fractal and fractional

Article

Accelerated Gradient Descent Driven by Lévy Perturbations

Yuquan Chen [1,*], Zhenlong Wu [2], Yixiang Lu [3], Yangquan Chen [4] and Yong Wang [5]

[1] Department of Automation, Hohai University, Nanjing 210024, China
[2] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China; wuzhenlong2020@zzu.edu.cn
[3] Anhui Engineering Laboratory of Human Robot Integration System and Equipment, School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China; lyxahu@ahu.edu.cn
[4] School of Engineering, University of California, Merced, CA 95343, USA; ychen53@ucmerced.edu
[5] Department of Automation, University of Science and Technology of China, Hefei 230026, China; yongwang@ustc.edu.cn
* Correspondence: cyq@mail.ustc.edu.cn

Abstract: In this paper, we mainly consider two kinds of perturbed accelerated gradient descents driven by Lévy perturbations, which is of great importance for enhancing the global search ability. By using Lévy representation, Lévy perturbations can be divided into two parts: small jumps and large jumps, whose properties are then carefully discussed. By introducing the concept of attraction domain for local minima, Makovian transition properties are proven for the proposed two perturbed accelerated gradient descents with different infinitesimal matrices. Finally, all the results are extended to the vector case and two simulation examples are provided to validate all the conclusions.

Keywords: accelerated gradient descent; Lévy perturbations; global optimization; Markovian transition property

Citation: Chen, Y.; Wu, Z.; Lu, Y.; Chen, Y.; Wang, Y. Accelerated Gradient Descent Driven by Lévy Perturbations. *Fractal Fract.* **2024**, *8*, 170. https://doi.org/10.3390/fractalfract8030170

Academic Editor: Vassili Kolokoltsov

Received: 17 January 2024
Revised: 1 March 2024
Accepted: 12 March 2024
Published: 14 March 2024

1. Introduction

Nowadays, optimization problems can be found almost everywhere, for instance, (1) Modeling: minimize some given cost function to find the optimal parameters for describing a system [1]; (2) Optimal control: find the optimal controller parameters for different goals, such as most energy-saving, most time-saving, or the fastest response rate [2]; (3) Machine learning: an important task of machine learning is to train the neural network to find a set of parameters that minimize the mismatching rate [3].

The key to an optimization problem is the optimization algorithm. Among all the existing optimization algorithms, gradient descent (GD) [4,5] is a basic but popular optimization algorithm and plays an important role in all kinds of problems. Moreover, to overcome the shortcomings of GD, many variants have been proposed. For instance, accelerated GDs (AGDs) including GD with momentum [6,7] and Nesterov acceleration [8], have been proposed to increase the convergence speed. Unlike AGD which is accelerated by introducing one more variable, second-order algorithms such as Newton method and quasi-Newton method [9], increase the convergence speed by modifying the iteration direction. Moreover, it has been shown that AGDs can be described by a second-order transfer function and its properties can be derived using system theory [7,10]. Some methods for estimating the Hessian matrix have also been provided to make Newton method more applicable [11,12].

As is known, GD is a local algorithm since gradient information is local, for which GD can easily fall into a local minimum point and it will be quite tough for escaping from a saddle point. To increase the global search ability of GD for non-convex optimization, perturbed GDs (PGDs) have then been proposed. The PGD driven by Brownian noise can be viewed as the discrete form of the Langevin equation in physics [13], which is used to interpret particle's transition between different potential wells [14]. Each potential well plays the same role as an attraction domain of a local minimum point in optimization, thus

PGD can help jumping out a local minimum and existing PGD driven by Brownian motion has been proven to jump out a local minimum but the first exist time depends exponentially on the depth of the well [15]. Moreover, PGD can also help escaping the saddle point and the property for faster escaping from saddle point was carefully discussed in [16,17]. To help jumping out the local minima more efficiently, the perturbation is then replaced by the Lévy perturbations, which is a heavy-tailed distribution and has been proven to be efficient in global searching for intelligent algorithms. Compared with the Brownian motion, the variance of Lévy perturbations is infinite and more frequently large jump sizes will take place, which contributes to jumping out the local minima a lot [18]. In [19,20], dedicated analyses were given to the simulated annealing driven by Lévy perturbations by dividing the Lévy perturbations into large jump sizes and small jump sizes. In [21], it is proven that the first exit time for escaping from a local minimum point driven by Lévy perturbations is polynomial.

Stochastic GD (SGD) was originally proposed to increase the training speed in deep learning, where it uses a random mini-batch instead of the whole data to generate an update [22,23]. SGD has the same format of the commonly used GDs in convex optimization and looks different from PGD, but they are the same in essence. Ref. [24] provided the detailed deduction for the relation between SGD and PGD by assuming that the error of the mini-batch gradient and the real gradient was distributed according to some distribution. Recently, in [24], it is found that the stochastic noise can be better described with a Lévy distribution. Moreover, in [25,26], it is found that the first exist time from a local minimum is only polynomial with the width of the well, which is conflict to the existing results. Thus GD driven by Lévy perturbations is then get increasing attention. In [27], the Langevin Monte Carlo driven by Lévy perturbations was carefully discussed and by carefully designing the drift item, the convergence to the invariant measure was proven. On this basis, in [28], the results were extended to the overdamped Langevin Monte Carlo, where an additional momentum item was included to accelerate the convergence speed to the invariant measure.

