

Article

Fractional Stochastic Search Algorithms: Modelling Complex Systems via AI

Bodo Herzog ^{1,2} 

¹ Economic & Data Science Department, ESB Business School, Reutlingen University, Alteburgstr. 150, 72762 Reutlingen, Germany; bodo.herzog@reutlingen-university.de

² RRI-Reutlingen Research Institute, Reutlingen University, 72762 Reutlingen, Germany

Abstract: The aim of this article is to establish a stochastic search algorithm for neural networks based on the fractional stochastic processes $\{B_t^H, t \geq 0\}$ with the Hurst parameter $H \in (0, 1)$. We define and discuss the properties of fractional stochastic processes, $\{B_t^H, t \geq 0\}$, which generalize a standard Brownian motion. Fractional stochastic processes capture useful yet different properties in order to simulate real-world phenomena. This approach provides new insights to stochastic gradient descent (SGD) algorithms in machine learning. We exhibit convergence properties for fractional stochastic processes.

Keywords: fractional Brownian motion; fractional stochastic gradient descent; machine learning; stochastic gradient descent; complex systems

MSC: 37M05; 60G18; 60G22; 62M45; 68T05; 68Q01; 65Y04; 68M07



Citation: Herzog, B. Fractional Stochastic Search Algorithms: Modelling Complex Systems via AI. *Mathematics* **2023**, *11*, 2061. <https://doi.org/10.3390/math11092061>

Academic Editors: Juan Eduardo Nápoles Valdes and Miguel Vivas-Cortez

Received: 2 April 2023

Revised: 24 April 2023

Accepted: 25 April 2023

Published: 26 April 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The gradient descent methodology is not computationally efficient in all applications. Sometimes, optimization algorithms become stuck in the flat regions of manifolds. In those cases, the optimization algorithm requires a long time to escape. This is the challenge of a vanishing gradient where, for instance, $\nabla z(\theta)$ is almost zero (see Section 4). The method of stochastic gradient descent (SGD) generally overcomes this problem.

Recent advancements in the field of fractional stochastic processes exhibit the theoretical benefits of modeling complex systems [1–4]. Yet, so far no literature exists on fractional stochastic gradient descent (fSGD) or fractional stochastic networks. This paper sketches the potential of such new literature for modeling complex systems. Moreover, we exhibit that fractional stochastic processes are an advancement in machine learning (ML) and artificial intelligence (AI).

The methodology of fractional stochastic gradient descent and the role of stochastic neural networks are based on a generalized assumption of randomness. Mandelbrot and Van Ness defined a fractional Brownian motion (fBM), B_t^H , together with a Hurst parameter in 1968 [5]. For $H = \frac{1}{2}$, we obtain a standard Brownian motion. Yet, for $H \neq \frac{1}{2}$, we obtain new forms of randomness or stochastic processes that match real-world phenomena.

The new feature of a fractional Brownian motion (fBM) is that increments are inter-dependent. In the literature, this is called self-similarity. A self-similar stochastic process reveals invariance with respect to the time scale (scaling invariance). A standard Brownian motion or a Lévy process displays different properties. They have independent increments and belong to the famous class of Markov processes.

However, in science, there is ubiquitous evidence that fractional stochastic processes are of relevance. For instance, we frequently observe probability densities with sharp peaks, which is related to the phenomena of long-range interdependence. In many real-world observations and applications, we find the presence of interdependence, too. This pattern can be captured by fractional stochastic processes.

Nonetheless, some phenomena are even more complicated and require further generalization towards sub-fractional stochastic processes. The literature on sub-fBM's demonstrates that those stochastic processes are useful in scientific applications [6]. A sub-fractional Brownian motion provides a nexus between a Brownian motion and fractional stochastic process. Those processes were introduced by Tudor et al. [7,8] and Bojdecki et al. [9]. Note that, as sub-fractional stochastic processes are not martingale processes, the basic tools of stochastic analysis are insufficient. However, researchers have developed new machinery to handle fractional stochastic processes, such as [10] or [11–15].

In this paper, our purpose is to develop and study the idea of fractional stochastic gradient descent algorithms. Our approach generalizes the existing literature on stochastic gradient descent (SGD) and stochastic neural networks. For instance, Hopfield [16] developed neural networks consisting of several perceptrons with randomness. Similarly, a Boltzmann network is a type of stochastic neural network wherein the output of the activation function is interpreted as a probability.

A study already exists about stochastic gradient descent and its challenges in machine learning [17,18]. Recent developments in the theory and applications of stochastic gradient descent are discussed in the following papers: Schmidt et al. [19], Haochen and Sra [20], Gotmare et al. [21], Curtis and Scheinberg [22], de Roos et al. [23]. The focus of our research is the motivation of fractional stochastic gradient descent (fSGD) algorithms. Thus, our research is beyond the scope of current literature and focuses on the possibility of fractional stochastic gradient descent in theory. We neglect potential computational limitations in machine learning.

The paper is organized as follows. Section 2 provides preliminary definitions. Subsequently, we introduce the foundations of fractional stochastic processes in Section 3. Section 4 introduces the idea of fractional stochastic search algorithms and derives the convergence results in general. Finally, in Section 5, we apply the method to two different cases. Section 6 concludes the paper.

2. Preliminaries

Machine learning is mainly based on neural networks and efficient optimization algorithms. The most primitive neural model is inspired by the work of Rosenblatt [24]. In the following section, we define the major elements from a machine learning perspective.

