DISCRETE AND CONTINUOUS
DYNAMICAL SYSTEMS SERIES S
Volume **15**, Number **4**, April **2022**

doi:10.3934/dcdss.2021088

pp. **877–892**

ANALYTIC CONTINUATION OF NOISY DATA USING ADAMS BASHFORTH RESIDUAL NEURAL NETWORK

Xuping Xie

New York University, New York, NY, 10012

Feng Bao*

Florida State University, Tallahassee, FL, 32304

Thomas Maier

Oak Ridge National Laboratory, Oak Ridge, TN, 37830

Clayton Webster

University of Tennessee-Knoxville, Knoxville, TN, 37916

Abstract. We propose a data-driven learning framework for the analytic continuation problem in numerical quantum many-body physics. Designing an accurate and efficient framework for the analytic continuation of imaginary time using computational data is a grand challenge that has hindered meaningful links with experimental data. The standard Maximum Entropy (MaxEnt)-based method is limited by the quality of the computational data and the availability of prior information. Also, the MaxEnt is not able to solve the inversion problem under high level of noise in the data. Here we introduce a novel learning model for the analytic continuation problem using a Adams-Bashforth residual neural network (AB-ResNet). The advantage of this deep learning network is that it is model independent and, therefore, does not require prior information concerning the quantity of interest given by the spectral function. More importantly, the ResNet-based model achieves higher accuracy than MaxEnt for data with higher level of noise. Finally, numerical examples show that the developed AB-ResNet is able to recover the spectral function with accuracy comparable to MaxEnt where the noise level is relatively small.

1. **Introduction.** Analytic continuation is a challenging problem appears in pure mathematics, applied physics, and other branches of applied sciences. The problem can be formulated mathematically in the realm of complex analysis. Given a complex function f with domain $\Omega \subset \mathbb{C}$,

$$f : \Omega \subset \mathbb{C} \to \mathbb{C},$$

analytic continuation process is to find an analytic complex function $\tilde{f} : \tilde{\Omega} \subset \mathbb{C} \to \mathbb{C}$ satisfies,

$$\tilde{f}(z) = z, \forall z \in \Omega$$

Analytic continuation can be found in a broad range of physical studies, such as quantum field theory, condensed matter physics, image reconstruction, etc. For

2020 *Mathematics Subject Classification.* Primary: 45B05; Secondary: 32W50, 49N30.

Key words and phrases. Analytic continuation, inverse problem, stochastic optimization, machine learning, neural network.

* Corresponding author.

many situations mentioned above, the knowledge of $f(\Omega)$ can be affected by uncertainties that come from numerical or experimental determination of the value of the function. In this paper, we focus on one type of physical applications arises in condensed matter physics where analytic continuation originate from a mapping between real time and imaginary time:

$$f(t) \longleftrightarrow f(-i\tau).$$

Calculations for the imaginary-time propagator are generally well behaved. Methods to compute imaginary-time correlation functions, e.g., quantum Monte Carlo (QMC), are crucial for the study of strongly correlated physical systems. QMC methods are intrinsically formulated in imaginary time and yield estimation of correlation functions. It is thus challenging to perform the analytic continuation to infer real-time properties. Specifically, one typical problem in this context is the estimation of spectral functions of many-body quantum systems starting from imaginary-time correlation functions. In this work, we introduce a novel data-driven framework with state-of-the-art deep neural network method for the analytic continuation process of density spectral function estimation.

The spectral function of a many body physical system is defined as follows,

$$A(\omega) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{i\omega t} \langle e^{\frac{i\hat{H}t}{h}} \hat{A} e^{-\frac{i\hat{H}t}{h}} \hat{B} \rangle dt, \tag{1}$$

where $\langle \cdot \rangle$ denotes expectation value on the state, $\hat{A}, \hat{B}$ are given operators acting on the Hilbert space of the system whose Hamiltonian operator is $\hat{H}$, and h is Planck's constant. The imaginary-time correlation function for the many-body system is defined as

$$G(\tau) = \langle e^{\frac{\hat{H}\tau}{h}} \hat{A} e^{-\frac{\hat{H}\tau}{h}} \hat{B} \rangle. \tag{2}$$

The two equations (1) and (2) are related by the celebrated *Wick rotation*, a mapping between the quantum mechanical evolution operator and the imaginary-time propagator:

$$e^{-\frac{i\hat{H}t}{h}} \longleftrightarrow e^{-\frac{\tau\hat{H}}{h}}.$$

For analytic continuation, the spectral function $A(\omega)$ is what are looking for while $G(\tau)$ corresponds to the observations. These two functions are related from a Fourier transform and the Wick rotation mapping. Thus, an inverse problem can be formulated by a Fredholm integral equation

$$G(\tau) = \int_{-\infty}^{\infty} \mathcal{K}(\tau, \omega) A(\omega) d\omega, \tag{3}$$

where the kernel function $\mathcal{K}(\tau, \omega) = \theta(\omega) e^{-\tau\omega}$ can be complemented by a prioir knowledge and defined in different ways for some properties with $\theta(\omega)$ being the Heaviside distribution [5, 7]. We use Fermionic kernel distribution in this work.

Quantum Monte Carlo (QMC) methods are widely used to study the finite temperature physics of strongly interacting electron systems. The underlying algorithms are generally formulated on the imaginary time axis to treat the finite temperature dynamics of the many-body system. To extract the real time dynamics, an additional analytic continuation of the imaginary time τ data to the real time or frequency ω-axis is required to extract the quantity of interest. This process is a highly ill-conditioned inverse problem so that small perturbations of the input data

result in large uncertainties in the resulting spectral function $A(\omega)$. The challenge is rooted in the integral equation

$$G(\tau) = \int_{-\infty}^{\infty} \frac{e^{-\tau\omega}}{1+e^{-\beta\omega}} A(\omega) d\omega \,. \tag{4}$$

Here $G(\tau)$ is the imaginary time QMC data for a fermionic observable such as the single-particle Green's function, $K(\tau,\omega) = \exp(-\tau\omega)/(1+\exp(-\beta\omega))$ is the Fermionic kernel function with $\beta = 1/T$ the inverse temperature, and $A(\omega)$ is the quantity of interest. The process of inverting this equation is numerically unstable because of the exponentially small tails in the kernel function for large ω, and is especially sensitive to the Monte Carlo sampling error in $G(\tau)$ [27].

