Theory Uncertainties.

(aka The Ugly)

Frank Tackmann

Deutsches Elektronen-Synchrotron

PHYSTAT Systematics, Nov. 02, 2021

European Research Council
Established by the European Commission

Disclaimers and Apologies.

- I'm not an experimentalist let alone a statistics expert, so apologies if some things are too pedestrian and others too complicated ...
- Half (if not most) of the talk will actually be about "how to estimate", because
  - We don't actually have good methods to properly do that (yet)
  - It's important to understand the limitations before we can talk about implementation
- I will focus on perturbative QCD predictions and theory uncertainties due to missing higher corrections
  - Many things (likely) carry over to other types of theory predictions
- While I have tried to capture the general state of affairs, for expedience, illustrative examples are taken from my own work
  (and since I'll be a bit critical, I want to avoid "bashing" the work of others)

Overview.

What Am I Talking About?

Pendulum example

- We have a formula to obtain the quantity of interest (g) from the observed quantities (number of swings/time)
- This formula is the *theory prediction*
- The *theory uncertainty* is due to the fact that in many cases the formula itself is not fully exact (e.g. derived in some approximation)
  - It is *not* the inexact knowledge of parameters needed in the (otherwise exact) formula (like the length of the pendulum)

    These are the usual systematics (parametric uncertainties)
  - Note: Sometimes certain parametric uncertainties are also called a theory uncertainty just because they primarily enter via the theory predictions (e.g. parton distribution functions). For this talk these are not theory uncertainties.

What Am I Talking About?

Pendulum example

- We have a formula to obtain the quantity of interest (g) from the observed quantities (number of swings/time)
- This formula is the *theory prediction*
- The *theory uncertainty* is due to the fact that in many cases the formula itself is not fully exact (e.g. derived in some approximation)
  - It is *not* the inexact knowledge of parameters needed in the (otherwise exact) formula (like the length of the pendulum)

    These are the usual systematics (parametric uncertainties)
  - Note: Sometimes certain parametric uncertainties are also called a theory uncertainty just because they primarily enter via the theory predictions (e.g. parton distribution functions). For this talk these are not theory uncertainties.

⇒ The Challenge: How to account for the inexactness of the formula itself?

- The theory uncertainty is different from other systematics because a priori there is no auxiliary measurement to improve inexactness
- But wait until the end of the talk ...

In Collider Physics

There is an almost continuous spectrum of interpretation steps with typically increasing dependence on theory predictions

In Collider Physics

Data → Measured cross sections (or limits) → Lagrangian parameters

↑ *Theory predictions* ↑ *Theory predictions*

There is an almost continuous spectrum of interpretation steps with typically increasing dependence on theory predictions

- Often it is separated somewhere in the middle
- Let us consider the simplest case (toward the right)

$$\sigma_i^{\text{measured}} = \sigma_i^{\text{predicted}}(x)$$

  - where x denotes the parameter(s) of interest to be determined
  - Precise method to obtain x is not relevant for now (e.g. the fitting methodology)
- We *never* know the exact formula for $\sigma_i^{\text{predicted}}(x)$
  - In fact, often (toward the left) we do not even have a formula, just a program

Sources of Inexactness.

All the approximations we have to make when deriving $\sigma^{\text{predicted}}$

- Perturbative expansion in coupling constants: $\alpha_s, \alpha_{\text{em}}$

$$\sigma^{\text{predicted}} = \underbrace{c_0}_{\text{LO}} + \underbrace{\alpha\, c_1}_{\text{NLO}} + \underbrace{\alpha^2\, c_2}_{\text{NNLO}} \underbrace{+ \alpha^3 c_3 + \cdots}_{\text{neglected}}$$

  - Usually the most relevant (QCD), so I will entirely focus on this one

- Various other expansions (usually used implicitly at their lowest order)
  - Kinematic power expansions: p_T/Q (e.g. in parton showers, resummation)
  - Nonperturbative power expansions: Λ/Q
  - Mass expansions: m_q/Q
  - Usually less relevant for uncertainty
    (and thus frequently ignored)

Sources of Inexactness.