Definition 1. A stochastic neuron is defined by n -inputs $\mathbf{X} = (x_1, \dots, x_n)^T$, n -weighting factors $\mathbf{W} = (w_1, \dots, w_n)^T$ and an n -dimensional vector of biases \mathbf{B} together with a sigmoid activation function $\sigma_c(\zeta) = \frac{1}{1+e^{-c\zeta}}$ with $c \geq 0$ and a stochastic output Y

$$Y = \sigma(\mathbf{W}^T \mathbf{X} + \mathbf{B}),$$

where $\sigma(\mathbf{W}^T \mathbf{X} + \mathbf{B}) \in (0, 1)$. Hence, we define the output for $Y = 1$ by $\text{Prob}(\mathbf{Y} = \mathbf{1}; \mathbf{X}) = \sigma(\mathbf{W}^T \mathbf{X} + \mathbf{B})$ and for $Y = 0$ by the inverse probability: $\text{Prob}(\mathbf{Y} = \mathbf{0}; \mathbf{X}) = 1 - \sigma(\mathbf{W}^T \mathbf{X} + \mathbf{B}) = \sigma(-\mathbf{W}^T \mathbf{X} - \mathbf{B})$.

Note that, if the activation potential is greater than zero, such as $\mathbf{W}^T \mathbf{X} + \mathbf{B} > 0$, then this neural network is not necessarily activated according to Definition 1. An activation value of one only occurs with the probability of the activation function.

In machine learning, the gradient descent algorithm is omnipresent in all optimization problems. Yet, it does not provide robust solutions in each case. There are computational obstacles, such as when the algorithm becomes stuck in a local minimum or lost in a plateau from which it takes a long time to get out. A plateau is defined as a flat surface region where the gradient $\nabla \sigma_c$ is very small (or almost zero).

The optimization algorithm of a neural network always has the goal of finding the optimal weighting parameters $\theta = (\mathbf{W}, \mathbf{B})$. The standard algorithm used to optimize the parameters is frequently reformulated in order to minimize the cost function. This is called the gradient descent method. This method is closely related to Newton's algorithm in

numerical computing. The following definition summarizes the algorithm from a machine learning vantage point.

Definition 2. The gradient descent algorithm is defined by

$$\theta_{t+1} = \theta_t - \lambda_t \mathbf{C}_t \mathbf{g}_t, \quad (1)$$

where $\mathbf{g}_t = \nabla \mathcal{L}(\theta_t)$ is the gradient of a cost function, \mathbf{C}_t is an optional conditioning matrix, and λ_t is the learning rate.

The stochastic gradient descent (SGD) method overcomes the obstacle if the gradient is close to zero. Indeed, SGD reaches a minimum along a non-linear stochastic process. In the following sections, we first discuss the literature and then generalize the approach to fractional stochastic processes.

3. Fractional Stochastic Processes

3.1. General Definitions

Consider a stochastic process with a Hurst parameter H . Subsequently, we define the elementary tools in fractional calculus.

Definition 3. Let $a, b \in \mathbb{R}$, $a < b$. Let $f \in L^1(a, b)$ and $\alpha > 0$. The left- and right-sided fractional integrals of f of order α are defined for $x \in (a, b)$, respectively, as

$${}_a D_x^{-\alpha} f(x) = {}_a I_x^{\alpha} f(x) = \frac{1}{\Gamma(\alpha)} \int_a^x (x-u)^{\alpha-1} f(u) du \quad -\infty \leq a \leq x,$$

and

$${}_b D_x^{-\alpha} f(x) = {}_b I_x^{\alpha} f(x) = \frac{1}{\Gamma(\alpha)} \int_x^b (u-x)^{\alpha-1} f(u) du \quad -\infty \leq x \leq b.$$

This is the fractional integral of the Riemann–Liouville type. In the same vein, we define fractional derivatives where we distinguish between left- and right-sided derivatives.

Definition 4. The fractional left- and right-sided derivatives, for $f \in I_a^{\alpha}(L^p)$ and $0 < \alpha < 1$, are defined by

$${}_a I_x^{-\alpha} f(x) = {}_a D_x^{\alpha} f(x) = \frac{1}{\Gamma(1-\alpha)} \left(\frac{d}{dx} \right) \int_a^x (x-u)^{-(\alpha+1)} f(u) du \quad (2)$$

and

$${}_b I_x^{-\alpha} f(x) = {}_b D_x^{\alpha} f(x) = \frac{(-1)^{\alpha}}{\Gamma(1-\alpha)} \left(\frac{d}{dx} \right) \int_x^b (u-x)^{-(\alpha+1)} f(u) du, \quad (3)$$

for all $x \in (a, b)$ and $I_a^{\alpha}(L^p)$ is the image of $L^p(a, b)$.

