Recovering True PDF Likelihoods By Reweighting

Ali Al Kadhim & Harrison B. Prosper

Data are assumed to be normally-distributed

$$P(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}[\mathbf{x} - g(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-\mathbf{1}} [\boldsymbol{x} - \mathbf{g}(\boldsymbol{\theta})]\right\}$$

- **And the likelihood is given by**

$$L(\boldsymbol{\theta}) = P(\boldsymbol{D}|\boldsymbol{\theta})$$

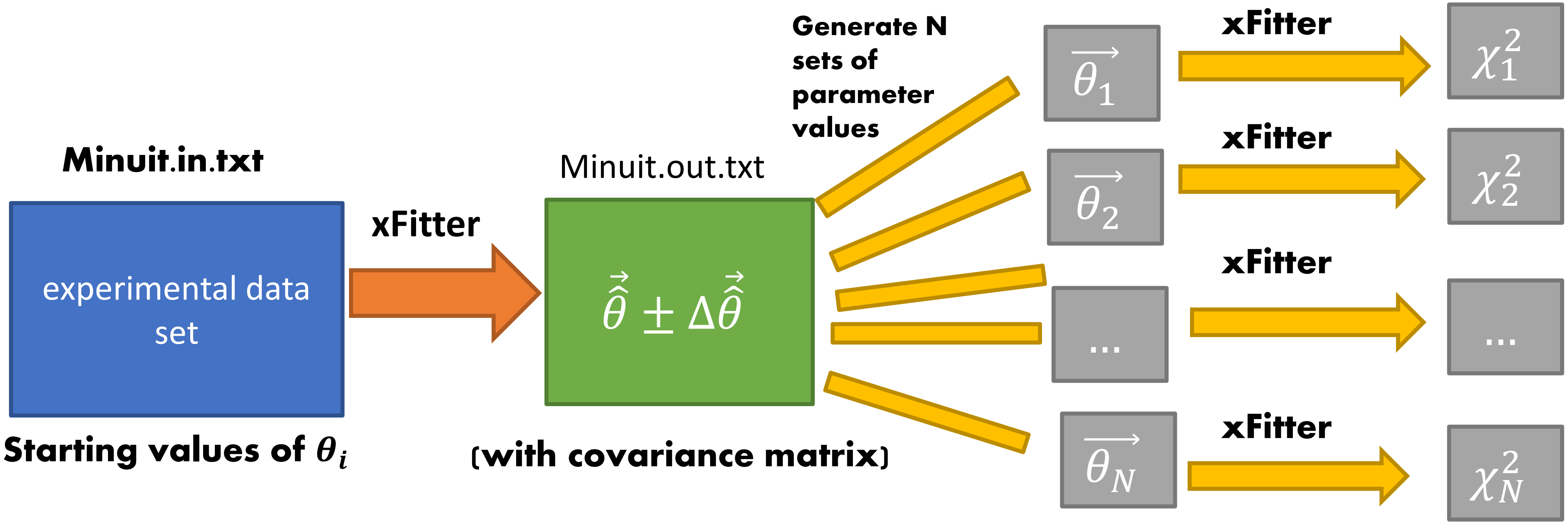
where D are the actual observations.

- **The best-fit values are found by minimizing $-2\log L(\boldsymbol{\theta}) = \chi^2$.**
- **xFitter finds $\widehat{\theta}$ (best-fit parameter values) and returns $\widehat{\Sigma}$ by solving** $\Delta\chi^2 = 1$.

- **Although the data are assumed to be normally-distributed, the likelihood function may not be a multivariate gaussian.**
- **One of the goals is to map out the true shape of $L(\boldsymbol{\theta})$.**

Procedure

With default HERAPDF parameterization: $xf(x,\mu_0^2) = Ax^B(1-x)^C[1+Dx+Ex^2] - A'x^{B'}(1-x)^{C'}$



We Generate N sets of parameter values sets according to $\boldsymbol{\theta_i} \sim \mathcal{N}(\boldsymbol{\mu_i} = \widehat{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma_i} = \widehat{\boldsymbol{\Sigma}}_i)$
Hence we approximate the likelihood as $L'(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu_i} = \widehat{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma_i} = \widehat{\boldsymbol{\Sigma}}_i)$.

