SYSTEMATIC AWARE LEARNING

a case study in High Energy Physics

Victor Estrade, Cecile Germain, Isabelle Guyon, David Rousseau

July 10, 2018

LRI, University of Paris-Sud, University of Paris-Saclay

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

enable fast experimentation

Experimental results

deep study of the proposed technics

Conclusions and Perspectives

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

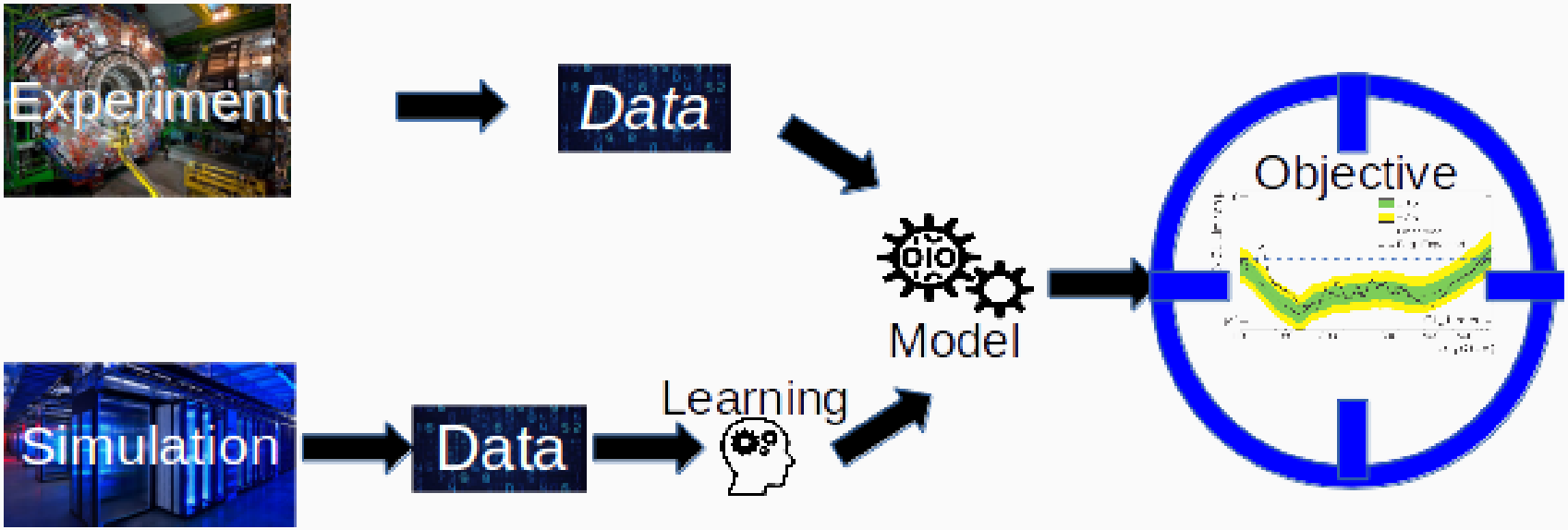
enable fast experimentation

Experimental results

deep study of the proposed technics

Conclusions and Perspectives

A common scenario in experimental science



- Simulations are needed to understand data
- Machine learning is central in the pipeline

Error sources

$$measure = value \pm \sigma_{stat} \pm \sigma_{syst}$$

Statistical error

- Empirical $\neq$ asymptotic (Lack of data)
- Noise

Systematic error

"known unknowns"

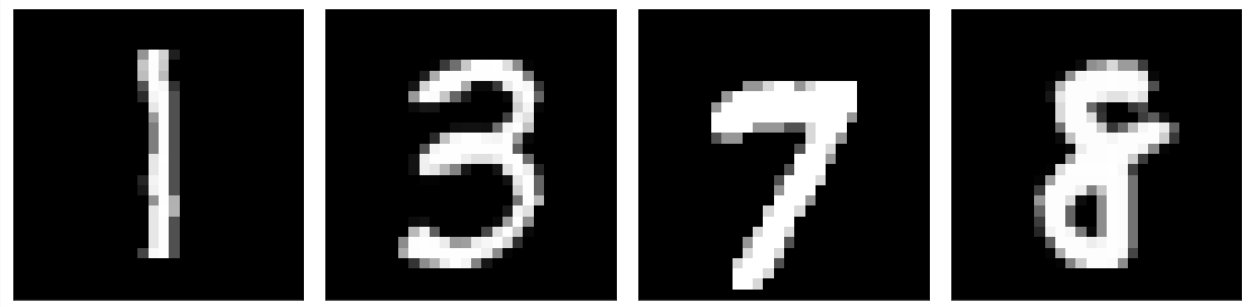
- Apparatus imperfection
- Simulation imperfection
- Lack of theoretical knowledge
- Bugs

Model selection

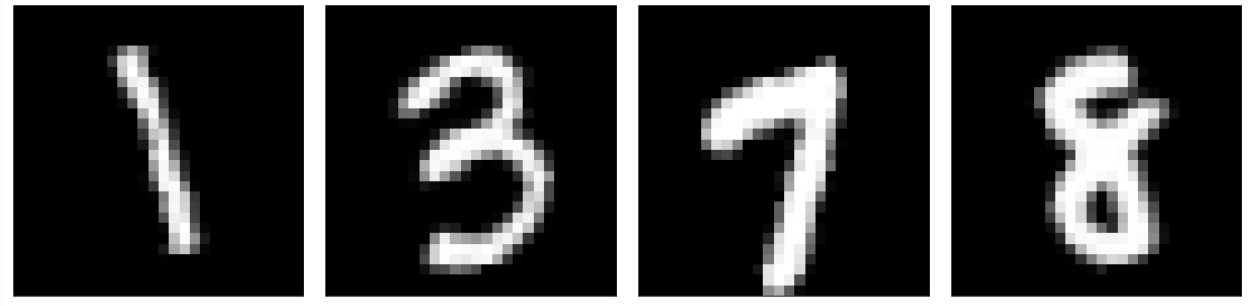
- Limited model capacity
- Biased model choice
- Limited computation resources

Example of systematic effect : camera rotation

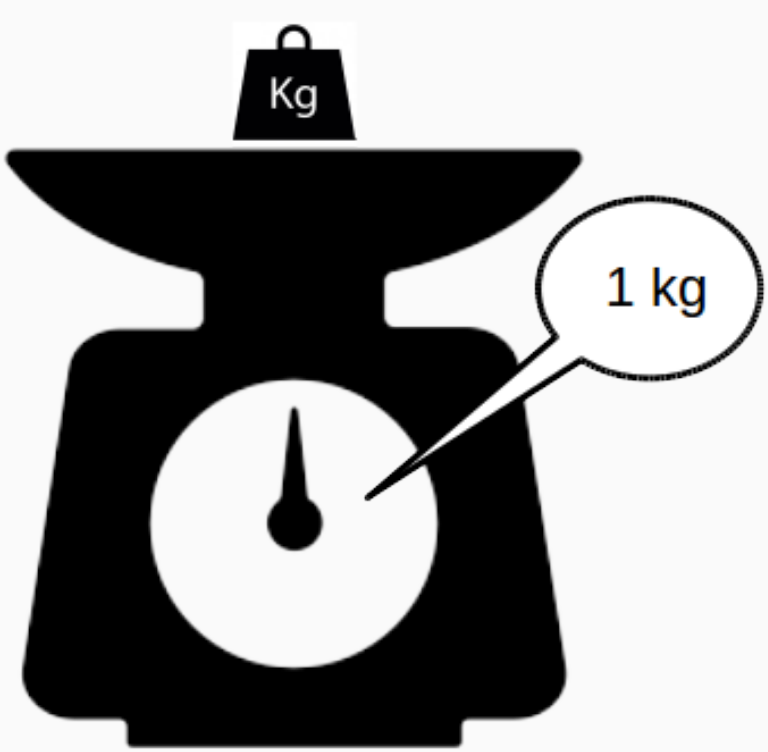
Original digits

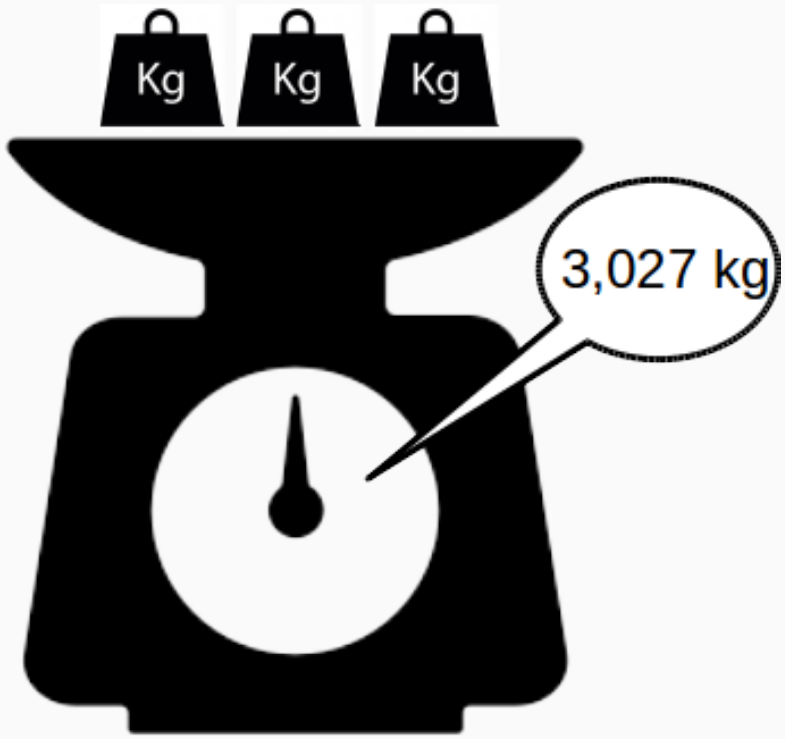


Apparatus imperfection slightly rotates the image



Example of systematic effect : Scale problem





The scaling issue : true value $= (1 + z) \times$ measured value

Caracterize systematics

Skewing function : $d(x, z)$

- Rotation of image input
- Rescaling

Nuisance parameter : z

- Angle of the rotation
- Scale factor

In real life there is several nuisance parameters with different impact on the data

Baby steps here : let's start with just one

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

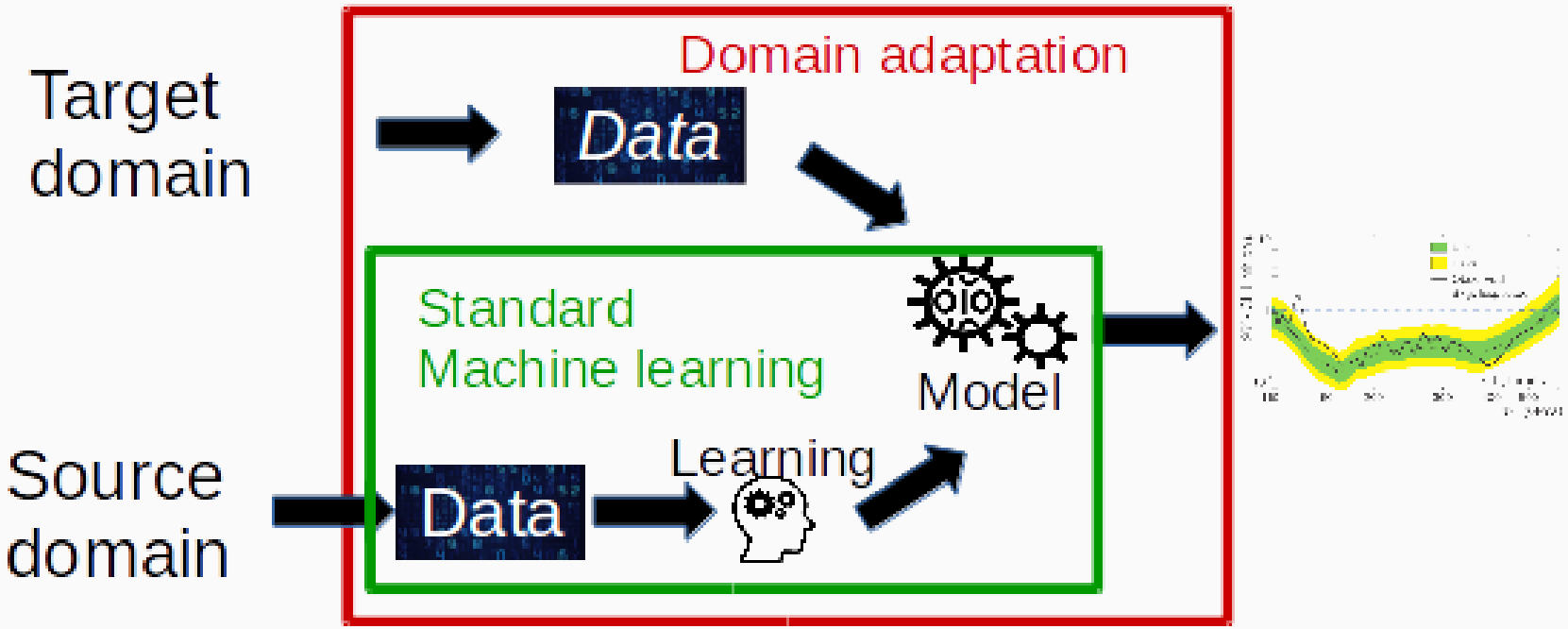
enable fast experimentation

Experimental results

deep study of the proposed technics

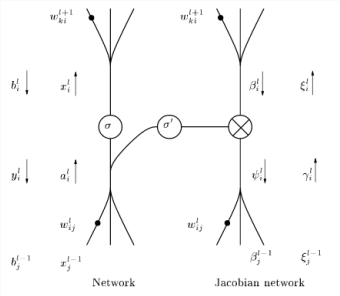
Conclusions and Perspectives

Domain Adaptation (transfer learning)



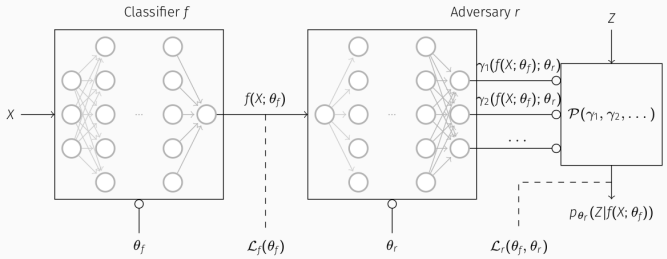
Domain adaptation helps machine learning to be accurate on similar data