Let us assume $f \in I_a^1(L^1)$, then we obtain

$${}_a D_x^{\alpha} {}_a D_x^{1-\alpha} f(x) = Df(x), \quad {}_b D_x^{\alpha} {}_b D_x^{1-\alpha} f(x) = Df(x). \quad (4)$$

Notably, $D^{\alpha} f(x)$ exists for all $f \in C^{\beta}([a, b])$ if $\alpha < \beta$. Given those definitions, we are ready to define a Brownian motion:

Definition 5. Let H be $0 < H < 1$, and let B_0 be an arbitrary real number. We call $B^H(t, \omega)$ a fractional Brownian motion (fBM) with Hurst parameter H and starting value B_0 at time 0, such as

1. $B^H(0, \omega) = B_0$, and;

2. $B^H(t, \omega) - B^H(0, \omega) = \frac{1}{\Gamma(H+\frac{1}{2})} \left[\int_{-\infty}^0 [(t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}}] dB(s, \omega) + \int_0^t (t-s)^{H-\frac{1}{2}} dB(s, \omega) \right]$ [Wyle fractional integral];
3. Equivalent to the Riemann–Liouville integral:
 $B^H(t, \omega) - B^H(0, \omega) = \frac{1}{\Gamma(H+\frac{1}{2})} \int_0^t (t-s)^{H-\frac{1}{2}} dB(s, \omega).$

Next, let us consider the following corollary:

Corollary 1. Consider $H = \frac{1}{2}$ and $B_0 = 0$. Then the Brownian motion is of $B(t, \omega) = B^{\frac{1}{2}}(t, \omega)$.

Proof. Let $H = \frac{1}{2}$, we find $B^{\frac{1}{2}}(t, \omega) - B^{\frac{1}{2}}(0, \omega) = \frac{1}{\Gamma(1)} \int_0^t dB(s, \omega) = B(t, \omega)$. \square

In the literature, there exists an alternative, yet useful, definition:

Definition 6. A fractional Brownian motion is a Gaussian process $B^H(t)$ for $t \geq 0$ defined by the following covariance function

$$R^{fBM}(t, s) = \mathbb{E}[B^H(t)B^H(s)] = \frac{1}{2}[|t|^{2H} + |s|^{2H} - |t-s|^{2H}], \quad (5)$$

where the Hurst index is denoted by $H \in (0, 1)$.

Since the covariance of a Brownian motion is given in the literature, it is easy to extend the definition to an fBM with Hurst index H , such as

$$\begin{aligned} \text{Var}[B(t) - B(s)] &= \mathbb{E}[(B(t) - B(s))^2] = |t - s| \\ \Leftrightarrow \text{Var}[B^H(t) - B^H(s)] &= \mathbb{E}[(B^H(t) - B^H(s))^2] = |t - s|^{2H}, \end{aligned}$$

where we obtain the definition of a Brownian motion for $H = \frac{1}{2}$. Following Herzog [15], we derive the covariance step-by-step:

$$\begin{aligned} \text{Cov}[B^H(t)B^H(s)] &= \mathbb{E}[(B^H(t) - \mathbb{E}[B^H(t)])(B^H(s) - \mathbb{E}[B^H(s)])] = \mathbb{E}[B^H(t)B^H(s)] \\ &= \frac{1}{2} \left[\mathbb{E}[B^H(t)^2] + \mathbb{E}[B^H(s)^2] - \mathbb{E}[(B^H(t) - B^H(s))^2] \right] \\ &= \frac{1}{2} [|t|^{2H} + |s|^{2H} - |t-s|^{2H}]. \end{aligned}$$

Corollary 2. Consider a fractional Brownian motion. The expectation values of non-overlapping increments are $\mathbb{E}[B^H(t) - B^H(s)] \neq 0$ and the variance is of $\mathbb{E}[(B^H(t) - B^H(s))^2] = |t-s|^{2H}$ for all $t, s \in \mathbb{R}$

Proof. See [15]. \square

3.2. Properties

Next, we consider the properties of the fBM over time for different Hurst parameters. Suppose $0 < H < \frac{1}{2}$ or $\frac{1}{2} < H < 1$. If we assume that the Hurst parameter is of $0 < H < \frac{1}{2}$, we say the fractional stochastic process has a short memory. Conversely, if $\frac{1}{2} < H < 1$, we obtain the property of long-range dependence. Figure 1 illustrates sample processes for the three ranges of the Hurst parameters, H .

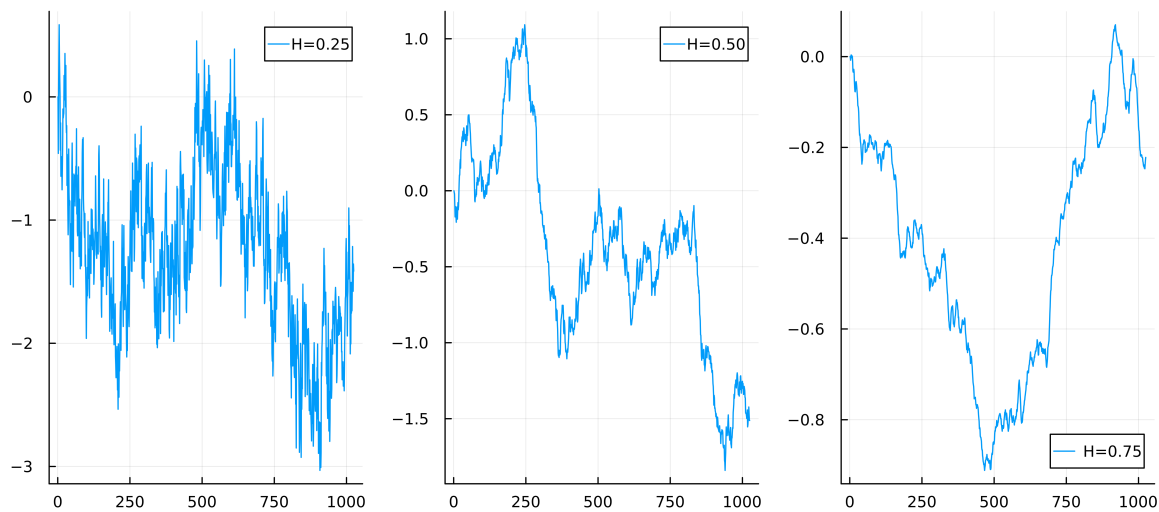


Figure 1. Different fractional Brownian motions with the following Hurst index: (Left-panel) $H = 0.25$, (middle-panel) $H = 0.50$ (standard BM), and (right-panel) $H = 0.75$.

