The terascale tutorial

Konstantinos Theofilatos[1]

Department of Physics
National and Kapodistrian University of Athens

Abstract

This note summarizes the lectures given in the tutorial session of the *Introduction to the Terascale* school at DESY on March 2023. The target audience are advanced bachelor and master physics students. The tutorial aims to best prepare the students for starting an LHC experimental physics thesis. The cross section of $t\bar{t}$ pair production is detailed alongside with the reconstruction of the invariant masses of the top quark as well as of the W and Z bosons. The tutorial uses ideas and CMS open data files from the `CMS HEP Tutorial` written by C. Sander and A. Schmidt, but is entirely rewritten so that it can be run in `Google` Colab Cloud in a columnar style of analysis with python. In addition, a minimal `C`/C++ version of a simple event-loop analysis relying on `ROOT` is exampled. The code is kept as short as possible with emphasis on the transparency of the analysis steps, rather than the elegance of the software, having in mind that the students will in any case need to rewrite their own custom analysis framework.

[1] comments to konstantinos.theofilatos@cern.ch, http://www.cern.ch/theofil

Contents

1 Introduction

In an learning by example approach, we will discuss how to measure the cross section of a physics process, which is known as top quark anti-quark pair ($t\bar{t}$) production.

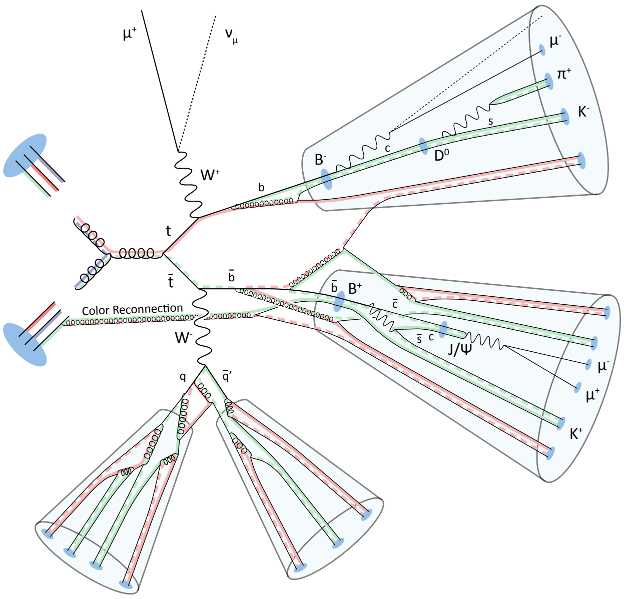

Figure 1: Artistic visualization of a $t\bar{t}$ produced by two colliding protons (on the left in blue) and decaying into a μ^{-} and hadrons that are later on clustered as jets. (Image credits: B. Stieger.)

We will make use of a pocket-size data sample that comes with the CMS HEP Tutorial [1], comprising of just a small fraction of pp collision data of 50 pb^{-1}at $\sqrt{s} = 7$ TeV. All data & MC simulation files as well as the code can be found in:

- http://theofil.web.cern.ch/theofil/cmsod/files/
- https://github.com/theofil/I2TheTerascale

The data have been selected among the many pp collisions occurring every second at the LHC, such as at least one muon is present in the collision debris. This type of selection has been made using the so-called single muon trigger of the CMS detector (see Sec. 8). Instead of a lengthy intro on the LHC and how a particle physics detector works, few video links below that need to be appreciated before moving forward.

- LHC YouTube video, absolute must see!
- CMS YouTube video, all what you need to know for getting started

In addition, a very nice introduction for collider physics has been written by M. D. Schwartz [2], should the students wish to dive deeper into the physics.

2 Physics Analysis

The most basic quantity we are interested in particle physics is called cross section (σ) for a particular particle interaction to occur. You could think the cross section of a process as the analogous of the probability for that process to take place, but instead being a pure number it is measured in units of area, 1 barn $= 10^{-28}m^2$.

The sample size of the LHC pp collision data is quantified by what is known as (integrated) luminosity measurement (L) and has units of inverse area (e.g., pb^{-1}), where p stands for the pico $= 10^{-12}$ order of magnitude. Smaller cross section area implies smaller chance for the interaction to occur. On the other hand, more pp collisions on tape, means more L. So we can probe a process of small σ if L is sufficiently large, provided that we have a way to select pp events enriched with the process of interest as in most cases a pp collision results into a "boring" final state.

2.1 The master equation: $N = \epsilon\sigma L$

The number of events (N) we expect for a specific process with known cross section (σ) in a data sample of known (integrated) luminosity (L) is:

$$N = \epsilon\sigma L \tag{1}$$

where ϵ is the total selection efficiency for recording this process, including both kinematic and geometric acceptance of the detector.

2.2 Physics Processes

While we have some control of the initial state, e.g., the center of mass energy of the colliding protons, we don't really control what comes out in the final state. It is like rolling a dice with an unknown number of faces and different frequencies for each of the possible outcomes. Provided that there is sufficient energy in the initial state, all possible paths (particle interactions) will be taken by nature with probabilities that governed (we believe) by the laws of quantum mechanics. During LHC Run II, for $\sqrt{s} = 13$ TeV and 20 nb$^{-1}/s$ *instantaneous* luminosity the production rate for different physics processes is shown below.

process	rate (Hz)
$W^{\pm}$	4000
Z^0	1200
$t\bar{t}$	17
h^0	1
h^0h^0	(0.007 ?)

