

Extending the Fully Bayesian Unfolding with Regularization Using a Combined Sampling Method

Petr Baron[†]  and Jiří Kvita^{*,†} 

Joint Laboratory of Optics of Palacký University and Institute of Physics AS CR, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic; petr.baron@upol.cz

* Correspondence: jiri.kvita@upol.cz

† These authors contributed equally to this work.

Received: 13 November 2020; Accepted: 15 December 2020; Published: 17 December 2020



Abstract: Regularization extensions to the Fully Bayesian Unfolding are implemented and studied with an algorithm of combined sampling to find, in a reasonable computational time, an optimal value of the regularization strength parameter in order to obtain an unfolded result of a desired property, like smoothness. Three regularization conditions using the curvature, entropy and derivatives are applied, as a model example, to several simulated spectra of top-pair quark pairs that are produced in high energy pp collisions. The existence of a minimum of a χ^2 between the unfolded and particle-level spectra is discussed, with recommendations on the checks and validity of the usage of the regularization feature in Fully Bayesian Unfolding (FBU).

Keywords: unfolding; Bayes theorem; regularization; MCMC

1. Introduction

This study is motivated by unstable results of unfolding in the case of spectra with a large number of bins or with a poorly diagonal migration matrix. In order to obtain the expected smooth results, a bias is carefully introduced via a regularization function. We focused on the Fully Bayesian Unfolding (FBU) method [1].

Other unfolding methods than FBU have already been introduced, including regularization features. Because FBU has the ability to provide the full probability posterior of the result and thus stands out from other methods, it is interesting to also provide FBU with the regularization option.

This study uses several over-binned spectra, so the unfolding without regularization results in fluctuating, high-variance unfolded spectra. These were constructed by fine binning, so that the migration matrix is under-represented on diagonal. Five typical spectra considering the production of top quark pairs in pp collisions were chosen, namely the pseudorapidity $\eta^{\text{t, had}}$ and transverse momentum $p_T^{\text{t, had}}$ of the hadronically decaying top quark, and pseudorapidity $\eta^{\text{t\bar{t}}}$, transverse momentum $p_T^{\text{t\bar{t}}}$, and mass $m^{\text{t\bar{t}}}$ of the $t\bar{t}$ system.

The regularization feature was introduced as a prior in the likelihood. However, the problem which remains is how strongly should user apply the regularization to get the optimal unfolded spectrum, which should be close to the particle spectrum, but, at the same time, respects the desired regularization property, e.g., low curvature. The power of regularization is governed by the regularization strength parameter τ . The higher the τ , the stronger the regularization.

An algorithm for finding the optimal value of τ in a reasonable time via a combined sampling is given so that the best unfolding result is achieved, representing the main message of this paper.

2. Unfolding

The unfolding procedure corrects measured spectra for the resolution, efficiency, and acceptance effects of the detector usually to the particle level. The principles of the Fully Bayesian Unfolding method are based on the Bayes theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \quad (1)$$

where the conditional probability $P(A|B)$ reads the probability of A , given as B . Fully Bayesian Unfolding provides the full probability posterior $P(A|B)$ of the result, which gives extra information when compared to other unfolding methods [2]. Rewriting Equation (1), the probability of the truth spectrum $\mathbf{T}$, given the data $\mathbf{D}$, is

$$P(\mathbf{T}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{T}) \cdot \pi(\mathbf{T})}{\text{Norm.}}, \quad (2)$$

where the Norm. is a normalization factor, $P(\mathbf{D}|\mathbf{T})$ is the likelihood function, and $\pi(\mathbf{T})$ is a prior probability density for $\mathbf{T}$. In the case of $\pi(\mathbf{T}) = 1$ the algorithm is non-regularized. By inserting an arbitrary regularization function instead of $\pi(\mathbf{T})$, one can intentionally bias the final result towards a suitable property, e.g., to be smooth, and unfolding is then called regularized. The convenient way is to use an exponential function with regularization strength parameter τ and an inner function $S(\mathbf{T})$

$$P(\mathbf{T}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{T}) \cdot e^{-\tau S(\mathbf{T})}}{\text{Norm.}}. \quad (3)$$

If the parameter $\tau = 0$, the prior $\pi(\mathbf{T}) = 1$, and no regularization is applied; on the other hand, the higher value of τ , the more dominant the regularization term.

2.1. Simulated Spectra

The input pseudo-data spectrum $\mathbf{D}_{\text{sim}}$ is simulated for the process $pp \rightarrow t\bar{t}$ at $\sqrt{s} = 14$ TeV while using MadGraph5 [3] software, Pythia8 generator [4], and Delphes [5] detector simulation with an ATLAS card while using basic kinematics cuts in order to obtain a realistic sample of top quark pairs decaying in a semi-leptonic way $t\bar{t} \rightarrow bWbW \rightarrow \ell + \text{jets}$, with more details being described in [6].

2.2. Defining Equations of Unfolding

The necessary ingredients for unfolding are the data spectrum $\mathbf{D}$ and the migration matrix M , which is normalized, so that it maps the probability of an event at the truth (particle) level bin i to be reconstructed (measured) at the detector level bin j .

The schematic formula describing unfolding for the case of a simple migration matrix inversion is

$$\hat{T}_i = \frac{1}{\epsilon_i} \cdot (M^{-1})_{ij} \cdot \eta_j \cdot (D - B)_j. \quad (4)$$

In this paper, the background spectrum $\mathbf{B}$ is not taken into account. The efficiency and acceptance corrections are derived from projections of the migration matrix M divided by the simulated particle level $\mathbf{T}$ or data spectrum $\mathbf{D}$, as

$$\epsilon = \frac{\hat{\mathbf{T}}_{\text{proj. from M}}}{\mathbf{T}}; \quad \eta = \frac{\tilde{\mathbf{D}}_{\text{proj. from M}}}{\mathbf{D}}. \quad (5)$$

Because our data are simulated we set $\mathbf{D} = \mathbf{D}_{\text{sim}}$ and Equation (4) becomes

$$\hat{\mathbf{T}} = \frac{1}{\epsilon} \cdot M^{-1} \cdot \eta \cdot \mathbf{D}_{\text{sim}}. \quad (6)$$

The production of spectra was divided into two statistically independent sets A and B, where set A is used for unfolding components and set B is taken as pseudo data $\mathbf{D}_{\text{simB}}$ input to unfolding

$$\hat{\mathbf{T}}_B = \frac{1}{\epsilon_A} \cdot M_A^{-1} \cdot \eta_A \cdot (\mathbf{D}_{\text{simB}}). \quad (7)$$

The result of unfolding $\hat{\mathbf{T}}_B$ is compared to the corresponding particle level spectrum $\mathbf{T}_B$ (green line in Figure 1a). The χ^2 test between the unfolded and particle spectrum divided by the number of degrees of freedom (number of the bins), χ^2/ndf , is calculated in the standard way, as follows

$$\chi^2/\text{ndf} = \frac{1}{N} \sum_{i=1}^N \frac{(\hat{T}_B^i - T_B^i)^2}{\sigma_{\hat{T}_B^i}^2 + \sigma_{T_B^i}^2}. \quad (8)$$

The input components that are needed for the unfolding process are visualized for the case of the $\eta^{t\bar{t}}$ spectrum presented in Figure 1.

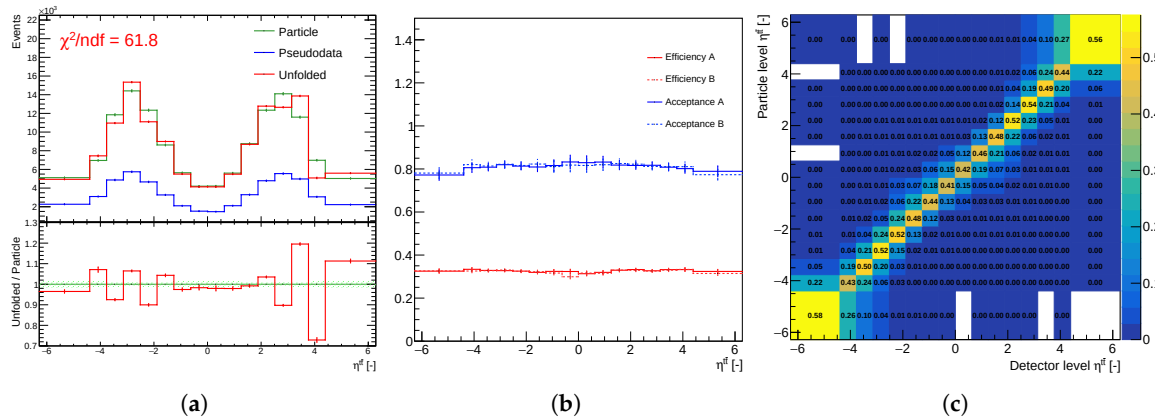


Figure 1. Unfolding components of the spectrum $\eta^{t\bar{t}}$: (a) particle spectra (green), pseudo data (blue), unfolding result $\hat{\mathbf{T}}$ (red), (b) efficiency ϵ and acceptance η corrections of statistically independent sets A and B, and (c) normalized migration matrix M_A .

In essence, the FBU unfolding technique exploits the response matrix to fold a probe truth-level spectrum, obtaining the detector-level spectrum, and computing a likelihood between such folded spectrum and the actual data. This elegantly avoids issues of singularities in explicit matrix inversion or their regularization, as e.g., in the SVD technique [7]. The tricky part of the FBU is how to effectively sample the truth space in the region giving large likelihood values.

