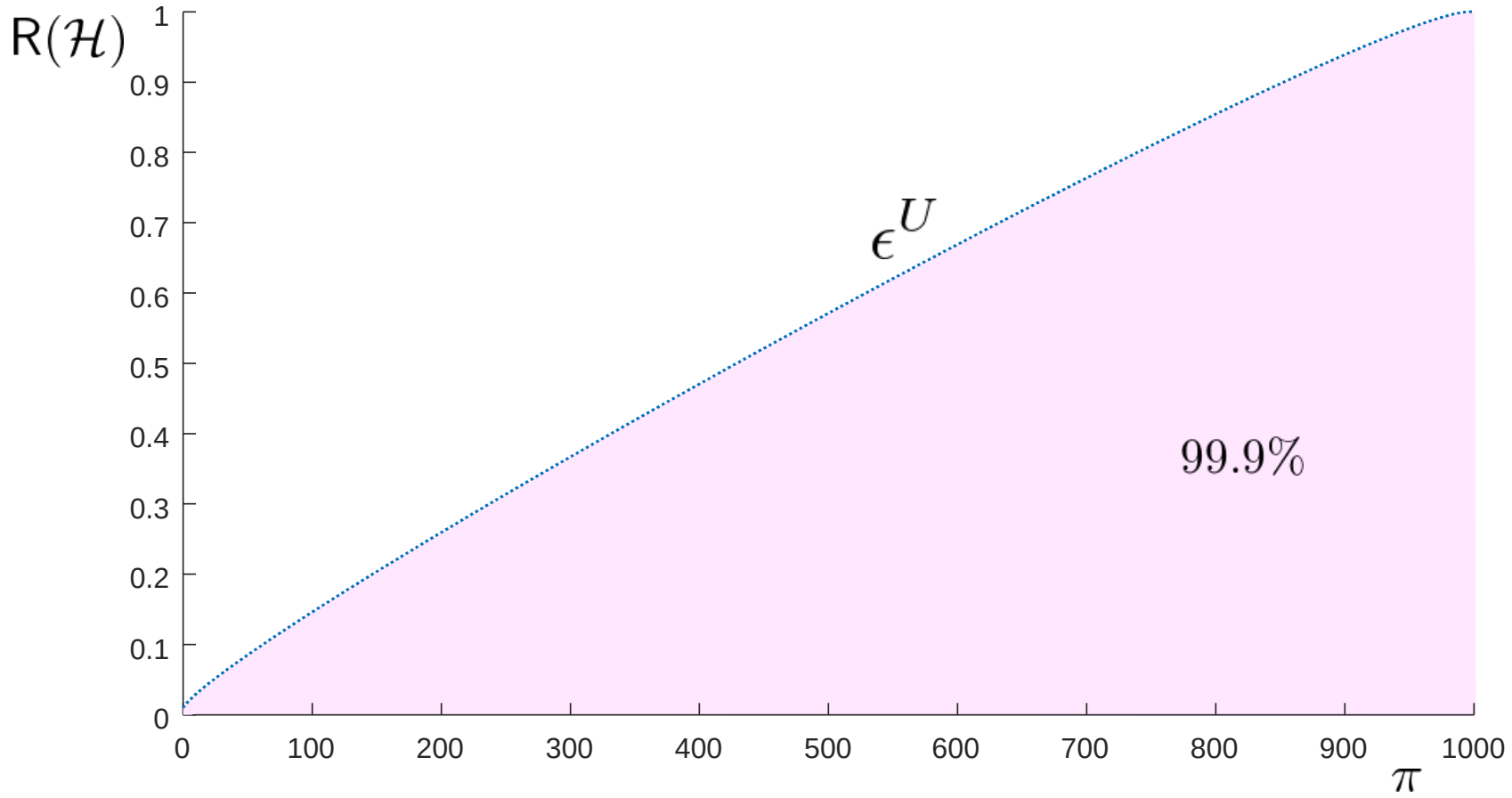
Main result (graphical illustration)

$$\mathbb{P}^N \left\{ \delta_1, \ldots, \delta_N : \ \mathsf{R}(\mathcal{H}) \leq \epsilon^U(\pi) \right\} \geq 1 - \beta \qquad N = 1000$$



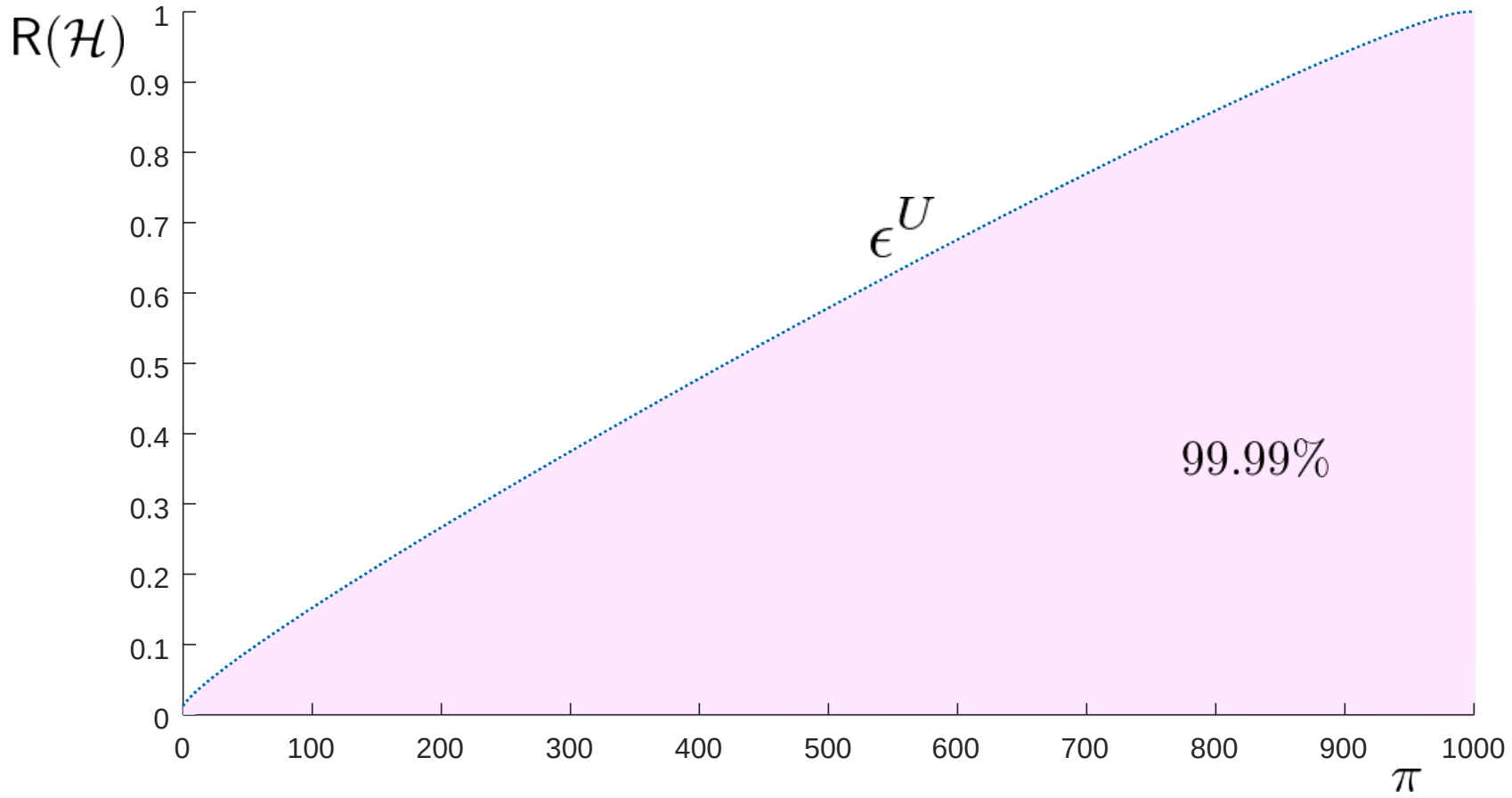
Main result (graphical illustration)

$$\mathbb{P}^N \left\{ \delta_1, \ldots, \delta_N : \ \mathsf{R}(\mathcal{H}) \leq \epsilon^U(\pi) \right\} \geq 1 - \beta \qquad N = 1000$$



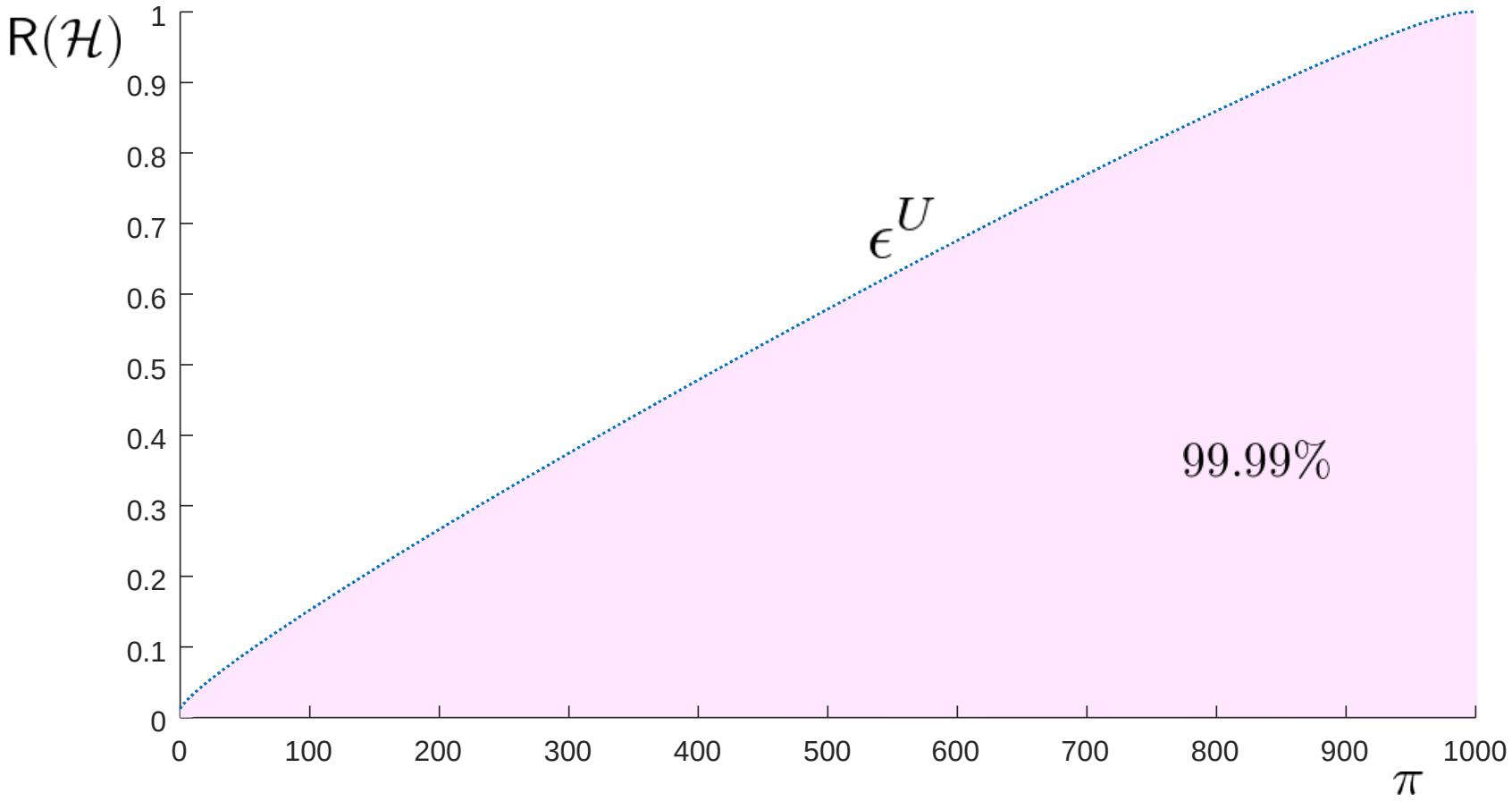
Main result (graphical illustration)

$$\mathbb{P}^N \left\{ \delta_1, \ldots, \delta_N : \ \mathsf{R}(\mathcal{H}) \leq \epsilon^U(\pi) \right\} \geq 1 - \beta \qquad N = 1000$$



