Approximating Decomposed Likelihood Ratios using Machine Learning

Juan G. Pavez S.
Universidad Técnica Federico Santa María
In collaboration with Kyle Cranmer (NYU)
ACAT 2016-Valparaíso
Outline

• **The problem**: How can we make likelihood based inference when the likelihood function is unknown?.

• **Solution**: Approximating likelihood ratios using machine learning and decomposing likelihood ratios.

• **Applications**: Signal vs. Background hypothesis testing, maximum likelihood estimation of parameters.

• **EFT Morphing**: Estimating coupling parameters for Higgs EFT Morphing.
The problem

Likelihood based inference when the likelihood function is unknown.
Statistics for Discovery

- **Likelihood ratios** are one of the main tools used in HEP when reporting results from an experiment.

- They also allow the incorporation of **systematics** effects (using the profile likelihood ratio).

5 SIGMA -> DISCOVERY!

Higgs signal on pair of photon mass.
The **Neyman-Pearson Lemma**:

\[
\varphi(x) = \frac{p(x | \theta_0)}{p(x | \theta_1)} < k_\alpha
\]

Given a null hypothesis \( \theta_0 \) and an alternative hypothesis \( \theta_1 \), this test is the most powerful test.

We can define the **Likelihood ratio** test as:

\[
\Lambda(x) = \frac{L(\theta_0 | x)}{L(\theta_1 | x)}
\]

Where \(-2 \ln \Lambda(x)\) asymptotically follows a \( \chi^2_p \) distribution with degrees of freedoms equals to the difference on size of \( \theta_1 \) and \( \theta_0 \).
Statistics for Discovery

- Likelihood ratios are used extendedly on HEP:
  - Search and discovery (Hypothesis testing).
  - Parameter estimation (Maximum Likelihood).
  - Limits (Confidence Intervals).
- The problem:
  - Most of the times the Likelihood function is unknown.
  - We work with complex simulated multidimensional data where estimation is computationally intractable.

Parameter estimation of the Higgs mass using log-likelihood ratios.
Solution

Approximating decomposed likelihood ratios using machine learning.
Machine Learning in HEP

- Classification algorithms have become a standard tool on HEP.
- The goal: Classify Signal events vs. Background events.
- **TMVA** is the common choice when applying machine learning in HEP.
- In recent years the collaboration between both fields (**ML** and **HEP**) has increased a lot:
  - Higgs Boson Challenge on Kaggle.
  - Flavours of Physics (charged lepton flavour violation) on Kaggle.
  - ALEPH workshop at NIPS.

How different classifiers classify the same data.
Approximating Likelihood Ratios using Machine Learning

• Noticeable we can use the all the power of machine learning to approximate Likelihood Ratios (Kyle Cranmer, 2015):
  
  • Given a classifier score \( s(x; \theta_0, \theta_1) \) trained to classify between signal and background data.
  
  • Let \( p(s(x; \theta_0, \theta_1) \mid \theta) \) the probability distribution of the score variable conditioned by \( \theta \).
    – This distribution can be estimated using histograms or any estimation technique.

\[
\begin{align*}
X &= [x_1, x_2, \ldots] \\
\text{Classifier} &\quad s(X)
\end{align*}
\]

Dimensionality Reduction
Approximating Likelihood Ratios using Machine Learning

• Then it can be proved that the likelihood ratio:

\[ T(D) = \prod_{i=1}^{N} \frac{p(x_i | \theta_0)}{p(x_i | \theta_1)} \]

• Is equivalent to the ratio:

\[ T(D) = \prod_{i=1}^{N} \frac{p(s(x; \theta_0, \theta_1) | \theta_0)}{p(s(x; \theta_0, \theta_1) | \theta_1)} \]

• If the function \( s(x; \theta_0, \theta_1) \) is monotonic with the ratio.

\[ s(x) \approx monotonic \left( \frac{p(x | \theta_1)}{p(x | \theta_0)} \right) \]

Most of the commonly used classifiers approximate a monotonic function of the ratio!
Deecomposing Likelihood Ratios

- Using this result we can derive another result very useful in many applications.
- Often want to separate a signal from various backgrounds, where the signal is only a small perturbation of the only backgrounds hypothesis.
- Another application is to test for
  - Null: SM Higgs + Bkg.
  - Alternate: BSM Higgs + Bkg.
Decomposing Likelihood Ratios

• Formally we can define a **mixture model** as:

\[
p (x \mid \theta) = \sum_i w_i(\theta)p_i(x \mid \theta)
\]

• Where \(w_i(\theta)\) define the contribution of each distribution to the full model.

