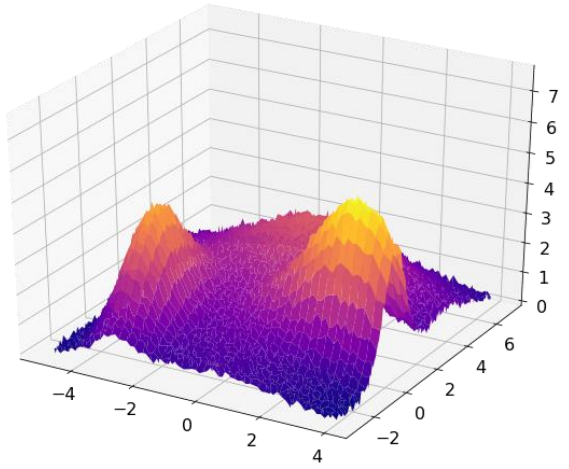


LGADUtils Update: Fit infrastructure Optimization

V. Gkougkousis

Institut de Física d'Altes Energies

FOR INTERNAL REVIEW ONLY – DO NOT SHARE

LHCb Group Meeting – 9 / 9 / 2020

•Fits infrastructure

Available fitting options

File: LGADFits.cxx

- Root multi-iterative automatic fitting for:

I. Gauss **II. Gauss X Landau**

```
int IterativeFit (std::vector<double> *w, std::pair<double, double> &gmean, std::pair<double, double> &gsigma,
                  TH1D* &FitHist, double &minchi2, std::string methode = "Gauss", std::pair<int, int> points =
                  std::make_pair(-1, -1))
```

- Unpinned 2-dimentional Linear fitting through RooFit and Minuit:

```
int LinearFit (std::vector<double>* vec, std::pair<double, double> &slope, std::pair<double, double> &intersept,
               std::vector<double>* vecErr = NULL);
```

- Roofit Convolution fitting (no iterative readjustment) for:

I. Gauss X Landau **II. Gauss X Linear**

```
int  RooConvFit  (std::vector<double>*  vec,  std::pair<double,  double>  &magMPV,  std::pair<double,  double>
                 &magSigma, std::string conv);
```

- Tow point linear interpolation:

```
double LinearInter(double x1, double y1, double x2, double y2, double y3);
```

- Fast Furrier transform algorithm: `double FFT(std::vector<double> *w, Long64_t snrate, int start, int stop);`

•Average pulse shape (RooFit)

Starting point, Non-optimized RooFit Convolution

I. Average calculated from 100 events

II. Each waveform is time aligned at 20% CFD

III. For all events, the same point of each waveform projected in TH1F
- ✓ as many TH1F as points in waveform
- ✓ each with as many entries as events (100 here)

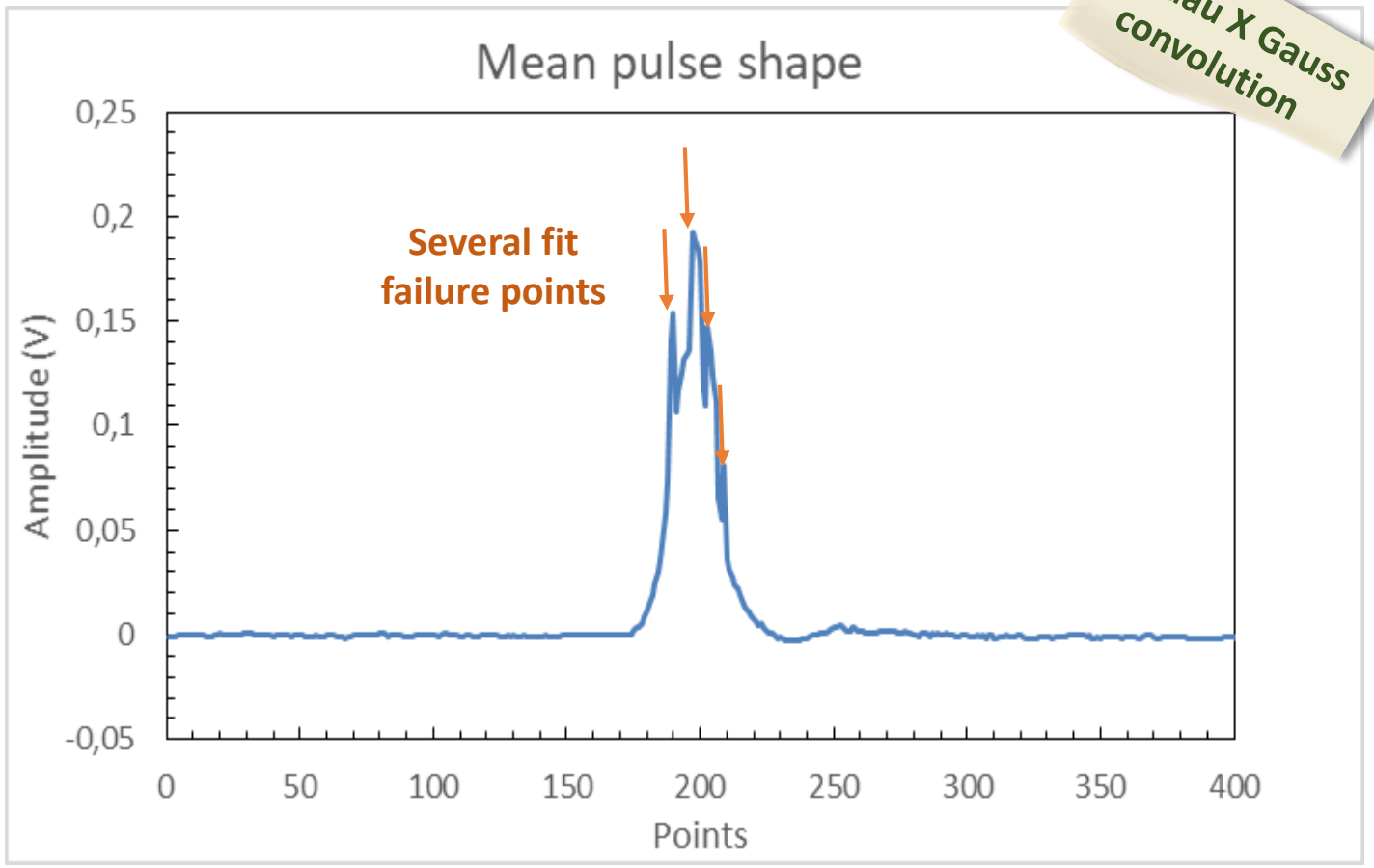
IV. Each TH1F fitted with a Landau X Gauss distribution