AGD has already been widely used in deep learning. Similar to the procedures provided in [24] where stochastic GD is reformulated as the conventional PGD, accelerated PGD (PAGD) can be derived. There have already been some work about the properties of PAGD driven by Brownian motion [29]. However, properties of PAGD driven by Lévy fligths have not been analyzed yet, which is of great importance in deep learning [24]. Therefore, in this study, two different types of PAGDs will be proposed and their corresponding continuous case will also be given. Divide Lévy perturbations into two parts: small jumps and large jumps and their properties are then carefully discussed. On this basis, by introducing the concept of attraction domain for local minima, Makovian transition properties will be proven for the proposed PAGDs with different infinitesimal matrices. Finally, two simulation examples are provided to validate all the conclusions. The main contribution of the paper is summarized as follows

- By dividing the Lévy perturbation into small jumps and large jumps, a general framework for analyzing PAGDs is given;
- Convergence performance of PAGDs are analyzed under small perturbations and large perturbations respectively;
- By introducing the concept of attraction domain for local minima, Makovian transition properties are proven for PAGDs.

2. Preliminaries

In this paper, consider the following unconstrained nonconvex optimization problem

$$\min_{x\in\mathbb{R}^n} f(x)$$

where $f(x)$ is differentiable and has multi local minimum points $\{m_i\}_{i=1}^r$.

2.1. Accelerated Gradient Descent

The commonly used AGD with momentum item [10] can be formulated as

$$\begin{cases} y_{k+1} = \lambda y_k - \rho \nabla f(x_k), \\ x_{k+1} = x_k + y_{k+1}, \end{cases} \tag{1}$$

and its corresponding continuous form [30] is

$$\begin{cases} dy = (\lambda - 1)ydt - \rho \nabla f(x)dt, \\ dx = ydt, \end{cases} \tag{2}$$

where $0 < \lambda < 1$, $\rho > 0$ is the step size, and $\nabla f(x_k)$ denotes the gradient of $f(x)$ at x_k. In all the following, it is assumed that $f(x)$ satisfies the following condition locally

$$\mu \|x - y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{3}$$

where $\mu > 0$ and $L > 0$.

2.2. Perturbed Gradient Descent

The PGD can be derived by adding a perturbation item

$$x_{k+1} = x_k - \rho \nabla f(x_k) + \varepsilon \eta_k, \tag{4}$$

where, $\varepsilon > 0$ is the scaling parameter and η_k is the perturbation generated from some given distribution, such as Gaussian distribution or Lévy distribution introduced in the following. PGD has played an important role in interpreting the excellent performance of stochastic GD in machine learning, where the target function is always non-convex [24].

2.3. Stable Process and Lévy Perturbations

Lévy distribution, a symmetric unbiased stable distribution, $F(x)$ is defined by its characteristic function [31]

$$F(k) = e^{-\gamma |k|^{\alpha}}, 0 < \alpha < 2, \tag{5}$$

where $\gamma > 0$ is the scaling parameter and α is called Lévy index.

A symmetric stable process $L_t, t \geq 0$ is a Markov process with independent stationary increments and marginal with a Lévy distribution. Stable process L_t can be described by its characteristic function [32]

$$E\left\{e^{ikL_t}\right\} = e^{-at|k|^{\alpha}}, 0 < \alpha < 2, \tag{6}$$

where $a > 0$ is the scaling parameter. The characteristic function of L_t can be formulated in a integral form as

$$E\left\{e^{ikL_t}\right\} = \exp\left\{t \int\limits_{\mathbb{R}\backslash\{0\}} \left[e^{iky} - 1 - iky\mathbf{I}\{|y| \leq 1\}\right] \frac{dy}{|y|^{1+\alpha}}\right\}. \tag{7}$$

where $\mathbf{I}\{|y| \leq 1\}$ is the index function. It is concluded that Lévy process is a compound Poisson process and the Lévy measure of the stochastic process L_t is given by

$$v(A) = \int\limits_{A\backslash\{0\}} \frac{dy}{|y|^{1+\alpha}}. \tag{8}$$

Lemma 1 ([20])**.** *Define $dL_t = L_t - L_{t-}$ as the jump size of L_t at time t, and the number of jumps on the time interval $(0, t]$ whose jump size belongs to the Borel set A has a Poisson distribution with mean $tv(A)$.*

Remark 1. *When $\alpha = 2$, Lévy distribution and stable process L_t will reduce to a Gaussian distribution and Brownian motion, respectively. Ref. [18] provided simple procedures to generate random numbers according to Lévy distribution.*

According to Lemma 1 and the fact that a Poisson process can be divided into several independent Poisson processes, we will divide the Lévy flights into two parts: one is the large jump size where $|dL_t| = |\eta_k| > \kappa$ and the other one is the small jump size where $|dL_t| = |\eta_k| \leq \kappa$.

For the large jump size, the Lévy measure is finite and one has that

$$\beta_\kappa = \int_{|x|\geq\kappa} \frac{dy}{|y|^{1+\alpha}} = \frac{2}{\alpha\kappa^\alpha}. \tag{9}$$

Denote τ_k and W_k as the jump arrival times and jump sizes. Then the internal-arrival times $T_k = \tau_k - \tau_{k-1}$ are identically independent and exponentially distributed with mean β_κ^{-1}. According to the property of Poisson process, it is known that thr probability density function of τ_k satisfies a Gamma distribution

$$f(\beta_\varphi, k) = \frac{\beta_\varphi \mathrm{e}^{-\beta_\varphi t}(\beta_\varphi t)^{k-1}}{(k-1)!}. \tag{10}$$

Moreover, perturbations in between two successive large step sizes are all small jump sizes (bounded by κ).

Remark 2. *The above analyses indicate that Lévy index could control both the frequency and amplitude of large jumps. The smaller Lévy index, the more frequent large jumps and the larger amplitude.*

3. PAGDs Driven by Lévy Perturbations

Combining the AGD with a perturbation item, one can derive two kinds of PAGDs respectively as

$$\begin{cases} y_{k+1} = \lambda y_k - \rho\nabla f(x_k), \\ x_{k+1} = x_k + y_{k+1} + \varepsilon\eta_k, \end{cases} \tag{11}$$

and

$$\begin{cases} y_{k+1} = \lambda y_k - \rho\nabla f(x_k) + \varepsilon\eta_k, \\ x_{k+1} = x_k + y_{k+1}. \end{cases} \tag{12}$$

where the perturbation $\varepsilon\eta_k$ is added to different positions.