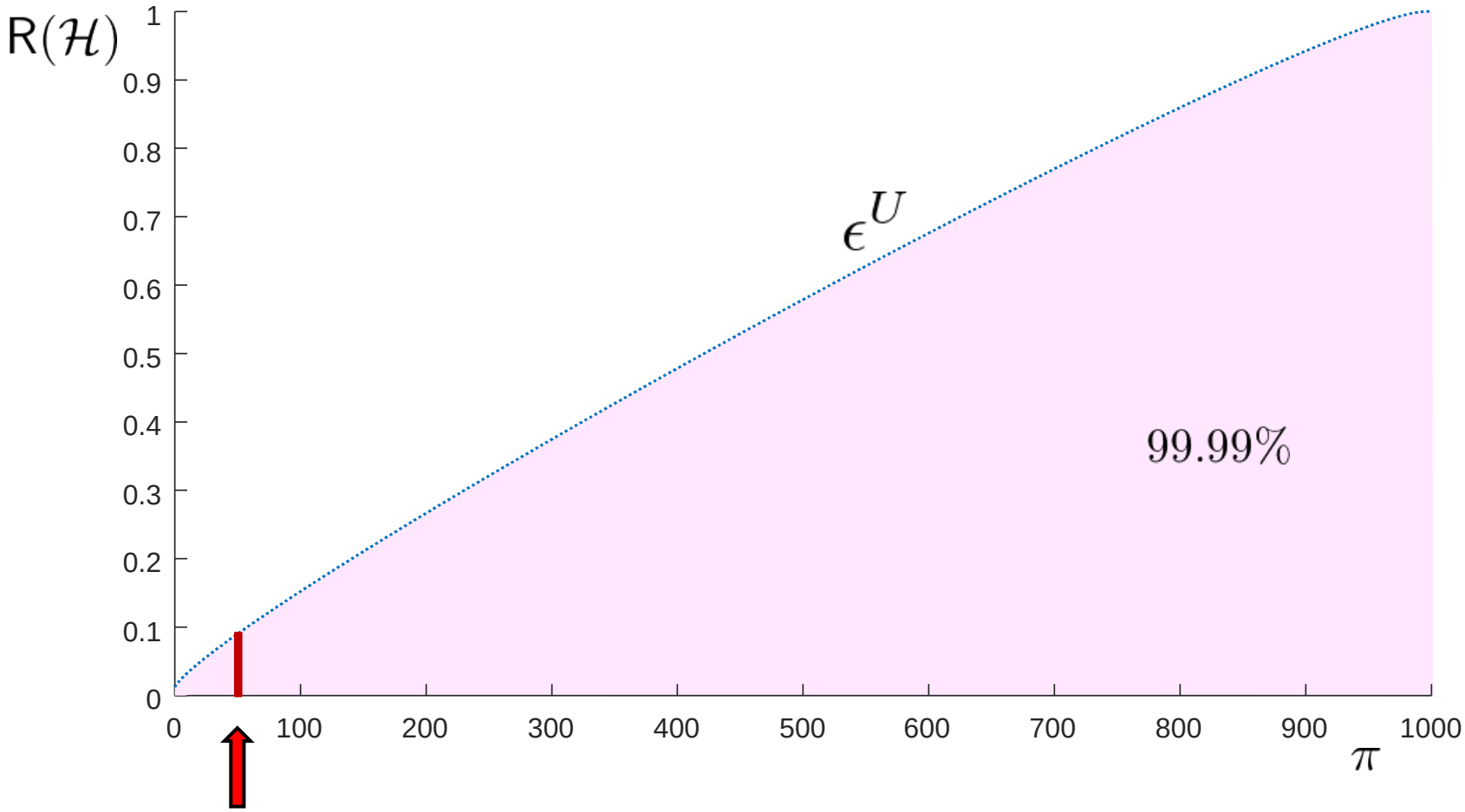
Main result (graphical illustration)

$\mathbf{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



Main result (graphical illustration)

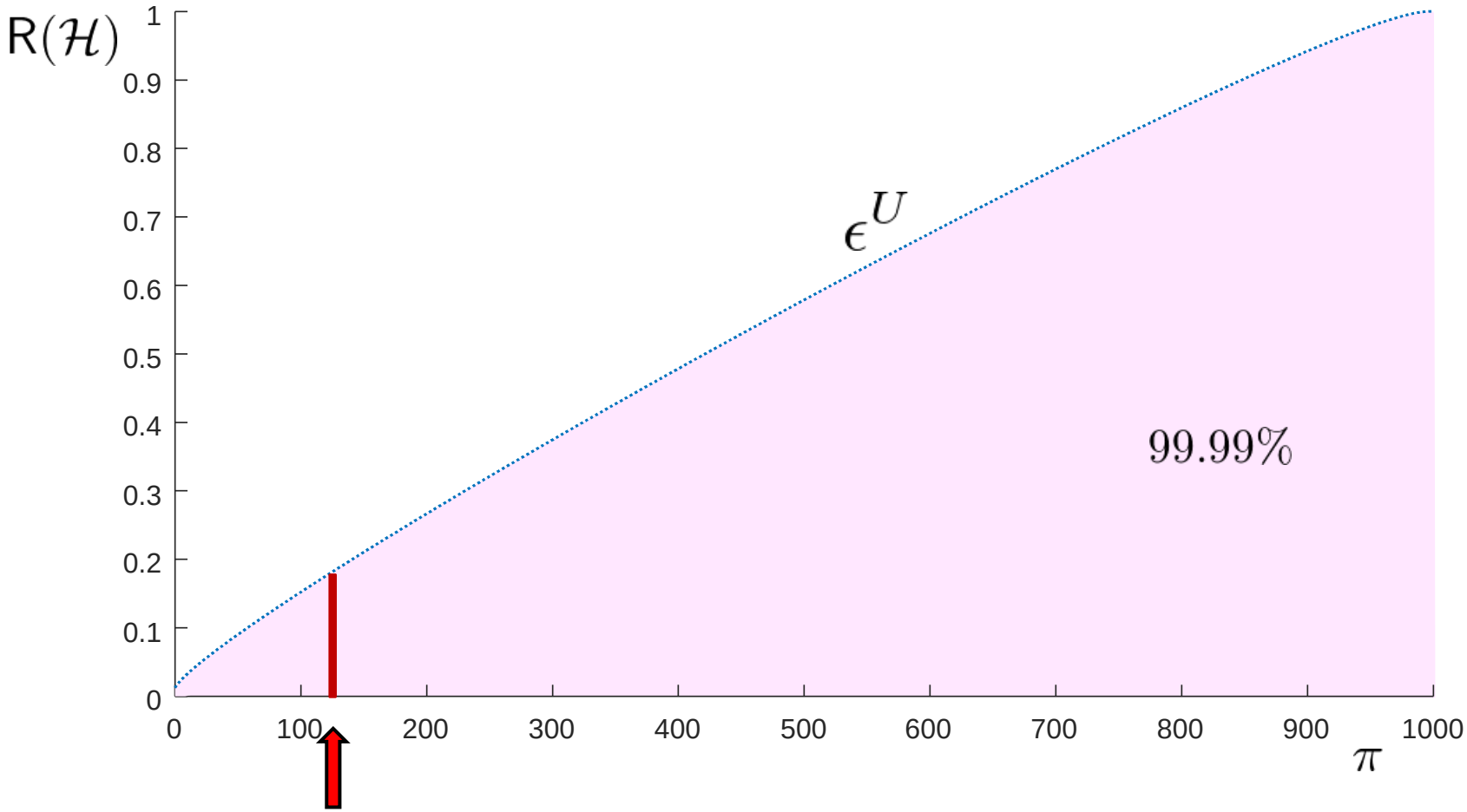
$\mathbf{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



$\pi = 50 \rightarrow \mathrm{R}(\mathcal{H}) \leq 9\%$

Main result (graphical illustration)

$\mathbf{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



Main result (cont'd)

Theorem (with M. Campi)

Assume preference and coherence + additional assumptions

Choose $\beta \in (0,1)$ (confidence parameter)

Let $\epsilon_L(k), \epsilon^U(k)$ be the unique roots in (0,1) of polynomials

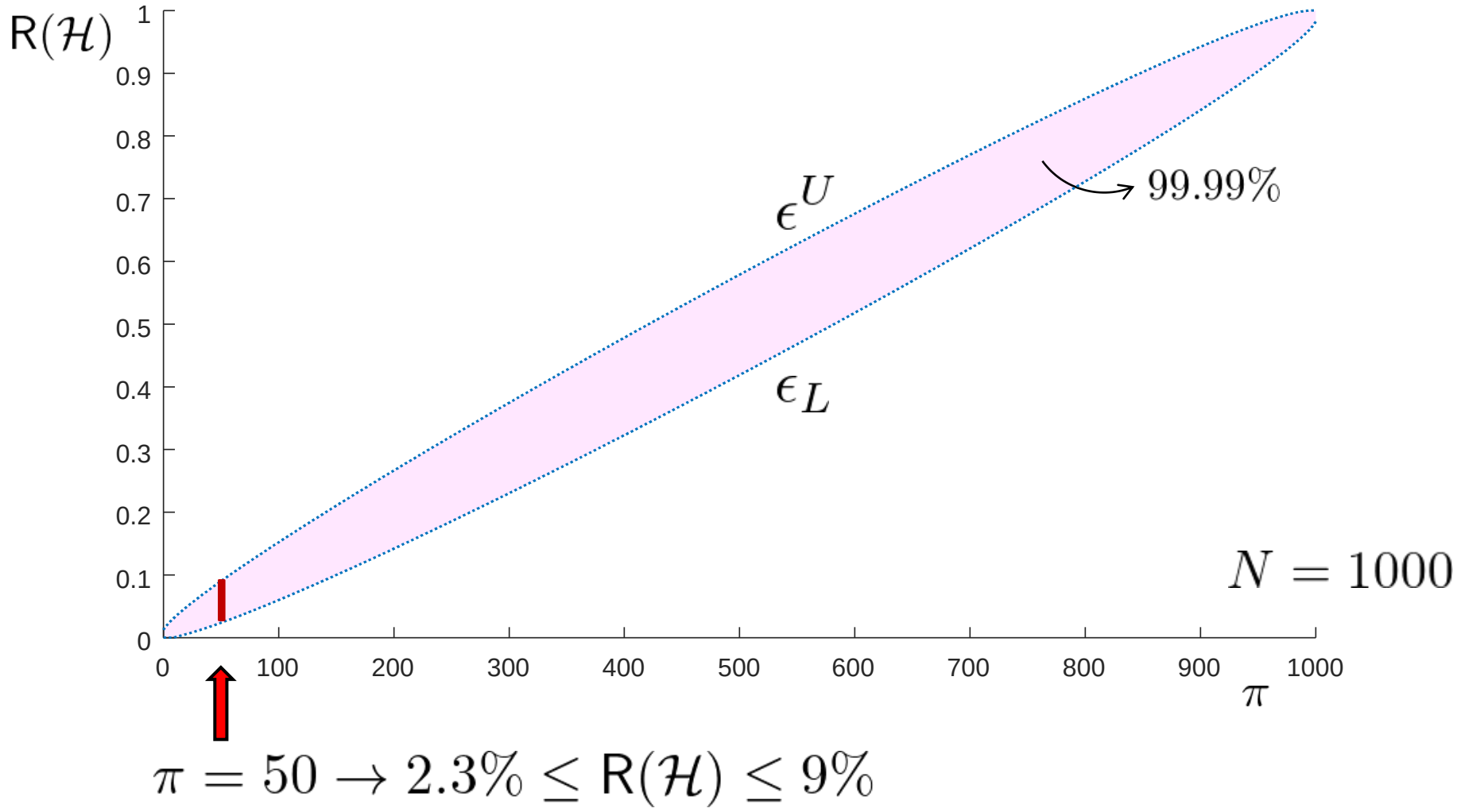
- $\binom{N}{k}(1-\epsilon)^{N-k} - \frac{\beta}{2N}\sum_{m=k}^{N-1}\binom{m}{k}(1-\epsilon)^{m-k}$
- $\binom{N}{k}(1-\epsilon)^{N-k} - \frac{\beta}{2N}\sum_{m=N+1}^{2N}\binom{m}{k}(1-\epsilon)^{m-k}$

