entropy

MDPI

Article

Information Field Theory and Artificial Intelligence

Torsten Enßlin [1,2]

[1] Max Planck Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, 85748 Garching, Germany; ensslin@mpa-garching.mpg.de
[2] Physics Department, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

Abstract: Information field theory (IFT), the information theory for fields, is a mathematical framework for signal reconstruction and non-parametric inverse problems. Artificial intelligence (AI) and machine learning (ML) aim at generating intelligent systems, including such for perception, cognition, and learning. This overlaps with IFT, which is designed to address perception, reasoning, and inference tasks. Here, the relation between concepts and tools in IFT and those in AI and ML research are discussed. In the context of IFT, fields denote physical quantities that change continuously as a function of space (and time) and information theory refers to Bayesian probabilistic logic equipped with the associated entropic information measures. Reconstructing a signal with IFT is a computational problem similar to training a generative neural network (GNN) in ML. In this paper, the process of inference in IFT is reformulated in terms of GNN training. In contrast to classical neural networks, IFT based GNNs can operate without pre-training thanks to incorporating expert knowledge into their architecture. Furthermore, the cross-fertilization of variational inference methods used in IFT and ML are discussed. These discussions suggest that IFT is well suited to address many problems in AI and ML research and application.

Keywords: information field theory; artificial intelligence; generative models; variational inference

Citation: Enßlin, T. Information Field Theory and Artificial Intelligence. *Entropy* **2022**, *24*, 374. https://doi.org/10.3390/e24030374

Academic Editors: Wolfgang von der Linden and Sascha Ranftl

Received: 18 December 2021
Accepted: 4 March 2022
Published: 7 March 2022

1. Motivation

Determining the concrete configuration of a field from measurement data is an ill-posed inverse problem, as physical fields have an infinite number of degrees of freedom (DoF), whereas data sets are always finite in size. Thus, the data provide a finite number of constraints for only a subset of the infinitely many DoF of a field. In order to infer a field, the remaining of its DoF need, therefore, to be constrained via prior information. Fortunately, physics provides such prior information on fields. This information might either be precise, like $\nabla \cdot B = 0$ in electrodynamics, or more phenomenological, in the sense that a field shaped by a certain process can often be characterized by its n-point correlation functions. Having knowledge on such correlations can be sufficient to regularize the otherwise ill-posed field inference problem from finite and noisy data such that meaningful statements about the field can be made.

As a formalism for this, information field theory (IFT) was introduced [1,2] following and extending earlier lines of work [3,4]. IFT is information theory for fields. It investigates the space of possible field configurations and constructs probability densities over those spaces in order to permit Bayesian field inference from data. It has been applied successfully to a number of problems in astrophysics [5–18], particle physics [19–21], and elsewhere [22–26]. Here, the relation of IFT with methods and concepts used in artificial intelligence (AI) and machine learning (ML) research are outlined, in particular with generative neural networks (GNNs) and in the usage of variational inference. The presented line of arguments summarizes a number of recent works [27–32].

The motivation for this work is twofold. On the one hand, understanding conceptual relations between IFT, ML, and AI techniques allows us to transfer computational methods

between these domains and to develop synergistic approaches. This article will discuss such. On the other hand, the current success of deep learning techniques for neural networks has let them appear as a synonym for AI in the public perception. This has consequences for decisions about which kind of technologies get scientific funding. The point this paper is trying to make is that if deep learning qualifies as AI in this respect, then this should also apply to a number of other techniques, including those based on IFT.

The paper is organized as follows. IFT is briefly introduced in Section 2 in its most modern incarnation in terms of standardized, generative models. These are shown to be structurally similar to GNNs in Section 3. The structural similarity of IFT inference and GNN training problems allows for a common set of variational inference methods, as discussed in Section 4. Section 5 concludes on the relation of IFT methods and those used in AI and ML and gives an outlook on future synergies.

2. Information Field Theory

2.1. Basics

IFT allows us to deduce fields from data in a probabilistic way. In order to be able to apply probability theory onto the space of field configurations, a measure in this space is needed. Although no canonical mathematical measure on function spaces exists, for IFT applications, the usage of Gaussian process measures [33], which are mathematically well defined [34,35], is usually fully sufficient. Gaussian processes can also be argued to be a natural starting point for reasoning on fields with known finite first and second order moments, as we will discuss now.

To be specific, let $\varphi : \Omega \to \mathbb{R}$ be a scalar field over some domain $\Omega \subset \mathbb{R}^u$ and our prior knowledge on φ be the first and second moments of the field, e.g.,

$$\langle \varphi^x \rangle_{(\varphi)} = \overline{\varphi}^x \text{ and} \tag{1}$$

$$\langle (\varphi - \overline{\varphi})^x (\varphi - \overline{\varphi})^y \rangle_{(\varphi)} = \Phi^{xy} \text{ for all } x, y \in \Omega, \tag{2}$$

with $\varphi^x := \varphi(x)$ denoting a field value and $\langle f(\varphi) \rangle_{(\varphi)} := \int \mathcal{D}\varphi \, \mathcal{P}(\varphi) f(\varphi)$ a prior expectation value for some function f of the field. If only the first and second field moments are given as prior information, it follows from the maximum entropy principle that the least informative probability distribution function encoding this information is a Gaussian with these moments. Thus, using this Gaussian

$$\begin{aligned} \mathcal{P}(\varphi|I) &\equiv \mathcal{G}(\varphi - \overline{\varphi}, \Phi) \\ &:= \frac{1}{\sqrt{|2\pi\Phi|}} \exp\left(-\frac{1}{2} (\varphi - \overline{\varphi})^\dagger \Phi^{-1} (\varphi - \overline{\varphi}) \right) \end{aligned} \tag{3}$$

as a prior with background information $I = (\langle \varphi \rangle_{(\varphi)} = \overline{\varphi}, \langle (\varphi - \overline{\varphi})(\varphi - \overline{\varphi})^\dagger \rangle_{(\varphi)} = \Phi)$ is a conservative choice, as it makes the least additional assumptions about the field except for the moments specified in I.

In many applications, however, the field of interest, the signal s, is not a Gaussian field, but may be related to such via a non-linear transformation. For example, in astronomical applications of IFT, the sky brightness field s is the quantity of interest, which is strictly positive, and therefore cannot be a Gaussian field. However, the logarithm of a brightness can be positive and negative and may therefore be modeled as a Gaussian process. In such a case, one could assign, e.g., $s^x(\varphi) = s_0 \exp(\varphi^x)$ as a model for a diffuse (spatially correlated) sky emission component, with s_0 a reference brightness, chosen such that for example $\langle \varphi^x \rangle_{(\varphi)} = 0$ holds.

Having established a field prior, Bayesian reasoning on the field φ, and therefore on the signal of interest $s = s(\varphi)$, based on some data d and its likelihood $\mathcal{P}(d|\varphi, I)$ is possible. The field posterior

$$\mathcal{P}(\varphi|d, I) = \frac{\mathcal{P}(d|\varphi, I)\mathcal{P}(\varphi|I)}{\mathcal{P}(d|I)} \tag{4}$$

is defined as well as the prior and permits us to answer questions about the field, like its most probable configuration $\varphi_{\mathrm{MAP}} = \mathrm{argmax}_{\varphi}\mathcal{P}(\varphi|d, I)$ (MAP = maximum a posteriori), its posterior mean $m = \langle\varphi\rangle_{(\varphi|d,I)}$, or its posterior uncertainty dispersion $D = \left\langle(\varphi - m)(\varphi - m)^{\dagger}\right\rangle_{(\varphi|d,I)}$. IFT exploits the formalism of quantum and statistical field theory to calculate such posterior expectation values [1,28,36–38]. These formal calculations, however, should not be the focus here. Instead, it should be the formulation of IFT inference problems in terms of generative models, as these can be interpreted as GNNs.

