J Glob Optim
DOI 10.1007/s10898-008-9397-4

Convergence analysis of a global optimization algorithm using stochastic differential equations

Panos Parpas · Berç Rustem

Received: 14 December 2008 / Accepted: 17 December 2008

Abstract We establish the convergence of a stochastic global optimization algorithm for general non-convex, smooth functions. The algorithm follows the trajectory of an appropriately defined stochastic differential equation (SDE). In order to achieve feasibility of the trajectory we introduce information from the Lagrange multipliers into the SDE. The analysis is performed in two steps. We first give a characterization of a probability measure (Π) that is defined on the set of global minima of the problem. We then study the transition density associated with the augmented diffusion process and show that its weak limit is given by Π.

Keywords Stochastic global optimization · Simulated annealing · Stochastic differential equations · Fokker-Planck equation · Laplace's method · Projection algorithms

1 Introduction

The global optimization of general smooth functions subject to linear constraints, using a system of SDEs was discussed in [18]. The main aim of this paper is to extend the method so that it can be applied to general non-linearly constrained problems. The proposed algorithm has already been used for the solution of portfolio optimization [15], and chance constrained problems [17]. However, the convergence of the algorithm has not yet been established. We address the issue in this paper by providing the theoretical foundations for the method.

Consider the following general non-linear programming problem:

$$P^* = \min_x f(x)$$
$$\text{s.t. } g(x) = 0. \tag{1.1}$$

P. Parpas (✉) · B. Rustem
Department of Computing, Imperial College, London SW7 2AZ, UK
e-mail: pp500@doc.ic.ac.uk

Published online: 16 January 2009

Where f and g are general non-linear functions. In particular we assume that $f : \mathbb{R}^n \to \mathbb{R}$, and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable. The feasible region will be denoted by $\mathcal{F}$.

A well known method for obtaining a solution to an unconstrained optimization problem is to consider the following Ordinary Differential Equation (ODE):

$$dX(t) = -\nabla f(X(t))dt. \tag{1.2}$$

By studying the behavior of $X(t)$ for large t, it can be shown that $X(t)$ will eventually converge to a stationary point of the unconstrained problem. A review of, so called, continuous-path methods can be found in [20]. More recently, application of this method to large scale problems was considered by Li-Zhi et al. [12]. A deficiency of using (1.2) to solve optimization problems is that it will get trapped in local minima. In order to allow the trajectory to escape from local minima, it has been proposed by various authors (e.g. [1,4,6,7,11]) to add a stochastic term that would allow the trajectory to "climb" hills. One possible augmentation to (1.2) that would enable us to escape from local minima is to add noise. One then considers the *diffusion process*:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2T(t)}dB(t). \tag{1.3}$$

Where $B(t)$ is the standard Brownian motion in $\mathbb{R}^n$. It has been shown in [4,6,7], under appropriate conditions on f, that if the *annealing schedule* is chosen as follows:

$$T(t) \triangleq \frac{c}{\log(2+t)}, \quad \text{for some } c \geq c_0, \tag{1.4}$$

where c_0 is a constant positive scalar (the exact value of c_0 is problem dependent). Under these conditions, as $t \to \infty$, the transition probability of $X(t)$ converges (weakly) to a probability measure Π. The latter, has its support on the set of global minimizers. A characterization of Π was given by Hwang in [10]. It was shown that Π is the weak limit of the following, so called, *Gibbs density*:

$$p(t,x) = \left[\exp\left\{-\frac{f(x)}{T(t)}\right\}\right]\left[\int_{\mathbb{R}^n} \exp\left\{-\frac{f(x)}{T(t)}\right\} dx\right]^{-1}. \tag{1.5}$$

Discussion of the conditions for the existence of Π, can be found in [10]. A description of Π in terms of the Hessian of f can also be found in [10].

The process in (1.3), was extended to handle linear constraints in [18]. Where the following projected SDE was studied:

$$dX(t) = P[-\nabla f(X(t)) + \mu X(t)^{-1}]dt + \sqrt{2T(t)}PdB(t), \tag{1.6}$$

where A is the matrix associated with the linear constraints, and $P = I - A^T(AA^T)^{-1}A$. The $\mu X(t)^{-1}$ term arises from the interior point methodology adopted in [18].

Theoretical work on the convergence of (1.3) to the global minimum of f, appeared in [4,6,7]. In [6], the convergence of (1.3) was established for the special case of f defined on a compact set. This assumption was lifted in [4], and in [7]. Analysis of the method when only noisy measurements of the function, and its gradient are available, was given in [5], and [11]. Numerical experience with (1.3) was reported in [1,2,15,17–19]. From numerical experience reported in [19] and [15], it appears that the method compares favorably with other stochastic methods. To the authors' knowledge, there appears to be very little work on extending this useful method to the constrained case. The contribution of this paper is to extend the algorithm described above to the general non-linearly constrained set of problems. The analysis exploits the fact that SDEs and Partial Differential Equations (PDEs), are closely intertwined. The link we shall exploit in this work is provided by *Kolmogorov's forward equation*. This PDE is also known as the *Fokker-Planck equation*. Gidas [7] first proposed to study the asymptotic behavior of the Fokker-Planck equation in order to establish the convergence of the method described above. In Sect. 4 we apply his technique of studying the asymptotic behavior of the PDE associated with an unconstrained system of SDEs, to the constrained case. Numerical integration of the projected SDE, details of our implementation and numerical results are discussed in our recent work [15,17].

We end this section by fixing some notation that will be used in the rest of this paper. The dimensions of the various gradients associated with our problem are assumed to be as follows: $\nabla g_i \in \mathbb{R}^{1\times n}$, $i = 1, \ldots, m$, $\nabla g \in \mathbb{R}^{m\times n}$, $\nabla f \in \mathbb{R}^{n\times 1}$. We will also make use of the Laplacian of the constraints:

$$\triangle g_i = \sum_{j=1}^{n} \frac{\partial^2 g_i}{\partial x_j^2}.$$

By $\triangle g \in \mathbb{R}^m$, we denote the vector of Laplacians associated with the constraints. B will be used to represent an n-dimensional Brownian motion:

$$B = \{B(t) = [B_1(t), \ldots, B_n(t)]^T,\ \mathcal{F}(t),\ t \geq 0\},$$

on the probability space $(\Omega, \Im, P)$.

