On fraudulent stochastic algorithms

Laurent Miclo

Toulouse School of Economics
Institut de Mathématiques de Toulouse

Plan of the talk

Plan

Global optimization

Problem of the global minimization of a function $U : M \to \mathbb{R}$.

Here: M is a compact Riemannian manifold of dimension $m \geqslant 1$, U is smooth.

Denote

$$\mathcal{U} := \left\{ x \in M : U(x) = \min_M U \right\}$$

We are happy if we find points close to $\mathcal{U}$.

The simplest generic approach: simulated annealing. More sophisticated methods are based on interacting particles. When U has particular features, there are more specific algorithms: gradient descent or Newton's method for convex optimisation, moment method for polynomial optimisation, ...

Simulated annealing

Consider the (time-inhomogeneous) stochastic algorithm

$$dZ(t) \quad = \quad -\gamma_t \nabla U(Z(t))\, dt + \sqrt{2}\, dB(t)$$

where $B(t)$ is a M-valued Brownian motion.

Appropriate inverse temperature schemes $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ lead to convergence in probability toward the global minima: for any neighborhood $\mathcal{N}$ of $\mathcal{U}$,

$$\lim_{t \to +\infty} \mathbb{P}[Z(t) \in \mathcal{N}] \quad = \quad 1$$

Almost sure convergence does not hold in general.

Fraudulent algorithms

The above algorithms do not require the knowledge of the minimal value of U.

Fraudulent algorithms: require $\min_M U$.

• Cheating? Folklore: to know the minimal value of a function is equivalent to know a global minimum.

• Interests:
– Useful to find other global minima, once one is known.
– Approximation of the large-time limit behavior of the mean-field swarm algorithms.
– Suggests the design of non-fraudulent interacting particles systems, evaluating on-line the minimal value.

A fraudulent algorithms

Assume that U is a Morse function and that $\min_M U = 0$.

Consider the (time-homogeneous) stochastic algorithm whose evolution is driven by

$$dX(t) \quad = \quad -\beta \nabla U(X(t))\, dt + \sqrt{2U(X(t))}\, dB(t) \qquad (1)$$

where a priori $\beta \in \mathbb{R}$.

Well-defined, fraudulent and $\mathcal{U}$ is absorbing.

The associated generator is

$$L_\beta \quad := \quad U\triangle \cdot - \beta \langle \nabla U, \nabla \cdot \rangle$$

Plan

Eigenvalues

Under the compactness and Morse assumptions, $\mathcal{U}$ consists of a finite set of points, say $y_1, y_2, ..., y_N$, with $N \in \mathbb{N}$. For each $n \in [\![N]\!] := \{1, 2, ..., N\}$, denote $\lambda_1(n) \leqslant \lambda_2(n) \leqslant \cdots \leqslant \lambda_m(n)$ the positive eigenvalues of the Hessian of U at y_n. Introduce the condition

$$\beta > \max_{n \in [\![N]\!]} \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_1(n)} - 1 \tag{2}$$

The main result

Theorem 1

Whatever the initial condition $X(0)$, *under* (2) *the limit* $X(\infty) := \lim_{t\to+\infty} X(t)$ *exists a.s. and belongs to* $\mathcal{U}$. *Furthermore, when* $m \geqslant 2$, *if* X *starts from a point* $x_0 \notin \mathcal{U}$, *we have for any* $y \in \mathcal{U}$,

$$\mathbb{P}_{x_0}[X(\infty) = y] \quad > \quad 0$$

When $m = 1$, *denote* y_1 *and* y_2 *the boundary points of the connected component of* $M\backslash\mathcal{U}$ *containing* x_0 *(when* $\mathcal{U}$ *is a singleton, we get* $y_1 = y_2$*). Then we have*

$$\mathbb{P}_{x_0}[X(\infty) = y_1] \ > \ 0, \quad \mathbb{P}_{x_0}[X(\infty) = y_2] \ > \ 0,$$
$$\forall\ y \in \mathcal{U}\backslash\{y_1, y_2\}, \quad \mathbb{P}_{x_0}[X(\infty) = y] = 0$$

A converse result

Theorem 2

Assume that

$$\beta < \min_{n \in [\![N]\!]} \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_m(n)} - 1 \tag{3}$$

Whatever the initial condition $X(0)$, we have

$$\mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] = 0$$

A "critical" consequence

Corollary 3

Assume that for each $n \in \mathbb{N}$ we have $\lambda_1(n) = \lambda_m(n)$, then we have

$$\beta > \frac{m}{2} - 1 \quad \Rightarrow \quad \mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] = 1$$

$$\beta < \frac{m}{2} - 1 \quad \Rightarrow \quad \mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] = 0$$

In particular, in dimension 1 (corresponding to the circle), the critical value is $\beta = -1/2$.

