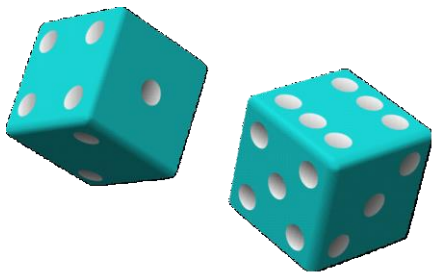
UMC: Unfinished Business

Donald L. Smith (ANL, retired)

Roberto Capote (IAEA)

Denise Neudecker (LANL)



CSEWG Meeting

2 – 4 November 2015

UMC

Unified Monte Carlo

The Concept

- Assume that **knowledge** about a set of nuclear **observable parameters** employed in nuclear system analyses can be represented by a multi-variate master **probability function**.

- This function should be **constructed** by incorporating the best available information from **theory** and **experiments**.

- The master function is then sampled using **Monte Carlo** methods to generate a **Markov Chain** of **random** observable parameter **vectors** that ultimately can be employed for a variety of **practical applications** such as generating **evaluations** and analyzing the **behavior** of derived nuclear **system parameters**.

Master Probability Density Function

Bayes Theorem: $p(\mathbf{x}\,|\,\mathcal{I},\mathcal{E}) = p_0(\mathbf{x}\,|\,\mathcal{I})\ \mathrm{L}(\mathbf{x}\,|\,\mathcal{I},\mathcal{E})$

- "$\mathcal{I}$" signifies **prior** information based on **theory** (modeling).
- "$\mathcal{E}$" denotes independent information from **experiments** that serves to **improve** (or augment) prior theoretical knowledge $\mathcal{I}$.
- "$\mathbf{x}$" represents **random vectors** corresponding to possible values of the nuclear **observables** (e.g., cross sections).
- The **prior** probability function p_0 is based on **theory**, while **likelihood L** is a probability function that **quantifies** the **consistency** of data from **theory** and **experiments** used to construct the master (posterior) probability function $p(\mathbf{x}\,|\,\mathcal{I},\mathcal{E})$.

Prior Probability Function $p_0(\mathbf{x}|\mathcal{I})$

- $\mathcal{I}$ is a complicated **algorithm** that **maps** theoretical **model parameters** $\mathbf{q}$ to **calculated observables** $\mathbf{x}$, i.e., $\mathbf{x} = \mathcal{I}(\mathbf{q})$.

- By applying Monte Carlo techniques, a Markov Chain of **vectors** $\mathbf{q}_k$ can be generated by **random sampling** of parameters in space $\mathcal{S}(\mathbf{q})$ governed by a **probability** function $r_0(\mathbf{q})$. Usually, **mean values** $\mathbf{q}_0$ and **covariance matrix** $\mathbf{V}_q$ are specified. Then, Maximum Entropy suggests r_0 should be a **normal probability** function.