2.3. Likelihood Function $P(\mathbf{D}|\mathbf{T})$

The likelihood function is given by the product of Poisson distributions in case of counts as

$$P(\mathbf{T}|\mathbf{D}) \propto L(\mathbf{D}|\mathbf{T}) \cdot \pi(\mathbf{T}) = \left(\prod_{i=1}^N \frac{\lambda_i^{k_i}}{k_i!} e^{-\lambda_i} \right) e^{-\tau_{\text{abs}} S(\mathbf{T})}, \quad (9)$$

where

$$\lambda_i = \sum_{j=1}^N M_{ij} T_j \quad (10)$$

and

$$k_i = \eta_i(D - B)_i. \quad (11)$$

After unfolding, the probability $P(\mathbf{T}|\mathbf{D})$ is divided by the efficiency correction ϵ . The acceptance correction η that is shown in Figure 1b takes on values [0,1] and, thus, the factorial is computed using the Gamma function as

$$x! = \Gamma(x + 1). \quad (12)$$

In the simplified case of no background ($\mathbf{B} = 0$), pseudo data $\mathbf{D} = \mathbf{D}_{\text{sim}}$, and introducing the natural logarithm, since the sampling of the functions is more numerically stable, Equation (9) becomes

$$\begin{aligned} \log P(\mathbf{T}|\mathbf{D}_{\text{sim}}) &\propto \log(L(\mathbf{D}_{\text{sim}}|\mathbf{T}) \cdot \pi(\mathbf{T})) = \\ &= \sum_{i=1}^N \log \left(\frac{\lambda_i^{[\eta_i(\mathbf{D}_{\text{sim}})_i]}}{[\eta_i(\mathbf{D}_{\text{sim}})_i]!} e^{-\lambda_i} \right) - \tau_{\text{abs}} S(\mathbf{T}), \end{aligned} \quad (13)$$

which can be rewritten as

$$\log P(\mathbf{T}|\mathbf{D}_{\text{sim}}) \propto \log L(\mathbf{T}) - \tau_{\text{abs}} S(\mathbf{T}). \quad (14)$$

This expression has the advantage of allowing one to find the maximum of $\log P(\mathbf{T}|\mathbf{D}_{\text{sim}})$ by sampling the log-likelihood function $\log L(\mathbf{T})$ and the regularization function $S(\mathbf{T})$ separately and tune the regularization parameter τ , so that the optimal τ can be found in a much shorter time than sampling the full likelihood for every possible value of τ .

In this study, both of the approaches were used: the one when sampling was applied to the function $P(\mathbf{T}|\mathbf{D}_{\text{sim}})$ as a whole, denoted as full sampling, and the one when sampling of the two components $\log L(\mathbf{T})$ and $S(\mathbf{T})$ is performed separately, denoted as combined (fast) sampling. These methods will be described in detail in Section 2.6.

Figure 2a shows an example of the log-likelihood, sampling of the function $\log L(\mathbf{T})$ for the spectrum η^{if} and its projection to the 6th and 9th bin $\log L(T_6, T_9)$. Figure 2b is created while using the highest points of the $\log L(T_6, T_9)$ and, thus, represents an envelope of the $\log L(T_6, T_9)$, while Figure 2c represents the exponential of the envelope of the original $L(T_6, T_9)$, with the posteriors plotted along the x and y axes (red lines), which are the marginalized binned probability densities that are analytically given as

$$p_i(T_i|\mathbf{D}) = \int \cdots \int P(\mathbf{T}|\mathbf{D}) dT_1 \dots dT_{i-1} dT_{i+1} \dots dT_N. \quad (15)$$

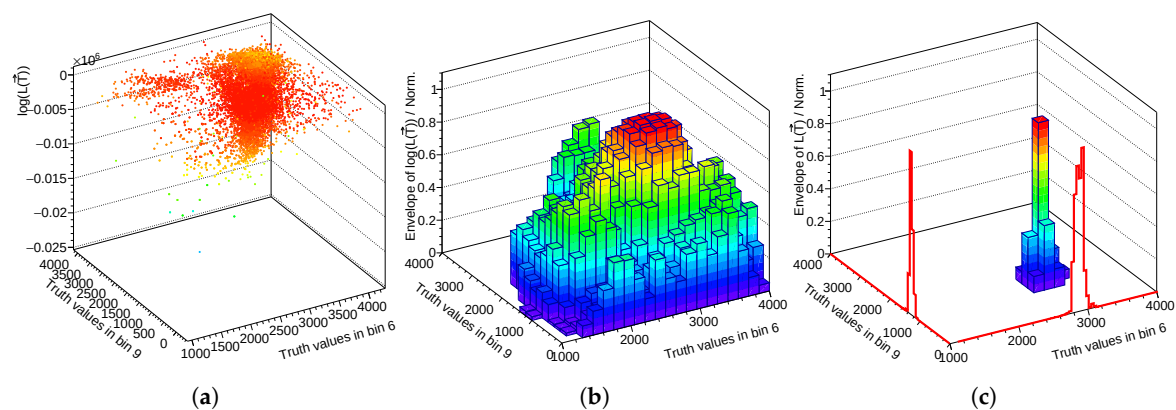


Figure 2. (a) View of a part of the 16-dimensional log-likelihood $\log L(\mathbf{T})$ function as function of the 6th and 9th bin (b) normalized maximal values of $\log L(T_6, T_9)$, and (c) normalized maximal values of $L(T_6, T_9)$.

The mean and standard deviation of the posteriors are taken as the unfolded result in each bin in this study in order to construct the unfolded spectrum and its statistical uncertainty. Another approach is to fit the posteriors and use parameters from the fit, see Figure 3, but this method is not used in this paper, since the regularization can change the shape of the posterior to non-Gaussian.

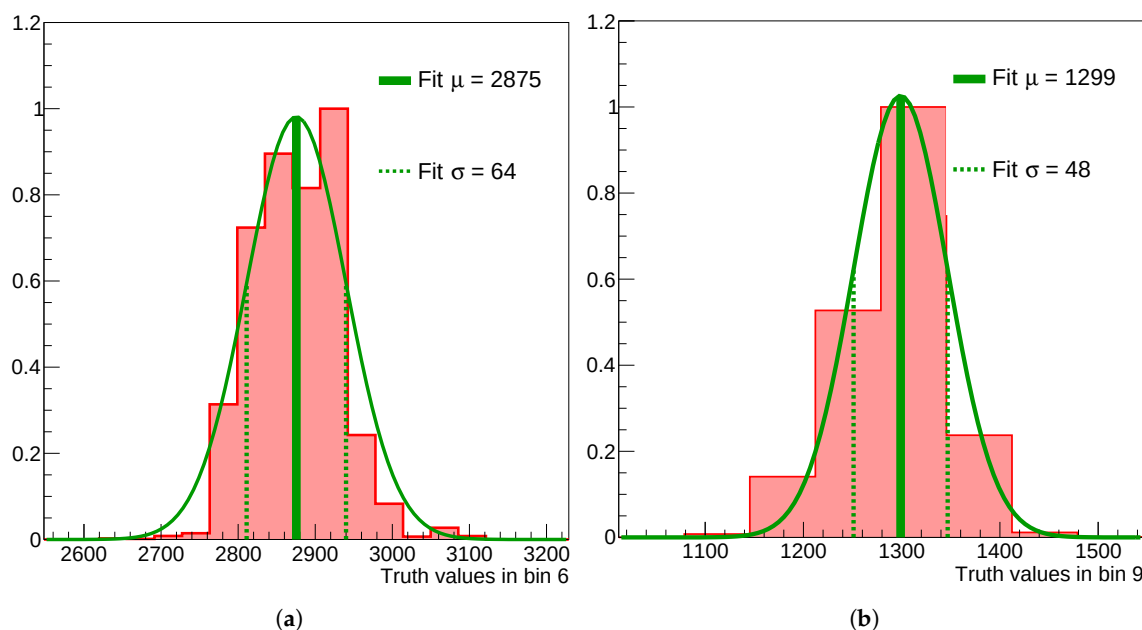


Figure 3. Marginalized 1D posteriors in the 6th (a) and 9th (b) bin of the η^H spectrum without regularization applied.

The advantage of the FBU in comparison to other unfolding methods is in providing the whole probability distribution. Marginalizing this probability to one-dimensional posteriors the user gains the probability distribution of the unfolded result for each bin and it can decide a custom statistical property to define the unfolding result and its uncertainty in the particular bin while other methods usually provide only a single number with some uncertainty. Thus, FBU provides more user control over the unfolded output interpretation. Furthermore, marginalizing the whole probability into two dimensions enables one to study cross-bin correlations, accessible only via dedicated pseudo experiments in other methods. Another advantage is the possibility of a natural inclusion of systematic uncertainties variation by introducing prior distributions for corresponding nuisance parameters directly into the likelihood function and performing a simultaneous fit, which is a procedure known as profiling.

2.4. Motivation for the Regularization

Figure 4 illustrates the unfolding process and proceeds from the (blue) pseudo data line to the (red) unfolded line in Figure 4a, which fluctuates due to low occupied diagonal of the migration matrix M (Figure 1c). One can improve the variance of the result by introducing a regularization term aiming to diminish the curvature of the unfolding result (Figure 4b) with the optimal strength parameter $\tau_{\text{opt}} = 2089$ derived using the full sampling algorithm discussed in Section 2.3 (Figure 5), so that the resulting χ^2/ndf between particle and unfolded spectra is minimal.

On the other hand, if one sets the strength of regularization to a very large number, e.g., $\tau \approx 10^5$, the corresponding χ^2/ndf is high and the unfolded result departs from the particle level spectrum, see Figure 4c.

This implies that $\chi^2(\tau)/\text{ndf}$ may have a minimum as function of τ (see Figure 5) and the aim of this study is to provide a way to find such a minimum in a reasonable computational time. In the next section, we describe three regularization functions used in this study.

