Implementation study of DISCOPOLIS

an algorithm for uniform sampling of metabolic flux distributions via iterative sequences of linear programs

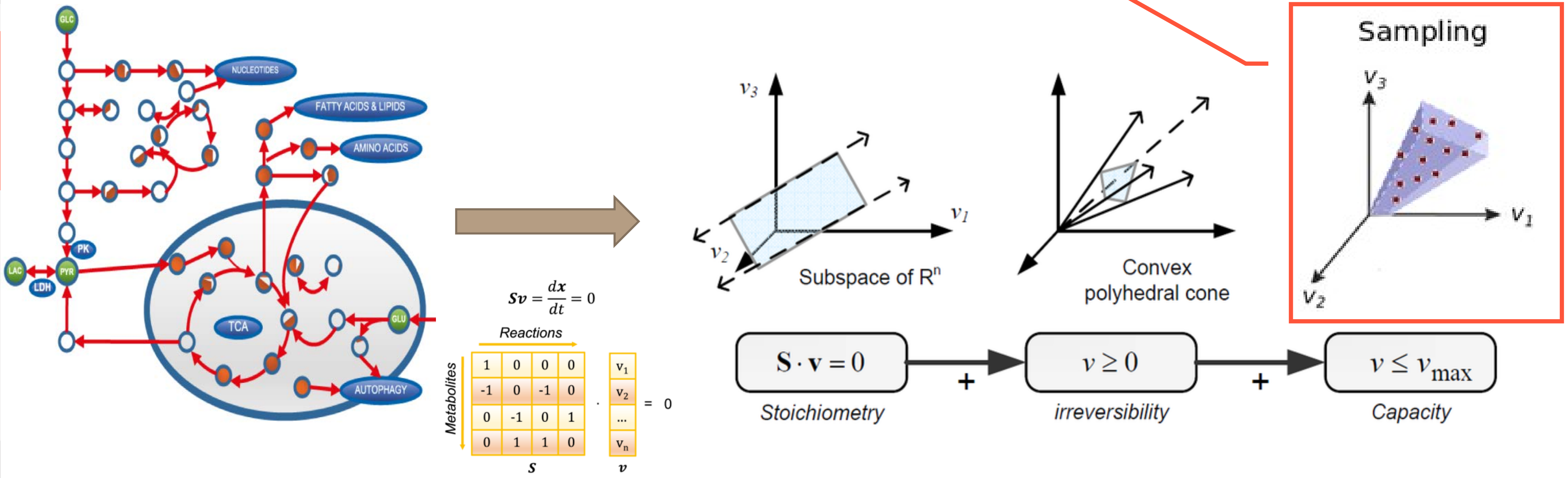
Name: Hongxing NIU

Promotor: Prof. Philippe Bogaerts

We are interested in metabolic networks, most of which are constrained under-determined systems

Problem

Solving constrained under-determined systems by random sampling



(source: http://2014.igem.org/Team:Valencia_UPV/Modeling/fba)

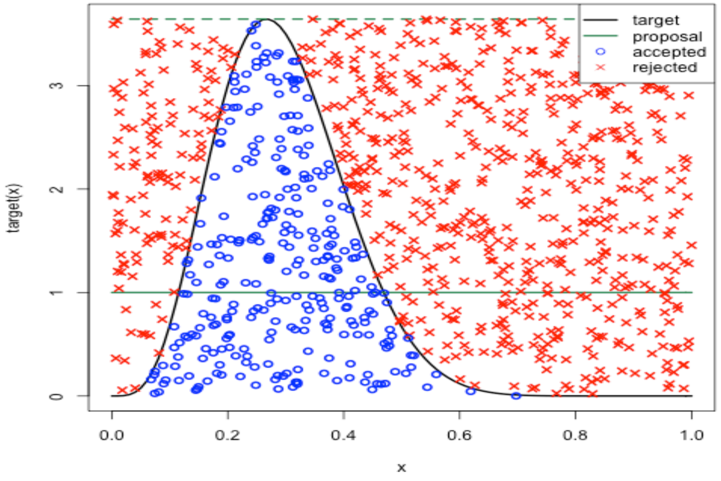
Approaches to the under-determined system by random sampling

Problem

Solving constrained under-determined systems by random sampling

(1). Acceptance-Rejection

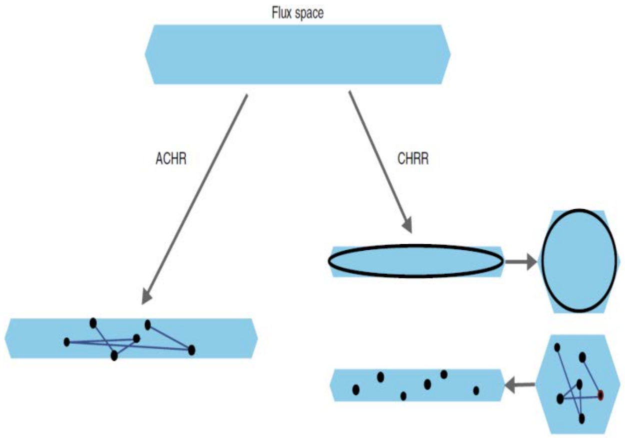
High computation load under high-dimen



(2). Hit-and-Run (MCMC, Metropolis-Hastings)

Getting stuck under high-dimen

(Schellendberger, 2011; Haraldsdottir, 2017)



(3). DISCOPOLIS

Discrete Sampling of COnvex POlytope via Linear program Iterative Sequences

(Bogaerts and Rooman, 2019)

Approaches to the under-determined system by random sampling

Problem

Solving constrained under-determined systems by random sampling

(3). DISCOPOLIS : Discrete Sampling of COnvex POlytope via Linear program Iterative Sequences (Bogaerts and Rooman, 2019).

