Global Optimizations by Intermittent Diffusion

Shui-Nee Chow*, Tzi-Sheng Yang† and Haomin Zhou‡

Abstract

We propose an intermittent diffusion(ID) method to find global minimizers of a given function $g : \mathbb{R}^n \to \mathbb{R}$. The main idea is to add intermittent, instead of continuously diminishing, random perturbations to the gradient flow generated by g, so that the trajectories can quickly escape from one minimizer and approach other minimizers. The associated Fokker-Planck equations for existing global optimization algorithms that use continuously diminishing random perturbations are of parabolic types. For the ID method, its Fokker-Planck equation is degenerate and alternates between hyperbolic and parabolic types. It is because of this alternation, we have a numerical algorithm which is efficient. We prove that with probability arbitrarily close to 1, one can find by using the ID algorithm, a good approximation to the global minimizer in a finite time T provided T is sufficently large. We also prove that for any given finite set of minimizers of g, any trajectories of the ID method will visit an arbitrary small neighborhood of each minimizer with positive probability. Numerical simulations show that the proposed method achieves significant improvements in terms of the time and the frequencies of visiting the global minimizers over some existing global optimization algorithms.

*School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, U.S. chow@math.gatech.edu

†Department of Mathematics, Tunghai University, Taichung 40704, Taiwan. tsyang@thu.edu.tw

‡Georgia Institute of Technology, Atlanta, GA 30332, U.S. hmzhou@math.gatech.edu, this author is partially supported by NSF Faculty Early Career Development (CAREER) Award DMS-0645266

1 Introduction

Global optimization is a classical issue appeared in numerous research fields and applications, such as inverse problems, optimal design and digital image processing, just to name a few. In this paper, we will investigate a new strategy to find global minimizers for a general objective functional

$$\min_x g(x), \quad x \in \Omega \tag{1.1}$$

where Ω is an admissible set for $x \in \mathbb{R}^n$. The objective functional $g(x)$ is often defined by an energy functional or a cost function in many applications. For examples, $g(x)$ is the double-well potential energy in the 2-phase composite material problem, the quality factor in optical spectrometer design, and the total distance in optimal path planning in unmanned vehicle navigation.

As one of the oldest applied mathematics problems, finding minimizers for (1.1) has been intensively studied. Numerous research results have been reported. We refer to some books for more details and references on this rich subject [5, 16, 19, 22]. Among the existing studies in the field, one of them is on convex optimizations in which the objective functional $g(x)$ is convex on a convex admissible set Ω. In that case, we have a rather complete theory on nearly every aspect of the problem. For instances, it is clear that there exists a unique minimizer provided that $g(x)$ is reasonable well behaved. In addition, there are many efficient numerical algorithms to find the minimizer. for example, the gradient flow given by

$$\dot{x}(t) = -\nabla g(x(t)), \tag{1.2}$$

will lead to the minimizer when t goes to infinity.

Despite the existence of extensive literature, finding global minimizers for general g, if it is possible, can still be extremely challenging, especially when the dimension of x is large and the level sets of g are complicated. One of the most notorious difficulties that any global optimization method has to face is how to escape local minimizers when the solution is trapped by them. This becomes more serious if negative gradient flow is used, because the gradient flow is guaranteed to stuck at the (possibly local) minimizer of a valley. For many problems, global minimum is reachable only if the initial guess is close enough to it. For this reason, many of the existing algorithms very much rely on good selections of initial guess. For examples, the inverse media scattering methods proposed in [3, 4] can be viewed as gradient flows for minimizing a regularized data fitting objective, and it is crucial to have a good starting point.

Among the existing global optimization methods, the celebrated Metropolis random walks [18] and simulated annealing method [6, 17] are generic stochastic based strategies to identify global optimal solutions for a broad range of discrete and continuous problems. An essential idea in simulated annealing is

to iteratively use sampling procedures which generate a new admissible state according to a probability associated with $g(x)$. The new state is compared to the concurrent best sample state, and replace it if the new state is better. After enough samples, the best sample state is considered as an approximation to the global optimal solution.

The basic idea of simulated annealing is intuitive. It has been broadly used in many problems with remarkable successes. In particular, there are many examples that simulated annealing can give reasonable good approximations to the global minimizers while other methods fail to provide anything close to the global minima. However, it is also well-known that the original simulated annealing may not be very efficient in many applications. To improve the efficiency, Szu and Hartley [20] proposed the "fast simulated annealing", which generates a new state according to the Cauchy density that has unbounded variance. Later, Ingber [15] generalized this idea and suggested a "non-local generating" of a new state in Metropolis algorithm, which is called "very fast simulated annealing". The studies in [20] heuristically provides a sufficient condition to guarantee that the state-generating is infinite often in time (i.o.t.), i.e., with probability one that any state x in $\mathbb{R}^n$ will be generated for infinitely many times. Another study for the analysis of fast and very fast simulated annealing can be found in [23].

The original simulated annealing method does not have to explicitly use the gradient information in searching for the next best samples. Later, some efforts have been devoted to use simulated annealing ideas together with the gradient flow for global optimizations, especially for objective functions g that are continuously depending on the state variables x. For example, the studies in [1] and [9], suggest to implement the idea of Metropolis algorithm by running the stochastic differential equation:

$$dx(t,\omega) = -\nabla g(x(t,\omega))dt + \sigma(t)dW(t), t \in [0,\infty], \tag{1.3}$$

where $W(t)$ is the Brownian motion in $\mathbb{R}^n$, ω is a random event (a random path) in the Wiener space of $W(t)$, and the diffusion coefficient function $\sigma(t)$ is continuously decreasing to zero. For convenience, we shall call this implementation the *continuous diminishing diffusion* (CDD) method in this paper. It is proved in [7, 9] that if $\sigma(t)$ is given by

$$\sigma(t) = c/\sqrt{\log t}, \tag{1.4}$$

for large $c > 0$, the solution of (1.3) will converge weakly to a distribution concentrated at the global minima of g. More approaches and analysis can be found in [10, 13, 14].

Inspired by simulated annealing and some recent developments in random dynamical systems, we propose a new strategy to find the global optimal solution. The main idea is to combine the advantages of gradient flow, which may

quickly lead to local minimizers, and stochastic perturbations that can promote the trajectories out of the traps set by local minimizers. More precisely, we consider $\sigma(t)$ as a piecewise smooth function (we use piecewise constants in this paper) of t with alternating positive and zero values. When $\sigma(t) = 0$, (1.3) corresponds to the gradient descent algorithm (1.2). When $\sigma(t) > 0$ the trajectory of (1.3) may jump off any local minimum with a probability controlled by $\sigma(t)$. Then we repeat such a process multiple times.

An interesting viewpoint for this new strategy can be explained by studying the probability distribution of the trajectories $x(t, \omega)$ of (1.3). Its density function $p(t, x)$ satisfies the Fokker-Planck equation,

$$p_t = \nabla \cdot (\nabla g(x) p) + \frac{1}{2}\sigma(t)^2 \Delta p. \tag{1.5}$$

When $\sigma(t)$ is positive, this is a parabolic equation. When $\sigma(t) = 0$, the equation (1.5) becomes a hyperbolic equation. We call this new method the *intermittent diffusion* (ID) method owing to the diffusion coefficient being intermittently set to zero.