Then, irrespective of $\mathbb{P}$,

$$\mathbb{P}^N\left\{\delta_1,\ldots,\delta_N:\ \epsilon_L(\pi) \le \mathsf{R}(\mathcal{H}) \le \epsilon^U(\pi)\right\} \ge 1-\beta$$

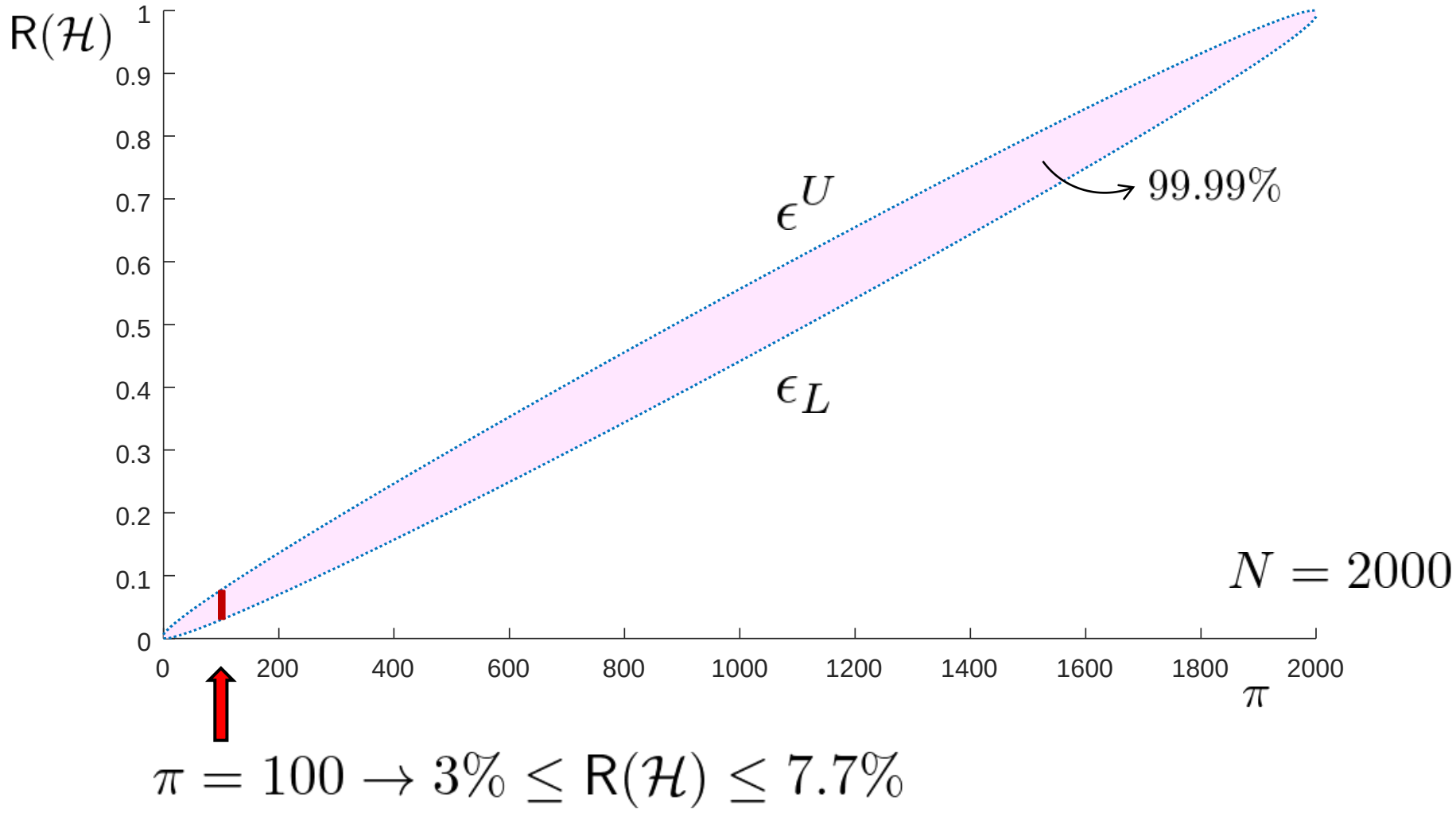
Main result (cont'd)

$\epsilon_L(\pi) \leq \mathsf{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



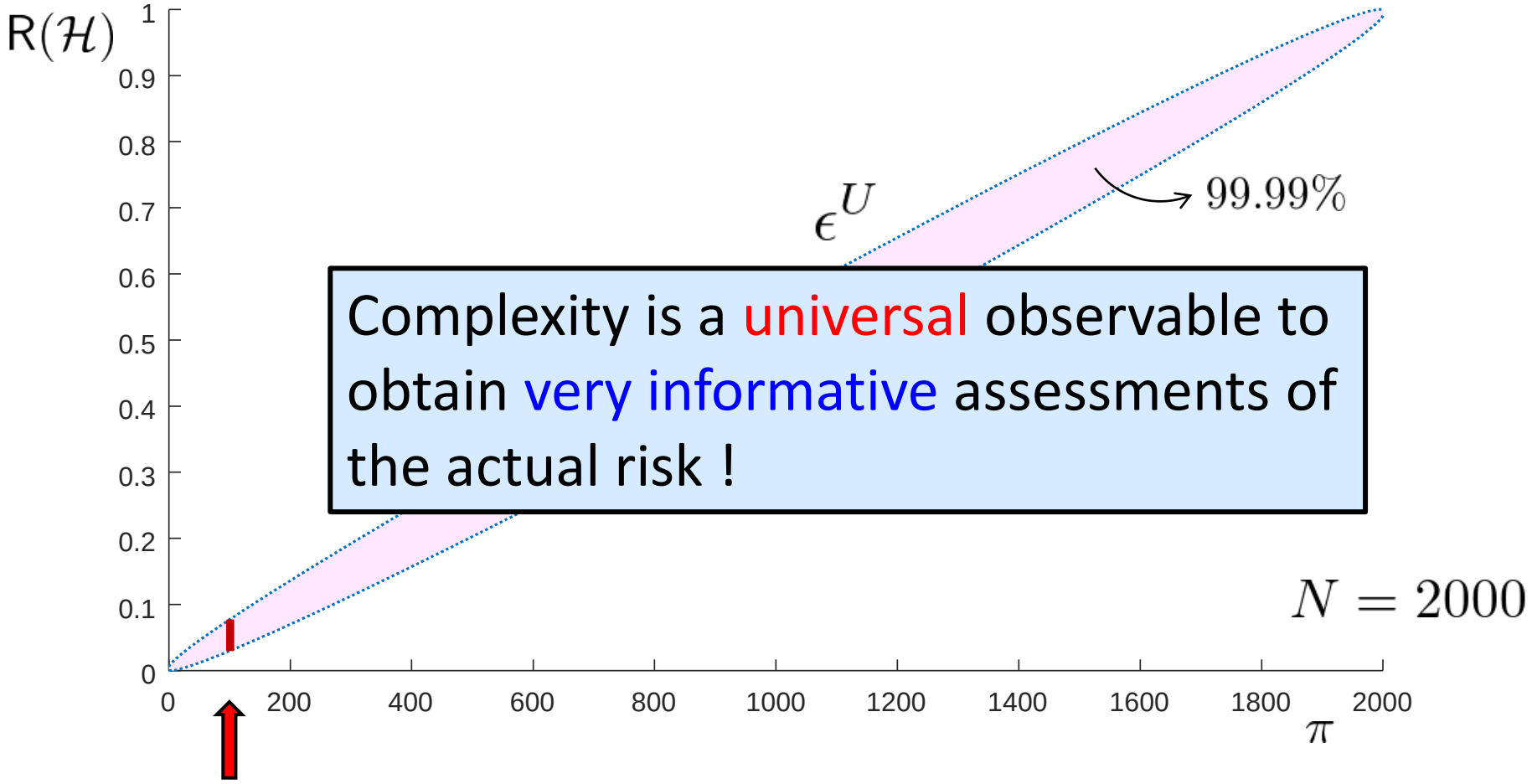
Main result (cont'd)

$\epsilon_L(\pi) \leq \mathsf{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



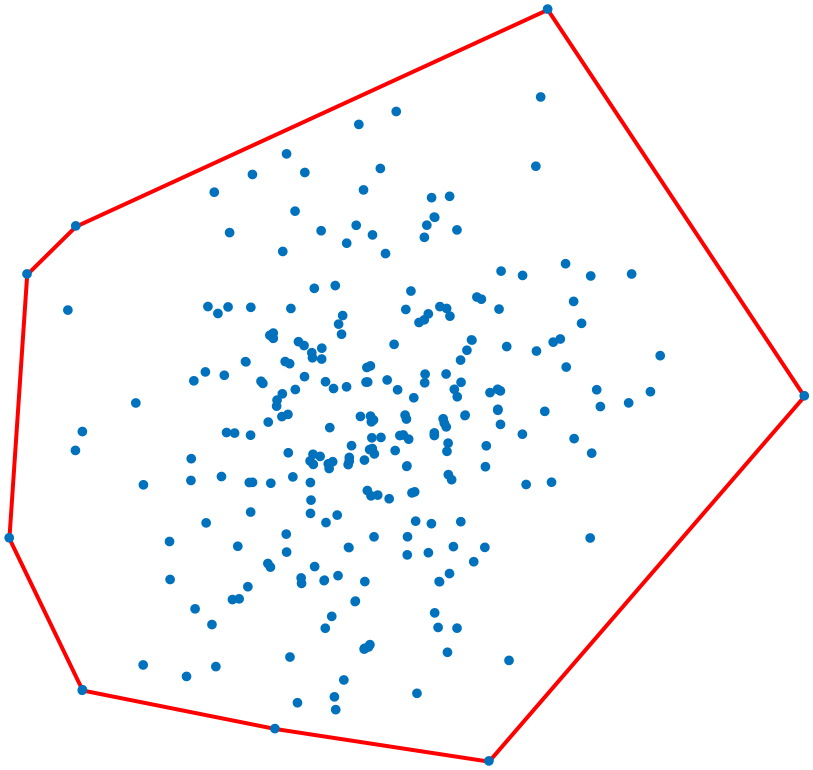
Main result (cont'd)

