Optimising HEP parameter fits: event-by-event sensitivities, weight derivative regression

Andrea Valassi (CERN IT-DI)

CHEP2019, 7 November 2019 – Adelaide, Australia
https://indico.cern.ch/event/773049/contributions/3476059
This is a follow-up of my CHEP2018 talk about *binned fits of a parameter* \( \theta \)

**Evaluation and training metrics:**
- Fisher Information Part

**Previous CHEP2018 talk**
- Event selection
- Binary classification
- Bin-by-bin sensitivity to \( \theta \)
- Cross-section fits (FIP1, FIP2)
- Medical Diagnostics (AUC), Information Retrieval (F1)

**This CHEP2019 talk**
- Event partitioning
- Non-binary *regression* (WEIGHT DERIVATIVE REGRESSION)
- Event-by-event sensitivity to \( \theta \)
- Mass fits, Coupling fits (FIP3)
- Meteorology (MSE, Brier), Medical Prognostics

**Compare to and learn from other domains**

Talk: [https://doi.org/10.5281/zenodo.1303387](https://doi.org/10.5281/zenodo.1303387)
Paper: [https://doi.org/10.1051/epjconf/201921406004](https://doi.org/10.1051/epjconf/201921406004)
Outline

• 1 - HEP parameter fits and Weight Derivative Regression

• 2 - Learning from others

• Conclusions

This talk only provides some maths and some literature review

No toy model or concrete applications are presented
There are two handles to minimize the statistical error $\Delta \theta$:

1. Event selection
   Signal-background discrimination

2. Event partitioning
   Variable(s) for the distribution fit

$m_W = 81.30 \pm 0.47 \text{ (stat.)} \pm 0.11 \text{ (syst.) GeV} / c^2$

I only discuss the **statistical error $\Delta \theta$ in this talk**
(I ignore systematic errors, even if at LHC they are the limitation)
Fisher Information \( \frac{1}{(\Delta \theta)^2} \) from bin-by-bin sensitivities

For a given partitioning scheme with \( K \) bins

\( n_k \) is the number of selected events in bin \( k \):

Statistical errors: 
information adds up 
(independent bins)

Bin-by-bin sensitivity to \( \theta \)

Minimizing \( \Delta \theta \) is equivalent to 
maximizing \( I_{\theta} \)

\[
I_{\theta} = \frac{1}{(\Delta \theta)^2} = \sum_{k=1}^{K} \frac{1}{(\Delta \theta)^2_k} = \sum_{k=1}^{K} n_k \left( \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2
\]
1 – Binned fit of a parameter $\theta$

Fisher Information Part (FIP)

There are two handles to minimize the statistical error $\Delta \theta$:

1. Event selection
   Signal-background discrimination

2. Event partitioning
   Variable(s) for the distribution fit

My CHEP2018 talk:
FIP evaluation of event selection

For a given data set and given partitioning, FIP compares $I_{\theta}$ to $I_{\theta}^{(\text{ideal})}$ for the ideal selection (select all signal, reject all bkg)

This CHEP2019 talk:
FIP evaluation of event partitioning

For a given data set, FIP compares $I_{\theta}$ to $I_{\theta}^{(\text{ideal})}$ for the ideal partitioning (and the ideal selection)

But what is the smallest statistical error achievable on a given data set with ideal partitioning and selection? Enter event-by-event sensitivities

Recap CHEP2018 talk

Fisher Information Part (FIP): the fraction of the information available “in an ideal case” retained by a given analysis

\[
\text{FIP} = \frac{I_{\theta}}{I_{\theta}^{(\text{ideal})}} = \frac{(\Delta \theta^{(\text{ideal})})^2}{(\Delta \theta)^2} \leq 100\%
\]

FIP is a metric between 0 and 1 – higher is better
### 1 – Binned fit of a parameter $\theta$

**Event-by-event Monte Carlo reweighting**

Fit for $\theta \rightarrow$ Compare data in bin k to model prediction $n_k$ as a function of $\theta$

$$n_k(\theta) = \sum_{i \in k} w_i(\theta) = \sum_{i \in k} w_i(\theta) + \sum_{i \in k} w_i = s_k(\theta) + b_k$$

---

**1. Generate signal sample at $\theta_{\text{ref}}$, with $w_i(\theta_{\text{ref}}) = 1$**
(By definition, background does not depend on $\theta$)

**2. Full detector simulation**
(MC truth event properties $x_i^{(\text{true})}$ → observed event properties $x_i$)

**3. Reweight each event by matrix element ratio**

$$w_i(\theta) = \frac{\text{Prob}(\theta)(x_i^{(\text{true})})}{\text{Prob}(\theta_{\text{ref}})(x_i^{(\text{true})})} = \frac{|\mathcal{M}(\theta, x_i^{(\text{true})})|^2}{|\mathcal{M}(\theta_{\text{ref}}, x_i^{(\text{true})})|^2}$$

---

Monte Carlo reweighting: used extensively at LEP
Simpler than Matrix Element Method (no integration)
[see Gainer2014, Mattelaer2016 for hadron colliders]