Input : solution polytope defined by A and b; number of samples N; number of grid points S; minimum and maximum values of the fluxes v_i^{MIN} and v_i^{MAX} ($i \in [1,n]$) obtained with Flux Variability Analysis

Output : N samples $v(k) \in \mathbb{R}^n$ ($k \in [1,N]$) with their weights $w(k)$

```
A_eq = ∅; b_eq = ∅;  /* initialize empty matrices for equality constraints
L_i = (v_i^MAX - v_i^MIN) / (S - 1);  /* compute for each flux v_i the interval
    between 2 grid points
for k = 1 to N do
  w(k) = 1;  /* initialize weight of the k^th sample
  I = [1,n];  /* set of all indexes i of all the fluxes v_i ∈ v
  Randomly select an index i in I;
  Remove index i from set I;
  Generate one index g from a uniform distribution on [1,S];
  v_i = v_i^MIN + (v_i^MAX - v_i^MIN) * (g - 1) / (S - 1);  /* discrete uniform
      sampling of v_i corresponding to the g^th grid point
  while I ≠ ∅ do
    Augment A_eq and b_eq to account for last fixed v_i;
    Randomly select an index i in I;
    Remove index i from set I;
    v_i^MINnew = min_v v_i computed with LP subject to A*v ≤ b
        and A_eq*v = b_eq;
    v_i^MAXnew = max_v v_i computed with LP subject to A*v ≤ b
        and A_eq*v = b_eq;
    S^new = 1 + floor ((v_i^MAXnew - v_i^MINnew) / L_i);  /* number of grid
        points remaining in the new constrained solution interval
    if S^new > 1 then
      Generate one index g from a uniform distribution on [1,S^new];
      v_i = v_i^MINnew + (v_i^MAXnew - v_i^MINnew) * (g - 1) / (S^new - 1);  /* discrete
          uniform sampling of v_i corresponding to the g^th grid point
    else
      v_i = (v_i^MAXnew + v_i^MINnew) / 2;  /* use of the center of the new
          solution interval in case of only 1 remaining grid point
    end
    w(k) = w(k) * S^new / S;  /* update weight of the k^th sample
  end
end
```

→ *N* iterations (*nSample*)

→ Discretized with *S* grid points (*nGrid*)

Iteratively, discrete sampling of convex polytope over which the objective function is optimized:

$$v_{Min,Max}(i) = Min, Max[v(i)] \qquad \forall\, i \in [1,n]$$

Problem

Solving constrained under-determined systems by random sampling

Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings:
1)monitoring convergence;
2)getting a "fit" solution distribution by appropriate discretization.

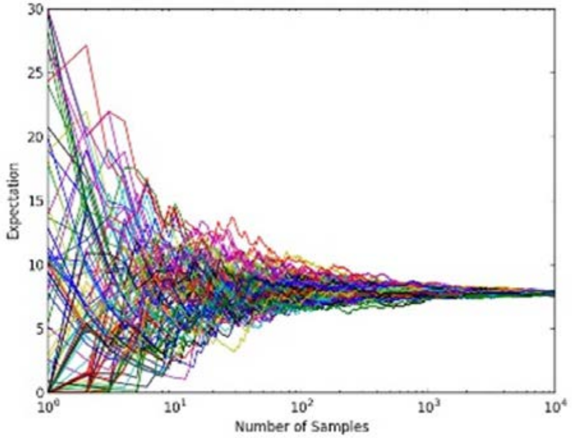
1) Monitoring convergence of averages (*nSample*)

----Univariate Gelman and Rubin diagnostic by
potential scale reduction factor (PSRF)

$$\text{PSRF} = \hat{R} = \sqrt{\frac{\widehat{var}(X)}{W}},$$

with

$$\widehat{var}(X) = \frac{N-1}{N}W + \frac{1}{N}B$$



----Multivariate extension (MPSRF)

$$\text{MPSRF} = \hat{R}^n = \sqrt{\frac{N-1}{N} + \frac{\lambda_{max}(\boldsymbol{W}^{-1}\boldsymbol{B})}{N}},$$

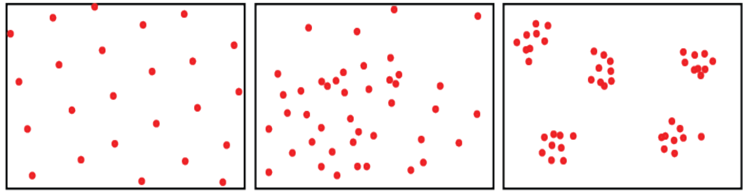
$\lambda_{max}()$ is the largest eigenvalue

2) Choosing an appropriate number of grid points (*nGrid*)

----Generalized variance (the determinant of covariance matrix $\boldsymbol{S}$ of fluxes)

----Total sample variance ($trace(\boldsymbol{S})$)

----P (99.9% weight)
(Bogaerts and Rooman, 2019 ,
the percentage of samples whose sum of weights =99.9% total sum of weights)



Uniform dispersion — Under-dispersed
Random dispersion
Clumped dispersion — Over-dispersed

Problem

Solving constrained under-determined systems by random sampling

Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings:
1)convergence;
2)discretized sampling

Case Studies

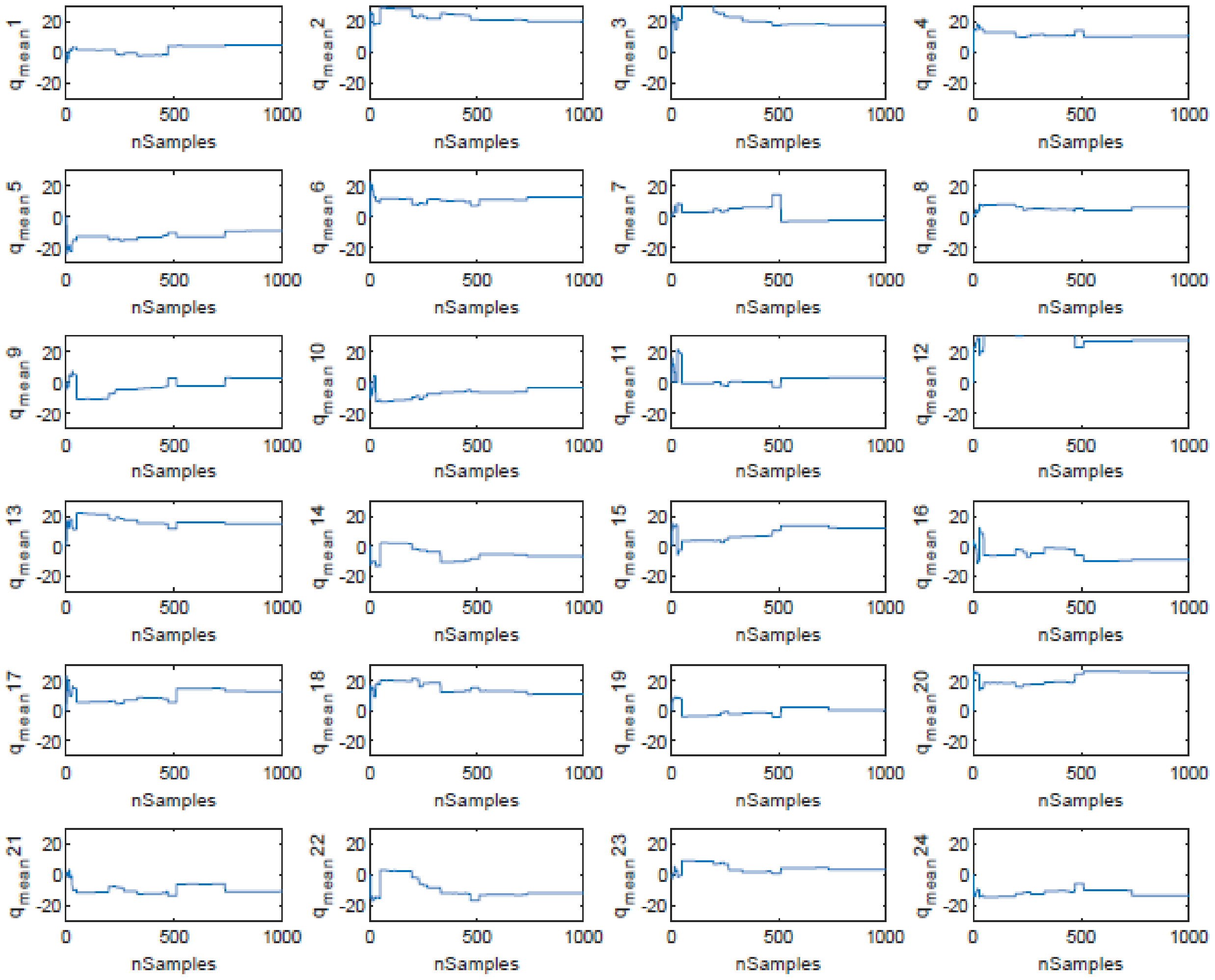
Toy example (extra slides);