For this purpose, the likelihood is expressed in terms of a measurement equation

$$\begin{aligned} d &= R(\varphi) + n, \text{ with} && (5)\\ R(\varphi) &:= \langle d\rangle_{(d|\varphi)}, && (6)\\ n &:= d - R(\varphi), \text{ and} && (7)\\ \mathcal{P}(d|\varphi, I) &\equiv \mathcal{P}(n = d - R(\varphi)|\varphi), && (8) \end{aligned}$$

which is always possible if the data can be embedded into a vector space and the data expectation value $\langle d\rangle_{(d|\varphi)}$ exists. Here and in the following, we omit the background information I in probabilities. This rewriting of the likelihood in terms of a mean instrument response $d' = R(\varphi)$ to the field φ and a noise process $\mathcal{P}(n|\varphi)$, which summarizes the fluctuations around that mean d', allows us to regard the data as the result of a noisy generative process that maps field values φ and associated noise realizations n onto data d according to Equation (5).

In case the instrument response and noise processes are provided for the signal s instead of the Gaussian field φ as $R'(s) := \langle d\rangle_{(d|s)}$ and $\mathcal{P}(n|s)$, their respective pull backs $R(\varphi) := \langle d\rangle_{(d|s(\varphi))} = R'(s(\varphi))$ and $\mathcal{P}(n|\varphi) := \mathcal{P}(n|s(\varphi))$ provide the necessary response and noise statistics w.r.t. the field φ.

All this provides a generative model for the signal s and data d via $\varphi \hookleftarrow \mathcal{G}(\varphi, \Phi)$, $s = s(\varphi)$, $n \hookleftarrow \mathcal{P}(n|s)$, and $d = R'(s) + n$, which should now be standardized. The standardization introduces a generic latent space that permits better comparison to GNNs used in AI and ML and simplifies the usage of variational inference methods discussed later on.

2.2. Prior Standardization

Standardization of a random variable φ refers to finding a mapping from a standard normal distributed random variable $\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1})$ to φ that reproduces the statistics of $\mathcal{P}(\varphi)$. For a Gaussian field φ, this is just a mapping of the form

$$\varphi(\xi) := \overline{\varphi} + \Phi^{\frac{1}{2}}\xi, \tag{9}$$

where $\Phi^{\frac{1}{2}}$ refers to a square root of Φ, which always exists for a covariance matrix that is positive definite. For the large class of band diagonal and therefore translational invariant covariance matrices Φ, which are very relevant for applications as we argue below, the square root of Φ can be explicitly constructed.

2.3. Power Spectra

In many signal inference problems, no spatial location is singled out a priori, before the measurement. This means that the field covariance between two locations only depends on the distance between these positions, but not on their absolute positions. Thus,

$\Phi^{xy} = C_\varphi(x-y)$. As a consequence of the Wiener–Khinchin theorem, such a translational invariant field covariance becomes diagonal in harmonic space,

$$\widetilde{\Phi}^{kq} = \mathcal{F}^k_x \Phi^{xy} (\mathcal{F}^\dagger)^q_y = (2\pi)^u \delta(k-q) P_\varphi(k) = \widehat{P_\varphi}^{kq}. \tag{10}$$

Here and in the following, $\mathcal{F}$ denotes a harmonic transform (a u-dimensional Fourier transform $\mathcal{F}^k_x = \exp(ik\cdot x)$ in case of an Euclidean space, as we assume in the following), $\dagger$ the adjoint (complex conjugate and transposed of a matrix or vector), $P_\varphi(k) := \mathcal{F}^k_{x'} C^{x'}_\varphi$ is the so called power spectrum of φ, the Einstein convention for repeated indices is used, as in $\widetilde{\varphi}^k := \mathcal{F}^k_x \varphi^x \equiv \int dx^u \exp(ik\cdot x)\, \varphi(x)$, and $\widehat{\phi} = \mathrm{diag}(\phi)$ denotes a diagonal operator in the space of the field ϕ with the values of ϕ on the diagonal.

Thanks to this diagonal representation of the field covariance in harmonic space, an explicit standardization of the field is given via

$$\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1}), \tag{11}$$

$$\varphi = \overline{\varphi} + \mathcal{F}^{-1} A_\varphi\, \xi, \text{ and} \tag{12}$$

$$A_\varphi = \widehat{P_\varphi^{1/2}}, \tag{13}$$

where the latter is an amplitude operator that is diagonal in harmonic space and that imprints the right amplitudes onto the Fourier modes of φ. This can be seen via a direct calculation,

$$\begin{aligned} \left\langle (\varphi(\xi)-\overline{\varphi})(\varphi(\xi)-\overline{\varphi})^\dagger \right\rangle_{(\xi)} &= \mathcal{F}^{-1} A_\varphi \left\langle \xi\xi \right\rangle_{(\xi)} A^\dagger_\varphi \mathcal{F}^{-1\dagger} \\ &= \mathcal{F}^{-1} A_\varphi\, \mathbb{1} A^\dagger_\varphi \mathcal{F}^{-1\dagger} \\ &= \mathcal{F}^{-1} \widetilde{\Phi} \mathcal{F}^{-1\dagger} = \Phi \\ &= \left\langle (\varphi-\overline{\varphi})(\varphi-\overline{\varphi})^\dagger \right\rangle_{(\varphi)}. \end{aligned} \tag{14}$$

In case no direction of the space is singled out a priori, the two-point correlation function and the power spectrum of φ become isotropic, $\Phi^{xy} = C_\varphi(|x-y|)$ and $\widetilde{\Phi}^{kq} = (2\pi)^u \delta(k-q) P_\varphi(|k|)$, respectively. In this case, only a one-dimensional power spectrum needs to be known. Such power spectra are often smooth functions on a double logarithmic scale in Fourier space, since any sharp feature in them would correspond to a (quasi-) periodic pattern in position space, which would be very unnatural for most signals. Thus, introducing the logarithmic Fourier space scale variable $\kappa(k) := \ln k/k_0$ w.r.t. some reference scale k_0, we expect

$$\psi(\kappa) := \ln\left(P_\varphi(k_0 e^\kappa)/P_0\right) \tag{15}$$

to be a field itself, in the sense that it is sufficiently smooth. Here, P_0 is a pivot scale for the power spectrum.

2.4. Amplitude Model

Often, the power spectrum as parameterized through ψ is not known a priori for a field φ, but statistical homogeneity, isotropy, and the absence of long range quasi-periodic signal variations make a Gaussian field prior for ψ plausible, $\mathcal{P}(\psi) = \mathcal{G}(\psi - \overline{\psi}, \Psi)$. This log-log-power spectrum may exhibit fluctuations $\chi := \psi - \overline{\psi}$ around a non-zero mean $\overline{\psi}(\kappa)$. The latter might, e.g., encode a preference for falling spectra and therefore for a spatially

smooth field φ. In this case, just another layer for χ of a standardized generative model has to be added,

$$\begin{aligned} \eta &\hookleftarrow \mathcal{G}(\eta, \mathbb{1}) && (16)\\ \chi(\eta) &:= A_\psi \eta, \text{ with} && (17)\\ A_\psi A_\psi^\dagger &:= \Psi, \text{ and} && (18)\\ \psi(\eta) &:= \overline{\psi} + \chi(\eta). && (19) \end{aligned}$$

Again, a prior for a field, here the only one dimensional $\chi(\kappa)$, is needed. A detailed description of how this amplitude model can be provided efficiently is given by [15]. This reference also provides a generative model for the case that the signal domain Ω is a product of sub-spaces, like position space and an energy spectrum coordinate, each requiring a different correlation structure, and the total correlation being a direct product of those. Assuming a direct product for the correlation structures might be possible for many field inference problems [15,39].