$\epsilon_L(\pi) \leq \mathrm{R}(\mathcal{H}) \leq \epsilon^U(\pi)$ **is true with confidence** $1 - \beta$



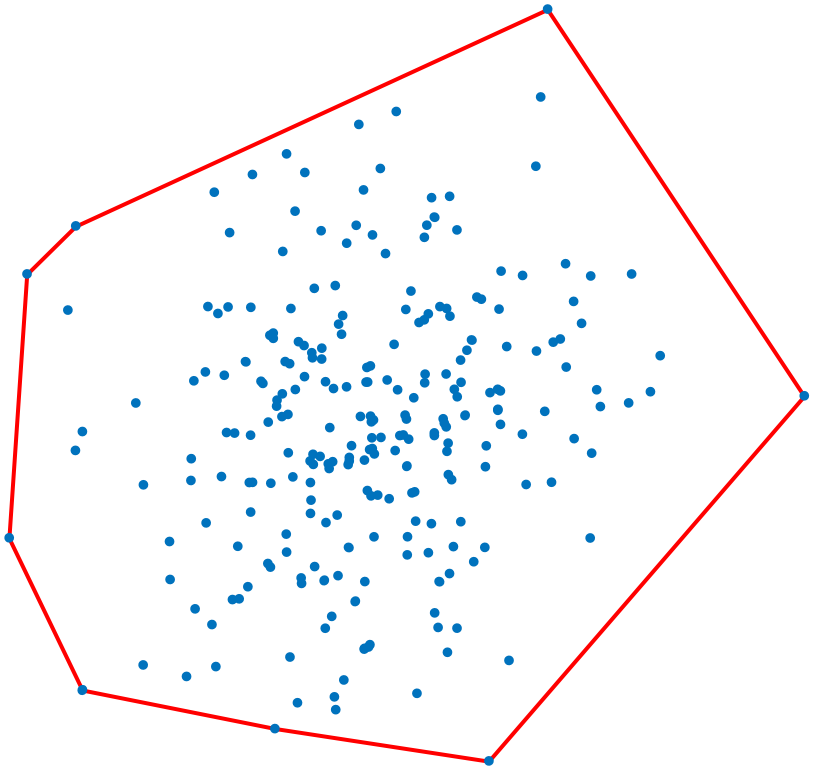
An example: convex hull

An example: convex hull



$$\begin{aligned} \min_{Q} \quad & \text{volume}(Q) \\ \text{s.t.} \quad & p^{(i)} \in Q, \\ & i = 1, \ldots, N \end{aligned}$$

An example: convex hull

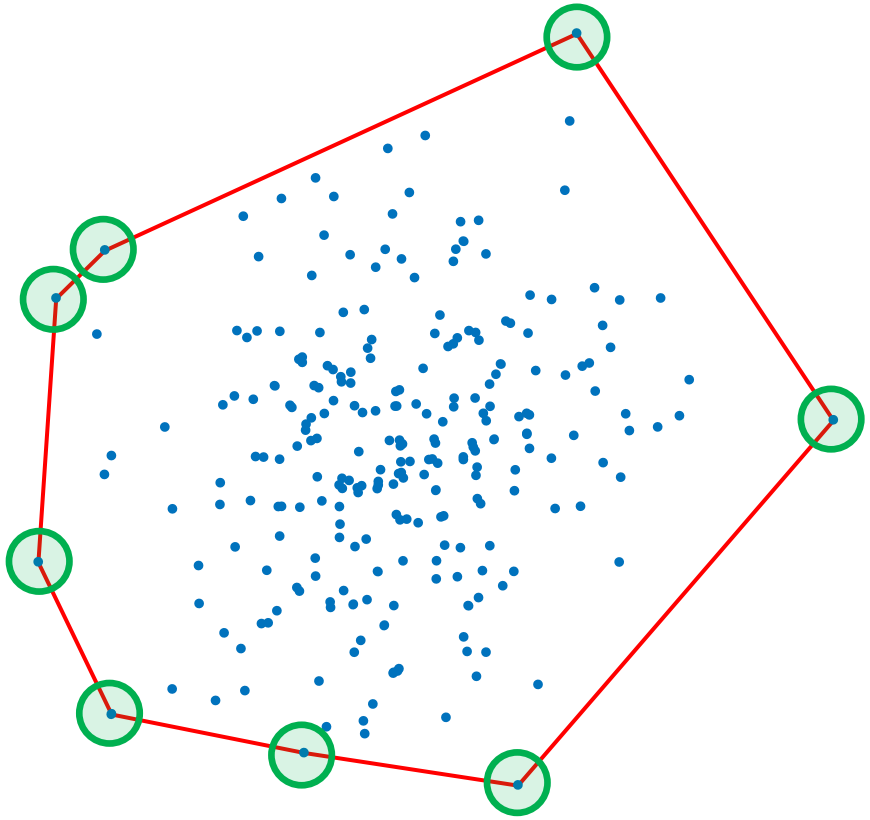


$$\begin{aligned} \min_{Q} \quad & \text{volume}(Q) \\ \text{s.t.} \quad & p^{(i)} \in Q, \\ & i = 1, \ldots, N \end{aligned}$$

Q inappropriate for p if p remains outside

Risk = mass outside

An example: convex hull



$$\min_{Q} \quad \text{volume}(Q)$$
$$\text{s.t.} \quad p^{(i)} \in Q,$$
$$i = 1, \ldots, N$$