Several approaches have been proposed to address the analytic continuation problem. The most commonly used framework based on Bayesian inference is the MaxEnt method [17], pioneered in the works [22, 39] for the analytic continuation problem given by Eq. (4). The MaxEnt method regularizes the inversion problem through the introduction of an entropy-like term that measures the deviation from a default spectrum and then determines the most probable spectrum $A(\omega)$ using deterministic optimization. A related method that uses consistent constraints for the regularization was introduced in [37]. Both methods have the drawback that prior information about the possible spectrum $A(\omega)$ is needed for the regularization.

An alternative idea, which in principle does not rely on prior information is based on stochastic optimization. The work [38] uses Monte Carlo sampling of possible spectra weighted by Boltzmann weights with a fictitious temperature. This method was later related to MaxEnt in a certain limit in the paper [5, 16]. The effort [16] showed that Bayesian inference can be used to eliminate the ficticious temperature in a similar fashion as the regularization parameter of the MaxEnt approach is removed. Moreover, the work [34] developed a stochastic optimization based method to randomly sample possible optimal solutions $A(\omega)$, which implicitly regularizes the problem by allowing less optimal solutions. Also, stochastic inference approaches based on bayesian statistic for the analytic continuation of QMC data was studied in [1, 16, 22]. A more accessible and less complex variant of this approach that uses a Gaussian process for implicit regularization was recently introduced in [4, 13] and shown to provide spectra similar to MaxEnt.

With the modern development and success of deep learning, the data-driven discovery of physical systems becomes extremely popular for many applications. The artificial neural network has been widely used to study physics-related problems. Much recent work has been proposed on the mathematical connections between residual neural network (ResNet) and differential equations, see, e.g., [6, 29, 31, 33]. The work [11] introduced ODE-net which parametrize the derivative of the hidden state using deep ResNet. Other efforts [9, 10, 18, 32] proposed the dynamical system view of ResNet and provide connections between numerical ODE and deep ResNet architecture. Moreover, the functional approximation ability of ResNet has also been explored [14, 40, 41]. The work [30] proved ResNet can be considered as a universal approximator with one hidden layer and has certain advantages to fully connected neural networks. Data-driven models with machine learning for the analytic continuation problem have been introduced recently. [35] proposed a sparse modeling approach to eliminated redundant degrees of freedom for solving the ill-conditioned analytic continuation. The paper [15] presented a general framework for building an artificial neural network to approximate the kernel of the inversion. [44]

introduced the convolution neural network based machine learning method with stochastic gradient descent optimizer to train the continuation kernel. Motivated by the recent development of residual networks [19], we propose a Multistep residual network architecture with Adams-Bashforth scheme to generate a more stable inversion of the kernel under high-noise data for the analytic continuation problem.

This paper is strucuted as follows: Section 1 introduces the background information about analytic continuation in details. In Section 2, we describe the general computational framework for analytic continuation with details about the MaxEnt method in Section 2.2. In Section 3, we present the recent mathematical interpretation of ResNet and our new network architecture. In Section 4, we demonstrate the effectiveness of our method by using numerical experiments. In Section 5, we discuss some further works need to investigate for our model.

2. Analytic continuation.

2.1. Bayesian statistics.

To computationally solve the equation (4) for spectral $A(\omega)$, we discretize the real frequency axis into N intervals, $\{\omega_n\}_{i=0}^N$, the numerical inversion of (4) becomes,

$$G_i = \sum_{i=0}^{N} K_{il} A_l, \tag{5}$$

where the discretized kernel $K_{il} = \Delta\omega_l/(i\omega_n - \omega_l)$, A_l is the discretized spectral function, G_n is an observable single particle Green's function measured in frequency $i\omega_n$, and $\Delta\omega_l = \omega_{l+1} - \omega_l$ is the frequency interval. For the linear system (5), the matrix K_{il} is ill-conditioned which causes large errors in the quantity of interest A_l with small deviations of G_i. There exist several approaches for solving the analytic continuation problem that regularize the problem by making use of prior knowledge. Among these methods, the most common approach is the maximum entropy (MaxEnt) method, which is based on Bayesian statistical inference. For the equation (5), one can consider the Bayesian formula,

$$P(A|G) \propto P(G|A)P(A), \tag{6}$$

with $P(A|G)$ proportional to the posterior probability of the spectrum A given the data G, $P(A)$ is the prior probability contains prior information about A, and $P(G|A)$ is the likelihood function that measures the quality of the fit between G and KA. The MaxEnt method is to find the most probable spectrum A that maximizes the conditional probability of $P(A|G)$, which is equivalent to optimizing the likelihood function and prior,

$$\max P(A|G) \propto \max P(G|A)P(A). \tag{7}$$

2.2. Maximum entropy.

In MaxEnt, prior information $P(A)$ is added by specifying a default distribution $A(\omega)$ that corresponds to the expected results in the absence of data. The algorithm iteratively searches for a distribution that maximizes the entropy with respect to $A(\omega)$. It can be formulated as a least square fitting. The likelihood function defined according to central limit theorem as

$$P(G|A) = e^{-\chi^2[A]/2}, \tag{8}$$

where

$$\chi^2[A] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{G_i - \sum_{i=1}^{N} K_{il}A_l}{\sigma_i}\right)^2 \tag{9}$$

represents the quality of the fit of G computed by the spectrum distribution $A(\omega)$ in (5). G_i is the mean value of the total number of M different quantum Monte Carlo (QMC) samples, i.e.