2 The augmented SDE

According to the algorithmic framework adumbrated in the introduction, we could obtain a solution to our problem by following the trajectory of (1.3). In order to enforce the constraints into our SDE we propose to compute the minimum force we need to add to (1.3) so that $X(t)$ will satisfy the constraints. In other words, we need to find a $Z(t)$ such that if $X(t)$ is defined by:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2T(t)}dB(t) + Z(t)dt,$$

then $g(X(t+s)) = 0$, provided that $g(X(t)) = 0$ a.s.[1] In [18] when we only had to deal with linear constraints, Z was given by:

$$\begin{aligned} -\int_t^{t+s} Z(u)du = &\int_t^{t+s} (A^T(AA^T)^{-1}A)(-\nabla f(X(u)))du \\ &+ \int_t^{t+s} \sqrt{2T(u)}(A^T(AA^T)^{-1}A)dB(u). \end{aligned}$$

Unfortunately, for the general case, Z must depend on the current point $X(t)$, and is not easy to derive. For this reason we introduce information from the Lagrangian in order to enforce feasibility. It will be shown later (Theorem 2.2) that the Lagrange multipliers can be suitably defined so that feasibility is enforced.

The augmented SDE we study is given by:

$$dX(t) = -\nabla f(X(t)) - \nabla g(X(t))^T \lambda(X(t), t)dt + \sqrt{T(t)}dB(t) \tag{2.1}$$

[1] It will be clear from the context when a statement will hold almost surely. Henceforth, we drop the a.s qualification from relevant statements.

where:

$$\lambda(x,t) \triangleq [\nabla g(x)\nabla g(x)^T]^{-1}[g(x) + T(t) \triangle g(x) - \nabla g(x)\nabla f(x)]. \tag{2.2}$$

The intuition behind (2.1) is as follows: the first term tries to move along a direction in which f is reduced. The second term enforces feasibility. The last tern ensures that we will not get trapped in local minima.

Note that the Lagrange multipliers are similar to the ones used in gradient projection methods [13,14]. The only difference is the inclusion of the Laplacian term in order to take into consideration the second order variation from the Brownian motion. The Lagrangian at time t will be denoted by:

$$\mathcal{L}(x,t) \triangleq f(x) + \lambda(x,t)^T g(x)$$

We now state the assumption that will be used in the rest of this paper.
Assumptions

1. f, g and λ are twice continuously differentiable.
2. The matrix $\nabla g_i(x)\nabla g_i(x)^T$ is linearly independent for all x.
3. The transition measure of $X(t)$ has a density $p_t(x,y)$ i.e. that:

$$E^x[h(X_t)] = \int h(y)p_t(x,y)dy \quad h \in C^2$$

4. $\exp\left(-\frac{\mathcal{L}(x,t)}{T(t)}\right) \in \mathcal{L}^2$
5. $\mathcal{L}(x,t)$ is uniformly bounded in $(\mathbb{R}^n \times \mathbb{R})$.
6. There exists a T_0 such that:

$$\nabla g(x(t))\nabla\lambda(x(t),t) \neq I, \forall t \geq T_0,\ x(t) \in \{y \in \mathbb{R}^n \mid \nabla\mathcal{L}(y,t) = 0\}.$$

Lemma 2.1 *Let $\{\tau_m\}$ denote the sequence of first exit times of $X(t)$ from balls of radius m about the origin. As m goes to infinity, we denote the limit of $\{\tau_m\}$ by τ_∞. For the process defined in* (2.1)

$$P(\tau_\infty < \infty) = 0$$

Proof Under our assumptions, this is a direct consequence of existence and uniqueness results for SDEs, see for example Theorem 5.2.1 in [16]. □

Theorem 2.2 *There exists a t_0 such that:*

$$\lim_{t\to\infty} g_i(X(t)) = 0 \text{ } P\text{-a.s.} \quad i = 1, \ldots, m. \tag{2.3}$$

Proof From (2.1) and Itô's Lemma, and the definition of λ we have:

$$\begin{aligned} dg_i(X(t)) &= -\nabla g_i(X(t))\nabla f(X(t))dt - g_i(X(t))dt - T(t) \triangle g_i(x)dt \\ &\quad + \sqrt{2T(t)}\nabla g_i(X(t))dB(t) + T(t) \triangle g_i(x)dt + \nabla g_i(X(t))\nabla f(X(t))dt \\ &= -g_i(X(t))dt + \sqrt{2T(t)}\nabla g_i(X(t))dB(t). \end{aligned} \tag{2.4}$$

Let $\phi(t)$, denote the solution of the initial value problem:

$$d\phi(t) = \phi(t)dt, \quad \phi(0) = 1,$$

and therefore $\phi(t) = e^t$. Using Itô's Lemma, and the fact that the quadratic variation of $\phi(t)$ is zero, we have:

$$\begin{aligned} d[g_i(X(t))\phi(t)] &= g_i(X(t))d\phi(t) + \phi(t)dg_i(X(t)) \\ &= \sqrt{2T(t)}\phi(t)\nabla g_i(X(t))dB(t). \end{aligned}$$

Using the preceding equation, we obtain:

$$g_i(X(t)) = \phi(t)^{-1}\left[g_i(X(0)) + \int_0^t \sqrt{2T(\tau)}\phi(\tau)\nabla g_i(X(\tau))dB(\tau)\right].$$

