Optimizing HEP parameter fits: Fisher Information ML metrics and Weight Derivative Regression (WDR)

Andrea Valassi (CERN IT)

Les Houches, 16 June 2023

Based on previous presentations (no new work since CHEP2019!):

IML Jul2021 - slides (https://indico.cern.ch/event/1054595)

CHEP2019 – paper (doi:10.1051/epjconf/202024506038), slides(doi:10.5281/zenodo.3523164)

CHEP2018 – paper (doi:10.1051/epjconf/201921406004), slides (doi:10.5281/zenodo.1303387)

IML Jan2018 – slides (doi:10.5281/zenodo.1300684), including a few toy examples

Overview: several related topics here

- In particular:
  - Machine Learning metrics: training (loss functions) and evaluation
  - My interest (from ALEPH times!): parameter measurements, e.g. EFTs
  - My work now: Madgraph matrix elements on GPU... useful for ME reweighting!

- **Disclaimer: I only present some ideas, not any concrete applications**
  - I am not doing physics analysis in any experiment now, I have no data
  - I am happy to discuss and collaborate if you want to try this out...

Different problems require different tools (and different ML metrics...)

I focus here on HEP measurements of a parameter θ

(and in particular on statistically limited measurements - but what I say has applications to systematics too)

The final goal:
minimize the error $\Delta\theta$ on the measurement
i.e.
maximize the Fisher information $(1/\Delta\theta)^2$ about θ

$\rightarrow$ the idea: use Fisher information
as both the training and evaluation metric for any ML tool we use

(Aside: my personal opinion is that the AUC is utterly irrelevant in HEP... we can discuss that!)

Measurement of a parameter θ

Binned fit for θ → Compare data in bin k to
model prediction n_k as a function of θ

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

You need samples of MC events for different values of θ
There are two solutions:

EITHER: Generate N different samples
- Expensive: N x detector simulation
- θ-dependency affected by MC statistics

OR: Generate 1 sample + Reweighting
+ Cheaper: 1 x detector simulation
+ *θ-dependency for each event*

Model prediction for $n_k(\theta)$

Event-by-event Monte Carlo reweighting

1. Generate signal sample at θ_{ref}

2. Full detector simulation

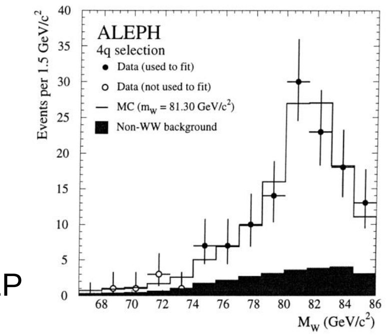
(MC truth $\mathbf{x}_i^{(\mathrm{true})} \rightarrow$ observed $\mathbf{x}_i$)

*3. **Reweight each event from θ to θ_{ref} by the MC matrix element ratio***

$$w_i(\theta) = \frac{|\mathcal{M}(\theta, \mathbf{x}_i^{(\mathrm{true})})|^2}{|\mathcal{M}(\theta_{\mathrm{ref}}, \mathbf{x}_i^{(\mathrm{true})})|^2}$$

ALEPH Collaboration, *Measurement of the W mass by direct reconstruction in e^+e^- collisions at 172 GeV*, Phys. Lett. B 422 (1998) 384. doi:10.1016/S0370-2693(98)00062-8

Example from LEP



Note: in Madgraph5_aMC@NLO, we are working on improving the infrastructure for reweighting samples of LHE event files (e.g. for EFT studies)

Event-by-event sensitivity to θ

MC-truth weight derivative γ_i

For every MC event: compute and store the

derivative with respect to θ of the MC weight w_i

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

Here assume unweighted sample at $\theta=\theta_{ref}$ hence $w_i(\theta_{ref})=1$

$\partial w/\partial\theta$ is related to the Fisher score
(but the latter is the derivative
of a probability normalized to 1)

I argue that this is

the most important MC-truth property of an event

in a fit for θ

Rephrase: if you want to "unfold"
some generator-level event properties from detector-level observable,
you only really need to unfold this single variable!
More on this later (Weight Derivative Regression)

The goal: partition by the evt-by-evt sensitivity γ_i

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 γ_{i_1} and γ_{i_2} are different

$$\Delta\mathcal{I}_\theta = \gamma_{i_1}^2 + \gamma_{i_2}^2 - 2\left(\frac{\gamma_{i_1}+\gamma_{i_2}}{2}\right)^2 = \frac{1}{2}(\gamma_{i_1}-\gamma_{i_2})^2$$

Goal of a distribution fit: partition (RESOLVE!) events by their different MC-truth event-by-event sensitivities γ_i to θ

What counts in your fit is your ***sharpness***:
how good you are at separating (resolving) events with different γ_i
(a term coming from probabilistic metrics – Brier score – in Meteorology)

$\mathtt{I}_\theta$ with an ideal detector
(for a given luminosity)

IF
you were able to know the true (generator-level) weight derivative of each event,
THEN
you would be able to achieve the IDEAL MESUREMENT OF θ

$$\mathcal{I}_\theta = \sum_{k=1}^{K} n_k \left(\frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2 = \sum_{k=1}^{K} n_k \langle \gamma \rangle_k^2$$

Maximum achievable I_θ with an ideal detector

(1) Ideal acceptance
(2) Ideal resolution on γ_i
(3) Ideal background rejection

$$\boxed{\mathcal{I}_\theta^{(\text{ideal})} = \sum_{i=1}^{N_{\text{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\text{tot}}} \gamma_i^2}$$

Minimum achievable error for a given luminosity

Easy to calculate for any measurement (including EFT)

Fisher Information Part (FIP)

Fisher Information Part (FIP)
fraction of "ideal" information
retained by a given analysis

$$\boxed{\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \frac{(\Delta\theta^{(\mathrm{ideal})})^2}{(\Delta\theta)^2} \leq 100\%}$$

FIP is a metric between 0 and 1 – higher is better

Applications?

(1) Quantitatively understand why $\Delta\theta$ is larger than the minimum ideal $\Delta\theta$

(2) May be used as both an evaluation metric and a training metric for ML analyses

Weight Derivative Regression (WDR): train q_i for γ_i

Goal of a distribution fit: separate events with different MC-truth event-by-event sensitivities γ_i to θ

But γ_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 a mix of signal and background MC events

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

Training metric: maximize FIP*
Evaluation metric: maximize FIP

(* ~equivalent to **minimizing MSE for γ_i**)

WDR vs. Optimal Observables

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 γ_i to θ
(OO does not use MC weight derivatives but it is similar)

D. Atwood, A. Soni, *Analysis for magnetic moment and electric dipole moment form factors of the top quark via $e^+e^- \to t\bar{t}$*, Phys. Rev. D 45 (1992) 2405. doi:10.1103/PhysRevD.45.2405,
M. Davier, L. Duflot, F. LeDiberder, A. Rougé, *The optimal method for the measurement of tau polarization*, Phys. Lett. B 306 (1993) 411. doi:10.1016/0370-2693(93)90101-M
M. Diehl, O. Nachtmann, *Optimal observables for the measurement of three-gauge-boson couplings in $e^+e^- \to W^+W^-$*, Z. Phys. C 62 (1994) 397. doi:10.1007/BF01555899
O. Nachtmann, F. Nagel, *Optimal observables and phase-space ambiguities*, Eur. Phys. J. C40 (2005) 497. doi:10.1140/epjc/s2005-02153-9

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

MC-truth
sensitivity $\gamma_i^{(\text{MC truth})}$

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

MC observable
event properties $x_i^{(\text{MC})}$

***MC-truth to MC-truth* functional dependence**

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

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

Fit optimal observable

$$OO_i^{(\text{DATA})} = f(\, x_i^{(\text{DATA})}\,)$$
$$OO_i^{(\text{MC})} = f(\, x_i^{(\text{MC})}\,)$$

The difference is in where the
effect of detector resolution
is taken into account

Detector-level MC to MC-truth
Weight Derivative Regression

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

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

Fit WDR regressor

$$q_i^{(\text{DATA})} = q(\, x_i^{(\text{DATA})}\,)$$
$$q_i^{(\text{MC})} = q(\, x_i^{(\text{MC})}\,)$$

WDR vs. other related methods (slide needs updating...)

- Matrix Element Method
  - MEM: needs event-by-event integrals (convolution of detector response function)
  - *WDR: no event-by-event integrals* (full detector simulation + reweighting)

- MadMiner, "Mining gold", Sally, “Learning the score” (Brehmer, Cranmer et al.)
  - arXiv:1805.00020, arXiv:1907.10621, arXiv:2010.06439 ...
    - Plus Brehmer et al.’s earlier work on information geometry arXiv:1612.05261
  - Sally learns the Fisher score – very similar to WDR which learns $(\partial w/\partial\theta)/w$
  - *WDR: derived from predictions of real & ideal $\Delta\theta$; focuses on metrics (FIP, MSE)*
  - I know Sally too little to give more comments on the similarities or differences...

- ThickBrick (Matchev, Shyamsundar) – doi:10.1007/JHEP03(2021)291
  - Part 1 binary classifiers (signal/bkg), Part 2 (θ fits) was in preparation in 2021

- Quadratic Classifier (Chen, Glioti, Panico, Wulzer) – arXiv:2007.10356
  - Learn $w(\theta)/w(\theta_{SM})$ – while WDR learns $(\partial w/\partial\theta)/w$

- Boosted Information Trees (Chatterjee, Schoefbeck, Schwarz et al.)
  - arXiv:2107.10859, arXiv.org:2205.12976 - similar to WDR in its training metrics

...Aside...
ML metrics: us (HEP) and them (other sciences)

- Again: **<u>different problems require different tools and metrics!</u>**
  - Solutions developed in other sciences may work for us, or they may not
  - And there is no one-size-fits-all solution (in this talk: parameter measurements!)