$$G_i = \frac{1}{M}\sum_{n=1}^{S} G_i^n, \tag{10}$$

and the variance

$$\sigma_i^2 = \frac{1}{S-1}\sum_{n=1}^{M}(G_i^n - G_i)^2. \tag{11}$$

The above formulation assumes there are no correlations between different frequencies presented in the QMC sample data G_i. In general, when correlations are considered, the covariance matrix has to be diagonalized with both the QMC data G_i and the kernel K have to be rotated into diagonal representation [36]. The quality of fitting for the samples A_i and the corresponding G_i becomes,

$$\chi^2[A^i] = \sum_{n,m=1}^{M}(G_i^m - G^n)C_{mn}^{-1}(G_i^m - G^n), \tag{12}$$

with C_{mn} being the diagonal covariance matrix. The entropy term, also named as Kullback-Leibler (KL) divergence, which measures the difference between distributions is defined relative to a positive definite and normalized function $D(\omega)$,

$$\begin{aligned} S[A] &= -\int [A(\omega) - D(\omega) - A(\omega)\ln\frac{A(\omega)}{D(\omega)}]d\omega \\ &= -\sum_{l=1}^{N}[A(\omega_l) - D(\omega_l) - A(\omega_l)\ln\frac{A(\omega_l)}{D(\omega_l)}]\Delta\omega_l \end{aligned} \tag{13}$$

The prior distribution in (7) is given by

$$P(A) = e^{\alpha S[A]}, \tag{14}$$

where α is a positive constant representing the regularization parameter for the optimization problem. The MaxEnt method uses least-square to minimize χ^2 with KL divergence as the regularization $S[A]$, namely:

$$S[A] = -\int d\omega A(\omega)\ln\left(\frac{A(\omega)}{d(\omega)}\right). \tag{15}$$

Therefore, instead of maximizing the posterior probability $P(A|G)$, the MaxEnt method minimizes the following least square function with the following regularization

$$Q[A] = \frac{1}{2}\chi^2[A] - \alpha S[A] \tag{16}$$

It is obvious that the Bayesian optimization of the posterior probability $P(A|G)$ is a deterministic optimization for the regularized least square fitting $Q[A]$. The parameter α controls the weights between $\chi^2[A]$ and the prior information contained in the entropy $S[A]$ to prevent over-fitting. There are several methods for fixing α, which often yield different results when applied in practice. In our numerical tests, the MaxEnt results are obtained by averaging over the optimal spectra A_α with various α.

2.3. **Data-driven learning.** With the fast development of machine learning, data-driven modeling is becoming increasingly important in research areas such as quantum mechanics, Monte Carlo methods, and computational physics. The MaxEnt approach based on Bayesian statistical methods for solving Fredholm integral equation (4) has been successfully applied in many situations. Recently, several works have shown that the machine learning approach is suitable for solving inverse problems in analytic continuation [2, 12, 23, 28]. The main idea of these data-driven methods is to distill the prior knowledge into simulated training datasets allowing higher flexibility in the regularization of the dataset compared to the MaxEnt method. [12] introduced a reinforcement learning framework for solving the Fredholm inverse integral equation. [2] utilizes the data-driven approach for the fermionic spectral density function (4). A database of spectral functions that resemble experimental data and the corresponding Green's function was considered in their work. A Kernel ridge regression performed on the database for training yielded results comparable to those obtained via MaxEnt. In [15], an artificial neural network was used for solving a physically more relevant scenario with known Hamiltonian and the data of interest obtained from QMC simulations.

The general framework of data-driven learning approach for solving the analytic continuation problem (4) is illustrated in Fig. 1. Unlike the classical MaxEnt

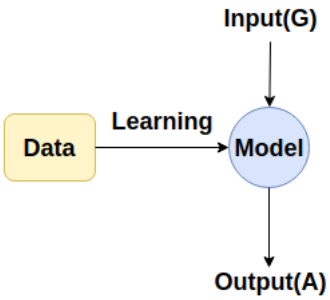

FIGURE 1. Illustration of data-driven learning framework for analytic continuation

method, the data-driven method produces the quantity of interest, i.e. the spectral density function A, directly from a learned model given the input data G. The model is trained offline from a large dataset. One advantage of this approach is that the prior information is implicitly embedded in the training dataset, whereas the MaxEnt method uses prior knowledge as the regularizer in the optimization. We can see that the dataset and model selection are crucial to the performance of the data-driven approach as they will determine the accuracy of the output. We use a neural network based model as a convenient framework in this paper. First, the universal approximation theorem ensures that neural networks can approximate any kind of continuous functions under mild assumptions. Moreover, the availability of powerful libraries allows for an efficient implementation of different network architecture that can take advantage of data structures, thus making the neural network a very versatile tool. Inspired by the recent work of the neural ordinary differential equation, we propose a novel multi-step neural network architecture for training the model in this work. The details of the neural network structure will be discussed in Section 3.

3. Adams-Bashforth (AB) residual network.

3.1. Artificial neural network.

The drastic improvements in computing power make deep neural networks become state-of-the-art technology for a wide range of practical data-driven tasks such as image classification, face recognition, natural language processing, system prediction etc. see, e.g., [20, 21, 24, 25, 26, 42, 43]. A neural network effectively implements a mapping approximating a function that is learned based on a given set of input-output pairs, typically through the back-propagation algorithm. The basic structure of the simple artificial neural network consists of an input layer, one or more hidden layers, and a final layer of output. Each of these layers has an associated transfer function, and each unit cell (neuron) has an associated bias. Any input to the neuron has a bias added to it followed by activation through the transfer function. Fig. 2 shows a typical single hidden layer network architecture, where lines connecting neurons are also shown. To describe

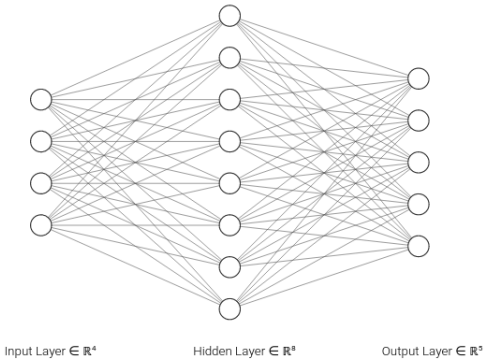

FIGURE 2. Single hidden layer neural network structure

this process using equation, we have for the output, h_i, of neuron i in the hidden layer is

$$h_i = \sigma(\sum_{j=1}^{N} w_{ij}x_j + b_i), \tag{17}$$

where $\sigma(\cdot)$ is called transfer (activation) function, N is the number of neurons in the layer, w_{ij} are the weights, x_j are the inputs from the previous layer, and b_i is the bias term. The weights and biases will be updated by solving the optimization process with backpropagation algorithm to compute the gradient. In theory, any differentiable function can qualify as an activation function, however, only a small number of functions which are bounded, monotonically increasing, and differentiable are used for this purpose [45].