Proposition 1. Given a fractional Brownian motion, we obtain the following properties:

1. The fBM has stationary increments: $B_t^H - B_s^H \stackrel{\text{dis.}}{=} B_u^H - B_s^H$;
2. The fBM is H -self-similar, such as $B^H(at) = a^H B^H(t)$;
3. The fBM is H -self-similar, such as $B^H(at) = a^H B^H(t)$;

Proof. The proof follows Herzog [15]. In order to prove the stationary of increments, we set $t_1 < t_2 < t_3 < t_4$. The equality of the covariance implies $Y := B^H(t_2) - B^H(t_1)$. Moreover, it has the same distribution, such as $X := B^H(t_4) - B^H(t_3)$. Subsequently, we find

$$\begin{aligned}\mathbb{E}[(B^H(t_2) - B^H(t_1))^2] &= (t_2 - t_1)^{2H} = (\Delta t)^{2H} \\ \mathbb{E}[(B^H(t_4) - B^H(t_3))^2] &= (t_4 - t_3)^{2H} = (\Delta t)^{2H},\end{aligned}$$

where $t_1 < t_2$ and $t_3 < t_4$ with $\Delta t = t_2 - t_1 = t_4 - t_3$. This demonstrates that the increments and the time evolution of the increments are the same at any given point. Consequently, we obtain stationary increments.

The second property of Proposition 1 is self-similarity. Consider the following definition,

$$\begin{aligned}\mathbb{E}[(B^H(at))^2] &= \frac{1}{2}[(at)^{2H} + (at)^{2H} - (at - at)^{2H}] = (at)^{2H} = a^{2H}t^{2H} \\ &= a^{2H}\mathbb{E}[(B^H(t))^2].\end{aligned}$$

Here, we find that $(B^H(at))^2 = a^{2H}(B^H(t))^2$ and $B^H(at) = a^H B^H(t)$. Part (3) is already given in Corollary 2. \square

3.3. Definition of Sub-Fractional Processes

In a recent paper, Herzog [15] described a sub-fractional Brownian motion (sub-fBM) as an intermediate between a Brownian motion and a fractional Brownian motion. Without loss of generality, a sub-fBM is a self-similar Gaussian process. Note that both the fBM and sub-fBM have the properties of self-similarity and long-range dependence, yet a sub-fBM does not have stationary increments [9].

Any Brownian motion is uniquely defined by its covariance. For the sub-fBM we denote covariance by $\text{Cov}(\zeta_t^H, \zeta_s^H)$.

Definition 7. Consider a sub-fractional Brownian motion with Hurst parameter H and a centered mean zero Gaussian process $\xi^H = \{\xi_t^H, t \geq 0\}$ with the following covariance function

$$R^{sfBM}(t, s) := \mathbb{E}[\xi_t^H \xi_s^H] = s^{2H} + t^{2H} - \frac{1}{2}[(s+t)^{2H} + |s-t|^{2H}], \quad (6)$$

where $\xi_0^H = 0$ and $\mathbb{E}[\xi_t^H] = 0$.

Note, a fractional Brownian motion coincides with a Brownian motion if the Hurst parameter is $H = \frac{1}{2}$. Thus, a Brownian motion on the real line has a covariance of $\text{Cov}(\xi_t^H, \xi_s^H) = s \wedge t := \min[s, t]$. The process ξ_t^H has the following representation for $H > \frac{1}{2}$ (see [25]):

$$\xi_t^H = \int_0^t K^H(t, s) dW_s, \quad (7)$$

$$K^H(t, s) = c_H \left(H - \frac{1}{2} \right) s^{1/2-H} \int_s^t (u-s)^{H-3/2} u^{H-1/2} du. \quad (8)$$

The kernel function of a sub-fractional Brownian motion is given by

$$\phi^{sfBM}(s, t) = \frac{\partial^2 \text{Cov}(\xi_t^H, \xi_s^H)}{\partial s \partial t} = H(2H-1) \left[|s-t|^{2H-2} - (s+t)^{2H-2} \right]. \quad (9)$$

3.4. Properties of Sub-Fractional Processes

In this subsection, we reiterate useful properties of sub-fractional Brownian motions such as those described in Herzog [15].

Lemma 1. Consider ξ_t^H be a sub-fBM for all t . The properties of the sub-fBM are:

1. $\mathbb{E}[(\xi_t^H)^2] = (2 - 2^{2H-1})t^{2H}$.
2. $\mathbb{E}[(\xi_t^H - \xi_s^H)^2] = -2^{2H-1}(t^{2H} + s^{2H}) + (t+s)^{2H} + (t-s)^{2H}$.
3. If $H \neq \frac{1}{2}$, then $\xi_t^H - \xi_s^H \stackrel{\text{dis.}}{\neq} \xi_u^H - \xi_s^H$, i.e., the increments are non-stationary.