Q inappropriate for p if p remains outside

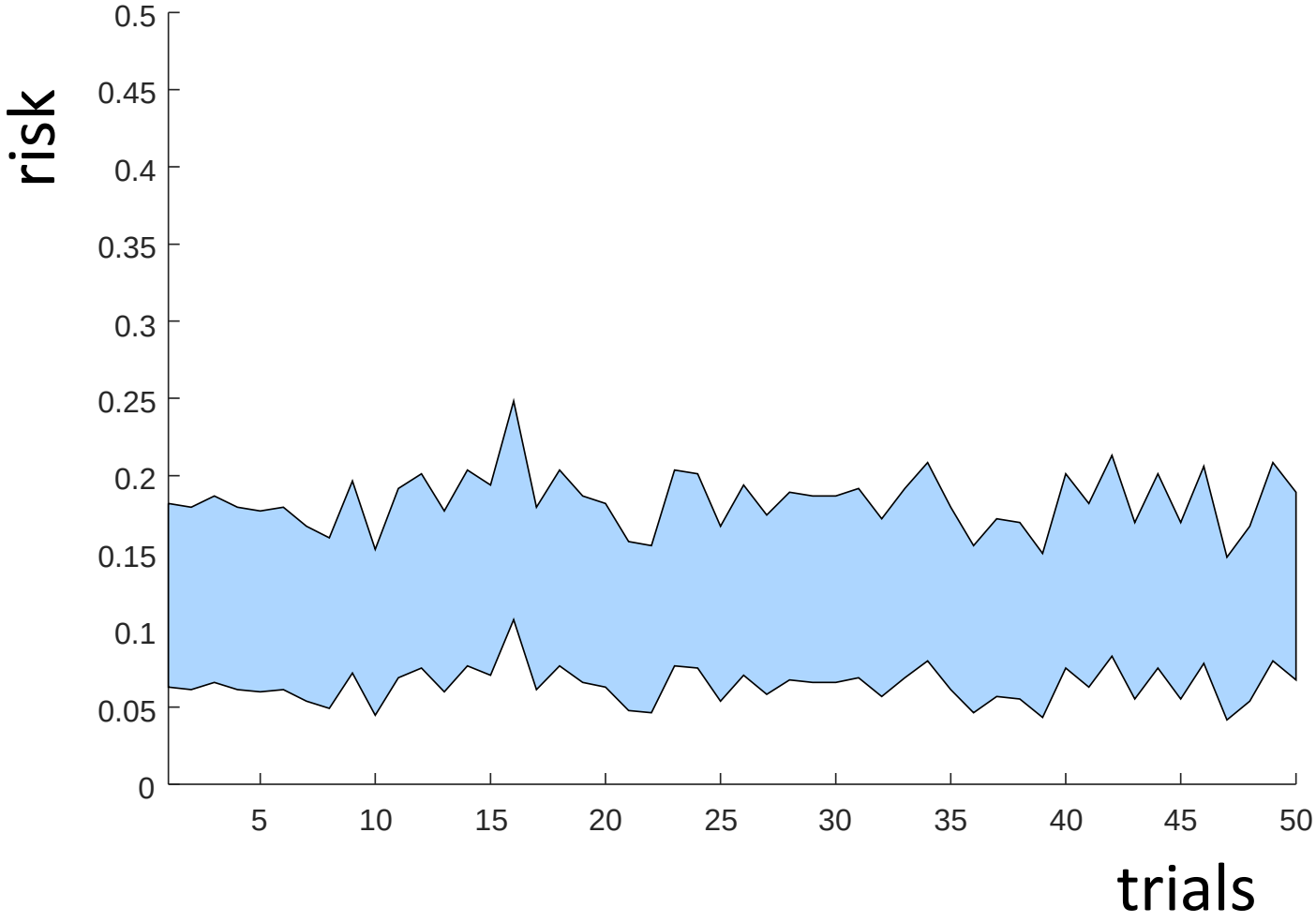
Risk = mass outside

Compression = vertexes of the convex hull

π = no. of vertexes

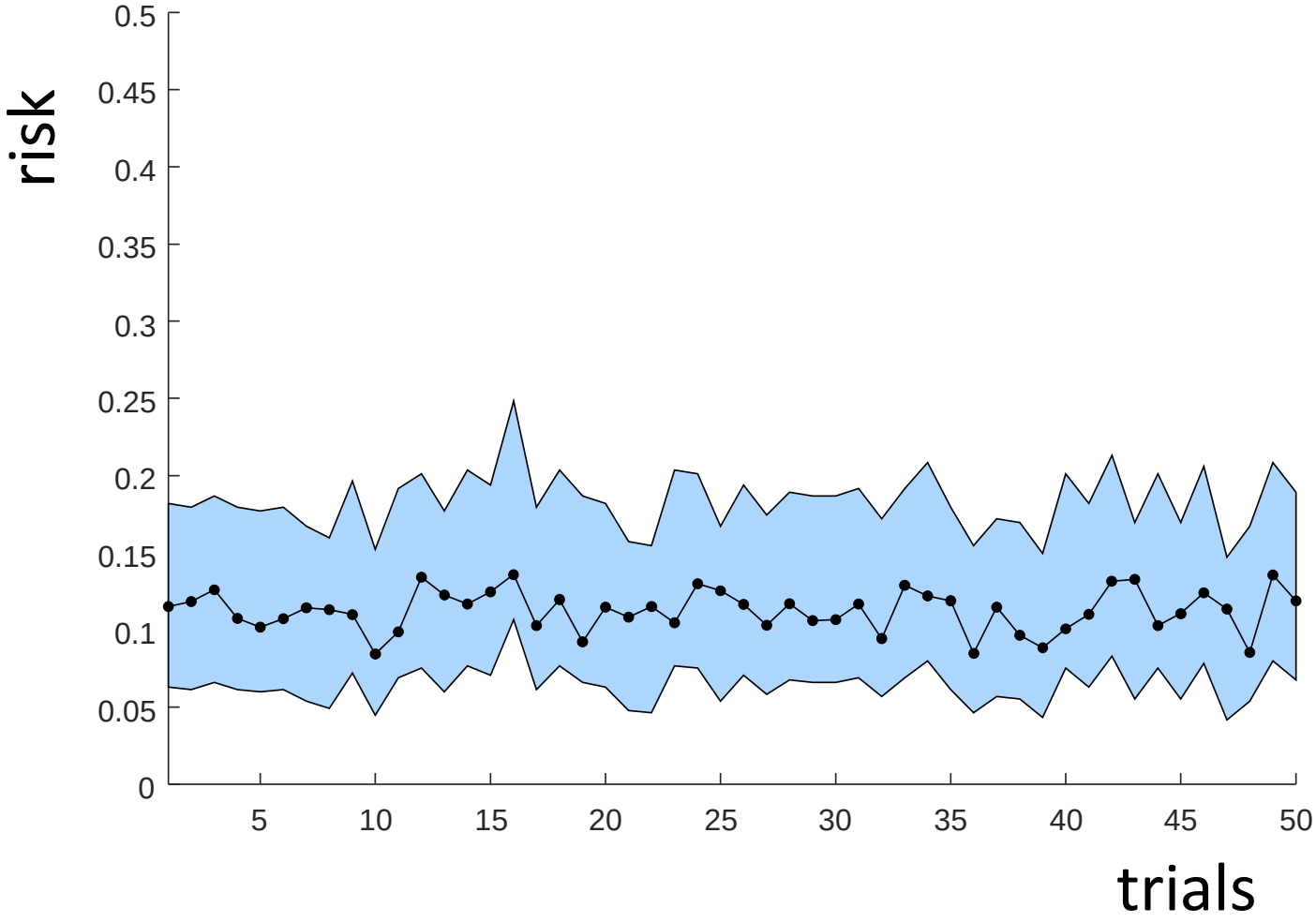
An example: convex hull

N = 500, risk assessment via sample compression theory



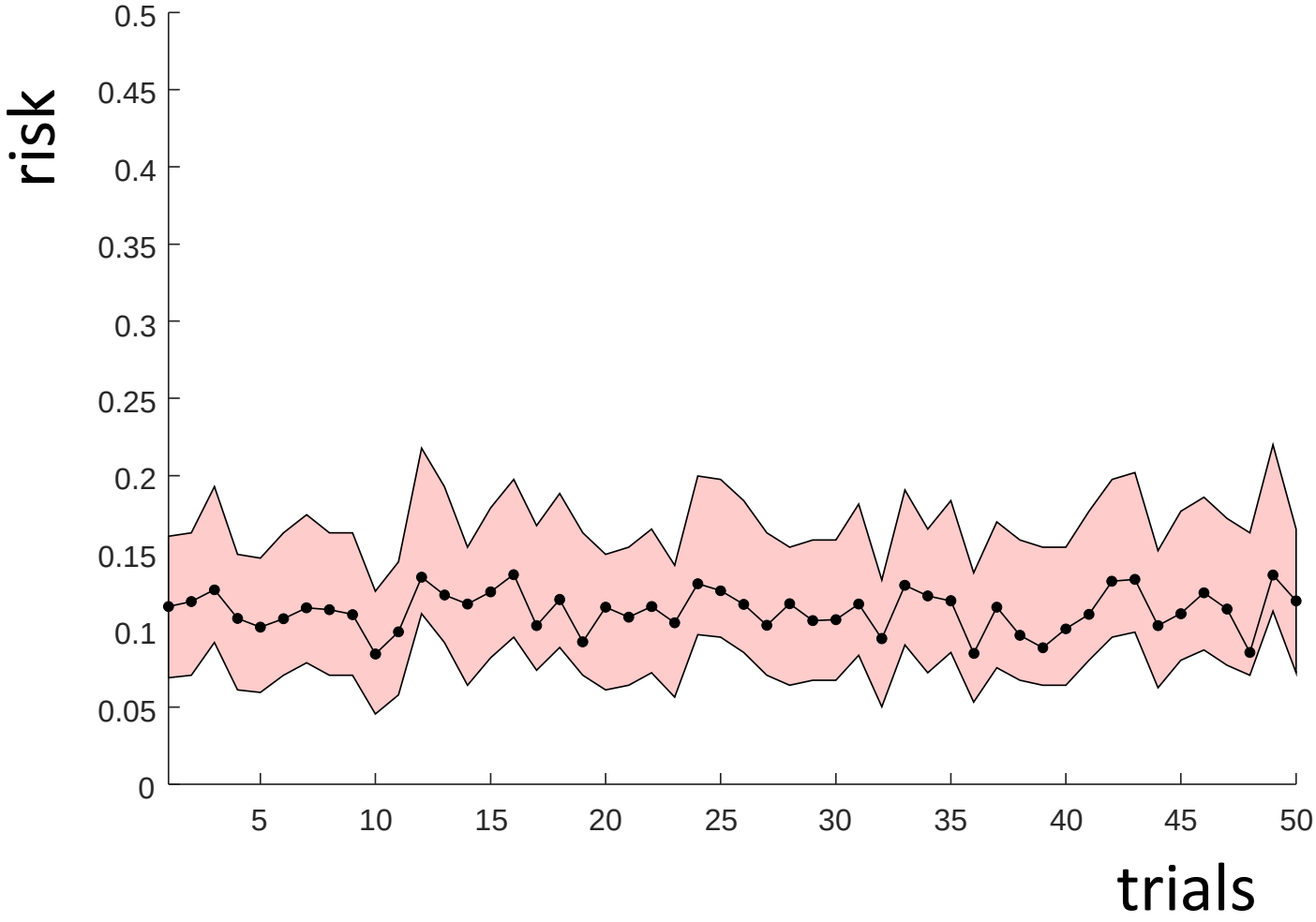
An example: convex hull

N = 500, risk assessment via sample compression theory



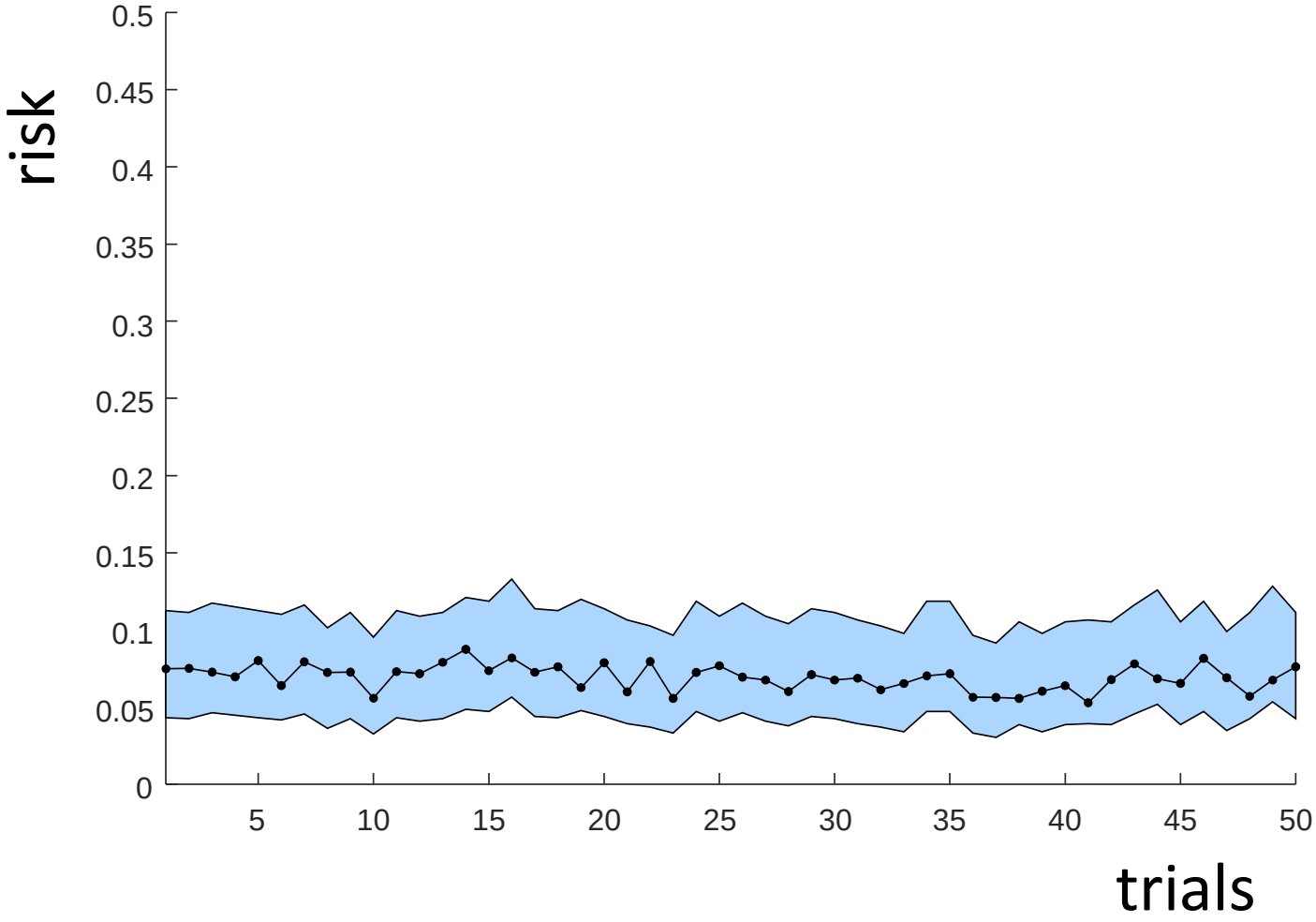
An example: convex hull

N = 500, risk assessment via test dataset (new 500 scenarios)

