Large scale data analysis in pursuit of gravitational waves

Vladimir Dergachev
Max Planck AEI/Golm
Continuous gravitational waves

• We know of many rotating neutron stars with frequencies from below 1 Hz to more than 700 Hz
• Gravitational radiation is expected to be emitted at twice the frequency
• Not all rotating neutron stars have to emit radio waves or X rays
• How do we find a nearby source?
Observation of a single target

• Detection of signals from a single target is relatively straightforward:
  • Record data
  • Compute correlation with known waveform
  • Analyze significance
• This procedure is not computationally limited, so we can rely on well-known statistical techniques.
Blind search

- Blind search requires large computing power:
  - Record data
  - Compute correlation with known waveform
  - Analyze significance
  - Sweep large area of parameter space
- Pure noise triggers events at $5\sigma$ level, need to pay close attention to computational efficiency
Blind search of large datasets:

- Record data
- Use carefully chosen algorithm for signal detection
- Analyze significance
- Sweep large area of parameter space

- Pure Gaussian noise triggers events at 6σ level
- Real data produces >7σ artifacts
- Signal template computation becomes challenging
- Computation efficiency becomes as important as statistical efficiency.

Phys. Rev. D 94, 042002, ...
Computational power

• In the near future we can expect access to clusters of ~1000 CPUs delivering ~1 TFlop each

• One year of run time provides $3 \cdot 10^{22} = 0.05$ mol of floating point operations.

• How can we use this computational power wisely?
  • For weak signals we really have to look at each template.
    Figure of merit: cycles per template.
  • Storing (even in RAM) is expensive.
    Aggregate data for multiple templates.
Loosely coherent search

- Analyze data for a set of nearby templates at once.
- Report cumulative statistics for the entire set:
  - Is there a signal?
  - Data quality?
- Interesting mathematical problem. Best algorithm depends on
  - set of signal templates
  - collected statistics
  - computing hardware

V Dergachev 2010 Class.
Quantum Grav. 27 205017
Example: search for CW gravitational signals

- Searches for continuous gravitational wave signals typically use 6-24 months of data (2 TB)
- We search millions of sky templates, millions of frequency templates and thousands of auxiliary parameters. **Computationally limited.**

Semi-coherent version runs on Einstein@Home

<table>
<thead>
<tr>
<th></th>
<th>Cycles per template</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComputeFStatistic</td>
<td>~ 1000000</td>
</tr>
<tr>
<td>(plain matched filter)</td>
<td></td>
</tr>
<tr>
<td>Resampling</td>
<td>~ 20000</td>
</tr>
<tr>
<td>Loosely coherent δ=0</td>
<td>&lt;1400</td>
</tr>
</tbody>
</table>

Loosely coherent searches for sets of well-modeled signals


Loosely coherent algorithm can cover $2 \cdot 10^{19}$ templates using 1 year of petaflop cluster CPU time. Efficiency increases with dimension of parameter space.
“Universal statistics”  
(Phys. Rev. D 87, 062001)

- It is not practical to design a separate statistics for each non-Gaussian artifact – usually there are too many artifacts to follow up and there might not be enough data.

- Instead, we need a “universal statistic” that would compute a valid figure, such as an upper limit, regardless of underlying distribution and would be close to optimal for the common case of Gaussian noise.

- Such statistic can be designed starting with Chebyshev's or Markov's inequality.
Are universal statistics possible?

- Example: Markov's inequality:

\[ E |X| = \int |X| dF_x \geq a \ P(|X| \geq a) \]

\[ P(|X| \geq a) \leq \frac{E|X|}{a} \]

This is just one step in Lebesgue integration.

The formula is valid for any random variable \( X \). But not efficient for Gaussian \( X \) – a 0.95% upper limit would set a 20\( \sigma \) threshold. It is exact for a Bernoulli variable.
A practical solution

We approximate a step function with a continuous monotonic function:

- Let \( Z = \frac{-X - \text{EX}}{(L \times \text{sd}(X))} \), where \( L \) is 95% quantile for Gaussian distribution
- Let \( Y = 0 \) when \( Z < 1 \), otherwise we define \( Y = \frac{Z - 1}{Z + 2} \)
- Note that we defined \( Y \) and \( Z \) to produce a limit on negative tail of distribution alone – this is optimal for establishing upper limits in power.
- Then by Markov's inequality

\[
P(Y \geq a) \leq \frac{EY}{a}
\]

After computing \( EY \) we pick \( a \) so that \( EY/a = 0.95 \) and then backtrack actual bound on \( X \) from definition of \( Y \).

This produces upper limits that are within 0.5% of what one would get with a conventional confidence level for Gaussian distribution.
A test of universal upper limit statistic

- Universal upper limit is efficient to compute
- Quantile
- Conventional sd and mad based upper limits
Summary

- Large scale analysis exposes non-Gaussian behaviour
- Can construct *universal statistics* that produce reliable upper limits without assuming specific distribution of data.
- Can make *universal statistics* close to optimal for the Gaussian case (common in data), while being correct for arbitrarily distributed data points and computationally efficient.
- *Loosely coherent* algorithms are used for computationally limited analysis of big data.
End of talk
Some *Loosely Coherent* ideas

- Consider a subspace containing a single set of nearby templates.
- The subspace is usually hard to describe exactly, but one can construct an over-determined basis that covers it.
- Evaluate groups of templates as a single linear operator in this basis (for example a convolution).
- Speed of operator evaluation determines efficiency of the search.

Loosely coherent searches for sets of well-modeled signals
Markov's inequality is exact for Bernoulli distribution

Suppose Y is 0 with probability \((1-\varepsilon)\) and 1 with probability \(\varepsilon\). Then

\[ P(|Y|\geq 1) = \varepsilon = E|Y| \]

To make an optimal upper limit for Gaussian case we need to convert Gaussian variable X to a Bernoulli-like form.

We introduce Y which is 0 whenever \( |X - EX| \leq L \text{sd}(X) \) and 1 otherwise. Then

\[ P(|X - EX| \geq L \text{sd}(X)) \leq EY \]

In theory this formula provides nice exact bounds for Gaussian X. In practice, expectations are computed from real data and we need to iterate over L until we have \( EY = 0.95 \)
Upper limit validation for universal statistic

- 17000 injections into frequency bands from 400-1500 Hz
- Random positions on the sky
- Gaussian and non-Gaussian bands
- Red line – injection strength
- Blue circles – upper limits
A test of universal upper limit statistic.

- For Gaussian bands (green circles) our universal statistic has very close upper limit
- For non-Gaussian (blue square) it provides a valid result
- Correct upper limit for hardware injection (pulsar 2)
Observation of multiple targets

- Multiple targets are slightly trickier:
  - Record data
  - Compute correlation with known waveform
  - Analyze significance
  - Repeat for all targets of interest
- Computing power starts to scale, need to pay attention to trials factor