It has been shown that a single hidden layer neural network can approximate any computable function. Numbers given to the input neurons are independent variables and those returned from the output neurons are dependent variables to the function being approximated by the neural network. The powerful computing resources can afford us to train a very deep neural network, which has been successfully applied in many supervised learning applications. Supervised learning is the task of learning the correspondence between input data X and output data Y from a training set of input-output pairs(x_i, y_i). There are two categories for supervised

learning: regression problems, for which the outputs take continuous values, and classification problems, consisting in the prediction of categorical labels. The neural network adopted in this paper for solving the analytic continuation problem is a regression task.

3.2. **ODE representation of neural network.** In this section, we describe recent mathematical representation of deep Residual Neural Network (ResNet). For a comprehensive introduction see, e.g., [10, 18, 32]. We outline the most important part of deep ResNet which is the forward propagation. For notational convenience, we stack the training features and target row-wise into matrices $\boldsymbol{X}_0 = [G^1, G^2, ..., G^s]^T \in \mathcal{R}^{s\times n}$ and $\boldsymbol{A} = [A^1, A^2, ..., A^s] \in \mathcal{R}^{s\in N}$. We consider a simplified version of ResNet model that has been successful in classifying images. The input values of forward propagation in the ResNet is given by

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \sigma(\boldsymbol{X}_t\boldsymbol{W}_t + b_t) \;\; t = 0, ..., N-1, \tag{18}$$

where N is the number of layers in the network architectures, $X_0 \in \mathcal{R}^{s\times n}$ is the initial input value. This propagation is parametrized by the nonlinear activation function $\sigma : \mathcal{R}^{s\times n} \to \mathcal{R}^{s\times n}$ and affine transformations represented by their weights, $\boldsymbol{W}_0, ..., \boldsymbol{W}_{N-1} \in \mathcal{R}^{n\times n}$, and biases $b_0, ..., b_{N-1} \in R^{1\times n}$. Fig. 3 shows the structure

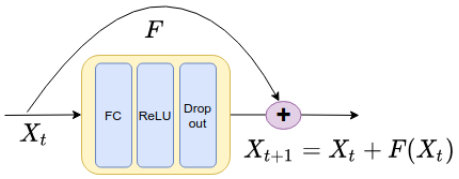

FIGURE 3. Residual neural network block

of a single ResNet block consists of fully connected layer, activation layer, and a dropout layer to prevent overfitting. The values $\boldsymbol{X}_t$ are called hidden layers and $\boldsymbol{X}_N$ is the final output layer. The activation function is applied element-wise and is typically smooth and non-decreasing. Two commonly used examples are hyperbolic tangent (tanh) and the Rectified Linear Unit (ReLU) activations. For simplicity, we only consider the ReLU activation in our model, i.e.

$$\sigma_{ReLU} = max(0, \boldsymbol{X}). \tag{19}$$

The final output layer predicts the values using the hypothesis function $h(\boldsymbol{X})$. For our problem, we assume the spectral function (output of the network) satisfies multinomial distributions so that we can use the softmax function in the output layer,

$$h(\boldsymbol{X}) = \frac{exp(\boldsymbol{X})}{exp(\boldsymbol{X})\boldsymbol{e}_m}, \tag{20}$$

where $\boldsymbol{e}_m \in \mathcal{R}^m$ denotes the $m-$dimensional vector of all ones.

The learning problem is to estimate the parameters of the forward propagation so that the deep ResNet can accurately approximate the training data set. This learning process can be solved by the following optimization problem

$$\min L(\tilde{\boldsymbol{A}}, \boldsymbol{A}) + \lambda R(\boldsymbol{W}, b), \tag{21}$$

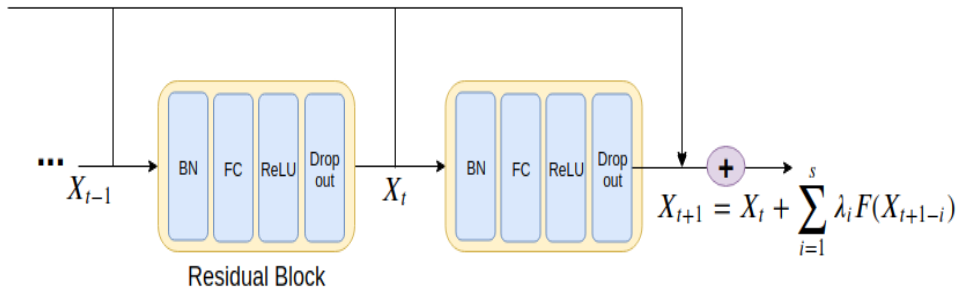

FIGURE 4. Multistep neural network architecture

where the loss function $L(\tilde{\boldsymbol{A}}, \boldsymbol{A}) = 1/2\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2$ is the sum of squared differences and the convex regularizer R penalizes undesirable parameters and can prevent overfitting.

3.3. **Adams-Bashforth scheme.** Much recent work has motivated us to view the ResNet as a dynamicals system [32, 18]. A significant piece of work on connecting numerical ordinary differential equations (ODEs) and deep neural networks [10] adopts the dynamical systems point-of-view and analyzes the lesioning properties of ResNet both theoretically and experimentally. The effort [11] introduced the ODE-net that can interpret and solve the ResNet using ODE solver, which provides memory efficiency for deep ResNet. In this article, motivated by the previous work, we propose a new Adams-Bashforth ResNet architecture for the analytic continuation problem.