---

A. Valassi – HEP parameter fits and Weight Derivative Regression

CHEP2019, Adelaide – 7 Nov 2019
1 – Binned fit of a parameter $\theta$

**Event-by-event sensitivities $\gamma_i$: MC weight derivatives**

Bin-by-bin model prediction $n_k(\theta)$

$$n_k(\theta) = \sum_{i \in k} w_i(\theta) = \sum_{i \in k} w_i^{\text{Sig}}(\theta) + \sum_{i \in k} w_i^{\text{Bkg}}(\theta) = s_k(\theta) + b_k$$

Define the **event-by-event sensitivity $\gamma_i$ to $\theta$** as the derivative with respect to $\theta$ of the MC weight $w_i$

$$\gamma_i|_{\theta} = \left( \frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right)_{\theta} \quad \rightarrow \quad \gamma_i = \gamma_i|_{\theta=\theta_{\text{ref}}} = \left( \frac{\partial w_i}{\partial \theta} \right)_{\theta=\theta_{\text{ref}}}$$

(normalized by $1/w_i$, but $w_i(\theta_{\text{ref}})=1$ at the reference $\theta=\theta_{\text{ref}}$)

The **bin-by-bin sensitivity** to $\theta$ in bin $k$ is the average in bin $k$ of the event-by-event sensitivity $\gamma_i$ to $\theta$

$$\left( \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)_{\theta=\theta_{\text{ref}}} = \frac{1}{n_k} \sum_{i \in k} \gamma_i = \langle \gamma \rangle_k = \frac{1}{n_k} \frac{\partial n_k}{\partial \theta}$$
### Background events have $\gamma_i=0$

*because by definition they are insensitive to $\theta$*

$$\gamma_i = \left( \frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right) = 0, \quad \text{if } i \in \{\text{Background}\}$$

$$\gamma_i = \left( \frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right) \in \{-\infty, +\infty\}, \quad \text{if } i \in \{\text{Signal}\}$$

Signal events may have sensitivity $\gamma_i>0$, $\gamma_i=0$ or $\gamma_i<0$

(special case: cross-section fit $\gamma_i=1/\sigma_s$)

---

**For what concerns statistical errors in a parameter fit, there is no distinction between background events and signal events with low sensitivity ($|\gamma_i| \sim 0$)**

---

**Bin-by-bin sensitivity $\phi_k$**

of signal events alone:

$$\phi_k = \langle \gamma \rangle_{k, \text{Sig}} = \frac{1}{s_k} \sum_{i \in k} \gamma_i = \frac{1}{s_k} \frac{\partial s_k}{\partial \theta}$$

**Bin-by-bin purity $\rho_k \leq 1$:**

$$\delta_i = \begin{cases} 1 & \text{if } i \in \{\text{Signal}\} \\ 0 & \text{if } i \in \{\text{Background}\} \end{cases}$$

$$\rho_k = \frac{s_k}{s_k + b_k} = \frac{s_k}{n_k} = \frac{\sum_{i \in k} \delta_i}{n_k} = \langle \delta \rangle_k$$

**Bin-by-bin sensitivity $\langle \gamma \rangle_k$**

of signal + background:

$$\langle \gamma \rangle_k = \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} = \frac{\rho_k}{s_k} \frac{s_k}{\partial \theta} = \rho_k \phi_k$$

**Effect of background:**

it dilutes by a factor $\rho_k \leq 1$

the bin-by-bin sensitivity and information for signal events alone

---

**Information from all bins for signal + background:**

$$I_0 = \sum_{k=1}^{K} n_k (\gamma)_k^2 = \sum_{k=1}^{K} n_k (\rho_k \phi_k)^2 = \sum_{k=1}^{K} s_k \rho_k^2 \phi_k^2$$
1 – Binned fit of a parameter $\theta$

**Ideal case: partition by the evt-by-evt sensitivity $\gamma_i$**

Information $I_\theta$ in terms of average bin-by-bin sensitivities:

\[
I_\theta = \sum_{k=1}^{K} n_k \left( \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2 = \sum_{k=1}^{K} n_k \langle \gamma \rangle_k^2
\]

There is an **information gain** in partitioning two events $i_1$ and $i_2$ in two 1-event bins rather than one 2-event bin if their sensitivities $\gamma_{i_1}$ and $\gamma_{i_2}$ are different.

\[
\Delta I_\theta = \gamma_{i_1}^2 + \gamma_{i_2}^2 - 2 \left( \frac{\gamma_{i_1} + \gamma_{i_2}}{2} \right)^2 = \frac{1}{2} \left( \gamma_{i_1} - \gamma_{i_2} \right)^2
\]

**Goal of a distribution fit:** partition events by their different MC-truth event-by-event sensitivities $\gamma_i$ to $\theta$

How to achieve this in practice: next two slides (WDR)

**Knowing one’s limits: maximum achievable information with an ideal detector**
- Ideal acceptance, select all signal events $S_{\text{sel}} = S_{\text{tot}}$
- Ideal resolution, measured $\gamma_i$ is that from MC truth (implies ideal rejection of background events, $\gamma_i = 0$)

Use $I_\theta^{\text{ideal}}$ to compute FIP: following two slides

\[
I_\theta^{\text{ideal}} = \sum_{i=1}^{N_{\text{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\text{tot}}} \gamma_i^2
\]
**1 – Binned fit of a parameter $\theta$**

**Weight Derivative Regression (WDR):** train $q_i$ for $\gamma_i$

---

Goal of a distribution fit: separate events with different MC-truth event-by-event sensitivities $\gamma_i$ to $\theta$

