Extension of sample size in Latin Hypercube Sampling with correlated variables

Miroslav Vořechovský

Institute of Structural Mechanics
Faculty of Civil Engineering
Brno University of Technology
Czech Republic

Outline

- Motivation
- Correlation control over samples of random vectors
- Adding samples for LHS
  - Random variable
  - Random vector
- Numerical examples (convergence)
- Conclusions

Motivation for sample extension

- Function of a random vect. $\mathbf{X} \in \mathbb{R}^{N_V}$ $\quad Z = g(\mathbf{X}), \quad Z \in \mathbb{R}$

  $\mathbf{Z} = \mathbf{g}(\mathbf{X}), \quad \mathbf{Z} \in \mathbb{R}^{N_R}$
- Analytical transformation $\mathbf{X} \xrightarrow{g} Z$ is not possible/feasible
- $g(\mathbf{X})$ is either a numerical of physical experiment – a function that is expensive to evaluate (e.g. FEM model)

- Perform statistical, sensitivity and reliability analysis of $g(\mathbf{X})$
- approximation to various integrals involving $g(\mathbf{X})$ and $f(\mathbf{X})$, $\mathbf{X} \sim f(\mathbf{X})$

Motivation for sample extension

- Function of a random vect. $\mathbf{X} \in \mathbb{R}^{N_V}$ $Z = g(\mathbf{X}), \quad Z \in \mathbb{R}$

  $\mathbf{Z} = \mathbf{g}(\mathbf{X}), \quad \mathbf{Z} \in \mathbb{R}^{N_R}$
- Analytical transformation $\mathbf{X} \xrightarrow{g} Z$ is not possible/feasible
- $g(\mathbf{X})$ is either a numerical of physical experiment

Inputs (have):

- N_V marginal densities of random vector $\mathbf{X}$
- Target correlation structure $\mathbf{T}$
- Function $\mathbf{g}$

Outputs (want):

- Statistical analysis of Z
- Sensitivity of Z on $\mathbf{X}$
- Reliability analysis (e.g. probability of failure integral $p_f = \mathrm{P}(Z < 0)$

Motivation for sample extension

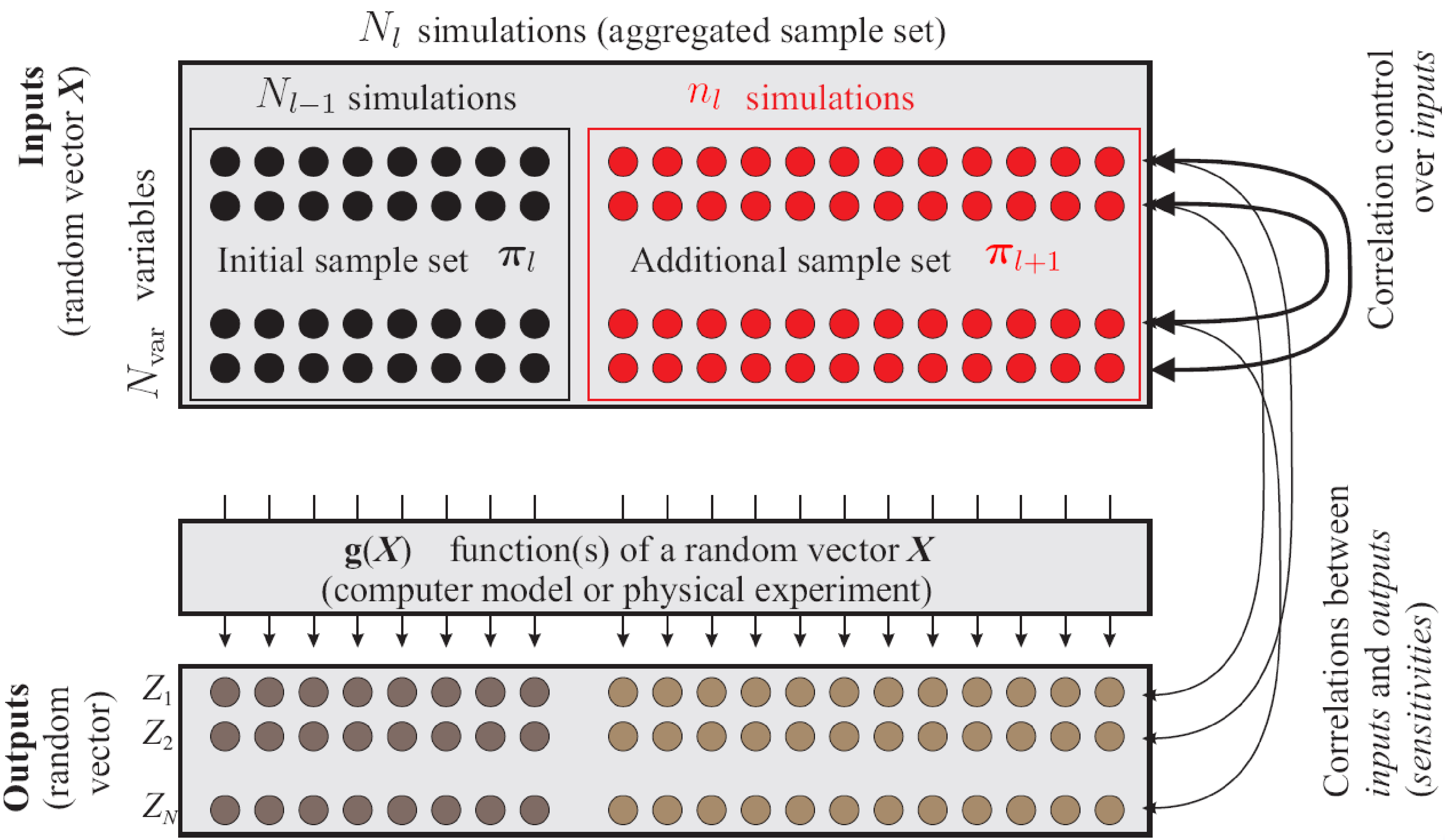
- Function of a random vect. $\mathbf{X} \in \mathbb{R}^{N_V}$ $Z = g(\mathbf{X}),\quad Z \in \mathbb{R}$

  $\mathbf{Z} = \mathbf{g}(\mathbf{X}),\quad \mathbf{Z} \in \mathbb{R}^{N_R}$
- Analytical transformation $\mathbf{X} \xrightarrow{g} Z$ is not possible/feasible
- $g(\mathbf{X})$ is either a numerical of physical experiment

<u>Problem:</u>

- Number of realizations needed is not known in advance
- Need to be able to
  - start with a small sample size and,
  - gradually refine the results by adding more simulations while keeping the old samples and results
  - achieve variance reduction

Merging of two subsets



Statistical correlation

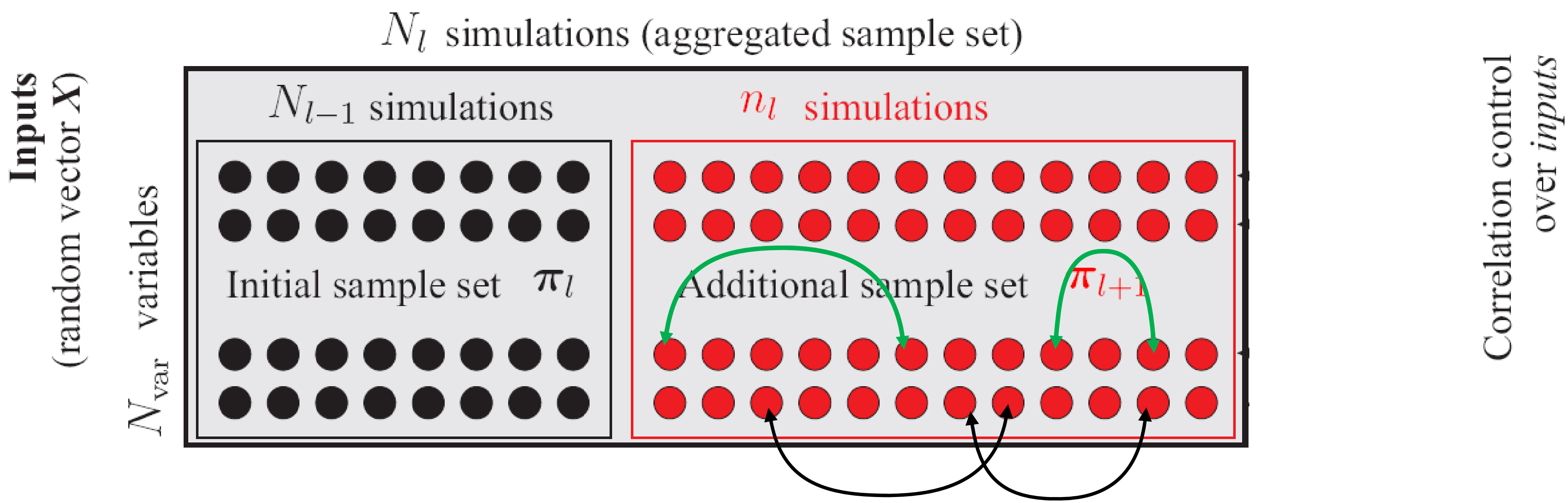
- To obtain meaningful estimates of statistics of $\mathbf{Z}$ and correlations $\mathbf{X}-\mathbf{Z}$, it is often important to precisely capture the input correlations structure $\mathbf{T}$ of the input
- We request match between $\mathbf{T}$ and $\mathbf{A}$
- Exploit combinatorial optimization algorithm based on SA developed previously by Vořechovský a Novák (2002,2009)
- **A** is estimated using e.g. Pearson or Spearman correlation coefficient, or Kendall's tau, copula,...