For the convenience, the following property analyses will mainly based on their continuous forms and Lévy perturbations are used, which can be described respectively as

$$\begin{cases} dy = (\lambda-1)ydt - \rho\nabla f(x)dt, \\ dx = ydt + \varepsilon dL_t, \end{cases} \tag{13}$$

and

$$\begin{cases} dy = (\lambda-1)ydt - \rho\nabla f(x)dt + \varepsilon dL_t, \\ dx = ydt. \end{cases} \tag{14}$$

Remark 3. *PGD (4) is the discrete form of the overdamped Langevin equation while PAGD (14) is known as the underdamped Langevin equation. We have to mention that PAGD (11) is newly proposed in this study. One may refer the work of [30] for more details about the discrete and continuous forms of AGDs.*

Remark 4. *In neural network training, stochatic GD is often used and it has been shown in [24] that the gradient estimation error can be better modeled by a Lévy perturabtion. In the following, we will focus on the property analyses of PAGDs driven by Lévy perturabtion.*

Convergence Analyses for Small Perturbations

Theorem 1. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, the trajectories of PAGDs (13) and (14) will both converge to that of AGD (2) for small perturbations bounded by κ.*

Proof. The solution of differential Equation (2) is

$$\bar{x}(t) = x_0 + \int_0^t \left[e^{(\lambda-1)T} y_0 - \rho \int_0^T e^{(\lambda-1)(T-\tau)} \nabla f(\bar{x}) d\tau \right] dT, \tag{15}$$

and the solution of differential Equation (14) is

$$x(t) = x_0 + \int_0^t \left[e^{(\lambda-1)T} y_0 - \rho \int_0^T e^{(\lambda-1)(T-\tau)} \nabla f(x) d\tau \right] dT + \varepsilon L_t. \tag{16}$$

Then the absolute error between $x(t)$ and $\bar{x}(t)$ is

$$\begin{aligned} \|\bar{x}(t) - x(t)\| &= \|\rho \textstyle\int_0^t \int_0^T e^{(\lambda-1)(T-\tau)} [\nabla f(x) - \nabla f(\bar{x})] d\tau dT + \varepsilon L_t\| \\ &\le \rho L \textstyle\int_0^t \int_0^T e^{(\lambda-1)(T-\tau)} \|\bar{x}(\tau) - x(\tau)\| d\tau dT + \|\varepsilon L_t\| \\ &= \rho L \textstyle\int_0^t \|\bar{x}(\tau) - x(\tau)\| \int_\tau^t e^{(\lambda-1)(T-\tau)} dT d\tau + \|\varepsilon L_t\| \\ &\le \tfrac{\rho L}{1-\lambda} \textstyle\int_0^t \|\bar{x}(\tau) - x(\tau)\| d\tau + \|\varepsilon L_t\|, \end{aligned}$$

where condition (3) is used.

As $\varepsilon \to 0$, one has that $\|\varepsilon d L_t\| \le \varepsilon^{1-\gamma} \to 0$ and

$$\|\bar{x}(t) - x(t)\| \le \frac{\rho L}{1-\lambda} \int_0^t \|\bar{x}(\tau) - x(\tau)\| d\tau.$$

Using Gronwall's inequality [33], we have that

$$\|\bar{x}(t) - x(t)\| \le 0,$$

which indicates the results of the given theorem.

For the differential Equation (14), the result still holds and the proof procedures are similar. □

Remark 5. *Theorem 1 indicates that if ε is set sufficiently small, the small perturbations (bounded by κ) will not influence the convergence trajectory a lot. In other words, small perturbations in Lévy perturbations will not enhance the global search ability while the large perturbations will do. Therefore, we will focus on the analyses on large perturbations in the following.*

Assumption 1. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, it is assumed that the mean waiting time of successive large jumps is much longer than the convergence time of AGD (2) to the ε-neighbourhood of a local minimum point i.e., $\|x(t) - x^*\| \le \varepsilon$.*

According to aforementioned discussion, perturbations in between two successive large jumps are all smaller than κ. Then the convergence time to a local minimum point of the conventional GD from its corresponding attraction domain could be derived. Choose Lyapunov function $V = \|x - x^*\|^2$ and take first-order time derivative, yielding,

$$\dot{V} = -2\rho \nabla^{\mathrm{T}} f(x)(x - x^*) \le -2\rho\mu \|x - x^*\|^2 = -2\rho\mu V.$$

where condition (3) is used. It is concluded that

$$\|x - x^*\|^2 \le \|x_0 - x^*\|^2 \mathrm{e}^{-2\rho\mu t}. \tag{17}$$

Therefore, if $t \ge \frac{1}{2\rho\mu} \ln\left(\frac{\|x_0 - x^*\|^2}{\varepsilon} \right)$, conventional GD will converge to the ε-neighbourhood of local minimum point. Besides, the mean waiting time of two successive large jumps is

$\frac{\alpha}{2\varepsilon^{\gamma\alpha}}$. As $\varepsilon \to 0$. it is concluded that $\frac{\alpha}{2\varepsilon^{\gamma\alpha}}$ is much larger than $\frac{1}{2\rho\mu} \ln\left(\frac{\|x_0 - x^*\|^2}{\varepsilon}\right)$. Since AGD converges faster than the conventional GD by carefully tuning the parameters, Assumption 1 can then be shown reasonable.