**But $\gamma_i$ is not observable on real data events!**

---

**Weight Derivative Regression:**

- **train a regressor** $q_i = q(x_i)$
  - on detector-level MC observables $x_i$
  - against the MC-truth $\gamma_i = \partial w_i / \partial \theta$
  - for signal and background MC events

Then determine $\theta$ by the 1-D fit of $q(x_i)$ for real data events $x_i$

---

Some of many caveats:

- Dependency of weight derivative on reference $\theta_{ref}$:
  - WDR easier for coupling fits than for mass fits?
- How feasible is it to compute and store MC-truth weight derivatives?
- How useful is this for measurements limited by systematics?
- Train $q$ on signal + background and 1-D fit of $q$, or train $q$ on signal alone and 2-D fit on $q$ and scoring classifier?
- How to deal with simultaneous fits of many parameters?

---

**Training metric:** maximize FIP

**Evaluation metric:** maximize FIP

(or equivalently minimize MSE? see final slides)
The WDR idea was inspired by the Optimal Observables (OO) method. Both OO and WDR partition data by an approximation of a MC-truth sensitivity $\gamma_i$ to $\theta$. (OO does not use MC weight derivatives but it is similar)

**MC-truth functional dependence**

$\gamma_i^{(MC \text{ truth})} = f(x_i^{(MC \text{ truth})})$

**Fit optimal observable**

$OO_i^{(DATA)} = f(x_i^{(DATA)})$

$OO_i^{(MC)} = f(x_i^{(MC)})$

**Weight Derivative Regression**

$\gamma_i^{(MC \text{ truth})} \sim q(x_i^{(MC)})$

**Fit WDR regressor**

$q_i^{(DATA)} = q(x_i^{(DATA)})$

$q_i^{(MC)} = q(x_i^{(MC)})$

**Data observable event properties** $x_i^{(DATA)}$

**MC-truth event properties** $x_i^{(MC \text{ truth})}$

**MC observable event properties** $x_i^{(MC)}$

Like OO, WDR can be useful in coupling/EFT fits (more than in mass fits)

Some similarities also with the MadMiner approach

"Constraining effective field theories with ML"
1 – Binned fit of a parameter $\theta$

**FIP decomposition: efficiency, sharpness, purity**

Numerator: Information retained by a given analysis using $N_{\text{sel}}=\sum n_k$ events with the given detector

Denominator: maximum theoretically available information from the given sample of $N_{\text{tot}}$ events ($S_{\text{tot}}$ signal events) if the true $\gamma_i$ were known for each event (ideal detector)

$$FIP_3 = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\text{ideal})}} = \frac{\sum_{k=1}^{K} S_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\text{tot}}} \gamma_i^2} = \text{FIP}_{\text{eff}} \times \text{FIP}_{\text{sha}} \times \text{FIP}_{\text{pur}}$$

Sensitivity-weighted signal **efficiency**: keep $S_{\text{sel}}$ of $S_{\text{tot}}$ events

**Sharpness** in separating signal events with different sensitivities: partition $S_{\text{sel}}$ signal events into $K$ bins

“sharpness” as in meteorology: see later why

Sensitivity-weighted signal **purity** or equivalently **sharpness** in separating signal events from background events: dilution of signal sensitivity caused by bin-by-bin purity $\rho_k$
### Limited detector acceptance
(detector geometry, trigger rate):
- factor this out in \( FIP_{\text{ACC}} \leq 1 \)

### Limited detector resolution
In the multi-dimensional space of event observables \( \mathbf{x} \),
it is impossible to resolve:
- signal events with high sensitivity \( \gamma_i \)
- from signal events with low sensitivity \( \gamma_i \):
  - average sensitivity is \( \phi(\mathbf{x}) \)
- signal events \( \delta_i=1 \)
- from background events \( \delta_i=0 \):
  - average purity is \( \rho(\mathbf{x}) \)

\[
FIP_{\text{ALL}} = FIP_{\text{ACC}} \times FIP_{3}
\]
\[
FIP_{3} = FIP_{\text{eff}} \times FIP_{\text{SHA}} \times FIP_{\text{PUR}}
\]

\[0 \leq FIP_{3} \leq FIP_{3}^{(\text{max})} \leq 1\]

\[
S_{\text{ALL}}, \gamma_i, \delta_i
\]
\[
I_\theta^{(\text{ideal}, S_{\text{ALL}})} = \sum_{i=1}^{S_{\text{ALL}}} \gamma_i^2
\]

\[
FIP_{\text{ACC}} \leq 1
\]

\[
S_{\text{tot}}, \gamma_i, \delta_i
\]
\[
I_\theta^{(\text{ideal})} = \sum_{i=1}^{S_{\text{tot}}} \gamma_i^2
\]

\[
FIP_{3}^{(\text{max})} = \frac{I_\theta^{(\text{ideal})}}{I_\theta^{(\text{ideal})}}
\]

\[
S_{\text{tot}}, \phi(\mathbf{x}), \delta_i
\]
\[
I_\theta = \int s(\mathbf{x}) \phi(\mathbf{x})^2 d\mathbf{x}
\]

\[
FIP_{\text{eff}} = \frac{I_\theta}{I_\theta^{(\text{ideal})}}
\]