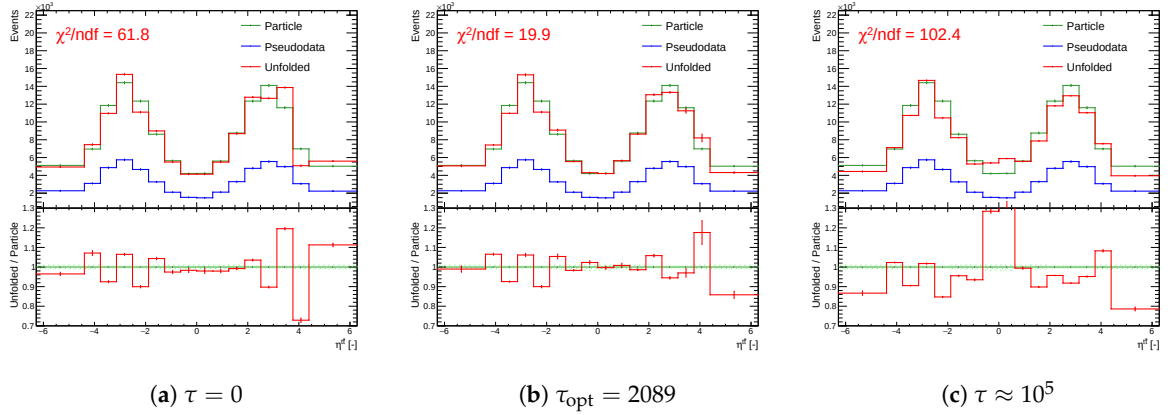


Figure 4. Unfolding the double-peaked η^{ff} over-binned spectrum for different values of the regularization strength parameter τ . The parameter τ is normalized, such that $\tau = \tau_{rel}$, see Section 3.

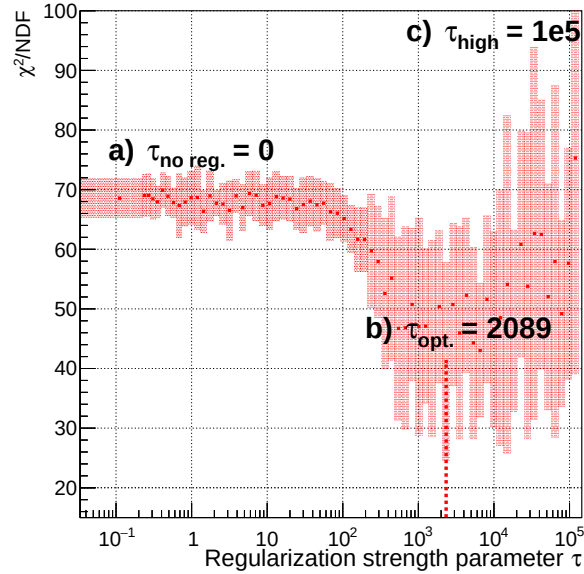


Figure 5. Relative $\chi^2(\tau)/ndf$ as function of the regularization strength parameter τ and its minimum at $\tau_{opt} = 2089$. The parameter τ is normalized, such that $\tau = \tau_{rel}$, see Section 3. The vertical line represents the minimum of χ^2/ndf .

2.5. Regularization Functions $S(\mathbf{T})$

While the regularization function can be taken as any arbitrary function, we motivate several choices [1]. The aim of the negative entropy, curvature and derivatives is to smoothen the unfolded spectra.

- Entropy regularization

$$S_1(\mathbf{T}) = - \left[- \sum_{t=1}^N \frac{T_t}{\sum T_{t'}} \log \left(\frac{T_t}{\sum T_{t'}} \right) \right]. \quad (16)$$

- Curvature regularization

$$S_2(\mathbf{T}) = \sum_{t=2}^{N-1} (\Delta_{t+1,t} - \Delta_{t,t-1})^2, \quad (17)$$

where

$$\Delta_{t_1,t_2} = T_{t_1} - T_{t_2}. \quad (18)$$

- First derivative regularization

$$S_3(\mathbf{T}) = \sum_{t=2}^{N-1} \frac{|\delta_{t+1,t} - \delta_{t,t-1}|}{|\delta_{t+1,t} + \delta_{t,t-1}|}, \quad (19)$$

where

$$\delta_{t_1,t_2} = \frac{\frac{T_{t_1}}{W_{t_1}} - \frac{T_{t_2}}{W_{t_2}}}{C_{t_1} - C_{t_2}} \quad (20)$$

and N is the number of bins, W_t is the width of t th bin, and C_t is the bin center of the t th bin.

Each of the regularization functions has a different effect on the full likelihood, see Figure 6. While sampling the likelihood according to gradient of $L_{\text{reg}}(\mathbf{T}) = L(\mathbf{T}) - S(\mathbf{T})$, i.e., leaving $\tau = 1$ for the moment, the values of $S(\mathbf{T})$ are stored.

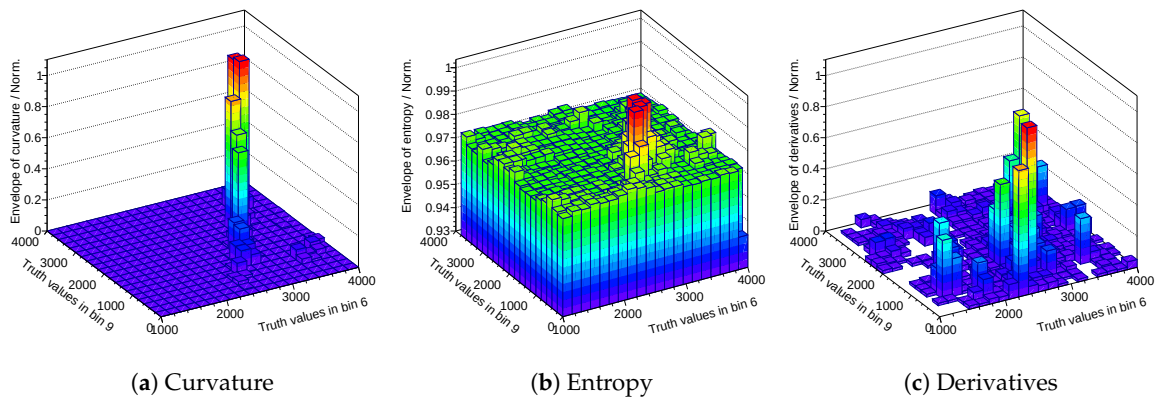


Figure 6. The envelope of normalized regularization functions $S(\mathbf{T})$ in the 6th and 9th bin. For sampling purposes, the gradient of $L(\mathbf{T}) - S(\mathbf{T})$ was used.

2.6. Sampling the Likelihood Function

Effective sampling plays a crucial role in finding extremes of the likelihood function. In this paper, a private implementation of the Markov chain Hamiltonian Monte Carlo [8] method is used when each point of a function is derived from the previous point, thus forming a chain. The problem of finding the maximum of the likelihood function is transformed to generate a sample via the study of motion of a virtual particle in the potential represented by the likelihood function in an N -dimensional space.

The Hamiltonian of a particle with momentum $\mathbf{p}$ in potential $V(\mathbf{x})$, depending on position vector $\mathbf{x}$, can be reformulated from the classical expression

$$H(\mathbf{p}, \mathbf{x}) = \frac{\mathbf{p} \cdot \mathbf{p}}{2} + V(\mathbf{x}) \quad (21)$$

to the studied problem, where the position vector $\mathbf{x}$ is substituted by the vector of particle-level truth values $\mathbf{T}$ and the potential $V(\mathbf{x})$ by the negative log-likelihood function $-\log L(\mathbf{T})$, so that

$$H(\mathbf{p}, \mathbf{T}) = \frac{\mathbf{p} \cdot \mathbf{p}}{2} - \log L(\mathbf{T}). \quad (22)$$

The first Hamiltonian equation of motion

$$\frac{\partial H}{\partial \mathbf{x}} = -\frac{d\mathbf{p}}{dt} \quad (23)$$

becomes in the space of truth values and momentum

$$\frac{\partial H}{\partial \mathbf{T}} = -\frac{d\mathbf{p}}{dt} \quad (24)$$

and numerically

$$\nabla_{\mathbf{T}} \log L(\mathbf{T}) = \frac{(\mathbf{p}_{t+\varepsilon/2} - \mathbf{p}_t)}{\varepsilon/2} \quad (25)$$

where ε is the time interval. Each point in momentum space is updated as

$$\mathbf{p}_{t+\varepsilon/2} = \mathbf{p}_t + \frac{\varepsilon}{2} \nabla_{\mathbf{T}} \log(L(\mathbf{T})). \quad (26)$$

Similarly, the second equation of motion

$$\frac{\partial H}{\partial \mathbf{p}} = \frac{d\mathbf{x}}{dt} \quad (27)$$

becomes

$$\mathbf{p} = \frac{(\mathbf{T}_{t+\varepsilon} - \mathbf{T}_t)}{\varepsilon} \quad (28)$$

so the truth (position) value is updated in time, according to

$$\mathbf{T}_{t+\varepsilon} = \mathbf{T}_t + \varepsilon \mathbf{p}_{t+\varepsilon/2} \quad (29)$$

and the momentum as

$$\mathbf{p}_{t+\varepsilon} = \mathbf{p}_{t+\varepsilon/2} + \frac{\varepsilon}{2} \nabla_{\mathbf{T}} \log L(\mathbf{T}_{t+\varepsilon}). \quad (30)$$

Equations (26), (29) and (30) provide one and two-step updates for $\mathbf{T}$ and $\mathbf{p}$, respectively. The time interval ε is adapted in first few steps and it then stays constant for the rest of the sampling. The initial momentum $\mathbf{p}$ is chosen randomly according to normal distribution and the initial truth vector $\mathbf{T}$ is randomly chosen while using uniform distribution within limits set by user or using a default expected range in i -th bin $\left[0, 3 \cdot \frac{(D_{\text{sim}})_i \eta_i}{\epsilon_i}\right]$. These limits create a hyper-cube, in which the sampling is performed. The algorithm details are described in [8].