Table 1: Expected production rate of different processes at the LHC Run II with $\sqrt{s} = 13$ TeV and 20 nb$^{-1}/s$ instantaneous luminosity. The last process has yet been confirmed and is one of the main goals of the LHC as it is particularly sensitive to the Higgs self-coupling.

In fact, those particles are produced by nature in an effortless manner for the given instantaneous luminosity and center-of-mass energy of the pp collisions. However, the particle detectors don't detect directly the very short lived particles listed above, but rather their decay products. We simply cannot speak at an event-by-event level that this event is Higgs, this is a W and so on, although surely we will hear people saying so when they look into beautiful event displays.

By applying selection criteria (analysis cuts) on the pp data, one can increase the efficiency of selecting a specific process (call it signal: S) against other processes (call them backgrounds: B) that will also satisfy the applied criteria mimicking the signal. Ideally, we would want the signal efficiency to be 100% while the backgrounds to have 0% efficiency. Unfortunately, this is almost never the case and there is always some background contribution in the sub-sample of data we selected to focus our attention. The amount of background events in our signal-enriched sample has to be estimated and MC simulation might be used for that purpose. It is therefore typical that together with the MC simulation of the signal we do also consider the background simulation, which is usually much more difficult to get correctly (i.e., having larger uncertainty on the predicted event yields due to theory uncertainties in its σ).

3 Event Weights, MC and Statistical Uncertainty

In practice, multiple processes contribute to the signal region, the number of events expected from MC (N_{MC}) is

$$N_{\mathrm{MC}} = \sum_{\mathrm{i}} \epsilon_{\mathrm{i}} \sigma_{\mathrm{i}} L \tag{2}$$

where the index i enumerates all simulated physics processes. Without any event selection, the total number of events that have been generated for the processes i is

$$N^{\mathrm{tot}}_{\mathrm{MC,i}} = \sigma_{\mathrm{i}} L_{\mathrm{i}} \tag{3}$$

where L_{i} is the simulated luminosity for the specific sample, which in general varies as function of the total number of computing hours used for the MC generation. In order to normalize all samples to the luminosity of data $L = 50\ \mathrm{pb}^{-1}$, we need to assign them appropriate weights

$$w_{\mathrm{i}} = \frac{\sigma_{\mathrm{i}} L}{N^{\mathrm{tot}}_{\mathrm{MC,i}}} = \frac{L}{L_{\mathrm{i}}} \tag{4}$$

where the index i is, as before, enumerating the simulated physics processes. To give an example, if $w_{\mathrm{i}} = 5$ we would need to count each entry (1 unweighted event) of the MC sample as 5 weighted MC events, when comparing simulation with data. On the contrary, if $w_{\mathrm{i}} = 0.1$ we need to count every 10 entries (10 unweighted events) of the MC as 1 weighted MC event. The statistical uncertainty of weighted (Poisson in nature) MC events is not just as simple $\sqrt{N}$ but is rather given by

$$\delta N^{\mathrm{sel}}_{\mathrm{MC}} = \sqrt{\sum_{j} w_j^2} \tag{5}$$

where now the index j counts all entries (unweighted events) of the MC processes (i) that contribute to a desired event selection.

3.1 Exercises

Assuming that we have B = 1000 (weighted) MC events, when applying the event selection of our signal region. Calculate what would be $\delta B/B$ if

1. all MC events have $w_j = 0.1$
2. all MC events have $w_j = 10$

assuming that δB is dominated by uncertainties of statistical nature, neglecting systematic uncertainties.

process	σ[pb]	triggerBit
data	–	always true
TTbar	165	true or false
WJets	31300	always true
DYJets	15800	always true
WW	4580	always true
WZ	3367	always true
ZZ	2421	always true
SingleTop	5684	always true
QCD	$\sim 10^8$	always true

Table 2: Cross section and trigger information for the MC samples [1].

4 Data and MC samples

In total for this tutorial, we have $N = 469384$ data events satisfying the single muon trigger, for an integrated luminosity of 50 pb^{-1}of pp collisions at $\sqrt{s} = 7$ TeV. Ideally, we would have wanted to have at least ×10 MC simulated events (i.e., 500 pb^{-1}of simulated luminosity), in order for the MC statistical uncertainty to be less than the one of the data. If that would have been the case, assign each MC event a weight of $w = 0.1$, i.e., counting each entry found in MC as 0.1 events. Unfortunately, this is not possible for processes with very large cross section where in practice we are only able to simulate much less events than those expected for 50 pb^{-1}. For these processes, the simulated luminosity is smaller than 50 pb^{-1}. Below follow the available weighted and unweighted events for data and all MC processes we will use in analysis.

```
Data: 469384.0 ± 685.1   [entries: 469384]
MC  : 331407.3 ± 55461.7 [entries: 240601]
-----------------------------
WJets  209576.7 ± 689.2  [entries: 109737]
DYJets  34113.2 ± 145.6  [entries: 77729]
TTbar  7928.6 ± 45.5     [entries: 36941]
WW  229.9 ± 3.7          [entries: 4580]
WZ  69.9 ± 1.3           [entries: 3367]
ZZ  16.9 ± 0.4           [entries: 2421]
Single Top  311.6 ± 4.4  [entries: 5684]
QCD  79160.5 ± 55457.2   [entries: 142]
-----------------------------
```

The first number corresponds to the number of weighted events and their uncertainty, while in [$entries$: ...] the number of entries (or unweighted events if you wish) is given. By construction we have $w = 1$ for data and the number of weighted events is equal to the number of unweighted events (entries) in this case. Note that when $w = 1$, the statistical uncertainty given by Eq. 5 reduces to $\sqrt{N}$. The QCD background has by far the largest event weight, for which 142 entries (unweighted events) correspond to as much as $79k$ events with very large statistical uncertainty $\sim 55k$. Already at this point, we get warned that this type of background should be filtered away by some event selection (cuts).