\[
S_{\text{tot}}, \phi(\mathbf{x}), \rho(\mathbf{x})
\]
\[
I_\theta^{(\text{max})} = \int s(\mathbf{x}) \phi(\mathbf{x})^2 \rho(\mathbf{x}) d\mathbf{x}
\]

\[
FIP_{\text{SHA}}
\]

\[
S_{\text{sel}}, \gamma_i, \delta_i
\]
\[
I_\theta = \sum_{i=1}^{S_{\text{sel}}} \gamma_i^2
\]

\[
FIP_{\text{pur}}
\]

\[
S_{\text{sel}}, \phi_k, \delta_i
\]
\[
I_\theta = \sum_{k=1}^{K} s_k \phi_k^2
\]

\[
FIP > FIP^{(\text{max})}\text{ while training } q_i
\text{ implies overtraining...}
Different problems in different domains require different metrics and tools…
2 – Learning from others

Evaluating the evaluation metrics

Evaluation metrics of (binary and non-binary) classifiers have been analysed and compared in many ways.

There are two approaches which I find particularly useful:

1. **Studying the symmetries and invariances of evaluation metrics**


   **Example:** *(ir)relevance of True Negatives: in my CHEP2018 talk*

2. **Separating threshold, ranking and probabilistic metrics**

   R. Caruana, A. Niculescu-Mizil, *Data mining in metric space: an empirical analysis of supervised learning performance criteria*, Proc. 10th Int. Conf. on Knowledge Discovery and Data Mining (KDD-04), Seattle (2004). doi:10.1145/1014052.1014063


   **Example:** AUC (ranking) vs. MSE (probabilistic): in this CHEP2019 talk (next 3 slides)

MSE decomposition: Validity and Sharpness

MSE = \frac{1}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} (q_i - \gamma_i)^2

MSE is a probabilistic metric for both evaluation and training.

MSE decomposition
(if the \( N_{\text{tot}} \) events are split into \( K \) partitions, with \( q_i = q_{(k)} \) \( \forall i \in k \):

\[
\text{MSE} = \frac{1}{N_{\text{tot}}} \left[ \sum_{k=1}^{K} n_k (q_{(k)} - \langle \gamma \rangle_k)^2 \right] + \frac{1}{N_{\text{tot}}} \left[ \left( \sum_{i=1}^{N_{\text{tot}}} \gamma_i^2 \right) - \left( \sum_{k=1}^{K} n_k (\langle \gamma \rangle_k)^2 \right) \right]
\]

Validity, Reliability, Calibration

Validity: in a partition with given true average sensitivity \( <\gamma_k> \), is the predicted sensitivity \( q_{(k)} \) well calibrated?

~0 in training by construction
~0 in evaluation if there are no systematics

Sharpness, Resolution, Refinement

Sharpness: how well do we separate events with different true sensitivities \( \gamma_i \)?

This is what determines the statistical error on the measurement of \( \theta \): related to FIP!
FIP is related to Sharpness:

In the ideal case: \( \text{MSE}_{\text{sha}} = 0 \) and \( \text{FIP} = 1 \) (events with different \( \gamma_i \) can be resolved)

**Practical implication for Weight Derivative Regression:**

\( \text{MSE is the most appropriate loss function for training the WDR regressor} \)
2 – Learning from others: HEP does not need ranking, or ranking metrics  

**HEP needs partitioning, and probabilistic metrics**

### Ranking, and ranking metrics

Pick two events at random and rank them

### Partitioning, and probabilistic metrics

Group events and make a forecast on each subset

#### Medical Diagnostics → ranking evaluation of diagnostic prediction

Patient A is diagnosed as more likely sick than B: how often am I right?

#### Meteorology → probabilistic evaluation of weather prediction

Rain forecast was 30% for these 10 days: actual rainy days?

#### Medical Prognostics → probabilistic evaluation of survival prediction

5yr survival forecast was 90% for these 10 patients: actual survivors?

#### HEP parameter fits → probabilistic evaluation of measurement of $\theta$

MC forecast for #events in this bin is 10 (20) for $\theta=1$ (2): actual data?

<table>
<thead>
<tr>
<th>Validity, Reliability, Calibration</th>
<th>Sharpness, Resolution, Refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{MSE} = \frac{1}{N_{\text{tot}}} \left[ \sum_{k=1}^{K} n_k \left( \bar{q}(k) - \langle \bar{q} \rangle_k \right)^2 \right] + \frac{1}{N_{\text{tot}}} \left[ \sum_{i=1}^{N_{\text{tot}}} \gamma_i^2 - \left( \sum_{k=1}^{K} n_k \langle \gamma \rangle_k^2 \right) \right]$</td>
<td></td>
</tr>
</tbody>
</table>

**AUC (Area Under the ROC Curve):** probability that a randomly chosen diseased subject is correctly rated or ranked with greater suspicion than a randomly chosen non-diseased subject

**Sharpness (from MSE):** how well can I resolve days with 10% and 90% chance of rain?  
Patients with 10% and 90% 5yr survival rate?  
Signal events with high sensitivity to $\theta$ from (signal or background) events with low sensitivity?