V. MPV, sigma and uncertainty extracted

VI. Fitting performed in RooFit using RooFit Convolution and Minuit

VII. No starting parameters or optimization

VIII. Plot the MPVs of each point in a single waveform



•Average pulse shape (RooFit)

Initial RooFit Optimization

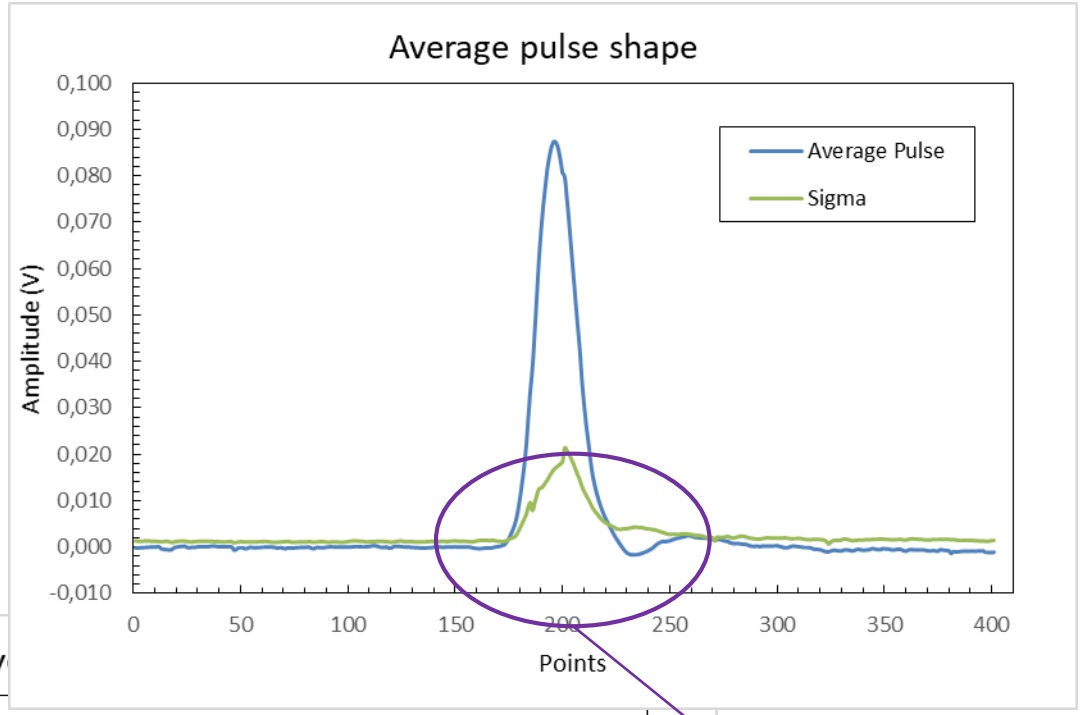
I. Fit Binning:
- ✓ Symmetric fit limits: $\mathbf{x_{av.} \pm (5 \times \sigma)}$
- ✓ Bin width defined as: $\boldsymbol{\sigma /3}$
- ✓ High statistics re-optimization, increase no. of bins if: $\mathbf{N_{points} > 1.5 \times N_{bins}}$

II. Parameter constraints:
- ✓ Asymmetric limits for Landau MVP: $\mathbf{-4 \times |x_{RMS}| < x_{fit} < 2 \times |x_{RMS}|}$
- ✓ Asymmetric limits for Landau sigma: $\mathbf{-0.001 \times \sigma < \sigma_{fit} < 2 \times \sigma}$
- ✓ Symmetric limits for Gauss mean: $\mathbf{-2 \times \sigma < \sigma_{fit} < 2 \times \sigma}$