4.1 Exercises

If we define as our signal the final state of $t\bar{t} \to b\bar{b}q\bar{q}\mu^+\nu$, sketch in a piece of paper possible ways for which the background simulated process (DY+jets, TTbar, WW, WZ, ZZ, SingleTop, QCD) can mimic our signal.

5 Opening ROOT files

The LHC experiments use ROOT to analyze and *store* the information recorded during hadronic collisions. So, if we would want to study the interactions taking place during *pp* collisions, we need to learn how to read the ROOT files produced by the experiments. There are many ways to open a ROOT file, the most popular are:

1. Install ROOT[2]

2. Install uproot and awkward arrays.

In addition a third way we developed here, is to use the *Google Colab* suite and install there all python packages needed to run an analysis. This approach is the least optimal but has the fastest time-to-analysis for the students, since it does not need any installation to a local computer. See how this works, by opening openROOTFile.ipynb. The code above can be easily modified to open and analyze any ROOT file.

5.1 Exercises

1. Open "data.root" from http://theofil.web.cern.ch/theofil/cmsod/files/

2. Check what's inside the "TTree" named "events".

3. Count how many events have exactly 1 μ and at least 2 jets.

6 What's inside the ROOT files?

Inside the ROOT file we can find all the information needed to build Lorentz four-vectors

$$p^\mu = (E, p_\mathrm{x}, p_\mathrm{y}, p_\mathrm{z})$$

of the particles detected by the CMS detector (Fig 6). We assume as known the masses, of muons, electrons as well of the pions. We measure their momenta ($\vec{p}$) using the deposits they leave as they go through the detector. Particles are also grouped into jets

$$p^\mu_\mathrm{jet} = \sum_\mathrm{i} p^\mu_i$$

using a clustering algorithm to decide which particles (i) will be grouped together. The particle jets are usualy interpreted as the evolution of the partons (q, g) produced in the hard scatter, but it should be kept in mind that their four-momenta is not $1 - 1$ even in MC truth, due to the QCD color confinement as well as the ambiguities arising from the clustering itself.

[2] For Windows, see also these instructions in case you do not find your way with the official ones.

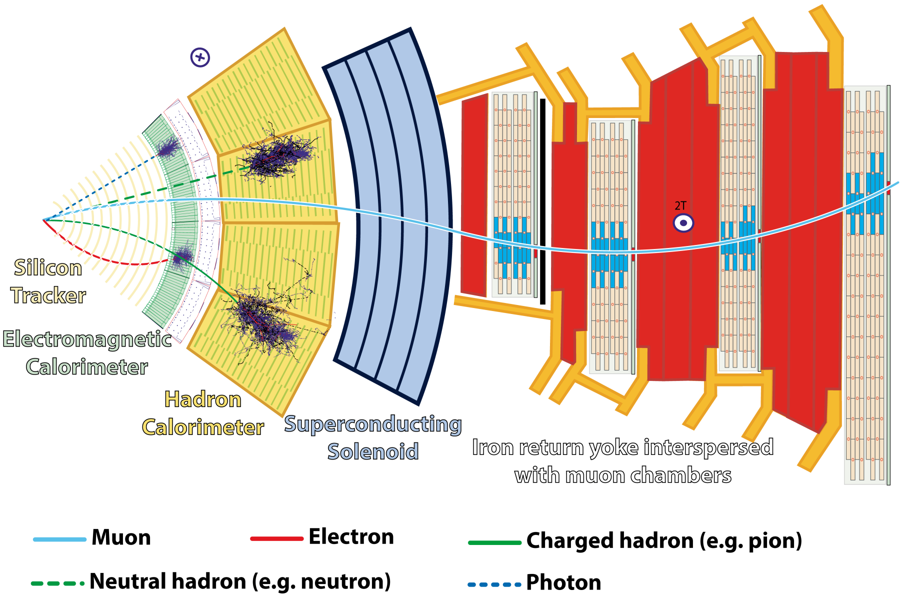
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with muon chambers
2T
Muon
Electron
Charged hadron (e.g. pion)
Neutral hadron (e.g. neutron)
Photon

Figure 2: Particles seen by the CMS detector.

The transverse momentum imbalance, with its magnitude best known as missing transverse energy $\not{E}_T$ or simply MET, is defined as

$$\vec{p}_{\mathrm{Tmiss}} = -\sum_{\mathrm{i}} \vec{p}_{\mathrm{T,i}}$$

where the index i (usually) runs over all visible particles that satisfy the experimental thresholds and pass some predefined identification criteria. More information on the contents of the ROOT files can be found here [1]

While the naming convention might be different, other data formats storing information by ATLAS and CMS typically give access to similar type of information. Getting familiar on how to use the ones given here, makes evident how to do the same type of job with other types of data.

6.1 Exercises

Go to the ATLAS open data and CMS open data, find your favorite dataset and open it using a modified version of the openROOTFile.ipynb.