Core metabolic network of *Escherichia coli* .

Core metabolic network of *E.coli*:

after elimination of equality constraints,

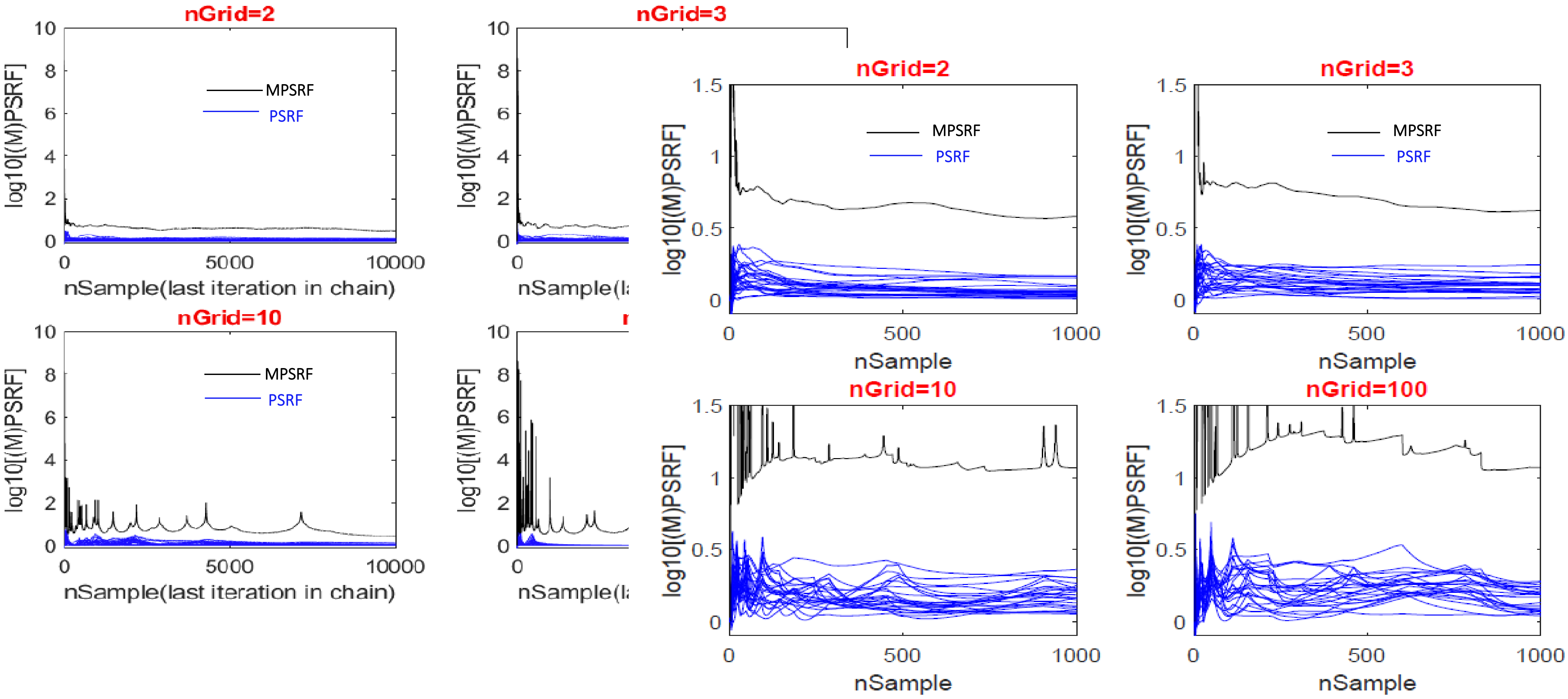
$\boldsymbol{A'q \le b'}$**, where** $\boldsymbol{A' \in \mathcal{R}^{172\times 24}}$