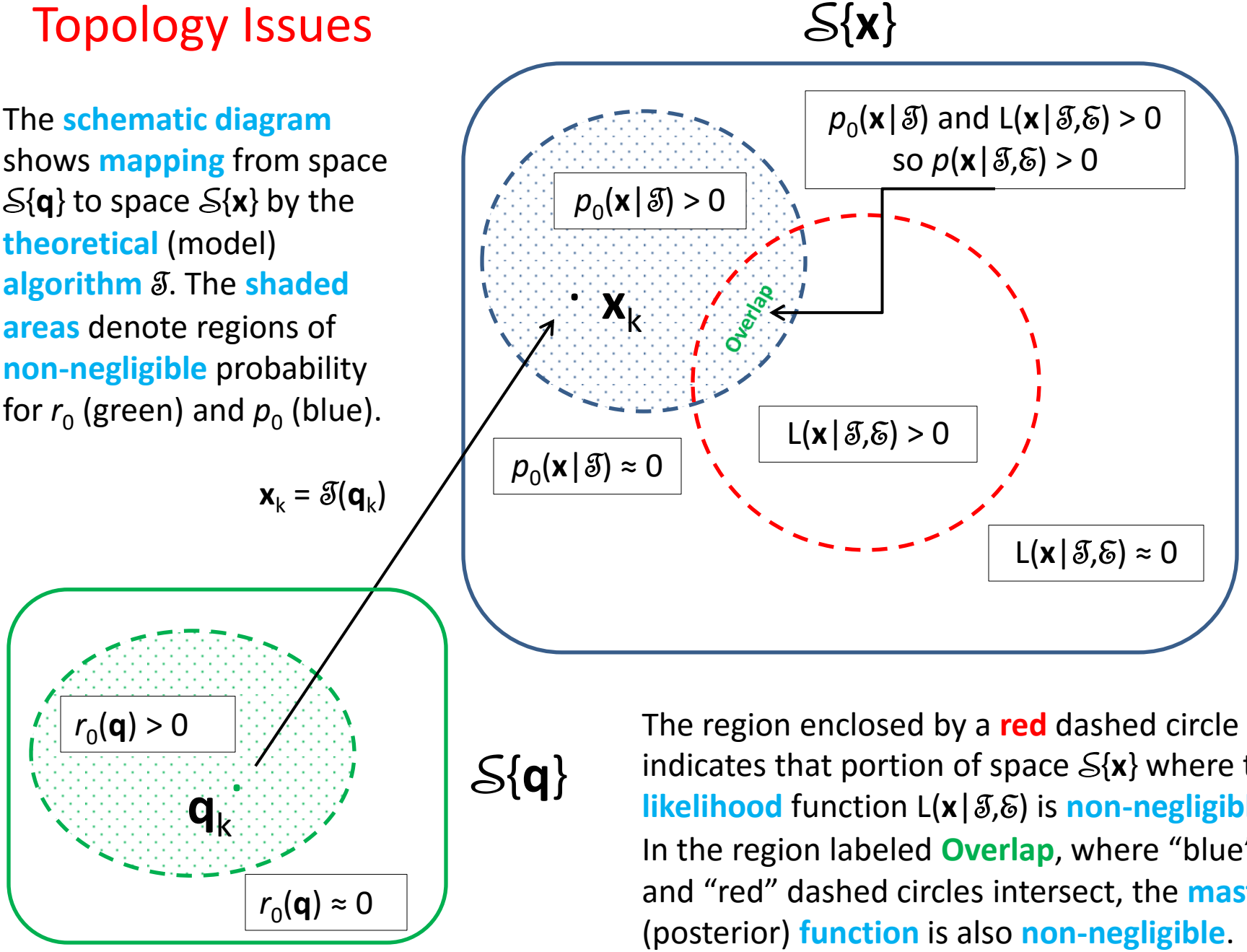
- A Markov chain of **values** $\mathbf{x}_k$ in **observables space** $\mathcal{S}(\mathbf{x})$ is generated by Monte Carlo sampling according to $\mathbf{x}_k = \mathcal{I}(\mathbf{q}_k)$.

- The collection $\{\mathbf{x}_k\}$ **reflects** the **prior probability** function p_0, but **rarely** (if ever) can p_0 be **expressed** explicitly as an **analytical function** that can be sampled in a conventional way!

Topology Issues

The **schematic diagram** shows **mapping** from space $\mathcal{S}\{\mathbf{q}\}$ to space $\mathcal{S}\{\mathbf{x}\}$ by the **theoretical** (model) **algorithm** $\mathfrak{I}$. The **shaded areas** denote regions of **non-negligible** probability for r_0 (green) and p_0 (blue).

$\mathbf{x}_k = \mathfrak{I}(\mathbf{q}_k)$



The region enclosed by a **red** dashed circle indicates that portion of space $\mathcal{S}\{\mathbf{x}\}$ where the **likelihood** function $L(\mathbf{x}|\mathfrak{I},\mathfrak{E})$ is **non-negligible**. In the region labeled **Overlap**, where "blue" and "red" dashed circles intersect, the **master** (posterior) **function** is also **non-negligible**.