Remark 6. *Assumption 1 indicates that if ε is set sufficiently small, both PAGDs (13) and (14) will fall into a ε-neighbourhood of a local minimum point, which is important for the following analysis.*

4. Convergence Properties for PAGD (13)

In this section, the properties of two PAGDs will be studied. Before moving, the attraction domain of a local minimum point must be defined. For a given target function $f(x)$, if the optimization algorithm is determined, then it can be viewed as a nonlinear feedback system, where a local minimum point is a steady equilibrium of the system. By using the definition of attraction domain in nonlinear systems, the attraction domain of a local minimum point can be defined.

Definition 1. *The attraction domain of a local minimum point m_i is the collection of all the initial states such that the optimization algorithm finally converges to the local minimum point m_i.*

For AGD (2) where $f(x)$ has multi local minimum points $\{m_i\}_{i=1}^r$, it has r steady equilibriums, i.e., $(x, y)^T = (m_i, 0)^T$ whose corresponding attraction domain are defined as Ω_i.

4.1. Convergence Properties for PAGD (13)

In this subsection, PAGD (13) is considered, where the perturbation is added to x. Then define a new state space $\{\Omega_i\}_{i=1}^r$ and denote σ_{ij} as the first time to jump from attraction domain Ω_i to attraction domain Ω_j.

According to Lemma 1, it is concluded that $(x, y)^T$ has converged to a small neighbourhood of $(m_i, 0)^T$ before the next large jump arrives and then the transition can be approximated by transition from $(m_i, 0)^T$ to attraction domain Ω_j. Then according to the the fact that the arrival of large perturbations constructs a Poisson process, the mean of σ_{ij} can be calculated as

$$\begin{aligned} E\{\sigma_{ij}\} &= \sum_{k=1}^{\infty} E\{\tau_k\} P(\sigma_{ij} = \tau_k) \\ &= \sum_{k=1}^{\infty} k E\{T_1\} P(\sigma_{ij} = \tau_k) \\ &\approx \frac{P\big((m_i + \varepsilon W_k, 0) \in \Omega_j\big)}{\beta_\varphi} \sum_{k=1}^{\infty} k[1 - P((m_i + \varepsilon W_1, 0) \notin \Omega_i)]^{k-1} \\ &= \frac{P\big(m_i + \varepsilon W_1 \in \Omega_j\big)}{\beta_\varphi P^2(m_i + \varepsilon W_1 \notin \Omega_i)}. \end{aligned} \tag{18}$$

where the definitions and properties of W_k and τ_k are given in Section 2.3, and

$$\begin{aligned} P(\sigma_{ij} = \tau_k) &\approx P\left(\bigcap_{i=1}^{k-1} (m_i + \varepsilon W_1, 0)^T \in \Omega_i, (m_i + \varepsilon W_k, 0) \in \Omega_j\right) \\ &= (1 - P((m_i + \varepsilon W_1, 0) \notin \Omega_i))^{k-1} P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big). \end{aligned}$$

Besides, the probability distribution function of σ_{ij} can be calculated as

$$
\begin{aligned}
P\big(\sigma_{ij} > u\big) &= \sum_{k=1}^{\infty} P(\tau_k > u) P\big(\sigma_{ij} = \tau_k\big) \\
&= \sum_{k=1}^{\infty} P(\tau_k > u) P\big(\sigma_{ij} = \tau_k\big) \\
&= \sum_{k=1}^{\infty} \int_u^{\infty} \frac{\beta_\varphi \mathrm{e}^{-\beta\varphi t}\big(\beta_\varphi t\big)^{k-1}}{(k-1)!} \mathrm{d}t P\big(\sigma_{ij} = \tau_k\big) \\
&\approx \beta_\varphi P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big) \int_u^{\infty} \sum_{k=1}^{\infty} \frac{\mathrm{e}^{-\beta\varphi t}\big(\beta_\varphi t\big)^{k-1} P_{\bar{\Omega}_i}^{k-1}}{(k-1)!} \mathrm{d}t \\
&= \frac{P\big((m_i+\varepsilon W_1,0)\in\Omega_j\big)}{P((m_i+\varepsilon W_1,0)_1 \notin \Omega_i)} \exp\{-u\beta_\varphi P((m_i + \varepsilon W_1, 0) \notin \Omega_i)\},
\end{aligned}
\tag{19}
$$

with $P_{\bar{\Omega}_i} = 1 - P((m_i + \varepsilon W_1, 0) \notin \Omega_i)$. Equation (19) indicates that σ_{ij} is approximately distributed to an exponential function with mean $E\{\sigma_{ij}\}$.

Theorem 2. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, the transition of $x(t)$, the solution of PAGD (11), among the state space $\{\Omega_i\}_{i=1}^{r}$ constructs an approximate continuous Markov chain whose infinitesimal matrix is $Q = \{q_{ij}\}_{i,j=1}^{m}$ with*

$$
q_{ij} := \beta_\varphi P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big), i \neq j,
$$

$$
q_{ii} = -\sum_{j \neq i} q_{ij}.
$$

Proof. According to Theorem 1, the influence of small perturbations can be totally ignored. Moreover, according to Asssumption 1, if $x(t)$ falls into Ω_i, it will soon converge to its corresponding equilibrium point $(m_i, 0)^T$ before the next large jump arrives. The distribution of σ_{ij} is calculated in (19), which follows a exponential distribution. On this basis, for $i \neq j$, q_{ij} in infinitesimal matrix can be derived as

$$
\lim_{u \to 0} \frac{\mathrm{d}}{\mathrm{d}u}\big(1 - P\big(\sigma_{ij} > u\big)\big) = \beta_\varphi P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big),
$$