Squaring and taking expectations with respect to P in the equation above, it follows by the Itô isometry that:

$$E[g_i(X(t))^2] = \phi(t)^{-2}\left[g_i(X(0))^2 + E\left[\int_0^t 2T(\tau)\phi(\tau)^2\nabla G_i(X(\tau))dt\right]\right], \quad (2.5)$$

where $\nabla G_i(x) = \sum_l \nabla g_i(x)_l^2$. By Lemma 2.1, the process $X(t)$ does not explode in finite time. Together with the continuity of ∇G_i, Lemma 2.1 implies that

$$\nabla G_i(X(t)) < c_1 \quad t < \infty,$$

for some finite scalar c_1. Let $\epsilon > 0$ be an arbitrary given constant, and choose t_1 to be large enough so that:

$$\frac{c_1\Lambda}{\ln(3+t_1)} \leq \frac{\epsilon}{3}.$$

For such a t_1 we have:

$$\begin{aligned} \int_0^t 2T(\tau)\phi(\tau)^2\nabla G_i(X(\tau))dt &= \int_0^{t_1} 2T(\tau)\phi(\tau)^2\nabla G_i(X(\tau))d\tau + \int_{t_1}^t 2T(\tau)\phi(\tau)^2\nabla G_i(X(\tau))d\tau \\ &\leq \frac{c_1\Lambda e^{2t_1}}{\ln(3)} + \frac{c_1\Lambda e^{2t}}{\ln(3+t_1)} \end{aligned}$$

Using the relationship above, we obtain the following estimate: It follows from that (2.5) that:

$$E[g_i(X(t))^2] \leq e^{-2t}g_i(X(0))^2 + \frac{c_1\Lambda e^{2(t_1-t)}}{\ln(3)} + \frac{c_1\Lambda}{\ln(3+t_1)} \quad (2.6)$$

Let $t_2,\ t_3$ be large enough so that:

$$e^{-2t_2}g_i(X(0))^2 \leq \frac{\epsilon}{3}\ ;\ \ \frac{c_1\Lambda e^{2(t_1-t)}}{\ln(3)} \leq \frac{\epsilon}{3},$$

and define $t_0 = \max\{t_2, t_3\}$. Then it follows from (2.6) that:

$$E[g_i(X(t))^2] \leq \epsilon \quad \forall t > t_0,$$

which implies the result. □

3 Properties of the density

This section builds the foundation of the convergence proof by establishing a density that, asymptotically, places positive mass only on the constrained global minima of the problem. In the next section, we show that the transition density function of (2.1) (asymptotically) behaves exactly like the density described in this section. The end result is that the stochastic process that solves (2.1) only visits the constrained global minima of the problem.

Let H be the set of global optima:

$$H = \left\{x \mid x = \arg\min_{y}\{f(y) \mid g(y) = 0\}\right\}.$$

Our aim is to introduce a probability measure Π, such that $\Pi(x) > 0$ if $x \in H$, and zero otherwise. A natural way to define such a measure is by using Laplace's method [3,10]. The latter, is a useful technique to study the asymptotic behavior of integrals. Hwang [10] was first to propose the applicability of the Laplace method for defining a probability measure with the desired properties. In [18] the method was extended to cover constrained problems. The aim here is to specialize the method to the Lagrangian formulation proposed in this paper. We start with two Lemmas that will be useful later on.

Lemma 3.1 *Let $x_l(t)$ be any local minimum of $\mathcal{L}(x,t)$, and suppose that $\lim_{t\to\infty} x_l(t) = x_l$. Then*

$$\lim_{t\to\infty} g(x_l(t)) = 0.$$

Proof Since $\mathcal{L}(x,t)$ is twice consciously differentiable and bounded, it achieves its local minimum at a point x_l such that:

$$\nabla f(x_l(t)) + \nabla g(x_l(t))^T \lambda(x_l(t), t) + \nabla\lambda(x_l(t), t) g(x_l(t)) = 0.$$

Multiplying the preceding equation with $\nabla g(x_l)$, we obtain:

$$(I - \nabla g(x_l(t))\nabla\lambda(x_l(t), t)) g(x_l(t)) = -T(t) \triangle g(x_l(t)).$$

Let $\epsilon > 0$, be any arbitrary given scalar. Let t_0 be large enough so that assumption 6 is satisfied. Then there must exist a $t_1 > t_0$ such that:

$$\begin{aligned}&|T(t) \triangle g(x_l(t))| \leq \epsilon \ \ \forall t > t_1\\ &(I - \nabla g(x_l(t))\nabla\lambda(x_l(t), t)) = c(t) \neq 0 \ \ \forall t > t_1\end{aligned}$$

Therefore,

$$\lim_{t\to\infty} \sum_i c_{ij}(x_l(t), t) g_j(x_l(t)) = 0 \ \ j = 1, \ldots, m,$$

where $c_{ij}(x,t) = [I - \nabla g(x)\nabla\lambda(x,t)]_{ij}$. By assumption 6 we must have:

$$\lim_{t\to\infty} \sum_i c_{ij}(x_l(t), t) = c_j \neq 0.$$

The preceding limit implies the result since we have:

$$0 = c_j \lim_{t\to\infty} g_j(x_l(t)).$$

□

Lemma 3.2 *Let:*

$$\mathcal{L}(x) = f(x) + \lambda(x)^T g(x),$$

where

$$\lambda_i(x) = [\nabla g_i(x)\nabla g_i(x)^T]^{-1}[g_i(x) - \nabla g_i(x)\nabla f(x)] \quad i = 1, \ldots, m.$$

Then $\hat{x}$ *is a global minimum of* (1.1) *if and only if* $\hat{x}$ *is a global minimum of* $\mathcal{L}(x)$.

Proof Suppose that $\hat{x}$ is the global minimum of $\mathcal{L}(x)$. We first show that $\hat{x}$ is feasible in (1.1). Indeed, since $\hat{x}$ is a global minim of $\mathcal{L}(x)$ it must also satisfy:

$$\nabla\mathcal{L}(\hat{x}) = \nabla f(\hat{x}) + \nabla g(\hat{x})\lambda(\hat{x}) + \nabla\lambda(\hat{x})\nabla g(\hat{x}) = 0.$$

Multiplying the preceding equation by $\nabla g(\hat{x})$, we obtain:

$$(I + \nabla\lambda(\hat{x})\nabla g(\hat{x}))g(\hat{x}) = 0.$$

By assumption $\nabla\lambda(\hat{x})\nabla g(\hat{x}) \neq -I$, which implies that $g(\hat{x}) = 0$. Moreover, since $\hat{x}$ is a global minimum of $\mathcal{L}$ we also have that:

$$f(\hat{x}) \leq f(x) + \lambda(x)^T g(x) \ \ \forall x.$$

Suppose that there exists an x_1 such that:

$$f(x_1) < f(\hat{x}) \text{ and } g(x_1) = 0.$$

Then,

$$f(\hat{x}) > f(x_1) = f(x_1) + \lambda(x_1)^T g(x_1),$$

contradicting that $\hat{x}$ is a global minimum of $\mathcal{L}$. We conclude that $\hat{x}$ must also be a global minimum of (1.1).