2.5. Dynamical Systems

Let us take a brief detour to fields shaped by dynamical systems. Dynamical systems, typically exhibit correlation structures that are not direct products of the spatial and temporal sub-spaces, as was proposed above. Here, the full spatial and temporal Fourier power spectrum $P_\varphi(k, \omega)$, with ω being the temporal frequency, encodes the full dynamics of a linear, homogeneous, and autonomous system. For example, a stochastic wave field $\varphi(x, t)$ may follow the dynamical equation

$$\left(\frac{\partial^2}{\partial t^2} + \eta \frac{\partial}{\partial t} - c^2 \frac{\partial^2}{\partial x^2} \right) \varphi(x, t) = \xi(x, t), \tag{20}$$

where c is the wave velocity and η a damping constant. The field dynamics are determined by a response operator (or Green's function) G that is a convolution of the exciting noise field ξ with a kernel g,

$$\varphi = G\,\xi = g * \xi, \tag{21}$$

where $*$ denotes convolution. In Fourier-space, this kernel can be applied by a direct point wise multiplication, $(\mathcal{F}\varphi)^{(k,\omega)} = (\mathcal{F}g)^{(k,\omega)}\,(\mathcal{F}\xi)^{(k,\omega)}$and is given by

$$(\mathcal{F}g)^{(k,\omega)} \quad = \quad (\omega^2 - i\eta\omega - c^2k^2)^{-1} =: P_G(k, \omega). \tag{22}$$

If the excitation of field fluctuations is caused by a white, stochastic noise field $\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1})$, the resulting field has a power spectrum of

$$P_\varphi(k, \omega) = |P_G(k, \omega)|^2 P_\xi(k, \omega) = \frac{1}{(\omega^2 - c^2k^2)^2 + \eta^2\omega^2}. \tag{23}$$

In this case, the spectrum is an analytical function in ω and k. This results from Equation (20) being a linear, homogeneous, and autonomous partial differential equation.

Linear integro-differential equations, however, can still be solved by convolutions, in which case the kernel might not have an analytically closed form any more, if the equation is still homogeneous and autonomous. For example, in neural field theory [40–43], a macroscopic description of the brain cortex dynamics, the neural activity $\varphi(x, t)$ might be described by

$$\frac{\partial}{\partial t}\varphi = -\varphi + w * (f \circ \varphi) + \xi. \tag{24}$$

Here, w is a spatial–temporal convolution kernel (that usually contains a delta function in time), $f : \mathbb{R} \to \mathbb{R}$ an activation function that is applied point wise to the field, $(f \circ \varphi)(x, t) = f(\varphi(x, t))$, and we added an input term ξ. In case f is linear, the system responds linearly

to inputs. Then, the input response is a convolution with a kernel g that has in general a non-analytical spectrum,

$$(\mathcal{F}g)^{(k,\omega)} = \left(1 + i\omega - (\mathcal{F}w)^{(k,\omega)}\, f'\right)^{-1}, \tag{25}$$

where f' is the slope of f and $\mathcal{F}w$ the Fourier transformed kernel of the dynamics.

An inference of such non-analytical and highly structured response spectra from data is possible with IFT and can be used to learn the system dynamics from noisy system measurements [26,44]. It just requires a more complex spectral prior than discussed here. Let us now return to our main line of argumentation.

2.6. Generative Model

To summarize, the field inference problems of IFT can often be stated in terms of a standardized, generative model for the signal and the data. For the illustrative case outlined above, where the probabilistic model is given by

$$\begin{aligned}
\mathcal{P}(d, \varphi, \psi) &= \mathcal{P}(d|\varphi)\mathcal{P}(\varphi|\psi)\mathcal{P}(\psi), && (26)\\
\mathcal{P}(\psi) &= \mathcal{G}(\psi - \overline{\psi}, \Psi), && (27)\\
\mathcal{P}(\varphi|\psi) &= \mathcal{G}(\varphi, \Phi(\psi)), && (28)\\
\Phi(\psi) &= \mathcal{F}^{-1}\widehat{P_\varphi}\mathcal{F}^{-1\dagger}, && (29)\\
P_\varphi(k) &= P_0 \exp(\psi(\ln(|k|/k_0))), \text{ and} && (30)\\
\mathcal{P}(d|\varphi) &= \mathcal{P}(d|s(\varphi)), && (31)
\end{aligned}$$

the corresponding standardized generative model is

$$\begin{aligned}
\zeta &:= (\xi, \eta) \hookleftarrow \mathcal{G}(\zeta, \mathbb{1}), && (32)\\
\psi(\eta) &:= \overline{\psi} + A_\psi \eta, && (33)\\
P_\varphi(k) &:= P_0 \exp(\psi(\ln(|k|/k_0)), is && (34)\\
\varphi(\xi, \psi) &:= \mathcal{F}^{-1}\widehat{P_\varphi^{1/2}}\xi, && (35)\\
s(\varphi) &:= s_0 \exp(\varphi), && (36)\\
n &\hookleftarrow \mathcal{P}(n|s), \text{ and} && (37)\\
d &= R'(s) + n. && (38)
\end{aligned}$$

This generative model is illustrated in Figure 1. Variants of it are used in a number of real world data applications [5–21]. Its performance in generative and reconstruction mode is illustrated for synthetic data in Figures 2 and 3.

For the noiseless data $d' = R'(s)$ the generative model reads

$$\begin{aligned}
d'(\zeta) &:= R'(s(\varphi(\xi, \psi(\eta)))) \\
&= (R' \circ s \circ \varphi \circ f)(\zeta), \text{ with} && (39)\\
f(\zeta) &:= (\xi, \psi(\eta)). && (40)
\end{aligned}$$

This way, the full model complexity as given by Equations (26)–(31) is transferred into an effective response function $d' = R' \circ s \circ \varphi \circ f$. For this latent variable vector, the prior is simply $\mathcal{P}(\zeta) = \mathcal{G}(\zeta, \mathbb{1})$, whereas the likelihood $\mathcal{P}(d|\zeta) = \mathcal{P}(n = d - d'(\zeta)|\zeta)$ has absorbed the full model complexity. This so called reparametrization trick [45] was introduced to IFT by [29] to simplify numerical variational inference.

At this point, it is essential to realize that this generative model consists of a latent space white noise process $\mathcal{P}(\zeta) = \mathcal{G}(\zeta, \mathbb{1})$ that generates an input vector ζ and a sequence of non-local linear and local non-linear operations that is applied to it. The Fourier transform $\mathcal{F}^{-1}$ and A_ψ are examples of non-local linear operations within the model. Among the non-linear operations are the exponential functions and the application of the ψ-dependent

amplitude operator $A_\varphi(\psi)$ to the latent space excitations ξ, as there the two components of $\zeta = (\xi, \eta)$ are multiplied together. Furthermore, the instrument response $R'(s)$ might also be decomposed into sequences of non-local linear and local non-linear operations, as physical processes in measurement devices can often be cast into the propagation of a quantity (an operation that is linear in the quantity) and the local interactions of the quantity (an operation non-linear in it), respectively.