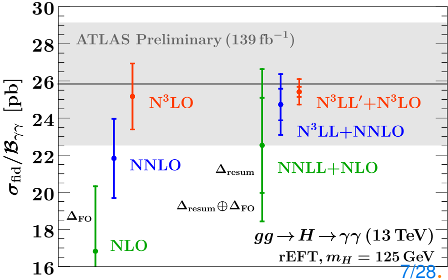
All the approximations we have to make when deriving $\sigma^{\text{predicted}}$

- Perturbative expansion in coupling constants: $\alpha_s, \alpha_{\text{em}}$

$$\sigma^{\text{predicted}} = \underbrace{c_0}_{\text{LO}} + \underbrace{\alpha\, c_1}_{\text{NLO}} + \underbrace{\alpha^2\, c_2}_{\text{NNLO}} \; \underbrace{+\, \alpha^3 c_3 + \cdots}_{\text{neglected}}$$

  - Usually the most relevant (QCD), so I will entirely focus on this one

- Various other expansions (usually used implicitly at their lowest order)
  - Kinematic power expansions: p_T/Q (e.g. in parton showers, resummation)
  - Nonperturbative power expansions: Λ/Q
  - Mass expansions: m_q/Q
  - Usually less relevant for uncertainty (and thus frequently ignored)



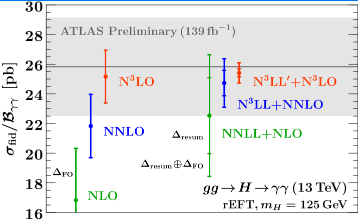
To account for inexactness, we quote an uncertainty for our prediction

$$\sigma^{\text{predicted}} = \sigma_{\text{order}} \pm \Delta\sigma_{\text{order}}$$

What Should $\Delta\sigma$ Actually Represent or Mean?

$$\sigma^{\text{predicted}} = \underbrace{c_0}_{\text{LO}} + \underbrace{\alpha c_1}_{\text{NLO}} + \underbrace{\alpha^2 c_2}_{\text{NNLO}} + \underbrace{\alpha^3 c_3 + \cdots}_{\text{neglected}}$$

$$= \sigma \pm \Delta\sigma$$



1. *Estimate* of difference to true result: $\Delta\sigma \approx |\sigma^{\text{true}} - \sigma|$

2. *Estimate* of missing next order(s): $\Delta\sigma \approx \alpha^3 c_3$
   - Same as above if series converges well (uncertainty on uncertainty is small)
   - Only condition we can check, so how most theorists tend to think about it
   - I'm happy when uncertainty covers highest-order result, unhappy when not

3. However, in implementation it is practically always used as some "1σ"
   - $|\sigma^{\text{true}} - \sigma| \leq \Delta\sigma$ with 68% "probability"
   - But "probability" in what sense?
   - And what probability distribution?

And How Is It Distributed?

Theorist: "Do not use a Gaussian, it should be a flat distribution"

Translation: "The central value shouldn't be the most likely"

- A flat box of size $\Delta\sigma$ makes no sense (obviously too aggressive)
- How about a central flat region with some (gaussian) tails?
  - How large is the flat vs. tail region? What does $\Delta\sigma$ cover?

And How Is It Distributed?

Theorist: "Do not use a Gaussian, it should be a flat distribution"

Translation: "The central value shouldn't be the most likely"

- A flat box of size $\Delta\sigma$ makes no sense (obviously too aggressive)
- How about a central flat region with some (gaussian) tails?
  - How large is the flat vs. tail region? What does $\Delta\sigma$ cover?

My opinion: Use whatever distribution suits you (Gaussian, log-normal)

- Until someone demonstrates that the choice actually matters
  - And if it does matter, you're so sensitive to theory uncertainties that you have much bigger problems ...
- And if a theorist complains, you can go ahead and easily measure their true mental distribution, by asking:

  "Which percentage of [citations on paper, monthly salary, postdoc funding, ...] are you willing to loose if the next order is outside your uncertainty? 68%? 95%?"

  (I'm only half-joking ... just please never ask me this question)

Even Bigger Challenge: Correlations.

Correlations can be crucial as soon as several predictions σ_i are used simultaneously

- Prototype of extrapolation that happens in many data-driven methods

$$\underbrace{\sigma^{\mathrm{SR}}(X)}_{\text{needed}} = \underbrace{\Big[\sigma^{\mathrm{CR}}(Y)\Big]_{\text{measured}}}_{\text{measure precisely}} \times \underbrace{\left[\frac{\sigma^{\mathrm{SR}}(X)}{\sigma^{\mathrm{CR}}(Y)}\right]_{\text{predicted}}}_{\text{theory uncertainties cancel}}$$

  - Cancellation of theory uncertainties is often taken for granted, but in fact crucially relies on precise correlations

- Whenever we deal with differential spectrum
  - Integrated cross section often more precisely predicted than spectrum (There are theoretical reasons for that.)
  - Any shape uncertainty which cancels in integral inherently requires (long-range) anticorrelation across spectrum

Scale Variations.