Deep learning architectures (a few samples)



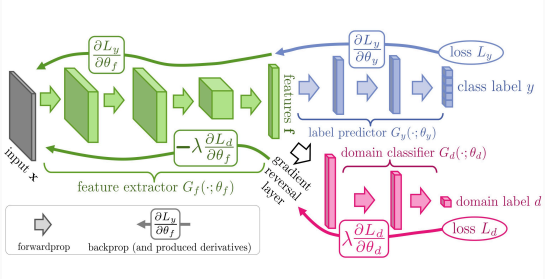
Tangent Propagation (TP)

[Simard et al., 1991]
[Rifai et al., 2011]



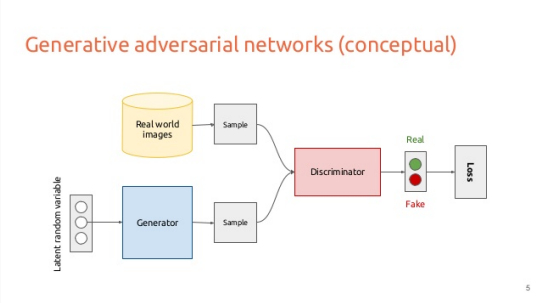
Pivot Adversarial Network (PAN)

[Louppe et al., 2016]



Domain Adversarial Network (DAN)

[Ganin et al., 2015]



Generative Adversarial Network (GAN)

[Goodfellow et al., 2014]

Tangent Propagation in a nutshell

$$d(\text{  }, z) = \text{  }$$

- Regularize the derivative of the model according to the transformation.

$$loss = E_{standard} + \lambda \sum_{x \in Data} \left| \frac{\partial f(d(x,z);\theta)}{\partial z} \right|^2_{z=0}$$

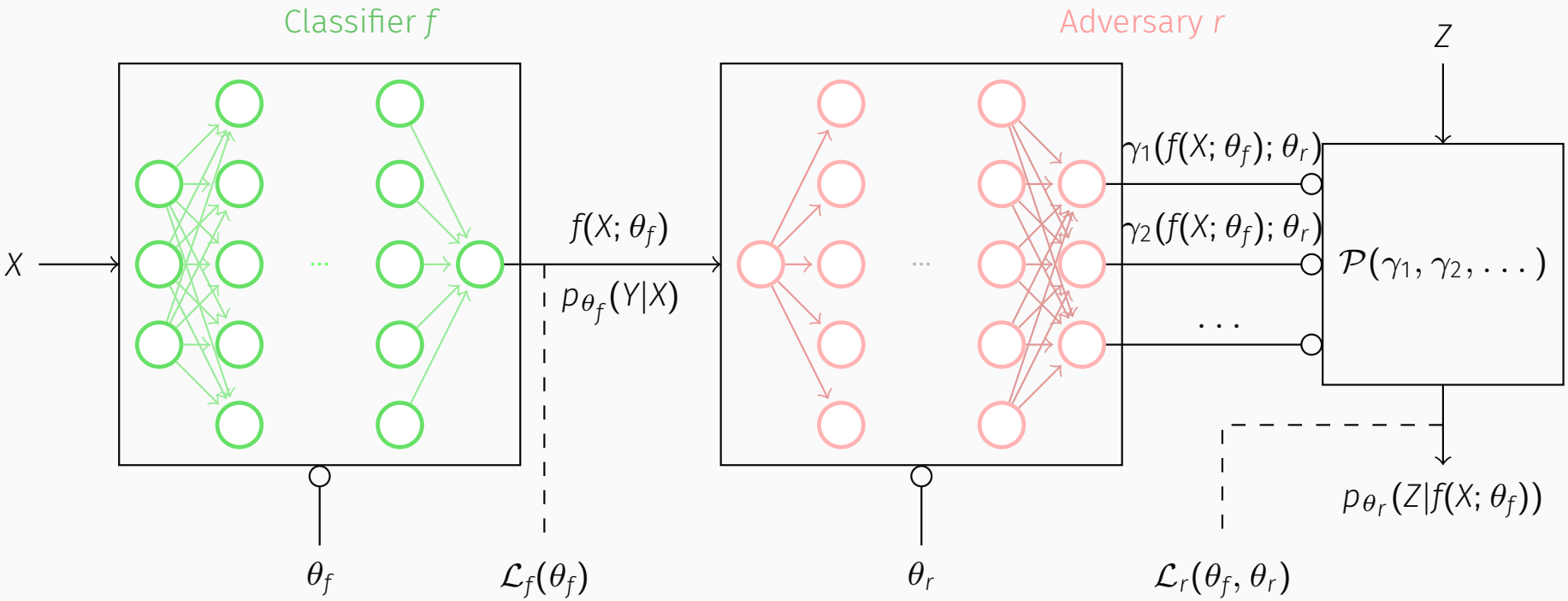
- Less data intensive than data augmentation
- Jacobian vector product trick : compute this derivative with a forward propagation through a "linearised" network.

$$\left. \frac{\partial f(d(x,z);\theta)}{\partial z} \right|_{z=0} = \nabla_x f(x;\theta).\ \left. \frac{\partial d(x,z)}{\partial z} \right|_{z=0}$$

[Simard et al., 1991] [Rifai et al., 2011]

Pivot Adversarial Neural Network

[Louppe et al., 2016]



- Learn the loss
- Makes it impossible to reconstruct Z from the output of the model
- Can include knowledge about nuisance parameter distribution

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

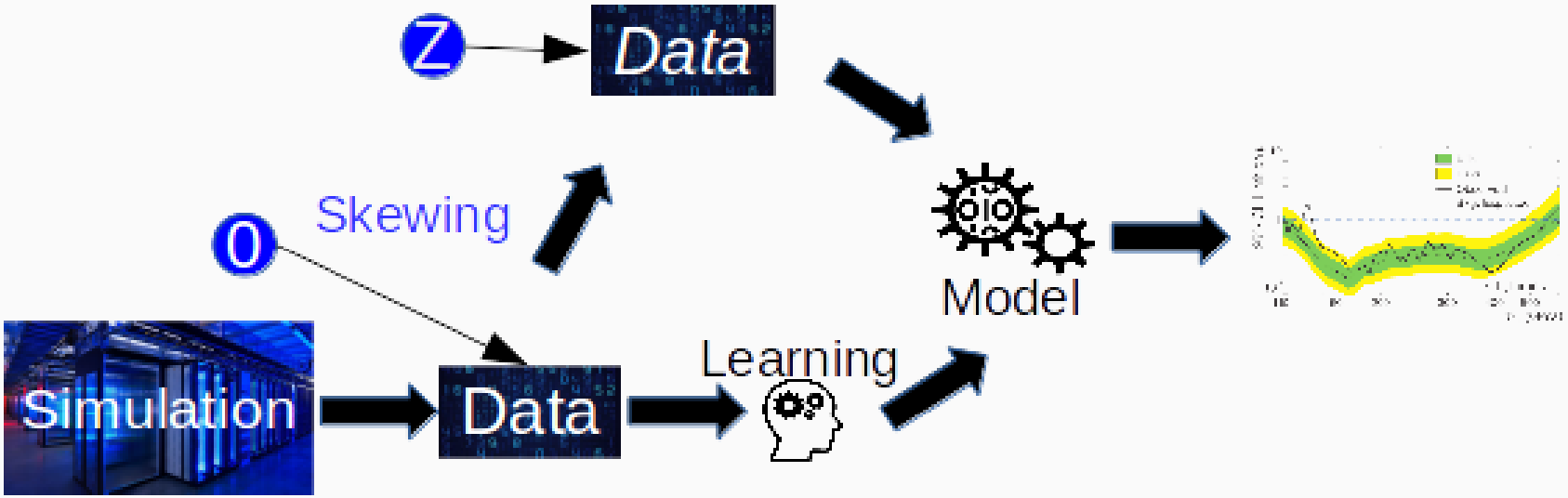
enable fast experimentation

Experimental results

deep study of the proposed technics

Conclusions and Perspectives

HEP benchmark : Simulation data



- Simulation of the $H \to \tau\tau$ decay
- Nuisance parameter : τ energy scale ($\pm[1\%, 10\%]$)
- Data from HiggsML challenge [Adam-Bourdarios et al., 2014][1]
- Data from [Baldi et al., 2014][2]

[1] Available `http://opendata.cern.ch/record/328`
[2] Available `https://archive.ics.uci.edu/ml/datasets/HIGGSs`

Estimate statistic and systematic error

Final objective : measuring a cross section

Number of events N_z follow a Poisson distribution

The nuisance parameter induce a multiplicative error

After addition of the logarithmic derivatives we get :

$$\frac{\sigma_\mu}{\mu} = \sqrt{\left(\frac{\sqrt{s_0 + b_0}}{s_0}\right)^2 + \left(\frac{(s_z - s_0) + (b_z - b_0)}{s_0}\right)^2}$$

- $s = \sum_{S, score_i > t} w_i$, selected signals (True positives)
- $b = \sum_{B, score_i > t} w_i$, selected backgrounds (False positives)
- $*_0$, on the nominal data
- $*_z$, on the skewed data