The introduction of the canonical momentum and shifting the N dimensional problem to $2N$ dimensions is one of the key features of the Hamiltonian MC chain generator. Classical mechanics Hamiltonian equations of motions are then employed, with the negative likelihood playing the role of a potential for a particle moving in the generalized phase-space, under the conservation of the total energy. The invariance of the system w.r.t. time translations is the motivation for exploring the trajectory of such a particle for different initial conditions, which leads to effectively sampling a larger phase space and more of the classical trajectories, thus increasing the probability of reaching the global minimum of the potential, i.e., the likelihood maximum.

3. Regularization

This section provides an overview of unfolding characteristics as function of the regularization strength parameter τ . Throughout this paper, the original τ_{abs} parameter is normalized while using the number of bins N and the curvature, entropy, or derivatives of the particle spectrum that are generally symbolized by $S(\mathbf{T}_{\text{part}})$ as

$$\tau_{\text{abs}} = \frac{\tau_{\text{rel}}}{(N-2) \cdot S(\mathbf{T}_{\text{part}})} \quad (31)$$

and, in the τ dependence, figures $\tau \equiv \tau_{\text{rel}}$ are used, so different spectra and regularization functions acquire similar values of τ , which are then more comparable.

3.1. Evolution of Posteriors with τ

Two effects come into shaping the posteriors with stronger regularization applied. First is the shifting of the peak, which is welcome in obtaining the expected modified unfolded spectrum. The second effect decreases the posteriors width and this effect decreases the uncertainty in the mean for each bin sometimes down to an artificially infinitesimal small width with extremely high τ and, thus, undesirably somewhat increases the χ^2 when comparing unfolded and truth spectra.

Because these are two competing effects, a minimum of the $\chi^2(\tau)/\text{ndf}$ may or may not appear. Figure 7 shows the posterior in the 15th bin of the η^{ti} spectrum for three scenarios. First, when no regularization is applied, so $\tau = 0$, then with regularization with optimal $\tau = 2089$, so the final χ^2/ndf is the smallest, and finally with an extremely high $\tau \approx 10^4$ when the posterior becomes almost a Dirac δ -function.

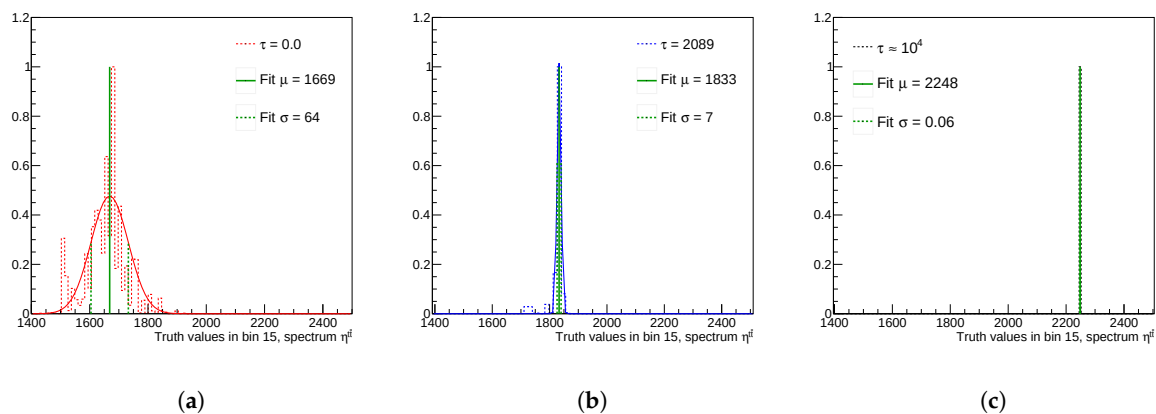


Figure 7. Posterior shifting and narrowing with increasing the regularization strength parameter τ in a selected single bin: (a) no regularization applied $\tau = 0$ (b) $\tau = 2089$ (c) $\tau \approx 10^4$.

3.2. Evolution of Curvature, Entropy, and Derivatives with τ

One of the first checks whether the regularization works properly is to study the curvature, entropy and derivatives of the unfolded spectra with respect to the increasing regularization strength τ . The expected result is that, with higher τ , the curvature decreases and the unfolded spectrum is smoother. The same characteristics also holds for entropy and derivatives. Figure 8 shows the decrease of curvature, entropy, and derivatives with increasing τ , as expected.

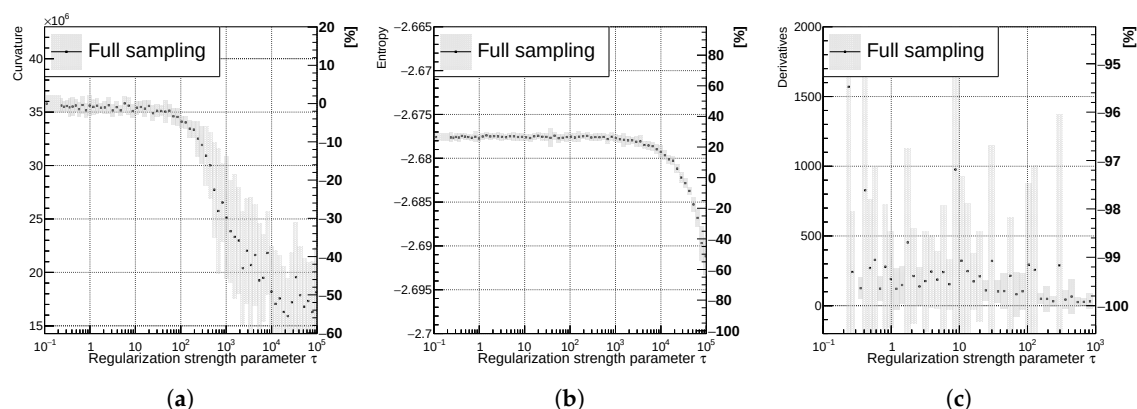


Figure 8. Mostly decreasing (a) curvature, (b) entropy, and (c) derivatives of the unfolded spectrum η^{ti} with respect to τ . The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds.

3.3. Evolution of Bin Cross-Correlations with τ

Using the likelihood projection to two different bins, the cross-bin correlations can be derived. Figure 9 shows the overall matrix built from these correlation factors. While, in case of (a) ($\tau = 0$), the bin correlations are minimal, the progressing trend in case (b) ($\tau = 912$), and finally (c) ($\tau \approx 10^4$) is evident. Thus, the stronger regularization is the more bins of the unfolded spectrum are correlated between each other.

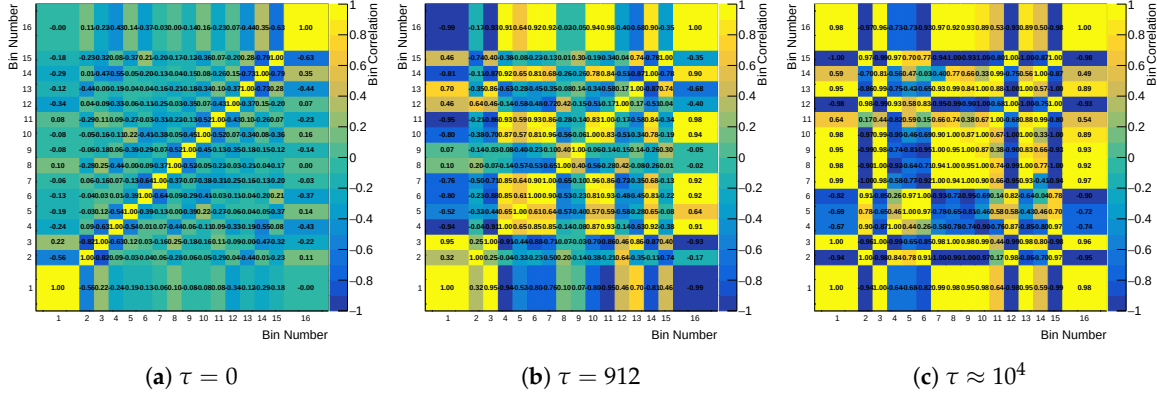


Figure 9. Cross-bin correlation matrix built from the correlation factor of likelihood, $L(T_i, T_j)$ while using the curvature regularization for three different values of τ .

In order to quantify the increase of correlations, the averaged sum of off-diagonal elements of bin correlation matrix A is shown for various τ , given by

$$\bar{C} \equiv \frac{1}{N_{\text{off}}} \sum_{i \neq j} A_{ij} \quad (32)$$

and also its absolute value version to study the sum of correlations and anti-correlations

$$\bar{C}_{\text{abs}} \equiv \frac{1}{N_{\text{off}}} \sum_{i \neq j} |A_{ij}|. \quad (33)$$

These bin-averaged cross correlations $\bar{C}$ and $\bar{C}_{\text{abs}}$ are shown in Figure 10, and the increasing trend proves that, the stronger regularization, the more correlated the unfolding result over bins.