This completes the proof. □

4.2. Convergence Properties for PAGD (14)

In this subsection, PAGD (14) is considered, where the perturbation is added to y. Similar to the analyses for PAGD (14), denote σ_{ij} as the first time to jump from attraction domain Ω_i to attraction domain Ω_j. One can then calculate the mean of σ_{ij} and probability distribution respectively as

$$
E\{\sigma_{ij}\} = \frac{P\big((m_i, \varepsilon W_1) \in \Omega_j\big)}{\beta_\varphi P^2((m_i, \varepsilon W_1) \notin \Omega_i)}
\tag{20}
$$

and

$$
P\big(\sigma_{ij} > u\big) = \frac{P\big((m_i, \varepsilon W_1) \in \Omega_j\big)}{P((m_i, \varepsilon W_1) \notin \Omega_i)} \exp\{-u\beta_\varphi P((m_i, \varepsilon W_1) \notin \Omega_i)\}
\tag{21}
$$

Remark 7. *The analyses for PAGD (11) and (12) are quite similar. Compare (19) and (21), it is found that the transition probability from $(m_i, 0)^T$ to attraction domain Ω_j for PAGD (11) is $P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big)$ while it is $P\big((m_i, \varepsilon W_1) \in \Omega_j\big)$ for PAGD (12), since the perturbations are added to different positions of PAGD (11) and (12).*

Theorem 3. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, the transition of $x(t)$, the solution of PAGD (11), among the state space $\{\Omega_i\}_{i=1}^{r}$ constructs an approximate continuous Markov chain whose infinitesimal matrix is $Q = \{q_{ij}\}_{i,j=1}^{m}$ with*

$$q_{ij} := \beta_{\varphi} P\big((m_i, \varepsilon W_1) \in \Omega_j\big), i \neq j,$$

$$q_{ii} = -\sum_{j \neq i} q_{ij}.$$

4.3. Conclusive Remarks

- The results here can be extended to many other kinds of optimization algorithms, such as the conventional PGD driven by Lévy perturbations. Moreover, our results will be more precise than that in [19] by using the attraction domain as the state space rather than using the local minimum points.
- For conventional PGD, the attraction domain for a local minimum point of $f(x)$ is the interval constructed by its two adjacent maximum points. Then the probability of $P\big((m_i + \varepsilon W_1, 0) \in \Omega_j\big)$ and $P\big((m_i, \varepsilon W_1) \in \Omega_j\big)$ can be calculated as shown in [19].
- Here, ε is set sufficiently small to guarantee the establishment of Theorem 1 and Assumption 1. For practical usage, one can set the small jumps as zero to derive the truncated Lévy perturbations and guarantee the establishment of Theorem 1. If κ is chosen such that β_{κ}^{-1} is much larger than the convergence time to a small neighbourhood of a local minimum point, then the results in Theorems 2 and 3 still hold for PAGDs driven by truncated Lévy perturbations.

5. Extensions to Vector Case

The aforementioned results only hold for scalar case. By using the vector form of Lévy process, the aforementioned analyses can be directly extended to the vector case. According to the presented heavy-tailed noise in [28], it is shown that the norm of the gradient noise is heavy-tailed. Therefore, according to Theorem 6.17 in [32], the characteristic function of the vector Lévy process can be formulated as follows

$$E\Big\{e^{i\langle k,L_t\rangle}\Big\} = \exp\Big\{t\int_{\mathbb{R}^n\setminus\{0\}}\Big[e^{i\langle k,y\rangle} - 1 - i\langle k,y\rangle\mathbf{I}\{\|y\| \leq 1\}\Big]v(dy)\Big\}.$$

Since the norm of the gradient noise is heavy-tailed, the Lévy measure can be described as the mentioned measure (8) multiplied by an uniform direction stochastic variable, which can be formulated as

$$v(y\theta : y \in A, \theta \in B) = \int_{\theta \in B}\int_{y \in A\setminus\{0\}} \frac{dy}{|y|^{1+\alpha}} M(d\theta),$$

and $M(d\theta)$ denotes the measure of θ.

The analyses for the vector case is almost the same as the scalar case and the main difference concentrates on using $W_k\Theta_k$ to replace W_k for the jump size, where W_k and Θ_k are two independent stochastic variables (W_k denotes the jump size and Θ_k denotes the jump direction) with

$$P(W_1 \in A) = \frac{1}{\beta}\int_{y \in A\setminus\{0\}} \frac{dy}{|y|^{1+\alpha}}$$

and

$$P(\Theta_k \in A) = M(A).$$

Then define σ_{ij} as the first exit time from attraction domain Ω_i to Ω_j and following equations hold

$$E\{\sigma_{ij}\} = \frac{P\big((m_i, \varepsilon W_1\Theta_1) \in \Omega_j\big)}{\beta P^2\big((m_i, \varepsilon W_1\Theta_1) \notin \Omega_i\big)},$$

and

$$P(\sigma_{ij} > u) = \frac{P\big((m_i, \varepsilon W_1\Theta_1) \in \Omega_j\big)}{P\big((m_i, \varepsilon W_1\Theta_1) \notin \Omega_i\big)} \exp\{-u\beta P((m_i, \varepsilon W_1\Theta_1) \notin \Omega_i)\}.$$

Define a new state space $\{\Omega_i\}_{i=1}^r$ where Ω_i denotes the attraction domain for local minimum point m_i, and follow theorem for vector PAGD (14) can similarly derived.

Theorem 4. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, the transition of $x(t)$, the solution of PAGD (14), among the state space $\{\Omega_i\}_{i=1}^r$ constructs an approximate continuous Markov chain whose infinitesimal matrix is $Q = \{q_{ij}\}_{i,j=1}^m$ with*

$$q_{ij} := \beta_\kappa P\big((m_i, \varepsilon W_1\Theta_1) \in \Omega_j\big), i \neq j,$$

$$q_{ii} = -\sum_{j\neq i} q_{ij}.$$

Furthermore, the transition property for PAGD (13) can be similarly derived as follows.