The flux means over *nSample* by the algorithm with *nGrid*=10

Convergence monitoring of means with changing *nGrid* w.r.t *nSample.*

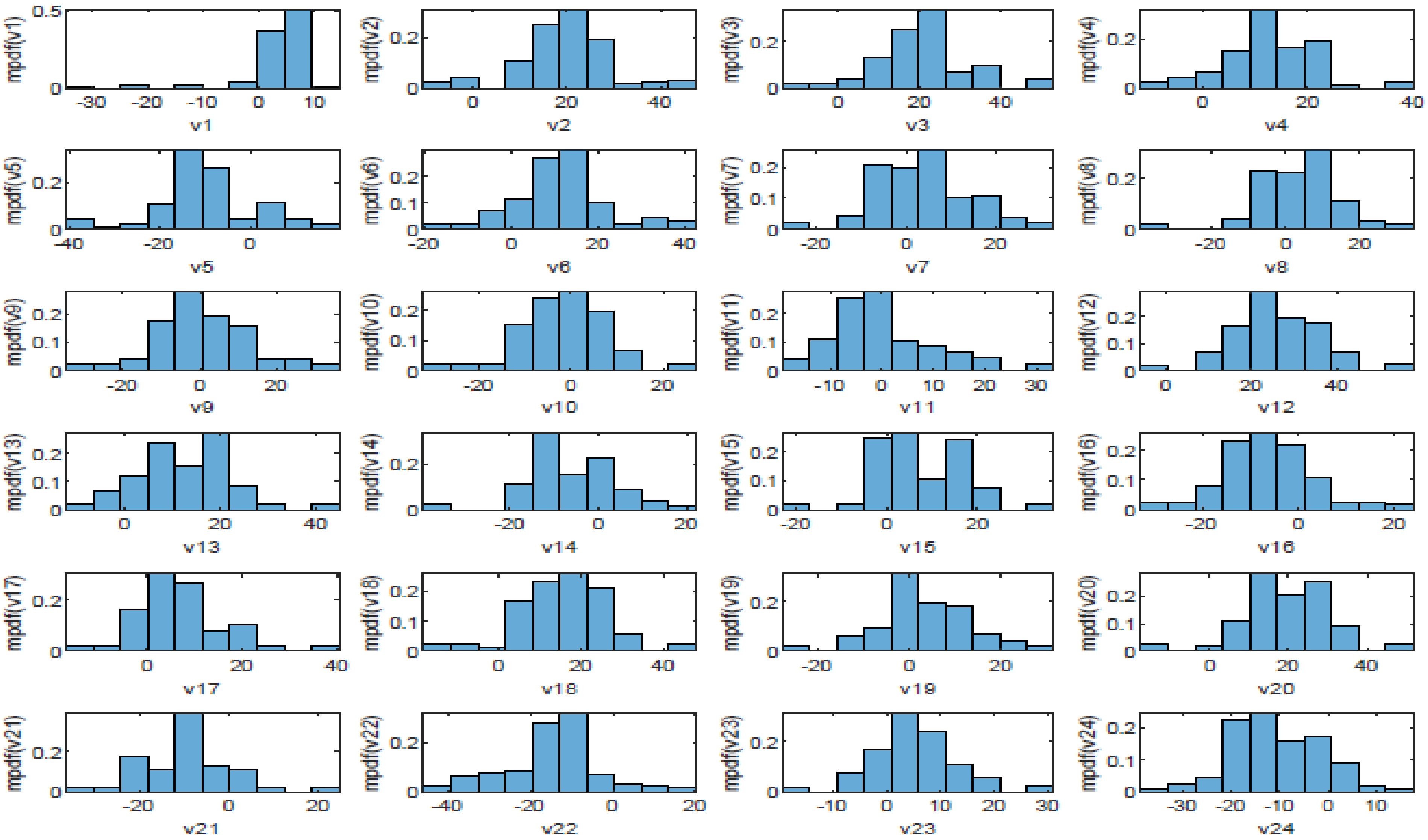
10 random seeds



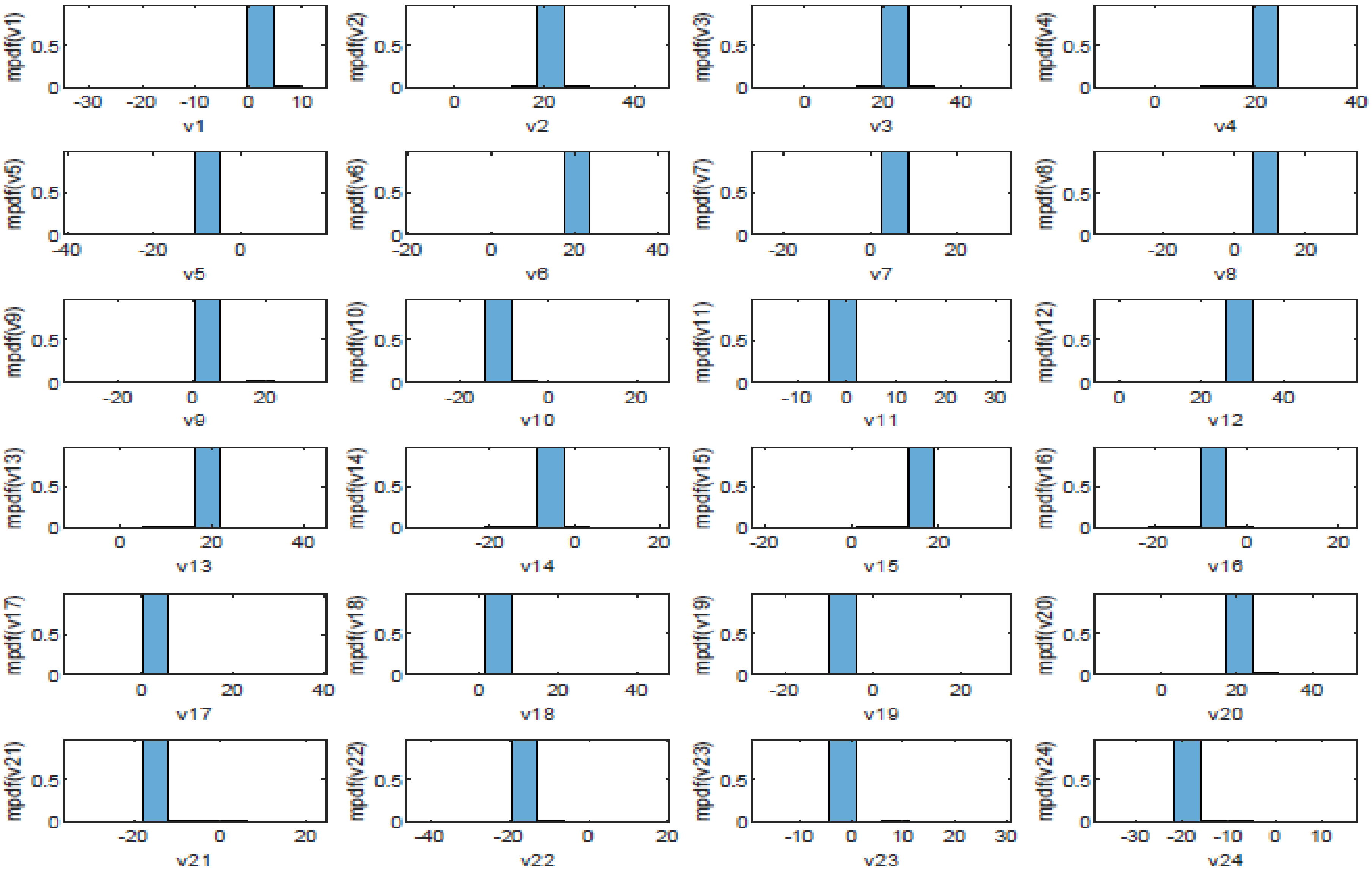
MPSRF>>*PSRF* in high-dimen, thus *MPSRF* is a conservative termination criterion.