By selecting $\sigma(t)$ as a discontinuous function (1.5) becomes degenerate. However, it possesses some interesting properties that are not shared by the standard Fokker-Planck equations for regular diffusion processes in which $\sigma(t)$ is taken as positive values for all time t. More precisely, in ID method, when $\sigma(t) > 0$, it is the standard diffusion process that gives a positive probability for the trajectories to go everywhere. When $\sigma(t) = 0$, the diffusion term drops from (1.5), and the equation becomes a backward hyperbolic equation with the coefficient $\nabla g(x)$ that compresses p toward point distributions (Dirac delta functions). The compression speed is quicker if $\nabla g(x)$ has larger magnitude. This indicates that probability density function p will cluster around the minima. This is actually consistent with the gradient descent process that every initial state will go to a minimum point. Our examples indicate that a repetitive implementation of such diffusion-compression process helps to cluster the probability density function toward the global minima in a much quicker pace than that of CDD method.

Another significant difference between ID method and CDD method is that CDD method requires the diffusion coefficient $\sigma(t)$ goes to zero as t tends to infinity, while ID sets $\sigma(t)$ as piecewise constants that do not go to zero. ID method can find many minima, including global minima, during the process while CCD method does not give any minima during the process and only settles near the global minimum asymptotically. More importantly, we shall prove theoretically that with probability arbitrarily close to 1, ID method can find the global minimum in a finite time T provided T reasonably large. We also prove that for any given finite sequence of (local or global) minimizers, with a positive probability the trajectory of ID method will visit an arbitrarily small neighborhood of each member of the sequence. Experimental trials shows

that within finite time intervals, the frequency that ID method reaches the global minimizer is significant more than CDD method with (1.4) does in many examples.

This paper is arranged as follows. In the next section, we present the intermittent diffusion method and give two simple examples to illustrate how it is used. A theoretical study is given in Section 3. And we show more numerical examples and comparisons in Section 4.

2 The Intermittent Diffusion Algorithm

In this section, we present the Intermittent Diffusion Algorithm. In contrast to the existing work on diffusion for global optimizations [7, 9], which gradually decreases the diffusion strength, we employ the deterministic property of (1.3) that with $\sigma = 0$ the ω-limit set of a trajectory of (1.3) is a (local) minimizer of $g(x)$ ([12]). Our idea is: (i) to allow the trajectory randomly jump off a local minimizer to a stable manifold of another local minimizer by setting $\sigma > 0$; (ii) If all eigenvalues of the linear part of the Hessian matrix have negative real parts (i.e., each local minimizer is hyperbolic), the trajectory will reach a local minimizer within short time once σ is set to 0. In fact, we can simply realize this by computing the stochastic perturbed gradient flow (1.3) with a discontinuous diffusion $\sigma(x,t)$ defined by,

$$\sigma(x,t) = \sum_{j=1}^{N} \sigma_j I_{[S_j,T_j]}(t), \tag{2.6}$$

where $0 = S_1 < T_1 < S_2 < T_2 \cdots < S_N < T_N < S_{N+1} = T$, and $I_{[S_j,T_j]}(t)$ is the characteristic function of interval $[S_j, T_j]$.

The discontinuous function $\sigma(x,t)$ actually "turns off" the diffusion in the time intervals $[T_j, S_{j+1}]$, so that the flow becomes a gradient flow. This will allow the state to approach local minimizer in a more efficient way. If the local minimizer is hyperbolic, the flow will approach the local minimizer exponentially fast.

On the other hand, when $\sigma(x,t)$ takes non-zero values, the stochastic differential equation (1.3) becomes active in diffusion and the trajectory will not rest at the stationary points. It has been shown that in many scenarios, the trajectories will eventually escape the traps of the stationary points provided sufficient noise is added to the system. We use this property of diffusion to promote the trajectories getting out of the traps of local minimizers. For this purpose, we do not want to add noise in decreasing strength. In our numerical experiments, we set $\sigma(x,t)$ to random positive constants for simplicity. The intervals $[S_j, T_j]$ are also picked with random lengths, i.e., $T_j - S_j$ is a random positive number to help the trajectories move from one stable manifold to another. In Figure

1, we illustrate the sample function σ that we used in our simulations. For convenience, we call the trajectories in the time interval $[S_j, S_{j+1}]$ one *segment* of the (random) dynamical systems.

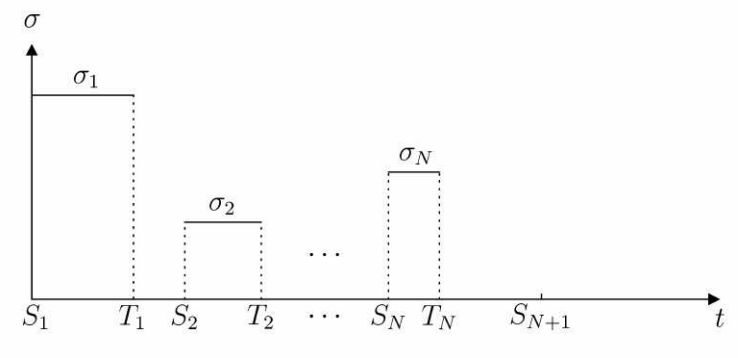

Figure 1:

Following this idea, we present the intermittent diffusion algorithm:

1. Set $\alpha \equiv$ the scale for diffusion strength, and $\gamma \equiv$ the scale for diffusion time.

2. Generate a random initial state $x_0 \in \mathbb{R}^n$, and set the optimal state $X_{opt} = x_0$.

3. Generate two positive random numbers d, t within $[0, 1]$ where d is for diffusion strength and t is for diffusion time.

4. Let $\sigma := \alpha d$, and $T := \gamma t$.

5. Compute the stochastic equation for $t \in [0, T]$

$$dx(t, \omega) = -\nabla g(x(t, \omega)dt + \sigma dW(t),\ x(0, \omega) = x_0, \tag{2.7}$$

   and record final state $x_T := x(T, \omega)$.

6. Compute the solution for the following system until a convergence criterion is satisfied,

$$\dot{x(t)} = -\nabla g(x(t)),\ x(0) = x_T, \tag{2.8}$$

   and record the final state as X_i. If $g(X_i) < g(X_{opt})$, set $X_{opt} = X_i$. This finishes one *segment* of ID method.

7. Repeat Step 3 to Step 6 for N times to obtain N segments of the trajectories, which obtains up to N local minimizers $X_1, \cdots, X_N$. For large enough N, X_{opt} is considered as an approximation of the global minimizer.

Remark 2.1.

1. Different schemes may be used to solve (2.7) and (2.8). For example, one can use Euler-Maruyama scheme for (2.7) and a Runge-Kutta scheme for (2.8).

2. The convergence criterion in Step 6 can be set in different ways. For example, a sample way that we use in our numerical experiments is to stop the iteration if the absolute value of the difference between two successive iterates is less than a prescribed tolerance $\epsilon > 0$.

