SIAM/ASA J. Uncertainty Quantification
Vol. 4, pp. 1345–1371

Convergence and Error Propagation Results on a Linear Iterative Unfolding Method*

András László†

Abstract. Unfolding problems often arise in the context of statistical data analysis. Such problematics occur when the probability distribution of a physical quantity is to be measured, but it is randomized (smeared) by some well-understood process, such as a nonideal detector response or a well-described physical phenomenon. In such case it is said that the original probability distribution of interest is folded by a known response function. The reconstruction of the original probability distribution from the measured one is called unfolding. That technically involves evaluation of the nonbounded inverse of an integral operator over the space of L^1 functions, which is known to be an ill-posed problem. For the pertinent regularized operator inversion, we propose a linear iterative formula and provide proof of convergence in a probability theory context. Furthermore, we provide formulae for error estimates at finite iteration stopping order which are of utmost importance in practical applications: the approximation error, the propagated statistical error, and the propagated systematic error can be quantified. The arguments are based on the Riesz–Thorin theorem mapping the original L^1 problem to L^2 space, and subsequent application of ordinary L^2 spectral theory of operators. A library implementation in C of the algorithm along with corresponding error propagation is also provided. A numerical example also illustrates the method in operation.

Key words. unfolding, convergence, error propagation, probability theory, statistics, functional analysis, Riesz–Thorin theorem

AMS subject classifications. 46E30, 46E27, 62H99

DOI. 10.1137/15M1035744

1. Introduction.

In analysis of experimental data one commonly faces the problem that the probability density function (pdf) of a given physical quantity of interest is to be measured, but some random physical process, such as the intrinsic behavior of the measurement apparatus, smears it. The reconstruction of the pertinent unknown pdf of interest based on the observed smeared pdf and on the known response function of the measurement procedure is called unfolding.

More specifically, one of the most common unfolding scenarios turning up in experimental data analysis is the following. Let $x \mapsto f(x)$ be the unknown pdf which we intend to reconstruct, let $(y, x) \mapsto \rho(y|x)$ be the known response function of the smearing effect, and we assume that $y \mapsto g(y) = \int \rho(y|x)\, f(x)\, \mathrm{d}x$ is the measured pdf after the smearing effect,

*Received by the editors August 18, 2015; accepted for publication (in revised form) September 30, 2016; published electronically November 29, 2016.
http://www.siam.org/journals/juq/4/M103574.html
Funding: This work was supported in part by the Momentum ("Lendület") program of the Hungarian Academy of Sciences under grant LP2013-60. The author also received support from a János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

†Wigner Research Centre for Physics, Konkoly-Thege M.u. 29-33, Budapest, H-1121, Hungary (laszlo.andras@wigner.mta.hu).

called folding. In practice, actually often only a statistical estimator of g can be measured. Or, putting it differently, g often contains an additional error term $y \mapsto e(y)$ originating from statistical counting and unaccounted systematic measurement distortions, in which case one has $y \mapsto g(y) = \int \rho(y|x)\, f(x)\, \mathrm{d}x + e(y)$ as the measured pdf estimator. The task of unfolding is to provide some close estimate for $x \mapsto f(x)$, given $y \mapsto g(y)$ and $(y, x) \mapsto \rho(y|x)$ along with some estimate on $y \mapsto e(y)$, i.e., to solve the above linear integral equation. It is quite well known in the literature that such a problem is numerically ill-posed. The primary reason for this is Banach's closed graph theorem: due to the pertinent theorem a generic folding operator maps certain distant pdfs to close ones whose differences after the folding are shadowed by the contribution of the measurement error term e. That quite well understood phenomenon is summarized, e.g., in [1, 2, 3, 4, 5, 6, 7, 8].

The problematics of unfolding can also be formulated using a language possibly more familiar to statisticians [9, 10]. Let $x_1, \ldots, x_n$ be statistical instances of a probability variable x, i.e., independent identically distributed random variables, each having the same but unknown pdf f. In the experimental setting, merely the random variables $y_i = x_i + \varepsilon_{x_i,i}$ $(i = 1, \ldots, n)$ are observed, i.e., the original x_i $(i = 1, \ldots, n)$ random variables corrupted by an x-dependent, but otherwise independent identically distributed error variable ε_x, having a known x-dependent pdf $\varepsilon_x \mapsto \rho(\varepsilon_x + x|x)$ for each fixed value of x as a condition. Given all these, the task of unfolding is to provide an estimator for the pdf f of the undistorted probability variable x. In some real experimental situation, it also happens that the individual observed samples $y_i = x_i + \varepsilon_{x_i,i}$ $(i = 1, \ldots, n)$ are not published, only their pdf estimator g is made available, for instance, because there is some correction procedure on the pdf level, e.g., for inefficiencies. Also, our model $(y, x) \mapsto \rho(y|x)$ for the response function might be systematically inaccurate, for which inaccuracy only an upper bound might be known. Therefore, often not the sample based observational model but rather the previously discussed pdf estimator based observational model is more practical to handle. But whichever way the problem is formulated—based on individual samples or on pdfs—the task remains ill-posed.

In order to overcome the ill-posedness of the unfolding problem, all the methods use restrictions on the unknown pdf, and in some special cases properties of the response function can also be used to improve the situation. For instance, in the field of image or signal processing, the shape of the response function is translationally invariant in an exact manner, i.e., $\forall\, x, y, z$ one has $\rho(y|x + z) = \rho(y - z|x)$, and thus the unfolding reduces to the problematics of deconvolution. In the language of statistical samples, this would correspond to the observational model when $y_i = x_i + \varepsilon_i$ $(i = 1, \ldots, n)$ are observed, with independent identically distributed random variables ε_i of a known distribution, not depending on x. Due to the applicational importance of the special case of deconvolution problems, that branch has a whole stream of literature [9, 10, 11, 12, 13, 14, 15, 16]. The statistical deconvolution methods heavily rely on the applicability of convolution theorem for the Fourier transformed pdfs, which is possible due to the translational invariance of the shape of the response function, i.e., it relies on the fact that the probability variables ε_i $(i = 1, \ldots, n)$ are independent identically distributed and are independent from x. The ill-posedness of the problem, similarly to the case of any generic unfolding method, is regularized by finding and approximative solution. The optimal approximation is controlled by the application of the minimax principle: for a given estimate of the true deconvolved pdf, a loss (penalty) function is defined, and the minimum of the worst

case expected loss is looked for as a function of the regularization parameters. It is worth noting that most of the advanced statistical deconvolution methods can work on unbinned samples, i.e., they do not need an a priori histograming of the observed data. In section 6 an illustrative numerical unfolding toy model application is presented, which also tries to clarify that in an experimental context more general approaches than deconvolution are also needed in order to handle real measurement situations.

Also in the case of generic—i.e., nondeconvolution—unfolding problems a regularization method must be applied [1, 3, 4, 5, 6, 7, 8, 16, 17, 18, 19] and an approximate solution of the folding integral equation within a reduced set of allowed pdfs is searched for. The approximation is controlled by some regularization parameters whose particular value brings in a certain degree of arbitrariness to the unfolded pdf (approximation error), which is often difficult to quantify. There are basically three main widespread ways in the literature addressing the problem of regularization.

(i) In certain data analysis problems a parametric ansatz for the unknown pdf f is justified. In that case, one can construct the folded version of f by the response function ρ numerically, and that can be fitted to the observed folded pdf g, for instance, via a maximum likelihood method. Such method is used, for instance, in inclusive particle identification in experimental high energy particle physics (see, for instance, [20]). Due to the ill-posedness of the unfolding problem, one may run into a situation in particular cases when the fit is insensitive to some details of the parametrically given f. In other words, the log-likelihood function (χ^2) may be flat in the direction of certain parameters of the ansatz for f.

(ii) Bin-by-bin fitting of the histogramed f, such that when numerically folding it by ρ the result gets close to the observed folded pdf g, e.g., in a maximum likelihood sense. This is very similar to approach (i) with every bin amplitude of the histogramed f being a fit parameter. This method is basically equivalent to the naive inversion of the discretized folding operator as a matrix. Due to the ill-posedness of the unfolding problem, this is not satisfactory in itself. The usual procedure is to add some artificial penalty function to the log-likelihood function (χ^2) in order to suppress the large local gradients. If that is performed, the method can deliver meaningful answers, but the introduced systematic bias by the additional penalty function is difficult to quantify. In addition, similarly to the method (i), the fit can be slightly insensitive to the details of f due to the ill-posedness of the problem. The so-called SVD methods [17] are implementations of this idea.

(iii) There are also iterative methods which intend to approximate the true pdf f, given the measured folded pdf g and the response function ρ. One of the most popular and most promising methods is the method of convergent weights, also called iterative Bayesian unfolding. It was first discovered and applied by Richardson [21] and Lucy [22] for image processing. Later it was rediscovered and applied to tomography problems by Shepp and Vardi [23] and by Kondor [24]. The first serious mathematical scrutiny of the method was done by Mülthei and Schorr [25, 26]. In the mid 1990s d'Agostini rediscovered and popularized the algorithm in the high energy physics community [18]. Recently, Zech [19] studied possible optimal iteration stopping criteria for the algorithm. One of the main advantages of the method of convergent weights or

Bayesian unfolding is that it takes into account the nonnegativity and the unitness of the integral of the true pdf f in an exact manner. Furthermore, if the measured folded pdf g is a histogram, i.e., its values fluctuate according to Poisson counting statistics, then the iterative approximants to f have increasing likelihood [25], i.e., the algorithm is a realization of a maximum likelihood approximation. Most unfortunately, despite the research efforts [25], there are no results stating that the method is convergent, although numerical evidence suggests its convergent nature. Moreover, there are no exact error propagation formulae available.

In case of a consistent method the approximation error should converge to zero when the regularization parameters are relaxed. In case of an iterative method, an approximating sequence $(f_N)_{N\in\mathbb{N}_0}$ to the unknown f is constructed and the regularization parameter is merely the iteration stopping order $N_{\max}$, i.e., a threshold index in the approximating sequence. When an iterative unfolding method is consistent, the approximation error, i.e., the distance of f_N to the true unknown f, must converge to zero with increasing number of iterations N. Although the above consistency property is an obvious minimal requirement for any unfolding method, often this is not easy to show analytically.

In a previous paper [1] we proposed a linear iterative unfolding method, discussed its pros and cons in comparison to other techniques, provided a detailed description from the practical point of view for experimentalists, and provided a set of relevant application examples. In the present paper we provide formal mathematical proofs for the claims therein for the proposed unfolding method:

(i) proof of consistency, i.e., that the approximation error converges to zero with increasing number of iterations,
(ii) explicit formula for the approximation error at finite iteration order,
(iii) explicit formula for the propagated statistical errors on the unfolded pdf at finite iteration order given the statistical errors of the measured folded pdf,
(iv) explicit formula for the propagated systematic errors on the unfolded pdf at finite iteration order given the systematic errors of the measured folded pdf or of the response function.

Because of (ii)–(iv) the competing error terms become calculable, and therefore these can be used to define an optimal iteration stopping criterion. In addition, the pertinent error terms can be determined at this optimum. The quantification of these are of utmost importance when presenting unfolded experimental results and is generally an unresolved task for other widely used unfolding methods. The key mathematical ingredient of the proofs is mapping our originally L^1 problem to the L^2 space using the Riesz–Thorin theorem and using spectral representation of the operators therein. The actual iteration formula is formally motivated by a preconditioned Neumann–Landweber–Richardson series, but these are not automatically convergent in the case of L^1 problems: our specific preconditioning makes the iteration convergent in the L^1 setting, given some quite generic conditions. The proposed method also does not rely on an inherent discretization of the pdfs: it does work also in the continuum limit or with any type of density estimators.[1]

[1]Some unfolding methods rely on an inherent discretization of pdfs in the problem and use the assumed discretization as an implicit regularization. Our method does not use such a trick.