(The ugly present)

Scale Variations in a Nutshell.

The prevalent method to estimate $\Delta\sigma$ in perturbative QCD predictions

$$\begin{aligned}\sigma &= c_0 + \alpha(\mu_0)\, c_1 + \alpha^2(\mu_0)\, c_2 + \cdots \\ &= c_0 + \alpha(\mu)\, c_1 + \underbrace{\alpha^2(\mu)(c_1\, b_0 \ln \mu/\mu_0 + c_2) + \cdots}_{\text{neglected}}\end{aligned}$$

- The scale μ is our choice of how we perform the expansion
  - All-order result does not depend on this choice, i.e. it is μ-independent
  - The truncated series does depend on this choice, and this residual μ dependence is cancelled by neglected higher-order terms

Scale Variations in a Nutshell.

The prevalent method to estimate $\Delta\sigma$ in perturbative QCD predictions

$$\begin{aligned}\sigma &= c_0 + \alpha(\mu_0)\, c_1 + \alpha^2(\mu_0)\, c_2 + \cdots \\ &= c_0 + \alpha(\mu)\, c_1 + \underbrace{\alpha^2(\mu)(c_1\, b_0 \ln \mu/\mu_0 + c_2) + \cdots}_{\text{neglected}}\end{aligned}$$

- The scale μ is our choice of how we perform the expansion
  - All-order result does not depend on this choice, i.e. it is μ-independent
  - The truncated series does depend on this choice, and this residual μ dependence is cancelled by neglected higher-order terms
- *Scale variations* exploit this by taking the difference between two choices to estimate the typical size of neglected higher-order terms

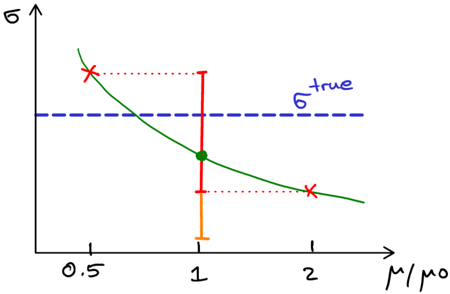
$$\begin{aligned}\sigma_{\mathrm{NLO}} \equiv \sigma|_{\mu_0} \qquad &= c_0 + \alpha(\mu_0)\, c_1 \\ \Delta\sigma_{\mathrm{NLO}} = \sigma|_{\mu} - \sigma|_{\mu_0} &= \alpha^2(\mu)\, c_1\, b_0 \ln \mu/\mu_0 + \cdots\end{aligned}$$

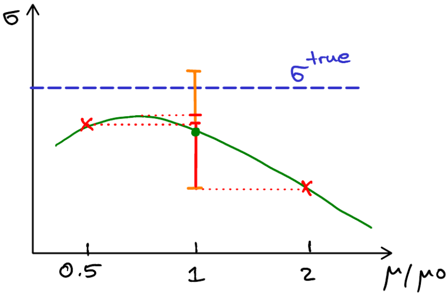
  - Basically just a convenient way to estimate size of missing c_2 from known c_1

Scale Variations Are Not Very Reliable.

- Basic problem: Nothing really guarantees that $c_2 \approx c_1 b_0$ is a good approximation
  - Many known examples where resulting $\Delta\sigma$ is underestimated
  - σ_1 can be accidentally small due to internal cancellations
  - There can be new structures in c_2 that are not yet present in c_1
- Most people would probably agree that it would be good to have something better
  - But we have gotten used to the ugliness and very good at ignoring it
  - And despite everything, scale variations are extremely convenient and we don’t really have anything better either
  - And anyway, there is no way to really know $\Delta\sigma$, so instead of loosing sleep over it, it is much more satisfying to just go on and calculate c_2
- There have been a few efforts to develop alternative methods
  - Have their own limitations, none has gotten much traction

Typical Pitfalls in Practice.





- Prone to various pitfalls
  - Underestimation, fake asymmetry, extreme case: one-sidedness
  - Bound to happen somewhere in a spectrum
- Minimal fix: Take maximal absolute deviation as symmetric uncertainty
  - Unfortunately often not done, instead scale variations get silently interpreted as “uncertainties”
  - There are thousands of theory papers quoting asymmetric (or even one-sided) “scale uncertainties”

Conceptual Limitations.