The forward propagation of (18) can be considered as the forward Euler discretization of the initial value ODE given by

$$\dot{\boldsymbol{X}}(t) = \boldsymbol{F}(\boldsymbol{X}(t), \boldsymbol{W}(t), b(t)), \ \boldsymbol{X}(0) = \boldsymbol{X}_0, \ 0 \le t \le T, \tag{22}$$

where time t corresponds to the direction from input to output, $\boldsymbol{X}(0)$ is the initial input feature, and $\boldsymbol{X}(T)$ is the output of the network. Thus, the problem of learning the network parameters, $\boldsymbol{W}$ and b, is equivalent to solving a parameter estimation problem or optimal control problem involving the ODE in (22). Note that the time step size Δt in the fully discretized ODE $\frac{\boldsymbol{X}_{t+1} - \boldsymbol{X}_t}{\Delta t} = \boldsymbol{F}(\boldsymbol{X}_t)$, is implicitly absorbed by the residual module in the original formulation of ResNet (18). Instead, we intend to use a multistep Adams-Bashforth (AB) method to discretize (22). As mentioned before, the standard ResNet can be considered as the forward Euler discretization, whereas multistep AB method has higher accuracy in numerical methods of ODE [3]. The fully discretized schemd is shown in Fig. 4 and

$$\boldsymbol{X}_{t+s} = \boldsymbol{X}_{t+s-1} + \Delta t \sum_{i=1}^{s} \lambda_i \boldsymbol{F}(\boldsymbol{X}_{t+s-i}), \tag{23}$$

where $\sum_{i=1}^{s} \lambda_i = 1$. The formula can be derived from Taylor's theorem. As an example, we use two-step method (AB2) to illustrate, i.e.,

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \boldsymbol{X}(t) + \Delta t((1-\lambda)\dot{\boldsymbol{X}}(t) + \lambda(\dot{\boldsymbol{X}}(t) \\ &- \Delta t\ddot{\boldsymbol{X}}(t) + \mathcal{O}(\Delta t^2))) \\ &= \boldsymbol{X}(t) + \Delta t\dot{\boldsymbol{X}}(t) - \lambda\Delta t^2\ddot{\boldsymbol{X}}(t) + \mathcal{O}(\Delta t^3). \end{aligned} \tag{24}$$

Then applying Taylor expansion tos the true solution, i.e.s

$$\boldsymbol{X}(t+1) = \boldsymbol{X}(t) + \Delta t \dot{\boldsymbol{X}}(t) + \frac{1}{2}\Delta t^2 \ddot{\boldsymbol{X}}(t) + \mathcal{O}(\Delta t^3), \tag{25}$$

we obtain the numerical schemes associated to AB2 and AB3 as following

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \frac{3}{2}\boldsymbol{F}(\boldsymbol{X}_t, \boldsymbol{W}_t, b_t) - \frac{1}{2}\boldsymbol{F}(\boldsymbol{X}_{t-1}, \boldsymbol{W}_{t-1}, b_{t-1}), \tag{26}$$

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \boldsymbol{X}_t + \frac{23}{12}\boldsymbol{F}(\boldsymbol{X}_t, \boldsymbol{W}_t, b_t) - \frac{4}{3}\boldsymbol{F}(\boldsymbol{X}_{t-1}, \boldsymbol{W}_{t-1}, b_{t-1}) \\ &+ \frac{5}{12}\boldsymbol{F}(\boldsymbol{X}_{t-2}, \boldsymbol{W}_{t-2}, b_{t-2}). \end{aligned} \tag{27}$$

The AB2 method has second order $\mathcal{O}(\Delta t^2)$ accuracy. Standard ResNet is considered a AB1 method which has first order $\mathcal{O}(\Delta t)$ accuracy. According to the stability analysis of linear multistep explicit methods, the AB3 method is strongly stable while AB2 and AB1 is conditionally stable. This stability property drives us to apply the AB method to obtain a more robust deep network architectures that can provide a model with better performance for noisy data. The family of linear multistep method is large. To shorten the discussion in this work, we focus on the AB2 and AB3 method in our numerical tests.

4. Numerical results.

4.1. Dataset.

In this section, we present the numerical results from our Adams-Bashforth ResNet model (AB-ResNet). The training data can be collected from experimental measurements or simulated according to a theoretical model. In this work, we choose to simulate spectral density functions that always have a quasi-particle peak close to $\omega = 0$, as often encountered when considering correlated metals. In the data generation, the spectral densities $A(\omega)$ are defined as a sum of uncorrelated Gaussian distributions:

$$A^i(\omega) = \frac{1}{R_i}\sum_{k=0}^{R_i} exp\left(-\frac{(\omega-\mu_k)^2}{2\sigma_k^2}\right), \tag{28}$$

where the frequencies $\omega \in [-10, 10]$, the centers of the peaks $\mu_k \in [-5, 5]$, the number of Gaussian distributions $R_i \in [1, ..., 21]$, and $\sigma_k \in [0.1, 1]$. Parameters R_i, μ_k, σ_k are uniformly sampled over the above-mentioned ranges. The Green's functions are then computed by Eq. (4). The discretization of the Green's function is generally over $\mathcal{O}(10^3)$. The amount of data necessary to approximate a function grows exponential with the number of dimensions. To reduce the effect of the curse of dimensionality, we use the orthogonal Legendre polynomials to represent the Green's function data which can facilitate the learning process of the model. The compact representation is given by

$$G(\tau) = \sum(2l+1)G_l P_l(2\frac{\tau}{\beta} - 1), \tag{29}$$

where P_l are the legendre polynomials. In the experiments, 64 basis are used to ensure the accurate approximation of the data, with similar strategies found in [15]. Three noise levels $\epsilon = 10^{-5}, 10^{-3}, 10^{-2}$ are added to the dataset, such that, $G^{train} = G + \epsilon$.