The obtained results can be particularly interesting as the proposed method can be considered as the "linearized" version of the method of convergent weights or iterative Bayesian unfolding [18, 19, 21, 22, 23, 24, 25, 26]. By understanding the convergence conditions and error propagation for the proposed method, the studies of Mülthei and Schorr [25] could eventually be completed on the Bayesian iteration, which would be a significant improvement in the field.

The paper is organized as follows. In section 2 the problem of unfolding is introduced in a mathematically rigorous way, and the basic properties of generic folding operators are discussed. In section 3 our proposed unfolding method is introduced and proofs are provided for its above listed properties. In section 4 we generalize a bit our results for the case of probability measures which are not described by pdfs. In section 5 we restrict our results to the special case when the unfolding problem is discrete: this presentation may be better understood by statisticians or experimental physicists not specialized in functional analysis. In section 6 a concrete numerical example is shown. Finally, in section 7 we summarize.

2. Mathematical properties of folding operators and the unfolding. In the text cpdf is the notion of conditional probability density function. We shall rely on the usual terminology in functional analysis and measure theory [27, 28]. As such, the notion of Lebesgue almost everywhere or Lebesgue almost every shall be abbreviated by a.e.

Let X and Y be finite dimensional real vector spaces equipped with the Lebesgue measure—unique up to a global positive normalization factor. Let $L^1(X)$ and $L^1(Y)$ denote the Banach spaces of $X \to \mathbb{C}$ and $Y \to \mathbb{C}$ Lebesgue integrable function equivalence classes, respectively, where the equivalence of functions is defined by being a.e. equal. As usual in functional analysis texts, we shall call these function equivalence classes simply functions. We shall also use the notion of essential bound for such a function which is the smallest upper bound valid a.e.

Definition 2.1. *Let $\rho : Y \times X \to \mathbb{R}_0^+$, $(y, x) \mapsto \rho(y|x)$ be a cpdf over the product space $Y \times X$, i.e., a nonnegative Lebesgue measurable function which satisfies $\forall x \in X : \int \rho(y|x)\,\mathrm{d}y = 1$. Then, the linear operator*

$$A_\rho \,:\, L^1(X) \to L^1(Y),\ (x \mapsto f(x)) \,\mapsto\, \left(y \mapsto \int \rho(y|x)\, f(x)\,\mathrm{d}x\right) \tag{2.1}$$

is called the folding operator by ρ, where the function ρ is called the response function of the folding.

Remark 2.1. The following basic properties of folding operators are direct consequences of the definition.

(i) A possible usual generalization of the notion of folding operator is when inefficiencies are also allowed, i.e., the less restrictive condition $\forall x \in X : \int \rho(y|x)\,\mathrm{d}y \leq 1$ is required for the response function ρ of the folding operator A_ρ. The results throughout the paper are also valid for that case.

(ii) By Fubini's theorem, a folding is a well-defined linear operator.

(iii) It is also quite evident [2] that such operator is continuous in the L^1 operator norm (i.e., in the probabilistic sense); moreover $\|A_\rho\|_{L^1(X)\to L^1(Y)} = 1$, while $\|A_\rho\|_{L^1(X)\to L^1(Y)} \leq 1$ whenever inefficiencies are allowed.

It is seen that such a folding operator A_ρ is quite well behaved: it is linear and is continuous in the probabilistic sense, i.e., close pdfs are mapped to close pdfs in the L^1 sense [1].

A quite important class of folding operators is convolutions, in which case the shape of the response function is translationally invariant.

Definition 2.2. *A folding operator A_ρ is called convolution whenever the response function ρ is translationally invariant in the sense that $Y = X$ and $\forall x, y, z \in X : \rho(y|x+z) = \rho(y-z|x)$.*

Remark 2.2. The following properties of convolution operators are well-known results [2, 29, 30].

(i) In case a folding operator A_ρ is a convolution, the response function ρ may be expressed by the single pdf $\eta := \rho(\cdot|0)$ in the form $\forall x, y \in X : \rho(y|x) = \eta(y-x)$. The alternative notation $\eta \star f := A_\rho f$ is often used in such case ($f \in L^1(X)$). Note that convolution is commutative, i.e. one has $\eta \star f = f \star \eta \ \forall \eta, f \in L^1(X)$.

(ii) A convolution operator is not onto, and its image is not closed.

(iii) The image of a convolution operator is dense if and only if the Fourier transform of the convolver function is nowhere zero (Wiener's approximation theorem).

(iv) A convolution operator is one-to-one if and only if the Fourier transform of the convolver function is a.e. nonzero.

(v) Consequently, the inverse of a convolution operator, whenever it exists, cannot be continuous. This is because a convolution is everywhere defined on the closed set $L^1(X)$, it is continuous, and therefore it has closed graph by Banach's closed graph theorem, but since the inverse operator's domain is not closed, again by Banach's closed graph theorem, it cannot be continuous.

Since the convolution operators form a quite large example class of folding operators, we can state that a generic folding operator's inverse, whenever it exists, is not continuous. This finding is often referred to as follows: the inversion of a generic folding operator is ill-posed. The argument goes as follows: we have an unknown pdf f, a known response function ρ, and a measured pdf $g = A_\rho f + e$, where e represents a small measurement error term. Then, when one would set $A_\rho^{-1} g = f + A_\rho^{-1} e$, the error term e contains modes not in the domain of A_ρ^{-1}, in which case $A_\rho^{-1} e$ is not meaningful, or when approximated numerically, this term shall diverge. Note that even if all modes of e were in the domain of A_ρ^{-1}, the smallness of $A_\rho^{-1} e$ is not guaranteed even though e is small. The ill-posedness of a generic unfolding problem may also be stated as follows: if f_1 and f_2 are distant pdfs, then $g_1 := A_\rho f_1 + e_1$ and $g_2 := A_\rho f_2 + e_2$ may be close pdfs, i.e., we lose discrimination power on pdfs after a folding [1]. The presented argument also warns us against relying solely on the so-called closure test when verifying an unfolding algorithm: whenever some unfolding method gives some estimate $\hat{f}$ for the unknown pdf f, it is usually argued that $A_\rho \hat{f} \approx A_\rho f$ confirms the validity of the estimate $\hat{f}$. Clearly, in light of our observations this is not enough, as $\hat{f}$ may still be far from f in the probabilistic distance.

Due to the ill-posedness of the unfolding problem, any unfolding method needs to use some kind of regularization—some assumption on the original (unknown) pdf—and a way to search for an approximative solution depending on some regularization parameters. Furthermore, the convergence to the original pdf when relaxing these parameters can usually be achieved

only in some weak sense, not in the probabilistic norm of $L^1(X)$. The most commonly applied unfolding strategies are summarized in [1, 3, 4, 5, 6, 7, 8, 16, 17, 18, 19].

3. A linear iterative unfolding method. Since the folding equation (2.1) is linear, it is quite natural to try applying some iterative inversion methods known in functional analysis when approximating the true solution f. One such self-suggesting method is the Neumann series [27, 28], which guarantees that whenever for a continuous linear operator A over a Banach space one has $\|I - A\| < 1$ (I being the identity operator), then $A^{-1} = \sum_{n=0}^{\infty}(I - A)^n$, where the convergence holds in the operator norm. That convergence requirement, however, cannot be satisfied in case of a probability theory folding operator because for such an operator one has $\|I - A_\rho\|_{L^1 \to L^1} = 2$ as shown in [2]. The Richardson iteration, based on similar requirements, does not work for the same reason. Another evident choice would be the Landweber iteration [31] known in the theory of Fredholm integral equations [27, 28]. This assumes, in the first place, that the unknown function f and the result of the folding g resides in the space of square integrable functions $L^2(X)$, and furthermore that the response function ρ satisfies the regularity condition $\int\int \left|\rho(y|x)\right|^2 \mathrm{d}y\,\mathrm{d}x < \infty$. The latter regularity condition, unfortunately, is violated in case of a generic cpdf, contrary to the common belief in the literature.[2]

Despite the fact that neither the Neumann series nor the Richardson iteration nor the Landweber iteration can be directly applied to an unfolding problem, they provide a possible starting point. Motivated by these algorithms we proposed a linear iterative unfolding method for a probability theory context, i.e., for the L^1 space [1]. The section is continued by recalling notions necessary for studying the pertinent algorithm.

In the following we shall denote by $L^p(X)$ the Banach space of $X \to \mathbb{C}$ functions [27, 28] which are Lebesgue integrable of the pth power ($1 \le p \le \infty$). The special case $L^\infty(X)$ for $p = \infty$ is defined as the Banach space of the $X \to \mathbb{C}$ essentially bounded functions with their norm being the essential bound.

Remark 3.1. The argument in the following relies on some known results.

(i) The Riesz–Thorin theorem [32] states that if $1 \le q \le r \le \infty$ and $F \subset L^q(X) \cap L^r(X)$ is a dense linear subspace in both $L^q(X)$ and $L^r(X)$, and furthermore a linear operator $T : F \to L^q(X) \cap L^r(X)$ is bounded both in the $L^q(X)$ and $L^r(X)$ norm, then $\forall q \le p \le r$ values $F \subset L^p(X)$, it is dense in $L^p(X)$, $T[F] \subset L^p(X)$ and T is bounded in the $L^p(X)$ norm. Thus, T is uniquely extendable as an $L^p(X) \to L^p(X)$ bounded linear operator. In addition we have that

$$\|T\|_{L^p \to L^p} \le \max\left(\|T\|_{L^q \to L^q}, \|T\|_{L^r \to L^r}\right) \tag{3.1}$$

holds for the operator norms.

(ii) An important consequence of the Riesz–Thorin theorem is that a convolution operator $\eta \star (\cdot)$ by a function $\eta \in L^1(X)$ is well-defined and continuous in $L^p(X)$ $\forall 1 \le p \le \infty$ and its operator norm is bounded by $\|\eta\|_{L^1}$. This obviously holds for the $p = 1$ and

[2] It is evidently seen that this regularity condition does not hold for any convolution. It is also seen at the price of some calculation that this situation cannot be repaired by a compactification mapping, i.e., if we map the support set of our pdfs and response function into a compact region of Y and X.

$p = \infty$ case due to Hölder's inequality, and then it is implied $\forall 1 < p < \infty$ as well by the pertinent theorem. As a consequence, using the commutativity of convolution, it also follows that if $\varphi \in L^p(X)$ and $\eta \in L^1(X)$, then $\varphi \star \eta \in L^p(X)$, i.e., pdfs may be mapped into $L^p(X)$ via convolution by pdfs integrable on the pth power.

(iii) We shall use in the following the spectral representation [28] of normal operators over complex separable Hilbert spaces. Let T be a normal operator over the pertinent space, i.e., a densely defined linear operator with closed graph, satisfying $T^*T = TT^*$, $(\cdot)^*$ being the adjoint. Then there exists a unique projection valued measure P over the Borel sets of the spectrum set of T, $\mathrm{Sp}(T)$, such that

$$T = \int_{\lambda \in \mathrm{Sp}(T)} \lambda \, \mathrm{d}P(\lambda) \tag{3.2}$$

holds, where the integral is defined in the weak sense. That is, for all elements f, g in the Hilbert space one has a complex valued Borel measure $\langle f, P(\cdot)g \rangle$ such that

$$\langle f, Tg \rangle = \int_{\lambda \in \mathrm{Sp}(T)} \lambda \, \mathrm{d} \, \langle f, P(\lambda)g \rangle \, . \tag{3.3}$$

In addition, one has that if M is a polynomial, then $M(T)$ is also normal operator, and furthermore

$$M(T) = \int_{\lambda \in \mathrm{Sp}(T)} M(\lambda) \, \mathrm{d}P(\lambda) \tag{3.4}$$

is satisfied in the same sense.