Marginal distributions of the fluxes by the DISCOPOLIS algorithm, *nSample=10,000.*

(a) *nGrid*=2, the solutions could move to the tails

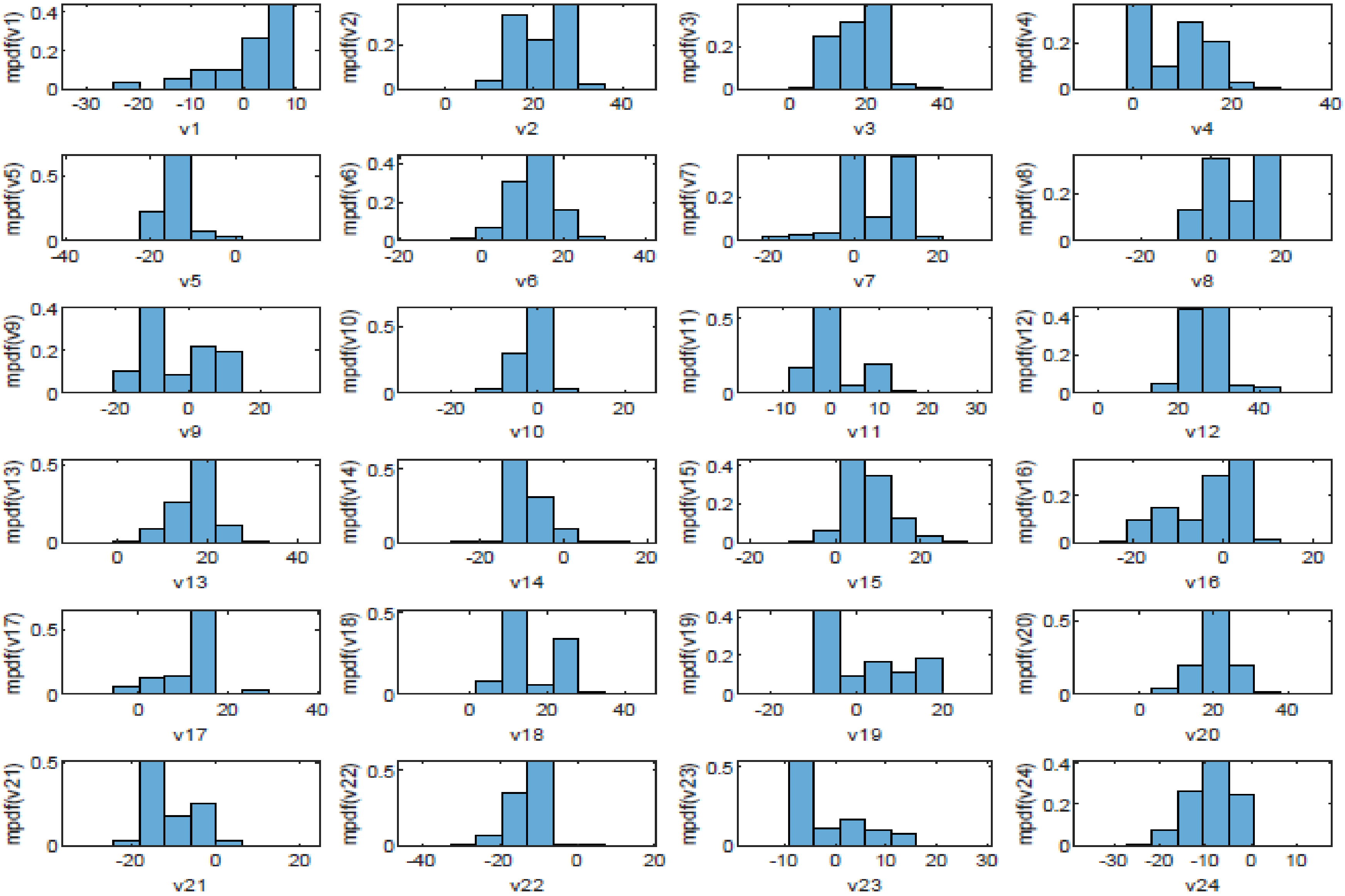


Marginal distributions of the fluxes by the DISCOPOLIS algorithm, *nSample=10,000.*
(b) *nGrid*=1000, under-dispersion, more regular than random