Even if with sufficient care scale variations give a reasonable size for $\Delta\sigma$, they have a much bigger conceptual problem/limitation

- Scales are *not* physical parameters with a true but uncertain value
  - The choice of μ is not the actual source of uncertainty
  - Varying μ is not a propagation of its uncertainty
  - At higher orders, μ does not become better known, rather truncated series becomes less dependent on it
  - At any given order, there might be no (sensible) value of μ at all that captures the true result

- They cannot be used to capture or derive correlations between σ_i
  - Scales used for different σ_i a priori have nothing to do with each other

$\Rightarrow$ Best we can do is come up with theoretically motivated (but still more-or-less ad hoc) model for correlations between σ_i

Implementation Options.

✗✗ Worst: Treat μ itself as nuisance parameter

- Usually out of desire to capture correlations or shape uncertainty
- All of the above pitfalls apply

✗ Less bad: $\sigma_i \pm \theta_i \Delta\sigma_i$

- with $\Delta\sigma_i$ from max-abs envelope (at least avoids pitfalls)
- Treat θ_i as nuisance parameters
- Main issue: Missing proper correlations

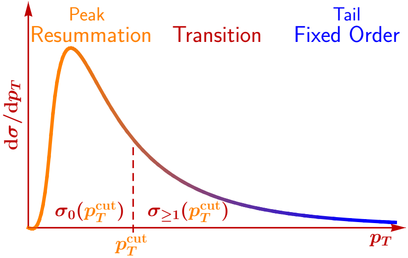
(✓) Least bad: $\sigma_i \pm \theta_a \Delta_{ia} \pm \theta_b \Delta_{ib} \pm \cdots$

- As above, but including some theoretically motivated correlation model
- Try to identify and separate independent uncertainty “sources” $a, b, \ldots$
- Δ_{ia} estimated from (max-abs envelope of) suitably chosen scale variations
- $\theta_{a,b,\ldots}$ are mutually independent nuisance parameters

Example: Correlation Model in Jet Binning.

[Stewart, FT, arXiv:1107.2117]

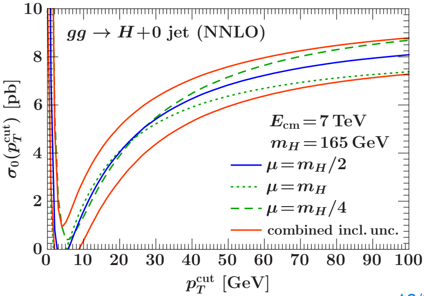
$$\sigma_{\text{tot}} = \underbrace{\int_0^{p_T^{\text{cut}}} \mathrm{d}p_T \frac{\mathrm{d}\sigma}{\mathrm{d}p_T}}_{\sigma_0(p_T^{\text{cut}})} + \underbrace{\int_{p_T^{\text{cut}}}^{\infty} \mathrm{d}p_T \frac{\mathrm{d}\sigma}{\mathrm{d}p_T}}_{\sigma_{\geq 1}(p_T^{\text{cut}})}$$



- Naive scale variation fails
- Instead, parametrize in terms of
  - yield: overall normalization
  - migration: induced by binning cut

	σ_0	$\sigma_{\geq 1}$	σ_{tot}
θ_{y}	$\Delta_{0\text{y}}$	$\Delta_{1\text{y}}$	$\Delta_{0\text{y}} + \Delta_{1\text{y}}$
θ_{cut}	Δ_{cut}	$-\Delta_{\text{cut}}$	0

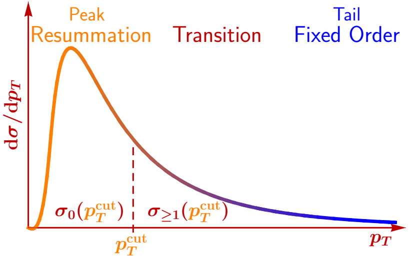
  - $\Delta_{i\text{y}}$ and Δ_{cut} can be estimated at FO or via resummation



Example: Correlation Model in Jet Binning.