3. Although we only consider constant σ_i in this paper, it can be functions of (t, x) in general. Ideally, $\sigma_i(t, x)$ shall be used to prevent repetitive visits to the same local minima and take the most efficient route to the global minimizer. But how to select $\sigma_i(t, x)$ in practice is still under investigation.

The time for convergence to a local minimizer at Step 6 is usually shorter than that of (1.3) with positive $\sigma(x, t)$. The strong noise will promote the trajectories to escape the stable manifolds quicker. In addition, ID method finds the best approximation in the entire trajectory and does not require the stopping state as the best solution, while the trajectories of CDD method only settle down near the global minimizer asymptotically. These are among the reasons that ID method is more effective than the existing diffusion method for global minimization. We present some numerical comparisons in Section 4.

In order to illustrate the algorithms, we consider two examples from [1] and [20] to show how intermittent diffusion algorithm works.

Example 1. *A Quartic Function:* The following function

$$g(x) = x^4 - 16x^2 + 5x; \tag{2.9}$$

has two local minima as shown in Figure 2. The trajectories of the algorithm move back and forth from the valleys of one minimum to the other and converge to the circles when the diffusion is turned off. They are close approximations to the local minimizers. ID method records the left one as the global minimizer because it has smaller g value.

Example 2. *Goldstein's Function:* A function

$$g(x) = x^6 - 15x^4 + 27x^2 + 250; \tag{2.10}$$

has three local minima as shown in Figure 3. The trajectories spend more time around the left and right minima than in the middle because it has higher value. The circle points are the convergent locations. ID method records both points as global minimizers because they have the same value.

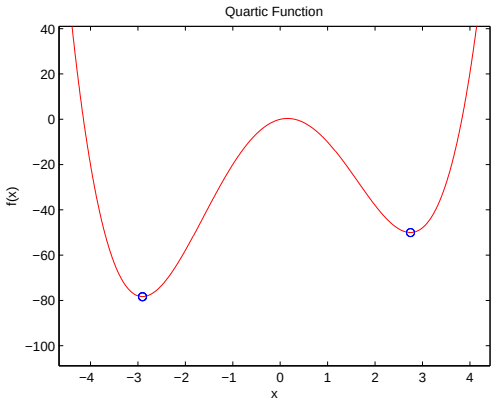

Figure 2: Numerical experiment for Example 1. The blue circles mark the convergent points at the non-diffusion step.

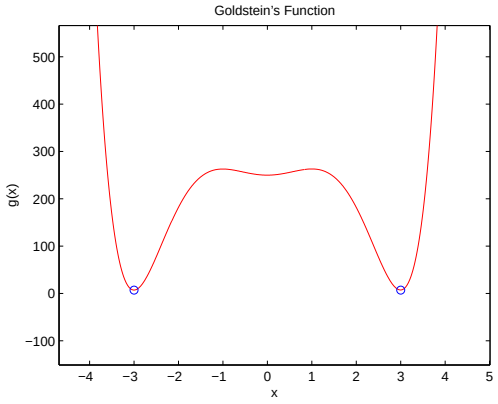

Figure 3: Numerical experiment for Example 2. The blue circles mark the convergent points at the non-diffusion step.

3 An analysis of ID Method

In this section, we consider some theoretical properties of the ID algorithm. For convenience, we focus our analysis on the case where σ is taken as piecewise constant functions as shown in Figure 1.

Theorem 3.1. *Consider a gradient flow*

$$\dot{x}(t) = -\nabla g(x(t)),$$

and its ID process generated by

$$dx(t,\omega) = -\nabla g(x(t,\omega))dt + \sigma(t)dW_t,$$

where $\sigma(t) = \sum_{j=1}^{N} \sigma_j I_{[S_j,T_j]}(t)$. *Suppose* $g(x)$ *has a finite number of minima. Let* Q *be the set of global minimizers,* U *a small neighborhood of* Q*, and* X_{opt}

the optimal solution obtained by the ID process. Then for any given $\varepsilon > 0$, there exist $\tau > 0$, $\sigma_0 > 0$ and $N_0 > 0$ such that if $T_i - S_i > \tau$, $\sigma_i < \sigma_0$ (for $i = 1, \cdots, N$), and $N > N_0$,

$$\mathbb{P}(X_{opt} \in U) \geq 1 - \varepsilon. \tag{3.1}$$

Proof. We give the proof for the case assuming there is one global minimum point $\{x^*\}$ and $U = B(x^*, \gamma)$, a small ball centered at x^* with small enough radius γ. For other situations including multiple global minima, a similar proof can be given with minor modifications and we will omit it in this paper.

Let us consider one segment of the ID process in $[S_i, S_{i+1}]$ first. Since the trajectory of ID rests on a local minimizer of g at the end of each gradient descent flow, we may assume $x(S_i) = q_i$, where q_i is a minimizer. For $t \in [S_i, T_i]$, (2.7) is a homogeneous diffusion process with initial value $x(S_i) = q_i$. We denote the probability density function of the ID trajectory $x(t, \omega)$ visiting x at time $t \in [S_i, T_i]$ by $p_{\sigma_i}(t - S_i, x; q_i)$ which depends on perturbation strength σ_i and initial state q_i. It is well known that p_{σ_i} satisfies the Fokker-Planck equation given by

$$(p_{\sigma_i})_t = \nabla \cdot (\nabla g(x) p_{\sigma_i}) + \frac{1}{2}\sigma_i^2 \Delta p_{\sigma_i}, \tag{3.2}$$

with initial condition as the Dirac delta function concentrated at q_i. As a solution of this linear convection diffusion equation, $p_{\sigma_i} \in C^\infty$ continuously depends on $t > 0$ and $\sigma_i > 0$.

It is straightforward to verify that (3.2) has a steady state solution which is given by the famous Gibbs distribution,

$$\bar{p}_{\sigma_i}(x) = A e^{-\frac{2g(x)}{\sigma_i^2}}, \tag{3.3}$$

where A is a constant defined as

$$A = \left(\int_{\mathbb{R}^n} e^{-\frac{2g(x)}{\sigma_i^2}} dx \right)^{-1} \tag{3.4}$$

to normalize p_{σ_i} to be a probability density function. Furthermore, it is also easy to verify [7, 9] that $\bar{p}_{\sigma_i}(x)$ converges to a point distribution concentrated at x^*, i.e.

$$\lim_{\sigma_i \to 0} \bar{p}_{\sigma_i}(x) = \delta_{x^*}(x). \tag{3.5}$$

In other words, (3.5) implies that

$$\lim_{\sigma_i \to 0} \lim_{t \to \infty} \int_U p_{\sigma_i}(t, x; q_i) dx = 1. \tag{3.6}$$

For a minimum point ξ, let us define the attraction set of ξ as

$$K(\xi) = \{x : (2.8) \text{ with } x \text{ as the initial state converges to a point in } B(\xi, \gamma)\}.$$

Obviously, we have $B(\xi,\gamma) \subseteq K(\xi)$ if γ is small enough. Thus, by the continuity of p_{σ_i} with respect to t and σ_i, for any given $\alpha \in (0,1)$ there exist $\tau(\alpha, q_i)$ and $\sigma(\alpha, q_i)$ such that for $T_i - S_i > \tau(\alpha, q_i)$ and $\sigma_i < \sigma(\alpha, q_i)$, we have

$$\mathbb{P}\big(x(S_{i+1}) \in U|\ x(S_i) = q_i\big) = \mathbb{P}\big(x(T_i) \in K(x^*)|\ x(S_i) = q_i\big) > \alpha. \tag{3.7}$$

This is true because the flow is deterministic on $[T_i, S_{i+1}]$.