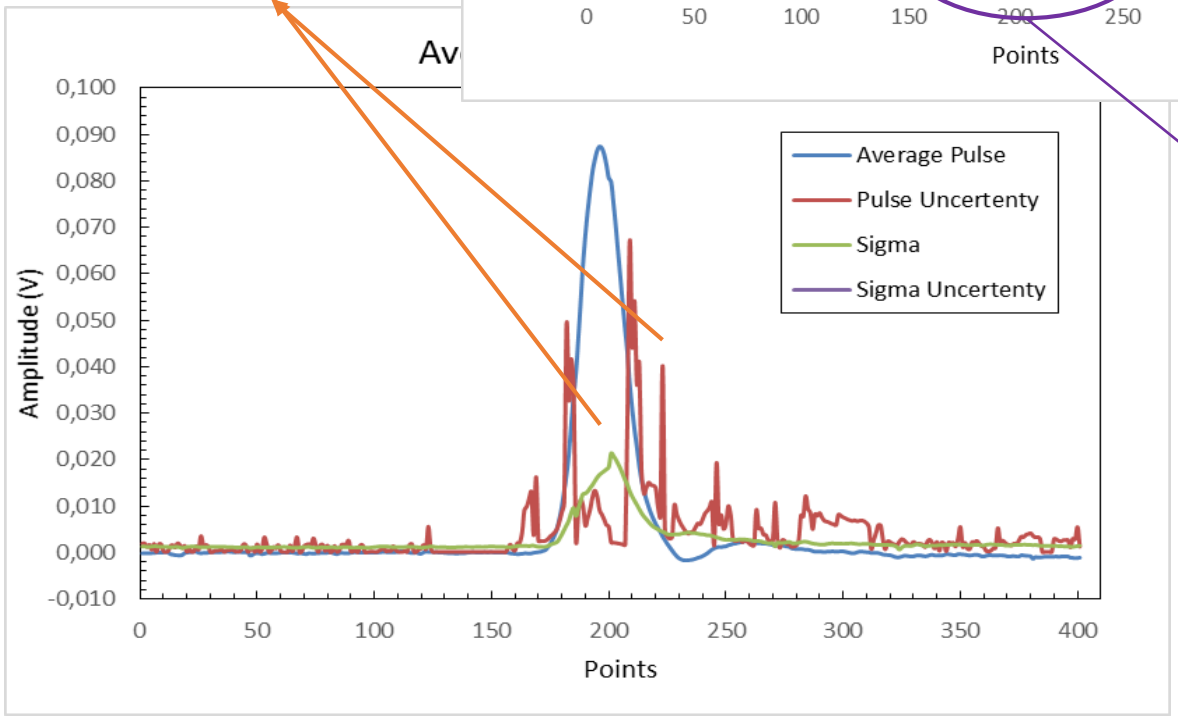
III. Parameter Initial values:
- ✓ Landau $\mathbf{MPV = x_{RMS}}$
- ✓ Gauss **sigma = 0**