Proof. See [15]. \square

Finally, we follow Herzog [15] and prove the following proposition:

Proposition 2. Let B_t^H be a fractional Brownian motion and ξ_t^H be a sub-fractional Brownian motion. For $H \in (\frac{1}{2}, 1)$, the following holds:

1. $\mathbb{E}[(\xi_t^H)^2] < \mathbb{E}[(B_t^H)^2]$;
2. $R_{\xi_t^H}(s, t) \leq R_{B_t^H}(s, t)$.

Proof. Obviously, an fBM has the following variance: $\text{Var}[B_t^H] = |t|^{2H}$. Similarly, we obtain the variance of $\text{Var}[\xi_t^H] = (2 - 2^{2H-1})|t|^{2H}$ for a sub-fBM. Subsequently, we have $0 < (2H-1) \ln 2$ if $H > \frac{1}{2}$.

The second part follows for $s, t > 0$:

$$s^{2H} + t^{2H} - \frac{1}{2}[(s+t)^{2H} + |t-s|^{2H}] \leq \frac{1}{2}[|t|^{2H} + |s|^{2H} - |t-s|^{2H}]$$

$$s^{2H} + t^{2H} \leq (s+t)^{2H}.$$

In the case of $s = t = 0$ or $s = 0, t \neq 0$, we have equality. \square

4. Fractional Stochastic Search

Let X_t be an m -dimensional stochastic process driven by a fractional Brownian motion B_t^H , where $H = \frac{1}{2}$. The respective stochastic process X_t is as follows:

$$dX_t = a(X_t)dt + \sigma(X_t)dB_t^H, \quad \text{and} \quad X_0 = x_0, \quad (10)$$

where X_0 is the initial value and $B_t^H = (B_1^H(t), \dots, B_m^H(t))$ is an m -dimensional fractional Brownian motion. Next, consider a cost function $z : \mathbb{R}^m \rightarrow \mathbb{R}$ which needs to be optimized. Hence, we study the vector field for which the auxiliary function $z(X(t))$ is decreasing. This requires us to find the expectation value:

$$\kappa(t) = \mathbb{E}[z(X(t))]. \quad (11)$$

Thus, the function $z(X(t))$ is stochastic and dependent on time t . In general, an optimization algorithm of a neural network minimizes the expectation value of this function. Utilizing the machinery of stochastic analysis, Dynkin's formula, among others, and following the approach described in [26], we obtain

$$\begin{aligned} \kappa(t) &= z(X_0) + \int_0^t \mathbb{E}[\mathcal{A}(z(X(s)))]ds \\ \kappa(t+dt) &= z(X_0) + \int_0^{t+dt} \mathbb{E}[\mathcal{A}(z(X(s)))]ds, \end{aligned} \quad (12)$$

where the operator $\mathcal{A} = \sum_k a_k \frac{\partial}{\partial X_k} + \frac{1}{2} \sum_{i,j} (\sigma \sigma^T)_{ij} \frac{\partial^2}{\partial X_i \partial X_j}$. The usage of Taylor-series approximation and the differencing of Equations (12) yields

$$\begin{aligned} \Delta\kappa(t) &= \kappa(t+dt) - \kappa(t) = \int_t^{t+dt} \mathbb{E}[\mathcal{A}(z(X(s)))]ds \\ &= \mathbb{E}[\mathcal{A}(z(X(t)))] + \mathcal{O}(dt^2). \end{aligned} \quad (13)$$

The method of steepest descent computes the gradient of $\Delta\kappa(t)$ such that the process X_t is as negative as possible. However, if X_t is a stochastic process, we need to study the expectation of the gradient, particularly where the value $\Delta\kappa(t)$ is as negative as possible, such as $\mathbb{E}[\mathcal{A}(z(X(t)))] < 0$.

In order to construct a stochastic process X_t with this property, we specify $a(X_t)$ and $\sigma(X_t)$ in Equation (10), respectively. Next, we specify the diffusion term in Equation (10), σ , or the product $\sigma\sigma^T$, which is a matrix, such that the algorithm in Equation (1) converges efficiently. Indeed, if we set the term of $\sigma\sigma^T$ as being inversely proportional to the Hessian matrix, \mathcal{H} , then $(\sigma\sigma^T)_{ij} = \tau(\mathcal{H}_z)_{ij}^{-1}$ with $\tau > 0$. Through this process we can show the convergence of the algorithm and the existence of the solution.

Given that function z is of class \mathcal{C}^2 and strictly convex, then, according to [27], the Hessian matrix \mathcal{H}_z is symmetric, real, positive definite, and non-degenerate. This guarantees that the Hessian matrix \mathcal{H}_z has an inverse, which is also positive definite. Efficient computation can be achieved by utilizing the Cholesky decomposition. One can show that the diffusion term σ is a lower triangular matrix satisfying $(\sigma\sigma^T)_{ij} = \tau(\mathcal{H}_z)_{ij}^{-1}$. Under those conditions, we compute $\mathcal{A}z(x)$

$$\begin{aligned} \mathcal{A}z(x) &= \sum_k a_k \frac{\partial z(x)}{\partial X_k} + \frac{1}{2} \tau \sum_{i,j} (\mathcal{H}_z^{-1})_{ij} (\mathcal{H}_z)_{ij} \\ &= \langle \nabla z(x), a(x) \rangle + \frac{n}{2} \tau. \end{aligned}$$