- Example: "everyone" in ML (and HEP?) seems to use the AUC
  - Who invented the AUC? Why? Is it *relevant* for us in HEP? *Read, read, read!*
  - The AUC comes from Psychophysics and Medical Diagnostics: it is the *"probability that a randomly chosen diseased subject is correctly <u>ranked</u> with greater suspicion than a randomly chosen non-diseased subject"*
  - In HEP, then, the AUC is the *"probability that a randomly chosen signal event is correctly <u>ranked</u> more signal-like than a randomly chosen background event"...*
    - ...so what?? is this relevant for us?... (and what about insensitivity of the AUC to prevalence? or crossing ROCs? or the irrelevance of Rejected Background in HEP?)
  - In my opinion: in HEP we do not need ranking metrics, we need probabilistic metrics that take into account partitioning... just like in Meteorology!
    - FIP metrics have many similarities to the Brier score and other Meteorology metrics...

Conclusions

- Metrics, metrics, metrics!
  - Each problem needs different evaluation and training metrics: which ones for you?
  - Do not be afraid to build your own metrics (pen, paper, no laptop...)

- Read, read, read!
  - Understand why others developed some tools that we use (is AUC relevant?)
  - Get new ideas from others that may help us (who would dream of Meteorology?!)

- What I have presented – which may be relevant for EFT measurements
  - A formalism to discuss the expected $\Delta\theta$ using evt-by-evt MC weight derivatives γ_i
    - *First and foremost: separating events with different γ_i is the most important goal in a fit*
    - A quantitative prediction of the *minimum achievable $\Delta\theta$ with an ideal detector*
    - A quantitative description (FIP) of how much we lose from efficiency, sharpness, purity
  - A proposal to *use a regressor of γ_i as the "optimal observable" for fits*
    - Probably many similarities to the "learning the Fisher score" and BIT approaches
  - A proposal to use FIP (or maybe MSE) as both training and evaluation metric

- What I have not presented: results from concrete applications
  - If you are interested to try this out, do not hesitate to contact me!

THANK YOU! QUESTIONS?

Reading Room, British Museum

Reading is a revolutionary act
(Inge Feltrinelli, 1930-2018)

Backup slides

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

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002

A. Luque, A Carrasco, A. Martin, J. R. Lama, *Exploring Symmetry of Binary Classification Performance Metrics*, Symmetry 11 (2019) 47. doi:10.3390/sym11010047.

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

C. Ferri, J. Hernández-Orallo, R. Modroiu, *An Experimental Comparison of Classification Performance Metrics*, Proc. Learning 2004, Elche (2004). http://dmip.webs.upv.es/papers/Learning2004.pdf

C. Ferri, J. Hernández-Orallo, R. Modroiu, *An Experimental Comparison of Performance Measures for Classification*, Pattern Recognition Letters 30 (2009) 27. doi:10.1016/j.patrec.2008.08.010

Example: AUC (ranking) vs. MSE (probabilistic): in my CHEP2019 talk

The sensitivity γ_i depends on θ

It is a derivative! $\gamma_i|_\theta = \left(\frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \right)_\theta$

Dependency on θ is very strong if θ is a mass

Dependency on θ is lower if θ is a coupling (e.g. EFT?)

In the following discussion of expected errors,
may assume that it is calculated at the true value of θ

Compute the expected $\Delta\theta$ from the sensitivity γ_i

Express $\Delta\theta$ in terms of the

Fisher Information about θ

$$I_\theta = 1/(\Delta\theta)^2$$

Minimizing $\Delta\theta$ means maximizing I_θ

Easy to compute the statistical error $\Delta\theta$
in terms of average bin-by-bin sensitivities:

$$\boxed{\mathcal{I}_\theta = \sum_{k=1}^{K} n_k \left(\frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2 = \sum_{k=1}^{K} n_k \langle \gamma \rangle_k^2}$$

Beyond the signal-background dichotomy
(hence: from binary classification to non-binary regression)

Background events have γ_i=0
because by definition they are insensitive to θ

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

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

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 ($|\gamma_i|$~0)

- Signal events with low sensitivity are a nuisance just as much as background events
  - Mixing high-γ_i signal events with background or low-γ_i signal *dilutes* their sensitivity!

- NB: binary classification (signal/background) extensively discussed in CHEP2018 talk
  - cross section measurement by counting experiments: FIP1 metric
  - cross section measurement by scoring classifier fits: FIP2 metric

There are three ways in which your analysis may be suboptimal!

FIP decomposition: efficiency, sharpness, purity

$$\mathrm{FIP}_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2} = \mathrm{FIP}_{\mathrm{eff}} \times \mathrm{FIP}_{\mathrm{sha}} \times \mathrm{FIP}_{\mathrm{pur}}$$
$$= \boxed{\frac{\sum_{i=1}^{S_{\mathrm{sel}}} \gamma_i^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2}} \times \boxed{\frac{\sum_{k=1}^{K} s_k \phi_k^2}{\sum_{i=1}^{S_{\mathrm{sel}}} \gamma_i^2}} \times \boxed{\frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{k=1}^{K} s_k \phi_k^2}}$$

Sharpness: how well do we separate events with different sensitivities γ_i? (term from Meteorology)

- (1) <u>γ_i^2-weighted signal efficiency</u>
  - –you are losing some signal events (especially bad if they have high γ_i)
- (2) <u>γ_i resolution (sharpness!) for signal events</u>
  - –you are mixing signal events with different γ_i (diluting those of high γ_i)
- (3) <u>γ_i^2-weighted signal purity</u>: γ_i resolution (sharpness) for background events
  - –you are mixing background (γ_i=0) and signal (especially bad if high γ_i)

- Again, this is true no matter which analysis method you use...

Limits to knowledge: a realistic detector

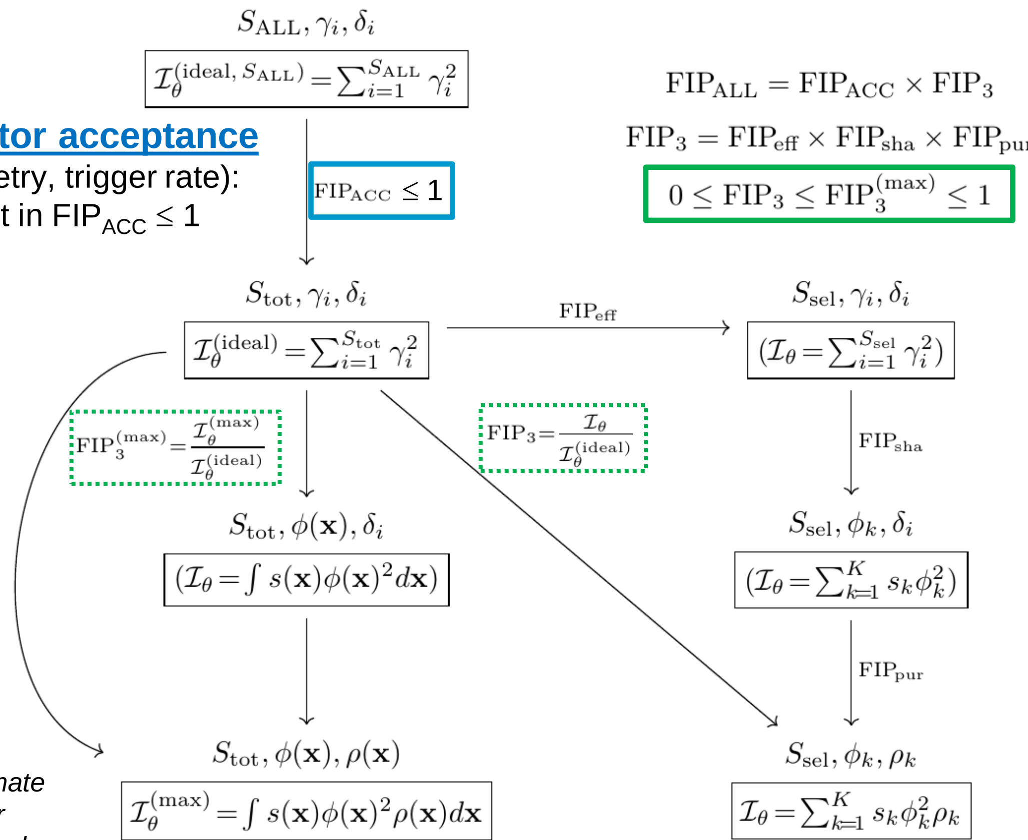
Limited detector acceptance
(detector geometry, trigger rate):
factor this out in $\mathrm{FIP}_{\mathrm{ACC}} \leq 1$

Limited detector resolution
In the multi-dimensional space of event observables **x**,
it is impossible to resolve:

- signal events with high γ_i from signal events with low γ_i: average sensitivity is $\phi(\mathbf{x})$

- signal events δ_i=1 from background events δ_i=0: average purity is $\rho(\mathbf{x})$

Note: will not discuss here if/how you can/should estimate $\mathrm{FIP}^{(max)}$ from $\phi(\mathbf{x})$ and $\rho(\mathbf{x})$ – would need a detector response convolution as in the matrix element method



FIP is a metric in [0,1], but the detector acceptance and resolution limit it to $0 \leq \mathrm{FIP} \leq \mathrm{FIP}^{(max)} < 1$

⇨

If you think you achieved $\mathrm{FIP} > \mathrm{FIP}^{(max)}$ in ML training then you are **overtraining** (you are resolving events better than your detector allows you to do)

Maximizing FIP = Minimizing MSE
(in Decision Trees)

Mean Squared Error (regressor q_i vs true γ_i)

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

MSE decomposition: **MSE = $\mathrm{MSE}_{\mathrm{cal}}$ (calibration) + $\mathrm{MSE}_{\mathrm{sha}}$ (sharpness)**
Paraphrases the Brier score decomposition in Meteorology!