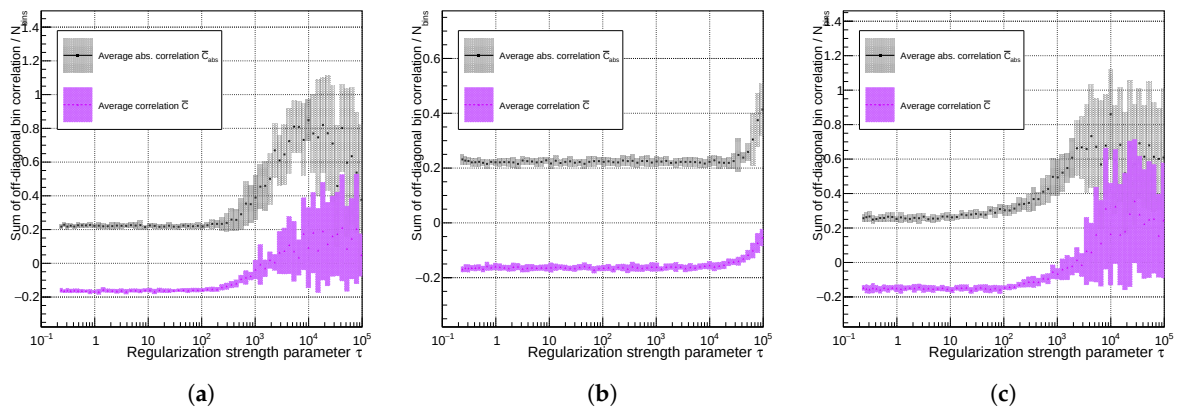


Figure 10. The averaged cross bin correlations $\bar{C}_{\text{abs}}$ (black) and $\bar{C}$ (pink) using (a) curvature, (b) entropy, and (c) derivative regularization for the η^{ff} spectrum. The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds.

3.4. Evolution of χ^2/ndf and Bin Uncertainties with τ

The χ^2/ndf is calculated as

$$\chi^2/\text{ndf} \equiv \frac{1}{N} \sum_i \frac{(T_i - \hat{T}_i)^2}{\sigma_{T_i}^2 + \sigma_{\hat{T}_i}^2}; \quad (34)$$

where each $\hat{T}_i$ for one particular bin is taken as the mean of the posterior and the uncertainty $\sigma_{\hat{T}_i}$ as the standard deviation of the posterior. Because the width of the posterior is observed to decrease with rising τ , the χ^2/ndf also has to rise. In order to naively separate effects of spectra smoothing and posteriors narrowing, two variables χ_{num} and χ_{denom} were studied, being defined as

$$\chi_{\text{num}} \equiv \frac{1}{N} \sum_i (T_i - \hat{T}_i)^2 \quad (35)$$

$$\chi_{\text{denom}} \equiv \frac{1}{N} \sum_i (\sigma_{T_i}^2 + \sigma_{\hat{T}_i}^2), \quad (36)$$

and shown in Figures 11 and 12.

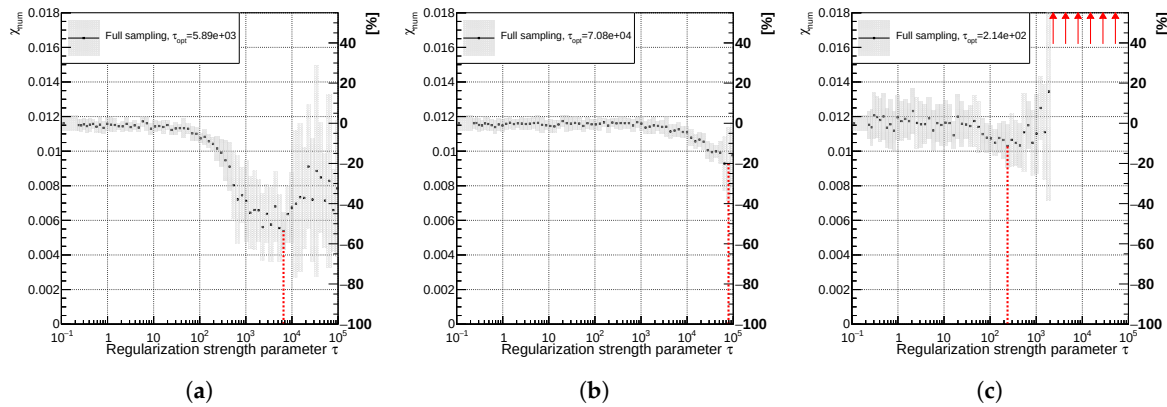


Figure 11. Variable $\chi_{\text{num}}(\tau)$ of the spectrum η^{ff} using (a) curvature (b) entropy and (c) derivative regularization illustrating the effect of spectra smoothing. The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds. The red line indicates the minimal value of the $\chi_{\text{num}}(\tau)$.

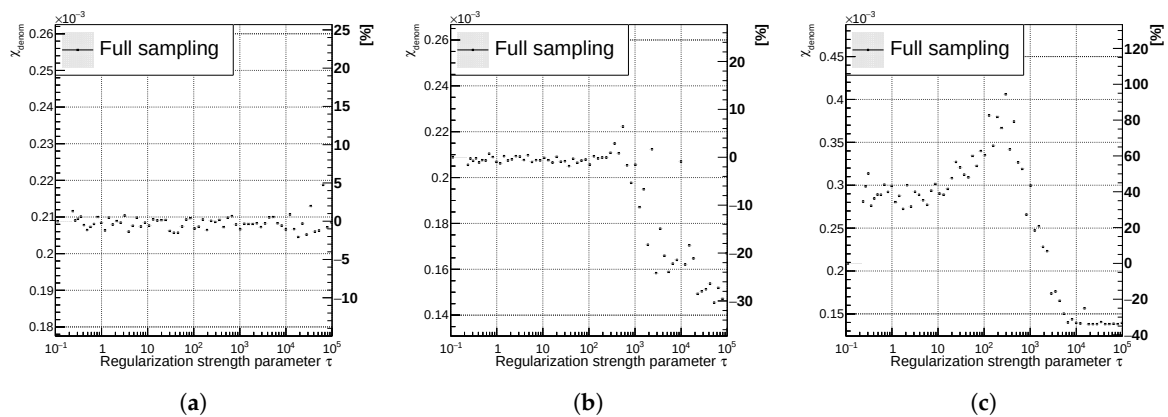


Figure 12. Variable $\chi_{\text{denom}}(\tau)$ of the spectrum η^{ff} using (a) curvature, (b) entropy, and (c) derivative regularization illustrating the effect of narrowing the posteriors and decreasing the uncertainty. The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds.

If the effects of shifting and narrowing the posterior are independent then χ^2/ndf and ratio $\chi_{\text{num}}/\chi_{\text{denom}}$ should be comparable

$$\chi^2/\text{ndf} \equiv \frac{1}{N} \sum_i \frac{(T_i - \hat{T}_i)^2}{\sigma_{T_i}^2 + \sigma_{\hat{T}_i}^2} \rightarrow \frac{\chi_{\text{num}}}{\chi_{\text{denom}}} \equiv \frac{\frac{1}{N} \sum_i (T_i - \hat{T}_i)^2}{\frac{1}{N} \sum_i \sigma_{T_i}^2 + \sigma_{\hat{T}_i}^2} \quad (37)$$

which can be seen in Figure 13.

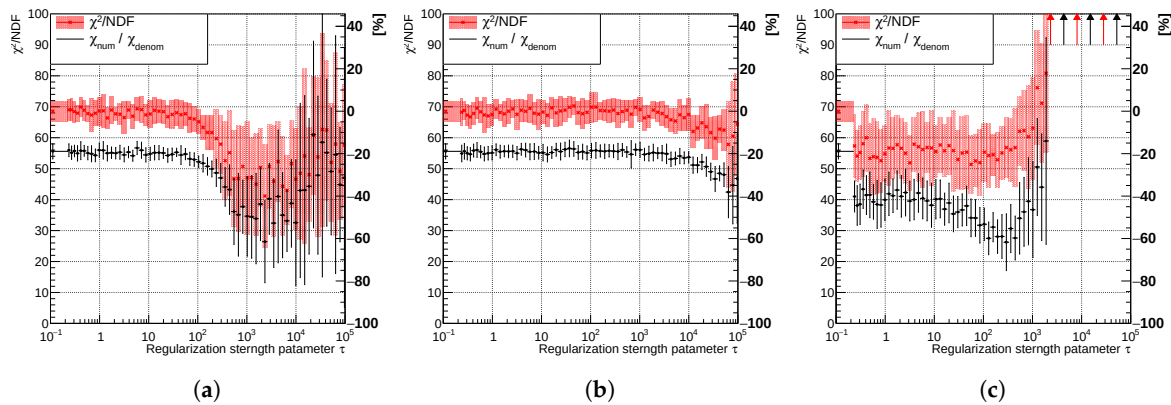


Figure 13. Variables χ^2/ndf and $\chi_{\text{num}}/\chi_{\text{denom}}$ of the η^{H} spectrum using (a) curvature, (b) entropy, and (c) derivative regularization showing good correspondence. The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds.

The problem of finding the optimal τ by finding the minimal χ^2/ndf is the time that is needed to create such a graph as in Figure 13, while each point of the graph is obtained by running the full unfolding. In the next section, an algorithm is described, which enables one to estimate the minimum of χ^2/ndf in a much faster way.

4. Combined Sampling as a Faster Algorithm

In order to speed up the algorithm which estimates the minimum χ^2/ndf value, the algorithm of combined sampling is proposed. The main idea is to sample separately the function $\log L_0(\mathbf{T})$ and the regularization function $S(\mathbf{T})$ and then vary the τ parameter to obtain the overall likelihood

$$\log L_\tau(\mathbf{T}) = \log L_0(\mathbf{T}) - \tau S(\mathbf{T}). \quad (38)$$

A problem arises with sampling the regularization function $S(\mathbf{T})$ since this function is not smooth and sampling algorithm can end up mapping only the local minima. To avoid this is, an auxiliary function

$$\log L_1(\mathbf{T}) = \log L_0(\mathbf{T}) - S(\mathbf{T}) \quad (39)$$

is sampled, where $\tau = 1$ so the global minimum is dominated by $L_0(\mathbf{T})$. The function $\log L_1(\mathbf{T})$ stands in sampling only for the purpose of providing a correct gradient, while values of $\log L_0(\mathbf{T})$ and $S(\mathbf{T})$ are stored separately, so post-unfolding these values can be used to compute in a fast way the full desired posterior of

$$\log L_{\text{comb}}(\mathbf{T}) = \log L_0(\mathbf{T}) - \tau S(\mathbf{T}) \quad (40)$$

for any τ .