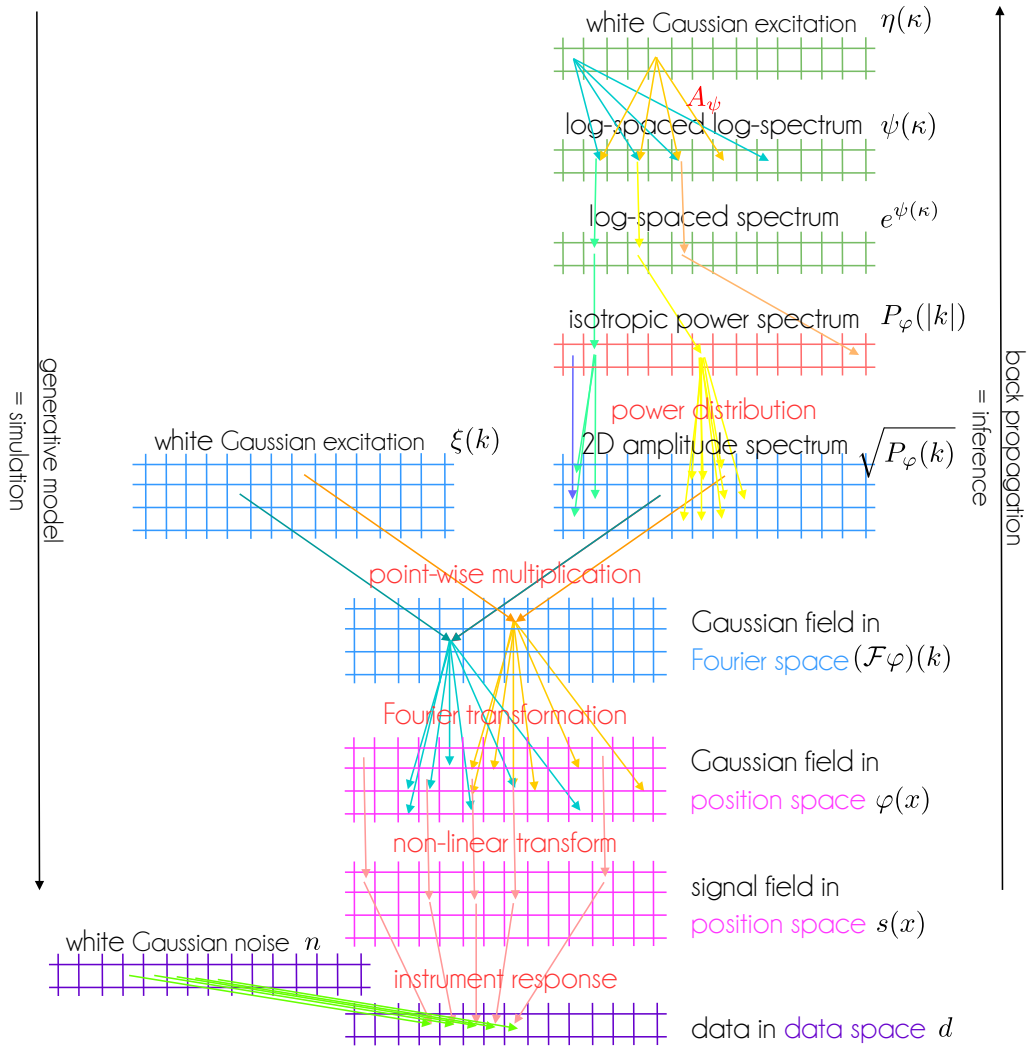

Figure 1. An IFT model for a 2D Gaussian random field also with generated homogeneous and isotropic correlation structure and its measurement according to Equations (32)–(38) displayed as a GNN. Layers with identical shapes are given identical colors. Note that all layers have a physical interpretation and the architecture of this GNN encodes expert knowledge on the field. Inserting random numbers into the latent spaces and executing the network from top to bottom corresponds to a simulation of signal and data generation. "Learning" the latent space variables from bottom to top via back propagation of data space residuals with respect to observed data corresponds to inference.

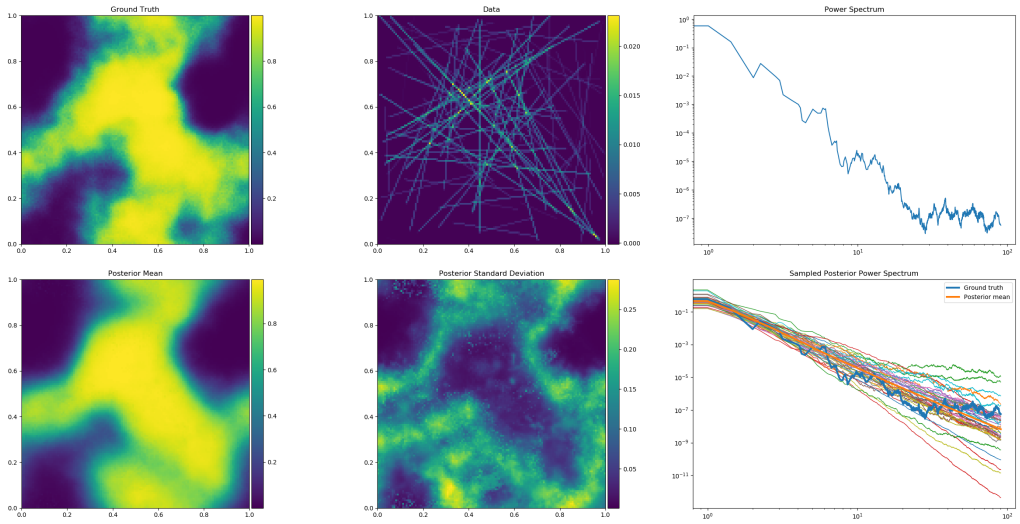

Figure 2. Output of a generative IFT model for a 2D tomography problem in simulation (**top row**) and reconstruction (**bottom rows**) mode. The model is depicted in Figure 1 and described by Equations (32)–(38) with the modification that in Equation (36) the exp-function is replaced by a sigmoid function to obtain more cloud-like structures. Run in simulation mode, the model first generates a non-parametric power spectrum (**top right panel**) from which a Gaussian realization of a statistical isotropic and homogeneous field is drawn (**top left**, after procession by the sigmoid function). This is then observed tomographically (**top middle**), by measurements that integrate over (here randomly chosen) lines of sight. The data values include Gaussian noise and are displayed at the locations of their measurement lines. Fed with this synthetic data set, the model run in inference mode (via geoVI) reconstructs the larger scales of the signal field (**bottom left**), the initial power spectrum (thick orange line in middle right panel; thick blue line is ground truth), and provides uncertainty information on both quantities (signal uncertainty is given at bottom middle, the power spectrum uncertainty is visualized by the set of thin lines at **bottom right**). The presented plots are the direct output of the `getting_started_3.py` script enclosed in the Numerical Information Field Theory (NIFTy) open source software package NIFTy8, downloadable at https://gitlab.mpcdf.mpg.de/ift/nifty (accessed on 17 December 2021) [46–48] that supports the implementation and inference of IFT models.

3. Artificial Intelligence

3.1. Neural Networks

AI and ML are vast fields. AI aims at building artificial cognitive systems that perceive their environment, reason about its state and the systems' best actions, and learn to improve their performance. ML can be regarded as a sub-field of AI, embracing many different methods like self-organized maps, Gaussian mixture models, deep neural networks, and many others. Here, the focus should be on specific neural networks, GNNs, as those have a close relation to the generative IFT models introduced before.

GNNs transform a latent space variable $\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1})$ into a signal or data realization, $s = s(\xi)$ or $d' = d'(\xi)$. A neural network is a function $g(\xi)$ that can be decomposed in terms of n layer processing functions g_i with

$$g = g_n \circ g_{n-1} \circ \dots g_1. \tag{41}$$

Any of the layer processing functions $g_i : \xi_i \mapsto \xi_{i+1}$ with $\xi_1 \equiv \xi$ consists typically of a non-local, affine linear transformation $l_i(\xi_i) := L_i\xi_i + b_i$ of the input vector ξ_i of layer i followed by a local, point wise application of non-linear, so-called activation functions $\sigma_i : \mathbb{R} \to \mathbb{R}$. Thus, the output vector ξ_{i+1} of layer i is

$$\xi_{i+1} \quad = \quad (\sigma_i \circ l_i)(\xi_i), \tag{42}$$

where σ_i acts component wise. The set $\eta = (L_i, b_i)_{i=1}^n$ of all coefficients of the l_is (the matrix elements of the L_i matrices, and the components of the b_i vectors) determines the function the network represents. Putting the input values and network coefficients into a single vector $\zeta := (\xi, \eta)$ a GNN can be regarded as a function of both, latent variables ξ and network parameters η, $d'(\zeta) = g(\xi; \eta)$.

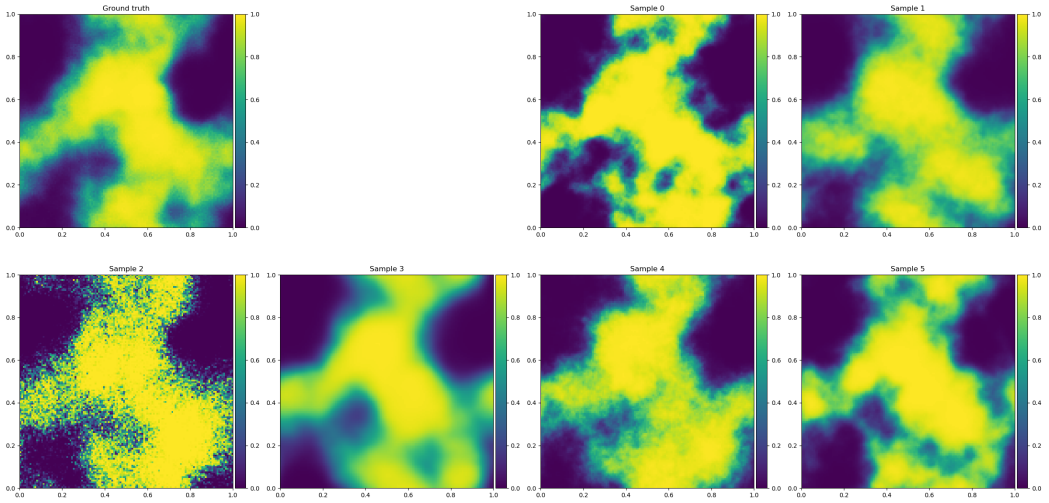

Figure 3. Signal ground truth (**top left panel**) and some signal posterior samples (**other panels**) of the field reconstructed in Figure 2. Note the varying granularity of the field samples due to the remaining posterior uncertainty of the power spectrum on small spatial scales as shown in Figure 2 at bottom right.