7 Data vs MC, histograms and histogram stacks

The most standard way to compare Data/MC is to make histograms for observables of interest. An example variable of interest here, is the muon multiplicity in our data and how they compare with the MC simulation (Fig. 3). The events data are binned in a histogram counting how many (offline muons) are present in each of our events $N = 469384$. Conventionally the data histogram is shown with black circular points (or sometimes squares) with $\sqrt{N}$ error bar if they are (pure) event counts. Events from each of the MC are binned in separate histograms with different colors and then stacked on top of each other to compute the expected event yield from simulation. All MC processes are normalized to the luminosity of data and each event has its own event weight. The statistical uncertainty of the MC estimation is then estimated by Eq. 5.

We can experiment on making such graphics using:

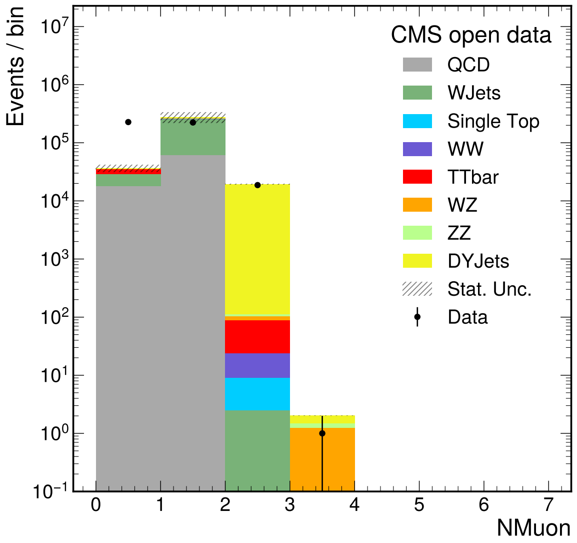

Figure 3: Muon multiplicity from the data and MC files used in the `CMS HEP Tutorial`.

- makePlot.C
- LazyHEPTutorialColab.ipynb

Doing such graphics synopsizes all the event counts we have in Data and MC, in a very economic manner. But we should not forget that our program knows more details and we should be able to be more verbose if required. We do this once for Fig. 3, as an example.

```
### printing number of events for each bin and its estimated uncertainty ###
###       disable this if you wish by setting printOut = False           ###

Data [ 227265.0,  223411.0,  18707.0,  1.0,  0.0,  0.0,  0.0,  ]
DataError [ 476.7,  472.7,  136.8,  1.0,  0.0,  0.0,  0.0,  ]
MCTot = [ 36534.6,  275505.5,  19365.2,  2.0,  0.1,  0.0,  0.0,  ]
MCTotError = [ 5041.7  55231.1  103.7  0.0  0.0  0.0  0.0  ]

### detailed analysis of MC ###

QCD = [ 18058.3,  61102.2,  0.0,  0.0,  0.0,  0.0,  0.0,  ]
QCDError = [ 5039.1,  55227.8,  0.0,  0.0,  0.0,  0.0,  0.0,  ]
WJets = [ 11070.9,  198503.2,  2.5,  0.0,  0.0,  0.0,  0.0,  ]
WJetsError = [ 154.8,  671.6,  2.2,  0.0,  0.0,  0.0,  0.0,  ]
Single Top = [ 13.8,  291.1,  6.7,  0.0,  0.0,  0.0,  0.0,  ]
Single TopError = [ 0.9,  4.3,  0.6,  0.0,  0.0,  0.0,  0.0,  ]
WW = [ 11.6,  203.8,  14.6,  0.0,  0.0,  0.0,  0.0,  ]
WWError = [ 0.8,  3.5,  0.9,  0.0,  0.0,  0.0,  0.0,  ]
TTbar = [ 6589.8,  1272.8,  65.9,  0.0,  0.0,  0.0,  0.0,  ]
TTbarError = [ 41.4,  18.3,  4.2,  0.0,  0.0,  0.0,  0.0,  ]
```

```
WZ = [ 2.6,  52.2,  13.9,  1.2,  0.0,  0.0,  0.0,  ]
WZError = [ 0.3,  1.1,  0.6,  0.2,  0.0,  0.0,  0.0,  ]
ZZ = [ 0.3,  6.5,  9.8,  0.2,  0.1,  0.0,  0.0,  ]
ZZError = [ 0.0,  0.2,  0.3,  0.0,  0.0,  0.0,  0.0,  ]
DYJets = [ 787.3,  14073.6,  19251.8,  0.5,  0.0,  0.0,  0.0,  ]
DYJetsError = [ 23.5,  99.9,  103.2,  0.5,  0.0,  0.0,  0.0,  ]
```

Already at this point we see that the QCD MC sample is a trouble maker. It contributes many events, with huge relative uncertainty. Even worse, MC is in large disagreement with data for the first two bins. We would like to restrict our analysis in a suitable subsample, applying an event selection that will eliminate the QCD contribution hoping that the Data/MC will become more reasonable.

We can experiment with the code to make it select only events for which the single muon trigger has fired[3] and the muon (offline) transverse momentum is above 24 GeV, which is the trigger threshold, requiring that $p_\mathrm{T} > 25$ GeV. In addition, we can further restrict the event selection, requiring that in parallel jets and other physics objects are present in the final state and compare again Data/MC for that subsample.

7.1 Exercises

a) Study the muon p_T distribution in bins of 1 GeV for the range $0 < p_\mathrm{T} < 50$ GeV.

b) Repeat a) for when `triggerIsoMu24` is true

c) Remake the muon multiplicity for when the muon $p_\mathrm{T} > 25$ GeV and trigger is fired.

8 Triggering

The LHC is designed to produced almost 1 pp bunch crossing event every 25 ns. In RAW format, the event size is $O(1)$ Mb. Recording all the pp collision events would require to write on tape ~ 40 Tb/s. It is thus a necessity for ATLAS and CMS to use an almost real time (online) decision system for selecting which of the LHC bunch crossing are the most interesting ones to be kept on tape for offline analysis.