G. W. Brier, *Verification of forecasts expressed in terms of probability*, Weather Rev. 78 (1950) 1. doi:10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2
F. Sanders, *On Subjective Probability Forecasting*, J. Applied Meteorology 2 (1963) 191. https://www.jstor.org/stable/26169573

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

FIP is related to $\mathrm{MSE}_{\mathrm{sha}}$

$$\boxed{\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \left(1 - \frac{N_{\mathrm{tot}} \times \mathrm{MSE}_{\mathrm{sha}}}{\mathcal{I}_\theta^{(\mathrm{ideal})}} \right)}$$

Decision Tree training : $\mathrm{MSE}_{\mathrm{cal}}$=0 by construction; maximizing FIP is equivalent to minimising MSE!

Consequence: reasonable to minimize MSE also to train other regressors (NNs)?

Further information in previous talks/papers

- Cross section $\theta=\sigma_s$ measurements: binary classification (CHEP2018 talk)
  - Signal events are all equivalent (same event-by-event sensitivity $\gamma_i=1/\sigma_s$ to $\theta=\sigma_s$)
  - Counting experiments: metric FIP1 (efficiency * purity)
  - Scoring classifier fit: metric FIP2
    - Equivalence between FIP2 and Gini for decision tree training
    - FIP2 as an integral from the ROC and PRC – difference with areas AUC and AUCPR
  - Learning from others: Medical Diagnostics, Information Retrieval, ML
    - Comparison with AUC, Accuracy, F1
    - Extensive literature on AUC limitations and crossing ROC curves
    - Symmetries and invariances (TNs are irrelevant in HEP)
    - Beyond binary classification: DCG, example-dependent cost-sensitive classification...

- Parameter θ measurements: from binary classification to regression (CHEP2019 talk)
  - Signal events are not all equivalent (different event-by-event sensitivity γ_i to θ)
  - More complete discussion of parameter fits, WDR: metric FIP3 (this talk)
  - Learning from others: Meteorology, Medical Prognostics
    - Three types of metrics: threshold, ranking, probabilistic
    - HEP needs probabilistic metrics (just like Meteorology and Medical Prognostics)

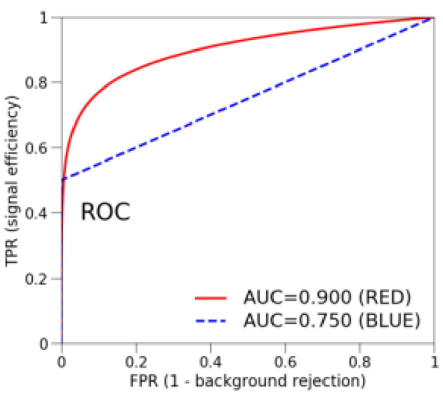
- Many details are available in the backup slides at the end of this slide deck

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

Medical Diagnostics → *ranking evaluation of diagnostic prediction*
Patient A is diagnosed as more likely sick than B: how often am I right?



D. M. Green, *General Prediction Relating Yes-No and Forced-Choice Results*, J. Acoustical Soc. Am. 36 (1964) 1042. doi:10.1121/1.2143339
D. J. Goodenough, K. Rossmann, L. B. Lusted, *Radiographic applications of signal detection theory*, Radiology 105 (1972) 199. doi:10.1148/105.1.199
J. A. Hanley, B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology 143 (1982) 29. doi:10.1148/radiology.143.1.7063747
A. P. Bradley, *The use of the area under the ROC curve in the evaluation of Machine Learning algorithms*, Pattern Recognition 30 (1997) 1145. doi:10.1016/S0031-3203(96)00142-2

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

IRRELEVANT FOR HEP PARAMETER FITS?

Partitioning, and probabilistic metrics
Group events and make a forecast on each subset

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

$$\mathrm{MSE} = \overbrace{\boxed{\frac{1}{N_{\mathrm{tot}}}\left[\sum_{k=1}^{K} n_k \left(q_{(k)} - \langle\gamma\rangle_k\right)^2\right]}}^{\text{Validity, Reliability, Calibration}} + \overbrace{\boxed{\frac{1}{N_{\mathrm{tot}}}\left[\left(\sum_{i=1}^{N_{\mathrm{tot}}} \gamma_i^2\right) - \left(\sum_{k=1}^{K} n_k \langle\gamma\rangle_k^2\right)\right]}}^{\text{Sharpness, Resolution, Refinement}}$$

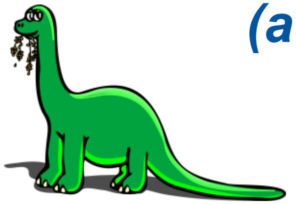
Sharpness (from MSE): how well can I resolve days with 10% and 90% chance of rain? Patients with 10% and 90% 5yr survival rate? Signal events with high sensitivity to θ from (signal or background) events with low sensitivity?

ESSENTIAL FOR HEP PARAMETER FITS!

More backup slides

Foreword – a classic ML problem: regressor training

(a frequentist dinosaur's view of Machine Learning)



https://openclipart.org

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

Choosing a ML methodology implies several choices:

0. The true variable $\gamma(x)$ to regress

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

2. The training metric: a "distance" of $R_\gamma(\mathbf{x}_i)$ to $\gamma(\mathbf{x}_i)$ or γ_i to minimize, or a property of $R_\gamma(\mathbf{x}_i)$ to maximize
Examples: Gini, Shannon entropy/information, MSE...

3. The evaluation metric: how good is $R_\gamma(\mathbf{x})$? is it better than $R'_\gamma(\mathbf{x})$?
Examples: ROC, AUC, MSE, Brier...

For parameter fits:
Weight Derivative
Regression (WDR)

*(I focus on **Decision Trees** because of the similarities to binned distribution fits; but the idea applies also to NNs et al...)*

<u>Always use the same metric for training and evaluation!</u>

For parameter fits:
***Fisher Information Part** (FIP)*

Executive summary: WDR in a nutshell (one slide!)

- Goal: minimize statistical error $\Delta\theta$ in fit of a single parameter θ (e.g. EFT coupling)
  - i.e. maximize Fisher information $\mathcal{I}_\theta = \frac{1}{(\Delta\theta)^2}$ about θ

- Write analytical formula for expected $\Delta\theta$ (in terms of I_θ) in a binned fit
  - I_θ depends on the event-by-event MC-truth weight derivative $\gamma_i|_\theta = \left(\frac{1}{w_i}\frac{\partial w_i}{\partial\theta}\right)_\theta$
  - Note: for background, γ_i=0; for signal, $\gamma_i \in [+\infty,-\infty]$ (but γ_i=1 for cross section fits)

- *Easy to prove that I_θ is maximized if events are binned according to their true γ_i*
  - Derive a formula for the maximum achievable $\mathcal{I}_\theta^{(\mathrm{ideal})} = \sum_{i=1}^{N_{\mathrm{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2$ in an "ideal" case
  - Evaluation metric (quality of the measurement): FIP=$\mathrm{I}_\theta/\mathrm{I}_\theta^{(\mathrm{ideal})}$ (with FIP$\in$[0,1])
  - Bonus, factorize FIP into: (1,2) γ_i-weighted efficiency, purity; (3) signal γ_i-resolution

- <u>*ML implications: for each event i, you only need to know the MC-truth value of γ_i*</u>
  - From detector level observables **x**, build a regressor R_γ(**x**) for γ(**x**); fit θ from R_γ(**x**)
  - Training metric if R_γ is a decision trees (DT): maximise FIP (i.e. minimize $\Delta\theta$)
  - Easy to see FIP is related to MSE for DTs: minimize MSE to train R_γ if it is a NN

- Bonus, compare HEP evaluation/training metrics to those of other domains
  - FIP is a *probabilistic metric*, as Brier score in Meteorology (same decomposition!)
  - We do needs *threshold metrics*, but only in counting experiments (e.g. FIP1=$\varepsilon\rho$)
  - IMO, *ranking metrics* (e.g. AUC from Medical Diagnostics) are irrelevant in HEP...

