Genetic Algorithms for HEP Event Reconstruction

Cho, Won Sang

University of Florida

MC4BSM-2014 at KAIST/IBS
Daejeon, 5. 20. 2014

HEP Event Reconstruction

- Given a hypothetical event decay topology, missing kinematic information to be reconstructed in a HEP event :
  - 1) Combinatorics : Cut-based selection in a good variable space
  - 2) Invisible momenta : Solving full constraining equations in (combined) events / Optimizing mass variables over all unknowns, subject to available constraints (MAOS/M_2)
- These classical methods try to reconstruct the information, based on the decision in an event-by-event basis.

Stochastic Optimization for a (joint) Distribution

- We introduce another approach using the collective behavior of event set.
- Stochastic optimization introduces randomness into the search process to accelerate progress.
- Imagin a string of events consist of known and unknown (random) data blocks !

1	2	...	N_{event}
$\{p_{vis}\}^1$	$\{p_{vis}\}^2$	...	$\{p_{vis}\}^{N_{event}}$
$\{q_{inv}\}^1,\ C(topol.)^1$	$\{q_{inv}\}^2,\ C^2$	...	$\{q_{inv}\}^{N_{event}},\ C^{N_{event}}$
$m(p_{vis}, q_{inv}\|C)^1$	$m(p_{vis}, q_{inv}\|C)^2$	...	$m(p_{vis}, q_{inv}\|C)^{N_{event}}$

$\rightarrow$ Consider the distribution of a function, $m(p_{vis}, q_{inv}|C)$ using a random sequence of missing momenta, $\{q^1_{inv}, q^2_{inv}, ..q^{N_{event}}_{inv}\}$ and combinatorics, $\{C^1, C^2, ...C^{N_{event}}\}$.

- Basically, one can try to search for the best sequence among all of the possible combinations of random variables in the unknown blocks, by which the corresponding $m(p_{vis}, q_{inv}|C)$ distribution best fits into a designated its physical (target) distribution in proportional to likelihood given model parameters.

- However, in general, one may end up with a huge set of degenerated solutions, as long as the target distribution of m itself is not quite case sensitive over the unknowns.

What kind of m (and its target function) do we use?

- The best target function would be a singular function, where a wrong value in an unknown block can lead a large distortion on the distribution, while the highest singularity is achieved when all the values are correct.
  $\rightarrow$ $m \equiv$ invariant mass of an on-shell resonance with random variables
- To be maximally model independent, one can simply test a stochastic optimization toward the maximal density assuming a narrow-width.

Which stochastic optimization algorithm do we use?

- **Genetic Algorithm !**

Genetic Algorithms for On-shell Singularity

What is the genetic algorithm ?

- Genetic algorithms are search and optimization techniques based on Darwin's Principle of *Natural Selection*.
- "Exchange good genes, Select the best individuals with good fitness, and Discard the rest"
- Optimizes a large number of (continuous/discrete) parameters with extremely complex objective function. It can easily jump out of a local minimum
- Does not require derivative information
- May encode the parameters so that the optimization is done with the encoded parameters
- A good explanation on why GAs work well can be found in the *schema-theorem* proposed by J. Holland (1975).

Simple GA works with the binary encoding for each variable.

$$
\begin{array}{rcl}
\textit{gene}_i & = & \textit{an encoded unknown value } (q_i \textit{ or } C_i) \\
 & & \textit{of an event block}
\end{array}
$$

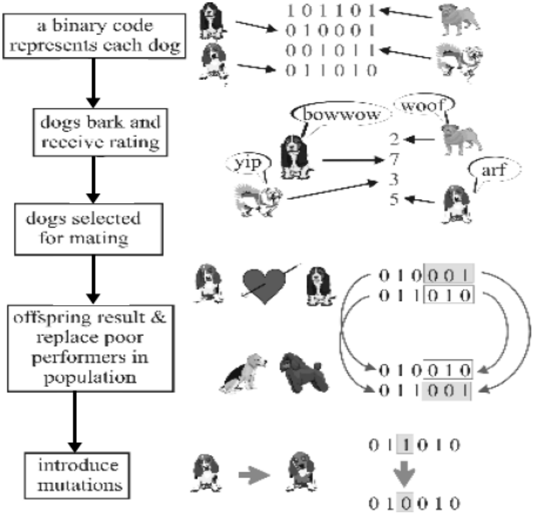
$$
\begin{array}{rcl}
 & & \begin{array}{cccc} \textit{Gene}_1 & \textit{Gene}_2 & .... & \textit{Gene}_{N_{event} \textit{ or } N_{var}} \end{array} \\
\textit{chromosome} & = & \left(\begin{array}{cccc} 010101 & 101010 & ..... & 110010 \end{array} \right) \\
(= \textit{individual}) & = & \textit{a set of genes} = \textit{a set of encoded unknowns} \\
 & \leftrightarrow & \{q_1/C_1,\ q_2/C_2, ..., \ q_{N_{event}}/C_{N_{event}}\} \\
 & \rightarrow & \textit{A distribution of } \{m(p_i, q_i | C_i)\}
\end{array}
$$