The first level of selection is known as Level-1-trigger and is made by very fast algorithms encoded in FPGAs. The L1 algorithms have to be quicker than $3.2\mu s$ and do partial and coarse reconstruction of physics objects like (jets, e/γ, MET, μ, τ, b-jets ...) that are used to decide if the event will be kept on tape for offline analysis. Events that are not firing the L1-trigger, are lost for ever. Further qualification criteria are imposed by algorithms running on a computing farm, known as High Level Trigger. Events that have passed the two levels of triggering (L1 and HLT) are available for offline studies, where typically the person doing the analysis defines further (offline) criteria to define how the signal should look like.

A data event has to be “triggered” to be kept on tape. The selection efficiency of the triggering system for the signals of interest, is of great importance and is among the dominant experimental uncertainties. In MC simulation, we can emulate the trigger system and study the triggering efficiency for the physics signals we are interested into. For that we will need MC simulated data (mock data) that include also events that in reality would not pass the trigger requirements. Here, the corresponding event flag accompanying each event entry is named “triggerIsoMu24“ and is available as a TRUE

[3] `triggerIsoMu24 == true, see Sec. 8`

(1) or FALSE (0) bit inside the data and MC root files. However only for the $t\bar{t}$ MC events with "triggerIsoMu24==0" are available[4].

8.1 Exercises

1. Open "ttbar.root" from http://theofil.web.cern.ch/theofil/cmsod/files/ and measure the efficiency of the "triggerIsoMu24==1" selection.

2. Repeat the efficiency measurement as function of the generated p_{T} of the μ.

3. Quantify the MC statistical uncertainty of the measured efficiencies.

9 Cross Section Measurements

Perhaps the most fundamental type of an LHC physics analysis is a cross section measurement. This can be found turning around Eq. 1. We measure L from data and estimate ϵ for the signal (S) using MC simulation (sometimes corrected with data-driven scale factors).

Assuming that MC predicts with good accuracy the amount of background we expect in the signal region (B), the measured signal yield in data should be just $N - B$. In the ideal case (with no uncertainties of any type) we would expect by construction that $N = S + B$, with N being the measured event counts in the signal region of data and S and B the expected signal and background in the signal region, which here we will get solely from MC simulation. Dividing our signal candidate events in data ($N - B$) by the factor ϵL, gives an estimate of signal's cross section in data, which could be compared with the σ expected from theory. In its simplest incarnation an LHC cross section measurement[5] is as simple as an event counting experiment, provided that we know accurately B, ϵ and L.

The uncertainty of L is at the level of $2-3$ percent, so all the analysis challenge boils to finding a way to define the signal region such that the uncertainties $\delta\epsilon$ and δB come out small, while (ideally) $\epsilon \to 1$ and $B \to 0$. Nowadays, neural networks of several hidden layers are used to define the event selection of the signal region, but keep in mind that the increase in sensitivity might come with increased systematic uncertainty, which is nontrivial to estimate.

9.1 Exercises

Measure the $t\bar{t}$ cross section for a signal region that you will define, to achieve good signal significance using the $S/\sqrt{B}$ as figure of merit. To calculate the signal selection efficiency ϵ we need to count how many $t\bar{t}$ (weighted) events we have at our disposal in total ($N_{\mathrm{gen}}^{\mathrm{tot}}$) inside the `ttbar.root` file. The efficiency will simply be $\epsilon = N_{\mathrm{sel}}/N_{\mathrm{gen}}^{\mathrm{tot}}$, where N_{sel} is the total number of (weighted) events passing the selection cuts of our signal region.

Assume that the relative uncertainty for the signal selection efficiency is purely of statistical origin (in reality this is a major systematic uncertainty) and that the luminosity L comes with 5% uncertainty. For the background estimation B, assume that is only as large as the corresponding MC statistical uncertainty reported by your program. Compare your measurement with the first measurement that CMS ever made [4], using pretty much the same data. What's the main differences among them and how they compare with yours in terms of precision?

[4] By construction real data have always "triggerIsoMu24==1" here, while the rest of the MC simulated processes but the $t\bar{t}$ have been filtered using the "triggerIsoMu24==1" to reduce the size of the data sample.

[5] A more sophisticated approach would be to minimize the likelihood (fit) of all signal and background samples constrained taking into account normalization and shape uncertainties.

10 Projects

10.1 Trigger efficiency as function of the μ p_{T}

Study the trigger efficiency of the signal, defined here as the semileptonic decay of $t\bar{t}$ pairs, leading to final states with $N_{\mu} >= 1$. For this project you will mostly need to open just the `ttbar.root` , as it's the only sample we have available for which the events that do not pass the `triggerIsoMu24` bit, i.e., events having `triggerIsoMu24 == false` are also stored inside the `ROOT` file.

Calculate $\epsilon_{\mathrm{trigger}}$ = pass/total in bins of the generated muon (MC truth) p_{T}. Select events that have one muon generated `fabs(MCleptonPDGid) == 13` and calculate the MC generated p_{T} of the muon using the `MClepton_px` and `MClepton_py` branches. Estimate the efficiency for a generated muon to pass the CMS trigger "triggerIsoMu24 == true" as a function of its p_{T} (i.e., in bins of pt), starting with very fine binning e.g., 0.25 or 0.5 GeV in width and increasing it to 1-20 GeV widths at high p_{T} for when the available statistics start to be an issue.