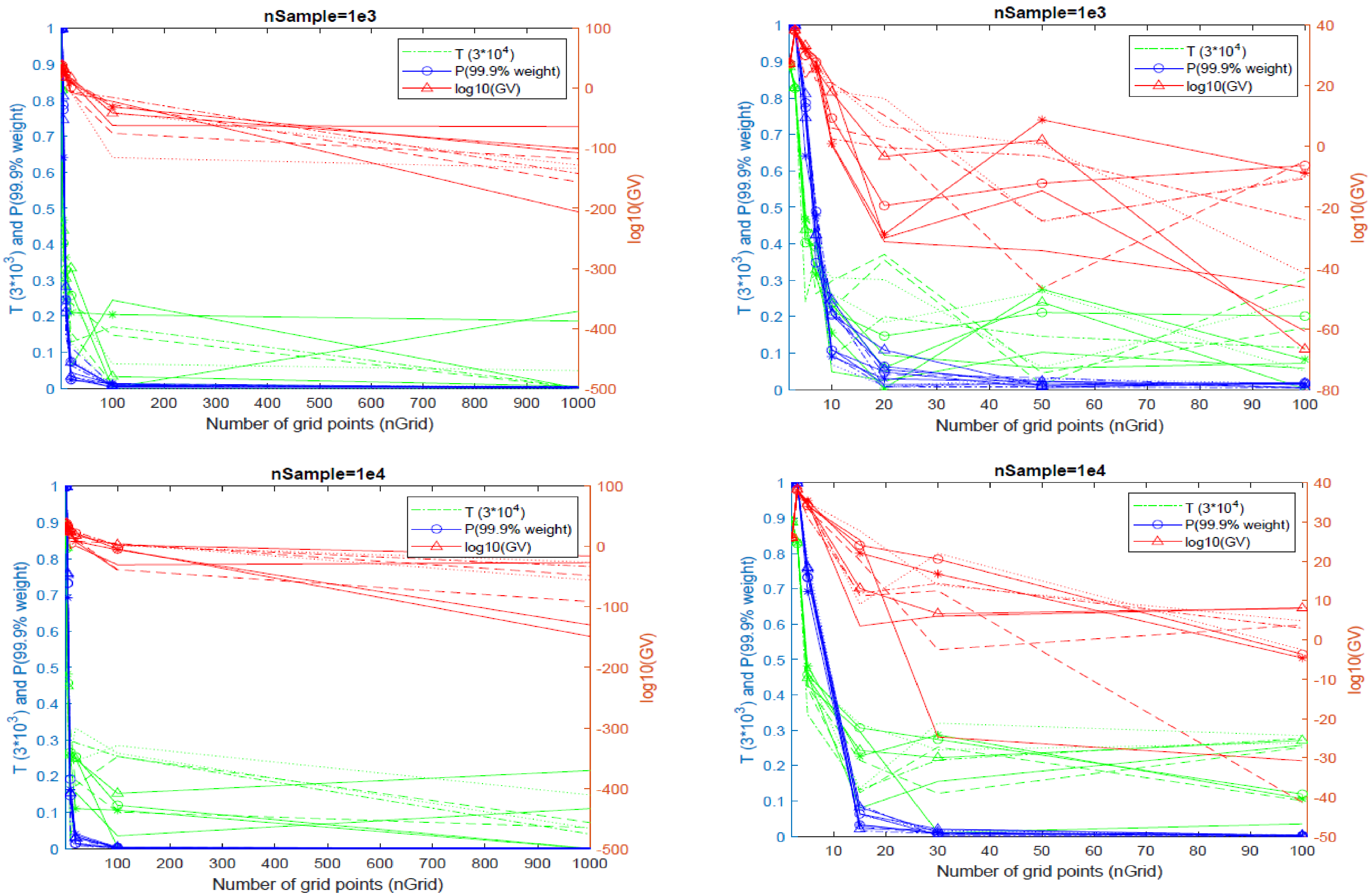


Marginal distributions of the fluxes by the DISCOPOLIS algorithm, *nSample=10,000.*

(c) *nGrid*=10



GV, T, and P (99.9% weight) are calculated and plotted against nGrid. 10 random seeds are taken under each scenario.



Summary

Problem

Solving constrained under-determined systems by random sampling

Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings:
1)convergence;
2)discretized sampling

Case Studies

1)Toy example;
2)Core metabolic network of *Escherichia coli*

Conclusions

The configuration of DISCOPOLIS algorithm, *i.e., nSamples + nGrid* were tuned to improve its performance

1) Gelman and Rubin diagnostic: the trend of *PSRF/MPSRF* from the $\overline{\boldsymbol{v}}^{T}$ tells if and when the results converge;

2)Two measures (*GV* and *T*) quantify the statistical dispersion of solution cloud: in the two cases the optimal *nGrid*=10~20 with a neither over-dispersed nor under-dispersed flux distribution.

THANK YOU

Extra Slides

Structure Overview

Problem

Solving constrained under-determined systems by random sampling

Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings: 1)convergence; 2)discretized sampling

Case Studies

Toy example (extra slides);

Core metabolic network of *Escherichia coli.*

Conclusions

The configuration of DISCOPOLIS algorithm, *i.e., nSamples + nGrid* were tuned to improve its performance

Structure Overview

Problem

Solving constrained under-determined systems by random sampling

Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings: 1)convergence; 2)discretized sampling

Case Studies

Toy example (extra slides);

Core metabolic network of *Escherichia coli.*

Conclusions

The configuration of DISCOPOLIS algorithm, *i.e., nSamples + nGrid* were tuned to improve its performance

1) Toy example: $A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 2 \end{bmatrix}$

Problem

Solving constrained under-determined systems by random sampling

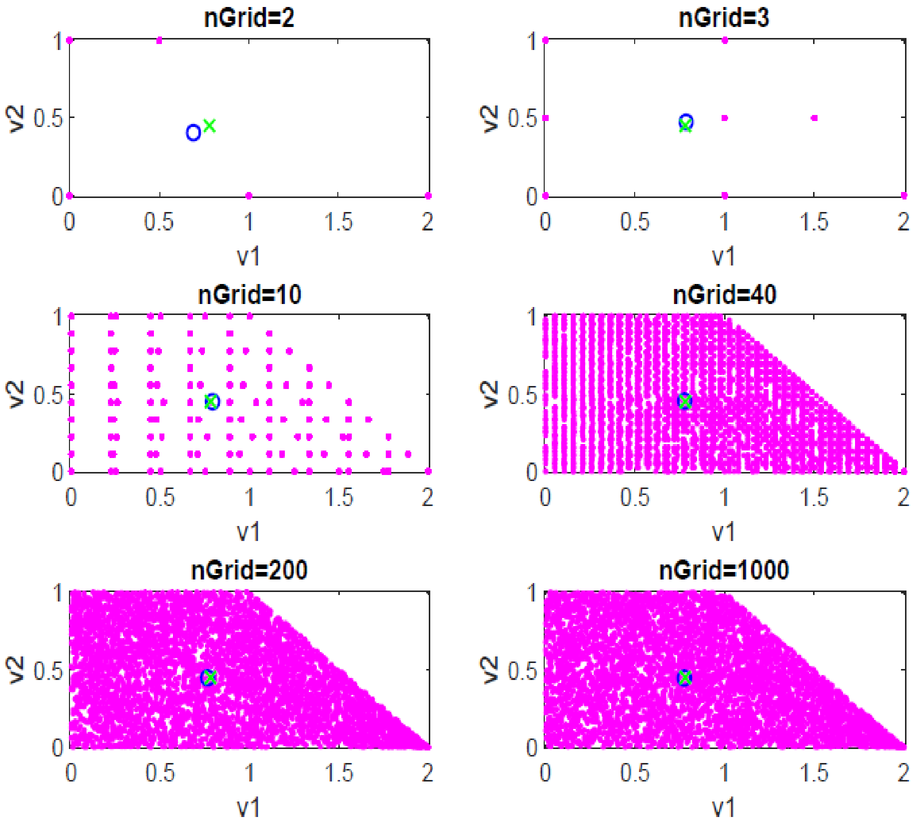
Objectives & Methods

Investigating the DISCOPOLIS algorithm and optimizing the parameter settings:
1)convergence;
2)discretized sampling