HEP cross-section in a counting experiment

- Measurement of a total cross-section σ_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 ε_s*ρ* ("common knowledge" in the LEP2 experiments) $\rightarrow$ *"FIP1"*
  - NB: This metric only makes sense for this specific HEP optimization problem!

$$\mathcal{I}_{\sigma_s} = \frac{1}{(\Delta\sigma_s)^2} = \frac{1}{{\sigma_s}^2}\,\epsilon_s\,\varrho S_{\mathrm{tot}} = \frac{1}{{\sigma_s}^2}\left(\frac{S_{\mathrm{sel}}^2}{S_{\mathrm{sel}}+B_{\mathrm{sel}}}\right)$$

$$\mathcal{I}_{\sigma_s}^{(\mathrm{ideal})} = \frac{S_{\mathrm{tot}}}{\sigma_s^2}\,,\ \mathrm{if}\ \varrho = 1\ \mathrm{and}\ \epsilon_s = 1$$

$$\Rightarrow\quad \boxed{\mathrm{FIP}_1 = \frac{\mathcal{I}_{\sigma_s}}{\mathcal{I}_{\sigma_s}^{(\mathrm{ideal})}} = \epsilon_s\,\varrho}$$

By the way: $\rho/\varepsilon_s=1$ where $\partial\mathrm{FIP1}/\partial\rho=\partial\mathrm{FIP1}/\partial\varepsilon_s$ (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

$$\boxed{\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}}$$

- **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

$$\boxed{F_1 = \frac{2\ \mathrm{TP}}{2\ \mathrm{TP}+\mathrm{FP}+\mathrm{FN}}}$$

- **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

$$\boxed{\mathrm{FIP}_1 = \frac{\mathrm{TP}^2}{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})}}$$

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.]

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002

A. Luque, A Carrasco, A. Martin, J. R. Lama, *Exploring Symmetry of Binary Classification Performance Metrics*, Symmetry 11 (2019) 47. doi:10.3390/sym11010047.

HEP: cross section in a counting experiment

(maximize FIP1 – the AUC is misleading!)

To minimize the statistical error $\Delta\sigma$:

Maximize $\boxed{\mathrm{FIP}_1 = \epsilon_s \varrho}$

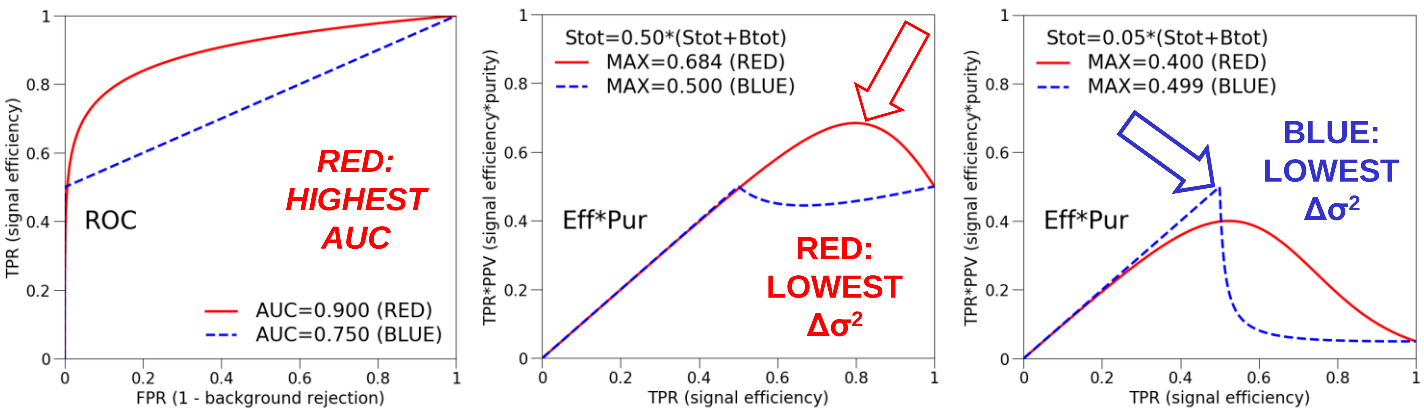
Choice of operating point is simple:
- Plot $\varepsilon_s \times \rho$ as a function of ε_s
- Choose the point where $\varepsilon_s \times \rho$ is maximum

Choice between two classifiers is simple:
- Determine max ($\varepsilon_s \times \rho$) 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…



	FIP1	AUC
Range in [0,1]	YES	YES
Higher is better	YES	NO
Numerically meanigful	YES	NO

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

$$\mathrm{FIP}_2 = \frac{\mathcal{I}_{\sigma_s}}{\mathcal{I}_{\sigma_s}^{(\mathrm{ideal})}} = \frac{\sum_k s_k \rho_k}{\sum_k s_k} = \frac{\sum_k s_k^2 / n_k}{\sum_k s_k} = \frac{\sum_k n_k \rho_k^2}{\sum_k s_k}$$

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: $\mathrm{FIP2} = \frac{\sum_{i=1}^{m} \rho_i s_i}{\sum_{i=1}^{m} s_i}$

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

- FIP2 from the ROC (+prevalence $\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}} + B_{\mathrm{tot}}}$):

$$\begin{array}{l} S_{\mathrm{sel}} = S_{\mathrm{tot}}\, \epsilon_s \\ B_{\mathrm{sel}} = B_{\mathrm{tot}}\, \epsilon_b \end{array} \Longrightarrow \begin{array}{l} s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\, d\epsilon_s \\ b_i = dB_{\mathrm{sel}} = B_{\mathrm{tot}}\, d\epsilon_b \end{array} \Longrightarrow \boxed{\rho_i = \frac{1}{1 + \frac{B_{\mathrm{tot}}}{S_{\mathrm{tot}}} \frac{d\epsilon_b}{d\epsilon_s}}} \Longrightarrow \boxed{\mathrm{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s} \frac{d\epsilon_b}{d\epsilon_s}}}$$

Compare FIP2(ROC) to AUC

$$\boxed{\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s}$$

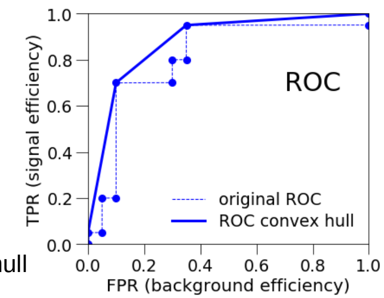
- FIP2 from the PRC:

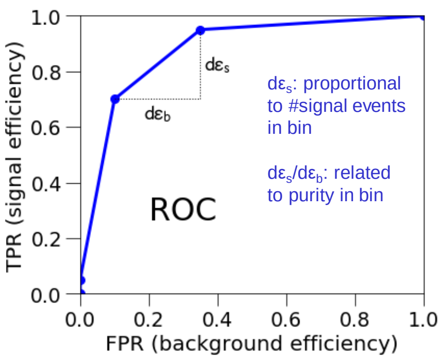
$$\begin{array}{l} S_{\mathrm{sel}} = S_{\mathrm{tot}}\, \epsilon_s \\ B_{\mathrm{sel}} = S_{\mathrm{sel}} \left(\frac{1}{\rho} - 1\right) \end{array} \Longrightarrow \begin{array}{l} s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\, d\epsilon_s \\ b_i = dB_{\mathrm{sel}} = S_{\mathrm{tot}} \left[d\epsilon_s \left(\frac{1}{\rho} - 1\right) - \epsilon_s \frac{d\rho}{\rho^2} \right] \end{array} \Longrightarrow \rho_i = \frac{\rho}{1 - \frac{\epsilon_s}{\rho} \frac{d\rho}{d\epsilon_s}} \Longrightarrow \boxed{\mathrm{FIP2} = \int_0^1 \frac{\rho\, d\epsilon_s}{1 - \frac{\epsilon_s}{\rho} \frac{d\rho}{d\epsilon_s}}}$$

