Direct optimization of discovery significance

A. Mohamed, D. Krüecker, I. Melzer-Pellmann, I. Pidhurskyi, O. Turkot

August 2019

Contents

Reminder

Statistical significance approximation

- $s \approx$ True Positive (TP) rate
  $b \approx$ False Negative (FN) rate
  σ_b – background uncertainty
- Asimov significance:

$$Z_A = \sqrt{2\left((s+b)ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2}ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)}$$

- given $b \gg s,\ \sigma_b \propto b$: $Z_A \to \frac{s}{\sqrt{s+b}}$

Loss

Given these approximations for significance, loss functions are defined as:

- $Z_A \rightarrow l_{Asimov} = \frac{1}{Z_A^2}$
- $\frac{s}{\sqrt{s+b}} \rightarrow l_{s/\sqrt{s+b}} = \frac{s+b}{s^2}$ – *but it has been shown that crossentropy is better*

Summary

Given these approximations for significance, loss functions are defined as:

- $Z_A \rightarrow l_{Asimov} = \frac{1}{Z_A^2}$
- ~~$\frac{s}{\sqrt{s+b}} \rightarrow l_{s/\sqrt{s+b}} = \frac{s+b}{s^2}$~~

DNN:

- Two hidden layers with ReLU as activation function.
- Binary classification:
  - Sigmoid in output layer.
- Multiclass
  - 4 signal classes: tt semi-leptonic, tt di-leptonic, W jets and actual signal, tt $\tilde{\chi}^o$.
  - Softmax(4) for output layer.

Summary

Given these approximations for significance, loss functions are defined as:

- $Z_A \rightarrow l_{Asimov} = \frac{1}{Z_A^2}$
- $\frac{s}{\sqrt{s+b}} \rightarrow l_{s/\sqrt{s+b}} = \frac{s+b}{s^2}$

DNN:

- Two hidden layers with ReLU as activation function.
- Binary classification:
  - Sigmoid in output layer.
- Multiclass
  - 4 signal classes: tt semi-leptonic, tt di-leptonic, W jets and actual signal, tt $\tilde{\chi}^o$.
  - Softmax(4) for output layer.

To train DNN:

1. Pretrain with crossentropy.
2. Finalize with l_{Asimov}.

Sample weights

Definition for s and b

- $s = W_s \sum_{sig} y^{pred}$, $W_s = \frac{L\sigma_{sig}\epsilon}{N_{sig}}$
  $b = W_b \sum_{bg} y^{pred}$, $W_b = \frac{L\sigma_{bg}\epsilon}{N_{bg}}$
  (definition used in research of A. Eldwood and D. Krücker)

Definition for s and b

- $s = W_s \sum_{sig} y^{pred}$, $W_s = \frac{L\sigma_{sig}\epsilon}{N_{sig}}$
  $b = W_b \sum_{bg} y^{pred}$, $W_b = \frac{L\sigma_{bg}\epsilon}{N_{bg}}$
- Resulting equations for s and b:
  $s = L\sigma_{sig}\epsilon < y^{pred}_{sig} >$
  $b = L\sigma_{bg}\epsilon < y^{pred}_{bg} >$

Definition for s and b

- $s = W_s \sum_{sig} y^{pred}$, $W_s = \frac{L\sigma_{sig}\epsilon}{N_{sig}}$
  $b = W_b \sum_{bg} y^{pred}$, $W_b = \frac{L\sigma_{bg}\epsilon}{N_{bg}}$
- Resulting equations for s and b:
  $s = L\sigma_{sig}\epsilon < y^{pred}_{sig} >$
  $b = L\sigma_{bg}\epsilon < y^{pred}_{bg} >$
- But in our case we have separate weights per event $\Rightarrow$ *must modify equations.*

Apply sample weights

- $s = W_s \sum_{i=\{sig\}} w_i y_i^{pred}$, $W_s = \frac{L}{N_{sig}}$
  $b = W_b \sum_{i=\{bg\}} w_i y_i^{pred}$, $W_b = \frac{L}{N_{bg}}$

Apply sample weights

- $s = W_s \sum_{i=\{sig\}} w_i\, y_i^{pred}$, $W_s = \frac{L}{N_{sig}}$
  $b = W_b \sum_{i=\{bg\}} w_i\, y_i^{pred}$, $W_b = \frac{L}{N_{bg}}$
- Resulting equations for s and b:

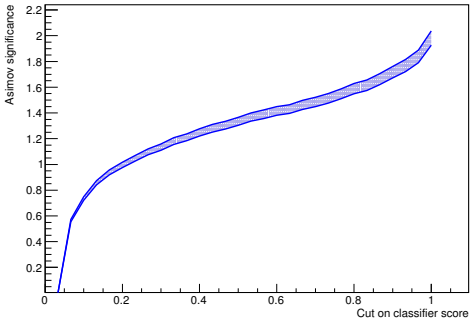
$$s = L\frac{\sum_{i=\{sig\}} w_i y_i^{pred}}{N_{sig}} = L < w_{sig} > \frac{\sum_{i=\{sig\}} \tilde{w}_i y_i^{pred}}{N_{sig}}$$

$$= L < w_{sig} >< y_{sig}^{pred} >_{weighted} = L < \sigma_{sig}\epsilon >< y_{sig}^{pred} >_{weighted}$$

$$b = L < \sigma_{bg}\epsilon >< y_{bg}^{pred} >_{weighted}$$

Apply sample weights

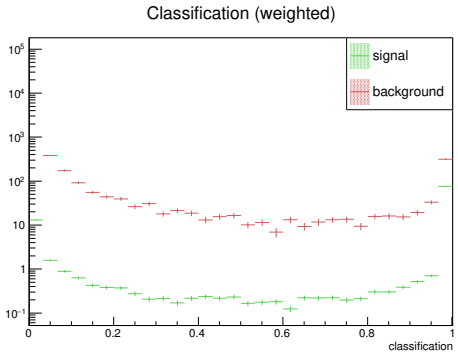
- $s = W_s \sum_{i=\{sig\}} w_i\, y_i^{pred}$, $W_s = \frac{L}{N_{sig}}$
  $b = W_b \sum_{i=\{bg\}} w_i\, y_i^{pred}$, $W_b = \frac{L}{N_{bg}}$
- Resulting equations for s and b:
  $s = L < \sigma_{sig}\epsilon >< y_{sig}^{pred} >_{weighted}$
  $b = L < \sigma_{bg}\epsilon >< y_{bg}^{pred} >_{weighted}$
- *. . . but it does not work*



$M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$

Apply sample weights

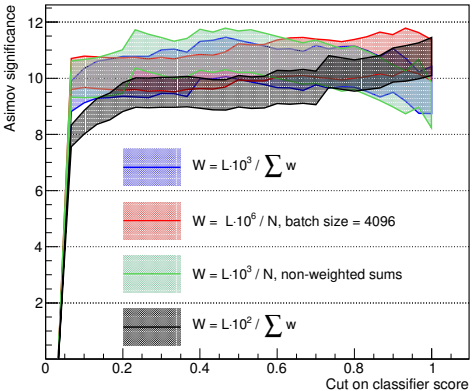
- $s = W_s \sum_{i=\{sig\}} w_i\, y_i^{pred}$, $W_s = \frac{L}{N_{sig}}$
  $b = W_b \sum_{i=\{bg\}} w_i\, y_i^{pred}$, $W_b = \frac{L}{N_{bg}}$
- Resulting equations for s and b:
  $s = L < \sigma_{sig}\epsilon >< y_{sig}^{pred} >_{weighted}$
  $b = L < \sigma_{bg}\epsilon >< y_{bg}^{pred} >_{weighted}$
- *. . . but it does not work*



$M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$

Black magic

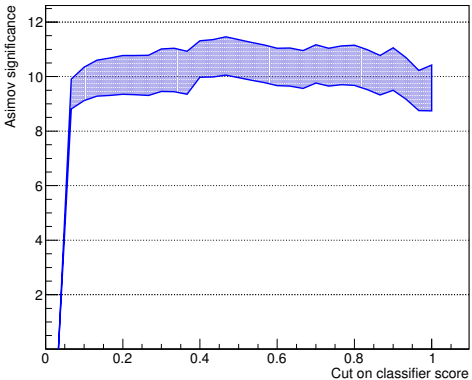
Powerful spells



$M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$

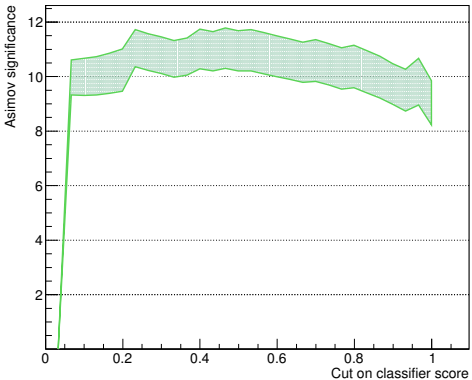
- $s = L\frac{\sum_{i=\{sig\}} w_i y_i^{pred}}{\sum_{i=\{sig\}} w_i} = L < y_{sig}^{pred} >_{wtd}$
- $s = L < \sigma_{sig}\epsilon >< y_{sig}^{pred} >_{wtd} \cdot 10^3$
- $s = L\frac{\sum_{i=\{sig\}} y_i^{pred}}{N_{sig}} = L < y_{sig}^{pred} >$
- $s = L < y_{sig}^{pred} >_{wtd} \cdot 10^{-1}$