**IRRELEVANT FOR HEP PARAMETER FITS?**

**ESSENTIAL FOR HEP PARAMETER FITS!**
Conclusions – HEP measurement of a parameter $\theta$

- **MC weight derivatives** (event-by-event sensitivities $\gamma_i$ to $\theta$) may be used:
  - To determine the **ideal partitioning strategy**: partition by $\gamma_i$
  - To derive the **minimum error on the measurement of $\theta$** (ideal detector)
    \[
    I^{(\text{ideal})}_\theta = \sum_{i=1}^{N_{\text{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\text{tot}}} \gamma_i^2
    \]
  - To derive **training and validation metrics** to optimize the measurement
    \[
    \text{FIP} = \frac{I_\theta}{I^{(\text{ideal})}_\theta} = \frac{\sum_{k=1}^{K} n_k (\gamma)^2_k}{\sum_{i=1}^{S_{\text{tot}}} \gamma_i^2} = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\text{tot}}} \gamma_i^2}
    \]
  - To train a **regressor $q_i$ of $\gamma_i$** (optimal observable) for a 1-D fit of $\theta$

- HEP parameter fits are closer to **Meteorology** than to Medical Diagnostics
  - They use **partitioning** and need **probabilistic metrics** (sharpness, MSE)
    \[
    \text{FIP} = \frac{I_\theta}{I^{(\text{ideal})}_\theta} = \left(1 - \frac{N_{\text{tot}} \times \text{MSE}_{\text{sha}}}{I^{(\text{ideal})}_\theta}\right)
    \]
  - They do not use ranking and do not need ranking metrics (AUC)

**Evaluation and training metrics: FIP**

**Compare to and learn from other domains**
Backup slides
Non-dichotomous truth: examples

- **Medical Diagnostics** → continuous scale gold standard
  - The Obuchowski measure, e.g. five stages of liver fibrosis,

- **Information Retrieval** → graded relevance assessment and DCG
  - Discounted Cumulated Gain
  
  \[
  \text{DCG}[k] = \sum_{i=1}^{k} \frac{G[i]}{\min(1, \log_2 i)}
  \]

- **ML (for finance)** → example-dependent cost-sensitive classification
  - Payoff matrix for transaction $x$:

<table>
<thead>
<tr>
<th>refuse</th>
<th>approve</th>
</tr>
</thead>
<tbody>
<tr>
<td>$20$</td>
<td>$-x$</td>
</tr>
<tr>
<td>$-20$</td>
<td>$0.02x$</td>
</tr>
</tbody>
</table>

- **Meteorology** → probabilistic evaluation of weather forecasts
  - Rain forecast was 30% for these 10 days: actual rainy days?

- **Medical Prognostics** → probabilistic evaluation of survival forecasts
  - 5yr survival forecast was 90% for these 10 patients: actual survivors?

- **HEP measurement of $\theta$** → evt-by-evt sensitivity to $\theta$
1 – Binned fit of a parameter $\theta$

Weight Derivative Regression – in practice

- **Compute event-by-event sensitivities $\gamma_i$ from signal MC weight derivatives**
  - Possibly at various reference values of $\theta$

- **Pre-select events to remove most backgrounds**
  - Possibly maximizing a sensitivity-weighted signal efficiency?

- **Train a regressor $q_i$ for the MC-truth $\gamma_i$ from measured event properties**
  - Possibly using MSE as the loss function in the training (see next slides)

- **Determine $\theta$ from a 1-D fit on the optimal observable $q_i$**
  - Or possibly a 2-D fit on $(q_i, D_i)$ including the pre-selection classifier $D_i$

Some of the many **limitations of this approach**
- MC weight derivative depend on $\theta$: coupling fits easier than mass fits
- I ignored systematic errors
- I only discussed fits of a single physics parameter at a time
  - *But I still find this approach better than maximizing an AUC…*

(Note: I did not try a real measurement – I did a few tests with a toy model, but I am not presenting them today)
Estimation of parameter $\theta$ in a binned distribution fit

Weight derivative regressors and their training

*(a frequentist dinosaur’s view of Machine Learning)*

Classic ML problem: create a model $q(x) = R_\gamma(x)$ to predict the value of $\gamma(x)$ in a multi-dimensional space of variables $x$

Choosing a ML methodology mainly implies two choices:

1. **The shape of the function** $R_\gamma(x)$:
   i.e. how we choose to model $\gamma(x)$
   Examples: decision tree (sparsely uniform), neural network (sigmoids), linear discriminant

2. **The training metric**: a “distance” of $R_\gamma(x_i)$ to $\gamma(x_i)$ or $\gamma_i$ to minimize, or a property of $R_\gamma(x_i)$ to maximize
   Examples: Gini, Shannon entropy/information, MSE

I focus on **Decision Trees** because of the similarities to binned distribution fits

I suggest to use $I_\theta$ or FIP both for training and for evaluation
(1) The scoring classifier D for signal/background discrimination is related to the average purity $p(x)$: it would be a pity to use it only for a yes/no decision. It can be used both for measuring cross-sections (1-D fit of D) or for measuring a mass or coupling (2-D fit against another variable).

Use the scoring classifier D to partition events, not only to accept or reject events.

(2) Signal events with zero or low sensitivity to $\theta$ and background events are equally irrelevant.