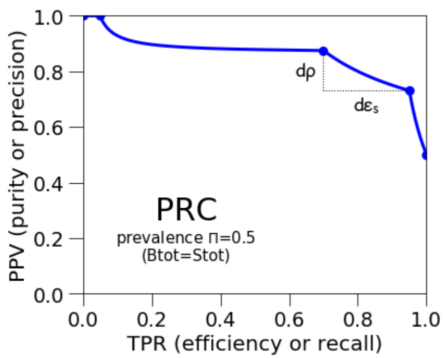
Compare FIP2(PRC) to AUCPR

$$\boxed{\mathrm{AUCPR} = \int_0^1 \rho\, d\epsilon_s}$$

- 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





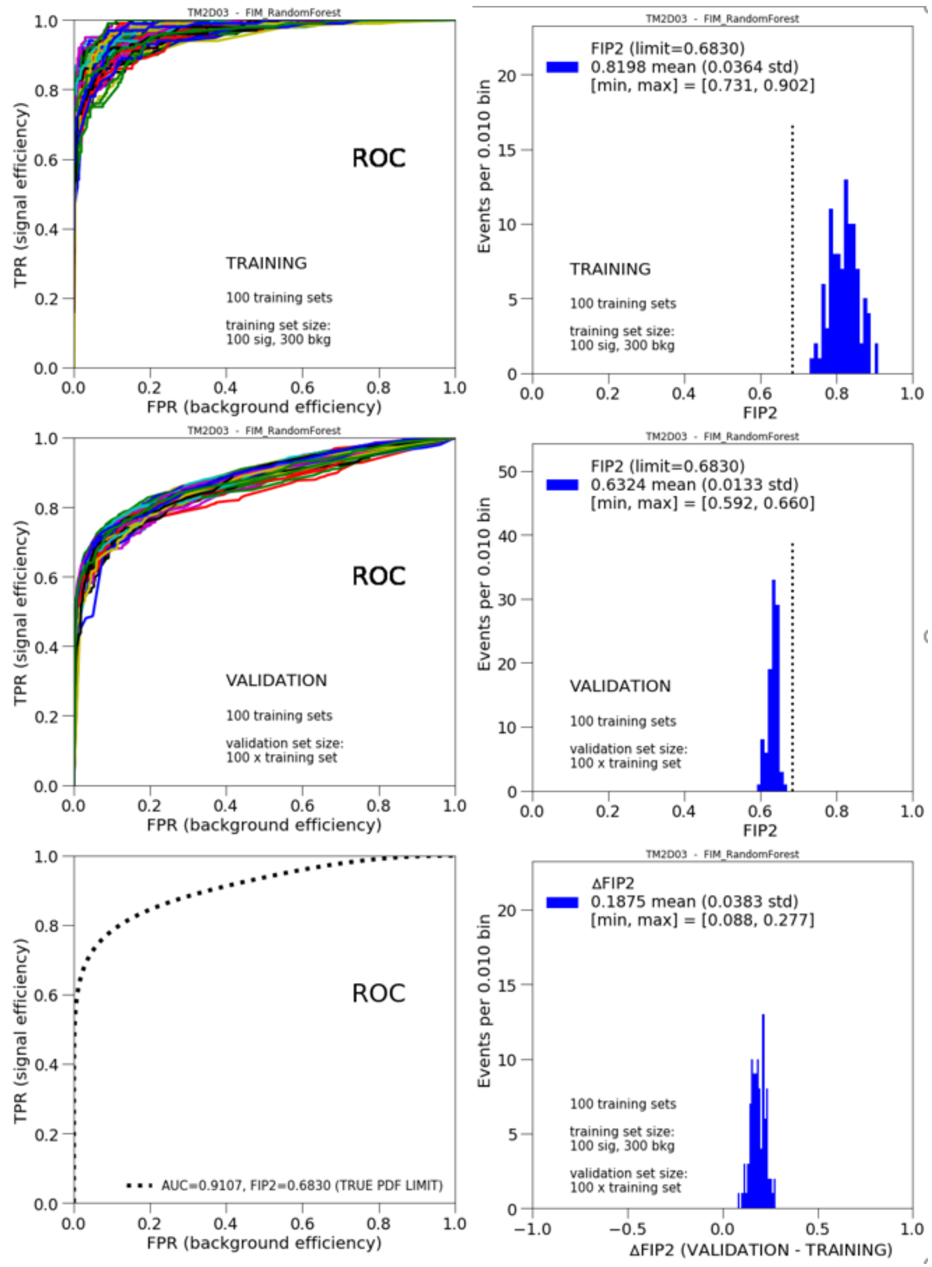


*Convert ROC to convex hull
- ensure decreasing slope
- avoid staircase effect that would artificially inflate FIP2
  (bins of 100% purity: only signal or only background)

HEP estimation of parameter θ in a binned distribution fit

FIP2$^{(max)}$ example

(and overtraining)



FIP2 is a metric in [0,1]
but the detector resolution
effectively determines a FIP2$^{(max)}$ < 1

FIP1 and FIP2 revisited

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

$$\mathrm{FIP}_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_\mathrm{tot}} \gamma_i^2} = \mathrm{FIP_{eff}} \times \mathrm{FIP_{sha}} \times \mathrm{FIP_{pur}}$$
$$= \boxed{\frac{\sum_{i=1}^{S_\mathrm{sel}} \gamma_i^2}{\sum_{i=1}^{S_\mathrm{tot}} \gamma_i^2}} \times \frac{\sum_{k=1}^{K} s_k \phi_k^2}{\sum_{i=1}^{S_\mathrm{sel}} \gamma_i^2} \times \boxed{\frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{k=1}^{K} s_k \phi_k^2}}$$

$$\mathrm{FIP}_1 = \epsilon_s \varrho$$

FIP1:
$\mathrm{FIP_{eff}}$ =ε
$\mathrm{FIP_{pur}}$=ρ

$$\mathrm{FIP}_2 = \frac{\mathcal{I}_{\sigma_s}}{\mathcal{I}_{\sigma_s}^{(\mathrm{ideal})}} = \frac{\sum_k s_k \rho_k}{\sum_k s_k} = \frac{\sum_k s_k^2/n_k}{\sum_k s_k} = \frac{\sum_k n_k \rho_k^2}{\sum_k s_k}$$

FIP2:
$\mathrm{FIP_{eff}}$ =1
$\mathrm{FIP_{pur}}$=FIP2

$$\mathrm{FIP}_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_\mathrm{tot}} \gamma_i^2} = \mathrm{FIP_{eff}} \times \mathrm{FIP_{sha}} \times \mathrm{FIP_{pur}}$$
$$= \frac{\sum_{i=1}^{S_\mathrm{sel}} \gamma_i^2}{\sum_{i=1}^{S_\mathrm{tot}} \gamma_i^2} \times \frac{\sum_{k=1}^{K} s_k \phi_k^2}{\sum_{i=1}^{S_\mathrm{sel}} \gamma_i^2} \times \boxed{\frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{k=1}^{K} s_k \phi_k^2}}$$

Non-dichotomous truth: examples

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

N. A. Obuchowski, *An ROC-Type Measure of Diagnostic Accuracy When the Gold Standard is Continuous-Scale*, Statistics in Medicine 25 (2006) 481. doi:10.1002/sim.2228
M. J. Pencina, R. B. D'Agostino, *Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation*, Statistics in Medicine 23 (2004) 2109. doi:10.1002/sim.1802
J. Lambert et al., *How to Measure the Diagnostic Accuracy of Noninvasive Liver Fibrosis Indices: The Area Under the ROC Curve Revisited*, Clinical Chemistry 54 (2008) 1372. doi:10.1373/clinchem.2007.097923

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

Response: partitioning + ranking

K. Järvelin, J. Kekäläinen, *IR evaluation methods for retrieving highly relevant documents*, Proc. 23rd ACM SIGIR Conf. (SIGIR 2000), Athens (2000). doi:10.1145/345508.345545
J. Kekäläinen, K. Järvelin, *Using graded relevance assessments in IR evaluation*, J. Am. Soc. Inf. Sci. 53 (2002) 1120. doi:10.1002/asi.10137
K. Järvelin, J. Kekäläinen, *Cumulated gain-based evaluation of IR techniques*, J. ACM Trans. on Inf. Sys. (TOIS) 20 (2002) 422. doi:10.1145/582415.582418

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

	fraudulent	legitimate
refuse	\$20	$-$\$20
approve	$-x$	$0.02x$

Response: yes/no decision

B. Zadrozny, C. Elkan, *Learning and making decisions when costs and probabilities are both unknown*, Proc. 7th Int. Conf. on Knowledge Discovery and Data Mining (KDD-01), San Francisco (2001). doi:10.1145/502512.502540
C. Elkan, *The Foundations of Cost-Sensitive Learning*, Proc. 17th Int. Joint Conf. on Artificial Intelligence (IJCAI-01), Seattle (2001).

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

G. W. Brier, *Verification of forecasts expressed in terms of probability*, Weather Rev. 78 (1950) 1. doi:10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2
F. Sanders, *On Subjective Probability Forecasting*, J. Applied Meteorology 2 (1963) 191. https://www.jstor.org/stable/26169573

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

HEP-like: probabilistic!

D. J. Spiegelhalter, *Probabilistic prediction in patient management and clinical trials*, Statist. Med. 5 (1986) 421. doi:10.1002/sim.4780050506
F. E. Harrell, K. L. Lee, D. B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*, Statist. Med. 15 (1996) 361. 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

- **HEP measurement of θ** $\rightarrow$ *evt-by-evt sensitivity to θ*

Signal and background are not dichotomous classes

(with one exception: cross section measurements)

Background events by definition are insensitive to θ
Signal events may have positive, zero or negative sensitivity

θ: mass, coupling
NON-DICHOTOMOUS

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

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

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

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

Changing the signal cross section ~is a global rescaling of all differential distributions

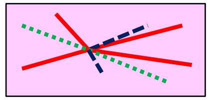
$$s_k(\sigma_s) = \frac{\sigma_s}{\sigma_{s,\text{ref}}} \times s_k(\sigma_{s,\text{ref}})$$

In a cross section measurement
All background events are equivalent to one another
All signal events are equivalent to one another

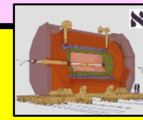
$$\boxed{\gamma_i = \frac{1}{\sigma_s}\delta_i = \begin{cases} \frac{1}{\sigma_s} & \text{if}\, i \in \{\text{Signal}\}\,, \\ 0 & \text{if}\, i \in \{\text{Background}\}\,, \end{cases} \qquad \text{if}\, \theta \equiv \sigma_s}$$

θ: cross section σ_s
DICHOTOMOUS

From CRLB to Fisher Information Part (FIP)



Particles produced in beam collisions



Detector
Trigger

Raw data events

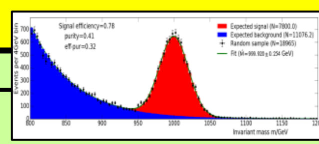
Data processing

Analysis object data

PHYSICS ANALYSIS
Event selection (Sig vs Bkg)
Event partitioning

Event counts in individual bins of a distribution

Max likelihood fit



Measured value of the parameter M ± ΔM

$$\mathcal{I}_\theta^{(\mathrm{ideal})}$$

$$\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \frac{(\Delta\theta^{(\mathrm{ideal})})^2}{(\Delta\theta)^2} \leq 100\%$$