For the reverse direction, suppose that $\hat{x}$ is a global minimum of (1.1). Suppose that $\hat{x}$ is not a global minimum of $\mathcal{L}$, and let x_1 be any global minimum of $\mathcal{L}$. Then we must have:

$$f(x_1) + \lambda(x_1)^T g(x_1) < f(\hat{x}) + \lambda(\hat{x})^T g(\hat{x}).$$

The same argument as in the beginning of the proof can be used to show that $g(x_1) = 0$, and therefore we must have:

$$f(x_1) < f(\hat{x}) \ \ g(x_1) = 0,$$

contradicting that $\hat{x}$ is the global minimum of (1.1). □

Lemma 3.3 *Let* $x^*(t)$ *be any global minimum of* $\mathcal{L}(x, t)$*, and suppose that* $\lim_{t\to\infty} x^*(t) = x^*$. *Furthermore, assume that* $\mathcal{L}(x, t)$ *has compact support. Then,*

$$\lim_{t\to\infty} -T(t)\ln c(t) = P^*.$$

Where

$$c(t) = \int \exp\left(\frac{1}{T(t)}(f(x) + \lambda(x, t)^T g(x))\right) d\Lambda(x)$$

Proof Let $x^*(t)$ be a global minimum of $\mathcal{L}(x, t)$. Then there must exist a scalar $\beta(t) > 0$, so that:

$$\begin{aligned} H(\beta(t), t) = \Bigg\{ & x \in K \mid \beta(t) \exp\left(-\frac{1}{T(t)}(f(x^*(t)) + \lambda(x^*(t))g(x^*(t)))\right) \\ & \leq \exp\left(-\frac{1}{T(t)}(f(x) + \lambda(x)g(x))\right),\ 0 < \beta(t) < 1\Bigg\}, \end{aligned}$$

where K is some compact subset of $\mathbb{R}^n$. Using the property above, we can obtain the following estimate:

$$\begin{aligned} & \int \exp\left(-\frac{\mathcal{L}(x,t)}{T(t)}\right) d\Lambda(x) \\ & = \int\limits_{\overline{H(\beta(t),t)}} \exp\left(-\frac{\mathcal{L}(x,t)}{T(t)}\right) d\Lambda(x) + \int\limits_{H(\beta(t),t)} \exp\left(-\frac{\mathcal{L}(x,t)}{T(t)}\right) d\Lambda(x) \\ & \geq \Lambda(H(\beta(t), t))\beta(t) \exp\left(-\frac{1}{T(t)}(f(x^*(t)) + \lambda(x^*(t))g(x^*(t)))\right). \end{aligned}$$

Let $S(t)$ denote the support of $\mathcal{L}(x, t)$, which by assumption is compact. Using the estimate above, we obtain:

$$\begin{aligned} & \Lambda(H(\beta(t), t))\beta(t) \exp\left(-\frac{\mathcal{L}(x^*(t), t)}{T(t)}\right) \\ & \leq \int \exp\left(-\frac{\mathcal{L}(x,t)}{T(t)}\right) d\Lambda(x) \\ & \leq \Lambda(S(t)) \exp\left(-\frac{\mathcal{L}(x^*(t), t)}{T(t)}\right). \end{aligned}$$

The preceding equation implies that:

$$T(t) \ln \Lambda(H(\beta(t), t))\beta(t) + \mathcal{L}(x^*(t), t) \geq -T(t) \ln c(t) \geq T(t) \ln \Lambda(S(t)) + \mathcal{L}(x^*(t), t) \tag{3.1}$$

By the continuity of $\mathcal{L}(x^*(t), t)$, and Lemma 3.2 we have:

$$\lim \mathcal{L}(x^*(t), t) = \mathcal{L}(x^*) = P^*.$$

Since $\Lambda(H(\beta(t), t))\beta(t)$ and $\Lambda(S(t))$ are bounded, and $T(t)$ tends to zero, taking the limit as $t \to \infty$ in (3.1) the desired result is obtained. □

Motivated by the result above, and by the results concerning the unconstrained case from [10]. We consider the following Radon-Nikodym derivative as the starting point for our definition:

$$\frac{dP_t}{d\Lambda} = \frac{\exp\left\{-\frac{1}{T(t)}(f(x) + \lambda(x,t)^T g(x))\right\}}{\int \exp\left\{-\frac{1}{T(t)}(f(x) + \lambda(x,t)^T g(x))\right\} d\Lambda(x)}. \tag{3.2}$$

Where Λ is some probability measure on $(\mathbb{R}^n, \mathcal{B})$. The rest of this Section is devoted to the study of P_ϵ as ϵ approaches zero. The results presented here are extensions of the results from [10,18].

Lemma 3.4 *Let $\kappa(t)$ be defined as follows:*

$$\kappa(t) = \min \mathcal{L}(x,t),$$

and suppose that $\kappa(t)$ belongs to a compact set. Then the sequence P_t is tight.