[Stewart, FT, arXiv:1107.2117]

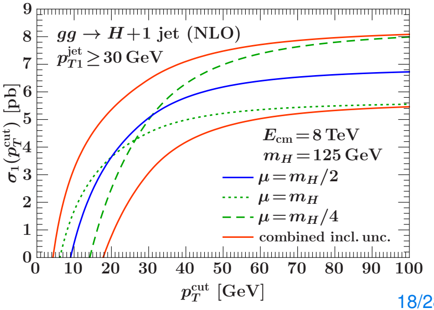
$$\sigma_{\rm tot} = \underbrace{\int_0^{p_T^{\rm cut}} \mathrm{d}p_T \frac{\mathrm{d}\sigma}{\mathrm{d}p_T}}_{\sigma_0(p_T^{\rm cut})} + \underbrace{\int_{p_T^{\rm cut}}^{\infty} \mathrm{d}p_T \frac{\mathrm{d}\sigma}{\mathrm{d}p_T}}_{\sigma_{\geq 1}(p_T^{\rm cut})}$$



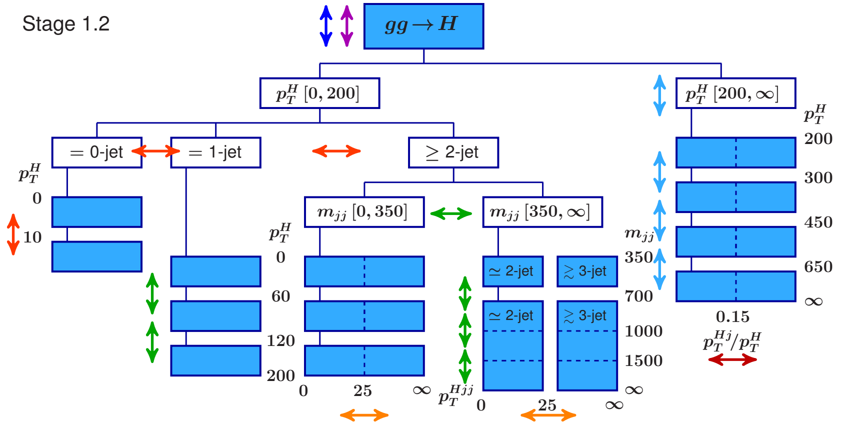
- Naive scale variation fails
- Instead, parametrize in terms of
  - yield: overall normalization
  - migration: induced by binning cut

	σ_0	$\sigma_{\geq 1}$	$\sigma_{\rm tot}$
$\theta_{\rm y}$	$\Delta_{0\rm y}$	$\Delta_{1\rm y}$	$\Delta_{0\rm y} + \Delta_{1\rm y}$
$\theta_{\rm cut}$	$\Delta_{\rm cut}$	$-\Delta_{\rm cut}$	0

  - $\Delta_{i\rm y}$ and $\Delta_{\rm cut}$ can be estimated at FO or via resummation



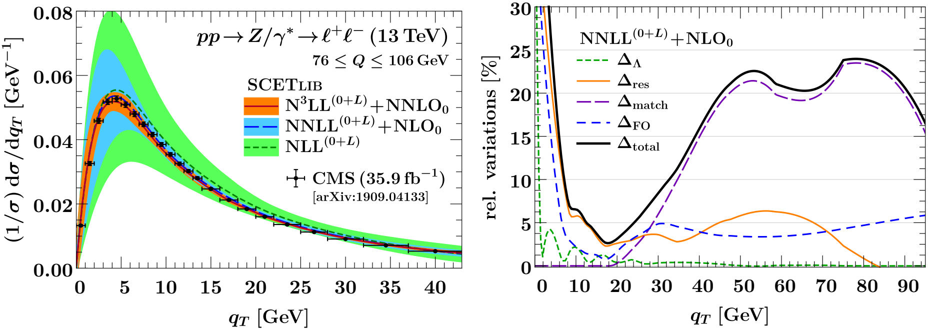
Example: STXS Uncertainty Scheme for $gg \to H$.



- Parametrize in terms of migration unc. across various bin boundaries
- Becomes more and more arbitrary with more bins
  - How to separate Δ_{cut} for given boundary among subbins
  - Which bin boundaries to consider independent vs. correlated
  - Danger of overestimation/double-counting with too many small bins

Example: Scale Variations for Z p_T Spectrum.