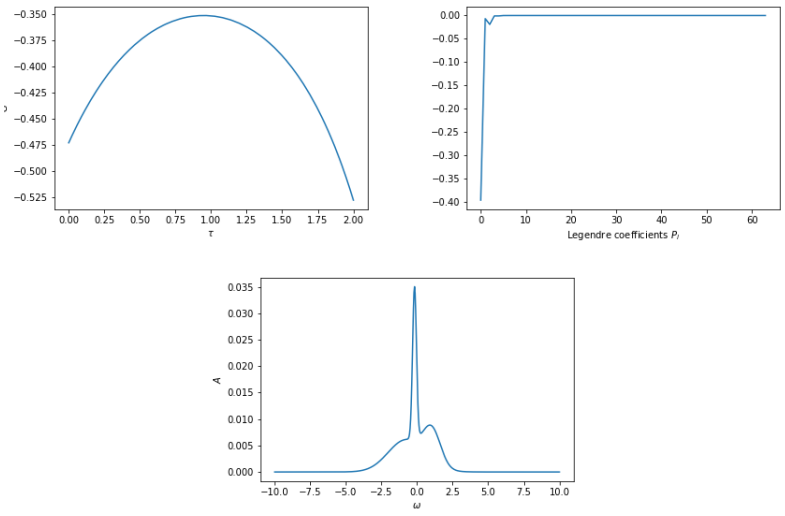

FIGURE 5. One data sample from the training set $G(\tau)$ (top left), Legendre representation G_l (top right), and target spectral density $A(\omega)$ (bottom)

4.2. **Numerical test.** The network architecture of our model, shown in Fig. 4, consists of an input layer connected to a residual block, followed by eight repetitions of the residual block. For each residual block, the first layer is a batch normalization followed by a fully collected dense layer with ReLU activation. Then it's followed by a dropout layer that helps to avoid the overfitting issue by randomly dropping units. The output is computed using a softmax layer, which ensure the similarity to a probability density function. The training is performed on a dataset of size 100,000 with validation and test sets, both of size 1000, used in our numerical experiment. The code implementation is based on PyTorch where the Adams optimizer and the KLD loss function were used.

We have investigated AB-ResNet approaches to improving the robustness of our model against noisy data. As mentioned before, the AB1 (ResNet) and AB2 method is conditionally stable whereas AB3 is strongly stable. In order to study the stability of the network architecture numerically, we trained each model on the dataset with different magnitude of noisy, i.e., 10^{-5}, 10^{-3}, and 10^{-2}. Fig. 6 shows the training performance from each network and, as expected, the AB3 network has a better learning behavior than AB1 and AB2.

Fig. 7 provides a qualitative comparison of the results of our AB-ResNet method and the MaxEnt method where we plot three samples from test set for illustration purposes. In these examples, both methods predict $A(\omega)$ accurately for the lowest level of noise. However, at noise $\epsilon = 10^{-2}$, MaxEnt is not able to recover the peaks in the predicted spectral function. While in the case of AB-ResNet, our model is able to correctly identify most peaks. Hence, it clearly shows that our AB-ResNet model generates better results compared to the classical MaxEnt. Fig. 8 shows the comparison of the prediction between each AB network model from three different samples. The average mean absolute error on the test dataset are $6.8e-4, 3.8e-4, 2.6e-4$ for AB1, AB2 and AB3, respectively. This is consistent with the

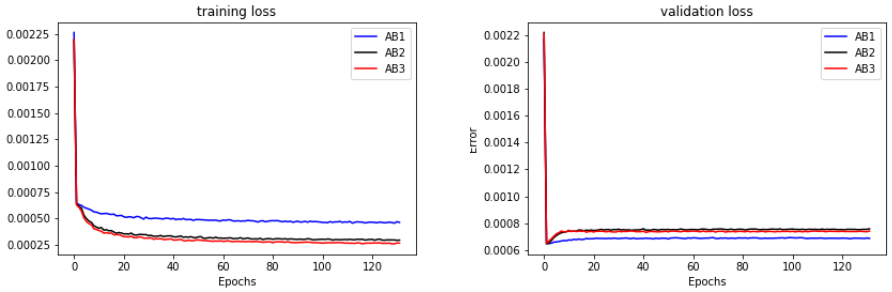

FIGURE 6. The training performance from AB1-ResNet, AB2-ResNet, and AB3-ResNet structure with data noise 10^{-2}

numerical ODE analysis. That is, higher order methods provide higher accuracy results. Then, we studied the computational efficiency of our model compared to MaxEnt. AB-ResNet model allows a direct mapping between Green's function and the spectral densities. In contrast, the MaxEnt method is an iterative method which requires generating trail functions until convergence is reached. For the computation cost, the CPU time for AB-ResNet model is $\mathcal{O}(10)$ second while for MaxEnt is $\mathcal{O}(10^3)$ second. So, the new model is computationally more efficient compared to the MaxEnt method.

5. **Conclusions.** In summary, we have developed the AB-ResNet that solves the kernel inversion with noisy data for the analytic continuation problem. The numerical experiments show that our AB-ResNet model can recover the spectral function with an accuracy similar to that of the commonly used MaxEnt approach under low level of noises. The new model gives much better results than MaxEnt under high level of noises at a fraction of its computational cost. Adding more training data and using larger step network architecture could further improve the model performance. Other inverse problem can apply our model the same way given the great representative capacity of deep AB-ResNet.

Some future work should consider the limitations of the proposed model. One main drawback of the method is that the model is learned for a particular inverse temperature, i.e., $\beta = 2$, whereas the MaxEnt method can provide it as a parameter. So, the MaxEnt method has the generality with respect to different β. To extend our model to arbitrary values of β, we can train a separate network for each parameter. Another approach for this issue would be to add β as an input parameter to the model and train it on a large collection of dataset. These approaches can improve the robustness of our model with respect to the inverse temperature. Another direction of improving the model would be using stochastic neural networks to address the generalization issue of deterministic neural networks [8].