Throughout the argument we will need the notion of transpose folding which is introduced below.

Definition 3.1. *If A_ρ is a folding operator such that the response function $\rho(\cdot|x)$ is square-integrable $\forall x \in X$, then $\forall k \in L^2(Y)$ the expression*

$$A_\rho^T k := \left(x \mapsto \int k(y) \, \rho(y|x) \, \mathrm{d}y \right) \tag{3.5}$$

is meaningful and defines a linear map from $L^2(Y)$ to the Lebesgue measurable functions $X \to \mathbb{C}$. We call the linear operator A_ρ^T the transpose folding.

3.1. The iterative approximation. Equipped with the listed notions, we can introduce the following approximating sequence for solution of the unfolding problem. Let $g = A_\rho f$ be our unfolding problem where f is to be determined, with g and ρ being known. We try to approximate the solution in the form

$$\begin{aligned} K_\rho &:= \sup_{x \in X} \int \int \rho(y|z) \, \rho(y|x) \, \mathrm{d}y \, \mathrm{d}z, \\ f_0 &:= K_\rho^{-1} A_\rho^T g, \\ f_{N+1} &:= f_N + \left(f_0 - K_\rho^{-1} A_\rho^T A_\rho f_N \right) \\ &\qquad (N \in \mathbb{N}_0). \end{aligned} \tag{3.6}$$

This is, formally, the iterative expression for Neumann series after preconditioning by $K_\rho^{-1}A_\rho^T$, i.e., for the composite operator $K_\rho^{-1}A_\rho^T A_\rho$.

3.2. Convergence conditions. The following theorem shows that under quite generic conditions the approximating sequence $(f_N)_{N\in\mathbb{N}_0}$ in terms of (3.6) is well defined and converges to f whenever A_ρ is one-to-one, and it converges to the closest possible function to f whenever A_ρ is not one-to-one.

Theorem 3.2 (convergence). *Let A_ρ be a folding operator and assume that its response function ρ has the property that $\forall x \in X$ the function $\rho(\cdot|x)$ is square-integrable, and furthermore $K_\rho < \infty$. Assume that the unknown pdf f in the unfolding problem $g = A_\rho f$ is square-integrable. Then,*

(i) *for any compact set $U \subset X$,*

$$\lim_{N\to\infty} \frac{1}{\text{Volume}(U)} \int_{x\in U} \left(f - \mathcal{P}_{\text{Ker}(A_\rho)} f - f_N\right)(x)\,\mathrm{d}x = 0, \tag{3.7}$$

where $\mathcal{P}_{\text{Ker}(A_\rho)}$ is the L^2 orthogonal projection onto the kernel set of A_ρ;

(ii) *we have that*

$$\lim_{N\to\infty} \left\| f - \mathcal{P}_{\text{Ker}(A_\rho)} f - f_N \right\|_{L^2} = 0 \tag{3.8}$$

and the convergence is monotone.

Proof. It is seen that whenever the regularity condition $\forall x \in X : \rho(\cdot|x) \in L^2(Y)$ holds, the function

$$\alpha : X \times X \to \mathbb{R}_0^+,\ (z,x) \mapsto \alpha(z,x) := \int \rho(y|z)\rho(y|x)\,\mathrm{d}y \tag{3.9}$$

is well defined. By construction, it is symmetric, i.e., $\forall z, x \in X : \alpha(z,x) = \alpha(x,z)$. Furthermore, because of $K_\rho < \infty$ and symmetricity,

$$\sup_{x\in X} \int_{z\in X} \alpha(z,x)\,\mathrm{d}z = \sup_{z\in X} \int_{x\in X} \alpha(z,x)\,\mathrm{d}x = K_\rho < \infty \tag{3.10}$$

holds. With this, we see that the operator $A_\rho^T A_\rho$ is well defined as $L^1(X) \to L^1(X)$ and is bounded, its $L^1 \to L^1$ operator norm being K_ρ. This is because for any $f \in L^1(X)$

$$\begin{aligned} \left\|A_\rho^T A_\rho f\right\|_{L^1} &= \int \left| \int \alpha(z,x) f(x)\,\mathrm{d}x \right| \mathrm{d}z \\ &\le \int \int \alpha(z,x)\,|f(x)|\,\mathrm{d}x\,\mathrm{d}z = \int \left(\int \alpha(z,x)\,\mathrm{d}z \right) |f(x)|\,\mathrm{d}x \\ &\le \sup_{x\in X} \left(\int_{z\in X} \alpha(z,x)\,\mathrm{d}z \right) \int_{x\in X} |f(x)|\,\mathrm{d}x = K_\rho \left\|f\right\|_{L^1} \end{aligned} \tag{3.11}$$

due to monotonicity of integration, Fubini's theorem, and Hölder's inequality. It is also seen that the operator $A_\rho^T A_\rho$ is well defined as $L^\infty(X) \to L^\infty(X)$ and is bounded, its $L^\infty \to L^\infty$

operator norm being K_ρ. That is because for any $f \in L^\infty(X)$

$$\begin{aligned}\left\|A_\rho^T A_\rho f\right\|_{L^\infty} &= \sup_{z\in X}\left|\int \alpha(z,x) f(x)\,\mathrm{d}x\right| \\ &\le \sup_{z\in X}\int \alpha(z,x)\,|f(x)|\,\mathrm{d}x \le \sup_{z\in X}\left(\int \alpha(z,x)\,\mathrm{d}x \sup_{x\in X}|f(x)|\right) \\ &= \sup_{z\in X}\left(\int_{x\in X}\alpha(z,x)\,\mathrm{d}x\right)\sup_{x\in X}|f(x)| = K_\rho\,\|f\|_{L^\infty} \end{aligned} \tag{3.12}$$

due to monotonicity of integration and Hölder's inequality.

Now, using the Riesz–Thorin theorem we have that the operator $A_\rho^T A_\rho$ is well defined as $L^2(X) \to L^2(X)$ and is bounded, its $L^2 \to L^2$ operator norm being bound by K_ρ. It is also easily seen that for any $f \in L^2(X)$ one has $\langle f, A_\rho^T A_\rho f\rangle = \langle A_\rho f, A_\rho f\rangle \ge 0$; therefore it is a self-adjoint and positive operator in $L^2(X)$. Thus, its spectrum lies within the interval $[0, K_\rho]$. For brevity, we introduce the notation $A := K_\rho^{-1} A_\rho^T A_\rho$ for the renormalized composite folding operator.

Let us observe that the iterative formula (3.6) may also be written in the series expansion form $f_N = \sum_{n=0}^N (I-A)^n f_0$, where we have that $f_0 = Af$, f being the unknown pdf. This form is particularly useful because then we see by induction that $\sum_{n=0}^N (I-A)^n A = I - (I-A)^{N+1}$, i.e., we have the explicit formula $f - f_N = (I-A)^{N+1} f$ for the residual term.

By the observed properties of A it is quite evident that $\mathrm{Sp}(A) \subset [0,1]$. Thus, there exists a unique projection valued measure P on the Borel sets of $[0,1]$ such that

$$A = \int_{\lambda\in[0,1]} \lambda\,\mathrm{d}P(\lambda) \tag{3.13}$$

in the weak sense. This implies that for any $h \in L^2(X)$ we have

$$\begin{aligned}\langle h, f - f_N\rangle &= \int_{\lambda\in[0,1]} (1-\lambda)^{N+1}\,\mathrm{d}\,\langle h, P(\lambda) f\rangle \\ &= \int_{\lambda\in\{0\}} (1-\lambda)^{N+1}\,\mathrm{d}\,\langle h, P(\lambda) f\rangle \\ &\quad + \int_{\lambda\in]0,1]} (1-\lambda)^{N+1}\,\mathrm{d}\,\langle h, P(\lambda) f\rangle\,. \end{aligned} \tag{3.14}$$

Since $\int_{\lambda\in\{0\}} (1-\lambda)^{N+1}\,\mathrm{d}P(\lambda) = \mathcal{P}_{\mathrm{Ker}(A_\rho)} \forall N \in \mathbb{N}_0$, we arrive at the identity

$$\left\langle h, f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N\right\rangle = \int_{\lambda\in]0,1]} (1-\lambda)^{N+1}\,\mathrm{d}\,\langle h, P(\lambda) f\rangle\,, \tag{3.15}$$

and by the monotonicity of integration

$$\left|\left\langle h, f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N\right\rangle\right| \le \int_{\lambda\in]0,1]} |1-\lambda|^{N+1}\,\mathrm{d}\,|\langle h, P(\lambda) f\rangle| \tag{3.16}$$

also holds, where the symbol $|\cdot|$ when applied to complex valued measures denotes variation, which is analogous to absolute value of complex valued functions. The measure $\langle h, P(\cdot) f\rangle$ on

$[0,1]$ has finite variation and the function sequence $\lambda \mapsto (1-\lambda)^{N+1}$ $(N \in \mathbb{N}_0)$ is bounded independently of N and converges pointwise to zero on $]0,1]$; therefore by Lebesgue's theorem of dominated convergence [27, 28] we have that the sequence of integrals converges to zero. Thus, the first part of the theorem is proved by setting $h := \frac{1}{\text{Volume}(U)}\,\chi_U$.

The second part of the theorem is proved by observing that

$$\begin{aligned} \left\| f - \mathcal{P}_{\text{Ker}(A_\rho)} f - f_N \right\|_{L^2}^2 &= \left\langle f, \left((I-A)^{N+1} - \mathcal{P}_{\text{Ker}(A_\rho)} \right)^2 f \right\rangle \\ &= \int_{\lambda \in]0,1]} (1-\lambda)^{2N+2}\, \mathrm{d}\, \langle f, P(\lambda) f \rangle \,, \end{aligned} \tag{3.17}$$

where $\langle f, P(\cdot) f \rangle$ is a nonnegative valued finite measure and the integrand, which is also nonnegative, has a bound independent of N; furthermore it monotonically decreases at each point to zero with increasing N. Therefore, by Lebesgue's theorem of dominated convergence and by the monotonicity of integration we have that the pertinent expression converges to zero with increasing N in a monotonically decreasing way. ∎

Remark 3.2. The following remarks clarify the meaning of Theorem 3.2 in the context of a probability theory setting.

(i) For any folding operator A_ρ the response function may be conditioned to have the regularity condition $\forall x \in X : \rho(\cdot|x) \in L^2(X)$ by convolving it with a square-integrable pdf η whose Fourier transform is nowhere vanishing. Namely, one can solve the modified problem $\eta \star g = A_{\eta\star\rho} f$ for f instead of the original form $g = A_\rho f$. In that way, the transpose folding operator can always be made well defined. When such a treatment is applied, the iteration modifies as

$$\begin{aligned} K_{\eta\star\rho} &:= \sup_{x\in X} \int\int (\eta \star \rho)(y|z)\, (\eta \star \rho)(y|x)\, \mathrm{d}y\, \mathrm{d}z, \\ f_0 &:= K_{\eta\star\rho}^{-1} A_{\eta\star\rho}^T\, \eta \star g, \\ f_{N+1} &:= f_N + \left(f_0 - K_{\eta\star\rho}^{-1} A_{\eta\star\rho}^T A_{\eta\star\rho} f_N \right) \\ &\qquad (N \in \mathbb{N}_0) \end{aligned} \tag{3.18}$$

with the very same convergence properties as in the previous theorem.

(ii) The regularity condition $K_\rho < \infty$ (or $K_{\eta\star\rho} < \infty$) holds for a quite large class of response functions in a probability theory context. Namely, it is easy to check that if A_ρ is a convolution, then $K_\rho = 1$. For other practical cases, this condition may be checked numerically as done in [1]. It is shown, e.g., that for the response function of particle energy measurement with a typical calorimeter device, one has $K_\rho \approx 1.4$. Also the response function of particle momentum measurement using bending in a magnetic field has the pertinent regularity property.