This faster combined sampling approach is shown in blue in Figure 14, while the red points stand for the full sampling, where, for each point in τ , a new sampling had to be performed. Figure 14 shows a very good agreement between combined and full sampling methods in terms of their minima.

In the case of derivative regularization in Figure 14c the full sampling method (red) does not cover the highest values of τ in the range of approximately $[5 \times 10^3, 10^5]$. The reason is that the regularization function of derivatives, as compared to the curvature and entropy, is very narrow, approaching for high τ almost a δ -function, see Figure 6c and, thus, the sampling either takes infinite time to find such a peak in the wide domain of the function or is stopped by reaching the maximum number of iterations and the resulting χ^2/ndf is very high as the global maximum is not found.

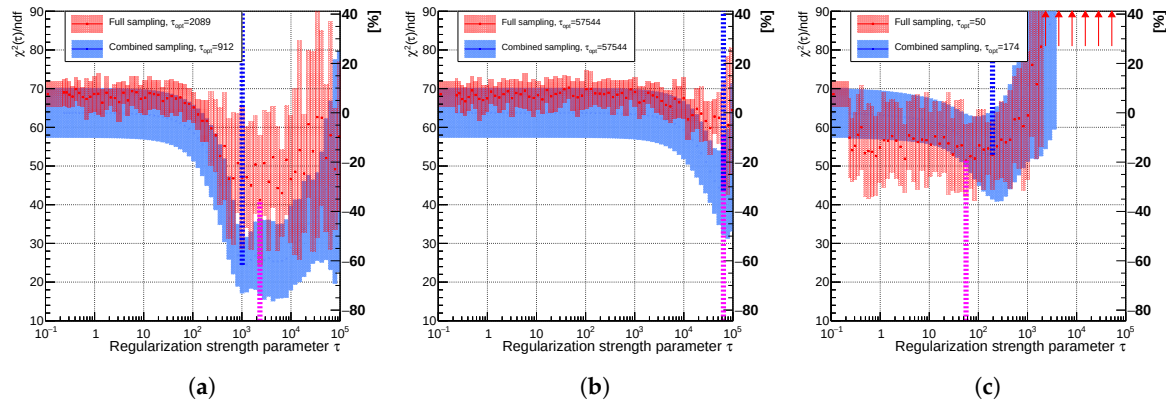


Figure 14. Variable $\chi^2(\tau)/\text{ndf}$ of the spectrum η^{tf} using (a) curvature, (b) entropy, and (c) derivative regularization comparing combined (faster) sampling (blue) and full sampling (red). The uncertainty band is evaluated as a standard deviation over 20 independent unfolding runs initiated with different random seeds. The vertical dotted lines indicate positions of χ^2/ndf minima for each sampling case.

5. Accidental Minima of $\chi^2(\tau)/\text{ndf}$

In the case of unfolding without regularization being effective enough so no smoothing is actually needed, the narrowing of posteriors can cause an accidental minimum in the $\chi^2(\tau)/\text{ndf}$ function. The example of the m^{tf} spectrum shows the variable χ_{num} as a flat function with increasing tail in high τ values, see Figure 15a. At the same time, the χ_{denom} decreases rapidly and the ratio of these components as well as χ^2/ndf exhibit a minimum at $\tau \approx 7000$ solely due to the narrowing of posteriors.

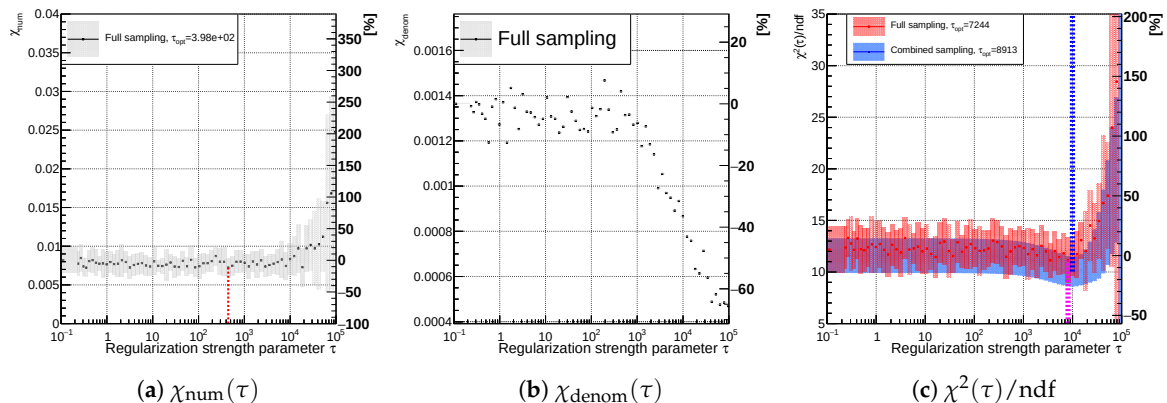


Figure 15. Variables (a) $\chi_{\text{num}}(\tau)$, (b) $\chi_{\text{denom}}(\tau)$ for the full sampling method; and (c) $\chi^2(\tau)/\text{ndf}$ using full (red) and combined (blue) sampling of the m^{tf} spectrum with an accidental minimum at $\tau \approx 7000$ (curvature regularization). The vertical dotted lines indicate positions of χ^2/ndf minima for each sampling case.

This accidental minimum does not improve the smoothness of the spectrum, as illustrated in Figure 16. The unfolded spectrum while using regularization in Figure 16b only decreases the uncertainties (red line) when compared to Figure 16a while not providing a smoother spectrum, as can be seen in the ratio and only a minor decrease of χ^2/ndf .

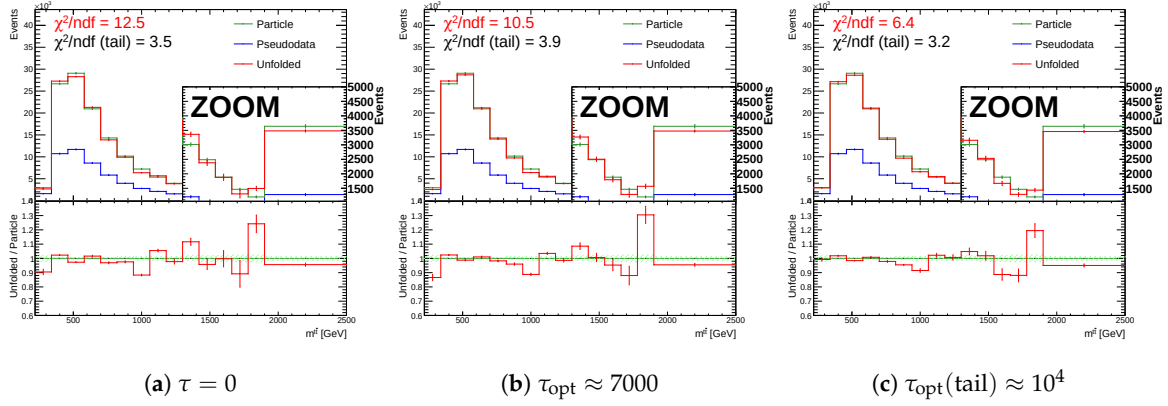


Figure 16. The result of unfolding (a) without regularization (b) with regularization and (c) with regularization applied only at second half of the spectrum m^{fi} while using the curvature in the case of a accidental minimum in $\chi^2(\tau)/\text{ndf}$ for one representative random seed.

The regularization is not effective due to the fact that the tail of the m^{fi} spectrum has lower number of events compared to the first half of the spectrum. The contribution to the curvature is thus small in the tail. The solution would be to switch on the regularization only for the second half of the spectrum as shown in Figure 16c where $\chi^2(\tau)/\text{ndf}(\text{tail})$ computed in the tail region of the spectrum decreases when compared to Figure 16a,b, also affecting the full χ^2/ndf via a change of the global maximum of the full likelihood.

6. Hidden Minima of $\chi^2(\tau)/\text{ndf}$

If the smoothing effect that is expressed by χ_{num} in Figure 17a has a minimum and if the narrowing effect described by χ_{denom} in Figure 17b starts to decrease with a similar or sharper slope than χ_{num} ($\tau \approx 10^2$), then the real minimum in χ^2/ndf in Figure 17c vanishes and flattens, so finding the optimal τ for the p_T^{had} spectrum in Figure 18 cannot be identified by minimizing χ^2/ndf as a function of τ . Figure 18b shows the improved unfolded spectrum that was obtained while using the minimum of χ_{num} . However, due to the narrowing of posteriors, the χ^2/ndf increases as compared to Figure 18a (no regularization).