Separating signal events with high sensitivity to $\theta$ from background events is as important as

Separating signal events with high sensitivity to $\theta$ from signal events with low sensitivity to $\theta$.
Fisher information (about a parameter $\theta$)

- **Fisher information $I_\theta$** is a useful concept because
  - 1. It refers to the parameter $\theta$ that is being measured
  - 2. It is additive: the information from independent measurements adds up
  - 3. The higher the information $I_\theta$, the lower the error $\Delta \theta$ achievable on $\theta$

  \[
  \text{Cramer-Rao lower bound CRLB (lowest achievable variance $\Delta \theta^2$)} \quad (\Delta \hat{\theta})^2 = \text{var}(\hat{\theta}) \geq \frac{1}{I_\theta}
  \]

- Some estimators achieve the CRLB and are called efficient
  - Example: a maximum likelihood fit (given the event counts in a given partitioning scheme)

- In the following I will express statistical error $\Delta \theta$ in terms of information $I_\theta$

  \[
  I_\theta = \frac{1}{(\Delta \theta)^2}
  \]

i.e. I will treat errors $\Delta \theta$ and information $I_\theta$ as equivalent concepts

NB: Shannon information is a very different metric!

HEP cross-section in a counting experiment

- Measurement of a total cross-section $\sigma_s$ in a counting experiment
- A distribution fit with a single bin
- Well-known since decades if final goal is to minimize statistical error $\Delta \sigma_s$
  - *Maximise $\epsilon_s \rho$* (“common knowledge” in the LEP2 experiments) $\rightarrow$ “FIP1”
  - NB: This metric only makes sense for this specific HEP optimization problem!

$$I_{\sigma_s} = \frac{1}{(\Delta \sigma_s)^2} = \frac{1}{\sigma_s^2} \epsilon_s \varrho S_{\text{tot}} = \frac{1}{\sigma_s^2} \left( \frac{S_{\text{sel}}^2}{S_{\text{sel}} + B_{\text{sel}}} \right)$$

$$I_{\sigma_s}^{(\text{ideal})} = \frac{S_{\text{tot}}}{\sigma_s^2}, \text{ if } \varrho = 1 \text{ and } \epsilon_s = 1$$

$$FIP_1 = \frac{I_{\sigma_s}}{I_{\sigma_s}^{(\text{ideal})}} = \epsilon_s \varrho$$

By the way: $\rho/\epsilon_s=1$ where $\partial FIP1/\partial \rho = \partial FIP1/\partial \epsilon_s$ (just like for F1)
A brief comparison of MD, IR and HEP

• **Medical Diagnostics**
  – *All patients are important, both truly ill (TP) and truly healthy (TN)*
  – e.g. ACC metric depends on all four categories: average over TP+TN+FP+FN

  \[ \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \]

• **Information Retrieval**
  – Based on *qualitative distinction between “relevant” and “non relevant” documents*
  – e.g. F1 metric *does not depend on True Negatives*
    • Rejected “irrelevant” documents are utterly irrelevant

  \[ F_1 = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}} \]

• **HEP (cross section measurement by counting)**
  – Based on *qualitative distinction between signal and background*
  – e.g. FIP1 metric *does not depend on True Negatives*
    • Measured cross section cannot depend on how many background events are rejected

  \[ \text{FIP}_1 = \frac{\text{TP}^2}{(\text{TP} + \text{FN})(\text{TP} + \text{FP})} \]

**HEP is more similar to Information Retrieval than to Medical Diagnostics**
*(qualitative asymmetry between positives and negatives)*

Invariance under TN change is only one of many useful symmetries to analyse

[Sokolova-Lapalme, Luque et al.]


HEP: cross section in a counting experiment (maximize FIP1 – the AUC is misleading!)

To minimize the statistical error $\Delta \sigma$:
\[ \text{Maximize } FIP_1 = \varepsilon_s \rho \]

Choice of operating point is simple:
- Plot $\varepsilon_s \rho$ as a function of $\varepsilon_s$
- Choose the point where $\varepsilon_s \rho$ is maximum

Choice between two classifiers is simple:
- Determine max ($\varepsilon_s \rho$) for each
- Choose the classifier with the higher max

NB1: The choice depends on prevalence
[which is fixed by physics and approximately known in advance]

NB2: AUC is misleading and irrelevant in this case

But there are better ways than a counting experiment to measure a total cross section in this case…

<table>
<thead>
<tr>
<th>Range in [0,1]</th>
<th>FIP1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Higher is better</th>
<th>FIP1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td></td>
<td>NO</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Numerically meaningful</th>
<th>FIP1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td></td>
<td>NO</td>
</tr>
</tbody>
</table>
Binary classifier metrics outside HEP – scoring classifiers

HEP: cross section by a fit to the score distribution

Use the scoring classifier D to partition events, not to accept or reject events

This is the most common method to measure a total cross section (example: a BDT or NN output fit)

Keep all Stot events and partition them in K bins

\[
FIP_2 = \frac{\mathcal{I}_{\sigma_s}}{\mathcal{I}_{\sigma_s}^{(\text{ideal})}} = \frac{\sum_k s_k \rho_k}{\sum_k s_k} = \frac{\sum_k s_k^2/n_k}{\sum_k s_k} = \frac{\sum_k n_k \rho_k^2}{\sum_k s_k}
\]

There is a benefit in partitioning events into subsets with different purities because

\[
\Delta \mathcal{I}_{\sigma_s} = \frac{n_1 n_2}{n_1 + n_2} (\rho_1 - \rho_2)^2
\]