Multivariate Gaussian Approximation to Likelihood

- **So we approximate the likelihood of the parameters as a Multivariate Gaussian from the best fit values.**
- $L(\boldsymbol{\theta})$: **true likelihood,** $L'(\boldsymbol{\theta})$: **approximate likelihood.**

$$L'(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}\right) = \frac{\mathbf{1}}{\sqrt{(\mathbf{2}\boldsymbol{\pi})^{d}\left|\hat{\Sigma}\right|}} \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\right]^T \hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\left[\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\right]\right\}$$

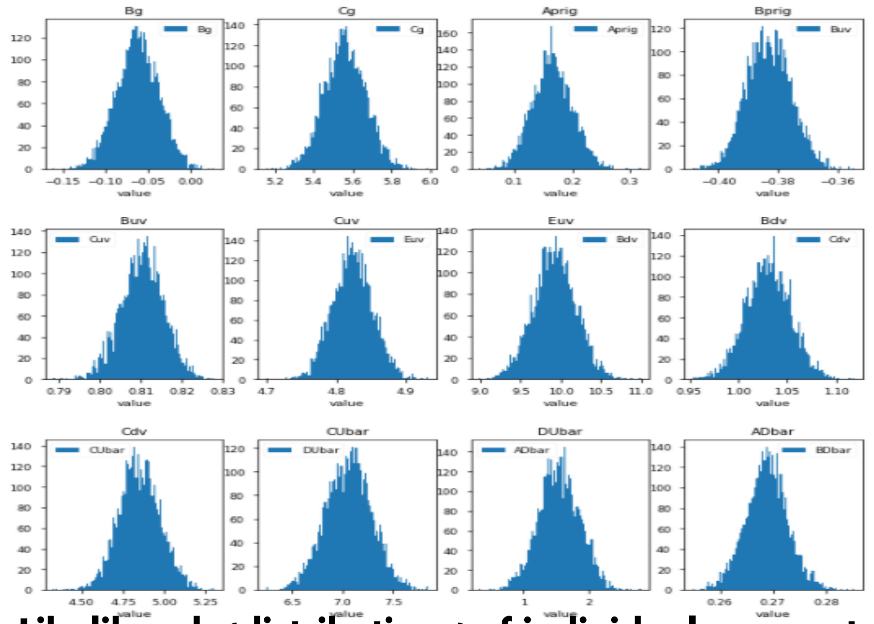
Data: HERA I & II + ZEUS combined

HERAPDF parameterization :

$$xf(x, \mu_0^2) = Ax^B(1-x)^C(1 + Dx + Ex^2) - A'x^{B'}(1-x)^{C'}$$

Parameter	xFitter Name	Starting Value	Step Size	Best-Fit Value	Approximate Error
B_g	Bg	-0.061953	0.027133	-00.61856	0.25134E-01
C_g	Cg	5.562367	0.318464	5.5593	0.10838
A'_g	Aprig	0.166118	0.028009	0.16618	0.34574E-01
B'_g	Bprig	-0.383100	0.009784	-0.38300	0.76253E-02
B_{u_v}	Buv	0.810476	0.016017	0.81056	0.53604E-02
C_{u_v}	Cuv	4.823512	0.063844	4.8239	0.29342E-01
E_{u_v}	Euv	9.921366	0.835891	9.9226	0.27481
B_{d_v}	Bdv	1.029995	0.061123	1.0301	0.23240E-01
C_{d_v}	Cdv	4.846279	0.295439	4.8456	0.12584
$C_{\bar{U}}$	CUbar	7.059694	0.809144	7.0603	0.22306
$D_{\bar{U}}$	DUbar	1.548098	1.096540	1.5439	0.31340
$A_{\bar{D}}$	ADbar	0.268798	0.008020	0.26877	0.39536E-02
$B_{\bar{D}}$	BDbar	-0.127297	0.003628	-0.12732	0.17428E-02
$C_{\bar{D}}$	CDbar	9.586246	1.448861	9.5810	0.60834