In order to minimize the gradient, $\mathcal{A}z(x)$, we have to minimize the first-term, because the second term is a constant. Choosing $a(x) = -\lambda \nabla z(x)$ and assuming $\mathcal{A}z(x) < 0$ obtains the following condition:

$$\begin{aligned}
\mathcal{A}z(x) &= \langle \nabla z(x), -\lambda \nabla z(x) \rangle + \frac{n}{2} \tau \\
&= -\lambda \|\nabla z(x)\|^2 + \frac{n}{2} \tau < 0 \\
\frac{\tau}{\lambda} &< \frac{2}{n} \|\nabla z(x)\|^2.
\end{aligned}$$

Using the square vector norm and the assumption of $\xi = \inf_x \|\nabla z(x)\|^2 \neq 0$, it is sufficient to set τ and λ as the main parameters in the SGD algorithm, such that $\tau = \frac{2\xi}{n} \lambda$. In the sequel, we apply this algorithm to fractional stochastic search problems.

5. Application of Fractional Stochastic Search

In this section, we demonstrate the working of a fractional stochastic search. We exhibit the convergence of a fractional stochastic search within neural networks.

5.1. Stochastic Search: Case I

Suppose that we have a neural network with the following cost function: $z(x) = \frac{1}{2}x^2$ for $x \in (a_1, a_2)$ with $a_i > 0$ and $a_1 < a_2$. The stochastic gradient descent method searches the minimum of this cost function.

Mathematically, the solution is obvious for this problem. The first derivative is of $z'(x) = x$ for $x \in (a_1, a_2)$. Hence, the minimum is at $z'(a_1) = a_1$, and consequently, the minimum value is $z(a_1) = \frac{1}{2}a_1^2$. Next, we show that we can obtain the same value under a fractional stochastic search algorithm in a neural network.

In step one, we establish an adequate stochastic differential equation according to $dX_t = a(X_t)dt + \sigma(X_t)dB_t^H$ for $X_0 = x_0 \in (a_1, a_2)$. The gradient of the cost function, which is equal to the first derivative $\nabla z(x) = x$, as well as the Hessian of the cost function and the second derivative, is defined as $\mathcal{H}(x) = \nabla^2 z(x) = 1$. Both conditions enable us to compute the Lipschitz continuous coefficient functions. For $a(X_t) = -\lambda \nabla z(x) = -\lambda x$. For $\sigma^2(X_t) = \tau \mathcal{H}^{-1}(x) = \frac{\tau}{\nabla^2 z(x)} = \frac{\tau}{1} = \tau$. Hence, we obtain $\sigma(X_t) = \sqrt{\tau}$. The stochastic differential equation for $H = \frac{1}{2}$ has the form:

$$dX_t = -\lambda X_t dt + \sqrt{\tau} dB_t^{\frac{1}{2}}. \quad (14)$$

The SDE in Equation (14) is an Ornstein–Uhlenbeck process driven by a Brownian motion $B^H(t)$ with the Hurst parameter $H = \frac{1}{2}$ [28].

The solution is divided into two parts: In part one, we solve the non-stochastic problem $dX_t = -\lambda X_t dt$. This is an ordinary differential equation and has the solution $X_t = X_0 e^{-\lambda t}$. In part two, we define an auxiliary function $Y_t = X_t e^{\lambda t}$ and apply the Itô–Doebelin’s lemma:

$$\begin{aligned}
dY_t &= X_t e^{\lambda t} (\lambda) dt + e^{\lambda t} dX_t \\
&= X_t e^{\lambda t} (\lambda) dt + e^{\lambda t} [-\lambda X_t dt + \sqrt{\tau} dB_t^H] \\
&= \sqrt{\tau} e^{\lambda t} dB_t^H.
\end{aligned}$$

Note that, in this case, the derivation coincides with a standard Brownian motion. Next, integrating the last line yields $Y_t - Y_0 = \sqrt{\tau} \int_0^t e^{\lambda s} dB_s^H$. Hence, we obtain $X_t = Y_t e^{-\lambda t}$, which is

$$\begin{aligned}
X_t &= \left[Y_0 + \sqrt{\tau} \int_0^t e^{\lambda s} dB_s^H \right] e^{-\lambda t} \\
&= X_0 e^{-\lambda t} + \sqrt{\tau} \int_0^t e^{-\lambda(t-s)} dB_s^H.
\end{aligned} \quad (15)$$

Based on Equation (15), we find the expectation of $\mathbb{E}[X_t] = X_0 e^{-\lambda t}$. Note that the expected integral of a Brownian motion is zero. For $t \rightarrow \infty$, the expected value is

$\lim_{t \rightarrow \infty} \mathbb{E}[X_t] = 0$. Next, utilizing the general condition of $\tau = \frac{2\xi}{n}\lambda$ for $\xi = \inf \|\nabla z(x)\|^2$ and $n = 1$ in Section 4, we obtain

$$\tau = \frac{2\xi}{n}\lambda = \frac{2 \inf \|a_1\|^2}{1}\lambda = 2a_1^2\lambda \quad \Leftrightarrow \quad \frac{\tau}{2\lambda} = a_1^2.$$