(iii) The regularity condition for the unknown pdf f, i.e., that it has to be square-integrable, holds for a quite generic class of pdfs. This is automatic, for instance, for any pdf which is known to be essentially bounded.

(iv) When the convergence condition is satisfied, it is seen that if A_ρ is one-to-one, the approximating functions $(f_N)_{N\in\mathbb{N}_0}$ converge to the original unknown pdf f. When A_ρ is not one-to-one, then $(f_N)_{N\in\mathbb{N}_0}$ converge to the closest possible function $f - \mathcal{P}_{\text{Ker}(A_\rho)} f$.

(v) The meaning of convergence result (i) in the context of probability theory is that the approximating functions $(f_N)_{N\in\mathbb{N}_0}$ converge in the sense that the probability of each compact set $U \subset X$ is restored to the maximum possible extent, but the rate of convergence might be different for different sets. When the pdfs are measured or modeled by histograms, as is usual in statistical data processing, this means binwise convergence of the restored histograms, the convergence rate being possibly different for different histogram bins. The more global convergence result (ii) does not have a direct probability theory interpretation but shall have a role in the estimation of approximation error at finite iteration order N.

(vi) Note that whenever our pdfs are modeled by histograms, the operation of histogram binning may also be regarded as part of the folding operator as described in [1], and thus it is wise to include its effect in the folding operator A_ρ. This might be done, for instance, by modeling the true (unknown) pdf f and its iterative approximates f_N as histograms binned on a much wider domain with larger binning density than the measured pdf g. In such approximation the folding operator A_ρ may be thought of as a real matrix which is not square.

3.3. Estimation of approximation error. The convergence result means that the residual term (approximation error) $f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N$ of the approximating sequence defined by (3.6) decreases to zero with increased iteration order N in the sense that it decreases to zero when averaged over any compact set, i.e., we have binwise convergence in the language of histograms. However, it would be very useful to quantify the approximation error at finite N in order to define some stopping criterion. To achieve this, we need to recall a result from the theory of projection valued measures.

Remark 3.3. Let P be a projection valued measure of some separable Hilbert space over the Borel sets of $\mathbb{C}$. Then, whenever α and β are $\mathbb{C} \to \mathbb{C}$ measurable functions, while h and f are elements of the Hilbert space, one has

(3.19)
$$\left| \int_{\lambda\in\mathbb{C}} \alpha(\lambda)\,\beta(\lambda)\,\mathrm{d}\,\langle h, P(\lambda) f\rangle \right| \le \sqrt{\int_{\lambda\in\mathbb{C}} |\alpha(\lambda)|^2\,\mathrm{d}\,\langle h, P(\lambda) h\rangle} \sqrt{\int_{\lambda\in\mathbb{C}} |\beta(\lambda)|^2\,\mathrm{d}\,\langle f, P(\lambda) f\rangle}$$

and the same inequality also holds when α and β are interchanged [28]. This upper bound is in the analogy of the Cauchy–Schwarz inequality.

The following theorem helps to quantify the approximation error at a finite iteration order $N \in \mathbb{N}_0$.

Theorem 3.3 (approximation error). *Take the iterative solution for the unfolding problem as in* (3.6) *and assume that the convergence conditions of Theorem* 3.2 *hold. Then, the distance of an Nth iterate f_N from the closest possible function to the true unfolded pdf f in the average over a compact set $U \subset X$ has the following upper bounds:*

(i) *One has*

(3.20)
$$\left| \frac{1}{\mathrm{Volume}(U)} \int_{x\in U} \left(f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N\right)(x)\,\mathrm{d}x \right| \le \frac{1}{\sqrt{\mathrm{Volume}(U)}} \left\| f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N \right\|_{L^2}.$$

(ii) *Similarly, when* $\mathrm{Ker}(A_\rho)$ *is not projected out,*

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}(f-f_N)(x)\,\mathrm{d}x\right| \le \frac{1}{\sqrt{\mathrm{Volume}(U)}}\,\|f-f_N\|_{L^2}\,. \tag{3.21}$$

(iii) *In addition,*

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}\left(f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right)(x)\,\mathrm{d}x\right| \le \left\|f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f\right\|_{L^2}\,\|\xi_U - \mathcal{P}_{\mathrm{Ker}(A_\rho)}\xi_U-\xi_{U,N}\|_{L^2} \tag{3.22}$$

is valid, where $\xi_U := \frac{1}{\mathrm{Volume}(U)}\chi_U$ *and* $\xi_{U,N}$ *is the* N*th iterative approximation of* ξ_U *in terms of* (3.6)*. Namely,* $\xi_{U,0} := K_\rho^{-1}A_\rho^T\xi_U$ *and* $\xi_{U,N+1} := \xi_{U,N} + \left(\xi_{U,0} - K_\rho^{-1}A_\rho^TA_\rho\xi_{U,N}\right)$.

(iv) *Similarly, one has*

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}(f-f_N)(x)\,\mathrm{d}x\right| \le \|f\|_{L^2}\,\left\|\xi_U-\xi_{U,N}\right\|_{L^2} \tag{3.23}$$

when $\mathrm{Ker}(A_\rho)$ *is not projected out.*

(v) *The identity*

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}(f-f_N)(x)\,\mathrm{d}x\right| = \left|\int\left(\xi_U-\xi_{U,N}\right)(x)\,f(x)\,\mathrm{d}x\right| \tag{3.24}$$

also holds.

Proof. These are direct consequence of spectral representation of the operator $A := K_\rho^{-1}A_\rho^TA_\rho$ as in the proof of Theorem 3.2 from which

$$\begin{aligned}\left|\left\langle h, f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right\rangle\right| &= \left|\int_{\lambda\in]0,1]} 1\,(1-\lambda)^{N+1}\,\mathrm{d}\left\langle h,P(\lambda)f\right\rangle\right| \\ &\le \sqrt{\int_{\lambda\in]0,1]}|1|^2\,\mathrm{d}\left\langle h,P(\lambda)h\right\rangle}\,\sqrt{\int_{\lambda\in]0,1]}\left|(1-\lambda)^{N+1}\right|^2\,\mathrm{d}\left\langle f,P(\lambda)f\right\rangle}\end{aligned} \tag{3.25}$$

and

$$\begin{aligned}\left|\left\langle h, f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right\rangle\right| &= \left|\int_{\lambda\in]0,1]} 1\,(1-\lambda)^{N+1}\,\mathrm{d}\left\langle h,P(\lambda)f\right\rangle\right| \\ &\le \sqrt{\int_{\lambda\in]0,1]}|1|^2\,\mathrm{d}\left\langle f,P(\lambda)f\right\rangle}\,\sqrt{\int_{\lambda\in]0,1]}\left|(1-\lambda)^{N+1}\right|^2\,\mathrm{d}\left\langle h,P(\lambda)h\right\rangle}\end{aligned} \tag{3.26}$$

follows with arbitrary $h\in L^2(X)$. These may be rewritten as

$$\left|\left\langle h, f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right\rangle\right| \le \left\|h-\mathcal{P}_{\mathrm{Ker}(A_\rho)}h\right\|_{L^2}\left\|\left((I-A)^{N+1}-\mathcal{P}_{\mathrm{Ker}(A_\rho)}\right)f\right\|_{L^2} \tag{3.27}$$

and

$$\left|\left\langle h, f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N \right\rangle\right| \leq \left\|f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f\right\|_{L^2} \left\|\left((I-A)^{N+1} - \mathcal{P}_{\mathrm{Ker}(A_\rho)}\right) h\right\|_{L^2}. \tag{3.28}$$

Then by using the fact that $\left((I-A)^{N+1} - \mathcal{P}_{\mathrm{Ker}(A_\rho)}\right) f = f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N$ and $((I-A)^{N+1} - \mathcal{P}_{\mathrm{Ker}(A_\rho)})h = h - \mathcal{P}_{\mathrm{Ker}(A_\rho)} h - h_N$, where h_N is the iterative approximation of h in terms of (3.6), we see that

$$\left|\left\langle h, f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N \right\rangle\right| \leq \left\|h - \mathcal{P}_{\mathrm{Ker}(A_\rho)} h\right\|_{L^2} \left\|f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N\right\|_{L^2} \tag{3.29}$$

and

$$\left|\left\langle h, f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N \right\rangle\right| \leq \left\|f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f\right\|_{L^2} \left\|h - \mathcal{P}_{\mathrm{Ker}(A_\rho)} h - h_N\right\|_{L^2}. \tag{3.30}$$

By using $\left\|h - \mathcal{P}_{\mathrm{Ker}(A_\rho)} h\right\|_{L^2} \leq \|h\|_{L^2}$ and setting $h := \frac{1}{\mathrm{Volume}(U)} \chi_U$ we have proved (i) and (iii).

Quite obviously, the same argument can be repeated with the projection operator $\mathcal{P}_{\mathrm{Ker}(A_\rho)}$ excluded from the equations, which proves (ii) and (iv).

Point (v) is proved by observing that for any $h \in L^2(X)$ one has $\langle h, f - f_N \rangle = \left\langle h, (I-A)^{N+1} f \right\rangle$, since $f - f_N = (I-A)^{N+1} f$. Due to the self-adjointness of the composite folding operator A, one has that $\langle h, f - f_N \rangle = \left\langle (I-A)^{N+1} h, f \right\rangle$. Since the identity $(I-A)^{N+1} h = h - h_N$ holds, one arrives at $\langle h, f - f_N \rangle = \langle h - h_N, f \rangle$ and thus $|\langle h, f - f_N \rangle| = |\langle h - h_N, f \rangle|$ is valid. Then, (v) is proved by simply substituting $h := \xi_U$. ∎

Remark 3.4. The following remarks clarify the usability of Theorem 3.3.

(i) By statements (i) and (ii) it is implied that the residual error averaged over a compact set $U \subset X$ scales as $\frac{1}{\sqrt{\mathrm{Volume}(U)}}$. In the language of histograms it means that it scales as one per square-root of the histogram bin size.

(ii) The upper bounds (i), (iii) decrease monotonically to zero with increasing N. The upper bounds (ii) and (iv) decrease monotonically to the corresponding limits $\frac{1}{\sqrt{\mathrm{Volume}(U)}} \left\|\mathcal{P}_{\mathrm{Ker}(A_\rho)} f\right\|_{L^2}$ and $\|f\|_{L^2} \left\|\mathcal{P}_{\mathrm{Ker}(A_\rho)} \xi_U\right\|_{L^2}$, respectively. Since $\left\|\xi_U - \xi_{U,N}\right\|_{L^2}$ is fully calculable, upper bound (iv) can be used to test whether the inverse of A_ρ exists, i.e., whether $\mathcal{P}_{\mathrm{Ker}(A_\rho)} = 0$ holds, or if not, it may be used to quantify the contribution of the irrecoverable part $\mathcal{P}_{\mathrm{Ker}(A_\rho)} f$.

(iii) Via spectral representation it is easy to see that $\|f_N\|_{L^2}$ converges to the limit $\left\|f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f\right\|_{L^2}$ in a monotonically increasing way, i.e., may be used to approximate this unknown coefficient from below.

(iv) Again via using spectral representation, one can see that with fixed N and $M > N$, the expressions $\|f_M - f_N\|_{L^2}$ and $\left\|\xi_{U,M} - \xi_{U,N}\right\|_{L^2}$ tend to the corresponding limits $\left\|f - \mathcal{P}_{\mathrm{Ker}(A_\rho)} f - f_N\right\|_{L^2}$ and $\left\|\xi_U - \mathcal{P}_{\mathrm{Ker}(A_\rho)} \xi_U - \xi_{U,N}\right\|_{L^2}$ with increasing M, respectively, in a monotonically increasing way. Therefore, they can be used for approximation of these unknown coefficients from below.