3.2. Comparison with IFT Models

From this abstract perspective, a standardized, generative model $d'(\zeta)$ in IFT is structurally a GNN, as both consist of sequences of local non-linear and non-local linear operations on their input vector $\zeta = (\xi, \eta)$. The concrete architecture of an IFT model and a typical GNN might differ significantly, as GNNs often map a lower dimensional latent space into a higher dimensional data or feature space, whereas the dimension of the IFT model latent space can be very high, as it contains a subset of the virtually infinite many degrees of freedom of a field, see Figure 1.

Additionally, the way IFT-based models and GNNs are usually used differs a bit. Both can be used to generate synthetic samples of outputs by processing random latent space vectors $\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1})$. However, typically an IFT model $d'(\zeta)$ is applied to infer all latent space variables in ζ from data d. From the latent variables, the signal of interest can always be recovered via $s(\zeta)$.

For this inference the so-called information Hamiltonian, potential, or energy

$$\begin{aligned} \mathcal{H}(d, \zeta) &= -\ln \mathcal{P}(d, \zeta) = -\ln \mathcal{P}(d|\zeta) - \ln \mathcal{P}(\zeta) \\ &= \mathcal{H}(n = d - d'(\zeta)|\zeta) + \tfrac{1}{2}\zeta^\dagger \zeta + \text{const} \end{aligned} \tag{43}$$

is investigated with respect to ζ, where $\mathcal{H}(a|b) := -\ln \mathcal{P}(a|b)$. This quantity is introduced to IFT in analogy to statistical mechanics, it summarizes the full knowledge on the problem (as it is just a logarithmic coordinate transformation in the space of probabilities) and has the nice property, that it allows to speak about information as an additive quantity, as $\mathcal{H}(a, b) = \mathcal{H}(a|b) + \mathcal{H}(b)$.

Investigating the relevant information Hamiltonian for our IFT problem $\mathcal{H}(d, \zeta)$ can be done, for example, by minimizing it to obtain a MAP estimator for ζ or—as discussed in

the next section—via variational inference (VI). In case of a constant, signal independent Gaussian white noise statistics, the information Hamiltonian becomes

$$\mathcal{H}(d,\zeta) = \frac{|d-d'(\zeta)|^2}{2\sigma_n^2} + \frac{1}{2}\zeta^\dagger\zeta + \text{const.} \tag{44}$$

The training of an usual GNN is done with a training data set $\widetilde{d} = (d_i)_i$ to which a corresponding latent space vector set $\widetilde{\zeta} = (\zeta_i)_i$ and common network parameters η need to be found. For this a loss function of the form

$$\widetilde{\mathcal{H}}(\widetilde{d},\widetilde{\xi},\eta) = \sum_i \widetilde{\mathcal{H}}(d_i|\xi_i,\eta) + \widetilde{\mathcal{H}}(\widetilde{\xi}|\eta) + \widetilde{\mathcal{H}}(\eta), \tag{45}$$

$$\widetilde{\mathcal{H}}(\widetilde{\xi}|\eta) = \frac{1}{2}\sum_i \xi_i^\dagger\xi + \text{const, and} \tag{46}$$

$$\widetilde{\mathcal{H}}(d_i|\xi_i,\eta) = \frac{1}{2}\frac{|d_i-d'(\xi_i,\eta)|^2}{2\sigma_n^2} + \text{const} \tag{47}$$

might be minimized. Here, a typical GNN data loss function $\widetilde{\mathcal{H}}(\widetilde{d}_i|\xi_i,\eta)$ as used for the decoder part of an autoencoder (AE) [49] was assumed. In an generative adversarial network (GAN) [50], however, this data loss function is given in terms of the output of a discriminator network. The network parameter prior term $\widetilde{\mathcal{H}}(\eta)$ might be chosen to be uninformative ($\widetilde{\mathcal{H}}(\eta) = \text{const}$) or informative (e.g., $\widetilde{\mathcal{H}}(\eta) = \frac{1}{2}\eta^\dagger\eta$ in case of a Gaussian prior on the parameters).

Anyhow, by comparison of Equations (45)–(47) with Equations (43) and (44), it should be apparent that the network loss functions can be structurally similar to the IFT information Hamiltonian. Both consist of a standardized quadratic prior-energy and a likelihood-energy and both can have a probabilistic interpretation in terms of being negative log-probabilities, e.g.,

$$\mathcal{P}(d,\xi,\eta) = e^{-\mathcal{H}(d,\xi,\eta)} \text{ and} \tag{48}$$

$$\mathcal{P}(\widetilde{d},\widetilde{\xi},\eta) = e^{-\widetilde{\mathcal{H}}(\widetilde{d},\widetilde{\xi},\eta)}, \tag{49}$$

respectively. For this reason, we do not distinguish between an information Hamiltonian $\mathcal{H}$ and a network loss function $\widetilde{\mathcal{H}}$ by writing $\mathcal{H}$ for both in the following.

The IFT-GNN can operate with solely a single data vector d due to the domain knowledge coded into their architecture, whereas usual GNNs require sets of data vectors $\widetilde{d} = (d_i)_i$ to be trained. Recently, more IFT-like architectures for GNNs were proposed as well, which are also able to process data without training [51].

4. Variational Inference

4.1. Basic Idea

So far, it has been assumed here that MAP estimators are used to determine network parameters ζ for both, IFT-based models as well as traditional GNNs. MAP estimators are known to be prone to over-fitting the data, as they are not probing the adjacent phase-space volumes of their solutions. VI methods perform better in that respect, while still being affordable in terms of computational costs for the high dimensional settings of IFT-based field inference and traditional GNN training. They were used in most recent IFT applications [12–18,21,52] and are prominently present in the name of variational autoencoders (VAEs) [45] that are built on VI.

In VI, the posterior $\mathcal{P}(\zeta|d)$ is approximated by a simpler probability distribution $\mathcal{Q}(\zeta|d')$, in many applications by a Gaussian

$$\mathcal{Q}(\zeta|\theta,\Theta) = \mathcal{G}(\zeta-\theta,\Theta), \tag{50}$$

where $d' = (\theta, \Theta)$. The Gaussian is chosen to minimize the variational Kullback–Leibler (KL) divergence

$$\begin{aligned} \mathrm{KL}_\zeta(d', d) &:= \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) \\ &= \int \mathcal{D}\zeta \, \mathcal{Q}(\zeta|d') \ln \frac{\mathcal{Q}(\zeta|d')}{\mathcal{P}(\zeta|d)} \end{aligned} \tag{51}$$

with respect to the parameters of d', θ and Θ in our case.

Ideally, all degrees of freedom (DoF) of θ and Θ are optimized. In practice, however, this is often not feasible due to the quadratic scaling of the number of DoF of Θ with that of θ. Three approximate schemes for handling the high dimensional uncertainty covariance will be discussed in the following, leading to the ADVI, MGVI, and geoVI techniques introduced below, namely

- mean field theory, in which Θ is assumed to be diagonal, as used by ADVI
- the usage of the Fisher information to approximate Θ as a function of θ and thereby effectively removing the DoF of Θ from the optimization problem as used by MGVI
- a coordinate transformation of the latent space that approximately standardizes the posterior and therefore sets the covariance to the identity matrix in the new coordinates, as performed by geoVI.

