Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **2**, Number **1**, March **2020**
doi:10.3934/fods.2020001
pp. **1–17**

STOCHASTIC GRADIENT DESCENT ALGORITHM FOR STOCHASTIC OPTIMIZATION IN SOLVING ANALYTIC CONTINUATION PROBLEMS

Feng Bao*

Department of Mathematics, Florida State University
Tallahassee, Florida, USA

Thomas Maier

Center for Nanophase Materials Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Abstract. We propose a stochastic gradient descent based optimization algorithm to solve the analytic continuation problem in which we extract real frequency spectra from imaginary time Quantum Monte Carlo data. The procedure of analytic continuation is an ill-posed inverse problem which is usually solved by regularized optimization methods, such like the Maximum Entropy method, or stochastic optimization methods. The main contribution of this work is to improve the performance of stochastic optimization approaches by introducing a supervised stochastic gradient descent algorithm to solve a flipped inverse system which processes the random solutions obtained by a type of Fast and Efficient Stochastic Optimization Method.

1. **Introduction.**

In this work, we introduce a stochastic gradient descent algorithm to solve a type of analytic continuation problems in physics. The goal of analytic continuation that we are interested is to transfer theoretical Quantum Monte Carlo simulations to spectral functions that reflect physical properties of quantum materials. Quantum Monte Carlo (QMC) is a class of numerical methods that simulate exact description for interacting quantum many-particle systems such like spin models or strongly correlated electron systems. It is generally formulated on the imaginary time axis to treat the finite temperature dynamics. To derive real time dynamics that can be compared with physical experiments, the process of analytic continuation is required to obtain a real frequency spectrum. The challenge of the analytic continuation is that the process to extract the real frequency spectrum is an ill-posed inverse problem.

One of the most widely used approaches to address the challenge of ill-posedness in solving data based inverse problems is to adopt the Bayesian inference framework. In numerical methods for solving analytic continuation problems, the most well developed state-of-the-art tool is the Maximum Entropy (MaxEnt) method [5, 6, 10, 15]. The general concept of MaxEnt is to introduce a regularization term as the prior information in the form of entropy that measures the deviation from

2010 *Mathematics Subject Classification.* Primary: 49N45; Secondary: 49M37.

Key words and phrases. Stochastic optimization, stochastic gradient descent, analytic continuation, physics supervision.

The first author is supported by NSF grant DMS-1720222.

* Corresponding author: Feng Bao.

a given default spectrum, and then solve the Bayesian inverse problem through deterministic optimization procedures. Although the MaxEnt method could provide fairly good representations for the target real time frequency spectra, the main drawback of MaxEnt as a regularized optimization method is that the entropy-like regularization might over smooth useful features in the desired spectrum, and the performance of MaxEnt heavily depends on the prior knowledge for the spectrum which is usually not available in practice. Another Bayesian type method to solve analytic continuation problems is the stochastic optimization method (SOM), which randomly samples large amount of possible spectral functions as random solutions [2, 8, 11]. Each of these random solutions optimizes the deviation from the QMC data. Combining all the possible solutions together, SOM provides an implicit regularization mechanism for the analytic continuation problem. The major advantage of SOM compared with MaxEnt is that it considers multiple possible spectra that fit the QMC data, which would possibly lead to better estimation for the solution. In a recent work, the authors introduced a fast and efficient version of SOM and named it as *Fast and Efficient Stochastic Optimization Method* (FESOM) [1]. Instead of complex parameterization for the spectral function applied in standard SOM, the optimization procedure in FESOM uses usual discretization on the real frequency axis as a parameterization, which is easy to design and more flexible to implement.

However, although FESOM could describe the deviation of different sample spectra from the QMC data in a more effective way compared to SOM, it only utilizes the Monte Carlo average to combine all the random solutions as the estimate for the desired spectrum – just like SOM. Despite of its rigorous perform as a regularization procedure, the Monte Carlo method averages out the deviation information among optimized samples and it does not take sufficient consideration for the variance of sample spectra. In this work, we introduce a stochastic gradient descent (SGD) algorithm that focus on the exploration of sample spectra obtained in FESOM. The main novelty of the proposed SGD algorithm is that we flip the role of QMC data and the optimized sample spectra in stochastic optimization. Specifically, in this approach we consider the random solutions that we obtain in the analytic continuation problem as the source of data and use the original QMC data as the model to derive a synthetic inverse problem. Then, we use SGD to solve the derived inverse problem and the resulting solution can be applied back to the QMC data to get an estimate for the target spectral function. Since the underline framework formulates a synthetic problem which does not naturally follow the physical background of the analytic continuation in this work, during the SGD procedure we enforce physical supervisions to guide the optimization.

The rest of this paper is organized as follows. In Section 2, we give a problem statement for analytic continuation and briefly discuss the stochastic optimization approach. In Section 3, we introduce our SGD algorithm for analytic continuation problem. We carry out three numerical examples to validate the performance of the SGD algorithm in Section 4 and include concluding remarks in Section 5.

2. Analytic continuation and stochastic optimization.

In the analytic continuation problem, we consider the following integral equation

$$G(i\omega_n) = \int d\omega K(i\omega_n, \omega) A(\omega), \tag{1}$$

where $K(i\omega_n, \omega)$ is a kernel function defined as

$$K(i\omega_n,\omega)=\frac{1}{i\omega_n-\omega}. \tag{2}$$

The function $A(\omega)$ with respect to the real frequency ω in (1) is the quantity of interest in the analytic continuation problem, which represents a spectral function in physics. The behavior of the spectrum A usually gives physical properties of some materials. The function $G(i\omega_n)$ in (1) is the observational data, such like the single-particle Green's function calculated by the quantum Monte Carlo method (QMC), at discrete Matsubara frequencies ω_n on the imaginary axis.