Case Studies

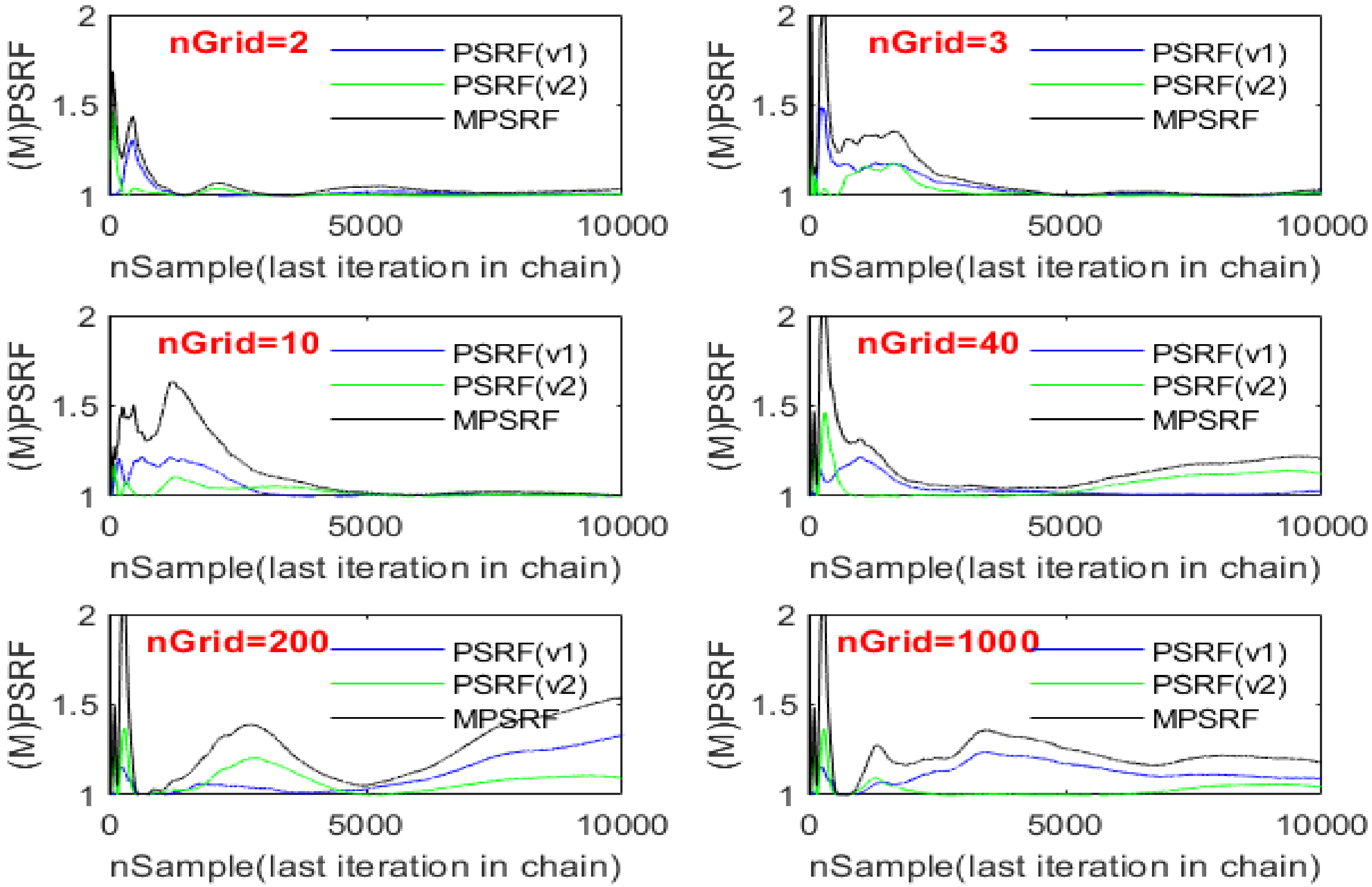
Toy example;