Proof In order to show that the sequence is tight, we need to find a compact set K, such that:

$$P_t(K) > 1 - \delta \;\; \forall \delta > 0.$$

Let $\delta_1 > 0$ be an arbitrary given constant, then the following estimate can be obtained:

$$\begin{aligned} P_t(\mathcal{L}(x,t) > \delta_1 + \kappa(t)) &= \int\limits_{\substack{\mathcal{L}(x,t) \\ > \delta_1 + \kappa(t)}} \exp\left\{-\frac{\mathcal{L}(y,t)}{T(t)}\right\} d\Lambda(y) \left[\int \exp\left\{-\frac{\mathcal{L}(y,t)}{T(t)}\right\} d\Lambda(y)\right]^{-1} \\ &\leq \Lambda(\mathcal{L}(x,t) > \delta_1 + \kappa(t)) e^{-\frac{\delta_1 + \kappa(t)}{T(t)}} \left[\int \exp\left\{-\frac{\mathcal{L}(y,t)}{T(t)}\right\} d\Lambda(y)\right]^{-1} \\ &\leq \left[\int \exp\left\{-\frac{\mathcal{L}(y,t) - \delta_1 - \kappa(t)}{T(t)}\right\} d\Lambda(y)\right]^{-1} \\ &\leq \left[\int\limits_{\substack{\mathcal{L}(x,t) \\ > \delta_1 + \kappa(t)}} \exp\left\{-\frac{\mathcal{L}(y,t) - \delta_1 - \kappa(t)}{T(t)}\right\} d\Lambda(y)\right]^{-1}. \end{aligned}$$

The preceding equation implies that:

$$\lim_{t \to \infty} P_t(\mathcal{L}(x,t) > \delta_1 + \kappa(t)) = 0.$$

Note that the set:

$$K(t) = \{x \mid \mathcal{L}(x,t) \leq \delta_1 + \kappa(t)\},$$

is compact. Moreover since $\kappa(t)$ belongs to some compact set, there must exist a κ such that, $\kappa(t) \leq \kappa$. The we can define a compact set K as follows:

$$K = \{x \mid \mathcal{L}(x,t) \leq \delta_1 + \kappa\},$$

Therefore, given any δ, then there exists a T, such that:

$$P_t(K(t)) \geq 1 - \delta_2 \quad \forall t > T.$$

Noting that $K(t) \subseteq K$, we also have:

$$P_t(K) \geq P_t(K(t)) \geq 1 - \delta_2 \quad \forall t > T,$$

as required. □

Corollary 3.5 *$\{P_\epsilon\}$ has a subsequence that weakly converges to Π, and the latter has its support in H.*

Proof The first part follows from Proposition 3.4, and Prokhorov's theorem. The second part is a consequence of Proposition 3.3. □

It is possible to get an explicit equation for P_ϵ. Such an analysis was carried out in [18]. Since the explicit form of P_ϵ is not used in the proof we just state the Theorem. The same proof found in [18] can easily be adapted to fit the assumptions of this paper.

Proposition 3.6 *Assume that H, the set of constrained global minima of* (1.1)*, consists of a finite number of points:*

$$H = \{x_1, x_2, \ldots, x_m\}.$$

Suppose that each $x_i \in H$, satisfies the second order KKT conditions for local optimality:

$$\begin{aligned}
&\nabla f(x_i) + \nu_i^T \nabla g(x_i) = 0\\
&\nu_{ij} g_j(x_i) = 0 \quad j = 1, \ldots, l\\
&g(x_i) \le 0 \quad \nu_i \ge 0\\
&d^T \left[\nabla^2 f(x_i) + \sum_{j=1}^{l} \nu_{ij} \nabla^2 g(x_i) \right] d > 0\\
&\forall d \in M = \{y \mid \nabla g(x_i)^T y = 0, \ \} \quad i = 1, \ldots, m,
\end{aligned}$$

where ν_i denotes the Lagrange multiplier vector associated with the ith minimizer. Let dx denote the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B})$, and suppose that there exists a continuous function h such that $h(x_i) > 0$, and $\frac{d\Lambda}{dx}(\cdot) = h(\cdot)$. Then:

$$\lim_{\epsilon \downarrow 0} P_\epsilon(x_i) = \frac{h(x_i) \det \left[\nabla^2 f(x_i) + \sum_{j=1}^{l} \nu_{ij} \nabla g_j(x_i) \right]^{-1/2}}{\sum_{k=1}^{m} h(x_k) \det \left[\nabla^2 f(x_k) + \sum_{j=1}^{l} \nu_{kj} \nabla g_j(x_k) \right]^{-1/2}} \triangleq \pi(x_i).$$

4 Convergence

In this section we complete the convergence proof of the proposed method. The argument is based on the asymptotic analysis of the Fokker-Planck PDE (see (4.1)). This type of analysis was initiated by Gidas in [7–9]. The results given here are based on the unconstrained proof given in [9], and the linearly constrained case in [18].

It is well known that, under some regularity assumptions, the transition density function of (2.1) satisfies the following PDE:

$$\begin{aligned}
\frac{\partial p(x,t)}{\partial t} &= \mathcal{A}^* p(x,t) = \sum_i \frac{\partial}{\partial x_i} \left[\left(\frac{\partial f(x)}{\partial x_i} + \sum_j \frac{\partial g_j(x)}{\partial x_i} \lambda_j(x,t) \right) p(x,t) \right]\\
&\quad + T(t) \sum_i \frac{\partial^2 p(x,t)}{\partial x_i^2} \qquad (4.1)
\end{aligned}$$

This PDE is known as the Fokker-Planck or Kolmogorov's forward equation. We will assume that the process commenced from a feasible point. We therefore use the following initial condition:

$$\lim_{t \downarrow 0} p(x_0, 0, x_t, t) = \delta(x - x_0), \quad x_0 \in \mathcal{F}.$$

In addition to the forward PDE, the SDE in (2.1) has the following (infinitesimal) generator associated with its dynamics:

$$\mathcal{A}p(x,t)=\sum_i\left(\frac{\partial f(x)}{\partial x_i}+\sum_j\frac{\partial g_j(x)}{\partial x_i}\lambda_j(x,t)\right)\frac{\partial p(x,t)}{\partial x_i}+T(t)\sum_i\frac{\partial^2 p(x,t)}{\partial x_i^2} \tag{4.2}$$

Where note that $\mathcal{A}^*$ is the adjoint of $\mathcal{A}$.