Acknowledgments. This material is based upon work supported in by: the Scientific Discovery through Advanced Computing (SciDAC) program, U.S. Department of Energy, Basic Energy Sciences, Division of Materials Sciences and Engineering; the U.S. Department of Energy, Office of Science, Early Career Research Program under award number ERKJ314; U.S. Department of Energy, Office of Advanced Scientific Computing Research under award numbers ERKJ331 and ERKJ345; the National Science Foundation, Division of Mathematical Sciences, Computational

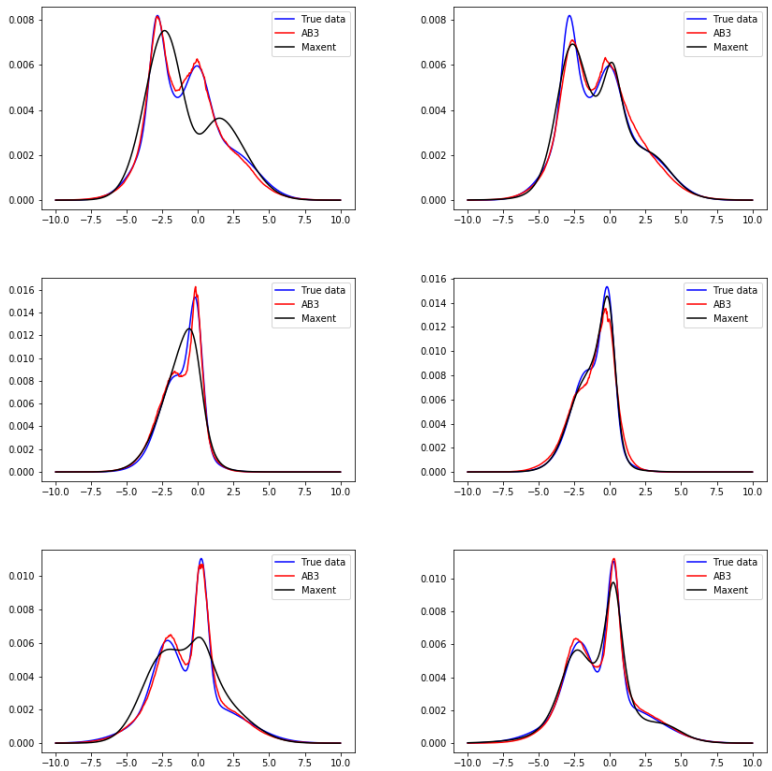

FIGURE 7. Three different spectral density function $A(\omega)$ generated from AB3-ResNet and Maxent (dark line). The left column represents results from dataset with noise level 10^{-2}, the right column shows results obtained from the dataset under noise level 10^{-3}

Mathematics program under contract number DMS1620280 and the contract number DMS-1720222; and by the Laboratory Directed Research and Development program at the Oak Ridge National Laboratory, which is operated by UT-Battelle, LLC., for the U.S. Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

[1] L.-F. Arsenault, R. Neuberg, L. A. Hannah and A. J. Millis, Projected regression methods for inverting fredholm integrals: Formalism and application to analytical continuation, *arXiv preprint* arXiv:1612.04895, 2016.

[2] L.-F. Arsenault, R. Neuberg, L. A. Hannah and A. J. Millis, Projected regression method for solving fredholm integral equations arising in the analytic continuation problem of quantum physics, *Inverse Problems*, **33** (2017), 115007.

[3] U. M. Ascher and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, volume **61**, SIAM, Philadelphia, PA, 1998.

[4] F. Bao, Y. Tang, M. Summers, G. Zhang, C. Webster, V. Scarola and T. A. Maier, Fast and efficient stochastic optimization for analytic continuation, *Physical Review B*, **94** (2016), 125149.

[5] K. S. D. Beach, Identifying the maximum entropy method as a special limit of stochastic analytic continuation, *arXiv preprint* arXiv:cond-mat/0403055, 2004.

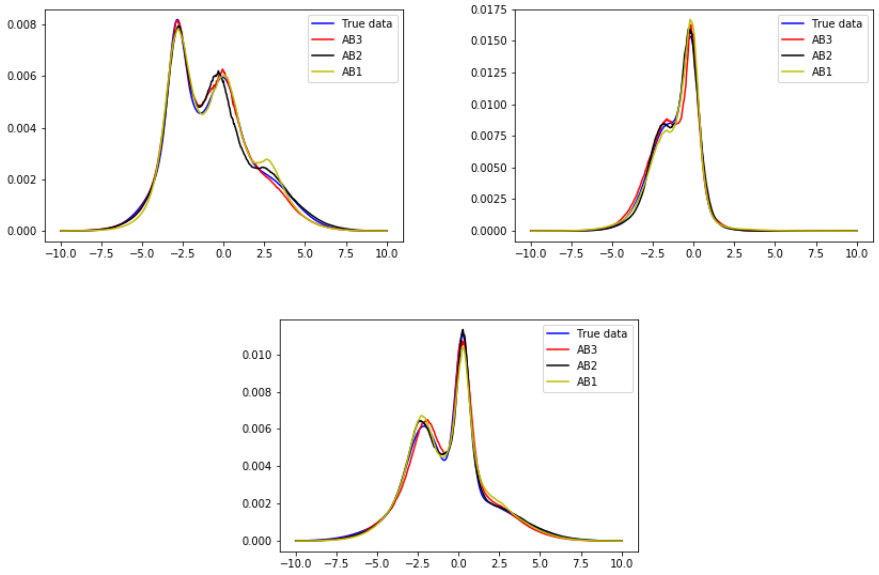

FIGURE 8. The comparison of predicted spectral function between different AB-ResNet

