On fraudulent stochastic algorithms

Laurent Miclo

Toulouse School of Economics
Institut de Mathématiques de Toulouse

Talk based on several collaborations with
Michel Benaïm, Jérôme Bolte and Stéphane Villeneuve

Plan of the talk

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

Plan

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

Global optimization

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

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

Denote

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

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.
– Appear as diffusive limits of mini-batch stochastic gradient descent algorithms in overparametrized machine learning, see Wojtowytsch [7].
– Approximation of the large-time limit behavior of the mean-field swarm algorithms.
– Suggest the design of non-fraudulent interacting particles systems, evaluating on-line the minimal value.

A fraudulent algorithms

Assume that U is a Morse function (in particular $\mathcal{U}$ is finite) and that $\min_M U = 0$.

Consider the (time-homogeneous) stochastic algorithm $X := (X(t))_{t \geqslant 0}$ whose evolution is driven by

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

where a priori $\beta \in \mathbb{R}$. When X is starting from $x \in M$, we will write X_x.

X is 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

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

The main result

Denote $\beta_0 := \frac{m}{2} - 1$.

Theorem 1

Assume that $\beta > \beta_0$. Whatever the initial condition $X(0)$, 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 $x \notin \mathcal{U}$, we have for any $y \in \mathcal{U}$,

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

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

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

A converse result

The value β_0 is critical for the above behavior:

Theorem 2

Assume that $\beta < \beta_0$ and $m \geqslant 2$. Whatever the initial distribution of $X(0)$ not charging $\mathcal{U}$, for large $t \geqslant 0$, $X(t)$ converges in law toward the invariant distribution π_β satisfying $\pi_\beta(\mathcal{U}) = 0$. In particular we get

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

For $\beta < \beta_0$ and $m = 1$, one gets similar results on the connected components of $M \backslash \mathcal{U}$.

More on the attractive minimizer case

For $y \in \mathcal{U}$, let $\lambda_{\min}(y)$ be the minimal eigenvalue of the Hessian of U at y, and

$$\lambda_{\min} \quad := \quad \min(\lambda_{\min}(y) \ : \ y \in \mathcal{U})$$

Denote δ the Riemannian distance.

Theorem 3

For $\beta > \beta_0$, we have a.s.,

$$\limsup_{t \to +\infty} \frac{\ln(\delta(X(t), X(\infty)))}{t} \leqslant -\lambda_{\min}(\beta - \beta_0)$$

More on the non-attractive minimizer case (1)

The invariant probability probability π_β is given by

$$\pi_\beta(dx) \quad \# \quad U(x)^{-1-\beta}\ell(dx)$$

where ℓ is the Riemannian probability on M.
The temporal ergodic theorem holds:

Theorem 4

For $\beta < \beta_0$, the diffusion X is positive recurrent on $M\backslash\mathcal{U}$ and for any $f \in \mathbb{L}^1(\pi_\beta)$, we have a.s.

$$\lim_{t\to+\infty} \frac{1}{t}\int_0^t f(X(s))\, ds \quad = \quad \pi_\beta[f]$$

(as soon as the initial law is not charging $\mathcal{U}$).

More on the non-attractive minimizer case (2)

The spatial ergodic theorem holds and can be quantified as:

Theorem 5

For $\beta < \beta_0$, there exist positive constants a, b, χ (depending on β) with $\chi < \beta_0 - \beta$, such that for any measurable $f : M \backslash \mathcal{U} \to \mathbb{R}$, we have for $x \in M \backslash \mathcal{U}$,

$$\forall\ t \geqslant 0, \qquad |\mathbb{E}[f(X_x(t))] - \pi_\beta[f]| \ \leqslant\ \frac{ae^{-bt}}{U(x)^\chi} \|f\|_\chi$$

with

$$\|f\|_\chi \ := \ \sup\{|f(x)|\, U(x)^\chi \,:\, x \in M \backslash \mathcal{U}\}$$

On the critical case

For $\beta = \beta_0$, we did not succeed in showing one of the two possible alternatives: X is null-recurrent or transient on $M \backslash \mathcal{U}$. But at least we have:

Proposition 6

For any neighborhood $\mathcal{O}$ of $\mathcal{U}$, we have a.s

$$\lim_{t \to +\infty} \frac{1}{t} \int_0^t \mathbb{1}_{X(s) \in \mathcal{O}} \, ds = 1$$

Plan

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

Preliminary remarks

A first proof of Theorems 1 and 2 under more restrictive conditions on β was obtained in [6] by comparing the stochastic process $(U^a(X(t)))_{t\geqslant 0}$, for appropriate exponents $a > 0$, with Bessel processes of negative and positive dimensions.