Statistical correlation

- To obtain meaningful estimates of statistics of $\mathbf{Z}$ and correlations $\mathbf{X}-\mathbf{Z}$, it is often important to precisely capture the input correlations structure $\mathbf{T}$ of the input
- We request match between $\mathbf{T}$ and $\mathbf{A}$
- Exploit combinatorial optimization algorithm based on SA developed previously by Vořechovský a Novák (2002,2009)
- **A** is estimated using e.g. Pearson:

$$\rho_{xy} = \frac{\sum_{i=1}^{N_S}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N_S}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{N_S}\left(y_i-\bar{y}\right)^2}}$$

Correlation control by rank shuffling



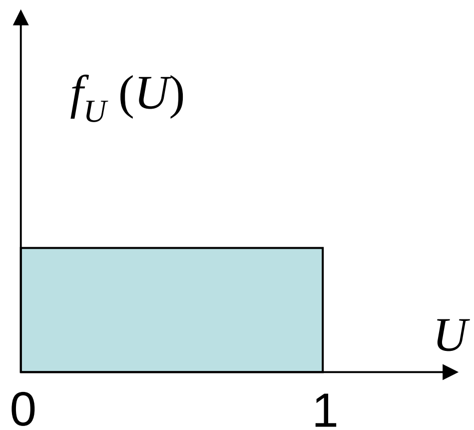
Hierarchical Subset
Latin Hypercube Sampling

- Why exploit LHS as basis for the proposed method?

  LHS never increases the variability of estimates as compared to crude MC.
  Moreover, the variability is usually significantly decreased (variance reduction technique).

HSLHS for a single random variable

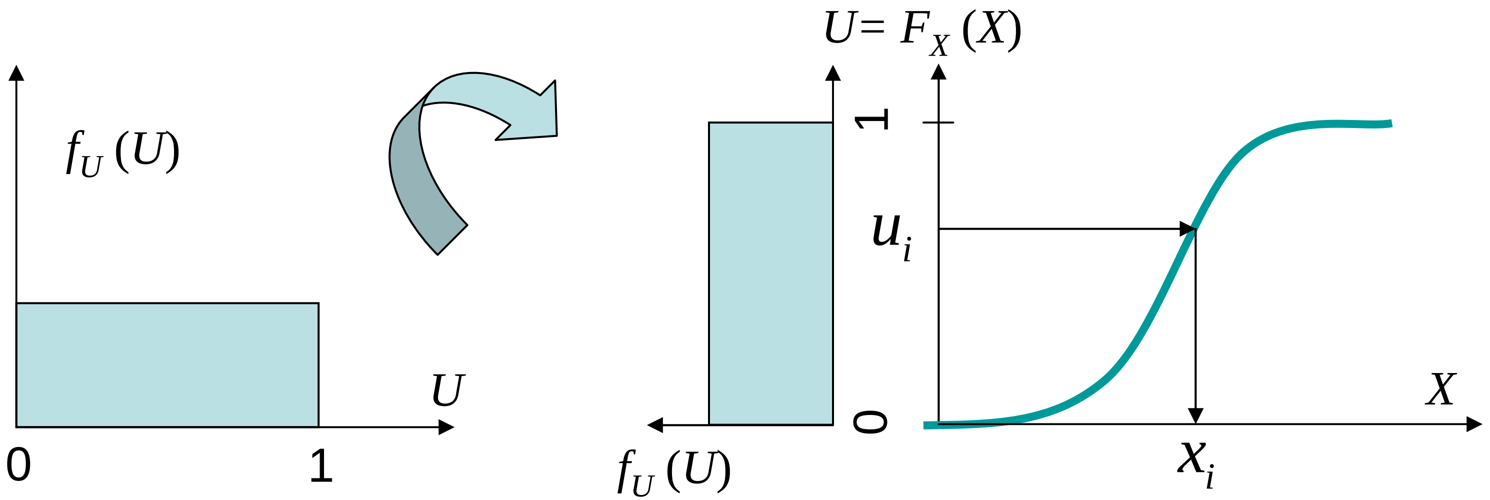
- Simplify the problem:
  Select samples from uniform distribution



$$U \sim R(0,1): \quad P(U \leq u) \equiv F_U(u)$$

HSLHS for a single random variable

- Illustration:



- sampling probabilities $u_i = \pi_i$ can be used in inverse transformation of dist'n function $X \; \square \; P(X \le x) \equiv F_X(x)$

$$x_i = F_X^{-1}(u_i) = F_X^{-1}(\pi_i)$$

HSLHS for a single random variable

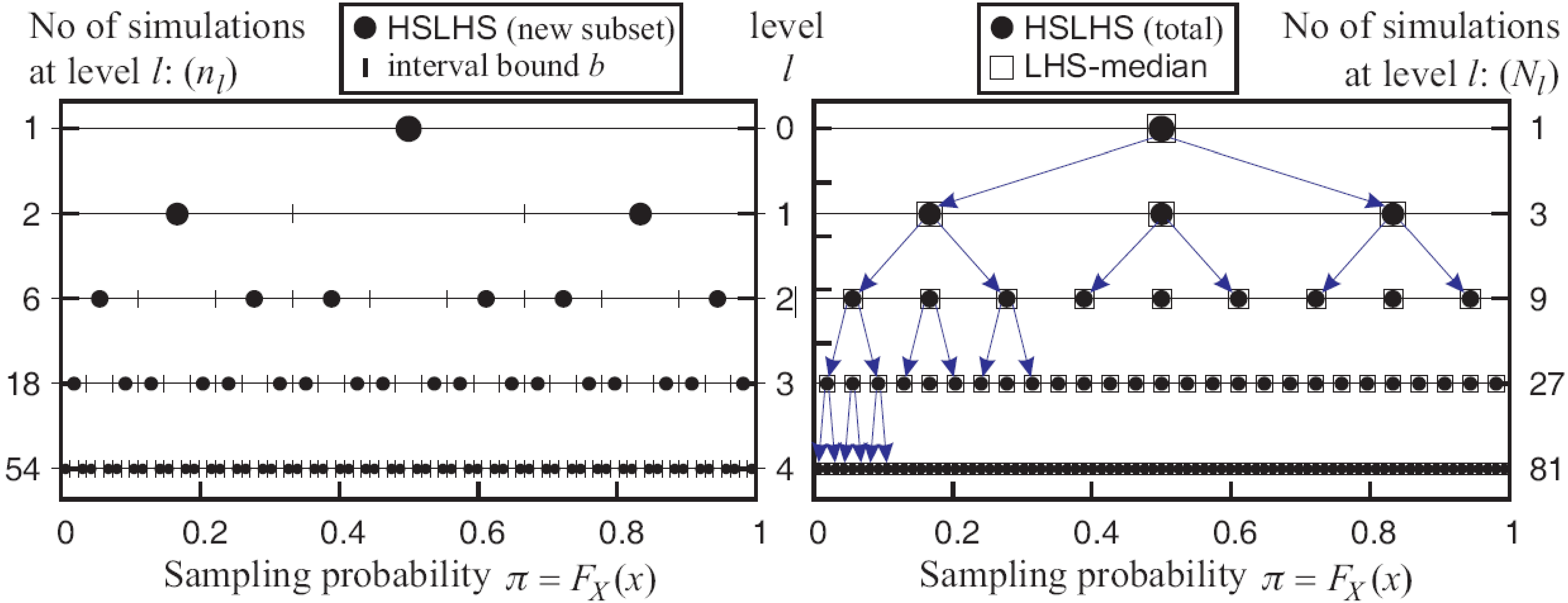
- levels l are associated with subsets of sampling probabilities $\Pi_l = \left\{ \pi_1^{(l)}, \pi_2^{(l)}, \ldots, \pi_i^{(l)}, \ldots, \pi_{n_l}^{(l)} \right\}$
- Sample size of a subset at each level is n_l
- The total sample size at level l is composed of the old sample and the new subset l: $N_l = N_{l-1} + n_l$
- The zero-level design is a regular LHS-median design: $\pi_i^{(l=0)} = \dfrac{i - 0.5}{n_0}$
- Each next subset doubles the previous total sample size:

$$N_l = N_{l-1} + n_l = N_{l-1} + 2N_{l-1} = 3N_{l-1} = N_0 3^l$$

HSLHS for a single random variable

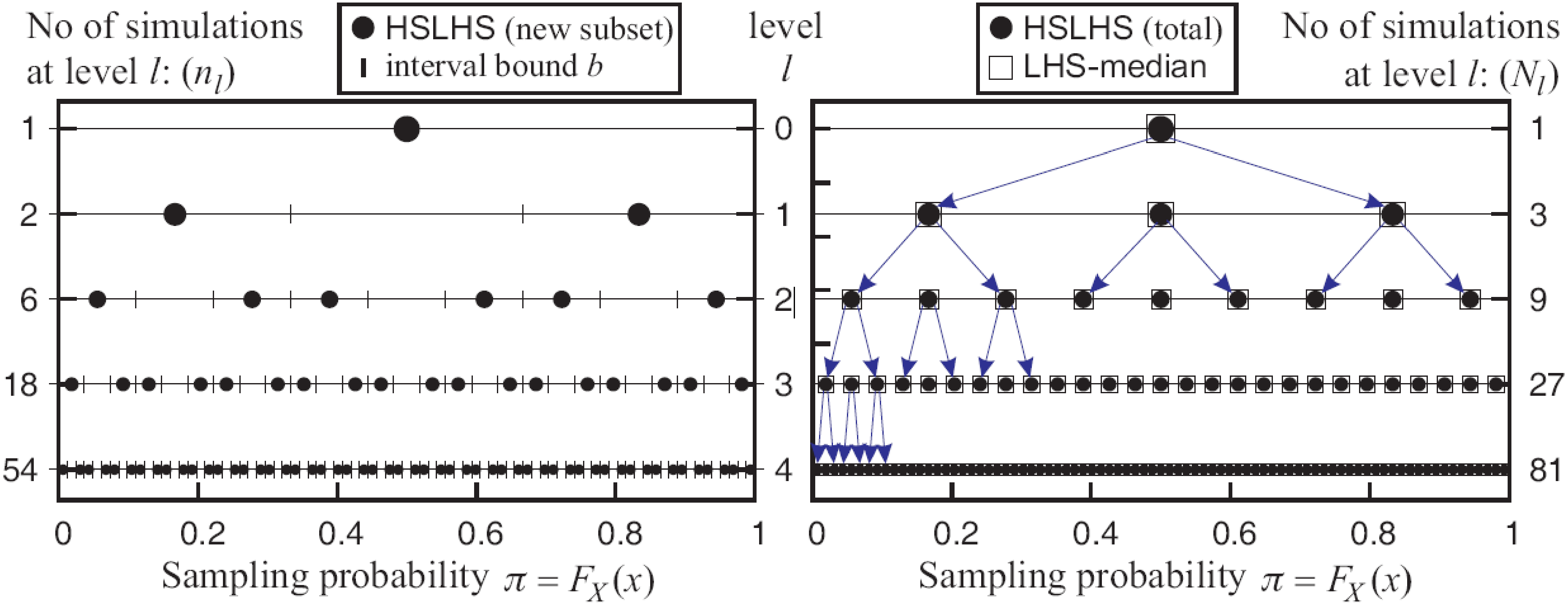
levels and sampling probabilities

(Analogy with refinement of integration points in numerics)



