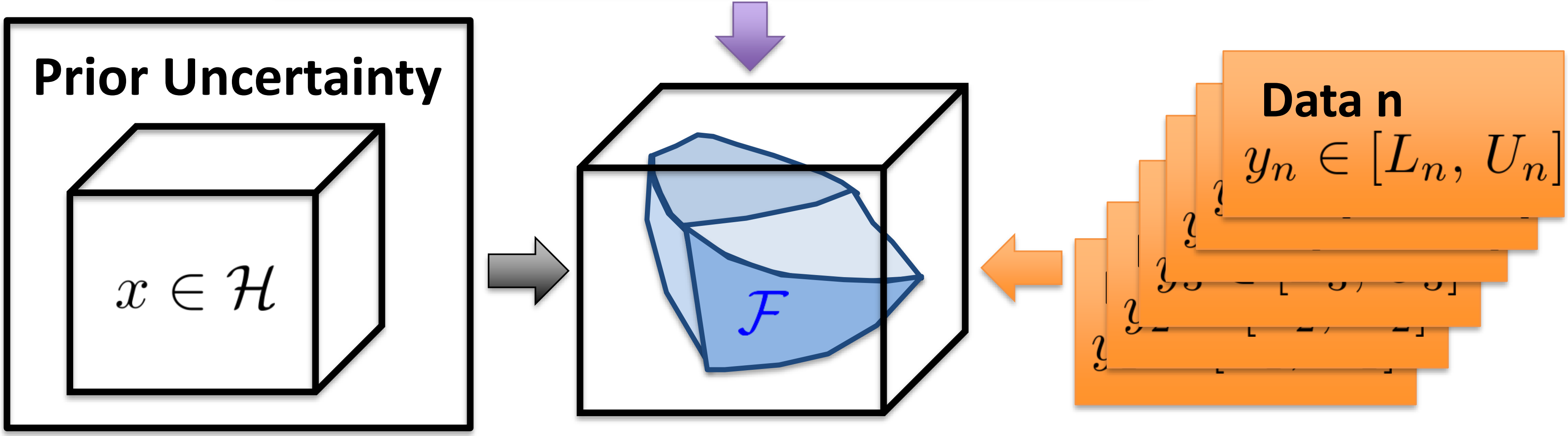
Uniform sampling of a feasible set and its application in UQ

Wenyu Li

Arun Hegde
Jim Oreluk
Andrew Packard
Michael Frenklach

Bound-to-Bound Data Collaboration (B2BDC)

Model: $M_e(x), \quad e = 1, 2, \ldots, n$



Feasible set

$$\{x \in \mathcal{H} : \; L_e \leq M_e(x) \leq U_e, \quad e = 1, 2, \ldots, n\}$$

Uniform sampling

Goal: uniform sampling of feasible set

- Sampling is useful in providing information about $\mathcal{F}$
- B2BDC makes **NO** distribution assumptions, but as far as taking samples, uniform distribution of $\mathcal{F}$ is reasonable
- Applying Bayesian analysis with **specific prior assumptions** also leads to uniform distribution of $\mathcal{F}$ as posterior (shown in next slide)

What Bayesian analysis leads to $\mathcal{U}(\mathcal{F})$

Deterministic model: $M_e(x)$

Prior distribution

$$X \sim \mathcal{U}(\mathcal{H})$$

$$f(x) = \begin{cases} \frac{1}{V(\mathcal{H})} & x \in \mathcal{H} \\ 0 & \textbf{else} \end{cases}$$

Measurement distribution

$$Y_e \sim \mathcal{U}([L_e, U_e])$$

$$f(y_e) = \begin{cases} \frac{1}{U_e - L_e} & y_e \in [L_e, U_e] \\ 0 & \textbf{else} \end{cases}$$

Bayesian analysis

$$p(x|y) \sim p(x)p(M_1(x)) \cdots p(M_n(x))$$

Posterior distribution

$$f(x|y) = \begin{cases} \frac{1}{V(\mathcal{F})} & x \in \mathcal{F} \\ 0 & \textbf{else} \end{cases}$$

B2BDC and Bayesian Calibration and Prediction (BCP)

Reference

[1] Frenklach, M., Packard, A., Garcia-Donato, G., Paulo, R. and Sacks, J., 2016. Comparison of Statistical and Deterministic Frameworks of Uncertainty Quantification. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), pp.875-901.

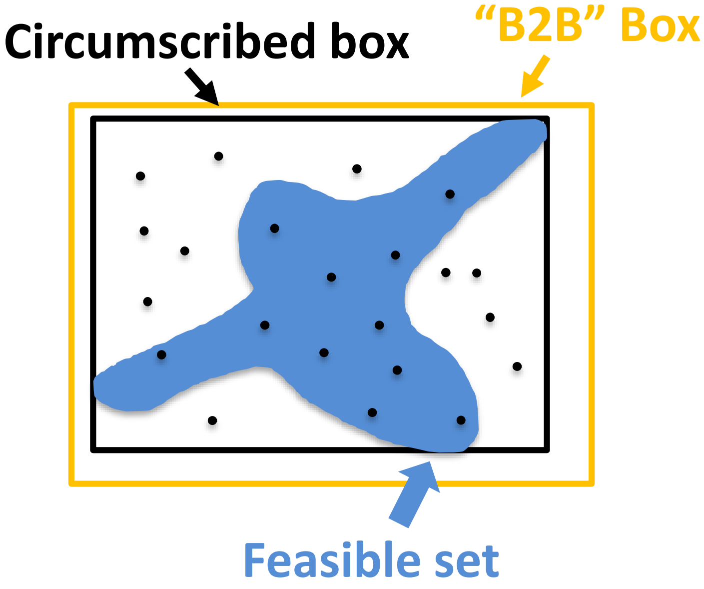
Nomenclature

- sampling efficiency ⟺ acceptance rate
- feasible set $\{x \in \mathcal{H} : \; L_e \leq M_e(x) \leq U_e, \quad e = 1, 2, \ldots, n\}$

Rejection sampling with box

Procedure:

- find a bounding box
  - available from B2BDC
- generate uniformly distributed samples in the box as candidates
- reject the points outside of feasible set



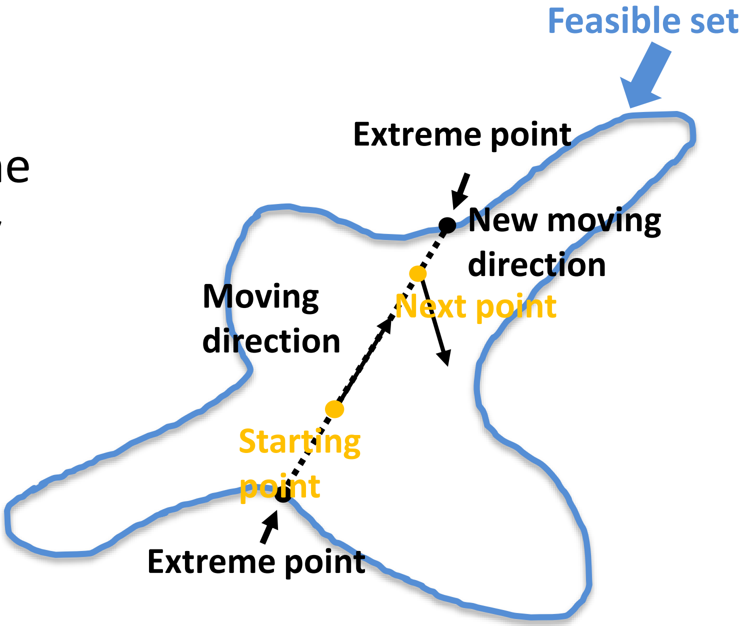
Pros & Cons

- **provably uniform** in the feasible set
- practical in **low** dimensions
- **impractical** in **higher** dimensions

Random walk (RW)

Procedure:

- start from a feasible point
  - available from B2BDC
- select a random direction, calculate extreme points and choose the next point uniformly
- repeat the process



Pros & Cons

- **NOT** limited by problem dimensions
- **NOT** necessarily uniform in the feasible set