IV. No re-iteration implementation

Smooth pulse shape



High uncertainties at the beginning and end of pulse

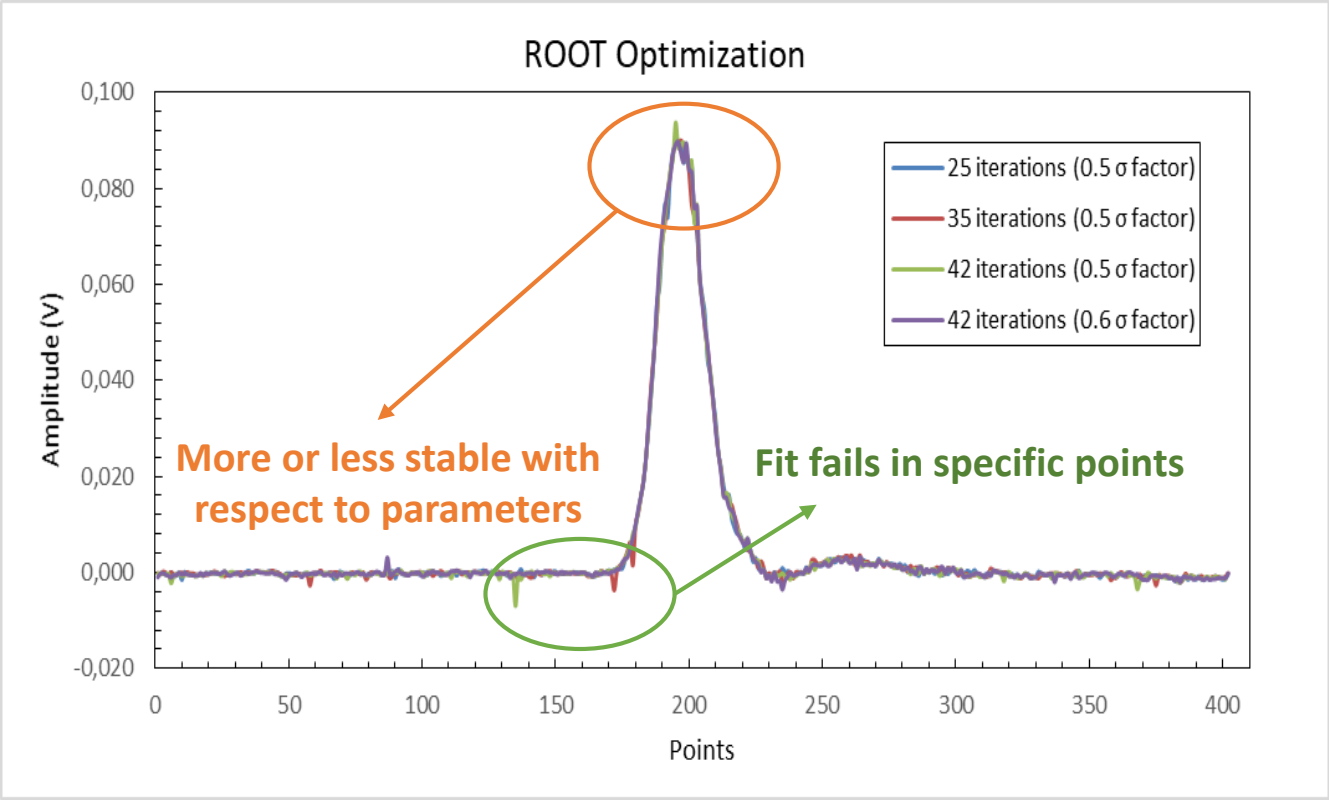


Reasonable sigma with respect to MPV, no outliers

•Average pulse shape (Root)

Root Optimization (no RooFit)

- Constraint parameter values but not fixed
- Manually defined convolution function
- 1000 convolution steps
- Select the fit with the best agreement (minimization of $|1-x^2/NDF|$)



I. Iterative approach

II. Fit Binning:

✓ Asymmetric fit limits:

lower: $\mathbf{x_{min} - ((n-5)/10) \times \sigma}$

upper: $\mathbf{x_{max} - ((n-5)/2) \times \sigma}$

where $7 < n < 2$ → **6 cases**

✓ Bin number defined at least as:

$\mathbf{N_{bins} = \sqrt{N_{events}}}$

augmented by:

$\mathbf{(x_{max} - x_{min})/(0.5 \times \sigma \times \alpha/4)}$

where $14 < \alpha < 7$ → **7 cases**

X

42 total iterations

III. Parameter constraints:

✓ Asymmetric limits for Landau MVP:

$\mathbf{x_{RMS} - 3\times\sigma < x_{RMS} < x_{RMS} + 3\times\sigma}$

✓ Function integral limits set to:

0.1×Int. < X < 10×Int.

IV. Initial values:

✓ Landau **MPV** = $\mathbf{x_{RMS}}$

✓ Convolution **Integral = distribution integral**

✓ Convoluted sigma set to : **σ / 4**

•Average pulse shape (Root)