Learning objective is to minimize the relative error $\frac{\sigma_\mu}{\mu}$

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

enable fast experimentation

Experimental results

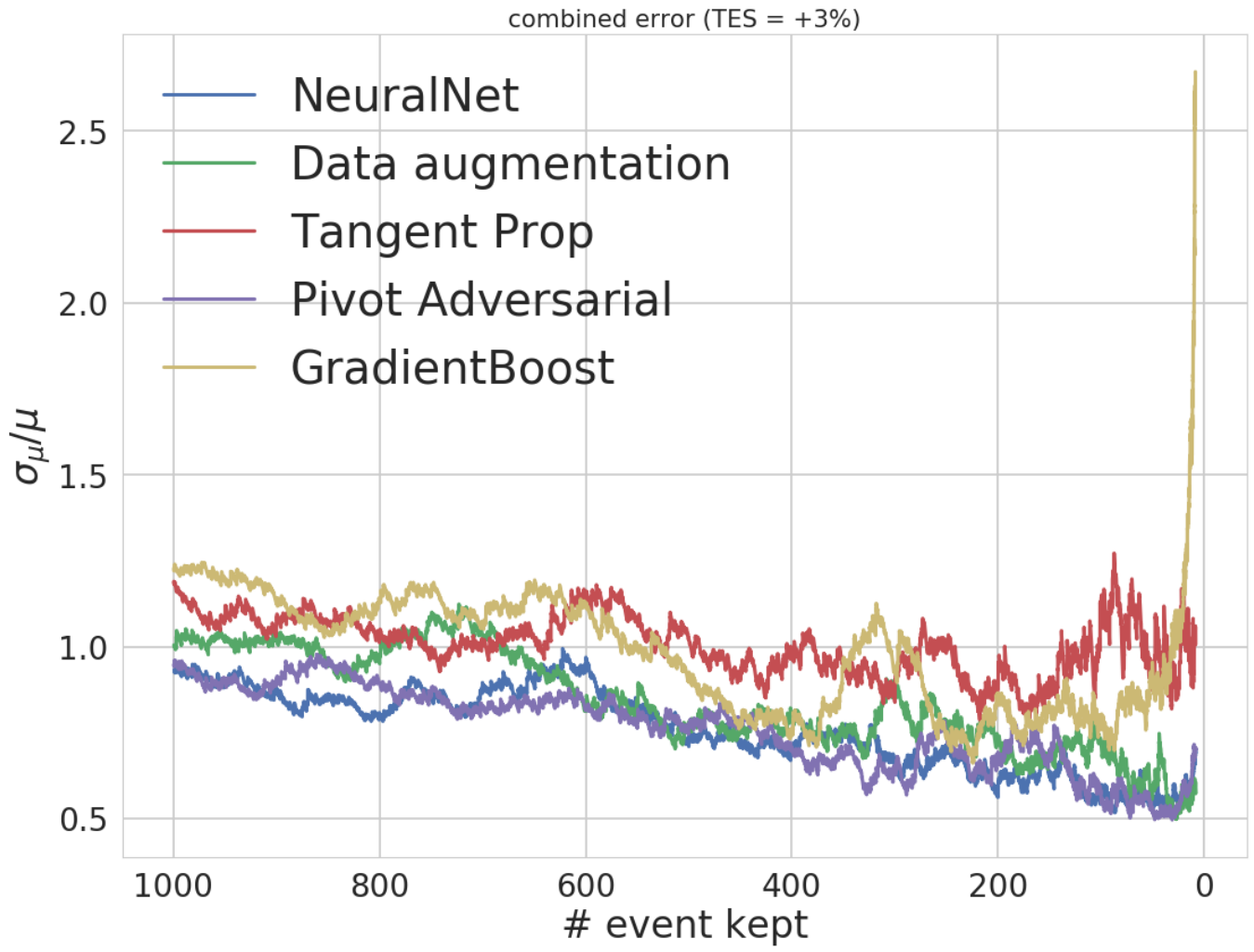
deep study of the proposed technics

Conclusions and Perspectives

Experimental settings

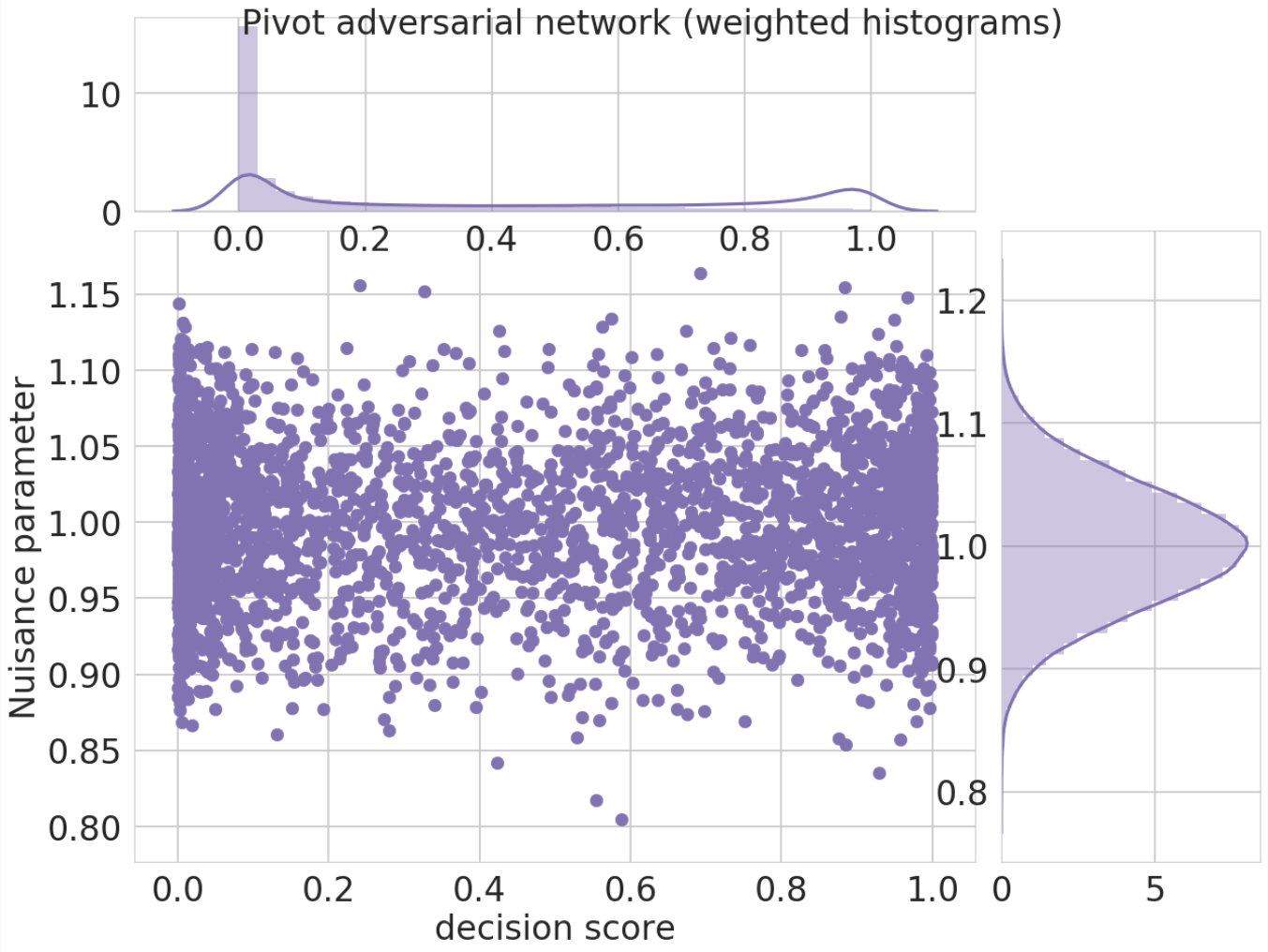
- Hyper parameters selected with grid search
- Models are compared with same structures (#neurons, non-linearities)
- Data augmentation and Pivot are fed with z drawn from a Gaussian distribution
- The others are trained with nominal data only
- Report estimated error along classification threshold

Results of the contest



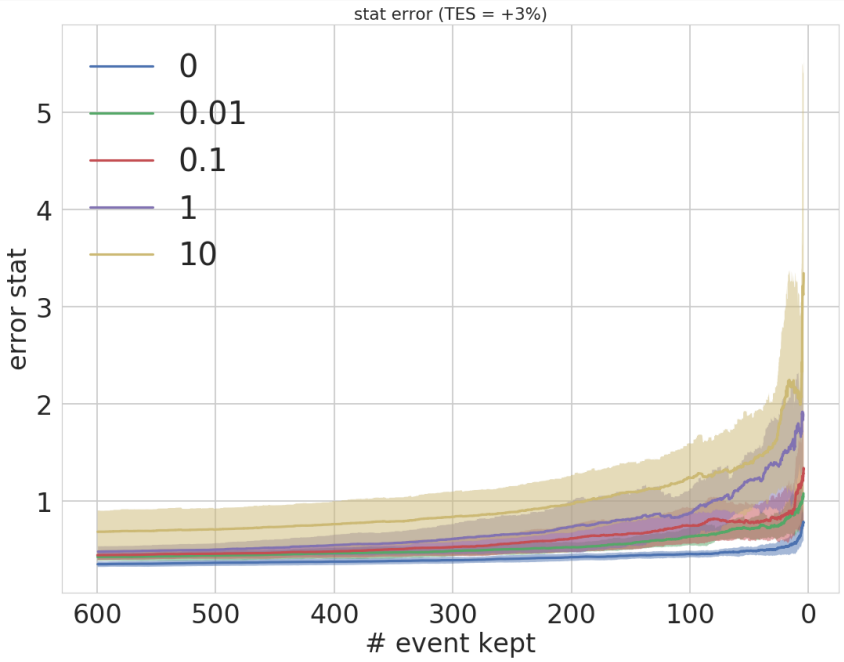
- Nothing beats the baseline (neural net)
- Tangent propagation is the worse.

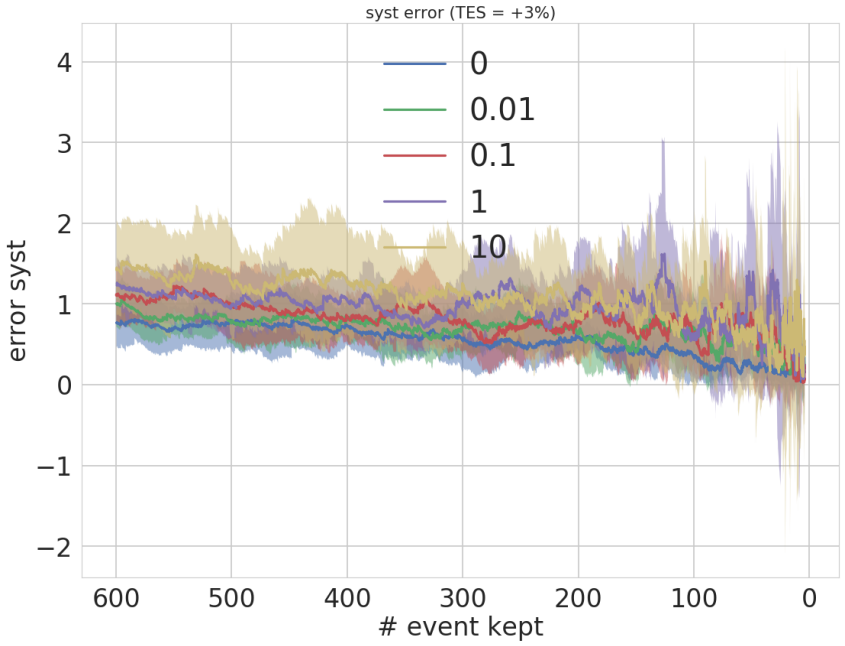
Mixture failure



- Data augmentation & Pivot imitate a basic Neural Net
- Nothing to learn from the new skewed data instances ?

Tangent Propagation : Loosing on stat only



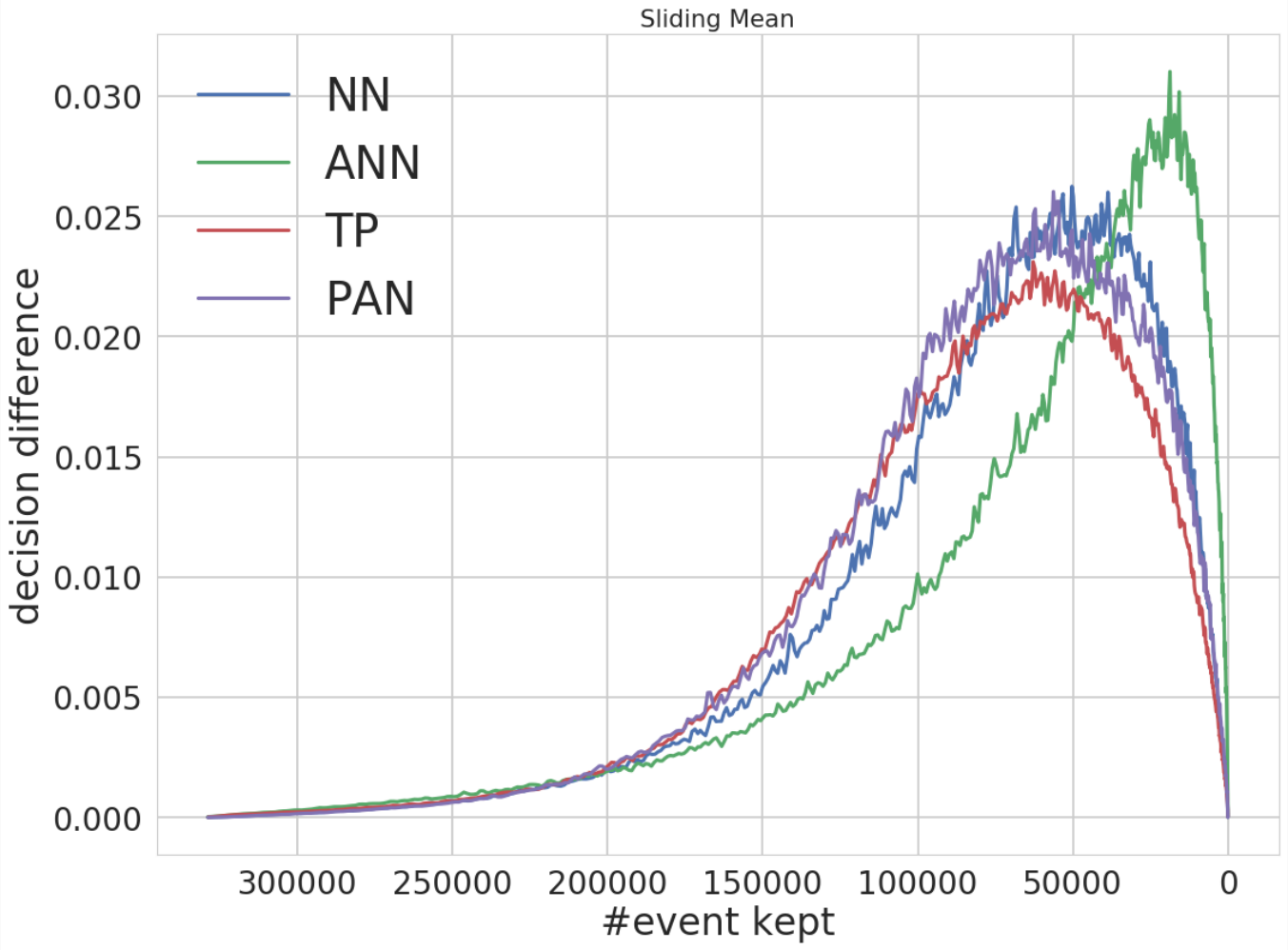


$$\frac{\sqrt{s_0+b_0}}{s_0}$$

$$\frac{(s_z-s_0)+(b_z-b_0)}{s_0}$$

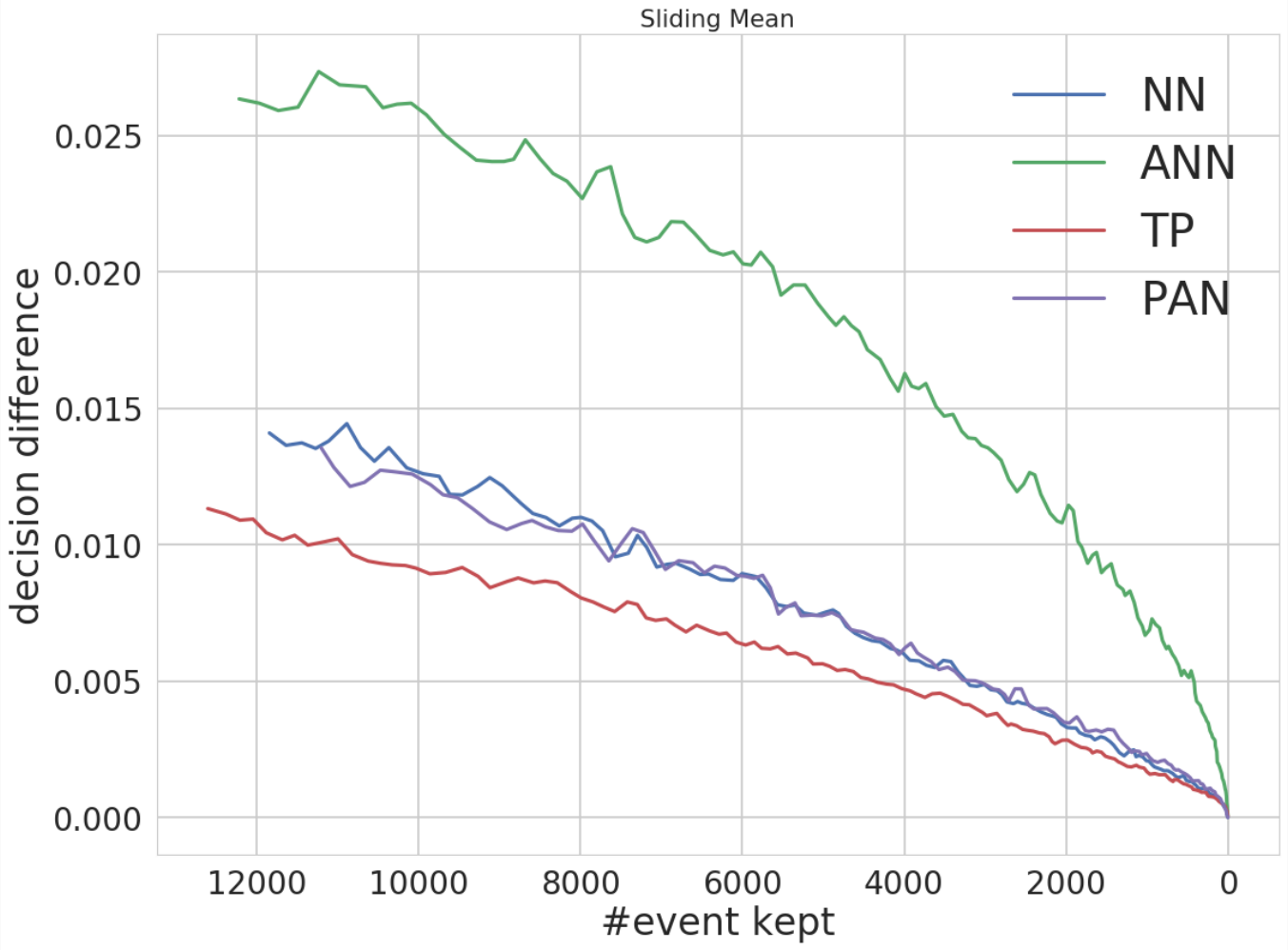
Performance loss mainly from statistical error

Decision shift vs score (+3%)



- Tangent Propagation is reducing decision shift
- Mixture models are not imitating neural net
- Why $\frac{\sigma_\mu}{\mu}$ is not showing this behaviour ?