Could be more complicated for different flavors



Likelihoods (distributions) of individual parameters

Reweighting

- **In order to correct the multivariate gaussian so that we have the true likelihood, we then weight each parameter point by the weight:**

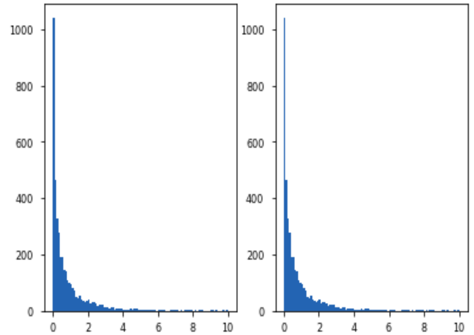
$$w_k^i = \frac{L(\boldsymbol{\theta})}{L'(\boldsymbol{\theta})}$$

- **If the weights alter the shape of the distributions in any way, we have non-gaussian distributions.**
- **In the limit of infinite number of sampled points, the weights will recover the true shapes of the likelihoods.**
- **We could explore different forms of $L'(\boldsymbol{\theta})$ (see backup.)**

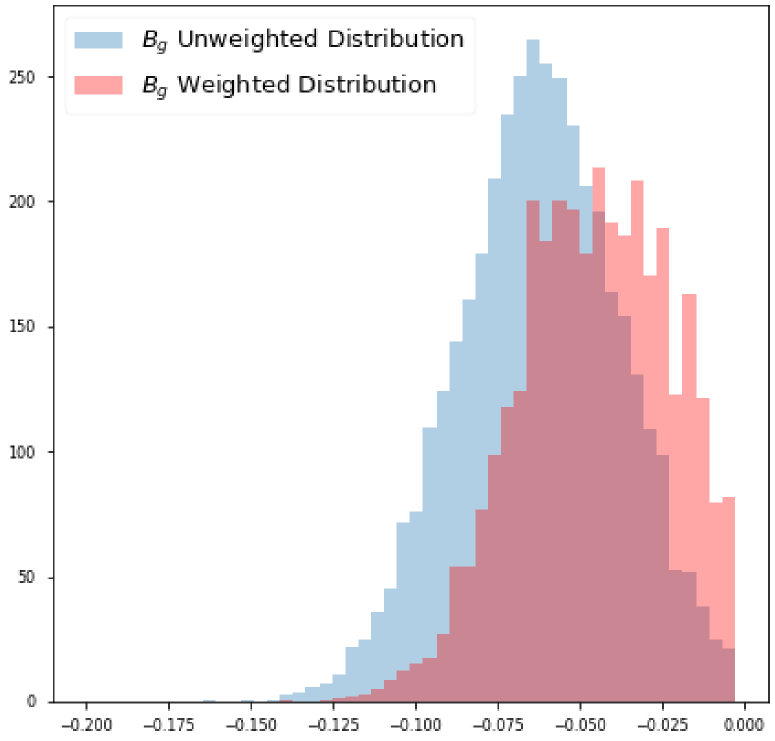
- **Clearly, if $L(\boldsymbol{\theta}) \propto L'(\boldsymbol{\theta})$, then all $w_k = \text{const.}$ and the distributions remain unchanged. If on the other hand w_k vary, then distribution shapes will potentially be altered.**
- **The weighted distributions are what we expect to arrive at if we could do a Markov chain sampling of $L(\boldsymbol{\theta})$.**