Let Θ be the set consisting of local minimizers of g, then

$$\mathbb{P}\big(x(S_{i+1}) \in U^c\big) = \sum_{q_i \in \Theta} \mathbb{P}\big(x(S_{i+1}) \in U^c|\ x(S_i) = q_i\big)\mathbb{P}(x(S_i) = q_i), \tag{3.8}$$

where U^c is the complement of U. If one chooses $\tau = \max_{q_i\in\Theta} \tau(\alpha, q_i)$ and $\sigma_0 = \min_{q_i\in\Theta} \sigma(\alpha, q_i)$, then by estimate (3.7), one has

$$\mathbb{P}\big(x(S_{i+1}) \in U^c|\ x(S_i) = q_i\big) < (1-\alpha) \text{ for all } q_i \in \Theta, \tag{3.9}$$

provided $(T_i - S_i) > \tau$ and $\sigma_i < \sigma_0$. This implies

$$\mathbb{P}\big(x(S_{i+1}) \in U^c\big) < (1-\alpha) \sum_{q_i\in\Theta} \mathbb{P}(x(S_i) = q_i) < (1-\alpha). \tag{3.10}$$

It is obvious that the estimate (3.10) can be made for all segments $i = 1, 2, \cdots, N$. If one repeats it for all intervals $[S_j, S_{j+1}]$, $j = 1, \cdots, N$, with the same fixed α, then the probability that the solution X_{opt} of the ID method does not belong to U satisfies,

$$\begin{aligned} \mathbb{P}(X_{opt} \in U^c) &= \mathbb{P}(\cup_{i=0}^{N} x(S_i) \in U^c) \\ &< (1-\alpha)^N. \end{aligned}$$

For any given $\varepsilon \in (0,1)$, there exists $N_0 > 0$, such as for $N \geq N_0$,

$$\mathbb{P}(X_{opt} \in U^c) < (1-\alpha)^N < \varepsilon,$$

which is equivalent to (3.1). □

It is desirable to extend the above results to more general situations such as there are infinitely many minima in g. One way to do so is to use the notion of isolated invariant sets in topological dynamic systems. The idea is to partition the minimizers into a finite number of neighborhoods of isolated invariant sets and study their statistical properties. More details along this line of investigations are reported in [8].

It is also possible to modify the proof of Theorem 3.1 for the case that g has infinitely many minimizers. In fact, it is easy to verify that the majority of the proof remain valid except two places.

1. If Q contains infinitely many global minimizers, there are two different cases: its Lesbegue measure is zero or positive. (3.5) is no longer true in either case. For example, if $|Q| > 0$, then the asymptotic distribution $\bar{p}_{\sigma_i}$ as $\sigma_i \to 0$ is uniform on Q. However, (3.6) still holds in both cases. Therefore the consequent (3.7) is still true.

2. It might not be possible to find finite values τ and σ_0 such that (3.9) and (3.10) holds for all $i = 1, \cdots, N$. This is because $\sup_{q_i \in \Theta} \tau(\alpha, q_i)$ may become ∞, and $\inf_{q_i \in \Theta} \sigma_i(\alpha, q_i)$ might be 0, if Θ contains infinitely many minimizers. However, for each fixed trajectory of ID method, any segment $[S_i, S_{i+1}]$ must have the following property,

$$\mathbb{P}\big(x(S_{i+1}) \in U^c\big) < (1 - \alpha),$$

provided $[T_i - S_i]$ is large and σ_i small enough. In particular, if one picks $\alpha \geq 1/2$, it implies that $x(S_{i+1})$ has larger probability in U than in the neighborhoods of other local minimizers. Then, one can repeat the estimate for all segments in this trajectory and obtain the bound (3.1), even though the conditions imposed on $(T_i - S_i)$ and σ_i might depend on the trajectories of ID process.

We summarize this extension in the following theorem.

Theorem 3.2. *Consider the ID process for the gradient flow stated in Theorem 3.1, and suppose g has infinitely many minimizers.*

(i) If $0 < \gamma << 1$, then

$$\mathbb{P}(x(S_{i+1}, \omega) \in U) \geq \mathbb{P}(x(S_{i+1}, \omega) \in B(q_j, \gamma)),$$

where q_j is any local minimizer, provided $T_i - S_i$ is large and σ_i small enough.

(ii) For any given $\varepsilon \in (0, 1)$, then

$$\mathbb{P}(X_{opt} \in U) \geq 1 - \varepsilon, \tag{3.11}$$

provided σ_i is small and $T_i - S_i$ large for all $i = 1, 2, \cdots, N$, and the number segments N also large enough in the ID process.

Remark 3.3. From the proof of Theorem 3.1, we can also observe the following points:

1. We note that in ID method, the convergence does not require to have $\sigma(t) \to 0$ as t tends to infinity. This is different from the diffusion coefficient requirements of CDD.

2. The convergence of ID method is monotone with respect to number of segments N. In fact, it can be seen from the proof that the convergence rate (in probability sense) of ID method is a geometric series based on the factor α that is related to the stable manifold that contains the global minimum. This ultimately determines good selections of $[S_i, T_i]$ and σ_i, which is still under investigation.

Theorem 3.4. *Consider the ID process for the gradient flow stated in Theorem 3.1, and suppose that g satisfies tthe following conditions:*

(H1) g has a global minimum and $\lim_{|x|\to\infty} g(x) = \infty$.

(H2) ∇g has a Lipschtiz constant L on $\mathbb{R}^n$.

(H3) All critical points of g are hyperbolic.

Given $0 < \epsilon$, $T > 0$ and any finite sequence $\{p_i\}_{i=1\cdots N}$ of local minimizers of g, there exist a $\Omega' \subset \Omega$ with $\mathbb{P}(\Omega') > 0$ such that for $\omega \in \Omega'$ the trajectory of ID visiting each $B(p_i, \epsilon)$ within $[0, T]$.

For convenience of following discussion, we denote $f := -\nabla g$, and $(\Omega, \mathscr{F}, \mathbb{P})$ the Wiener space generated by Brownian motion $W(t)$. To prove this theorem, we need some lemmas for the counterpart equation of (1.3):

$$dx = f(x)dt + \sigma dW, x(0) = x_0, \tag{3.12}$$

where σ is a constant. Denoted by $\phi(t, \omega)x_0$, $\omega \in \Omega$ the solution of (3.12).

Lemma 3.5. Assume (H1) and (H2) are true. The solution of (3.12) is a global solution for all $\omega \in \Omega$ (i.e., $\phi(t, \omega)x_0$ is defined for $t \in [0, \infty)$).