Powerful spells



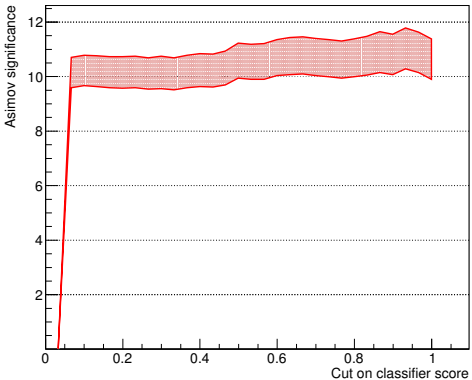
- $s = L\frac{\sum_{i=\{sig\}} w_i y_i^{pred}}{\sum_{i=\{sig\}} w_i} = L < y_{sig}^{pred} >_{wtd}$
- *For s to represent a number of events, we must have $s \propto L\sigma \Rightarrow s$ is **not** a number of events. And L here is just a magic hyper parameter:*
  - $L = 1 \rightarrow Z_A \approx 4.5$
  - $L = (35.9 \cdot 10^3) \cdot 10^{-1} \rightarrow max(Z_A) > 10$ ***(black plot)***

Powerful spells



- $s = L\frac{\sum_{i=\{sig\}} y_i^{pred}}{N_{sig}} = L < y_{sig}^{pred} >$
- *Issue is the same as for blue approach. But now we don't even use the weights.*
- *This is the way Z_A was applied for my previous reports. I.e. just pass sample weights to the Keras, close your eyes and hope it knows what to do.*

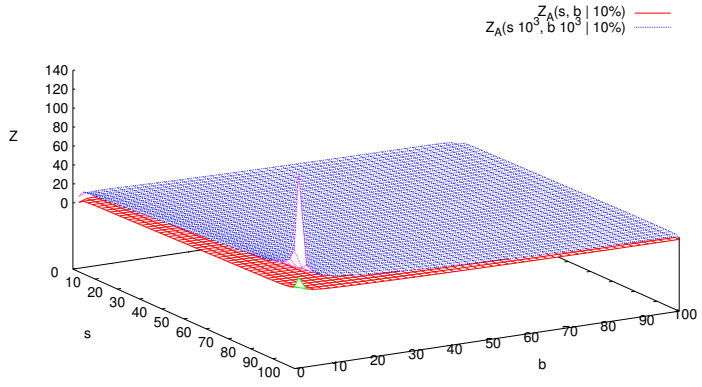
Powerful spells



- $s = L < \sigma_{sig}\epsilon >< y_{sig}^{pred} >_{wtd} \cdot 10^3$
- *Close to the correct definition, but...*
- *Issue is in the scale factor,* 10^3*:*
  Z_A *is non-liniar to the scale of* s *and* b*,*
  $\forall x, y, \lambda, \sigma_b : \; Z_A(\lambda x, \lambda y|\sigma_b) \neq \lambda' Z_A(x, y|\sigma_b)$

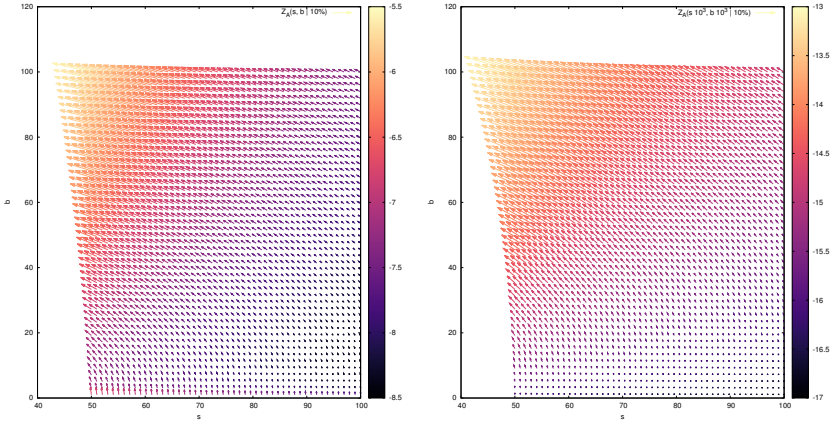
Scaling s and b

But is it really an issue?