For a partition on the real frequency axis, denoted by

$$\Pi=\{\omega_l|a=\omega_0\le\omega_1\le\omega_2\le\cdots\le\omega_l\le\cdots\le\omega_L=b\},$$

we consider the following discretized form of equation (1)

$$G_n=\sum_{l=1}^{L}K_{nl}A_l, \tag{3}$$

where $G_n=G(i\omega_n)$, $A_l=A(\omega_l)$, $K_{nl}=\frac{\Delta\omega_l}{i\omega_n-\omega}$, and $\Delta\omega_l$ is the frequency partition step-size. The major challenge of inverting equation (3) from the QMC data $G:=\{G_n\}$ and calculate $A:=\{A_l\}$ is that the kernel matrix $K:=\{K_{nl}\}$ is ill-conditioned, and small perturbations on the equivalence system would cause totally different results for the spectral function A. In this case, there are infinitely many possible solutions for the analytic continuation problem due to the fact that the QMC data G is always noise perturbed.

One of the most important approaches to address the aforementioned ill-posedness for the analytic continuation problem is the Bayesian statistical inferences. The Bayes' formula, given by the following

$$P(A|G)=\frac{P(G|A)P(A)}{C}, \tag{4}$$

provides a mathematical model to combine the prior information for the target spectrum A with the QMC data G and it describes the quantity of interest as a conditional probability density function (pdf). In (4), the probability distribution $P(A|G)$, which is also called the posterior pdf, gives the conditional distribution of the spectral function A given G; $P(G|A)$ is the likelihood function that measures the discrepancy between the data G and our choice of A; $P(A)$ is the prior distribution for A followed by the prior knowledge about the spectrum, which is very limited in the analytic continuation problem; and C is some normalization factor. There are several successful approaches to solve the analytic continuation problem through the Bayesian formula (4) [1, 2, 8, 11]. In this work, we shall introduce a supervised stochastic gradient descent method to improve the performance of the Fast and Efficient Stochastic Optimization Method (FESOM) introduced in [1].

In what follows, we briefly discuss the framework of FESOM that solves the analytic continuation problem (3). The central idea of FESOM is to use random walk to construct stochastic realizations of the target spectral function, denoted by $\{\tilde{A}^r\}_{r=1}^{R}$, that minimizes the χ^2 errors between $K\tilde{A}^r$ and the data G, i.e. $\chi^2[\tilde{A}^r]:=\frac{(K\tilde{A}^r-G)^2}{\sigma^2}$, where σ is the standard deviation of the QMC data. Specifically, we run a stochastic optimization procedure for each realization $\tilde{A}^r$. To this end, we choose some initial guess for the spectrum as $D(\omega_l)$, which is typically chosen as a Gaussian function in the absence of prior information for A, and set

$$\tilde{A}^r_0(\omega_l)=D(\omega_l),\quad l=0,1,\cdots,L.$$

Assume that we have the sample spectrum $\tilde{A}_i^r$ at an iteration step i, we add a Gaussian process λ_i^r to $\tilde{A}_i^r$ and get a proposed spectral function

$$\tilde{A}_{i+\frac{1}{2}}^r := \frac{1}{I}(\tilde{A}_i^r + \lambda_i^r),$$

where I is a normalization factor such that $\sum_{l=1}^{L} \tilde{A}_{i+\frac{1}{2}}^r(\omega_l)\Delta\omega_l = 1$ to meet the basic physical property for $\tilde{A}^r$ as a spectral function. If the proposed spectrum $\tilde{A}_{i+\frac{1}{2}}^r$ fits the data better, i.e. $\chi^2[\tilde{A}_{i+\frac{1}{2}}^r] < \chi^2[\tilde{A}_i^r]$, we accept the proposed sample and let $\tilde{A}_{i+1}^r = \tilde{A}_{i+\frac{1}{2}}^r$; otherwise we drop the proposed sample and let $\tilde{A}_{i+1}^r = \tilde{A}_i^r$. In this way, the χ^2 error decreases monotonically as the iteration number i increases and we stop the optimization procedure when the χ^2 reaches a threshold number ϵ at the j-th iteration step. Then, the r-th realization of the spectral function is chosen as $\tilde{A}^r := \tilde{A}_j^r$. In this way, the spectral sample set $\{\tilde{A}^r\}_{r=1}^R$ provides a representation for the likelihood distribution $P(G|A)$, and the posterior, i.e.$P(A|G)$, as desired in the Bayes' formula (4) can be derived by combining the likelihood with the prior distribution, i.e. $P(A)$, which is obtained from the knowledge of the spectrum. However, in most practical problems, the prior knowledge is minimal. Therefore, people typically choose $P(A)$ to be a uniform distribution, which leads to $P(A|G) \propto P(G|A)$. As a result, the random samples $\{\tilde{A}^r\}_{r=1}^R$ that describe the likelihood function $P(G|A)$ also provide the desired conditional pdf $P(A|G)$. It's worthy to mention that due to the nature of the stochastic optimization, each sample spectral function contains lots of random features with massive fluctuations in the spectral curve. In the FESOM approach, we use the average of $\tilde{A}^r$, i.e. $\bar{A} := \frac{1}{R}\sum_{r=1}^{R} \tilde{A}^r$, to be our estimated spectral function. This would also be considered as a regularization procedure through Monte Carlo average.

In the stochastic gradient descent method that we shall introduce in this work, we use optimized sample spectra obtained in FESOM as our "data" and then try to learn from the FESOM data to find a model that could better approximate the target spectral function A. From our previous study, we know that the FESOM could give accurate approximations for the spectral function when the spectral curve is smooth, just like the Maximum Entropy method (MaxEnt). At the same time, compared to MaxEnt, the FESOM could capture some features that are not easy to be captured by MaxEnt. In addition to the approximate spectral function, the FESOM also produces an approximation for the conditional distribution $P(A|G)$ by using empirical samples $\{\tilde{A}^r\}_{r=1}^R$. This would give us a confidence band surrounding the estimate $\bar{A}$. We want to point out that in the current version of FESOM, *the simple regularization procedure by averaging all the samples in $\{\tilde{A}^r\}_{r=1}^R$ did not consider the covariance information hidden in the empirical distribution of samples.* Therefore, the Monte Carlo type regularization procedure smoothes out many important features and ignored the information contained in the confidence band. At the same time, it's necessary to mention that each sample $\tilde{A}^r$ fits the data G well and the sample set $\{\tilde{A}^r\}_{r=1}^R$ provides a large pool of possible features for the spectral function A.

In the following section, we introduce our stochastic gradient descent (**SGD**) based algorithm that allows us to consider the covariance information contained in the FESOM spectral samples and construct better approximations for the spectral function A.