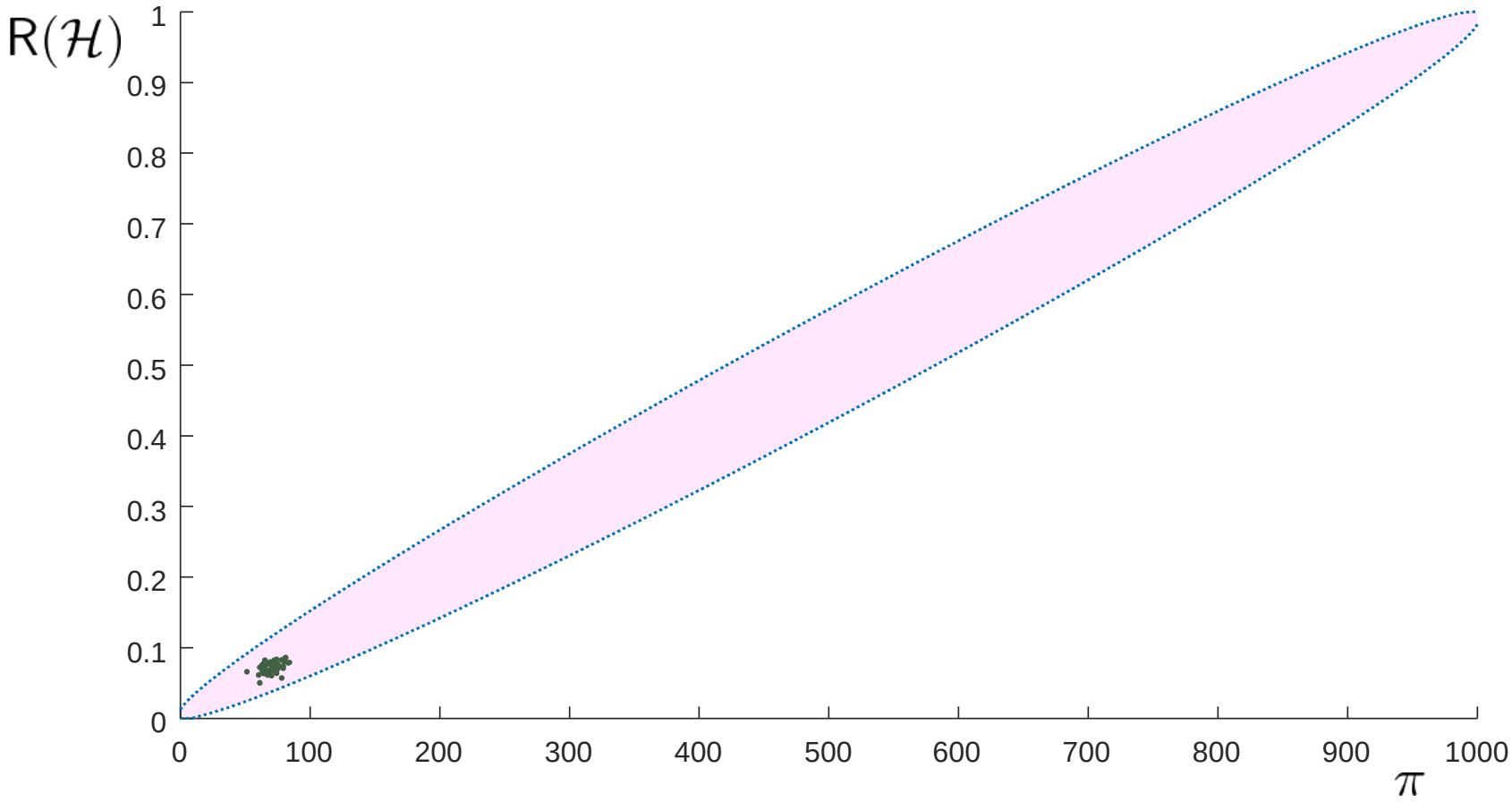


An example: convex hull

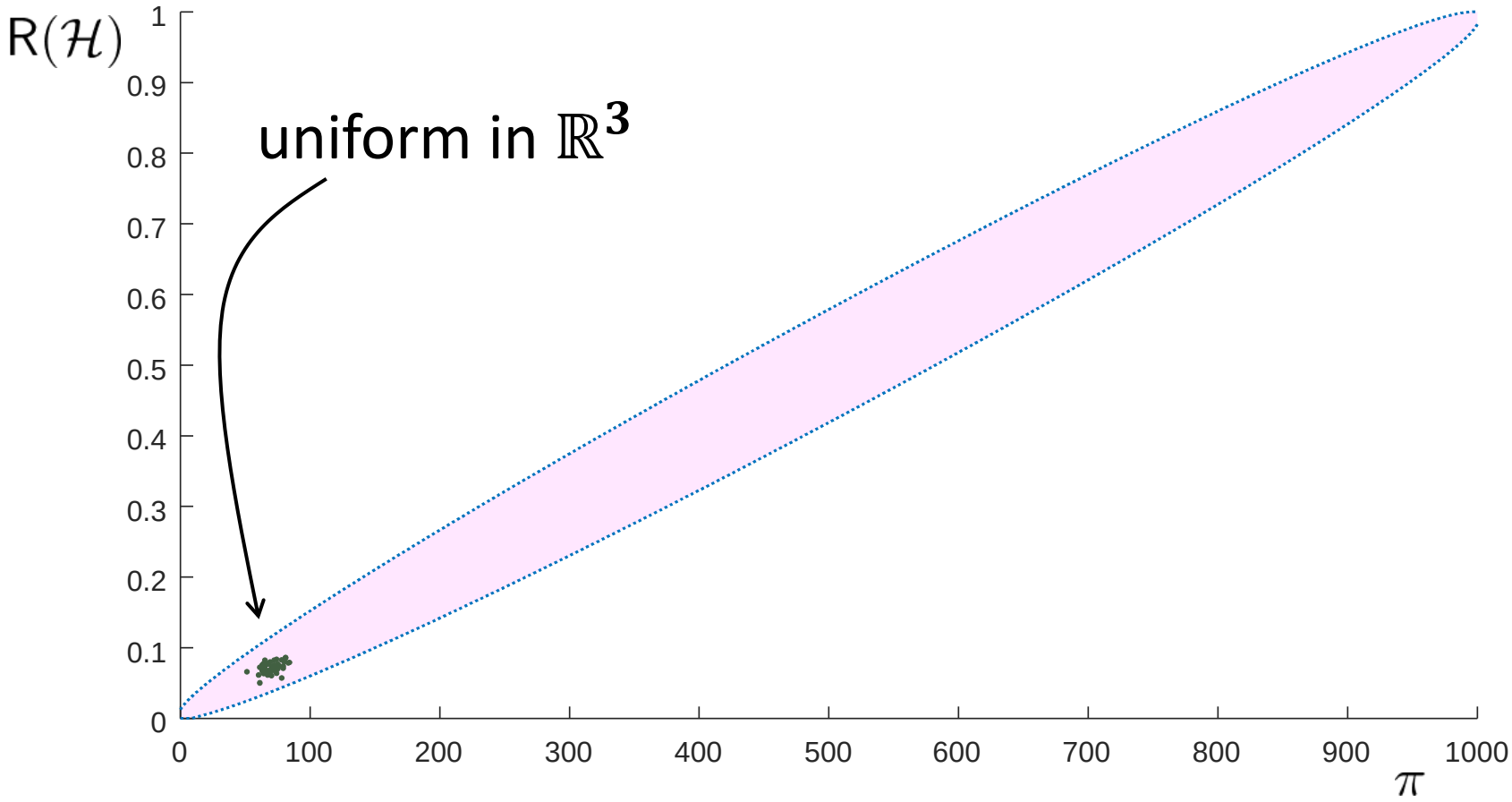
N = 1000, risk assessment via sample compression theory