A finer analysis of the attractiveness or repulsiveness of the global minimizers is obtained by resorting to the homogenization techniques of [3] for the study of extinction and persistence.

The starting point: investigation near a global minima $y \in \mathcal{U}$ and blow up of the point y into a sphere that is supporting a fast diffusion.

Euclidean computations

Let us first study the situation where $M = \mathbb{R}^m$ and $y = 0$.
Let $\epsilon > 0$ be small enough so the only critical point for U in the ball $B(0, \epsilon)$ is 0. For any $x \in B(0, \epsilon)\backslash\{0\}$ consider the polar decomposition $x = \rho\theta$ with $\rho \in (0, \epsilon)$ and $\theta \in \mathbb{S}^{m-1}$, the sphere of dimension $m - 1$. This decomposition induces the mapping

$$P : \mathcal{C}^2(B(0, \epsilon)) \ni f \quad \mapsto \quad P[f] \in \mathcal{C}^2((0, \epsilon) \times \mathbb{S}^{m-1})$$

with

$$\forall\ (\rho, \theta) \in (0, \epsilon) \times \mathbb{S}^{m-1}, \qquad P[f](\rho, \theta) \quad := \quad f(\rho\theta)$$

Using traditional polar change of variables, we get the intertwining relation

$$\mathrm{L}_\beta \circ P \quad = \quad P \circ L_\beta$$

with

$$\begin{aligned} \mathsf{L}_\beta \cdot \quad &:= \quad \mathsf{U}\left(\partial_\rho^2 \cdot + \frac{m-1}{\rho}\partial_\rho \cdot + \frac{1}{\rho^2}\triangle_\theta \cdot\right) \\ &\qquad -\beta\left((\partial_\rho \mathsf{U})\partial_\rho \cdot + \frac{1}{\rho^2}\langle \nabla_\theta \mathsf{U}, \nabla_\theta \cdot\rangle_\theta\right) \end{aligned}$$

where $\mathsf{U} := P[U]$.
Our assumptions on U imply, uniformly over $\theta \in \mathbb{S}^{m-1}$,

$$\begin{aligned} \lim_{\rho\to 0_+} \frac{\mathsf{U}(\rho,\theta)}{\rho^2} &= \frac{1}{2}\langle \theta, A\theta\rangle \\ \lim_{\rho\to 0_+} \frac{\partial_\rho \mathsf{U}(\rho,\theta)}{\rho} &= \langle \theta, A\theta\rangle \\ \lim_{\rho\to 0_+} \frac{\nabla_\theta \mathsf{U}(\rho,\theta)}{\rho^2} &= A\theta - \langle \theta, A\theta\rangle\,\theta \end{aligned}$$

with $A := \mathrm{Hess}\, U(0)$.

A diffusion on the sphere

It follows that for any $F \in \mathcal{C}^2([0,\epsilon) \times \mathbb{S}^{m-1})$,

$$\lim_{\rho \to 0_+} \mathsf{L}_\beta[F](\rho,\theta) \quad = \quad G_\beta[F(0,\cdot)](\theta)$$

where G_β is the diffusion generator on $\mathbb{S}^{m-1}$ given by

$$G_\beta \cdot \quad := \quad \langle \theta, A\theta \rangle \left(\frac{1}{2} \triangle_\theta \cdot - \beta \langle b(\theta), \nabla_\theta \cdot \rangle_\theta \right)$$

with

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad b(\theta) \quad := \quad \frac{A\theta - \langle \theta, A\theta \rangle \theta}{\langle \theta, A\theta \rangle}$$

The invariant measure of G_β

We have

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad b(\theta) \quad = \quad \frac{1}{2}\nabla_\theta \ln(\langle \theta, A\theta\rangle)$$

so the invariant measure associated to G_β is given by

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad \mu_\beta(d\theta) \quad \# \quad \langle \theta, A\theta\rangle^{-1-\beta}\ \sigma(d\theta)$$

where σ is the uniform probability measure on $\mathbb{S}^{m-1}$.
This is the first ingredient needed in the (more general) approach of [3]: on the boundary $\{0\} \times \mathbb{S}^{m-1}$ of $[0, \epsilon) \times \mathbb{S}^{m-1}$, the generator L_β coincides with a generator G_β for which we are able to compute the invariant measure μ_β.