(v) As a consequence, the approximation error may be estimated for a fixed iteration order N in the following way. For any $\varepsilon > 0$ there exists an iteration index threshold $M_{\varepsilon,N} > N$ such that $\forall M > M_{\varepsilon,N}$

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}\left(f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right)(x)\,\mathrm{d}x\right|\leq\frac{1}{\sqrt{\mathrm{Volume}(U)}}\,(1+\varepsilon)\,\|f_M-f_N\|_{L^2} \tag{3.31}$$

is valid. In addition, a closer, U-dependent estimate may be calculated: for any $\varepsilon>0$ there exists an iteration index threshold $M_{\varepsilon,U,N}>N$ for which $\forall M>M_{\varepsilon,U,N}$ the upper bound

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}\left(f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f-f_N\right)(x)\,\mathrm{d}x\right|\leq(1+\varepsilon)\,\|f_M\|_{L^2}\,\left\|\xi_{U,M}-\xi_{U,N}\right\|_{L^2} \tag{3.32}$$

holds. Alternatively,

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}\left(f-f_N\right)(x)\,\mathrm{d}x\right|\leq(1+\varepsilon)\,\|f_M\|_{L^2}\,\left\|\xi_{U}-\xi_{U,N}\right\|_{L^2} \tag{3.33}$$

is also valid whenever A_ρ is known to be one-to-one, which expression is slightly cheaper to calculate.

(vi) The identity (v) is particularly useful. In order to constructively evaluate it, one needs to use the fact that the sequence $(f_N)_{N\in\mathbb{N}_0}$ converges to $f-\mathcal{P}_{\mathrm{Ker}(A_\rho)}f$ in the L^2 sense. Thus, whenever A_ρ is invertible, it converges to f in the L^2 sense. In that case, the identity (v) can be rewritten as

$$\left|\frac{1}{\mathrm{Volume}(U)}\int_{x\in U}\left(f-f_N\right)(x)\,\mathrm{d}x\right|=\lim_{M\to\infty}\left|\int\left(\xi_U-\xi_{U,N}\right)(x)\,f_M(x)\,\mathrm{d}x\right|. \tag{3.34}$$

Technically, the right side of this identity may be approximated by the integral $\left|\int\left(\xi_U-\xi_{U,N}\right)(x)\,f_M(x)\,\mathrm{d}x\right|$ with large enough M. For large N, even $M:=N$ may be used for evaluation of this expression.

3.4. Estimation of statistical error. Armed with the approximation error estimates of Theorem 3.3 one can construct penalty functions which define optimal stopping criterion of the iteration, and one can quantify the error of the approximation at finite iteration order which decreases with increasing iteration order.

In practice, however, the unfolding problem $g=A_\rho f+e$ may also contain a small statistical error term e whose expectation value is zero; its exact value is unknown, but an estimate to the behavior of the random variable $e(x)$ for each $x\in X$ is available. Normally, the statistical covariance matrix $\mathrm{Cov}(e)$ is known along with the measured pdf g and the known response function ρ. If, for instance, g was a result of a measurement in the form of a histogram, then $\mathrm{Cov}(e)=\mathrm{Cov}(g)$ will be nothing but the diagonal matrix composed of the histogram bin entries. The question naturally arises: how can one quantify the propagated statistical error of the Nth iterative approximation of f, i.e., of f_N. In the following we show an exact formula for the case when g is measured as a histogram, i.e., when g can be regarded as an n-component vector of real probability variables with known covariance.

Remark 3.5. The following simple facts in probability theory will aid the argumentation of the statistical error propagation.

(i) If v is an n-component vector of real probability variables, then its covariance $\mathrm{Cov}(v)$ is an $n\times n$ real symmetric positive matrix. Therefore, for any $m\geq n$ there exists (not

necessarily uniquely) a real $n \times m$ matrix $\mathrm{Err}(v)$ such that

$$\mathrm{Cov}(v) = \mathrm{Err}(v)\mathrm{Err}(v)^T \tag{3.35}$$

holds, the symbol $(\cdot)^T$ denoting matrix transpose. Indeed, because of realness, symmetricity, and positivity of $\mathrm{Cov}(v)$ there exists uniquely a real symmetric positive $n \times n$ matrix satisfying (3.35), the square-root of $\mathrm{Cov}(v)$, and therefore $\mathrm{Err}(v) = \sqrt{\mathrm{Cov}(v)}$ may be chosen. Then, this may be extended to be $n \times m$ $(m \geq n)$ by zeros without affecting (3.35). In some special cases, however, there also exists such $n \times m$ $(m \leq n)$ real matrix $\mathrm{Err}(v)$ such that (3.35) still holds.

(ii) If v is an n-component vector of real probability variables and M is a real $m \times n$ matrix, then the standard error propagation formula

$$\mathrm{Cov}(Mv) = M\mathrm{Cov}(v)M^T \tag{3.36}$$

holds.

(iii) As a consequence of the previous observations, one can express the standard error propagation formula also in the form

$$\mathrm{Err}(Mv) = M\mathrm{Err}(v), \tag{3.37}$$

where $\mathrm{Err}(v)$ is any real $n \times n$ matrix satisfying (3.35), and the resulting real $m \times n$ matrix $\mathrm{Err}(Mv)$ shall obey $\mathrm{Err}(Mv)\mathrm{Err}(Mv)^T = \mathrm{Cov}(Mv)$.

(iv) In our unfolding problem the Nth iterative approximation of f, i.e., f_N, may be expressed in the form

$$f_N = \left(\sum_{n=0}^{N} \left(I - K_\rho^{-1} A_\rho^T A_\rho \right)^n \right) K_\rho^{-1} A_\rho^T g, \tag{3.38}$$

which is manifestly linear in the measured pdf g. This fact may be used in order to construct statistical error propagation formula in terms of the previous observations.

Armed with these equalities, we are ready to state the statistical error propagation formula for our unfolding method.

Theorem 3.4 (statistical error). *Take the iterative solution for the unfolding problem as in (3.6) and assume that the convergence conditions of Theorem 3.2 hold. Let* $\mathrm{Cov}(g)$ *be the* $n \times n$ *statistical covariance matrix of the measured pdf* g*, where* g *is given in the form of an* n*-bin histogram. If* f *and* f_N *are modeled as an* m*-bin histogram, then the* $m \times m$ *covariance matrix of* f_N*,* $\mathrm{Cov}(f_N)$*, may be obtained by the following iterative formula along with* f_N*:*

$$\begin{aligned} E_0 &:= K_\rho^{-1} A_\rho^T \mathrm{Err}(g), \\ E_{N+1} &:= E_N + \left(E_0 - K_\rho^{-1} A_\rho^T A_\rho E_N \right) \\ &\qquad (N \in \mathbb{N}_0), \end{aligned} \tag{3.39}$$

where $E_N {E_N}^T = \mathrm{Cov}(f_N)$ *holds for each* N*.*

Proof. This is a simple consequence of the linearity of the unfolding method (3.6) and of Remark 3.5(iv) combined with (iii) and then reexpressing it via iterative form. ∎

Remark 3.6. The following remarks add some pieces of information about the practical usage of the statistical error propagation theorem.

(i) If the measured pdf g is a histogram, then each component obeys Poisson distribution, and thus $\mathrm{Cov}(g) = \mathrm{diag}(g)$. Furthermore a real $n \times n$ matrix $\mathrm{Err}(g)$, satisfying $\mathrm{Err}(g)\mathrm{Err}(g)^T = \mathrm{Cov}(g)$, may be constructed by taking the componentwise square-root of $\mathrm{diag}(g)$. This can directly be used in calculation of E_0 in Theorem 3.4.

(ii) If f is modeled as a histogram with m bins, then for each iteration order N the real matrix E_N is of $m \times n$ type, i.e., $\mathrm{Cov}(f_N) = E_N {E_N}^T$ shall be of $m \times m$ type.

(iii) The square-root of the diagonal elements of the covariance matrix $\mathrm{Cov}(f_N)$ give the exact statistical errors of f_N, which then may be used to define an iteration stopping criterion, for instance, the sum of the statistical errors may be required to be under a predefined threshold. One should not forget, however, that this unfolding method—just as any other unfolding method—introduces pretty strong correlations and thus the nondiagonal elements of $\mathrm{Cov}(f_N)$ also play an important role when describing the characteristics of the statistical fluctuations of f_N.

3.5. Estimation of systematic error. It was shown that in the case of a statistical unfolding problem of the form $g = A_\rho f + e$ the quantification of the two competing error terms is possible: close upper bound to the convergent approximation error term was given, whereas exact error propagation formula to the divergent statistical error term was shown. A combination, such as the sum of these terms, may be considered as a penalty function and the iteration may be stopped when the penalty function is minimal; furthermore these terms may be quantified at this optimal iteration order with the shown formulae. In practice, however, one often faces the problem of systematic errors whenever the measured pdf contains some systematic distortion not accounted for in our model of response function, or equivalently, whenever our model of response function is slightly inaccurate. Formally we may write in such case that the actually measured pdf is $g + \delta g = A_{(\rho+\delta\rho)} f + e$, where $\delta\rho$ is the deviation of the true response function $\rho + \delta\rho$ from our model response function ρ. Since by definition $g = A_\rho f + e$ would be the measured pdf in the absence of $\delta\rho$, one arrives at the relation $\delta g = A_{\delta\rho} f$ between δg and $\delta\rho$. When applying the iterative solution (3.6) using ρ to the actually measured pdf $g + \delta g$, the Nth iterative estimate of the true unknown pdf f shall contain a propagated contribution δf_N which needs to be quantified. In experimental practice, the systematic error of the actually measured pdf is given in terms of some close upper estimate sg for which $|\delta g| \leq sg$ holds, or similarly as a close upper estimate $s\rho$ for which $|\delta\rho| \leq s\rho$ is valid. Our aim is to provide some upper estimate to $|\delta f_N|$ based on sg or $s\rho$, for any given iteration order $N \in \mathbb{N}_0$. For this, let us introduce the normalization factors

$$C_{\rho,sg} := \sqrt{\int \left(K_\rho^{-1} A_\rho^T sg\right)^2 (x)\, \mathrm{d}x} \tag{3.40}$$

if the systematic errors are known in terms of sg and

$$D_{\rho,s\rho} := \sqrt{\sup_{x\in X} \int\int \left(K_\rho^{-1} A_\rho^T s\rho\right)(y|z)\ \left(K_\rho^{-1} A_\rho^T s\rho\right)(y|x)\, \mathrm{d}y\, \mathrm{d}z} \tag{3.41}$$

if the systematic errors are known in terms of $s\rho$.

Theorem 3.5 (systematic error). *Take the iterative solution for the unfolding problem as in (3.6) and assume that the conditions of convergence hold. Then, the following upper bounds are valid on the systematic deviation δf_N of the Nth iterative approximation of f, f_N.*

(i) *For the average of δf_N over any compact set $U \subset X$ one has*

$$\left|\frac{1}{\text{Volume}(U)}\int_{x\in U}\delta f_N(x)\,\mathrm{d}x\right| \leq \left\|\Xi_{U,N}\right\|_{L^2} C_{\rho,sg}, \tag{3.42}$$

where $\xi_U := \frac{1}{\text{Volume}(U)}\chi_U$ and $\Xi_{U,N}$ is defined by the iteration

$$\begin{aligned}\Xi_{U,0} &:= \xi_U,\\ \Xi_{U,N+1} &= \Xi_{U,N} + \left(\Xi_{U,0} - K_\rho^{-1}A_\rho^T A_\rho \Xi_{U,N}\right)\\ &\qquad (N\in\mathbb{N}_0).\end{aligned} \tag{3.43}$$

(ii) *Alternatively,*

$$\left|\frac{1}{\text{Volume}(U)}\int_{x\in U}\delta f_N(x)\,\mathrm{d}x\right| \leq \left\|\Xi_{U,N}\right\|_{L^2} D_{\rho,s\rho}\left\|f\right\|_{L^2}. \tag{3.44}$$

(iii) *The upper bound*

$$\left|\frac{1}{\text{Volume}(U)}\int_{x\in U}\delta f_N(x)\,\mathrm{d}x\right| \leq \int \left|K_\rho^{-1}A_\rho\Xi_{U,N}\right|(y)\, sg(y)\,\mathrm{d}y \tag{3.45}$$

also holds.