Plan

The underlying idea

For any given global minimum y_n, with $n \in [\![N]\!]$, we can find a small radius $r_n > 0$ such that inside the ball $B(y_n, r_n)$, the evolution of $U(X(t))$ is comparable to a time-changed Bessel process with negative dimension, as soon as

$$\beta > 1 + \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_1(n)} \tag{4}$$

Since a Bessel process with negative dimension a.s. converges to zero in finite time, with positive probability, X stays in $B(y_n, r_n)$ forever and then converges to y_n, if it belongs to an appropriate neighborhood of y_n at some time. Nevertheless, the convergence is not expected to occur in finite time due to the time-change.

But this leads to a weaker version of Theorem 1: only for

$$\beta > 1 + \sup_{n \in [\![N]\!]} \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_1(n)} \tag{5}$$

Martingales problems

To avoid the geometric considerations under the stochastic differential equation (1), it is better to resort to the synthetic martingale problems.

For any smooth function $f : M \to \mathbb{R}$, the process $M^f := (M^f(t))_{t \geqslant 0}$ defined by

$$\forall\ t \geqslant 0, \qquad M^f(t) \ := \ f(X(t)) - f(X(0)) - \int_0^t L_\beta[f](X(s))\, ds$$

is a continuous martingale. Its bracket is given by

$$\forall\ t \geqslant 0, \qquad \left\langle M^f \right\rangle_t \ = \ 2 \int_0^t U(X(s)) \left\| \nabla f(X(s)) \right\|^2 ds$$

Time-changes

Consider the time change $(\tau_t^f)_{t\in[0,\varsigma)}$ uniquely defined through

$$\forall\ t \in [0, \varsigma^f), \qquad 2\int_0^{\tau_t^f} U(X(s))\,\|\nabla f(X(s))\|^2\ ds \quad = \quad t$$

where

$$\begin{aligned}
\varsigma^f &:= 2\int_0^{\sigma^f} U(X(s))\,\|\nabla f(X(s))\|^2\ ds \\
\sigma^f &:= \inf\{t \geqslant 0\ :\ U(X(t))\,\|\nabla f(X(t))\|^2 = 0\}
\end{aligned}$$

Define the process Y^f via

$$(Y^f(t))_{t\in[0,\varsigma^f)} \quad := \quad (f(X(\tau_t^f)))_{t\in[0,\varsigma^f)}$$

Evolution of Y^f

We compute:

$$dY^f(t) = \frac{1}{2}F^f(Y^f(t))\,dt + dW(t)$$

where for any $x \in M$ such that $U(x)\|\nabla f(x)\|^2 \neq 0$,

$$\begin{aligned} F^f(x) &:= \frac{L_\beta[f](x)}{U(x)\|\nabla f(x)\|^2} \\ &= \frac{\triangle f(x)}{\|\nabla f(x)\|^2} - \beta\frac{\langle \nabla U(x), \nabla f(x)\rangle}{U(x)\|\nabla f(x)\|^2} \end{aligned}$$

Particular functions f

To improve (5) and end up with (2), consider $f = U^a$, with $a \in (0, 1]$, instead of $f = U$.

It appears that

$$F^{U^a} = \frac{1}{a}\left(\frac{U \triangle U}{\|\nabla U\|^2} + a - 1 - \beta\right)\frac{1}{U^a}$$

Thus to get a comparison of the process Y^{U^a} with a Bessel process, we need to bound the ratio $\frac{U \triangle U}{\|\nabla U\|^2}$. Let us do it in sufficiently small neighborhoods around the global minima.