Proof. Let $v := x - \sigma W$, we have

$$\frac{dv}{dt} = f(v + \sigma W). \tag{3.13}$$

Then, the time-derivative of g along trajectories $v(t)$ is given by

$$\begin{aligned}
\frac{dg(v(t))}{dt} &= -f(v) \cdot f(x) && (3.14)\\
&= -(f(v) - f(x) + f(x)) \cdot f(x) && (3.15)\\
&\leq -|f(x)|^2 + |f(x)||f(x) - f(v)| && (3.16)\\
&\leq -|f(x)|^2 + L|f(x)||\sigma W| && (3.17)\\
&\leq (L|\sigma W(t)|)2/4, && (3.18)
\end{aligned}$$

which implies

$$g(v(t)) \le g(v(0)) + \int_0^t (L|\sigma W(s)|)2/4 := M(t).$$

Therefore $v(t)$ is bounded by $K_{M(t)}$ which is a bounded set. It follows that $v(t)$ is defined for $t \in [0, \infty)$, and so is $x(t)$. □

Let $\dot{\psi}(t)$ be a piecewise continuous function. We consider the following nonautonomous equation:

$$dy = f(y)dt + \sigma\dot{\psi}dt,\ y(0) = x_0. \tag{3.19}$$

Denoted by $S(t, \dot{\psi})x_0$ the solution of (3.19).

Lemma 3.6. Assume (H1) and (H2) are true. For any piecewise continuous $\dot{\psi}$, the solution of (3.19) is a global solution (i.e., $S(t, \dot{\psi})x_0$ is defined for t $[0, \infty)$).

Proof. Let $\psi(t) := \int_0^t \dot{\psi}(s)ds$, which is a continuous function of t. The proof is similar to that for Theorem 3.5 with $W(t)$ replaced by $\psi(t)$. □

Lemma 3.7. Assume (H1) and (H2) are true. Given $T > 0$, $\epsilon > 0$, x_0 and $q \in \mathbb{R}^n$, there exists a piecewise continuous $\dot{\psi}(t)$ such that solution of (3.19) satisfying the boundary condition $S(T, \dot{\psi})x_0 \in B(q, \epsilon)$.

Proof. We may assume $\sigma = 1$ for simplicity and $\epsilon < 1$. Let $N \in \mathbb{N}$ and

$$0 := t_0 < t_1 < \cdots < t_N := T, \qquad t_i - t_{i-1} = h := T/N \tag{3.20}$$

$$q_i := x_0 + i\big((p - x_0)/N\big), \qquad 1 \le i \le N. \tag{3.21}$$

Let V be the tube centered at the line segment from x_0 to q and with radius one, i.e.,

$$V := \bigcup_{t\in[0,1]} B((1-t)x_0 + tq, 1)$$

Let $C := \sup_{x\in V} |f(x)| < \infty$. We are going to construct $\psi(t)$ on each $(t_i, t_i + 1)$. According to Lemma 3.6 the solution are defined for $t \in (t_i, t_i + 1)$. For simplicity we set $\psi(t_i) = 0$ at each t_i.

(i) On (t_0, t_1):

Denote $q_0 := x_0$. Let

$$\psi(t) := (q_1 - q_0)t/h,\ t \in (t_0, t_1)$$

then $q_0 + \psi(t) \in V$, $t \in (t_0, t_1)$. Therefore,

$$\sup_{s\in(t_0,t_1)} |f(q_0 + \psi(s))| \le C.$$

From (3.19) we have

$$\begin{aligned} y(t) - (q_0 + \psi(t)) &= \int_{t_0}^{t} f(y(s)) - f(q_0 + \psi(s))ds \\ &\quad + \int_{t_0}^{t} f(q_0 + \psi(s))ds, \end{aligned}$$

which implies

$$\begin{aligned} |y(t) - (q_0 + \psi(t))| &\leq \int_{t_0}^{t} |f(y(s)) - f(q_0 + \psi(s))|ds \\ &\quad + \int_{t_0}^{t} |f(q_0 + \psi(s))|ds, \\ &\leq \int_{t_0}^{t} L|y(s) - (q_0 + \psi(s))|ds + hC. \end{aligned}$$

By Gronwall inequality, we have

$$|y(t_1) - (q_0 + \psi(t_1))| \leq hCe^{Lh} < \epsilon/2,$$

provided h small. Denote by $\tilde{q}_1 := y(t_1) \in V$.

(ii) On (t_{i-1}, t_i), $2 \leq i \leq N$:

To define $\psi(t)$ on each (t_{i-1}, t_i), we repeat the following process for $2 \leq i \leq N$. Let

$$\psi := (q_i - \tilde{q}_{i-1})(t - t_{i-1})/h, \ t \in (t_{i-1}, t_i)$$

then $\tilde{q}_{i-1} + \psi(t) \in V$, $t \in (t_{i-1}, t_i)$. Therefore,

$$\sup_{s \in (t_{i-1}, t_i)} |f(\tilde{q}_{i-1} + \psi(s))| \leq C.$$

Consider (3.19) with $t \in (t_{i-1}, t_i)$ and $y(t_{i-1}) = \tilde{q}_{i-1}$, then we have

$$\begin{aligned} y(t) - (\tilde{q}_{i-1} + \psi(t)) &= \int_{t_{i-1}}^{t} f(y(s)) - f(\tilde{q}_{i-1} + \psi(s))ds \\ &\quad + \int_{t_{i-1}}^{t} f(\tilde{q}_{i-1} + \psi(s))ds, \end{aligned}$$

which implies

$$\begin{aligned} |y(t) - (\tilde{q}_{i-1} + \psi(t))| &\leq \int_{t_{i-1}}^{t} |f(y(s)) - f(\tilde{q}_{i-1} + \psi(s))|ds \\ &\quad + \int_{t_{i-1}}^{t} |f(\tilde{q}_{i-1} + \psi(s))|ds, \\ &\leq \int_{t_{i-1}}^{t} L|y(s) - (\tilde{q}_{i-1} + \psi(s))|ds + hC. \end{aligned}$$

By Gronwall inequality,

$$|y(t_i) - (\tilde{q}_{i-1} + \psi(t_i))| \leq hCe^{Lh} < \epsilon/2,$$

provided h small. Denote by $\tilde{q}_i := y(t_i)$.

At the final stage $i = N$, we have

$$|y(T) - (\tilde{q}_{i-1} + \psi(T))| = |y(T) - q| < \epsilon/2$$

Thus we have $|S(T, \dot{\psi})x_0 - q| < \epsilon$. The proof is complete. □

We are going to prove that with positive probability of ω the solution $\phi(t, \omega)p$ of (3.12) can be approximated by solution $S(t, \dot{\psi})p$ of (3.7) in any finite time interval and any initial point p.

Lemma 3.8. Assume (H1) and (H2) are true. Given a piecewise continuous $\dot{\psi}$, $T > 0$ and $\gamma > 0$, the set A_γ defined by

$$A_\gamma := \{\omega : \sup_{p \in \mathbb{R}^n} \sup_{t \in [0,T]} |S(t, \dot{\psi})p - \phi(t, \omega)p| < \gamma\}, \tag{3.22}$$

has $\mathbb{P}(A_\gamma) > 0$.