UMC-G: Analytical Approximation to $p_0(\mathbf{x}|\mathcal{I})$

D.L. Smith, *Proceedings of AccApp'07*, Pocatello, ID, July 29 – August 2, 2007, Amer. Nucl. Soc. , p. 736.

- The collection of K calculated observable parameter **vectors** $\{\mathbf{x}_k\}$ generated by **Monte Carlo** (see preceding two slides), according to the mapping $\mathbf{x}_k = \mathcal{I}(\mathbf{q}_k)$, is used to calculate **mean values** $\mathbf{x}_0$ and **covariance matrix** $\mathbf{V}_x$ via the formulas:
  $x_{0i} \approx (\sum_{k=1,K} x_{ik}) / K$ and $(\mathbf{V}_x)_{ij} \approx [(\sum_{k=1,K} x_{ik}\, x_{jk}) / K] - x_{0i}\, x_{0j}$ (K is very large).

- The "true" **prior probability** function $p_0(\mathbf{x}|\mathcal{I})$ typically is **approximated** by a multi-variate **normal probability** function given by:
  $p_0(\mathbf{x}|\mathcal{I}) \approx C \exp\{-(½)[(\mathbf{x} - \mathbf{x}_0)^T \mathbf{V}_x^{-1} (\mathbf{x} - \mathbf{x}_0)]\}$ (C is a normalization constant).

Advantage: A lengthy Markov Chain of **sample values** is thus **replaced** by an **analytical approximation** having the **same mean values** and **covariance matrix**. This yields a master (posterior) function $p(\mathbf{x}|\mathcal{I},\mathcal{E})$ that can be **sampled** readily by **conventional Monte Carlo** methods, e.g., "Brute Force" or "Metropolis-Hastings".

Disadvantage: This approximation **discards** all information pertaining to **higher-order** distribution **moments** inherent in the Monte Carlo generated Markov chain $\{\mathbf{x}_k\}$. This **rejection** of information can lead to significant **biases** in cases where **non-linear** effects and distribution **skewness** and **kurtosis** are present.

UMC-B: Information in $p_0(\mathbf{x}|\mathcal{I})$ is Preserved

R. Capote et al., *Proceedings of ISRD-14*, Breton Woods, NH, May 22 – 27, 2011, ASTM STP-1550, p. 179.

- The **collection** of K calculated observable parameter **vectors** $\{\mathbf{x}_k\}$ generated by **Monte Carlo** (shown in two earlier slides), according to the mapping $\mathbf{x}_k = \mathcal{I}(\mathbf{q}_k)$, is **preserved**. Thus, **no information** on higher-order moments of p_0 is **discarded**.

- For each $\mathbf{x}_k$, a scalar **weighting factor** ω_k is generated according to the expression: $\omega_k = L(\mathbf{x}_k|\mathcal{I},\mathcal{E})$. Thus, the **worth** that is assigned to each MC sampled parameter vector $\mathbf{x}_k$ is based on its **consistency** with available **experimental data**, as reflected in the **likelihood** function.

- For very large K, it is assumed that **mean values** and **covariance matrix** for the master (posterior) probability function $p(\mathbf{x}|\mathcal{I},\mathcal{E})$ are estimated from:
  $x_{0i} \approx (\sum_{k=1,K} \omega_k x_{ik}) / (\sum_{k=1,K} \omega_k)$ and $(\mathbf{V}_x)_{ij} \approx [(\sum_{k=1,K} \omega_k x_{ik} x_{jk}) / (\sum_{k=1,K} \omega_k)] - x_{0i} x_{0j}$