Iteration Optimization

I. Number of iterations is reduced by varying the limits of **n** and **α** parameters

II. Limits are modified symmetrically

III. Limits are modified in both of them and in each one separately

IV. From each family of fits the best one is selected based on $|1-x^2/NDF|$ minimization

V. Relative uncertainties for each point are plotted for selected fit

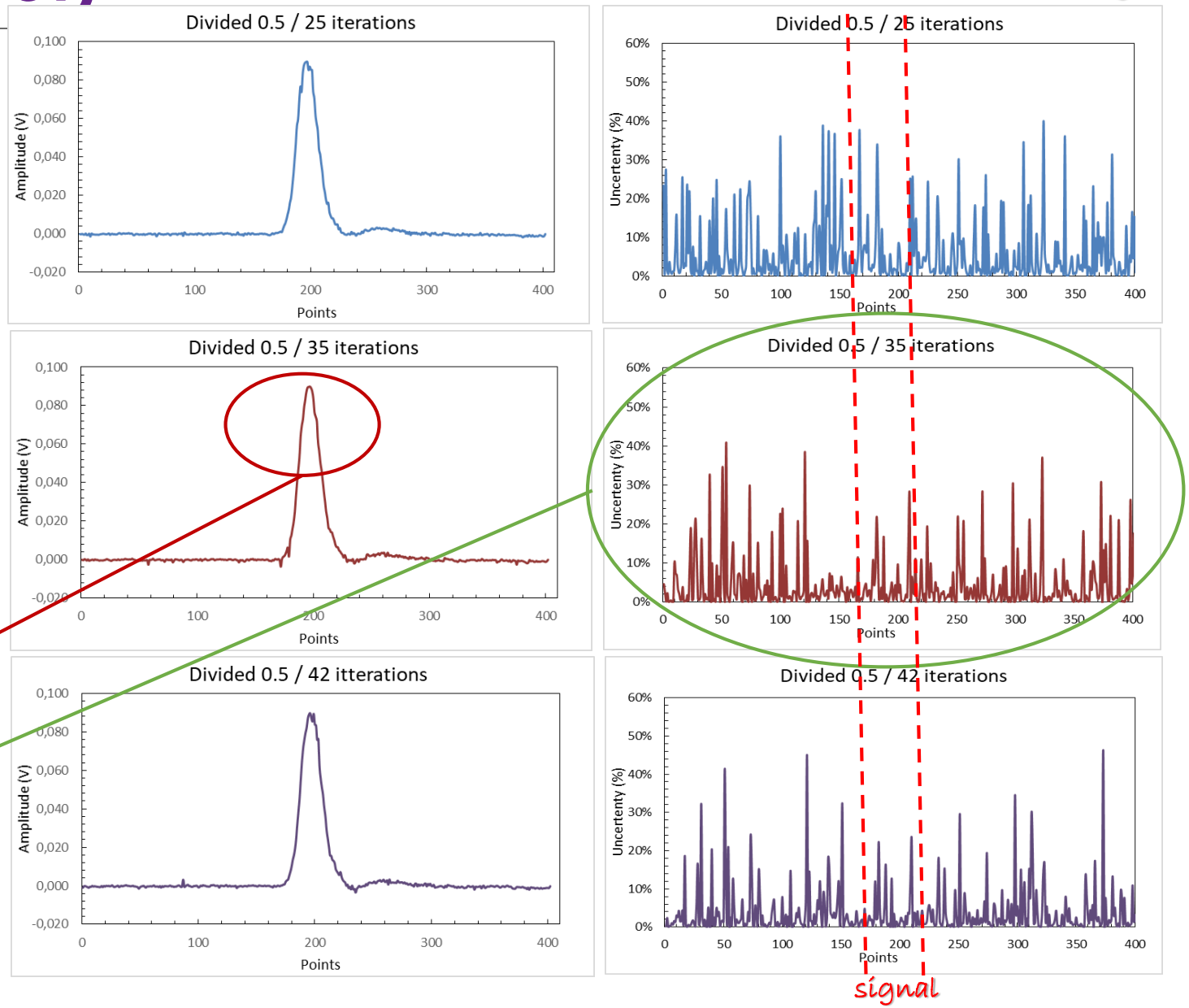
More is not always better

Most stable case across the board, closer to RooFit result

No outliers over 40% error, small relative error within the signal

Average pulse shape

Relative point uncertainty



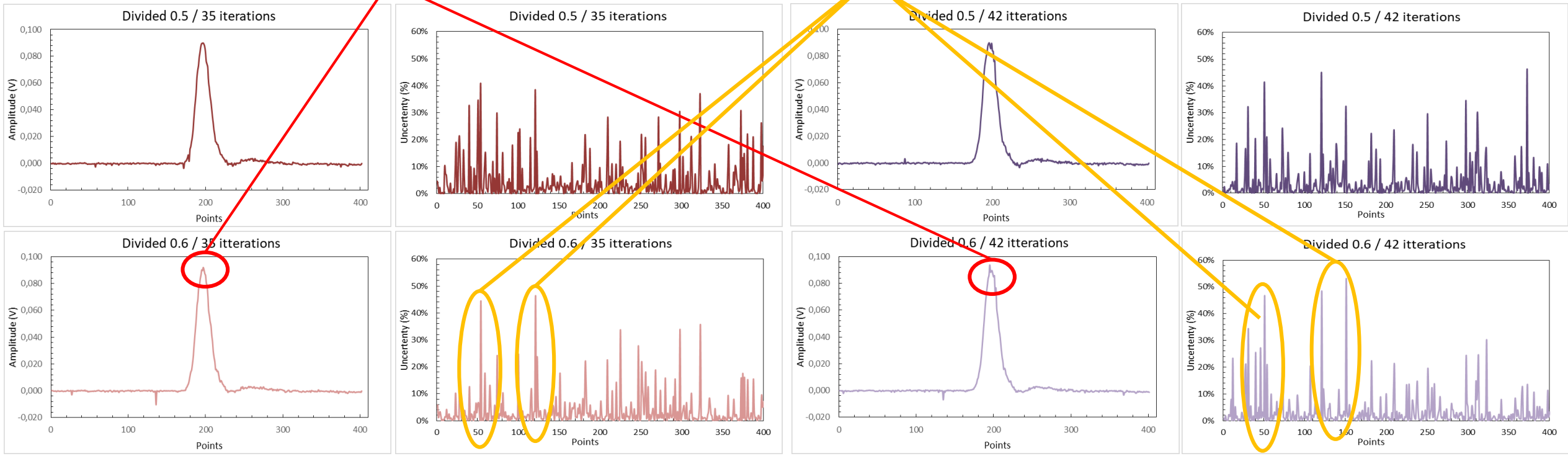
•Average pulse shape (Root)

Bin size optimization

I. Bins size adjusted by modifying the multiplication factor on bin number denominator
II. Tests are performed for 35 and 42 iterations
III. The effect of bin size variation seems to be constant in both 42 / 35 iteration cases
IV. Smaller bin size probes point fluctuations

Multi peak shape, sensitive to fluctuations

Substantial increase in single point uncertainty



•Average pulse shape (Root)

Dynamic iteration

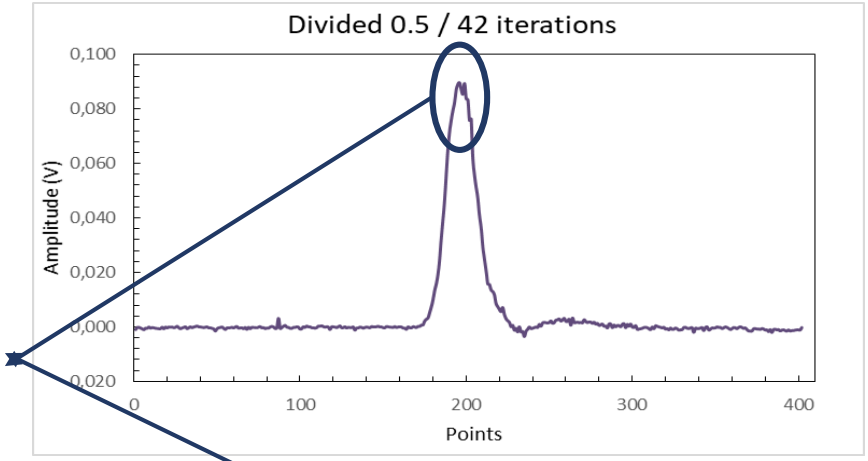
I. Instead of performing all iterative steps for each fit, stop when result is at least 97% good