Theorem 5. *Set $\kappa = \varepsilon^{-\gamma}$, $0 < \gamma < 1$. As $\varepsilon \to 0$, the transition of $x(t)$, the solution of PGD (13), among the state space $\{\Omega_i\}_{i=1}^r$ constructs an approximate continuous Markov chain whose infinitesimal matrix is $Q = \{q_{ij}\}_{i,j=1}^m$ with*

$$q_{ij} := \beta_\kappa P\big((m_i + \varepsilon W_1\Theta_1, 0) \in \Omega_j\big), i \neq j,$$

$$q_{ii} = -\sum_{j\neq i} q_{ij}.$$

Remark 8. *Compared with the scalar case, the convergence property performs almost the same, where the main difference is a direction variable Θ is introduced. Moreover, the large jump size can help jumping out a local minimum point, which can be viewed as re-initialization. We have to declare the proposed expression can better captured the characteristic of the gradient noise since it has been shown that the norm of the gradient noise is heavy-tailed in [28].*

6. Illustrative Examples

In this section, some simulation examples are provided to validate all the conclusions.

Example 1. *Consider function with two local minima, which is formulated as*

$$f(x) = x^2 - 10\cos(x). \tag{22}$$

Set the Lévy index $\alpha = 1.5$, step size $\rho = 0.01$, scaling parameter $\varepsilon = 0.01$, and $\lambda = 0.8$. Results are shown in Figures 1 and 2, where case 1 and case 2 indicates using PAGD (11) and PAGD (12) respectively. Then we have the following observations

- *The transition among different local minima always happens when the large jump arrives, and we have labeled some typical transition points in Figure 2, which shows the main contribution of large jumps for improving global search ability.*
- *For different PAGDs where perturbations are added to different positions, different transitions can be viewed with the same perturbations, where we have also labeled in Figure 2.*
- *As labeled in Figure 1, the arrival time of two successive large jumps can be very close. Therefore, Assumption 1 only holds in the mean of expectation and we can try to design the arrival time of large jumps in the future.*

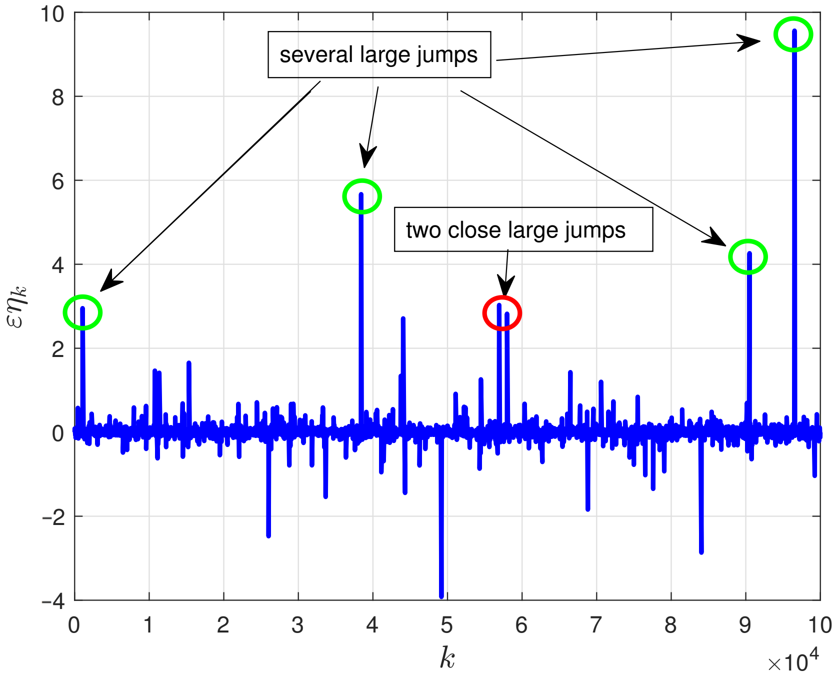

Figure 1. Lévy perturbations with order $\alpha = 1.5$.

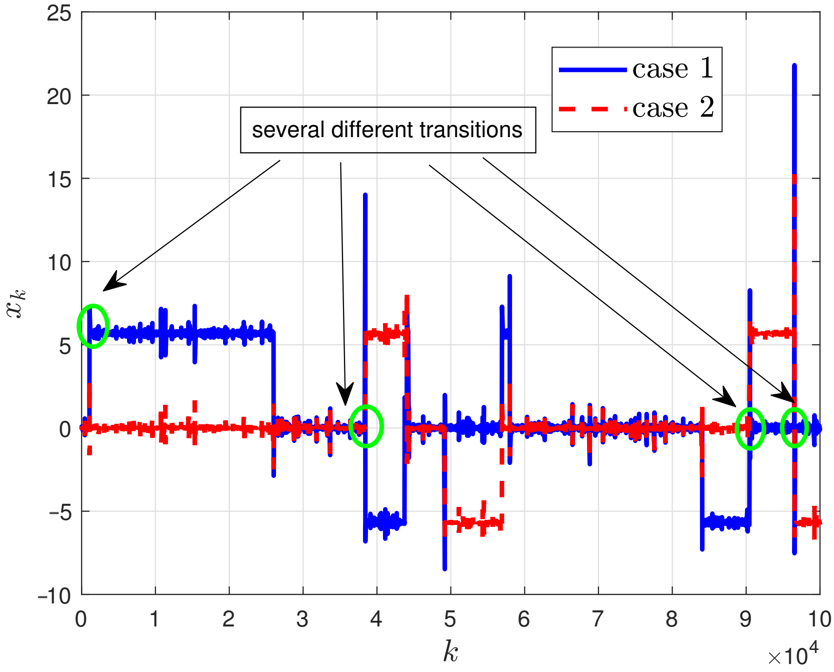

Figure 2. Comparison of PAGD (11) and PAGD (12).