Finally, it remains to show that the SDE in Equation (14) converges to the minimum value. Hence, we study the convergence sequence:

$$\begin{aligned} \kappa(t) &= \mathbb{E}[z(X_t)|X_0] = \frac{1}{2}\mathbb{E}[(X_t)^2|X_0] = \\ &= \frac{1}{2}\mathbb{E}\left[X_0^2e^{-2\lambda t} + 2X_0e^{-\lambda t}\sqrt{\tau}\int_0^t e^{-\lambda(t-s)}dB_s^H + \tau\left(\int_0^t e^{-\lambda(t-s)}dB_s^H\right)^2\right], \end{aligned}$$

where, for $\mathbb{E}[(X_t)^2]$, we have substituted Equation (15). Next, we use the property that the expected stochastic integral is zero and the variance of the Brownian motion is $\text{Var}(dB_s^H) = ds$. Thus, we obtain

$$\begin{aligned} \kappa(t) &= \frac{1}{2}\left[X_0^2e^{-2\lambda t} + \tau\int_0^t e^{-2\lambda(t-s)}ds\right] = \\ &= \frac{1}{2}\left[X_0^2e^{-2\lambda t} + \tau\frac{e^{-2\lambda(t-s)}}{2\lambda}\Big|_0^t\right] = \\ &= \frac{1}{2}\left[X_0^2e^{-2\lambda t} + \frac{\tau}{2\lambda}(1 - e^{-2\lambda t})\right]. \end{aligned}$$

In order to show the convergence, we compute the limit of the sequence for time to infinity. We obtain the following:

$$\lim_{t \rightarrow \infty} \kappa(t) = \lim_{t \rightarrow \infty} \frac{1}{2}\left[0 + \frac{\tau}{2\lambda}\right] = \frac{1}{2}\frac{\tau}{2\lambda} = \frac{1}{2}a_1^2.$$

Indeed, we find that the (fractional) stochastic algorithm converges to the same minimum value of our function $z(x) = \frac{x^2}{2}$ for $x \in (a_1, a_2)$.

5.2. Stochastic Search: Case II

Conversely, suppose a fractional stochastic differential equation with a Hurst index $H \neq \frac{1}{2}$ of the form

$$dX_t = \mu X_t dt + \lambda X_t dB_t^H, \quad (16)$$

where we define $\mu := -\eta$ and $X_0 = x > 0$. We search the minimum of the function $z(X_t) = X_t$, where X_t is the solution of the SDE in Equation (16). This equation can be rewritten in the fractional Hida space $(\mathcal{S})_H^*$ as

$$\begin{aligned} \frac{dX_t}{dt} &= \mu X_t + \lambda X_t \diamond W_t^H \\ &= (\mu + \lambda W_t^H) \diamond X_t, \end{aligned}$$

where \diamond is defined as the Wick product. Using Wick calculus, we find the solution as

$$X_t = xe^{\diamond(\mu t + \lambda \int_0^t W_s^H ds)} = e^{\diamond(\mu t + \lambda B_t^H)}, \quad (17)$$

where we have used the definition $B_t^H = \int_0^t W_s^H ds$. By applying the following definitions $e^{\diamond\langle w, f \rangle} = e^{(\int_{\mathbb{R}} z_t dB_t^H - \frac{1}{2}\|z\|_H^2)}$, where $\|z\|_H^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} z(s)z(t)\phi(s, t)dsdt$ and $\phi(s, t) = H(2H - 1)|s - t|^{2H-2}$ for $s, t \in \mathbb{R}$, we obtain the final solution:

$$X_t = xe^{(\mu t + \lambda B_t^H - \frac{1}{2}\lambda^2 t^{2H})}. \quad (18)$$

The solution of Equation (18) has an expectation of $\mathbb{E}[X_t] = X_0 e^{-\eta t}$. Hence, for $t \rightarrow \infty$, the expected value is zero: $\lim_{t \rightarrow \infty} \mathbb{E}[X_t] = 0$. It remains to show the convergence of the fractional SDE in machine learning:

$$\begin{aligned}\kappa(t) &= \mathbb{E}[z(X_t)|X_0] = \mathbb{E}\left[x e^{\left(\mu t + \lambda B_t^H - \frac{1}{2}\lambda^2 t^{2H}\right)} | X_0\right] \\ &= x \mathbb{E}\left[e^{\left(\mu t + \lambda B_t^H - \frac{1}{2}\lambda^2 t^{2H}\right)}\right] \\ &= x e^{-\eta t}.\end{aligned}$$

In order to show convergence, we compute the limit of the sequence for time to infinity. We equally obtain $\lim_{t \rightarrow \infty} \kappa(t) = \lim_{t \rightarrow \infty} x e^{-\eta t} = 0$.

There are notable limitations of fractional stochastic gradient descent in general. Fractional calculus is built around the Riemann–Liouville integral, which is a non-local operator, lacks in uniqueness, and relies on the initial conditions. Given that a fractional process is not a martingale, the common stochastic tools are not applicable. Whether those properties constrain fractional stochastic gradient descent remains an open research question. Computational aspects might also be a limiting factor. However, for the first time, this research studies the idea of fractional search analogous to stochastic gradient descent in machine learning.

6. Conclusions

This article discovers fractional stochastic gradient descent algorithms for the optimization of neural networks. In the standard case, the fractional stochastic approach follows the well-known stochastic gradient descent method in machine learning. We discuss two special cases. First, we exhibit that fractional stochastic algorithms find the minima. This result might enhance algorithmic optimization in machine learning. Second, we discover the generalized patterns and properties of fractional stochastic processes. These insights may create a universal optimization approach in machine learning and AI in the future. We highlight the need for further research in that direction, particularly for the computational issues.