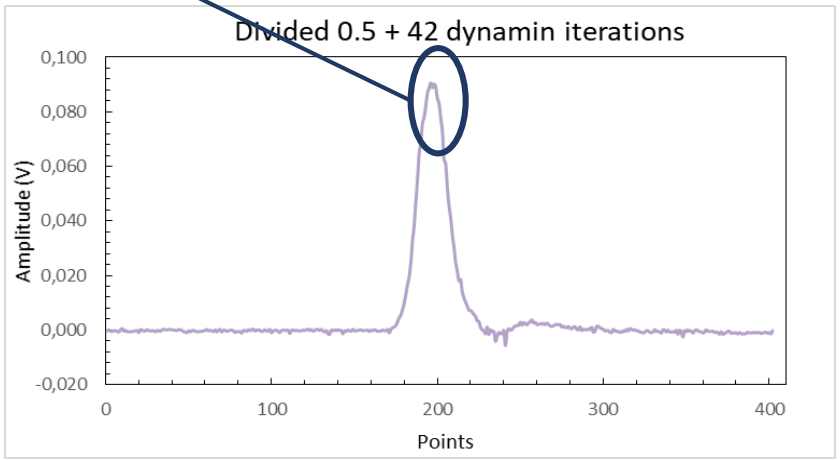
II. If a 97% not reached, perform all iterations and choose best result

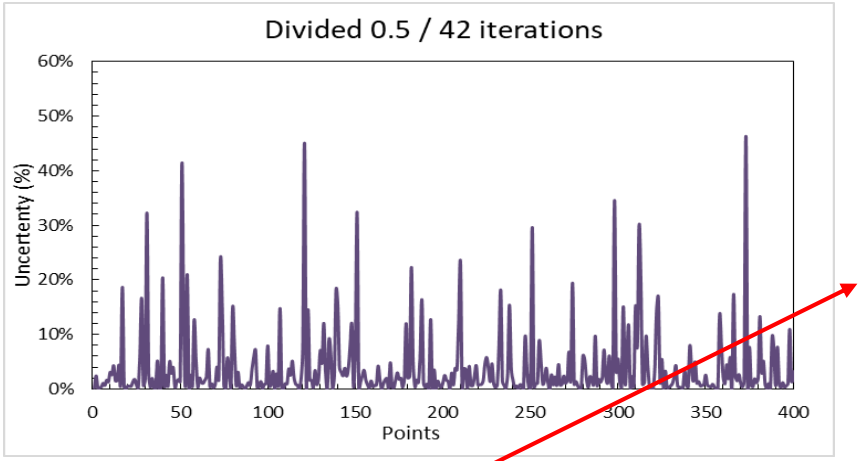
III. Goodness of the fit based on $|\mathbf{1-x^2/NDF}|$ minimization

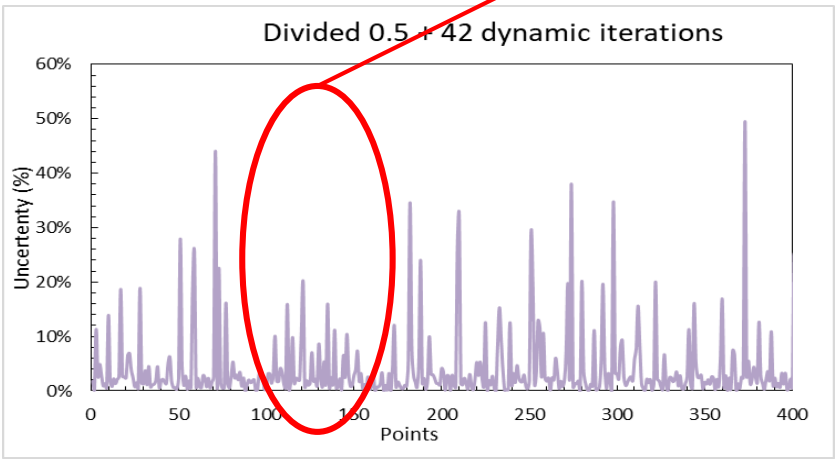
IV. Significantly reduce processing time (~50%)

More stable fit at the top, effect of binning fluctuations
As iterations progress, bin size becomes smaller and more sensitive to point fluctuations