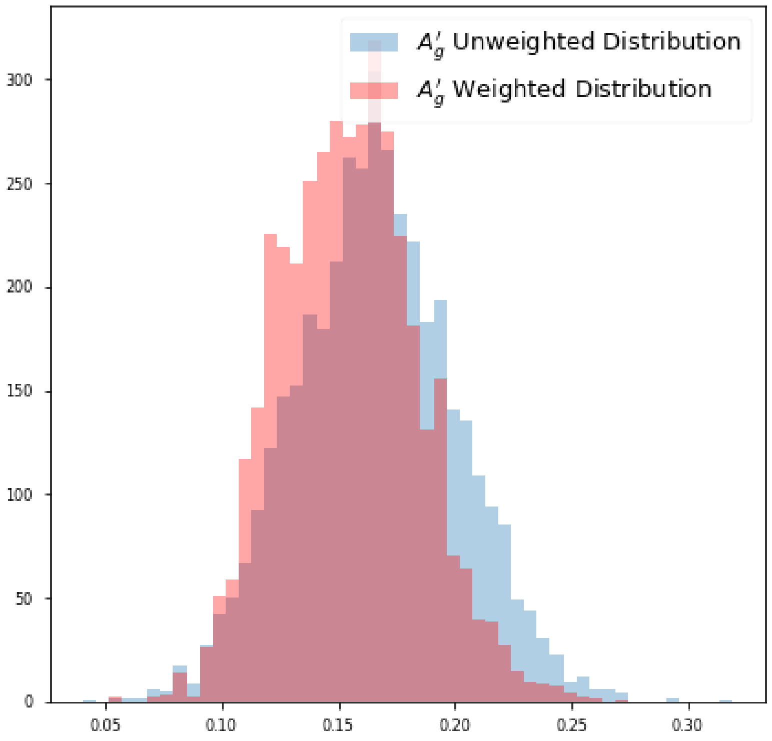
One dataset Reweighted Distributions

$$w_k^i > \overline{w^i}_k - 4 \times \sigma^i$$
$$w_k^i < \overline{w^i}_k + 4 \times \sigma^i$$



- **With one dataset, Gaussian approximations are close to reweighted likelihoods.**
- **Data: HERA I, II & ZEUS**