3. Stochastic gradient descent for analytic continuation.

We first recall the analytic continuation problem (3) and write it in the matrix form for the convenience of presentation

$$G = KA, \tag{5}$$

where the spectral function A is the quantity of interest, G is the QMC data and K is the kernel. A straight forward approach to get A is to compute the K inverse, i.e. K^{-1}, and multiply it to the data to get $A = K^{-1}G$. However, since the kernel K defined in (2) is ill-conditioned, it's not feasible to compute K^{-1} by inverting K. Even if we can derive very accurate approximation method to calculate K^{-1} from K, the terms that we truncate as approximation errors would have significant influence to A based on the data we receive. Therefore, directly approximating K^{-1} and simply multiplying it with the data G could not give us good estimation for A.

The major contribution of this work is that we develop a SGD based estimation method to calculate a data informed K^{-1}. In addition to the original QMC data in the analytic continuation framework, in this approach we consider the optimized spectral samples that we obtained from the FESOM as our data source and apply SGD as a machine learning type method to learn the K^{-1} from the sample spectral functions $\{\tilde{A}^r\}_{r=1}^R$. Then, the estimated spectral function A can be calculated by the product of the estimated K^{-1} learned from FESOM samples and the original QMC data G. The motivation of this SGD based algorithm is to explore the optimized spectral samples and learn more information contained in these samples which has been averaged out in the Monte Carlo regularization procedure in the FESOM.

Stochastic gradient descent method. To proceed, we first give a brief description for the SGD method that solves a data driven optimization problem. Let $\lambda \to F(\lambda, Z)$ be a random cost function, where λ is the optimization parameter performing as the quantity of interest, and Z is a random variable with uniform distribution representing the source of data. The goal of the optimization problem regarding the cost function F is to find $\min_\lambda E[F(\lambda, Z)]$. In practical applications, the random variable Z is represented by a set of data, denoted by $\mathcal{Z}$, and the original optimization problem becomes finding λ to satisfy the minimum of $\tilde{E}[F(\lambda, Z)] := \sum_{z\in\mathcal{Z}} F(\lambda, z)$, i.e. $\min_\lambda \sum_{z\in\mathcal{Z}} F(\lambda, z)$. The classic gradient descent method solves the optimization problem by the following iteration

$$\lambda_{i+1} = \lambda_i - \alpha \sum_{z\in\mathcal{Z}} \frac{\nabla F(\lambda_i, z)}{n}, i = 0, 1, 2, \dots, \tag{6}$$

where $\nabla\cdot$ is the gradient operator, λ_0 is the initial guess for the parameter, n is the size of $\mathcal{Z}$, and α is the step size moving forward to the gradient descent direction, which is also called the "learning rate" in machine learning. The above gradient descent approach uses the entire data set to estimate the evolution direction of the target optimization parameter. When the data set $\mathcal{Z}$ is large, it is very expensive to calculate $\sum_{z\in\mathcal{Z}} \frac{\nabla F(\lambda_i, z)}{n}$, which typically results the "big data" problem.

In the stochastic gradient descent method, instead of approximating the expectation with all the data $\{\nabla F(\lambda_i, z)\}_{z\in\mathcal{Z}}$, we use one random selection of $\nabla F(\lambda_i, z)$ to approximate $E[F(\lambda, Z)]$ and implement the gradient descent procedure. In this way, we update the parameter λ_i with the following scheme [7]

$$\lambda_{i+1} = \lambda_i - \alpha \nabla F(\lambda_i, z), \tag{7}$$

where $z \in \mathcal{Z}$ is a randomly selected sample in $\mathcal{Z}$, which means that $\nabla F(\lambda_i, z)$ explores the data in $\mathcal{Z}$ in a stochastic manner, and $\{\lambda_i\}_{i\geq 1}$ forms a stochastic process that approaches the optimal λ randomly. Although the SGD method was primarily developed to address the big data problem and save computational cost in calculating $\sum_{z\in\mathcal{Z}} \frac{\nabla F(\lambda_i,z)}{n}$, the random walking behavior of the stochastic dynamics (7) also provides a mechanism to get out of possible local minima in the cost function F that might trap the deterministic gradient descent dynamics implemented in (6), which is the main concern of the problem that we are interested in this work.

Apparently, the ill-conditioned kernel K in the analytic continuation (1) brings many local minima in the optimization problem, which also causes the ill-posedness in finding K^{-1}, and eventually in finding the spectrum A. These local minima make the deterministic gradient descent methods very difficult to explore a large parameter space due to the traps of local minima, and it's natural to apply the SGD method to address the ill-posedness problem in analytic continuation. Since in the analytic continuation, the SGD is primarily used to address the ill-posedness in the optimization procedure, we modify the classic SGD dynamics (7) by adding extra isotropic noises to get more flexibility to attack the local minima problem, and we introduce the following stochastic gradient Langevin dynamics (SGLD)

$$\lambda_{i+1} = \lambda_i - \alpha\nabla F(\lambda_i, z) + \beta\epsilon_i, \tag{8}$$

where $z \in \mathcal{Z}$ is also a sample in $\mathcal{Z}$, ϵ_i is the inflation random noise that gives the opportunities to let the above optimization procedure move out of the local minima, and β is the level of noise which is a user defined factor typically chosen as a constant proportional to the learning rate α. In this work, we shall apply the SGLD (8), instead of (7), to implement the SGD optimization. Extensive studies show that the above SGLD and its extensions could solve ill-posed optimization problem optimization problem well (see [12, 18] for example).

Implementation of SGD in analytic continuation. When solving the analytic continuation problem with SGD, we consider the original QMC data G as our model and use the optimized samples that we derive from FESOM as "data". Therefore, the sample spectra $\{\tilde{A}^r\}_{r=1}^R$ are the main source of information to guide us obtain the estimate for K^{-1}.