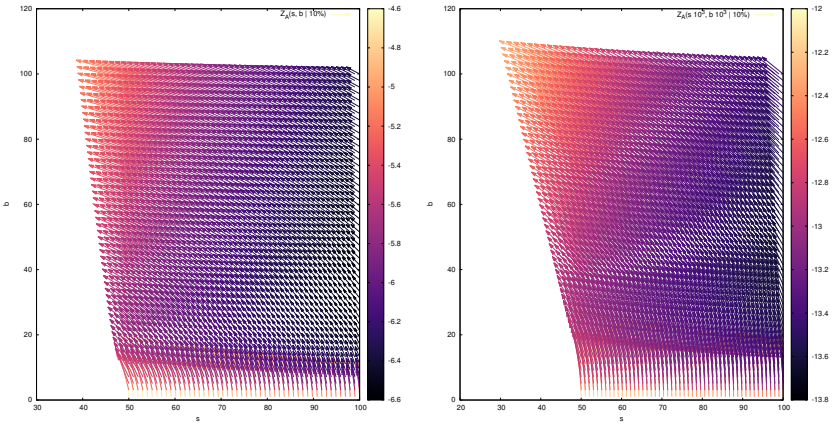
Asimov significance

Scaling s and b



Gradients for $l_{Asimov}(s, b|\sigma = 10\%)$ (left), and $l_{Asimov}(s \cdot 10^3, b \cdot 10^3|\sigma = 10\%)(right)$

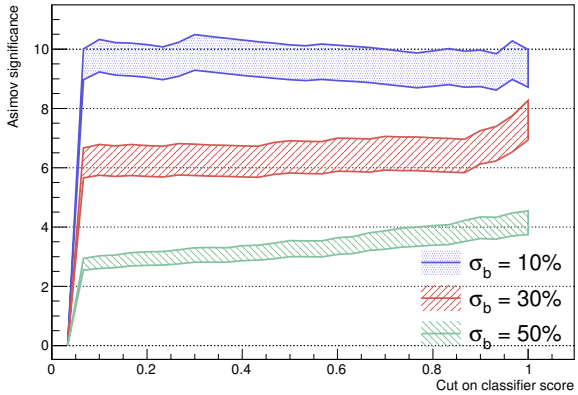
Scaling s and b



Gradients for $1/Z_A(s, b|\sigma = 10\%)$ (left), and $1/Z_A(s \cdot 10^3, b \cdot 10^3|\sigma = 10\%)(right)$

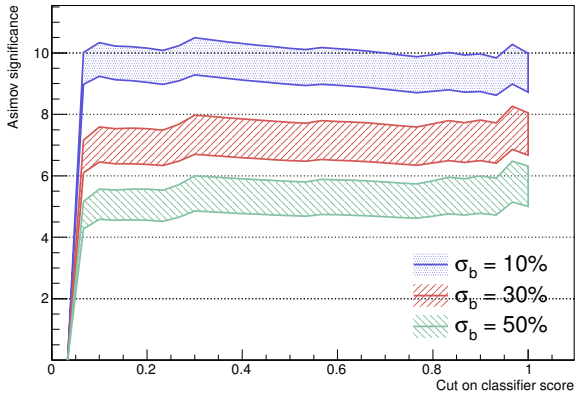
Binary classification

Red approach



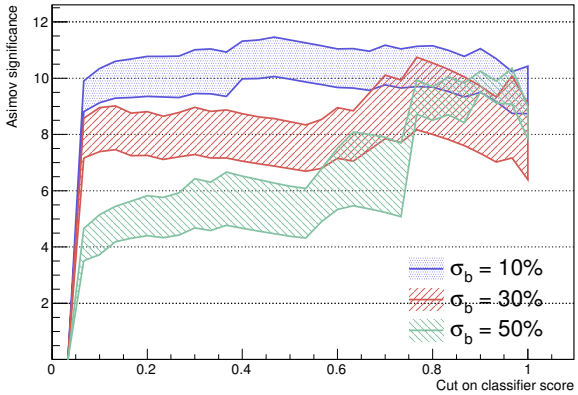
Results for different background uncertainties obtained with red approach.
$M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$.