(iv) *Alternatively,*

$$\left|\frac{1}{\text{Volume}(U)}\int_{x\in U}\delta f_N(x)\,\mathrm{d}x\right| \leq \int \left(K_\rho^{-1}A_{s\rho}^T\left|A_\rho\Xi_{U,N}\right|\right)(x)\ |f|\,(x)\,\mathrm{d}x. \tag{3.46}$$

(v) *More specifically,*

$$\left|\frac{1}{\text{Volume}(U)}\int_{x\in U}\delta f_N(x)\,\mathrm{d}x\right| \leq \|f\|_{L^1}\ \sup_{x\in X}\left(K_\rho^{-1}A_{s\rho}^T\left|A_\rho\Xi_{U,N}\right|\right)(x). \tag{3.47}$$

Here, whenever f is a pdf, then $\|f\|_{L^1} = 1$ automatically holds.

Proof. We begin the proof by recalling that because of (3.38) and its modified form

$$f_N + \delta f_N = \left(\sum_{n=0}^{N}\left(I - K_\rho^{-1}A_\rho^T A_\rho\right)^n\right) K_\rho^{-1}A_\rho^T(g+\delta g) \tag{3.48}$$

in the presence of systematic distortions, we have that

$$\delta f_N = \left(\sum_{n=0}^{N}\left(I - K_\rho^{-1}A_\rho^T A_\rho\right)^n\right) K_\rho^{-1}A_\rho^T\delta g \tag{3.49}$$

holds, where δg is the unaccounted systematic distortion of the measured pdf, which is related to the unaccounted systematic distortion of the response function $\delta\rho$ by $\delta g = A_{\delta\rho}f$.

Again, we use the notation $A := K_\rho^{-1} A_\rho^T A_\rho$ and use its spectral representation as in the proof of Theorem 3.2. With this, one has

$$\langle h, \delta f_N \rangle = \int_{\lambda\in[0,1]} 1 \sum_{n=0}^{N} (1-\lambda)^n \, \mathrm{d}\, \langle h, P(\lambda) K_\rho^{-1} A_\rho^T \delta g \rangle \tag{3.50}$$

for any $h \in L^2(X)$. From that, using Remark 3.3 we arrive at

$$\begin{aligned} |\langle h, \delta f_N \rangle| &\le \sqrt{\int_{\lambda\in[0,1]} \left| \sum_{n=0}^{N} (1-\lambda)^n \right|^2 \mathrm{d}\, \langle h, P(\lambda) h \rangle} \sqrt{\int_{\lambda\in[0,1]} |1|^2 \, \mathrm{d}\, \langle K_\rho^{-1} A_\rho^T \delta g, P(\lambda) K_\rho^{-1} A_\rho^T \delta g \rangle} \\ &= \left\| \sum_{n=0}^{N} (I-A)^n h \right\|_{L^2} \left\| K_\rho^{-1} A_\rho^T \delta g \right\|_{L^2} = \|H_N\|_{L^2} \left\| K_\rho^{-1} A_\rho^T \delta g \right\|_{L^2}, \end{aligned} \tag{3.51}$$

where the notation $H_N := \sum_{n=0}^{N} (I-A)^n h$ is introduced. It is quite evident that H_N may be calculated using the iterative form

$$\begin{aligned} H_0 &:= h, \\ H_{N+1} &:= H_N + (H_0 - A H_N) \\ &\qquad (N \in \mathbb{N}_0) \end{aligned} \tag{3.52}$$

in order to evaluate $\|H_N\|_{L^2}$.

An upper bound for $\left\| K_\rho^{-1} A_\rho^T \delta g \right\|_{L^2}$ may be readily constructed using the inequality

$$\left\| K_\rho^{-1} A_\rho^T \delta g \right\|_{L^2}^2 \le \left\| K_\rho^{-1} A_\rho^T s g \right\|_{L^2}^2 = C_{\rho, sg}^2, \tag{3.53}$$

which is seen to hold using Fubini's theorem and monotonicity of integration, where nonnegativity of ρ and sg is tacitly assumed as previously.

Now, by setting $h := \xi_U$, part (i) of the theorem is proved.

Part (ii) may be proved by using the relation $\delta g = A_{\delta\rho} f$, which implies that

$$\left\| K_\rho^{-1} A_\rho^T \delta g \right\|_{L^2}^2 = \left\| K_\rho^{-1} A_\rho^T A_{\delta\rho} f \right\|_{L^2}^2 \le \left\| K_\rho^{-1} A_\rho^T A_{s\rho} f \right\|_{L^2}^2 \tag{3.54}$$

again because of Fubini's theorem and monotonicity of integration, where one should note that ρ, $s\rho$, and f are assumed to be nonnegative as previously. Then, we see that

$$\left\| K_\rho^{-1} A_\rho^T A_{s\rho} f \right\|_{L^2}^2 = \left\langle f, K_\rho^{-1} A_{s\rho}^T A_\rho K_\rho^{-1} A_\rho^T A_{s\rho} f \right\rangle \le \|f\|_{L^2}^2 \left\| K_\rho^{-1} A_{s\rho}^T A_\rho K_\rho^{-1} A_\rho^T A_{s\rho} \right\|_{L^2 \to L^2} \tag{3.55}$$

holds. Realizing that the L^2 operator norm of the positive self-adjoint operator $K_\rho^{-1} A_{s\rho}^T A_\rho K_\rho^{-1} A_\rho^T A_{s\rho}$ can be bound via the Riesz–Thorin theorem similarly as for $K_\rho^{-1} A_\rho^T A_\rho$ in the proof of Theorem 3.2 we conclude that the pertinent operator norm is bound by $D_{\rho, s\rho}^2$.

Part (iii) is proved by using the self-adjointness of A and that the adjoint of A_ρ^T is A_ρ. Due to that, for any $h \in L^2(X)$, one has

$$\langle h, \delta f_N \rangle = \left\langle K_\rho^{-1} A_\rho H_N, \delta g \right\rangle \tag{3.56}$$

with the previous notation. Due to the monotonicity of integration, the identity $|\langle h, \delta f_N\rangle| \leq \left\langle \left|K_\rho^{-1} A_\rho H_N\right|, sg\right\rangle$ is obtained, since $|\delta g| \leq sg$ holds. When setting $h := \xi_U$ and correspondingly $H_N := \Xi_{U,N}$, this is nothing but (iii).

Part (iv) is proved by using (3.56) and $\delta g = A_{\delta\rho} f$ and furthermore that the adjoint of $A_{\delta\rho}$ is $A_{\delta\rho}^T$. With that, one has $\langle h, \delta f_N\rangle = \left\langle K_\rho^{-1} A_{\delta\rho}^T A_\rho H_N, f\right\rangle$. Using $|\delta\rho| \leq s\rho$ and the monotonicity of integration, one arrives at $|\langle h, \delta f_N\rangle| \leq \left\langle K_\rho^{-1} A_{s\rho}^T |A_\rho H_N|, |f|\right\rangle$. The upper bound (iv) is obtained whenever $h := \xi_U$ and $H_N := \Xi_{U,N}$ is set.

Part (v) is a consequence of (iv), applying Hölder's inequality, in addition. ∎

Remark 3.7. The following remarks provide some more explanation about the usability of the above results on upper estimation of the systematic errors of f_N originating from the systematic errors of the measured pdf g or of the response function ρ.

(i) For any given iteration order $N \in \mathbb{N}_0$ the upper estimate (i) of Theorem 3.5 bounds the systematic deviation of the unfolded pdf f_N averaged over any compact set, in a manifestly calculable way if the systematic errors of the measured pdf are given. In the language of histograms this means that a bin-by-bin upper bound to the systematic error of the unfolded pdf is available in terms of the systematic error of the measured pdf.

(ii) The upper estimate (ii) of Theorem 3.5 provides an alternative bound for the same quantity for the case when the systematic errors are known in terms of the systematic error of the response function. This, similarly to Theorem 3.3(iv), needs the unknown value of $\|f\|_{L^2}$ which may be circumvented in the analogy of Remark 3.4(v). Namely, for any $\varepsilon > 0$ there exists an iteration index threshold $M_\varepsilon \in \mathrm{N}_0$ such that $\forall M > M_\varepsilon$ one has

$$\left|\frac{1}{\mathrm{Volume}(U)} \int_{x\in U} \delta f_N(x)\,\mathrm{d}x\right| \leq \left\|\Xi_{U,N}\right\|_{L^2} D_{\rho,s\rho}\,(1+\varepsilon)\,\|f_M\|_{L^2} \tag{3.57}$$

whenever A_ρ is one-to-one, because then in light of Remark 3.4(iii), $\|f_M\|_{L^2}$ as a function of M converges to $\|f\|_{L^2}$ in a monotonically increasing way.

(iii) The right side of (3.46) may be approximated by

$$\int \left(K_\rho^{-1} A_{s\rho}^T \left|A_\rho \Xi_{U,N}\right|\right)(x)\ |f_M|\,(x)\,\mathrm{d}x \tag{3.58}$$

due to $|f| = f$ and because f_M converges to f as $M \to \infty$ in the L^2 sense, whenever A_ρ is invertible. For large N, the approximative formula with $M := N$ may be used.

4. Generalization to the context of probability measures. In rare cases one faces the problem that the distributions in question cannot be described in terms of pdfs, only in terms of probability measures instead.[3] Such practical cases may arise, for instance, when the folding operator represents kinematics of particle decays [2]. Therefore, it is interesting to ask the question whether the iterative unfolding method (3.6) applies in the framework of probability measures.

[3] A measure is a set function of the subsets of the probability base space. A common example of measures is the Dirac delta.

Remark 4.1. Let us recall some notions in measure theory [33].

(i) A complex measure F over X is a complex valued σ-additive set function on the Borel σ-algebra of the subsets of X. The variation of the complex measure F is the nonnegative valued measure $|F|$ defined by the following requirement: for a Borel set E the value of $|F|(E)$ is the supremum of $\sum_{k=0}^{K} |F(E_k)|$ for any splitting $E_1, \dots, E_K$ of E, i.e., for all such finite system of disjoint Borel sets $E_1, \dots, E_K$ whose union totals up to E. The measures with finite variation, i.e., which have $|F|(X) < \infty$, form a Banach space with the norm being $\|F\| := |F|(X)$. We shall denote this space by $M(X)$.

(ii) A probability measure F on X is a nonnegative measure on the Borel σ-algebra of X with the requirement $F(X) = 1$. Thus, quite naturally, a probability measure on X resides in $M(X)$.

We continue with the formal definition of folding operators whose response function is described by a measure rather than a function.

Definition 4.1. *A mapping $Q : X \to M(Y), x \mapsto Q(\cdot|x)$ is called a folding measure if for every $x \in X$ the measure $Q(\cdot|x)$ is a nonnegative measure on Y with $Q(Y|x) = 1$ (i.e., $Q(\cdot|x)$ is a probability measure $\forall x \in X$), and for every Borel set E in Y the function $x \mapsto Q(E|x)$ is measurable.*

Remark 4.2. A possible usual generalization is when inefficiencies are also allowed, i.e., the less restrictive condition $Q(Y|x) \leq 1$ is required $\forall x \in X$. The results throughout this paper also hold for that case.