Specifically, we rewrite the original analytic continuation problem (5) and consider the following equation

$$K^{-1}G = A, \tag{9}$$

where K^{-1} is the inverse of kernel matrix K which is the quantity of interest in our SGD optimization framework, G is the QMC data, and A is the spectral function. In this approach, we use the optimized sample spectra obtained in FESOM to be our data to represent A. Although none of the FESOM samples is the real spectrum, each sample $\tilde{A}^r$ is a reasonably good estimate for the real spectrum A since the error between $K\tilde{A}^r$ and G is very small. In this way, we let the cost function in our optimization problem to be the square error in comparing A in (9), i.e. let

$$F(K_i^{-1}, \tilde{A}^r) := \sum_n \Big(\sum_l \big((K_i^{-1})_{nl} G_n - \tilde{A}_l^r \big)^2 \Big) \tag{10}$$

in the SGD method. Since $\{\tilde{A}^r\}_{r=1}^R$ are samples in a collection of ω_l functions which contain covariance information along frequency axis ω, the potential F defined in (10) also reflects the covariance of $\{\tilde{A}^r\}_{r=1}^R$. The SGLD equation (8) related to the analytic continuation now is formulated as

$$K_{i+1}^{-1} = K_i^{-1} - 2\alpha \sum_n \sum_l |(K_i^{-1})_{nl} G_n - \tilde{A}_l^r| + \beta\epsilon_i, \quad i = 0, 1, 2, \ldots, N-1, \tag{11}$$

where $\tilde{A}^r$ is a random selection in $\{\tilde{A}^r\}_{r=1}^R$, and N is a user defined integer as the stopping criteria for iteration. Then, we let the approximate inverse kernel, denoted by $\hat{K}^{-1}$, to be $\hat{K}^{-1} := K_N^{-1}$. To initialize our optimization procedure, we let the initial condition for K^{-1}, i.e. K_0^{-1}, to be calculated by the single value decomposition (SVD) inverse of K. We want to point out that although SVD is a popular method to calculate the inverse of matrices, we don't rely on the accuracy of SVD in our approach due to the highly ill-posedness of the analytic continuation problem and the noise perturbation in the QMC data. Since the goal of this SGD approach is to get an estimate for the spectral function A, when we get $\hat{K}^{-1}$, we calculate the estimated spectrum, denoted as $\hat{A}$, by $\hat{A} := \hat{K}^{-1}G$.

One important property of the optimized stochastic spectral samples is that for some features, most samples have very similar behaviors. However, sample spectra may have very different behaviors over certain frequency regions. As a result, the estimated K^{-1} obtained in the SGD method would provide a spectral function that recovers the features which are suggested by most optimized samples. On the other hand, the diverse behaviors of samples in describing some features would expand the searching domain in the SGD method and provide more variety to discover features that could fit the model (QMC data) better. Actually, the ill-posedness of the analytic continuation problem occurs primarily at the parts that sample spectra provide different features, and being able to have better estimation on these frequency regions and capture some typical fine features would be very useful in the analytic continuation problem.

Supervision in the SGD. It's worthy to point out that the SGD optimization we discussed above would learn from the data and get an approximation for K^{-1} with perturbations of the inflation noises. However, since the ultimate quantity of interest in the analytic continuation problem is the spectral function A, which has specific physical meaning, we should also consider physics in the SGD optimization so that the randomness in the SGLD (11) would not lead us to some non-physical outcomes. To this end, we introduce a physical supervision procedure to guide the SGD optimization and formulate a supervised SGD. Specifically, when we select a sample $\tilde{A}^r$ and an inflation random noise ϵ_i, we apply some physical knowledge about reasonable behaviors of spectra as supervision guidelines. If the K_{i+1}^{-1} calculated by (11) generates a spectrum that violates the known physical knowledge, we reduce the learning rate α to reduce the influence of the random pair $(\tilde{A}^r, \epsilon_i)$. The reason that we don't drop the corresponding sample by letting the learning rate to be 0 is that the information contained in $\tilde{A}^r$ is still valuable and it might be a necessary intermediate step that leads to a better local optimal estimate.

In what follows, we list a few supervision conditions we may consider as examples to supervise the update from K_i^{-1} to K_{i+1}^{-1}.

- Since the spectrum should always be nonnegative, we reduce the learning rate α for the candidate K_{i+1}^{-1} that will cause negative values in A and redo the SGD step (11) with the same selection of sample.
- If we know from physics that the spectral function should have a certain peak in a certain frequency region and the candidate K_{i+1}^{-1} that we generated from the random pair $(\tilde{A}^r, \epsilon_i)$ does not indicate this physical behavior, we reduce the learning rate and redo the SGD step.
- All the optimal samples generated in the FESOM would build a confidence band. It is most likely that the true spectrum lies within the high probability

density region of the confidence band. For the optimized sample with certain parts lie outside of the confidence band, we reduce the learning rate for this sample and redo the SGD step.

Since the initial condition K_0^{-1} is calculated from SVD inversion, K_0^{-1} may violate many of supervision restrictions and we may not be able to effectively update K^{-1} at beginning. To improve the efficiency of the SGD algorithm, we allow several burn in steps at first and only use the plain SGD without supervision.

Convolution for the spectrum. Since we calculate our estimated spectral function from $\hat{K}^{-1}$, we might recover an irregular spectral curve. In order to get a smoother estimate for the spectral curve, we implement a convolution step to smooth the spectral function. We want to mention that the convolution step is different from the regularization in the MaxEnt and FESOM as we have the control to the level of smoothness we want in the spectrum.

Discussions on the algorithm. In what follows, we bring out some discussions on the SGD approach for analytic continuation which is introduced above. For the convenience of presentation, from here and in the following we generally call our approach the *SGD*.