Decision shift vs score (+3%)



- Tangent Propagation is reducing decision shift
- Mixture models are not imitating neural net
- Why $\frac{\sigma_\mu}{\mu}$ is not showing this behaviour ?

Plan

Systematics

what are systematic effects and systematic uncertainties ?

Domain apdapation

learning to be robust against distribution shifts

Benchmark

enable fast experimentation

Experimental results

deep study of the proposed technics

Conclusions and Perspectives

Conclusions and Perspectives

Rank vs Score

- We've been tackling the problem in the wrong way.
- We don't need robustness along the classification score.
- We need the rank to be constant.

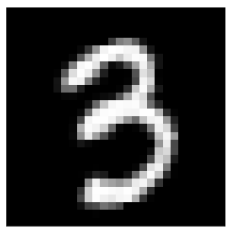
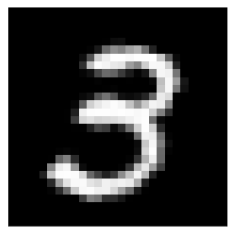
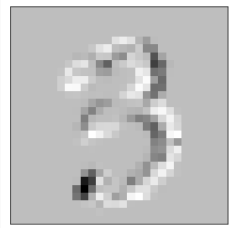
Perspective

- New toy dataset with controled properties
- Explore domain adaptation giving robust ranking

Thank you for your attention

Questions ?

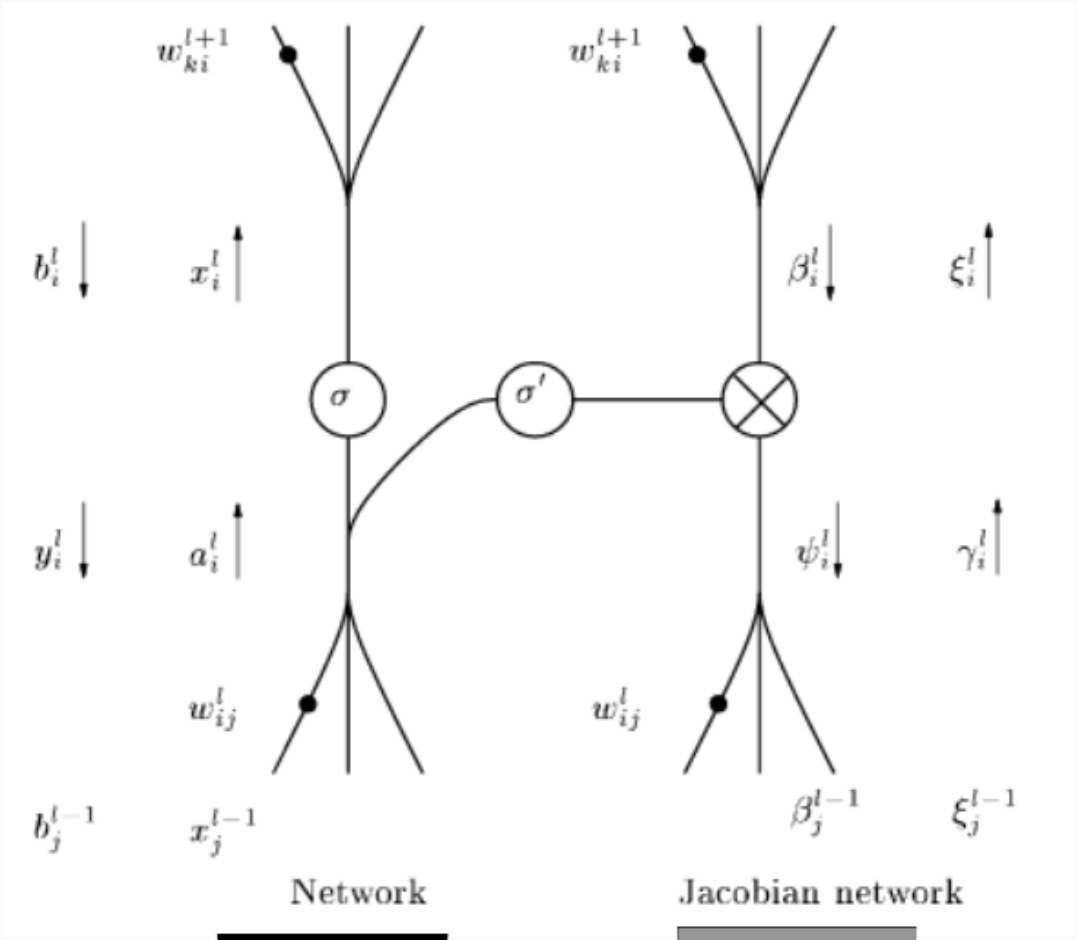
Tangent vector

$$\left.\frac{\partial d(\text{  }, z)}{\partial z}\right|_{z=0} = \frac{\text{  } - \text{  }}{2z} = \text{  }$$

Jacobian vector product trick

$p(Y=y|X=x)$

$\nabla_x f(x;\theta).\frac{\partial d(x,z)}{\partial z}$



Pivot Adversarial Neural Network

Generative process $p(X, Y, Z)$

Train a neural net $f(X; \theta_f)$ to estimate the probability density $p(Y|X)$ (Z is marginalized)

We want to have the pivotal condition :

$$\forall (z, z'), p(f(X; \theta_f) = score|z) = p(f(X; \theta_f) = score|z')$$

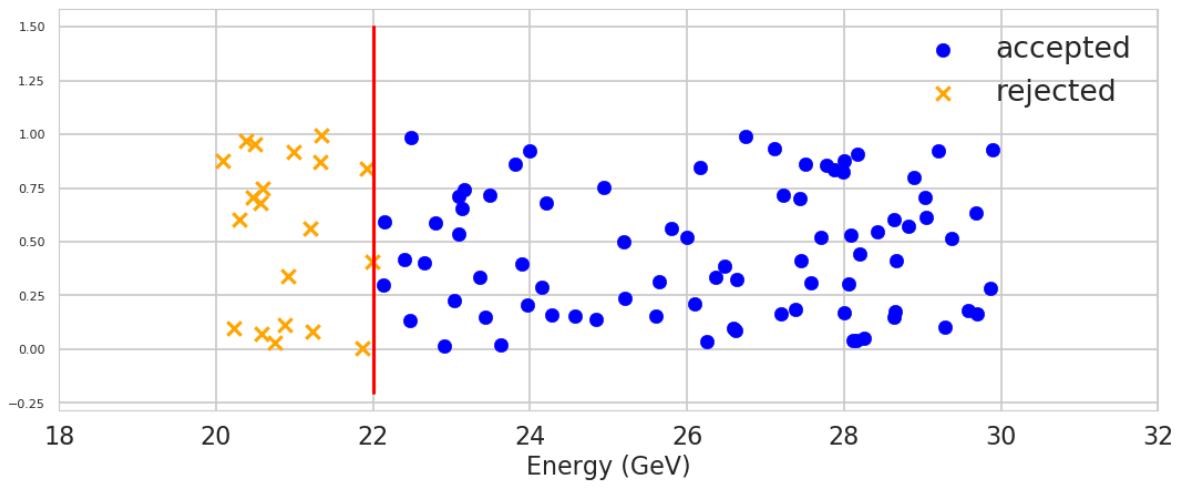
We want $f(X; \theta_f)$ and Z to be independent random variables

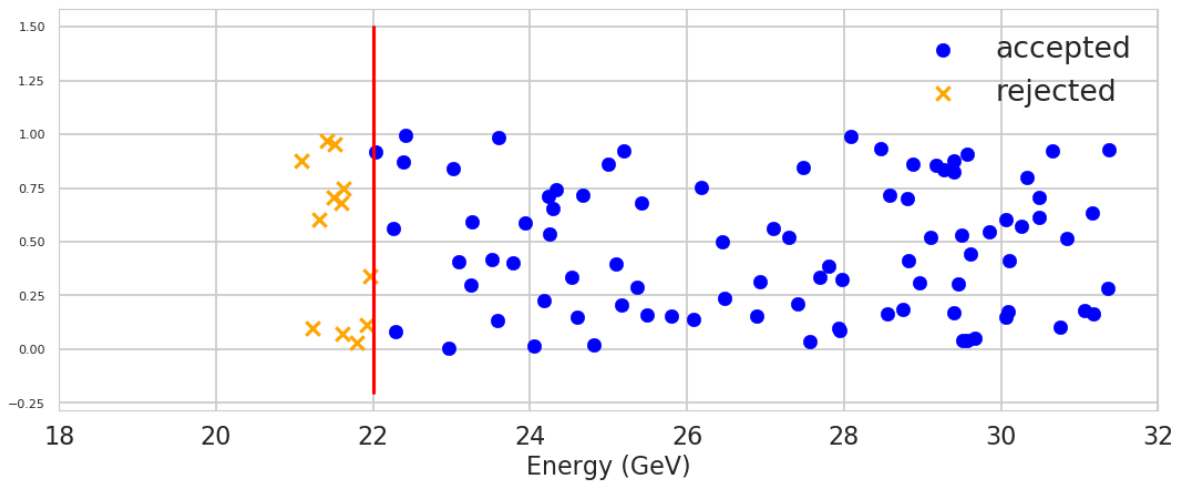
So we train an adversarial network to estimate $p(Z|f(X; \theta_f))$

[Louppe et al., 2016]

HEP benchmark : Tau energy scale

1. Scaling; 2. recompute derivate features; 3. cut data

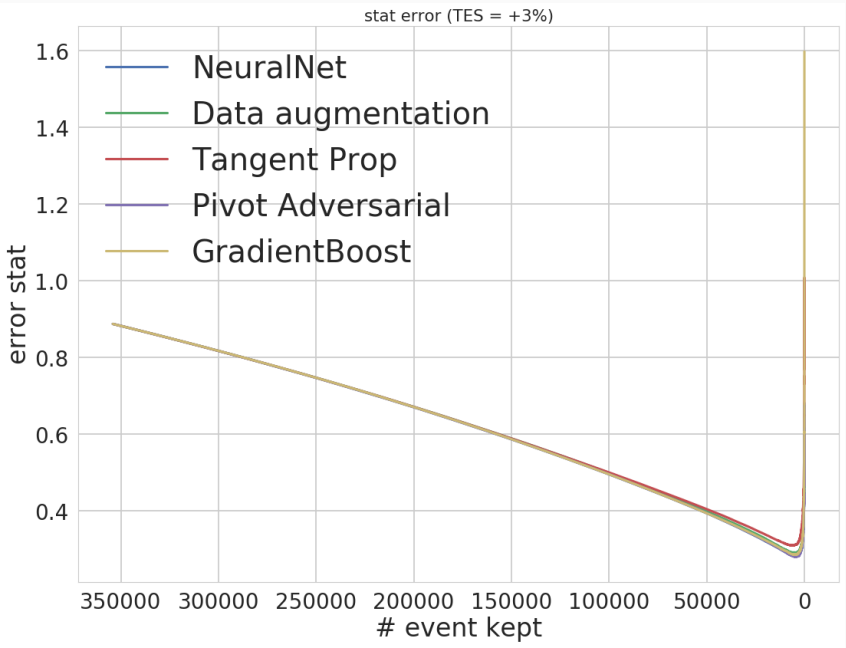


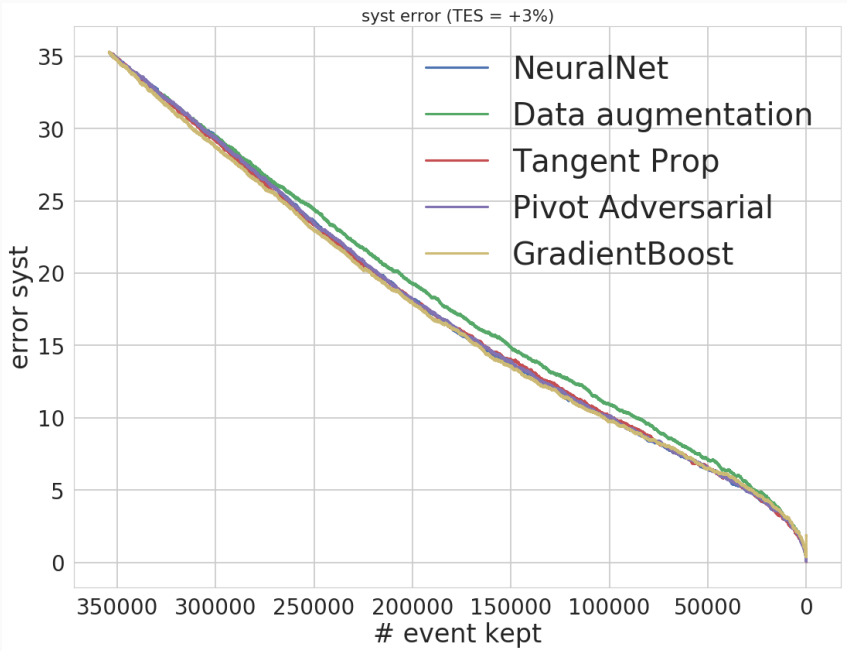


HEP benchmark : Measuring a cross section

- Cross section = counting positive class
- $N_z = s_z + b_z$ selected event (Positives)
- $s_z = \sum_{S, score_i > t} w_i$, selected signals (True positives)
- $b_z = \sum_{B, score_i > t} w_i$, selected backgrounds (False positives)
- $\hat{s}_z = N_z - b_0$
- Normalized cross section $\mu_z = \frac{\hat{s_z}}{s_0} = \frac{s_z + b_z - b_0}{s_0}$