Better than a counting experiment for two reasons
- All events are used, none are rejected
- Those which were previously in a single bin are now subpartitioned
FIP2 from the ROC (+prevalence) or from the PRC

- From the previous slide:
  \[ FIP2 = \sum_{i=1}^{m} \frac{\rho_i s_i}{S_{\text{tot}} + B_{\text{tot}}} \]

- FIP2 from the ROC (+prevalence): 
  \[ S_{\text{sel}} = S_{\text{tot}} \epsilon_s \quad B_{\text{sel}} = B_{\text{tot}} \epsilon_b \]
  \[ s_i = dS_{\text{sel}} = S_{\text{tot}} \, d\epsilon_s \quad b_i = dB_{\text{sel}} = B_{\text{tot}} \, d\epsilon_b \]
  \[ \rho_i = \frac{1}{1 + \frac{B_{\text{tot}} \, d\epsilon_b}{S_{\text{tot}} \, d\epsilon_s}} \]
  \[ FIP2 = \int_{0}^{1} \frac{d\epsilon_s}{1 + \frac{1 - \pi_s}{\pi_s} \frac{d\rho}{d\epsilon_s}} \]

- FIP2 from the PRC:
  \[ S_{\text{sel}} = S_{\text{tot}} \left( \frac{1}{\rho} - 1 \right) \quad B_{\text{sel}} = S_{\text{tot}} \, d\epsilon_s \]
  \[ s_i = dS_{\text{sel}} = S_{\text{tot}} \, d\epsilon_s \quad b_i = dB_{\text{sel}} = S_{\text{tot}} \left[ \epsilon_s \left( \frac{1}{\rho} - 1 \right) - \epsilon_s \frac{d\rho}{\rho^2} \right] \]
  \[ \rho_i = \frac{\rho}{1 - \epsilon_s \frac{d\rho}{\rho \, d\epsilon_s}} \]
  \[ FIP2 = \int_{0}^{1} \frac{\rho \, d\epsilon_s}{1 - \epsilon_s \frac{d\rho}{\rho \, d\epsilon_s}} \]

- Easier calculation and interpretation from ROC (+prevalence) than from PRC
  - region of constant ROC slope = region of constant signal purity
  - decreasing ROC slope = decreasing purity
  - technicality (my Python code): convert ROC to convex hull* first

FIP2: integrals on ROC and PRC, more relevant to HEP than AUC or AUCPR!
(well-defined meaning for distribution fits)

Compare FIP2(ROC) to AUC
\[ \text{AUC} = \int_{0}^{1} \epsilon_s \, d\epsilon_b = 1 - \int_{0}^{1} \epsilon_b \, d\epsilon_s \]

Compare FIP2(PRC) to AUCPR
\[ \text{AUCPR} = \int_{0}^{1} \rho \, d\epsilon_s \]

*Convert ROC to convex hull
- ensure decreasing slope
- avoid staircase effect that would artificially inflate FIP2
(bins of 100% purity: only signal or only background)
HEP estimation of parameter $\theta$ in a binned distribution fit

$\text{FIP}^2_{(\text{max})}$ example
(and overtraining)

$\text{FIP}^2$ is a metric in $[0,1]$ but the detector resolution effectively determines a $\text{FIP}^2_{(\text{max})} < 1$
HEP estimation of parameter $\theta$ in a binned distribution fit

**Fisher information $I_\theta$ about $\theta$ (statistical errors)**

For a given partitioning scheme with $K$ bins
($n_k$ is the number of selected events in bin $k$)

\[
I_\theta = \frac{1}{(\Delta \theta)^2} = \sum_{k=1}^{K} \frac{1}{(\Delta \theta)^2_k} = \sum_{k=1}^{K} n_k \left( \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2
\]

Statistical errors: information adds up

Each bin is an independent measurement with error

\[
(\Delta \theta)_k = \left( \frac{\partial n_k}{\partial \theta} \right)^{-1} \Delta n_k = \left( \frac{\partial n_k}{\partial \theta} \right)^{-1} \sqrt{n_k}
\]

(Combination more complex with systematic errors, or for searches)
HEP estimation of parameter $\theta$ in a binned distribution fit

Optimal partitioning

$$I_\theta = \frac{1}{(\Delta \theta)^2} = \sum_{k=1}^{K} \frac{1}{(\Delta \theta)_k^2} = \sum_{k=1}^{K} n_k \left( \frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2$$

Is there a benefit (information inflow) in splitting bin 0 into two bins 1, 2 with $n_0 = n_1 + n_2$?