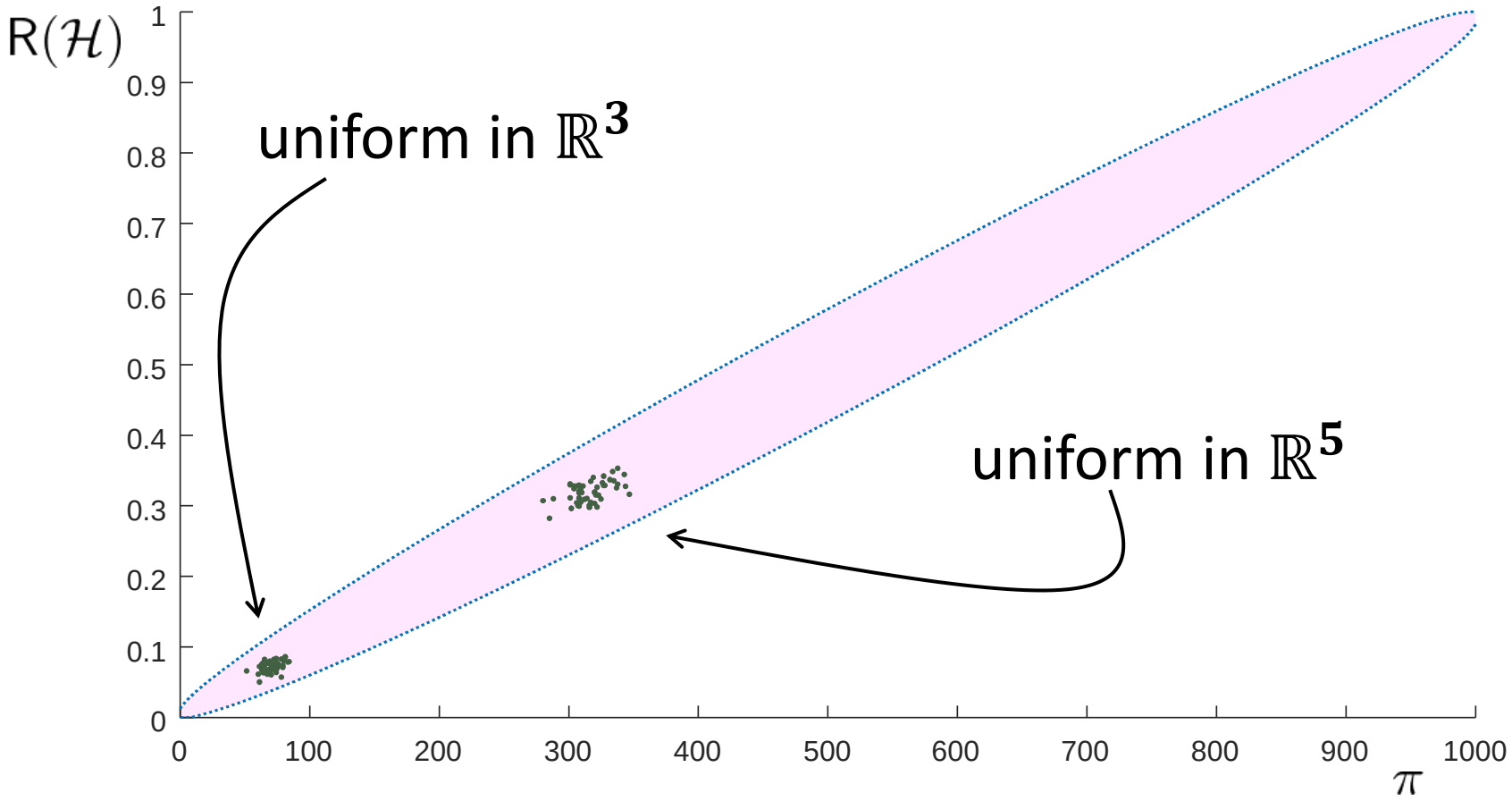
An example: convex hull



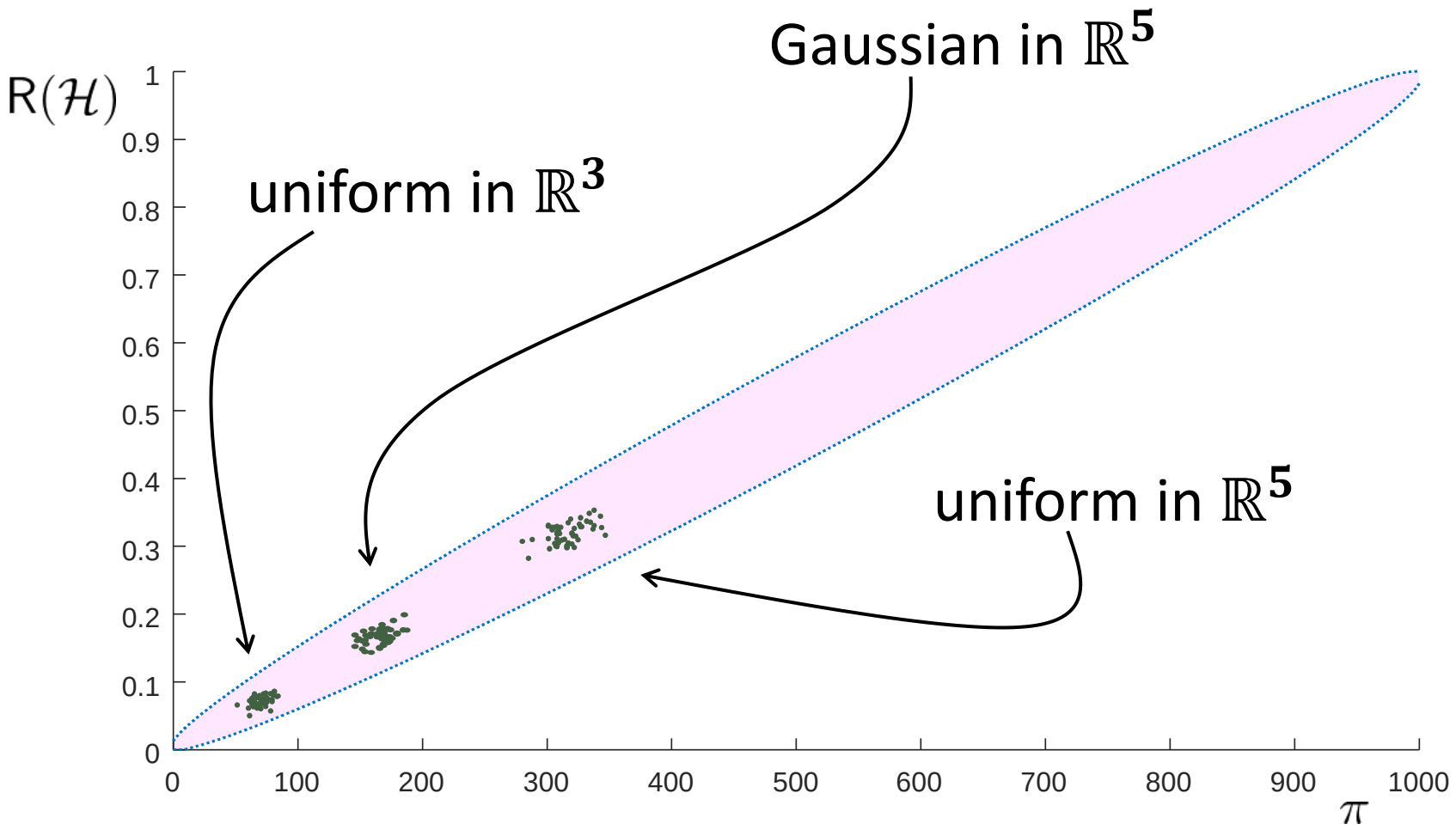
An example: convex hull



An example: convex hull



An example: convex hull



Preferent compression: range of applicability

Many learning algorithms (SV methods, all scenario optimization schemes…) naturally satisfy compression properties… many yet to be discovered…

However, many others do not… notably: Neural Networks

Preferent compression: range of applicability

Many learning algorithms (SV methods, all scenario optimization schemes…) naturally satisfy compression properties… many yet to be discovered…

However, many others do not… notably: Neural Networks

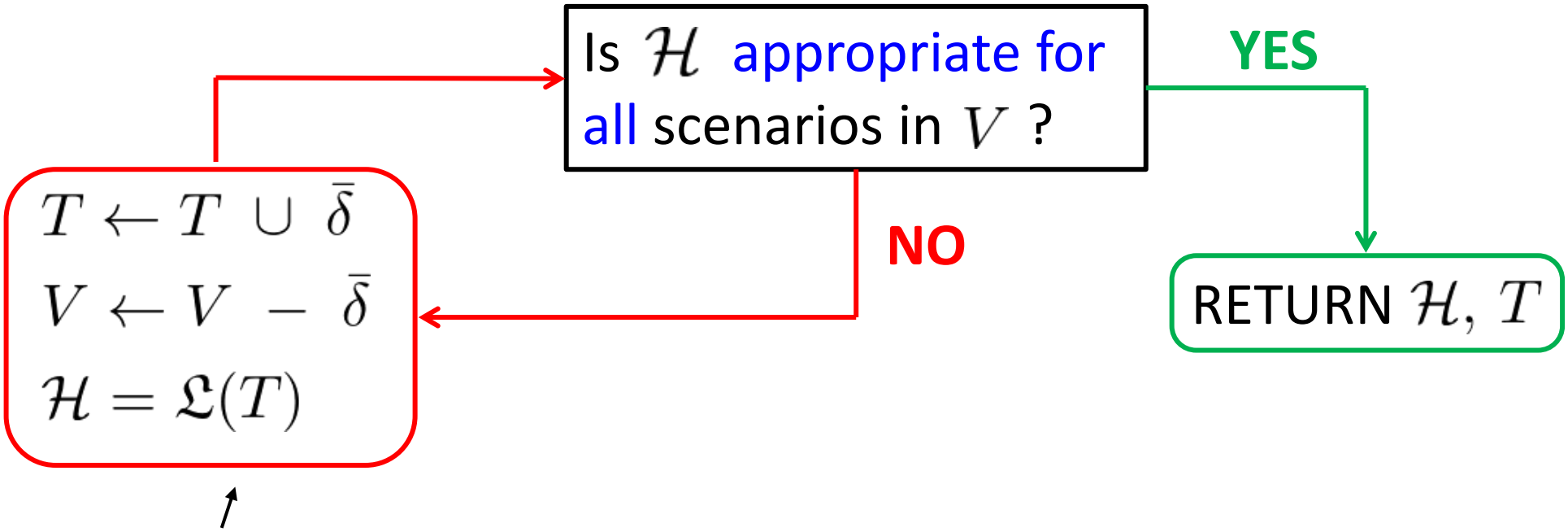
Idea – the Pick-to-Learn (P2L) algorithm:

construct a meta-algorithm that builds on the existing learning algorithm as a block-box to induce the compression properties

The Pick-to-Learn (P2L) algorithm

INPUT: scenarios $\delta_1, \delta_2, \ldots, \delta_N$, learning algorithm $\mathfrak{L}$, initial decision $\mathcal{H}_0$

Initialization: $T = \emptyset$, $V = (\delta_1, \ldots, \delta_N)$, $\mathcal{H} = \mathcal{H}_0$



$\bar{\delta}$ = element in V for which $\mathcal{H}$ is least appropriate

P2L: main features

P2L: $\delta_1, \ldots, \delta_N \rightarrow \mathcal{H}$

⇨ new learning-based decision scheme $\mathfrak{L}'$

P2L: $\delta_1, \ldots, \delta_N \rightarrow T$

⇨ compression function κ' associated to $\mathfrak{L}'$

P2L: main features

P2L: $\delta_1, \ldots, \delta_N \rightarrow \mathcal{H}$

⇨ new learning-based decision scheme $\mathfrak{L}'$

P2L: $\delta_1, \ldots, \delta_N \rightarrow T$

⇨ compression function κ' associated to $\mathfrak{L}'$

Theorem (with D. Paccagnan and M. Campi)

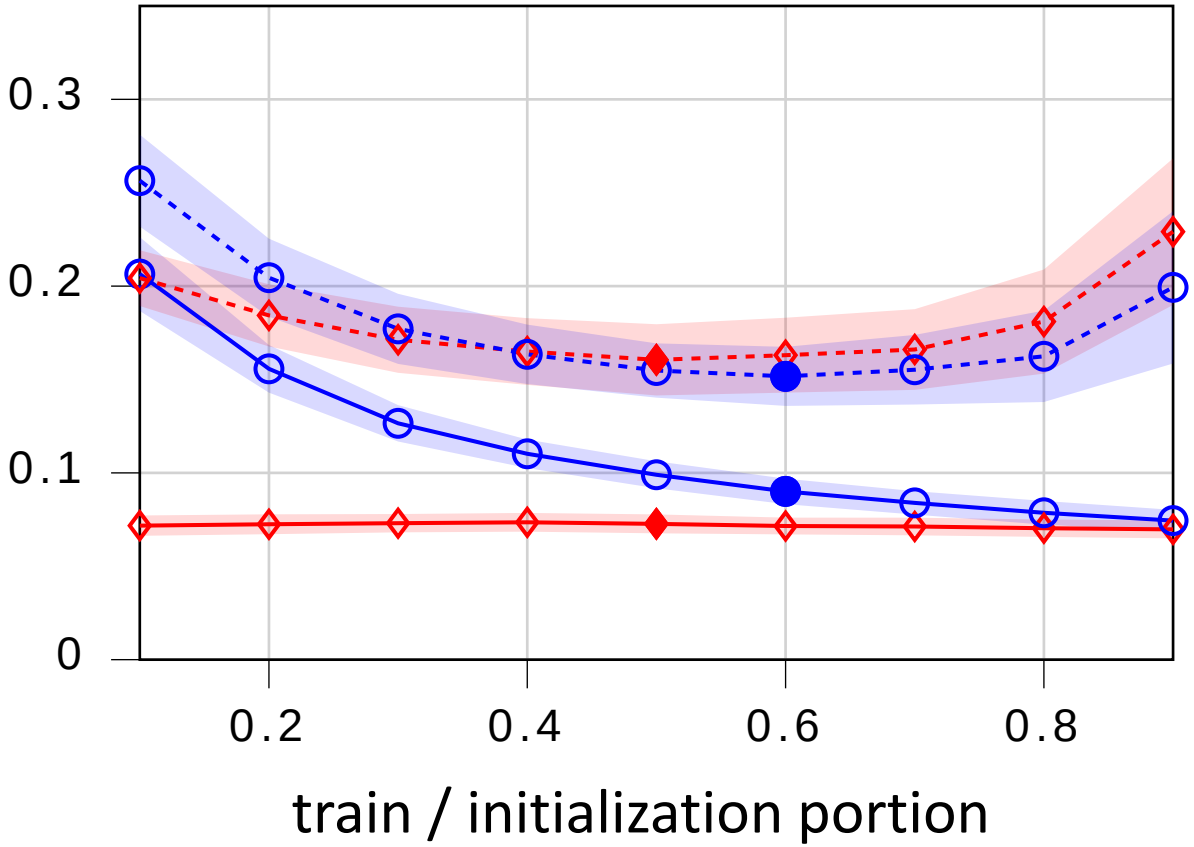
Preference and coherence hold true!

⇨ the risk of $\mathcal{H} = \mathfrak{L}'(\delta_1, \ldots, \delta_N)$ can be assessed via the size of T

Numerical example: classification of MNIST dataset

binary digit classification – $\mathfrak{L}$ = SGD for NN

P2L vs test-set validation (TSV)