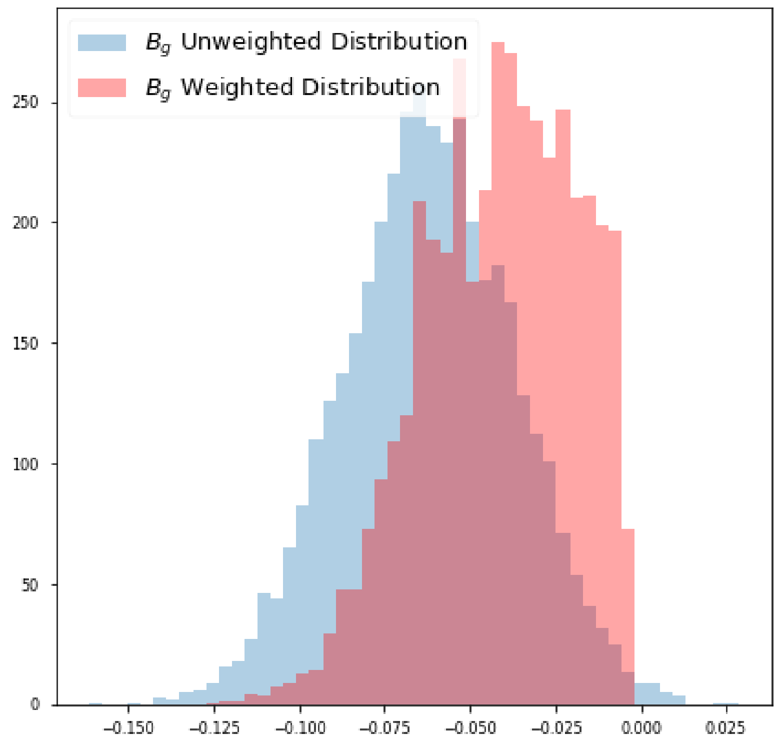


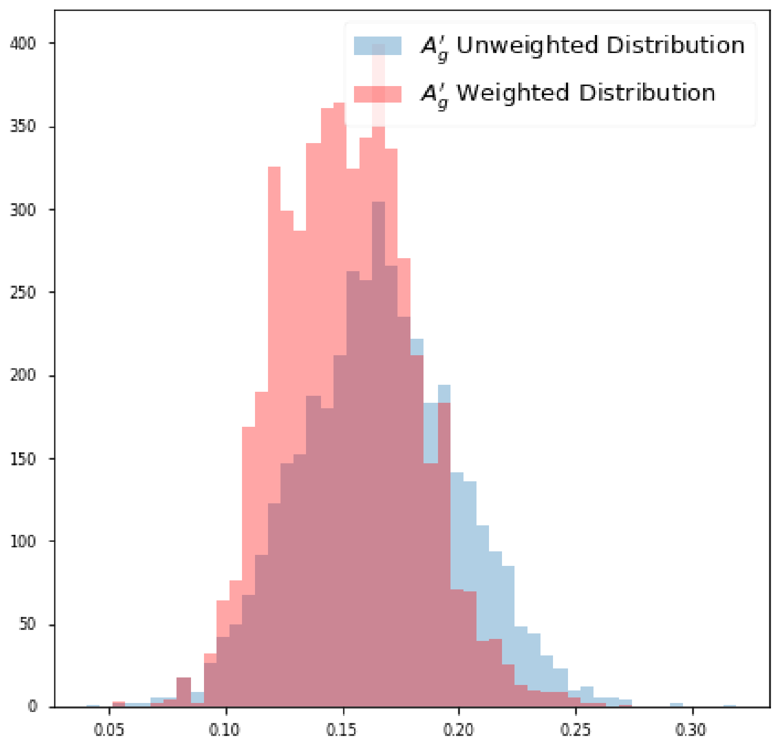


4,000 data points (parameter sets) HERAPDF parameterization

Multiple Data sets Reweighted Distributions

- **With multiple data sets (global fit), we see bigger discrepancies between the Gaussian approximation and the reweighted likelihoods. Reason: data sets are discrepant among themselves.**
- **Data: HERA I+II & ZEUS + CDF W asymmetry + D0 Run II cone jets**





4,000 data points (parameter sets)
HERAPDF parameterization

Next Steps and Suggestion for xFitter

- **Once we have a full mapping of $L(\theta)$ using multiple (all) the data sets, we construct Bayesian credible intervals for θ without using $\Delta\chi^2 = 1$.**
- **We really need more points and more data sets. This is computationally expensive as we are doing a whole fit in xFitter to get the χ^2 value.**

Add xFitter functionality to allow for calculation of the unminimized χ^2 given a parameter set.

- **Calculate Bayesian Credible intervals. Hypothesis: the effective size of the 68% intervals is much larger than the one we would obtain if we simply used a gaussian approximation.**
- **Hypothesis: the ratio of these two sizes would be the tolerance that is often used.**