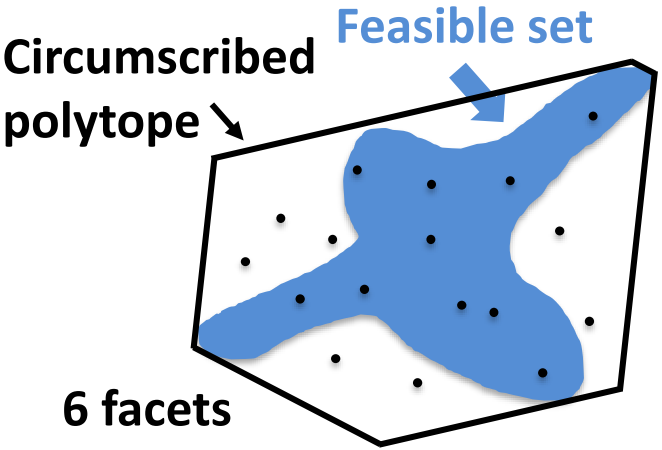
Rejection sampling with polytope

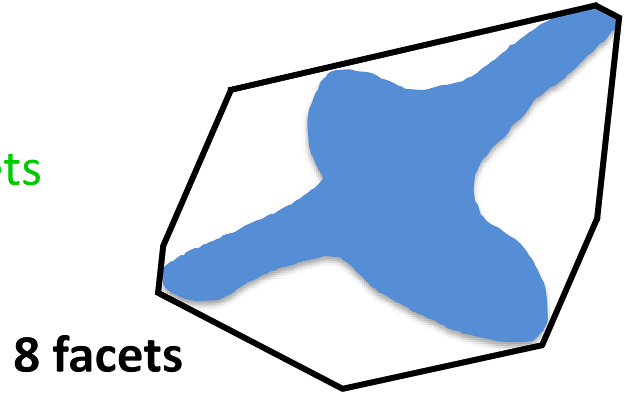
Procedure:

- find a bounding polytope
- generate candidate points by random walk
- reject the points outside of feasible set

Pros & Cons

- **provably uniform** in the feasible set
- increased efficiency with more polytope facets

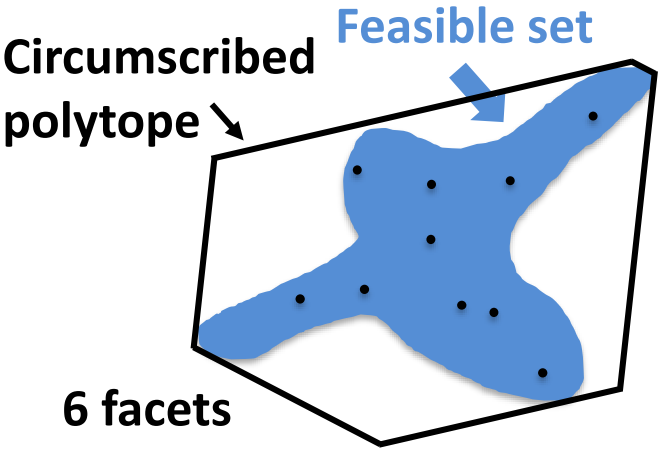




Rejection sampling with polytope

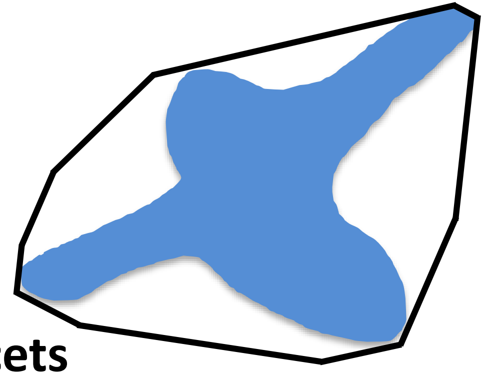
Procedure:

- find a bounding polytope
- generate candidate points by random walk
- reject the points outside of feasible set



Pros & Cons

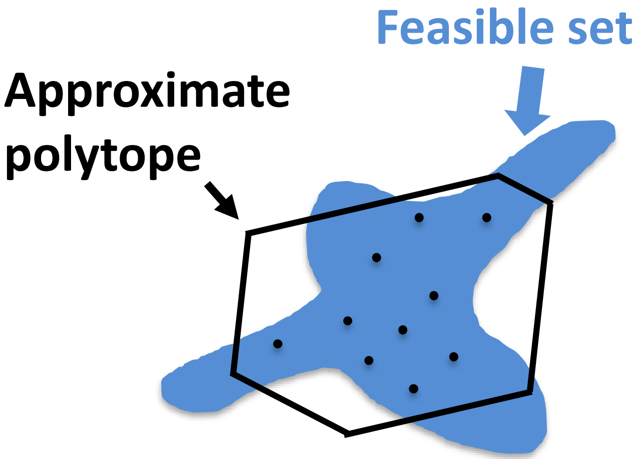
- **provably uniform** in the feasible set
- increased efficiency with more polytope facets
- practical in **low to medium** dimensions
- **limited** by computational resource



Approximation strategy

Procedure:

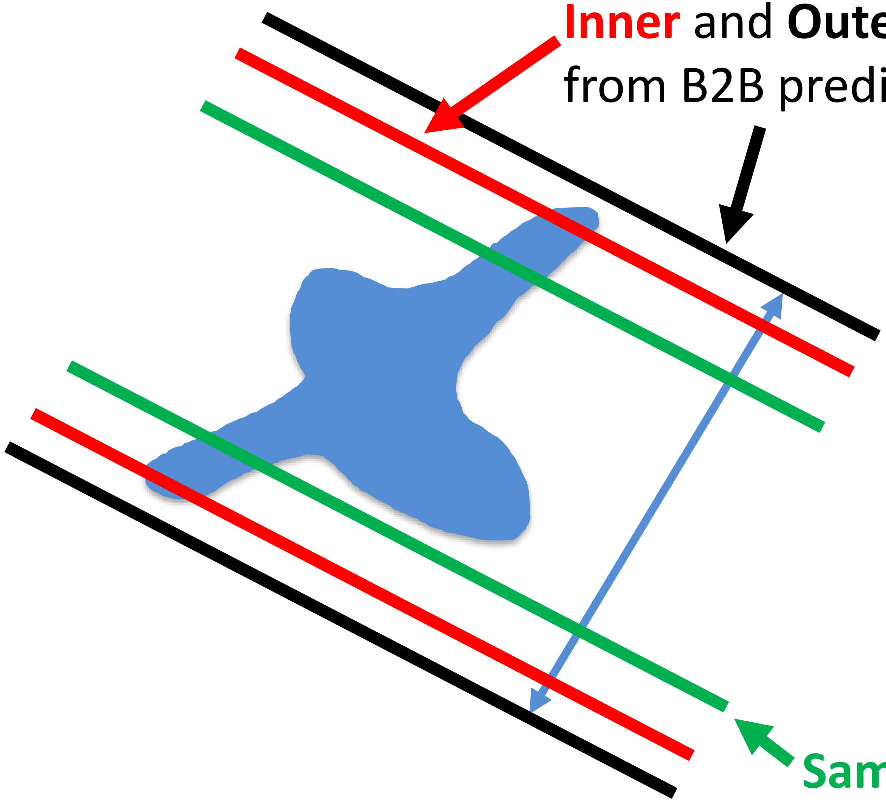
- relax the requirement that the polytope needs to contain the feasible set completely
- generate candidate points by random walk
- reject the points outside of feasible set



Pros & Cons

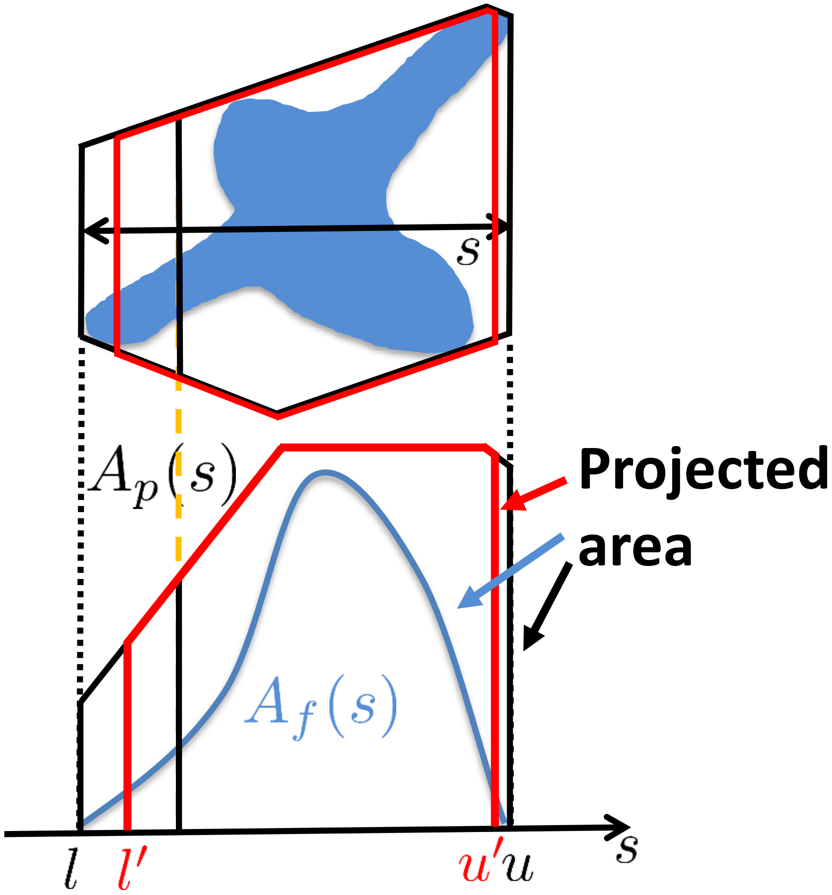
- practical in **medium to high** dimensions
- samples don't cover the **whole** feasible set