Systematics dominate



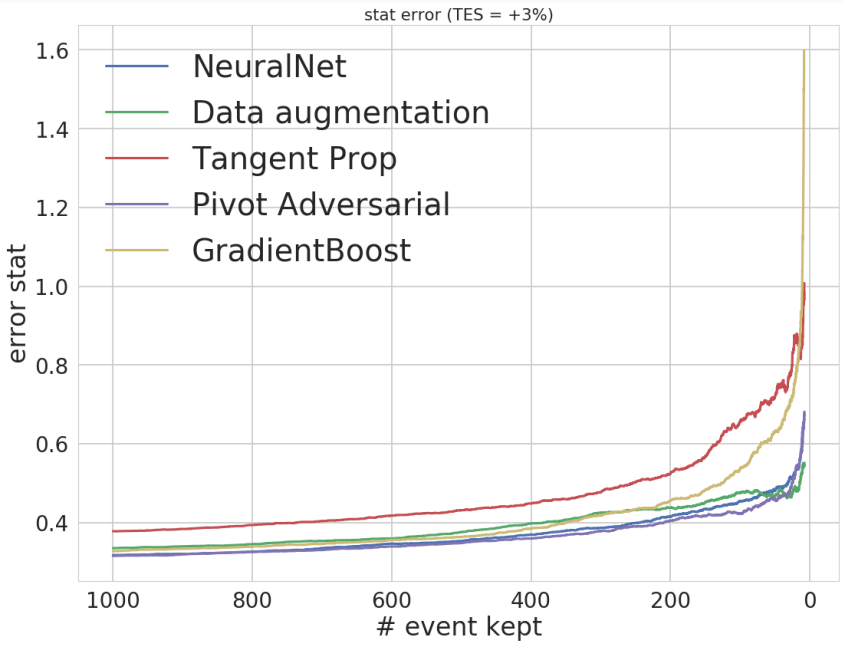


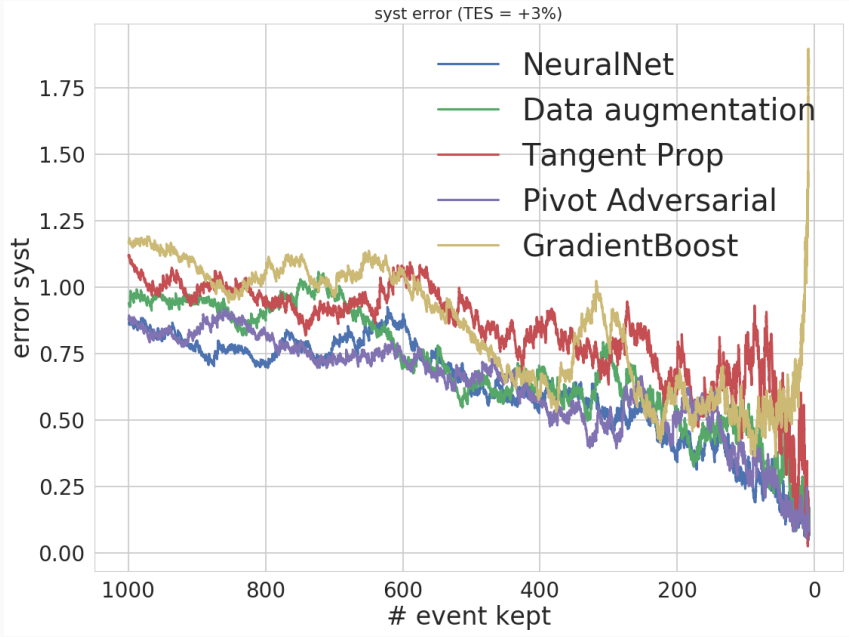
$$\frac{\sqrt{s_0+b_0}}{s_0}$$

$$\frac{(s_z-s_0)+(b_z-b_0)}{s_0}$$

- The systematics dominates the measurement error

Systematics dominate



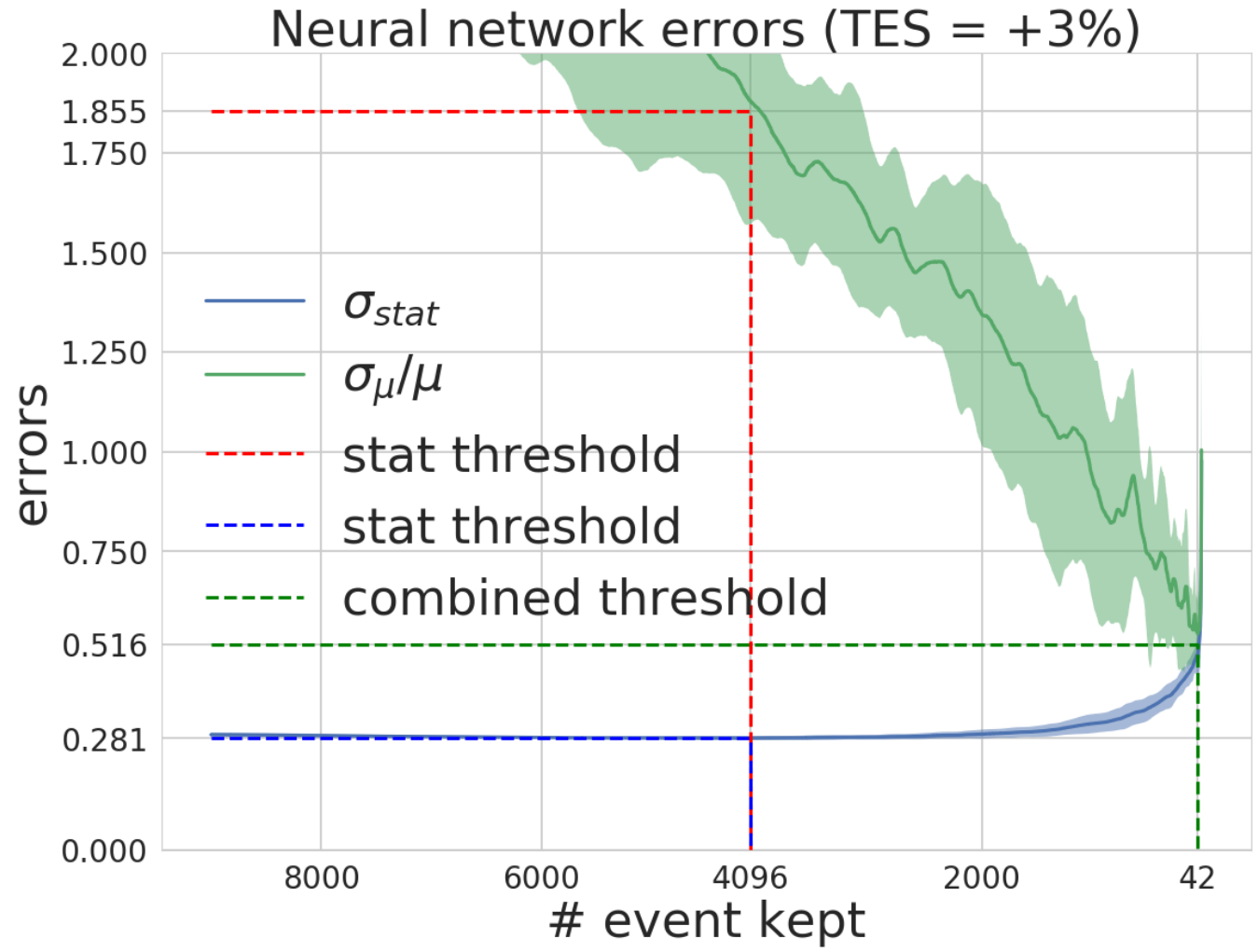


$$\frac{\sqrt{s_0+b_0}}{s_0}$$

$$\frac{(s_z-s_0)+(b_z-b_0)}{s_0}$$

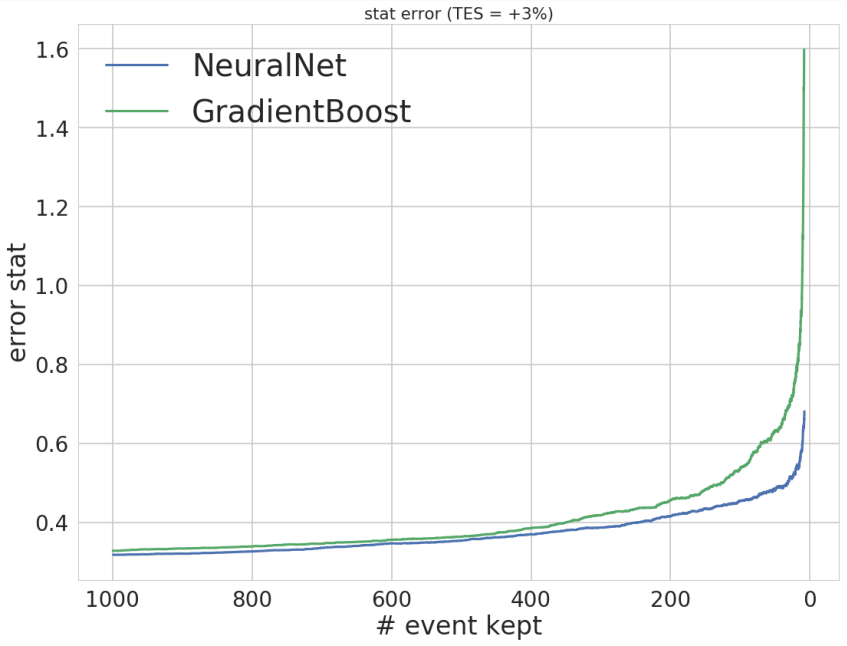
- The systematics dominates the measurement error
- But the statistical error cannot be ignored near minimum