HSLHS for a single random variable

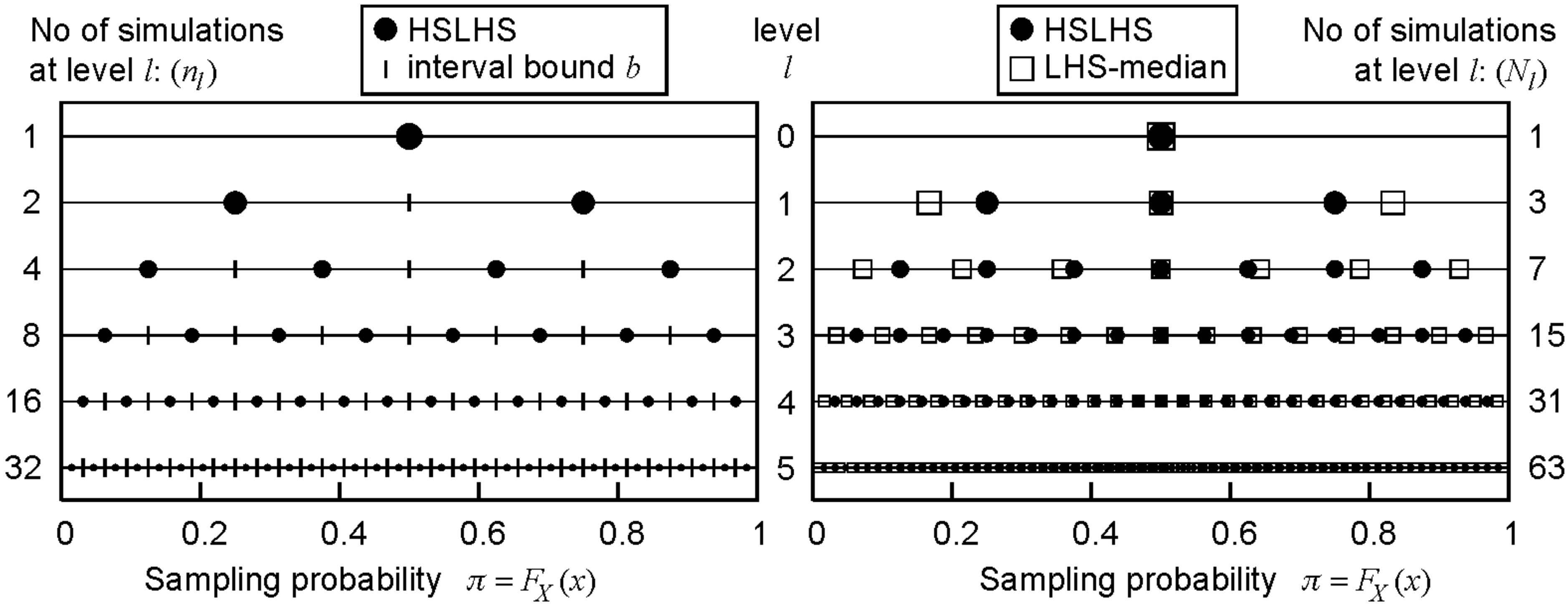
optimal coverage of the probabilistic space. Perfect regularity of sampling (LHS) guarantees good properties



HSLHS for a single random variable

levels and sampling probabilities

(Alternative design combining the true LHS subsets)

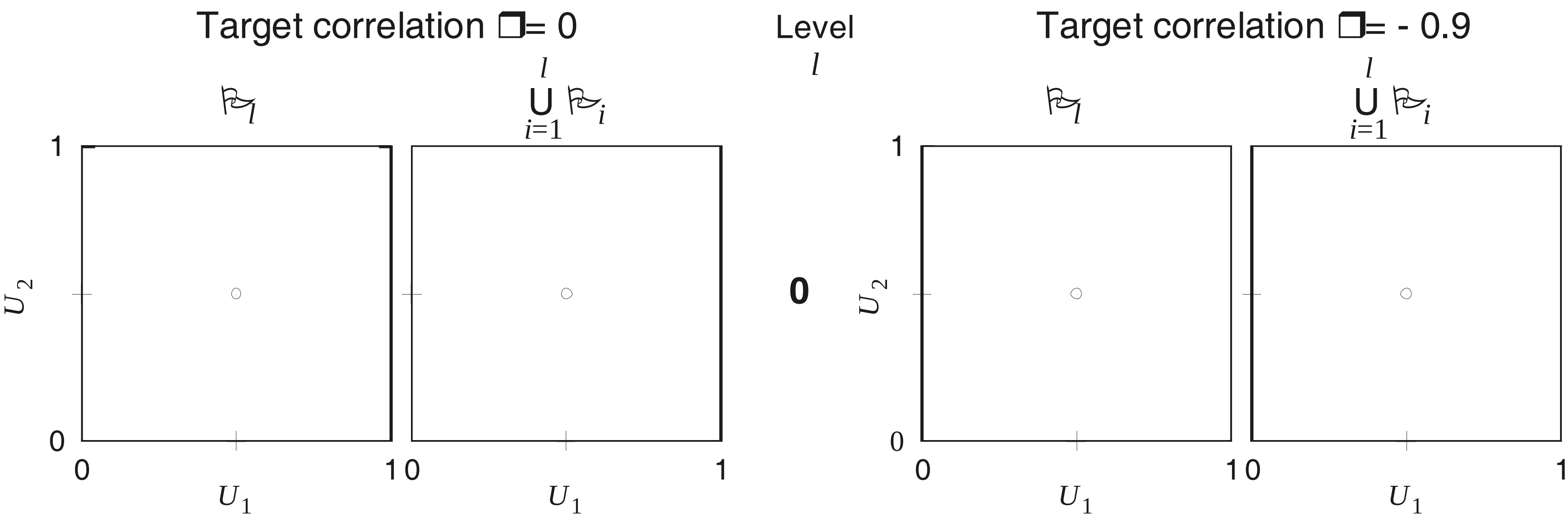


HSLHS for a random vector

- Combination of aforementioned components:
  - simulate marginals according to the described scheme
  - pair mutually all newly added samples such that $\mathbf{A} \to \mathbf{T}$

HSLHS for a random vector

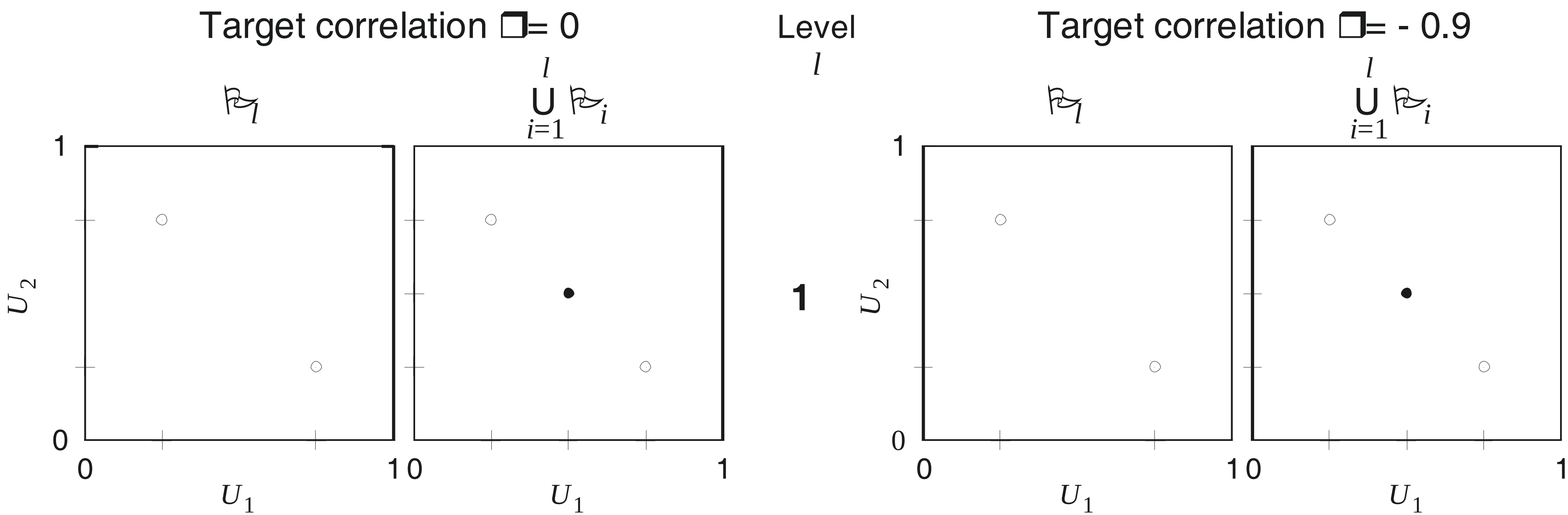
example: uniform distribution $\mathbf{U}=\{U_1,U_2\}: \quad U_i \,\square\, R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