Define the polytope: one facet



- **Outer** bound from optimization (**NO** approximation, **provably uniform**)
- **Inner** bound from optimization (less aggressive approximation, very close to circumscribed bound)
- **Sample** bound (more aggressive approximation, performance depends on problem)

Effect on sampling efficiency



Efficiency density function

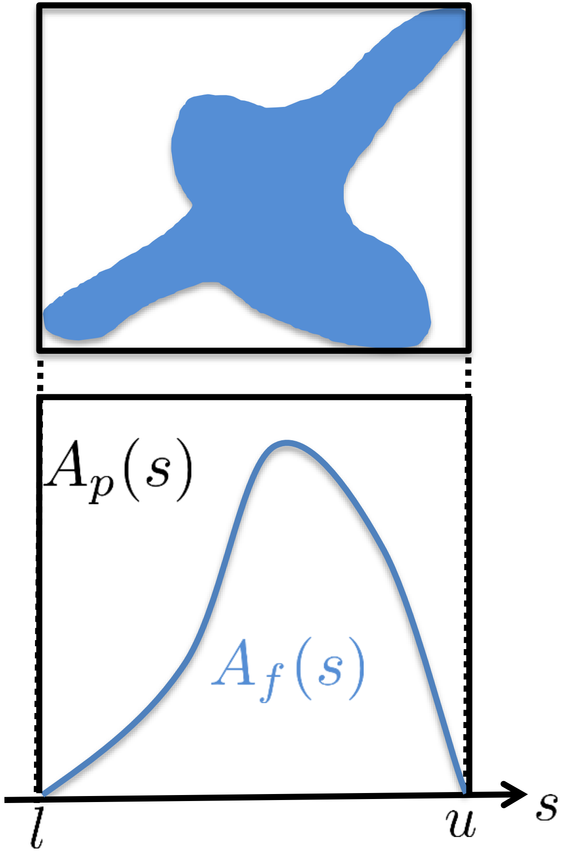
$$E(s) = \frac{A_f(s)}{A_p(s)}$$

$$e = \frac{1}{u-l} \int_l^u E(s)ds$$

Condition for improved efficiency

$$\frac{1}{u'-l'} \int_{l'}^{u'} E(s)ds > \frac{1}{u-l} \int_l^u E(s)ds$$

Effect on sampling efficiency



Special case with bounding box

$$A_p(s) = c \quad s \in [l,\ u]$$

$$A_f(s) \propto p(s)$$

$$E(s) = \frac{A_f(s)}{A_p(s)} \propto p(s)$$

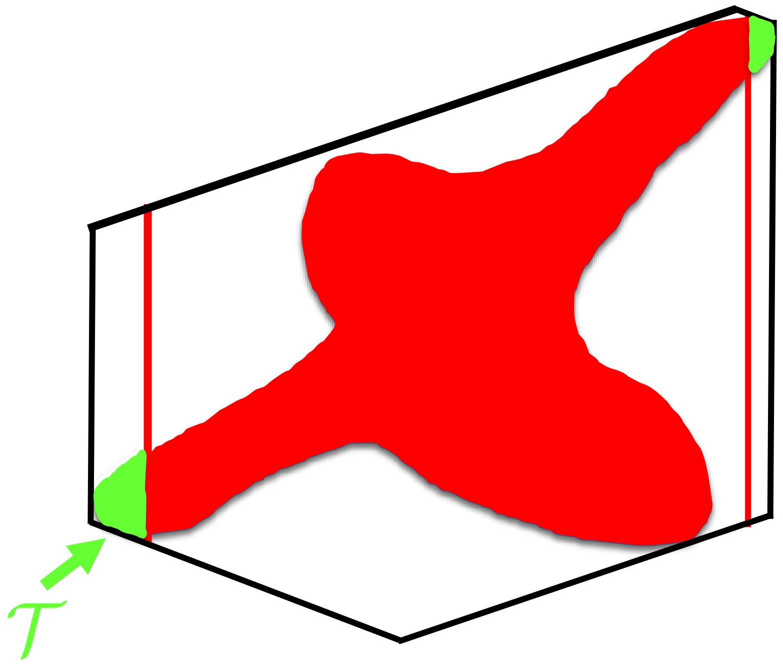
Assumption

$E(s) \propto p(s)$ **in the polytope case**

Posterior check

$$\frac{1}{u'-l'} \int_{l'}^{u'} p(s)ds > \frac{1}{u-l} \int_{l}^{u} p(s)ds$$

Effect on sampled distribution



Target distribution

$$p(x) = \begin{cases} \frac{1}{V(\mathcal{F}_b)} & \textbf{if } x \in \textbf{blue} \\ 0 & \textbf{else} \end{cases}$$

Approximated distribution

$$p'(x) = \begin{cases} \frac{1}{V(\mathcal{F}_r)} & \textbf{if } x \in \textbf{red} \\ 0 & \textbf{else} \end{cases}$$

Difference of mean for a function $Q(x)$

$$d = \left| \int Q(x)p'(x)dx - \int Q(x)p(x)dx \right|$$

$$\leq \left(\max_{x \in \mathcal{F}_b} |Q(x)|\right) \int_{\mathcal{T}} p(x)dx$$

Toy example

Test condition:

- 5 parameters, 30 constraints
- 1000 facets for each polytope
- Optimization and sample bounds
- 1000 sample points

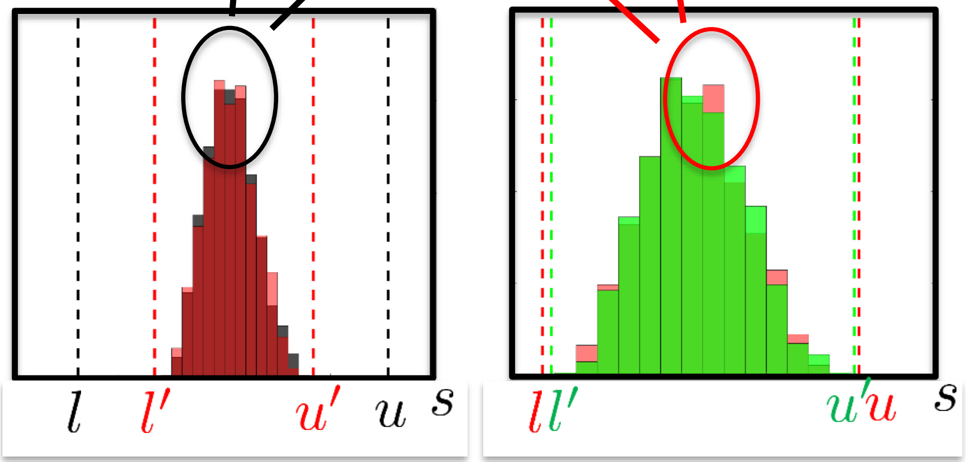
Polytope bound	Efficiency (%)
Outer bound	0.095
Inner bound	20.8
Sample bound	27.7

Posterior check

$$\frac{1}{u'-l'}\int_{l'}^{u'} p(s)ds > \frac{1}{u-l}\int_{l}^{u} p(s)ds$$

Outer -> Inner : 1.33 > 0.68

Inner -> Sample : 1.40 > 1.33

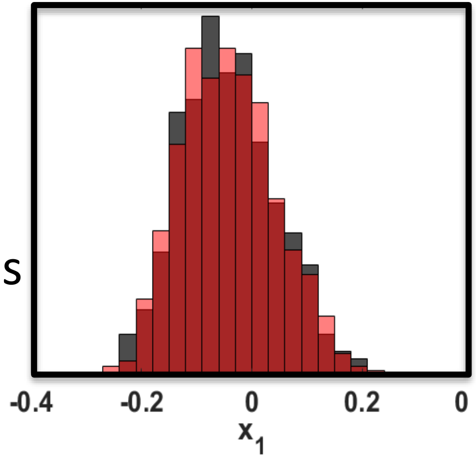


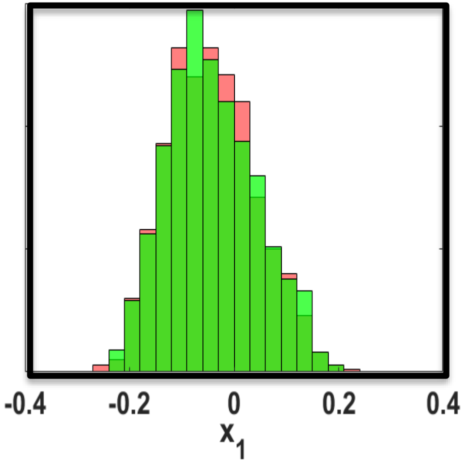
Toy example

Test condition:

- 5 parameters, 30 constraints
- 1000 facets for each polytope
- Optimization and sample bounds
- 1000 sample points

Polytope bound	Efficiency (%)
Outer bound	0.095
Inner bound	20.8
Sample bound	27.7



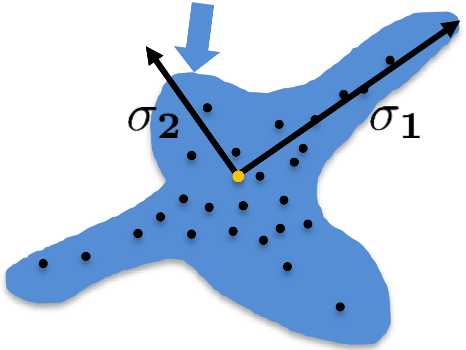


Passed the Kolmogorov-Smirnov test with 0.05 significance level

Principal component analysis (PCA)

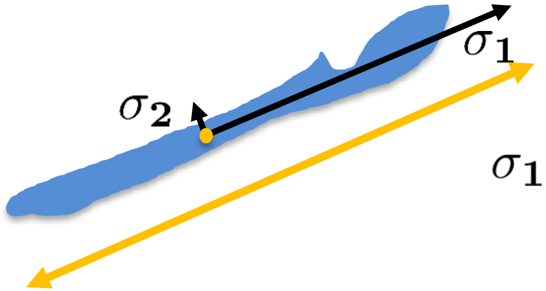
Procedure:

- collect RW samples from the feasible set
- conduct PCA on RW samples
- find a subspace based on PCA result
- generate uniform samples in the subspace



Pros & Cons

- **reduced** problem dimension
- works **only if** feasible set approximates lower-dimensional manifold/subspace



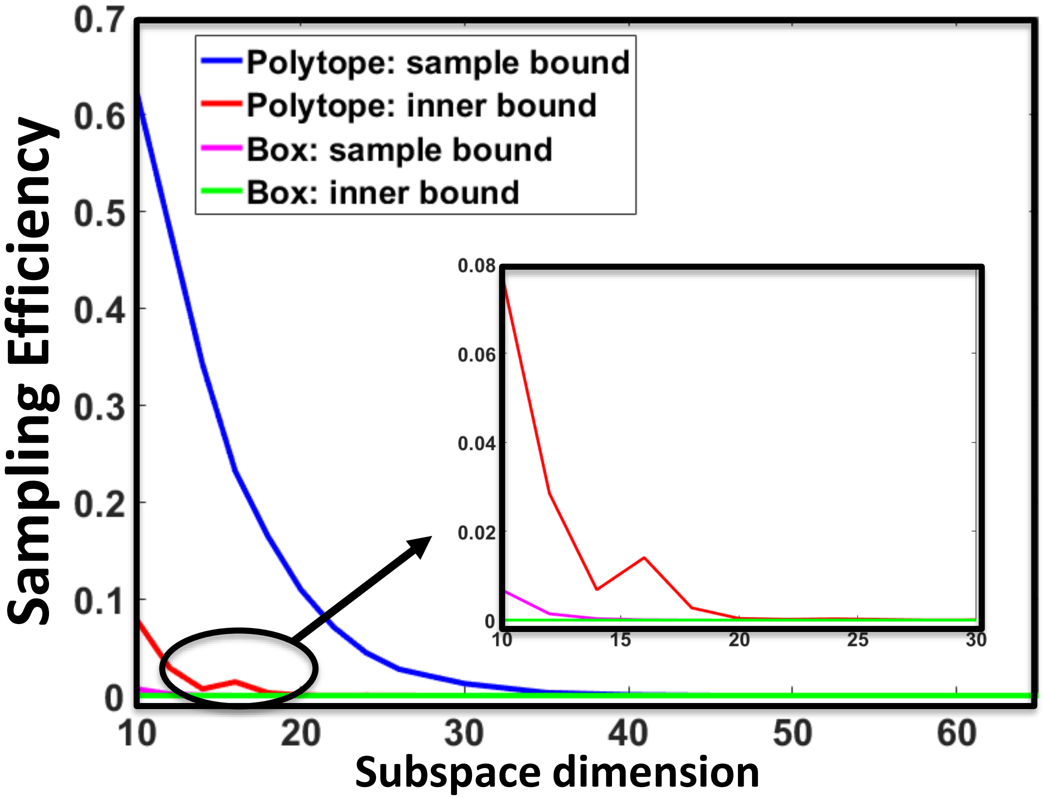
GRI-Mech

Test condition:

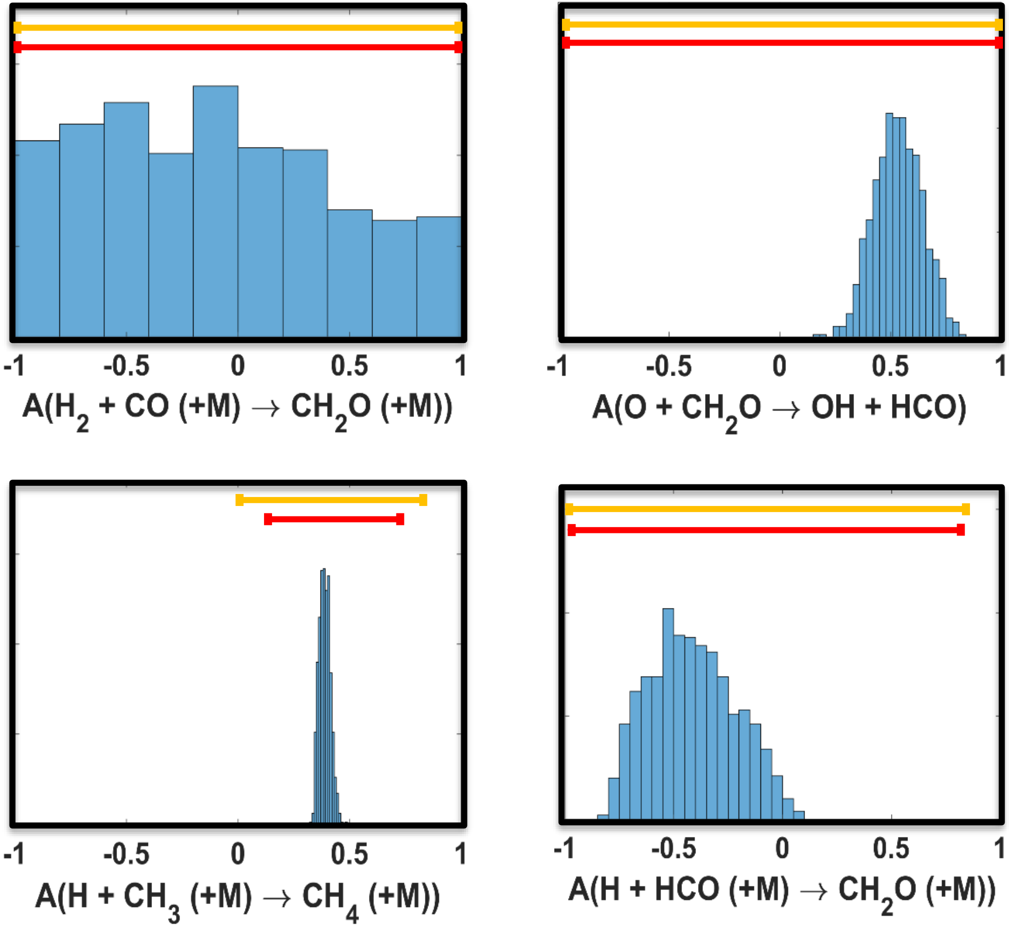
- 102 parameters
- 76 experimental data
- 10^7 RW samples for PCA
- 10-65 subspace dimension
  - 10^4 facets for each polytope
  - 10^7 candidate points for sampling

Test methods:

- polytope and box
- inner and sample bounds



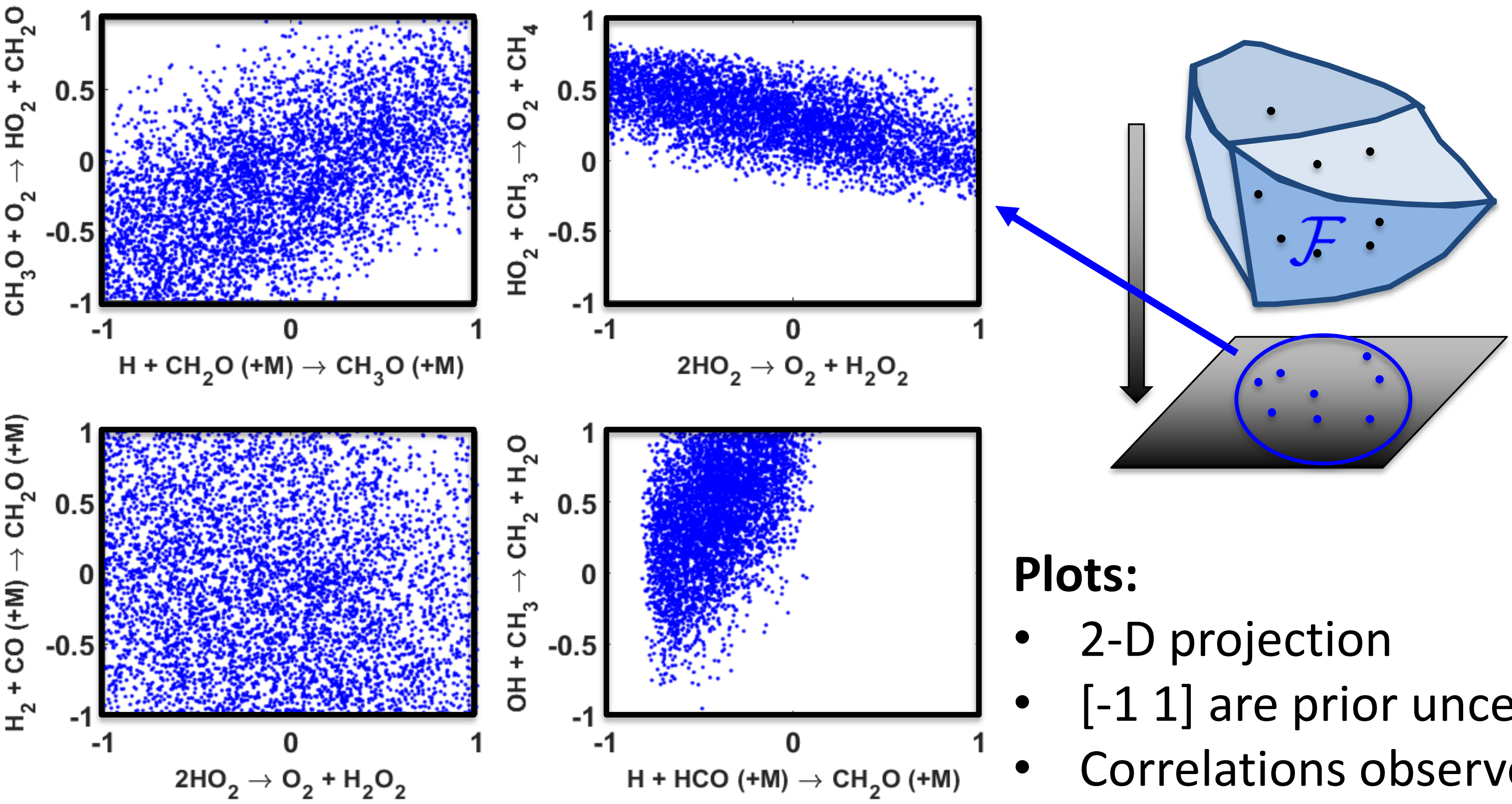
GRI-Mech: 1-D posterior marginal uncertainty



Test condition:

- 45 subspace dimension
- Polytope with sample bound
- 10^4 facets for the polytope
- 1000 sample points
- [-1, 1] are prior uncertainties

GRI-Mech: 2-D posterior joint uncertainty



Plots:

- 2-D projection
- [-1 1] are prior uncertainties
- Correlations observed

Summary

- We developed methods to generate uniformly distributed samples of a feasible set
- Approximation strategy and PCA further improves the practicality of rejection sampling method
- Hybrid statistical-deterministic uncertainty quantification process combining B2BDC prediction and uniform sampling

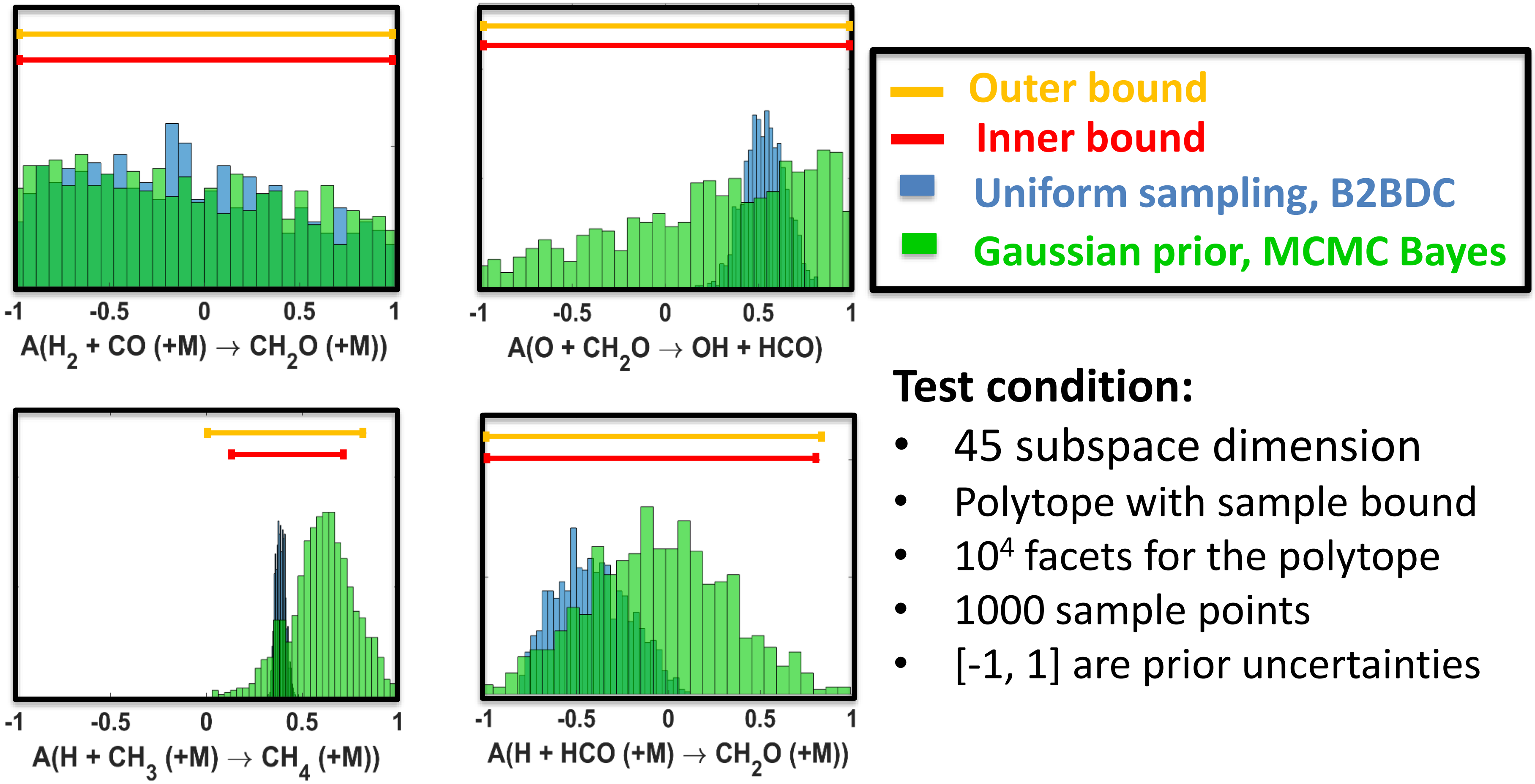
Acknowledgements

This work is supported as a part of the CCMSC at the University of Utah, funded through PSAAP by the National Nuclear Security Administration, under Award Number DE-NA0002375.

Thank you

Questions?