It follows from the definition that a folding measure Q may be viewed as a conditional probability measure over the product space $Y \times X$. Quite evidently, if ρ is a response function, then $Q_\rho(E|x) := \int_{y \in E} \rho(y|x)\, dy$ defines a folding measure.

Definition 4.2. *Let Q be a folding measure. Then, the linear map*

$$A_Q : M(X) \to M(Y),\ F \mapsto \left(\int Q(\cdot|x)\, dF(x) \right) \tag{4.1}$$

is called the folding operator by Q.

Remark 4.3. The remarks below follow from the definition [2].

(i) A folding operator A_Q is well-defined as for all points $x \in X$ and Borel sets E of Y the inequality $Q(E|x) \leq 1$ holds; thus the function $x \mapsto Q(E|x)$ is integrable by any measure with finite variation.

(ii) The monotonicity of integration implies that a folding operator is continuous and $\|A_Q\|_{M(X)\to M(Y)} = 1$, just as in the case of L^1 theory. If inefficiencies are allowed, $\|A_Q\|_{M(X)\to M(Y)} \leq 1$ holds.

(iii) The folding operators defined by folding measures is a generalization of the folding operators by response functions.

As in the L^1 theory, the convolutions represent an important class of folding operators.

Definition 4.3. *A folding operator A_Q is called a convolution if its folding measure is translationally invariant in the sense that $Y = X$ and $\forall x, z \in X$ and Borel sets E one has $Q(E|x + z) = Q(E - z|x)$.*

Remark 4.4. The following are important properties of convolution operators with measures [2].

(i) Whenever the folding operator A_Q by a folding measure Q is a convolution, Q may be expressed by a single probability measure $R := Q(\cdot|0)$ in the form of $Q(E|x) = R(E-x) \forall x \in X$ and Borel set E. The alternative notation $R \star F := A_Q F$ is often used in such case ($F \in M(X)$). Note that the convolution is commutative, i.e., one has $R \star F = F \star R \, \forall R, F \in M(X)$.

(ii) Fourier transformation of measures in $M(X)$ can also be defined and has similar properties as in the L^1 case, except that the Fourier transform functions do not decay at infinity, i.e., the Riemann–Lebesgue lemma does not hold. Only the boundedness of Fourier transforms are guaranteed.

(iii) Properties of convolution operators are similarly related to the Fourier transform of the underlying probability measure, as in the L^1 theory. For instance, a convolution operator is one-to-one if and only if its Fourier transform is nonzero almost everywhere.

(iv) It is easily seen that if $\varphi \in L^1(X)$ and $F \in M(X)$, then $\varphi \star F$ is a function in $L^1(X)$. Combining this with Remark 3.1(ii) we conclude that if $\varphi \in L^p(X) \cap L^1(X)$ then $\forall F \in M(X)$ the function $\varphi \star F \in L^p(X) \cap L^1(X)$ $(1 \leq p \leq \infty)$. That is, probability measures may be mapped into pdfs in $L^p(X)$ via convolution by a pdf integrable on the pth power.

Armed with the introduced notions we may try to ask the question whether one can generalize the results in section 3 to probability measures.

Remark 4.5. The following results are generalization of the results in section 3 for probability measures.

(i) The naive application of Neumann series fails to work similarly as in the L^1 framework. This is because as proved in [2] one has $\|I - A_Q\|_{M(X)\to M(X)} = 2$ whenever $Q(\{y\}|y) = 0$ for any point y—which is the generic case.

(ii) The convergence and error propagation results of Theorems 3.2, 3.3, 3.4, and 3.5 may be generalized in a similar manner to Remark 3.2(i)–(ii). Namely, instead of the original problem $G = A_Q F$ one may consider the modified version $\eta \star G = A_{\eta\star Q} F$ to be solved for F, where η is a square-integrable pdf whose Fourier transform is nowhere vanishing. In this case, the folding operator A_Q is mapped to be a folding operator by a response function $A_{\eta\star Q}$ instead, as we have $\eta \star A_Q F = A_{\eta\star Q} F$ for any $F \in M(X)$. Furthermore, for each $x \in X$ the pdf $\eta \star Q(\cdot|x)$ is square-integrable. Then, the iteration

$$\begin{aligned} K_{\eta\star Q} &:= \sup_{x\in X} \int\int (\eta \star Q)(y|z)\,(\eta \star Q)(y|x)\,\mathrm{d}y\,\mathrm{d}\mu(z), \\ F_0 &:= K_{\eta\star Q}^{-1} A_{\eta\star Q}^T\, \eta \star G, \\ F_{N+1} &:= F_N + \left(F_0 - K_{\eta\star Q}^{-1} A_{\eta\star Q}^T A_{\eta\star Q} F_N\right) \\ &\qquad (N \in \mathbb{N}_0). \end{aligned} \tag{4.2}$$

obeys the very same convergence and error propagation properties as stated in Theorems 3.2, 3.3, 3.4, and 3.5, whenever $K_{\eta\star Q} < \infty$ and when the unknown probability measure F corresponds to a square-integrable pdf with respect to some a priori given

non-negative valued measure μ over X. This latter requirement means that $F = f\mu$ needs to be satisfied with some nonnegative measure μ over X and with some μ-measurable function $f : X \to \mathbb{R}_0^+$ for which $\int |f|^2 (x)\, \mathrm{d}\mu(x) < \infty$ needs to hold.

The previous observations conclude that whenever the unknown distribution is described by a pdf which is square-integrable with respect to some volume measure, then the folding measure may be conditioned in a way that the iterative unfolding (3.6) applies to it.

5. The discrete case. For better illustration, we specialize our results in sections 3 and 4 to the case when the unknown probability distribution along with the response function and the measured probability distribution are discrete. In that case the measured pdf g and the unknown pdf f are a finite dimensional vector of nonnegative entries, and the folding operator A_ρ is simply a finite dimensional matrix with nonnegative entries as well. Our equation to solve is then the matrix equation $g = A_\rho f$ for f, or in the case of presence of measurement errors e, the matrix equation $g = A_\rho f + e$. We also assume that the entries of f, A_ρ, and g are probabilities, i.e., they are normalized such that $\sum_i g_i = 1$, $\sum_i f_i = 1$, and $\sum_j (A_\rho)_{ji} = 1$, or $\sum_j (A_\rho)_{ji} \leq 1$ in case of presence of inefficiencies.

Then, the iterative solution of our discrete unfolding problem reads as

$$\begin{aligned} K_\rho &:= \max_i \sum_j \sum_k (A_\rho)_{ji}\, (A_\rho)_{jk}\,, \\ f_0 &:= K_\rho^{-1} A_\rho^T g, \\ f_{N+1} &:= f_N + \left(f_0 - K_\rho^{-1} A_\rho^T A_\rho f_N\right) \\ &\qquad (N \in \mathbb{N}_0), \end{aligned} \tag{5.1}$$

where A_ρ^T is the matrix transpose of A_ρ. A simple observation shows that (5.1) is nothing but an iterative form of

$$\begin{aligned} K_\rho &:= \max_i \sum_j \sum_k (A_\rho)_{ji}\, (A_\rho)_{jk}\,, \\ f_N &:= \sum_{n=0}^{N} (I - K_\rho^{-1} A_\rho^T A_\rho)^n K_\rho^{-1} A_\rho^T g \\ &\qquad (N \in \mathbb{N}_0), \end{aligned} \tag{5.2}$$

I denoting the identity matrix. Due to the results of sections 3 and 4, the convergence of this approximation is monotonic in the l^2 vector norm and also holds entrywise, however, with possibly quite different convergence rates for different vector entries. Along with this, all the convergence and error propagation properties listed in sections 3 and 4 hold, independently of the fineness of the discretization. This decoupling from the discretization is quite important, as it shows that in the presented method the discretization does not become an important ingredient of the regularization procedure itself in the case when f, g, and ρ are in reality continuum distributions, modeled and measured as histograms.

6. Numerical example. In this section the performance of the proposed method is illustrated by a numerical example. The example calculation is implemented via the C library `libunfold` [34], also including the automatic approximation, statistical, and systematic error

propagation formulae presented in the paper. The shown example is also shipped with the pertinent library. The illustrative case was deliberately chosen in a way when the response function is not translationally invariant, i.e., when ordinary deconvolution methods are not sufficient.

Our simulated measurement scenario is the following. We would like to measure the true pdf of a quantity, namely, of the energy of produced charged particles in a high energy particle collision experiment. This true pdf used in our toy Monte Carlo shall be a parametrization of a real measurement at the LHC accelerator [35] at CERN. It is of the form

$$E \mapsto f(E) := \chi_{[0,\infty[}(E)\ \ |E|\,\frac{(n-1)(n-2)}{(n\,T)^2}\,\left(1+\frac{|E|}{n\,T}\right)^{-n} \tag{6.1}$$

with parameters $n = 6.6$ and $T = 0.145\,\mathrm{GeV}$. The response function

$$(E_{\mathrm{measured}},\, E_{\mathrm{true}}) \mapsto \rho(E_{\mathrm{measured}} \,|\, E_{\mathrm{true}}) \tag{6.2}$$

shall be such a cpdf that for each fixed value $E_{\mathrm{true}} > 0$ the pdf

$$E_{\mathrm{measured}} \mapsto \rho(E_{\mathrm{measured}} \,|\, E_{\mathrm{true}}) \tag{6.3}$$

shall be a Gaussian pdf with a mean of E_{true} and standard deviation of $a + \sqrt{b\,E_{\mathrm{true}}} + c\,E_{\mathrm{true}}$, with parameter values $a = 0.150\,\mathrm{GeV}$, $b = 0.7174\,\mathrm{GeV}$, $c = 0.074$. This response function models the behavior of a calorimeter device used for the energy measurement of particles, namely, of the HCAL calorimeter [36] of the CMS experiment at the LHC accelerator at CERN. In the simulated measurement scenario 10^4 Monte Carlo samples according to the pdf (6.1) was generated, and its corresponding smeared response according to (6.2) was generated. These responses were assumed to be collected with an inefficiency of

$$E_{\mathrm{measured}} \mapsto \frac{1}{2}\left(1+\tanh\left(\frac{E_{\mathrm{measured}}-\mathcal{E}}{\Delta}\right)\right)\,d \tag{6.4}$$

with parameters $\mathcal{E} = 1\,\mathrm{GeV}$, $\Delta = 1\,\mathrm{GeV}$, and $d = 0.05$, i.e., with an inefficiency not greater than 5% on the overall measurement domain. The collected responses were histogramed, providing the measured pdf g with our nonideal detector. By construction, the statistical covariance matrix of the histogram g shall be $\mathrm{diag}(g)$. The inefficiency profile (6.4) causing a systematic deviation of the measured pdf from the folded pdf by (6.2) is assumed to be not known quantitatively and therefore is not corrected for. It is assumed, however, that an overall 5% upper bound to this systematic deviation is known, being the systematic error of the measured pdf, i.e., one has $sg = 0.05\,g$. With these inputs, the linear iterative unfolding according to (3.6) was performed. The approximation errors were quantified using Remark 3.4(vi). The propagated statistical errors were calculated according to (3.39). The propagated systematic errors were quantified using Theorem 3.5(iii). The iteration was stopped when the combined statistical, approximation, and systematic error exceeded a predefined threshold of 7%. The result of the numerical test is shown in Figure 1. Note, that more optimalized stopping criteria can also be invented, using the estimates for the approximation error, statistical error, and systematic error. A natural candidate can be a double-threshold criterion: the approximation error needs to be below a threshold (sufficient shape restoration), whereas the combined