Before these are discussed, a note that applies to all of them is in order. Optimizing of the VI KL, Equation (51), is slightly sub-optimal from an information theoretical point of view as this minimizes the amount of information introduced by going from $\mathcal{P}$ to $\mathcal{Q}$. The expectation propagation (EP) KL with reversed arguments $\mathcal{D}_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q})$ would be better, as it minimizes the information loss from approximating $\mathcal{P}$ with $\mathcal{Q}$ [53]. VI is known to underestimate the uncertainties, whereas EP conservatively overestimates them. However, calculating the EP solution for θ and Θ would require integrating over the posterior. If this would be feasible, any posterior quantity of interest could be calculated as well and there would be no need to approximate $\mathcal{P}(\zeta|d)$ in the first place. Estimating and minimizing the VI KL $\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})$ is less demanding, as the integral over the simpler (Gaussian) distribution $\mathcal{Q}$ can very often be performed analytically, or by sample averaging using samples drawn from $\mathcal{Q}$.

4.2. ADVI and Mean Field Approximation

In all here discussed VI techniques, the posterior mean θ and the posterior uncertainty covariance Θ become parameters to be determined. The vector θ has the dimension N_{dim} of the latent space, whereas the posterior uncertainty covariance Θ has $N_{\mathrm{dim}}(N_{\mathrm{dim}} - 1)/2 = \mathcal{O}(N_{\mathrm{dim}}^2)$ independent DoF. For small problems, these might be solved for, however, for large problems with millions of DoF, these cannot even be stored in a computer memory. To circumvent this, the Automatic Differentiation Variational Inference (ADVI) algorithm [54] often invokes the so called mean field approximation (MFA). This assumes a diagonal covariance $\Theta_{\mathrm{MFA}} = \widehat{\theta'} \equiv \mathrm{diag}(\theta')$, with θ' being a latent space vector. Cross-correlations between parameters can not be represented by this, which is problematic in particular in combination with the tendency of VI to underestimate uncertainties.

4.3. MGVI and Fisher Information Metric

In order to overcome this limitation of ADVI that limits its usage in IFT contexts with their large number of DoF, the Metric Gaussian Variational Inference (MGVI) [30] algorithm approximates the posterior uncertainty of ζ with the help of the Fisher information metric

$$M(\zeta) := \left\langle \frac{\partial \mathcal{H}(d|\zeta)}{\partial \zeta} \frac{\partial \mathcal{H}(d|\zeta)}{\partial \zeta}^\dagger \right\rangle_{(d|\zeta)}. \tag{52}$$

The starting point for obtaining the uncertainty covariance Θ used in MGVI is the Hessian of the log-posterior

$$\begin{aligned}\frac{\partial^2 \mathcal{H}(\zeta|d)}{\partial\zeta\partial\zeta^\dagger} &= \frac{\partial^2 \mathcal{H}(d,\zeta)}{\partial\zeta\partial\zeta^\dagger} - \underbrace{\frac{\partial^2 \mathcal{H}(d)}{\partial\zeta\partial\zeta^\dagger}}_{=0} \\ &= \frac{\partial^2 \mathcal{H}(d,\zeta)}{\partial\zeta\partial\zeta^\dagger} \end{aligned} \tag{53}$$

as a first guess for the approximate posterior precision matrix Θ^{-1}. Using this evaluated at the minimum ζ_{MAP} of the information Hamiltonian $\mathcal{H}(d,\zeta)$ would correspond to the Laplace approximation, in which the posterior is replaced by a Gaussian obtained from doing a saddle point approximation at its maximum.

However, neither is the MAP solution ideal, as discussed above, nor would this be a good approximation at many locations ζ that differ from ζ_{MAP}. This is because positive definiteness of the Hessian is not guaranteed there, but it is an essential property of any correlation and precision matrix. For this reason, Θ^{-1} cannot directly be approximated by this Hessian.

It turns out that the likelihood averaged Hessian is strictly positive definite, and is therefore a candidate for an approximate posterior precision matrix for any guessed posterior mean θ. A short calculation shows that the likelihood averaged Hessian is indeed positive definite:

$$\begin{aligned}\Theta^{-1}(\theta) &\approx \left\langle \frac{\partial^2 \mathcal{H}(d,\zeta)}{\partial\zeta\partial\zeta^\dagger} \right\rangle_{(d|\zeta=\theta)} \\ &= \left\langle \frac{\partial^2 \mathcal{H}(\zeta)}{\partial\zeta\partial\zeta^\dagger} + \frac{\partial^2 \mathcal{H}(d|\zeta)}{\partial\zeta\partial\zeta^\dagger} \right\rangle_{(d|\zeta=\theta)} \\ &= \left\langle \mathbb{1} - \frac{\partial^2 \ln \mathcal{P}(d|\zeta)}{\partial\zeta\partial\zeta^\dagger} \right\rangle_{(d|\zeta=\theta)} \\ &= \mathbb{1} - \left\langle \frac{1}{\mathcal{P}(d|\zeta)} \frac{\partial^2 \mathcal{P}(d|\zeta)}{\partial\zeta\partial\zeta^\dagger} \right\rangle_{(d|\zeta=\theta)} \\ &\quad + \left\langle \frac{1}{\mathcal{P}^2(d|\zeta)} \frac{\partial \mathcal{P}(d|\zeta)}{\partial\zeta} \frac{\partial \mathcal{P}(d|\zeta)}{\partial\zeta}^\dagger \right\rangle_{(d|\zeta=\theta)} \\ &= \mathbb{1} - \int \mathrm{d}d \, \frac{\cancel{\mathcal{P}(d|\zeta)}}{\cancel{\mathcal{P}(d|\zeta)}} \frac{\partial^2 \mathcal{P}(d|\zeta)}{\partial\zeta\partial\zeta^\dagger} \\ &\quad + \left\langle \frac{\partial \mathcal{H}(d|\zeta)}{\partial\zeta} \frac{\partial \mathcal{H}(d|\zeta)}{\partial\zeta}^\dagger \right\rangle_{(d|\zeta=\theta)} \\ &= \mathbb{1} - \underbrace{\frac{\partial^2}{\partial\zeta\partial\zeta^\dagger} \underbrace{\int \mathrm{d}d \, \mathcal{P}(d|\zeta)}_{=1}}_{=0} + M(\theta) \\ &= \mathbb{1} + M(\theta) > 0. \end{aligned} \tag{54}$$

The last step follows because the Fisher metric $M(\theta)$ is an average over outer products ($v\,v^\dagger \geq 0$) of likelihood Hamiltonian gradient vectors $v = \partial\mathcal{H}(d|\zeta)/\partial\zeta$ and thereby positive semi-definite. Adding $\mathbb{1} > 0$ to the Fisher metric turns the approximate precision matrix into a positive definite matrix $\Theta^{-1}(\theta) > 0$, of which the inverse $\Theta(\theta)$ exists for all θ, and which is positive definite as well.

4.4. Exact Uncertainty Covariance

Being positive definite is of course not the only property an approximation of the posterior uncertainty covariance has to fulfill. It also has to approximate well. Fortunately, this seems to be the case in many situations. The likelihood averaged Laplace approximation actually becomes the exact posterior uncertainty in case of linear Gaussian measurement problems as is shown in the following. If it is exact in such linear situations, it should be a valid approximation in the vicinity of any linear case.

For linear measurement problems, the measurement equation is of the form $d = R\zeta + n$, the noise statistics $\mathcal{P}(n|\zeta) = \mathcal{G}(n, N)$, and the standardized prior is $\mathcal{P}(\zeta) = \mathcal{G}(\zeta, \mathbb{1})$. The corresponding posterior is known to be a Gaussian

$$\mathcal{P}(\zeta|d) = \mathcal{G}(\zeta - m, D) \tag{55}$$

with mean m and covariance D given by the generalized Wiener filter solution $m = D\, R^\dagger N^{-1} d$ and the Wiener covariance $D = (\mathbb{1} + R^\dagger N^{-1} R)^{-1}$, respectively (e.g., [1]). In this case, the Fisher information metric $M = R^\dagger N^{-1} R$ is independent of ζ. The approximate posterior uncertainty covariance as given by Equation (54) equals the exact posterior covariance, $\Theta = (\mathbb{1} + M)^{-1} = (\mathbb{1}^{-1} + R^\dagger N^{-1} R)^{-1} = D$. Thus indeed, the adopted approximation becomes exact in this situation. This should show why this approximation can hold sensible results in sufficiently well behaved cases, in particular when a linearization of the inference problem around a reference solution (e.g., a MAP estimate) is already a good approximation.