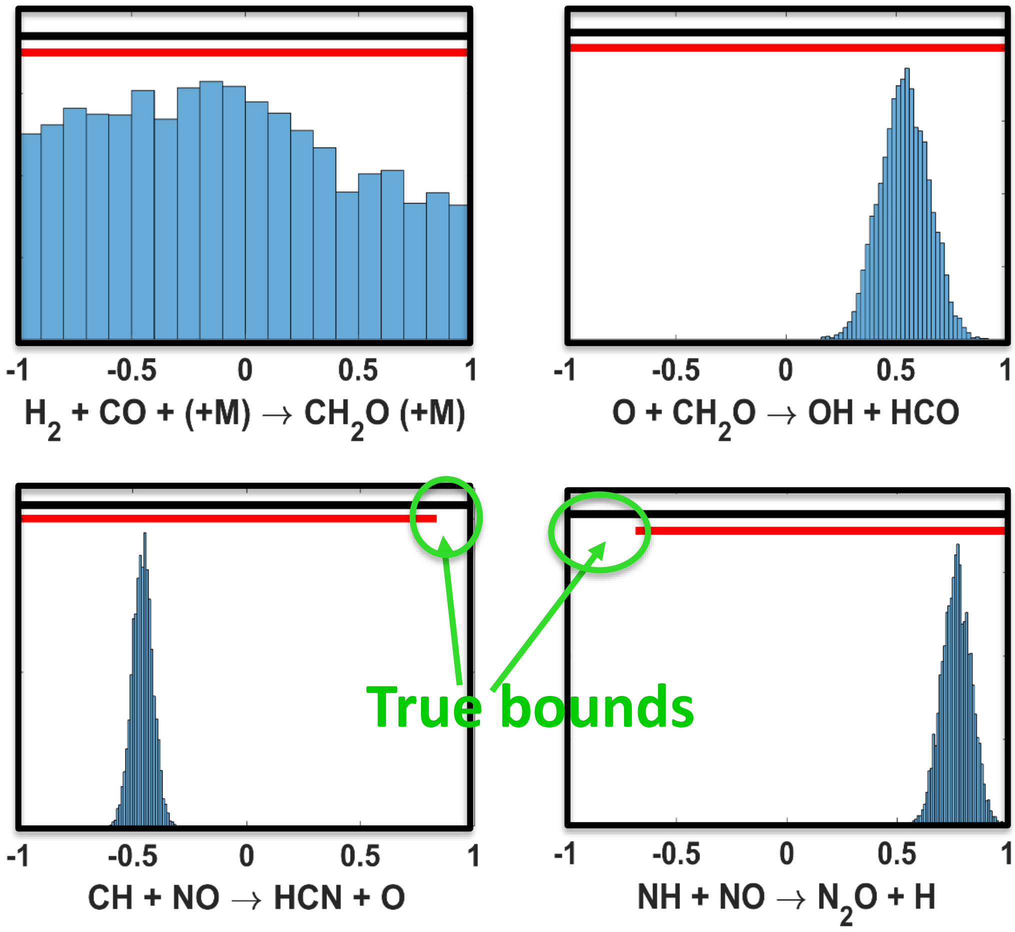
GRI-Mech: 1-D posterior marginal uncertainty

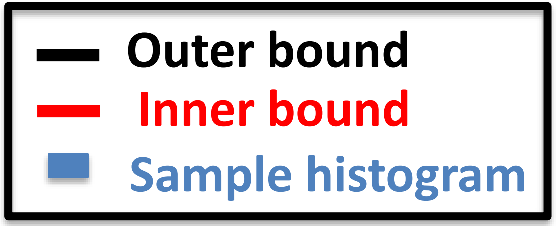


Test condition:

- 45 subspace dimension
- Polytope with sample bound
- 10^4 facets for the polytope
- 1000 sample points
- [-1, 1] are prior uncertainties

GRI-Mech: 1-D posterior marginal uncertainty

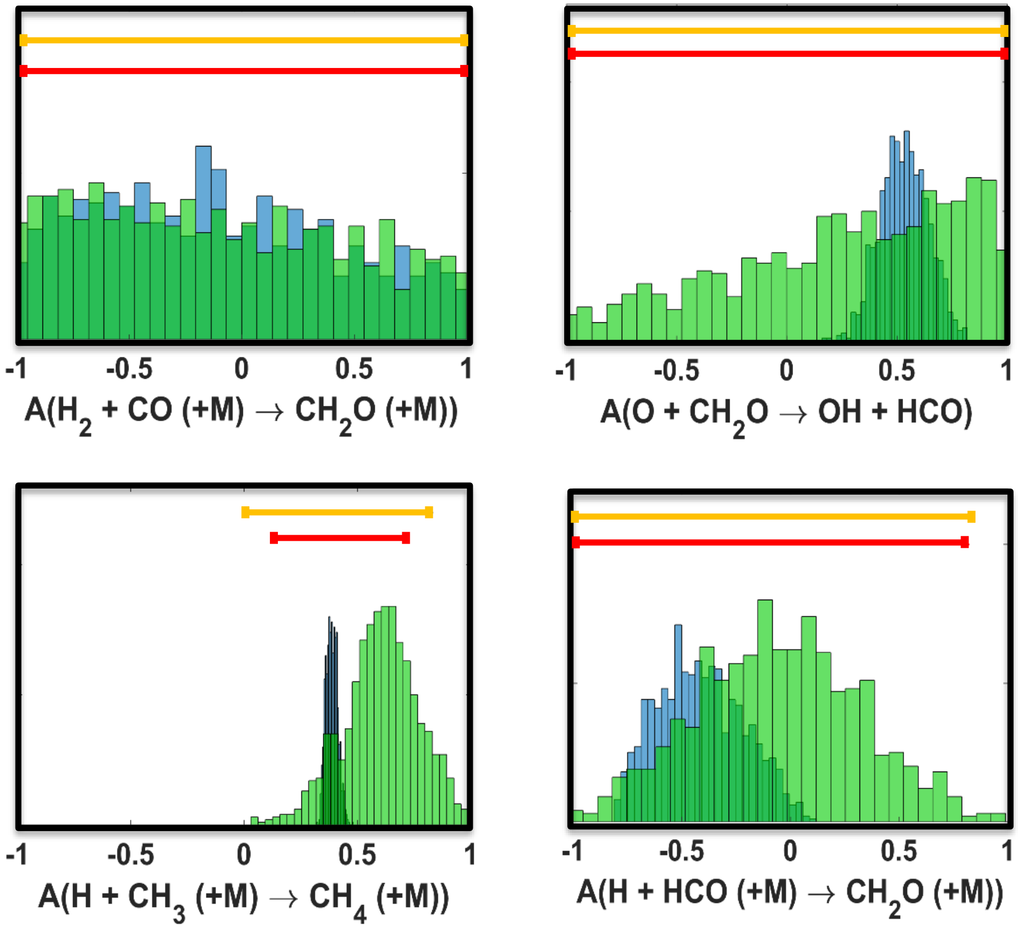




Test condition:

- 45 subspace dimension
- Polytope with sample bound
- 10^4 facets for the polytope
- 1000 sample points
- [-1, 1] are prior uncertainties

GRI-Mech: 1-D posterior marginal uncertainty



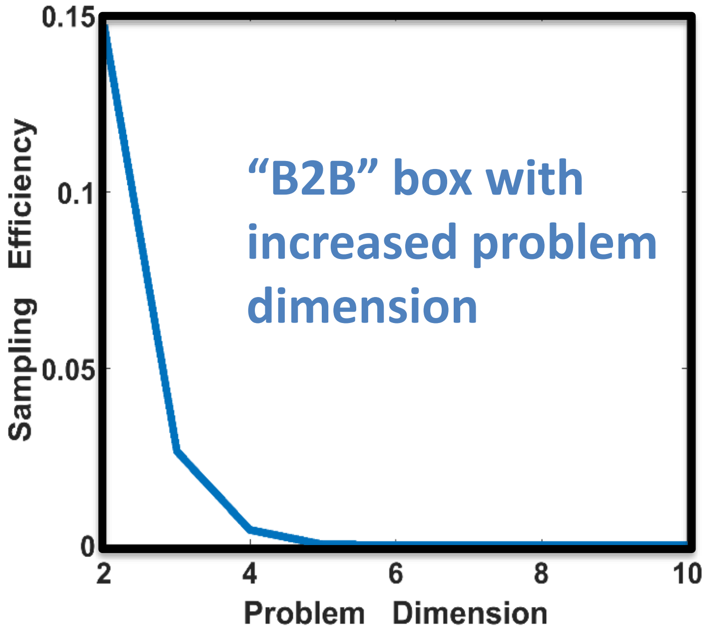
Test condition:

- 45 subspace dimension
- Polytope with sample bound
- 10^4 facets for the polytope
- 1000 sample points
- [-1, 1] are prior uncertainties

Rejection sampling with box

Procedure:

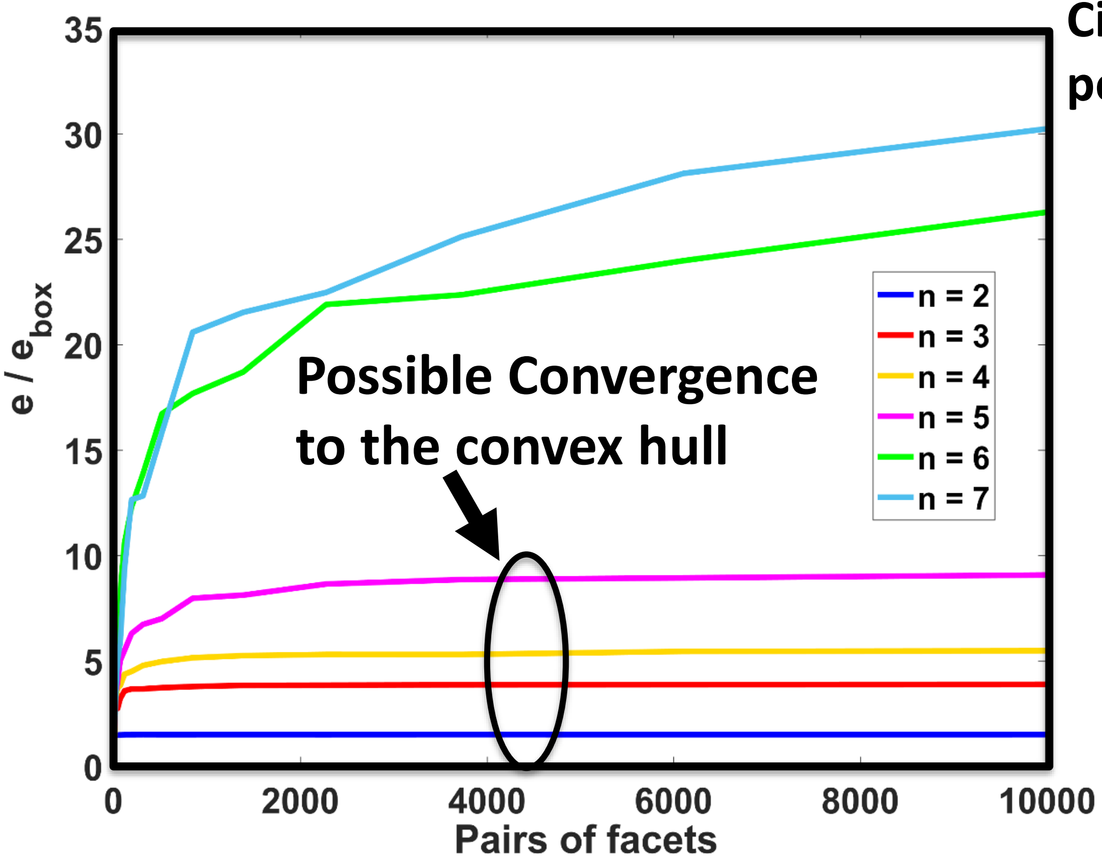
- find a bounding box
    - available from B2B
- generate uniformly distributed samples in the box as candidates
- reject the points outside of feasible set

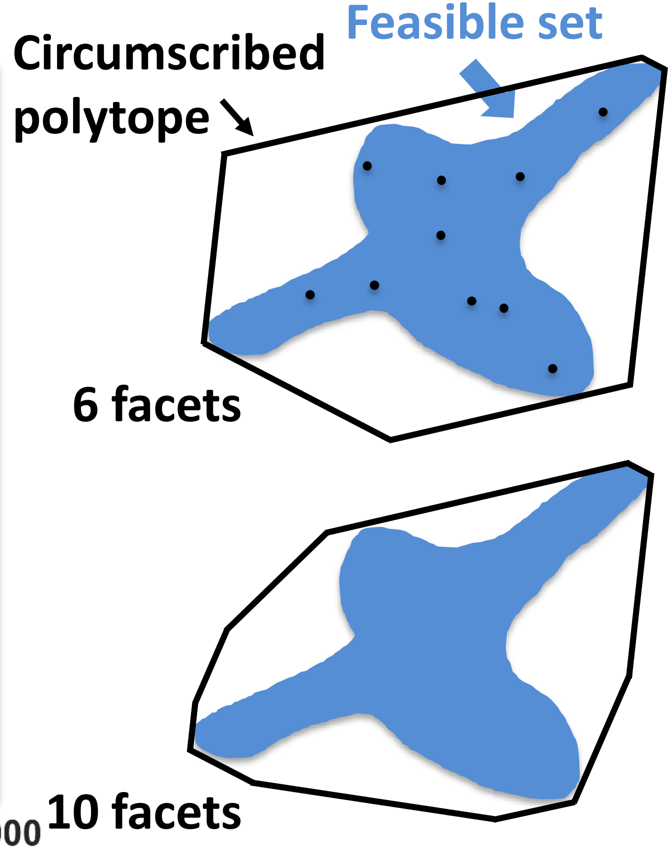


Pros & Cons

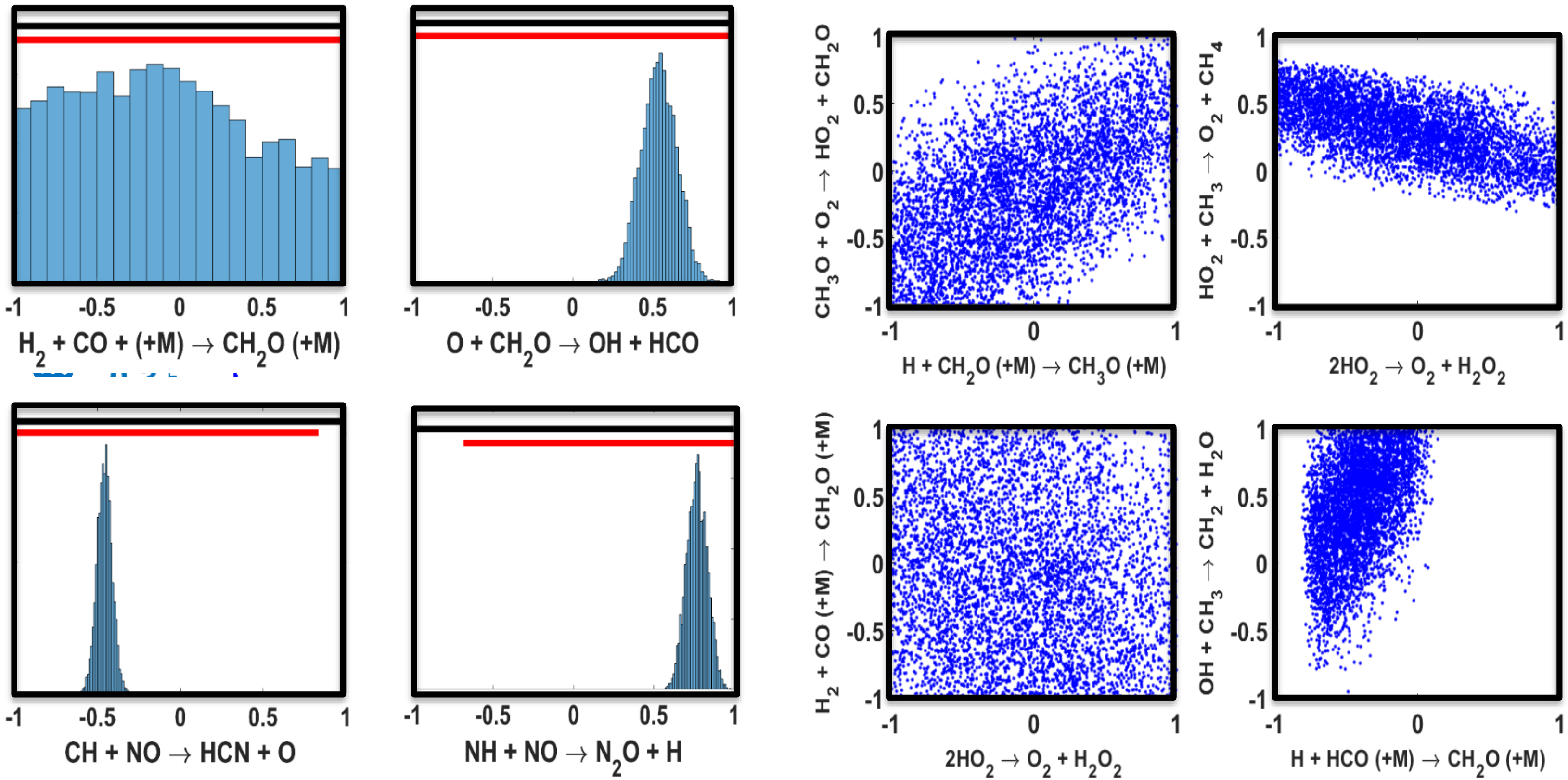
- **provably uniform** in the feasible set
- practical in **low** dimensions

Rejection sampling with polytope





Conclusion



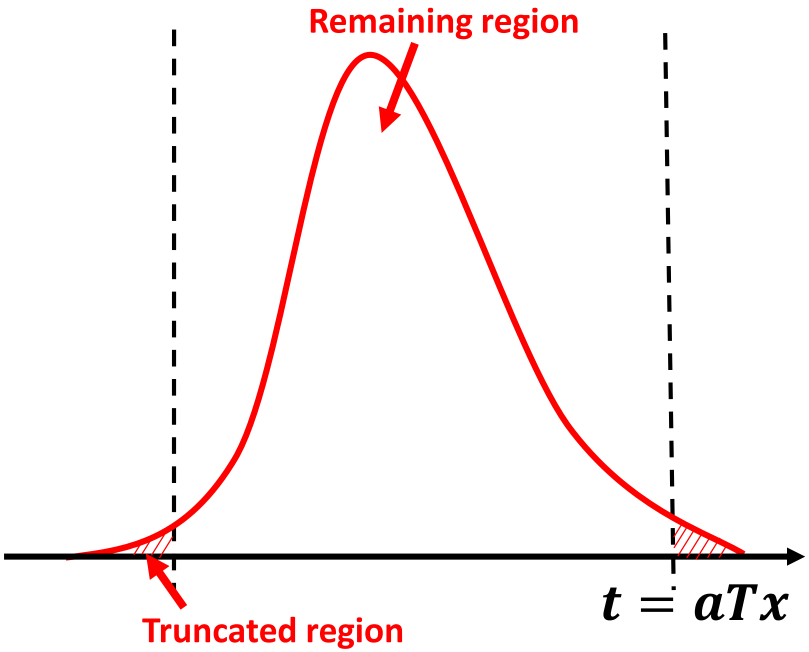
H2 + CO + (+M) → CH2O (+M)
O + CH2O → OH + HCO
CH + NO → HCN + O
NH + NO → N2O + H
CH3O + O2 → HO2 + CH2O
H + CH2O (+M) → CH3O (+M)
HO2 + CH3 → O2 + CH4
2HO2 → O2 + H2O2
H2 + CO (+M) → CH2O (+M)
2HO2 → O2 + H2O2
OH + CH3 → CH2 + H2O
H + HCO (+M) → CH2O (+M)