Motivated by the results of the previous sections we will use the following density as a trial solution to our PDE:

$$\theta(x,t)=\frac{\exp\left(-\frac{1}{T(t)}\left(f(x)+\lambda(x,t)^T g(x)\right)\right)}{\mathcal{Z}(t)}, \tag{4.3}$$

where,

$$\mathcal{Z}(t)=\int\exp\left(-\frac{1}{T(t)}\left(f(y)+\lambda(y,t)^T g(y)\right)\right)dy. \tag{4.4}$$

The aim is to show that, asymptotically, (4.3) solves (4.1). For later use note that:

$$\frac{\partial\theta}{\partial x_i}=-\frac{1}{\mathcal{Z}(T)T(t)}\left(\frac{\partial f}{\partial x_i}+\sum_j\frac{\partial g_j}{\partial x_i}\lambda_j(x,t)+\sum_j\frac{\partial\lambda_j}{\partial x_i}g_j(x)\right) \tag{4.5}$$

Finally, the CDF of p will be denoted by:

$$\Pi(x_s,s,B,t)=\int_B p(x_s,s,y,t)dy.$$

where p is the solution to (4.1).

Lemma 4.1 *Suppose that $k,\ h\in C^2$, and that they have their support in $\mathcal{F}$. Then, on $(L^2,\theta dx)$, the following holds:*

$$\langle k,Ah\rangle=-T(t)\int\theta(t,x)\sum_i\frac{\partial k}{\partial x_i}\frac{\partial h}{\partial x_i}dx$$

Proof Note that

$$\begin{aligned}&\frac{T(t)}{\theta(t,x)}\sum_i\frac{\partial}{\partial x_i}\left(\theta(t,x)\frac{\partial h}{\partial x_i}\right)+\sum_i\frac{\partial h}{\partial x_i}\left(\sum_j\frac{\partial\lambda_j}{\partial x_i}g_j(x)\right)\\&=T(t)\sum_i\frac{\partial^2 h}{\partial x_i^2}-\sum_i\frac{\partial h}{\partial x_i}\left(\frac{\partial f}{\partial x_i}+\sum_j\frac{\partial g_j}{\partial x_i}\lambda_j\right)\\&=\mathcal{A}h.\end{aligned} \tag{4.6}$$

Where to obtain the above relationship we used (4.5). It follows that:

$$\begin{aligned}\langle k,\mathcal{A}h\rangle=&\int T(t)k(x)\sum_i\frac{\partial}{\partial x_i}\left(\theta(t,x)\frac{\partial h}{\partial x_i}\right)dx\\&+\int k(x)\left(\sum_{ij}\frac{\partial h}{\partial x_i}\frac{\partial\lambda_j}{x_i}g_j(x)\right)\theta(t,x)dx.\end{aligned} \tag{4.7}$$

The second integral in the equation must be zero, since k has its support in $\mathcal{F}$. Integrating by parts the first integral, and noting that k has its support in a compact set, we obtain:

$$\langle k, \mathcal{A}h \rangle = -T(t) \int \theta(t, x) \sum_i \frac{\partial k}{\partial x_i} \frac{\partial h}{\partial x_i} dx, \tag{4.8}$$

as required. □

Lemma 4.2 *Suppose that p satisfies the PDE in* (4.1)*. Then for t large enough, there exists a function with its support in $\mathcal{F}$, such that p can be written as:*

$$p(t, x) = \eta(t, x)\theta(t, x).$$

Proof From Theorem 2.2 it follows that, for t large enough, $\pi(t, B)$ will have its support in $\mathcal{F}$. Let, $\eta(t, x) = p(t, x)/\theta(t, x)$. Clearly, for t large enough, η will have its support in $\mathcal{F}$. Therefore, the solution to (4.1), can be written as: $p(t, x) = \eta(t, x)\theta(t, x)$. □

Lemma 4.3 *Let $\eta(t, x) \in C^2$ be any function with its support in $\mathcal{F}$. Then,*

$$\int \theta(t, x) \sum_i \left(\frac{\partial \eta}{\partial x_i} \right)^2 dx \geq \Theta(t) \mathcal{Q}(K(t)) \int \eta(t, x)^2 dx,$$

where $0 < \Theta(t) < \infty$, and $\mathcal{Q}$, is some strictly positive constant depending on a compact set $K(t)$.

Proof Since, $\eta(t, x)$ vanishes outside a compact set, it follows that its derivative also vanishes outside a compact set. Let the latter set be denoted by $K_1(t)$, and let $K(t) = K_1(t) \cup \mathcal{F}$. Let,

$$\Theta(t) = \min_{x \in K(t)} \theta(t, x),$$

and note that, $0 < \Theta(t) < \infty$. The following bound can then be obtained:

$$\begin{aligned}\int \theta(t, x) \sum_i \left(\frac{\partial \eta}{\partial x_i} \right)^2 dx &\geq \Theta(t) \int_{K(t)} \sum_i \left(\frac{\partial \eta}{\partial x_i} \right)^2 dx \\ &\geq \Theta(t) \mathcal{Q}(K(t)) \int_{K(t)} \eta(t, x)^2 dx \\ &= \Theta(t) \mathcal{Q}(K(t)) \int \eta(t, x)^2 dx,\end{aligned}$$

where the second inequality was obtained by an application of the Poincare inequality. □

Lemma 4.4 *Let p denote the solution of* (4.1)*, and let η be a C^2 function with its support in $\mathcal{F}$. Then:*

$$\theta(t, x)^{-1} \mathcal{A}^* p = \mathcal{A}\eta + N(t, x), \tag{4.9}$$

where:

$$\begin{aligned} N(t,x) = &-2\sum_{ij}\frac{\partial\lambda_j}{\partial x_i}g_j(x) - \eta(t,x)\sum_{ij}\frac{\partial g_j}{\partial x_i}\frac{\partial\lambda_j}{\partial x_i} \\ &-\sum_{ij}\frac{\partial^2\lambda_j}{\partial x_i}g_j(x)\eta(t,x) + \frac{\eta(t,x)}{T(t)}\sum_{ij}\left(\frac{\partial\lambda_j}{\partial x_i}g_j(x)\right)^2 \\ &-\frac{\eta(t,x)}{T(t)}\sum_i\left(\frac{\partial f}{\partial x_i}+\sum_j\frac{\partial g_j}{\partial x_i}\lambda_j(x,t)\right)\left(\sum_j\frac{\partial\lambda_j}{\partial x_i}g_j(x)\right) \end{aligned}$$

Moreover,

$$\int N(t,y)dy = 0.$$

Proof Let p, and θ satisfy the conditions stated in the Theorem, and note that from Lemma 4.2 we have the following:

$$p(t,x) = \eta(t,x)\theta(t,x). \tag{4.10}$$

This proof is by direct calculation and we omit the details. In order to obtain the required expression, one proceeds as follows: calculate the first and second derivatives of θ with respect to x, and substitute the formulae obtained into $\theta(t,x)^{-1}\mathcal{A}^*p$. The (rather long) resulting expression can be simplified to (4.9) when one uses the parametrization in (4.10).