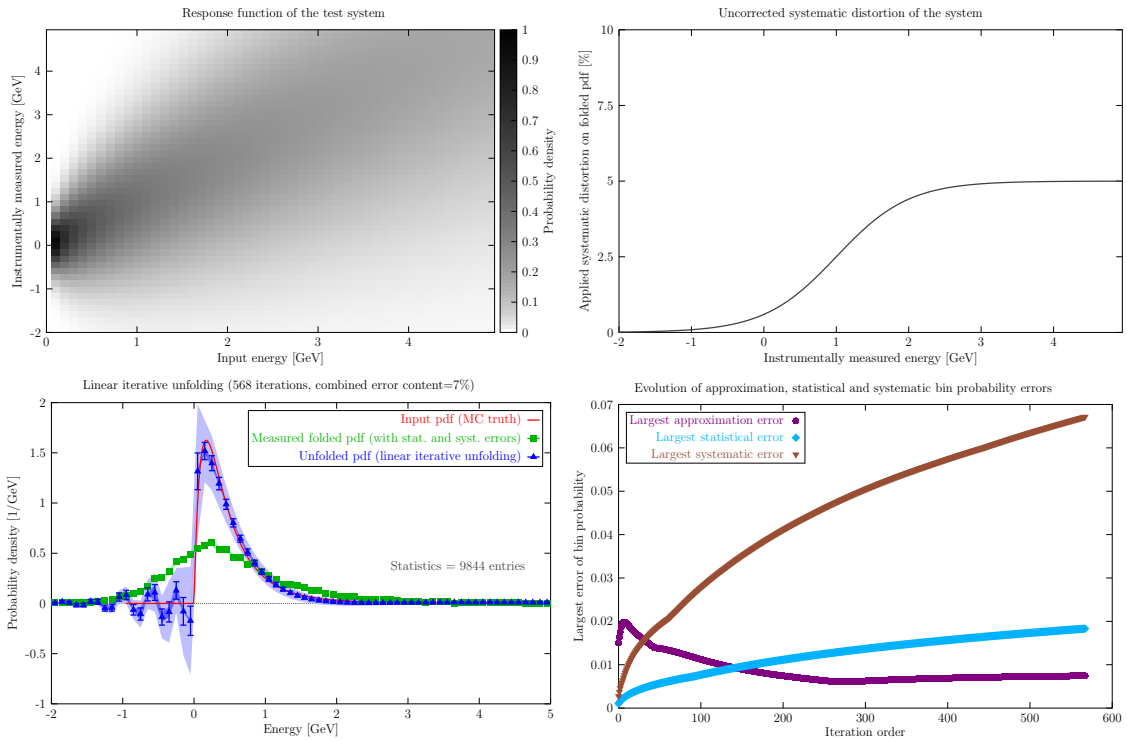

Figure 1. *(Color online) Top left: illustration of the response function of our test example. The color intensity indicates the probability density of the response function. Top right: illustration of the unaccounted systematic distortion applied to the folded pdf in our test example. The solid curve indicates the systematic distortion (an inefficiency, in our example) on the unfolded pdf, which is assumed to be not exactly quantifiable and therefore is not corrected for in the simulated measured pdf. Only an upper bound for the systematic distortion, called the systematic error, is assumed to be known for the simulated measured pdf. That is taken to be a constant* 5% *upper bound in the example. Bottom left: the true input pdf (solid line), the simulated measured pdf (squares), and the unfolded pdf (triangles) by the proposed method. The pdfs are shown together with their bin-by-bin statistical errors (error bars), systematic errors (error bands), and approximation errors (narrow error bands). Bottom right: evolution of the bin-by-bin maximum of the approximation error (circles), statistical error (diamonds), and systematic error (flipped triangles) as a function of the number of iterations. Note that the binwise approximation errors converge to zero, but not in a monotonic manner, which explains the slight increase of that term after about* 300 *iterations. If the iteration was continued, that term indeed converged to zero, but with several local minima, i.e., "waves" or "jumps" are seen in the convergence curve. On the other hand, the binwise statistical and systematic error terms are seen simply to diverge, as expected. The competition of these three error terms gives a possibility to define a stopping criterion.*

statistical and systematic error must stay below an upper bound (divergence regularization). Also, the iteration might be stopped at the error optimum: at the minimum of the combined approximation, statistical, and systematic error. Note, however, that one often might require a better shape reconstruction at the expense of increased statistical and systematic errors, as also seen in the example.

7. Concluding remarks. In this paper we presented mathematical proofs of convergence and error propagation formulae for a linear iterative unfolding method [1] in the probability theory context. It was shown that the pertinent method is convergent in the "binwise" sense

under quite generic conditions, which does hold in the case of many practical applications. Furthermore, explicit formulae for the three important error terms, the approximation error, the statistical error, and the systematic error, were derived. These can be used to define optimal iteration stopping criterion and quantification of errors therein. The key element of the proofs is the Riesz–Thorin theorem mapping the original L^1 problem to L^2 with a subsequent usage of spectral theory of L^2 operators. The typical use-cases of the method are those unfolding problems which cannot be handled by statistical deconvolution [9, 10], due to the absence of translational invariance of the response function. The possibility for propagation of the systematic errors is a special advantage, which deserves to be emphasized for experimental applications.

The pertinent method is also available as a C numerical library [34]. Using that, the method was demonstrated in a numerical example. The algorithm could be included in the ROOUnfold package [37] in the future or in the GNU Scientific Library [38].

The present paper can serve also as a good motivation to perform similar convergence and error propagation studies on another iterative unfolding method [18, 19, 21, 22, 23, 24, 25, 26], also called the method of convergent weights or iterative Bayesian unfolding. That method is nonlinear and therefore is somewhat more complicated to study; however, it can be more suitable for unfolding problems in probability theory as it conserves the integral and nonnegativity of pdfs. Although widely used and numerically very promising, so far little is known on the convergence properties of that algorithm, and nothing is known about its error propagation. Our proposed method can be considered as the "linearized" version of that method, and thus the presented results are expected to provide clues also for the convergence and error propagation properties of the method of convergent weights or iterative Bayesian unfolding.

Acknowledgments. The author would like to thank Tamás Matolcsi for valuable comments and for reading various versions of the manuscript and Dezső Varga for discussions on the physical applications and on the relevance of error propagation formulae, in particular for the systematic errors.

REFERENCES

[1] A. László, *A linear iterative unfolding method*, J. Phys. Conf. Ser., 368 (2012), 012043.
[2] A. László, *A robust iterative unfolding method for signal processing*, J. Phys. A, 39 (2006), 13621.
[3] G. Cowan, *Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics*, Durham, UK, 2002.
[4] V. Blobel, *Unfolding for HEP experiments*, presented at DESY Computing Seminar, http://www.desy.de/~blobel/DESYcompsem08.pdf (2008).
[5] G. Bohm and G. Zech, *Introduction to Statistics and Data Analysis for Physicists*, Verlag Deutsches Elektronen-Synchrotron, Hamburg, 2010.
[6] M. Kuusela and V. M. Panaretos, *Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification*, Ann. Appl. Stat., 9 (2015), pp. 1671–1705.
[7] M. Kuusela and P. B. Stark, *Shape-Constrained Uncertainty Quantification in Unfolding Steeply Falling Elementary Particle Spectra*, preprint, arXiv:1512.00905, 2015.
[8] H. P. Dembinski and M. Roth, *An algorithm for automatic unfolding of one-dimensional data distributions*, Nucl. Instr. Meth. A, 729 (2013), pp. 410–416.
[9] I. Dattner, A. Goldenshluger, and A. Juditsky, *On deconvolution of distribution functions*, Ann. Statist., 39 (2011), pp. 2477–2501, doi:1232.62056.

[10] I. Dattner, M. Reiß, and M. Trabs, *Adaptive quantile estimation in deconvolution with unknown error distribution*, Bernoulli, 22 (2016), pp. 143–192.
[11] J. Fan, *On the optimal rates of convergence for nonparametric deconvolution problems*, Ann. Statist., 19 (1991), pp. 1257–1272.
[12] C. H. Hesse, *Iterative density estimation from contaminated observations*, Metrika 64 (2006), pp. 151–165.
[13] F. Comte and C. Lacour, *Anisotropic adaptive kernel deconvolution*, Ann. Inst. H. Poincaré Probab. Stat., 49 (2013), pp. 569–609.
[14] M. C. Liu and R. L. Taylor, *A consistent nonparametric density estimator for the deconvolution problem*, Canad. J. Statist., 17 (1989), pp. 427–438.
[15] L. A. Stefanski and R. J. Carol, *Deconvoluting kernel density estimators*, Statistics, 21 (1990), pp. 169–184.
[16] J. Kalifa and B. Rouge, *Deconvolution by thresholding in mirror wavelet bases*, IEEE Trans. Image Process., 12 (2003), pp. 446–457.
[17] A. Hoecker and V. Kartvelishvili, *SVD Approach to data unfolding*, Nuclear Instr. Meth. A, 372 (1996), pp. 469–481.
[18] G. D'Agostini, *A multidimensional unfolding method based on Bayes' theorem*, Nuclear Instr. Meth. A, 362 (1995), pp. 487–498.
[19] G. Zech, *Iterative unfolding with the Richardson-Lucy algorithm*, Nuclear Instr. Meth. A, 716 (2013), pp. 1–9.
[20] C. ALT et al., *High transverse momentum Hadron spectra at* $\sqrt{s_{NN}} = 17.3\,\text{GeV}$*, in* Pb+Pb *and* p+p *Collisions*, Phys. Rev. C, 77 (2008), 034906.
[21] W. H. Richardson, *Bayesian-based iterative method of image restoration*, J. Opt. Soc. Amer. A, 62 (1972), pp. 55–59.
[22] L. B. Lucy, *An iterative technique for the rectification of observed distributions*, Astronomi. J., 79 (1974), p. 745.
[23] L. A. Shepp and Y. Vardi, *Maximum likelihood reconstruction for emission tomography*, IEEE Trans. Med. Imag., 1 (1982), pp. 113-122.
[24] A. Kondor, *Method of convergent weights – An iterative procedure for solving Fredholm's integral equations of the first kind*, Nuclear Instr. Meth., 216 (1983), pp. 177–181.
[25] H. N. Mülthei and B. Schorr, *On an iterative method for a class of integral equations of the first kind*, Math. Methods Appl. Sci., 9 (1987).
[26] H. N. Mülthei and B. Schorr, *On an iterative method for the unfolding of spectra*, Nuclear Instr. Meth. A, 257 (1987), pp. 371–377.
[27] P. D. Lax, *Functional Analysis,* Chichester, Wiley-Interscience, 2002,
[28] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.
[29] G. Arfken, *Fourier convolution theorem*, in Mathematical Methods for Physicists, 7th ed., Elsevier, Amsterdam, 2013.
[30] P. Bracewell, *Convolution theorem*, in The Fourier Transform and Its Applications, 3rd ed., McGraw-Hill, New York, 1999, pp. 108–112.
[31] L. Landweber, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624.
[32] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Wiley-Interscience, New York, 1999.
[33] N. Dinculeanu, *Vector Measures*, Elsevier, New York, 1967.
[34] A. László, *The Libunfold Package*, http://www.rmki.kfki.hu/~laszloa/downloads/libunfold.tar.gz (2011).
[35] V. Khachatryan et al., *Transverse-momentum and pseudorapidity distributions of charged Hadrons in pp collisions at* $\sqrt{s} = 7$ TeV, Phys. Rev. Lett., 105 (2010), 022002.
[36] E. Yazgan, *The CMS barrel calorimeter response to particle beams from* 2 *to* 350 *GeV/c*, J. Phys. Conf. Ser., 160 (2009), 012056.
[37] T. Adye et al., The ROOUnfold Package, http://hepunx.rl.ac.uk/~adye/software/unfold/RooUnfold.html.
[38] GNU Scientific Library, http://www.gnu.org/software/gsl.