Proof. According to Lemma 3.5 and Lemma 3.6, $S(t, \dot{\psi})p$ and $\phi(t, \omega)p$ are defined for $t \in [0, T]$. From (3.12) and (3.19) we have

$$d(x - y) = (f(x) - f(y))dt - \sigma\dot{\psi}(t)dt + \sigma dW(t). \tag{3.23}$$

We may assume $\psi(0) = 0$ since it doesn't affect the solution of (3.19). Let $\tilde{W}(t) := W(t) - \psi(t)$, $t \in [0, T]$, then we have

$$d\tilde{W} = -\dot{\psi}(t)dt + dW. \tag{3.24}$$

By Girsanov's Theorem ([11]), $\tilde{W}$ is a Brownian motion under probability $\tilde{\mathbb{P}}$, where

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \exp(\int_0^T \dot{\psi}(s) \cdot dW(s) - \frac{1}{2}|\dot{\psi}(s)|^2 ds) > 0. \tag{3.25}$$

Let $\tilde{\Lambda}_\varepsilon := \{\omega : \sup_{t \in [0,T]} |\tilde{W}(t, \omega)| < \varepsilon\}$ and $\Lambda_\varepsilon := \{\omega : \sup_{t \in [0,T]} |W(t, \omega)| < \varepsilon\}$, then $\tilde{\mathbb{P}}(\tilde{\Lambda}_\varepsilon) > 0$, which implies $\mathbb{P}(\Lambda_\varepsilon) > 0$ by (3.25). It follows that for $\omega \in \Lambda_\varepsilon$

$$\begin{aligned} |\phi(t, \omega)p - S(t, \dot{\psi})p| &\leq \int_0^t |f(\phi(s, \omega)p) - f(S(s, \dot{\psi})p)|ds + \sigma|\tilde{W}(t)| \\ &\leq \int_0^t L|\phi(s, \omega)p - S(s, \dot{\psi})p|ds + \sigma\varepsilon. \end{aligned} \tag{3.26}$$

By Gronwall inequality, we have

$$\sup_{p\in\mathbb{R}^n}\sup_{t\in[0,T]}|S(t,\dot{\psi})p-\phi(t,\omega)p|<\sigma\varepsilon e^{LT},\ \omega\in\Lambda_\varepsilon.$$

The assertion follows by chose ε small enough. □

Proof of Theorem 3.4:
We still denote by $\phi(t,\omega)$ the flow for (1.3) with piecewise constant diffusion $\sigma(t)$ shown in Figure 1. Let q_0 be a initial point. We may assume $B(p_i,\epsilon)$ lying in the stable manifold of p_i for $i=1,\cdots,N$.

Let $q_i\in\mathbb{R}^n$. For each $1\le i\le N$, by Lemma 3.7, there exists a piecewise continuous $\dot{\psi}_i(t)$, $t\in[S_i,T_i]$ such that the solution of

$$dy^{(i)}=f(y^{(i)})dt+\sigma(t)\dot{\psi}_i dt,\ y^{(i)}(S_i)=q_{i-1}, \tag{3.27}$$

satisfying

$$y^{(i)}(T)\in B(p_i,\epsilon/2).$$

Since $\sigma(t)=0$ and $y^{(i)}(T_i)$ lies in the stable manifold of p_i, it follows that

$$y^{(i)}(t)\in B(p_i,\epsilon/2),\ t\in[T_i,S_{i+1}],$$

Selecting $q_i=y^{(i-1)}(S_{i-1})$, $2\le i\le N$, and let $\dot{\psi}(t)$ be the piecewise continuous function defined by

$$\dot{\psi}(t)=\begin{cases}\dot{\psi}_i(t), & t\in[S_i,T_i],\ 1\le i\le N,\\ 0, & \text{otherwise.}\end{cases} \tag{3.28}$$

Then we have

$$S(t,\dot{\psi})q_0\in B(p_i,\epsilon/2),\ t\in[T_i,S_{i+1}],\ 1\le i\le N,$$

By Lemma 3.8, the set

$$\{\omega:\sup_{t\in[0,T]}|S(t,\dot{\psi})q_0-\phi(t,\omega)q_0|<\epsilon/2\}, \tag{3.29}$$

has positive probability, which implies the set

$$\Omega':=\{\omega:\phi(t,\omega)q_0\in B(p_i,\epsilon_i),t\in[T_i,S_{i+1}],1\le i\le N\}, \tag{3.30}$$

has $\mathbb{P}(\Omega')>0$. The proof is complete.

4 Numerical Examples

In this section, we will demonstrate the ID method on four standard test examples for global optimization methods. They are designed to have many local minima that are similar to the unique global minimum and they can easily trap the solutions. In all the experiments, we set the parameters $\alpha = 10$, $\gamma = 10$ and $N = 10$.

Problem 1. *Penalized Shubert Function.* Let

$$g_1(x) = \sum_{i=1}^{5} i \cos((i+1)x + 1) \tag{4.31}$$

be the standard Shubert function. Then the penalized Shubert function is defined by

$$g(x) := g_1(x) + u(x, 10, 100, 2) \tag{4.32}$$

where $u(x, a, k, m)$ is the penalization function defined by

$$u(x, a, k, m) := \begin{cases} k(x-a)^m, & x > a \\ 0 & -a \le x \le a \\ k(-x-a)^m & x < -a \end{cases}. \tag{4.33}$$

As shown in Figure 4, the function g has 19 local minima in the region $\{x : |x| < 10\}$ and three of them are global minima. One realization of the ID method gives the circled points, which is the convergent points of the gradient flow after the diffusion is turned off. It finds the global minimum in the middle.

Problem 2. *Two-Dimensional Penalized Shubert Function* is defined by

$$\begin{aligned} g(x,y) &= \left\{\sum_{i=1}^{5} i \cos((i+1)x+1)\right\}\left\{\sum_{i=1}^{5} i \cos((i+1)y+1)\right\} \\ &+ u(x, 10, 100, 2) + u(y, 10, 100, 2). \end{aligned} \tag{4.34}$$

It has 760 local minima in the region $\{(x, y) : |x| < 10, |y| < 10\}$ and 18 of them are global minima as shown in Figure 5. It also shows that an arbitrary realization of ID algorithm gives the circled points. Among them, it finds three global minima.

Problem 3. *Two-Dimensional Penalized Shubert Function with Parameter* β is defined by

$$\begin{aligned} g(x,y) &= \left\{\sum_{i=1}^{5} i \cos((i+1)x+1)\right\}\left\{\sum_{i=1}^{5} i \cos((i+1)y+1)\right\} \\ &+ u(x, 10, 100, 2) + u(y, 10, 100, 2) \\ &+ \beta[(x - 6.0835)^2 + (y + 5.8581)^2], \end{aligned} \tag{4.35}$$

where $\beta > 0$ is a constant and $(x, y) = (6.0835, -5.8581)$ is a global minimizer of g with $\beta = 0$. The function behaves roughly the same with the function considered in Problem 2, but has only one unique global minimizer at $(x, y) = (6.0853, -5.8581)$ as shown in Figure 6. We show the results of convergent points (circled points) of an arbitrary ID realization. Similar to the previous problems, the ID method is able to reach the global minima.