FIP

$$\mathcal{I}_\theta$$

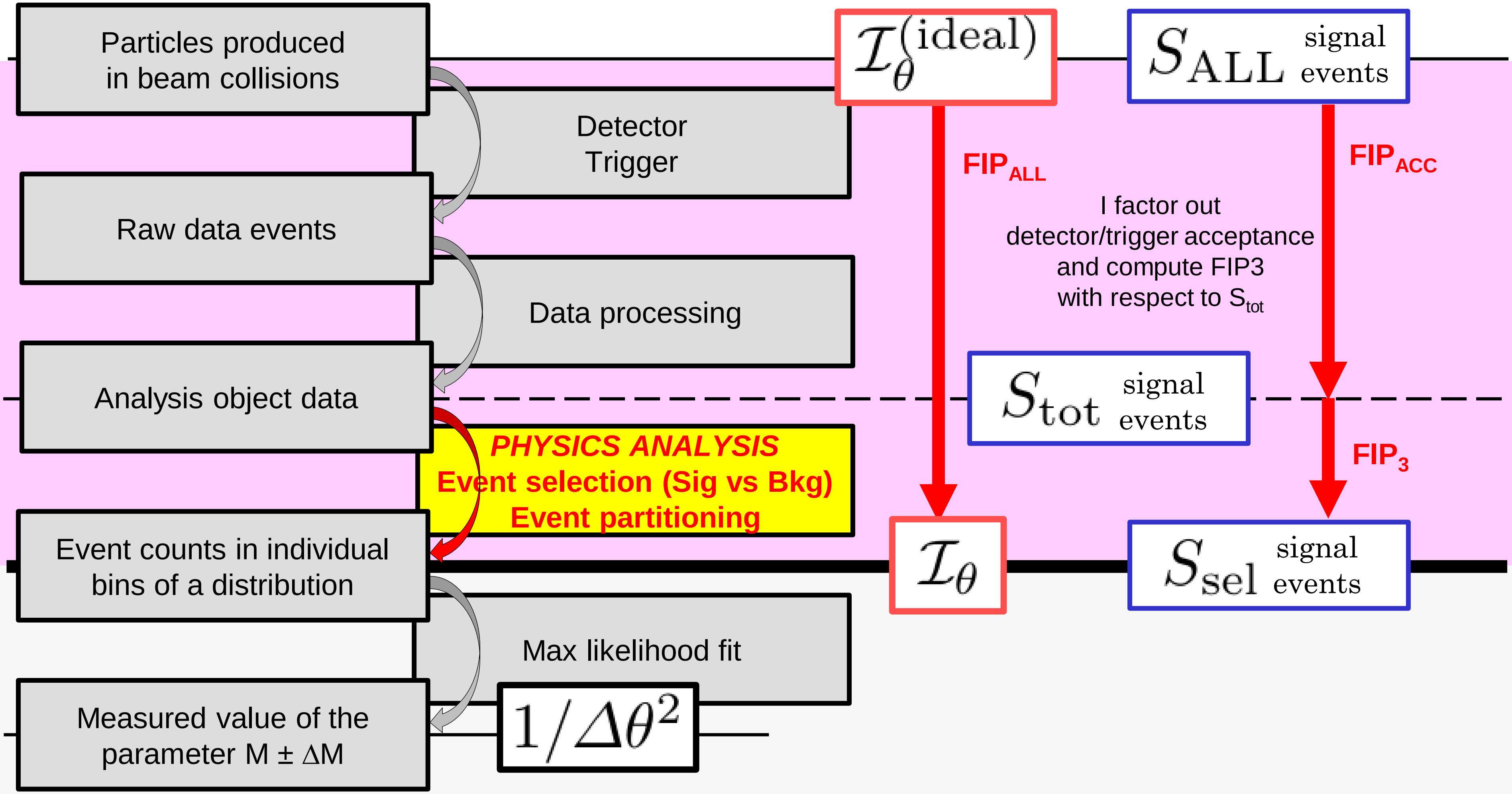
$(1/\Delta\theta^2)/\mathcal{I}_\theta$ is 100%

$$1/\Delta\theta^2$$

A max likelihood fit is 100% efficient: it achieves the CRLB, *for the given event selection and event partitioning*

CRLB

Two optimization handles: event selection and partitioning



Fisher information (about a parameter θ)

NB: Shannon information is a very different metric!

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

F. James, *Statistical Methods in Experimental Physics*, 2nd edition, World Scientific (2006).

Cramer-Rao lower bound CRLB
(lowest achievable variance $\Delta\theta^2$)

$$(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$$

- 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 Δθ in terms of information I_θ***

i.e. I will treat errors Δθ and information I_θ as equivalent concepts

$$\boxed{\mathcal{I}_\theta = \frac{1}{(\Delta\theta)^2}}$$

Fisher information I_θ about θ (statistical errors)

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

Bin-by-bin sensitivity to θ

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

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)

Backup slides – CHEP2019 slides

(CHEP 2019 slides: https://zenodo.org/record/3715951)

This is a follow-up of my CHEP2018 talk about *binned fits of a parameter θ*

Evaluation and training metrics:
Fisher Information Part

Previous CHEP2018 talk

Event selection
Binary classification

Bin-by-bin sensitivity to θ

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 θ

MINIMUM ERROR WITH AN IDEAL DETECTOR

Mass fits, Coupling fits (FIP3)

Meteorology (MSE, Brier),
Medical Prognostics

Talk: https://doi.org/10.5281/zenodo.1303387
Paper: https://doi.org/10.1051/epjconf/201921406004

Compare to and learn from other domains

Outline

- 1 - HEP parameter fits and Weight Derivative Regression

- 2 - Learning from others

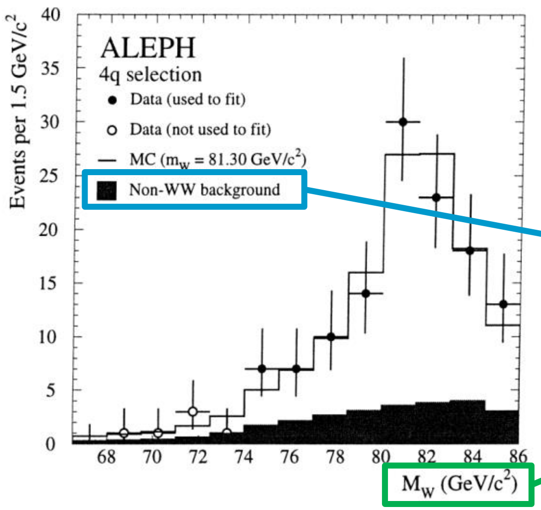
- Conclusions

This talk only provides some maths and some literature review

No toy model or concrete applications are presented

1 – Binned fit of a parameter θ

ALEPH Collaboration, *Measurement of the W mass by direct reconstruction in e^+e^- collisions at 172 GeV*, Phys. Lett. B 422 (1998) 384. doi:10.1016/S0370-2693(98)00062-8



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: event selection

This CHEP2019 talk: event partitioning
(selection is a special case of partitioning)

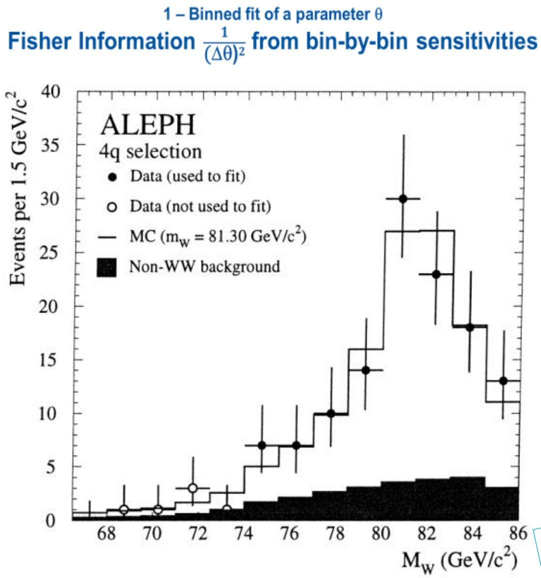
$$m_W = 81.30 \boxed{\pm 0.47(\mathrm{stat.})} \pm 0.11(\mathrm{syst.})\ \mathrm{GeV}/c^2$$

*I only discuss the **statistical error $\Delta\theta$** in this talk*

(I ignore systematic errors, even if at LHC they are the limitation)

1 – Binned fit of a parameter θ

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 θ

Recap CHEP2018 talk

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

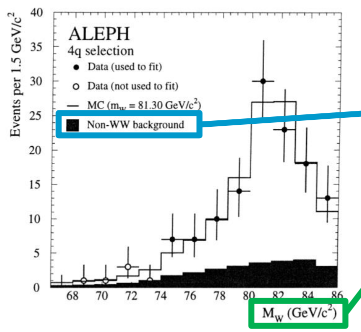
$$m_W = 81.30 \boxed{\pm 0.47(\text{stat.})} \pm 0.11(\text{syst.})\ \text{GeV}/c^2$$

Minimizing $\Delta\theta$ is equivalent to maximizing I_θ

1 – Binned fit of a parameter θ

Fisher Information Part (FIP)

ALEPH Collaboration, *Measurement of the W mass by direct reconstruction in e^+e^- collisions at 172 GeV*, Phys. Lett. B 422 (1998) 384. doi:10.1016/S0370-2693(98)00062-8



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_θ to $I_\theta^{(ideal)}$ 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_θ to $I_\theta^{(ideal)}$ for the **ideal partitioning** (and the ideal selection)*

Recap CHEP2018 talk

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

$$\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\boxed{\mathcal{I}_\theta^{(\mathrm{ideal})}}} = \frac{(\Delta\theta^{(\mathrm{ideal})})^2}{(\Delta\theta)^2} \le 100\%$$