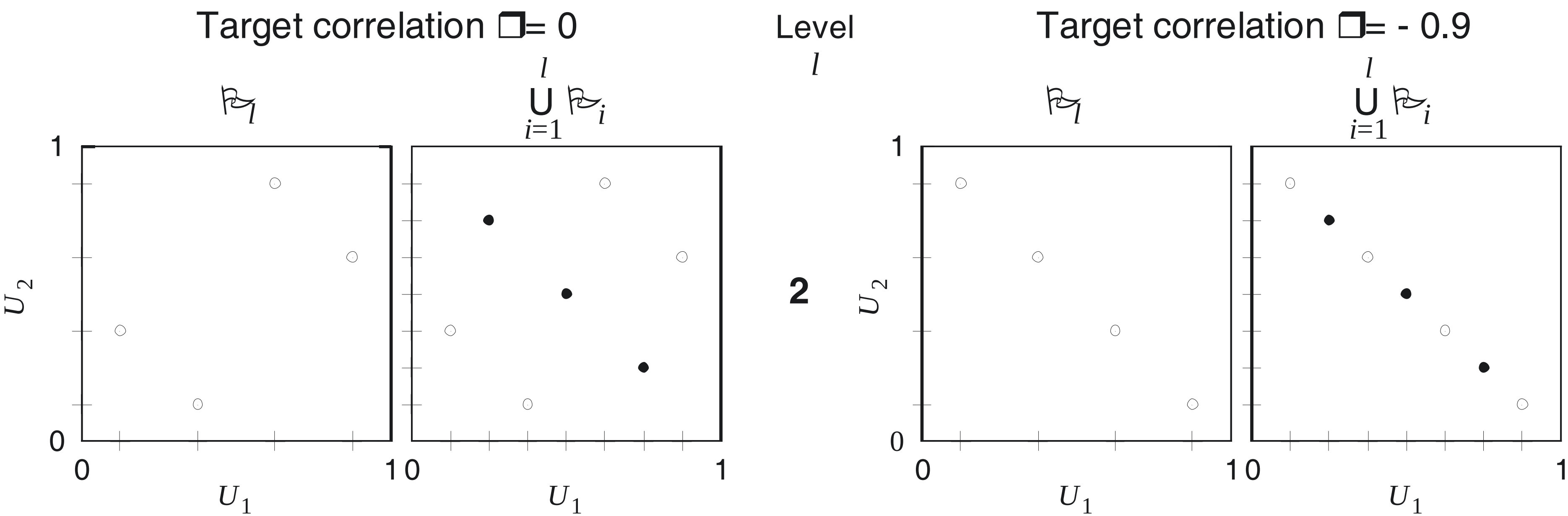
example: uniform distribution $\mathbf{U}=\{U_1,U_2\}: \quad U_i \sim R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

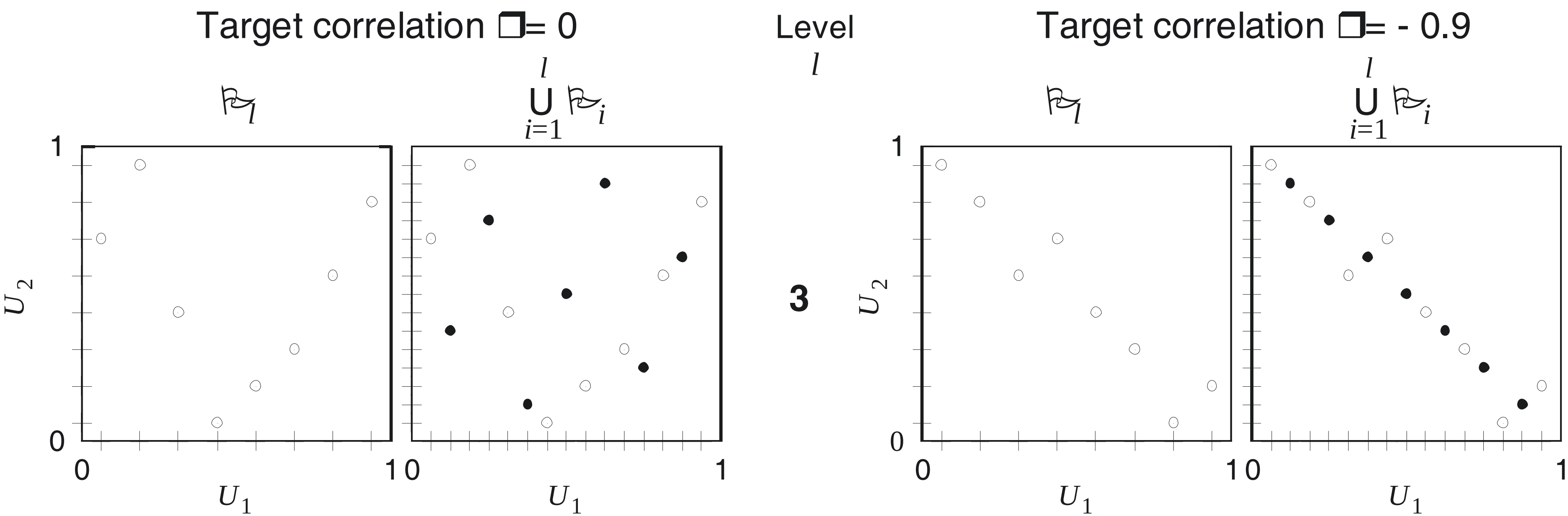
example: uniform distribution $\mathbf{U} = \{U_1, U_2\}: \quad U_i \sim R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

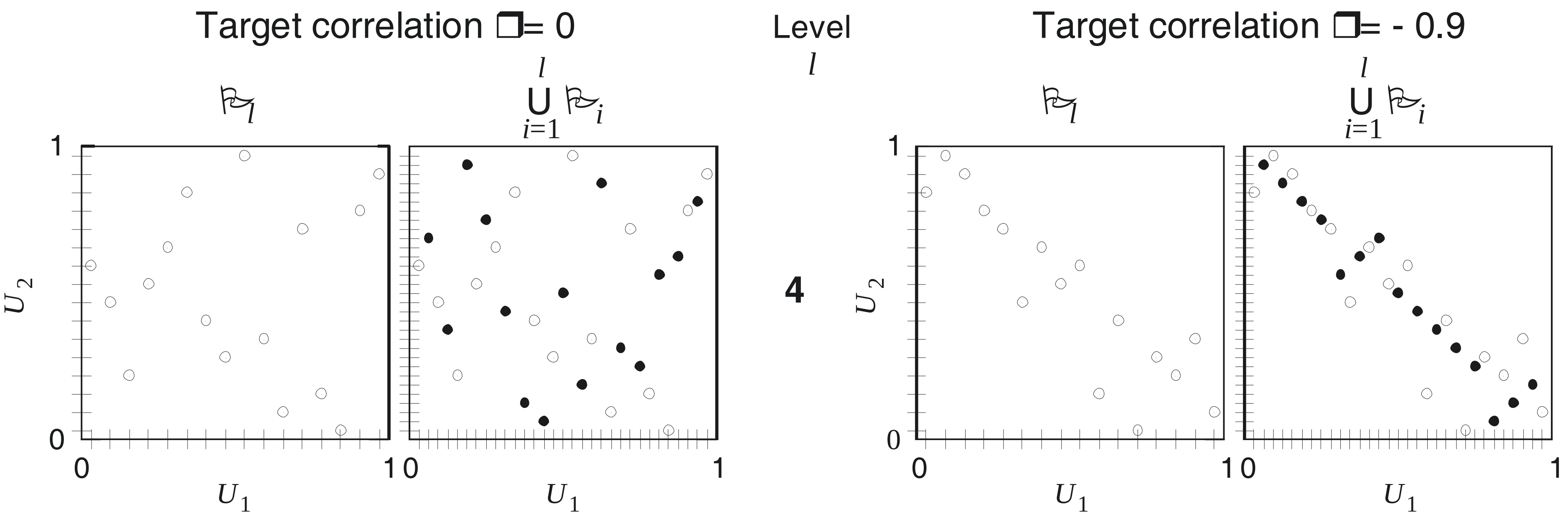
example: uniform distribution $\mathbf{U} = \{U_1, U_2\}: \quad U_i \sim R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