Heuristic approximation strategy (continued…)

- Consider the statistical quality of samples returned with heuristic approximation by estimating the difference in its statistical inference of a function Q(x). Denote the truncated and remaining area as $\mathcal{T}$ and $\mathcal{R}$, then

$$
\begin{aligned}
d &= |\int_{\mathcal{R}} Q(x)p'(x)dx - \int_{\mathcal{S}} Q(x)p(x)dx| \\
&\leq (c-1)\hat{Q_{\mathcal{R}}} \int_{\mathcal{R}} p(x)dx + \hat{Q_{\mathcal{T}}} \int_{\mathcal{T}} p(x)dx \\
&= (\hat{Q_{\mathcal{R}}} + \hat{Q_{\mathcal{T}}}) \int_{\mathcal{T}} p(x)dx \leq \hat{Q_{\mathcal{S}}} \int_{\mathcal{T}} p(x)dx
\end{aligned}
$$

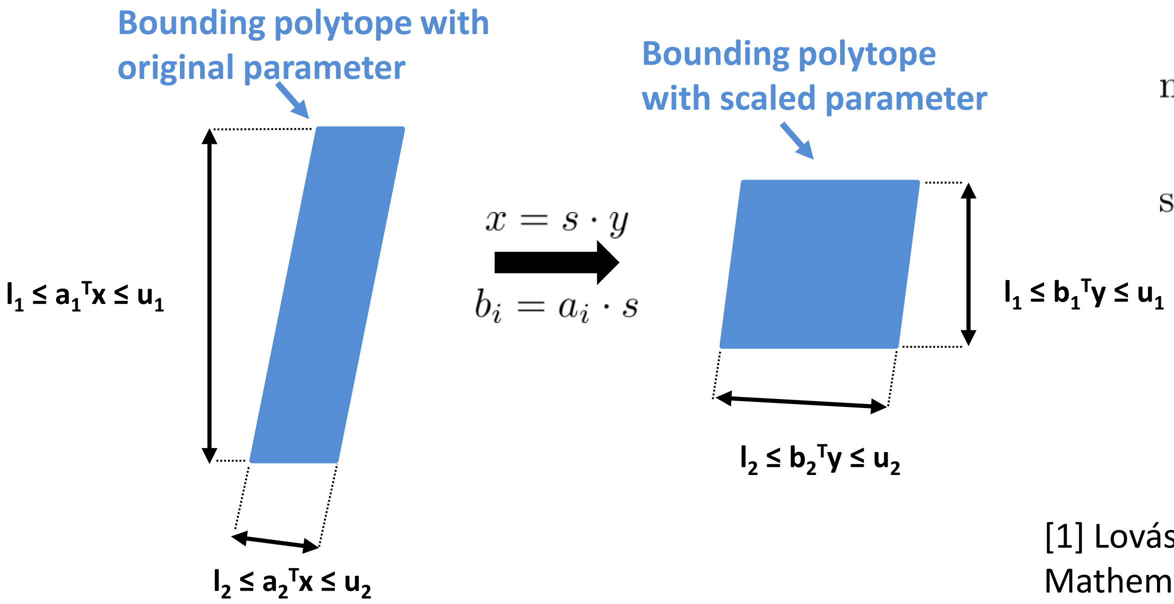
- *Hypothesis*. If the target distribution has a small integrated probability in the truncated region, the inferring difference of the returned samples are likely to be small compared to the target distribution



Rejection sampling with polytope (continued…)

Parameter scaling

- scales the parameters so the polytope with the scaled parameters is more isotropic
- a 2-D example is given in the following figure for illustration
- RW performs better (converges faster) with a more isotropic polytope[1]



$$\begin{aligned}
\underset{s,t_1,t_2}{\text{minimize}} \quad & t_2 - t_1 \\
\text{subject to} \quad & t_1 \le \frac{a_j^T s}{u_j - l_j} \quad j = 1, 2, 3\ldots \\
& t_2 \ge \frac{a_j^T s}{u_j - l_j} \quad j = 1, 2, 3\ldots \\
& s \ge 1
\end{aligned}$$

[1] Lovász, L., 1999. Hit-and-run mixes fast. Mathematical Programming, 86(3), pp.443-461.

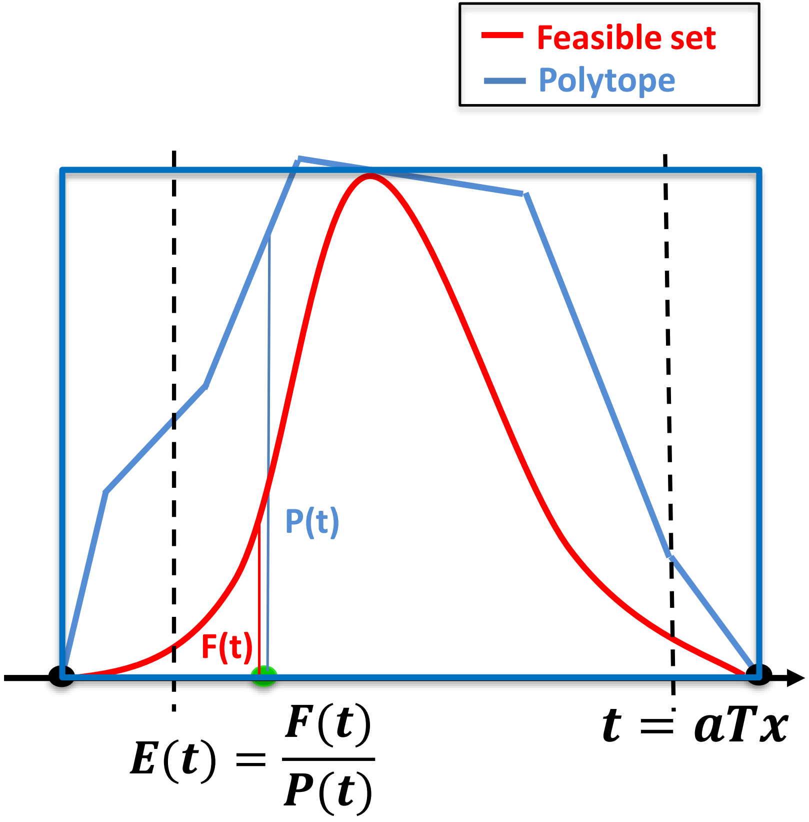
Acknowledgement

We gratefully acknowledge the support by U.S. Department of Energy, National Nuclear Security Administration, under Award Number DE-NA0002375.

National Nuclear Security Administration

Heuristic approximation strategy (continued…)

- A sufficient condition that the sampling efficiency will increase with the heuristic approximation is derived:
- *Hypothesis*. Parameterize the direction as $t = a^T x$ and specify the efficiency density function $E(t)$ as $E(t) = \int_{a^T x = t} I(F)dx / \int_{a^T x = t} I(P)dx$. Denote the truncated region as $\mathcal{T}$ and the remaining region as $\mathcal{R}$. If $\int_{t \in \mathcal{R}} E(t)dt / \int_{t \in \mathcal{R}} dt > \int_{t \in \mathcal{T}} E(t)dt / \int_{t \in \mathcal{T}} dt$ the sampling efficiency will increase with the approximation
- *Conjecture*. If the target distribution approximates a high-weight center, low-weight tail shape along the directions selected for heuristic approximation, then the efficiency is likely to increase.



Motivation of uniform sampling of the feasible set

- **We don't know the distribution of returned points if the feasible set is not convex (and in general it isn't).**
- **Only qualitative conclusions can be made.**
- **To make the analysis quantitatively valid, we assume the uniform distribution of the feasible set.**
- **This is also the posterior distribution from Bayesian method if we assume uniform prior distributions on both parameter and measurement uncertainties**

Generate uniform samples of a feasible set

and its application in uncertainty quantification

Random walk application in DLR dataset

Test condition:

- 55 parameters
- 244 constraints
- 10^6 samples
- 2-D projection
- Bounds are prior