Example 2. *In this example, we will compare results of PAGD (11) driven by Lévy perturbations and truncated Lévy perturbations. Consider function with multi local minima, which is formulated as*

$$f(x) = x^2 - 100\cos(x). \tag{23}$$

All the parameters are the same as in Example 1. The truncated Lévy perturbations are derived by truncating Lévy perturbations with $\kappa = 50$. Results are shown in Figures 3 and 4, where case 1 and case 2 indicates using PAGD Lévy perturbations and truncated Lévy perturbations respectively.

- *Figure 3 shows the comparison of Lévy perturbations and truncated Lévy perturbations (red circles) with threshold $\kappa = 50$. Moreover, it is found that transition among local minimum* *points always happens when the large jump arrives, and Markovian transition property can also be observed.*

- *Figure 4 shows that PAGD using truncated Lévy perturbations has almost the same transitions as PAGD using Lévy perturbations. Besides, PAGD (11) using truncated Lévy perturbations has a more convergence accuracy since it has no small perturbations.*

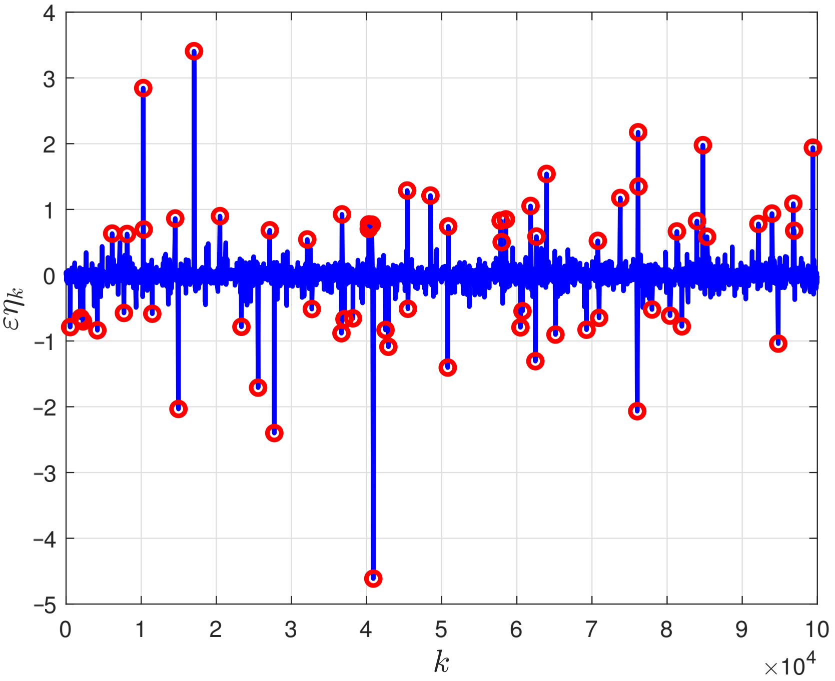

Figure 3. Lévy perturbations and truncated Lévy perturbations.

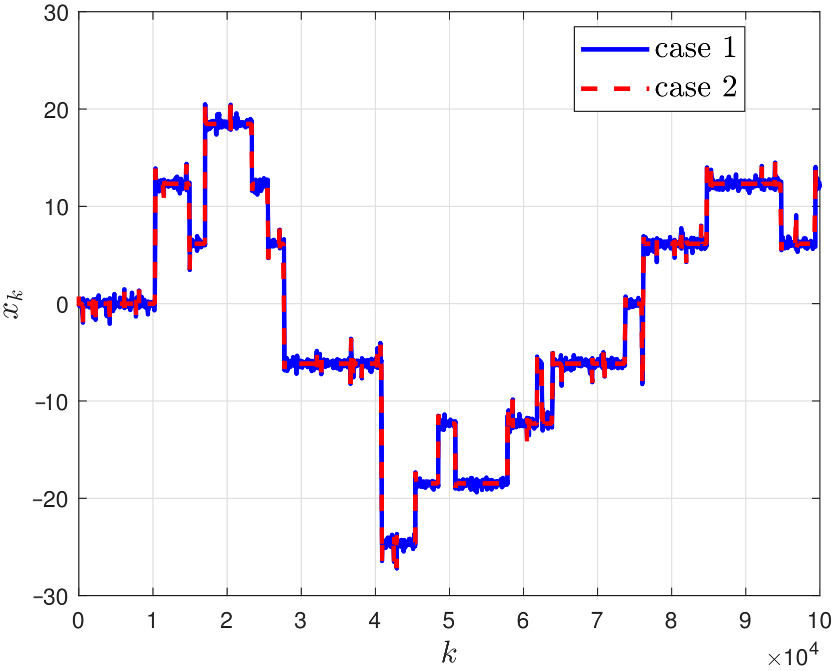

Figure 4. Comparison of PAGD (11) and truncated PAGD (11).

Remark 9. *The simulation results implies that the transitions always happen when the large jumps occure. In neural network training, a sudden performance improvement or deterioration is often found, where the algorithm jumps from a local minimum point to another one.*

7. Conclusions and Future Topics

In this study, we have presented two types of PAGDs driven by Lévy flights and analyzed their properties. By dividing Lévy perturbations into small perturbations and large perturbations, the properties of Lévy perturbations are then carefully analyzed. On this basis, the Markovian transition properties for different PAGDs are given by introducing the concept of attraction domain for local minima. The main difference for different PAGDs is concentrated on different infinitesimal matrices. All the conclusions are finally validated by simulation examples. Some promising research topics are listed as follows

- extend the results to PAGDs with adaptive Lévy index for better convergence performance;
- extend the results to the optimization algorithms driven by truncated Lévy perturbations;
- apply the proposed PAGDs in non-convex optimization problems such as neural networks training.