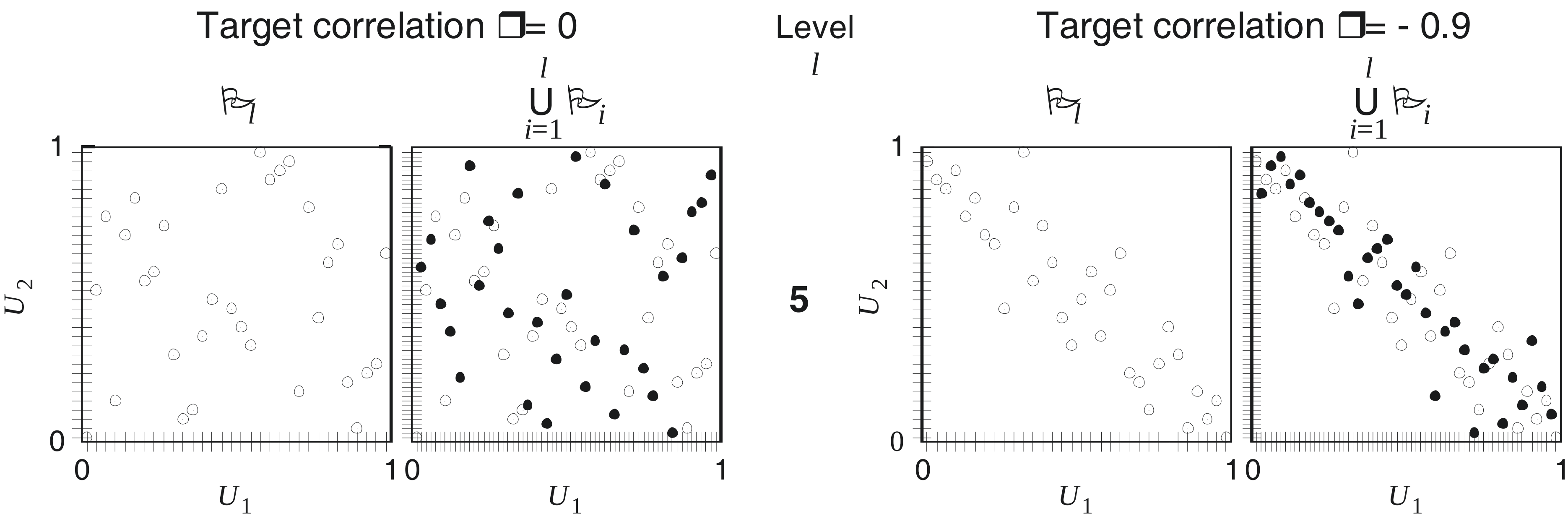
example: uniform distribution $\mathbf{U} = \{U_1, U_2\}: \quad U_i \,\square\, R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

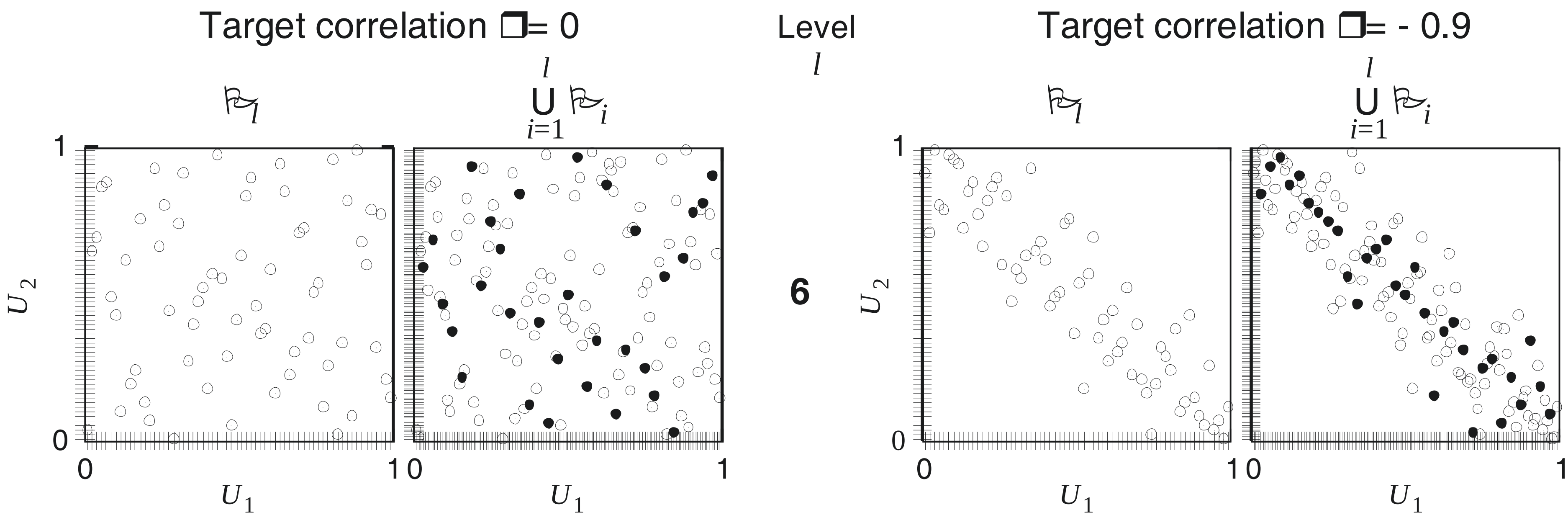
example: uniform distribution $\mathbf{U} = \{U_1, U_2\}: \quad U_i \sim R(0,1)$



(e.g. sampling probabilities)

HSLHS for a random vector

example: uniform distribution $\mathbf{U} = \{U_1, U_2\}: \quad U_i \square R(0,1)$

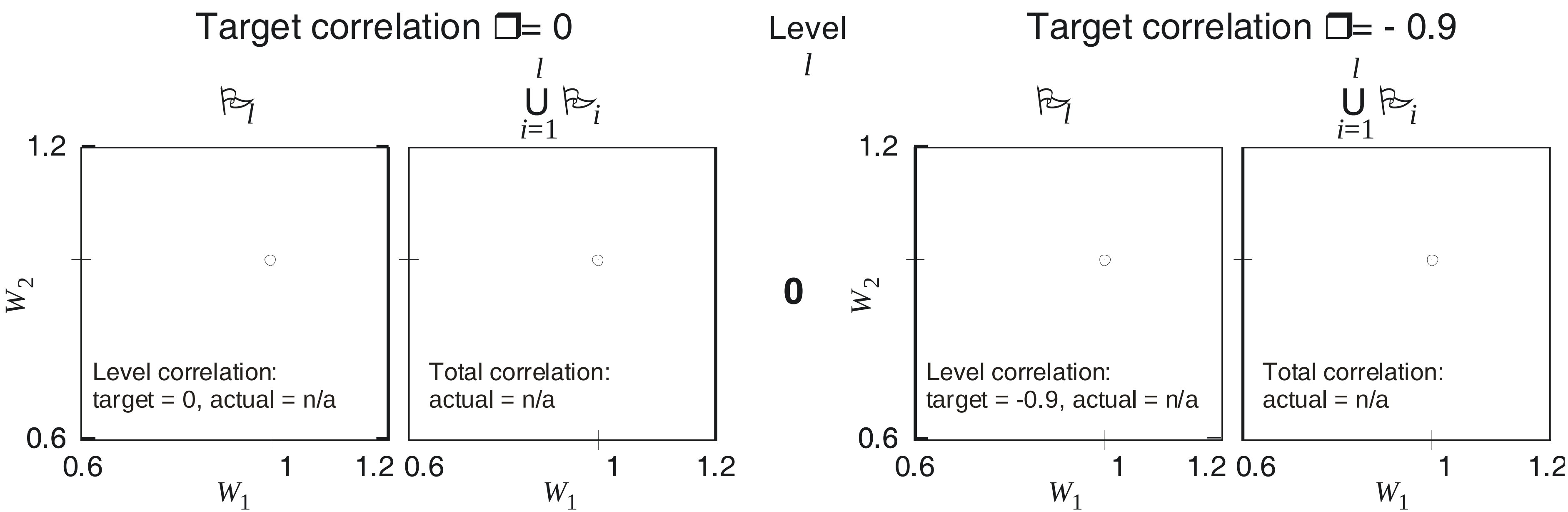


(e.g. sampling probabilities)

HSLHS for a random vector

correlation convergence

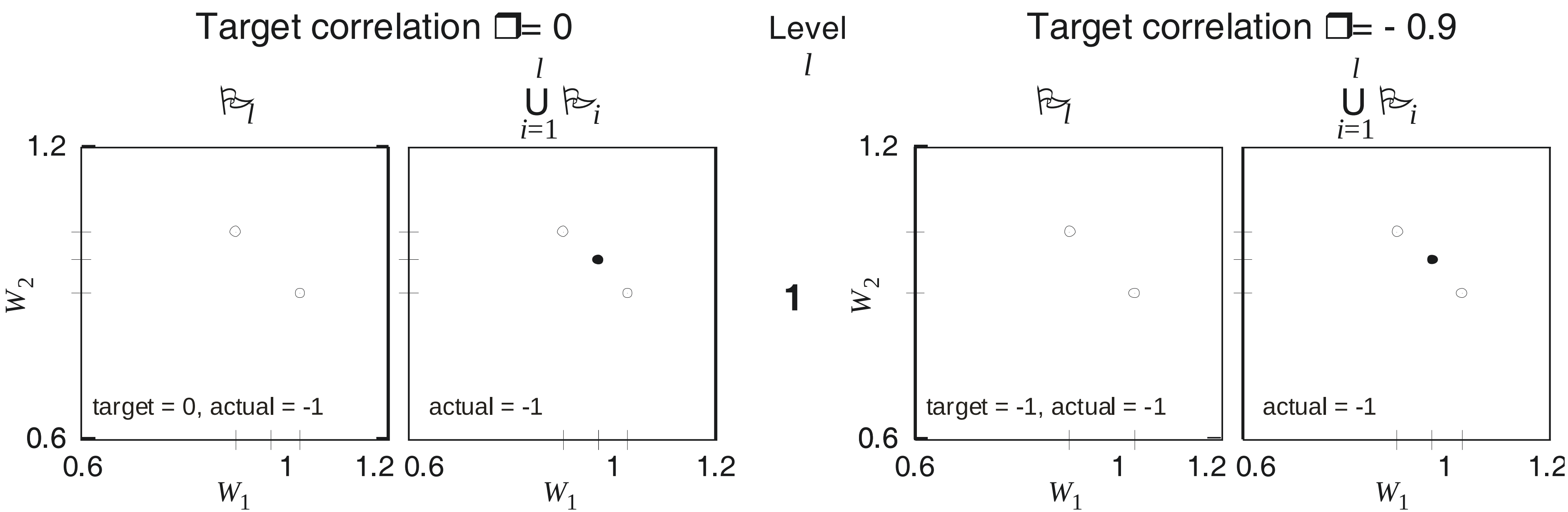
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m,s)$