Core metabolic network of *Escherichia coli*

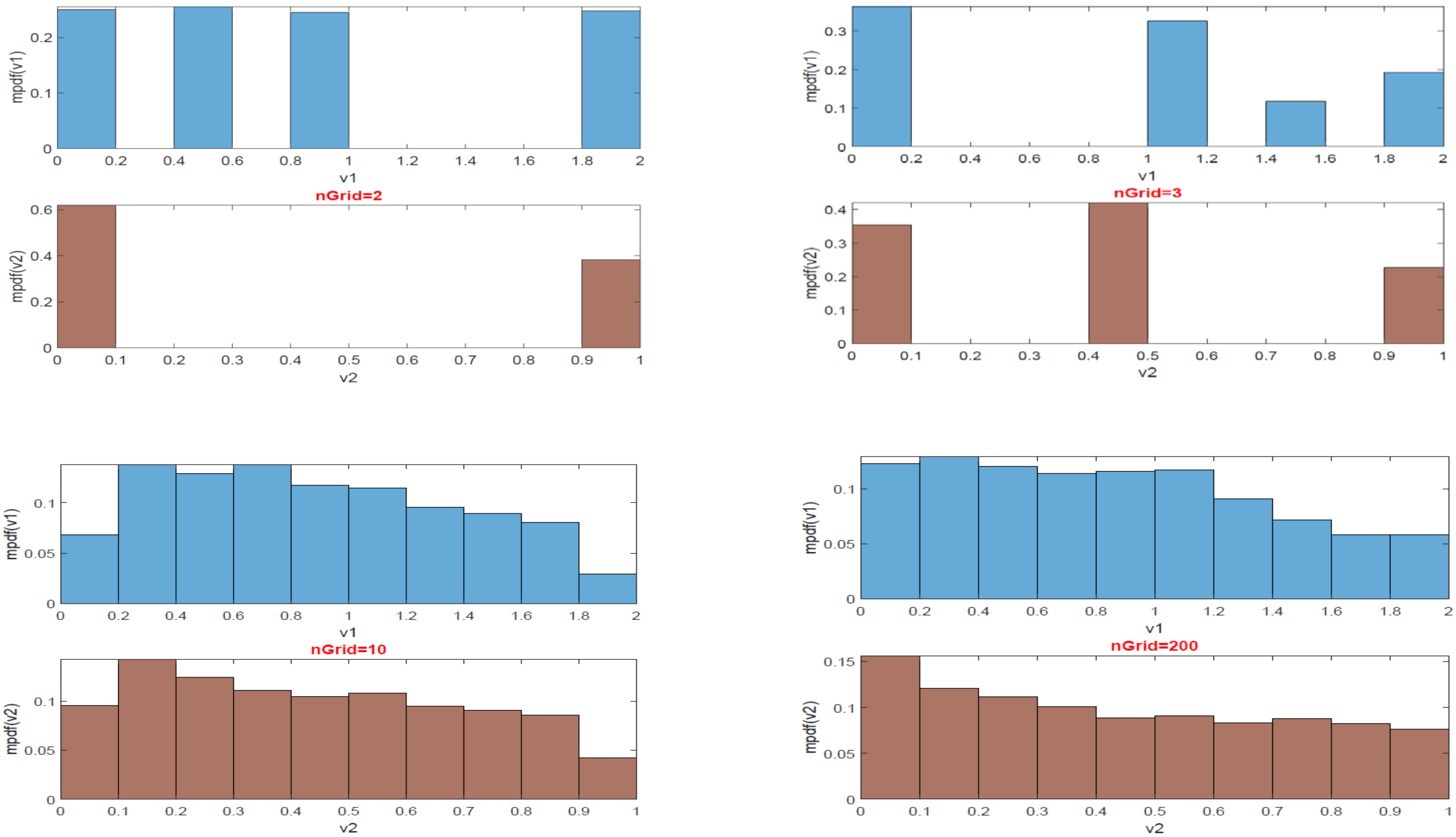


Solutions (dots) by DISCOPOLIS, *nSample*=5,000 , O =the means; $\bar{v}^T = [0.78 \;\; 0.45]$, by the rejection algorithm shown by ✖.

The sequences of PSRF and MPSRF indicate that they approach 1~1.1 over iterations (*nSample*=5,000) under scenarios of *nGrid*=2, 3 or10, showing fast convergence.

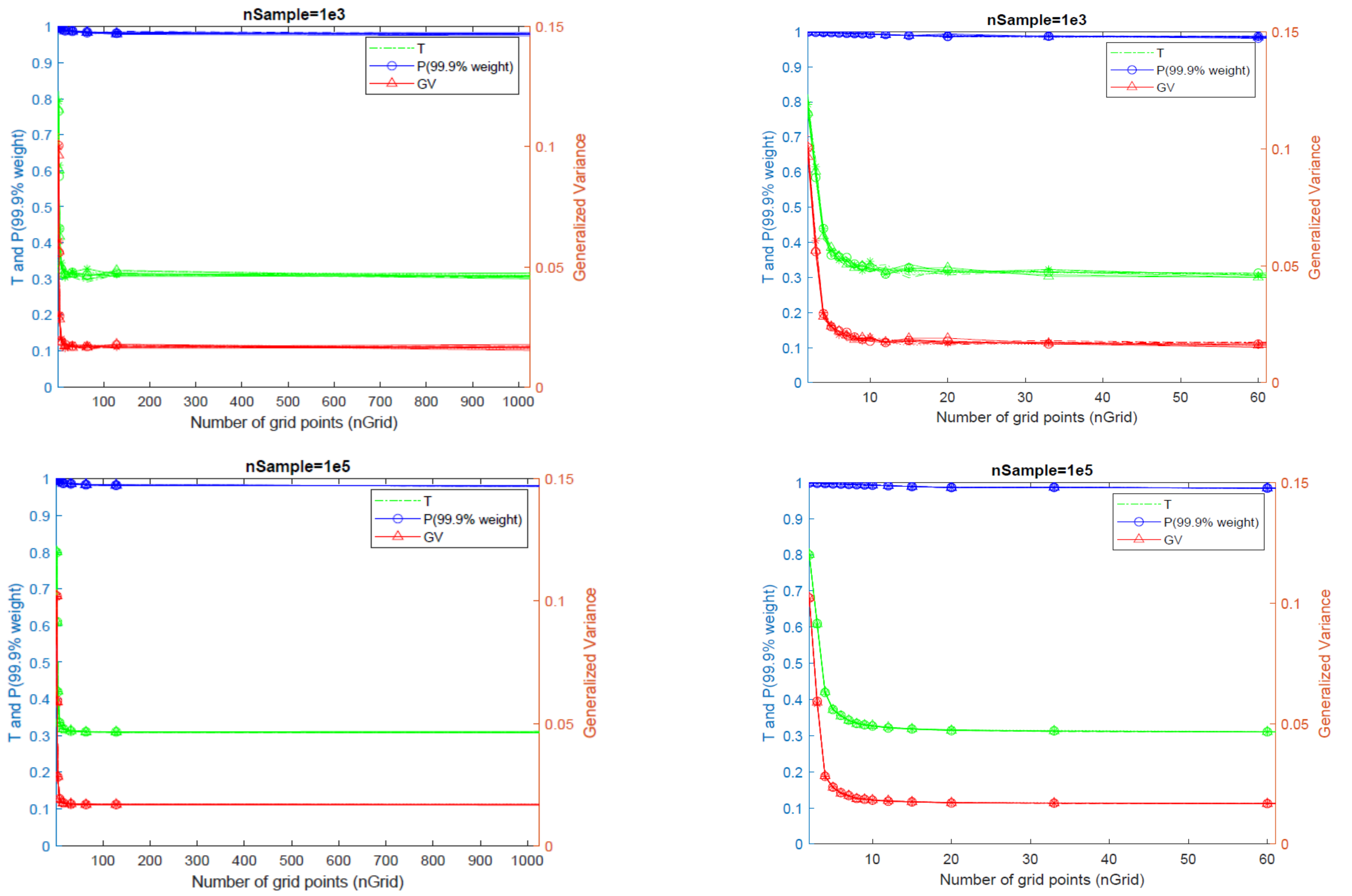


Marginal distributions of the fluxes of the toy example by the DISCOPOLIS algorithm with changing *nGrid*. *nSample*=5,000.



“Generalized variance” (GV) and “total sample variance” (T) are calculated and plotted against *nGrid* with changing *nSample*. P (99.9% weight) is also shown.

10 random seeds are taken under each scenario.



Flux mean values of the flux distribution obtained with optimal setting of the DISCOPOLIS algorithm