Red approach



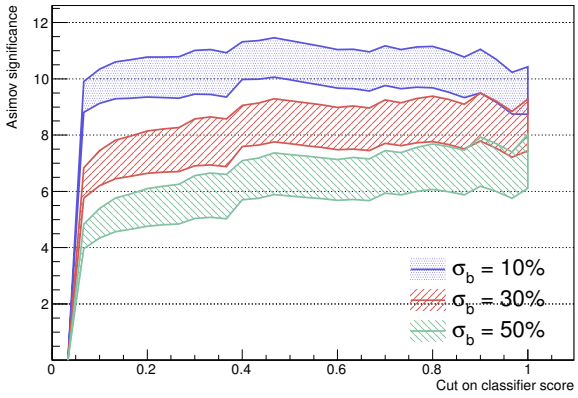
Now all plots are done for DNN trained with $\sigma_b = 10\%$.

Blue approach



Results for different background uncertainties obtained with blue approach.
$M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$.

Blue approach



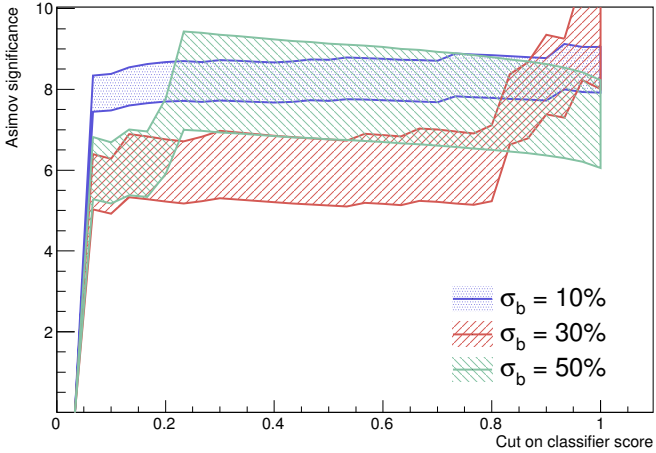
Now all plots are done for DNN trained with $\sigma_b = 10\%$.

Multiclass

Multiclass

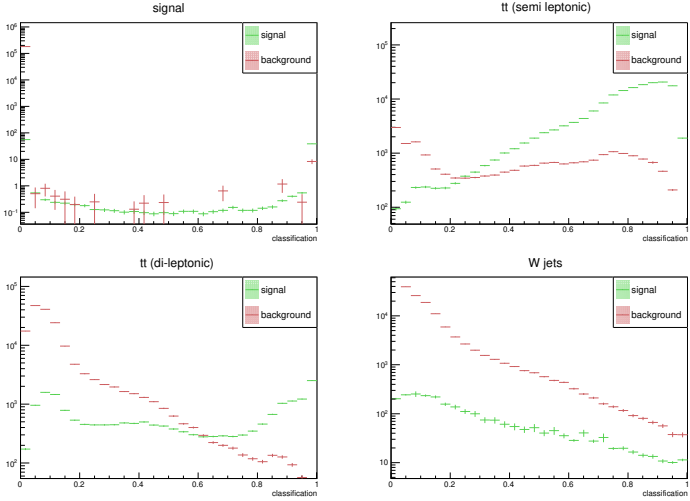
- $loss = l_{Asimov} \cdot 10^{10} + crossentropy \cdot 10^{-3}$
- l_{Asimov} is defined as in the blue approach.
- Coefficients for l_{Asimov} and $crossentropy$ are chosen to weight values of both parts to be of the same order.

Multiclass



Results with different background uncertainties. $M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$.

Multiclass



Classifier output for different classes. $M_{LSP} = 1000GeV$, $M_{\tilde{g}} = 1900GeV$, $\sigma_b = 10\%$.

Summary

Summary

- Binary classification
  - Asimov significance is decided to be used with:
    $s = L\frac{\sum_{i\in sig} w_i y_i^{pred}}{\sum_{i\in sig} w_i},\ b = L\frac{\sum_{i\in bg} w_i y_i^{pred}}{\sum_{i\in bg} w_i}$
  - *...unless during this meeting it will be decided to use another one.*
  - Obtained significance for test-data set is $\approx 10\sigma$, which impress, since it has only ≈ 99 signal events vs $\approx 182\ 627$ background events.
  - Try $l_{Asimov} = 1/Z_A$.
- Multiclass
  - Applied combination l_{Asimov} and crossentropy with l_{Asimov} defined as in binary classification.
  - Obtained significance is lower compared to binary classification -approach. But...
  - total loss in multiclass is very sensitive to the weighting-coefficients, thus...
  - hyper-parameter-optimization is the next step.
  - Maybe “W jets”-class should be removed.

...and sorry for black magic.