Criteria for attractiveness/repulsiveness of 0

The second ingredient is a function V on $(0, \epsilon) \times \mathbb{S}^{m-1}$ such that $\lim_{\rho \to 0_+} V(\rho, \theta) = +\infty$ uniformly in $\theta \in \mathbb{S}^{m-1}$, $\Gamma_{\mathsf{L}_\beta}[\mathsf{V}]$ is bounded on $(0, \epsilon) \times \mathbb{S}^{m-1}$ and $\mathsf{L}_\beta[\mathsf{V}]$ can be extended into a continuous function H_β on $[0, \epsilon) \times \mathbb{S}^{m-1}$. Recall that the carré du champ is given by

$$\Gamma_{\mathsf{L}_\beta}[\mathsf{V}] \quad := \quad \mathsf{L}_\beta[\mathsf{V}^2] - 2\mathsf{V}\mathsf{L}_\beta[\mathsf{V}]$$

Then we get the criteria for $m \geqslant 2$ and any $x \notin \mathcal{U}$,

$$\mu_\beta[H_\beta(0, \cdot)] > 0 \quad \Rightarrow \quad \mathbb{P}\left[\lim_{t \to +\infty} X_x(t) = 0\right] > 0$$

$$\mu_\beta[H_\beta(0, \cdot)] < 0 \quad \Rightarrow \quad \mathbb{P}\left[\lim_{t \to +\infty} X_x(t) = 0\right] = 0$$

A good function V

A function satisfying the previous assumptions is $\mathsf{V} := -\ln(\rho)$. Using the polar expression of L_β, we end up with

$$\mathsf{L}_\beta[\mathsf{V}] \quad = \quad (2-m)\frac{\mathsf{U}}{\rho^2} + \beta\frac{\partial_\rho \mathsf{U}}{\rho}$$

leading to

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad H_\beta(0,\theta) \quad = \quad \left(\beta - \frac{m}{2} + 1\right)\langle \theta, A\theta\rangle$$

It follows that the sign of $\mu_\beta[H_\beta(0,\cdot)]$ is that of $\beta - \beta_0$, explaining the pivotal role of β_0 for the attractiveness/repulsiveness of 0.

A.s. convergence for $\beta > \beta_0$

The previous computations can be extended to the Riemannian setting, in the neighborhood of each $y \in \mathcal{U}$.
End of the qualitative argument for $\beta > \beta_0$: each time X enters a sufficiently small ball $B(y, \epsilon_y)$, it has a positive probability to converge to y. We deduce the desired convergence, since L_β is elliptic on $M \setminus \cup_{y \in \mathcal{U}} B(y, \epsilon_y)$ and thus X always ends up exiting it.
The quantitative estimate comes from a more careful local analysis in the line of [1] and [3].

The results for $\beta < \beta_0$ rely on Lyapounov arguments in the spirit of Meyn and Tweedie [5], [1] and [3].

More general sets $\mathcal{U}$

The Morse assumption on U can be relaxed, in particular 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.

Plan

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

A non-linear p.d.e.

Consider the non-linear evolution equation

$$\frac{d}{dt}\rho_t \quad = \quad \mathrm{div}(\rho_t[\gamma_t \nabla U + \nabla \varphi'(\rho_t)]) \tag{2}$$

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

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

where

- ρ stands both for a probability measure from $\mathcal{P}(M)$ and its density with respect to ℓ,
- the term $\int_M U\, d\rho$ 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 (2), whose evolution is described by

$$dY(t) \quad = \quad -\gamma_t \nabla U(Y(t)) + \sqrt{2\alpha(\rho_t(Y(t)))}\, dB(t) \qquad (3)$$

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 $\varphi = \varphi_1$, (2) 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 $\varphi = \varphi_b$, $b > 1$, $M = \mathbb{R}^m$ and $U = 0$, (2) 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 $b < 0$), $M = \mathbb{R}^m$ and $U = 0$, (2) 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 to behave at large times like the diffusion X described by (1) with $\beta = b/(1-b)$.
Indeed, if ρ_t is replaced by μ_{γ_t} in (3), we recover the evolution (1) up to a time-change, due to

$$\lim_{\gamma \to +\infty} \frac{1}{\gamma} \alpha(\mu_\gamma(x)) \quad = \quad \frac{1-b}{b} U(x)$$

It suggests that we should take

$$b \quad > \quad 1 - \frac{2}{m}$$

Plan

1. Global optimization
2. Results
3. Sketch of proofs
4. Stochastic swarm algorithms
5. References

References

- Michel Benaïm. Stochastic persistence. Unpublished manuscript, May 2019.
- Michel Benaïm and Laurent Miclo. The asymptotic behavior of fraudulent algorithms. arXiv preprint, January 2024.
- Michel Benaïm and Edouard Strickler. Random switching between vector fields having a common zero. Ann. Appl. Probab., 29(1):326–375, 2019.
- Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the mean-field limit case. *Math. Program.*, 205(1-2 (A)):661–701, 2024.
- Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009.
- Laurent Miclo. On the convergence of global-optimization fraudulent stochastic algorithms. HAL preprint, April 2023.
- Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part ii: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):16, 2023.