Local expansions

Fix a global minimum y_n. Consider an exponential system of coordinates $(x_1, x_2, ..., x_m)$ on a neighborhood of y_n with $x_1(y_n) = x_2(y_n) = \cdots = x_m(y_n) = 0$ and such that the vectors $(\partial_{x_i})_{i \in [\![m]\!]}$ forms an orthonormal basis of the tangent space at y_n consisting of eigenvectors of the Hessian of U respectively to the eigenvalues $(\lambda_i(n))_{i \in [\![m]\!]}$. For any given $\epsilon \in (0, 1)$, we can further reduce the neighborhood to a ball $B(y_n, r_n)$, such that for any $x \in B(y_n, r_n)$ (identified with its coordinates $(x_1, x_2, ..., x_m)$),

$$(1-\epsilon) \sum_{i \in [\![m]\!]} \lambda_i(n) \leqslant \triangle U(x) \leqslant (1+\epsilon) \sum_{i \in [\![m]\!]} \lambda_i(n)$$

$$(1-\epsilon) \sum_{i \in [\![m]\!]} \lambda_i^2(n) x_i^2 \leqslant \|\nabla U\|^2(x) \leqslant (1+\epsilon) \sum_{i \in [\![m]\!]} \lambda_i^2(n) x_i^2$$

$$\frac{1}{2}(1-\epsilon) \sum_{i \in [\![m]\!]} \lambda_i(n) x_i^2 \leqslant U(x) \leqslant \frac{1}{2}(1+\epsilon) \sum_{i \in [\![m]\!]} \lambda_i(n) x_i^2$$

End of the argument

We deduce an asymptotic comparison, as ϵ goes to zero, of Y^{U^a} with a Bessel process of dimension

$$\delta_a \quad = \quad 2 + \frac{1}{a}\left(\frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_1(n)} - 1 - \beta\right)$$

when Y^{U^a} is close to y_n.

Under (2), whatever $n \in [\![N]\!]$, the term inside the parenthesis is negative, thus by choosing $a > 0$ sufficiently small, we get a negative dimension, leading to the wanted result.

Theorem 2 uses reversed bounds and the fact that Bessel processes of dimension larger than 2 diverge to $+\infty$ (without hitting 0 when they start from a positive value).

Plan

Varying β

In practice (2) may be difficult to check. An alternative to the appropriate choice of the coefficient β is to let it depend on the current value of U. Consider $\zeta : (0, +\infty) \to (0, +\infty)$ a smooth function such that

$$\lim_{u \to 0_+} \zeta(u) = +\infty \tag{6}$$

$$\lim_{u \to 0_+} \sqrt{u}\zeta(u) = 0 \tag{7}$$

We replace the generator L_β by

$$L_\zeta := U\triangle \cdot - \zeta(U)\langle \nabla U, \nabla \cdot \rangle \tag{8}$$

with the convention $\zeta(U(x))\nabla U(x) = 0$ for $x \in \mathcal{U}$.

Extension of Theorem 1

This generator corresponds to the Itô stochastic differential equation

$$dX(t) \quad = \quad -\zeta(U(X(t))\nabla U(X(t))\,dt + \sqrt{2U(X(t))}\,dB(t)$$

It can be shown that whatever the initial distribution, there is a unique strong diffusion X associated to (8) in the sense of martingale problem.

For this process X, Theorem 1 still holds, without Assumption (2).

More general sets $\mathcal{U}$

The Morse assumption on U can be relaxed (the non-degeneracy of the Hessians was only used on $\mathcal{U}$).

Typically when $\mathcal{U}$ consists of finite number of connected and disjoint submanifolds, say $\mathcal{U}_1$, $\mathcal{U}_2$, ..., $\mathcal{U}_N$, with non-degenerate Hessians of U in the orthogonal directions. The simplest case is when for each $n \in [\![N]\!]$, orthogonal vector fields on $\mathcal{U}_n$ corresponding to the eigenvectors of the orthogonal Hessians can be found.

Plan

A non-linear p.d.e.