$$\frac{C.L._{Reweighted}^{68\%}}{C.L._{Gaussian}^{68\%}} \approx T$$

Backup

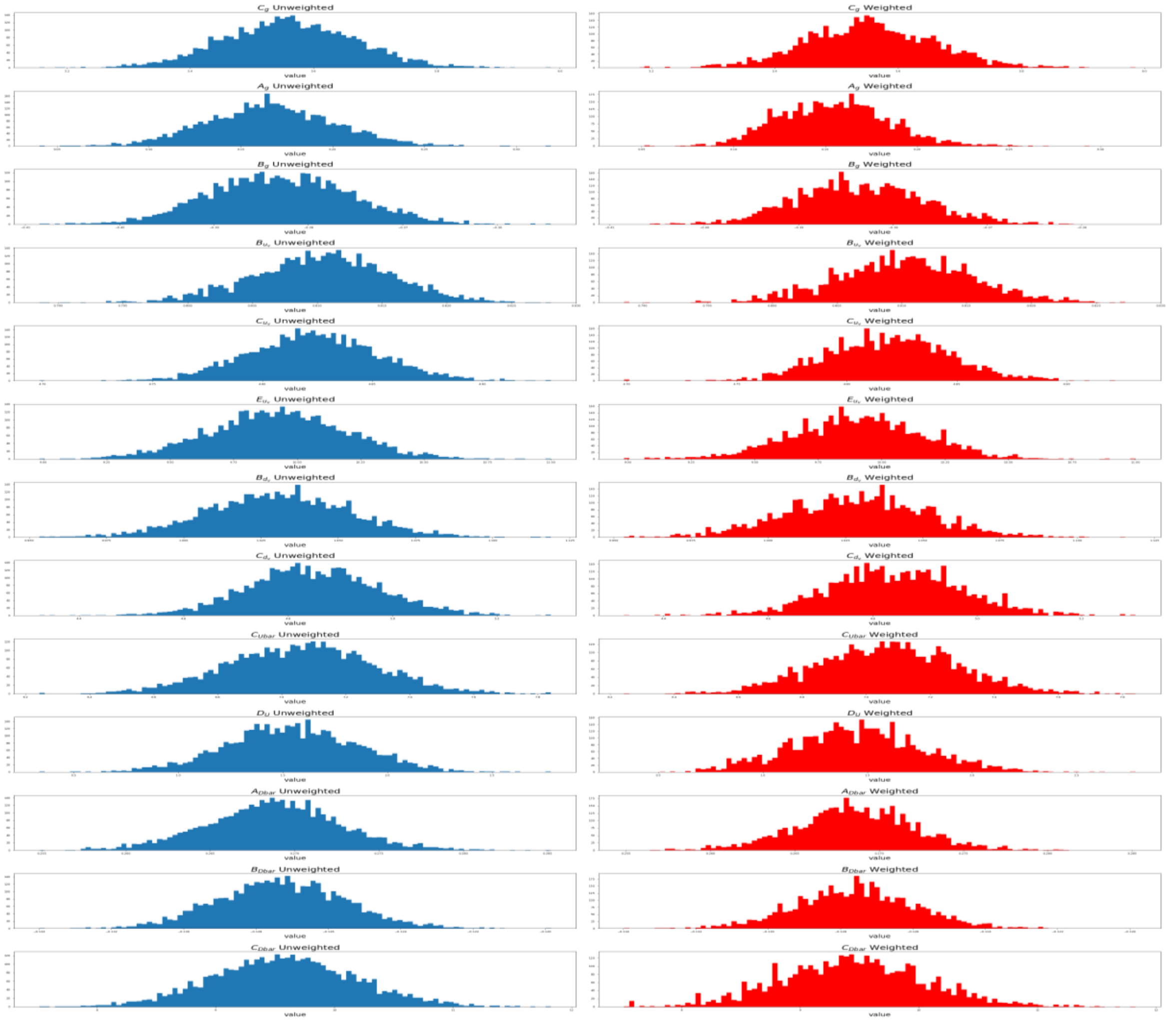
- **All code is available at: https://github.com/AliAlkadhim/NNPDF_Uncertainty/**
- **If we approximate $L'(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{D};\ \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{i}})$, then**

$$w_k = \frac{L(\boldsymbol{\theta})}{L'(\boldsymbol{\theta})} = \frac{N_{samples}}{\sum_{k=1}^{N_{samples}} w_k} \frac{e^{-\frac{1}{2}\chi_k^2}}{\mathcal{N}(\boldsymbol{D};\ \widehat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}),} = \begin{cases} 1, \text{Gauss. Approx. holds for } \mathrm{L}(\boldsymbol{\theta}) \\ \text{else,} \qquad \mathrm{L}(\boldsymbol{\theta}) \text{ is non} - \text{Gauss.} \end{cases}$$

- **If the likelihood for θ is multivariate normal, the likelihood of a single observation is of the form**
- $L(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}[\mathbf{x} - g(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-\mathbf{1}}[\boldsymbol{x} - \mathbf{g}(\boldsymbol{\theta})]\right\} = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\chi^2\right\} \longrightarrow \log L(\boldsymbol{\theta}) = -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\chi}^{\mathbf{2}}$
- **68% confidence intervals are obtained by finding points where $\Delta\chi^2 = 1$, i.e.**
- $-2\Delta \log[\mathrm{L}] = -2[\log[L(\boldsymbol{\theta}_{\pm}|\boldsymbol{x})] - \log[L(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})] = 1 \longrightarrow (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{-}, \widehat{\boldsymbol{\theta}} + \boldsymbol{\theta}_{+})$ **but this assumes normal sampling of data.**
- **The tolerance $T = \sqrt{\Delta\chi^2_{global}}$, ideally $T = 1$, but this assumes ideal gaussian errors & well defined theory.**
  - **In global fits, $T > 1$ to account for discrepant data sets (e.g. see arxiv: 1410.8849).**

All parameter Distributions

One Dataset



Multiple Datasets