• Then the Likelihood ratio between two mixture models is:

\[
\frac{p (x \mid \theta_0)}{p (x \mid \theta_1)} = \frac{\sum_i w_i(\theta_0)p_i(x \mid \theta_0)}{\sum_j w_j(\theta_1)p_j(x \mid \theta_1)}
\]

• Which is equivalent to *(Cranmer, 2015)*:

\[
\frac{p (x \mid \theta_0)}{p (x \mid \theta_1)} = \sum_i \left[ \sum_j \frac{w_j(\theta_1)}{w_i(\theta_0)} \frac{p_j(s_{i,j}(x;\theta_0,\theta_1) \mid \theta_1)}{p_i(s_{i,j}(x;\theta_0,\theta_1) \mid \theta_0)} \right]^{-1}
\]
Decomposing Likelihood Ratios

• Now, the likelihood ratio is decomposed into distributions of pairwise trained classifiers.
• Moreover, in the common case that the pairwise distributions $p(s_{i,j}(x; \theta_1, \theta_0) \mid \theta)$ are independent of $\theta$ the only free parameters are $w_i(\theta)$.
• It is possible to estimate using maximum likelihood the signal or background contributions. Keeping $\theta_0$ fixed:

$$\hat{\theta}_1 = \arg\max_{\theta_1} \prod_{e=1}^{n} \frac{p(x_e \mid \theta_1)}{p(x_e \mid \theta_0)}$$
Applications

Signal vs. Background hypothesis testing, maximum likelihood estimation of parameters.
Applications

- First, consider a **simple mixture model** of 1-dim distributions.
- One of the distributions correspond to **signal** while the others are **background**.

![Graphs showing single and composed distributions](image)
Applications

- We fit the **pairwise classifiers** and a **single classifier** (both MLPs) and then compute the approximated likelihood ratios and compare it to the **true ratio**.

```plaintext
Juan Pavez (UTFSM)  ACAT 2016
```
Applications

- We do the same but now with a **much harder** model: Three 10-dim distributions, each one is a mixture of Gaussians.
Applications

• Again we can vary the signal contribution and observe that the decomposed model has optimal results even for very small signal (0.5%).

<table>
<thead>
<tr>
<th>AUC</th>
<th>Truth</th>
<th>Dec.</th>
<th>F0/F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.86</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>5%</td>
<td>0.86</td>
<td>0.85</td>
<td>0.72</td>
</tr>
<tr>
<td>1%</td>
<td>0.84</td>
<td>0.83</td>
<td>0.56</td>
</tr>
<tr>
<td>0.5%</td>
<td>0.80</td>
<td>0.79</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Applications

- It is possible to fit the signal contribution or background contributions (or both) values using maximum likelihood on the approximated likelihood ratios.

Fit for one single pseudo-experiment for:
A) Signal contribution.
B) Signal and Bkg.

Histogram of fit of many pseudo-experiments for:
A) Signal contribution.
B) Bkg contribution.
Applications

- An open source **Python** package (**Carl**) has been implemented by **Gilles Louppe** allowing to easily use approximated likelihood ratio inference.

[github.com/diana-hep/carl]
Estimating coupling parameters for Higgs EFT Morphing.
Higgs EFT Morphing

• It is possible to expand the **Standard Model Lagrangian** adding **non-SM** couplings of the Higgs boson to SM particles using an **effective field theory (EFT)** approach.

\[
L_0^V = \left\{ c_\alpha \kappa_{SM} \left[ \frac{1}{2} \bar{g}_{HZZ} Z_{\mu} Z^{\mu} + \bar{g}_{HWW} W_{\mu}^+ W^{-\mu} \right] \right. \\
- \frac{1}{4} \left[ c_\alpha \kappa_{HYY} \bar{g}_{HYY} A_{\mu \nu} A^{\mu \nu} + s_\alpha \kappa_{AYY} \bar{g}_{AYY} A_{\mu \nu} A^{\mu \nu} \right] \\
- \frac{1}{2} \left[ c_\alpha \kappa_{HZY} \bar{g}_{HZY} Z_{\mu \nu} A^{\mu \nu} + s_\alpha \kappa_{AZY} \bar{g}_{AZY} Z_{\mu \nu} A^{\mu \nu} \right] \\
- \frac{1}{4} \left[ c_\alpha \kappa_{Hgg} \bar{g}_{Hgg} G^a_{\mu \nu} G^{a \mu \nu} + s_\alpha \kappa_{Agg} \bar{g}_{Agg} G^a_{\mu \nu} G^{a \mu \nu} \right] \\
- \frac{1}{2} \left[ c_\alpha \kappa_{HZZ} Z_{\mu \nu} Z^{\mu \nu} + s_\alpha \kappa_{AZZ} Z_{\mu \nu} Z^{\mu \nu} \right] \\
- \frac{1}{2} \left[ c_\alpha \kappa_{HWW} W_{\mu}^+ W^{-\mu} + s_\alpha \kappa_{AWW} W_{\mu}^+ \bar{W}^{-\mu} \right] \\
- \frac{1}{4} \left[ \kappa_{H0y} Z_{\mu} \partial_{\mu} A^{\mu} + \kappa_{H02} Z_{\mu} \partial_{\mu} Z^{\mu} + \kappa_{H0W} (W_{\mu}^+ \partial_{\mu} W^{-\mu} + h.c.) \right] \chi_0.
\]

**Effective Lagrangian of spin-0 particle to gauge bosons.**

• This allows to search for deviation from **SM** predictions for Higgs boson properties.