To prove the last statement we also proceed by direction calculation using integration by parts. □

Theorem 4.5 *Supposed that the diffusion term in* (2.1) *is given by:*

$$T(t) = \frac{c}{\ln(3+t)}, \tag{4.11}$$

for some constant c. Furthermore, let p denote the solution of (4.1)*, so that:*

$$\Pi(x_0,0,B,t) = \int_B p(x_0,0,y,t)dy,$$

where B is any Borel set in $\mathbb{R}^n$. Let r be any bounded and continuous function, then

$$\lim_{t\to\infty}\int_B r(x)p(x_0,0,y,t)dy = \int_B r(x)\theta(t,y)dy,$$

where θ is the density function of a probability measure that assigns all its mass to the global minima of (1.1).

Proof Let p, and θ satisfy the conditions stated in the Theorem, and note that from Lemma 4.2 we have the following:

$$p(t,x) = \eta(t,x)\theta(t,x). \tag{4.12}$$

Using the relationship above, the following bound can be derived:

$$\begin{aligned}\left|\int r(y)p(t,y)dy-\int r(y)\theta(t,y)dy\right| &= \left|\int r(y)[\theta(t,y)\eta(t,y)-\theta(t,y)]dy\right|\\ &= \left|\int r(y)\sqrt{\theta(t,y)}\sqrt{\theta(t,y)}[\eta(t,y)-1]dy\right|\\ &\le \left(\int r(y)^2\theta(t,y)dy\right)^{\frac{1}{2}}\left(\int \theta(t,y)(\eta(t,y)-1)^2dy\right)^{\frac{1}{2}}.\end{aligned}$$

Let:

$$\Delta(t) \triangleq \int \pi(t,y)(\eta(t,y)-1)^2dy.$$

The result would follow if $\Delta(t)$ vanishes for t large enough. This is what we show next. The plan of the proof is to find a useful form for the derivative of $\Delta(t)$. We then show that, for large t, this derivative is negative. We well then show, that this condition alone is enough to guarantee that $\Delta(t)$ will eventually reach its lower bound. The lower bound is by construction zero.

In order to implement the plan outlined above, we will first need to place the derivative of $\Delta(t)$ in a convenient form. We will need the following derivatives:

$$\frac{d\lambda_i(x,t)}{dt} = c_i(x,t)\frac{dT(t)}{dt},$$

where $c_i(x) = [(\nabla g(x)\nabla g(x)^T)^{-1} \triangle g(x)]_i,\ i = 1,\ldots,m$. It can easily be seen that:

$$\frac{d\mathcal{Z}(t)}{dt} = -\frac{1}{T(t)^2}\frac{dT}{dt}\int \Gamma(t,x)\exp(\mathcal{L}(y,t))dy,$$

where,

$$\Gamma(t,x) = -\mathcal{L}(x,t) + T(t)c(x,t)^T g(x).$$

By direct calculation we find:

$$\frac{\partial\theta}{\partial t} = \frac{dT}{dt}\frac{\theta(t,x)}{T(t)^2}(\langle\Gamma(t)\rangle - \Gamma(t,x))$$

where,

$$\langle\Gamma(t)\rangle = \frac{1}{\mathcal{Z}(t)}\int \Gamma(t,y)\exp\left(\frac{\mathcal{L}(y,t)}{T(t)}\right)dy.$$

It follows from (4.12), that:

$$\Delta(t) = \int \theta(t,y)\eta(t,y)^2dy - 1.$$

Using the preceding equation, and Lemma 4.4, the following estimate can be made:

$$\begin{aligned}\frac{1}{2}\frac{d\Delta}{dt} &= \int \theta(t,y)\eta(t,y)\frac{\partial\eta}{\partial t}dy - \frac{1}{2}\int \eta(t,y)^2\frac{\partial\theta(t,y)}{\partial t}dy\\ &= \int \theta(t,y)\eta(t,y)\mathcal{A}\eta(t,y) + N(t,y)dy\\ &\quad -\frac{1}{2T(t)^2}\frac{dT}{dt}\int \theta(t,y)\eta(t,y)^2(\langle\Gamma(t)\rangle - \Gamma(t,y))dy \end{aligned} \tag{4.13}$$

Using Lemma 4.3, on the weighted space $(\mathcal{R}^n, \theta dx)$, we can bound the first term above as follows:

$$\begin{aligned}\int \theta(t,y)\eta(t,y)\mathcal{A}\eta(t,y)dy &= -T(t)\int \theta^2(t,y)\sum_i\left(\frac{\partial\eta}{\partial x_i}\right)^2 dy\\ &\leq -c_1 T(t)\int \eta(t,y)^2\theta(t,y)dy\\ &\leq -c_1 T(t)(\Delta(t)+1)\\ &\leq -c_1 T(t)\Delta(t)\end{aligned} \tag{4.14}$$