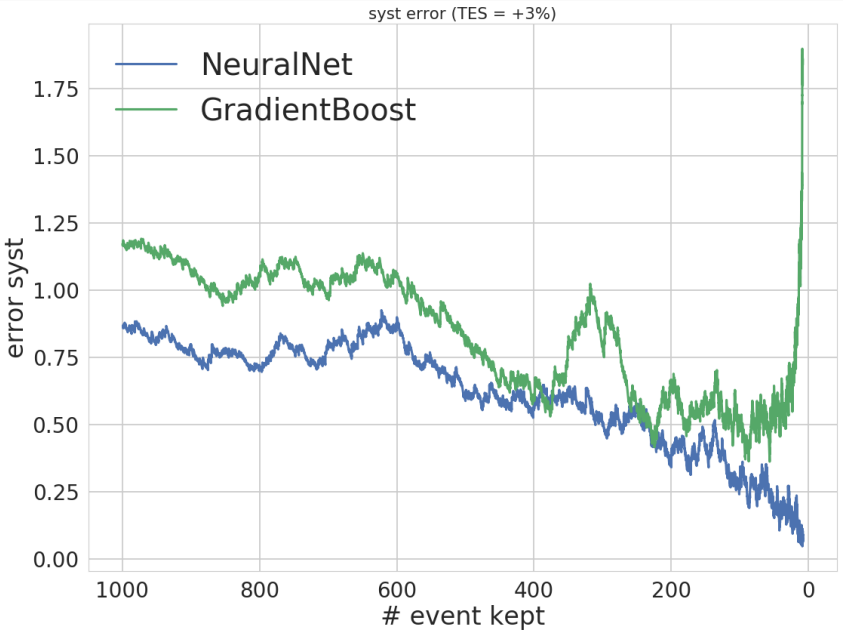
Threshold optimization



- Taking systematic effect into account drastically changes nb of event kept

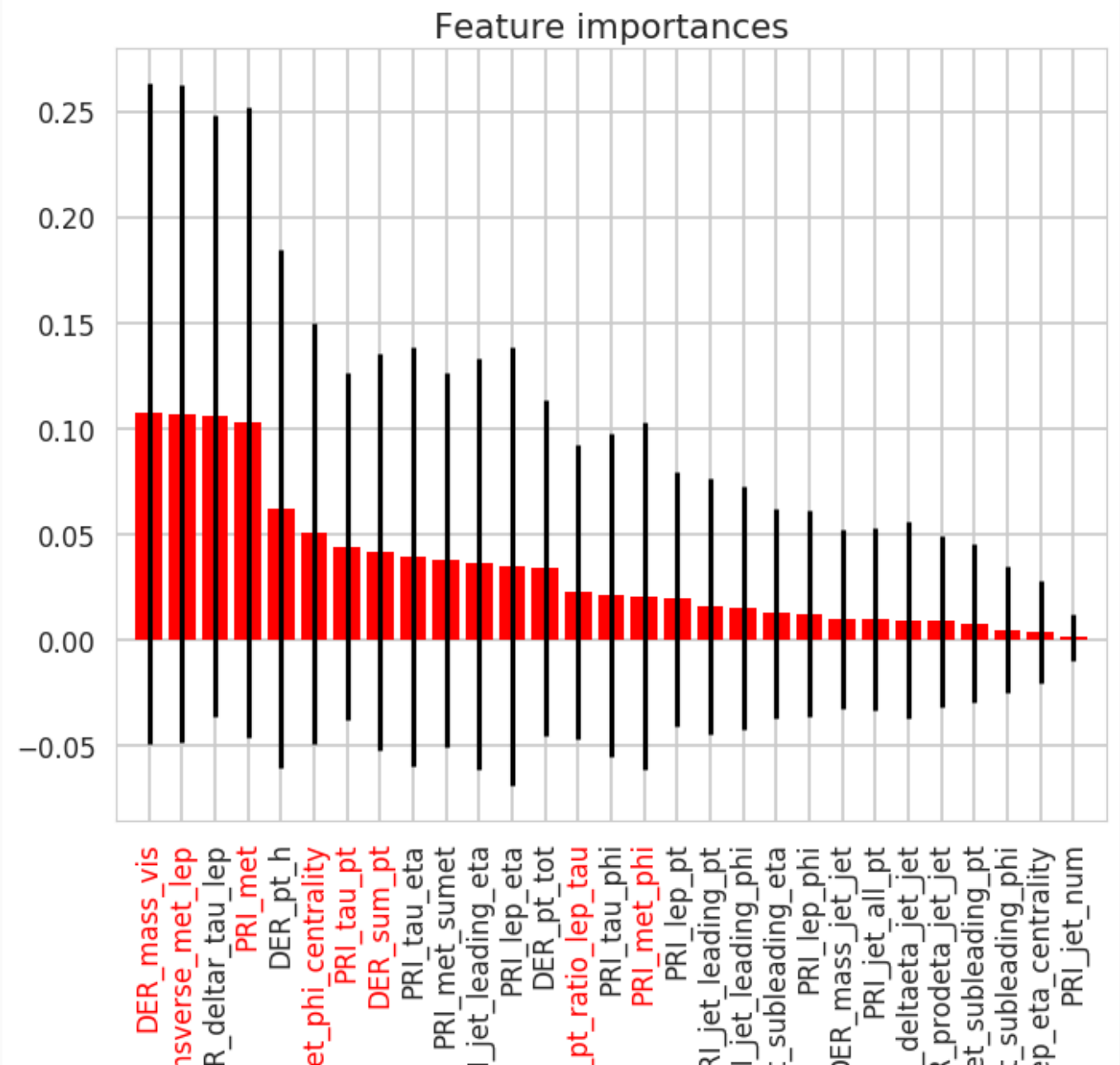
Gradient Boosting : details



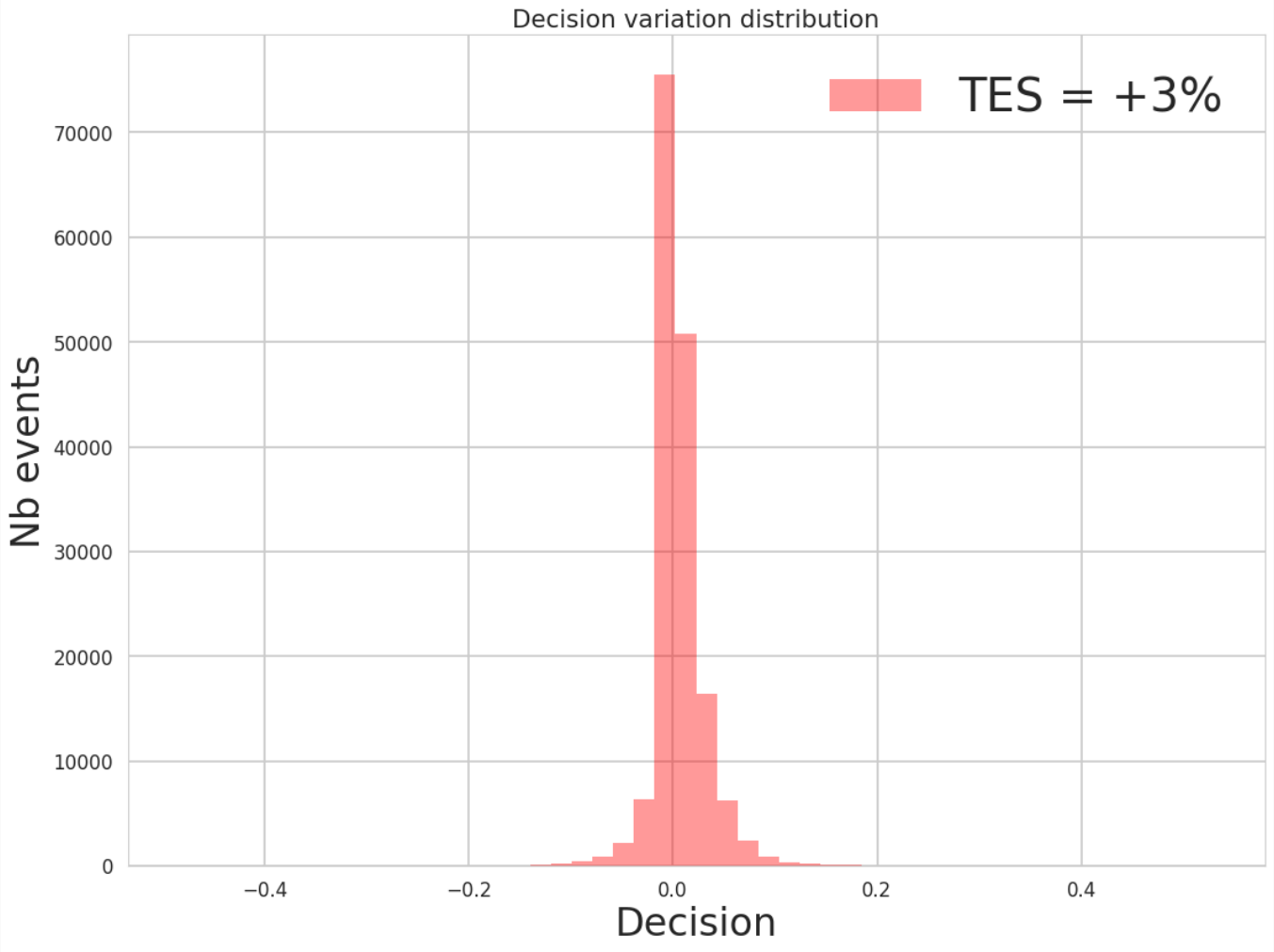


- sklearn 0.18
- 1000 trees
- maximum depth is 3 (i.e. 6 nodes maximum)
- Train only on nominal data ($z = 0$)

Gradient Boosting : using skewed features ?

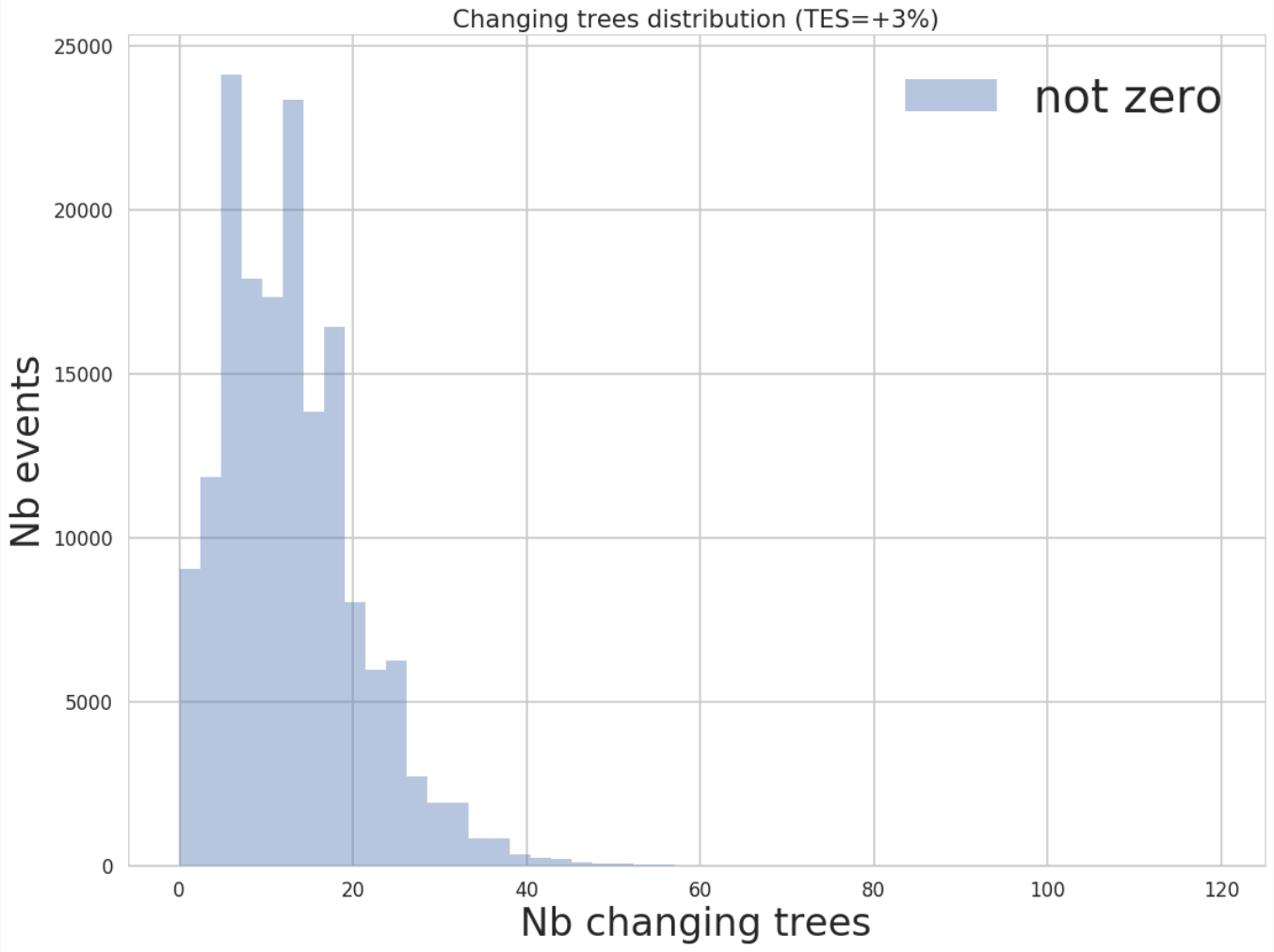


Gradient Boosting : small decision variation



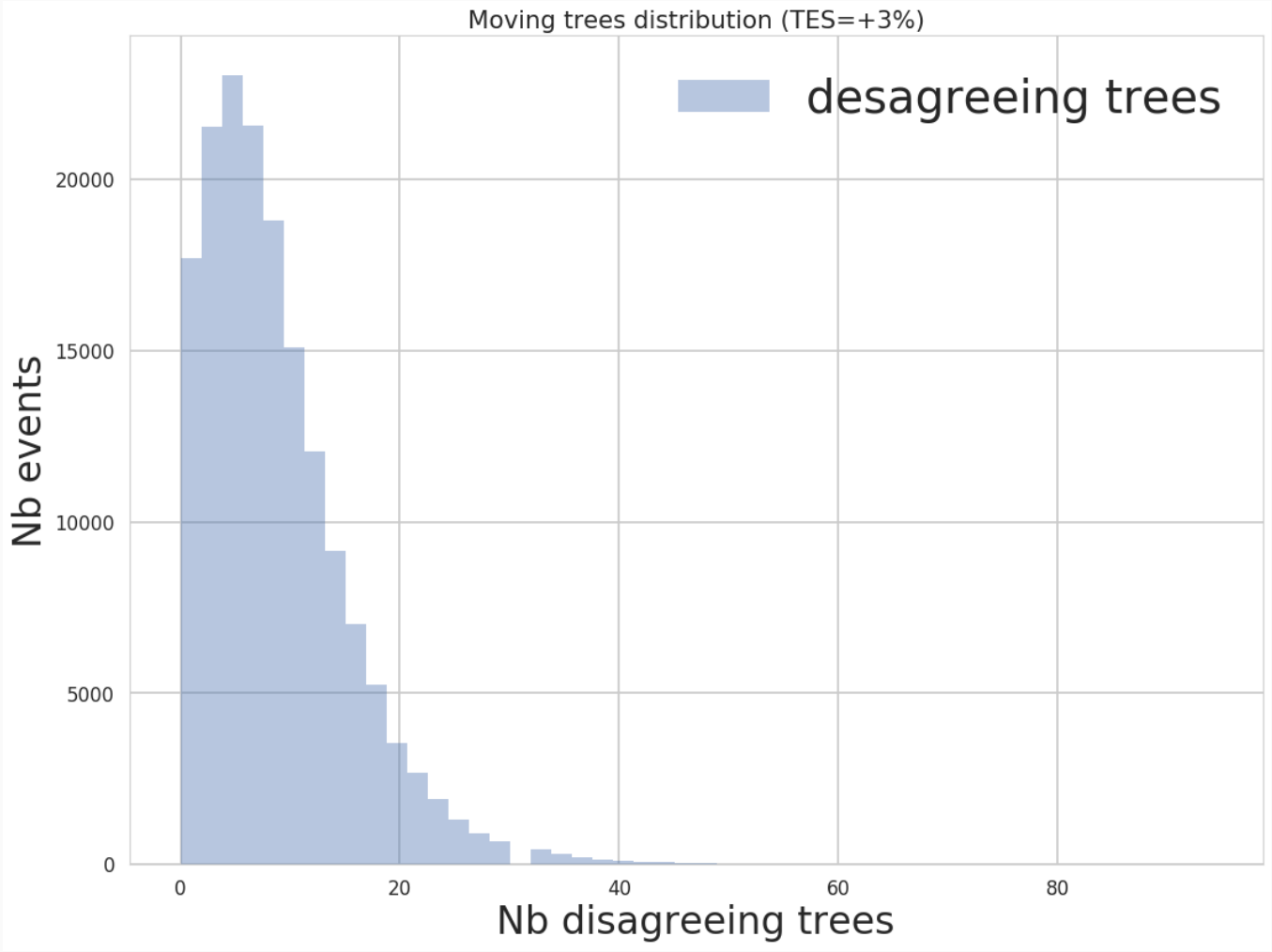
Score variation are small and goes both way

Gradient Boosting : few trees are affected



Only a few trees among the 1000 are changing

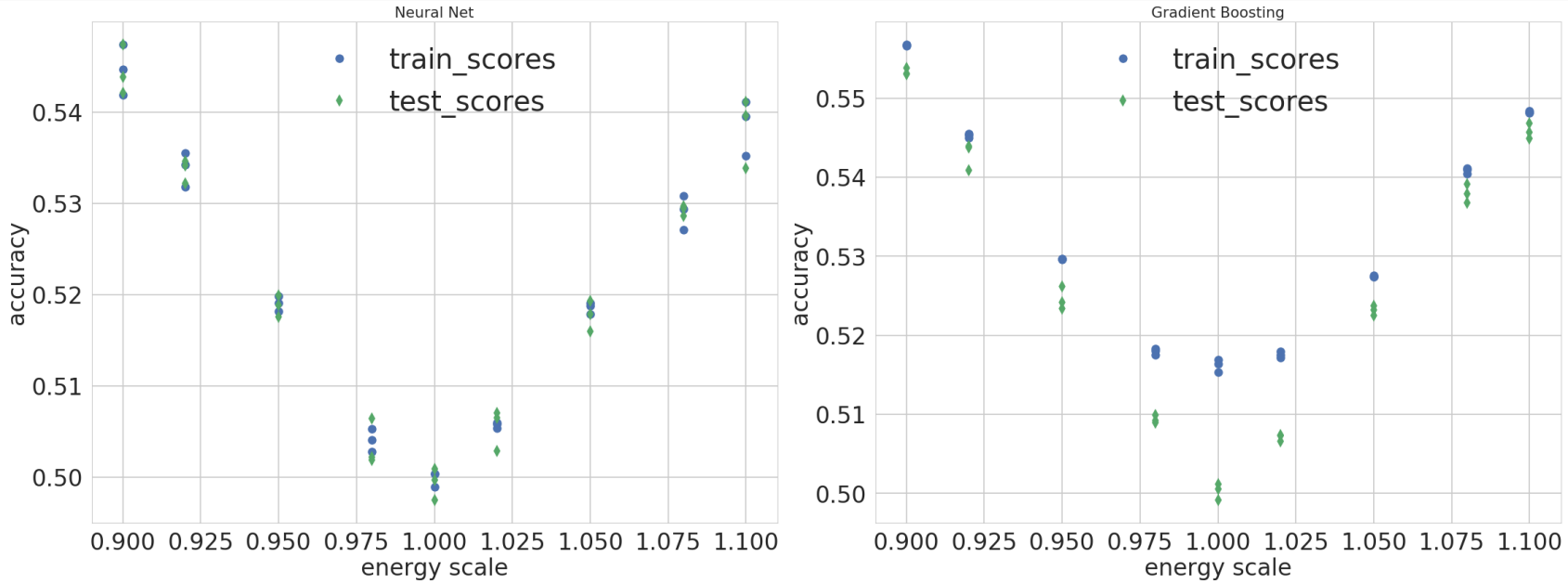
Gradient Boosting : tree disagreement



Gradient Boosting : conclusion

- Based on small trees : decision function is constant almost everywhere
- Many trees disagree among themselves