risk bound P2L

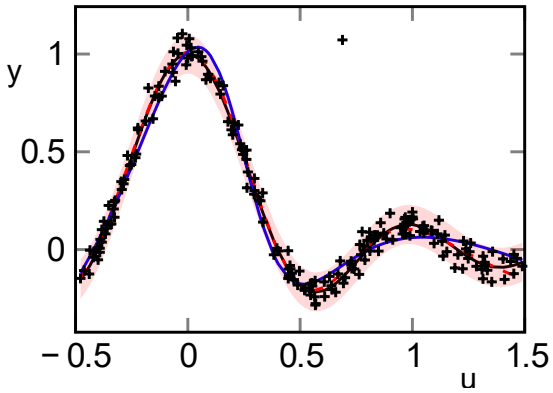
risk bound TSV

actual risk P2L

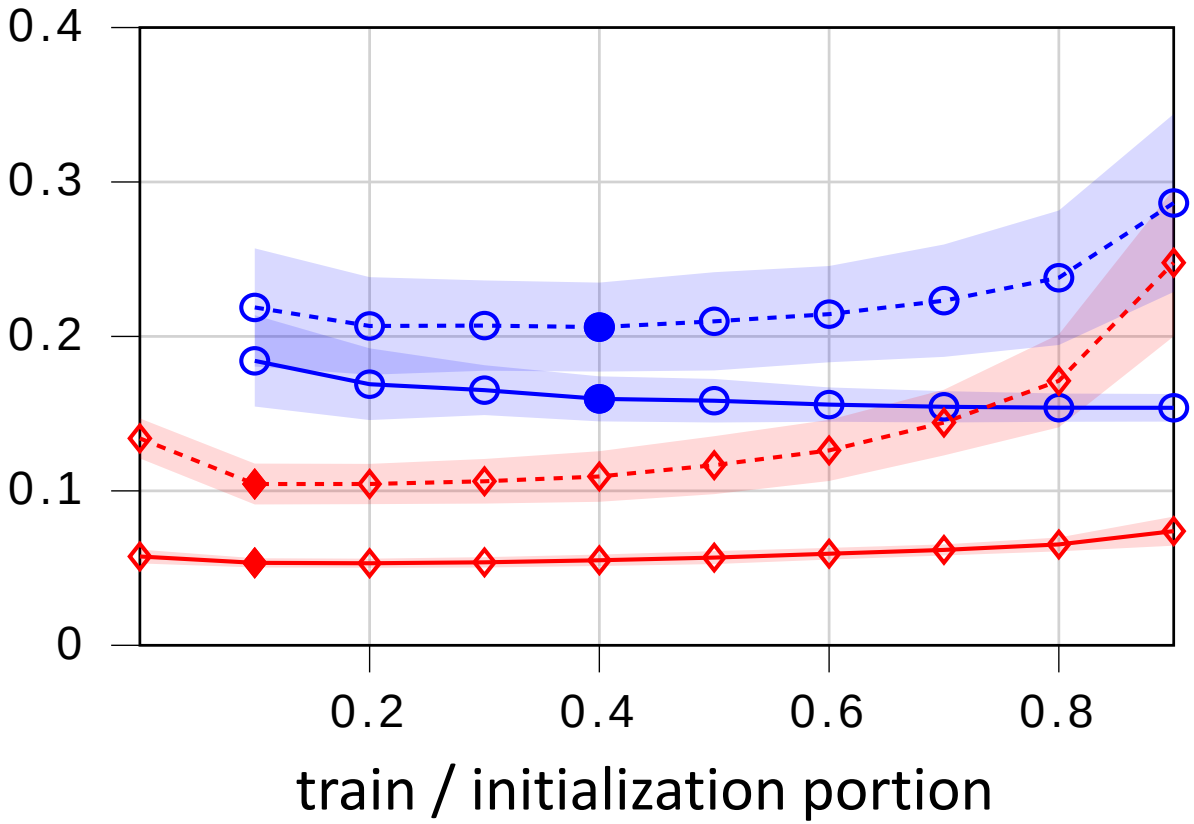
actual risk TSV

Numerical example: non-linear regression

non-linear regression – $\mathfrak{L}$ = SGD for NN



P2L vs test-set validation (TSV)



risk bound P2L

risk bound TSV

actual risk P2L

actual risk TSV

Acknowledgments

Many thanks to prof. Nicolò Cesa-Bianchi for insightful and inspiring discussions

This research is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence)

Relevant articles

- *M.C. Campi, S. Garatti.* Compression, Generalization and Learning. *Journal of Machine Learning Research,* 24(339):1-74, 2023.
- *D. Paccagnan, M.C. Campi, S. Garatti,* The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds and Improved Post-training Performance. *In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 2023.

Acknowledgments

Many thanks to prof. Nicolò Cesa-Bianchi for insightful and inspiring discussions

This research is
project, fu
scheme (M

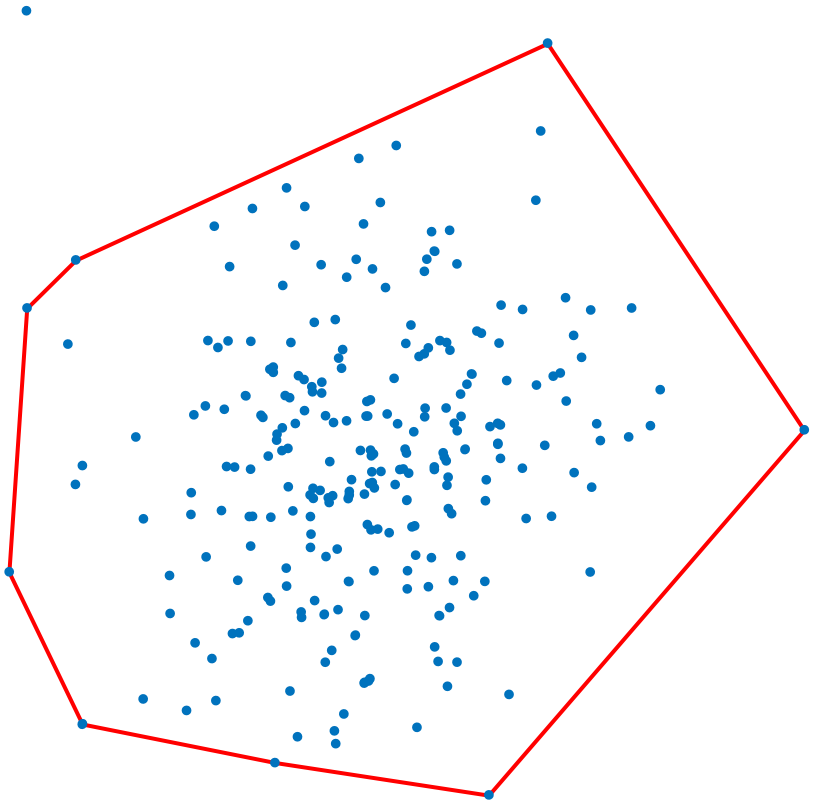
Thank you !

FUTURE AI RESEARCH

Relevant a

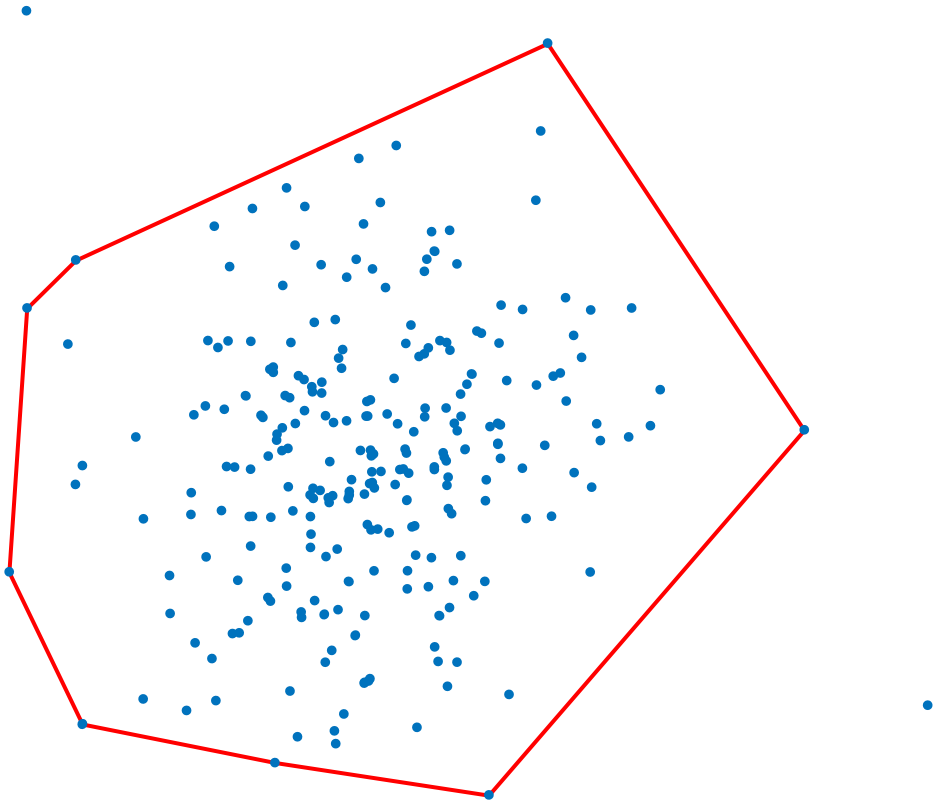
- *M.C. Campi, S. Garatti.* Compression, Generalization and Learning. *Journal of Machine Learning Research,* 24(339):1-74, 2023.
- *D. Paccagnan, M.C. Campi, S. Garatti,* The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds and Improved Post-training Performance. *In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 2023.

An example: convex hull



$$\begin{aligned} \min_{Q} \quad & \text{volume}(Q) + \rho \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & \text{dist}(p^{(i)}, Q) \leq \xi_i \\ & i = 1, \ldots, N \end{aligned}$$

An example: convex hull

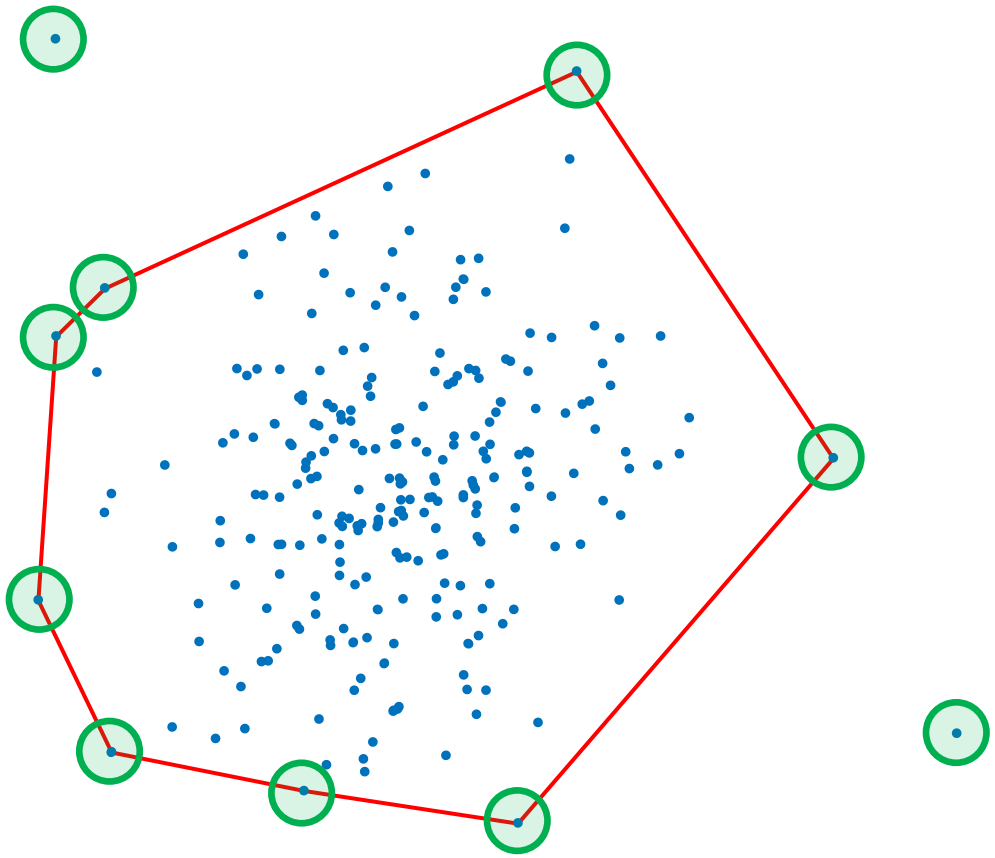


$$\begin{aligned} \min_{Q} \quad & \text{volume}(Q) + \rho \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & \text{dist}(p^{(i)}, Q) \leq \xi_i \\ & i = 1, \ldots, N \end{aligned}$$

Q inappropriate for p if p remains outside

Risk = mass outside

An example: convex hull



$$\min_{Q} \quad \text{volume}(Q) + \rho \sum_{i=1}^{N} \xi_i$$
$$\text{s.t.} \quad \text{dist}(p^{(i)}, Q) \leq \xi_i$$
$$i = 1, \ldots, N$$

Q inappropriate for p if p remains outside

Risk = mass outside

Compression = vertexes of the convex hull + violated

π = no. of vertexes + violated