$$
\begin{array}{rcl}
\textit{population} & = & \textit{set of chromosomes} \leftrightarrow \textit{set of solutions}
\end{array}
$$

$$
\begin{array}{rcl}
\textit{fitness} & = & \textit{quality of a solution}
\end{array}
$$

Generic Flow-chart of GA

1. Encode Chromosomes
    1.1 Binary encoding
    1.2 Permutation encoding
    1.3 Continous encoding
    1.4 Tree encoding
2. Generating initial population
3. Evaluation of fitness values of all chromosomes
4. Selection and Mating
    4.1 Roulette Wheel Selection in proportion to its fitness
5. Reproduction (Crossover, Mutation)
6. Convergence test (go to 3, if not converged)



a binary code represents each dog

dogs bark and receive rating

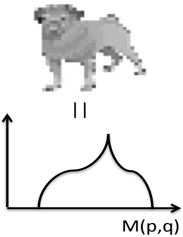
dogs selected for mating

offspring result & replace poor performers in population

introduce mutations

[101101 ...]

= {q_1, q_2, ... q_N}



Reproduction operators

- Crossover : random points are chosen on the individual chromosomes and the genes are exchanged at this point. This is the first operation where the GA explores another points in the variable space.
e.g.) single point crossover

$$\begin{array}{c|c} 1010101 & 0001 \\ 0101010 & 1110 \end{array} \tag{1}$$

$$\rightarrow \quad \begin{array}{c|c} 1010101 & 1110 \\ 0101010 & 0001 \end{array} \tag{2}$$

- Mutation : this is the process by which a bit/string is deliberately changed so as to maintain diversity in the population set

$$\begin{aligned} \#\, of\ mutations &= \mu \times (N_{pop} - N_{elite}) \times N_{bits}, \\ \mu &= mutation\, rate \end{aligned}$$

GA for the reconstruction of HEP events

1. Grouping 4 visible momenta into 2 pairs - (a1,b1), (a2,b2) by their origins - 2 on-shell mother particles, each decays to (a_i, b_i), without the knowledge of M_{mother}.
   ($N_{event} = 100$, $N_{pop} = 1000$, Elite rate $= 0.2$, $\mu = 0.25$)
   (Fitness = Height of the peak)
   1st generation :
   Best chromosome : 0001110........2112020
   fitness : 98 , Nfalse : 51 , NEvent : 100
   ....
   42nd generation :
   Best chromosome : 0000..1..00000000000
   fitness : 198 , Nfalse : 1 , NEvent : 100
   Running time - Real Time : 3.846, Cpu : 3.610 s

Gain :

1. Naive search : # of fitness estimation $\sim 3^{100} = 59049^{10}$
   → may not be possible before the end of the universe.

2. Stochastic search using GA : # of fitness estimation $\sim N_{pop}/2 \times N_{gen} = 500x42 = 21000$!!!
   → just 3-sec using an old single core.

Conclusion

- We are testing various collective stochastic optimizations by employing genetic algorithms (GA), in search for the best random sequence (of HEP event unknowns) which best fits a physical target distribution (from likelihoods).

- In this process, the best random sequence can have good correlation with the true sequence. We showed such a good solution (as an extremum searched by GA) really exists, in the case of (singular) on-shell mass functions of a random sequence realizing the maximal singularity(density) collectively.

- The GA, based on the principe of *Natural Selection*, is found to be very efficient and powerful for implementing the (complicated) collective stochastic optimizations for HEP event reconstruction.