Funding: This research received no external funding except basic financial support from RRI—Reutlingen Research Institute, Reutlingen University. I appreciate the support for the advancement of scientific research and the betterment of society for the future.

Data Availability Statement: All data are available in the paper or upon request from the author.

Acknowledgments: I thank three anonymous reviewers for helpful comments.

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

References

1. Padhi, S.; Graef, J.; Pati, S. Multiple Positive Solutions for a boundary value problem with nonlinear nonlocal Riemann–Stieltjes Integral Boundary Conditions. *Fract. Calc. Appl. Calc.* **2018**, *21*, 716–745. [\[CrossRef\]](#)
2. Ruiz, W. Dynamical system method for investigating existence and dynamical property of solution of nonlinear time-fractional PDEs. *Nonlinear Dyn.* **2019**, *99*, 1–20.
3. Kamran, J.W.; Jamal, A.; Li, X. Numerical Solution of Fractional-Order Fredholm Integro-differential Equation in the Sense of Atangana–Baleanu Derivative. *Math. Probl. Eng.* **2020**, *2021*, 6662808.
4. Guariglia, E. Fractional calculus, zeta functions and Shannon entropy. *Open Math.* **2021**, *19*, 87–100. [\[CrossRef\]](#)
5. Mandelbrot, B.; van Ness, J. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Rev.* **1968**, *10*, 422–437. [\[CrossRef\]](#)
6. Monin, A.; Yaglom, A. *Statistical Fluid Mechanics: Mechanics of Turbulence*; Dover Publication: New York, NY, USA, 2007; Volume 2.
7. Tudor, C. On the Wiener integral with respect to sub-fractional Brownian motion on an interval. *J. Math. Anal. Appl.* **2009**, *351*, 456–468. [\[CrossRef\]](#)
8. Tudor, C.; Zili, M. Covariance measure and stochastic heat equation with fractional noise. *Fract. Calc. Appl. Anal.* **2014**, *17*, 807–826. [\[CrossRef\]](#)

9. Bojdecki, T.; Gorostiza, L.; Talarczyk, A. Sub-fractional Brownian motion and its relation to occupation times. *Statist. Probab. Lett.* **2004**, *69*, 405–419. [[CrossRef](#)]
10. Duncan, T.; Hu, Y.; Pasik-Duncan, B. Stochastic Calculus for Fractional Brownian Motion. *SIAM J. Control Optim.* **2000**, *38*, 582–612. [[CrossRef](#)]
11. Shen, G.; Yan, L. The stochastic integral with respect to the sub-fractional Brownian motion with $H > \frac{1}{2}$. *J. Math. Sci. Sci. Adv.* **2010**, *6*, 219–239.
12. Yan, L.; Shen, G.; He, K. Itô's formula for a sub-fractional Brownian motion. *Commun. Stoch. Anal.* **2011**, *5*, 135–159. [[CrossRef](#)]
13. Liu, J.; Yan, L. Remarks on asymptotic behavior of weighted quadratic variation of subfractional Brownian motion. *J. Korean Stat. Soc.* **2012**, *41*, 177–187. [[CrossRef](#)]
14. Prakasa, R. On some maximal and integral inequalities for sub-fractional Brownian motion. *Stoch. Anal. Appl.* **2017**, *35*, 2017.
15. Herzog, B. Adopting Feynman–Kac Formula in Stochastic Differential Equations with (Sub-)Fractional Brownian Motion. *Mathematics* **2022**, *10*, 340. [[CrossRef](#)]
16. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [[CrossRef](#)] [[PubMed](#)]
17. Bottou, L.; Curtis, F.E.; Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.* **2018**, *60*, 223–311. [[CrossRef](#)]
18. Kochenderfer, H.; Wheeler, T. *Algorithms for Optimization*; MIT Press: Cambridge, MA, USA, 2019.
19. Schmidt, M.; Roux, N.L.; Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Program.* **2017**, *162*, 83–112. [[CrossRef](#)]
20. Haochen, J.; Sra, S. Random Shuffling Beats SGD after Finite Epochs. *Proc. Mach. Learn. Res.* **2019**, *97*, 2624–2633.
21. Gotmare, A.; Keskar, N.S.; Xiong, C.; Socher, R. A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation. *arXiv* **2018**, arXiv:1810.13243.
22. Curtis, F.E.; Scheinberg, K. Adaptive Stochastic Optimization: A Framework for Analyzing Stochastic Optimization Algorithms. *IEEE Signal Process* **2020**, *37*, 32–42. [[CrossRef](#)]
23. de Roos, F.; Jidling, C.; Wills, A.; Schön, T.; Hennig, P. A Probabilistically Motivated Learning Rate Adaptation for Stochastic Optimization. *arXiv* **2021**, arXiv:2102.10880.
24. Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386–408. [[CrossRef](#)] [[PubMed](#)]
25. Alòs, E.; Mazet, O.; Nualart, D. Stochastic Calculus with Respect to Gaussian processes. *Ann. Probab.* **2001**, *29*, 766–801. [[CrossRef](#)]
26. Calin, O. *Deep Learning Architectures*; Springer: Berlin/Heidelberg, Germany, 2020.
27. Golub, G.H.; van Loan, C. *Matrix Computations*; Johns Hopkins Press: Baltimore, MD, USA, 1996.
28. Ornstein, L.; Uhlenbeck, G. On the theory of Brownian motion. *Phys. Rev.* **1930**, *36*, 823–841.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.