[6] C. Beck, E. Weinan and A. Jentzen, Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations, *Journal of Nonlinear Science*, **29** (2019), 1563–1619.
[7] G. Bertaina, D. E. Galli and E. Vitali, Statistical and computational intelligence approach to analytic continuation in quantum monte carlo, *Advances in Physics: X*, **2** (2017), 302–323.
[8] Y. Cao, H. Zhang, R. Archibald and F. Bao, A backward sde method for uncertainty quantification in deep learning, *arXiv preprint arXiv:2011.14145*, 2021.
[9] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert and E. Holtham, Reversible architectures for arbitrarily deep residual neural networks, In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[10] B. Chang, L. Meng, E. Haber, F. Tung and D. Begert, Multi-level residual networks from dynamical systems view, In *International Conference on Learning Representations*, 2018.
[11] T. Chen, Y. Rubanova, J. Bettencourt and D. K. Duvenaud, Neural ordinary differential equations, In *Advances in Neural Information Processing Systems*, 2018, 6571–6583.
[12] K. Dahm and A. Keller, Learning light transport the reinforced way, In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, **241**, Springer, 2018, 181–195.
[13] F. Bao and T. Maier, Stochastic gradient descent algorithm for stochastic optimization in solving analytic continuation problems, *Foundations of Data Science*, **2** (2020), 1–17.
[14] W. E and Q. Wang, Exponential convergence of the deep neural network approximation for analytic functions, *Sci. China Math.*, **61** (2018), 1733–1740. *arXiv preprint arXiv:1807.00297*, 2018.
[15] R. Fournier, L. Wang, O. V. Yazyev and Q. Wu, Artificial neural network approach to the analytic continuation problem, *arXiv preprint arXiv:1810.00913*, 2018. *Phys. Rev. Lett.*, **124** (2020), 056401, 6 pp.
[16] S. Fuchs, T. Pruschke and M. Jarrell, Analytic continuation of quantum Monte Carlo data by stochastic analytical inference, *Physical Review E*, **81** (2010), 056701.
[17] S. F. Gull and J. Skilling, Maximum entropy method in image processing, *IEE Proceedings F (Communications, Radar and Signal Processing)*, **131** (1984), 646–659.
[18] E. Haber and L. Ruthotto, Stable architectures for deep neural networks, *Inverse Problems*, **34** (2017), 014004.

[19] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–778.
[20] R. Hecht-Nielsen, Theory of the backpropagation neural network, In *Neural Networks for Perception*, Elsevier, 1992, 65–93.
[21] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, **29** (2012), 82–97.
[22] M. Jarrell and J. E. Gubernatis, Bayesian inference and the analytic continuation of imaginary-time quantum Monte Carlo data, *Physics Reports*, **269** (1996), 133–195.
[23] K. H. Jin, M. T. McCann, E. Froustey and M. Unser, Deep convolutional neural network for inverse problems in imaging, *IEEE Transactions on Image Processing*, **26** (2017), 4509–4522.
[24] A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, **25** (2012), 1097–1105.
[25] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444.
[26] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, et al, Comparison of learning algorithms for handwritten digit recognition, In *International Conference on Artificial Neural Networks*, volume **60**. Perth, Australia, 1995, 53–60.
[27] R. Levy, J. P. F. LeBlanc and E. Gull, Implementation of the maximum entropy method for analytic continuation, *Computer Physics Communications*, **215** (2017), 149–155.
[28] H. Li, J. Schwab, S. Antholzer and M. Haltmeier, Nett: Solving inverse problems with deep neural networks, *Inverse Problems*, **36** (2020), 065005.
[29] Q. Li, C. Tai and W. E, Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations, *Journal of Machine Learning Research*, **20** (2019), Paper No. 40, 47 pp.
[30] H. Lin and S. Jegelka, Resnet with one-neuron hidden layers is a universal approximator, In *Advances in Neural Information Processing Systems*, 2018, 6169–6178.
[31] Z. Long, Y, Lu, X. Ma and B. Dong, PDE-net: Learning PDEs from data, In *Proceedings of the 35th International Conference on Machine Learning*, 2018, 3208–3216.
[32] Y. Lu, A. Zhong, Q. Li and B. Dong, Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations, In *Proceedings of the 35th International Conference on Machine Learning*, 2018, 3282–3291.
[33] C. Ma, J. Wang and Weinan E, Model reduction with memory and the machine learning of dynamical systems, *arXiv preprint* `arXiv:1808.04258`, 2018. *Commun. Comput. Phys.*, **25** (2019), 947–962.
[34] A. S. Mishchenko, N. V. Prokof'ev, A. Sakamoto and B. V. Svistunov, Diagrammatic quantum Monte Carlo study of the Fröhlich polaron, *Physical Review B*, **62** (2000), 6317–6336.
[35] J. Otsuki, M. Ohzeki, H. Shinaoka and K. Yoshimi, Sparse modeling approach to analytical continuation of imaginary-time quantum monte carlo data, *Physical Review E*, **95** (2017), 061302.
[36] E. Pavarini, E. Koch, F. Anders and M. Jarrell, Correlated electrons: From models to materials, *Reihe Modeling and Simulation*, **2** (2012).
[37] N. V. Prokof'ev and B. V. Svistunov, Spectral analysis by the method of consistent constraints, *JETP Lett.*, **97** (2013), 649–653.
[38] A. W. Sandvik, Stochastic method for analytic continuation of quantum Monte Carlo data, *Physical Review B*, **57** (1998), 10287–10290.
[39] R. N. Silver, J. E. Gubernatis, D. S. Sivia and M. Jarrell, Spectral densities of the symmetric Anderson model, *Physical Review Letters*, **65** (1990), 496–499.
[40] B. Wang, X. Luo, Z. Li, W. Zhu, Z. Shi and S. Osher, Deep neural nets with interpolating function as output activation, In *Advances in Neural Information Processing Systems*, 2018, 743–753.
[41] L. Wu, C. Ma and W. E, How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective, In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 2018, 8289–8298.
[42] X. Xie, C. Webster and T. Iliescu, Closure learning for nonlinear model reduction using deep residual neural network, *Fluids*, **5** (2020), 39.

[43] X. Xie, G. Zhang and C. G. Webster, Non-intrusive inference reduced order model for fluids using deep multistep neural network, *Mathematics*, **7** (2019), 757.
[44] H. Yoon, J.-H. Sim and M. J. Han, Analytic continuation via domain knowledge free machine learning, *Physical Review B*, **98** (2018), 245101.
[45] G. Zhang, B. Eddy Patuwo and M. Y. Hu, Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting*, **14** (1998), 35–62.

Received February 2021; revised April 2021, early access August 2021.

E-mail address: xxie@nyu.edu
E-mail address: fbao@fsu.edu
E-mail address: maierta@ornl.gov
E-mail address: cwebst13@utk.edu