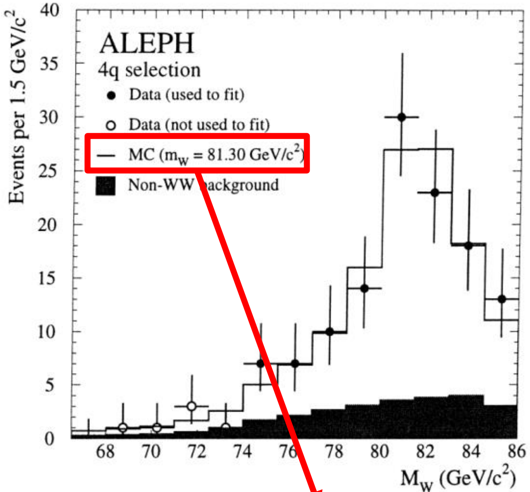
FIP is a metric between 0 and 1 – higher is better

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

1 – Binned fit of a parameter θ

Event-by-event Monte Carlo reweighting

ALEPH Collaboration, *Measurement of the W mass by direct reconstruction in e^+e^- collisions at 172 GeV*, Phys. Lett. B 422 (1998) 384. doi:10.1016/S0370-2693(98)00062-8



$$w_i(m_W, \Gamma_W) = \frac{|\mathcal{M}(m_W, \Gamma_W, p_i^1, p_i^2, p_i^3, p_i^4)|^2}{|\mathcal{M}(m_W^{MC}, \Gamma_W^{MC}, p_i^1, p_i^2, p_i^3, p_i^4)|^2}$$

Fit for θ → Compare data in bin k to *model prediction n_k as a function of θ*

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

1. *Generate signal sample at θ_{ref}, with $w_i(\theta_{ref})=1$*
   (By definition, background does not depend on θ)
2. *Full detector simulation*
   (MC truth event properties $\mathbf{x}_i^{(true)}$ → observed event properties $\mathbf{x}_i$)
3. ***Reweight each event by matrix element ratio***

$$w_i(\theta) = \frac{\mathrm{Prob}_{(\theta)}(\mathbf{x}_i^{(\mathrm{true})})}{\mathrm{Prob}_{(\theta_{\mathrm{ref}})}(\mathbf{x}_i^{(\mathrm{true})})} = \frac{|\mathcal{M}(\theta, \mathbf{x}_i^{(\mathrm{true})})|^2}{|\mathcal{M}(\theta_{\mathrm{ref}}, \mathbf{x}_i^{(\mathrm{true})})|^2}$$

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

J. S. Gainer, J. Lykken, K. T. Matchev, S. Mrenna, M. Park, *Exploring theory space with Monte Carlo reweighting*, JHEP 2014 (2014) 78. doi:10.1007/JHEP10(2014)078

O. Mattelaer, *On the maximal use of Monte Carlo samples: re-weighting events at NLO accuracy*, Eur. Phys. J. C 76 (2016) 674. doi:10.1140/epjc/s10052-016-4533-7

Event-by-event sensitivities γ_i: MC weight derivatives

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

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

Define the **event-by-event sensitivity γ_i to θ** as the *derivative with respect to θ of the MC weight w_i*

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

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

Aside: ∂w/∂θ is closely related to the *Fisher score* (but the latter is defined as the derivative of a probability normalized to 1)

The bin-by-bin sensitivity to θ in bin k is the *average in bin k of the event-by-event sensitivity γ_i to θ*

$$\left(\frac{1}{n_k}\frac{\partial n_k}{\partial\theta}\right)_{\theta=\theta_{\mathrm{ref}}}=\frac{1}{n_k}\sum_{i\in k}\gamma_i=\boxed{\langle\gamma\rangle_k=\frac{1}{n_k}\frac{\partial n_k}{\partial\theta}}$$

Beyond the signal-background dichotomy

Background events have γ_i=0
because by definition they are insensitive to θ

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

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

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 (|γ_i|~0)***

Bin-by-bin sensitivity ϕ_k of signal events alone:

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

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

$$\delta_i = \begin{cases} 1 & \text{if } i \in \{\text{Signal}\} \\ 0 & \text{if } i \in \{\text{Background}\} \end{cases} \qquad \rho_k = \frac{s_k}{s_k+b_k} = \frac{s_k}{n_k} = \frac{\sum_{i\in k}\delta_i}{n_k} = \langle\delta\rangle_k$$

Bin-by-bin sensitivity $\langle\gamma\rangle_k$ of signal + background:

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

Information from all bins for signal + background:

$$\mathcal{I}_\theta = \sum_{k=1}^{K} n_k\langle\gamma\rangle_k^2 = \sum_{k=1}^{K} n_k(\rho_k\phi_k)^2 = \sum_{k=1}^{K} s_k\boxed{\rho_k}\phi_k^2$$

Effect of background: it dilutes by a factor $\rho_k \leq 1$ the bin-by-bin sensitivity and information for signal events alone

Ideal case: partition by the evt-by-evt sensitivity γ_i

Information I_θ in terms of average bin-by-bin sensitivities:

$$\mathcal{I}_\theta = \sum_{k=1}^{K} n_k \left(\frac{1}{n_k} \frac{\partial n_k}{\partial \theta} \right)^2 = \sum_{k=1}^{K} n_k \langle \gamma \rangle_k^2$$

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 γ_{i_1} and γ_{i_2} are different*

$$\Delta\mathcal{I}_\theta = \gamma_{i_1}^2 + \gamma_{i_2}^2 - 2\left(\frac{\gamma_{i_1}+\gamma_{i_2}}{2}\right)^2 = \frac{1}{2}(\gamma_{i_1}-\gamma_{i_2})^2$$

Goal of a distribution fit: partition events by their different MC-truth event-by-event sensitivities γ_i to θ

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

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

Knowing one's limits: maximum achievable information with an ideal detector

- Ideal acceptance, select all signal events $S_{sel}=S_{tot}$
- Ideal resolution, measured γ_i is that from MC truth

(implies ideal rejection of background events, $\gamma_i=0$)

$$\mathcal{I}_\theta^{(\mathrm{ideal})} = \sum_{i=1}^{N_{\mathrm{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2$$

Weight Derivative Regression (WDR): train q_i for γ_i

Goal of a distribution fit: separate events with different MC-truth event-by-event sensitivities γ_i to θ

But γ_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

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

Training metric: maximize FIP
Evaluation metric: maximize FIP

(or equivalently minimize MSE? see final slides)

Some of many caveats:
- *Dependency of weight derivative on reference θ_{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?*

1 – Binned fit of a parameter θ

WDR and Optimal Observables

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 γ_i to θ
(OO does not use MC weight derivatives but it is similar)

D. Atwood, A. Soni, *Analysis for magnetic moment and electric dipole moment form factors of the top quark via $e^+e^- \to t\bar{t}$*, Phys. Rev. D 45 (1992) 2405. doi:10.1103/PhysRevD.45.2405,
M. Davier, L. Duflot, F. LeDiberder, A. Rougé, *The optimal method for the measurement of tau polarization*, Phys. Lett. B 306 (1993) 411. doi:10.1016/0370-2693(93)90101-M
M. Diehl, O. Nachtmann, *Optimal observables for the measurement of three-gauge-boson couplings in $e^+e^- \to W^+W^-$*, Z. Phys. C 62 (1994) 397. doi:10.1007/BF01555899
O. Nachtmann, F. Nagel, *Optimal observables and phase-space ambiguities*, Eur. Phys. J. C40 (2005) 497. doi:10.1140/epjc/s2005-02153-9

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

Some similarities also with the MadMiner approach
See CHEP 2019 contribution #506
"Constraining effective field theories with ML"

MC-truth
sensitivity $\gamma_i^{(\text{MC truth})}$

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

MC observable
event properties $x_i^{(\text{MC})}$

MC-truth functional dependence

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

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

Fit optimal observable

$\mathbf{OO_i^{(DATA)} = f(\, x_i^{(DATA)}\,)}$
$\mathbf{OO_i^{(MC)} = f(\, x_i^{(MC)}\,)}$

The difference is in where the ***effect of detector resolution*** *is taken into account*

Weight Derivative Regression

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

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

Fit WDR regressor

$\mathbf{q_i^{(DATA)} = q(\, x_i^{(DATA)}\,)}$
$\mathbf{q_i^{(MC)} = q(\, x_i^{(MC)}\,)}$

FIP decomposition: efficiency, sharpness, purity

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

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

$$\mathrm{FIP}_3 = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \frac{\sum_{k=1}^{K} n_k \langle\gamma\rangle_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2} = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2}$$

$$\mathrm{FIP}_3 = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2} = \mathrm{FIP}_{\mathrm{eff}} \times \mathrm{FIP}_{\mathrm{sha}} \times \mathrm{FIP}_{\mathrm{pur}}$$

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

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

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

Sensitivity-weighted signal **purity** or equivalently **sharpness** in separating signal events from background events: dilution of signal sensitivity caused by bin-by-bin purity ρ_k

"sharpness" as in meteorology: see later why

Limits to knowledge: FIP for a realistic detector

Limited detector acceptance
(detector geometry, trigger rate):
factor this out in $\mathrm{FIP}_{\mathrm{ACC}} \le 1$

$S_{\mathrm{ALL}}, \gamma_i, \delta_i$

$$\boxed{\mathcal{I}_\theta^{(\mathrm{ideal},\, S_{\mathrm{ALL}})} = \sum_{i=1}^{S_{\mathrm{ALL}}} \gamma_i^2}$$

$$\boxed{\mathrm{FIP}_{\mathrm{ACC}} \le 1}$$

$$\mathrm{FIP}_{\mathrm{ALL}} = \mathrm{FIP}_{\mathrm{ACC}} \times \mathrm{FIP}_3$$

$$\mathrm{FIP}_3 = \mathrm{FIP}_{\mathrm{eff}} \times \mathrm{FIP}_{\mathrm{sha}} \times \mathrm{FIP}_{\mathrm{pur}}$$

$$\boxed{0 \le \mathrm{FIP}_3 \le \mathrm{FIP}_3^{(\max)} \le 1}$$

$S_{\mathrm{tot}}, \gamma_i, \delta_i$

$$\boxed{\mathcal{I}_\theta^{(\mathrm{ideal})} = \sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2}$$