HIGGS : SEPARABILITY ISSUE

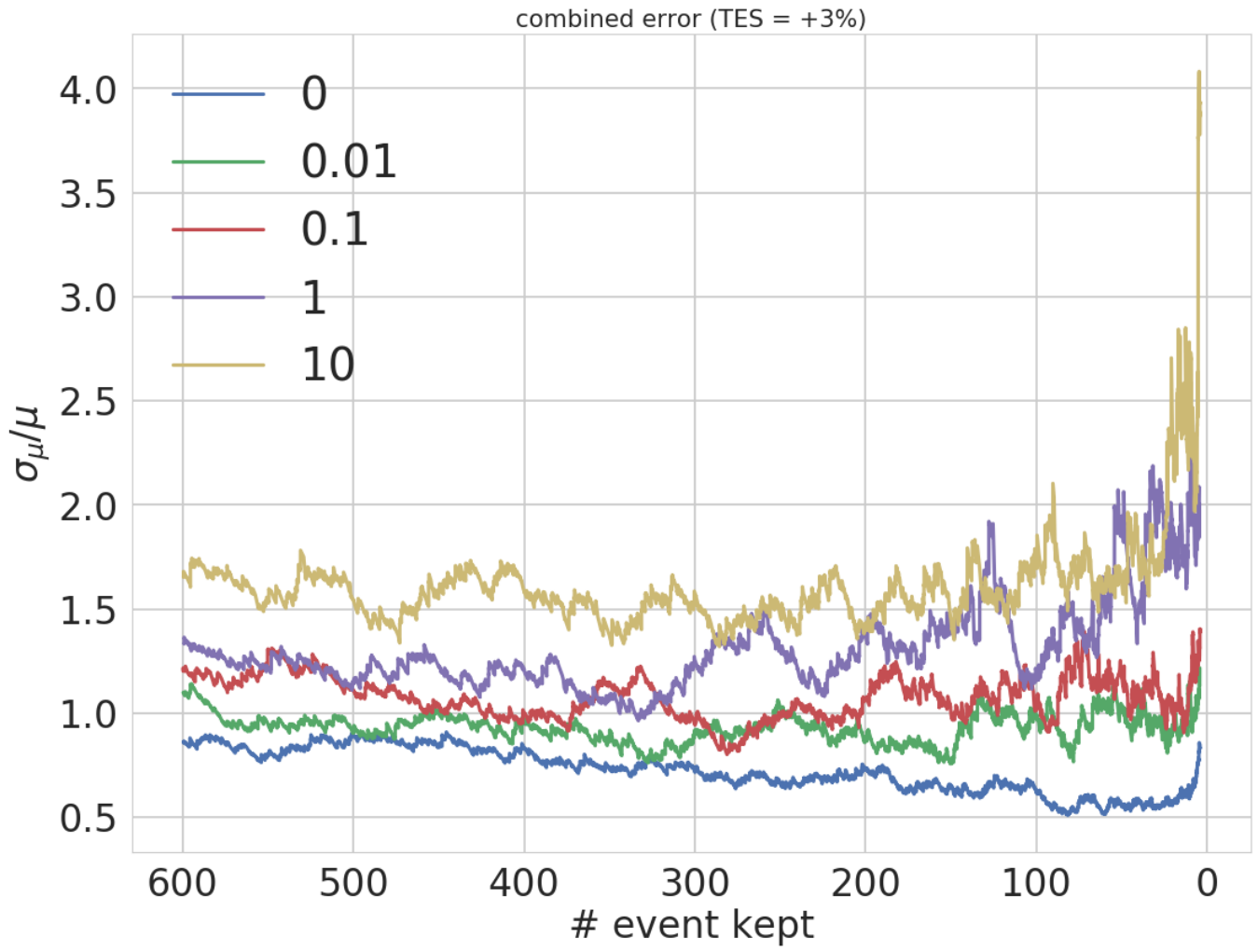


- low $\mathcal{H}$-divergence [Ben-David et al., 2010]

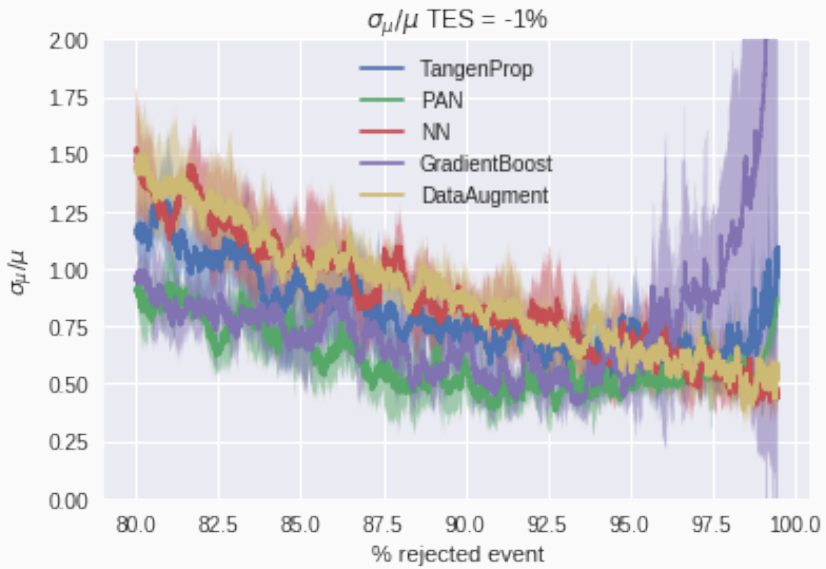
Tangent Propagation : Details

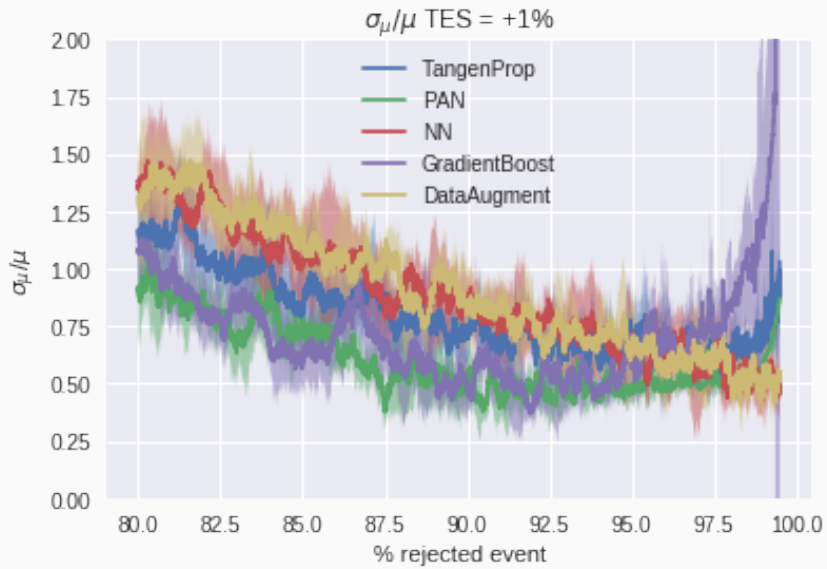
- 3 hidden layers
- 120 neurons each
- Adam optimizer
- Train on nominal data + tangent vectors

Tangent Propagation : Loosing for every λ

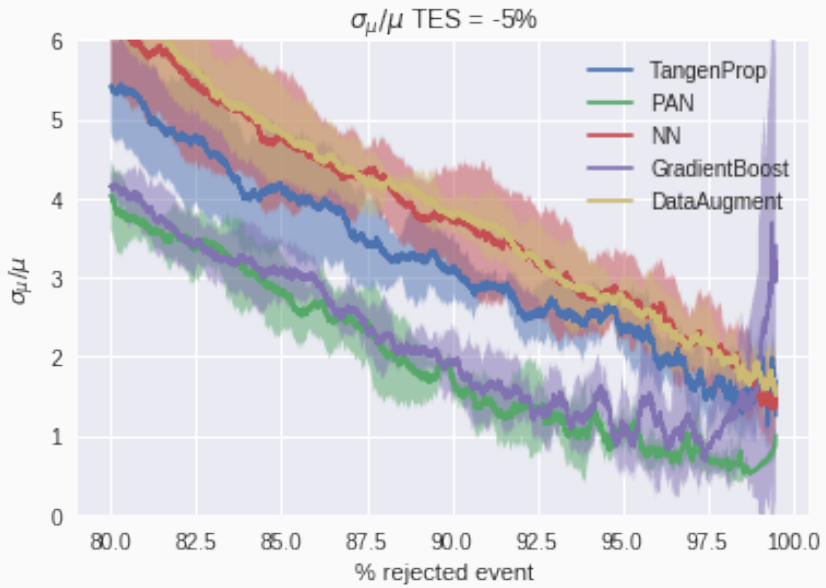


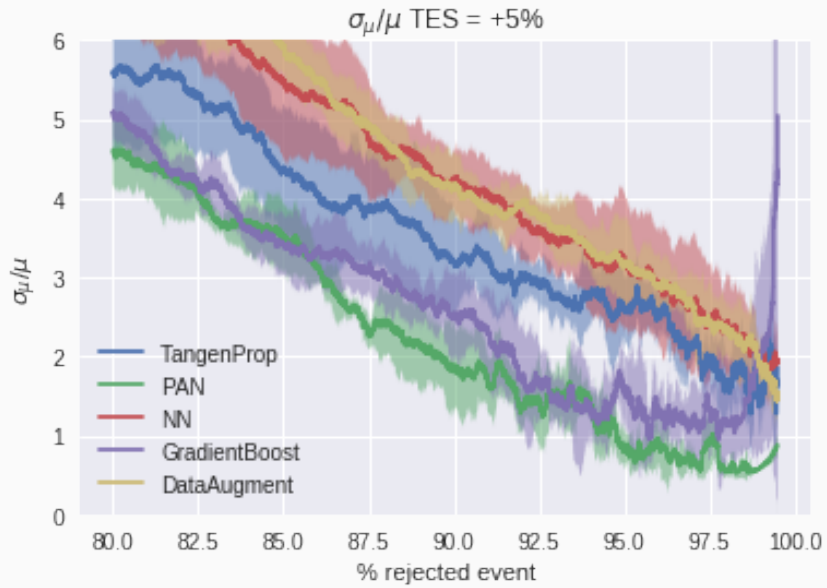
Symmetry ? (1%)


σ_μ/μ TES = -1%
TangenProp
PAN
NN
GradientBoost
DataAugment
σ_μ/μ
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
80.0
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
% rejected event

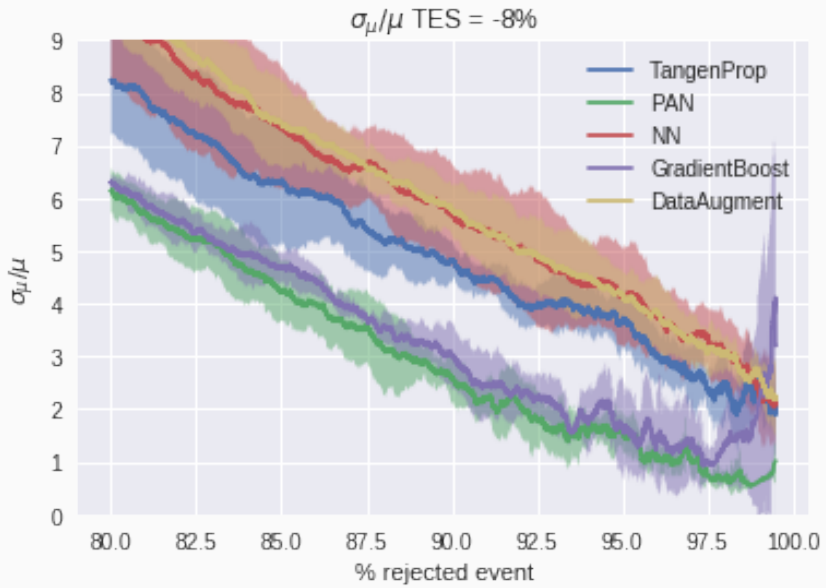

σ_μ/μ TES = +1%
TangenProp
PAN
NN
GradientBoost
DataAugment
σ_μ/μ
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
80.0
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
% rejected event

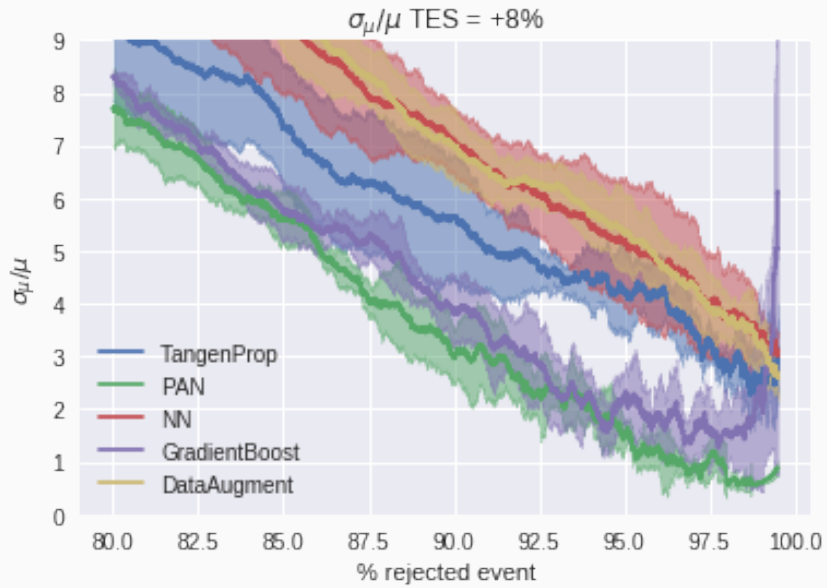
Symmetry ? (5%)