Author Contributions: Writing—original draft preparation, Y.C. (Yuquan Chen); Validation, Z.W. and Y.L.; Simulation, Y.C. (Yuquan Chen) and Z.W.; supervision, Y.C. (Yangquan Chen) and Y.W.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the open research subject of Anhui Engineering Laboratory of Human Robot Integration System and Equipment (No. RJGR202206), the National Nature Science Foundation of China (No. 62303158), and the "SCBS" plan of Jiangsu Province (No. JSSCBS20210243).

Data Availability Statement: No data were used to support this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Y.; Wei, Y.; Liang, S.; Wang, Y. Indirect model reference adaptive control for a class of fractional order systems. *Commun. Nonlinear Sci. Numer. Simul.* **2016**, *39*, 458–471. [CrossRef]
2. Lewis, F.L.; Vrabie, D.; Syrmos, V.L. *Optimal Control*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Amsterdam, The Netherlands, 2016.
4. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: New York, USA, 2004.
5. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
6. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151. [CrossRef]
7. Wilson, A.C.; Recht, B.; Jordan, M.I. A Lyapunov analysis of momentum methods in optimization. *arXiv* **2016**, arXiv:1611.02635.
8. Nesterov, Y.E. A method for solving the convex programming problem with convergence rate O (1/kˆ 2). *Dokl. Akad. Nauk SSSR* **1983**, *269*, 543–547.
9. Gill, P.E.; Murray, W. Quasi-Newton methods for unconstrained optimization. *IMA J. Appl. Math.* **1972**, *9*, 91–108. [CrossRef]
10. An, W.; Wang, H.; Sun, Q.; Xu, J.; Dai, Q.; Zhang, L. A PID controller approach for stochastic optimization of deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8522–8531.
11. Liu, L.; Liu, X.; Hsieh, C.J.; Tao, D. Stochastic second-order methods for non-convex optimization with inexact Hessian and gradient. *arXiv* **2018**, arXiv:1809.09853.
12. Prashanth, L.; Bhatnagar, S.; Bhavsar, N.; Fu, M.C.; Marcus, S. Random directions stochastic approximation with deterministic perturbations. *IEEE Trans. Autom. Control.* **2019**, *65*, 2450–2465. [CrossRef]
13. Stariolo, D.A. The Langevin and Fokker-Planck equations in the framework of a generalized statistical mechanics. *Phys. Lett. A* **1994**, *185*, 262–264. [CrossRef]
14. Kalmykov, Y.P.; Coffey, W.; Waldron, J. Exact analytic solution for the correlation time of a Brownian particle in a double-well potential from the Langevin equation. *J. Chem. Phys.* **1996**, *105*, 2112–2118. [CrossRef]
15. Visscher, P.B. Escape rate for a Brownian particle in a potential well. *Phys. Rev. B* **1976**, *13*, 3272. [CrossRef]
16. Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S.M.; Jordan, M.I. How to escape saddle points efficiently. In Proceedings of the International Conference on Machine Learning PMLR, Sydney, Australia, 6–11 August 2017; pp. 1724–1732.
17. Staib, M.; Reddi, S.; Kale, S.; Kumar, S.; Sra, S. Escaping saddle points with adaptive gradient methods. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 5956–5965.
18. Mantegna, R.N. Fast, accurate algorithm for numerical simulation of Lévy stable stochastic processes. *Phys. Rev. E* **1994**, *49*, 4677. [CrossRef] [PubMed]

19. Imkeller, P.; Pavlyukevich, I. Lévy flights: Transitions and meta-stability. *J. Phys. A Math. Gen.* **2006**, *39*, L237. [CrossRef]
20. Pavlyukevich, I. Lévy flights, non-local search and simulated annealing. *J. Comput. Phys.* **2007**, *226*, 1830–1844. [CrossRef]
21. Imkeller, P.; Pavlyukevich, I.; Wetzel, T. The hierarchy of exit times of Lévy-driven Langevin equations. *Eur. Phys. J. Spec. Top.* **2010**, *191*, 211–222. [CrossRef]
22. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 177–186.
23. Shamir, O.; Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Proceedings of the International Conference on Machine Learning, Miami, FL, USA, 4–7 December 2013; pp. 71–79.
24. Simsekli, U.; Sagun, L.; Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv* **2019**, arXiv:1901.06053.
25. Tzen, B.; Liang, T.; Raginsky, M. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In Proceedings of the Conference on Learning Theory PMLR, Stockholm, Sweden, 6–9 July 2018; pp. 857–875.
26. Zhang, Y.; Liang, P.; Charikar, M. A hitting time analysis of stochastic gradient langevin dynamics. In Proceedings of the Conference on Learning Theory PMLR, Amsterdam, The Netherlands, 7–10 June 2017; pp. 1980–2022.
27. Nguyen, T.H.; Simsekli, U.; Richard, G. Non-asymptotic analysis of Fractional Langevin Monte Carlo for non-convex optimization. In Proceedings of the International Conference on Machine Learning PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 4810–4819.
28. Simsekli, U.; Zhu, L.; Teh, Y.W.; Gurbuzbalaban, M. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In Proceedings of the International Conference on Machine Learning PMLR, Virtual, 13–18 June 2020; pp. 8970–8980.
29. Cheng, X.; Chatterji, N.S.; Bartlett, P.L.; Jordan, M.I. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv* **2017**, arXiv:1707.03663.
30. Li, Q.; Tai, C.; Weinan, E. Dynamics of stochastic gradient algorithms. *arXiv* **2015**, arXiv:1511.06251.
31. Samoradnitsky, G. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*; Routledge: New York, NY, USA, 2017.
32. Meerschaert, M.M.; Sikorskii, A. *Stochastic Models for Fractional Calculus*; Walter de Gruyter: Berlin, Germany 2011; Volume 43.
33. Ye, H.; Gao, J.; Ding, Y. A generalized Gronwall inequality and its application to a fractional differential equation. *J. Math. Anal. Appl.* **2007**, *328*, 1075–1081. [CrossRef]

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.