• While in **Run 1** a few number of coupling parameters was considered is expected than in **Run 2** many more **BSM** parameters will be considered.
**Problem**: Distributions of observables is dependant on the parameters of the EFT.

**Morphing Method**: Provides description of any observable as a linear combination of a minimal set of orthogonal base samples.

\[
S(g_1, g_2, \ldots) = \sum_{i=1}^{N} w^{(i)}(g_1, g_2, \ldots) S(g_1^{(i)}, g_2^{(i)}, \ldots)
\]

Simple morphing procedure for one BSM coupling in decay or production. A Morphing Technique for signal modelling in a Multidimensional space of non-SM coupling parameters.

cds.cern.ch/record/2066980
Higgs EFT Morphing

• The team working on Morphing has provide us samples for VBF production $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ and $H \rightarrow WW^* \rightarrow 4l$.

• **15 samples and 5 samples** are needed in the base.

• **Problem:** In areas of the coupling space not covered by the base of samples statistical fluctuations increase a lot. 
  
    $\Rightarrow$ This affect the fitting procedure

• **Solution:** Choose a pair of **orthogonal** sets of bases and sum as:

$$B_{full} = \alpha_1(g_1, g_2, g_3)B_1 + \alpha_2(g_1, g_2, g_3)B_2$$

• Good coverture of the coupling space while allowing a **smooth** change of bases when fitting.
**Higgs EFT Morphing**

$H \rightarrow WW^* \rightarrow 4l$

1 BSM

- Likelihood fit for coupling parameters using decomposed approximated likelihood ratios.
- Fitting sample $S(1.,1.5)$. Good agreement with real values.

**SM coupling**

![Graph of SM coupling](image)

**BSM coupling**

![Graph of BSM coupling](image)
Higgs EFT Morphing

\[ H \rightarrow WW^* \rightarrow e\nu\mu\nu \]

Preliminary Results

2 BSM

Likelihood ratio values for Kazz,KHzz, target=(-0.50,-0.33)

Likelihood ratio values for Kazz,KHzz, target=(0.33,0.14)

Likelihood ratio values for Kazz,KHzz, target=(0.33,0.20)

Likelihood ratio values for Kazz,KHzz, target=(1.00,0.50)
Conclusions
Conclusions & Future Work

• Machine Learning approximation of Likelihood ratios is a great alternative for Likelihood estimation methods, such as **Approximate Bayesian Computation (ABC)**.
  – This method allows to use all recent advances on Machine Learning (e.g. Deep Learning, Tree-based methods, SVMs).
• The technique is also an alternative to **MEM (Matrix Element Method)**, but faster since this approach is not event-based (but need many simulated samples).

• We need to understand how **errors** (e.g. training error, poison fluctuations from histogram estimation) affects the **final approximation**.
• Integration with common tools used in **HEP** might be needed (we have been working mainly with python frameworks).
Thank You!

All the code and plots on this presentation can be found at: github.com/jgpavez/systematics
Backup
• Datasets for VBF $H \rightarrow WW^* \rightarrow 4l$ with one BSM coupling.

\( S(1,0)/S(1,2) \)  
\( S(1,2)/S(1,3) \)  
\( S(1,1)/S(1,3) \)
• Score distributions for pairwise trained classifiers (BDTs) for VBF $H \rightarrow WW^* \rightarrow 4l$ datasets.
Backup

- Studies on how the quality of the classifier training affect the final approximation.
- Expected: More training data -> Better Approximation (Better fits).
- Results for 10-dim harder model.
### Backup

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>$L[y, f(x)]$</th>
<th>Minimizing Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial Deviance</td>
<td>$\log[1 + e^{-yf(x)}]$</td>
<td>$f(x) = \log \frac{\Pr(Y = +1</td>
</tr>
<tr>
<td>SVM Hinge Loss</td>
<td>$[1 - yf(x)]_+$</td>
<td>$f(x) = \text{sign}[\Pr(Y = +1</td>
</tr>
<tr>
<td>Squared Error</td>
<td>$[y - f(x)]^2 = [1 - yf(x)]^2$</td>
<td>$f(x) = 2\Pr(Y = +1</td>
</tr>
<tr>
<td>“Huberised” Square Hinge Loss</td>
<td>$-4yf(x), \quad yf(x) &lt; -1$</td>
<td>$f(x) = 2\Pr(Y = +1</td>
</tr>
<tr>
<td></td>
<td>$[1 - yf(x)]^2_+$</td>
<td></td>
</tr>
</tbody>
</table>

Classifiers minimizing functions. The Elements of Statistical Learning, Hastie et al.
Bibliography


• ATLAS Collaboration (Belyaev, K. et al.). A morphing technique for signal modelling in a multidimensional space of non-SM coupling parameters. cds.cern.ch/record/2066980.