Thinking needs to be placed for what would be the statistical uncertainty of $\epsilon_{\mathrm{trigger}}$. For simplicity we can calculate the uncertainty on the efficiency estimation, in the normal frequentist approximation. In this approximation we assume that the observed events that pass the selection (n) over the total events (N), $\epsilon = n/N$, is an estimate of the true efficiency ϵ_{true}. They uncertainty in the estimated efficiency in the normal approximation and the large-N limit is $\delta\epsilon = \sqrt{(\epsilon(1-\epsilon)/N)}$ [3].

1. Explain why this definition of $\delta\epsilon$ is reasonable, starting from the fact that $n \sim$ binomial$(N, p = \epsilon_{\mathrm{true}})$ and that we approximate the unknown true efficiency $p = \epsilon_{\mathrm{true}} \approx \epsilon = n/N$.

2. Furthermore, verify that the branching ratio we get in ttbar MC for events with exactly 1 μ and no other charged leptons in the final state (semi-leptonic final state in the muon channel), is what we expect given that $BR(W \to \mu\nu) \approx 10.6\%$.

Key figures to study:

- $\epsilon_{\mathrm{trigger}}$ in bins of the generated muon p_{T}, when taking into account the event weights but ignoring any uncertainty.
- $\epsilon_{\mathrm{trigger}}$ in bins of the generated muon p_{T}, without taking into account the event weights and estimating the corresponding uncertainty in the normal frequentist method.
- reconstructed muon p_{T} calculated from `Muon_Px[0]` and `Muon_Py[0]` for data and MC, without any threshold in the muon p_{T} for events selected `triggerIsoMu24 == true`. (For this plot you will need to modify `makePlot.C` analysis script and use fine binning of 1 GeV width.)

10.2 W control region and the W-boson transverse mass

The event preselection starts with requiring exactly one muon ($N_{\mu} = 1$) final state, for events with `triggerIsoMu24==1` true.

Study the reconstructed muon p_{T} calculated from `Muon_Px[0]` and `Muon_Py[0]` without any threshold starting from $p_{\mathrm{T}} = 0$, for events selected `triggerIsoMu24 == true`. (For this plot you will need to modify `makePlot.C` analysis script and use fine binning of 1 GeV width.) Show that is reasonable to select only those events with a leading muon having $p_{\mathrm{T}} > 25$ GeV.

For the selected events (i.e., preselection + $p_{\mathrm{T}} > 25$ GeV requirement), produce the transverse mass m_{T} and the MET distributions as well as the (N_j) and b-jet (N_{bj}) multiplicity distributions. Key figures:

- reconstructed muon p_{T} for data and MC, without any threshold in the muon p_{T} for events with $N_{\mu} == 1$ that fire the trigger, in bins of 1 GeV width.

- MET in bins of 10 GeV width
- transverse mass m_{T} in bins of 10 GeV width
- jet multiplicity N_j
- b-jet multiplicity N_{bj}
- event counting statistics summary

10.3 Drell–Yan control region and the Z boson mass

The event preselection starts with requiring $N_\mu \geq 2$ and leading muon $p_{\mathrm{T}} > 25$ GeV, for events with `triggerIsoMu24==1`. In addition, require that the two muons have opposite charge. Key figures to show in a presentation:

- invariant mass of the two muons $m(\mu^+, \mu^-)$ in bins of 0.25 GeV width in the range $[0, 20]$ GeV
- invariant mass of the two muons $m(\mu^+, \mu^-)$ in bins of 10 GeV width in the range $[20, 160]$ GeV
- MET in bins of 10 GeV width
- jet multiplicity N_j
- b-jet multiplicity N_{bj}
- event counting statistics summary

10.4 $t\bar{t}$ cross section in the μe final state

The event preselection starts with requiring $N_\mu \geq 1$ and leading muon $p_{\mathrm{T}} > 25$ GeV, at least one electron $N_{\mathrm{e}} \geq 1$, for events with `triggerIsoMu24==1` true. In addition, require that the two charged leptons have opposite charge. Key figures:

- invariant mass of the two leptons $m(l^+, l^-)$ in bins of 10 GeV width in the range $[0, 200]$ GeV
- MET in bins of 20 GeV width in the range $[0, 300]$ GeV
- jet multiplicity N_j
- event counting statistics summary

In this final state we expect significant contribution from the Drell-Yan (Z/γ^*) process, explain why and how the cut on the $m(l^+, l^-)$ might help getting rid of this process.

10.5 $t\bar{t}$ cross section in the $\mu + \mathrm{bjet} + \not{E}_T$ final state

The event preselection starts with requiring $N_\mu = 1$ and leading muon $p_{\mathrm{T}} > 25$ GeV, $N_{\mathrm{bj}} \geq 1$, for events with `triggerIsoMu24==1` true. Key figures:

- muon p_{T} in bins of 5 GeV width
- MET in bins of 10 GeV width
- jet multiplicity N_j
- b-jet multiplicity N_{bj}
- event counting statistics summary

10.6 $t\bar{t}$ cross section in the $\mu + 4\text{jet} + 2\text{bjet} + \not{E}_T$inal state

The event preselection starts with requiring $N_\mu = 1$ and leading muon $p_\mathrm{T} > 25$ GeV, $N_j \geq 4$, $N_\mathrm{bj} \geq 2$, for events with `triggerIsoMu24==1` true. Key figures:

- muon p_T in bins of 5 GeV width
- MET in bins of 10 GeV width
- jet multiplicity N_j
- jet multiplicity N_j when no cut on N_j is placed, but all other cuts are applied
- b-jet multiplicity N_bj
- b-jet multiplicity N_bj when no cut on N_j is placed, but all other cuts are applied
- event counting statistics summary

10.7 Reconstruction of the t-quark and W-boson masses

Reconstruct the t quark and W boson masses in the semi-leptonic $t\bar{t}$ final state, requiring $N_\mu = 1$ and leading muon $p_\mathrm{T} > 25$ GeV, $N_j \geq 4$, $N_\mathrm{bj} \geq 2$, for events with `triggerIsoMu24==1` true. Assume the final state $tt \to WWbb \to \mu\nu qqbb$ as fully resolved, where we have omitted charge and anti-particle notation for simplicity. Assume that the first four leading jets can be attributed to $qqbb$. The qq are the two jets that are not b-tagged, while for bb we assign the two jets that pass the b-tagging threshold.

We interpret the qq pair jets as coming from the hadronic decay of the W boson. Compute the invariant mass distribution of the two q jets m_qq as well as the invariant mass distribution of the three jet system m_qqb assuming that is coming from the same parent t quark decay. We don't know which of the two b-jets is the correct one to be paired with the qq, i.e., which of the bjets has the same t-quark parent as the q-jets. Try both combinations and name them m_{qqb_1} and m_{qqb_2}, where b_1 and b_2 is the leading and sub-leading b-jets. Key figures:

- muon p_T in bins of 5 GeV width
- MET in bins of 10 GeV width
- jet multiplicity N_j
- b-jet multiplicity N_bj
- m_qq in bins of 10 GeV width
- m_{qqb_1} in bins of 10 GeV width
- m_{qqb_2} in bins of 10 GeV width
- m_{qqb_1} and m_{qqb_2} stacked in the same histogram (e.g., fill the 2 entries in the same histogram for each event)
- event counting statistics summary

10.8 Charge asymmetry in $pp \to W^{\pm}$

Select a W boson control region, find how many of them are W^+ and W^-. What do we naively expect from the total charge of the initial state ? What our MC simulation predicts for the charge asymmetry ?

10.9 Charge asymmetry in $t\bar{t}$

Select semileptonic $t\bar{t}$events, find how many of them are W^+ and W^-. What do we naively expect and how it compares with what MC simulation predicts ?

11 Acknowledgments

I would like to thank Isabell Melzer-Pellman for organizing and inviting me to the Introduction to the Terascale School in DESY 2023 and to Andreas Meyer for reading and providing precision comments to the present document.

A Invariant Mass

To calculate the invariant mass of $X \to AB$ decays, given the four-vectors $p^{\mu}_{\mathrm{A}} = (E_{\mathrm{A}}, p_{\mathrm{Ax}}, p_{\mathrm{Ay}}, p_{\mathrm{Az}})$ and $p^{\mu}_{\mathrm{B}} = (E_{\mathrm{B}}, p_{\mathrm{Bx}}, p_{\mathrm{By}}, p_{\mathrm{Bz}})$ we use the square of four-momentum conservation $P^{\mu}_{\mathrm{X}} = p^{\mu}_{\mathrm{A}} + p^{\mu}_{\mathrm{B}}$, which gives

$$M^2_{\mathrm{X}} = (E_{\mathrm{A}} + E_{\mathrm{B}})^2 - (p_{\mathrm{Ax}} + p_{\mathrm{Ax}})^2 - (p_{\mathrm{By}} + p_{\mathrm{By}})^2 - (p_{\mathrm{Cz}} + p_{\mathrm{Cz}})^2. \tag{6}$$

Having computed M^2_{X} we only need to take its square root to end up to M_{X}.

B Transverse Mass

To calculate the transverse mass (m_{T}) in $W \to \ell\nu$ decays, we will work under the assumption that the visible MET is solely due to the transverse momentum of one escaping neutrino. We will neglect the muon and the neutrino masses and build their transverse 4 vectors such as they are light-like ($P^2 = 0$), using only the transverse component of their momentum

$$p^{\mu} = (E, p_{\mathrm{x}}, p_{\mathrm{y}}, p_{\mathrm{z}}) = (\sqrt{p^2_{\mathrm{x}} + p^2_{\mathrm{y}}}, p_{\mathrm{x}}, p_{\mathrm{y}}, 0). \tag{7}$$

We will sum the two transverse 4-vectors and calculate the magnitude (mass) of their sum, which is the definition of the m_{T}. Note that is not possible to calculate the ordinary invariant mass of the $m(\mu, \nu)$ system, since the p_{z} of the ν is unknown. The m_{T} is the closest quantity we could built to the invariant mass of the system of two particles, having as endpoint the $m(\mu, \nu)$ and being itself also invariant. See also Sec. 49.6 of PDG2022.

C Error bars in histogram bins

Histograms are the most usual way to quickly estimate the shape of the underlying probability density function of an observable. Suppose that the random variable X we wish to measure takes continuous values, as for example the p_{T} of a muon. We count how many events have a muon within a certain p_{T} range (bin of our histogram) and populate the event content of that particular bin.

We are interested to learn about the probability p_j that X is observed inside the boundaries $x_{\mathrm{min,j}} < x < x_{\mathrm{max,j}}$ of the j-th bin. In the limit of $N \to \infty$, an estimator of the desired probability is the $\hat{p_j} = n_j/N \approx p_j$, where N is the total number of events we have and n_j the subset of those that are observed within the bin range $[x_{\rm min,j}, x_{\rm max,j})$[6].