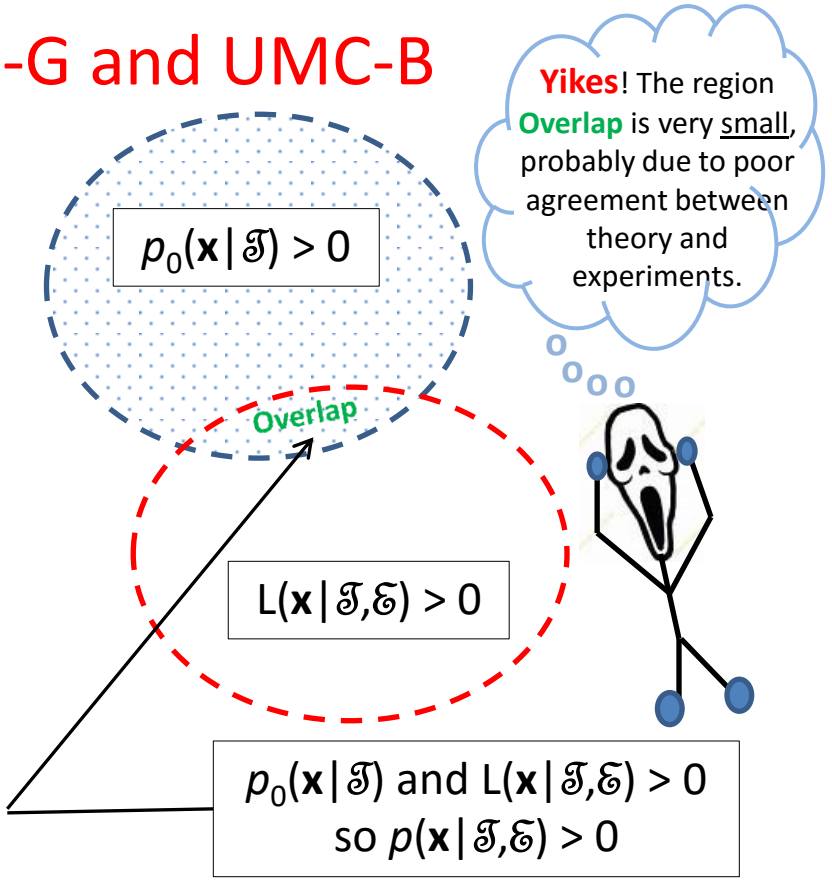
- The **Markov Chain** for **UMC-B** thus consists of the set of pairs $\{\mathbf{x}_k,\omega_k\}$. These values can be used for nuclear **systems applications** as well as **evaluations**.

Advantage: All **information** in function $p_0(\mathbf{x}|\mathcal{I})$ is clearly **preserved**, including that related to **non-linearity** as well as the distribution **skewness** and **kurtosis**.

A Closer Look at UMC-G and UMC-B

- The areas enclosed by “blue” and “red” dashed circles, respectively, indicate **regions** of **non-negligible probability** for the model-generated **prior probability** and the **likelihood function** that **quantifies** the **consistency** of theory and experiment.

- The **small** region **Overlap** of these two circles is indicative of **data inconsistency**. Such an outcome could have potentially **negative implications** for an application of the **UMC-B** method (e.g., **limited** or **biased sampling** of the sparsely sampled region **Overlap**). **Statistical inadequacy** is **not a problem** in applying the **UMC-G** approach, but it can **suffer** from significant **bias effects** due to the explicit **rejection** of **higher-order moments** of $p_0(\mathbf{x}|\mathfrak{I})$. **Data inconsistency** between theory and experiment will inevitably lead to evaluations and system analysis **results** that are very **questionable** and thus inherently **unreliable**.



Unfinished Business

- Further investigation of the mentioned sampling issues for the region **Overlap** in the UMC-B approach is warranted.
- Detailed inter-comparisons of GLS, UMC-G, and UMC-B predictions for extreme cases and inconsistent data are needed.

UMC-G Plus: A New Option?

- The **original UMC-G** formulation **discards** potentially valuable **information** about **higher moments** of the **prior probability** $p_0(\mathbf{x}|\mathfrak{I})$ that is **reflected** in the Markov Chain of **vectors** $\{\mathbf{x}_k\}$ generated by **Monte Carlo** sampling.

- It is **unlikely** that the **moments** of p_0 of **higher order** than mean values, covariances, skewness, and kurtosis will **affect applications** significantly.

- The **moments** of p_0 (mean values, covariances, skewness, and kurtosis) can be **estimated** using the collection of **sample vectors** $\{\mathbf{x}_k\}$.