Consider the non-linear evolution equation

$$\frac{d}{dt}\rho_t \quad = \quad \mathrm{div}(\rho_t[\gamma_t \nabla U + \nabla \varphi'(\rho_t)]) \tag{9}$$

where

- ρ_t is a probability density with respect to the Riemannian probability ℓ on M,
- $(\gamma_t)_{t \geqslant 0}$ is an inverse temperature scheme, assumed to be smooth and to increase to $+\infty$ in large times,
- $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly convex function satisfying $\varphi(1) = 0$ and is $\mathcal{C}^2$ on $(0, +\infty)$.

Gradient descent

At any given time $t \geqslant 0$, this evolution corresponds to an instantaneous gradient descent on the Wasserstein space $\mathcal{P}(M)$ with respect to the functional

$$\rho \quad \mapsto \quad \gamma_t \int_M U\rho \, d\ell + \int_M \varphi(\rho) \, d\ell$$

where

- the term $\int_M U\rho \, d\ell$ should be seen as an up-lift from M to $\mathcal{P}(M)$ of the mapping U,
- the last term is a penalized cost.

As soon as $\varphi'(0) = -\infty$, there exists a unique associated stationary density μ_{γ_t}.

Non-linear diffusion

A non-linear diffusion $Y := (Y(t))_{t \geqslant 0}$ is associated to (9), whose evolution is described by

$$dY(t) \quad = \quad -\gamma_t \nabla U(Y(t)) + \sqrt{2\alpha(\rho_t(Y(t)))}\, dB(t)$$

where

- ρ_t is the density of the law of $Y(t)$,
- the function $\alpha : (0, +\infty) \to \mathbb{R}_+$ is given by

$$\forall\ r > 0, \qquad \alpha(r) \quad := \quad \frac{1}{r} \int_0^r s\varphi''(s)\, ds$$

- $(B(t))_{t \geqslant 0}$ is a M-valued Brownian motion.

Particular situations (1)

For any $b \in \mathbb{R}$, define the convex function $\varphi_b : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$\forall\ r \geqslant 0, \qquad \varphi_b(r) \ := \ \frac{r^b - 1 - b(r-1)}{b(b-1)}$$

with the conventions that for any $r \in \mathbb{R}_+$,

$$\begin{aligned} \varphi_0(r) &:= -\ln(r) + r - 1 \\ \varphi_1(r) &:= r\ln(r) - r + 1 \end{aligned}$$

We will also be interested in hybrid versions: for any $b_1, b_2 \in \mathbb{R}$,

$$\forall\ r \geqslant 0, \qquad \varphi_{b_1,b_2}(r) \ := \ \begin{cases} \varphi_{b_1}(r) & , \text{ if } r \in (0,1], \\ \varphi_{b_2}(r) & , \text{ if } r \in (1,+\infty). \end{cases}$$

Particular situations (2)

- With $\varphi = \varphi_1$, (9) corresponds to the evolution of the time-marginal distributions of a simulated annealing algorithm. Then μ_γ is the Gibbs density associated to the potential U and the inverse temperature γ.
- With $\varphi = \varphi_b$, $b > 1$, $M = \mathbb{R}$ and $U = 0$, (9) corresponds to the porous media evolution equation. If we rather take U to be quadratic, then μ_γ is a Barrenblatt distribution, which has a compact support.
- With $\varphi = \varphi_b$, $b \in [0, 1)$ (respectively $b < 0$), $M = \mathbb{R}$ and $U = 0$, (9) corresponds to the fast (respectively, ultra-fast) diffusion evolution equation.

A convergence result

In [4], we proved the concentration around $\mathcal{U}$ of ρ_t for large time $t \geqslant 0$, for $\varphi = \varphi_{b,2}$ with $b \in (0, 1/2)$ and appropriate polynomial scheme γ, on the circle. The basic ingredient is a new functional inequality.

The link with the previous fraudulent algorithm, is that heuristically, Y is expected behaves at large times like the diffusion X described by (1) with $\beta = b/(1-b)$.

Plan

References


R. Holley, S. Kusuoka, and D. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2), p. 333-347, 1989.


Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

Characterization and convergence.


F. Otto. *The geometry of dissipative evolution equations: the porous medium equation*, Communications in Partial Differential Equations, 26, p. 101-174, 2001.


Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the density case. ArXiv preprint, April 2022.


Laurent Miclo. On the convergence of global-optimization fraudulent stochastic algorithms. HAL preprint, April 2023.