Furthermore, for all signal space directions around this reference point that are unconstrained by the data, this covariance approximation returns the prior uncertainty, as it should. Additional discussion of this approximation can be found in Knollmüller and Enßlin [30], where also its performance with respect to ADVI is numerically investigated.

The important point about this approximate uncertainty covariance Θ is that it is a function of the latent space mean estimate θ, i.e., $\Theta(\theta)$, and therefore does not need to be inferred as well. For many likelihoods, the Fisher metric is available analytically, alleviating the need to store Θ in a computer memory as an explicit matrix. It is only necessary that certain operations can be performed with Θ, like applying it to a vector or drawing samples from a Gaussian with this covariance. Relying solely on those memory inexpensive operations, the MGVI algorithm is able to minimize the relevant VI KL, namely $\mathrm{KL}_\zeta((\theta, \Theta(\theta)), d) = \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}, \mathcal{P})$, with respect to the approximate posterior mean θ. The result of MGVI are then the posterior mean θ, the uncertainty covariance $\Theta(\theta)$, and posterior samples $\{\zeta_i\}_i$ drawn according to this mean and covariance. These samples can then be propagated into posterior signal samples $s_i = s(\zeta_i)$, from which any desired posterior signal statistics can be calculated.

MGVI has enabled field inference for problems, which are too complex to be solved by MAP, in particular when multiple layers of hyperpriors were involved (e.g., [13,15]).

4.5. Geometric Variational Inference

ADVI's and MGVI's weak point, however, can be the Gaussian approximation of the posterior, which might be strongly non-Gaussian in certain applications. In order to overcome this, the geometrical variational inference (geoVI) algorithm [32] was introduced as an extension of MGVI. geoVI puts another coordinate transformation on top of the one used by MGVI, so that $\zeta = g_0(y)$ — with g_0 to be performed before any of the other IFT-GNN operations $g_1, \dots g_n$ — approximately standardizes the posterior, $\mathcal{P}(y|d) \approx \mathcal{G}(y, \mathbb{1})$. Astonishingly, this transformation can be constructed without the (prohibitive) usage of any explicit matrix or higher order tensor in the latent space, thus also allowing us to tackle very high dimensional inference problems, like MGVI. The transformation is basically a normalizing flow (network) [55], just with the difference to their usual usage in ML, that the geoVI flow does not need to be trained, but is derived from the problem statement in

form of its information Hamiltonian in an automated fashion. Specifically, the coordinate transformation g_0 is defined to solve the constraining equation

$$\left.\frac{\partial g_0}{\partial y}^\dagger \Theta(\zeta) \frac{\partial g_0}{\partial y}\right|_{\zeta=g_0(y)} \approx \mathbb{1} \quad \forall y, \tag{56}$$

which fully specifies g_0 up to an integration constant θ. This remaining constant is solved for by minimizing the VI KL with respect to θ to retrieve the optimal geoVI aproximation.

With geoVI, deeper hierarchical models, which more often exhibit non-Gaussian posteriors due to a larger number of degenerate parameters in them, can be approached via VI. The ability of geoVI to provide uncertainty information is illustrated in Figure 2 (bottom middle and right panels) and in Figure 3. Further details on geoVI and detailed comparisons of ADVI, MGVI, geoVI, and Hamiltonian Monte Carlo methods can be found in [32].

5. Conclusions and Outlook

This paper argues that IFT techniques can well be regarded as ML and AI methods by showing their interrelation with GNNs, normalizing flows, and VI techniques. This insight is not necessarily new, as this paper just summarizes a number of recent works [29–32] that suggested this before.

First, the generative models build and used in IFT are GNNs that can interpret data without initial training, thanks to the domain knowledge coded into their architecture [29]. Related architectures have very recently been proposed as image priors in the context of neural network architectures as well [51]. As IFT models and the newly proposed image priors do not obtain their intelligence from data driven learning, they are strictly not ML techniques, but might be characterized as (expert) knowledge-driven AI systems. From a technical point of view, however, such a distinction could be seen as splitting hairs.

Second, the VI algorithms used in IFT and AI to approximately infer quantities are a natural interface between these areas. Here, the related ADVI [54], MGVI [30], and geoVI [32] algorithms were briefly discussed, which can be used in classical ML and AI as well as in IFT applications.

And third, the common probabilistic formulation of IFT models and GNNs, as well as the common VI infrastructure of the two areas allows for combining pre-trained GNNs and other networks with IFT-style model components. In that respect, the possibility to perform Bayesian reasoning with trained neural networks as described in [31] might give an outlook on the potential to combine IFT with other ML and AI methods.

To summarize, IFT [1,2] addresses perception [5–21], reasoning [22–26,31], and adaptive inference [30,32] tasks. All these are central to the aims of AI and ML to build intelligent systems including such for perception, cognition, and learning.

Funding: This research received no external funding.

Data Availability Statement: The presented data are synthetically generated via the script `getting_started_3.py` enclosed in the open source software package NIFTy in its version 8, downloadable at https://gitlab.mpcdf.mpg.de/ift/nifty/-/tree/NIFTy_8 (accessed on 17 December 2021).