[Ebert, Michel, Stewart, FT, arXiv:2006.11382]

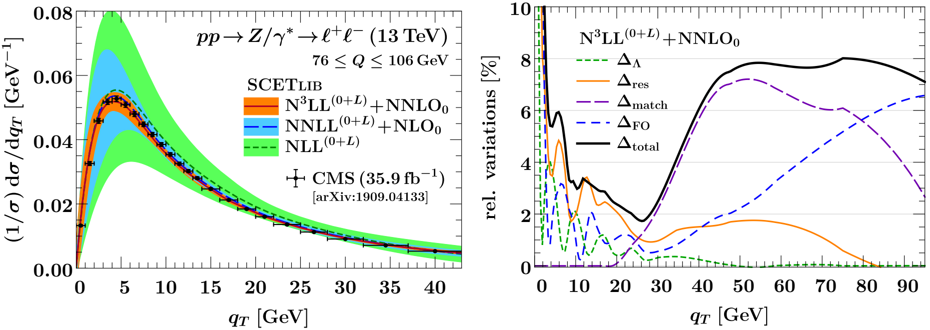


Define independent, and thus uncorrelated, sources of pert. uncertainties

- Estimate each from max-abs envelope of suitably chosen scale variations
  - In this context (resummation), we have up to six different scales to play with
- Added in quadrature to get total uncertainty band
  - ✓ Achieves desired decorrelation across spectrum
  - ✓ Lower-order bands cover best central value
  - ✗ Anticorrelations (shape) within each source still not captured

Example: Scale Variations for Z p_T Spectrum.

[Ebert, Michel, Stewart, FT, arXiv:2006.11382]



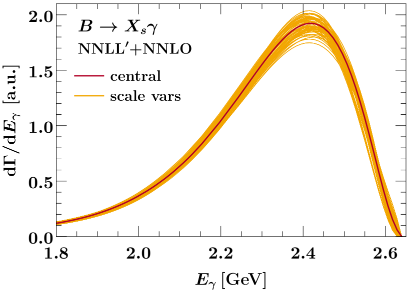
Define independent, and thus uncorrelated, sources of pert. uncertainties

- Estimate each from max-abs envelope of suitably chosen scale variations
  - In this context (resummation), we have up to six different scales to play with
- Added in quadrature to get total uncertainty band
  - ✓ Achieves desired decorrelation across spectrum
  - ✓ Lower-order bands cover best central value
  - ✗ Anticorrelations (shape) within each source still not captured

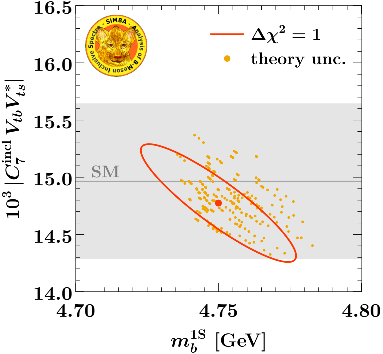
Another Implementation: Envelope Propagation.

[SIMBA, Bernlochner et al., arXiv:2007.04320]

Repeat fit with varied theory inputs from various scale choices



⇒



- Propagates the envelope, which can be useful in particular for spectra
  - ✓ Maintains behaviour of individual scale variations (i.e. some form of anticorrelated shape uncertainty)
  - ✓ Avoids overestimate from only taking edges of uncertainty band
  - ✗ Theory uncertainties cannot affect central fit result
  - ✗ Correlations could still be rather arbitrary

⇒ How to actually take and interpret envelope in fit results?

Theory Nuisance Parameters.

(The promise of a less ugly future)

What We Should be Doing.

$$\sigma = c_0 + \alpha(\mu_0)\, c_1 + \alpha^2(\mu_0)\, c_2 + \cdots$$

1) Identify the actual source of uncertainty

- The unknown, neglected higher-order terms: $c_2, c_3, \ldots$

What We Should be Doing.

$$\sigma = c_0 + \alpha(\mu_0)\, c_1 + \alpha^2(\mu_0)\, c_2 + \cdots$$

1) Identify the actual source of uncertainty

- The unknown, neglected higher-order terms: $c_2, c_3, \ldots$

2) Identify the knowns and parametrize the unknowns

- We typically know a lot about the general structure of c_2 even without explicitly calculating it
  - μ dependence, color structure, partonic channels, kinematic structure, ...
  - All we want is an uncertainty estimate, so it is sufficient to consider dominant contributions or limits
- Suitably parametrize the remaining unknown pieces/contributions
  - Best case: Unknowns are a few numbers
  - More generally, one or more unknown functions

What We Should be Doing.

$$\sigma = c_0 + \alpha(\mu_0)\, c_1 + \alpha^2(\mu_0)\, c_2 + \cdots$$

1) Identify the actual source of uncertainty

- The unknown, neglected higher-order terms: $c_2, c_3, \ldots$

2) Identify the knowns and parametrize the unknowns

- We typically know a lot about the general structure of c_2 even without explicitly calculating it
  - μ dependence, color structure, partonic channels, kinematic structure, ...
  - All we want is an uncertainty estimate, so it is sufficient to consider dominant contributions or limits
- Suitably parametrize the remaining unknown pieces/contributions
  - Best case: Unknowns are a few numbers
  - More generally, one or more unknown functions

3) Treat the remaining unknowns as nuisance parameters

- Figure out allowed range based on theory arguments

Advantages of Theory Nuisance Parameters.