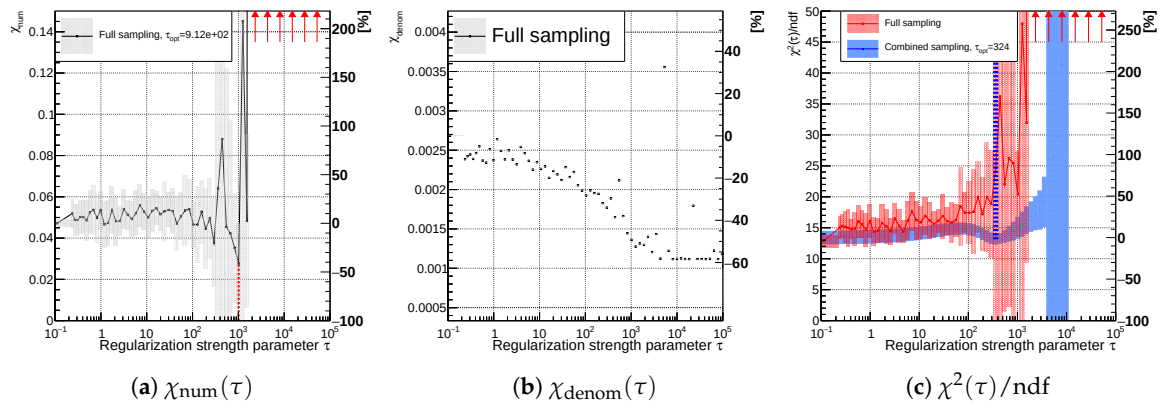


Figure 17. Variables (a) $\chi_{\text{num}}(\tau)$, (b) $\chi_{\text{denom}}(\tau)$ for the full sampling method; and, (c) $\chi^2(\tau)/\text{ndf}$ using full (red) and combined (blue) sampling of the p_T^{had} spectrum with a vanishing minimum (derivative regularization). The vertical dotted lines indicate positions of χ^2/ndf minima for each sampling case.

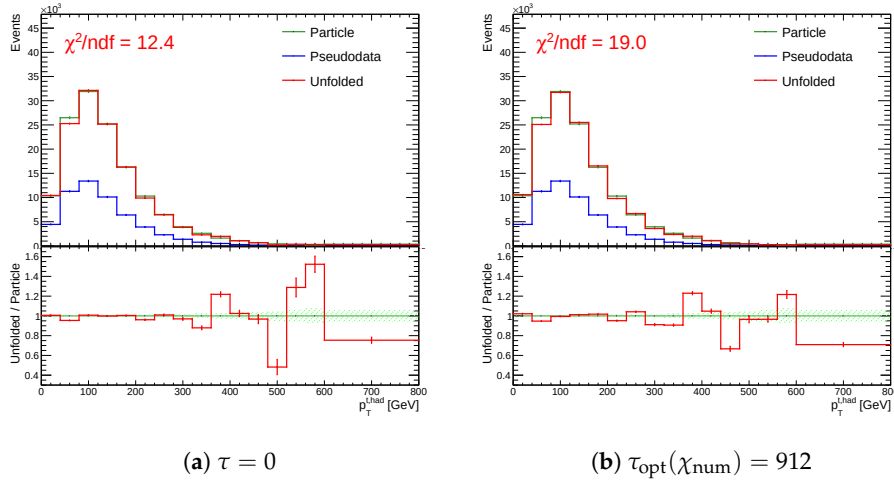


Figure 18. Result of unfolding (a) without regularization and (b) with regularization of the spectrum p_T^{had} using the derivatives in case of a hidden minimum in $\chi^2(\tau)/\text{ndf}$ for one representative random seed. Spectrum becomes smoother, but χ^2/ndf does not improve due to the narrowing of posteriors.

7. Real Minima of $\chi^2(\tau)/\text{ndf}$

If the smoothing effect expressed by χ_{num} in Figure 19a has a global minimum and, if the narrowing effect described by χ_{denom} in Figure 19b starts to decrease around the same region of τ values as χ_{num} ($\tau \approx 10^2$), then a real minimum in χ^2/ndf in Figure 19c is observed and it helps to improve the desired property the unfolded spectrum, as shown for the spectrum of η^{ff} in Figure 20.

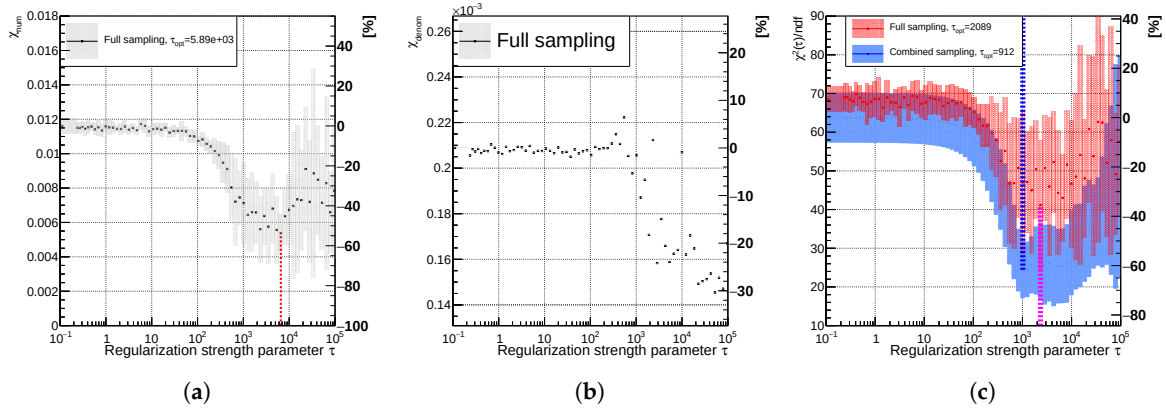


Figure 19. Variables (a) $\chi_{\text{num}}(\tau)$, (b) $\chi_{\text{denom}}(\tau)$ for the full sampling method; and (c) $\chi^2(\tau)/\text{ndf}$ using full (red) and combined (blue) sampling of the η^{ff} spectrum with the real minimum at $\tau \approx 900$ (curvature regularization). The vertical dotted lines indicate positions of χ^2/ndf minima for each sampling case.

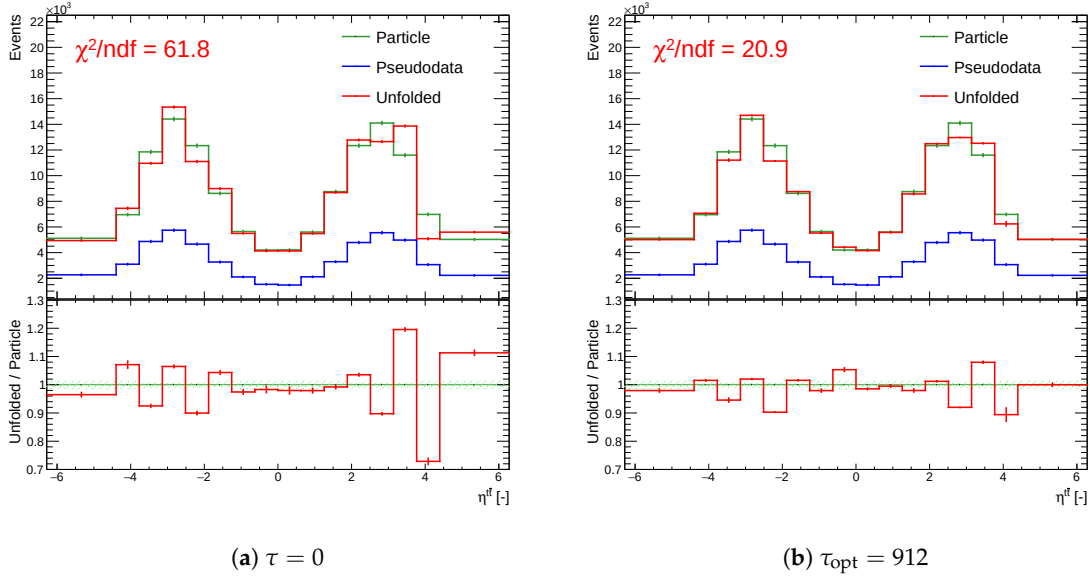


Figure 20. The result of unfolding (a) without regularization and (b) with regularization of the spectrum $\eta^{t\bar{f}}$ while using the curvature in case of a real minimum in χ^2/ndf for one representative random seed.

8. Results

The results for the spectra of $m^{t\bar{f}}$, $p_T^{t,\text{had}}$ and $\eta^{t\bar{f}}$ were already shown in Figures 16, 18 and 20. In this section, remaining two studied spectra $\eta^{t,\text{had}}$ and $p_T^{t\bar{f}}$ are shown in Figures 21 and 22, where, for $\eta^{t,\text{had}}$, the regularization is unnecessary.

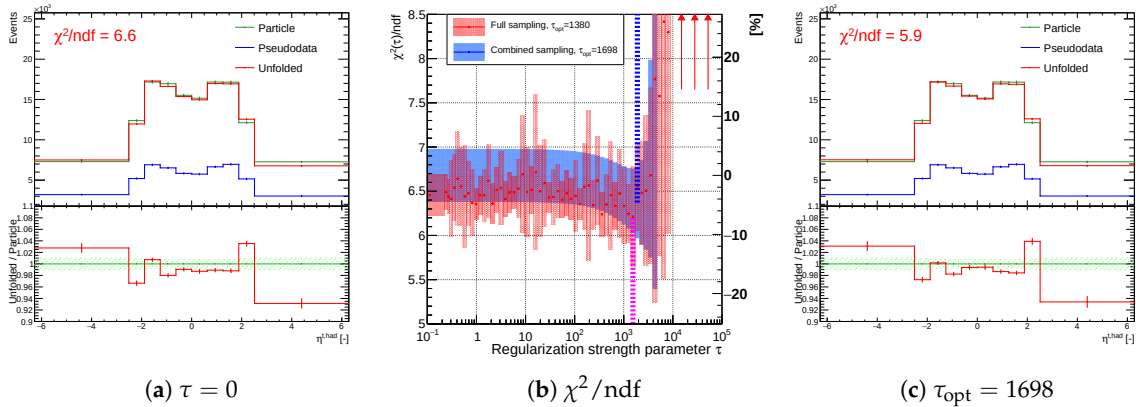


Figure 21. Result of unfolding (a) without regularization and (c) with regularization of the spectrum $\eta^{t,\text{had}}$ using the curvature regularization for one representative random seed. Variable $\chi^2(\tau)/\text{ndf}$ (b) while using full (red) and combined (blue) sampling of the $\eta^{t,\text{had}}$ spectrum with minimum at $\tau = 1698$ (curvature regularization). In this case, the regularization is not needed. The vertical dotted lines indicate positions of χ^2/ndf minima for each sampling case.