HSLHS for a random vector

correlation convergence

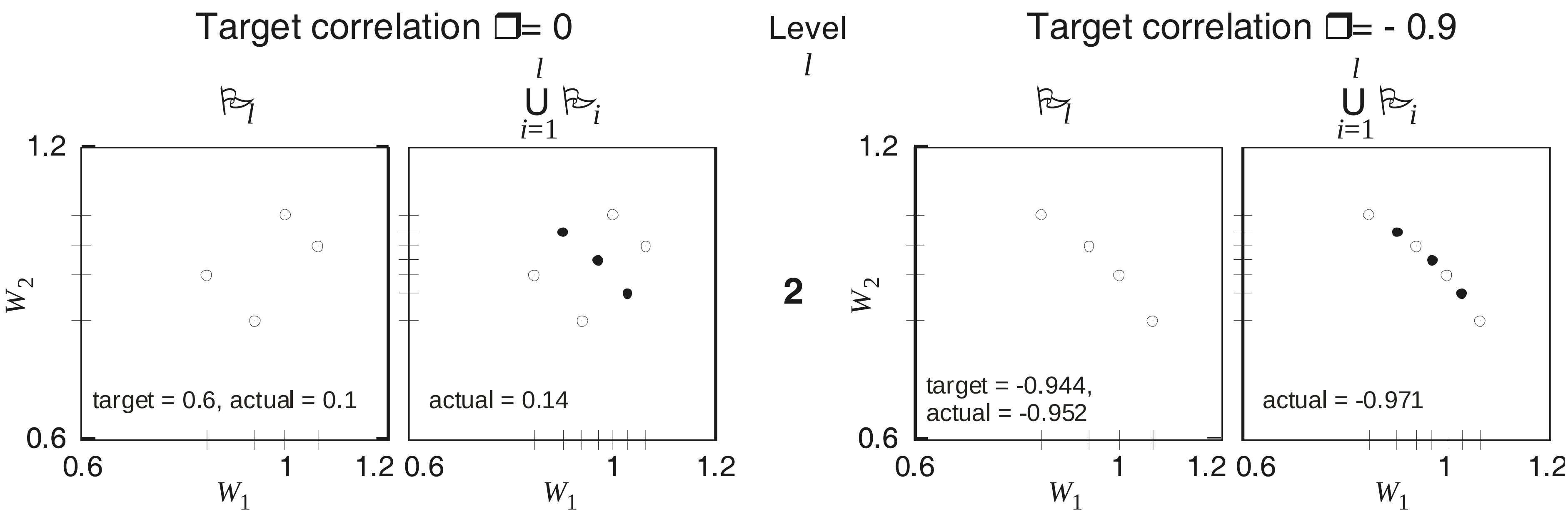
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m, s)$



HSLHS for a random vector

correlation convergence

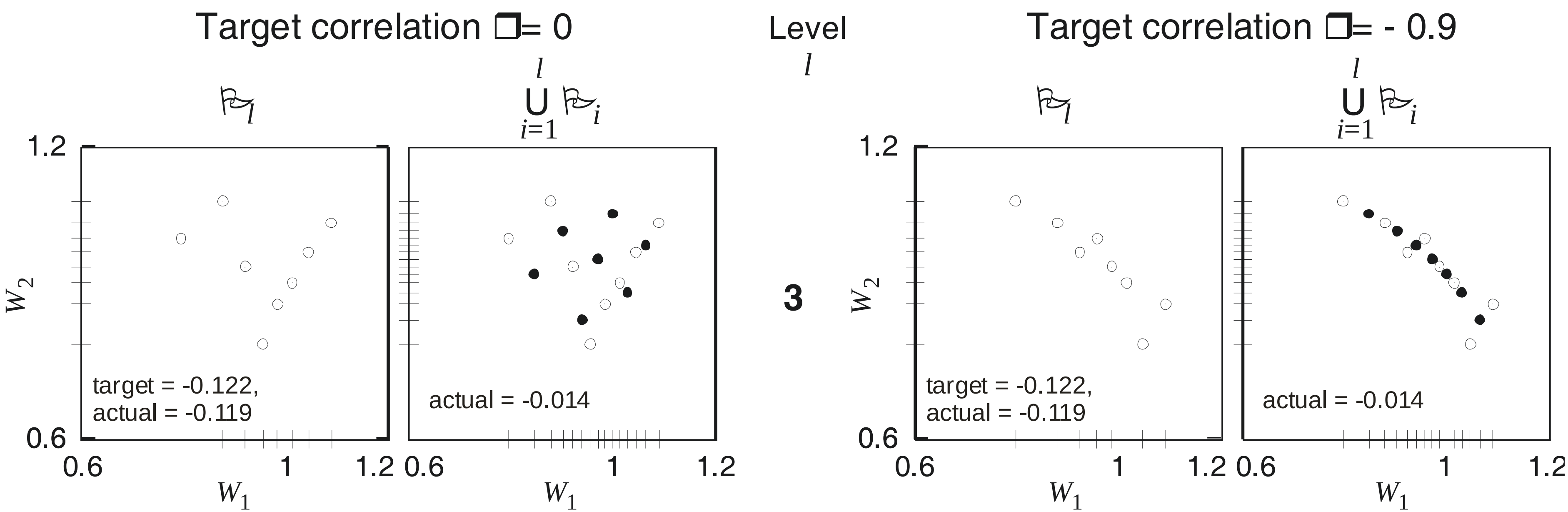
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m, s)$



HSLHS for a random vector

correlation convergence

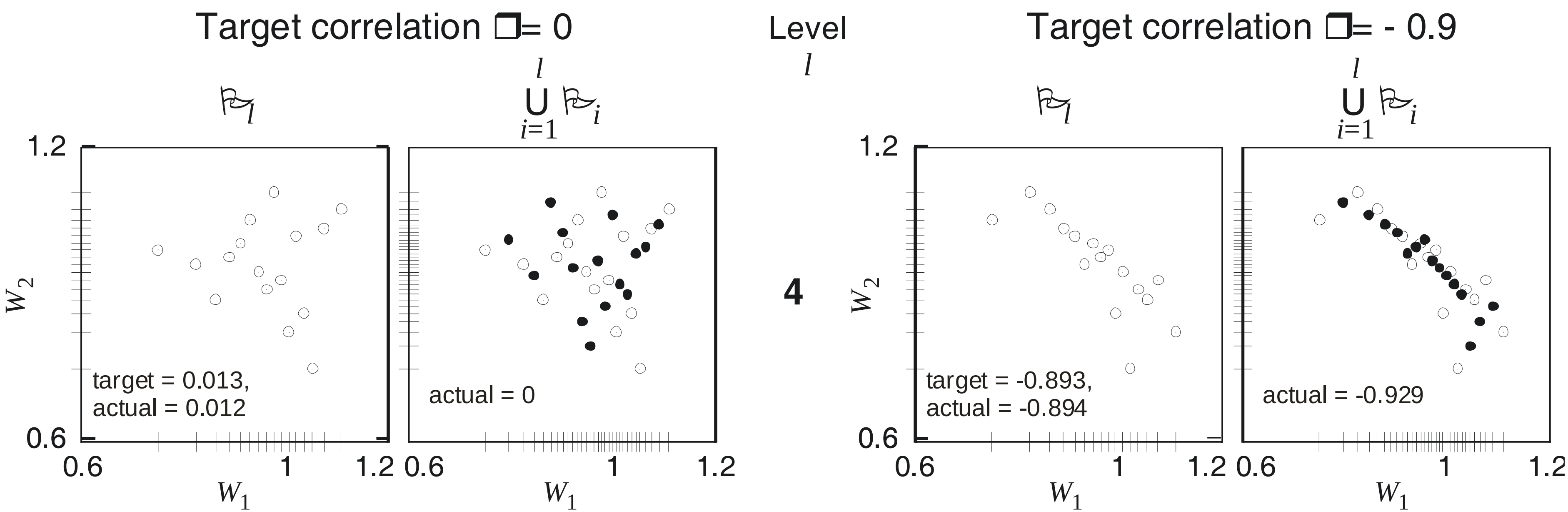
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m,s)$



HSLHS for a random vector

correlation convergence

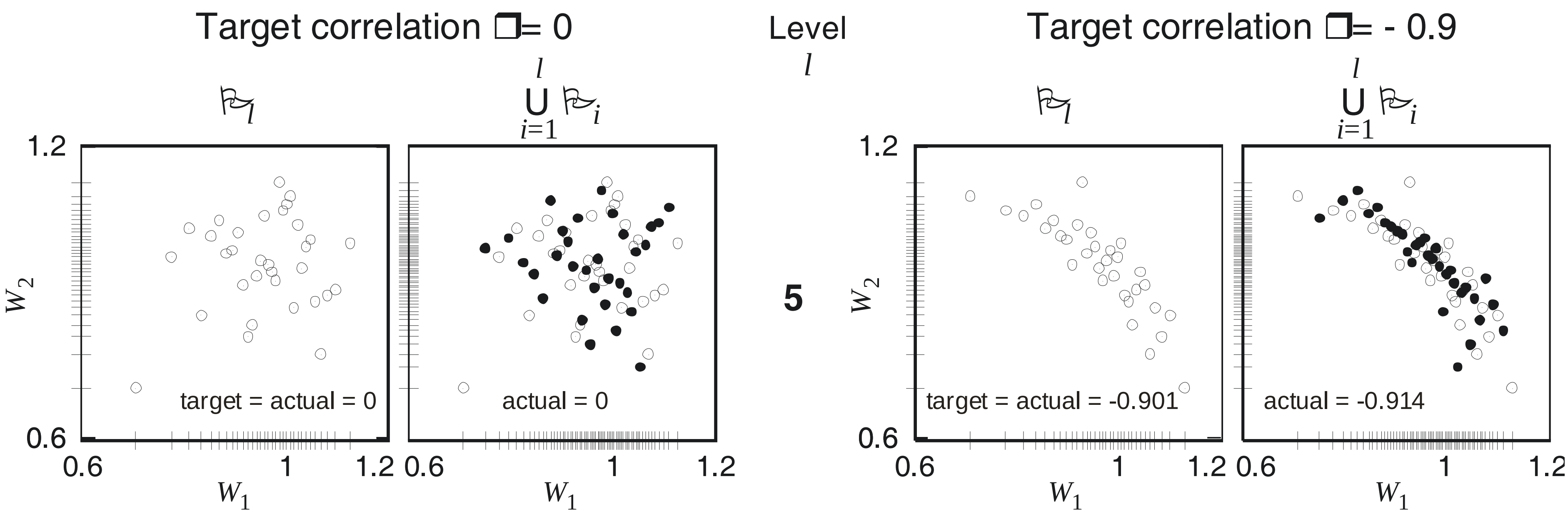
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m, s)$