- *Convergence*
  In machine learning algorithms, the convergence of SGD is an important topic. It has been shown that the aforementioned SGLD scheme could approach to the global minimum when the iteration step is large enough [9, 12, 13, 14, 17]. However, it has also been proved that when the dimension of the problem is high, the SGLD approaches to the global minimum exponentially slow and the global convergence would be especially difficult to achieve for ill-posed optimization problems [18]. Therefore, we should not expect that we will obtain the global minimum in the analytic continuation problem. On the other hand, it can be shown that after reasonable length of iterations, the proposed K inverse, K_i^{-1}, moves around the global minimum. This would also cause the randomness of our result. It's worthy to mention that in the high confident region of the FESOM samples, it's easier for SGLD to reach the convergent feature; and in the low confident region (with wide confidence band), it's less likely that SGLD will reach converged true features due to the large variance in the samples.
- *Usage of FESOM samples*
  In the SGLD equation (8), when we approximate ∇F by $\nabla F(w_i, z)|_{z\in\mathcal{Z}}$, we have considered the distribution information contained in the data Z through the gradient set $\{\nabla F(w_i, z)\}_{z\in\mathcal{Z}}$. Specifically, in the SGLD equation (11) for the analytic continuation problem, the learning procedure actually considers the variance of $\{\tilde{A}^r\}_{r=1}^R$ by exploring different choices of $\tilde{A}^r$. Apparently, this gives us more information than just taking the average of samples $\{\tilde{A}^r\}_{r=1}^R$. Indeed, different choices of samples would influence the transition from K_i^{-1} to K_{i+1}^{-1} in different ways. For example, the smaller variation of $\{\tilde{A}^r\}_{r=1}^R$ in certain region may cause smaller transition stepsize from K_i^{-1} to K_{i+1}^{-1} in some matrix components. On the other hand, the larger variation of $\{\tilde{A}^r\}_{r=1}^R$ in certain region may cause larger transition stepsize from K_i^{-1} to K_{i+1}^{-1} in some matrix components. This allows us to search wider range of K^{-1} values to better solve the inverse problem. In this way, the SGLD procedure could

explore adaptively admissible domain for K^{-1} that reflects the variance of $\{\tilde{A}^r\}_{r=1}^R$ samples. As a result, this gives us more opportunities to find a K^{-1} that could generate a spectral function with physically meaningful features and fit the QMC data better.

- *Randomness of the SGD learned K^{-1}.*

  It's important to point out that we do not require the SGD method to reach the global minimum and our approximated K^{-1} has random behavior. Therefore different realizations and different iteration steps of the SGD will generate different $\hat{K}^{-1}$ that result different spectra through (9). To provide a more rigorous estimate for the spectrum A, we need a spectrum selection step.

Spectrum selection. For a pre-chosen user defined positive integer M, we run the above SGD algorithm M times and get M realizations of approximated inverse kernel, i.e. $\{\hat{K}_m^{-1}\}_{m=1}^M$. Then, we derive $\{\hat{A}_m\}_{m=1}^M$ from $\{\hat{K}_m^{-1}\}_{m=1}^M$. Although all the estimated spectra $\{\hat{A}_m\}_{m=1}^M$ have considered the entire FESOM spectral samples and have similar structure, we still need a criteria to select one spectrum to reduce the uncertainty of our estimation. Since the goal of the analytic continuation problem is to minimize the error between KA and G, in this approach we define $Err := \|K\hat{A} - G\|_{L^2}$ as the L^2 error in fitting the data and use Err to be a criteria and pick the $\hat{A}$ among $\{\hat{A}_m\}_{m=1}^M$ with the smallest Err error as our estimated spectral function.

Summary of the algorithm. We summarize the SGD method as following.

```
Step 1: Input the learning rate $\alpha$, $\beta$, the number of iteration $N$, the number of SGD
    realizations $M$, and the initial inverse kernel $K_0^{-1}$
Step 2: For each realization index $m = 1, 2, \cdots, M$, do the following iteration:
    For $i = 0 \cdots, N-1$,
        1 Compute $K_{i+1}^{-1}$ from the SGLD (11).
        2 Carry out the supervision procedure to decide wether need to reduce the
          learning rate and redo the SGLD or not.
        3 Implement the convolution step to get a smoother estimation $\hat{A}_m$.
Step 4: Select a representative among $\{\hat{A}_m\}_{m=1}^M$ with the smallest error to be the
    estimated spectral function $\hat{A}$.
```

4. Numerical experiments.

In this section, we demonstrate the performance of our SGD algorithm by using three numerical examples. In the first example, we focus on comparing this SGD algorithm with the FESOM, in which we take average of all the optimized stochastic spectral samples to build up an estimate for the spectrum, and show the improvement of the SGD algorithm in solving a synthetically designed analytic continuation problem. In Example 2, we consider a two-dimensional Hubbard model and compare the SGD method with both FESOM and MaxEnt to demonstrate the effectiveness of SGD in solving a real physical problem. The MaxEnt method that we compare with in our numerical experiments is a regularized optimization method, which is the state-of-the-art method in analytic continuation, and the regularization term that we add to the χ^2 error is in the form of entropy[6]. Then, in Example 3 we compare our method with MaxEnt in a specially designed example. The purpose of the third example is to demonstrate that

SGD could recognize small changes in the data and provide some finer structures in spectral functions which are usually smoothed out in MaxEnt as a regularized deterministic optimization method.

Example 1. In the first example, we assume that we receive QMC data G which is corresponding to the "real" spectral function A as plotted by the black curve in Fig. 1. Using the data G, we carry out the FESOM algorithm and obtain a samples set consisting stochastic optimal spectra. By saying "optimal" in FESOM, it means that all the FESOM samples produce very low χ^2 errors in fitting the data. However, although each sample has low χ^2 error, different samples have different behavior. In Fig. 2 (a), we plot 10 realizations of FESOM samples and compare them with the true spectrum $A(\omega)$, where the FESOM samples are presented by dashed blue curves and the true spectrum is the black curve. From this subplot, we can see

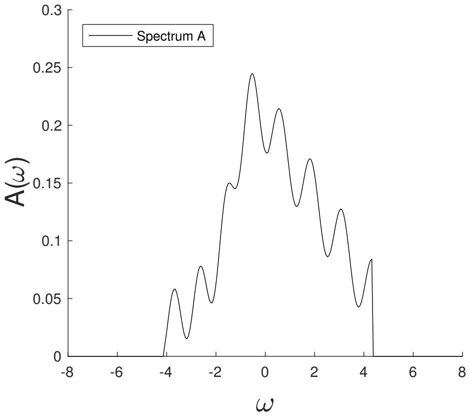
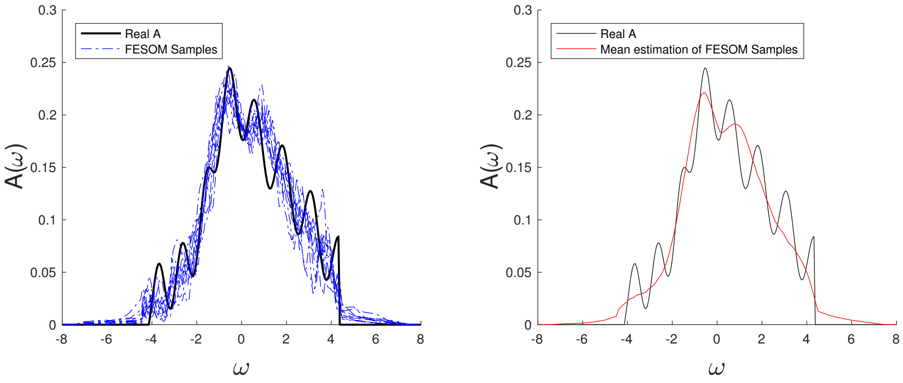