Theory nuisance parameters (TNPs) are genuine parameters with a true but uncertain value

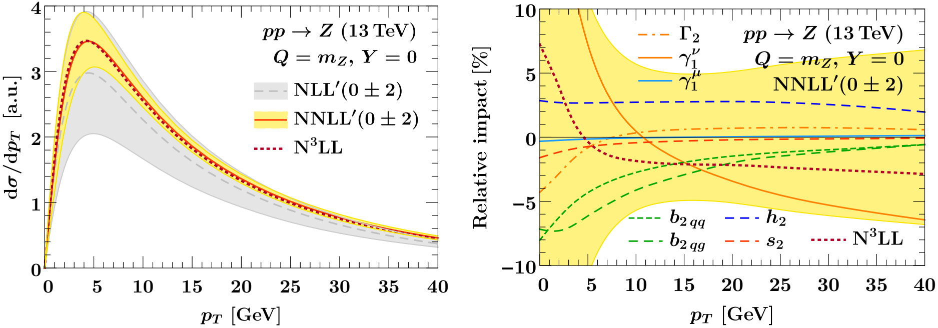
- Renders the whole problem much more well-defined
- We get all benefits of truly parametric uncertainties
  - ✓ Encode correct correlations, straightforward to propagate everywhere
  - ✓ Can be constrained by measurements (auxiliary and/or primary)
- There will typically be several parameters
  - Much safer against accidental underestimate of any one parameter
  - Total theory uncertainty becomes Gaussian due to central-limit theorem
- Can even lead to reduced theory uncertainties
  - Can fully exploit partially known higher-order information
  - Can also reduce theory uncertainties at a later time

Price to pay

- Predictions become quite a bit more complex
  - Need to implement complete next order in terms of unknown parameters

Example: $Z\ p_T$ Spectrum.

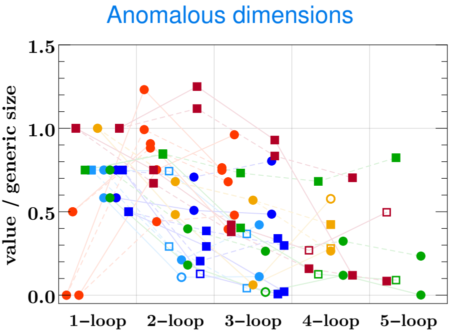
[FT, work in progress ...]

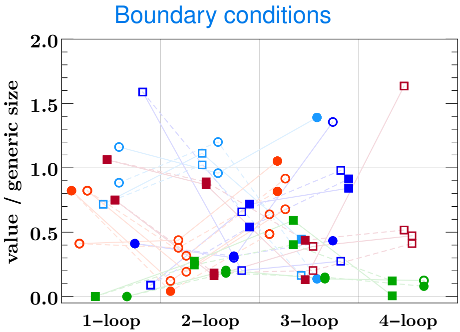


- Dependence on p_T is determined by resummation (RG structure)
  - ✓ Correlations in p_T spectrum are fully captured/predicted
  - ✓ Similarly, correlations in predictions for different Q, E_{cm}, processes
- Underlying TNPs are anomalous dimensions and boundary conditions required at each resummation order
  - ► Illustration: Show $\theta_i = (0 \pm 2)\theta_i^{\mathrm{true}}$ with known θ_i^{true} at this order

Estimating Allowed Size of TNPs.

[FT, work in progress ...]





In practice, one still to estimate the possible generic size of the TNPs

- Just the usual exercise for estimating possible size of a systematic
- Possible based on what we know about structure of perturbation theory
  - Illustration: Estimate based on leading color and n_f dependence
  - Works very well for many known perturbative series
  - → See backup for example of functional TNP

Summary.

Theory uncertainties are indeed ugly business

- Be aware of limitations of current methods like scale variations
  - Not particularly reliable, cannot predict correlations
  - → See backup for “Herwig vs.Pythia”
- Best way is to avoid theory uncertainties
  - Yes, but “avoiding” often secretly means “canceling” them, which relies on correlations, and we’re back to the previous point

Theory nuisance parameters can overcome these limitations

- A paradigm change, but the obvious way forward (at least to me)
- Just at the start, many things still to investigate, gathering experience, ...
  - Your feedback is most welcome ...