HSLHS for a random vector

correlation convergence

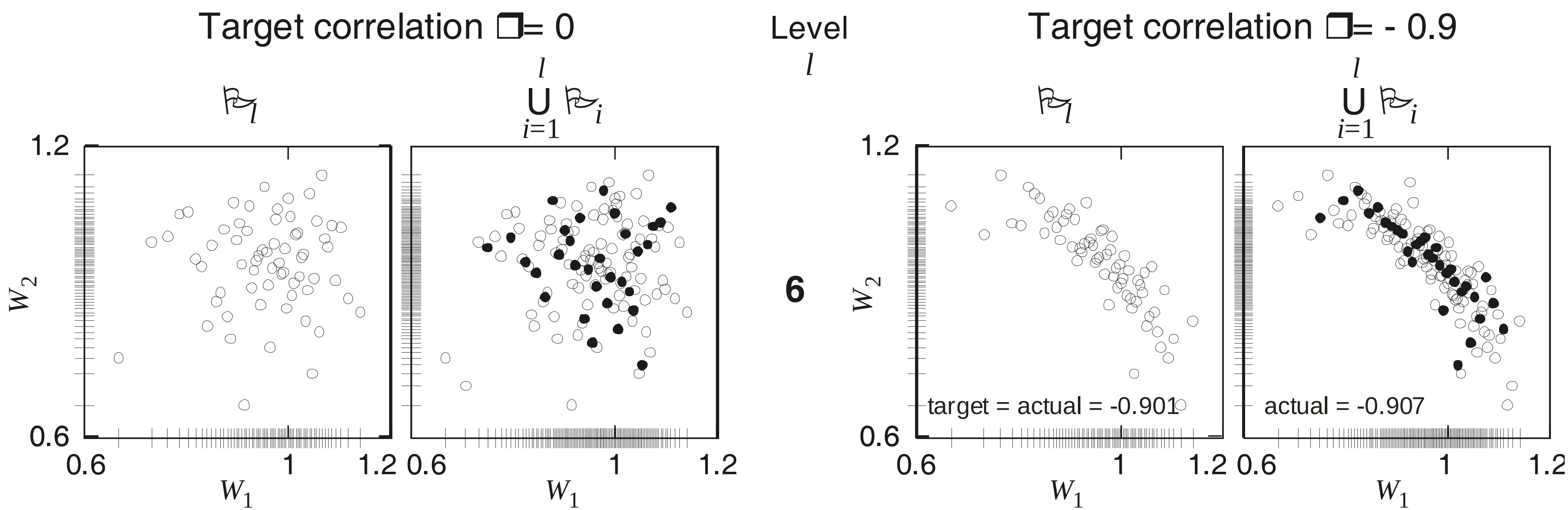
example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m,s)$



HSLHS for a random vector

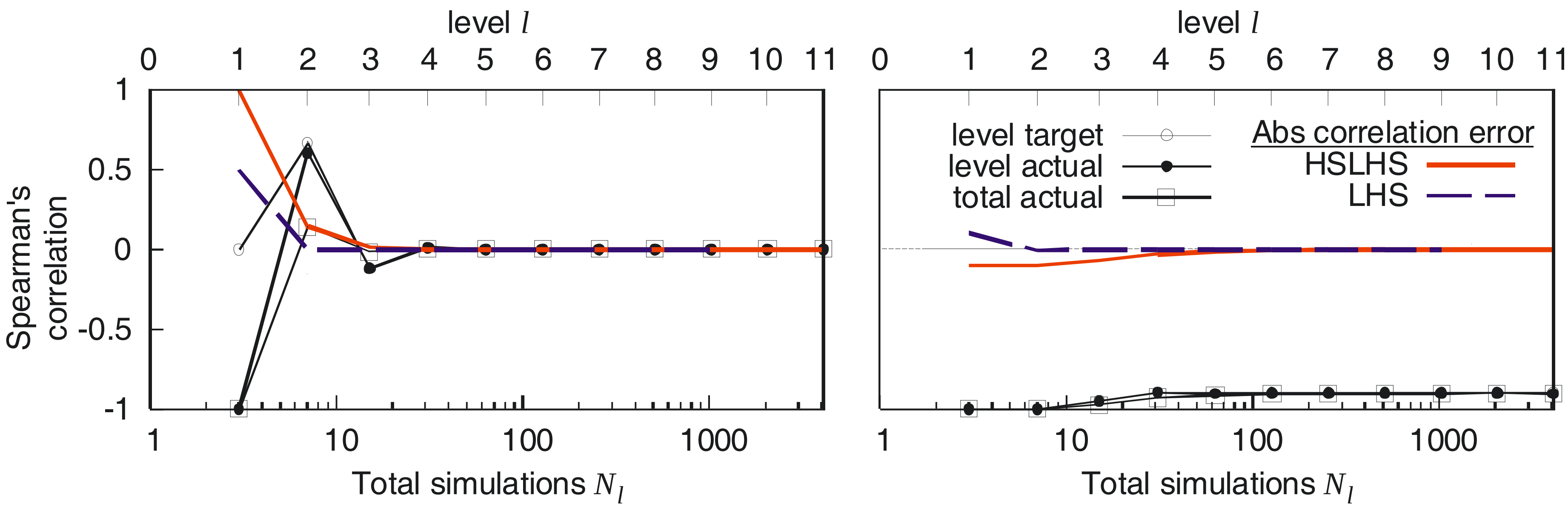
correlation convergence

example: Weibull distribution $\mathbf{W} = \{W_1, W_2\}: \quad W_i \sim W(m, s)$



HSLHS for a random vector

correlation convergence $\mathbf{A} \rightarrow \mathbf{T}$



Numerical examples: three functions of rvect.

- Three functions of a random vector (corr 0 and – 0.9)
  - Features three frequent transformations: addition, multiplication and minima (extremes)

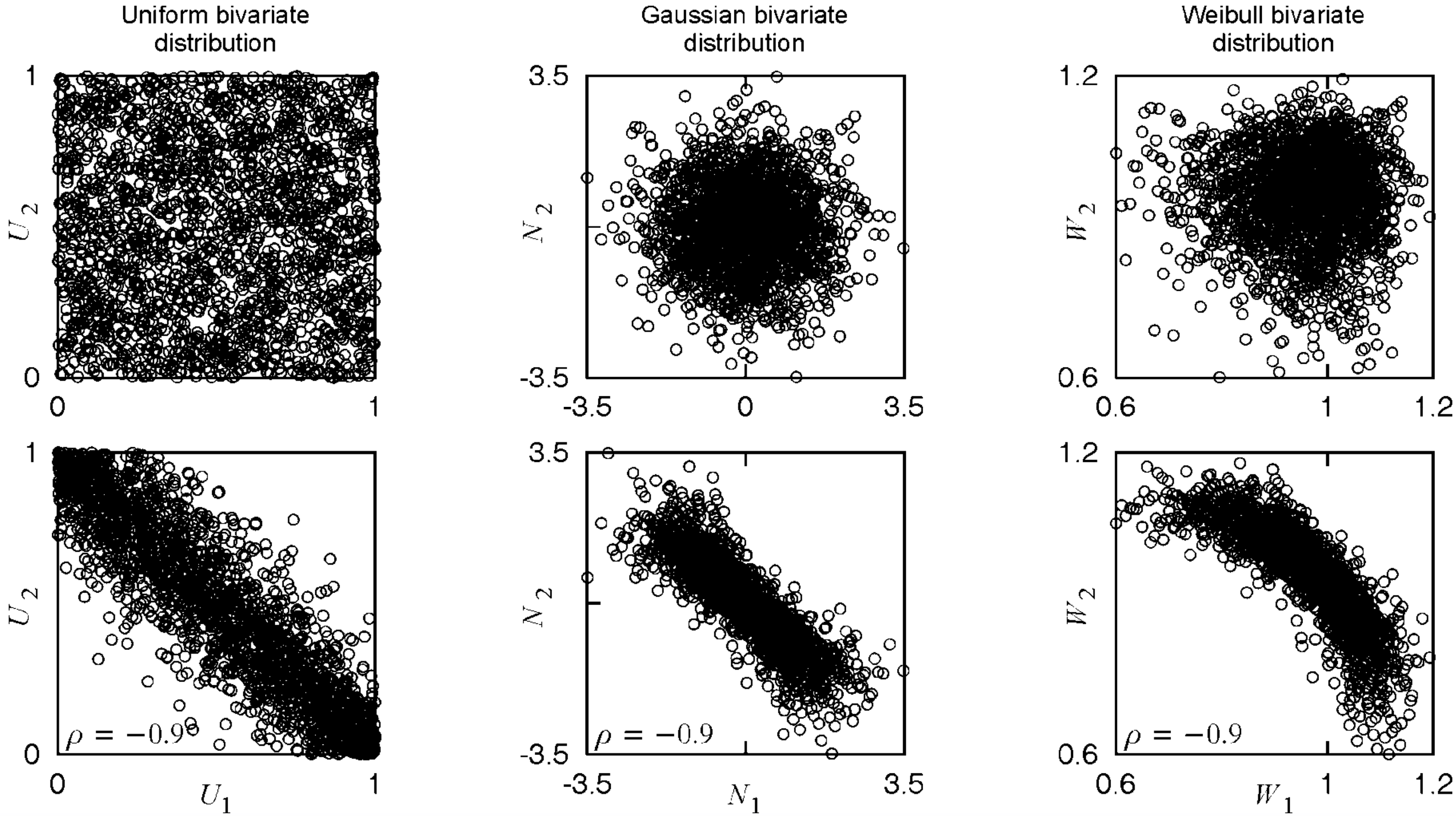
$$Z_1 = g_1(X, Y) = X + Y \qquad X, Y \sim N(0,1)$$