Acknowledgments: I am grateful to many colleagues and students that helped me to understand the relation of IFT and AI. The line of thoughts presented here benefited particularly from discussions with Philipp Arras, Philipp Frank, Jakob Knollmüller, and Reimar Leike. I thank Philipp Arras, Vincent Eberle, Gordian Edenhofer, Johannes-Harth-Kitzerow, Philipp Frank, Jakob Roth, and three constructive anonymous reviewers for detailed comments on the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Enßlin, T.A.; Frommert, M.; Kitaura, F.S. Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis. *Phys. Rev. D* **2009**, *80*, 105005. [CrossRef]
2. Enßlin, T.A. Information Theory for Fields. *Ann. Phys.* **2019**, *531*, 1800127. [CrossRef]
3. Bialek, W.; Zee, A. Statistical mechanics and invariant perception. *Phys. Rev. Lett.* **1987**, *58*, 741–744. [CrossRef] [PubMed]
4. Lemm, J.C. *Bayesian Field Theory*; JHU Press: Baltimore, MD, USA, 2003.
5. Oppermann, N.; Junklewitz, H.; Robbers, G.; Bell, M.R.; Enßlin, T.A.; Bonafede, A.; Braun, R.; Brown, J.C.; Clarke, T.E.; Feain, I.J.; et al. An improved map of the Galactic Faraday sky. *Astron. Astrophys.* **2012**, *542*, A93. [CrossRef]
6. Oppermann, N.; Junklewitz, H.; Greiner, M.; Enßlin, T.A.; Akahori, T.; Carretti, E.; Gaensler, B.M.; Goobar, A.; Harvey-Smith, L.; Johnston-Hollitt, M.; et al. Estimating extragalactic Faraday rotation. *Astron. Astrophys.* **2015**, *575*, A118. [CrossRef]
7. Junklewitz, H.; Bell, M.R.; Enßlin, T. A new approach to multifrequency synthesis in radio interferometry. *Astron. Astrophys.* **2015**, *581*, A59. [CrossRef]
8. Imgrund, M.; Champion, D.J.; Kramer, M.; Lesch, H. A Bayesian method for pulsar template generation. *Mon. Not. R. Astronmical Soc.* **2015**, *449*, 4162–4183. [CrossRef]
9. Selig, M.; Vacca, V.; Oppermann, N.; Enßlin, T.A. The denoised, deconvolved, and decomposed Fermi γ-ray sky. An application of the D^{3}PO algorithm. *Astron. Astrophys.* **2015**, *581*, A126. [CrossRef]
10. Dorn, S.; Greiner, M.; Enßlin, T.A. All-sky reconstruction of the primordial scalar potential from WMAP temperature data. *J. Cosmol. Astropart. Phys.* **2015**, *2015*, 041. [CrossRef]
11. Knollmüller, J.; Frank, P.; Enßlin, T.A. Separating diffuse from point-like sources—A Bayesian approach. *arXiv* **2018**, arXiv:1804.05591.
12. Arras, P.; Frank, P.; Leike, R.; Westermann, R.; Enßlin, T.A. Unified radio interferometric calibration and imaging with joint uncertainty quantification. *Astron. Astrophys.* **2019**, *627*, A134. [CrossRef]
13. Hutschenreuter, S.; Enßlin, T.A. The Galactic Faraday depth sky revisited. *Astron. Astrophys.* **2020**, *633*, A150. [CrossRef]
14. Leike, R.H.; Glatzle, M.; Enßlin, T.A. Resolving nearby dust clouds. *Astron. Astrophys.* **2020**, *639*, A138. [CrossRef]
15. Arras, P.; Frank, P.; Haim, P.; Knollmüller, J.; Leike, R.; Reinecke, M.; Enßlin, T. M87* in space, time, and frequency. *arXiv* **2020**, arXiv:2002.05218.
16. Arras, P.; Bester, H.L.; Perley, R.A.; Leike, R.; Smirnov, O.; Westermann, R.; Enßlin, T.A. Comparison of classical and Bayesian imaging in radio interferometry. *arXiv* **2020**, arXiv:2008.11435.
17. Hutschenreuter, S.; Anderson, C.S.; Betti, S.; Bower, G.C.; Brown, J.A.; Brüggen, M.; Carretti, E.; Clarke, T.; Clegg, A.; Costa, A.; et al. The Galactic Faraday rotation sky 2020. *arXiv* **2021**, arXiv:2102.01709.
18. Mertsch, P.; Vittino, A. Bayesian inference of three-dimensional gas maps. I. Galactic CO. *Astron. Astrophys.* **2021**, *655*, A64. [CrossRef]
19. Davis, J.H.; Enßlin, T.; Bœhm, C. New method for analyzing dark matter direct detection data. *Phys. Rev. D* **2014**, *89*, 043505. [CrossRef]
20. Huang, X.; Enßlin, T.; Selig, M. Galactic dark matter search via phenomenological astrophysics modeling. *J. Cosmol. Astropart. Phys.* **2016**, *2016*, 030. [CrossRef]
21. Welling, C.; Frank, P.; Enßlin, T.; Nelles, A. Reconstructing non-repeating radio pulses with Information Field Theory. *J. Cosmol. Astropart. Phys.* **2021**, *2021*, 071. [CrossRef]
22. Selig, M.; Oppermann, N.; Enßlin, T.A. Improving stochastic estimates with inference methods: Calculating matrix diagonals. *Phys. Rev. E* **2012**, *85*, 021134. [CrossRef] [PubMed]
23. Enßlin, T.A. Information field dynamics for simulation scheme construction. *Phys. Rev. E* **2013**, *87*, 013308. [CrossRef] [PubMed]
24. Leike, R.H.; Enßlin, T.A. Towards information-optimal simulation of partial differential equations. *Phys. Rev. E* **2018**, *97*, 033314. [CrossRef] [PubMed]
25. Kurthen, M.; Enßlin, T. A Bayesian Model for Bivariate Causal Inference. *Entropy* **2019**, *22*, 46. [CrossRef] [PubMed]
26. Frank, P.; Leike, R.; Enßlin, T.A. Field Dynamics Inference for Local and Causal Interactions. *Ann. Phys.* **2021**, *533*, 2000486. [CrossRef]
27. Enßlin, T.A.; Knollmüller, J. Correlated signal inference by free energy exploration. *arXiv* **2016**, arXiv:1612.08406.
28. Leike, R.; Enßlin, T. Optimal Belief Approximation. *Entropy* **2017**, *19*, 402. [CrossRef]
29. Knollmüller, J.; Enßlin, T.A. Encoding prior knowledge in the structure of the likelihood. *arXiv* **2018**, arXiv:1812.04403.
30. Knollmüller, J.; Enßlin, T.A. Metric Gaussian Variational Inference. *arXiv* **2019**, arXiv:1901.11033.
31. Knollmüller, J.; Enßlin, T.A. Bayesian Reasoning with Trained Neural Networks. *Entropy* **2021**, *23*, 693. [CrossRef]
32. Frank, P.; Leike, R.; Enßlin, T.A. Geometric Variational Inference. *Entropy* **2021**, *23*, 853. [CrossRef] [PubMed]
33. Edward, C. *Rasmussen and Christopher KI Williams. Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 211, p. 212.
34. Lassas, M.; Siltanen, S. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.* **2004**, *20*, 1537. [CrossRef]
35. Saksman, M.L.; Siltanen, S. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging* **2009**, *3*, 87.
36. Enßlin, T.A.; Frommert, M. Reconstruction of signals with unknown spectra in information field theory with parameter uncertainty. *Phys. Rev. D* **2011**, *83*, 105014. [CrossRef]

37. Enßlin, T.A.; Weig, C. Inference with minimal Gibbs free energy in information field theory. *Phys. Rev. E* **2010**, *82*, 051112. [CrossRef] [PubMed]
38. Westerkamp, M.; Ovchinnikov, I.; Frank, P.; Enßlin, T. Dynamical Field Inference and Supersymmetry. *Entropy* **2021**, *23*, 1652. [CrossRef]
39. Pumpe, D.; Reinecke, M.; Enßlin, T.A. Denoising, deconvolving, and decomposing multi-domain photon observations. The D^4PO algorithm. *Astron. Astrophys.* **2018**, *619*, A119. [CrossRef]
40. Nunez, P.L. The brain wave equation: A model for the EEG. *Math. Biosci.* **1974**, *21*, 279–297. [CrossRef]
41. Amari, S.I. Homogeneous nets of neuron-like elements. *Biol. Cybern.* **1975**, *17*, 211–220. [CrossRef]
42. Amari, S.i. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **1977**, *27*, 77–87. [CrossRef]
43. Coombes, S.; Potthast, R. Tutorial on neural field theory. In *Neural Fields*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1–43.
44. Frank, P.; Steininger, T.; Enßlin, T.A. Field dynamics inference via spectral density estimation. *Phys. Rev. E* **2017**, *96*, 052104. [CrossRef] [PubMed]
45. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
46. Selig, M.; Bell, M.R.; Junklewitz, H.; Oppermann, N.; Reinecke, M.; Greiner, M.; Pachajoa, C.; Enßlin, T.A. NIFTY—Numerical Information Field Theory. A versatile PYTHON library for signal inference. *Astron. Astrophys.* **2013**, *554*, A26. [CrossRef]
47. Steininger, T.; Dixit, J.; Frank, P.; Greiner, M.; Hutschenreuter, S.; Knollmüller, J.; Leike, R.; Porqueres, N.; Pumpe, D.; Reinecke, M.; et al. NIFTy 3—Numerical Information Field Theory: A Python Framework for Multicomponent Signal Inference on HPC Clusters. *Ann. Phys.* **2019**, *531*, 1800290. [CrossRef]
48. Arras, P.; Baltac, M.; Ensslin, T.A.; Frank, P.; Hutschenreuter, S.; Knollmueller, J.; Leike, R.; Newrzella, M.N.; Platz, L.; Reinecke, M.; et al. NIFTy5: Numerical Information Field Theory v5. 2019. Available online: https://ascl.net/1903.008 (accessed on 6 March 2022).
49. Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **1991**, *37*, 233–243. [CrossRef]
50. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
51. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. *Int. J. Comput. Vis.* **2020**, *128*, 1867–1888. [CrossRef]
52. Leike, R.H.; Enßlin, T.A. Charting nearby dust clouds using Gaia data only. *Astron. Astrophys.* **2019**, *631*, A32. [CrossRef]
53. Leike, R.H.; Enßlin, T.A. Operator calculus for information field theory. *Phys. Rev. E* **2016**, *94*, 053306. [CrossRef]
54. Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic differentiation variational inference. *J. Mach. Learn. Res.* **2017**, *18*, 430–474.
55. Rezende, D.; Mohamed, S. Variational Inference with Normalizing Flows. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 37, pp. 1530–1538.