FIGURE 1. Example 1. True spectrum



(A)

(B)

FIGURE 2. Example 1. (a) FESOM samples; (b) FESOM estimation

that on the $(-2,0)$ frequency region, all the samples have similar behavior; on the other hand, FESOM samples demonstrate different structures on the $(-4,-2)$ and $(0,4)$ frequency region which indicate that we might encounter more uncertainties on these intervals. In Fig. 2 (b), we plot the FESOM estimate by taking the average of all the samples – in this experiment we take the average of 300 samples. The red curve is the mean of samples as the FESOM estimate and the black curve is the real spectrum. We can see that the FESOM estimation captures the main trend of the real spectrum and provides reasonable estimation at the low frequency region. However, it ignores most fine features due to the Monte Carlo averaging of samples and only provides a smooth curve that pass through all the features. On the other hand, in Fig. 2 (a) we observe that FESOM samples actually could provide many features in the true spectrum.

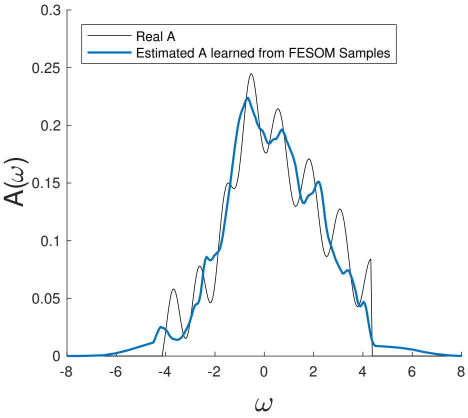


FIGURE 3. Example 1. Estimated spectrum learned from FESOM samples

In Fig. 3, we plot the estimated spectrum obtained by using the SGD method, where the black curve is the real spectrum and the blue curve is the estimated spectrum learned from FESOM samples. From this figure, we can see that the extra supervised learning procedure makes the SGD method capture some of the fine features in the estimated spectral function since running SGD through FESOM samples allows us to explore more spectral domain to better fit the data.

Example 2. In this example, we consider a two-dimensional Hubbard model on a square lattice with nearest-neighbor hopping t and Coulomb repulsion U described by the Hamiltonian

$$H = -t \sum_{\langle ij \rangle} c^{\dagger}_{i\sigma} c_{j\sigma} + U \sum_{i} n_{i\uparrow} n_{i\downarrow}, \tag{12}$$

where $c^{\dagger}_{i\sigma}$ creates and $c_{i\sigma}$ destroys an electron with spin $\sigma = \uparrow, \downarrow$ on site i and $n_{i\sigma} = c^{\dagger}_{i\sigma} c_{i\sigma}$ is the corresponding number operator. We use the dynamical mean-field theory (DMFT) [4] together with a non-crossing approximation (NCA) [3] to obtain the local spectral function $A(\omega)$ in the antiferromagnetic state as the true spectral function in this example. The local spectral function $A(\omega)$ we obtain is shown as the black line in Fig. 4. One can see that there are fine structure with

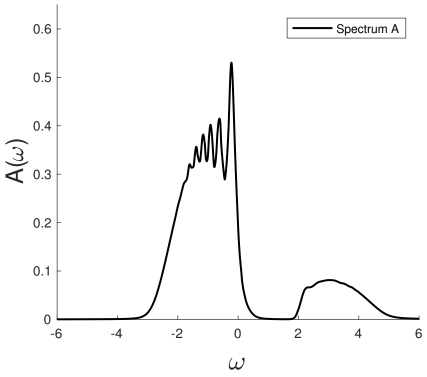
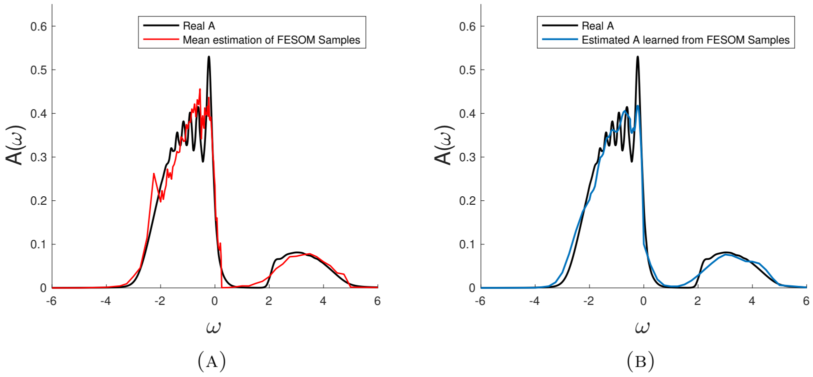


FIGURE 4. Example 2. True spectrum

multiple peaks in the lower Hubbard band (in the frequency interval $(-2, 0)$) and a major peak around the 0 frequency. These resonances reflect the bound states of a hole propagating in an antiferromagnetic background [16].

In Fig. 5, we present the estimations obtained by using FESOM and SGD. In each subplot, we use the black curve to represent the real spectrum; in Fig. 5 (a), the red curve is the FESOM estimate; and in Fig. 5 (b), the blue curve is the estimate learned from FESOM samples. We can see that the SGD method could present a better result compared with FESOM since SGD could explore more feasible spectral space to better fit the data. Actually, denoting $Err := \|K\hat{A} - G\|_{L^2}$ to be the L^2 error in matching the original data, we have that the error for FESOM is $Err = 0.0068$ and the error for SGD is only $Err = 0.0038$ which is much smaller than the FESOM error. Since analytic continuation is a highly ill-posed problem, a better description for the target spectral function with lower error at the same time indicates that by considering the entire FESOM sample set, SGD could provide better estimation results.