The results of all five spectra and their characteristics are summarized in Tables 1–3 in order to attempt finding a common feature that could predict the existence of real minima. Firstly, a characteristic of the migration matrix M was studied defined as

$$f_{\text{diag}} \equiv \frac{1}{N} \sum_i^N M_{ii} = \frac{1}{N} \text{Tr} M \quad (41)$$

where f_{diag} has the meaning of averaged diagonal elements. As a second characteristic, the correlation factor ρ is computed while using the ROOT's TH2D class and the GetCorrelationFactor() method [9].

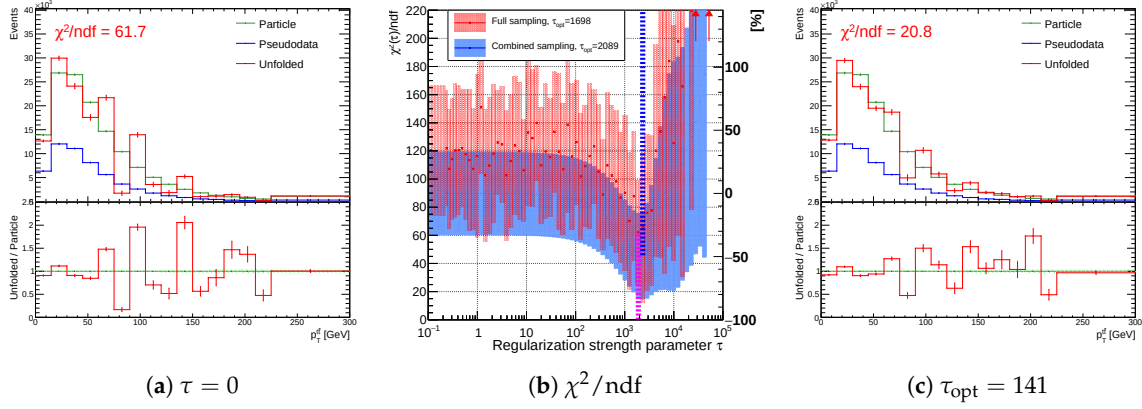


Figure 22. Result of unfolding (a) without regularization and (c) with regularization of the spectrum $p_T^{t\bar{t}}$ using entropy regularization for one representative random seed. Variable $\chi^2(\tau)/\text{ndf}$ (b) using full (red) and combined (blue) sampling of the $p_T^{t\bar{t}}$ spectrum with minimum at $\tau = 2089$ (entropy regularization). The vertical dotted lines indicate the positions of χ^2/ndf minima for each sampling case.

Table 1. Results of minima type and basic characteristics of the migration matrix M : averaged on-diagonal factor f_{diag} and correlation matrix ρ .

Spectrum	Type	Minimum	Method	f_{diag}	ρ
$m^{t\bar{t}}$	falling	Accidental	Curvature	0.51	0.92
$p_T^{t,\text{had}}$	falling	Hidden	Derivative	0.48	0.86
$\eta^{t,\text{had}}$	double-peaked	Real	Curvature	0.75	0.93
$\eta^{t\bar{t}}$	double-peaked	Real	Curvature	0.49	0.94
$p_T^{t\bar{t}}$	falling	Real	Entropy	0.29	0.86

Another characteristic that is shown in the Table 2 corresponds to the shape of the unfolded spectra. The curvature C_0 , entropy E_0 and derivatives D_0 of unfolded spectra without regularization are given relative to the curvature C_{ptcl} , entropy E_{ptcl} , and derivatives D_{ptcl} of the truth particle spectra.

Table 2. Relative curvature, entropy and derivatives of the unfolded spectra ($\tau = 0$) with respect to the curvature, entropy, and derivatives of the particle level spectra.

Spectrum	Minimum	$\frac{C_0}{C_{\text{ptcl}}}$	$\frac{E_0}{E_{\text{ptcl}}}$	$\frac{D_0}{D_{\text{ptcl}}}$
$m^{t\bar{t}}$	Accidental	1.1	1.0	2.7
$p_T^{t,\text{had}}$	Hidden	0.98	1.0	11
$\eta^{t,\text{had}}$	Real	1.1	1.0	0.96
$\eta^{t\bar{t}}$	Real	3.7	1.0	3.5
$p_T^{t\bar{t}}$	Real	14	0.97	11

Table 3. Time needed to produce $\chi^2/\text{ndf}(\tau)$ curve using combined sampling T_{comb} and full sampling T_{full} for 68 points of τ in the region $[0; 10^5]$. Time in seconds is rounded to hundreds.

Spectrum	Minimum	$T_{\text{comb}}[\text{s}]$	$T_{\text{full}}[\text{s}]$	$\frac{T_{\text{comb}}}{T_{\text{full}}}$
$m^{t\bar{t}}$	Accidental	900	30,700	0.029
$p_T^{t,\text{had}}$	Hidden	1200	41,300	0.030
$\eta^{t,\text{had}}$	Real	600	14,500	0.039
$\eta^{t\bar{t}}$	Real	1000	33,700	0.029
$p_T^{t\bar{t}}$	Real	1300	44,700	0.030

9. Discussion

We studied a method of finding an optimal regularization strength parameter τ and developed an independent implementation of two methods of adding regularization terms to the Fully Bayesian Unfolding and proposing a fast combined sampling method for finding an optimal value of the regularization strength parameter τ . The C++ version of the Fully Bayesian Unfolding was implemented into a private modification to RooUnfold [10], which is a widely used unfolding package in HEP.

The recommended option for the user is to use the combined sampling algorithm, which returns the $\chi^2/\text{ndf}(\tau)$ curve in a shorter time than the regular full sampling method, see Table 3.

The time that is needed to produce $\chi^2(\tau)$ curve while using the full sampling depends on the number of points in τ the user wishes to obtain (here, 67 points $\times$ three methods), while, using the combined method, the sampling algorithm is only run four times: first, to sample the log-likelihood function without the regularization part and afterwards three times for each of the regularization functions (curvature, entropy, and derivative), which roughly explains the time-saving ratio $4:(3 \times 67) \simeq 0.02$.

While no clear relation between the spectra characteristics in Tables 1 and 2 and real minima of χ^2/ndf was observed, the user is encouraged to evaluate χ^2/ndf , χ_{num} , and χ_{denom} him/herself in order to convince oneself of the validity and usefulness of the regularization.

While we provided an example for the spectra of reconstructed top quarks that are produced in pp collisions, the applicability of the method is not at all unique to spectra in high energy physics. Any inverse problem with finite migrations and a large number of bins is a potential candidate for benefitting from the fast sampling regularization that is introduced for the FBU method.

Author Contributions: Both authors contributed equally to this work. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors gratefully acknowledge the support from the project of the Czech Science Foundation project GAČR 19-21484S as well as the support from the project IGA_PrF_2020_007 of the Palacky University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Choudalakis, G. Fully Bayesian Unfolding. *arXiv* **2012**, arXiv:1201.4612.
2. Prosper, H.B.; Lyons, L. *PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*; CERN: Geneva, Switzerland, 2011. [CrossRef]
3. Alwall, J.; Frederix, R.; Frixione, S.; Hirschi, V.; Maltoni, F.; Mattelaer, O.; Shao, H.S.; Stelzer, T.; Torrielli, P.; Zaro, M. The automated computation of tree-level and next-to-leading order differential cross-sections, and their matching to parton shower simulations. *J. High Energy Phys.* **2014**, 2014, 79. [CrossRef]

4. Sjöstrand, T.; Ask, S.; Christiansen, J.R.; Corke, R.; Desai, N.; Ilten, P.; Mrenna, S.; Prestel, S.; Rasmussen, C.O.; Skands, P.Z. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.* **2015**, *191*, 159–177. [[CrossRef](#)]
5. De Favereau, J.; Delaere, C.; Demin, P.; Giammanco, A.; Lemaitre, V.; Mertens, A.; Selvaggi, M. DELPHES 3 Collaboration. DELPHES 3: A modular framework for fast simulation of a generic collider experiment. *J. High Energy Phys.* **2014**, *2014*, 57. [[CrossRef](#)]
6. Kvita, J. Study of methods of resolved top quark reconstruction in semileptonic $t\bar{t}$ decay. *Nucl. Instrum. Methods Phys. Res. Sect. A* **2018**, *900*, 84–100. [[CrossRef](#)]
7. Hoecker, A.; Kartvelishvili, V. SVD approach to data unfolding. *Nucl. Instrum. Methods Phys. Res. Sect. A* **1996**, *372*, 469–481. [[CrossRef](#)]
8. Matthew, D. Hoffman and Andrew Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623. Available online: <http://jmlr.org/papers/v15/hoffman14a.html> (accessed on 13 November 2020).
9. Brun, R.; Rademakers, F. ROOT—An Object Oriented Data Analysis Framework. *Nucl. Instrum. Methods Phys. Res. Sect. A* **1997**, *389*, 81–86. Available online: <http://root.cern.ch/> (accessed on 13 November 2020). [[CrossRef](#)]
10. Abye, T. CERN-2011-006. In *PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*; Prosper, H.B., Lyons, L., Eds.; CERN: Geneva, Switzerland, 2011; pp. 313–318.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).