$$Z_2 = g_2(X, Y) = XY \qquad X, Y \sim N(0,1)$$

$$Z_3 = g_3(X, Y) = \min(X, Y) \qquad X, Y \sim W(m, s)$$

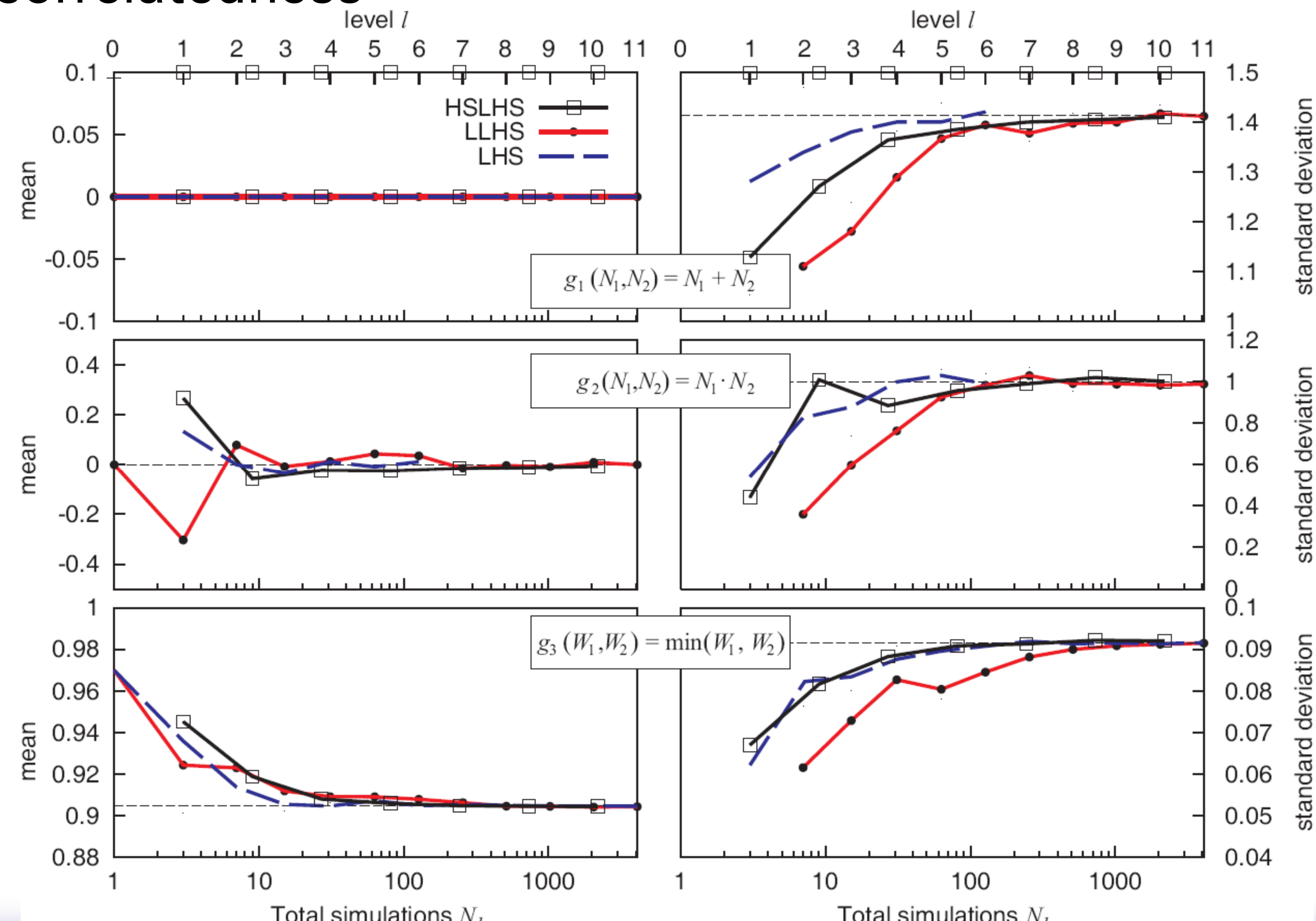
Numerical examples: three functions of rvect.

- Three functions of a random vector (corr 0 and −0.9)
  - Features three frequent transformations



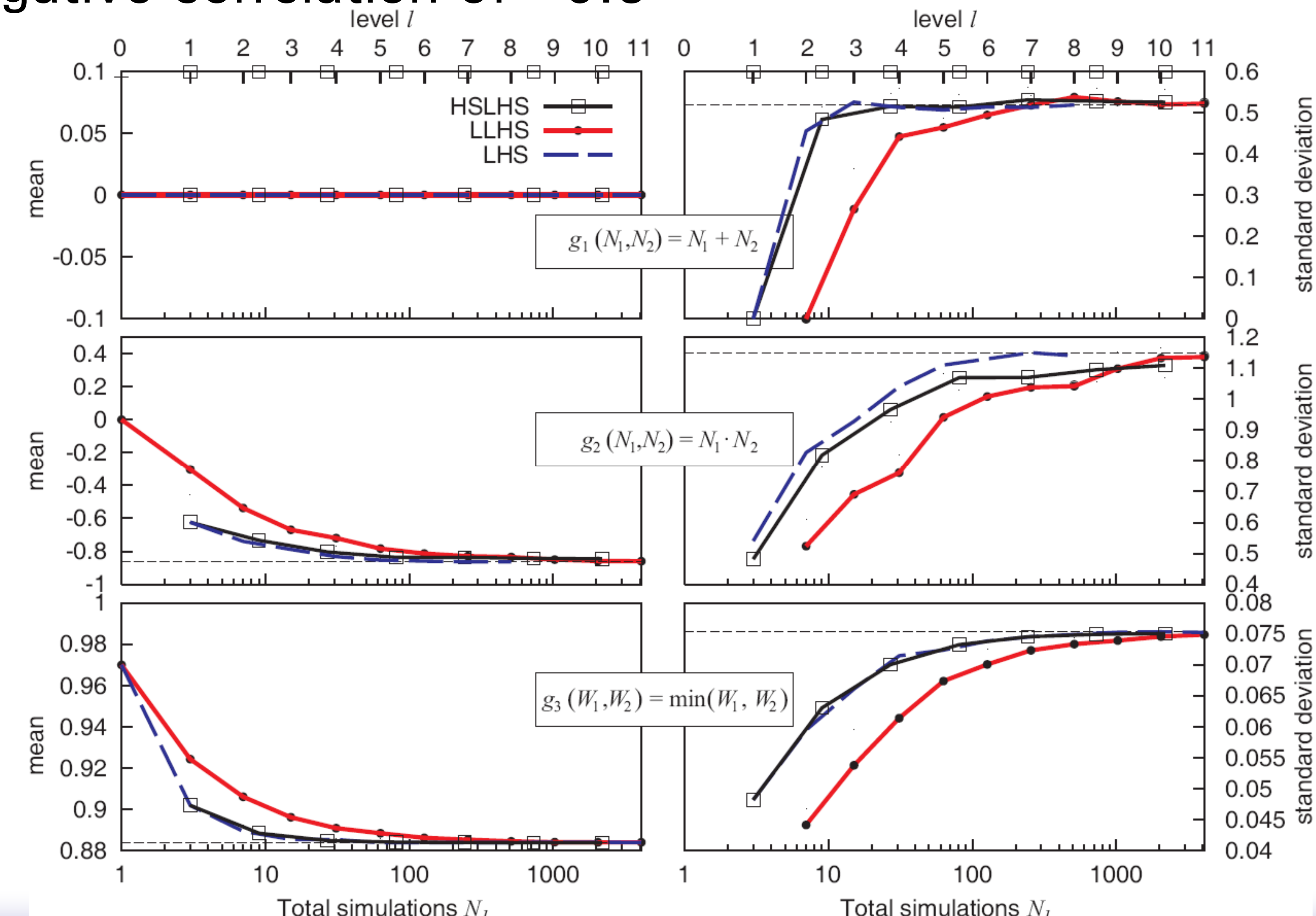
Numerical examples: three functions of rvect.

- Uncorrelatedness



Numerical examples: three functions of rvect.

- Negative correlation of −0.9



Conclusions

- Correlation control by combinatorial optimization can be used for sample extension within arbitrary version of MCS
- HSLHS for small sample sizes – subsets that can be combined into consistent sample sizes while retaining the variance reduction
- Adaptive refinement can be halted automatically:
  - User decision (e.g. material or computational resources)
  - Statistical significance of arbitrary parameter (intvrl estimator,..)
- Application areas: physical or numerical experiments, pilot studies, simulation of random fields, progressive learning of NN, response surface generation, adaptive importance sampling, FReET.