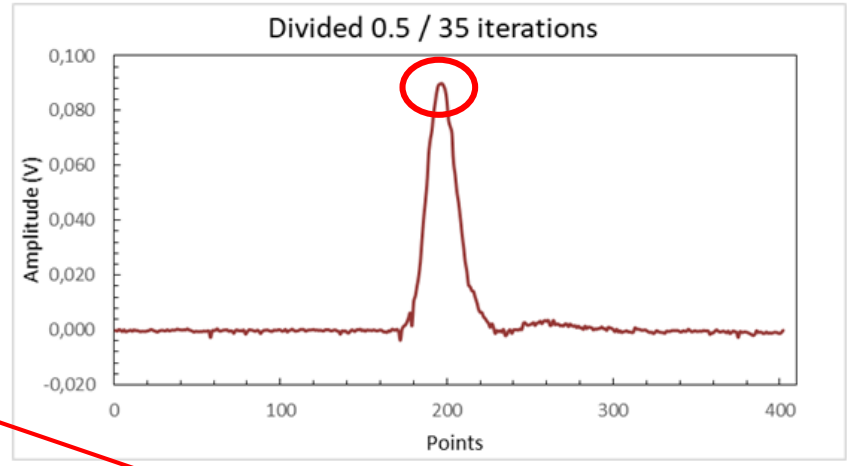
Lower uncertainties in the signal region

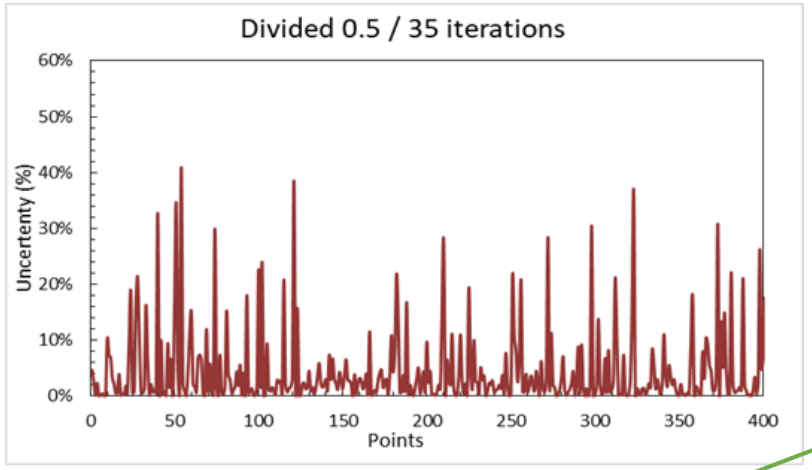
•Average pulse shape (Root)

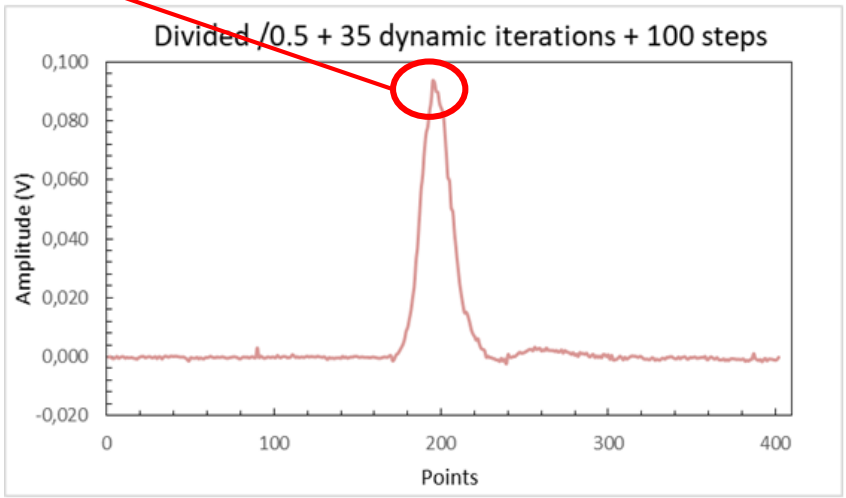
Convolution steps optimization

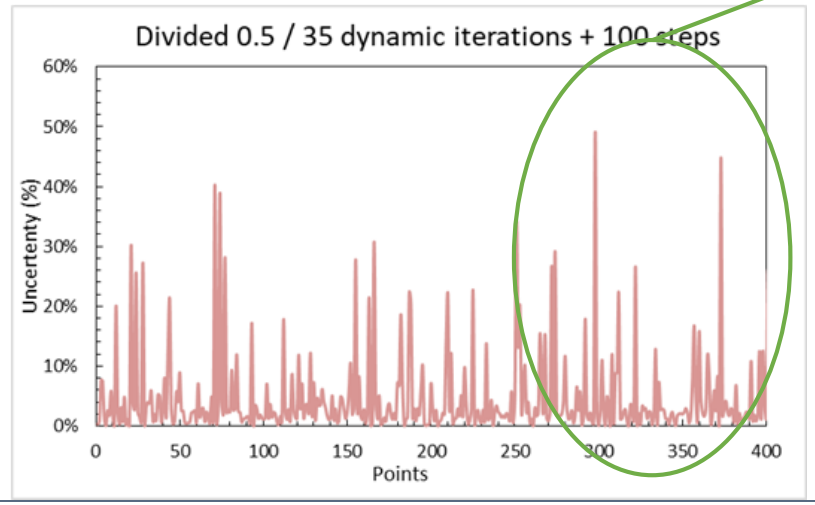
I. Initial configuration calls for 10k convolution steps