Problem 4. *A Multi-Dimensional Function.* Let

$$\begin{aligned} g(x) &= (\pi/n)\left\{k\sin^2(\pi y_1)\right. \\ &\quad \left. + \sum_{i=1}^{n-1}(y_i - A)^2[1 + k\sin^2(\pi y_{i+1})] + (y_n - A)^2\right\}, \end{aligned} \tag{4.36}$$

where

$$\begin{aligned} x &= (x_1, \cdots, x_n) \in \mathbb{R}^n, \\ y_i &= 1 + (x_i - 1)/4,\ i = 1, \cdots, n, \\ k &= 10,\ A = 1. \end{aligned}$$

The function has roughly 5^n local minima in the region $\{(x, y) : |x| < 10, |y| < 10\}$ and a unique global minimum located at

$$x_i = 1,\ i = 1, \cdots, n.$$

For $n = 3$, Figure 7 is the projection plot of the graph of $y = g(x_1, x_2, x_3)$ on the space $x_3 = 1$, which illustrates the global minimizers is surrounded by a lot of local minimizers. The convergent points visited in sequential order by an arbitrary ID realization for $n = 3$ and $n = 4$ are shown in following respectively:

$$\begin{bmatrix} -6.96 \\ 5 \\ -7 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 9 \\ 5 \\ -3 \end{bmatrix}, \begin{bmatrix} -10.94 \\ -7 \\ 5 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 9 \\ -7 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 9 \\ 5 \end{bmatrix}.$$

$$\begin{bmatrix} -6.96 \\ 8.99 \\ 8.99 \\ 4.99 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2.98 \\ -2.98 \\ -2.98 \\ -2.98 \end{bmatrix}, \begin{bmatrix} -8.96 \\ -2.99 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2.98 \\ 4.98 \\ -2.98 \\ 1 \end{bmatrix}, \begin{bmatrix} -6.96 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The ID method reaches the global minima for both cases. And the frequencies of finding the global minimum are obviously higher than that of other local minima.

We compare the ID method with the CDD method for global optimizations studied in [9]. We run both methods on Problem 3 for 100 independent realizations with the final stopping time 300 respectively. In each realization, the ID

and diffusion methods use the same initial point and Brownian motion path. We record the number of visits that ID and CDD reach a small neighborhood of the unique global minimizer respectively. Our experiments shows that the average number of visits in the 100 realizations for the ID method is 7.5. In contrast, it is 0 for the CDD method, which indicates that it fails to reach the small neighborhood of the global minimum in most of the realizations. This is not surprising since the convergence time for CDD is exponentially long, while the convergence time at non-diffusion step for ID is short because the gradient flow drives the trajectories to a small neighborhood of the minimizers.

We also compare the first time to reach a small neighborhood of the global minimizer for Problem 3 by both methods. The first time that ID and CDD methods enter a square neighborhood of the global minimizer with diameter $\epsilon = 0.001$ and $\epsilon = 0.0001$ are recorded respectively for 30 independent realizations. Figure 8 shows the log plot of the first arriving time for $\epsilon = 0.001$. The ID takes less time to reach the global minimizer than CDD does at most realizations. Figure 9 shows the log plot of the first arriving time for a much smaller neighborhood with $\epsilon = 0.0001$. The ID takes less time to hit the global minimizer than CDD does for all realizations except one, and the time difference is significant large. This is the evidence that CDD takes too much time to wander in the stable manifold, while ID method converges to the minimizer very quickly due to the deterministic gradient descent process.

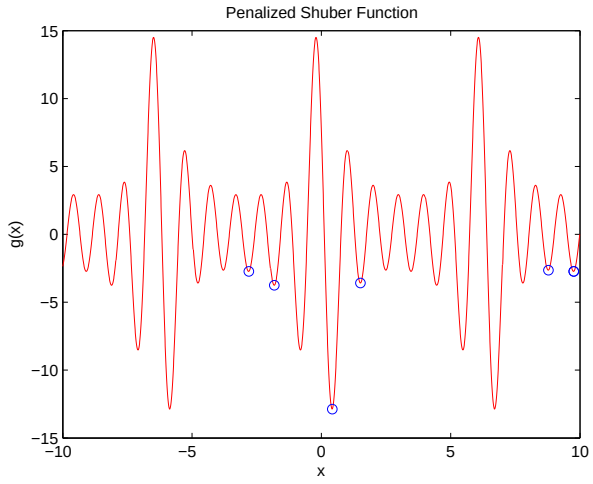

Figure 4: Numerical experiment for Problem 3. The blue circles mark the convergent points at the non-diffusion step.

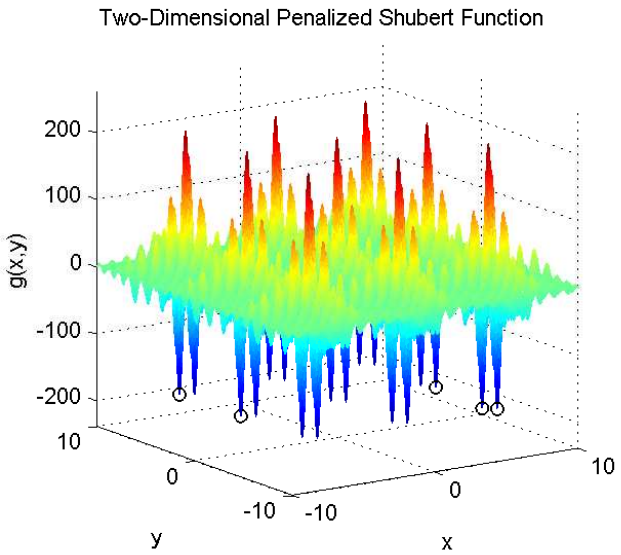

Figure 5: Numerical experiment for Problem 4. The black circles mark the convergent points at the non-diffusion step.

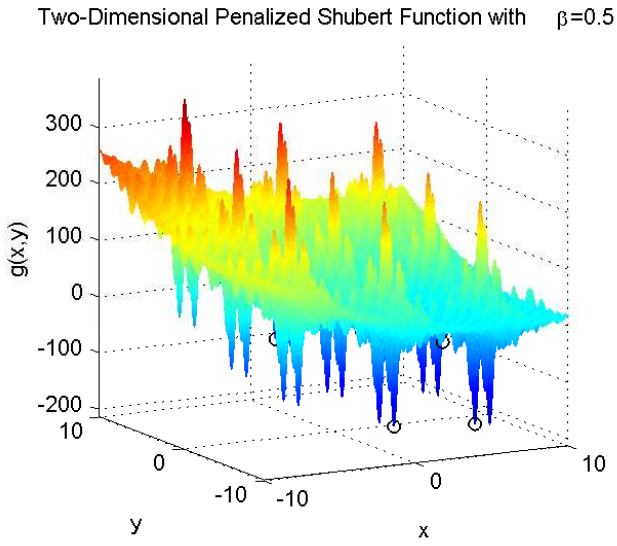

Figure 6: Numerical experiment for Problem 5. The black circles mark the convergent points at the non-diffusion step.