Summary.

Theory uncertainties are indeed ugly business

- Be aware of limitations of current methods like scale variations
  - Not particularly reliable, cannot predict correlations
  - → See backup for “Herwig vs.Pythia”
- Best way is to avoid theory uncertainties
  - Yes, but “avoiding” often secretly means “canceling” them, which relies on correlations, and we’re back to the previous point

Theory nuisance parameters can overcome these limitations

- A paradigm change, but the obvious way forward (at least to me)
- Just at the start, many things still to investigate, gathering experience, ...
  - Your feedback is most welcome ...

Thanks for your attention

(and apologies for running over time)

Additional Slides

2-Point Systematics: “Herwig vs. Pythia”.

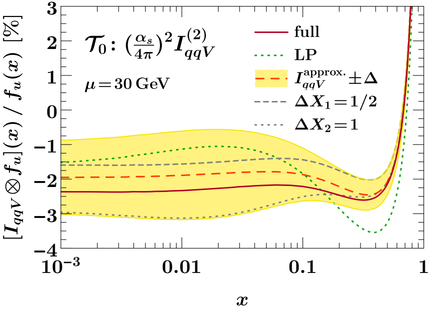
Take difference of two predictions as the uncertainty

- Usually done out of desperation for lack of anything better
- If the two are close: does not mean actual uncertainty is small
  - They might just be doing the same (possibly wrong) thing.
  - Completely underestimates
- If the two are very different: does not mean actual uncertainty is large
  - One might just be wrong or not as good as the other
  - Might just be comparing apples with bananas (not even oranges)
  - Completely overestimates
- If both can be considered equally good approximations: Treat just like a scale variation
  - Difference between two approximations may (or may not) give an estimate of size of neglected terms
  - All caveats/pitfalls of scale variations apply

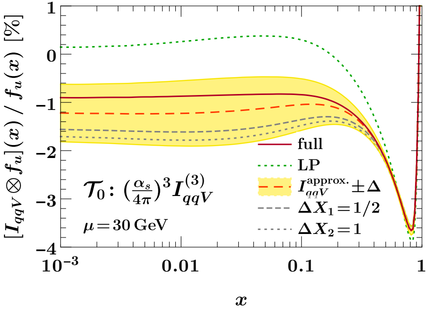
Functional TNPs.

- Strategy: Parametrize by exploiting known functional dependence and/or expanding in known limits
- Example: Beam function matching coefficients depend on parton momentum fraction x (similar to splitting functions)
  - Can construct a parametrization based on expanding around $x \to 1$
    [Billis, Ebert, Michel, FT, arXiv:1909.00811]

NNLO (full was known)



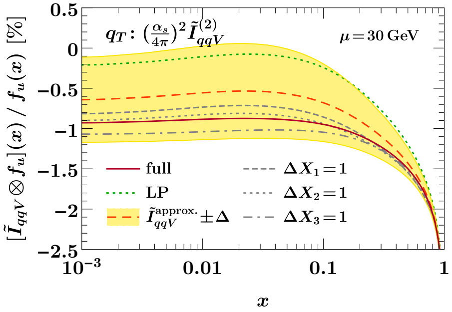
N^{3}LO (full was not yet known)



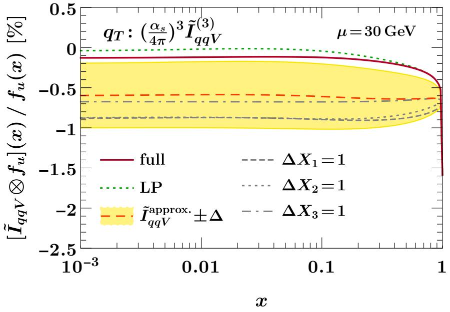
Functional TNPs.

- Strategy: Parametrize by exploiting known functional dependence and/or expanding in known limits
- Example: Beam function matching coefficients depend on parton momentum fraction x (similar to splitting functions)
  - Can construct a parametrization based on expanding around $x \to 1$
    [Billis, Ebert, Michel, FT, arXiv:1909.00811]

NNLO (full was known)



N^{3}LO (full was not yet known)



Acknowledgments.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101002090 COLORFREE)

erc

European Research Council

Established by the European Commission