σ_μ/μ TES = -5%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment

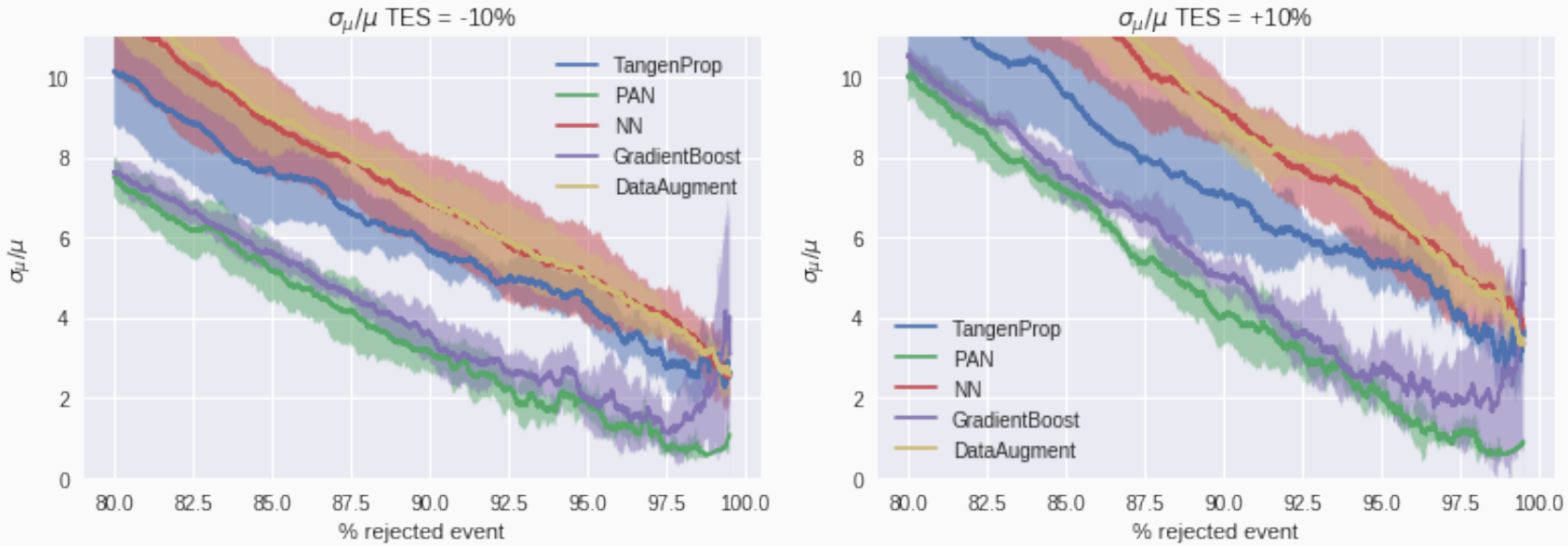

σ_μ/μ TES = +5%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment

Symmetry ? (8%)

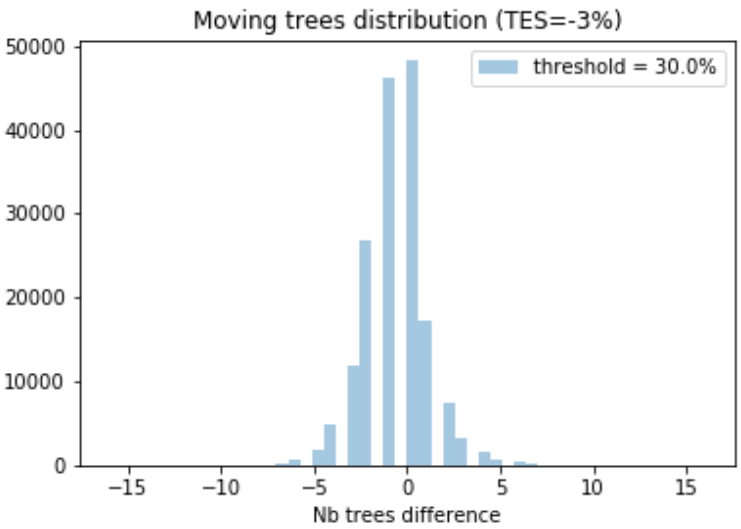

σ_μ/μ TES = -8%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment

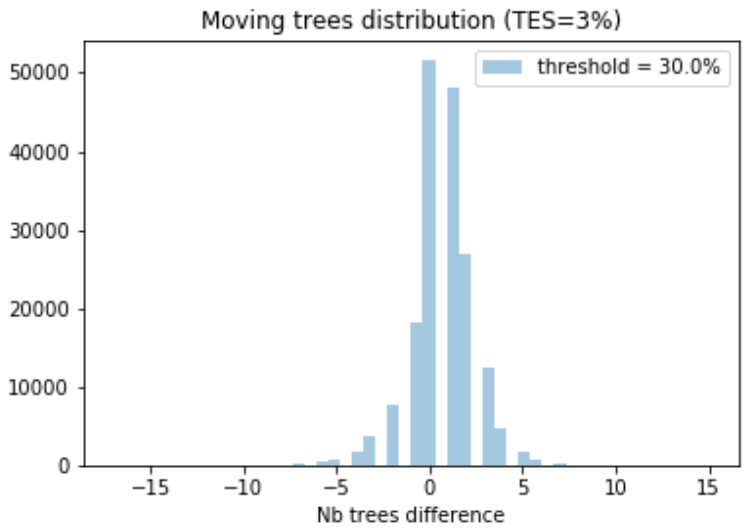

σ_μ/μ TES = +8%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment

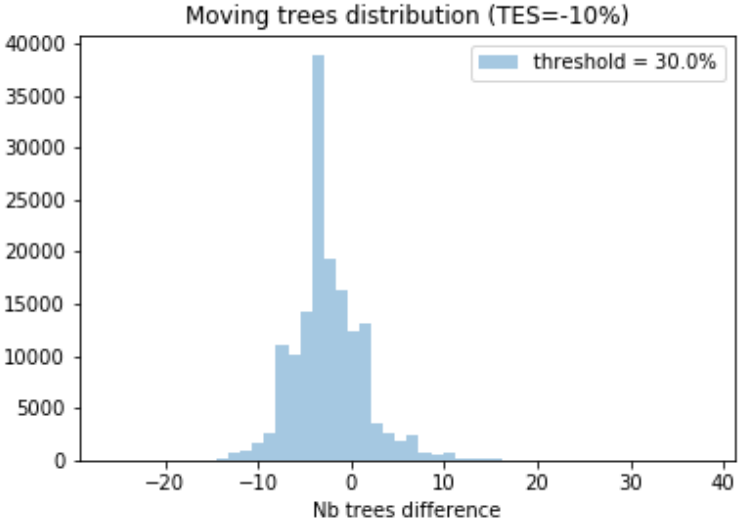
Symmetry ? (10%)

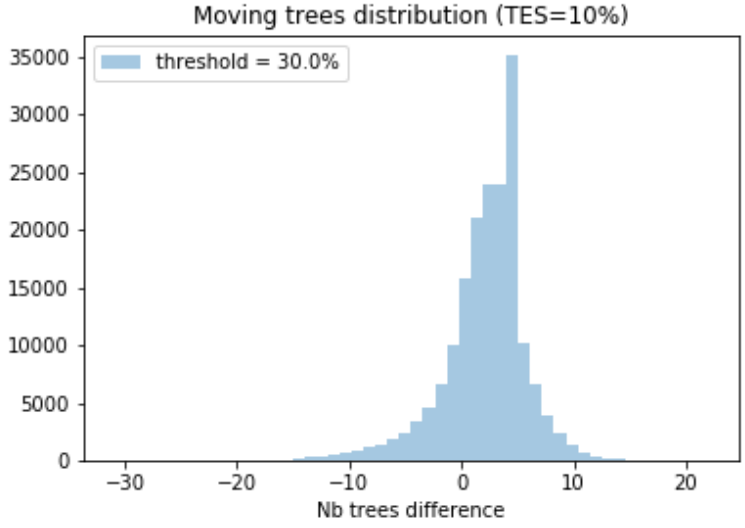

σ_μ/μ TES = -10%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment
σ_μ/μ TES = +10%
σ_μ/μ
% rejected event
TangenProp
PAN
NN
GradientBoost
DataAugment

Gradient Boosting : disagreement symmetry

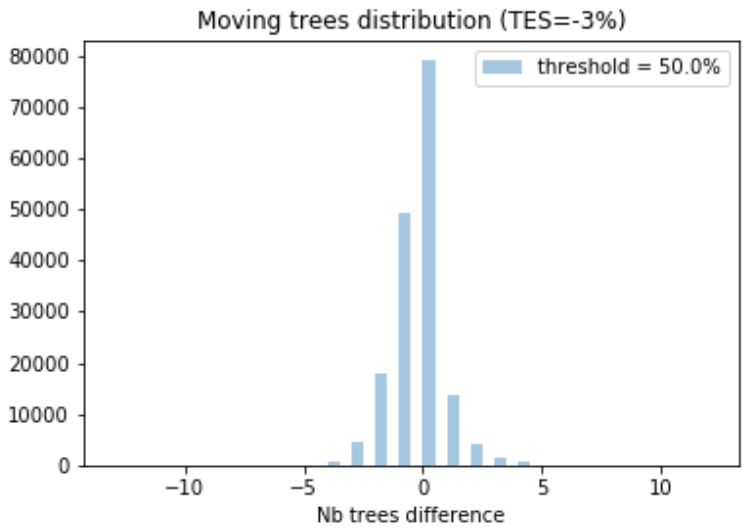


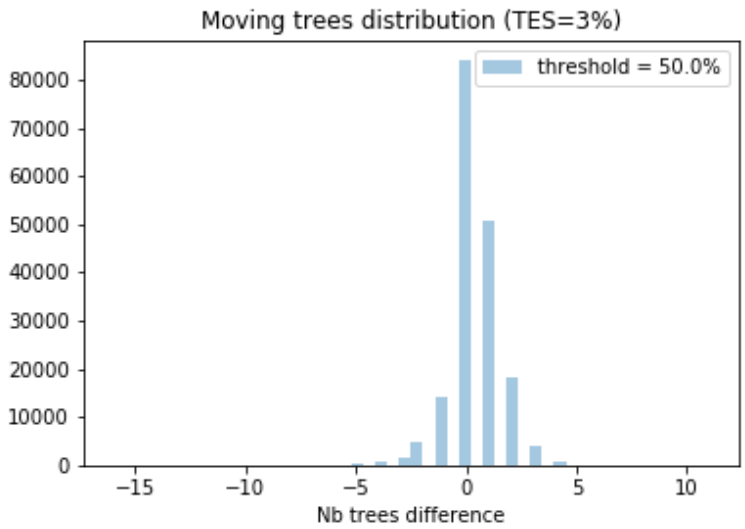


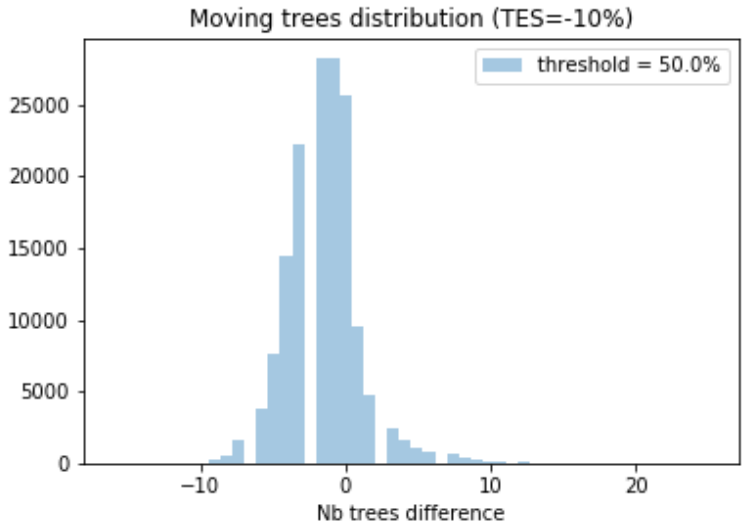


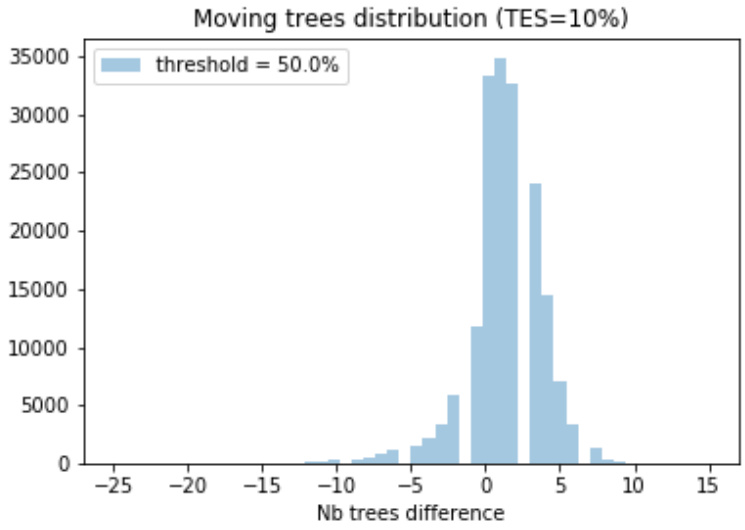


Gradient Boosting : strong disagreement symmetry









Discussion

invariant vs profiling.

Being invariant is too hard. In the end the nuisance params take only one value.

Ask David : ”In practice, a NP is η and p_T dependent and affect each events differently” So in the end the tau energy scale may differ a little bit event wise ? Answer : No, the NP is the same for all the events but its impact on each event depend on other variables

The DER features may contribute to the robustness of GB ?


Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D. (2014).
The Higgs boson machine learning challenge.
In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, volume 42 of *JMLR: Workshop and Conference Proceedings*, pages 19–55, Montreal, Canada.


Baldi, P., Sadowski, P., and Whiteson, D. (2014).
Enhanced Higgs to $\tau^+\tau^-$ Searches with Deep Learning.


Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010).
A theory of learning from different domains.
Machine Learning, 79(1-2):151–175.


Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2015).
Domain-Adversarial Training of Neural Networks.
arXiv:1505.07818 [cs, stat].
arXiv: 1505.07818.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative Adversarial Networks.
arXiv:1406.2661 [cs, stat].
arXiv: 1406.2661.

Louppe, G., Kagan, M., and Cranmer, K. (2016).
Learning to Pivot with Adversarial Networks.
arXiv:1611.01046 [physics, stat].
arXiv: 1611.01046.

Rifai, S., Dauphin, Y., Vincent, P., Bengio, Y., and Muller, X. (2011).
The Manifold Tangent Classifier.
In *NIPS*, volume 271, page 523.

Simard, P. Y., Victorri, B., LeCun, Y., and Denker, J. S. (1991).
Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network.
In Moody, J. E., Hanson, S. J., and Lippmann, R., editors, *NIPS*, pages 895–903. Morgan Kaufmann.