We may regard each bin of a histogram as an independent experiment, governed by the binomial probability

$$P(n_j) = \frac{N!}{n_j!(N-n_j)!} p_j^{n_j} (1-p_j)^{N-n_j} \tag{8}$$

as the event will either belong in bin j or not. Like when we flip a coin it is either heads or tails. But there, the two outcomes are equiprobable while for our observable one of the two outcomes might be very rare. The variance of n_j is

$$\sigma^2_{n_j} = E(n_j^2) - E(n_j)^2 = Np_j(1-p_j). \tag{9}$$

When our random variable $X \sim p(x)$ is *continuous* and distributes according to the $p(x)$ probability density function, we have

$$E(n_j) = N \int_{x_{min,j}}^{x_{max,j}} p(x)dx = Np_j \tag{10}$$

with $E(n_j)$ denoting the theoretically expected value of n_j. In the limit of $N \to \infty$ we can always chose the bin boundaries with fine segmentation (fine binning), such as $p_j \to 0$ while at the same time Np_j remains finite and equal to n_j. In this limit, Eq. 8 is reduced to a Poisson,

$$P(n_j) = \frac{\mu_j^{n_j}}{n_j!} \exp^{-\mu_j} \tag{11}$$

with $\mu_j = Np_j$ and $\sigma^2_{n_j} = Np_j(1-p_j) = Np_j = \mu_j$, since $1 - p_j \approx 1$ when $p_j \to 0$. Thus, the bin entries distribute with Poisson probability mass function having $\sigma^2_{n_j} = \mu_j$. Unfortunately, the Poisson confidence intervals are not simple in calculation and one has to use the quantile of the chi-square distribution (in `ROOT` it's the `TMath::ChisquareQuantile`) which give asymmetric error bars.

```
float statErrorN(float x){return x - 0.5*TMath::ChisquareQuantile(0.3173/2,2*x);}
float statErrorP(float x){return 0.5*TMath::ChisquareQuantile(1-0.3173/2,2*(x+1))-x;}
```

Doing error propagation with asymmetric error bars is tedious. People sometimes take the largest of the up/down error bar and symmetrize it calling this procedure conservative.

We can further simplify life making the binning such as the $E(n_j) > 20$, while at the same time $p_j \to 0$ and $N \to \infty$. We are one step before saying that $\delta n_j = \sigma_{n_j} = \sqrt{E(n_j)} \approx \sqrt{n_j}$ and treat it as a *Normal* symmetric confidence interval of 68% confidence level since for $E(n_j) \gg 20$ the Poisson distribution is well approximated by the Normal distribution. For *continuous* random variables and very large number of events, we can always chose fine binning such as $p_j \to 0$, thus rendering the above considerations reasonable. However, an extra leap is needed to use the *observed number of events* n_j as an estimator of the *unknown* $E(n_j) = \mu_j = Np_j$.

In the realm of statistics it can be disputable to attach error bars on the observed data of a counting experiment. Unarguably, we are 100% sure how many we have already observed. My personal take on the subject. For displaying error bars in a plot and back to the envelope calculations, using $\sqrt{n}$ as the error bar of n *observed* events is not unreasonable [7], provided that our bins have at least 20 events as content and we do not speak about 0 ± 0 or 1 ± 1. The number of events we have is related to the statistical precision of our data and the $\sqrt{n}$ prescription is in many times helpful when visualizing the data. However, we should keep in mind the error bars displayed in figures should not necessarily, and as a matter of fact are very often not, propagated as such to the final result.

[6] By convention, bin intervals are usually closed "[" on the lower end and open ")" on the higher.

[7] In the Gaussian limit large-N, the confidence belt is symmetric, i.e in the Neyman construction, the width of the confidence belt as obtained from the observed value is the same as that of the pdf of the true value.

D Poisson processes

The number of phone calls (n) a help line receives over a fixed time interval (e.g., per 5 minutes) is following the Poisson distribution,

$$P(n) = \frac{\mu^n}{n!} \exp^{-\mu} \tag{12}$$

with μ being the mean number of phone calls received per time interval. It is easy to show that the maximum likelihood estimator of μ is just the arithmetic mean (average) rate of phone calls per time interval. Same goes for the number of nuclear decays recorded in a certain time interval by a `Geiger-Müller` detector, or for the number of events recorded by the CMS detector satisfying specific selection criteria for a fixed amount of integrated luminosity.

You may wonder what is common between a telephone center and nuclear and particle physics. The answer is,

- we have *many* clients (pp collisions) each of them having the *very same small probability* $p \to 0$ calling the help center (passing the event selection criteria).
- each client (pp collision) is *independent* from the others and *memory-less*, i.e., future pp collisions don't care what was the outcome of past pp collisions and the probability that a client calls the help center is *constant* and *the same* for all clients.

The above conditions bring us to close to the Poisson limit of Eq. 8. Verify numerically the that Eq. 8 is well approximated by Eq. 12 when $N \to \infty$ and $p \to 0$ with $Np = \mu$ giving your own values to p and N.

References

[1] C. Sander and A. Schmidt, "CMS HEP Tutorial", http://opendata.cern.ch/record/50

[2] M. D. Schwartz, "TASI Lectures on Collider Physics", https://arxiv.org/abs/1709.04533

[3] https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

[4] CMS Collaboration first $t\bar{t}$ cross section measurements at $\sqrt{s} = 7$ TeV:
http://arxiv.org/abs/1010.5994
http://arxiv.org/abs/1105.5661
http://arxiv.org/abs/1106.0902
http://arxiv.org/abs/1108.3773