For some constant c_1 that only depends on $\mathcal{F}$. The second term in (4.13) can be reorganized as follows:

$$\begin{aligned}&\frac{1}{2T(t)^2}\frac{dT}{dt}\int \theta(t,y)\eta(t,y)^2(\langle\Gamma(t)\rangle-\Gamma(t,y))dy\\ &=\frac{1}{2T(t)^2}\frac{dT}{dt}\int(\langle\Gamma(t)\rangle-\Gamma(t,y))\theta(t,y)(\eta(t,y)-1)^2dy\\ &\quad+\frac{1}{T(t)^2}\frac{dT}{dt}\int(\langle\Gamma(t)\rangle-\Gamma(t,y))(\eta(t,y)\theta(t,y)-1)dy\\ &\leq\frac{1}{T(t)^2}\frac{dT}{dt}\left(c_2\int\theta(t,y)(\eta(t,y)-1)^2dy+c_3\int\theta(t,y)|\eta(t,y)-1|dy\right)\\ &\leq\frac{c_4}{T(t)^2}\frac{dT}{dt}\left(\int\theta(t,y)(\eta(t,y)-1)^2dy+\sqrt{\int\theta(t,y)(\eta(t,y)-1)^2dy}\right)\\ &=\frac{c_4}{T(t)^2}\frac{dT}{dt}\left(\Delta(t)+\sqrt{\Delta(t)}\right)\end{aligned} \tag{4.15}$$

Using (4.14) and (4.15) in (4.13) we finally obtain:

$$\frac{\partial\Delta(t)}{\partial t}\leq -c_1T(t)\Delta(t)-\frac{c_4}{T(t)^2}\frac{dT}{dt}\left(\Delta(t)+\sqrt{\Delta(t)}\right). \tag{4.16}$$

It follows that for t large enough we must have:

$$\frac{\partial\Delta(t)}{\partial t}\leq 0$$

It follows that if t is large enough $\Delta(t)$ is a monotonically decreasing sequence that is bounded from below. Note that we made use of the fact that $T(t)$ has the functional form given by (4.11), and that its first derivative is negative. Eventually, for some t_c we must have:

$$\frac{\partial\Delta(t_c)}{\partial t}=0.$$

It follows that,

$$\frac{\partial\Delta(t)}{\partial t}=0.\forall t\geq t_c$$

then,

$$0\leq\Delta(t)\leq\frac{c_4}{T(t)^3}\frac{dT}{dt}\left(\Delta(t)+\sqrt{\Delta(t)}\right)\leq\frac{c_5}{T(t)^3}\frac{dT}{dt}$$

The r.h.s. goes to zero as $t\rightarrow\infty$. Therefore for t large enough Δ is zero, as required. □

5 Conclusions

We discussed a stochastic algorithm for the solution of a general class of global optimization problems. We showed that the transition density associated with the stochastic differential equation used to specify the algorithm converges weakly to a density that assigns positive mass only to the global minima of the problem. The numerical integration the SDE, details of our implementation and numerical results can be found in [15,17].

Acknowledgements The work of the first author was supported by an EPSRC grant GR/T02560/01. The authors also wish to acknowledge helpful discussions during the Global Optimization workshop.

References

1. Aluffi-Pentini, F., Parisi, V., Zirilli, F.: Global optimization and stochastic differential equations. J. Optim. Theory Appl. **47**(1), 1–16 (1985)
2. Aluffi-Pentini, F., Parisi, V., Zirilli, F.: A global optimization algorithm using stochastic differential equations. ACM Trans. Math. Softw. **14**(4):345–365 (1989)
3. Bender, C.M., Orszag, S.A.: Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory. Springer-Verlag, Berlin (1999)
4. Chiang, T.S., Hwang, C.R., Sheu, S.J.: Diffusion for global optimization in $\mathbf{R}^n$. SIAM J. Control Optim. **25**(3), 737–753 (1987)
5. Gelfand, S.B., Mitter, S.K.: Recursive stochastic algorithms for global optimization in $\mathbf{R}^d$. SIAM J. Control Optim. **29**(5), 999–1018 (1991)
6. Geman, S., Hwang, C.R.: Diffusions for global optimization. SIAM J. Control Optim. **24**(5), 1031–1043 (1986)
7. Gidas, B.: The Langevin equation as a global minimization algorithm. In: Disordered Systems and Biological Organization (Les Houches, 1985), Vol. 20 of NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci., pp. 321–326. Springer, Berlin (1986)
8. Gidas, B.: Simulations and global optimization. In: Random Media (Minneapolis, Minn., 1985), Vol. 7 of IMA Vol. Math. Appl., pp. 129–145. Springer, New York (1987)
9. Gidas, B.: Metropolis-type Monte Carlo simulation algorithms and simulated annealing. In: Topics in Contemporary Probability and Its Applications, Probability Stochastics Series, pp. 159–232. CRC, Boca Raton, FL (1995)
10. Hwang, C.R.: Laplace's method revisited: weak convergence of probability measures. Ann. Probab. **8**(6), 1177–1182 (1980)
11. Kushner, H.J.: Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo. SIAM J. Appl. Math. **47**(1), 169–185 (1987)
12. Li-Zhi, L., Liqun, Q., Hon, W.T.: A gradient-based continuous method for large-scale optimization problems. J. Glob. Optim. **31**(2), 271 (2005)
13. Luenberger, D.G.: The gradient projection method along geodesics. Manag. Sci. **18**, 620–631 (1972)
14. Luenberger, D.G.: Linear and Nonlinear Programming. 2nd edn. Kluwer Academic Publishers, Boston (2003)
15. Maringer, D., Parpas, P.: Global optimization of higher order moments in portfolio selection. J. Glob. Optim. doi: 10.1007/s10898-007-9224-3
16. Oksendal, B.: Stochastic Differential Equations, an Introduction with Applications, 6th edn. Springer, New York
17. Parpas, P., Rustem, B., Pistikopoulos, E.N.: Global optimization of robust chance constrained problems. J. Glob. Optim. doi: 10.1007/s10898-007-9244-z
18. Parpas, P., Rustem, B., Pistikopoulos, E.N.: Linearly constrained global optimization and stochastic differential equations. J. Glob. Optim. **36**(2), 191–217 (2006)
19. Recchioni, M.C., Scoccia, A.: A stochastic algorithm for constrained global optimization. J. Glob. Optim. **16**(3), 257–270 (2000)
20. Zirilli, F.: The use of ordinary differential equations in the solution of nonlinear systems of equations. In: Nonlinear Optimization, 1981 (Cambridge, 1981), NATO Conference Series II: Systems Science, pp. 39–46. Academic Press, London (1982)