II. Study if this can be reduced to 1k for speed









Sharper peak, sensitive to extreme values

Higher relative uncertainty

•New Re-Binning method

Dataset Type	Statistics Case	Bin number array (7 cases)		
		Lower 3 bin number variations	Optimum Bin number	Higher 3 bin number variations
Discrete Datasets	$\frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} < \sqrt{N_{elements}} < N_{bins\_max}$	$\left\lceil \sqrt{N_{elements}} \right\rceil - n \times \left\lceil \frac{\sqrt{N_{elements}} - \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma}}{3} \right\rceil$ with $1 < n < 3$	$\left\lceil \sqrt{N_{elements}} \right\rceil$	$\left\lceil \left\lceil \sqrt{N_{elements}} \right\rceil + n \times \left\lfloor \frac{N_{bins\_max} - \sqrt{N_{elements}}}{3} \right\rfloor \right\rceil$ with $1 < n < 3$ **
	$\sqrt{N_{elements}} \leq \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} < N_{bins\_max}$	Lowest bin number	Rest of the bin number array	
	$\sqrt{N_{elements}} \leq N_{bins\_max} < \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma}$	$\left\lfloor \sqrt{N_{elements}} \right\rfloor$	$\left\lfloor \sqrt{N_{elements}} \right\rfloor + n \times \left\lfloor \left\lvert N_{bins\_max} - \sqrt{N_{elements}} \right\rvert / 6 \right\rfloor$ with $1 < n < 6$ **	
	$N_{bins\_max} \leq \sqrt{N_{elements}}$	$n \times \lfloor N_{bins\_max}/7 \rfloor$ with $1 < n < 7$		
Non – Discrete Datasets	$\frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} < \sqrt{N_{elements}}$	Lower 3 bin number variations	Optimum Bin number	Higher 3 bin number variations
		$\left\lceil \sqrt{N_{elements}} \right\rceil - n \times \left\lceil \frac{\sqrt{N_{elements}} - \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma}}{3} \right\rceil$ with $1 < n < 3$	$\left\lceil \sqrt{N_{elements}} \right\rceil$	$\left\lceil \left\lceil \sqrt{N_{elements}} \right\rceil + n \times \left\lfloor \frac{\frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma}}{3} \right\rfloor \right\rceil$ with $1 < n < 3$
	$\sqrt{N_{elements}} \leq \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma}$	$\left\lceil \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} \right\rceil - n \times \left\lceil \frac{\frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} - \sqrt{N_{elements}}}{3} \right\rceil$ with $1 < n < 3$	$\left\lceil \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} \right\rceil$	$\left\lceil \frac{\left\lvert \lim_{fit} High - \lim_{fit} Low \right\rvert}{\sigma} + n \times \left\lfloor \frac{\sqrt{N_{elements}}}{3} \right\rfloor \right\rceil$ with $1 < n < 3$

•General Updates

Latest Implementations

I. Framework remanded from `HGTDUtils` to `LGADUtils`

II. `DUTChannel` Class completed with:

- ✓ Get() functions for individual event variables
- ✓ Get() functions for fits and x^2 coefficients

III. Fitting option choises added (`m_fitopt`) using global functions:

- ✓ `HGTDUtils::GetFitMethode()`
- ✓ `HGTDUtils::SetFitMethode()`

User can choose between: `RooFit`, `RooInt`, `Root`

IV. Linear fit implementation switched to un-binned approach:

```
int HGTDUtils::LinearFit(std::vector<double>* vec, std::pair<double, double> &slope,
                         std::pair<double, double> &intersept, std::vector<double>* vecErr)
```

V. Bayesian uncertainties implementation for efficiencies and variables following binomial laws:

```
template <typename V, typename T> double HGTDUtils::BayesianErr(V *w, T value)
```

(https://indico.cern.ch/event/66256/contributions/2071577/attachments/1017176/1447814/EfficiencyErrors.pdf)

VI. Multi-level debugging infrastructure (`m_verbose`) implemented with 0, 1 and 2 options and the global functions:

- ✓ `int HGTDUtils::GetVerbosity()`
- ✓ `Void HGTDUtils::SetVerbosity()`

VII. Optional calculation of average pulse shape:

```
int DUTChannel::updateChProperties(bool waveshape)
```

default false to avoid shape calculation if not desired

Latest version:
https://gitlab.cern.ch/ifaepix/lgad-timing-analysis

•General Updates

Known issues

I. Linear interpolation fails when consecutive points are identical (user does not see this, code just rejects the event, but nevertheless should not happen)

II. In extremely low statistics (< 4 values) fit will fail while returning empty pointers. Protection implemented but depending on root version might not work

III. In iterative fitting, symmetric limit variation may cause parameter starting values to be out of range. Fixed with safeguard, but in principles should not happen for Gaussian fits (user does not see this)

IV. I have a doubt on the low voltage optimization performed for irradiated sensor waveforms, probably coed over-optimized, this needs to be re-visited by looking at results (basically baseline optimization might be too aggressive)

V. Average pulse shape calculation based on storing CFD time-realigned and baseline voltage corrected waveforms in a global vector. Approach fails in high event counts (<50k) as we hit limitation of ROOT CINT. Not crucial for testbeam, not used at the moment, but a better solution to be found.