Information increases if the two new bins have different sensitivities to $\theta$

$$\Delta I_\theta = \frac{1}{n_1} \left( \frac{\partial n_1}{\partial \theta} \right)^2 + \frac{1}{n_2} \left( \frac{\partial n_2}{\partial \theta} \right)^2 - \frac{1}{n_1 + n_2} \left( \frac{\partial (n_1 + n_2)}{\partial \theta} \right)^2$$

$$= \frac{n_1 n_2}{n_1 + n_2} \left[ \left( \frac{1}{n_1} \frac{\partial n_1}{\partial \theta} \right) - \left( \frac{1}{n_2} \frac{\partial n_2}{\partial \theta} \right) \right]^2$$

$$\Delta I_\theta > 0 \iff \left( \frac{1}{n_1} \frac{\partial n_1}{\partial \theta} \right) \neq \left( \frac{1}{n_2} \frac{\partial n_2}{\partial \theta} \right)$$

Goal of a distribution fit: partition events into subsets with different bin-by-bin sensitivities to $\theta$
Background events by definition are insensitive to $\theta$

Signal events may have positive, zero or negative sensitivity

$$\gamma_i = \left( \frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right) = 0, \quad \text{if } i \in \{\text{Background}\}$$

$$\gamma_i = \left( \frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right) \in \{-\infty, +\infty\}, \quad \text{if } i \in \{\text{Signal}\}$$

$$\delta_i = \begin{cases} 1 & \text{if } i \in \{\text{Signal}\} \\ 0 & \text{if } i \in \{\text{Background}\} \end{cases}$$

The distinction between signal events with low ($|\gamma_i|\sim 0$) sensitivity and background events is blurred (example: events far from an invariant mass peak)

In a cross section measurement

All background events are equivalent to one another

All signal events are equivalent to one another

$$\gamma_i = \frac{1}{\sigma_s} \delta_i = \begin{cases} \frac{1}{\sigma_s} & \text{if } i \in \{\text{Signal}\}, \\ 0 & \text{if } i \in \{\text{Background}\}, \end{cases} \quad \text{if } \theta \equiv \sigma_s$$

Changing the signal cross section $\sim$ is a global rescaling of all differential distributions

$$s_k(\sigma_s) = \frac{\sigma_s}{\sigma_{s,\text{ref}}} \times s_k(\sigma_{s,\text{ref}})$$

$\theta$: cross section $\sigma_s$ DICHOTOMOUS

$\theta$: mass, coupling NON-DICHOTOMOUS
HEP estimation of parameter $\theta$ in a binned distribution fit

**FIP1 and FIP2 revisited**

$FIP_{sha} = 1$ for both
(dichotomous, all signal events are equivalent)

\[
FIP_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{\gamma_i} S_{tot}^2} = FIP_{eff} \times FIP_{sha} \times FIP_{pur}
\]

\[
= \frac{\sum_{i=1}^{S_{sel}} \gamma_i^2}{\sum_{i=1}^{S_{tot}} \gamma_i^2} \times \frac{\sum_{k=1}^{K} s_k \phi_k^2}{\sum_{i=1}^{S_{sel}} \gamma_i^2} \times \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{tot}} \gamma_i^2}
\]

$FIP_1 = \epsilon_s \varrho$

**FIP1:**
$FIP_{eff} = \epsilon$
$FIP_{pur} = \rho$

\[
FIP_2 = \frac{I_{\sigma_s}}{I_{\sigma_s}^{(ideal)}} = \frac{\sum_k s_k \rho_k}{\sum_k s_k} = \frac{\sum_k s_k^2/n_k}{\sum_k s_k} = \frac{\sum_k n_k \rho_k^2}{\sum_k s_k}
\]

**FIP2:**
$FIP_{eff} = 1$
$FIP_{pur} = FIP_2$

\[
FIP_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{\gamma_i} S_{tot}^2} = FIP_{eff} \times FIP_{sha} \times FIP_{pur}
\]

\[
= \frac{\sum_{i=1}^{S_{sel}} \gamma_i^2}{\sum_{i=1}^{S_{tot}} \gamma_i^2} \times \frac{\sum_{k=1}^{K} s_k \phi_k^2}{\sum_{i=1}^{S_{sel}} \gamma_i^2} \times \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{tot}} \gamma_i^2}
\]
HEP estimation of parameter $\theta$ in a binned distribution fit

**From CRLB to Fisher Information Part (FIP)**

- **Particles produced in beam collisions**
- **Raw data events**
- **Analysis object data**
- **Event counts in individual bins of a distribution**
- **Measured value of the parameter $M \pm \Delta M$**

**Detector Trigger**

**Data processing**

**PHYSICS ANALYSIS**
- Event selection (Sig vs Bkg)
- Event partitioning

**Max likelihood fit**

**$I(\mathit{ideal})$**

**FIP**

$$\text{FIP} = \frac{I_\theta}{I_\theta^{(\mathit{ideal})}} = \left(\frac{\Delta\theta^{(\mathit{ideal})}}{\Delta\theta}\right)^2 \leq 100\%$$

A max likelihood fit is 100% efficient: it achieves the CRLB, for the given event selection and event partitioning.
**HEP estimation of parameter $\theta$ in a binned distribution fit**

**Two optimization handles: event selection and partitioning**

- **Particles produced in beam collisions**
- **Detector Trigger**
- **Data processing**

**PHYSICS ANALYSIS**
- **Event selection (Sig vs Bkg)**
- **Event partitioning**

**Max likelihood fit**
- $1/\Delta \theta^2$

**Signal events**
- $S_{\text{ALL}}$
- $S_{\text{tot}}$
- $S_{\text{sel}}$

**FIP handles**
- $FIP_{\text{ALL}}$
- $FIP_{\text{ACC}}$
- $FIP_3$

I factor out detector/trigger acceptance and compute FIP3 with respect to $S_{\text{tot}}$. 

**Measured value of the parameter $M \pm \Delta M$**