Suggestion: Perhaps **analytical functions** might be found whose **parameters** can be **adjusted** to **approximate** the distribution **moments** deduced from $\{\mathbf{x}_k\}$. It could then be employed to serve as a **surrogate** for the **master** (posterior) **probability function** $p(\mathbf{x}|\mathfrak{I},\mathfrak{E})$, and it would then be **sampled** using **conventional Monte Carlo** methods.

Unfinished Business

- Investigate the structure of realistic MC-generated distributions $p_0(\mathbf{x}|\mathfrak{I})$ with the intent of quantifying typical mean values, covariances, skewness, and kurtosis.
- Identify families of analytical mathematical functions that might serve as surrogates for representing the MC-generated distributions $p_0(\mathbf{x}|\mathfrak{I})$ with greater fidelity than using a simple normal distribution (as in UMC-G).

Thoughts on Likelihood Functions $L(\mathbf{x}|\mathcal{I},\mathcal{E})$

- Available **experimental data** are usually comprised of **mean values** and (far less often) **covariances**. Therefore, **comparisons** between theoretically **calculated observables** and **experimental observables** should **involve** <u>at most</u> **mean values** and **covariances**.

- Consequently, the **likelihood** function $L(\mathbf{x}|\mathcal{I},\mathcal{E})$, in accordance with **Maximum Entropy**, should be an appropriately constructed **normal probability** function. In particular, it should have the form:

$$L(\mathbf{x}|\mathcal{I},\mathcal{E}) = C \exp\{-(\tfrac{1}{2})[(\mathbf{y}-\mathbf{y}_E)^T \mathbf{V}_E^{-1} (\mathbf{y}-\mathbf{y}_E)]\} \quad \text{(C is a normalization constant)}$$

<u>Note</u>: $\mathbf{y}_E$ is an **experimental data vector** with **covariance matrix** $\mathbf{V}_E$. Furthermore, $\mathbf{y} = \mathbf{f}(\mathbf{x})$, since what is measured ($\mathbf{y}$) **may not correspond** directly to the **observable parameters** ($\mathbf{x}$) that are being considered. The function collection "**f**" establishes how $\mathbf{x}$ and $\mathbf{y}$ are related.

- It may be very **difficult** to **construct** a rigorous **likelihood function** $L(\mathbf{x}|\mathcal{I},\mathcal{E})$ in any given situation due to one or more of the following limitations: i) **incomplete data**, ii) **discrepant** (wrong) **data**, iii) **weak sensitivity** relationships between the data ($\mathbf{y}$) and parameters of interest ($\mathbf{x}$), and iv) excessive **computational overhead**.

<u>Alternatives</u>: Because of these **limitations**, some investigators (notably A. Koning and D. Rochman) for **pragmatic reasons** have investigated using simpler **alternative likelihood** functions $L(\mathbf{x}|\mathcal{I},\mathcal{E})$.

<u>Unfinished Business</u>

- The impact of experimental data quality and availability on applications of UMC needs to be investigated thoroughly.
- Improve experimental covariance data.

The UMC Approach at a Crossroads?

➔ There are unresolved **technical issues** and unanswered **questions**. The **way forward** to further develop UMC must be **clarified**.

- Would **UMC-G** and **UMC-B** be truly **comparable** if $p_0(\mathbf{x}|\mathcal{I})$ could be expressed exactly as an **analytical function**?
- Can more **sophisticated** analytical function **approximations** to a MC prior than the normal distribution be found (e.g., **UMC-G Plus**)?
- Are the available **experimental data** sufficiently **accurate** and **comprehensive** to be useful in practice for **applying UMC**?
- Can **better** theoretical **models** be developed to **reduce** the **discrepancies** between **theory** and quality **experimental data**?
- If not, can **model-defects formalisms** be developed as practical measures to **cope** with model vs. experimental data **discrepancies**?
- How much extra **"value"** does **UMC contribute**, **compared** with **GLSQ**, to justify the additional **effort** and **computational burden**?

The End