$\xrightarrow{\mathrm{FIP}_{\mathrm{eff}}}$

$S_{\mathrm{sel}}, \gamma_i, \delta_i$

$$\boxed{(\mathcal{I}_\theta = \sum_{i=1}^{S_{\mathrm{sel}}} \gamma_i^2)}$$

$$\boxed{\mathrm{FIP}_3^{(\max)} = \frac{\mathcal{I}_\theta^{(\max)}}{\mathcal{I}_\theta^{(\mathrm{ideal})}}}$$

$$\boxed{\mathrm{FIP}_3 = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}}}$$

$\downarrow \mathrm{FIP}_{\mathrm{sha}}$

$S_{\mathrm{tot}}, \phi(\mathbf{x}), \delta_i$

$$\boxed{(\mathcal{I}_\theta = \int s(\mathbf{x})\phi(\mathbf{x})^2 d\mathbf{x})}$$

$S_{\mathrm{sel}}, \phi_k, \delta_i$

$$\boxed{(\mathcal{I}_\theta = \sum_{k=1}^{K} s_k \phi_k^2)}$$

$\downarrow \mathrm{FIP}_{\mathrm{pur}}$

$S_{\mathrm{tot}}, \phi(\mathbf{x}), \rho(\mathbf{x})$

$$\boxed{\mathcal{I}_\theta^{(\max)} = \int s(\mathbf{x})\phi(\mathbf{x})^2 \rho(\mathbf{x}) d\mathbf{x}}$$

$S_{\mathrm{sel}}, \phi_k, \rho_k$

$$\boxed{\mathcal{I}_\theta = \sum_{k=1}^{K} s_k \phi_k^2 \rho_k}$$

Limited detector resolution
In the multi-dimensional space
of event observables **x**,
it is impossible to resolve:

- signal events
with high sensitivity γ_i
from signal events
with low sensitivity γ_i:
average sensitivity is $\phi(\mathbf{x})$

- signal events δ_i=1
from background events δ_i=0:
average purity is $\rho(\mathbf{x})$

FIP is a metric in [0,1], but
the detector acceptance and resolution
limit it to $0 \le \mathrm{FIP} \le \mathrm{FIP}^{(\max)} < 1$

⇨ *$\mathrm{FIP} > \mathrm{FIP}^{(\max)}$ while training q_i*
*implies **overtraining**…*

2 – Learning from others

Reading Room, British Museum

Reading is a revolutionary act
(Inge Feltrinelli, 1930-2018)

Different problems in different domains require different metrics and tools…

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

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002

A. Luque, A Carrasco, A. Martin, J. R. Lama, *Exploring Symmetry of Binary Classification Performance Metrics*, Symmetry 11 (2019) 47. doi:10.3390/sym11010047.

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)

C. Ferri, J. Hernández-Orallo, R. Modroiu, *An Experimental Comparison of Classification Performance Metrics*, Proc. Learning 2004, Elche (2004). http://dmip.webs.upv.es/papers/Learning2004.pdf

C. Ferri, J. Hernández-Orallo, R. Modroiu, *An Experimental Comparison of Performance Measures for Classification*, Pattern Recognition Letters 30 (2009) 27. doi:10.1016/j.patrec.2008.08.010

2 – Learning from others: Meteorology

MSE decomposition: Validity and Sharpness

MSE (mean squared error) of regressor prediction q_i versus the true γ_i for event *i*:

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

MSE is a probabilistic metric for both evaluation and training

MSE decomposition

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

Paraphrases the "Brier score" decomposition in Meteorology

G. W. Brier, *Verification of forecasts expressed in terms of probability*, Weather Rev. 78 (1950) 1. doi:10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

F. Sanders, *On Subjective Probability Forecasting*, J. Applied Meteorology 2 (1963) 191. https://www.jstor.org/stable/26169573

Validity, Reliability, Calibration

Sharpness, Resolution, Refinement

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

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: ***how well do we separate events with different true sensitivities γ_i?***

This is what determines the statistical error on the measurement of θ: related to FIP!

FIP is related to Sharpness (MSE)

(Validity, Reliability, Calibration) **MSE$_{sha}$** (Sharpness, Resolution, Refinement)

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

$$\frac{1}{N_{\mathrm{tot}}}\left[\mathcal{I}_\theta^{(\mathrm{ideal})} - \mathcal{I}_\theta\right]$$

$$\mathcal{I}_\theta^{(\mathrm{ideal})} = \sum_{i=1}^{N_{\mathrm{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2$$

$$\mathcal{I}_\theta = \sum_{k=1}^{K} n_k \left(\frac{1}{n_k}\frac{\partial n_k}{\partial\theta}\right)^2 = \sum_{k=1}^{K} n_k \langle\gamma\rangle_k^2$$

FIP is related to Sharpness:

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

$$\boxed{\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \left(1 - \frac{N_{\mathrm{tot}} \times \mathrm{MSE}_{\mathrm{sha}}}{\mathcal{I}_\theta^{(\mathrm{ideal})}}\right)}$$

Practical implication for Weight Derivative Regression:

MSE is the most appropriate loss function for training the WDR regressor

2 – Learning from others: HEP does not need ranking, or ranking metrics

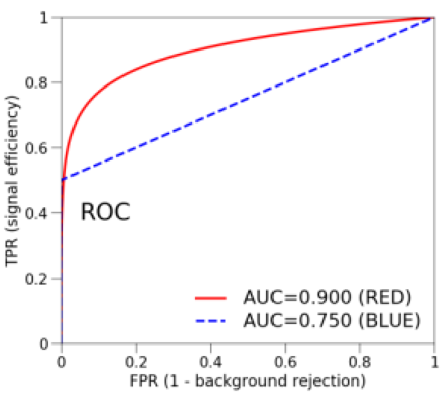
HEP needs partitioning, and probabilistic metrics

Ranking, and ranking metrics

Pick two events at random and rank them

Medical Diagnostics → *ranking evaluation of diagnostic prediction*

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



D. M. Green, *General Prediction Relating Yes-No and Forced-Choice Results*, J. Acoustical Soc. Am. 36 (1964) 1042. doi:10.1121/1.2143339

D. J. Goodenough, K. Rossmann, L. B. Lusted, *Radiographic applications of signal detection theory*, Radiology 105 (1972) 199. doi:10.1148/105.1.199

J. A. Hanley, B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology 143 (1982) 29. doi:10.1148/radiology.143.1.7063747

A. P. Bradley, *The use of the area under the ROC curve in the evaluation of Machine Learning algorithms*, Pattern Recognition 30 (1997) 1145. doi:10.1016/S0031-3203(96)00142-2

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

IRRELEVANT FOR HEP PARAMETER FITS?

Partitioning, and probabilistic metrics

Group events and make a forecast on each subset

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 θ*

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

$$\mathrm{MSE} = \overbrace{\boxed{\frac{1}{N_{\mathrm{tot}}}\left[\sum_{k=1}^{K} n_k \left(q_{(k)} - \langle\gamma\rangle_k\right)^2\right]}}^{\text{Validity, Reliability, Calibration}} + \overbrace{\boxed{\frac{1}{N_{\mathrm{tot}}}\left[\left(\sum_{i=1}^{N_{\mathrm{tot}}}\gamma_i^2\right) - \left(\sum_{k=1}^{K} n_k \langle\gamma\rangle_k^2\right)\right]}}^{\text{Sharpness, Resolution, Refinement}}$$

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 θ from (signal or background) events with low sensitivity?

ESSENTIAL FOR HEP PARAMETER FITS!

Conclusions – HEP measurement of a parameter θ

- **MC weight derivatives** (event-by-event sensitivities γ_i to θ) may be used :
  - To determine the **ideal partitioning strategy**: partition by γ_i
  - To derive the **minimum error on the measurement of θ** (ideal detector)

$$\boxed{\mathcal{I}_\theta^{(\mathrm{ideal})} = \sum_{i=1}^{N_{\mathrm{tot}}} \gamma_i^2 = \sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2}$$

  - To derive **training and validation metrics** to optimize the measurement

$$\boxed{\mathrm{FIP} \;=\frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \frac{\sum_{k=1}^{K} n_k \langle\gamma\rangle_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2} = \frac{\sum_{k=1}^{K} s_k \rho_k \phi_k^2}{\sum_{i=1}^{S_{\mathrm{tot}}} \gamma_i^2}}$$

Evaluation and training metrics: FIP

  - To train a **regressor q_i of γ_i (optimal observable)** for a 1-D fit of θ

- HEP parameter fits are closer to **Meteorology** than to Medical Diagnostics
  - They use **partitioning** and need **probabilistic metrics** (sharpness, MSE)

$$\boxed{\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(\mathrm{ideal})}} = \left(1 - \frac{N_{\mathrm{tot}} \times \mathrm{MSE}_{\mathrm{sha}}}{\mathcal{I}_\theta^{(\mathrm{ideal})}}\right)}$$

Compare to and learn from other domains

  - They do not use ranking and do not need ranking metrics (AUC)