(A)

(B)

FIGURE 5. Example 2. (a) FESOM estimation; (b) Estimated spectrum learned from FESOM samples.

In addition to the comparison with FESOM, in this example we also want to present the comparison between SGD and MaxEnt. In Fig. 6, we plot the MaxEnt estimate and the SGD estimate in the same figure, where the black curve is the real spectrum, the green curve is the MaxEnt estimate and the blue curve is the estimated spectrum learned from the FESOM samples. From this figure, we can see that both MaxEnt and SGD could capture the main peak around the 0 frequency region. However, the MaxEnt missed the fine structure in the lower Hubbard band. Instead, it gives a wrong peak around the -2 frequency. On the other hand, the SGD accurately describes the main trend of the fine structure as well as a hint for another major peak besides the main peak around the 0 frequency.

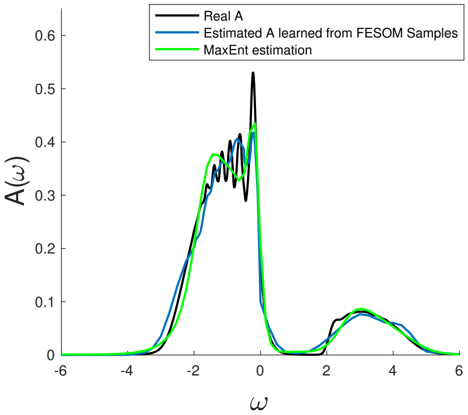


FIGURE 6. Example 2. Comparison between SGD and MaxEnt

Example 3. In this example, we compare SGD with MaxEnt in a synthetic analytic continuation problem. The main purpose of this example is to show that the SGD algorithm could provide some fine structure that might be ignored by MaxEnt.

We first consider a spectrum given in Fig. 7. From the figure, we can see that the

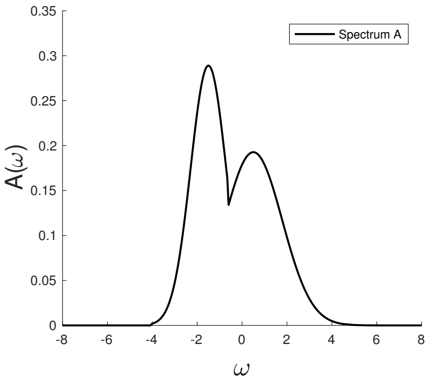


FIGURE 7. True spectrum

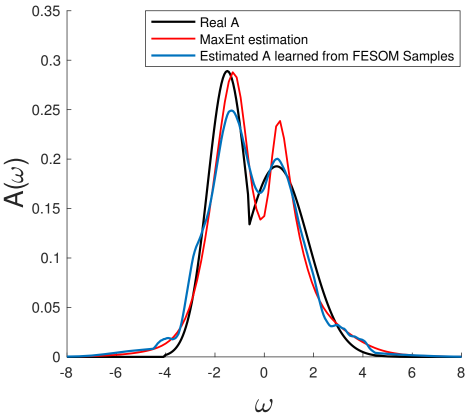


FIGURE 8. Example 3. Estimations for the spectrum.

spectrum has two smooth peaks at the -2 and the 0 frequencies. Then, we apply both MaxEnt and SGD to process the data corresponding to the spectrum in Fig. 7 and plot the performance of MaxEnt and SGD estimations in Fig. 8. The black curve is the true spectrum, the red curve is the estimate obtained by using MaxEnt and the blue curve is the estimate obtained by using the SGD. We can see from Fig. 8 that indeed both MaxEnt and SGD work well in estimating the spectrum and they could capture both the main features at the -2 and the 0 frequencies.

On the other hand, we consider another spectral function given by Fig. 9, which is very similar to the spectrum in Fig. 7 except that there's a fine feature in the positive frequency region. For the clarification of presentation, we name the spectrum in

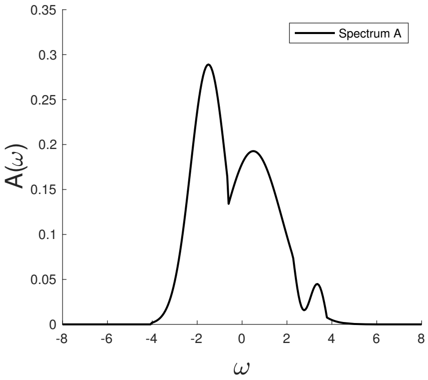
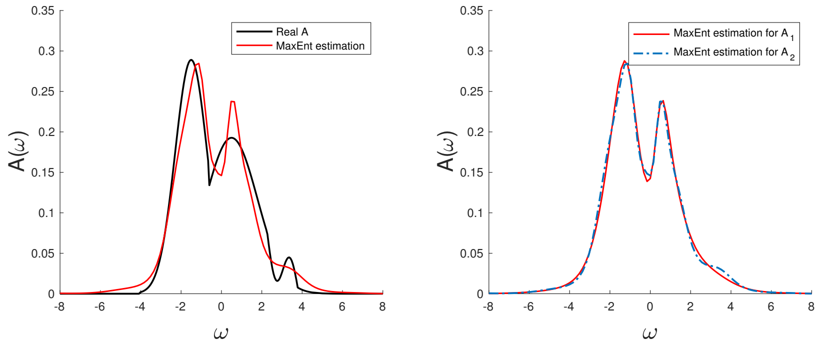


FIGURE 9. Example 3. Spectrum with fine feature in positive frequency region.