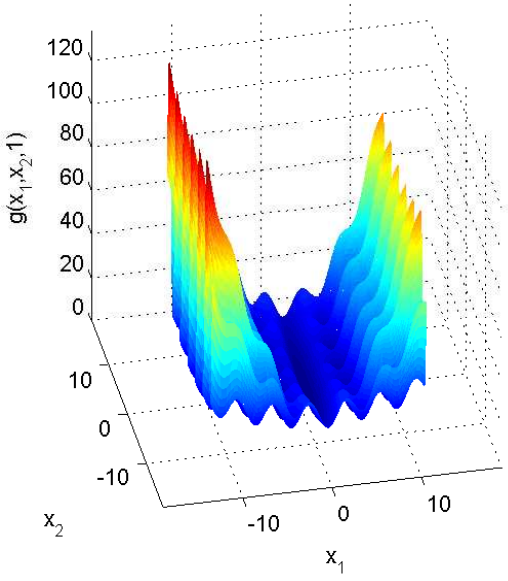

Figure 7: the projection plot of the graph of $y = g(x_1, x_2, x_3)$ on the space $x_3 = 1$ for Problem 6.

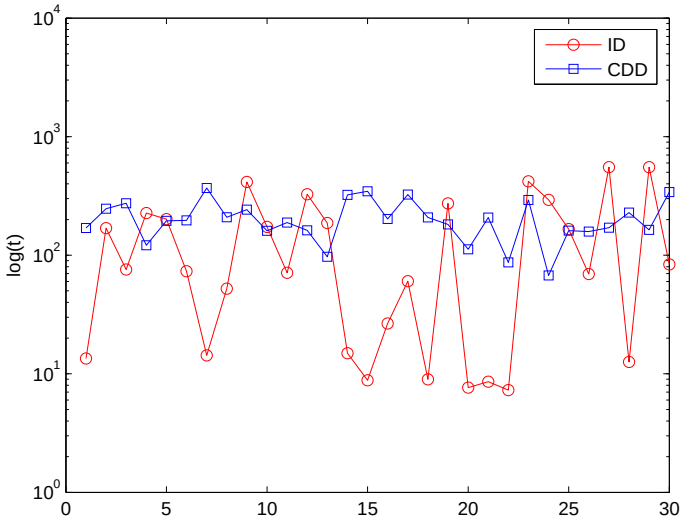

Figure 8: The log plot of the The first time to enter a square neigborhood of the global minimizer with diameter $\epsilon = 0.001$ by ID and CDD for 30 independent realizations.

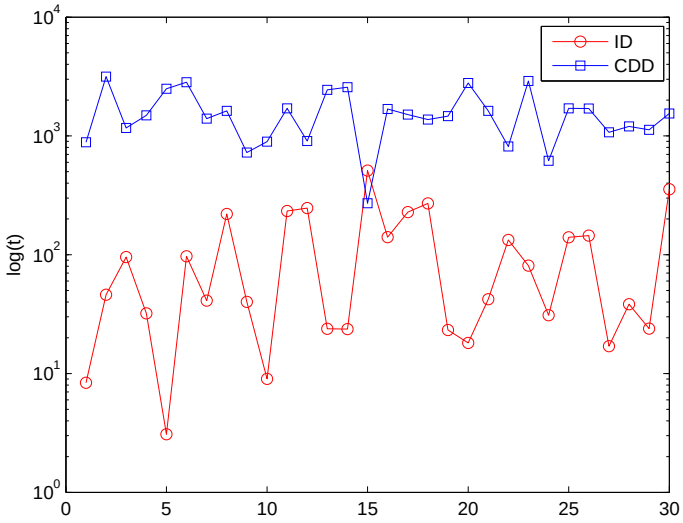

Figure 9: The log plot of the The first time to enter a square neighborhood of the global minimizer with diameter $\epsilon = 0.0001$ by ID and CDD for 30 independent realizations.

Acknowledgement: The authors would like to thank Luca Dieci, Wen Huang, Liangda Huang and Kening Lu for helpful comments and discussions in this project.

References

[1] F. Aluffi-Pentini, V. Parisi and F. Zirilli, Global optimization and stochastic differential equations, *J. Optmiz. Theory App.*, 47 (1), 1-16, 1985.

[2] L. Arnold, *Random Dynamical Systems*, Springer Verlag, Berlin, 1998.

[3] G. Bao and P. J. Li, Inverse medium scattering for the Helmholtz equation at fixed frequency, *Inverse Problems*, 21 (2005), pp1621-1641.

[4] G. Bao and P. J. Li, *Inverse medium scattering problems for electromagnetic waves*, SIAM J. Appl. Math., 65 (2005), pp2049-2066.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] V. Cerny, A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *J. Optmiz. Theory App.*, 45(1), 41-51, 1985.

[7] T.-S. Chiang, C.-R. Hwang and S.-J. Sheu, Diffusion for global optimization in $\mathbb{R}^n$, *SIAM J. Control and Optim.*, 25 (3), 737-753, 1986.

[8] S. N. Chow, W. P. Li, Z. X. Liu and H. M. Zhou, Conley-Markov Connection Matrices for Gradient-like Flows with white noises, in preparation.

[9] S. Geman and C.-R. Hwang, Diffusion for global optimization, *SIAM J. Control and Optim.*, 24 (5), 1031-1043, 1986.

[10] B. Gidas, Metropolis-type Monte Carlo simulation algorithms and simulated annealing, in: J.L. Snell (Ed.), *Topics in Contemporary Probability and Its Applications, in: Probability and Stochastics Series*, CRC Press, Boca Raton, FL, 1995, pp. 159 - 232.

[11] I. V. Girsanov, On transforming a certain class of stochastic processes by absolutely continuous substitution of measures, Theory Probab. Appl., 5, No. 3 (1960), 285-301.

[12] Jack K. Hale, *Asymptotic behavior of dissipative systems (Mathematical Surveys and Monographs, 25)*, Amer Mathematical Society

[13] R. Holley, S. Kusuoka and D. Strook, Asymptotics of the spectral gap with applications to the theory of simulated annealing, *J. Func. Anal.*, 83(2), 333-347, 1989.

[14] R. Holley and D. Strook, Simulated annealing via Sobolev inequalities, *Commun. Math. Phys.*, 115 (4), 553-569, 1988.

[15] L. Ingber, Very fast simulated re-annealing. *Math. Comput. Modelling*, 12 (8), 967–973, 1989.

[16] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, 1999

[17] S. Kirkpatrick, S.; C. D. Gelatt, Jr. and M. P. Vecchi, Optimization by simulated snnealing, *Science*, Vol. 220 No. 4598, pp. 671-680, 1983.

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equations of state calculations by fast computing machines, *J. Chem. Phys.*, vol 21, 1953, pp 1087-1091.

[19] J. Nocedal and S. J. Wright *Numerical Optimization*, Springer, 1999.

[20] H. Szu and R. Hartley, Fast simulated annealing, *Phys. Lett. A*, 122 (3-4), 157-162. 1987,

[21] Pavlyukevich, Ilya, Simulated annealing for Le'vy-driven jump-diffusions, *Stochastic Process. Appl.* 118 (2008), no. 6, 1071–1105.

[22] A. Torn and A. Zilinskas, *Global optimization*, Springer Verlag, Berlin, 1989.

[23] R. L. Yang, Convergence of the simulated annealing algorithm for continuous global optimization, *J. Optim. Theory Appl.*, 104 (3), 691–716, 2000.