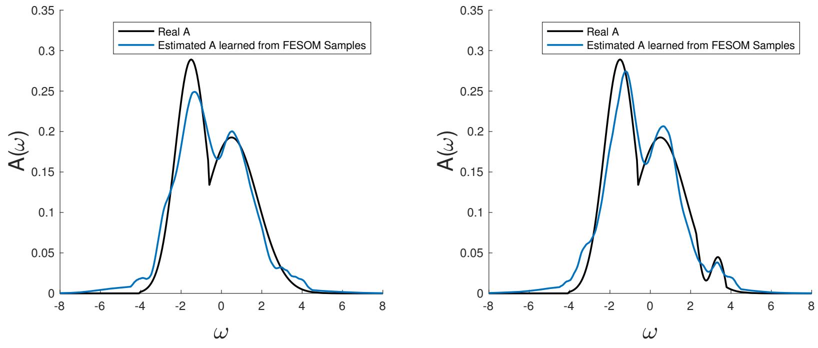
Fig. 7 by A_1 and the spectrum in Fig. 9 by A_2. Since A_2 is similar to A_1, the QMC data corresponding to A_2 would be very similar to the QMC data corresponding to A_1. In this way, it's difficult for state-of-the-art methods, such like MaxEnt, to recognize the difference in the data due to the standard regularization procedures which could smooth out this difference. In Fig. 10 (a), we present the MaxEnt estimate based on the data corresponding to the spectrum A_2, where the black curve is the true spectrum and the red curve is the estimate obtained by MaxEnt. We can see that MaxEnt could still capture both main features well. But, as expected, it does not provide the fine feature on frequency interval $(3, 4)$. In Fig. 10 (b), we compare the MaxEnt estimations for the data corresponding to the spectrum A_1 and the spectrum A_2, where the red curve is the estimate obtained by MaxEnt for the spectrum A_1 and the dashed blue curve is the estimate obtained by MaxEnt for the spectrum A_2. We can see from this subplot that the MaxEnt estimation has very similar behaviors although it estimates spectra corresponding to different data sets.



(A) (B)

FIGURE 10. Example 3. (a) MaxEnt estimation for A_2; (b) Comparison of MaxEnt in estimating A_1 (red) and A_2 (blue).

Then, in Fig. 11 (a) and (b), we show SGD estimates for the data corresponding to the spectra A_1 and A_2, respectively. We can see that the SGD method could



(A) (B)

FIGURE 11. Example 3. (a) SGD estimation for A_1; (b) SGD estimation for A_2.

capture the main features for both spectra A_1 and A_2. At the same time, from Fig. 11 (b) we can see that SGD could give a hint for the fine feature on frequency interval $(3, 4)$ and we can clearly tell that there's some difference from the data corresponding to A_1 and the data corresponding to A_2.

5. Conclusion.

In this paper, we introduced a stochastic gradient descent algorithm for stochastic optimization in solving analytic continuation problems. The SGD algorithm could learn fine features for targeting spectral functions in the analytic continuation problem based on the QMC data and the optimized stochastic samples that we obtain in the *fast and efficient stochastic optimization method* (FESOM). The SGD algorithm for stochastic optimization is an extension of FESOM. Instead of taking the Monte-Carlo average of samples in FESOM, SGD considers the information contained in the optimized spectral samples through stochastic gradient descent schemes. In this way, SGD takes the advantage of FESOM, which provides variance of spectral samples. As a result, the SGD algorithm could explore a larger spectral space to better fit the QMC data. In the numerical experiments, we showed that SGD out-performs FESOM, and it could capture fine features that are typically ignored by the maximum entropy method (MaxEnt).

Acknowledgments. This work is partially supported by the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research through CompFUSE project and FASTMath Institute. The first author also acknowledges support by U.S. National Science Foundation under Contract DMS-1720222.

REFERENCES

[1] 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.

[2] S. Fuchs, T. Pruschke and M. Jarrell, Analytic continuation of quantum monte carlo data by stochastic analytical inference, *Physical Review E*, **81** (2010), 056701.

[3] A. Georges, G. Kotliar, W. Krauth and M. J. Rosenberg, Self-consistent large-n expansion for normal-state properties of dilute magnetic alloys, *Physical Review B*, 1988, page 2036.

[4] A. Georges, G. Kotliar, W. Krauth and M. J. Rosenberg, Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions, *Reviews of Modern Physics*, **68** (1996), 13–125.

[5] S. F. Gull and J. Skilling, Maximum entropy method in image processing, *IEE Proceedings F*, **131** (1984), 646–659.

[6] M. Jarrell and J. Gubernatis, Bayesian inference and the analytic continuation of imaginary-time quantum monte carlo data, *Physics Reports*, **269** (1996), 133–195.

[7] 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.

[8] A. S. Mishchenko, N. V. Prokof'ev and A. Sakamoto, Diagrammatic quantum monte carlo study of the fröhlich polaron, *Physical Review B*, **62** (2000), 6317–6336.

[9] D. Needell, N. Srebro and R. Ward, Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm, *Mathematical Programming*, **155** (2016), 549–573.

[10] N. V. Prokof'ev and B. V. Svistunov, Spectral analysis by the method of consistent constraints, *Jetp Lett.*, **97** (2013), 649–653.

[11] A. Sandvik, Stochastic method for analytic continuation of quantum monte carlo data, *Physical Review B*, (1998), 10287–10290.

[12] I. Sato and H. Nakagawa, Convergence analysis of gradient descent stochastic algorithms, *Proceedings of the 31st International Conference on Machine Learning*, (2014), 982–990.

[13] O. Shamir and T. Zhang, Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes, *Proceedings of the 30th International Conference on Machine Learning*, 2013, p28.

[14] A. Shapiro and Y. Wardi, Convergence analysis of gradient descent stochastic algorithms, *Journal of Optimization Theory and Aplications*, **91** (1996), 439–454.

[15] R. N. Silver, J. E. Gubernatis, D. S. Sivia and M. Jarrell, Spectral densities of the symmetric anderson mode, *Physical Review Letters*, 1990, 496–499.

[16] R. Strack and D. Vollhardt, Dynamics of a hole in the t-j model with local disorder: Exact results for high dimensions, *Physical Review B*, 1992, 13852.

[17] L. Wu, C. Ma and W. E, How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective, *NeurIPS 2018*, 2018, 8289–8298.

[18] Y. Zhang, P. Liang and M. Charikar, A hitting time analysis of stochastic gradient langevin dynamics, *Conference on Learning Theory*, 2017, 1980–2022.

E-mail address: bao@math.fsu.edu
E-mail address: maierta@ornl.gov